UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE BIOLOGIA

**ROBERTO HIROCHI HERAI**



# METODOLOGIAS DE BIOINFORMÁTICA PARA DETECÇÃO E ESTUDO DE SEQUÊNCIAS REPETITIVAS EM *LOCI* GÊNICOS DE TRANSCRITOS QUIMÉRICOS



Tese apresentada ao Instituto de Biologia para obtenção do Título de Doutor em Genética e Biologia Molecular na Área de Bioinformática.

**Orientador: Prof. Dr. Michel Eduardo Beleza Yamagishi**

**Campinas, 2010**

**FICHA CATALOGRÁFICA ELABORADA PELA**
**BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP**

**H412m**

Herai, Roberto Hirochi
Metodologias de bioinformática para detecção e estudo de sequências repetitivas em loci gênicos de transcritos quiméricos / Roberto Hirochi Herai. – Campinas, SP: [s.n.], 2010.

Orientador: Michel Eduardo Beleza Yamagishi.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Bioinformática - Metodologia. 2. Transcritos quiméricos. 3. Sequências repetitivas do ácido nucléico. I. Yamagishi, Michel Eduardo Beleza. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

(rcdt/ib)

**Título em inglês:** Bioinformatics methodologies for detection and study of repetitive sequences in gene loci of chimeric transcripts.
**Palavras-chave em inglês**: Bioinformatics - Methodology; Chimeric transcripts; Nucleic acid repetitive sequences.
**Área de concentração:** Bioinformática.
**Titulação:** Doutor em Genética e Biologia Molecular.
**Banca examinadora**: Michel Eduardo Beleza Yamagishi, Gonçalo Amarante Guimarães Pereira, José Andrés Yunes, Guilherme Correa de Oliveira, Roberto Hiroshi Higa.
**Data da defesa**: 23/02/2010.
**Programa de Pós-Graduação:** Genética e Biologia Molecular.

Campinas, 23 de fevereiro de 2010.

**Banca Examinadora**

Prof. Dr. Michel Eduardo Beleza Yamagishi (Orientador) _____
Assinatura

Prof. Dr. José Andrés Yunes _____
Assinatura

Prof. Dr. Guilherme Correa de Oliveira _____
Assinatura

Prof. Dr. Gonçalo Amarante Guimarães Pereira _____
Assinatura

Profa. Dra. Poliana Fernanda Giachetto _____
Assinatura

Profa. Dra. Paula Regina Kuser Falcão _____
Assinatura

Dr. Roberto Hiroshi Higa _____
Assinatura

Prof. Dr. Michel Georges Albert Vincentz _____
Assinatura

O presente trabalho foi desenvolvido no Laboratório de Bioinformática Aplicada, da unidade Embrapa Informática Agropecuária da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), Campinas, São Paulo, com apoio financeiro do CNPQ (processo #140730/2008-7) e da FAPESP (processo #2008/02647-3).

*"Acreditar é monótono, duvidar é apaixonante,*

*manter-se alerta: eis a vida.*"

(Oscar Wilde)

*"Investir em conhecimentos rende sempre melhores juros."*

(Benjamin Franklin)

*Mãe (in memoriam), pela graça de me dar a vida e, mais ainda, pelo amor que sempre teve pela família enquanto esteve ao nosso lado.*

*Pai, por ter lutado por nós.*

*Meus irmãos Paulo e Marisa, seu esforço e luta me deram tranquilidade para a realização deste sonho.*

*Minha pequena e querida família, dedico esta tese a vocês, sem os quais jamais teria a chance de aprender e vencer os desafios deste trabalho.*

# Agradecimentos

Ao também amigo, meu orientador, professor Dr. Michel E. B. Yamagishi, um pesquisador com idéias vislumbrantes que permitem contagiar qualquer pessoa no mundo da ciência. Ter sido seu primeiro orientando de pós-graduação foi um grande privilégio, pois o trabalho, apesar de alguns caminhos tortuosos, seguiu sempre na direção correta e com grande objetividade.

Ao amigo e professor Dr. Marco Aurélio A. Henriques, pelo perfeccionismo e objetividade, os quais me permitiram seguir com tranquilidade na área de pesquisa.

Ao amigo e professor Dr. João U. Furquim de Souza, pelo grande exemplo de vida e pela oportunidade de me apresentar ao mundo da pesquisa de forma ética e correta.

Aos colegas da EMBRAPA, em especial aos do Laboratório de Bioinformática Aplicada (LBA), os quais me deram amplo suporte para o desenvolvimento das atividades deste trabalho.

À Dra. Paula Regina Kuser Falcão, chefe do LBA, que tanto me incentiventou e me fez crescer intelectualmente, além de sua contribuição efetiva com o desenvolvimento deste trabalho.

À Dra. Poliana Fernanda Giachetto, pelas breves discussões associadas aos conceitos da biologia e genética.

Às famílias dos amigos Rafael e Helmann, que tornaram Campinas minha segunda casa.

A todos os meus amigos, que sempre foram fundamentais em minha vida.

À minha doce e querida Solange, seu carinho e companheirismo nos momentos finais deste trabalho foram muito importantes em minha vida.

Aos meus irmãos, irmãs, pai, sobrinhos e cunhados, que residem no Brasil e no Japão.

À minha mãe e irmãos Renato e Lúcia, *in memorian*, vocês me ajudaram na realização deste grande sonho.

À empresa EMBRAPA Informática Agropecuária, pela infra-estrutura fornecida.

À Universidade Estadual de Campinas (Unicamp), que acolheu-me para desenvolver este trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo importante suporte financeiro que tornou possível a elaboração desta pesquisa.

# Resumo

A grande quantidade de dados biológicos gerados recentemente permitiu verificar que os genomas são repletos de sequências repetitivas (SR), como microsatélites e elementos genéticos móveis, altamente improváveis de ocorrer estatisticamente se os genomas fossem gerados a partir de uma distribuição aleatória de nucleotídeos. Tal comprovação motivou a classificação de tais sequências e também a construção de diversas ferramentas de bioinformática, além de mecanismos de armazenamento baseados em sistemas de gerenciamento de bancos de dados (SGBD) para permitir localizá-las e armazená-las para posterior estudo. Entretanto, foi com a comprovação biológica da importância das SR, como no mecanismo de interferência por RNAi (SR reversa complementar), que as SR despertaram maior interesse por parte da comunidade científica. Atualmente, já há fortes evidências que associam as SR com fenômenos biológicos bastante interessantes, como o processamento de RNA por *cis-splicing* e a formação de transcritos quiméricos, frequentes em organismos inferiores e muito raro em organismos superiores. Tais tipos de transcritos podem ser gerados a partir de *trans-splicing* ou, como conjecturamos nesse trabalho, pela transposição de elementos genéticos móveis (como por exemplo *transposons* ou *retrotransposons*). Em virtude disso, este projeto propõe a construção de metodologias de Bioinformática, disponibilizadas na WEB, para detectar transcritos quiméricos em genomas de organismos, tanto em versões *draft* ou em alta qualidade, e também estudar as SR que ocorrem no *locus* gênico dos transcritos envolvidos na formação de uma sequência quimérica. As ferramentas propostas permitiram identificar, a partir de bibliotecas de transcritos de *full-length* cDNA, tanto de humanos quanto de bovinos, novos transcritos quiméricos provenientes de células de tecidos normais, e que não seguem *splice-sites* canônicos na região de fusão dos transcritos envolvidos. Além disso, as sequências encontradas apresentam uma elevada taxa de concentração de pares de SR do tipo reverso complementar no *locus* gênico dos dois transcritos que formam a sequência quimérica. As ferramentas propostas podem ser utilizadas para outros organismos e direcionar trabalhos experimentais para tentar comprovar em bancada novos transcritos quiméricos, tanto em organismos inferiores quanto em superiores.

Palavras-chave: metodologias de bioinformática, transcritos quiméricos, sequências repetitivas, *full-length* cDNA.

# Abstract

The recent availability of a huge amount of biological data allowed to know about the high concentration of repetitive sequences (SR) like microsatellites and genetic mobile elements in different genomes. Repetitive sequences are improbable to occur statistically if genome data were generated by a random distribution of nucleotides. Such observation motivated the classification of repetitive sequences, and the construction of several bioinformatics tools. Furthermore, several mechanisms to store repetitive sequences, which are based on data base management systems (DBMS) were proposed and created. They can be used to search for specific sequences to make a posteriori study. However, it was with the biological confirmation of the importance of repetitive sequences, like by the RNA interference (reverse complement, or inverted repeat) mechanism, that the scientific community gained more interest by such sequences. Actually, there is strong evidence that associates the repetitive sequences with some interesting biological phenomena, like in RNA processing by cis-splicing, and in chimeric transcript formation mechanism. This last one is very frequently in inferior organism, but rare in superior organisms. Such types of transcripts can be generated by trans-splicing, or like conjectured in this work, by the retrotransposition of mobile genetic elements (like transposons or retrotransposons). In this way, this work proposed the construction of several Bioinformatics methodologies, available in the WEB, to detect new evidences of chimeric transcripts in genomes of different organisms, both in draft genome and in high quality genome assemblage. We also studied repetitive sequences in gene *loci* of the involved transcripts in a chimeric sequence formation. The proposed tools allowed us to identify, using a full-length cDNA databank, new chimeric transcript candidates in human and in bovine genome. They are from cells of normal tissues, and do not follow canonical splice-sites in the fusion region of the involved transcripts. Moreover, it was possible to show that the detected sequences have high concentration pairs of reverse complement type of repetitive sequences in gene *loci* of the two involved transcripts, which originated a new chimeric transcript candidate. The created bioinformatics tools can be used in other organisms in addition to the one used in this work, leading to the proposition of new experimental work to try to prove in vivo new chimeric transcripts, both in superior organism and in inferior organism.

Key-words: bioinformatics methodology, chimeric transcripts, repetitive sequences, full-length cDNA.

# Sumário

# Lista de figuras

# Lista de tabelas

# Lista de siglas

| Sigla | Significado |
|---|---|
| A | Adenine |
| AC | Accession Number |
| ADAR | Adenosina Deaminase |
| ALU | Arthrobacter Luteus |
| BD | Banco de Dados |
| BGD | Bovine Genone Database |
| BLAST | Basic Local Alignment Sequence Tool |
| BLAT | Basic Local Alignment Tool |
| C | Citosina |
| cDNA | Complementar DNA |
| CDS | Coding Sequence |
| CIB | Center for Information Biology |
| CPU | Central processing unit |
| DDBJ | DNA Database of Japan |
| DDL | Data Definition Language |
| DNA | Deoxyribonucleic Acid |
| EBI | European Bioinformatics Institute |
| EMBL | European Molecular Biology Laboratory |
| EMBRAPA | Empresa Brasileira de Pesquisa Agropecuária |
| EST | Expressed Sequence Tag |
| ET | Evidência de Transcrição |

| | |
|---|---|
| FCM | Faculdade de Ciências Médicas |
| Fl-cDNA | *Full-length* cDNA |
| G | Guanina |
| Gb | Giga base |
| GR | Genoma Real |
| HERV-K | Human Endogenous Retrovirus type k |
| H-InvDB | Human Invitational Database |
| HIV | Human Immunodeficiency Vírus |
| IAC | International Advisory Committee |
| ICM | International Collaborative Meeting |
| IDE | Integrated Development Environment |
| JDBC | Java Data Base Connectivity |
| Kpb | Kilo bases |
| LLA | Linfoma Linfóide Aguda |
| LINE | Long Interspersed Nuclear Elements |
| LTR | Long Terminal Repeat |
| MER | Modelo Entidade Relacionamento |
| MER | Modelo Entidade Relacionamento |
| MGC | Mammalian Gene Collection |
| miRNA | Micro RNA |
| MVC | Model View Controller |
| NCBI | National Center of Biotechnology |
| ncRNA | Non Coding RNA |

| | |
|---|---|
| NGS | Next Generation Sequencing |
| NIG | National Institute for Genetics |
| NIH | National Institutes of Health |
| NIST | National Institute of Standards and Technology |
| NLM | National Library of Medicine |
| pb | Pares de base |
| PCR | Polimerase Chain Reaction |
| PMRD | Plant miRNA Database |
| RC | Repetição Complementar |
| RD | Repetição Direta |
| RefSeq | Reference sequence |
| RNA | Ribonucleic Ccid |
| RNAi | RNA de Interferência |
| RR | Repetição Reversa |
| RRC | Repetição Reversa Complementar |
| rRNA | Ribossomal RNA |
| SAGE | Serial Analysis of Gene Expression |
| SDL | Store Definition Language |
| SGBD | Sistema de Gerenciamento de Banco de Dados |
| SINE | Small Interspersed Nuclear Elements |
| SL | Spliced-lider |
| SMART | Spliceosome Mediated RNA *Trans-splicing* |
| SNP | Single Nucleotide Polimorphism |

| | |
|---|---|
| SQL | Single Query Language |
| SR | Sequência Repetitiva |
| sRNA | Small RNA |
| T | Timina |
| TQ | Transcrito Quimérico |
| TR | Transcrito Real |
| tRNA | Transporter RNA |
| TSR | Trans-spliced Region |
| UCSC | University of California Santa Cruz |
| UNICAMP | Universidade Estadual de Campinas |
| UTR | Untranslated Region |
| XML | Extensible Markup Language |

# Artigos derivados deste trabalho

HERAI, RH; YAMAGISHI, MEB. **RepGraph: Web-based tool for Identification and Visualization of Repetitive Sequences.** Bioinformatics, Oxford Journals, 2010 (em correção);

HERAI, RH; YAMAGISHI, MEB. **Are there "*forbidden*" biological sequences?** Journal of Computational Biology and Chemistry, Elsevier, 2010 (em correção);

HERAI, RH; YAMAGISHI, MEB. **A Bioinformatics Approach To Detect Interchromosomal Trans-Splicing In Bovine Full Length cDNA Databanks**. Plant and Animal Genome XVIII Conference (XVIII PAG), San Diego, USA, 2010;

HERAI, RH; YAMAGISHI, MEB. **Detection of Human Interchromosomal Trans-Splicing in Sequence Databanks**. Briefings in Bioinformatics, Oxford Journals, 2009;

HERAI, RH; YAMAGISHI, MEB. **Identificação de transcritos quiméricos em bibliotecas de *FlcDNA* de organismos com genomas "*draft*": estudo de caso em bovinos**. I Workshop em Genômica Animal – Rede Genômica Animal da Embrapa, Fortaleza, Ceará, Brasil, 2009;

HERAI, RH; YAMAGISHI, MEB. **Metodologia de bioinformática para detecção de erros de montagem em regiões gênicas usando bibliotecas de *Fl-cDNA*: estudo de caso em bovinos**. I Workshop em Genômica Animal – Rede Genômica Animal da Embrapa, Fortaleza, Ceará, Brasil, 2009;

HERAI, RH; YAMAGISHI, MEB. **A web based tool for identification and visualization of repetitive sequences within gene *loci***. 5th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-Meeting 2009), Angra dos Reis, RJ, Brasil, 2009;

HERAI, RH; SANTOS, EH; YAMAGISHI, MEB. **Algoritmos de bioinformática para detecção de *nullomers* no transcriptoma humano**. II Encontro Acadêmico de Modelagem Computacional (IIEAMC) do Laboratório Nacional de Computação Científica (LNCC), Petrópolis, RJ, Brasil, 2009;

HERAI, RH; YAMAGISHI, MEB. **Uma ferramenta WEB para identificação e visualização de sequências repetitivas em *lócus* gênico de animais e plantas**. V Mostra de Trabalhos de Estagiários e Bolsistas da EMBRAPA, Embrapa Informática Agropecuária, Campinas, SP, Brasil, 2009;

HERAI, RH; YAMAGISHI, MEB. **Proposta de uma metodologia In silico na detecção de genes transcritos por *trans-splicing* inter-cromossomal em genomas de animais e plantas**. V Mostra de Trabalhos de Estagiários e Bolsistas da EMBRAPA, Embrapa Informática Agropecuária, Campinas, SP, Brasil, 2008;

HERAI, RH; YAMAGISHI, MEB. **Detecting evidences of inter-chromosomal trans-spliced genes using a bioinformatic approach**. 4th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-Meeting 2008), Salvador, Ba, Brasil, 2008;

# Capítulo 1 Introdução

A era pós-genômica permitiu sequenciar dezenas de organismos com o intuito de se estudar e tentar obter informações a respeito das estruturas genéticas que compõem os seres vivos, tanto superiores quanto inferiores. Para tal, grande esforço tem sido empregado para se tentar identificar novos padrões biológicos, e que possam ser utilizados para garantir a continuidade dos organismos, bem como para gerar produtos com características genéticas específicas e, consequentemente, de valor agregado cada vez maior. Uma das áreas que tem despertado grande interesse é o estudo de sequências repetitivas (SR), tanto de DNA quanto de RNA, e sua relação com fenômenos biológicos, os quais podem estar relacionados com funções metabólicas ou regulatórias dentre os mais interessantes nos organismos. Este capítulo apresenta uma breve contextualização do trabalho, associado intimamente com o estudo das SR, bem como sua importância. Além disso, são apresentados os objetivos do trabalho, suas contribuições e a forma com que está organizada.

## 1.1 Código genético dos seres vivos

Os seres vivos sejam superiores ou inferiores, procariotos ou eucariotos, são formados por componentes celulares que possuem, internamente, uma estrutura chamada DNA. Ela é responsável por armazenar as informações genéticas de um organismo, de forma que as mesmas sejam transmitidas para seus descendentes ou utilizadas pelo próprio ser vivo para sua constituição (Watson & Berry, 2003).

O DNA foi isolado pela primeira vez no ano de 1869 pelo bioquímico alemão Johann Friedrich Miescher (1844–1895) e é a estrutura molecular que carrega as informações genéticas dos seres vivos (Avery et al., 1944). A partir de tal descoberta, surgiram muitas pesquisas relacionadas ao DNA, entre elas a de Watson e Crick, que conseguiram identificar a estrutura tridimensional em hélice da molécula de DNA, conforme Figura 1 (Watson & Crick, 1953).

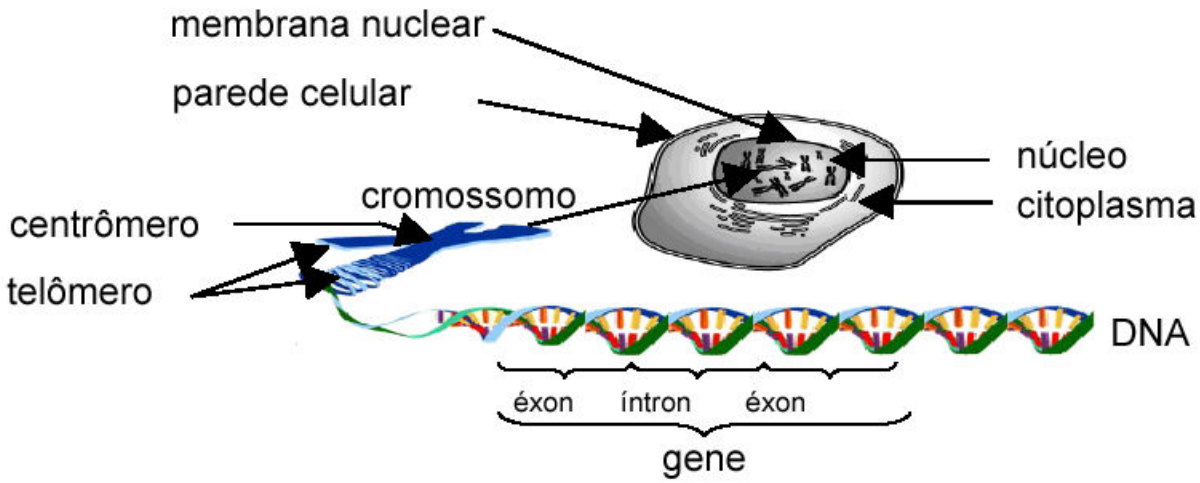


*Figura 1 - Estrutura de uma célula de um organismo eucarioto.*
*Fonte: da própria pesquisa, baseado em Loodish et al. (1999).*

Os eucariotos, conforme Figura 1, possuem uma estrutura celular composta basicamente por um núcleo, uma membrana nuclear que reveste o núcleo, o citoplasma que é composto por organelas responsáveis pelo metabolismo celular e que são importantes no processo de diferenciação celular, e uma parede celular (Kohler & Hurt, 2007). O DNA encontra-se interno ao núcleo em uma estrutura chamada cromossomo que é subdivido em duas partes principais: a primeira é um estrangulamento que separa as cromátides e é chamado de centrômero, e a segunda os telômeros, que formam as extremidades de um cromossomo (Watson & Berry, 2003).

É na molécula de DNA que se encontram os genes, que estão, em geral, relacionados a alguma característica de um organismo. Muitas doenças possuem relação

direta com genes ou com sua regulação. Além disso, o estudo de genes tem se tornando muito importante devido a capacidade de modificações genéticas em organismos superiores ou inferiores de interesse comercial. Já há no mercado diversos exemplos de organismos que foram modificados geneticamente. Em plantas, alguns tipos de modificações consistem em fornecer a elas resistência contra a ação de pragas ou a condições climáticas e geográficas antes consideradas desfavoráveis. Na Figura 2 é apresentado um exemplo de um fruto que foi gerado a partir da planta geneticamente modificada do mamão, o qual passou a adquirir resistência a um vírus da espécie *Papaya ringspot*, responsável por gerar manchas e enfraquecer a planta e, consequentemente, a qualidade e durabilidade do fruto (Gonsalves et al., 2004).

*Figura 2 – Modificação genética da fruta do mamão.*
*Fruta da espécie Carica papaya, conhecido popularmente como mamão: a) fruto infectado com o vírus Papaya ringspot; b) fruto geneticamente modificado, resistente ao vírus.*
*Fonte: Gonsalves et al., 2004.*

Em animais, há também casos comprovados de modificações genéticas com o intuito de melhorar a qualidade de um produto, de forma que o mesmo apresente melhores condições econômicas para sua comercialização. Na Figura 3 é apresentado um exemplo do peixe GloFish, o qual representa uma versão geneticamente modificada do peixe zebra (*Danio rerio*), popularmente conhecido por paulistinha.

*Figura 3 – Modificação genética do peixe paulistinha.*
*Peixe da espécie Danio rerio, conhecido popularmente como peixe zebra.*
*a) peixe original; b) peixe geneticamente modificado, com diferentes fluorescências.*
*Fonte: http://www.glofish.com/images/*

Em ambos os casos, para que a modificação genética fosse possível, foi feito um estudo do gene associado à resistência contra vírus (mamão) ou à característica de interesse (cor do peixe), de forma a permitir que uma nova característica fosse expressa em cada organismo considerado.

## 1.2 Sequenciamento e montagem de genomas

Dada a grande importância na identificação de genes, e da sua localização dentro de um genoma, realizaram-se o sequenciamento do genoma de vários organismos, principalmente os de interesse comercial. Um dos grandes fatores que tem motivado e facilitado isso é o rápido avanço das tecnologias de sequenciamento e das ferramentas de bioinformática utilizadas para montagem e análise de dados genômicos.

As novas tecnologias de sequenciamento existentes, conhecidas popularmente por NGS (*Next Generation Sequencing*) (Schuster, 2008), permitem sequenciar em questão de algumas horas, cerca de até 1 giga de pares de base (Gb), dependendo da tecnologia empregada. Estas novas tecnologias baseiam-se na geração de pequenas sequências (*reads*) de nucleotídeos, os quais são posteriormente analisados por programas de bioinformática de

tal forma que um genoma ou transcriptoma possa ser montado, ou para que novos SNPs sejam descobertos, os quais são, atualmente, algumas das aplicações mais frequentes. As três principais plataformas de sequenciamento NGS do mercado são descritas a seguir (Hurd & Nelson, 2009):

- 454 Roche: tal plataforma, conhecida como GS FLX Titanium, é baseada em sequenciamento por síntese. Ela gera sequências de aproximadamente 500 pares de base (pb), e em apenas uma análise ("corrida") permite gerar uma ordem de até 500Mb de sequências. Dentre as tecnologias NGS, a 454 é a que gera as maiores sequências, o que permite aplicar um método de montagem conhecido como "*de novo*" (Li et al., 2010), no qual é necessário que haja, pelo menos, um conjunto de *mate-pairs* suficientes para tal ao invés de um genoma de referência para que seja gerada uma montagem do genoma considerado. Um dos problemas da tecnologia é, assim como nas demais, sequênciar regiões de um genoma que são repletas de SR;

- Solexa/Illumina Genome Analyzer: tal plataforma é também baseada em sequenciamento por síntese. Ela gera sequências de tamanho entre 30 a 75 pares de base, e em apenas uma análise permite gerar entre 3 a 7.5 Gb de sequências. Devido ao tamanho reduzido das sequências, esta tecnologia de sequenciamento torna-se mais apropriada para genomas já anotados. O uso de métodos de montagem, como "*de novo*", apesar de possível, pode gerar uma montagem com alto grau de redundância e erros. Com relação às sequências repetitivas, seu mecanismo químico de sequenciamento permite sequenciar com melhor qualidade sequências repetitivas de um genoma.

- SOLiD (Applied Biosysytems): tal plataforma é baseada em sequenciamento por ligação. A tecnologia propõe-se a fornecer até 99.94% de precisão nas bases seqüenciadas. Uma simples corrida no sequenciador pode gerar até 2 Gb de DNA, cada uma com cerca de 35 pb, o que equivale a aproximadamente 15 Gb de sequências.

Outra classe de tecnologia de sequenciamento, conhecida como 3ª Geração, já possui previsão de lançamento para este ano de 2010. A tecnologia é baseada em nanoporos construídos com nitrato de silício, que por meio de um campo magnético gerado em seu interior permite atrair as sequências de DNA para que produzam uma assinatura eletrônica distinta de cada tipo de base nitrogenada presente no genoma (Wanunu et al., 2009).

Dependendo do tamanho de um genoma, a quantidade de *reads* gerados pode facilmente chegar a alguns milhões. Com o uso de ferramentas de bioinformática, os *reads* são analisados para que seja gerada uma versão rascunho ("*draft*") do genoma montado. Um exemplo fictício da forma com que isso é feito é apresentado na Figura 4.

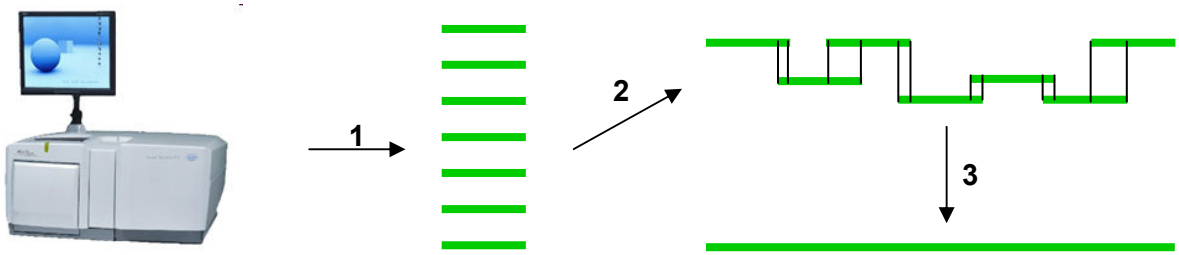


*Figura 4 – Sequenciamento e montagem de sequências de dados biológicas.*
*Montagem de sequências (reads) geradas por sequenciadores: 1) reads produzidos pelo equipamento sequenciador; 2) sobreposição, com o uso de uma ferramenta de bioinformática, dos reads, para identificação de regiões de consenso entre as sequências; 3) geração de uma sequência consenso, equivalente ao genoma, ou de parte dele, montado.*
*Fonte: da própria pesquisa.*

Apesar da alta tecnologia empregada para o sequenciamento do genoma de distintos organismos, ainda é fato que, mesmo com o uso das chamadas NGS, a qualidade das

montagens a partir de regiões ricas em sequências repetitivas não é tão precisa, e possivelmente apresenta erros. Desta forma, há algumas propostas baseadas em bioinformática que foram criadas com o intuito de tentar amenizar os erros de sequenciamento e de montagem ocasionados pela presença de sequências repetitivas em um dado genoma. Há também outros tipos de abordagens, que são discutidas neste trabalho, que foram construídas para tentar detectar e corrigir erros de montagem em genomas que ainda não foram concluídos, montagens estas denotadas como "*draft*".

## 1.3 Abundância de sequências repetitivas em genomas

A recente conclusão de diversos projetos genoma deu início à fase de análise das sequências genômicas. Tais estudos tem se concentrado não somente nas regiões genômicas que codificam proteínas, mas também em todos os demais trechos do genoma que, até pouco tempo atrás, eram pejorativamente denominados "*garbage DNA*" ou "*junk DNA*" (Gunter & Thomas, 2004). Com o avanço das pesquisas, outras regiões genômicas começaram a despertar interesse na comunidade científica. Em função disso, novos trabalhos surgiram com o intuito de identificar novas características, sendo uma delas associada à abundância de sequências repetitivas (SR) nos genomas de animais e plantas, e também em microorganismos. Em plantas, aproximadamente, 50% do genoma é composto por sequências repetitivas (Bannert & Reinhard, 2004), e em algumas espécies, como em trigo (*Triticum* spp.), podem alcançar até 90% (Haig & Kazazian, 2004; SanMiguel et al., 1996).

A importância das SR nos genomas ganhou grande importância por haver evidências de estarem envolvidas em fenômenos de regulação gênica, como o mecanismo de RNAi proposto por Fire et al. (1998), o qual rendeu-lhe o prêmio Nobel de Medicina e Fisiologia no ano de 2006. Há também evidências que associam o fenômeno de *cis-splicing* com a presença de sequências repetitivas (Gal-Mark et al., 2008), e mais recentemente, a formação de transcritos quiméricos (Gingeras, 2009), considerado um fenômeno muito raro em organismos superiores, também foi associado com a presença de sequências repetitivas no *locus* gênico das sequências envolvidas. Entretanto, por serem raros, faltam casos de transcritos quiméricos anotados, sejam experimentais ou baseados em bioinformática, e que permitam estudar o papel de SR em sua formação. Vale ressaltar que até pouco tempo atrás, a ocorrência de tais tipos de transcritos estava associada apenas ao câncer.

Apesar da conjectura de que a presença de SR possa de fato mediar a ocorrência de alguns fenômenos biológicos, não há ainda trabalhos na literatura que estudem e relacionem a frequência de tais SR com a ocorrência de transcritos quiméricos. Para a formação de transcritos quiméricos, por exemplo, é ignorado se a ocorrência de pares de SR do tipo reverso complementar, comumente chamadas de *inverted repeats*, é abundante ou não nos *loci* gênicos de transcritos que estão envolvidos na formação de um transcrito quimérico, e se essa frequência é diferente de regiões genômicas quaisquer.

Apesar da comprovada importância das SR na ocorrência de fenômenos biológicos dentre os mais interessantes, como RNAi, faltam sistemas de bancos de dados que permitam armazená-las e anotá-las de forma extensiva, independente do tipo, para facilitar seu estudo de forma global. A maioria das abordagens utiliza formas primitivas de armazenamento, como o uso de arquivos do tipo texto (consumo de recursos, propenso a

erros, dificuldade na análise). Além disso, os bancos de dados, quando existem, são construídos com fins específicos que limitam seu acesso; estão disponíveis apenas para consulta (maioria), e com o uso de diferentes tecnologias. Alguns exemplos de bancos de dados para armazenamento de SR e que foram construídos para fins específicos são: AluGene (Dagan et al., 2004) para sequências ALU; mirBase (Griffiths-Jones et al., 2006) para sequências de micro RNA; TREP (Wickera et al., 2002) para elementos repetitivos da planta Triticea; RepPop (Zhou & Xu, 2009) para o organismo *Populus trichocarpa*; RepBase (Jurka et al., 2005) para organismos eucariotos, entre outros. Todos eles serão discutidos na seção de revisão da literatura. Uma alternativa de manipular todos estes dados que já se encontram armazenados seria, dentre várias tecnologias, o uso de WebServices. Do ponto de vista estritamente computacional, uma SR possui 4 formas distintas: (*i*) repetição direta (RD), (*ii*) repetição reversa (RR), (*iii*) repetição complementar (RC) e (*iv*) repetição reversa complementar (RRC). As metodologias existentes para a detecção e armazenamento das SR levam em consideração apenas os casos (*i*) e (*iv*). Ferramentas como BLAST e os BLAST-like limitam-se a somente estes dois tipos de SR, sendo necessária a construção de uma metodologia que permita localizar também os tipos de sequências dos casos (*ii*) e (*iii*) em um dado organismo.

## 1.4 Objetivos

Com base na ampla importância das SR, desde sua relação com as NGS no sequenciamento e montagem de genomas, até sua associação com fenômenos biológicos dentre os mais importantes, foram propostos os seguintes trabalhos nesta tese:

1. analisar possível relação entre as sequências repetitivas (SR) e o interessante fenômeno de formação de transcritos quiméricos, como por exemplo, o processamento de RNA por *trans-splicing* intercromossomal:
    a. criar metodologia para detecção de transcritos quiméricos candidatos com base nas evidências experimentais até então reportadas na literatura;
    b. identificar SR que podem estar associadas com a mediação da ocorrência de *trans-splicing*;
2. construir um portal WEB com ferramentas de bioinformática para:
    a. analisar e detectar SR entre diferentes tipos de pares de transcritos, como aqueles que podem estar relacionados com um transcrito quimérico, considerando suas respectivas regiões genômicas;
    b. ilustrar graficamente a relação entre dois transcritos envolvidos na formação de um quimérico por meio de SR dos 4 tipos mencionados na seção anterior: repetição direta (RD), repetição reversa (RR), repetição complementar (RC) e repetição reversa complementar (RRC);
3. estudar a presença de SR em pares de transcritos e que formam um transcrito quimérico. Além disso, estudar a frequência de cada um dos tipos de SR existentes entre diferentes tipos de pares de sequências;
4. criar um novo banco de dados (BD) biológico, considerando:
    a. definir as fases de análise, projeto e implementação do BD, com o uso de uma metodologia Relacional;
    b. anotar SR dos 4 tipos mencionados;

c. construir uma ferramenta de migração de bases de dados que estejam disponíveis na WEB, em formato de dados primitivos para o banco de dados proposto;

5. criar uma metodologia para detecção de erros de montagem em regiões gênicas de genomas do tipo "*draft*" para permitir que transcritos quiméricos candidatos também sejam encontrados em genomas ainda não finalizados.

## 1.5 Contribuições

As contribuições deste trabalho foram:

- uma nova metodologia para busca e detecção de transcritos quiméricos em transcriptomas, considerando o uso de genomas, tanto "*draft"* quanto de alta qualidade, para posterior confirmação experimental;
- disponibilização de uma ferramenta WEB para permitir identificar pares de sequências repetitivas entre pares de transcritos. Isso permite analisar a possibilidade de algum tipo de correlação entre os tipos de sequências repetitivas encontradas com a formação de um transcrito quimérico;
- uma metodologia que permite estudar a frequência de ocorrência de pares de sequências repetitivas no *locus* gênico entre diferentes tipos de transcritos;
- um banco de dados biológico, desenvolvido a partir de um modelo de BD Relacional, para anotação de 4 tipos distintos de SR: repetição direta (RD), repetição reversa (RR), repetição complementar (RC) e repetição reversa complementar (RRC). O banco também permite anotar transcritos quiméricos. Tais características do banco de dados,

ausentes nas demais propostas da literatura, permitem estudar as SR para obtenção de novos padrões biológicos associados às SR;

- uma nova metodologia para detecção de erros de montagem em regiões gênicas de genomas "*draft*", a partir do uso de bibliotecas de transcritos, como de *full-length* cDNA, ou até mesmo EST.

## 1.6 Organização do trabalho

Para apresentar a forma com que as metodologias e ferramentas de bioinformática propostas neste trabalho foram desenvolvidas, e também discutir os resultados obtidos, este trabalho foi subdividido nos seguintes capítulos:

- Capítulo 1: que se apresenta, fornece uma breve contextualização do trabalho, da sua proposta, objetivos específicos, contribuições e um breve roteiro da organização desta tese;
- Capítulo 2: apresenta uma breve revisão da literatura com base na ocorrência de alguns fenômenos biológicos que foram comprovados pela presença de SR, como a interferência pelo mecanismo de RNAi, ou que já apresentam fortes evidências da relação com as SR, como o processamento de transcritos de pré-mRNA por *cis-splicing*, ou pela formação de transcritos quiméricos;
- Capítulo 3: assim como no Capítulo 2, apresenta uma breve revisão da literatura, mas desta vez tratando de conceitos teóricos, os quais se concentram na descrição de algumas tecnologias empregadas para o seqüenciamento de genomas e transcriptomas,

na classificação e nos tipos de SR, e aos mecanismos e tipos atuais de sistemas de bancos de dados construídos com a finalidade de armazenar e anotar as SR;

- Capítulo 4: apresenta as metodologias e ferramentas de bioinformática que foram desenvolvidas neste trabalho. Para cada uma das metodologias descritas, é feita uma breve citação da implementação e da forma de funcionamento de cada uma delas;
- Capítulo 5: apresenta os resultados obtidos computacionalmente com o uso das metodologias de bioinformática propostas pelo trabalho. A partir dos resultados, foi feita uma breve discussão da forma com que foram obtidos, e também uma série de análises para verificar a qualidade dos mesmos, para comprovar a utilidade das ferramentas desenvolvidas;
- Capítulo 6: apresenta uma breve conclusão das principais contribuições deste trabalho, bem como trabalhos futuros que podem ser propostos com base no que foi desenvolvido.

# Capítulo 2 As sequências repetitivas e sua relação com fenômenos biológicos

O estudo das sequências repetitivas (SR) ganhou maior interesse da comunidade científica após a confirmação de que tais sequências são responsáveis pela mediação de fenômenos biológicos associados à vários níveis de regulação gênica: transcrição, *splicing* etc. Neste capítulo serão apresentados alguns fenômenos que foram comprovados ou que já possuem fortes evidências de que podem estar sendo mediados pelas SR. Para tal, é feita uma breve revisão da literatura com informações que dão suporte aos objetivos propostos por este trabalho, principalmente com relação ao fenômeno de formação de transcritos quiméricos por meio de *trans-splicing* e sua mediação por sequências repetitivas do tipo reverso complementar.

## 2.1 O DNA como fonte de material genético

Os cromossomos encontram-se em uma estrutura quimicamente estável e condensada em associação com proteínas envolvidas pela molécula de DNA, composto de vários genes (vide Figura 1), que codificam ou não proteínas, e outras sequências não codificantes de nucleotídeos com funções especificas, porém ainda pouco conhecidas (Wieczorek & Stortz, 2002; Stewart, 2007). É importante destacar que, atualmente, a definição do termo "gene" tem sido motivo de debate na comunidade científica. Há uma divisão de uma parte dela que considera um gene como sendo qualquer parte do genoma que é transcrito, já há outros que consideram um gene como sendo apenas as partes que são codificadas em proteína

(definição também adotada neste trabalho). Uma breve discussão disso pode ser encontrada nos trabalhos de Scherrer & Jost (2007), de Gros (2009) e de Forsdyke (2009).

Para que os genes desempenhem sua função em um ser vivo, eles precisam passar por um processo chamado transcrição, que transforma a molécula de DNA em uma de RNA, e, para alguns genes, por um processo de tradução, que converte o RNA em uma proteína (Figura 5).

A transcrição é um processo que ocorre no núcleo da célula, seja ele codificante ou não, que foi catalisado pela enzima RNA polimerase (RNApol ou RNAP), gera uma sequência específica de RNA, chamado transcrito. Tais transcritos podem se transformar em uma proteína ou atuar no processo de transcrição e de tradução de um gene que codifica uma proteína (Stortz, 2002). Cada tipo de transcrito diferencia-se dos demais na estrutura e na função. Alguns dos mais conhecidos são:

a. RNA transportador (tRNA): Encaminha os aminoácidos dispersos no citoplasma ao local onde ocorrerá a síntese das proteínas;
b. RNA ribossomal (rRNA): Faz parte da estrutura dos ribossomos (organelas citoplasmáticas) onde a síntese de proteínas ocorre;
c. micro RNA (microRNA ou miRNA): participa do processo de regulação de outros genes; e
d. RNA mensageiro (mRNA): único tipo de transcrito que pode ser traduzido para uma proteína. Ele é composto por uma região codificante (CDS) e, nas extremidades, é composto por duas sequências distintas que não são traduzidas, chamadas 5’UTR à esquerda, e 3’UTR à direita (Figura 5).

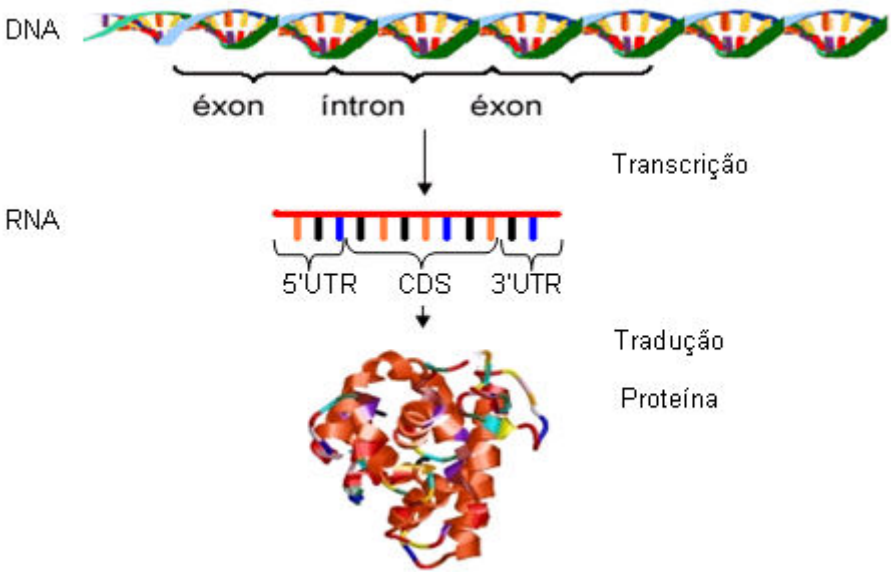


*Figura 5 – Transcrição e tradução de um gene.*
*O DNA é composto por regiões gênicas que, quando transcritas, geram moléculas formadas por éxons e íntrons. Tais moléculas são processadas, de forma que os íntrons são removidos e os éxons sejam mantidos, para a formação de um transcrito maduro. Se a molécula formada for de mRNA, o transcrito pode sofrer tradução, que corresponde a formação de uma proteína.*
*Fonte: da própria pesquisa.*

Com exceção do mRNA, todos os outros tipos de RNA transcritos não podem ser traduzidos para uma proteína e são conhecidos como RNA não codificante ou ncRNA (em bactérias são sRNA) (Stortz, 2002). Apesar da identificação dos tipos de RNA listados, o número total de mRNA e de ncRNAs codificado por um genoma ainda é desconhecido.

## 2.2 Sequênciamento de transcriptomas

Devido à importância biológica associada ao tipo de expressão dos transcritos, surgiram diversos projetos de pesquisa com o intuito de mapear o transcriptoma de vários tipos de organismos. As técnicas empregadas para a realização de tal atividade concentram-se em alguns modelos principais. Técnicas como SAGE e RNAseq são aplicadas para se tentar avaliar e analisar os níveis de expressão de genes (Høgh & Nielsen, 2007). Uma

delas, baseada em EST (*expressed sequence tag*), consiste no sequenciamento de sequências curtas (entre 200 a 500 pb) de transcritos, das extremidades 5′ ou 3′ de cada molécula. Outra é baseada em *full-length* cDNA (FlcDNA, DNA complementar), consiste no sequênciamento total, ao invés de parcial, de um transcrito e, desta forma, o tamanho da sequência varia conforme o tamanho do transcrito (Lodish et al., 1999).

Apesar de amplamente utilizada, a técnica de identificação de transcritos baseada em EST apresenta resultados modestos, pois as sequências geradas são menores e com menor qualidade que aquelas geradas por meio do uso de FlcDNA. Consequentemente, a cobertura das sequências é menor, e sua qualidade é inferior. Apesar da diferença de qualidade e de cobertura entre ambas, a definição do tipo de tecnologia utilizada, apesar de parecer vantajosa somente para FlcDNA, depende da finalidade. Em trabalhos cujo objetivo é descobrir novos tipos de trancritos, o uso de bibliotecas de EST é mais vantajoso, pois seu custo é menor do que FlcDNA, e o tempo de obtenção das sequências é reduzido.

Quando um gene sofre *splicing* alternativo no qual os exons terminais das extremidades 5’ e 3’ dos transcritos gerados permacem inalterados, a tecnologia de EST é menos sensível à detecção de tais variações do gene. Por meio de *full-length* cDNA, pelo contrário, pode-se identificar todas as variações geradas pelo *splicing* alternativo.

## 2.3 SR em genomas de animais e plantas

Embora sejam as grandes responsáveis, na grande maioria dos casos, pela geração de erros de montagem ou pela impossiblidade de se gerar *contigs* a partir de *reads*, as SR possuem um papel comprovadamente muito importante em diversos organismos. Além

disso seu estudo mostra-se importante para tentar entender os motivos pelas quais tais sequências são altamente frequentes em um genoma. Alguns fenômenos biológicos possuem relação com as SRs e serão discutidos nas seções seguintes.

## 2.4 SR e sua relação com fenômenos biológicos

O estudo das SR ganhou maior importância porque foi descoberto que trechos de DNA transcritos, quando formados por SR do tipo reverso complementar (*inverted repeat*) (exatamente ou parcialmente complementares entre si), independentemente do tamanho, podem se auto complementar (ligar) formando estruturas conhecidas como *stem-loop* (também conhecidas como *hairpin*, *hairpin-loop* ou grampo) (Figura 6 (a)); *stem-loop* aninhado, chamado *pseudoknot* (Figura 6 (b)) (Goodrich & Kugel, 2006), e um outro tipo que envolve o pareamento de pelo menos duas sequências distintas (Figura 6 (c)) (Fire et al., 1998). Tais estruturas podem ocorrer tanto em regiões gênicas quanto em regiões intergênicas ou intragênicas, independentemente das espécies até então observadas (Forsdyke, 1995 a,b,c). Também, recentemente, Lewin (2007) verificou em um caso isolado que existe a possibilidade da ocorrência de um *stem-loop* (neste caso chamados de terminadores) durante a transcrição, na qual se inibe a ação da enzima RNA-polimerase, fazendo com que a transcrição pare, sem que seja gerado um transcrito completo.

Apesar da possibilidade de ocorrência de *stem-loops* em transcritos de DNA, o principal fator que fez aumentar esforços voltados ao entendimento da função das SR do tipo *inverted repeat* surgiu, principalmente após a descoberta do mecanismo de regulação gênica de RNA de interferência, descoberto por Fire et al. (1998). Existem associações que

também ligam a ocorrência dos fenômenos de *cis-splicing* e *trans-splicing* (Gingeras, 2009) com a presença de SR do tipo *inverted repeats*.

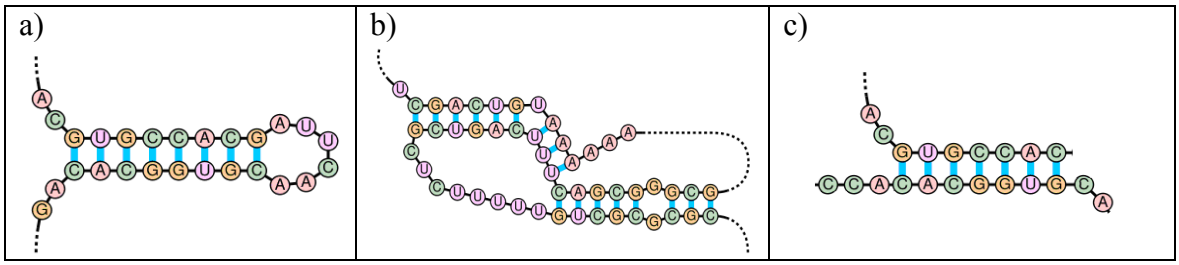


*Figura 6 – Complementaridade entre sequências de nucleotídeos.*
*Sequências de nucleotídeos com bases complementares (inverted repeats) entre si e auto-ligadas: a) stem-loop: somente uma auto-ligação; b) pseudoknot: seqüência com stem-loops aninhados; c) complementaridade entre sequências independentes formando inverted repeat.*
*Fonte: da própria pesquisa.*

### 2.4.1 Interferência por RNAi

RNAi é um mecanismo na qual uma sequência é desencadeada a partir de regiões repetitivas do tipo reverso e complementar (RRC) com relação ao mRNA, de RNA não codificante de proteínas (ncRNA) (Robinson, 2004). Há evidências concretas de que RNAi pode atuar como um elemento chave na regulação de processos relacionados ao início do desenvolvimento de um organismo: em plantas (Reinhart et al., 2000), na proliferação e morte (apoptose) das células (Brennecke et al., 2003), diferenciação celular (Dostie et al., 2003; Chen, 2004), e infecção viral associadas por ligações entre ncRNA e doenças causadas por vírus e câncer (Pfeffer et al., 2004).

Assim como as SR do tipo reverso e complementar, os trechos de DNA que compõem o RNAi podem ser provenientes de inúmeras partes do genoma, porém principalmente a partir de regiões não codificantes (ncRNA) (Kim & Nam, 2006; Rana, 2007), conforme esquema da Figura 7.

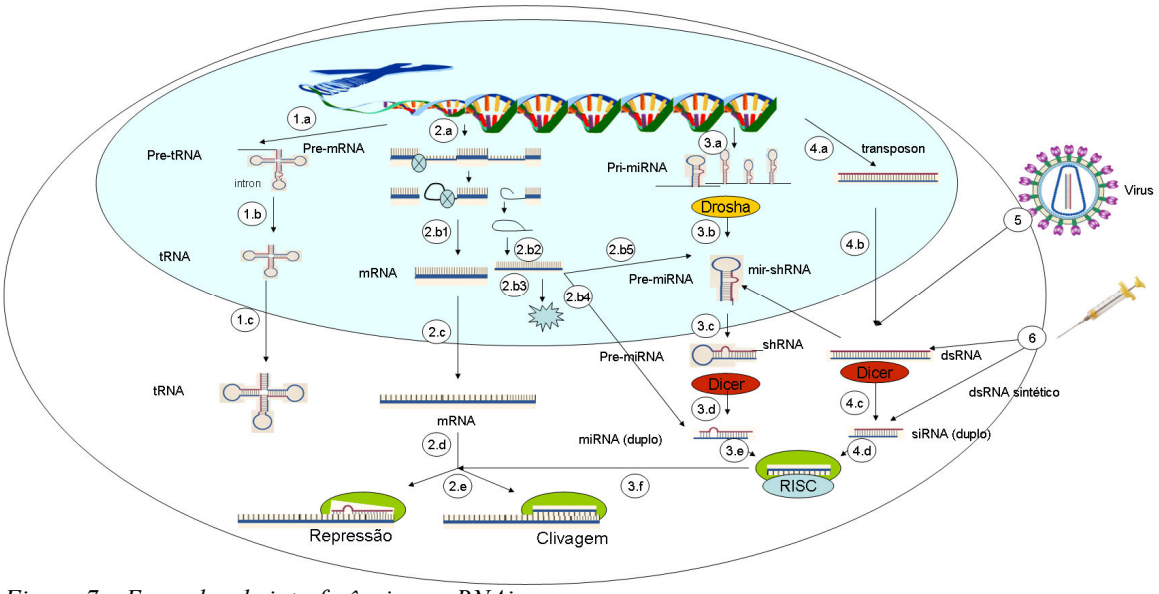


*Figura 7 – Exemplos de interferência por RNAi.*
*Processos celulares gerados por transcrição e tradução de DNA, dependentes ou não da ação de interferência por meio de RNAi: 1.a) transcrição de pre-tRNA; 1.b) remoção de íntrons da região do anticódon ativo, quando existem, e ação de enzimas; 1.c) exportação do tRNA para o citoplasma; 2.a) geração do pré-mRNA; 2.b1) remoção de íntrons e geração do mRNA; 2.b2) remoção dos íntrons; 2.b3) degradação do íntron; 2.b4) íntron pode ser exportado para o citoplasma e se transforma em miRNA; 2.b5) íntron que forma stem-loop pode se transformar em um pre-miRNA; 2.c) exportação do mRNA para o citoplasma; 2.d) interferência por RNAi, para repressão ou clivagem do mRNA, quando ocorre paremanto parcial ou total, respectivamente; 3.a) transcrição do pri-miRNA; 3.b) enzima endonuclease Drosha RNase III corta a seqüência de pri-miRNA próximo do stem-loop e gera outro stem-loop de pré-miRNA; 3.c) pré-miRNA é exportado para o citoplasma; 3.d) Dicer reconhece o pré-miRNA, cortando e removendo o grampo, gerando pequenas sequências de miRNA de dupla fita; 3.e) miRNA é associado com o complexo RISC que dissocia e elimina um dos lados do dsRNA; 3.f) reação de interferência por RNA; 4.a) transcrição das duas fitas de um elemento genético móvel: transposon; 4.b) transcrito é exportado para o citoplasma; 4.c) longa cadeia de dsRNA é cortada, pela enzima Dicer, em siRNA fita dupla; 4.d) siRNA duplas são incorporadas no complexo RISC, que as separa 5) introdução de dsRNA viral; 6) injeção de shRNA; dsRNA ou siRNA de meio externo.*
*Fonte: da própria pesquisa.*

Na interferência por RNAi, o ncRNA, devido a sua instabilidade e alto poder de ligação (Stark et al., 2008; Fire et al., 1998), junta-se, em geral, ao mRNA causando redução no nível de expressão (pareamento parcial), ou clivagem (pareamento exato) no gene. Apesar dos sucessos obtidos em diversas pesquisas até então realizadas com RNAi, faltam muitas informações para que se possa entender de maneira conclusiva outros tipos de elementos que também podem estar envolvidos no processo de transcrição, e também a forma com que as SR atuam no nível de expressão dos genes. Neste sentido, todas as SR

com poder de ligação são importantes, pois ainda não se tem idéia clara, por exemplo, de qual dos lados de uma sequência de DNA pode ser gerado um transcrito (Stark et al, 2008), se um íntron de qualquer tipo de transcrito primário pode atuar na interferência por RNA ( Di Segni et al., 2005), se um íntron poderá formar *stem-loop*, mesmo que seja composto por uma sequência repetitiva do tipo reversa complementar (RRC).

Atualmente há também fortes evidências que associam as sequências repetitivas do tipo *inverted repeats*, entre elas algumas do grupo das ALUs que podem gerar tais tipos de sequências, com o processamento de transcritos primários de mRNA (pré-mRNA), o *cis-splicing*, porém raramente com o *trans-splicing*. Ambos são brevemente discutidos a seguir.

### 2.4.2 Processamento de RNA por *splicing* constitutivo e alternativo e sua relação com *Inverted Repeats*

*Splicing alternativo* é um processo pós-transcricional exclusivo dos seres eucariotos no qual um único transcrito primário (pré-mRNA), formado por estruturas do tipo éxon-íntron, pode gerar diferentes moléculas de mRNA. Quando são transcritos para gerar pré-mRNAs, sequências de íntrons são removidas, e éxons adjacentes são ligados entre si (Anderson & Staley, 2008). Tal processo, conhecido como *splicing* constitutivo, ou *cis-splicing*, é catalisado por um complexo de RNAs e proteínas chamado *spliceossomo*, componentes estes que reconhecem e clivam as extremidades 5' e 3' de cada íntron, realizando a junção dos éxons. Ao final, na extremidade 3' do transcrito, na maioria dos casos, uma enzima chamada RNA polimerase realiza a adição de uma sequência do tipo poli(A) (Hui et al., 2005).

Entretanto, para uma grande quantidade de genes, o *splicing* pode funcionar de maneiras alternativas, conforme Figura 8, o que permite aumentar a capacidade de codificação de um gene quando é formado por múltiplos íntrons. Em função disso, podem ser produzidas formas distintas de mRNA e, consequentemente, a síntese de isoformas de proteínas, estruturalmente e funcionalmente distintas (Hui et al., 2005).

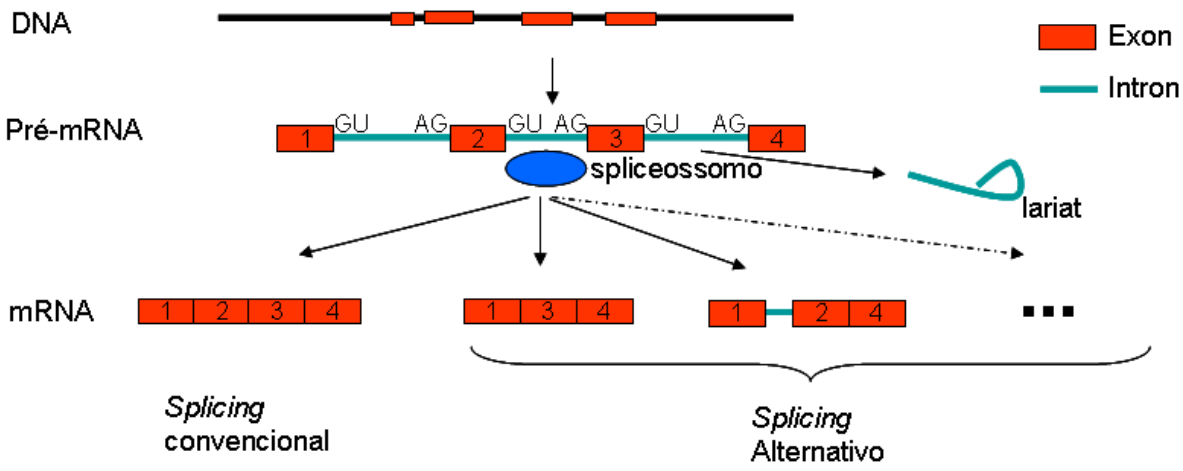


*Figura 8 – Transcrição e splicing a partir do DNA.*
*Etapas do processo de transcrição primária e final a partir de uma região genomica. O processamento do transcrito primário pode ser do tipo constitutivo ou convencional (esquerda) ou alternativo (direita).*
*Fonte: da própria pesquisa.*

A ocorrência do mecanismo de *splicing* é mediado por meio de sequências de dinucleotídeos específicas e bem conhecidas, presentes nas extremidades de um íntron, e, em mais de 99% dos casos, começam por uma sequência **GU** e terminam com a sequência **AG** (posição 5' para 3'), e costumam ser referenciados pelos termos *splice donor* e *splice acceptor*, respectivamente. Entretanto, apenas as sequências nas extremidades de um íntron não são suficiente para indicar sua presença. Outra sequência que é muito importante, chamada de *branch site*, é "CU(A/G)**A**(C/U)", sendo **A** conservado em todos os genes. Ela localiza-se a aproximadamente 20 e 50 bp da região *upstream* (5') do *splice acceptor* (Hui et al., 2005). Em outros casos, em menos de 1% deles, há os sinais de *splice-sites* ditos raros, os quais ainda são pouco conhecidos e que, segundo Steitz et al. (2008), as moléculas

de pré-mRNA podem sofrer *splicing*, inclusive, no citoplasma, não somente da forma convencional, em que o pré-mRNA sofre o *splicing* no núcleo da célula. Ambos os casos são mediados por spliceossomos, porém distintos.

Em alguns casos (como no gene fl(2)d, da *Drosophila melanogaster*), um sinal de *splicing* pode ser mascarado por uma proteína regulatória, resultando em *splicing* alternativo. Entretanto, há casos raros de genes que possuem uma grande quantidade de sinais de *splicing* ambíguos (como em genes de HIV), dos quais apenas alguns produzem moléculas de mRNA (Faustino & Cooper, 2003).

Apesar deste mecanismo convencional de *splicing* ser muito estudado e com muitas informações já descobertas, há ainda muitos estudos que tentam identificar novas particularidades associadas à ocorrência de tal fenômeno, entre eles sua relação com a presença de sequências repetitivas do tipo *inverted repeats*. Um dos primeiros trabalhos que fizeram tal associação foi proposto por Munroe (1993), porém as análises realizadas foram pouco conclusivas devido a falta de dados e qualidade dos mesmos, principalmente pelo fato de que a tecnologia empregada na época era mais propensa a erros. Em outro trabalho desenvolvido por Melquist & Bender (2004), desta vez com o uso de um procedimento experimental, verificou-se em genes da planta *Arabidopsis* a ocorrência de *inverted repeats* que, após a formação de um *stem-loop*, passaram a inibir a transcrição, na qual ocorre uma poliadenilação prematura do transcrito. Um estudo muito similar foi feito por Lian & Garner (2005) em humanos, na qual verificaram que a formação de *stem-loops* por meio de *inverted repeats* em íntrons distintos afetaram diretamente o processo de *splicing* alternativo.

Em outro trabalho desenvolvido por (Gal-Mark et al.**,** 2008), foi feito um estudo na qual sequências repetitivas da classe das ALUs em regiões intrônicas podem fazer com que tais regiões adquiram mutações que geram sinais funcionais de *splice-sites*, em um processo conhecido como exonização. A maioria deles ocorre na extremidade direita de elementos ALU. Foi verificado também que, sem a extremidade esquerda, a exonização da extremidade direita modifica a forma de *splicing*, que passa a ser constitutivo, na qual a região do novo éxon sempre fará parte do transcrito final.

Com o intuito de analisar a conexão entre *splicing* alternativo e a exonização de elementos do tipo ALU, Sorek et al. (2002) utilizaram uma base de dados de ESTs (etiqueta de sequência expressa) e de cDNAs que foi alinhado, com a ferramenta BLAST, contra o genoma humano. Foi utilizado um conjunto de transcritos formados por dois tipos distintos de éxons: um com 1176 transcritos formados por éxons que participam de *splicing* alternativo, e outro formado por 4151 transcritos formados por exons constitutivos. Na análise, a grande maioria, 84%, de éxons contendo ALUs geram um sinal de *frame-shift*, ou de um *stop-códon* terminal prematuro. Estes resultados comprovam que éxons que possuem sequências do tipo ALU são, predominantemente, afetados pelo *splicing* alternativo.

Recentes análises realizadas por DeCerbo & Carmichael (2005) utilizaram uma metodologia baseada em bioinformática na qual observaram que a maioria dos transcritos humanos são editados por uma enzima chamada *adenosina deaminase* (ADAR), que converte adenosinas para inosinas. A maioria dessas edições ocorre em elementos ALU, presentes nos transcritos. Os autores relatam que tal mecanismo de edição pode estar envolvido em funções como a regulação do mecanismo de *splicing*.

Apesar de corresponder a um processo biológico ainda em amplo estudo, o mecanismo de *cis-splicing* já apresenta diversos padrões conhecidos. Diferentemente disso, o *trans-splicing* é um mecanismo relativamente raro em organismos superiores e pouco entendido e, por este motivo, tem gerado trabalhos com o intuito de se entender melhor este tipo de fenômeno biológico, e que provavelmente está entre os mais interessantes no processo de formação dos genes.

## 2.5 Formação de transcritos quiméricos e sua relação com *Inverted Repeats*

A formação de um transcrito quimérico corresponde a um evento que pode se dar por meio de *trans-splicing*, em que dois transcritos, oriundos de um mesmo cromossomo (intracromossomal, Figura 9 (a)) ou de cromossomos distintos (intercromossomal, Figura 9 (b)), fundem-se para formar o transcrito quimérico.

*Trans-splicing* é uma forma rara de *splicing* de RNA, na qual dois transcritos contribuem para a formação de um único mRNA. O trans-splicing foi classificado por Bonen (1993), que propôs duas categorias de *trans-splicing*: "*spliced leader*" (SL) e "*discontinuous group II intron.*" O primeiro é comum em organismos inferiores, como em nematóides (Blumenthal, 2005) e kinetoplastideos (Mayer & Floeter-Winter, 2005), em que uma pequena sequência líder (ou *spliced lider*) é anexada na região 5'UTR de um transcrito de mRNA. O outro tipo é encontrado em plantas, cloroplastos de algas e em mitocôndrias de plantas, e envolve a junção de sequências codificantes independentes. Ambas as categorias respeitam as regiões de consenso dos *splice-sites* e são catalisadas pelos spliceossomos (Blumenthal, 2005).

Apesar de corresponder a um fenômeno ainda pouco compreendido, já foi relatado há bastante tempo atrás por Agabian (1990). A categoria de trans-splicing associada aos SL apresenta uma quantidade muito maior de trabalhos realizados pelo simples fato de estarem associadas a microorganismos de interesse comercial e, principalmente, a diversos patógenos causadores de algum tipo de doença em organismos superiores, como humanos.

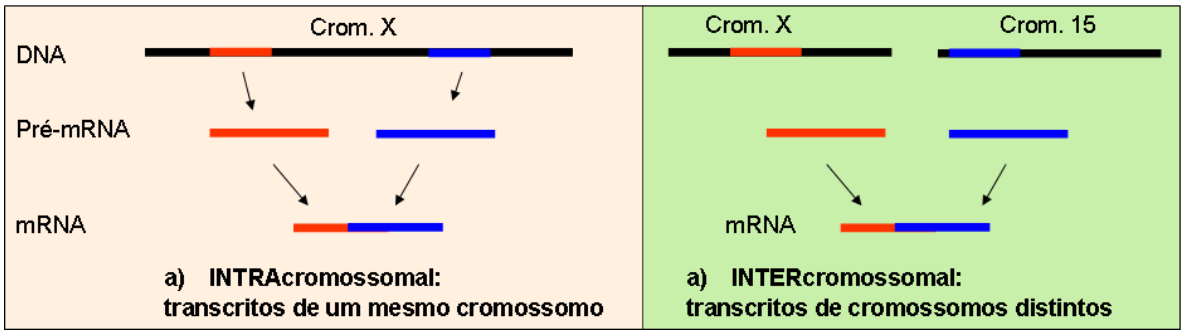


*Figura 9 – Tipos de transcritos quiméricos.*
*Formação de transcritos quiméricos por meio de trans-splicing. a) intracromossomal e b) intercromossomal.*
*Fonte: da própria pesquisa.*

Neste mecanismo, o SL é um tipo especial de transcrito formado por um dinucleotídeo GU em sua extremidade 3', que liga-se a uma adenosina localizada na região 5' de um pré-mRNA alvo. Após o processamento da nova molécula, o sequência do SL, anterior à região do dinucleotídeo, fará parte do transcrito final de mRNA (vide Figura 10) (Stover et al., 2006).

Cheng et al. (2006) apresenta uma discussão que compara as sequências de SL em vermes. Nas análises, ele verificou que a sequência AUG da região terminal 3' do SL é a responsável pelo processo de tradução dos transcritos. Uma breve comparação é apresentada na Figura 11, na qual as sequências estão alinhadas entre si.

Por mais que este tipo de categoria de *trans-splicing* ainda não tenha sido provado experimentalmente, ele tem sido adotado como um modelo pela maioria dos trabalhos que estudam o fenômeno em organismos inferiores.

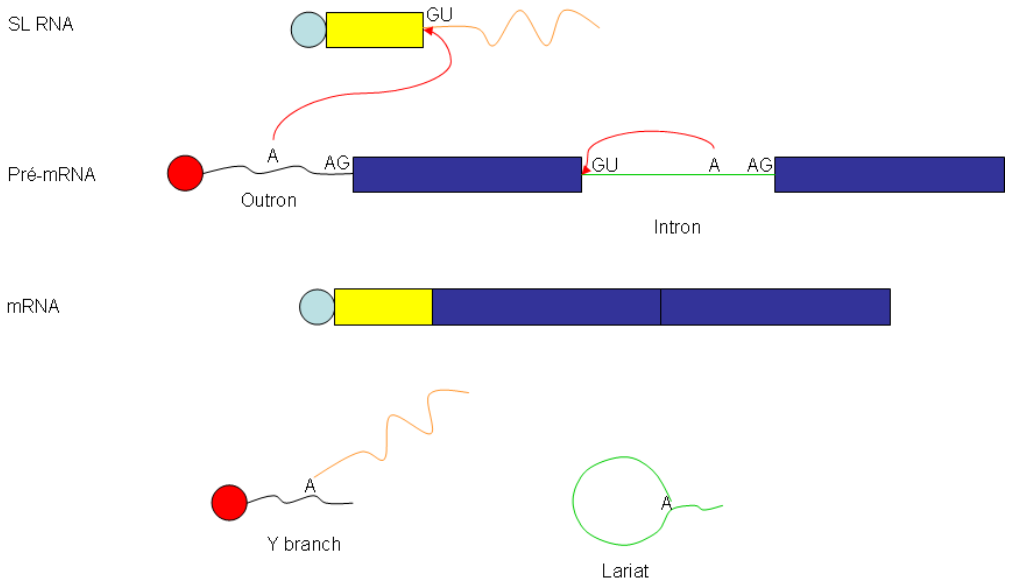


*Figura 10 – Modelo de formação de transcritos quiméricos em organismos pequenos.*
*A sequência lider (caixa amarela) de um SL RNA liga-se à uma adenosina da região 5' do primeiro exon (caixa laranha) de uma molécula de pré-mRNA por meio de uma reação de trans-splicing. O outron do pré-mRNA e a porção íntron-like do SL RNA forma um subproduto chamado Y, similar a uma estrutura de um lariat, formado durante o processo de splicing. A região azul do SL RNA é similar à estrutura promotora encontrada em transcritos normais (caixa vermelha). Tais regiões fornecem, parcialmente, novas propriedas para a geração de transcritos quiméricos em organismos inferiores.*
*Fonte: da própria pesquisa.*

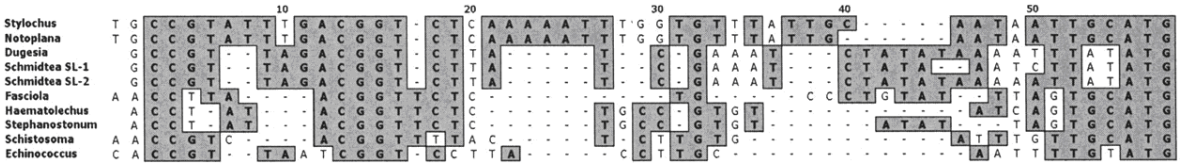


*Figura 11 – Splice-leader de diferentes organismos.*
*Fonte: baseado em Cheng et al. (2006).*

Em organismos superiores, ainda não há sequer modelos, pois sua ocorrência é tão rara de forma natural que a quantidade de evidências ainda é muito pequena. Não se tem idéia se a ocorrência deste fenômeno está associada com fatores fisiológicos ou se é algo aleatório ou ligado a algum tipo de mecanismo evolutivo. Desta forma, embora muito

menos frequente do que o mecanismo de *cis-splicing*, *trans-splicing* tem gerado grande interesse da comunidade científica devido as suas potenciais aplicações. Por exemplo, *trans-splicing* já foi aplicado para corrigir defeitos genéticos na expressão de genes com efeitos indesejáveis (Tahara et al., 2004), ou foi utilizado em uma variedade de aplicações voltadas para a saúde humana, descritas na literatura especializada (Pergolizzi et al., 2003; Schlesinger et al., 2003; Garcia-Blanco, 2003; Garcia-Blanco et al., 2004; Mansfield et al., 2003; Chao et al., 2003; Otto et al., 2003). Nos últimos anos, foram reportadas evidências em outros organismos, mostrando que, embora seja um fenômeno raro em organismos superiores, *trans-splicing* pode ser mais comum do que o esperado.

Em insetos, Robertson et al. (2007) encontrou a primeira evidencia até então reportada de *trans-splicing* que envolve éxons internos provenientes de um *locus* diferente do gene de origem. *Trans-splicing* também foi reportado em mamíferos, como ratos (Caudevilla et al., 1998; Rigatti et al., 2004), e bovinos (Roux et al, 2006). Em humanos, o fenômeno costuma ser associado a doenças como o câncer, como mostrou Chen et al. (2005) no estudo do gene MYC, Hahn et al. (2004) por meio de uma biblioteca de EST, e no trabalho feito por Shao et al. (2006), que baseou-se na análise de um banco de dados de EST por meio de uma metodologia de bioinformática. Apenas em Romani et al. (2003) propôs-se um trabalho para buscar por *trans-splicing* em transcritos normais, por meio de uma metodologia baseada em bioinformática (ISTREs), que encontrou diversas evidências de *trans-splicing*, porém obrigatoriamente formadas por *splice-sites* canônicos. É importante destacar que em tais estudos, as regiões de consenso de *splice-sites* foi sempre observada, ou simplesmente imposta pelas metodologias propostas.

As primeiras evidências de *trans-splicing* intercromossomal em humanos foi reportado por Breen & Ashcroft (1997), em uma isoforma do gene *CaM kinase II*, formado por um transcrito pertencente ao cromossomo 10 e outro ao 18, e por Li et al. (1999) no mRNA híbrido *ACAT-1*, cuja 5'UTR foi mapeada no cromossomo 7, enquanto que o restante da sequência foi mapeada no cromossomo 1. Em ambos, curiosamente, não haviam sítios de *splice-sites* canônicos. Isto pode ser explicado pela não ocorrência de *splice-sites* canônicos, ou pela existência de algum mecanismo de *trans-splicing* que não seja mediado pelo spliceossomo.

Em outros trabalhos, Finta and Zaphiropoulos (2002) mostraram que o fígado humano tem o potencial de produzir uma variedade de moléculas quiméricas de mRNA *CYP3A* (*cytochrome P450 3A*) a partir de 4 genes compartilhando um alto grau de similaridade, *CYP3A4*, *CYP3A5*, *CYP3A7* e *CYP3A43*. Em ratos, 3 variantes de mRNA do gene *Msh4* (*mutS homologue 4*) são produzidos por *trans*-splicing (Hirano & Noda, 2004). Tasic et al. (2002) demonstrou, em uma exaustiva análise do gene *protocadherin* (*Pcdh*), que *trans-splicing* intergênico ocorre somente entre pré-mRNAs derivados e associados com o cluster de genes *Pcdh* e um gene próximo, *mDia1* (homólogo da *Drosophila diaphanous*).

Apesar de poucas evidências naturais da ocorrência em organismos superiores, já há procedimentos de terapia gênica que tiram proveito do mecanismo de *trans-splicing* para corrigir falhas em genes. Um dos procedimentos mais conhecidos é o SMART$^{TM}$, desenvolvido por Puttaraju et al. (1999). Nele, moléculas sintéticas de pré-mRNA são sintetizadas em laboratório, os quais possuem uma região complementar para que possam hibridizar-se a um gene alvo, conforme Figura 12, para gerar um transcrito quimérico de

pré-mRNA, que é processado pelo complexo spliceossomo para a geração de uma molécula de mRNA madura.

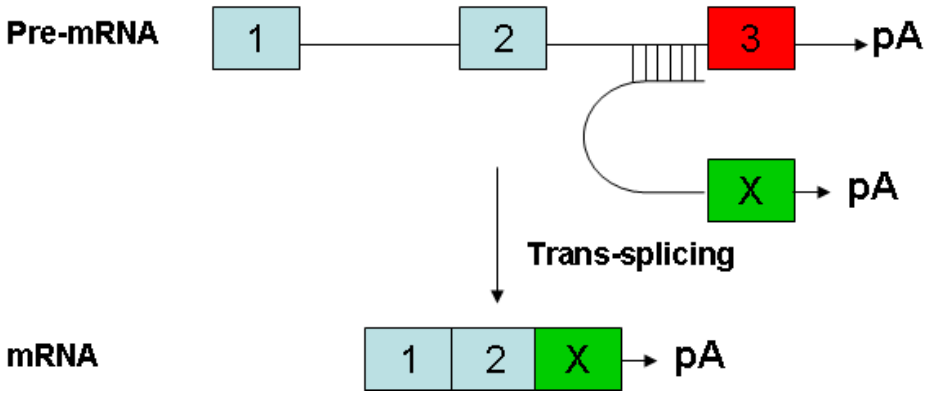


*Figura 12 – Trans-splicing pelo método SMART™.*
*Diagrama esquemático do processo de trans-splicing mediado por spliceossomo (SMART™). Uma molécula de pré-trans-splicing é mostrada ligando-se a um transcrito de pre-mRNA, na região 3' do segundo intron do pre-mRNAalvo. Cis-splicing deste alvo geraria um mRNA composto por 3 exons (1, 2 e 3) são unidos (isso não é mostrado). Uma reação de trans-splicing entre o splice site 5' adjacente ao exon 2 da molécula alvo e o Splice site 3' da molécula de pré-trans-splicing resulta em uma molécula de mRNA composta pelos exons 1 e 2, associadas com a sequência codificante X. Este processo pode ser feito com qualquer junção exon-intron de uma molécula alvo. Desta forma, poderia ser sintetizada uma molécula para que também os exons 1 ou 2 fossem removidos da sequência de pré-mRNA original.*
*Fonte: baseado em Mansfield et al. (2004).*

Há diversas evidências que sugerem uma relação entre as sequências repetitivas do tipo *inverted repeat* com o mecanismo de formação de transcritos quiméricos por meio de *trans-splicing*, similar ao mecanismo sintético proposto pelo SMART™.

Em Drosophila, *trans-splicing* foi reportado por Labrador et al. (2001) e por Horiuchi & Aigaki (2006), que observaram que o fenômeno de *trans-splicing* ocorria após a união de transcritos de pré-mRNA independentes, formando um RNA de dupla fita por meio de sequências complementares (*inverted repeat*) no gene *mod(mdg4)*. Recentemente, um mecanismo similar foi descrito, em que sequências repetitivas do tipo *inverted repeat* flanqueavam dois genes distintos, aproximando seus respectivos pré-mRNAs para permitir a ocorrência do evento de trans-splicing (Fischer et al., 2008).

Em outro trabalho, Dixon et al. (2007) estudou transcritos com éxons duplicados, apresentado na Figura 13, cujas duplicações não aparecem no genoma. Em uma análise computacional de 48 transcritos deste tipo, ele identificou duas regiões distintas. Uma região rica em pirimidina e mais comum na região *upstream* do íntron, e outra altamente rica em purinas e mais comum na região *downstream* de um íntron. Como as duas regiões são complementares entre si, o modelo propõe que a duplicação de éxons é resultado de um *trans-splicing* entre dois transcritos de pré-mRNA de um mesmo gene e são ligadas entre si durante a transcrição por meio das sequências intrônicas complementares. Mais ainda, a maioria dessas regiões é formada por trechos repletos de elementos genéticos móveis, tais como as ALUs, possivelmente responsáveis por favorecer a ocorrência do fenômeno de *trans-splicing*.

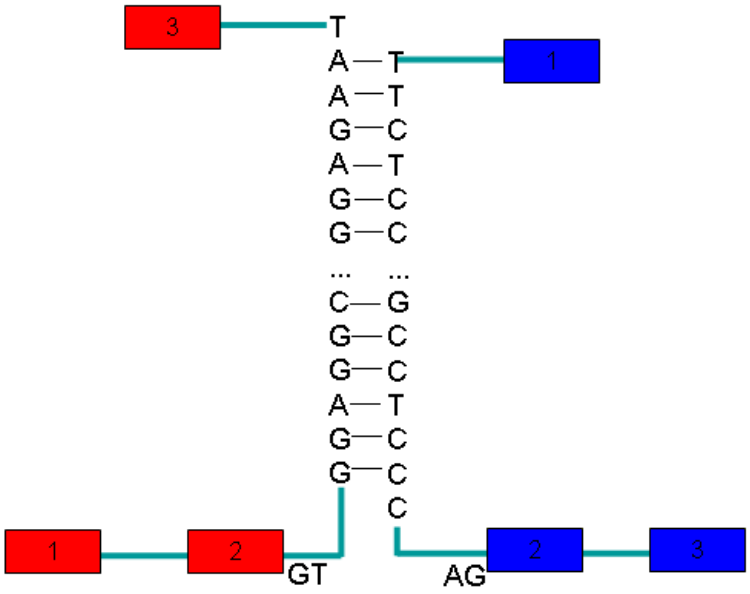


*Figura 13 – Transcrito quimérico formado por sequências reversas complementares.*
*Formação de um transcrito quimérico que pode ter sido mediado pela presença de sequências repetitivas do tipo reverso complementar (RRC).*
*Fonte: ilustração baseada em Dixon et al. (2007).*

Também, de forma muito breve, Di Segni et al. (2008) cita que uma possível explicação para a ocorrência do fenômeno de *trans-splicing* de sequências de mRNAs por meio de endonucleases de tRNA (vide Figura 14) é a presença de sequências repetitivas do tipo *inverted repeat*.

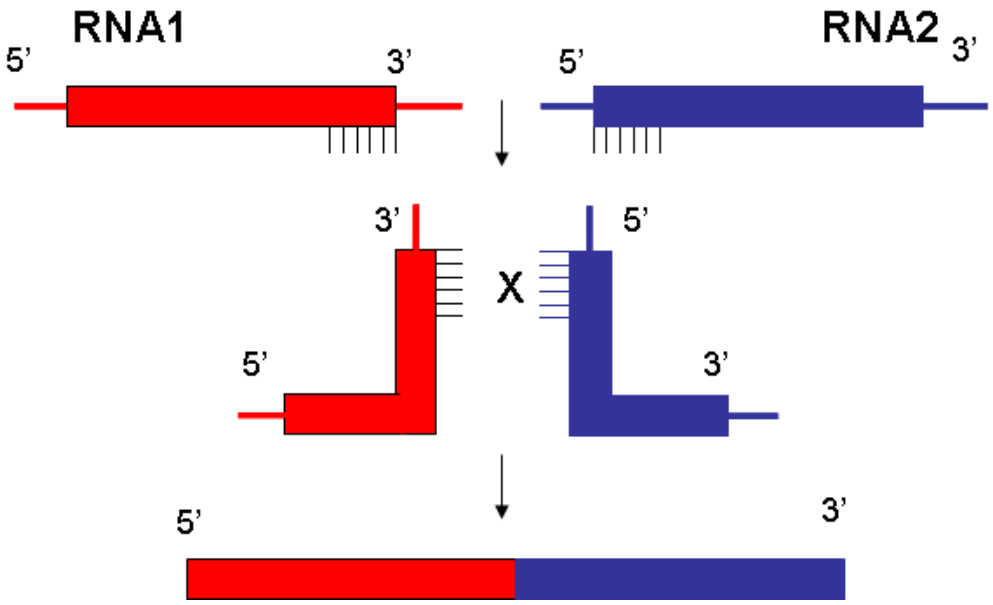


*Figura 14 – Transcrito quimérico de tRNA mediado por sequências repetitivas.*
*Formação de endonucleases de tRNA a partir de dois transcritos independentes. RNA1 possui uma sequência repetitiva em sua extremidade 3' que é reversa complementar da extremidade 3' da sequência RNA2. O par de sequências repetitivas hibridiza para formar uma nova molécula de tRNA.*
*Fonte: adaptação de Di Segni et al. (2008).*

Apesar de poucas informações conclusivas a respeito da relação direta entre *trans-splicing* e as sequências repetitivas do tipo *inverted repeat*, as informações preliminares geradas pelas pesquisas até então realizadas demonstram que possa existir tal relação. Além disso, a não ocorrência das regiões de *splice-sites*, como foi verificado no gene ACAT-1, demonstra a possibilidade da existência de algum outro mecanismo de processamento de RNA por meio de *splicing* que não seja, necessariamente, mediado pelos spliceossomos. As abordagens de bioinformática até agora sempre exigiram a ocorrência de *splice-sites*

canônicos, deixando de lado as melhores evidências experimentais, pelo menos no que se refere a humanos.

Além disso, como resultado de nossos experimentos, conjecturamos que um mecanismo até agora não relatado de formar transcritos quiméricos pode ser por meio de retrotransposição de elementos genéticos móveis, como retrotransposons. Essa conjectura será retomada nos capítulos vindouros.

## 2.6 Qualidade e detecção de erros de montagem

É importante ressaltar que, apesar da elevada tecnologia empregada para a geração de sequêcias de genomas e transcriptomas, um grande desafio reside ainda na geração de sequências isentas de erros de montagens, de forma que as pesquisas associadas com a identificação de novos padrões biológicos sejam mais confiáveis. Tais erros podem ser proveninentes de algum tipo de contaminação por enzimas de restrição, DNA mitocondrial, vetores de clonagem, entre outros, ou provenientes da própria tecnologia empregada. As tecnologias atuais, como aquelas baseadas em NGS, ainda costumam ser incapazes de sequenciar com alta precisão regiões maiores do que 1kb sem o uso de mecanismos, em sua maioria estatísticos, de sobreposição dos *reads* para a montagem de um genoma. Além disso, trechos compostos por sequências repetitivas longas podem gerar sequências com baixa confiabilidade. Em função disso, foram propostos alguns trabalhos para tentar detectar erros em montagens de genomas "*draft"*.

Cheung et al. (2003), desenvolveu uma heurística computacional veloz baseada na análise dos alinhamentos gerados pela ferramenta BLAST para detectar segmentos

duplicados do genoma humano (versão de 2002). Em tal trabalho, foi observado que cerca de 107.4 Mb (3.53%) do genoma continha duplicações segmentais erradas.

Sundquist *et al.* (2007) propôs um mecanismo de detecção de falsas sobreposiçoes durante a montagem de *contigs* de um genoma. Ele baseia-se no uso de informações a respeito de elementos correlacionados ou repetitivos.

Em outro trabalho, Gajer *et al.* (2004), a partir do uso de informações da montagem de um genoma, desenvolveu um sistema chamado AutoEditor que é capaz de, significantemente, fornecer maior qualidade na montagem com relação à uma montagem gerada a priori.

## 2.7 Considerações do capítulo

Este capítulo discutiu, de forma breve, algumas características importantes das sequências repetitivas. Foi visto no Capítulo 1, que tais sequências, quando são abundantes em um genoma, dificultam o processo de montagem do genoma, principalmente com o uso das chamadas novas tecnologias (NGS). Apesar disso, foi com a comprovação biológica de sua função em um organismo (como em RNAi) que as pesquisas sobre SR ganharam maior interesse da comunidade científica em busca dos segredos que circundam a forma com que a informação genética de um organismo é armazenada e utilizada para determinar as características fundamentais de um ser vivo. Atualmente, conforme discutido, já há fortes evidências que relacionam a ocorrência de fenômenos biológicos dentre os mais interessantes, e que podem estar associados com a presença de sequências repetitivas em um genoma. O próximo capítulo apresenta uma breve discussão disso. Em função da

importância biológica das SR elas foram classificadas e diversas ferramentas de bioinformática, tanto para busca quanto para armazenamento e anotação foram propostas. O próximo capítulo discute tais questões de forma detalhada.

# Capítulo 3 Caracterização das sequências repetitivas

Este capítulo apresenta brevemente uma classificação das SR e um conjunto de abordagens de bioinformática para detecção, anotação e armazenamento de SR em sistemas de gerenciamento de bancos de dados, tanto para consultas quanto para novas anotações. Tais trabalhos foram motivados pela necessidade de se entender qual a proporção de um genoma ou transcriptoma que é composto por SR, e também pela comprovação biológica de que tais sequências possuem função comprovada dentro de um organismo.

## 3.1 Classificação das SR

Grosso modo, as SR, do ponto de vista biológico, podem ser classificadas em *tandem repeats* e *interspersed repeats* (Jurka et al., 2007). As *tandem repeats* são caracterizadas por fragmentos de DNA que aparecem imediatamente adjacentes ou muito próximas entre si, e em humanos representam cerca de 3% de todo o genoma (Lander et al., 2001; Boby et al., 2005). Elas podem ser agrupadas em três subgrupos (vide Figura 15):

- microsatélites (2 a 10 pb): são utilizados como marcadores moleculares de plantas e animais devido ao alto nível de polimorfismo encontrado em seus *loci*, proporcionando sua utilização em diversos propósitos de estudo populacional, permitindo analisar desde indivíduos até espécies proximamente relacionadas. Mutações em tais sequências podem também estar associadas com a evolução de alguns seres vivos (Moxon & Wills, 1999);

- minisatélites (11 a 99 pb): assim como os microsatélites, minisatélites são muito importantes como marcadores moleculares em estudos de melhoramento genético de plantas e animais para identificação de genes de interesse científico e econômico (Lander & Botstein, 1989; Sharma et al., 2007), ou para testes de paternidade (Rocheta et al., 2007; Lander & Botstein, 1989);
- DNA satélite (ou telômero) (sequências grandes de até 170 pb) (Gur-Arie et al., 2000; Charlesworth et al., 2002): associadas com a estabilidade dos cromossomos, elas concentram-se nas regiões dos centrômeros e telômeros.

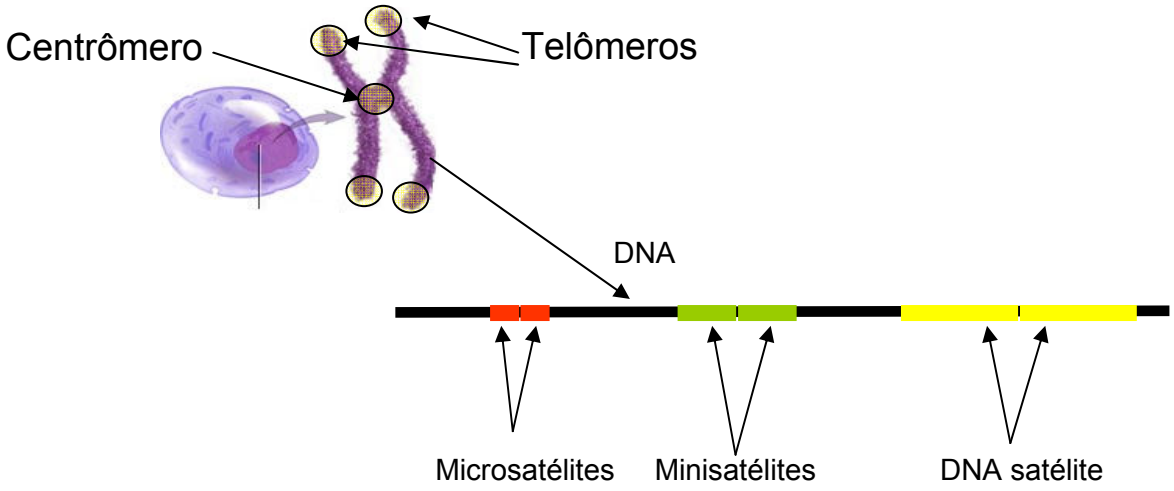


*Figura 15 – Tipos de SR dentro de um genoma.*
*Tandem repeats no DNA de um cromossomo. Tais SR aparecem adjacentes entre si, e encontram-se distribuídas no genoma sob diferentes formas.*
*Fonte: da própria pesquisa.*

Embora as *tandem repeats* tenham aplicações instrumentais em melhoramento genético, há também diversos trabalhos isolados que identificaram padrões específicos de tais SR em problemas neurológicos, doença de Huntington's, síndrome do cromossomo X, distrofia miotônica entre outros (Helvik et al., 2007; Kim & Nam, 2006), sendo todas elas relacionadas a uma expansão anormal das SR. Instabilidades com *tandem repeats* também foram associados com o desenvolvimento de câncer (Kiss, 2004).

O outro grupo de SR, as *interspersed repeats*, são fragmentos de DNA presentes em posições aparentemente randômicas de um genoma. As *interspersed repeats* costumam ser mais inativas e, frequentemente, são cópias incompletas de elementos derivados de *transposons* (*transposable elements*, *jumping genes* ou elementos genéticos móveis), presentes no DNA genômico. Em humanos, eles representam cerca de 40% de todo o genoma (Lander et al., 2001). As *interspersed repeats* podem também ser agrupadas em três subgrupos (Smit, 1999) (vide Figura 16):

- DNA *transposons*: elementos que codificam proteínas, chamadas de transposases, capazes de mover-se de uma posição para outra dentro do genoma. Para tal, eles não necessitam da enzima transcriptase reversa (TR), pois são capazes de se auto-propagar por meio das transposases, que catalisam o processo de excisão e posterior reintegração ao genoma, sem a presença de um RNA intermediário. São também caracterizados por SR reversas complementares (*inverted repeats*) em suas extremidades e que não são encontradas em outros *transposons*. No genoma humano, por exemplo, há pelo menos 14 famílias distintas de tais elementos, cujo tamanho varia entre 180 a 1200 pb, sendo algumas vezes relacionados a *transposons* que caracterizam eventos da evolução humana (Smit and Riggs, 1996). *Transposons* possuem aplicações na regulação gênica e terapia genética (Ivics & Izsvak, 2006).

- LTR *retrotransposons*: elementos caracterizados por uma região com várias centenas de pares de base, chamados *Long Terminal Repeat* (LTR), que encontram-se nas extremidades da região em que aparecem. Alguns desses elementos autônomos são estruturalmente semelhantes aos retrovírus (como HIV), porém sem um envelope

funcional o que os torna incapazes de sobreviver fora da célula, chamados de retrovírus endógenos (como o *HERV-K*). LTR *retrotransposons* correspondem a um dos principais grupos de retroelementos e estão entre os mais abundantes constituintes dos genomas eucariotos. As LTRs são repetições diretas que flanqueiam regiões codificantes, que se incluem proteínas estruturais e enzimáticas (Havecker et al., 2004).

- elementos não-LTR *retrotransposons*: elementos que se subdividem em duas classes, que de acordo com seu tamanho classificam-se em:

  - LINEs (*long interspersed nuclear element*, média de 6.1 kpb), ou elementos não-LTR, são sequências longas (6–8 kpb), com baixa quantidade de dinucleotídeos GC e que codificam uma endonuclease de função desconhecida e um polipeptídio de TR. Análises filogenéticas dos domínios das TR identificaram 11 tipos distintos de LINE. No genoma humano, elementos L1 representam o grupo de LINEs mais abundantes. A TR codificada por L1s foi proposta por estar envolvida na retrotransposição de ALUs (Jurka et al., 2007).

  - SINE (*short interspersed nuclear element*, 100 a 400 pb), em geral, utiliza a TR de uma LINE para mover-se no genoma (Deininger & Batzer, 2002). Há diferentes tipos de SINE para cada ramo da taxonomia de seres vivos. Primatas, por exemplo, possuem SINE do tipo ALU (300pb). Elas acumulam-se, preferencialmente, em regiões ricas em GC, ao contrário das LINES do tipo L1, que aparecem em regiões pobres (Jurka et al., 2007). ALUs são subdivididas em 3 grupos, sendo cada uma delas classificadas em subfamílias, que estão relacionadas com o tempo de existência de cada elemento, partindo do mais antigo (Jo e Jb), para o intermediário

(Sq, Sp, Sx, Sc, Sg, Sg1), até os mais recentes (Yb8, Ya5, Ya8) (Batzer et al., 1996). Estima-se que cerca de 10.7% do genoma humano é formado por sequências do tipo ALU (Roy-Engel et al., 2001). Embora a maioria das LINEs e SINEs estejam localizadas em regiões consideradas como não gênicas, há alguns em regiões intrônicas, como é o caso do gene da *retinoblastoma* (RB) (Álvarez, 2008).

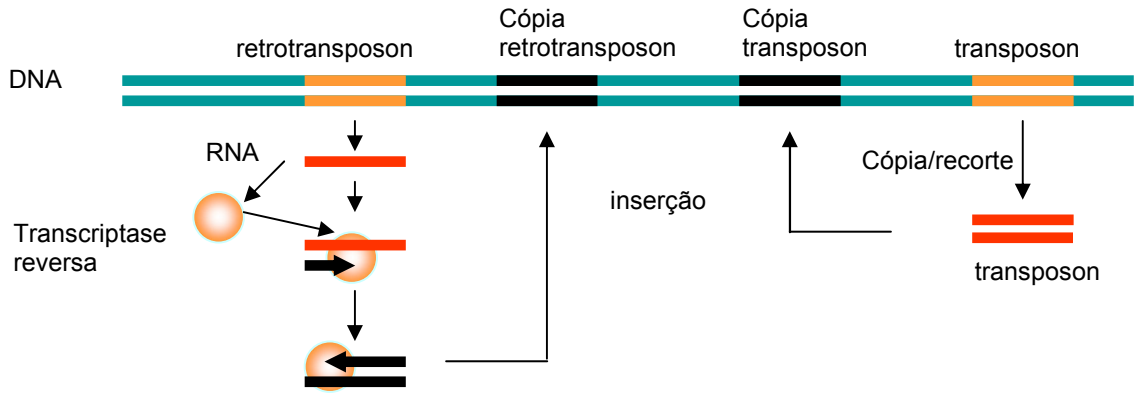


*Figura 16 – Representação de elementos genéticos móveis.*
*Representação de elementos genéticos móveis: transposons e retrotransposons.*
*Fonte: da própria pesquisa.*

Tipos similares a *interspersed* derivados de *transposons* são os genes cujo produto final é uma molécula de rRNA pertencente a uma família de multigenes, que é uma das mais repetidas no genoma humano, aparecendo de 30 a 40 vezes em segmentos de DNA com 45 kb (Strachan & Read, 1999).

A partir da classificação e da importância biológica das SR, as próximas seções apresentam um conjunto de ferramentas associadas com sua detecção, anotação e armazenamento a partir de sequências de dados biológicas.

## 3.2 Ferramentas de detecção de SR

As observações feitas a partir das SR do tipo reverso complementar (*inverted repeats*), que possuem potencial de atuação como em RNAi, fizeram com que um grande número de métodos computacionais fossem desenvolvidos para identificá-las em sequências de genomas inteiros. Métodos computacionais como miRAlign (Wang et al., 2005), ProMiR (Kim, 2005) e microHARVESTER (Dezulian et al., 2006) são técnicas baseadas em homologia de sequências e sempre partem de um padrão inicial ou específico. Há também programas de busca baseadas em perfil, como ERPIN (Legendre et al., 2005) que pode ser utilizado para buscar por miRNA homólogos a algum já existente a partir de um banco de dados. Baseado também em homologia de sequências de EST, Wang et al. ( 2005) propôs um método que permite identificar sequências de miRNA de uma forma especial, sem que o genoma esteja disponível. Alguns métodos, como o MirScan (Lim et al., 2003) são baseados em identificar miRNAs que pertençam a famílias específicas de miRNAS por meio da análise de genomas distintos e que mostram alto grau de conservação. MiRSeeker (Lai, 2003) utilizou um outro tipo de abordagem para identificar genes de miRNA por meio da análise de regiões intrônicas e intergênicas entre duas espécies de *Drosophila*. Apesar destas ferramentas baseadas em diversas metodologias, a limitação de quase todas elas é o fato pela qual procuram encontrar SR a partir de alguma já conhecida, ou a partir de um padrão pré-estabelecido, fazendo com que, juntamente com métodos baseados em homologia, não possam ser utilizados para detectar miRNAs únicos e raros. Um outro problema é que ainda não há teorias conclusivas a respeito da organização e características das sequências de RNAi, apenas casos particulares que foram

identificados. Isto torna falho o uso de abordagens fortemente baseadas em filtragem de dados com heurísticas e estatísticas.

Outras abordagens de construção de ferramentas buscaram desenvolver técnicas baseadas nos tipos de sequências em *tandem* e *interspersed repeat*, sem se preocupar com a natureza e importância biológica das sequências que atuam no processo de interferência por RNAi. As ferramentas Tandem Repeats Finder (Bensom, 1999) e Inverted Repeat Finder (Warburton et al., 2004) são muito utilizadas para identificação de *tandem repeats* porém, suas versões mais recentes, restringem as buscas para sequências de tamanho limitado. REPEATMASKER (Nishimura, 2000) é um software muito utilizado para mascarar regiões do DNA com SR que representam ruídos introduzidos por elas na busca por regiões com alta similaridade, porém identifica *tandem repeats* de forma muito limitada (tamanhos pequenos). Outra ferramenta, proposta por Schattner (2004), busca por *tandem repeats* a partir de padrões já pré-selecionados, com a meta de comprimir a informação presente no DNA. Um algoritmo baseado em heurística e estatística, associado com o programa Tandem Repeat Finder é proposto por Nam et al. (2005). A técnica geral pode ser comparada com a utilizada pelo algoritmo BLAST, que identifica todas as sequências pequenas de *tandem repeat*, de 5 a 7 pb, e a partir deles procura por fragmentos maiores. Outras abordagens, como as de Wasserman et al. (2001), Stortz (2002) e a da ferramenta EQUICKTANDEM (Rice et al., 2000) propõem heurísticas para encontrar *tandem repeats* de um tamanho ou padrão definido a priori, muito similar ao proposto por Nam et al. (2005). De forma similar funcionam MReps (Kolpakov et al., 2003) e uma ferramenta proposta por Wang et al. (2004), porém não fazem uso de heurísticas e procuram por

*tandem repeats* aproximados ao invés de exatos, partindo também de padrões pré-determinados, como também faz TROLL (Castelo et al., 2002).

Nestas categorias de ferramentas e abordagens, também podem-se identificar diversas limitações, pois, em geral, a busca parte de sequências inicialmente conhecidas, com restrições de tamanho mínimo e máximo, uso de heurísticas e estatística para remover trechos das sequências analisadas, mesmo que ainda não existam teorias conclusivas a respeito do assunto. Além disso, as inversões cromossômicas exigem que a busca por SR seja a mais flexível e geral possível. Além dos problemas mencionados, outro grande problema, senão o motivo pelas quais as abordagens atuais apresentem limitações é a proporção da quantidade de dados de um genoma com relação ao tempo que se gasta para que sejam analisadas nos melhores sistemas computacionais da atualidade.

Outra categoria de ferramentas que podem ser utilizadas para busca por SR são aquelas amplamente utilizadas para o mapeamento de sequências contra um genoma de referencia, como BLAST e BLAT.

Devido ao tamanho dos genomas, necessidade do seu entendimento e do tempo gasto em sua análise por meio de técnicas de bioinformática, surgem novas metodologias, como a de *Suffix Arrays* (Lopez, 2007), que se destacam pela sua velocidade e precisão no processamento das informações provenientes de sequências biológicas. Uma ferramenta pertencente a este grupo é o Reputer (Kurtz, 2001).

Outra classe de propostas que estuda as SR, muito mais difícil de ser realizada, procura modelar o crescimento de tais sequências. Um exemplo disso é o trabalho de Dress et al. (2003) que propôs um modelo denominado FiboString para tentar compreender o

crescimento das *tandem repeats*. Tal modelo foi inspirado na fórmula de recorrência dos números de Fibonacci, porém foi abandonado por ser demasiadamente abstrato, ou seja, não levava em consideração que as sequências poderiam apresentar ruídos (mutações, inserções, deleções, etc). Isoladamente, Yamagishi & Shimabukuro (2008) comprovaram que, para o genoma humano (Build 35), a segunda regra de paridade de Chargaff (Chargaff, 1951) se aplica. A partir da análise da distribuição das proporções de repetições nos cromossomos, constatou-se neste trabalho que as frequências se concentravam em valores específicos, cuja razão entre eles envolve quocientes de números de Fibonacci, sendo potencialmente uma evidência indireta para o modelo FiboString de Dress et al. (2003), envolvendo a razão áurea (proporção de dois números que é igual a aproximadante 1,6180, e que está envolvido com a natureza de crescimento de alguns organismos ou de suas partes). Assumindo que o modelo de Yamagishi & Shimabukuro (2008) seja correto, a presença e a enorme frequência de SR nos genomas são fundamentais para a validade da segunda regra de paridade de Chargaff. Isto é especialmente verdadeiro para as *inverted repeats*, pois estas podem ser vistas como recombinações associadas a inversões que ocorreram ao longo da evolução, e este processo tende a equalizar a quantidade de bases complementares na mesma fita de DNA (segunda regra de paridade de Chargaff).

Com base nas abordagens de estudo das SR, a próxima seção trata de sistemas de bancos de dados, os quais são fundamentais para facilitar o estudo das SR.

## 3.3 Bancos de dados biológicos

Bancos de dados biológicos são ferramentas fundamentais para a bioinformática, pois é por meio delas que se tornou possível gerenciar grandes quantidades de dados para que possam ser representadas, armazenadas e porteriormente analisadas. Seu uso tornou-se indispensável na era conhecida como pós-genômica (a partir do ano 2000), devido ao início de uma grande explosão na geração de dados biológico. Conforme pode-se observar na Figura 17, a partir desta era, a quantidade de dados aumentou de forma exponencial e, hoje, já está próxima de ultrapassar a marca de 100 bilhões de pares de base.

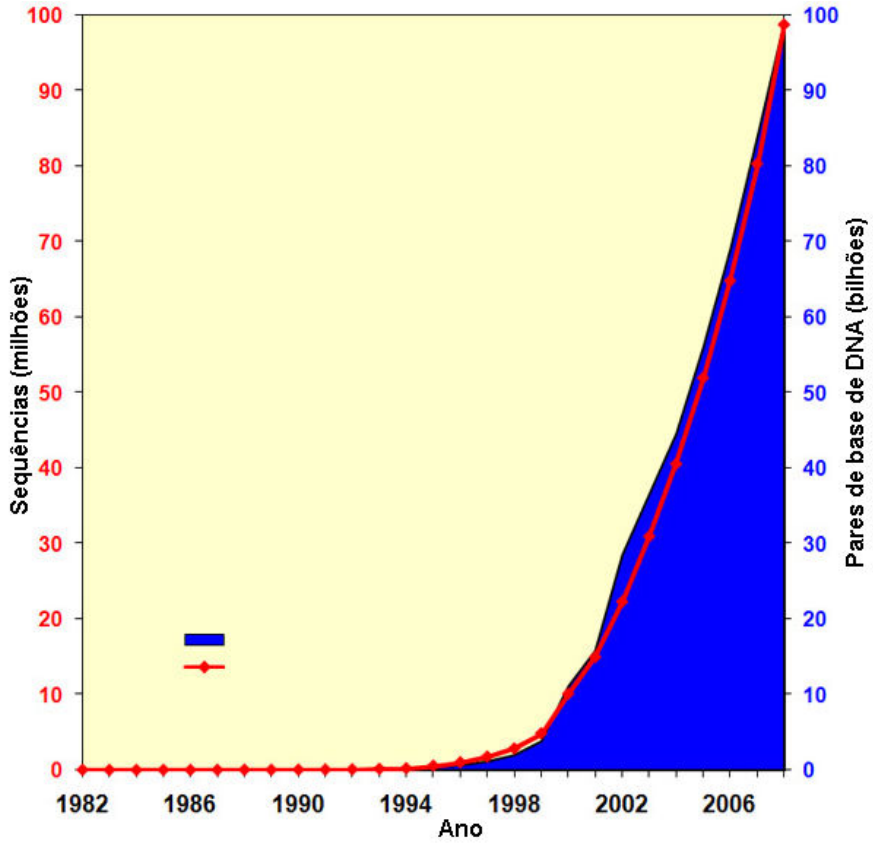


*Figura 17 – Sequências de dados biológicas no GenBank.*
*Quantidade de sequências armazenadas no GenBank, até o término do ano de 2008.*
*Fonte: NCBI - http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html*

Os bancos de dados biológicos foram desenvolvidos com base na teoria de banco de dados já amplamente difundida na área de ciência da computação e, para a área de bioinformática, foi necessário apenas adaptar a metodologia para um novo tipo de aplicação, entre eles para o armazenamento de SR. Para tal, as seções seguintes descrevem, de forma breve, os conceitos da modelagem relacional e alguns bancos de dados de SR existentes.

### 3.3.1 Modelo de especificação de bancos de dados

A construção de bancos de dados baseada em modelagem relacional é, atualmente, a mais empregada para representação de dados biológicos. Ela baseia-se em uma abordagem de análise e projeto *top-dow* hierárquica, conforme Figura 18 (Bornberg-Bauer & Paton, 2002), na qual se subdivide o projeto de um banco de dados em 3 etapas de projeto distintas (destacadas na figura): conceitual, lógica e física. A principal e mais importante delas é a conceitual, pois corresponde a uma aproximação do mundo real na forma de um diagrama, e que, com o uso de ferramentas computacionais, podem-se gerar rapidamente as demais fases envolvidas no projeto do banco de dados.

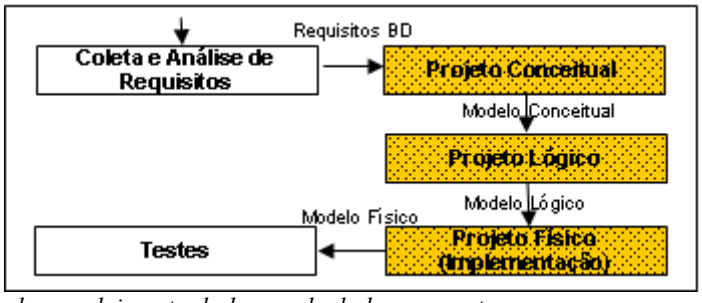


*Figura 18 – Fluxo de desenvolvimento do banco de dados proposto.*
*Fonte: da própria pesquisa.*

### 3.3.2 Banco de dados biológicos e sequências repetitivas

Em função da grande importância das SR, são fundamentais novas ferramentas e métodos que permitam que o entendimento destas sequências seja mais aprofundado. Além disso, é fundamental o uso de metodologias matemáticas e estatísticas que tentem relacionar a forma com que tais sequências estão organizadas, ou encontrar hipóteses confiáveis de algum tipo de distribuição, cujas informações serão muito úteis na tentativa de explicar alguns dos grandes desafios da ciência atual, na área de ciências biológicas e da saúde. Na genômica, por exemplo, grandes desafios são a anotação de todas as partes funcionais do DNA, entender como ocorre a regulação gênica, descobrir o papel de cada tipo de sequência repetitiva, descobrir quais as formas com que pode ocorrer a formação de um transcrito quimérico, entre outros.

Entretanto, primeiramente, é necessário que a busca e o tratamento das informações seja feita de forma clara e objetiva, sem que ocorra a remoção de trechos repetitivos de sequências genômicas, como fazem algumas propostas baseadas no mascaramento das mesmas (discutido na seção 3.2). Além disso, é também necessário considerar que o genoma, apesar da sua capacidade em gerar milhares de células distintas a partir de uma estrutura quimicamente muito simples, sofre constantes modificações, como aquelas comprovadas pelos *transposons* e pela ação de agentes externos, patógenos ou não, ou da própria evolução.

Em virtude disso, para que o uso de metodologias de análise e identificação de novos e prováveis comportamentos e/ou padrões biológicos seja viável, é imprescindível que as informações utilizadas estejam armazenadas e disponíveis por meio de um eficiente

sistema de armazenamento e acesso a dados, dada a imensidão de dados que são gerados por diversos laboratórios localizados em diferentes partes do mundo. Atualmente, já há alguns trabalhos que se propõem a fazer isso, que são baseados na criação de grandes bancos de dados públicos, como é o caso de um grupo de colaboração, que envolve três laboratórios em diferentes países: Japão (DDBJ), Estados Unidos (GenBank) e Europa (EMBL) (Figura 19) (Selzer et al., 2008). Cada um deles possui um sistema de gerenciamento de banco de dados (SGBD), cujo intuito é armazenar sequências anotadas, como transcritos (EST, RNA (ncRNA, tRNA, miRNA, iRNA, mRNA, snRNA)), e suas estruturas em um genoma. Tal colaboração corresponde a um grande sistema distribuído com serviços WEB que permitem que todos os dados de cada um dos SGBDs sejam acessados de maneira remota.

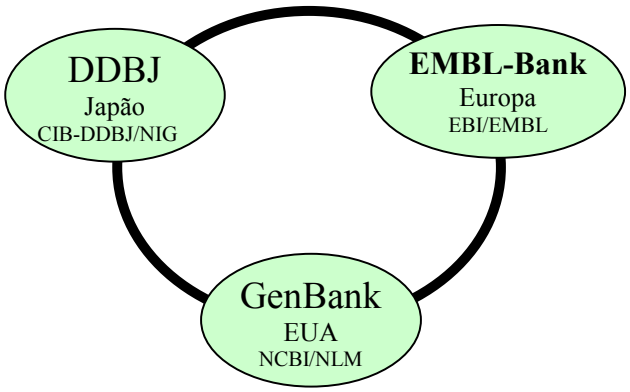


*Figura 19 – Bases de dados biológicas públicas.*
*Colaboração Internacional de bancos de dados públicos de sequências de nucleotídeos. DDBJ: DNA Data Bank of Japan. CIB-DDBJ: Center for Information Biology e DNA Data Bank of Japan. NIG: National Institute of Genetics. EBI: European Bioinformatics Institute. EMBL: European Molecular Biology Laboratory. NCBI: National Center for Biotechnology Information. NLM: National Library of Medicine. IAC: International Advisory Committee. ICM: International Collaborative Meeting.*
*Fonte: DDBJ - http://www.ddbj.nig.ac.jp/intro-e.html.*

Com a criação dos bancos de dados públicos, surgiram centenas de pesquisas que tiraram proveito das informações neles armazenados, e novas descobertas foram feitas a

respeito do funcionamento de diversos processos biológicos, tanto dos seres humanos quanto de outros seres vivos.

Além deste grande consórcio público, surgiram também dezenas de propostas de criação de BD com intuitos específicos. A seguir, são apresentadas algumas propostas de BD para armazenamento de SR.

Horng et al. (2003) propôs uma base de dados para armazenar repetições diretas, reversas complementares, e palindrômicas de sequências. Para tal base, há um sistema baseado em mineração de dados para detectar e analisar sítios de transcrição presentes nas SR armazenadas.

Loots & Ovcharenko (2007) propuseram o banco de dados ECRbase, cujo intuito é armazenar sequências repetitivas e outros tipos de sequências que estejam associadas com regiões conservadas, promotores, fatores de transcrição e pontos de ligação em genomas de vertebrados, todos associados com elementos funcionais dentro do genoma. Similarmente, FREPs (Nagashima et al., 2004) propôs um banco de dados de SR a partir de cDNA de ratos, os quais foram associados, com o uso de uma metodologia de bioinformática, com fatores de variação dos efeitos de transcrição, tradução, função da proteína ou envolvimento com doenças. Por este motivo, são chamadas de sequências repetitivas funcionais.

Outro banco de dados de sequências repetitivas, chamado Plant miRNA Database (PMRD) foi proposto por Zhang et al. (2010). PMRD integra dados de miRNA de plantas disponíveis em bancos de dados públicos, coletados da literatura e por laboratórios. O banco de dados contém informações de cada sequência, sua estrutura secundária, genes alvo, perfis de expressão e um sistema de navegação (genome browser). São 8433 miRNAs

coletados de 121 espécies de plantas no PMRD. Também em plantas, mas de forma mais ampla, RepPop (Zhou & Xu, 2009) foi desenvolvido para anotar os tipos repetição direta e repetição reversa complementar de sequência repetitiva da árvore *Populus trichocarpa*, sejam ou não funcionais.

Em outro trabalho, Grissa et al. (2007) propôs um banco de dados que armazena sequências repetitivas palindrômicas dispersas em genomas. Com o uso da ferramenta CRISPRFinder, as SR são automaticamente localizadas em novos genomas, e posteriormente inseridas em uma base de dados. O objetivo é mapear as sequências para construir um dicionário de sequências únicas, para permitir avaliar sua relação com algum tipo de fenômeno biológico. Le Flèche et al. (2001) desenvolveu um banco de dados de SR do tipo *tandem repeats*, a partir de genomas públicos de bactérias. Para garantir que as sequências possam ser analisadas por meio de procedimentos experimentais, o tamanho mínimo das sequências consideradas foi de 9 pb de comprimento. De forma similar, Jurka et al. (2005) propôs um banco de dados genérico para armazenar SR de vários eucariotos, incorporando sequências que foram anotadas e armazenadas por outras bases de dados.

Em Ruitberg et al. (2001), foi proposta a criação de um banco de dados de *tandem repeats* curtas para disponibilizá-las na Internet, chamado STRBase. Tal projeto, financiado pelo NIST (National Institute of Standards and Technology), é um repositório de informações para uso em genética forense, em que é armazenado um conjunto de pequenas *tandem repeats* que são comumente utilizadas para auxiliar as análises de processos judiciais. Além disso, o portal apresenta uma série de manuais que permitem a realização de experimentos práticos com vistas à comprovação de algum tipo de análise judicial que envolve análise de DNA.

Outro banco de dados foi proposto por Kennedy et al. (2007), chamado 3'-UTR SIRF (Short Interspersed Repeat Finder), o qual é especializado em armazenar SR da região 3´UTR de transcritos.

Chaparro et al. (2007), com o intuito de estudar a quantidade de elementos genéticos móveis (LTR *retrotransposons*) em arroz (*Oryza sativa*), propôs a criação do banco de dados chamado RetrOryza. Nele, são armazenados elementos genéticos móveis, abundantes ou não, com o intuito de permitir também que, por meio de uma interface WEB, novas sequências sejam anotadas por pesquisadores. Para humanos, Dagan et al. (2004) criou um mecanismo de BD para também armazenar sequências de elementos genéticos móveis de um tipo específico, as ALUs. Em tal banco, há também informações da região gênica de ocorrência de cada SR.

RNAiDB (Gunsalus et al., 2004) foi criado para armazenar, distribuir e analisar dados de fenotipagem a partir de análises em larga escala de RNAi em *Caenorhabditis elegans*. A base de dados contém um conjuto de dados associados com grandes bancos de dados públicos, e fornece informações a respeito de métodos experimentais e resultados de fenotipagem, incluindo dados originais na forma de imagens e vídeos. Por meio de ferramentas disponíveis em um sistema WEB, pode-se comparar intuitivamente resultados de ensaios de diferentes RNAi, e visualizar genes que possam sofrer interferência para cada RNAi. Além disso, há ferramentas de busca para diferentes tipos de características. Tipos similares de SR também são armazenadas pelos sistemas miRBase (Griffiths-Jones et al., 2006) e miRGen (Alexiou et al., 2010). O banco de dados miRBase apresenta uma interface integrada para análise e visualização de moléculas de microRNA, sua anotação e predição de sites em genes. O BD propõe um mecanismo que associa os registros de cada microRNA

com 3 pontos principais: atuar como um sistema para padronizar a nomenclatura e a anotação de cada microRNA descoberto, de forma que os nomes sejam atribuídos antes da publicação; repositório primário para anotação e dados de miRNA; e um novo banco de dados de genes com predições de regiões alvo de miRNA. miRGen 2.0 é uma das mais recentes propostas de banco de dados de sequências repetitivas específicas da literatura. Ele propõe anotar moléculas de microRNA de ratos e humanos, e fornecer informações precisas a respeito de sua posição em tais organismos. O intuito é anotar também as regiões regulatórias associadas a fatores de transcrição, incluindo dados preditos e experimentais. Outras informações associadas aos miRNA, como perfis de expressão em tecidos e linhagens celulares, SNPs, regiões alvo em genes codificantes de proteínas de alguma rede metabólica também são integradas na base de dados e podem ser acessadas por meio de uma interface gráfica.

Como pôde ser claramente observado, os bancos de dados para anotação de sequências repetitivas são construídos com fins específicos. Por este motivo, na maioria dos casos, não são fornecidas pelas bases de dados públicas disponíveis até então, sendo acessíveis apenas por meio de arquivos do tipo texto ou em algum outro tipo de formato similar (fasta, gff, genbank etc.), cuja manipulação requer tempo e alto custo de processamento.

## 3.4 Considerações do capítulo

Após a descoberta de que SR participam da regulação gênica, como em RNAi, ocorreu um grande aumento nos investimentos associados ao estudo e influência das SR

dentro dos genomas dos seres vivos. Desta forma, já foram observadas fortes evidências que associam as SR com outros fenômenos biológicos dentre os mais interessentes, como o processamento de RNA por *cis-splicing* e, inclusive, para a formação de transcritos quiméricos que, por serem raros, faltam ainda metodologias que permitam gerar novos candidatos para que novos estudos possam ser realizados nesta mesma direção. Dada a importância biológica das SR, este capítulo apresentou uma classificação das SR e descreveu diversas abordagens que tratam da sua detecção, anotação e armazenamento em sistemas de bancos de dados. Em função disso e do que foram discutidos nos capítulos anteriores, os próximos capítulos descrevem, de forma clara e suscinta, todas as etapas consideradas para a realização dos objetivos da seção 1.4, na qual são propostas metodologias e ferramentas de bioinformática, e um conjunto de experimentos *in silico* para demonstrar sua usabilidade e importância na análise de dados de genomas e transcriptomas

# Capítulo 4 Desenvolvimento de metodologias e de ferramentas de bioinformática

Os capítulos anteriores apresentaram uma breve revisão da literatura, com a descrição de alguns trabalhos relacionados com a importância das sequências repetitivas (SR) na regulação gênica de organismos inferiores e superiores. Devido à comprovada importância das SR, este capítulo apresenta de forma detalhada um conjunto de metodologias de bioinformática para detecção e estudo da relação das SR no *locus* gênico de transcritos quiméricos, além de outras que foram construídas para dar suporte ao seu uso. Tais metodologias, citadas na seção 1.5 do Capítulo 1, foram propostas nesta tese para construir novas ferramentas de bioinformática voltadas para a análise de dados de genomas e transcriptomas. Cada uma das metodologias é descrita nas seções seguintes.

## 4.1 Detecção de transcritos quiméricos e sua relação com SR

A identificação de sequências repetitivas (SR) associada ao fenômeno de formação de transcritos quiméricos requer que, antes, sejam encontrados transcritos quiméricos, particularmente aqueles formados por trechos oriundos de cromossomos diferentes e que não sejam explicados por rearranjos genômicos. Para tal, foram desenvolvidas duas metodologias *in silico* para detecção de evidências de tais tipos de transcritos. A primeira metodologia Fusion5Finder (Herai & Yamagishi (a), 2009) foi desenvolvida especialmente para encontrar instâncias similares àquelas reportadas pelos experimentos em humanos,

cuja 5'UTR veio de um cromossomo diferente do restante do transcrito. A segunda metodologia, FusionAllFinder (Herai & Yamagishi (a), 2010), é uma extensão da primeira. Para ambas as metodologias, três pontos são importantes: o primeiro é o banco de dados utilizado, o segundo é a adoção de uma estratégia de filtragem dos dados, e o terceiro corresponde à definição da metodologia propriamente dita. O intuito das metodologias é permitir que dados de genomas e transcriptomas sejam integrados e gerem informações que orientem futuros trabalhos experimentais que comprovem ou refutem as evidências encontradas *in silico*.

### 4.1.1 Escolha do banco de dados de transcritos

Para que os dados fornecidos pelas metodologias sejam minimamente confiáveis, é importante, dentre vários fatores, o uso de sequências de dados biológicas geradas por meio de uma técnica de sequênciamente do tipo *full-length* cDNA, ao invés do uso de EST, cujas sequências costumam ser menores, mais propensas a erros e mais redundantes, conforme discutido no Capítulo 2.

### 4.1.2 Filtragem dos dados

O intuito é criar uma estratégia de filtragem para identificar, em ambas as metodologias, evidências de transcritos quiméricos, possivelmente geradas por meio de *trans-splicing* intercromossomal em bases de dados de transcritos. Para tal, parte de um transcrito deve mapear em um único cromossomo (denotado neste trabalho por região TSR: *trans-spliced region*), e o restante em um cromossomo distinto (como nas melhores

evidências experimentais), também único, para evitar dificuldades na interpretação dos resultados. A região TSR pode corresponder a uma pequena parte do transcrito, dada pela região 5'UTR, ou mesmo a uma região maior, que englobe também parte da região codificante (CDS). Ela pode também corresponder somente à região 3'UTR, conforme ilustração da Figura 20. A primeira metodologia, Fusion5Finder, busca apenas por sequências com a estrutura da Figura 20 (a), cuja 5'UTR e o restante do transcrito mapeiam-se em cromossomos distintos. A segunda metodologia, FusionAllFinder, considera todos os casos da Figura 20 (a, b, c). Além disso, o processo de filtragem deve levar em consideração que os transcritos candidatos sejam oriundos de tecidos normais, visto que, conforme a literatura, o rearranjo cromossômico é comum em tecidos oriundos de câncer, enfraquecendo a hipótese de que os quiméricos sejam gerados por *trans-splicing*, observação esta feita por Li et al. (2008), na melhor evidência de formação de um transcrito quimérico em organismos superiores (humano).

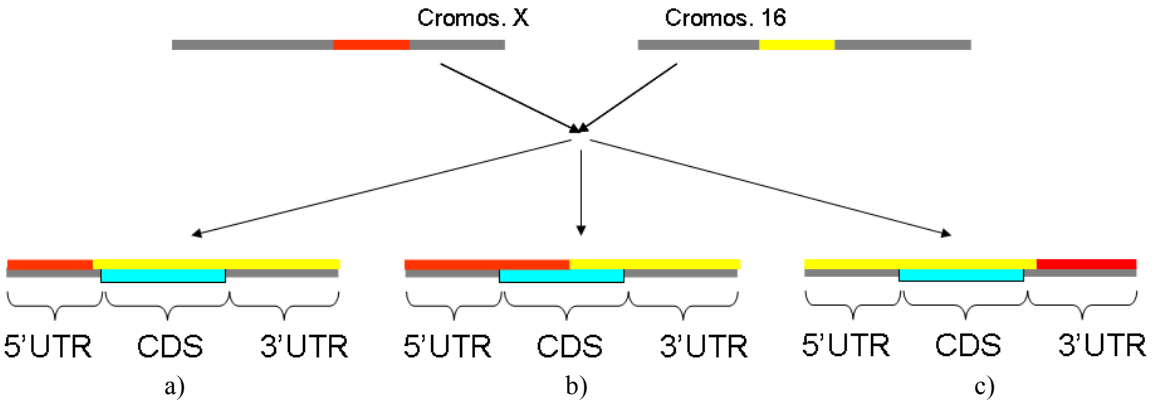


*Figura 20 – Regiões de fusão entre transcritos.*
*Estratégia de busca por transcritos candidatos, formados por* trans-splicing *intercromossomal. Um transcrito candidato é formado pela fusão de outros dois transcritos. A região em amarelo é chamada TSR, e a região em vermelho, corresponde ao gene propriamente dito. A TSR pode incluir 3 partes do gene: a) a região 5'UTR; b) a região CDS; e c) a região 3'UTR.*
*Fonte: da própria pesquisa.*

### 4.1.3 Metodologia

A metodologia proposta para detecção de evidências por *trans-splicing* é ilustrada na Figura 21. Inicialmente, os transcritos da base de dados de FLcDNA escolhida são filtrados (HInvFilter) para criar um arquivo em formato fasta FLcDNA_FILT que contém somente aquelas sequências com informações completas de 5'UTR, CDS, e 3'UTR, coletadas de células de tecidos normais. Para o caso da primeira metodologia, conforme Figura 21 (a), são criados 3 arquivos distintos FLcDNA_FILT_5UTR, FLcDNA_FILT_CDS, e FLcDNA_FILT_3UTR para as sequências de 5'UTR, CDS, e 3'UTR respectivamente. Utilizando NCBI-BLAST, mapeiam-se todas as sequências 5'UTR, CDS, e 3'UTR no genoma de referência (parâmetros do BLAST são detalhados na figura e confirmados com o uso da ferramenta BLAT (Kent, 2002)). O resultado é então armazenado em 3 arquivos de alinhamento em formato XML. Os dados do arquivo XML correspondente à região 5'UTR foram cruzados (GeneAlignmentAnalyser) com os dados dos outros dois arquivos em XML (CDS e 3'UTR) para reter somente aquelas sequências, cujas TSR foram localizadas em um único cromossomo. Isto é feito para evitar ambiguidades na interpretação dos dados. Somente estes mRNAs híbridos foram armazenados no arquivo TransSpl.

Para o caso da segunda metodologia, FusionAllFinder (Herai & Yamagishi (a), 2010), conforme Figura 21 (b), é criado um arquivo FLcDNA_FILT para as sequências da base de FlcDNA. Utilizando NCBI-BLAST, mapeiam-se todas as sequências de transcritos no genoma de referência (parâmetros do BLAST são detalhados na Figura 21 e verificados com o uso da ferramenta BLAT (Kent, 2002)). O resultado é então armazenado em um arquivo de alinhamento em formato texto. O dados de tal arquivo são analisados

(GeneAlignmentAnalyser) e somente os RNAs quiméricos candidatos são armazenados no arquivo TransSpl. Estes correspondem aos candidatos de *trans-splicing* para posterior confirmação experimental.

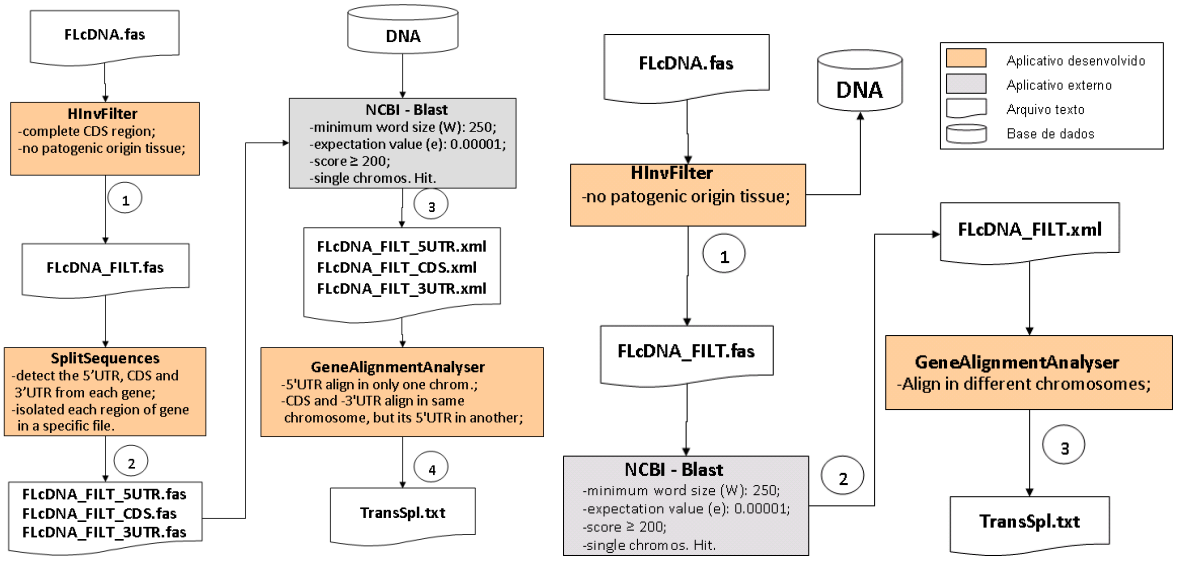


*Figura 21 – Metodologias para detecção de transcritos quiméricos.*
*a) Metodologia Fusion5Finder (Herai & Yamagishi, 2010); b) Metodologia FusionAllFinder (Herai & Yamagishi, 2010).*
*Fonte: da própria pesquisa.*

É importante destacar que, apesar da alta similaridade entre as duas metodologias, a primeira foi desenvolvida especialmente para encontrar transcritos quiméricos em genomas com montagem de alta qualidade, como de humanos. Além disso, ela busca por candidatos que sejam similares às melhores evidências reportadas em organismos superiores, como aquelas encontradas em humanos, em que parte ou toda a região 5'UTR de um transcrito mapeia em um cromossomo distinto do cromossomo em que são mapeadas as regiões CDS e 3'UTR. A segunda metodologia não parte da premissa de que as sequências quiméricas sejam similares às melhores evidências experimentais e, desta forma, permite que a região TSR inclua, de forma exclusiva ou não, as regiões CDS ou 3'UTR. Além disso, outro

grande diferencial é que tal metodologia permite a análise de genomas de baixa qualidade, em versões ainda *"draft"*, evitando, assim, a geração de uma grande quantidade de candidatos falso-positivos provenientes de partes do genoma que ainda não foram montados ou que apresentam possíveis erros de montagem.

### 4.1.4 Implementação

As duas metodologias fazem uso de ferramentas gratuitas como parte do mecanismo para detecção de candidatos de transcritos quiméricos gerados por *trans-splicing*.

As implementações das duas metodologias são versões com aplicações em linha de comando, que permitem analisar de forma automática um banco de dados inteiro na busca por prováveis candidatos de transcritos quiméricos. As ferramentas foram desenvolvidas usando uma metodologia orientada a objetos e a linguagem de programação Java, visando facilitar sua portabilização para um sistema WEB, e utilização em diferentes sistemas operacionais. O uso de outras linguagens de programação, como Perl, Python ou mesmo .Net da Microsoft também poderiam ser considerados, ou inclusive a integração com ferramentas prontas e funcionais em tais linguagens de programação citadas.

## 4.2 Metodologia para detecção de SR em *locus* gênico de sequências

Sequências repetitivas (SR) em *locus* gênico de animais e plantas podem estar envolvidas em diversos fenômenos biológicos, como interferência por RNAi (Fire et al., 1998) e *trans-splicing* (Di Segni et al., 2008). A maioria das SR localiza-se em regiões de

íntrons e, usualmente, seus estudos envolvem apenas sequências de mRNA, sendo necessário a estrutura éxon-íntron dos genes. Desta forma, é preciso realizar um mapeamento dos genes em seu respectivo genoma de referência para identificação de quatro tipos de SR: repetição direta (RD), repetição reversa (RR), repetição complementar (RC) e repetição reversa complementar (RRC). Embora tais tipos de sequências sejam facilmente descritas, sua implementação computacional requer o conhecimento e uso de algoritmos de bioinformática.

### 4.2.1 Metodologia

A metodologia empregada para a detecção e visualização de SR em lócus gênico é apresentada na Figura 22, na qual é ilustrado o fluxo de dados do núcleo da aplicação, sua relação com os aplicativos externos que foram utilizados para dar suporte à metodologia, e um formato fictício de visualização das SR presentes entre dois transcritos.

Conforme a Figura 22, inicialmente o usuário deve definir duas sequências de transcritos de entrada e associá-los com um genoma de algum organismo (Figura 22 (a)). Com o uso da ferramenta GMAP (Wu & Watanabe, 2005) e de um módulo chamado GenomeTools, os mRNAs devem ser mapeados no respectivo genoma selecionado para obtenção da estrutura éxon-íntron dos transcritos de entrada (Figura 22 (b)), e também suas respectivas sequências genômicas (Figura 22 (c)). Ao final, utilizando a ferramenta NCBI-Bl2Seq do pacote BLAST (Figura 22 (d)), os 4 tipos de SR são mapeados e cada uma delas é analisada por um módulo chamado AlignRepeats (Figura 22 (e)), que faz uso de um outro módulo chamado DrawAlignments (Figura 22 (f)). Tal módulo gera uma representação

gráfica das duas sequências de entrada e podem ser visualizadas por meio de um navegador WEB, com as estruturas éxon-íntron coloridas e os 4 tipos de SR, na forma de linhas com cores distintas, que associam os transcritos de entrada.

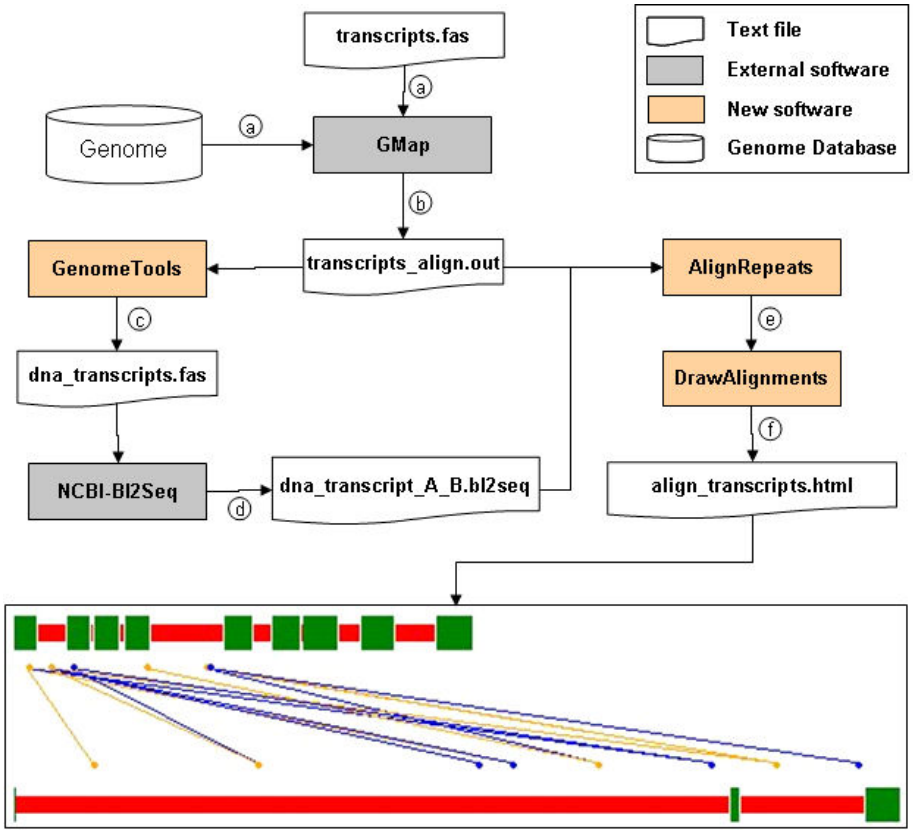


*Figura 22 – Metodologia RepGraph.*
*Metodologia de identificação e exibição de SR entre dois transcritos com o uso de ferramentas de bioinformática.*
*Fonte: da própria pesquisa.*

### 4.2.2 Implementação

Com o intuito de verificar a possibilidade de mediação da formação de sequências quiméricas por meio de SR, foi desenvolvida, com base na metodologia apresentada pela Figura 22, uma ferramenta chamada RepGraph (Herai & Yamagishi (b), 2010), para detecção de SR em *locus* gênico entre os transcritos quiméricos encontrados.

O desenvolvimento da ferramenta de visualização, RepGraph, baseia-se no uso de ferramentas gratuitas e em um modelo de desenvolvimento de software para a WEB cuja arquitetura do sistema é subdividida em 3 camadas distintas: interface WEB, regras de negócio e camada de persistência ou banco de dados. A interação entre tais camadas é feita por um modelo conceitual chamado MVC, que é composto por um componente Controlador, responsável por coordenar a comunicação entre as 3 camadas do modelo (Cavalcanti et al., 2005). Considerando tal arquitetura e modelo de interação, o sistema foi implementado com o uso da linguagem de programação Java v1.6 e com o framework MVC Apache Struts. A codificação foi feita com o uso da ferramenta Eclipse Ganymede, uma IDE Java com suporte ao desenvolvimento de aplicações WEB. Por se tratar de um sistema WEB, a ferramenta é armazenada em uma máquina com o servidor de aplicações Tomcat v6.0, responsável por tratar conexões de usuários. Recursos externos, como aplicações (GMap e NCBI-Bl2Seq) e bases de dados de genomas são definidos no sistema por meio de arquivos de configuração para facilitar a atualização de tais recursos externos, caso novas versões dos mesmos sejam disponibilizadas. Na Figura 23 é apresentada a interface WEB do sistema RepGraph.

O procedimento de obtenção das representações gráficas pode ser feito de forma simples e segue estritamente a metodologia descrita aqui. Inicialmente o usuário fornece duas sequências de entrada (em formato *fasta*, número de acesso AC ou na forma de um arquivo) e seleciona um genoma de interesse para que as sequências sejam mapeadas. Em seguida, é exibida uma lista de regiões genômicas em que cada sequência foi mapeada. Uma região correspondente dentre várias, caso o alinhamento não seja único, deve ser selecionada para cada uma das sequências.

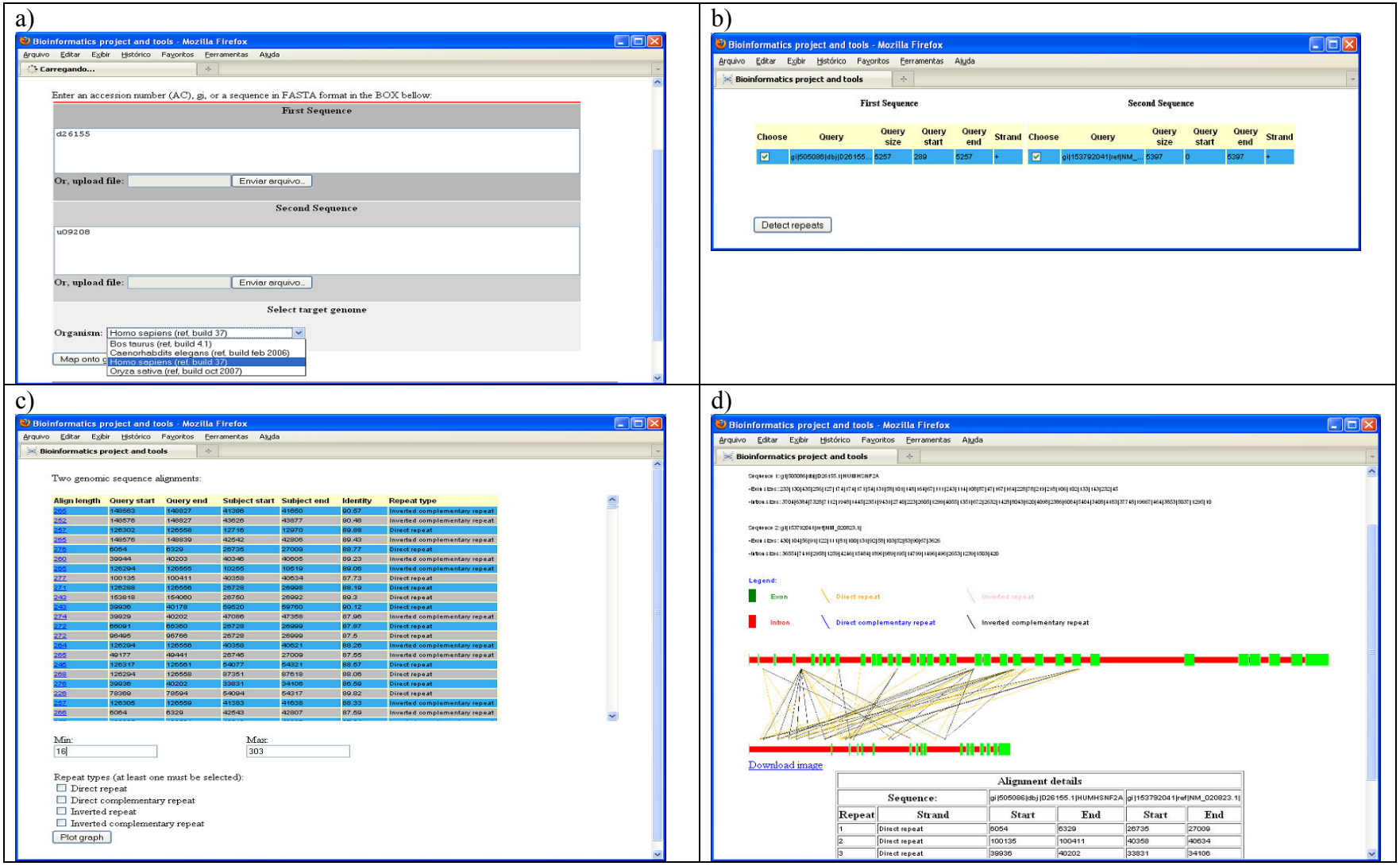


*Figura 23 - Telas do portal WEB RepGraph.*
*RepGraph ilustrando o uso da ferramenta para mapeamento, detecção e criação de imagens em sequências genômicas com SR entre transcritos distintos: a) Entrada das sequências; b) resultado do mapeamento de cada sequência em seu respectivo genoma de interesse; c) pares de sequências repetitivas detectadas e seleção dos tipos que serão ilustrados graficamente; d) ilustração gráfica dos pares de SR selecionados na etapa anterior em seu respectivo lócus gênico.*
*Fonte: da própria pesquisa.*

Na etapa seguinte, as SR dos 4 tipos mencionados são exibidas. O usuário pode selecionar o tipo de SR e especificar um intervalo de tamanho mínimo e máximo das sequências que serão exibidas pela figura.

RepGraph permite analisar SR que são mapeadas na estrutura éxon-íntron de um gene, e pode ajudar na identificação de novos padrões como quantidade, tamanho e região de ocorrência de cada tipo de SR. Além disso, a ferramenta também pode ser atualizada e aplicada em genomas de organismos que ainda não foram sequenciados.

## 4.3 Estudo de SR em lócus gênicos

O estudo da frequência de ocorrência de um tipo de SR é importante, por exemplo, para verificar se há predominância de um tipo ou regiões gênicas envolvidas na formação de transcritos quiméricos. Tais estudos são motivados, além das evidências biológicas já apresentadas, pelo fato de que a probabilidade da ocorrência de pares de SR de tamanho relativamente pequeno, na ordem de 20 pb, ser altamente improvável (1 ocorrência esperada em uma sequência de tamanho 20 é igual a $4^{20}$=1.099.511.627.776), supondo probabilidades iguais para os quatro tipos de nucleotídeos (A,C,G, T) e desconsiderando a existência de SR. Porém, para um genoma real de um organismo qualquer, já é fato que são repletos de sequências repetitivas, conforme revisão da literatura do Capítulo 3 e, desta forma, a probabilidade de se encontrar SR, se for comparada com um genoma fictício e gerado totalmente de forma aleatória, é maior. Tal observação é importante, caso contrário seria praticamente impossível encontrar (considerando um genoma de tamanho razoavelmente grande, do humano por exemplo, cujo tamanho estimado é de um pouco

menos de 3.5 bilhões de pares de base) um par de SR de tamanho 20, mesmo que entre o par de SR exista uma pequena proporção de *gaps*, *indels* ou até mesmo *mismatches*, inferiores a 10%.

Em função disso, na próxima seção é apresentada uma metodologia, chamada FreqRepeat, para estudar e analisar a frequência de SR que ocorrem entre diferentes tipos de pares de sequências de nucleotídeos, reais ou fictícios. O objetivo é verificar se algum tipo, principalmente as com poder de auto-ligação como as reversas complementares (RRC), é exclusiva de transcritos envolvidos na formação de sequências quiméricas. Neste caso, considera-se que um transcrito quimérico foi formado pela fusão de duas sequências (TQ e ET), que precisam, necessariamente ser transcritas para que a fusão possa ter ocorrido. A primeira (TQ) está anotada no mesmo *locus* gênico do transcrito quimérico, e inclusive é anotada com o mesmo nome. A segunda (ET) deve ser encontrada e, para tal, mapeia-se a TSR no genoma de referência do transcrito para verificar qual sequência está anotada na mesma região, o que representa uma evidência de transcrição da região TSR.

### 4.3.1 Metodologia

A metodologia FreqRepeat proposta para o estudo da frequência das SR em regiões gênicas de transcritos quiméricos envolve duas partes. A primeira permite detectar diferenças significativas de frequência entre um par (TQ,ET), formado por um candidato quimérico (TQ) e sua evidência de transcrição (ET) da região TSR, com outros três tipos de pares:

a) par (GR,GR) formado por duas sequências oriundas de regiões genômicas a partir de uma posição definida de forma aleatória;

b) par (TR,GR) formado por uma sequência TR do *locus* gênico de um transcrito, e uma sequência GR de uma região genômica selecionada aleatoriamente;

c) par (TR,TR) formado por duas sequências oriundas do *locus* gênico de dois transcritos distintos selecionados aleatoriamente de uma base de transcritos.

A segunda parte envolve a determinação dos tipos de SR que são detectados pela metodologia, que podem ser: repetição direta (RD), repetição reversa (RR), repetição complementar (RC) e repetição reversa complementar (RRC). É importante destacar que a busca é sempre feita aos pares, de forma que as sequências do par repetitivo apareçam em sequências de nucleotídeos distintos, conforme ilustração da Figura 24.

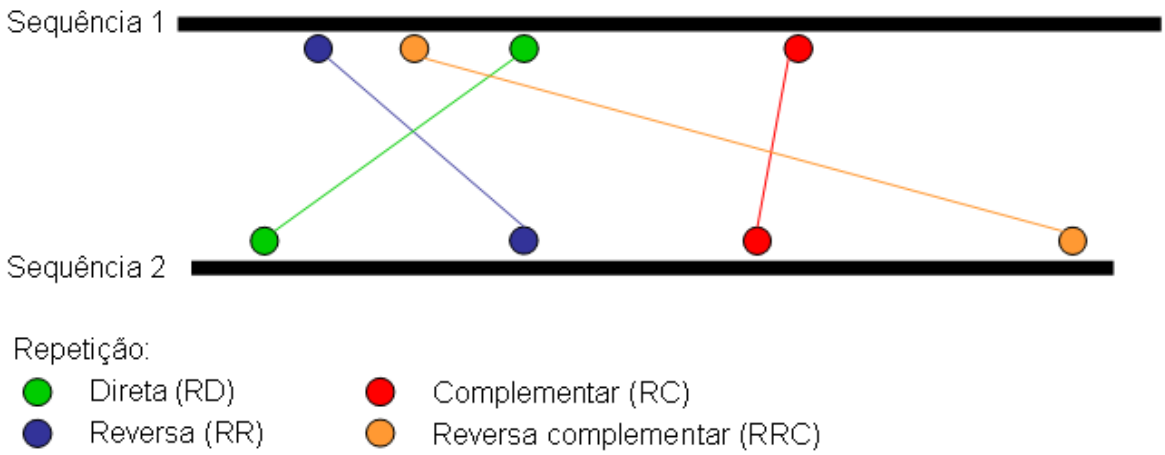


*Figura 24 – Tipos de pares de sequências repetitivas.*
*Fonte: da própria pesquisa.*

A comparação visa verificar se, entre os pares (TQ,ET) de sequências envolvidas na formação de um transcrito quimérico existe algum padrão de tipo ou tamanho de SR que não ocorre em outros tipos de pares. Esta metodologia é apresentada na Figura 25, e será explicada junto com a implementação, na próxima seção.

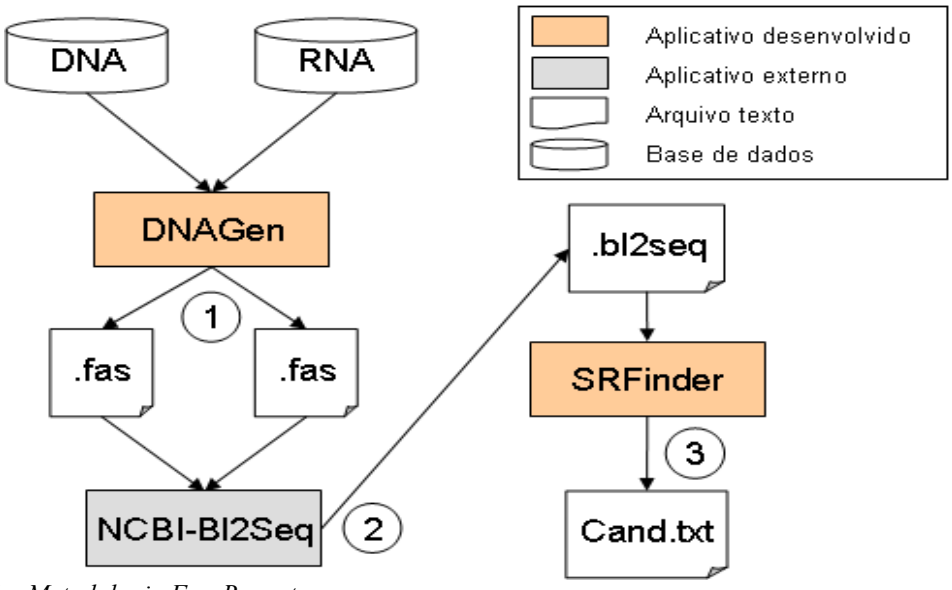


*Figura 25 – Metodologia FreqRepeat.*
*FreqRepeat: análise da frequência de ocorrência de SR entre duas sequências, reais ou fictícias.*
*Fonte: da própria pesquisa.*

### 4.3.2 Implementação

A implementação da metodologia FreqRepeat subdividiu-se em 3 partes. Na primeira, foi desenvolvida uma aplicação, DNAGen (Figura 25 (1)), que, a partir de duas sequências especificadas, gera (sequência fictícia) ou obtém (sequência real) suas respectivas sequências de nucleotídeos. Na segunda, as sequências são alinhadas entre si com a ferramenta externa NCBI-Bl2Seq (Figura 25 (2)), para que seja feita a detecção de pares de SR de tamanhos maiores do que 10 pb, e com cobertura mínima de 90%. O resultado é então armazenado em um arquivo no formato texto.

Na terceira parte, foi desenvolvida uma ferramenta, SRFinder (Figura 25 (3)), para filtrar os dados gerados pela segunda ferramenta. Para permitir que sejam portabilizadas para sistemas WEB e modificadas em ocasiões futuras, as ferramentas foram desenvolvidas em metodologia orientada a objetos, e com o uso da linguagem de programação Java. A

versão atual é em linha de comando, e os parâmetros devem ser informados no momento da execução de cada uma delas.

## 4.4 Banco de dados biológico para armazenamento de SR

Embora hoje já existam inúmeros outros trabalhos de criação de bancos de dados de biologia molecular para aplicações em bioinformática, como aqueles baseados na integração de fontes de dados, armazenamento específico de dados científicos, definição de protocolos e linguagens de acesso aos dados, modelos de bancos de dados *ad-hoc* e definição de transações e sistemas de *work-flow* (Davidson et al., 1995; Bornberg-Bauer & Paton, 2002; Buneman et al., 2004; Lemos, 2004; Cavalcanti et al., 2005), a grande maioria é destinada a trabalhos com objetivos específicos, o que torna seu acesso ainda muito limitado, causando seu desuso. Além disso, devido à inexistência de um SGBD específico para aplicações de Bioinformática, muitas das ferramentas criadas acessam dados diretamente de arquivos textos ou binários sem a utilização de um SGBD, o que os impede de beneficiar-se de mecanismos eficazes de armazenamento, acesso e segurança de dados fornecidos por tais sistemas.

Em função disso, foi proposto um banco de dados, chamado Rep4DB, cujo objetivo é armazenar informações das sequências repetitivas presentes em um genoma ou transcriptoma, para posteriormente acessá-las. Desta forma, as seguintes informações são fundamentais para esclarecimento dos requisitos do banco de dados proposto, que deve permitir igualmente o tratamento de diferentes genomas, tanto de organismos superiores quanto de inferiores:

— um genoma deve possuir informações relativas à sua espécie, gênero, reino, descrição, versão, ano de publicação e consórcio (grupo) que foi responsável pelo seu sequenciamento;

— um genoma é formado por um ou vários cromossomos, que possuem informações relativas ao seu nome, a localização do arquivo que possui sua correspondente sequência e uma breve descrição. Um cromossomo pode também representar um DNA plasmidial de bactérias, um DNA mitocondrial de organismos eucariotos, ou simplemente o material genético (DNA ou RNA) de vírus;

— um cromossomo, por sua vez, pode gerar diferentes transcritos primários, sendo estes formados por estruturas conhecidas como éxons e, não obrigatoriamente, íntrons. Ambos apresentam posições específicas no transcrito primário, e aparecem sempre de forma alternada;

— um transcrito primário, para que seja gerado a partir do DNA, possui alguns elementos que são importantes para iniciar o processo de transcrição. Entre eles estão os agentes promotores, os ativadores e os silenciadores que estão associados com tal processo;

— um transcrito primário, quando formado por éxons e íntrons, passa por um processo de remoção destes últimos, e posterior ligação dos éxons adjacentes, gerando um transcrito maduro;

— um transcrito maduro ou produto, pode corresponder a um gene, ou a outro tipo de molécula, como ncRNA, RNAi, snRNA, entre outros. Tal produto possui também informações relativas ao seu código de acesso em uma base de dados pública, sua

respectiva versão, uma breve descrição, sua sequência de nucleotídeos e sua orientação no genoma;

— um produto pode ainda ser formado por algumas partes específicas, como região codificante (CDS) e regiões não traduzíveis (5'UTR e 3'UTR), regiões específicas para anelamento de *primer* para uso em experimentos de reação de polimerização em cadeia (PCR). Além disso, um produto pode também ser associado a um cluster de genes, que compartilham similaridades em suas respectivas sequências e, consequentemente, em suas funções metabólicas;

— um produto, se traduzido, pode gerar uma proteína. Ambos podem ser de tipos específicos de compostos que, dada uma concentração, reagem entre si por um tempo determinado, por meio de alguma via metabólica. Cada via metabólica pode fazer parte de um grupo de reações ainda maior, chamado de rede metabólica. Tal informação é importante porque permite criar associações entre diferentes produtos e proteínas;

— uma sequência, seja do tipo transcrito primário ou produto, pode também ser anotado com outras informações, que permitem associá-las com alguma doença, algum polimorfismo referente a um único nucleotídeo (SNP) ou com alguma particularidade relacionada a uma sequência repetitiva;

— uma anotação deve possuir informações que indiquem sua localização, uma breve descrição e sua orientação na sequência, seja ela de um cromossomo, um transcrito primário ou de um produto. Além disso, deve ser possível associar uma sequência repetitiva com informações relativas à sua localização, família a que pertence a sequência repetitiva, descrição, orientação com relação à sequência de origem. As SR

que podem ser associadas a uma única sequência alvo, independentemente da quantidade, podem ser do tipo: repetição direta, repetição complementar, repetição reversa ou repetição reversa complementar;

— um genoma, um cromossomo ou um produto costumam ter suas informações publicadas em artigos científicos ou livros. Em função disso, informações referentes ao: título, ano de publicação, volume, edição, descrição e meio de impressão devem ser armazenadas;

— tanto produto quanto genoma ou cromossomo podem ter informações do laboratório responsável pelo seu sequenciamento, tais como: localização na WEB, nome do laboratório e país;

A descrição de cada uma das relações e das informações pertinentes ao banco de dados biológico proposto corresponde aos requisitos necessários para sua especificação na forma de modelos, que serão resumidamente descritos nas seções seguintes.

### 4.4.1 Projeto do banco de dados

Esta seção aborda as 3 etapas envolvidas na fase de projeto de um banco de dados baseado em uma metodologia relacional, que subdivide-se em: modelagem conceitual, modelagem lógica e modelagem física. Considerando os requisitos do banco de dados Rep4DB, cada uma das etapas da metodologia considerada é apresentada a seguir.

#### 4.4.1.1 Modelagem conceitual

A modelagem conceitual fez uso das informações coletadas durante o processo de levantamento de requisitos do sistema e produziu um modelo do tipo entidade relacionamento. Nele, como propõe o modelo conceitual, são ilustradas as relações entre todas as entidades do banco de dados, bem como seus relacionamentos e multiplicidades por meio de um modelo de entidades e relacionamentos (MER).

Para sua concepção inicial, o projeto do modelo conceitual foi subdividido em 4 partes principais com o intuito de permitir armazenar e, posteriormente, manipular as seguintes estruturas biológicas:

— transcrito primário e suas partes: éxons, íntrons, região promotora, silenciadores e ativadores (Figura 26);

— produto (genes, RNA não codificante, entre outros) resultante de uma transcrição e suas partes: regiões não traduzíveis (UTRs), região codificante (CDS) e cluster a que pertence (Figura 27);

— redes e vias metabólicas: reações e compostos que podem estar relacionados com prováveis proteínas, geradas pelos seus respectivos transcritos (Figura 28);

— anotações associadas com produtos, transcritos primários e cromossomos: sequências repetitivas, *primers*, polimorfismos de um único nucleotídeo e doenças (Figura 29);

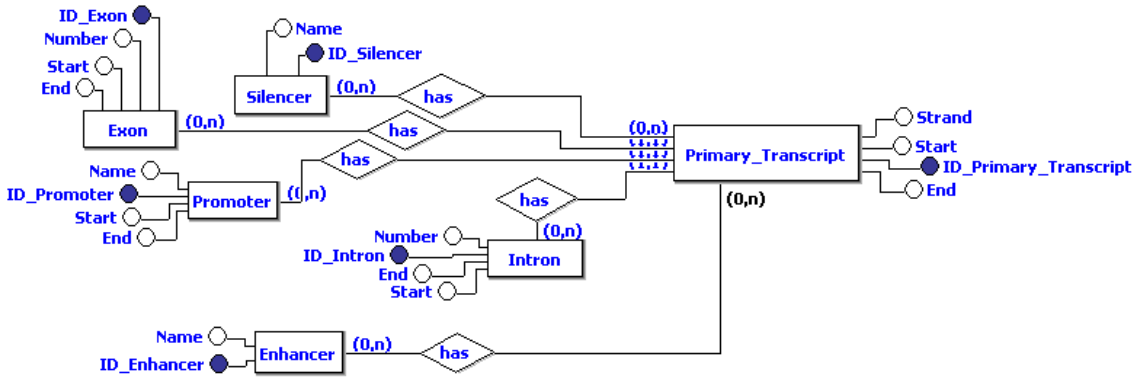


*Figura 26 – MER das tabelas representativas de um transcrito primário.*
*MER das tabelas representativas de um transcrito primário e suas respectivas partes.*
*Fonte: da própria pesquisa.*

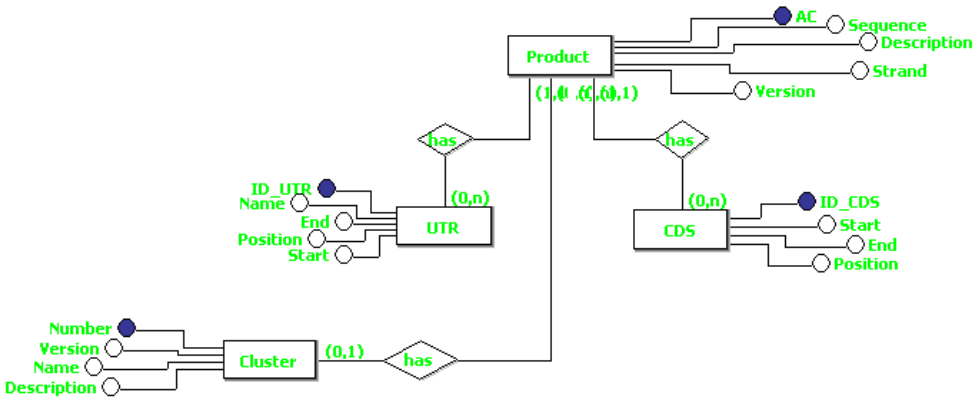


*Figura 27 – MER das tabelas representativas de um produto de transcrição.*
*MER das tabelas representativas de um produto de transcrição e suas respectivas partes.*
*Fonte: da própria pesquisa.*

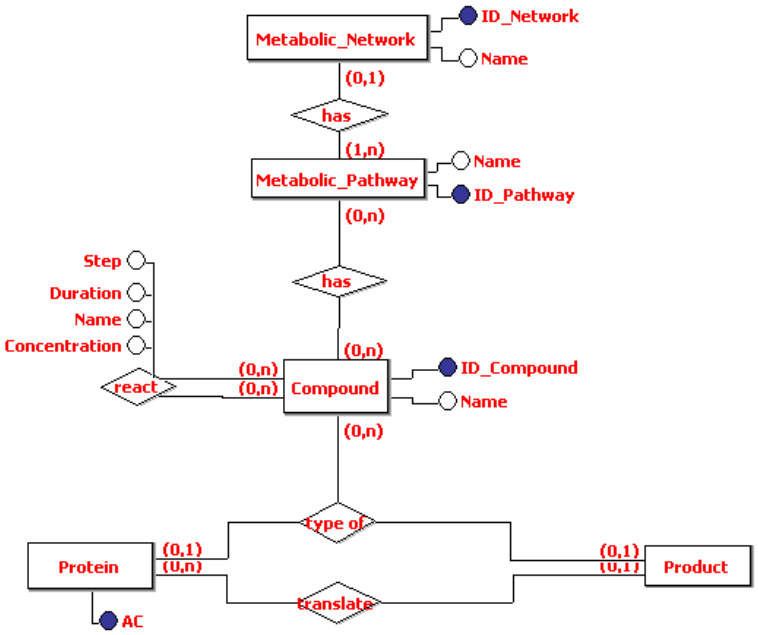


*Figura 28 – MER das tabelas representativas das redes e vias metabólicas.*
*Fonte: da própria pesquisa.*

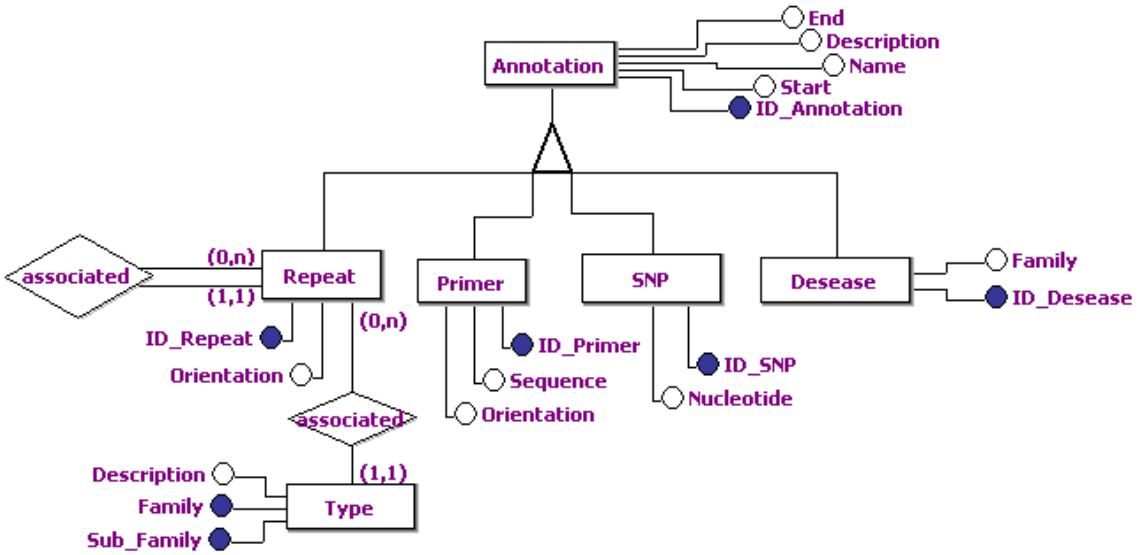


*Figura 29 – MER das tabelas para anotação de sequências.*
*MER das tabelas referentes aos tipos de anotação para alguma região de um cromossomo, transcrito primário, transcrito e produto de transcrição.*
*Fonte: da própria pesquisa.*

O modelo completo, conforme Figura 30, possui outras tabelas destacadas nas cores rosa e azul piscina, que complementam o diagrama entidade-relacionamento com informações relativas a estrutura do genoma, e a origem dos dados, respectivamente.

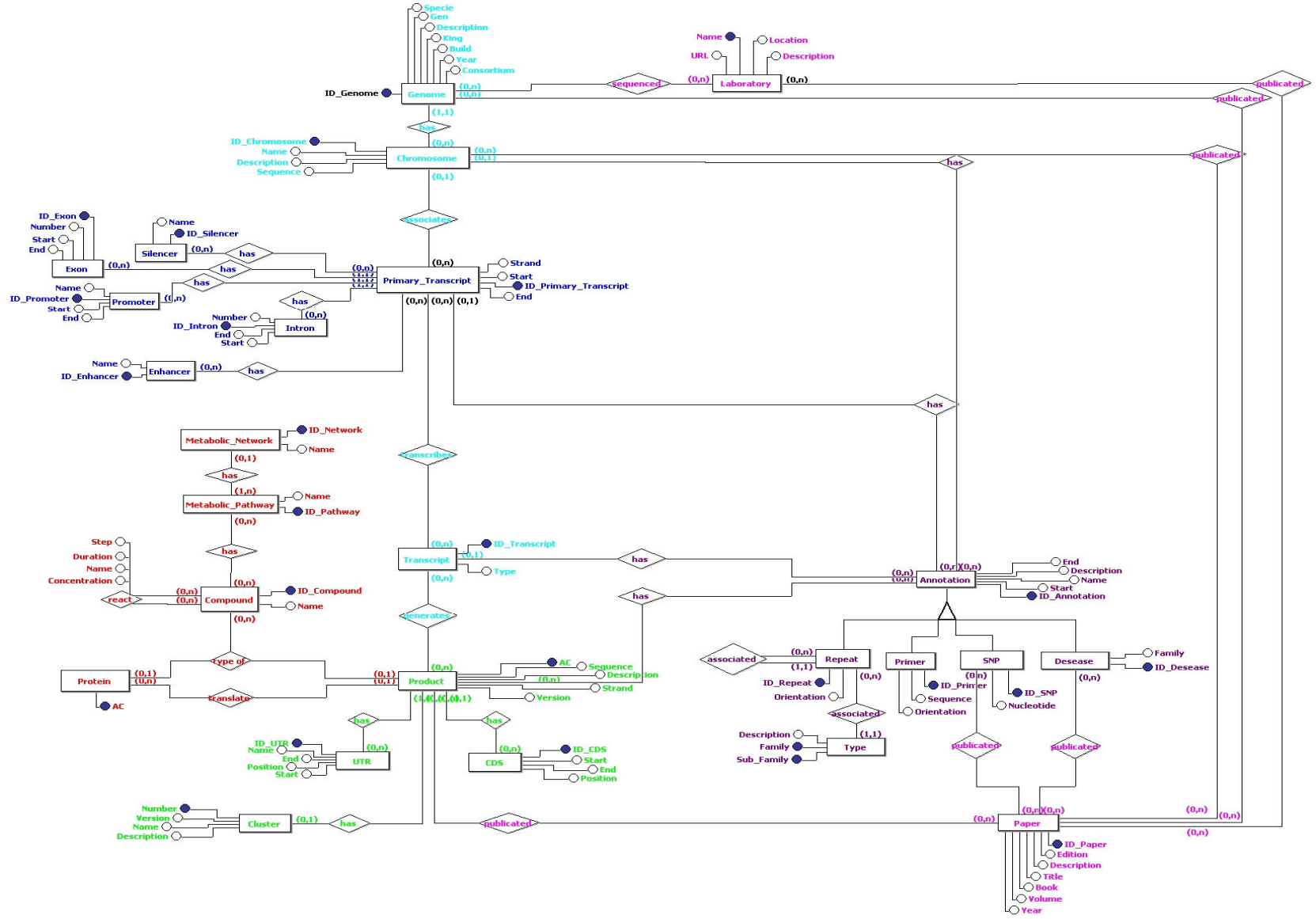


*Figura 30 – Modelo conceitual completo de Rep4DB.*
*Modelo conceitual do banco de dados Rep4DB proposto, desenvolvido com a ferramenta brModelo (Candido, 2005).*
*Fonte: da própria pesquisa.*

#### 4.4.1.2 Modelagem lógica

O modelo lógico é gerado diretamente a partir do modelo conceitual. Nesta fase do projeto, todas as informações presentes no modelo conceitual e que foram projetadas em função dos requisitos do sistema estão intimamente relacionadas com um SGBD específico, mas não implementadas. A Figura 35 apresenta uma versão completa do modelo lógico do banco de dados proposto, que baseia-se em um SGBD relacional, gratuito e de código aberto, chamado PostGreSQL (Stinson, 2001).

A geração do modelo lógico baseou-se, consequentemente, nas 4 partes do modelo conceitual previamente descrito:

— transcrito primário e suas partes (Figura 31);

— produto de transcrição (Figura 32);

— redes e vias metabólicas (Figura 33);

— anotações de sequências (Figura 34).

O modelo lógico de cada uma das 4 partes, apresentadas anteriormente, é ilustrado na Figura 35, de forma interligada, em que outras tabelas, destacadas nas cores rosa e azul-piscina, complementam o modelo lógico, todas elas geradas diretamente do modelo conceitual.

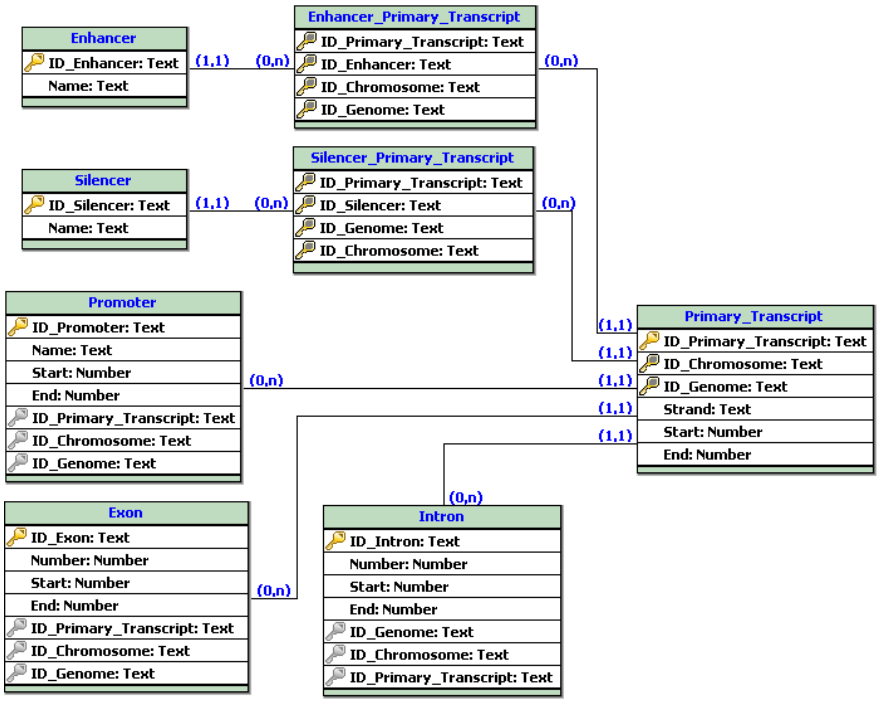


*Figura 31 – Modelo lógicas das tabelas associadas com transcrito primário.*
*Modelo lógico das tabelas representativas do MER de um transcrito primário e suas respectivas partes.*
*Fonte: da própria pesquisa.*

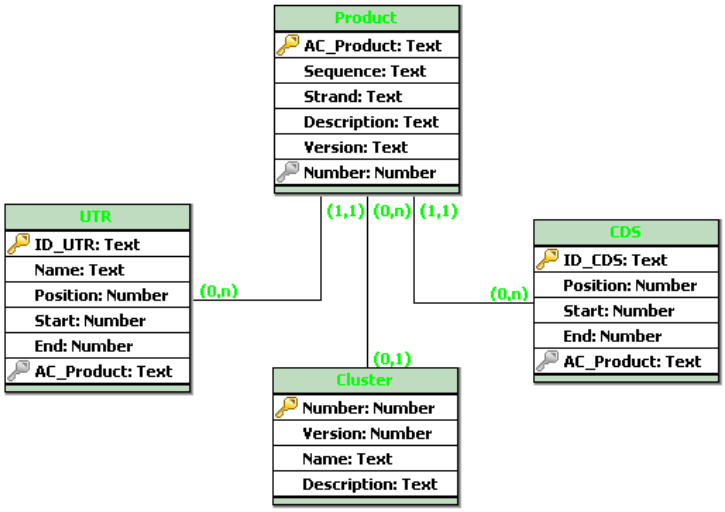


*Figura 32 - Modelo lógicas das tabelas associadas com produto.*
*Modelo lógico das tabelas representativas do MER de um produto de transcrição e suas respectivas partes.*
*Fonte: da própria pesquisa.*

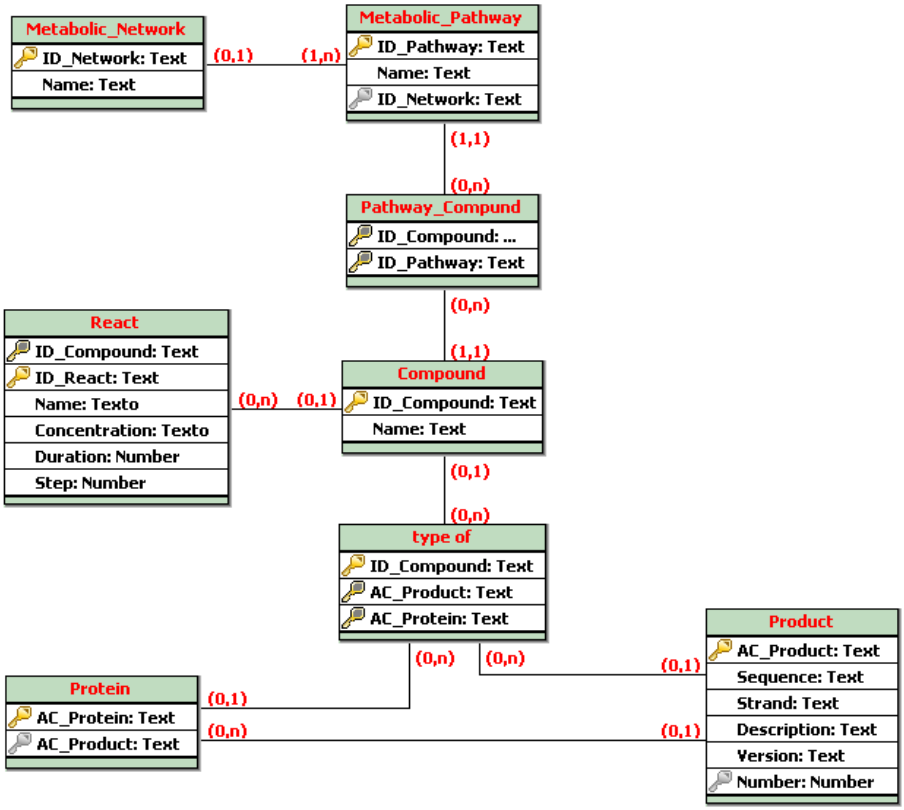


*Figura 33 – Modelo lógicas das tabelas associadas com o metabolismo.*
*Modelo lógico das tabelas representativas do MER das redes e vias metabólicas.*
*Fonte: da própria pesquisa.*

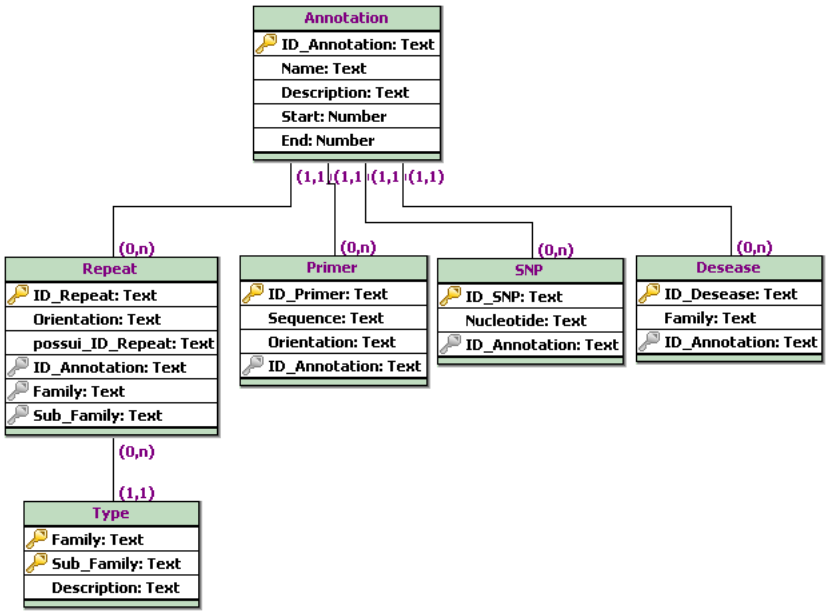


*Figura 34 – Modelo lógicas das tabelas associadas com anotação de sequências..*
*Modelo lógico das tabelas representativas ao MER referente aos tipos de anotação para alguma região de um cromossomo, transcrito primário, transcrito e produto de transcrição.*
*Fonte: da própria pesquisa.*

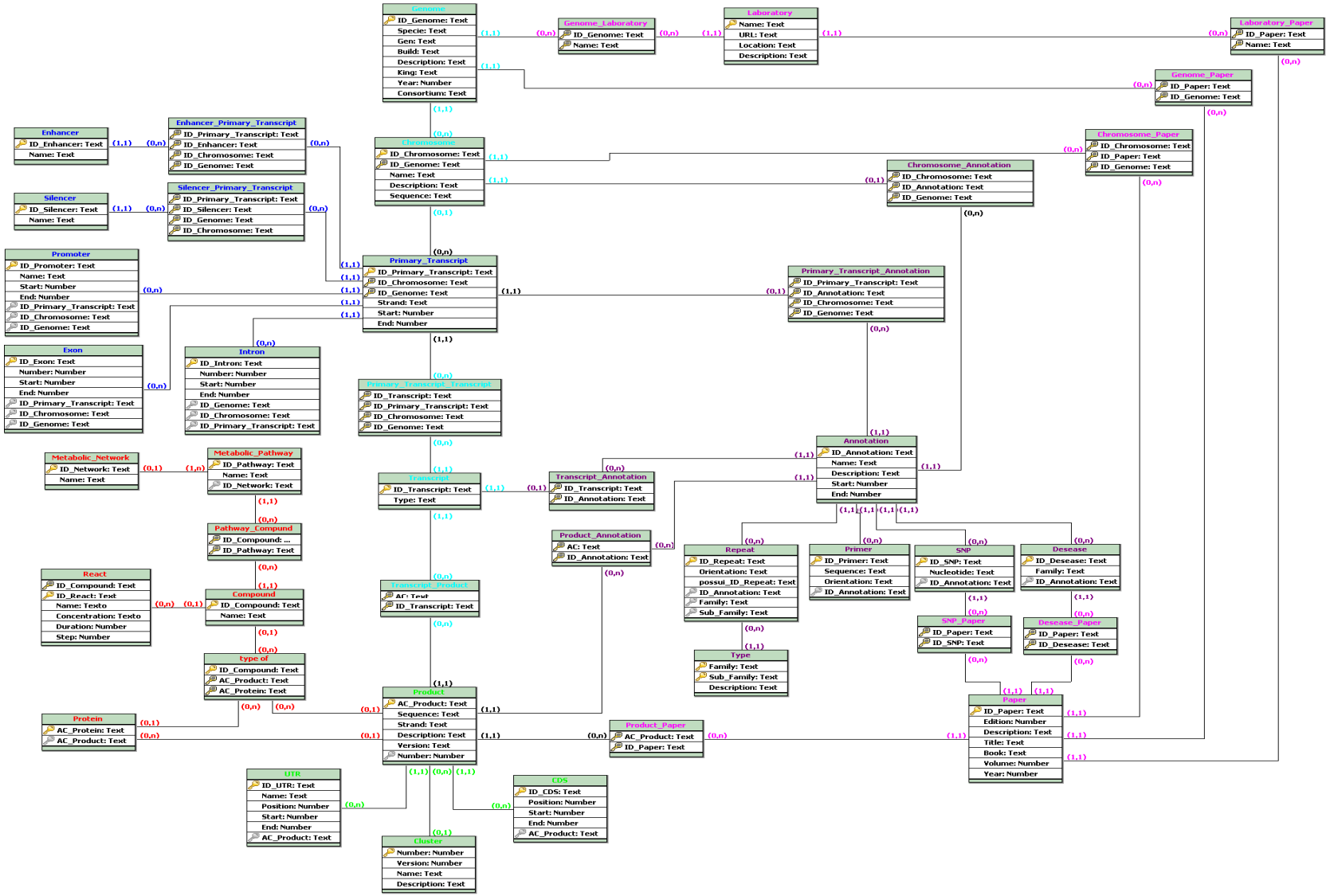


*Figura 35 – Modelo lógico do banco de dados Rep4DB.*
*Modelo lógico do banco de dados Rep4DB para armazenamento de sequências repetitivas e de outras estruturas genômicas, desenvolvido com a ferramenta brModelo (Candido, 2005). Modelo parcialmente normalizado.*
*Fonte: da própria pesquisa.*

#### 4.4.1.3 Modelagem física (implementação)

Após a conclusão do projeto lógico, passou-se para a etapa de implementação do sistema de banco de dados (modelagem física). Declarações DDL (linguagem de definição de dados), inclusive SDL (linguagem de definição de armazenamento), do SGBD PostGreSQL selecionado (outros SGBDs gratuitos poderiam ser utilizados), foram compiladas e usadas para criar e modificar a estrutura das tabelas no esquema do banco de dados (vazio) e os respectivos arquivos, ambos referentes ao modelo lógico considerado na seção anterior. Tal abordagem resume-se, basicamente, no uso de declarações baseadas na linguagem SQL, sendo as principais instruções: *CREATE TABLE*, *ALTER TABLE*, *ADD CONSTRAINT* e *DROP TABLE*. No Quadro 1 está uma representação em *DDL SQL*, para as tabelas do modelo lógico referente a um produto de transcrição e suas respectivas partes para o SGBD considerado: Cluster, UTR e CDS.

*Quadro 1 – Representação ilustrativa em DDL SQL.*
*Representação em DDL SQL das tabelas Product, Cluster, UTR e CDS do modelo de dados proposto.*

```
CREATE TABLE Product (
  AC_Product int PRIMARY KEY,
  Sequence Text,
  Strand Text,
  Description Text,
  Version int,
  Number int,
  FOREIGN KEY(Number) REFERENCES Cluster (Number)
)

CREATE TABLE Cluster (
  Number int PRIMARY KEY,
  Version int,
  Name Text,
  Description Text
)

CREATE TABLE UTR (
  ID_UTR int PRIMARY KEY,
  Name Text,
  Position int,
```

```
  Start int,
  End int,
  AC_Product int
)

CREATE TABLE CDS (
  ID_CDS int PRIMARY KEY,
  Position int,
  Start int,
  End int,
  AC_Product int,
  FOREIGN KEY(AC_Product) REFERENCES Product (AC_Product)
)
```

*Fonte: da própria pesquisa.*

Independentemente do tipo de sistema de banco de dados projetado e construído, como o proposto neste trabalho, direcionado principalmente para a anotação e armazenamento de sequências repetitivas, são fundamentais, na maioria das vezes, outros três tipos de sistemas para que o banco de dados seja mais util. O primeiro corresponde a um sistema de migração de dados, importante para garantir que dados, como os presentes em outros tipos de bases de dados (SGBDs ou arquivos), possam ser analisados e posteriormente armazenados no banco de dados proposto. O segundo refere-se à disponibilidade de uma ferramenta que permita a realização de consultas às informações disponíveis na base de dados, sem que exija conhecimento prévio da forma de manipulação e da estrutura do banco de dados. Um terceiro tipo de ferramenta muito importante são aquelas que permitem gerar novos dados que não existem em outras bases de dados, e que precisam ser inferidas. A ferramenta RepGraph, discutida anteriormente, apresenta tal tipo de funcionalidade. A primeira será discutida a seguir, e a segunda é tratada em um trabalho futuro.

## 4.5 Ferramenta de migração de dados

Uma ferramenta de migração de dados para sistemas de bancos de dados biológico requer, basicamente, uma fonte de dados e uma aplicação para analisar seu conteúdo para posterior armazenamento na base de dados proposta.

### 4.5.1 Metodologia

A metodologia para a construção de uma ferramenta de migração de dados biológica para o banco de dados Rep4DB, proposto neste trabalho, apresenta a estrutura da Figura 36. Nela, uma ferramenta, chamada MigDB, analisa um conjunto de dados de entrada proveniente de diferentes locais e formatos e armazena-os no banco de dados.

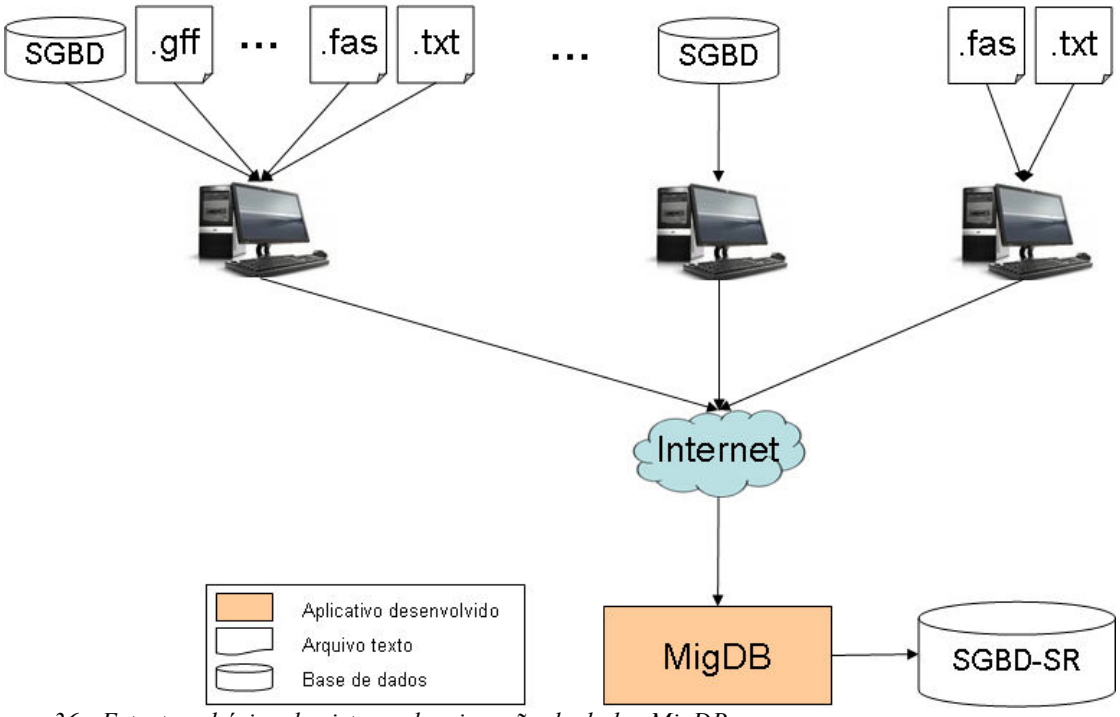


*Figura 36 - Estrutura básica do sistema de migração de dados MigDB.*
*Fonte: da própria pesquisa.*

### 4.5.2 Implementação

A implementação da ferramenta de migração de dados MigDB gerou uma aplicação em linha de comando, cuja versão atual analisa apenas sequências de dados biológicas em um formato texto, referenciado como *flatfile*. Tal formato é utilizado pelo DDBJ, que também faz parte do maior consórcio internacional de dados biológicos públicos, conforme descrito na seção 3.3.2, do Capítulo 3. A ferramenta foi desenvolvida em uma metodologia orientada a objetos com o uso da linguagem de programação Java. Tal aplicação utiliza o conector JDBC de banco de dados para o SGBD PostGreSQL, para conectar-se e gravar dados no banco de dados criado. Esta ferramenta, apesar de ter sido construída apenas para um formato de representação de dados (*flatile*), pode ser facilmente adaptada para outros formatos de bases de dados biológico, pois a estrutura da aplicação, que trata do acesso e da manipulação do banco de dados proposto, não sofrerá mudanças.

## 4.6 Detecção de erros de montagem em genomas "*draft*"

A detecção de erros de montagem de genomas *"draft"* (não finalizados) ainda é um dos grandes desafios, dentre vários, da era pós-genômica. Apesar da grande quantidade de metodologias propostas, conforme descritas nos capítulos anteriores, tal problema de checagem da montagem de um genoma ainda requer novos métodos que permitam minimizar a presença de possíveis erros de montagem, principalmente em regiões gênicas. Tais regiões, embora representem uma pequena fração do genoma, são essenciais em muitos estudos e devem, na medida do possível, estar isentas de erros de montagem. Para

tal, esta seção apresenta uma metodologia chamada DraftDNACheck (Herai & Yamagishi (b), 2009), que permite detectar possíveis erros de montagem em regiões gênicas de genomas. Ela baseia-se no uso de sequências adquiridas de forma independente, como, por exemplo, bases de ESTs ou bases de *Full length* cDNA (FlcDNA).

### 4.6.1 Metodologia

Esta metodologia para detecção de erros de montagem em regiões gênicas de genomas *"draft"* baseia-se no mapeamento de transcritos contra seu respectivo genoma de referência. Para tal, mapeamentos que não correspondem ao da Figura 37 (a), em que os éxons do transcrito mantêm a ordem e orientação de seu respectivo *locus* gênico, são considerados possíveis erros de montagem. Tais transcritos devem ser verificados posteriormente, tanto por outras estratégias de bioinformática quanto por técnicas laboratoriais, dado que podem, ao invés de corresponder de fato a erros de montagem, representar eventos biológicos raros ou que ainda não foram reportados na literatura. A Figura 37 (b), c) e d)) destaca os 3 tipos possíveis de erros que são considerados pela metodologia, que levam em consideração trechos de sequências que mapearam de forma desordenada quanto à orientação ou quanto à ordem dos éxons do transcrito, ou que não existem no genoma.

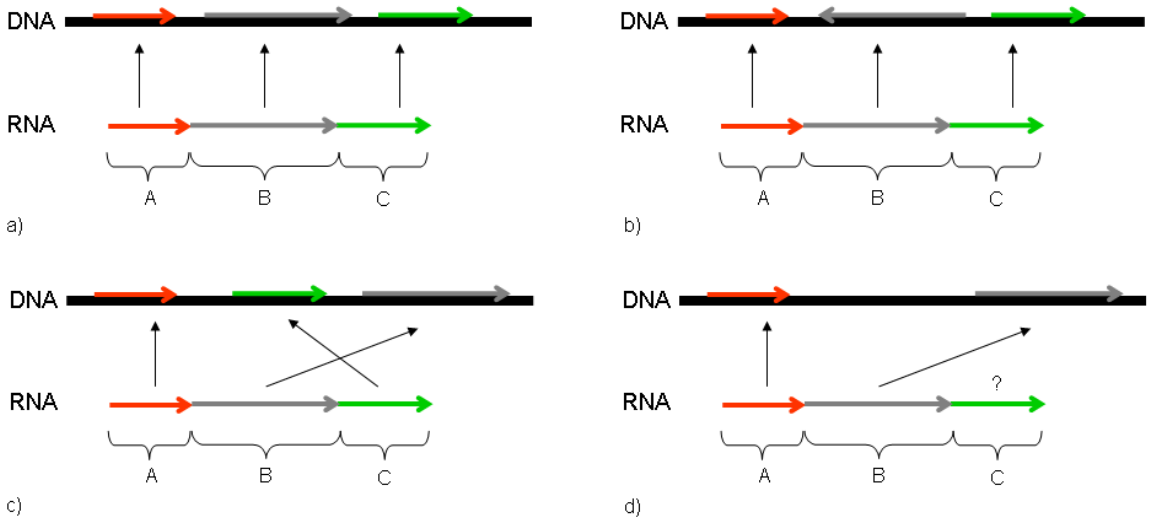


*Figura 37 – Mapeamentos de um transcrito contra um genoma.*
*Tipos de mapeamento de um transcrito contra seu respectivo genoma de referência, analisados para verificar erros de montagem: a) mapeamento esperado; b) mapeamento tratado como possível erro de montagem, em que a parte B do transcrito mapeia-se em sentido* antisense *no genoma, e as demais partes em sentido* sense*, em que a ordem dos éxons é preservada; c) mapeamento tratado como possível erro de montagem, cuja ordem dos éxons aparece na mesma orientação, porém invertida no genoma; d) mapeamento tratado como possível erro de montagem, na qual o éxon C não é mapeado no genoma.*
*Fonte: da própria pesquisa.*

A Figura 38 ilustra a metodologia proposta para detectar os casos de mapeamentos descritos pela Figura 37 (b), c) e d)).

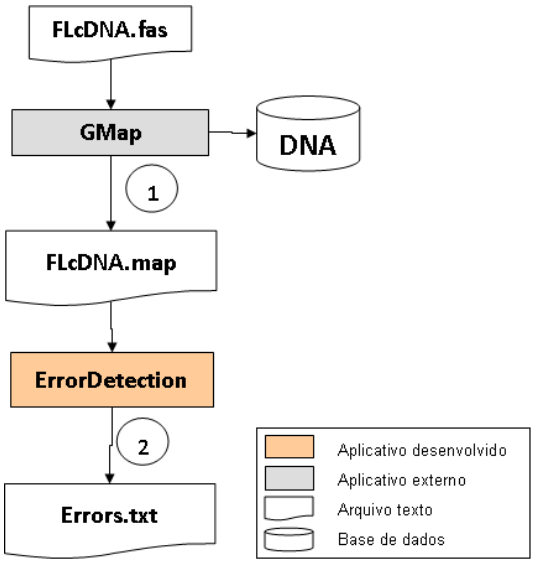


*Figura 38 – Metodologia da ferramenta DraftDNACheck.*
*Metodologia para detecção de erros de montagem em genomas "draft" com o uso de bibliotecas de transcritos. 1) GMap: mapeamento de transcritos contra seu respectivo genoma de referência; 2) ErrorDetection: análise dos mapeamentos para detecção de possíveis erros de montagem.*
*Fonte: da própria pesquisa.*

### 4.6.2 Implementação

A implementação da metodologia DraftDNACheck envolveu o uso de duas ferramentas. A primeira é uma ferramenta de mapeamento de sequências biológicas, o GMap (Wu & Watanabe, 2005), disponível gratuitamente. Tal ferramenta produz, como resultado, informações de cada porção de uma sequência que foi mapeada no genoma, tais como: orientação, posição de início e fim da sequência e do mapeamento do transcrito no genoma, percentual de cobertura, quantidade de *gaps*, entre outros.

A segunda parte da metodologia é a responsável por analisar os mapeamentos realizados pela ferramenta GMap para detectar possíveis erros de montagem. Para tal, o desenvolvimento da ferramenta DraftDNACheck produz um conjunto de informações referentes aos possíveis erros de montagem já descritos pela Figura 37, os quais podem ser relativos a 3 tipos: (*i*) parte de um transcrito mapeia-se em sentido *antisense* no genoma, e as demais partes em sentido *sense*, em que a ordem dos éxons é preservada; (*ii*) éxons aparecem na mesma orientação, porém de forma desordenada no genoma; (*iii*) pelo menos um dos éxons não é mapeado no genoma. Obviamente, para os 3 casos, uma análise a *posteriori* é fundamental para comprovar a suposição de erro. Para os casos de alinhamentos de uma mesma sequência que ocorra em lugares diferentes do genoma, é verificado se pelo menos um dos casos permite gerar um alinhamento satisfatório.

Em seu desenvolvimento, foi utilizado um modelo de software orientado a objetos e uma linguagem de programação gratuita e independente de plataforma, a Java. Será visto no capítulo de resultados que a metodologia já permitiu detectar e, inclusive, sugerir

correções na que é considerada hoje a melhor montagem do genoma bovino, o *Bos taurus* UMD 3.0.

A versão atual da ferramenta é uma aplicação em linha de comando. Entretanto, pelo fato de ter sido desenvolvida em Java, pode ser facilmente portada em um ambiente WEB com interface gráfica e com funcionalidades que podem ser inseridas de forma incremental.

## 4.7 Considerações do capítulo

Este capítulo apresentou uma descrição das metodologias de bioinformática propostas nos objetivos do projeto, e também algumas características de implementação de cada uma delas. As metodologias foram apresentadas separadamente, para facilitar a compreensão de cada ferramenta na realização de futuras análises de dados biológicos associados a genomas e transcriptomas, e sua relação com SR. Vale destacar que as metodologias e ferramentas construídas correspondem a uma importante contribuição deste trabalho. Para comprovar a validade e utilidade das ferramentas propostas, o próximo capítulo apresenta alguns resultados gerados a partir delas, e uma breve discussão das mesmas.

# Capítulo 5 Resultados e discussão

As ferramentas de bioinformática construídas principalmente para detecção e análise de transcritos quiméricos e sua relação com SR foram satisfatoriamente implementadas e testadas com dados reais de genomas e transcriptomas. Neste capítulo, é feita uma breve apresentação e discussão dos resultados gerados pelas ferramentas propostas, com o intuito de demonstrar sua aplicabilidade.

## 5.1 Identificação de transcritos quiméricos candidatos e sua relação com SR em *locus* gênico

A formação de transcritos quiméricos por meio de *trans-splicing* em tecidos normais, conforme reportado na literatura especializada, ainda encontra-se em fase de investigação, apesar de algumas comprovações experimentais já terem sido publicadas em humanos, como a de Li et al. (2009). Além disso, já há algumas evidências que tentam demonstrar que uma das formas possíveis de formação de transcritos quiméricos é por meio da presença de pares de SR do tipo reverso complementar (RRC) em transcritos distintos, como aquelas descritas por Dixon et. al (2007) e por Di Segni et al. (2008). Tais pares sofrem hibridização e juntam os dois transcritos, formando uma nova molécula de RNA. Em terapia gênica isso já foi comprovado pela técnica chamada SMART$^{TM}$ (Otto et al., 2003).

Embora já existam evidências, os casos de transcritos quiméricos reportados são poucos, e faltam novos transcritos candidatos que sirvam de subsídio para novos

experimentos laboratoriais. Em função disso, as metodologias Fusion5Finder e FusionAllFinder para detecção de transcritos quiméricos foram utilizadas para identificar novos transcritos quiméricos candidatos, tanto em humanos quanto em bovinos. Além disso, um estudo da frequência de SR do tipo reverso complementar foi realizado. As próximas seções definem as bases de dados utilizadas nos testes, e apresentam os resultados obtidos com uma breve discussão sobre os mesmos.

### 5.1.1 Transcritos quiméricos em Humano e SR

#### 5.1.1.1 Características dos dados

Para os testes de detecção de transcritos quiméricos em humanos em tecidos normais, foi utilizado um banco de dados curado formado por um conjunto de clones de cDNA de alta qualidade, composto de transcritos provenientes de humanos e que foi gerado por um projeto colaborativo chamado H-Invitational Database (H-InvDB DB) (Imanishi et al., 2004). A versão utilizada, a 5.0, é baseada no build 36.3 do genoma humano e possui 187.156 sequências de FL-cDNA, das quais apenas 113.202 possuem informações a respeito das regiões 5'UTR, CDS, e 3'UTR, e que são provavelmente provenientes de tecidos normais.

#### 5.1.1.2 Resultados e discussão de candidatos quiméricos em Humano

A metodologia Fusion5Finder procurou por candidatos quiméricos em humanos, semelhante às melhores evidências experimentais, cuja principal característica é o fato da

5'UTR, ou parte dela, ocorrer em um único cromossomo (denotado neste trabalho por região TSR: *trans-spliced region*), com suas regiões CDS e 3'UTR sendo encontradas em um outro cromossomo distinto, sem impor a existência dos *splice-sites* canônicos.

Foram encontrados 16 mRNAs híbridos (Tabela 1) com caracterísiticas que se assemelham às melhoras evidências experimentais de *trans-splicing* em humanos, oriundos supostamente de células de tecidos normais. Como descrito na seção 4.1.2, para evitar ambiguidades na interpretação, todos os candidatos possuem sua região TSR mapeada em somente um único *locus* cromossômico.

*Tabela 1 – Transcritos quiméricos candidatos em Humano.*
*Transcritos quiméricos candidatos que foram identificados na base de dados H-InvDB. As 16 sequências candidatas são listadas com as seguintes informações: Accession Number (AC), Trans-Spliced Region (TSR), regiões não traduzidas 5' e 3' (5' e 3' UTR), região codificante (CDS), tamanho, em pares de base, do transcrito de RNA mensageiro (mRNA), e tecido de origem (N/A indica que a informação não foi encontrada).*

| AC | TSR | 5'UTR | CDS | 3'UTR | mRNA | Tecido |
|---|---|---|---|---|---|---|
| [DDBJ:D26155] (*hsNF2a*) | [1-293] | 297 | 4719 | 241 | 5257 | Cérebro |
| [DDBJ:AL834489] (*DKFZp434F1431*) | [1-322] | 324 | 1056 | 2367 | 3747 | Testículo |
| [DDBJ:AB023216] (*KIAA0999*) | [1-302] | 437 | 3792 | 231 | 4460 | Cérebro |
| [DDBJ:AK124366] (*FLJ42375 fis*) | [1-224] | 302 | 255 | 2047 | 2604 | Útero |
| [DDBJ:AK226066] (*LAMP2*) | [1-342] | 539 | 1236 | 2333 | 4108 | Cérebro |
| [DDBJ:AF003522] (*Delta mRNA*) | [8-249] | 322 | 2172 | 668 | 3162 | N/A |
| [DDBJ:L33075] (*IQGAP1*) | [3-400] | 467 | 4974 | 2132 | 7573 | Placenta, pulmão |
| [DDBJ:AK130557] (*FLJ27047 fis*) | [1-578] | 678 | 1065 | 1118 | 2861 | Glândula salivar |
| [DDBJ:L14837] (*z. occludens ZO-1*) | [6-732] | 1190 | 5247 | 1450 | 7887 | N/A |
| [DDBJ:U09825] (*acid finger protein*) | [1-345] | 555 | 1620 | 1420 | 3595 | Rim |
| [DDBJ:AB007865] (*KIAA0405*) | [1-987] | 1124 | 1983 | 4420 | 7887 | Cérebro |
| [DDBJ:AB020656] (*KIAA0849*) | [4-233] | 446 | 2862 | 2106 | 5414 | Cérebro |
| [DDBJ:AF458052] (*GRM7*) | [8-303] | 451 | 2775 | 163 | 3389 | * |
| [DDBJ:AF458053] (*GRM7*) | [8-303] | 451 | 2736 | 135 | 3322 | * |
| [DDBJ:AF458054] (*GRM7*) | [8-303] | 451 | 2721 | 31 | 3203 | * |
| [DDBJ:U92458] (*GRM7*) | [8-303] | 451 | 2748 | 1113 | 4312 | Cérebro fetal |

* cérebro, traquéia, testículo, útero, grândula salivar.
*Fonte: da própria pesquisa.*

Embora o segundo caso experimental sugira que se observem transcritos com 5'UTR excepcionalmente longos (como *ACAT-1*, cuja 5'UTR possui 1.396 bp), nossa metodologia encontrou 4 sequências candidatas com 5'UTR menores do que 400 bp. Vale

destacar que aqueles que foram experimentalmente comprovados não foram detectados por não fazerem parte da base de dados utilizada.

De fato, há candidatos similares às evidências esperimentais com 5'UTR longas. Um exemplo ilustrativo é o do gene FLRT2 (candidato quimérico [DDBJ:AB007865]), cuja estrutura é apresentada na Figura 39. Ela possui uma região com 987 pb oriunda de um cromossomo distinto do restante do transcrito. Tal transcrito é o maior do seu respectivo cluster HIX0011865 do H-InvDB, conforme Figura 40.

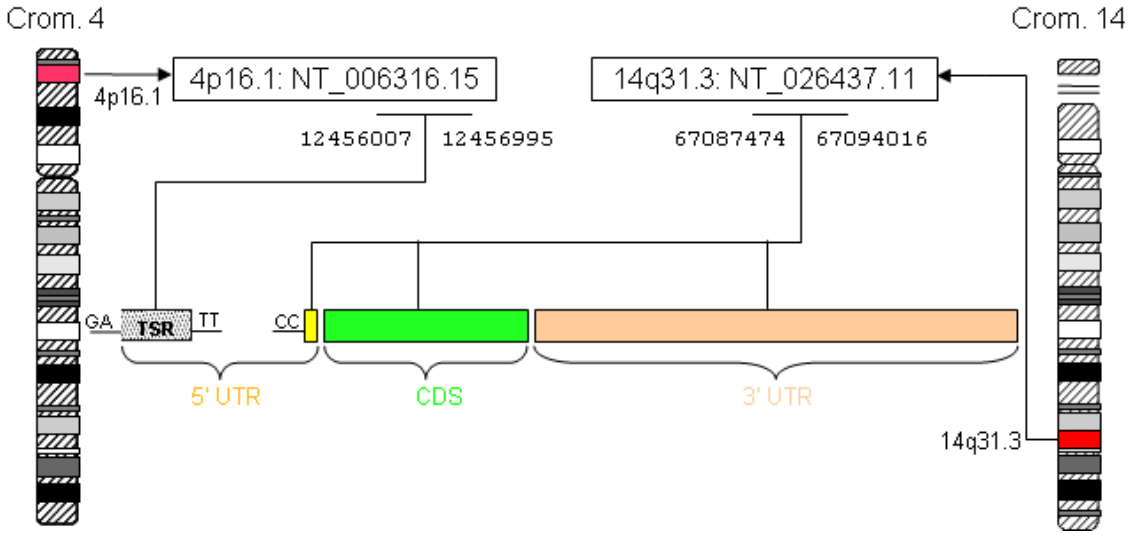


*Figura 39 – Loci gênicos do transcrito quimérico candidato [DDBJ:AB007865].*
*Estrutura do transcrito quimérico candidato [DDBJ:AB007865] (FLRT2), gerado possivelmente por meio de trans-splicing intercromossomal. A região 5'UTR possui 987 pb que são oriundos do cromossomo 4. O restante do transcrito pertence ao cromossomo 14.*
*Fonte: da própria pesquisa.*

Não há intersecção entre nossos candidatos e a lista do ISTReS' (Romani et al., 2003) que é uma alternativa à nossa metodologia. Diferenças na base de dados de cDNA pode ser uma explicação para isso. Entretanto, há outra razão que parece mais significante: não impomos *splice-sites* canônicos e, surpreendentemente, nenhuma de nossas sequências candidatas seguem tais *splice-sites* (Tabela 2)**.** Isto pode estar relacionado com o fato de que a metodologia Fusion5Finder foi projetada para encontrar um tipo muito específico de

*trans-splicing*: intercromossomal, cuja TSR pertence exclusivamente a 5'UTR. É possível que este tipo de *trans-splicing* não faça uso da maquinaria do complexo de enzimas spliceossomo. Outra hipótese que não pode ser descartada é de que *splice-sites* raros ocorram (Steitz et al., 2008), pois não seria improvável que *splice-sites* raros ocorressem num fenômeno igualmente raro.

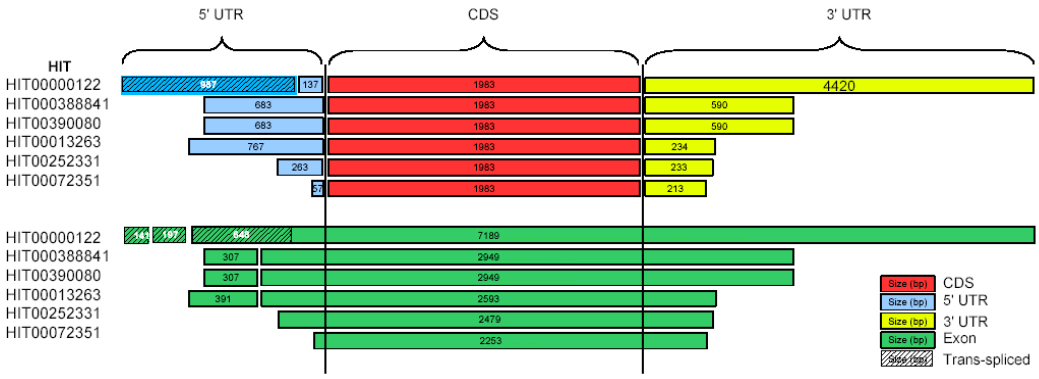


*Figura 40 – Transcritos do Cluster HIX0011865.*
*Na parte superior estão os seis mRNAs com suas regiões 5'UTR, CDS, e 3'UTR. Na parte inferior, o mesmo conjunto de genes e o tamanho de cada região, considerando a estrutura exon-intron de cada uma delas.O primeiro transcrito de cada uma das partes corresponde ao transcrito quimérico candidato [DDBJ:AB007865].*
*Fonte: da própria pesquisa.*

Também verificamos a qualidade das sequências com relação à possibilidade de contaminação externa, e nenhum tipo de contaminação por sequências de vetores, DNA mitocondrial ou bactérias foi encontrada. Além disso, embora H-InvDB seja uma base de dados altamente curada, não podemos excluir a possibilidade que alguns mRNAs híbridos foram produzidos por erro de clonagem ou outros tipos de artefatos experimentais. Entretanto, 4 candidatos gerados por dois laboratórios e bibliotecas independentes pertencem a um mesmo cluster do H-InvDB, o que minimiza a probabilidade de erros de clonagem ou outros artefatos experimentais.

*Tabela 2 – Splice sites das sequências quiméricas candidatas.*
*A primeira coluna contém o código de acesso (AC) das sequências candidatas, seguida pela sua localização em alguma banda de cromossomo (locus), o cromossomo cuja região TSR é proveniente (TSR Crom), o dinucleotídeo que representa os splice-sites de cada TSR (em negrito) e do restante do transcrito quimérico (CDS-3'UTR Sp, em negrito), respectivamente.*

| AC | Locus | TSR Crom | TSR Sp | CDS-3'UTR SP |
|---|---|---|---|---|
| [DDBJ:D26155] | 9p24.3 | 6 | **AA**-TSR-**TT** | **AG-**CDS-3'UTR |
| [DDBJ:AL834489] | 5q35.2 | 12 | **GG**-TSR-**AG** | **CG-**CDS-3'UTR |
| [DDBJ:AB023216] | 11q23.3 | 2 | **CG**-TSR-**GC** | **CT-**CDS-3'UTR |
| [DDBJ:AK124366] | 8q24.12 | 5 | **CT**-TSR-**CA** | **TG-**CDS-3'UTR |
| [DDBJ:AK226066] | Xq24 | 1 | **GT**-TSR-**AT** | **CT-**CDS-3'UTR |
| [DDBJ:AF003522] | 6q27 | 14 | **TT**-TSR-**AG** | **TC-**CDS-3'UTR |
| [DDBJ:L33075] | 15q26.1 | 4 | **CT**-TSR-**TC** | **CA-**CDS-3'UTR |
| [DDBJ:AK130557] | 4q13.3 | 19 | **AT**-TSR-**TA** | **TT**-CDS-3'UTR |
| [DDBJ:L14837] | 15q13.1 | 16 | **TT**-TSR-**AC** | **GC-**CDS-3'UTR |
| [DDBJ:U09825] | 6p21.33 | 1 | **AA**-TSR-**AA** | **CA-**CDS-3'UTR |
| [DDBJ:AB007865] | 14q31.3 | 4 | **GA**-TSR-**TT** | **CC-**CDS-3'UTR |
| [DDBJ:AB020656] | 16q12.1 | 11 | **GG**-TSR-**AG** | **TC-**CDS-3'UTR |
| [DDBJ:AF458052] | 3p26.1 | 19 | **CA**-TSR-**TT** | **CT-**CDS-3'UTR |
| [DDBJ:AF458053] | 3p26.1 | 19 | **CA-**TSR-**TT** | **CT-**CDS-3'UTR |
| [DDBJ:AF458054] | 3p26.1 | 19 | **CA**-TSR-**TT** | **CT-**CDS-3'UTR |
| [DDBJ:U92458]] | 3p26.1 | 19 | **CA**-TSR-**TT** | **CT-**CDS-3'UTR |

*Fonte: da própria pesquisa.*

Procuramos também por evidências de que as TSR são transcritas, pois a formação de um transcrito quimérico por *trans-splicing* requer que, além do transcrito principal, sua região TSR também seja transcrita. Utilizando a ferramenta UCSC BLAT (Kent, 2002), cada TSR foi mapeada no genoma humano e, utilizando o *UCSC Genome Browser*, identificamos genes anotados na região genômica das TSRs. Considerando as 16 TSRs da Tabela 3, foram observados dois casos (Tabela 3): (*i*) doze TSRs foram mapeadas em *loci* gênicos conhecidos (Figura 41 mostra as TSRs mapeadas em regiões de éxons, e a Figura 42, as TSRs mapeadas em regiões de íntrons); (*ii*) quatro TSRs (na realidade, somente uma TSR não redundante) foram mapeadas em uma região genômica de um retrovírus endógeno humano (*HERV-K*) (Figura 43).

Por exemplo, o transcrito com número de acesso [DDBJ:AK124366] é quase completamente mapeado no cromossomo 8. Entretanto, parte de sua 5'UTR vem do

cromossomo 5, e ele pertence à região 3'UTR de um gene que está em sentido *antisense*, cujo numero de acesso é [GenBank: NM002715]. Este padrão de transcrição em *sense/antisense* se assemelha ao mecanismo de *trans-splicing* mediado por sequências repetitivas (conforme reportado por Dixon et al. (2007) e por Di Segni et al. (2008)). Para verificar tal relação, foi utilizada a ferramenta RepGraph, a qual permitiu identificar SR no *locus* gênico dos transcritos, e que podem estar relacionadas com a mediação do evento de *trans-splicing*.

*Tabela 3 – Evidências de transcrição (ET) dos candidatos quiméricos em Humano.*
*Evidências de transcrição (ET) das regiões de TSR de cada uma das sequências quiméricas candidatas. Quando a TSR foi mapeada em genes conhecidos, foram encontrados dois casos: TSR em regiões de éxons e TSR em regiões de íntrons. O código de acesso dos transcritos associados é apresentado em parênteses, e a localização da TSR indicada entre parênteses, seja no éxon ou no íntron.*

| AC | ET |
|---|---|
| [DDBJ:D26155] | [GenBank:NM_020823] (*TMEM181*) [Éxon: 1-293] |
| [DDBJ:AL834489] | [GenBank:NM_175736] (*FMNL3*) [Éxon: 106-322] |
| [DDBJ:AB023216] | [GenBank:NM_016552] (*ANKMY1* ) [Éxon: 110-218] |
| [DDBJ:AK124366] | [GenBank:NM_002715] (*PPP2CA* ) [Éxon: 1-224] |
| [DDBJ:AK226066] | [GenBank:NM_001821] (*CHML* ) [Éxon: 1-342] |
| [DDBJ:AF003522] | [GenBank:NM_021136] (*RTN1-1*) [Éxon: 1-242] |
| [DDBJ:L33075] | [GenBank:BC032784] (*CAMK2D* ) [Íntron] |
| [DDBJ:AK130557] | [GenBank:BC136777] (*ZNF700*) [Íntron] |
| [DDBJ:L14837] | [GenBank:AK124977] (FLJ42987 fis) [Íntron] |
| [DDBJ:U09825] | [GenBank:NM_025106] (*SPSB*) [Íntron] |
| [DDBJ:AB007865] | [GenBank:NM_147182] (*KCNIP4*) [Íntron] |
| [DDBJ:AB020656] | [GenBank:BC030148] (*ARFGAP2*) [Íntron] |
| [DDBJ:AF458052] | [GenBank:Q9YNA8] *(HERV-K)* – 19q12 |
| [DDBJ:AF458053] | [GenBank:Q9YNA8] *(HERV-K)* - 19q12 |
| [DDBJ:AF458054] | [GenBank:Q9YNA8] *(HERV-K)* - 19q12 |
| [DDBJ:U92458] | [GenBank:Q9YNA8] *(HERV-K)* - 19q12 |

*Fonte: da própria pesquisa.*

Embora tenhamos encontrado centenas e quase perfeitas SR reversas complementares em todos os casos, para visualizar alguns exemplos, consideramos somente aquelas com pelo menos 250 pb e com alta identidade entre as sequências (Figura 44). Pode-se observar que, em alguns casos, há uma quantidade tão grande de SR do tipo reverso complementar o que dificulta a interpretação da representação gráfica. Tais sequências repetitivas possuem tamanho aproximado ao das SR do tipo ALUs.

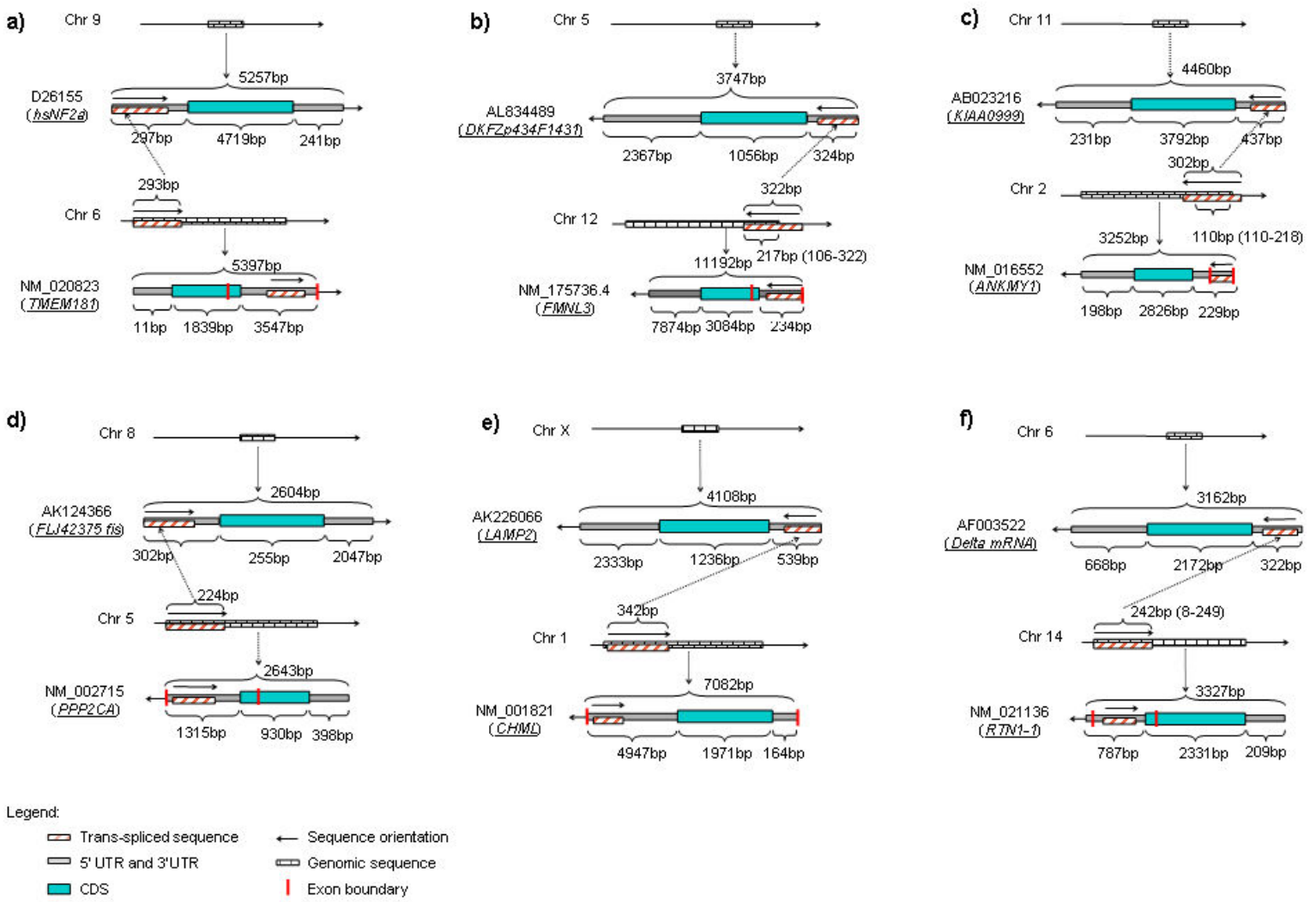


*Figura 41 – Candidatos quiméricos em Humano com TSR em éxons de outros transcritos..*
*Estrutura gênica de 6 candidatos quiméricos (5'UTR-CDS-3'UTR, incluindo o tamanho da sequência em pares de base) cuja região TSR mapeia-se em éxons de um transcrito. Cada transcrito está relacionado com um cromossomo do qual foi transcrito, e uma região TSR que localiza-se em sua porção 5'UTR, e que veio de um cromossomo distinto do restante do transcrito. Flechas indicam a orientação do transcrito e do cromossomo, em sentido sense ou antisense.*
*Fonte: da própria pesquisa.*

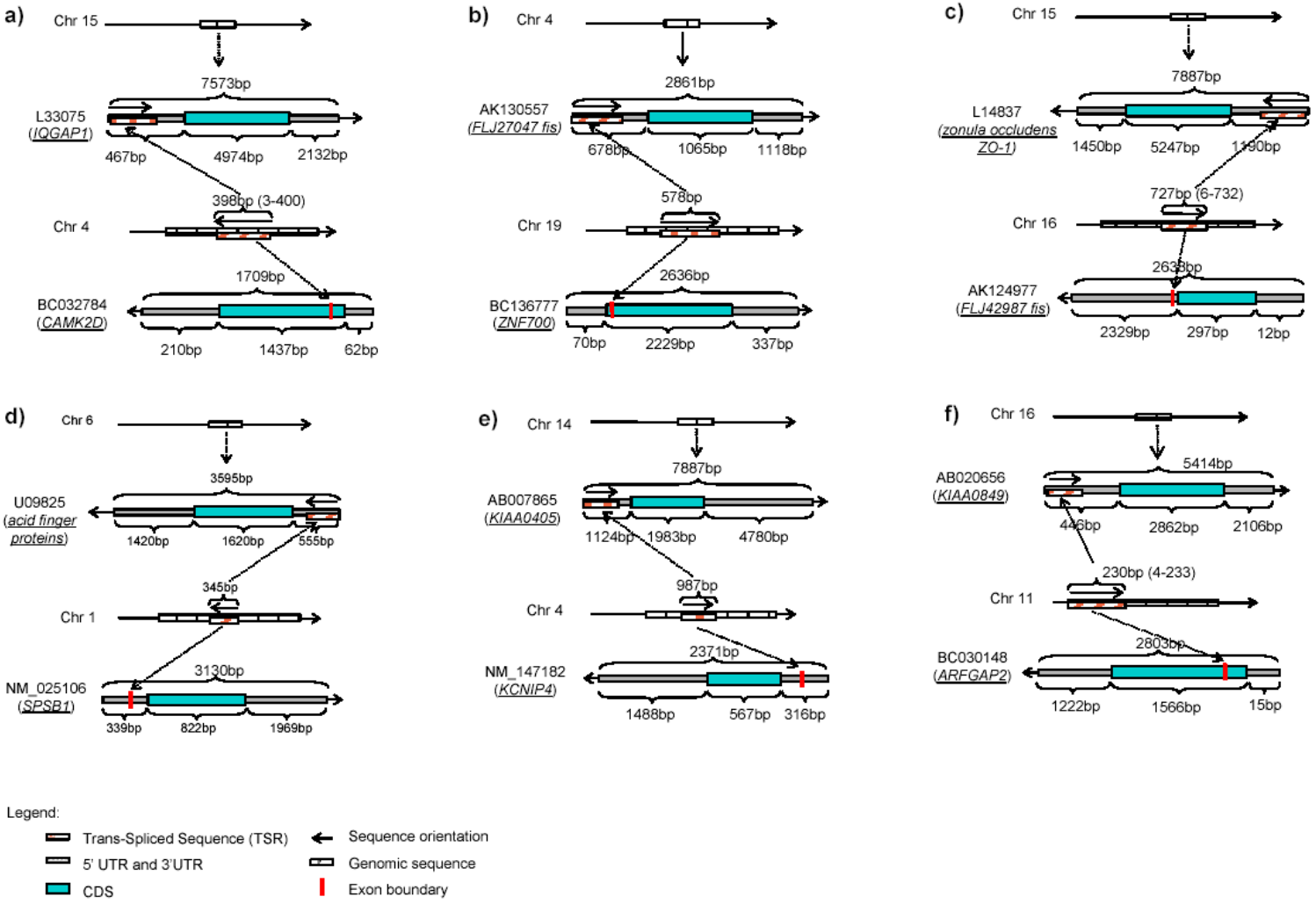


*Figura 42 - Candidatos quiméricos em Humano com TSR em íntrons de outros transcritos.*
*Estrutura gênica de 6 candidatos quiméricos (5'UTR-CDS-3'UTR, incluindo o tamanho da sequência em pares de base), cuja região TSR mapeia-se em íntrons de um transcrito. Cada transcrito está relacionado com um cromossomo do qual foi transcrito, e uma região TSR que localiza-se em sua porção 5'UTR, e que veio de um cromossomo distinto do restante do transcrito. Flexas indicam a orientação do transcrito e do cromossomo, em sentido sense ou antisense.*
*Fonte: da própria pesquisa.*

*Tabela 4 – Transcritos do cluster HIX0019725 da base de dados H-InvDB.*
*Para cada transcrito, é definido o Número de Acesso (AC), identificação no H-InvDB (HIT), tecido de origem (Tecido), e o tamanho das regiões 5' e 3' UTR, e CDS.*

| AC | HIT | Tecido | 5' UTR | CDS | 3' UTR |
|---|---|---|---|---|---|
| [DDBJ:AF458052] (*GRM7*) | HIT000079970 | * | 451 | 2775 | 163 |
| [DDBJ:AF458053] (*GRM7*) | HIT000079971 | * | 451 | 2736 | 135 |
| [DDBJ:AF458054] (*GRM7*) | HIT000079972 | * | 451 | 2721 | 31 |
| [DDBJ:U92458] (*GRM7*) | HIT000222625 | Cérebro de feto | 451 | 2748 | 1113 |

* cérebro, traquéia, testículo, útero, glandula salivar.
*Fonte: da própria pesquisa.*

As TSRs que foram mapeadas na região genômica do gene *HERV-K* (Figura 43) merecem especial atenção. A discussão a seguir levantou uma nova hipótese com relação à forma com que um transcrito quimérico pode ser formado, que até então costuma ser associada com *trans-splicing* ou rearranjo genômico. No *UCSC Genome Browser* estão indicados apenas elementos LINE no *locus* das TSR.

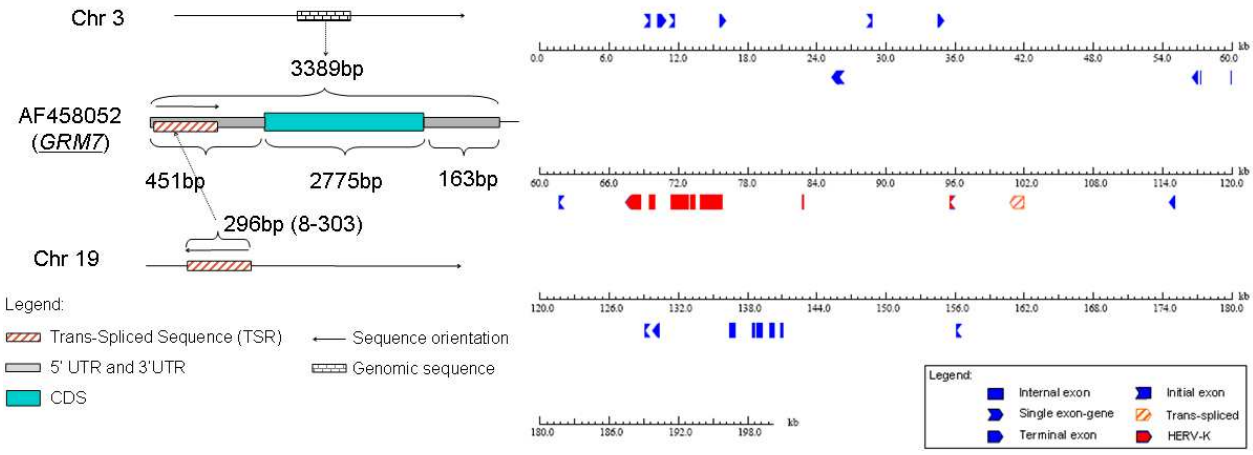


*Figura 43 - TSR do transcrito [DDBJ:AF458052] mapeado no cromossomo 19.*
*Na esquerda, a estrutura do gene. À direita, análise da região genômica da região TSR do transcrito candidato quimérico, na qual foi mapeado o retrovírus endógeno HERV-K.*
*Fonte: da própria pesquisa.*

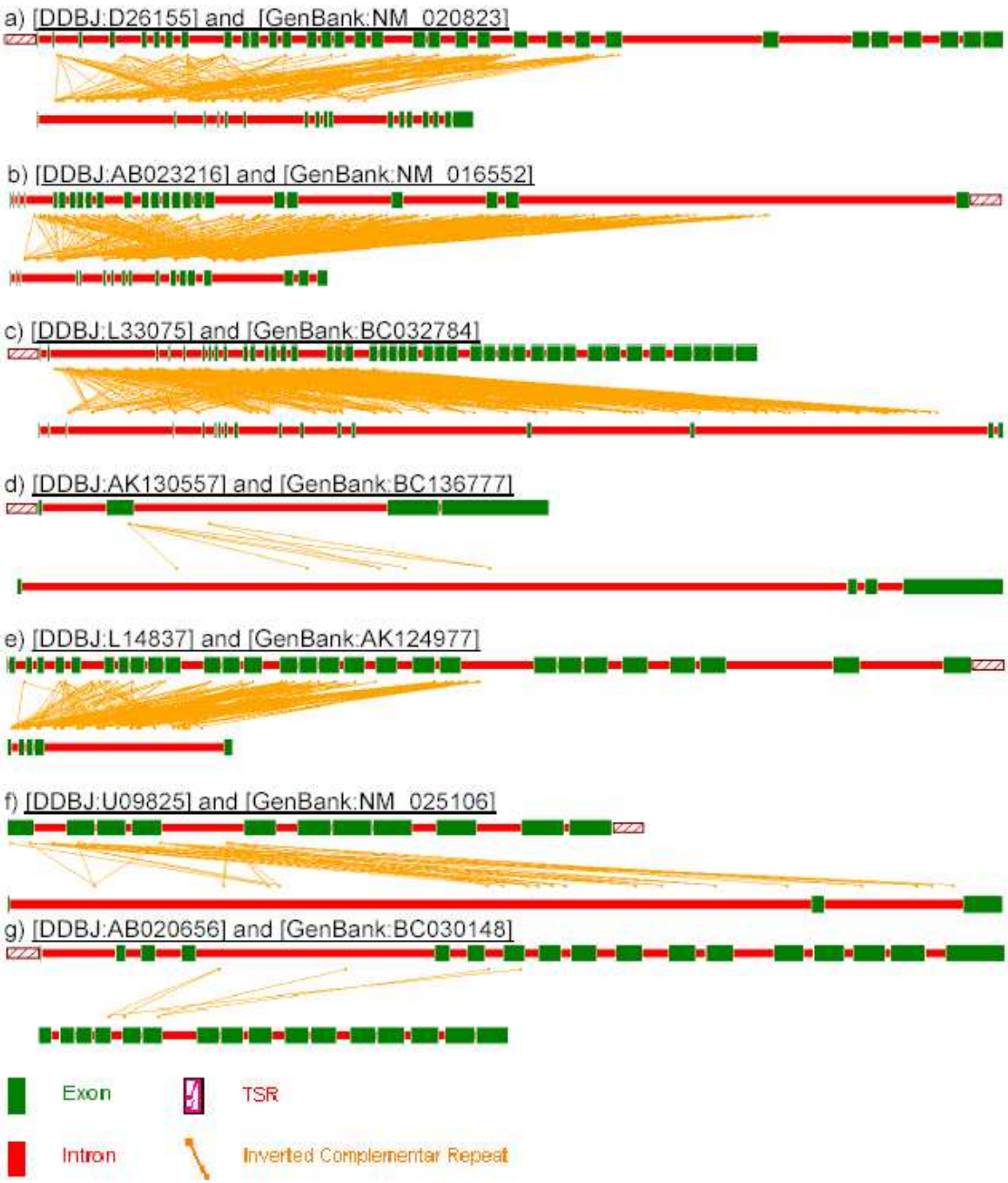


*Figura 44 – SR entre os candidatos quiméricos em Humano com uso de RepGraph.*
*Alinhamentos, com o uso da ferramenta RepGraph, entre os pares de transcritos envolvidos na formação das prováveis sequências quiméricas candidatas. Tal alinhamento ilustra que há SR complementares entre si, e que podem favorecer a fusão entre as duas moléculas de RNA envolvidas para a formação de uma molécula quimérica por meio de trans-splicing. Somente as SR do tipo reverso e complementar de tamanho maior ou igual a 250 pb são ilustradas na figura. Tais SR apresentam tamanho aproximado ao das ALUs, exclusivas de animais mamíferos.*
*Fonte: da própria pesquisa.*

Com o intuito de confirmar a presença de um transcrito ativo em tal região, com o uso do BLAT, mapeamos essas TSRs no genoma humano, com o intuito de obter sequências expandidas, considerando 100 kbp antes e depois de cada sequência mapeada. Aplicando um software validado para predição de genes, GeneMark (Lomsadze et al., 2005), encontramos diversas sequências putativas, alguns dos quais apresentam similaridade com a estrutura de um retrovírus (*gag-pol-env*), neste caso, o retrovírus endógeno humano *HERV-K*. Tal retrovírus é considerado o de maior atividade no organismo humano (Flockerzi et al., 2008). Evidentendente, isto não é uma evidência forte para a transcrição porque a informação de que o *HERV-K* é um elemento genético móvel e ativo pode sugerir que a TSR está de fato sendo transcrita. Entretanto, por considerarmos apenas o genoma de referência e por *HERV-K* ser uma sequência repetitiva, também é possível que uma nova inserção tenha ocorrido na região do transcrito candidato. Consequentemente, tal inserção invalidaria a hipótese de que o transcrito foi gerado após um evento de *trans-splicing*, e sim por retrotransposição do *HERV-K* que carrega consigo a região TSR da sequência candidata, conforme ilustração da Figura 45.

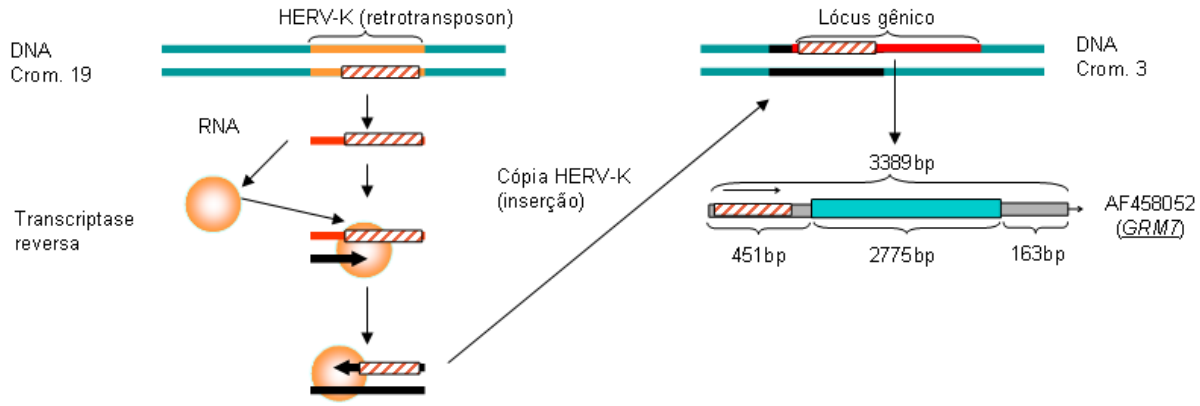


*Figura 45 – Hipótese de retrotransposição do retrovírus HERV-K.*
*Hipótese de retrotransposição do retrovírus HERV-K para a formação do transcrito quimérico [DDBJ:AF458052] por meio de retrotransposição ao invés da ocorrência do mecanismo de* trans-splicing.
*Fonte: da própria pesquisa.*

Para os transcritos deste cluster, em especial, a probabilidade da ocorrência de erros de clonagem ou da inserção de artefatos experimentais na geração das sequências é ligeiramente reduzida pelo fato de que foram geradas por dois laboratórios independentes. A sequência [DDBJ:U92458] foi gerada nos Estados Unidos (Wo et al., 1998), e as demais por um grupo da Alemanha (Schulz et al., 2002).

Com o intuito de tentar validar experimentalmente os transcritos candidatos obtidos em humano, reações de polimerização em cadeia (PCR (Regitano & Coutinho, 2001)) foram realizados com dois pesquisadores do Centro Infantil de Investigações Hematológicas Dr. Domingos A. Boldrini, o Dr. José Andrés Yunes e seu aluno de doutorado André Bortolini Silveira. Foi feito o desenho e aquisição de *primers* dos tipos *reverse* e *forward* de cada um dos transcritos candidatos detectados pela metodologia Fusion5Finder. O desenho de cada *primer* foi feito de forma que flanqueassem a junção TSR/transcrito, com o intuito de deixar a junção na região central do amplificado que se deseja obter para posterior sequenciamento. Para os experimentos, foram obtidos transcritos de cDNA extraídos a partir de amostras de linhagens celulares de leucemia linfóide aguda (LLA) do centro de pesquisa mencionado. A partir dos *primers* e dos transcritos de cDNA, foi realizado o PCR. Ao final dos experimentos, nenhum dos candidatos foi encontrado. Algumas hipóteses, dentre várias, que podem explicar isso são:

- impossibilidade de se utilizar células provenientes dos mesmos tecidos que os transcritos candidatos foram originalmente obtidos;

- transcritos candidatos não são tão freqüentes como aqueles que não são quiméricos, o que demandaria um experimento com enriquecimento de biblioteca de cDNA;

- transcritos candidatos não são expressos na LLA;
- candidatos descritos são oriundos de artefatos das bibliotecas de cDNA de que foram obtidas;
- estratégia adotada para o desenho de *primers* para os candidatos quiméricos não foi boa, entre outros.

### 5.1.2 Transcritos quiméricos em Bovino

Diferentemente de humanos, em que já há evidências experimentais da formação de transcritos quiméricos por meio de *trans-splicing* intercromossomal, em bovinos há apenas um caso descrito por Roux et al. (2006). Entretanto, conforme discutido no Capítulo 3, tal evidência pode ter sido resultante de um evento de *splicing* alternativo, o que parece mais provável.

Desta forma, as seções seguintes apresentam características dos dados utilizados e os resultados obtidos na busca por transcritos quiméricos em bovinos com o uso da metodologia estendida, chamada FusionAllFinder. Vale lembrar que tal metodologia estendida foi desenvolvida com o intuito evitar a geração de uma quantidade enorme de candidatos falso-positivos, como ocorreu com a metodologia Fusion5Finder. Esta última, além de limitar a região de busca, foi desenvolvida para dados com alta qualidade e, desta forma, produziu para este caso uma quantidade muito grande, na ordem de algumas centenas, de candidatos falso-positivos. Os candidatos detectados foram analisados de maneira similar com o que foi feito naqueles que foram descritos em humanos. Além disso,

em função da similaridade de parte dos resultados, a discussão para bovinos será resumida, pois o objetivo principal é demonstrar a aplicabilidade da metodologia proposta na detecção de possíveis transcritos quiméricos, para que possam direcionar pesquisas futuras para sua validação experimental.

#### 5.1.2.1 Características dos dados

Para testar a metodologia estendida na busca por transcritos quiméricos em bovinos, foi escolhido o genoma da espécie *Bos taurus*, versão UMD 3.0. Tal versão é a mais recente e, até então, possui cobertura superior a 95% (Zimin et al., 2009). Além disso, foram utilizadas 3 bibliotecas de transcritos de *full-length* cDNA geradas por consórcios internacionais independentes que totalizam 83.048 transcritos de cDNA. Tais bibliotecas são descritas a seguir:

1. MGC (*Mammalian Gene Collection*): gerado pelo projeto Genoma Canadá por meio de uma iniciativa liderada pela agência NIH (*National Institutes of Health*) do Departamento de Saúde dos Estados Unidos (Temple et al., 2009), tal base é composta por 8.604 transcritos;

2. BGD (*Bovine Genome Database*): gerada por um consórcio liderado pelas universidades de Georgetown, Adelaide e de Minnesota, todas localizadas nos Estados Unidos da América, foram também responsáveis pela montagem do genoma de referência da espécie bovina *Bos taurus*, build 4.1 (Elsik et al., 2009). Tal base de dados é composta por 50.467 transcritos;

3. NCBI-RefSeq (*NCBI Reference Sequence*): gerada pela agência NIH (*National Institutes of Health*) do Departamento de Saúde dos Estados Unidos (Pruitt et al., 2007), tal base é composta por 23.977 transcritos.

### 5.1.2.2 Resultados e discussão de candidatos quiméricos em Bovino

A partir da definição dos dados, a metodologia estendida FusionAllFinder encontrou satisfatoriamente 13 possíveis candidatos de transcritos quiméricos, gerados, possivelmente, por *trans-splicing* intercromossomal. Eles são listados na Tabela 5.

*Tabela 5 — Candidatos quiméricos em Bovino.*
*Transcritos quiméricos candidatos que foram identificados nas três bases de dados de transcritos de* full-length *cDNA bovino. As 11 sequências candidatas são listadas com as seguintes informações: Número de Acesso (AC), Trans-Spliced Region (TSR), regiões não traduzidas 5' e 3' (5' e 3' UTR), região codificante (CDS), tamanho do transcrito de RNA mensageiro (mRNA), e tecido de origem (N/A indica que a informação não foi encontrada).*

| AC | TSR | 5'UTR | CDS | 3'UTR | mRNA | Tecido |
|---|---|---|---|---|---|---|
| [RefSeq:NM_203358] | 1-341 | 376 | 729 | 481 | 1586 | N/A |
| [RefSeq:NM_174719] | 1-357 | 359 | 1488 | 109 | 1956 | N/A |
| [RefSeq:NM_001015598] | 1847-2104 | 38 | 1995 | 71 | 2104 | N/A |
| [RefSeq:NM_001080267] | 1-322 | 208 | 645 | 1440 | 2293 | N/A |
| [MGC: BC113235] | 638-771 | 162 | 528 | 75 | 765 | Utero |
| [RefSeq:NM_001078147] | 638-771 | 162 | 528 | 75 | 765 | N/A |
| [MGC:BC133483] | 1882-2810 | 25 | 2577 | 248 | 2850 | Ventrículo do coração |
| [RefSeq:NM_001038213] | 1882-2810 | 25 | 2577 | 248 | 2850 | N/A |
| [MGC:BC134601] | 1-1657 | 499 | 795 | 2869 | 4163 | Pele de feto |
| [MGC:BC148157] | 1125-1487 | 265 | 726 | 496 | 1487 | Hipotálamo |
| [MGC:BC112810] | 1198-1614 | 488 | 648 | 583 | 1719 | Ventrículo do coração |
| [MGC:BC105254] | 2095-3665 | 43 | 852 | 2794 | 3689 | Intestino grosso |
| [RefSeq:NM_174055] | 3674-4005 | 917 | 468 | 2620 | 4005 | N/A |

*Fonte: da própria pesquisa.*

A análise dos transcritos candidatos permitiu, em bovinos, detectar as três categorias de transcritos quiméricos descritos na Figura 20, cuja fusão entre a TSR e o gene anotado localiza-se na região 5'UTR, CDS ou 3'UTR. Conforme Tabela 5, dois candidatos possuem TSR na região 5'UTR (destacadas em branco), outros 6 possuem TSR na região codificante

CDS (destacadas em verde), e outros 5 possuem a TSR na região 3'UTR (destacadas em laranja).

A busca pelas sequências candidatas, assim como em humanos, também não impôs a existência dos *splice-sites* canônicos. É possível observar que, conforme dados da Tabela 6, somente um dos candidatos, o transcrito [MGC:BC133483], possui *splice-sites* canônicos. O fato dos demais candidatos não possuírem *splice-sites* canônicos pode ser um indício de que podem pertencem a uma categoria rara de *splice-sites*, visto que o fenômeno de *trans-splicing* também é raro. Outra hipótese que não pode ser descartada é a possibilidade das outras sequências terem sido formadas por algum outro tipo de mecanismo, como aquele mediado pelas SR reversa complementares, similar ao que foi apresentado em uma molécula de tRNA (Capítulo 3), ou até mesmo por retrotransposição, como no caso do *HERV-K* (Figura 45).

*Tabela 6 – Splice-sites das sequências quiméricas candidatas de bovinos.*
*A primeira coluna contém o Número de Acesso (AC) das sequências candidatas, seguida pela sua localização em algum cromossomo (AC Crom), o cromossomo cuja região TSR está localizada (TSR Crom), o dinucleotídeo que representa o* splice-site *de cada TSR (TSR SP, em negrito) e do restante do transcrito quimérico (Transcrito SP, em negrito), respectivamente*

| AC | AC Crom | TSR Crom | TSR SP | Transcrito SP |
|---|---|---|---|---|
| [RefSeq: NM_203358] | 1 | 19 | TSR-**AC** | **AC**-TSR |
| [RefSeq: NM_174719] | 21 | 5 | TSR-**AC** | **GG**-TSR |
| [RefSeq: NM_001015598] | 22 | 18 | **AT**-TSR | TSR-**AG** |
| [RefSeq: NM_001080267] | 4 | 13 | TSR-**GA** | **GC**-TSR |
| [MGC: BC113235] | 6 | 3 | **CC**-TSR | TSR-**AT** |
| [RefSeq:NM_001078147] | 6 | 3 | **CC**-TSR | TSR-**AT** |
| [MGC: BC133483] | 3 | 5 | **AG**-TSR | TSR-**GT** |
| [RefSeq:NM_001038213] | 3 | 5 | **AG**-TSR | TSR-**GT** |
| [MGC: BC134601] | X | 3 | TSR-**CT** | **AT**-TSR |
| [MGC: BC148157] | 22 | 20 | **GG**-TSR | TSR-**AC** |
| [MGC: BC112810] | 15 | 10 | **CC**-TSR | TSR-**AA** |
| [MGC: BC105254] | 27 | 26 | **GC**-TSR | TSR-**GT** |
| [RefSeq: NM_174055] | 7 | 3 | **AC**-TSR | TSR-**AT** |

*Fonte: da própria pesquisa.*

Com o objetivo de verificar se a região genômica da TSR é transcrita para que a sequência quimérica possa ser formada, foi feita uma análise das TSR dos 11 (2 sequências são redundantes) candidatos encontrados pela metodologia (evidências de transcrição ET são listadas na Tabela 7).

Isso permitiu identificar dois casos: (*i*) TSR que mapeia no lócus gênico de outros transcritos; (*ii*) TSR que mapeia numa região genômica de um elemento genético móvel (BCNT2) e de uma classe de gene de transcriptase reversa (RTLf).

*Tabela 7 – Candidatos quiméricos em Bovino e suas evidências de transcrição.*
*Transcritos quiméricos candidatos de* trans-splicing *intercromossomal detectados em bovinos pela ferramenta FusionAllFinder. A coluna AC indica o número de acesso do transcrito quimérico, e a coluna ET indica a evidência de transcrição da região TSR.*

| AC | ET |
|---|---|
| [RefSeq:NM_203358] | BC105140 |
| **[RefSeq:NM_174719]** | **-------** |
| [RefSeq:NM_001015598] | BT021600 |
| [RefSeq:NM_001080267] | BC149975 |
| [MGC: BC113235] | BC133513 |
| [RefSeq:NM_001078147] | BC133513 |
| [MGC:BC133483] | TRMU |
| [RefSeq:NM_001038213] | TRMU |
| [MGC:BC134601] | NM_001083790 |
| [MGC:BC148157] | BC102850 |
| [MGC:BC112810] | BC148136 |
| [MGC:BC105254] | BC103009 |
| [RefSeq:NM_174055] | BC140622 |

*Fonte: da própria pesquisa.*

Para o primeiro caso, conforme já visto em humanos, tem-se uma evidência que de fato a região genômica de origem de cada TSR sofre transcrição, para que a sequência quimérica possa ser formada. Além disso, com o uso da ferramenta RepGraph, foi possível identificar uma grande quantidade de pares de SR, como as reversas complementares que podem favorecer a fusão de dois transcritos, porém de comprimento menor. Apesar do

tamanho reduzido se comparado com os casos que ocorrem em humanos, é fato que, mesmo o menor tamanho encontrado entre todos os candidatos, de 15 pb, tais SR podem permitir que ocorra uma união entre os dois transcritos, como no mecanismo descrito em endonucleases de tRNA. Tamanhos de SR da ordem de 300pb, como ocorrem nos candidatos identificados em humanos não são esperados em bovinos, visto que SR deste tamanho costumam estar relacionadas com as ALUs, exclusivas de organismos primatas.

Os casos em que a região de fusão dos transcritos candidatos encontra-se na região codificante (CDS, destacados em verde na Tabela 5) merecem especial atenção, pois é possível que tal fenômeno, se de fato confirmado, tenha gerado uma nova sequência codificadora para uma proteína específica, que depende exclusivamente da existência do transcrito quimérico. Para verificar isso, duas etapas foram realizadas. Na primeira etapa, localizamos os dois genes que possivelmente formaram o transcrito quimérico candidato. Para tal, utilizamos o *UCSC Genome Browser*, mapeamos cada um dos 6 candidatos no genoma bovino para tentar identificar transcritos anotados na região genômica de cada um deles, e que compreendam uma região genômica maior. Dos 6 candidatos, apenas para o candidato [MGC:BC133483] (vide Figura 46) foi possível identificar outro gene em sua região genômica, o POGZ, cujo transcrito é o [RefSeq:NM_001163190], e que nomearemos de transcrito gerador.

Na segunda etapa, considerando o transcrito gerador [RefSeq:NM_001163190] e a evidência de transcrição da sua TSR, o transcrito [RefSeq:NM_001163188], analisamos a fusão entre ambas para verificar se uma nova proteína pode ter sido de fato gerada, ou se ocorreu algum evento que causasse um sinal de parada (*stop códon*) prematuro no processo de tradução do transcrito para uma proteína. A análise deste caso é feita com base na Figura

47. Para a formação do transcrito quimérico, pode-se observar claramente que 1881 pb do quimérico são provenientes de éxons completos do transcrito [RefSeq:NM_001163190]. O restante do transcrito quimérico, a partir da sua posição 1882, é proveniente do transcrito [RefSeq:NM_001163188], na qual é considerada uma região parcial de um éxon.

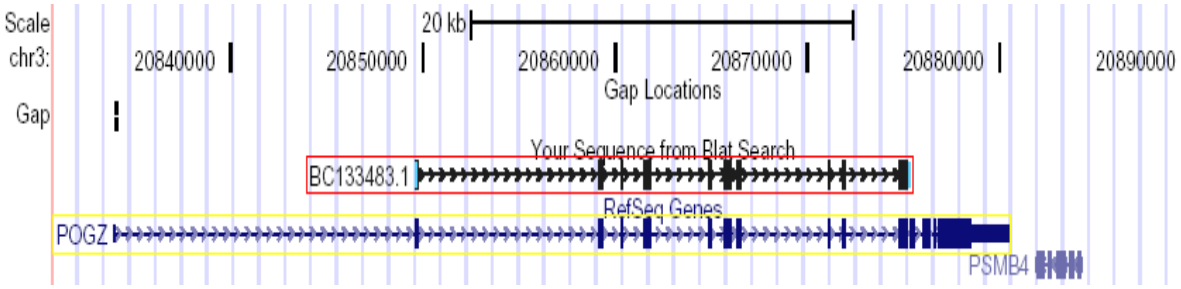


*Figura 46 – GBrownser do candidato quimérico [MGC:BC133483].*
*Mapeamento do transcrito candidato [MGC:BC133483] no genoma bovino para identificação de outro transcrito em sua mesma região genômica. Em vermelho, o mapeamento do transcrito quimérico candidato. Em amarelo, o gene POGZ cujo transcrito está em uma mesma região genômica do candidato quimérico.*
*Fonte: da própria pesquisa, gerado com o uso da ferramenta UCSC Genome Browser.*

A partir disso, buscamos as proteínas geradas pelos transcritos [RefSeq:NM_001163190] e [RefSeq:NM_001163188] e identificamos as proteínas anotadas [RefSeq:NP_001156662] e [RefSeq:NP_001156660], respectivamente. Analisamos também a sequência de resíduos gerada pela região codificante do transcrito quimérico candidato, e a mapeamos contra os resíduos das duas proteínas anteriores, e obtivemos como resultado a proteína [MGC:AAI33484], com 858 resíduos, e que se encontra anotada como uma proteína predita no NCBI. Pela análise, confirmamos que parte dela, entre os aminoácidos 1 a 618, é proveniente da proteína real [RefSeq:NP_001156662], e o restante, entre as posições 620 a 858, é proveniente da proteína real [RefSeq:NP_001156660]. Na posição 619, um novo resíduo foi formado, sendo necessária uma maior invetigação para verificar as implicações disso.

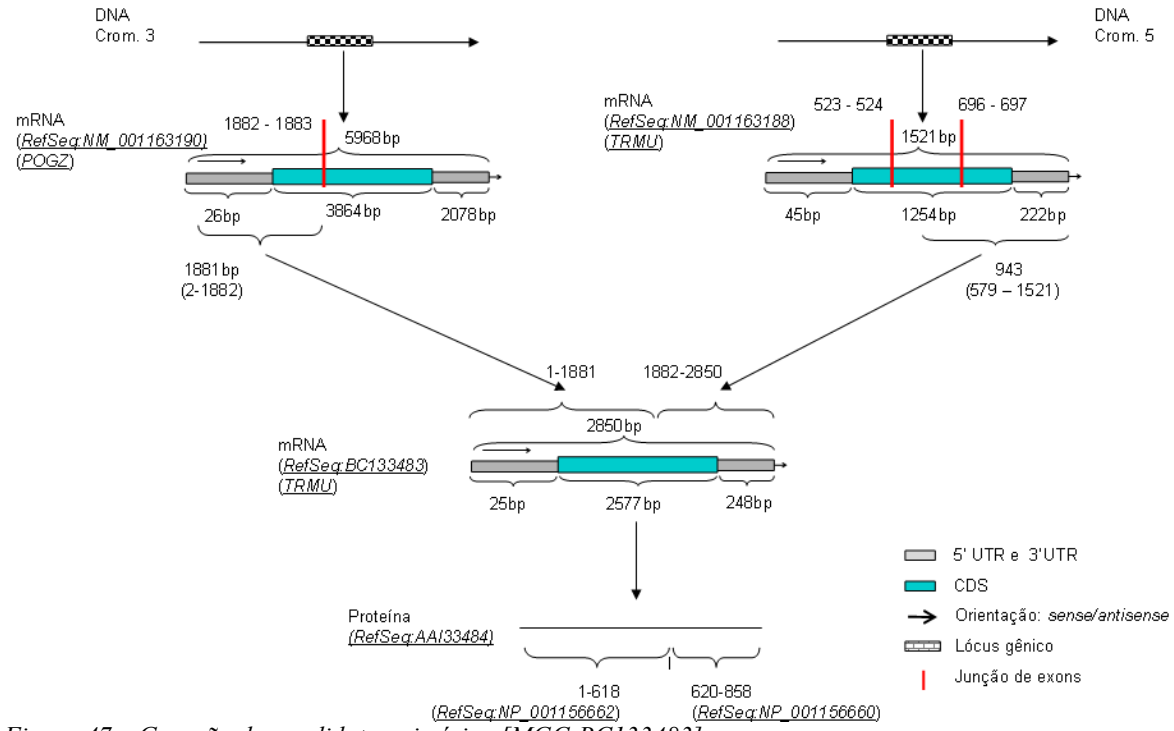


*Figura 47 – Geração do candidato quimérico [MGC:BC133483].*
*Transcrito quimérico candidato [MGC:BC133483] gerado possivelmente a partir da fusão dos transcritos [RefSeq:NM_001163190] e [RefSeq:NM_001163188]. Cada transcrito possui uma proteína relacionada, inclusive o transcrito quimérico candidato, em que parte de seus resíduos são da proteína [MGC:AAI33484] gerada pelo transcrito transcrito [RefSeq:NM_001163190] e outra parte da proteína do transcrito [RefSeq:NM_001163188]. A fusão entre ambas gera um novo aminoácido na posição 619 da proteína predita. Não foi verificado se a fusão ocorre antes ou depois do splicing.*
*Fonte: da própria pesquisa.*

Além da hipótese de *trans-splicing* levantada, outra explicação para a formação do transcrito quimérico candidato é a retrotransposição do transcrito [RefSeq:NM_001163190] para a região genômica do transcrito [RefSeq:NM_001163188]. Após o evento de retrotransposição, o transcrito quimérico seria formado. Este caso é bastante similar ao do transcrito candidato que será tratado no parágrafo seguinte.

Retomando os casos encontrados, no segundo, temos uma TSR do transcrito [RefSeq:NM_174719] que é mapeada na região genômica de 3 elementos distintos. Para que a identificação de tais elementos fosse possível, a TSR do candidato foi mapeada no genoma bovino, com o intuito de obter uma sequência expandida (100kpb antes e depois da

TSR mapeada). O tamanho da sequência expandida foi selecionada com base no tamanho genômico médio observado dos transcritos da base de dados. Entretanto, vale destacar que uma maior investigação é necessária para determinar o tamanho da sequência expandida a ser analisada. Aplicando novamente o software GeneMark, foram encontradas diversas sequências putativas de transcritos, porém é pouco esperado que tenha alta similaridade com sequências identificadas e confirmadas em laboratório, caso contrário já teriam sido anotadas. De todas elas, três sequências possuem alta identidade com três elementos presentes no organismo bovino. O primeiro é um elemento genético móvel, o BCNT2. A segunda, curiosamente, é de um gene de transcriptase reversa, o RTLf. Notadamente, ambos possuem forte relação, pois já é sabido que, para um elemento genético móvel poder copiar-se para diferentes partes de um genoma, como o NCNT2, necessariamente precisa de um gene de transcriptase reversa, como o RTLf, para que seja gerada uma molécula de DNA dupla-fita para posterior reintegração no genoma. Coincidentemente, tal observação é similar a um dos casos encontrado em humanos, quando se tratava de um retrovírus endógeno humano que é ativo, o HERV-K. Apesar desta observação, é necessária uma investigação mais aprofundada para verificar se de fato tal região pode ser transcrita, e também se o elemento genético móvel, caso realmente exista, está de fato ativo para que a formação de um transcrito quimérico possa ocorrer. É importante destacar que nesse caso, o transcrito quimérico pode ter sido formado por retrotransposição, e não por *trans-splicing*. Essa observação é equivalente ao que foi reportado na Figura 45.

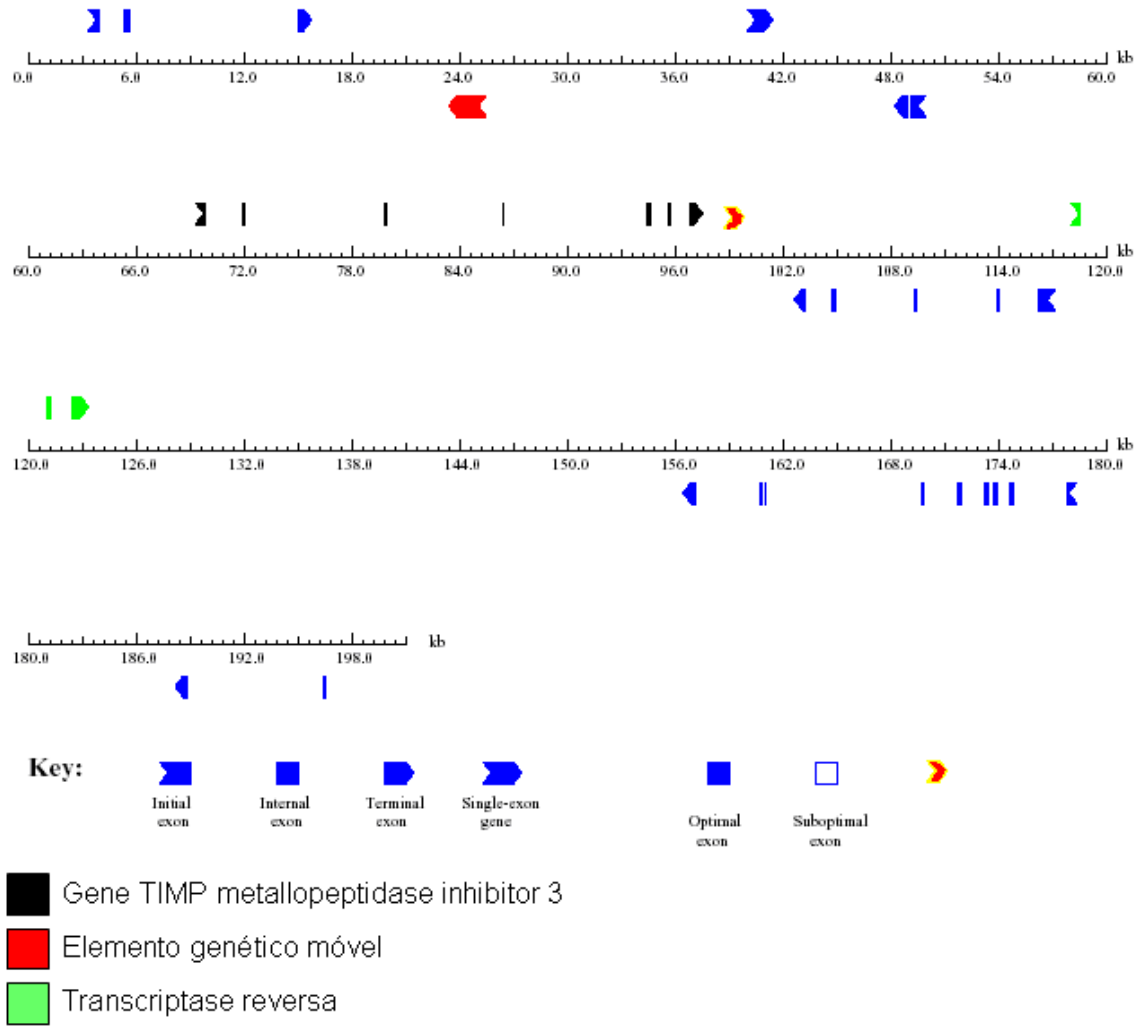


*Figura 48 – Região genômica do candidato quimérico Bovino [RefSeq:NM_174719].*
*Região genômica com 100kpb da porção* upstream *e* downstream *da região TSR do transcrito [RefSeq:NM_174719].*
*Fonte: da própria pesquisa.*

Um terceiro elemento encontrado e que também localiza-se na região genômica estendida da TSR é uma sequência que apresenta alta identidade com uma proteína chamada TIMP, responsável pela inibição de *metallopeptidase* do tipo 3 em bovinos. A Figura 48 ilustra a localização dos três elementos citados com relação à região genômica da TSR do transcrito [RefSeq:NM_174719].

Uma possibilidade de contaminação por vetores ou por artefatos experimentais para os transcritos candidatos bovinos é reduzida porque foi utilizada a ferramenta VecScreen e nada foi encontrado. Além disso, como nos casos de humanos, há casos de transcritos que foram anotados por laboratórios distintos, o que reduz consideravelmente a possibilidade de terem sido gerados por um mesmo tipo de erro. Duas sequências geradas pelo MGC: [MGC:BC113235] e [MGC:BC133483] também foram geradas pela fundação RefSeq do NCBI, sob os códigos de acesso [RefSeq:NM_001078147] e [RefSeq:NM_001038213], respectivamente.

Apesar das evidências encontradas de que SR podem favorecer a formação dos transcritos candidatos quiméricos detectados em humanos e bovinos, é necessário responder à pergunta se a frequência de SR, dos 4 tipos distintos, é predominante entre os pares de transcritos envolvidos na formação do quimérico ou é esperada para qualquer região genômica (ou entre pares de transcritos quaisquer). Para responder a essa pergunta foi feito um estudo, com base na metodologia apresentada na seção 4.3, para verificar a quantidade média das SR dos tipos direta, reversa, direta complementar e reversa complementar entre diferentes tipos de sequências que será apresentado na secção seguinte. Vale ressaltar que essa análise é apenas superficial, pois a quantidade de sequências quiméricas é muito pequena, o que dificulta muito o levantamento de novas hipóteses a respeito da formação de tais sequências.

Com o intuito de também tentar validar experimentalmente os transcritos candidatos obtidos em bovino, reações de polimerização em cadeia (PCR), como aquelas que foram realizadas em humanos, já estão programadas com duas pesquisadoras da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), das subdivisões Embrapa Informática

Agropecuária e Embrapa Pecuária Sudeste, Dra. Poliana F. Giachetto e Dra. Luciana C. A. Regitano, respectivamente. Para os experimentos envolvendo os candidatos quiméricos detectados em bovino, serão utilizadas células oriundas dos mesmos tecidos em que os transcritos foram anotados, para aumentar a probabilidade de detecção, visto que é possível que os candidatos existam somente nos tecidos que foram originalmente anotados.

## 5.2 Frequência de pares de SR em *loci* gênicos de transcritos quiméricos

Uma questão importante diz respeito à frequência de SR nas regiões gênicas dos transcritos quiméricos. Será essa frequência igual à frequência de SR em regiões genômicas quaisquer? Além disso, haverá diferença entre a frequência de SR em regiões gênicas de transcritos quiméricos e de não quiméricos? São algumas perguntas que procuramos responder por meio de algumas simulações computacionais.

Há, entretanto, três aspectos que devem ser considerados nesses tipos de estudo:

1. Tamanho das SR;
2. Tamanho dos transcritos envolvidos;
3. Quantidade de pares de transcritos.

É de se esperar que SR menores ocorram com maior frequência que as maiores. Entretanto, o número de SR de um determinado tamanho pode ou não ocorrer para um par específico de transcritos, o que dificultaria a comparação. Para evitar esse tipo de problema, as SR foram agrupadas em faixas de tamanho 50, sendo o último intervalo composto pelas SR maiores do que 400 pb.

A frequência de SR está diretamente relacionada ao tamanho dos transcritos envolvidos. Quanto maiores os transcritos, maior a frequência de SR. Como nossas sequências candidatas possuem transcritos geradores com tamanhos muito diferentes, seria incorreto agrupá-las indiscriminadamente. Além disso, a quantidade de transcritos candidatos, por ser menor que duas dezenas não é suficiente tanto em humano quanto em bovino. Por esses dois motivos, resolvemos tratar as frequências fixando-se o tamanho dos transcritos geradores. Para facilitar a compreensão, tomemos um exemplo real. O transcrito quimérico [DDBJ:U09825] é gerado por um par de transcritos que têm tamanho 20.312 pb e 76.652 pb, representados pelo candidato (TQ) e pela evidência de transcrição (ET) da região da TSR, respectivamente. Com base em tal transcrito e na evidência de transcrição ET da sua respectiva região TSR, foram realizadas, conforme metodologia apresentada na seção 4.3.1, três análises computacionais utilizando a ferramenta FreqRepeat para comparar as quantidades de SR do par (TQ,ET) com relação às médias de SR dos pares (GR,GR), (TR,GR) e (TR,TR), também definidos na seção 4.3.1. A partir disso, foram gerados 100 pares aleatórios, cujo tamanho da primeira sequência do par é igual ao de TQ, e da segunda igual ao de ET. Em seguida, computamos a média para cada um dos pares para comparar com a quantidade de SR observadas no par (TQ,ET). Para cada um dos candidatos quiméricos encontrado, tanto em humano quanto em bovino, é feita a comparação entre o par (TQ,TE) contra os outros tipos de pares de SR. Em humano, foram 12 testes, e em bovino 10, pois foram consideradas somente aquelas sequências candidatas que possuem uma evidência de transcrição ET da sua região TSR.

Para tal, foram utilizados os mesmos dados considerados pelos experimentos de detecção de transcritos quiméricos candidatos nos organismos humano e bovino.

Outras combinações, envolvendo sequências de nucleotídeos geradas de forma aleatória (sequências completamente aleatórias) também foram testadas, mas como esperado, nenhuma SR de tamanho considerável (maior do que 5 pb) foi encontrada, dada a natureza aleatória de tais sequências.

Os resultados obtidos permitiram comprovar que, para os dois organismos considerados, a frequência média de SR depende fortemente do tamanho dos transcritos envolvidos. Ou seja, quanto menor o tamanho de uma das sequências envolvidas na formação de um transcrito quimérico, menor a quantidade de SR encontrada. A *Tabela 8* lista as SR do tipo reversa complementar (RRC) entre os candidatos quiméricos de humano. O mesmo padrão, com poucas sequências, é encontrado nas SR do tipo direto, complementar e reverso.

*Tabela 8 – SR do tipo RRC nos candidatos quiméricos em Humano.*
*Quantidade de SR do tipo Repetição Reversa Complementar (RRC) entre os pares de transcritos que formam os candidatos em humanos. As sequências estão ordenadas pelo tamanho do loci* gênico *do transcrito candidato (tamanhos inferiores a 10kpb estão destacados em vermelho). AC: Accession Number; AC tamanho: tamanho da sequência AC; ET: tamanho da sequência da evidência de transcrição.*

| AC | AC tamanho | ET | 50 | 100 | 150 | 200 | 250 | 300 | >350 |
|---|---|---|---|---|---|---|---|---|---|
| [DDBJ:AK124366] | **2485** | 29816 | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| [DDBJ:AB007865] | **4860** | 1220142 | 23 | 0 | 0 | 0 | 0 | 0 | 0 |
| [DDBJ:AL834489] | **5362** | 69475 | 13 | 0 | 0 | 0 | 0 | 0 | 0 |
| [DDBJ:AF003522] | **8003** | 274865 | 11 | 0 | 0 | 0 | 0 | 0 | 0 |
| [DDBJ:AK130557] | 10805 | 25459 | 36 | 14 | 0 | 5 | 5 | 4 | 0 |
| [DDBJ:U09825] | 20312 | 76652 | 160 | 48 | 19 | 7 | 17 | 43 | 0 |
| [DDBJ:AK226066] | 43170 | **7068** | 15 | 0 | 0 | 0 | 0 | 0 | 0 |
| [DDBJ:L33075] | 113971 | 303954 | 2058 | 817 | 394 | 163 | 370 | 362 | 3 |
| [DDBJ:L14837] | 122364 | 51500 | 977 | 315 | 236 | 113 | 158 | 193 | 0 |
| [DDBJ:D26155] | 163977 | 99001 | 881 | 346 | 282 | 84 | 196 | 194 | 2 |
| [DDBJ:AB023216] | 252993 | 78568 | 1855 | 634 | 413 | 116 | 209 | 423 | 2 |
| [DDBJ:AB020656] | 564444 | 12569 | 602 | 234 | 140 | 66 | 81 | 61 | 0 |

*Fonte: da própria pesquisa.*

No gráfico da Figura 49 é ilustrada a média de SR entre os pares de sequências que formam o transcrito quimérico [DDBJ:AB023216] em humano. Um ponto importante a

ser observado no gráfico, é que a média é decrescente, entretanto, por volta do tamanho 300, volta a crescer e retoma o decrescimento até torna-se zero.

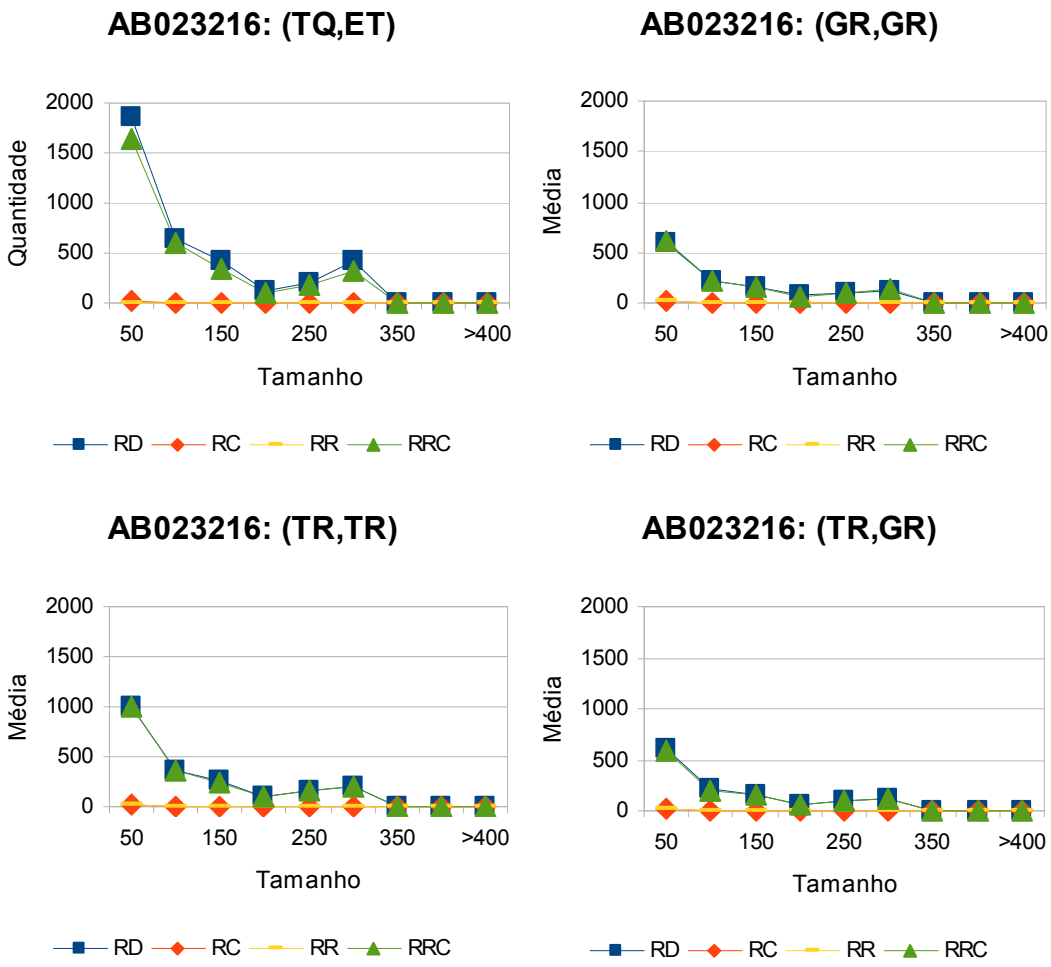


*Figura 49 – SR entre sequências relacionadas com o transcrito [DDBJ:AB023216].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:AB023216]. Os tipos detectados são: repetição direta (RD), repetição reversa (RR), repetição complementar (RC) e repetição reversa complementar (RRC). Conforme o gráfico: par (TQ,ET): quantidade de SR entre o transcrito quimérico candidato e a evidência de transcrição da sua região TSR; par (GR,GR): média de SR entre duas sequências reais do genoma, porém de posições aleatórias; par (TR,GR): média de SR entre o loci gênico de um transcrito real selecionado aleatoriamente e uma sequência genômica real de uma posição aleatória do genoma; par (TR,TR): média de SR entre o loci gênico de dois transcritos reais selecionados aleatoriamente de uma base de transcritos. Todos os pares são formados por sequências cujo tamanho da primeira é equivalente e TQ, e o da segunda equivalente a ET.*
*Fonte: da própria pesquisa.*

Esse crescimento pode ser explicado pela presença de sequências ALU-like. Este valor médio é valido para todos os casos, exceto quando o tamanho de uma das sequências

é inferior a cerca de 10 kpb. Além disso, foi possível observar que, em tais sequências, cujo tamanho é inferior a 10 kpb, a quantidade de SR do tipo RD e RRC é sempre menor se comparado com a média dos outros tipos de pares de sequências (vide gráficos do Apêndice A). A maior abundância de SR dos tipos RD e RRC encontra-se em investigação, e será tratada em um trabalho futuro a este.

O mesmo tipo de análise realizada para verificar a frequência de SR entre diferentes tipos de pares de sequências de humano foi aplicada nas sequências quiméricas candidatas de bovino. Nesse organismo, pode-se verificar que o tamanho do *locus* gênico das sequências quiméricas candidatas é, se comparada ao de humanos (conforme apresentado na Tabela 8), menor para a maioria dos casos. Dos 10 pares (TQ,ET) formados pelo transcrito quimérico e a respectiva evidência de transcrição da TSR, apenas dois são formados por um par onde ambas são maiores do que 10 kpb, mas que não ultrapassam o tamanho de 26 kpb (vide Tabela 9). Em consequência direta disso, a quantidade de SR detectadas foi bastante inferior às quantidades observadas em humano.

*Tabela 9 – SR do tipo RRC nos candidatos quiméricos em Bovino.*
*Quantidade de SR do tipo Repetição Reversa Complementar (RRC) entre os pares de transcritos que formarm os candidatos em bovinos. As sequências estão ordenadas pelo tamanho do lócus gênico do transcrito candidato (tamanhos inferiores a 10kpb estão destacados em vermelho).*

| AC | AC tamanho | ET - tamanho | 50 | 100 | 150 | 200 | 250 | 300 | >350 |
|---|---|---|---|---|---|---|---|---|---|
| [MGC:BC113235] | **646** | **7020** | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| [MGC: BC134601] | **2475** | **2469** | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| [MGC: BC148157] | **2963** | 22872 | 7 | 1 | 0 | 0 | 0 | 0 | 0 |
| [RefSeq: NM_001015598] | **6820** | 17198 | 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| [RefSeq: NM_001080267 | **7886** | 12000 | 11 | 0 | 0 | 0 | 0 | 0 | 0 |
| [MGC: BC112810] | 22771 | 17788 | 37 | 14 | 4 | 0 | 1 | 0 | 0 |
| [MGC: BC133483] | 25624 | 15392 | 16 | 4 | 0 | 0 | 0 | 0 | 0 |
| [MGC: BC105254] | 46564 | **9320** | 22 | 2 | 2 | 0 | 0 | 0 | 0 |
| [RefSeq: NM_174055] | 96192 | **2329** | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| [RefSeq: NM_203358] | 100934 | **7246** | 24 | 0 | 0 | 0 | 0 | 0 | 0 |

*Fonte: da própria pesquisa.*

Nas análises dos dados, foi possível observar que, para todos os casos, a quantidade de SR entre o par (TQ,ET) de cada transcrito quimérico candidato com relação a média de SR formada pelo par (GR,GR) foi sempre menor, porém um pouco próxima como no caso do gráfico da Figura 50, que ilustra um desses casos. Nele é possível comparar a quantidade de SR do par (TQ,ET) referente ao transcrito quimérico candidato [MGC:BC112810] com a média de SR, considerando diferentes tamanhos, dos pares de sequências (GR,GR), (TR,GR) e (TR,TR).

Outro padrão observado é o fato de que a média de SR no par (GR,GR) ter sido sempre maior do que no par (TR,TR) e também com relação ao par (TR,GR), com exceção de um caso. Já com relação ao par (TQ,ET), não há um padrão se compararmos a quantidade observada com relação às médias dos pares (TR,TR) e (TR,GR). Uma lista completa dos gráficos que ilustram as quantidade de SR dos 10 candidatos e a médias dos pares de SR (GR,GR), (TR,GR) e (TR,TR) pode ser observada no Apêndice B.

Vale lembrar que humanos, assim como outros primatas, possuem uma classe de SR exclusiva de tais organismos, e que possuem tamanho aproximado de 300 pb. Tal característica explica o fato de que somente em humanos a quantidade de SR, conforme os gráficos demonstram, apresenta sequências relativamente grandes.

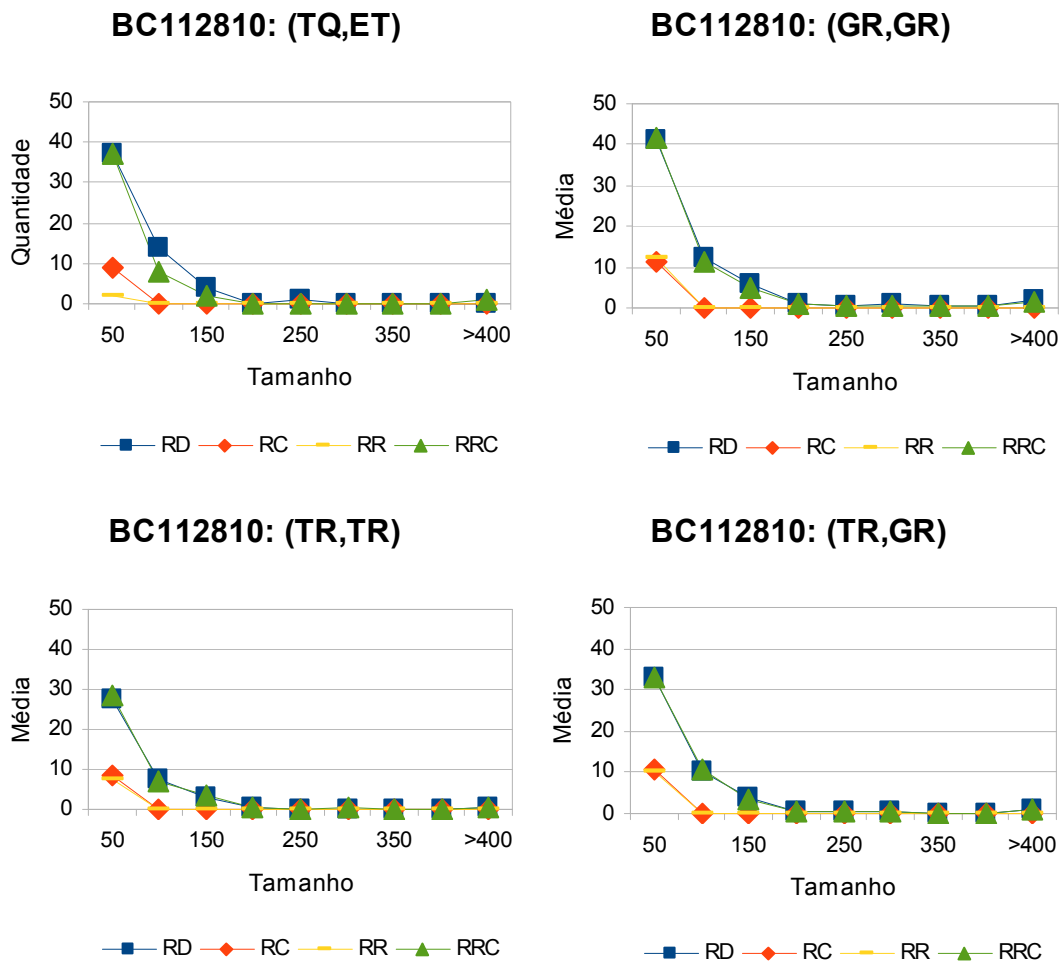


*Figura 50 – SR entre sequências relacionadas com o transcrito [MGC:BC112810].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [MGC:BC112810]. Os tipos detectados são: repetição direta (RD), repetição reversa (RR), repetição complementar (RC) e repetição reversa complementar (RRC). Conforme o gráfico: par (TQ,ET): quantidade de SR entre o transcrito quimérico candidato e a evidência de transcrição da sua região TSR; par (GR,GR): média de SR entre duas sequências reais do genoma, porém de posições aleatórias; par (TR,GR): média de SR entre o loci gênico de um transcrito real selecionado aleatoriamente e uma sequência genômica real de uma posição aleatória do genoma; par (TR,TR): média de SR entre o loci gênico de dois transcritos reais selecionados aleatoriamente de uma base de transcritos. Todos os pares são formados por sequências cujo tamanho da primeira é equivalente e TQ, e o da segunda equivalente a ET.*
*Fonte: da própria pesquisa.*

Para ambos os organismos considerados, humano e bovino, verificamos se, entre as frequências de SR observadas de cada par (TQ,ET) há diferença significativa com relação às freqüências dos pares (GR,GR), (TR,GR) e (TR,TR). Para tal, utilizamos uma análise estatística conhecida como teste do Chi-Quadrado de Pearson (Plackett, 1983). Para a aplicação de tal teste, consideramos os mesmos dados que foram utilizados para a geração

dos gráficos para análise das frequências de SR. Por meio da aplicação do teste mencionado, foi possível observar que há, tanto para humano quanto para bovino, uma diferença significativa entre as frequências existentes no par (TQ,ET) com relação aos outros 3 tipos de pares de sequências. Em humano a diferença é com relação à frequência de SR ser maior nos pares (TQ,ET) com relação aos demais tipos de pares, cujo *p-value* foi de no máximo 0,0002 em 12 dos 13 pares. Já em bovino, ocorre o oposto, em que os pares (TQ,ET) possuem frequência menor, cujo *p-value* foi de no máximo 0,05 em 10 dos 12 casos. Essa aparente contradição pode ser explicada pela forte correlação entre a frequência das SR e o tamanho das sequências geradoras, como mencionado anteriormente. Contudo, esses resultados nos impedem de concluir que, considerando apenas as frequências, as SR estejam desempenhando algum papel importante na formação de transcritos quiméricos. Mas, não podemos deixar de mencionar que bastaria uma SR ou uma classe de SR para mediar a formação de quiméricos, e o estudo de frequência não identificaria essa SR ou essa classe. Essa outra possibilidade será objeto de estudos futuros, bem como a verificação de outros pares de sequências formadas por, por exemplo, (TQ,TR) ou (ET,TR).

Outro tipo de investigação que não foi considerado, é pesquisar em genes ortólogos dos candidatos quiméricos, para ver se houve conservação nas sequências, ou se as SR foram carregadas a partir de transposons.

## 5.3 Banco de dados biológico para armazenamento de sequências repetitivas

Os testes preliminares realizados no BD biológico Rep4DB proposto foram feitos apenas para validar o uso da ferramenta de migração de dados, na qual foi utilizada uma

pequena base de dados do transcriptoma humano. Desta forma, as seções seguintes descrevem apenas esta etapa inicial do banco de dados, e que conforme será discutido na seção de trabalhos futuros, passará por novos testes para verificar sua utilização de forma prática ao que é proposto neste trabalho.

Para a realização dos testes preliminares do BD Rep4DB proposto, foram utilizadas declarações DML (linguagem de manipulação de dados) específicas da linguagem SQL, sendo os principais comandos: SELECT, INSERT, UPDATE, DELETE, COMMIT e ROLLBACK. Tais comandos foram utilizados para inicialmente inserir dados iniciais no banco de dados proposto. Para tal, foram definidos os seguintes itens:

- SGBD PostGreSQL: priorizou-se o uso de um sistema de gerenciamento de banco de dados gratuito, para facilitar, futuramente, sua distribuição em larga escala, evitando a necessidade de aquisição de licenças para sua utilização;

- Hardware: testes foram realizados com um computador convencional, composto por um Processador AMD Athlon dual core 64 X2 4200, 1.00 GHz, 512KB de memória cachê, 4 GB-RAM, HD 250GB;

- Software: foi utilizado o Sistema Operacional Linux, distribuição Fedora core 9. O software de migração desenvolvido neste trabalho foi utilizado para verificar sua validade e uso futuro para inserir dados no banco de dados;

- Fonte de dados: foi utilizada a base de transcritos de *full-length* cDNA H-InvDB, disponível atualmente somente em arquivo do tipo texto. Tal base é a mesma que foi utilizada nos experimentos de detecção de transcritos quiméricos em humanos. Ela é

composta por 187.156 sequências de transcritos, e ocupa um espaço físico de aproximadamente 2 GB.

### 5.3.1 Migração de dados

Foi feito um teste básico da ferramenta de migração MigDB. Ela segue a estrutura da Figura Figura 36, na qual os dados são provenientes de um meio externo e posteriormente são armazenados em um banco de dados.

Após a conclusão do processo de migração de dados para as tabelas do BD construído, o banco de dados passou a ocupar um tamanho aproximado de 394 MB no disco rígido do computador considerado. Na Tabela 10 são listadas algumas das principais tabelas do banco de dados e a quantidade de registros que foi inserida em cada uma delas. O tempo gasto pela aplicação, para completar a análise da fonte de dados e migrar seu conteúdo para o banco proposto foi de aproximadamente 29 horas. Uma análise simplificada do consumo de recursos do computador utilizado durante todo o processo de migração utilizou uma quantidade máxima de 115 MB de memória RAM, e a utilização da CPU não ultrapassou 10%.

Para o exemplo em questão, é importante mencionar que apesar de ter sido feito com uma fonte de dados do transcriptoma humano, sua inserção no banco de dados foi possível porque ela permite que regiões não anotadas também sejam cadastradas (novas sequências). Além disso, é possível realizar a inserção de trechos de cromossomos que não foram finalizados (no exemplo, foram anotados 47 trechos de cromossomos do genoma

humano, que possui apenas 22 pares de cromossomos diplóides, além dos cromossomos X e Y).

*Tabela 10 - Quantidade de registros nas tabelas que possuem associação com a base de dados.*

| Nome da tabela | Quantidade de registros |
|---|---|
| Transcript | 187.156 |
| Chromosome | 47 |
| SNP | 2.577.474 |
| Exon | 1.497.679 |
| UTR | 199.279 |
| Promoter | 219.765 |
| Product | 219.765 |
| Desease | 29.945 |
| Genome | 1 |
| LabInfo | 1 |

*Fonte: da própria pesquisa.*

Outro ponto importante é que, apesar do banco de dados Rep4DB proposto ter sido construído para anotação de sequências repetitivas. O mesmo pode também ser empregado para outras finalidades, como o armazenamento de outros tipos de dados de sequências biológicas, como transcriptomas (conforme teste preliminar).

## 5.4 Detecção de erros de montagem em regiões gênicas

O tratamento dos dados biológicos, principalmente os de genomas, requer na maioria das vezes verificar se de fato as sequências utilizadas não apresentam erros em sua geração. Em função disso, para verificar a qualidade dos candidatos quiméricos encontrados pela metodologia FusionAllFinder, este trabalho produziu também uma interessante ferramenta, DraftDNACheck (Herai & Yamagishi (b), 2009), para detectar um tipo específico de possíveis erros em regiões gênicas de montagem em genomas "*draft"*, conforme metodologia apresentada no Capítulo 4.

Para validação da ferramenta, foi necessário definir um genoma cuja montagem mais recente não tenha sido completamente finalizada. Em genomas cuja cobertura é baixa, é de se esperar que a metodologia detecte com sucesso regiões com erros de montagem. Entretanto, o objetivo é validar a ferramenta em genomas com altíssima cobertura, acima de 95%, pois desta forma será possível, também comprovar que ela, caso funcione, seja um procedimento padrão e adicional, dentre os vários existentes, para a validação de genomas em fase de montagem. Outro ponto importante é a disponibilidade de uma biblioteca de transcriptoma gerada a partir de técnicas que forneçam sequências com alta qualidade, como aquelas de *full-length* cDNA.

### 5.4.1 Configuração dos testes

Para a realização dos testes, foram utilizados os mesmos dados dos testes da ferramenta DraftDNACheck, descritos na seção 5.1, sendo 3 bibliotecas de transcritos de cDNA e o genoma bovino da espécie *Bos taurus*, versão UMD 3.0.

### 5.4.2 Resultados preliminares

Os dados descritos na seção anterior foram analisados pela ferramenta DraftDNACheck, que permitiu, satisfatoriamente, mesmo em um genoma com altíssima cobertura, identificar prováveis erros de montagem em regiões gênicas na versão mais recente do genoma bovino, *Bos taurus* UMD 3.0. Um possível erro de montagem foi

encontrado na região genômica do gene HDHD2 (*haloacid dehalogenase-like hydrolase domain containing 2*).

Tal erro foi reportado por DraftDNACheck porque parte do transcrito de tal gene, o [MGC:BC102232] (1629 pb), entre as posições 1 e 503 foi mapeado inteiramente no contig 965 da montagem (entre suas posições 6892 a 32797) em sentido *antisense* (Figura 51 (a)). O restante do transcrito, entre as posições 501 a 1629, foi mapeado no contig 963 (entre as posições 2303 e 11362) em sentido *sense* (Figura 51 (a)). Entretanto, se alinharmos este mesmo transcrito contra o genoma de referência *Bos taurus*, build 4.1, obtemos um alinhamento satisfatório em sentido *antisense* (Figura 51 (b)).

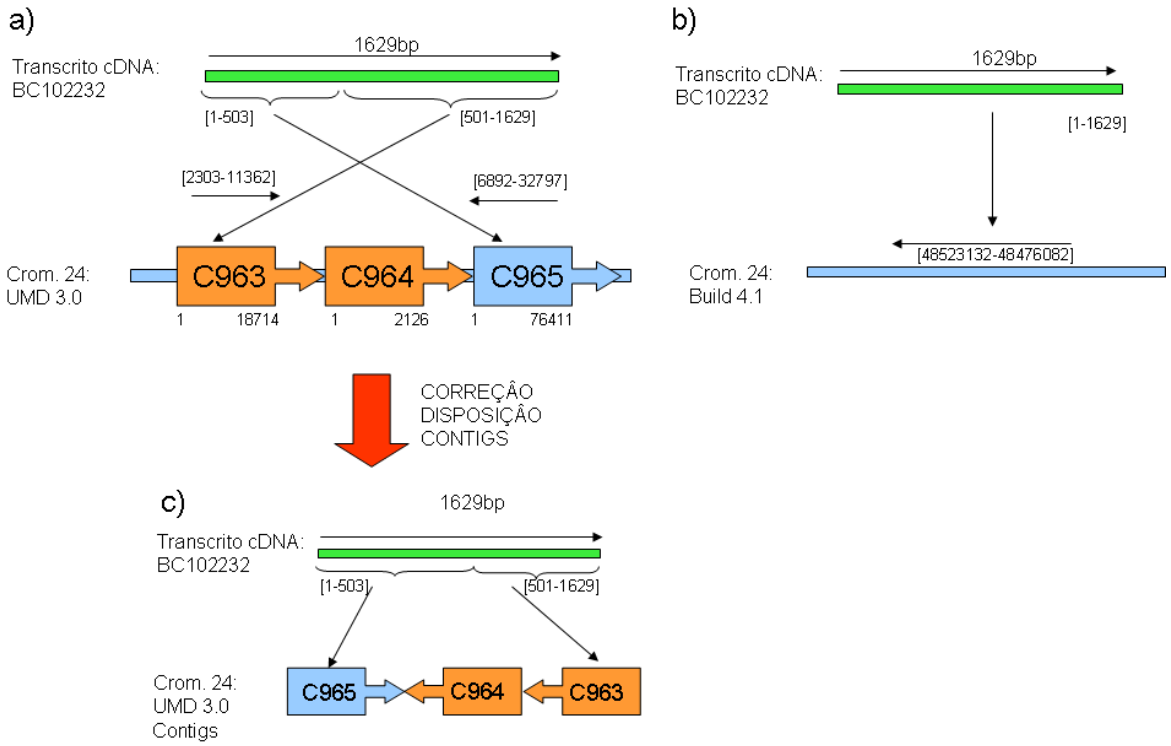


*Figura 51 - Correção de montagem do genoma Bos taurus, UMD 3.0.*
*Disposição correta que foi sugerida para os contigs C963, C964 e C965. Tais contigs estão anotados com os números 7180001732203, 7180001760913 e 7180002023644, respectivamente.a) mapeamento do transcrito contra a montagem UMD 3.0; b) mapeamento do transcrito contra a montagem de referência, build 4.1; c) correção sugerida para os contigs C963, C964 e C965 da montagem UMD 3.0.*
*Fonte: da própria pesquisa.*

Esse mesmo resultado é confirmado se comparado ao gene humano correspondente, o que reforça a possibilidade de que o genoma contenha de fato erros de montagem nos *contigs* citados. Em função disso, para este caso especial, foi possível inclusive, graças à ferramenta, propor uma correção na montagem da região mencionada. Assumindo-se que o transcrito alinha-se de fato em sentido *antisense* (assim como no genoma de referência, build 4.1), o *contig* 965 deverá manter sua orientação, devendo-se trocar a posição dos *contigs* 963 e 964, além de considerar a sequência reversa complementar correspondente, conforme Figura 51 (c).

Além do caso reportado pela Figura 51, outras evidências de erro foram encontradas e estão em análise no genoma dos organismos *Bos taurus* UMD 3.0 e *Mus musculus* ref. build 37.1.

## 5.5 Considerações do capítulo

Este capítulo discutiu o uso das metodologias de bioinformática propostas e especificadas no Capítulo 4 deste trabalho. Foi feita uma discussão a respeito da detecção de transcritos quiméricos em humanos e bovinos, gerados possivelmente por *trans-splicing* ou por retrotransposição de elementos genéticos móveis. Posteriormente, foi apresentado um estudo preliminar a respeito da média das SR, em que foi verificado que o estudo da frequência de SR, independentemente do tipo, não permite concluir que as SRs têm relevância no mecanismo de formação de transcritos quiméricos, sendo necessário outro estudo que procure identificar se existe uma classe de SR que possa mediar o fenômeno. Ao final, foi apresentado de forma muito breve a utilização da ferramenta de migração de

dados MigDB para o banco de dados proposto neste trabalho, que foi carregado com informações de uma base de transcritos humanos.

# Capítulo 6 Conclusão e trabalhos futuros

## 6.1 Conclusão

*Trans-splicing* intercromossomal é um fenômeno comum em alguns organismos inferiores, mas muito raro em organismos superiores. O fenômeno estava, na literatura, associado ao câncer, até o trabalho de Hui Li (Li et al., 2008) que pela primeira vez reportou *trans-splicing* em tecidos normais de humanos. Esse fenômeno ainda não é bem compreendido e seu mecanismo permanece obscuro. Uma conjectura afirma que o mecanismo é mediado por sequências repetitivas que poderiam aproximar os transcritos através da complementaridade de bases. Recentemente, essa conjectura ganhou força, pois em experimentos, *in vitro*, usando moléculas de tRNA, mostrou-se que era possível realizar um *trans-splicing* baseado nessa idéia. Entretanto, ainda há um longo caminho a ser percorrido e muitas dificuldades a serem vencidas. A principal talvez seja a identificação de transcritos quiméricos candidatos a *trans-splicing* que possam direcionar experimentos em bancada. Nossa contribuição foi criar ferramentas de Bioinformática para suprir exatamente essa demanda, e, ao mesmo tempo, tentar compreender o papel das SR na formação de transcritos quiméricos.

Para estudar a relação entre SR e a formação de transcritos quiméricos em organismos superiores, o primeiro desafio é encontrar transcritos que possam ser candidatos. Embora as bases de dados disponíveis tenham um número considerável de

sequências, como foi mencionado, o fenômeno é raro. Por isso, técnicas de filtragem devem ser cuidadosamente empregadas a fim de minimizar a detecção de falso-positivos e ao mesmo tempo não perder possíveis candidatos. Para verificar isso, foi desenvolvida uma metodologia para busca e identificação de mRNAs quiméricos, Fusion5Finder (Herai & Yamagishi (a), 2009) que encontrou candidatos em humano, não reportados anteriormente, que se assemelham às melhores evidências experimentais encontradas e que podem direcionar novos testes de bancada. Utilizamos uma base de dados curada de *FL-cDNA* de humanos gerada a partir de tecidos supostamente normais, aplicamos critérios de filtragem sugeridos pelas evidências experimentais, e satisfatoriamente encontramos 16 mRNAs quiméricos, sendo 4 transcritos redundantes e obtidos por dois laboratórios independentes. Observamos que os transcritos quiméricos encontrados não apresentavam *splice-sites* canônicos. Vale destacar que *splice-sites* canônicos foram extensivamente exigidos em outras estratégias de varredura reportadas na literatura, o que parcialmente explica o motivo pelo qual nossos candidatos não foram previamente identificados. Há pelo menos duas possíveis explicações para não ocorrerem *splice-sites* canônicos em transcritos quiméricos: como o fenômeno é raro, não é improvável que *splice-sites* pouco frequentes possam estar ocorrendo sobre condições especiais (como os *minor* spliceossomos, descritos por Steitz et al. (2008)). Uma segunda explicação seria a existência de um mecanismo de formação de quiméricos não mediado por spliceossomo, como o que foi recentemente descrito pelo mecanismo de *trans-splicing* que originam um tRNA (Di Segni et al., 2008), em que a fusão de dois transcritos foi mediada pela presença de SR do tipo reverso complementar, ou ainda por meio de retrotransposição que é uma conjectura levantada durante esta pesquisa a partir da análise de quiméricos de humanos e bovinos.

Para cada um dos 16 candidatos, procuramos por evidências de transcrição (ET) da região genômica em que a TSR está contida, e encontramos evidências em 12 dos 16 casos. Os outros 4 casos representam um único transcrito que foi mapeado numa região que contem um *HERV-K*, que é um retrovirus humano ativo. Para verificar a existência de SR entre o par (TQ,ET) formado pelo transcrito candidato quimérico e sua evidência de transcrição, foi desenvolvida a ferramenta RepGraph (Herai & Yamagishi (b), 2010) para mapear 4 tipos de pares de SR presentes no *locus* gênico de dois transcritos: repetição direta (RD), repetição reversa (RR), repetição complementar (RC) e repetição reversa complementar (RRC). Para os 12 candidatos, foi constatado que ambas as regiões genômicas dos candidatos e dos transcritos associados (ET das TSRs) possuem SR complementares entre si (no caso de humanos, RRC, predominantemente elementos com tamanhos equivalentes às ALUs). Tais sequências podem estar envolvidas em um possível mecanismo de formação de quiméricos não mediado por spliceossomo, como conjecturado por Di Segni et al. (2008). O estudo baseado na frequência de SR não foi conclusivo, pois as diferenças entre a frequência de SR em regiões gênicas de transcritos quiméricos e regiões genômicas aleatórias, embora estatisticamente significativos, dependiam fortemente do tamanho dos transcritos envolvidos, o que nos proibe realizar qualquer tipo de conclusão. Isso não exclui a possibilidade de que apenas uma única SR, ou uma pequena classe, esteja desempenhando um papel importante. Essa linha de estudo será um desdobramento futuro desse trabalho.

Como mencionado, os outros 4 candidatos são transcritos redundantes, e fazem parte de um mesmo cluster do H-InvDB. Com o intuito de procurar por alguma ET, utilizamos uma ferramenta de predição de genes (GenMark), que nos forneceu uma região

similar a um retrovírus endógeno humano ativo, o HERV-K. Conforme discutido no texto, tal evidência enfraquece a hipótese do transcrito quimérico ter sido gerado por um evento de *trans-splicing*, pois o elemento genético móvel pode ter se movido exatamente para a posição do transcrito candidato, levando consigo sua TSR. Essa observação juntamente com outras evidências encontradas em bovinos nos levou a conjecturar que talvez o transcrito quimérico seja produto de um mecanismo onde a retrotransposição desempenhe o papel principal. Essa conjectura também será estudada futuramente.

A metodologia Fusion5Finder também já está sendo utilizada em um projeto de doutorado para tentar identificar experimentalmente novos genes quiméricos associados a câncer. Tal projeto será desenvolvido pela aluna Msc. Danielle Ribeiro Lucon, do Instituto de Ciências Médicas da Universidade Estadual de Campinas, e teve início em agosto de 2009 (Identificação de novos genes quiméricos associados a tumores sólidos da criança e adolescente. UNICAMP, FCM).

Estendemos a metodologia a outros organismos, e retiramos a exigência da TSR ocorrer exclusivamente na 5'UTR, pois de fato ela pode ocorrer em qualquer região do transcrito. Essa nova metodologia foi chamada FusionAllFinder (Herai & Yamagishi (a), 2010). O principal desafio foi reduzir o número de falso-positivos, através de filtros específicos. Como organismo modelo, utilizamos o *Bos taurus*, montagem UMD 3.0, cuja cobertura é superior a 95%. Além disso, utilizamos também 3 bibliotecas de cDNA dos grupos MGC, BGD e NCBI-RefSeq. Mesmo com a aplicação de filtros específicos, a metodologia detectou ainda alguns candidatos, que após verificação manual auxiliada por ferramentas de bioinformática especialmente desenvolvidas para tal fim se confirmaram como falso-positivos, comprovando que a metodologia, apesar de funcional, ainda requer

novos critérios de filtragem para que seja utilizada para genomas cuja montagem não apresenta qualidade equivalente ao do humano. Vale ressaltar que a metodologia original, Fusion5Finder, havia sido utilizada com os mesmos dados, e mostrou-se inviável pela quantidade excessiva de candidatos falso-positivos gerada.

Com base nesta metodologia estendida, foram identificados 13 candidatos quiméricos em bovino, dos quais 2 possuem a TSR na região 5'UTR, outros 6 na região CDS e outros 5 na região 3'UTR. Verificamos os *splice-sites* e somente em um dos casos encontramos os *splice-sites* canônicos, sugerindo novamente que a formação da maioria dos transcritos quiméricos seja pela presença de *splice-sites* pouco frequentes, ou por um mecanismo similar ao do tRNA, reportado anteriormente. Desta forma, procuramos por evidências de transcrição que mostrem que a região gênica da TSR possa ser transcrita, para que o transcrito quimérico possa ser formado. Dos 13 candidatos, em 10 deles encontramos evidências de transcrição ET, sendo que há dois transcritos que são redundantes com relação a outros dois candidatos. A partir disso, com o intuito de verificar a existência de SR nos *loci* gênicos entre o par (TQ,ET) que possivelmente esteja gerando um transcrito quimérico, utilizamos novamente a ferramenta RepGraph para procurar por SR deste tipo de forma análoga ao que foi feito em humanos. Para o candidato em que não encontramos uma ET, assim como no cluster redundante de humano, foi utilizada a ferramenta GeneMark para predição de genes em uma região estendida da TSR e foram encontradas duas sequências que possuem uma elevada identidade com relação a um elemento genético móvel, o NCNT2, e a um gene de transcriptase reversa, o RTLf. O ponto importante é o fato que um elemento genético móvel precisa de uma enzima de transcriptase reversa para se copiar no genoma, e encontramos uma evidência deste par:

elemento genético móvel e transcriptase reversa. Em função disso, não podemos refutar, assim como em humano, para o caso do retrovírus endógeno *HERV-K*, que o candidato quimérico possa também ter sido gerado por um evento de retrotransposição, ao invés de *trans-splicing* intercromossomal.

Analisamos também os casos onde o ponto de fusão se deu dentro da região codificante para verificar a formação de novas proteínas. Em um caso, encontramos um transcrito quimérico que gerou uma nova proteína predita. Embora não se saiba qual mecanismo deu origem a esse transcrito quimérico, provavelmente, ele seja importante na geração de novas proteínas.

Algumas alternativas não mutuamente excludentes para a formação de transcritos quiméricos com base na literatura analisada até a presente data, e também com base nas observações realizadas no desenvolvimento desta tese são graficamente ilustradas na Figura 52. Nossa conjectura, conforme observado nos candidatos quiméricos detectados em humano e bovino, é que algumas sequências quiméricas tenham sido geradas por meio de uma retrotransposição de elementos genéticos móveis. Recentemente, foi adicionada mais uma alternativa que é a possibilidade de que um dos transcritos que formam a molécula quimérica seja oriunda de uma molécula externa. Tal tipo de molécula pode ter sido gerada de maneira sintética para algum tipo de terapia gênica (conforme proposto pelo mecanismo SMART (Otto et al., 2003), ou pode inclusive ser oriunda de algum agente externo, patógeno (Kikumori et al., 2002) ou não.

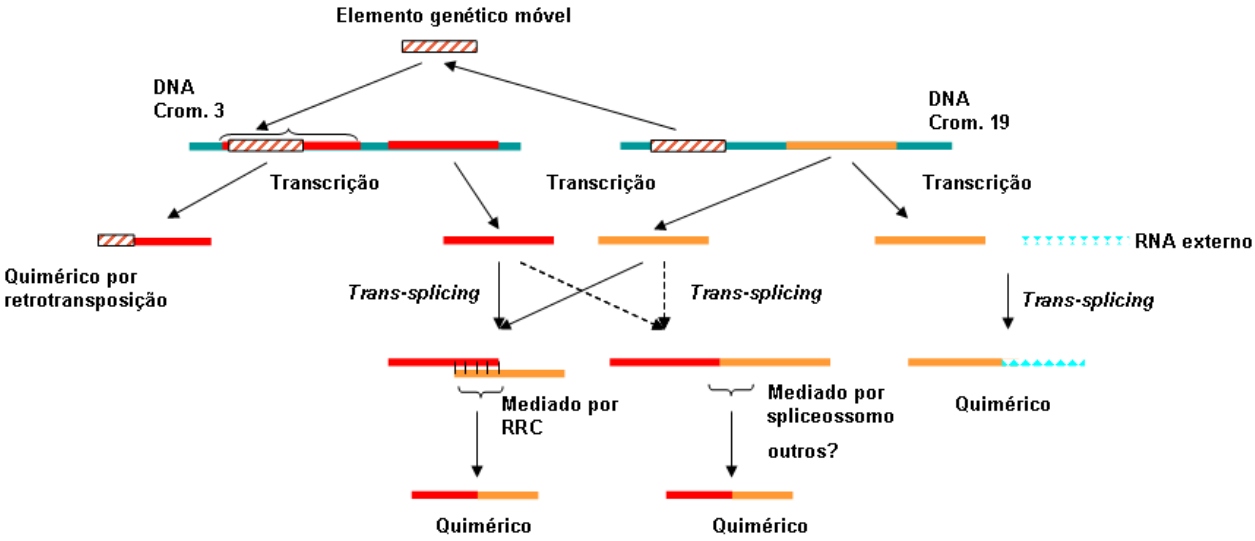


*Figura 52 – Hipóteses de formação de transcritos quiméricos.*
*Algumas das possíveis formas que um transcrito quimérico pode ser formado: por retrotransposição de elementos genéticos móveis (nossa conjectura); por material genético externo, oriundo de outros organismos, patogênicos ou não, ou sintéticos para tratamentos de terapia gênica; por trans-splicing intercromossomal (poderia ser também pelo tipo intracromossomal).*
*Fonte: da própria pesquisa.*

A metodologia estendida de detecção de transcritos quiméricos FusionAllFinder, por requerer a aplicação de filtros para remover a geração excessiva de candidatos falso-positivos, permitiu propor uma nova metodologia, DraftDNACheck (Herai & Yamagishi (b), 2009), para detecção de erros de montagem em regiões gênicas. DraftDNACheck foi testada com a que é considerada hoje a melhor montagem do genoma do bovino, a UMD 3.0 (idêntica em termos de sequências com relação à UMD 3.1), e permitiu detectar um provável erro com relação à ordem e orientação de um conjunto de contigs do cromossomo 24. No *XVIII Plant and Animal Genome Conference,* realizado na cidade de San Diego, nos Estados Unidos, relatamos tal erro a um dos responsáveis pela montagem de tal genoma, Dr Aleksey Zimin (Zimin et al., 2009). Outros casos de possíveis erros de montagem já foram detectados em bovino e também no genoma do rato (*Mus musculus*), os quais estão sendo verificados.

O banco de dados de SR Rep4DB, proposto neste trabalho, prevê a anotação de SR dos 4 tipos que podem ser localizados pela ferramenta RepGraph. Além disso, já foi construída uma ferramenta de migração de dados, MigDB, que pode ser facilmente estendida para que possa migrar bases de dados em formato texto além daqueles disponibilizados pelo H-InvDB, para que as informações a respeito não só das SR sejam agrupadas. A intenção é fornecer mais uma alternativa que permita estudar de maneira mais ampla as SR, pois sua importância já é comprovada, além de corresponderem, para alguns organismos (como trigo), a quase totalidade do material genético.

## 6.2 Trabalhos futuros

Com base no que foi observado a respeito da importância das SR, alguns trabalhos futuros são propostos e numerados a seguir:

1. buscar parceiros com experiência em Biologia Molecular para confirmação em bancada dos candidatos encontrados, tanto em humanos quanto em bovinos;
2. aplicar as metodologias Fusion5Finder e FusionAllFinder em outros organismos de interesse para tentar encontrar mais evidências de candidatos quiméricos. Além disso, dada a raridade do fenômeno em organismos superiores, utilizar os candidatos detectados para que possamos direcionar a sua confirmação experimental;
3. considerando as SR que foram encontradas entre os pares (TQ,ET), (GR,GR), (TR,GR) e (TR,TR), verificar se as SR envolvidas possuem algum padrão de similaridade e de ocorrência. Tal análise pode permitir, por exemplo, verificar se existe algum tipo de SR que é específico apenas entre pares de transcritos, e em regiões específicas dos mesmos;

4. pelo fato de que as principais ferramentas de mapeamento como BLAST e BLAST-like terem sido construídas para detectar somente SR dos tipos repetição direta (RD) e repetição reversa complementar (RRC), a metodologia RepGraph foi construída para permitir detectar também as SR de outros dois tipos: repetição reversa (RR) e repetição complementar (RC). A meta é realizar uma ampla análise em organismos para mapear estes 4 tipos de sequências repetitivas para que sejam anotados no banco de dados proposto neste trabalho. Além disso, é fundamental também a criação de uma plataforma WEB que permita visualizar e consultar a base de dados criada, e também permitir que a mesma seja livremente copiada para que outros grupos de pesquisa possam montar sua própria base de dados, mesmo que não tenham interesse em anotar esclusivamente as SR de um organismo;
5. copiar as outras bases de dados de SR e, com o uso da ferramenta MigDB, agrupar as informações no banco de dados Rep4DB proposto. Isso facilita categorizar os tipos de SR com base nos organismos existentes na base de dados, bem como para facilitar a associação e frequências das SR de acordo com sua função, caso existam, para cada organismo considerado;
6. utilizar a ferramenta DraftDNACheck para outros organismos já sequenciados e com bibliotecas de transcritos, preferencialmente as de *full-length* cDNA, para detectar novos e possíveis erros de montagem em regiões gênicas de outros genomas "*draft*", contribuindo, desta forma, para que a qualidade das montagens de genomas possa ser melhorada e, consequentemente, para que as evidências encontradas com base na organização de um genoma sejam mais confiáveis.

# Referências bibliográficas

Agabian, N. **Trans splicing of nuclear pre-mRNAs**. Cell **61**, 1157−1160, 1990;

Alexiou, Panagiotis; Vergoulis, Thanasis; Gleditzsch, Martin; Prekas, George; Dalamagas, Theodore; Megraw, Molly; Grosse, Ivo; Sellis, Timos; Hatzigeorgiou, Artemis G. **miRGen 2.0: a database of microRNA genomic information and regulation**. Nucleic Acids Research, 2010, Vol. 38, Database issue D137-D141;

Álvarez, C.S. **Molecular biology of retinoblastoma**. Journal Clinical and Translational Oncology, Springer Milan, ISSN 1699-048X, 10(7), 2008;

Anderson, A.M.; Staley, J.P. **Long-distance splicing**. PNAS 105(19): 6793–6794, 2008;

Avery, Oswald T.; MacLeod, Colin M.; McCarty, Maclyn. **Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III**. Journal of Experimental Medicine, Vol 79, 137-158, Copyright, 1944;

Bannert, N.; Reinhard, K. **Retroelements and the human genome: new perspectives on an old relation**. PNAS, 101: 14572-79, 2004;

Batzer, M.; Deininger, P.; Hellmann-Blumberg, U.; Jurka, J.; Labuda, D.; Rubin, C.; Schmid, C.; Zietkiewicz, E.; Zuckerkandl, E. **Standardized nomenclature for Alu repeats**. J. Mol. Evol., 42: 3–6, 1996;

Benson, G**. Tandem repeats finder: a program to analyze DNA sequences**. Nucleic Acids Research, 27, 573-580, 1999;

Blumenthal, T. **Community TCeR. Trans-splicing and operons**. WormBook, The C. elegans Research Community Edition, Pasadena, USA, 2005;

Boby, T.; Patch, A.M.; AveS, S.J. **TRbase: a database relating tandem repeats to disease genes for the human genome**. Bioinformatics, 21: 811-816, 2005;

Bornberg-Bauer, E.; Paton, N. **Conceptual data modelling for bioinformatics**. *Briefings in Bioinformatics*, 166–180, 2002;

Bonen, L. **Trans-splicing of pre-mRNA in plants, animals and protests**. FASEB J., **7**: 40-46, 1993;

Breen, M.A.; Ashcroft, S.J.H. **A truncated isoform of $Ca^{2+}$/calmodulin-dependent protein kinase II expressed in human islets of Langerhans may result from trans-splicing**. FEBS, **409**: 375-379, 1997;

Brennecke, J.; Hipfner, D.R.; Stark, A.; Russell, R.B.; Cohen, S.M. **Bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila**. Cell, 113: 25-36, 2003;

Buneman, P.; Khanna, S.; Tajima, K.; Tan, W.C. **Archiving scientific data**. *ACM Transactions on Database System*, 2–4, 2004;

Castelo, A.; Martins, W.; Gao, G. **TROLL - Tandem Repeat Ocurrence Locator**. Bioinformatics Journal, 18, 634-636, 2002;

Caudevilla, C.; Serra, D.; Miliar, A.; Codony, C.; Asins, G.; Bach, M.; Hegardt, F.G. **Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver**. Proc Natl Acad Sci U S A, **95(21)**: 12185–12190, 1998;

Cavalcanti, M.C.; Targino, R.; Baião, F.A.; Rossle, S.C.; Bisch, P.M.; Pires, P.F.; Campos, M.L.M.; Mattoso, M. **Managing structural genomic workflows using web services**. Data and Knowledge Engineering, 53(1):45–74, 2005;

Chao, H.; Mansfield, S.G.; Bartel, R.; Hiriyanna, S.; Mitchell, L.G.; Garcia-Blanco, M.A.; Walsh, C.E. **Phenotype Correction of Hemophilia A Mice by Spliceosome-Mediated RNA Trans-Splicing**. Nature Medicine, **9**: 1015-1019, 2003;

Chaparro, Cristian; Guyot, Romain; Zuccolo, Andrea; Piégu, Benoît; Panaud, Olivier. **RetrOryza: a database of the rice LTR-retrotransposons**. Nucleic Acids Research, 2007, Vol. 35, Database issue D66-D70;

Chargaff, E. **Structure and function of nucleic acids as cell constituents**. Federal Proceedings, 10, 654-659, 1951;

Charlesworth, B.; Sniegowski, P.; Stephan, W. **The evolutionary dynamics of repetitive DNA in eukaryotes**. Nature, 371: 215-220, 2002;

Chen, X.A. **MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development**. Science, 303: 2022-2025, 2004;

Chen, C.; Fossar, N.; Weil, D.; Guillaud-Bataille, M.; Danglot, G.; Raynal, B.; Dautry, F.; Bernheim, A.; Brison, O. **High frequency trans-splicing in a cell line producing spliced and polyadenylated RNA polymerase I transcripts from an rDNA-myc chimeric gene**. Nucleic Acids Research, **33(7)**: 2332-2342, 2005;

Cheng, Guofeng; Cohen, Leah; Ndegwa, David; Davis, Richard E. **The Flatworm Spliced Leader 3-Terminal AUG as a Translation Initiator Methionine**. The Journal of Biological Chemistry, 281, 733-743, 2006;

Cheung, Joseph; Estivill, Xavier; Khaja, Razi; MacDonald, Jeffrey R.; Lau, Ken; Tsui, Lap-Chee; Scherer, Stephen W. **Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence.** Genome Biol. 2003; 4(4): R25;

Schuster, Stephan C. **Next-generation sequencing transforms today's biology**. Nature Methods - 5, 16 - 18, 2008, doi:10.1038/nmeth1156;

Dagan, Tal; Sorek, Rotem; Sharon, Eilon; Ast, Gil; Graur, Dan. **AluGene: a database of Alu elements incorporated within protein-coding genes**. Nucleic Acids Research, 2004, Vol. 32, Database issue D489-D492;

Davidson, S.B., Overton, G.C., and Buneman, P. **Challenges in integrating biological data sources**. *Journal of Computational Biology*, 4:557–572, 1995;

DeCerbo, J.; Carmichael, G.G. **SINEs point to abundant editing in the human genome**. Genome Biology, 6:216, 2005;

Deininger, P.L.; Batzer, M.A. **Mammalian Retroelements**. Genome Research, 12: 1455-1465, 2002;

Dezulian, T.; Remmert, M.; Palatnik, J.F.; Weigel, D.; Huson, D.H. **Identification of plant microRNA homologs**. Bioinformatics, 22, 359-360, 2006;

Di Segni, G.; Borguese, L.; Sebastiani, S.; Tocchini-Valentini, G. P. **A pre-tRNA carrying intron features typical of Archaea is spliced in yeast**. RNA Society, 11: 70-76, 2005;

Di Segni, G.; Gastaldi, S.; Tocchini-Valentini, G.P. **Cis- and trans-splicing of mRNAs mediated by tRNA sequences in eukaryotic cells**. PNAS, 105(19): 6864–6869, 2008;

Dixon, R.J.; Eperon, I.C.; Samani, N.J. **Complementary intron sequence motifs associated with human exon repetition: a role for intragenic, inter-transcript interactions in gene expression**. Bioinformatics, 23(2):150-155, 2007;

Dostie, J.; Mourelatos, Z.; Yang, M.; Sharma, A.; Dreyfuss, G. **Numerous microRNPs in neuronal cells containing novel microRNAs**. RNA, 9: 180-6, 2003;

Dress, A.; Giegerich, R.; Grunewald, S.; Wagner, H. **Fibonacci-Cayley Numbers and Repetition Patterns in Genomic DNA**. Annals of Combinatorics, 7, 259-279, 2003;

Elsik, C.G.; Tellam, R.L.; Worley, K.C.; Gibbs, R.A. **The genome sequence of taurine cattle: a window to ruminant biology and evolution**. Science, 2009; 324(5926):522-8;

Faustino, N.A.; Cooper, T.A. **Pre-mRNA splicing and human disease**. Genes & Dev. 17: 419-437, 2003;

Fire, A.; Xu, S.; Montgomery, M.K.; Kostas, S.A.; Driver, S.E.; Mello, C.C. **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans**. Nature, 391: 806-811, 1998;

Finta, C.; Zaphiropoulos, P.G. **Intergenic mRNA molecules resulting from trans-splicing**. J. Biol. Chem., 277: 5882–5890, 2002;

Fischer, S.E.; Butler, M.D.; Pan, Q.; Ruvkun, G. **Trans-splicing in C. elegans generates the negative RNAi regulator ERI-6/7**. Nature, **455(7212)**:491-6, 2008;

Flockerzi, A.; Ruggieri, A.; Frank, O.; Sauter, M.; Maldener, E.; Kopper, B.; Wullich, B.; Seifarth, W.; Müller-Lantzsch, N.; Leib-Mösch, C.; Meese, E.; Mayer, J. **Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project**. BMC Genomics, 2008; 9:354;

Forsdyke, D.R.(a). **A stem-loop "kissing" model for the initiation of recombination and the origin of íntrons**. Molecular Biology and Evolution, 12: 949-958, 1995;

Forsdyke, D.R.(b). **Conservation of stem-loop potential in introns of snake venom phospholipase A2 genes: An application of FORS-D analysis**. Molecular Biology and Evolution, 12: 1157-1165, 1995;

Forsdyke, D.R.(c). **Relative roles of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species**. Journal of Molecular Evolution, 41: 573-581, 1995;

Forsdyke, D.R. **The gene concept in 2008**. Scherrer and Josts' symposium, 2009, DOI 10.1007/s12064-009-0071-2;

Gajer, Pawel; Schatz, Michael; Salzberg, Steven L. **Automated correction of genome sequence errors**. Nucleic Acids Research, 2004, Vol. 32, No. 2 562-569;

Gal-Mark, N.; Schwartz, S.; Ast, G. **Alternative splicing of Alu exons—two arms are better than one**. Nucleic Acids Research, 36(6): 2012-2023, 2008;

Garcia-Blanco, M.A. **Messenger RNA Reprogramming by Spliceosome-Mediated RNA Trans-Splicing**. Journal of Clinical Investigation, **112**: 474-480, 2003;

Garcia-Blanco, M.A.; Baraniak, A.P.; Lasda, E.L. **Alternative Splicing in Disease and Therapy**. Nature Biotechnology, **22**: 535-546, 2004;

Gingeras, T. R. **Implications of chimaeric non-co-linear transcripts**. Nature 461, 206-211, 2009;

Gonsalves, D.; Gonsalves, C.; Ferreira, S.; Pitz, K.; Fitch, M.; Manshardt, R.; Slightom, J. **Transgenic virus resistant papaya: from hope to reality for controlling payaya ringspot virus in Hawaii**. American Phytopathological Society, 2004;

Goodrich, A.; Kugel, J.F. **Non-coding-RNA regulators of RNA polymerase II transcription**. Nature Reviews Molecular Cell Biology, 7: 612-616, 2006;

Griffiths-Jones, Sam; Grocock, Russell J.; van Dongen, Stijn; Bateman, Alex; Enright, Anton J. **miRBase: microRNA sequences, targets and gene nomenclature**. Nucleic Acids Research, 2006, Vol. 34, Database issue D140-D144;

Grissa, Ibtissem; Vergnaud, Gilles; Pource, Christine. **The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats**. BMC Bioinformatics, 2007, 8:172doi:10.1186/1471-2105-8-172;

Gros, François. **Comments on the paper by K. Scherrer and J. Jost "Gene and genon" concept: coding versus regulation**. Theory Biosci., 2009, 128:155–156, DOI 10.1007/s12064-009-0070-3;

Gunsalus, K.C.; Yueh, W.C.; MacMenamin, P.; Piano, F. **RNAiDB and PhenoBlast: Web tools for genome-wide phenotypic mapping projects**. Nucl Acids Res, 2004, 32,Database issue:D406-D410;

Gunter, M.; Thomas, T. **Mechanisms of gene silencing by double-stranded RNA**. Nature, 431, 343-349, 2004;

Gur-Arie, R.; Cohen, C.J.; Eitan, Y.; Shelef, L.; Hallerman, E.M.; Kashi, Y. **Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism**. Genome Res., 10, 62-71, 2000;

Haig; H.; Kazazian, Jr. **Mobile Elements: Drivers of Genome Evolution**. Science, 303: 1626–1632, 2004;

Hahn, Y.; Bera, T.K.; Gehlhaus, K.; Kirsch, I.R.; Pastan, I.H.; Lee, B. **Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases**. Proc Natl Acad Sci USA, **101(36)**: 13257–13261, 2004;

Havecker E.R; Gao, X.; Voytas, D.F. **The diversity of LTR retrotransposons**. Genome Biology, 5:225, 2004;

Helvik, S. A.; Snove, O. Jr.; Saetrom, P. **Reliable prediction of Drosha processing sites improves microRNA gene prediction**. Bioinformatics, 23: 142-149, 2007;

Herai, R.H.; Yamagishi, M.E.B (a). **Detection of human interchromosomal trans-splicing in sequence databanks**. Briefings in Bioinformatics, Oxford Journals, 2009: doi:10.1093/bib/bbp041;

Herai, R.H.; Yamagishi, M.E.B (b). **Metodologia de bioinformática para detecção de erros de montagem em regiões gênicas usando bibliotecas de Fl-cDNA: estudo de caso em bovinos**. I Workshop em Genômica Animal – Rede Genômica Animal da Embrapa, Fortaleza, Ceará, Brasil, 2009.

Herai, R.H.; Yamagishi, M.E.B (a). **A Bioinformatics Approach To Detect Interchromosomal Trans-Splicing In Bovine Full Length cDNA Databanks**. Plant & Animal Genomes XVIII Conference, 2010, San Diego, USA;

Herai, R.H.; Yamagishi, M.E.B (b). **RepGraph: Web-based tool for Identification and Visualization of Repetitive Sequences.** Bioinformatics, Oxford Journals, 2010 (em correção);

Hirano, M.; Noda, T. **Genomic organization of the mouse Msh4 gene producing bicistronic, chimeric and antisense mRNA**. Gene, 342, 165–177, 2004;

Høgh, Annabeth Laursen; Nielsen, Kåre Lehmann. **SAGE and LongSAGE**. Serial Analysis of Gene Expression (SAGE): Methods and Protocols, Methods in Molecular Biology, v 387, 2007: 3-24, DOI: 10.1007/978-1-59745-454-4_1;

Horiuchi, T.; Aigaki, T. **Alternative trans-splicing: a novel mode of pre-mRNA processing**. Biol. Cell, **98(2)**: 135-140, 2006;

Horng, Jorng-Tzong; Lin, F.M.; Lin, J.H.; Huang, H.D.; Liu, B.J. **Database of Repetitive Elements in Complete Genomes and Data Mining Using Transcription Factor Binding Sites**. IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 7, NO. 2, 2003;

Hui, J.; Hung, L.; Heiner, M.; Schreiner, S.; Neumüller, N.; Reither, G.; Haas, S.A.; Bindereif, A. **Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing**. The EMBO Journal, 24: 1988–1998, 2005;

Hurd, P.J.; Nelson, C.J. **Advantages of next-generation sequencing versus the microarray in epigenetic research**. Briefings in Functional Genomics and Proteomics, 2009; 8(3):174-83;

Imanishi, T.; Ito, T.; Suzuki, Y.; et al., **Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones**. PLoS Biology, 2(6):856-875, 2004;

Ivics, Z.; Izsvak, Z. **Transposons for Gene Therapy!**. Current Gene Therapy, 6, 593-607, 2006;

Jurka, J.; Kapitonov, V.V.; Kohany, O.; Jurka, M.V. **Repetitive sequences in complex genomes: structure and evolution**. Review of Genomics Human Genetics, 8, 241-59, 2007;

Jurka, J.; Kapitonov, V.V.; Pavlicek, A.; Klonowski, P.; Kohany, O.; Walichiewicz, J. **Repbase Update, a database of eukaryotic repetitive elements.** Cytogenet Genome Res. 2005;110(1-4):462-7;

Kennedy, B.; Lim, I.; Benson, Gary; Vincent, J.; Ferenc, M.; Heinrich, B.; Jarzylo, L.; Man, H.-Y.; Deshler, J. **3'-UTR SIRF: A Database for Identifying Clusters of Short Interspersed Repeats in 3' Untranslated Regions**. BMC Bioinformatics 2007, 8:274 doi:10.1186/1471-2105-8-274;

Kent, W. James. **BLAT—The BLAST-Like Alignment Tool**. Genome Res. 2002. 12: 656-664;

Kikumori, Toyone; Cote, Gilbert J.; Gagel, Robert F. **Naturally occurring heterologous trans-splicing of adenovirus RNA with host cellular transcripts during infection**. FEBS 26197 FEBS Letters 522, 2002: 41-46;

Kim, V.N.; Nam, J.W. **Genomics of microRNA**. Trends in Genetics, 22: 165-173, 2006;

Kiss, A. **Human box H/ACA pseudouridylation guide RNA machinary**. Molecular Cell Biology, 24: 5797-5807, 2004;

Köhler, Alwin; Hurt, Ed. **Exporting RNA from the nucleus to the cytoplasm**. Nature reviews. Molecular cell biology 2007;8(10):761-73.

Kolpakov, R.; Bana, G.; Kucherov, G. **Mreps: efficient and flexible detection of tandem repeats in DNA**. Nucleic Acid Research, 31, 3672-3678, 2003;

Kurtz, S.; Choudhuri, J.V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerich, R. **REPuter: the manifold applications of repeat analysis on a genomic scale**. Nucleic Acids Research, 29, 4633-4642, 2001;

Labrador, M.; Mongelard, F.; Plata-Rengifo, P.; Baxter, E.M.; Corces, V.G.; Gerasimova, T.I. **Protein encoding by both DNA strands**. Nature, **409**(6823): 1000, 2001;

Lai, E.C. **Computational identification of Drosophila microRNA genes**. Genome Biol., 4, R42, 2003;

Lander, E.S.; Botstein, D. **Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps**. Genetics, 121, 185-199, 1989;

Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; et al. **Initial sequencing and analysis of the human genome**. Nature, 409, 860-921, 2001;

Le Flèche, Philippe; Hauck, Yolande; Onteniente, Lucie; Prieur, Agnès; Denoeud, France; Ramisse, Vincent; Sylvestre, Patricia; Benson, Gary; Ramisse, Françoise; Vergnaud, Gilles. **A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis**. BMC Microbiology 2001, 1:2doi:10.1186/1471-2180-1-2;

Legendre, M.; Lambert, A.; Gautheret, D. **Pro_le-based detection of microRNA precursors in animal genomes**. Bioinformatics, 21, 841, 2005;

Lemos, M. **Workflow para Bioinformatica**. PhD thesis, Departamento de Informatica da PUC-Rio, 2004;

Lewin, B. **Genes IX**. Sudbury, MA: Jones and Bartlett Publishers, ISBN 0-7637-4063-2, 2007;

Li, B-L.; Li, X-L.; Duan, Z-L.; Lee, O.; Lin, S.; Ma, Z-M.; Chang, C.C.Y.; Yang, X-Y.; Park, J.P.; Mohandas, T.K.; et al. **Human Acyl-CoA: Cholesterol Acyltransferase-1 (ACAT-1) Gene Organization and Evidence That the 4.3-Kilobase ACAT-1 mRNA Is Produced from two Different Chromosomes**. J. Biol. Chemistry, **274**: 11060-11071, 1999;

Li, Hui; Wang, Jinglan; Mor, Gil; Sklar, Jeffrey. **A Neoplastic Gene Fusion Mimics Trans-splicing of RNAs in Normal Cells**. Science, 2008, 321,1357-1361;

Li, Hui; Wang, Jinglan; Ma, Xianyong; Sklar, Jeffrey. **Gene fusions and RNA Trans-splicing in Normal and Neoplastic Human Cells**. Cell Cycle, 2009, 8:2, 1-5;

Li, Ruiqiang; Fan, Wei; Tian, Geng Tian; Zhu, Hongmei; et al. **The sequence and de novo assembly of the giant panda genome**. Nature 463, 311-317, 2010, doi:10.1038/nature08696;

Lian, Y.; Garner, H.R. **Evidence for the regulation of alternative splicing via complementary DNA sequence repeats.** Bioinformatics, 21(8):1358-1364, 2005;

Lim, L.P.; Glasner, M.E.; Yekta, S.; Burge, C.B.; Bartel, D.P. **Vertebrate microRNA genes**. Science, 299, 1540, 2003;

Lodish, Harvey; Berk, Arnold; Zipursky, S. Lawrence; Matsudaira, Paul; Baltimore, David; Darnell, James E. **Molecular Cell Biology**. New York: W. H. Freeman & Co., 1999;

Lomsadze, A.; Ter-Hovhannisyan, V.; Chernoff, Y.; Borodovsky, M. **Gene identification in novel eukaryotic genomes by self-training algorithm**. Nucleic Acids Research, 2005, Vol. 33, No. 20, 6494-6506;

Loots, G.G.; Ovcharenko, I. ECRbase: **Database of Evolutionary Conserved Regions, Promoters, and Transcription Factor Binding Sites in Vertebrate Genomes**. Bioinformatics, 23(1):122-4, 2007;

Lopez, A. **Association for Computational Linguistics Hierarchical Phrase-Based Translation with Suffix Arrays**. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 976–985, 2007;

Mansfield, S. Gary; Chao, Hengjun; Walsh, Christopher E. **RNA repair using spliceosome-mediated RNA trans-splicing**. Trends in Molecular Medicine, Volume 10, Issue 6, 2004, Pages 263-268;

Mansfield, S.G.; Hawkins-Clark, R.; Puttaraju, M.; Kole, J.; Cohn, J.A.; Mitchell, L.G.; Garcia-Blanco, M.A. **5' Exon Replacement and Repair by Spliceosome-Mediated RNA Trans-Splicing**. RNA, **9**: 1290-1297, 2003;

Martianov, I.; Ramadass, A.; Barros, A.S.; Chow, N.; Akoulitchev, A. **Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript**. Nature, 445: 666-670, 2007;

Mayer, M.G.; Floeter-Winter, L.M. **Pre-mRNA trans-splicing: from kinetoplastids to mammals, an easy language for life diversity**. Mem. Inst. Oswaldo Cruz, **100**: 501-513, 2005;

Melquist, S.; Bender, J. **An Internal Rearrangement in an Arabidopsis Inverted Repeat Locus Impairs DNA Methylation Triggered by the Locus**. Genetics, 166: 437–448, 2004;

Moxon, E.R.; Wills, C. **DNA microsatellites: Agents of Evolution?** Scientific American, 72-77, 1999;

Munroe, S.H. **A large inverted repeat sequence overlaps two acceptor splice sites in adenovirus**. Nucleic Acids Res., 11(24): 8891–8900, 1983;

Nagashima, Takeshi; Matsuda, Hideo; Silva, Diego G.; Petrovsky, Nikolai; Konagaya, Akihiko; Schönbach, Christian. **FREP: a database of functional repeats in mouse cDNAs**. Nucleic Acids Research, 2004, Vol. 32, Database issue D471-D475;

Nam, Jin-Wu; Kim, Jinhan; Kim, Sung-Kyu; Zhang, Byoung-Tak. **ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs**. Nucleic Acids Res. 2006, 34(Web Server issue): W455–W458.

Nam, J.W.; Shin, K.R.; Han, J.; Lee, Y.; Kim, V.N.; Zhang, B.T. **Human microRNA prediction through a probabilistic co-learning model of sequence and structure**. Nucleic Acids Res., 33, 3570-3581, 2005;

Nishimura, D. **RepeatMasker**. Biotech Software & Internet Report, 1, 36-39, 2000;

Otto, E.; Temple, G.F.; McGarrity, G.J. **Re-Programming Gene Expression Using Spliceosome-Mediated RNA Trans-Splicing (SMaRT™)**. Current Drug Discovery, **6**: 37-42, 2003;

Pergolizzi, R.; Ropper, A.; Dragos, R.; Reid, A.; Nakayama, K.; Tan, Y.; Ehteshami, J.; Coleman, S.; Silver, R.; Hackett, N.; et al. **In Vivo Trans-splicing of 5' and 3' Segments of Pre-mRNA Directed by Corresponding DNA Sequences Delivered by Gene Transfer**. Molecular Therapy, **8**: 999-1008, 2003;

Pfeffer, S.; Zavolan, M.; Grasser, F.A.; Chien, M.; Russo, J.J.; Ju, J.; John, B.; Enright, A.J.; Marks, D.; Sander, C.; Tusch, T. **Identification of virus-encoded microRNAs**. Science, 304: 734-6, 2004;

Pruitt, Kim D.; Tatusova, Tatiana; Maglott, Donna R. **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. Nucleic Acids Res., 2007, 35(Database issue): D61–D65;

Puttaraju, M.; Jamison, S.F.; Mansfield, S.G.; Garcia-Blanco, M.A.; Mitchell, L.G. **Spliceosome-mediated RNA trans-splicing as a tool for gene therapy**. Nat. Biotech. 1999. 17:246-252;

Rana, T.M. **Illuminating the silence: understanding the structure and function of small RNAs**. Nature Review of Molecular Cell biology, 8, 23–36, 2007;

Regitano, L.C.A.; Coutinho, L.L. **Biologia molecular aplicada à produção animal**. Embrapa Informação Tecnológica, ISBN: 85-7383-122-7, 2001;

Reinhart, B.J.; Slack, F.A.; Basson, M.; Pasquinelli, A.E.; Bettinger, J.C.; Rougvie, A. C.; Horvitz, H.R.; Ruvkun, G. **The 21 nucleotide let-7 RNA regulates C. elegans developmental timing**. Nature, 403: 901, 2000;

Rice, P.;Longden, I.;Bleasby, A. **EMBOSS: The European Molecular Biology Open Software Suite**. Trends in Genetics, 14, 473-475, 2000;

Rigatti, R.; Jia, J-H.; Samani, N.J.; Eperon, I.C. **Exon repetition: a major pathway for processing mRNA of some genes is allele-specific**. NAR-Nucleic Acids Research, **32**: 441-446, 2004;

Robertson, H.M.; Navik, J.A.; Walden, K.K.O.; Honegger, H-W. **The Bursicon Gene in Mosquitoes: an Unusual Example of mRNA Trans-splicing**. Genetics, **176**: 1351-1353, 2007;

Rocheta, M.; Dionisio, F.M.; Fonseca, L.; Pires, A.M. **Paternity analysis in excel**. Computer Methods and Programs in Biomedicine, 88: 234-238, 2007;

Romani, A.; Guerra, E.; Trerotola, M.; Alberti, S. **Detection and analysis of spliced chimeric mRNAs in sequence databanks**. Nucleic Acids Research, **31**: 17-25, 2003;

Roux, M.; Levéziel, H.; Amarger, V. **Cotranscription and intergenic splicing of the PPARG and TSEN2 genes in cattle**. BMC Genomics, **4**: 7-71, 2006;

Roy-Engel, A.M.; Carroll, M.L.; Vogel, E.; Garber, R.K.; Nguyen, S.V.; Salem, A.H.; Batzer, M.A.; Deininger, P.L. **Alu Insertion Polymorphisms for the Study of Human Genomic Diversity**. Genetics, 159: 279–290, 2001;

Ruitberg, C.M.; Reeder, D.J.; Butler, J.M. **STRBase: a short tandem repeat DNA database for the human identity testing community**. Nucleic Acids Res. 29, 2001: 320-322;

SanMiguel, P.; Tikhonov, A.; Jin, Y.; Motchoulskaia, N.; Zakharov, D.; Melake-Berhan, A.; Springer, P.; Edwards, K.; Lee, M.; Avramova, Z.; Bennetzen, J. Nested **Retrotransposons in the Intergenic Regions of the Maize Genome**. Science, 274: 765–768, 1996;

Schattner, P. **Genome-wide searching for pseudourydilation guide snoRNAs: analysis of the saccharomyces cerevisiae genome**. Nucleic Acid Research, 32, 4281-4296, 2004;

Scherrer, K.; Jost, J. **Gene and genon concept: coding versus regulation. A conceptual and information-theoretic analysis of genetic storage and expression in the light of modern molecular biology**. Theory Biosci. 126, 2007: 65-113;

Schlesinger, J.; Arama, D.; Noy, H.; Dagash, M.; Belinky, P.; Gross, G. **In-Cell Generation of Antibody Single-Chain Fv Transcripts by Targeted RNA Trans-Splicing**. J. Immunological Methods, **282**: 175-186, 2003;

Schulz, H.L.; Stohr, H.; Weber, B.H. **Characterization of three novel isoforms of the metabotrobic glutamate receptor 7 (GRM7)**. Neurosci. Lett. 326 (1), 37-40, 2002;

Selzer, P.M.; Marhöfer, R.J.; Rohwer, A. **Applied Bioinformatics**. Springer Berlin Heidelberg, ISBN: 978-3-540-72799-6, 2008;

Shao, X.; Shepelev, V.; Fedorov, A. **Bioinformatic analysis of exon repletion, exon scrambling and trans-splicing in humans**. Briefings in Bioinformatics, **22**: 692-698, 2006;

Sharma, P.C.; Grover, A.; Kahl, G. **Mining microsatellites in eukaryotic genomes**, Trends in Biotecnology, 25: 490-498, 2007;

Smit, A.F.A. **Interspersed repeats and other mementos of transposable elements in mammalian genomes**. Current Opinion in Genetics & Development, 9(6): 657-663, 1999;

Smit, A.F.A.; Riggs, A.D. **Tiggers and other DNA transposon fossils in the human genome**. Proceedings of the National Academy of Sciences of the United States of America, 93(4): 1443-1448, 1996;

Sorek, R.; Ast, G.; Graur, D. **Alu-Containing Exons are Alternatively Spliced**. Genome Res., 12: 1060-1067, 2002;

Stark, A.; Bushati, N.; Jan, C.H. **A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands**. Genes & Development, 22: 8–13, 2008;

Stewart, M. **Molecular mechanism of the nuclear protein import cycle**. Nature Rev. Mol. Cell Biol. 8, 2007,195–208;

Steitz, Joan A.; Dreyfuss, Gideon; Krainer, Adrian R.; Lamond, Angus I.; Matera, A. Gregory; Padgett, Richard A.. **Where in the cell is the minor spliceosome?** Proc Natl Acad Sci U S A, 105(25), 2008: 8485–8486, DOI: 10.1073/pnas.0804024105;

Stinson, B. **PostgreSQL Essential Reference**. Editora New Rider, ISBN13: 9780735711211, 2001, pg. 400;

Stover, Nicholas A.; S. Kaye, Michelle; Cavalcanti, Andre R.O. **Spliced leader trans-splicing.** Current Biology, Volume 16, Issue 1, 2006, Pages R8-R9;

Strachan, T.; Read, A.P. **Human Molecular Genetics 2**. BIOS Scientific Publishers, Ltd, 1999;

Sundquist, A.; Ronaghi, M.; Tang, H.; Pevzner, P.; Batzoglou, S. **Whole-Genome Sequencing and Assembly with High-Throughput, Short-Read Technologies**. PLoS ONE 2(5), 2007: e484. doi:10.1371/journal.pone.0000484;

Tahara, M.; Pergolizzi, R.G.; Kobayashi, A.; Luettich, K.; Lesser, M.L.; Crystal, R.G. **Trans-splicing repair of CD40 ligand deficiency results in naturally regulated correction of a mouse model of hyper-IgM X-linked immunodeficiency**. Nature Medicine, **10**: 835-841, 2004;

Tasic, B.; Nabholz, C.E.; Baldwin, K.K.; Kim, Y.; Rueckert, E.H.; Ribich, S.A.; Cramer, P.; Wu, Q.; Axel, R.; Maniatis, T. **Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing**. Mol. Cell, 10, 21–33, 2002;

Temple, Gary; Gerhard, Daniela S.; Rasooly, Rebekah; Feingold, Elise A. **The completion of the Mammalian Gene Collection (MGC)**. Genome Res. , 2009, 19: 2324-2333;

Wang, Xiaowo; Zhang, Jing; Li, Fei; Gu, Jin; He, Tao; Zhang, Xuegong; Li, Yanda. **MicroRNA identification based on sequence and structure alignment**. Bioinformatics 2005 21(18):3610-3614; doi:10.1093/bioinformatics/bti562

Wang, J.; Zhang, J.; Zhang, H.; Li, J.; Liu, D.; Li, H.; Samudarala, R.; Yu, J.; Wong, G. K**. Neutral evolution of non-coding cDNA from the mouse transcriptome**. Nature, 431, 3016, 2004;

Warburton, P.E.; Giordano, J.; Cheung, F.; Gelfand, Y.; Benson, G. **Inverted Repeat Structure of the Human Genome: The X-Chromosome Contains a Preponderance of Large, Highly Homologous Inverted Repeats That Contain Testes Genes**. Genome Research, 14:1861-1869, 2004.

Wanunu, Meni; Morrison, Will; Rabin, Yitzhak; Grosberg, Alexander Y.; Meller, Amit. **Electrostatic Focusing of Unlabelled DNA into Nanoscale Pores Using a Salt Gradient**. Nature Nanotechnology, 2009, DOI: 10.1038/nnano.2009.379;

Watson, J.D.; Berry, A. **DNA: The Secret of Life**. Alfred A. Knopf: New York, New York, USA, 446 pp., ISBN: 0-375-41546-7, 2003;

Wickera, Thomas; Matthewsb, David E.; Keller, Beat. **TREP: a database for Triticeae repetitive elements**. Trends in Plant Science, Volume 7, Issue 12, 2002, 561-562;

Wieczorek, E.; Stortz, G. **An Expanding Universe of Non-coding RNAs**. Science, 2002, 296: 1260-1263;

Wu, S.; Wright, R.A.; Rockey, P.K.; Burgett, S.G.; Arnold, J.S.; Rosteck, P.R. Jr.; Johnson, B.G.; Schoepp, D.D.; Belagaje,R.M. **Group III human metabotropic glutamate receptors 4, 7 and 8: molecular cloning, functional expression, and comparison of pharmacological properties in RGT cells**. Brain Res. Mol. Brain Res. 53 (1-2), 88-97, 1998;

Yamagishi, M.E.B.; Shimabukuro, A.I. **Nucleotide Frequencies in Human Genome and Fibonacci Numbers**. Bulletin of Mathematical Biology, 70, 643-653, 2008;

Zhang, Huan; Campbell, David A.; Sturm, Nancy R.; Lin, Senjie. **Dinoflagellate Spliced Leader RNA Genes Display a Variety of Sequences and Genomic Arrangements**. Molecular Biology and Evolution, 2009, 26(8):1757-1771; doi:10.1093/molbev/msp083;

Zhang, Zhenhai; Yu, Jingyin; Li, Daofeng; Zhang, Zuyong; Liu, Fengxia; Zhou, Xin; Wang, Tao; Ling, Yi; Su, Zhen. **PMRD: plant microRNA database**. Nucleic Acids Research, 2010, Vol. 38, Database issue D806-D813;

Zhou, Fengfeng; Xu, Ying. **RepPop: a database for repetitive elements in Populus trichocarpa**. BMC Genomics. 2009; 10: 14, doi: 10.1186/1471-2164-10-14;

Zimin, Aleksey V.; Delcher, Arthur L.; Florea, Liliana; Kelley, David R.; et al. **A whole-genome assembly of the domestic cow, *Bos taurus***. Genome Biology, 2009, 10:R42doi:10.1186/gb-2009-10-4-r42;

# Apêndices

## Apêndice A

Os gráficos referentes às figuras Figura *53* a Figura *64* apresentam a quantidade de SR detectadas em diferentes tipos de sequências de humanos. Foram analisados os seguintes tipos de pares de sequências: par (TQ,ET): quantidade de SR entre o transcrito quimérico candidato e a evidência de transcrição da sua região TSR; par (GR,GR): média de SR entre duas sequências reais do genoma, porém de posições aleatórias; par (TR,GR): média de SR entre o loci gênico de um transcrito real selecionado aleatoriamente e uma sequência genômica real de uma posição aleatória do genoma; par (TR,TR): média de SR entre o loci gênico de dois transcritos reais selecionados aleatoriamente de uma base de transcritos. Os tamanhos das sequências são listados na Tabela 11.

*Tabela 11 – Tamanhos dos pares de transcritos (TQ,ET) em Humano.*
*Tamanho dos pares de transcritos (TQ,ET) envolvidos na formação das sequências quiméricas candidatas em humano.*

| Transcrito candidato | Tamanho TQ (pb) | Tamanho ET (pb) |
|---|---|---|
| AK130557 | 10805 | 25459 |
| AL834489 | 5362 | 69475 |
| U09825 | 20312 | 76652 |
| D26155 | 163977 | 99001 |
| AB023216 | 252993 | 78568 |
| L14837 | 122364 | 51500 |
| AB007865 | 4860 | 1220142 |
| AB020656 | 564444 | 12569 |
| AK226066 | 43170 | 7068 |
| L33075 | 113971 | 303954 |
| AK124366 | 2485 | 29816 |
| AF003522 | 8003 | 274865 |

*Fonte: da própria pesquisa.*

Todos os pares são formados por sequências cujo tamanho da primeira é equivalente e TQ, e o da segunda equivalente a ET. Por meio deles, é possível comparar, conforme discutido no corpo do trabalho, a quantidade de SR do par (TQ,ET) com relação aos pares (GR,GR), (TR,GR) e (TR,R). Os tipos de SR considerados pelos gráficos são dos tipos: repetição direta (RD), repetição reversa (RR), repetição complementar (RC) e repetição reversa complementar (RRC).

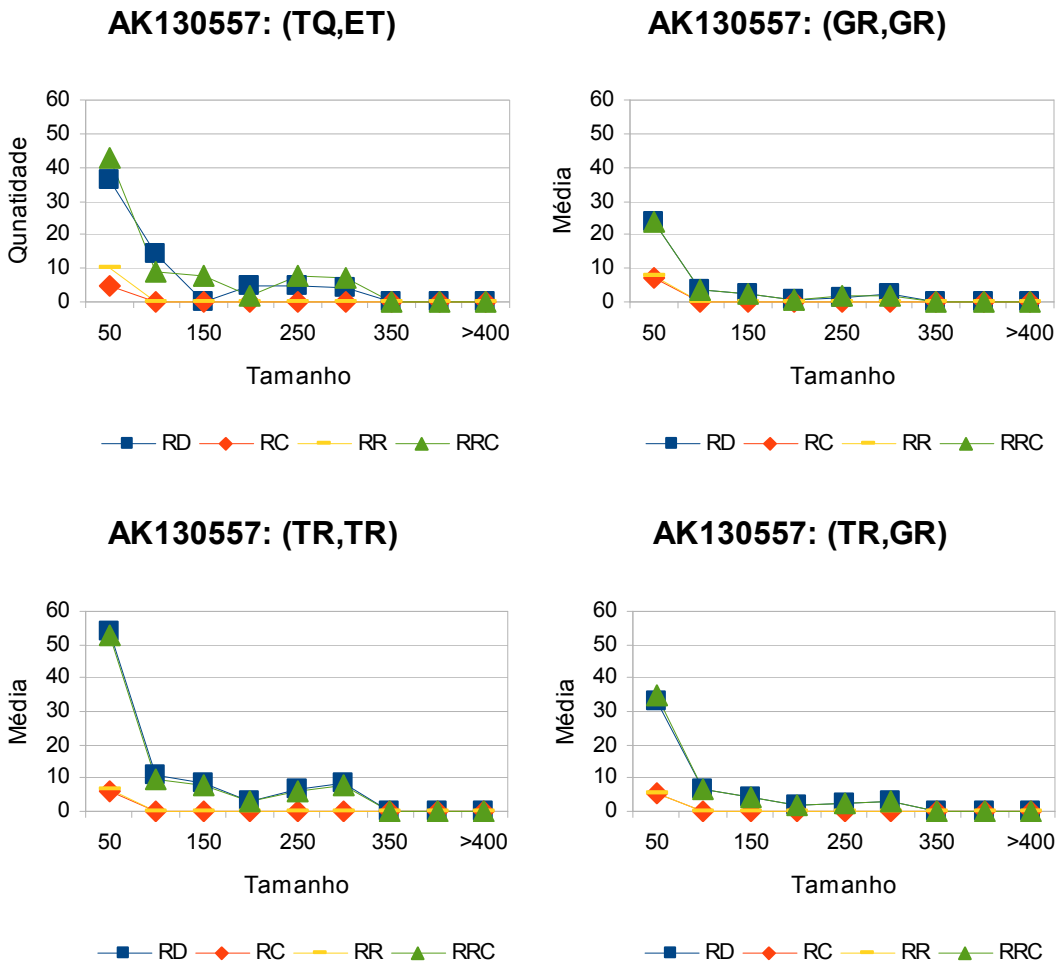


*Figura 53 – SR entre sequências relacionadas com o transcrito [DDBJ:AK130557].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:AK130557].*
*Fonte: da própria pesquisa.*

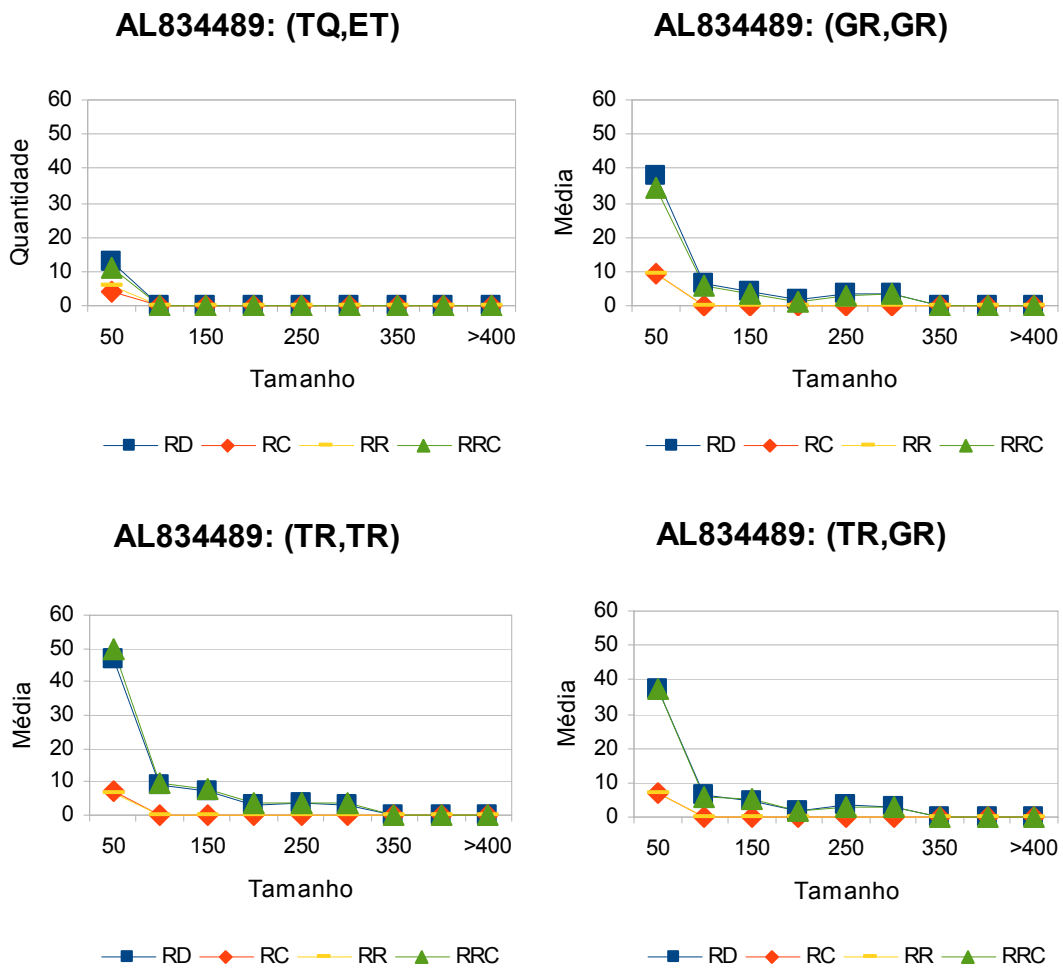


*Figura 54 - SR entre sequências relacionadas com o transcrito [DDBJ:AL834489].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:AL834489].*
*Fonte: da própria pesquisa.*

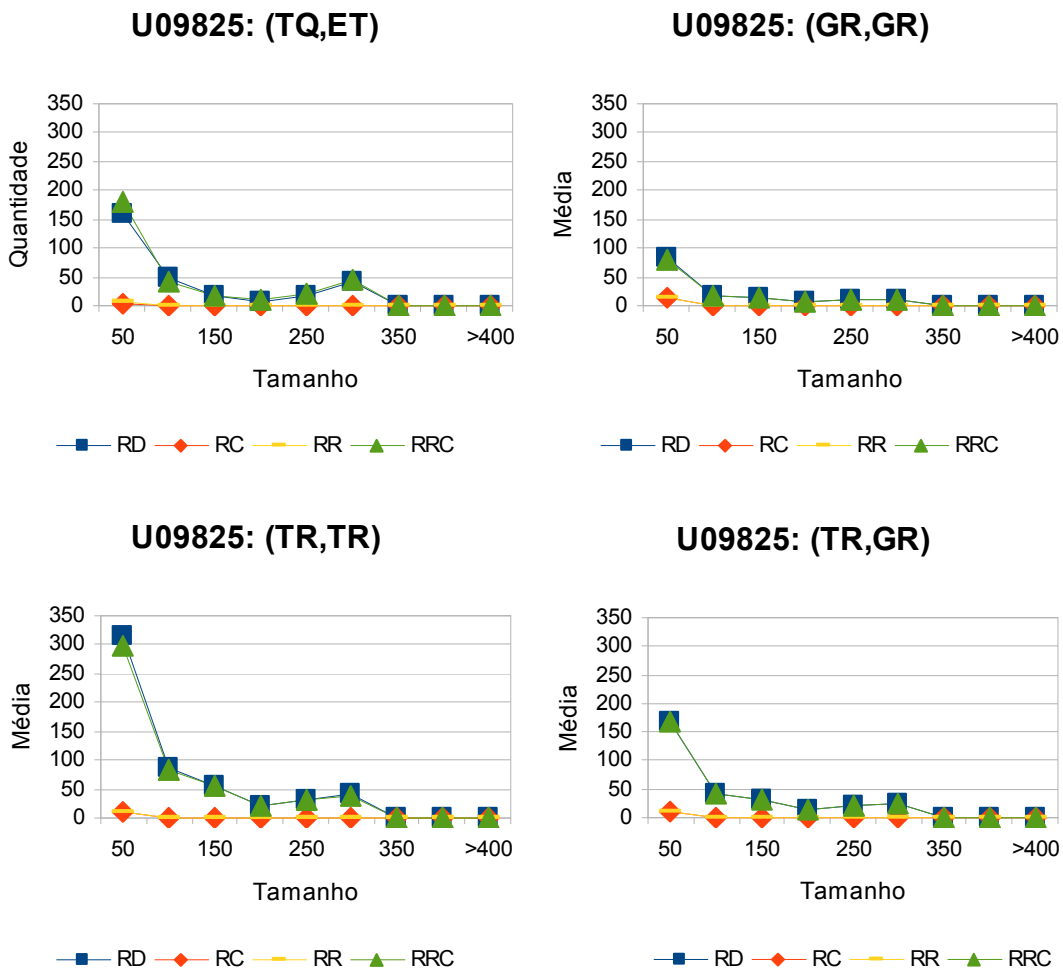


*Figura 55 – SR entre sequências relacionadas com o transcrito [DDBJ:U09825].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:U09825].*
*Fonte: da própria pesquisa.*

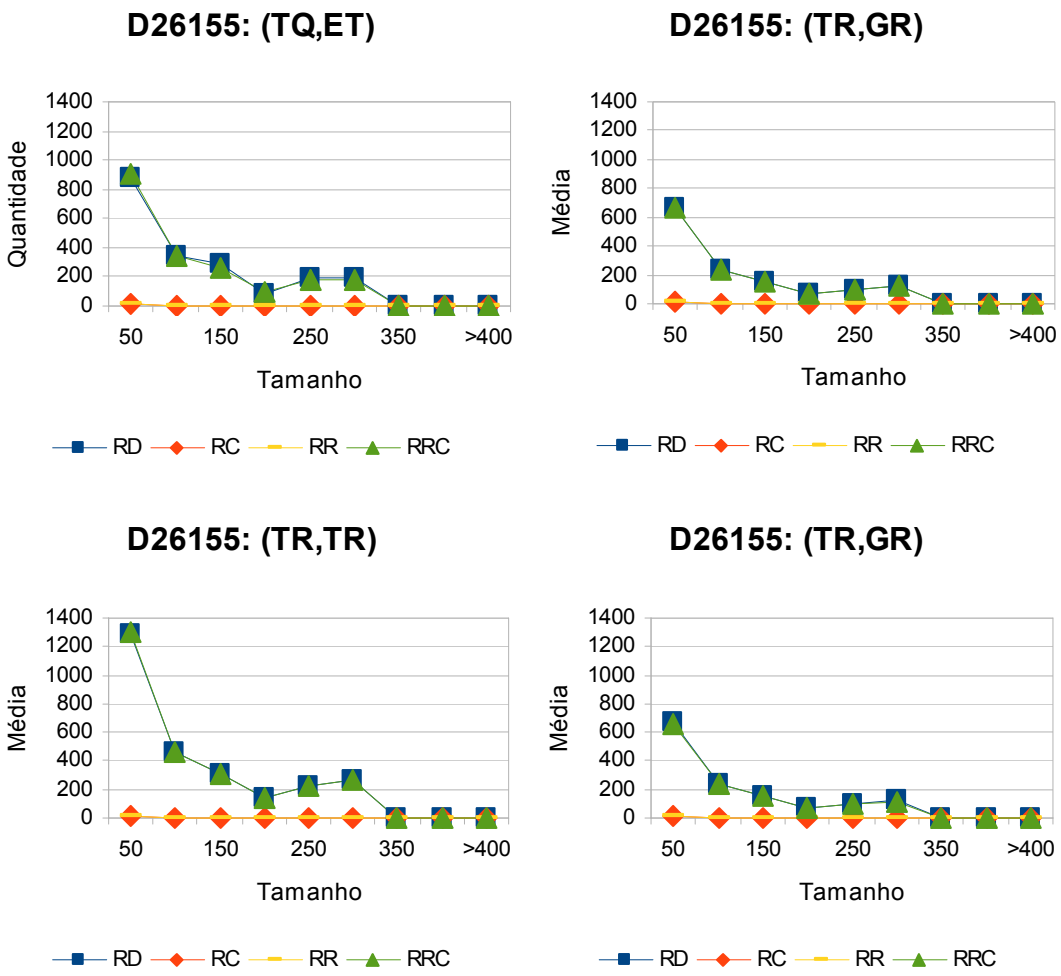


*Figura 56 – SR entre sequências relacionadas com o transcrito [DDBJ:D26155].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:D26155].*
*Fonte: da própria pesquisa.*

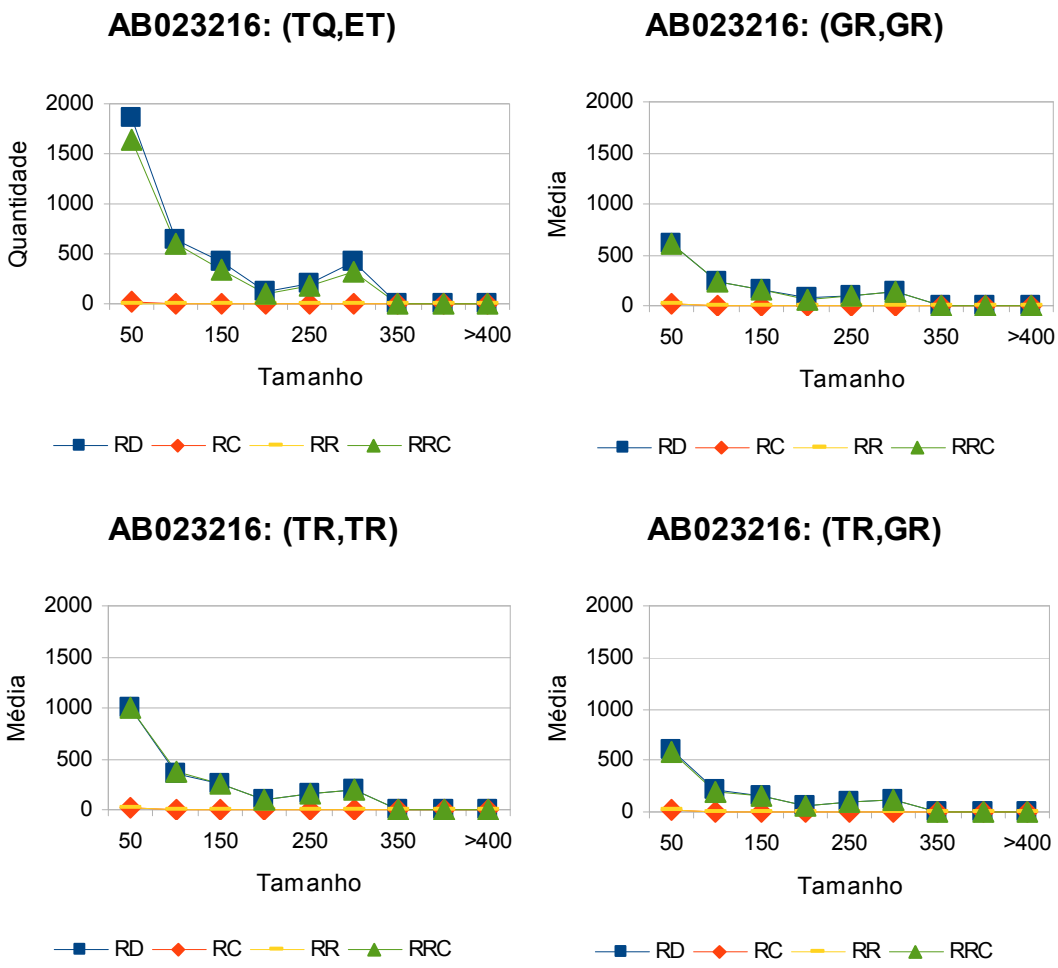


*Figura 57 - SR entre sequências relacionadas com o transcrito [DDBJ:AB023216].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:AB023216].*
*Fonte: da própria pesquisa.*

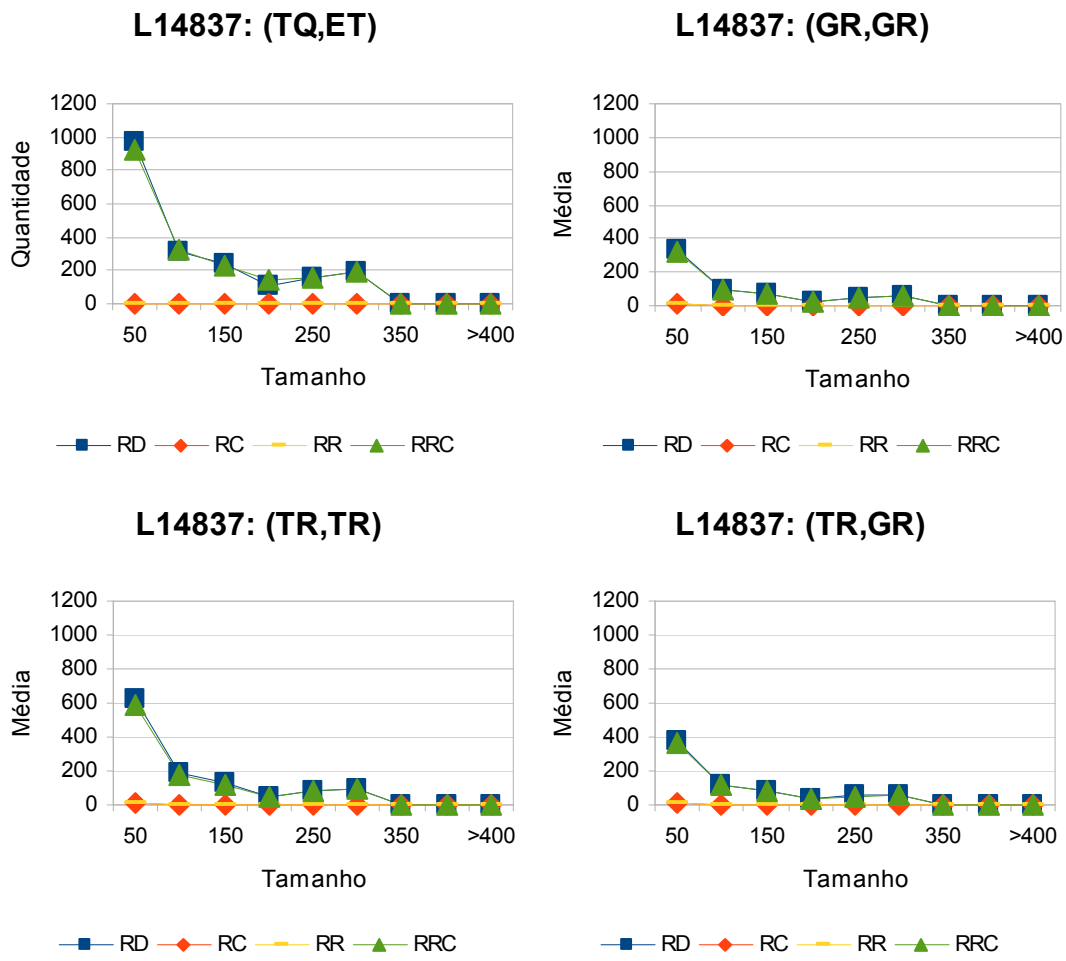


*Figura 58 – SR entre sequências relacionadas com o transcrito [DDBJ:L14837].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:L14837].*
*Fonte: da própria pesquisa.*

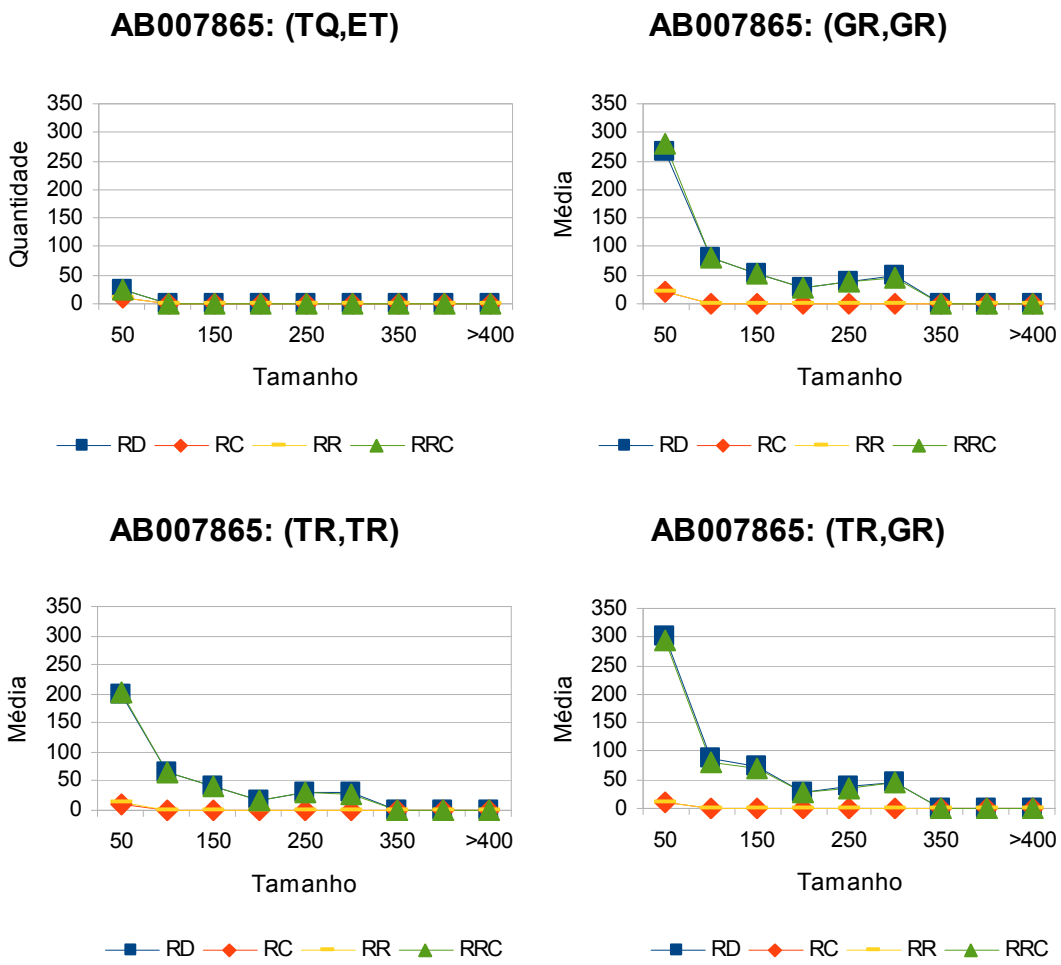


*Figura 59 – SR entre sequências relacionadas com o transcrito [DDBJ:AB007865].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:AB007865].*
*Fonte: da própria pesquisa.*

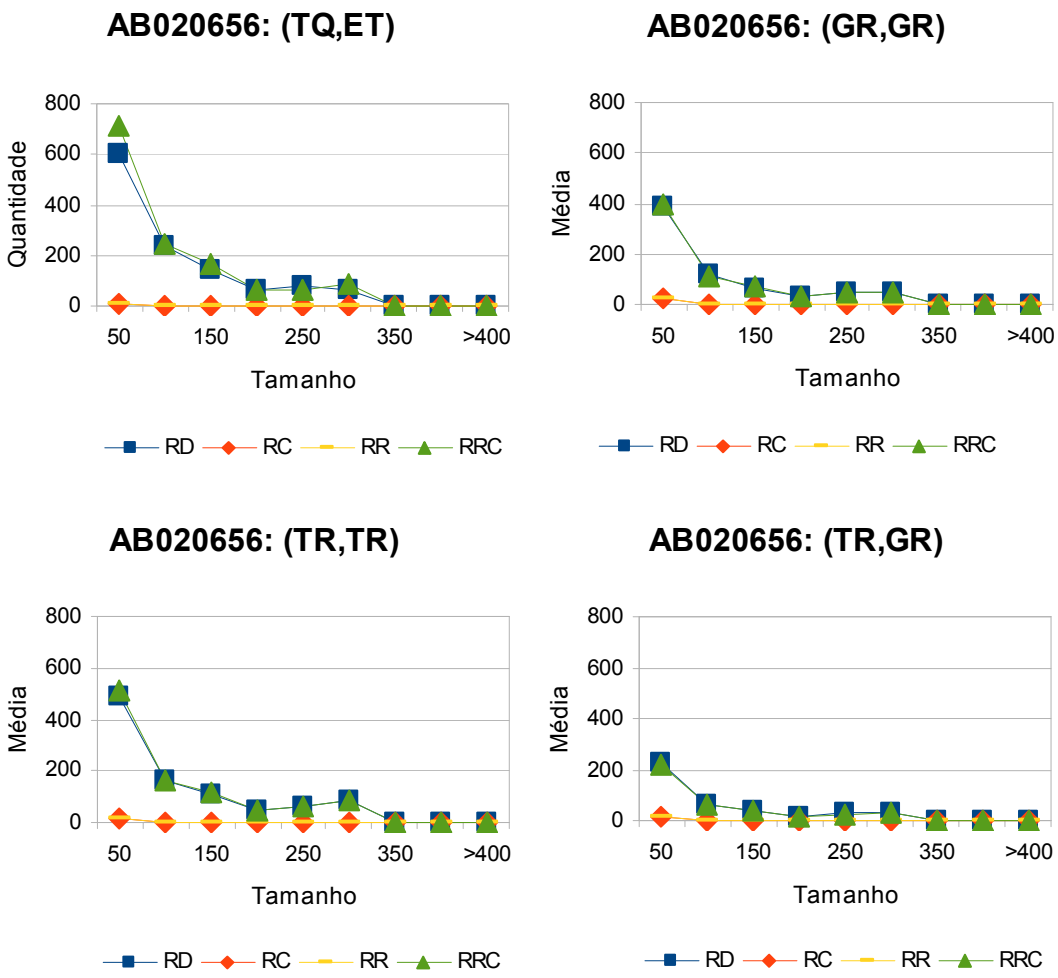


*Figura 60 - SR entre sequências relacionadas com o transcrito [DDBJ:AB020656].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:AB020656].*
*Fonte: da própria pesquisa.*

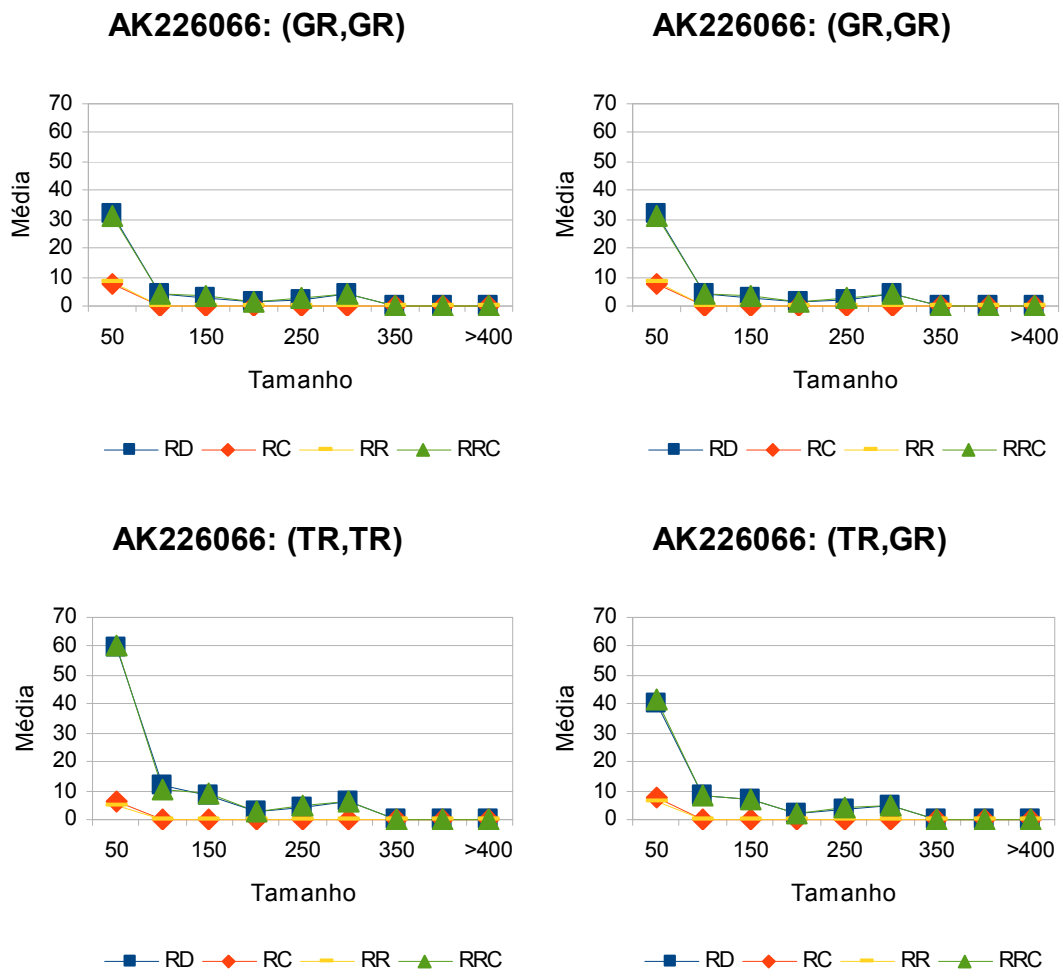


*Figura 61 – SR entre sequências relacionadas com o transcrito [DDBJ:AK226066].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:AK226066].*
*Fonte: da própria pesquisa.*

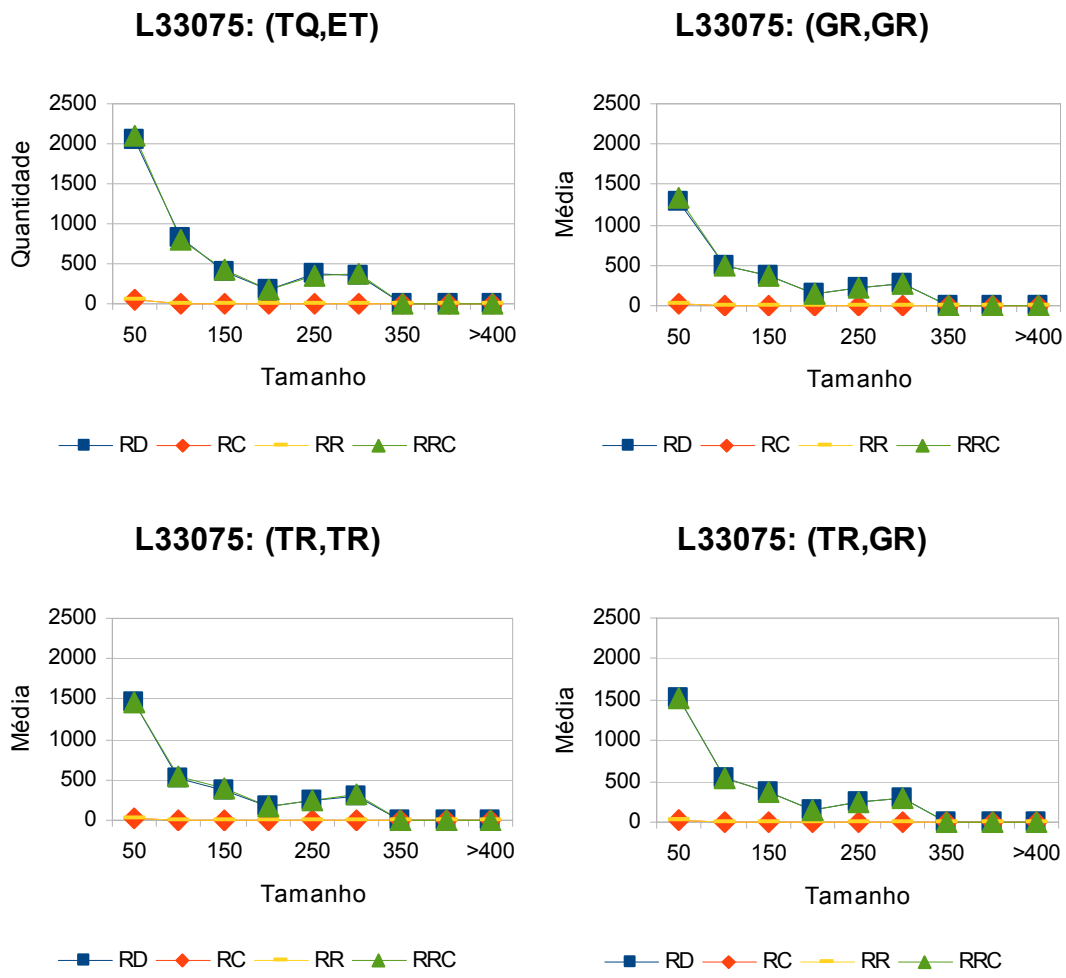


*Figura 62 - SR entre sequências relacionadas com o transcrito [DDBJ:L33075].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:L33075].*
*Fonte: da própria pesquisa.*

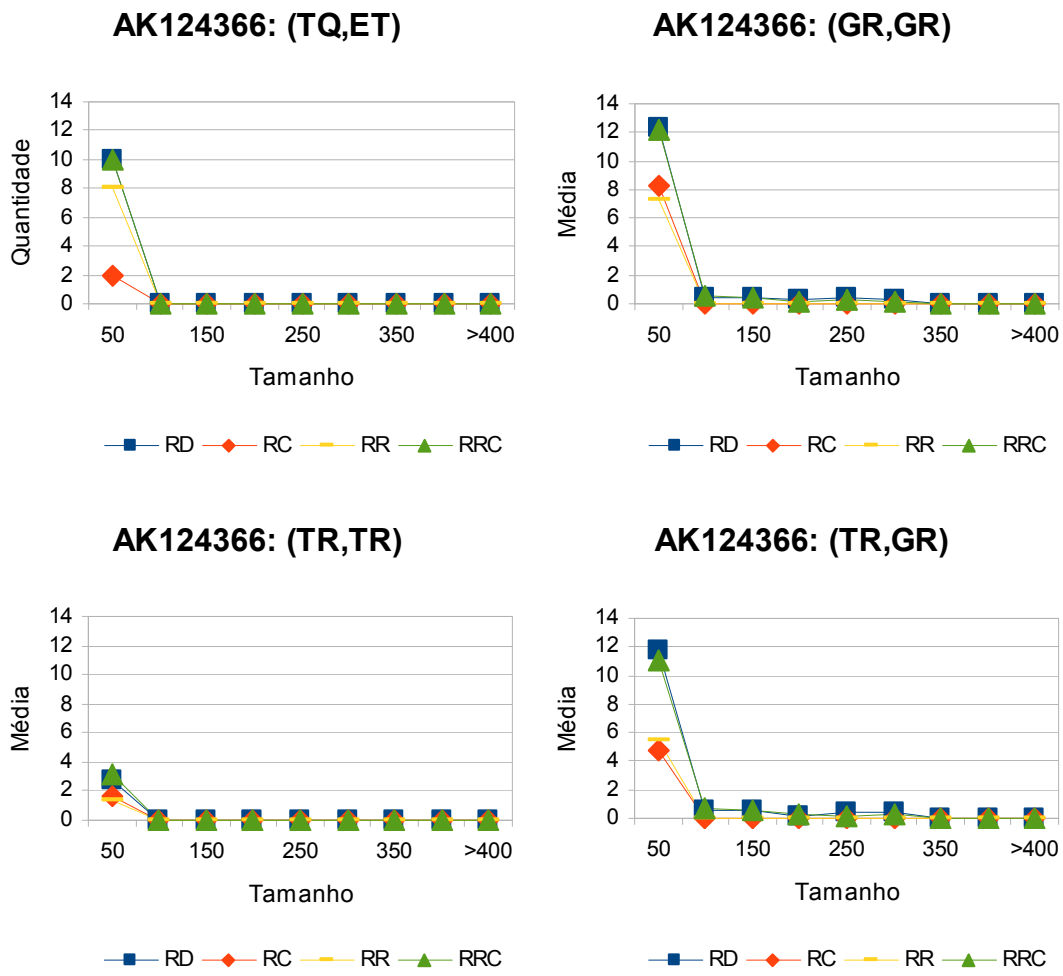


*Figura 63 – SR entre sequências relacionadas com o transcrito [DDBJ:AK124366].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:AK124366].*
*Fonte: da própria pesquisa.*

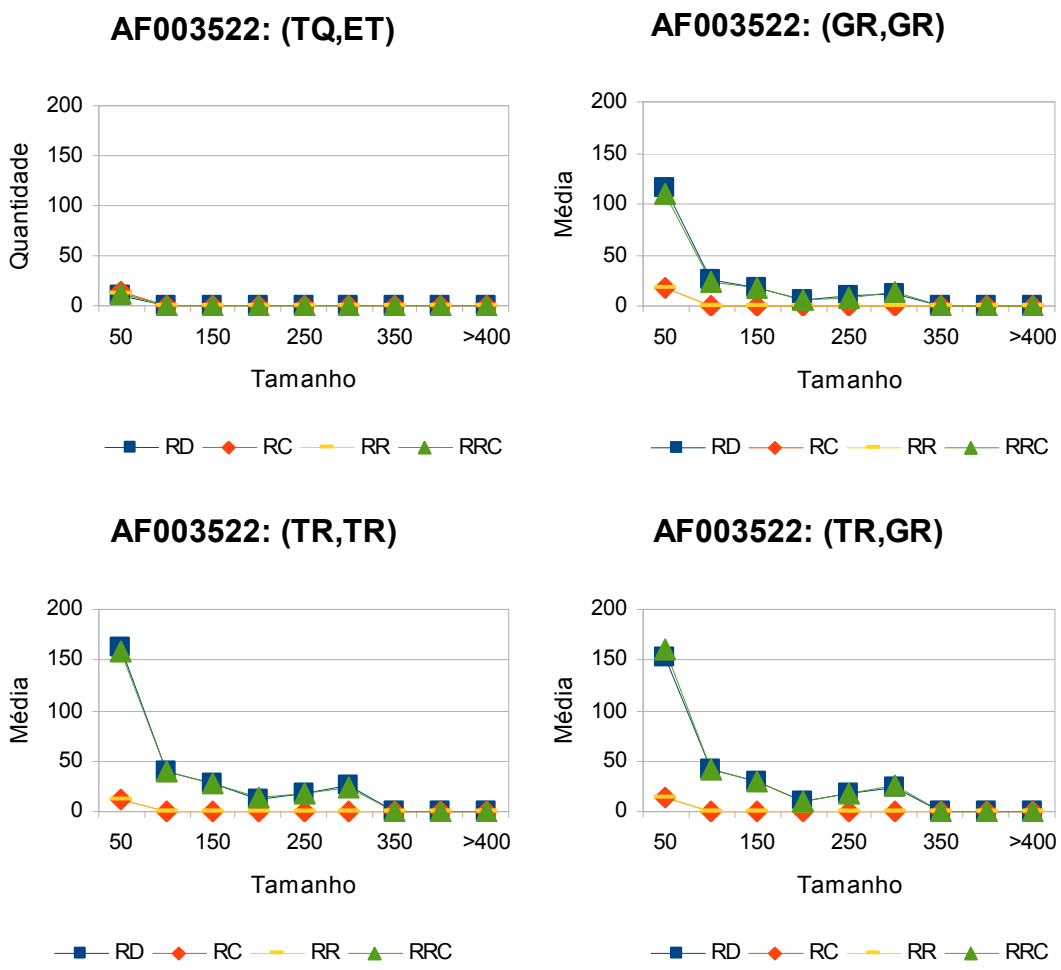


*Figura 64 - SR entre sequências relacionadas com o transcrito [DDBJ:AF003522].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico humano, [DDBJ:AF003522].*
*Fonte: da própria pesquisa.*

## Apêndice B

Os gráficos referentes às figuras Figura *65* a Figura *74* apresentam a quantidade de SR detectadas em diferentes tipos de sequências de bovinos. Foram analisados os seguintes tipos de pares de sequências: par (TQ,ET): quantidade de SR entre o transcrito quimérico candidato e a evidência de transcrição da sua região TSR; par (GR,GR): média de SR entre duas sequências reais do genoma, porém de posições aleatórias; par (TR,GR): média de SR entre o loci gênico de um transcrito real selecionado aleatoriamente e uma sequência genômica real de uma posição aleatória do genoma; par (TR,TR): média de SR entre o loci gênico de dois transcritos reais selecionados aleatoriamente de uma base de transcritos. O tamanho das sequências é listado na Tabela 12.

*Tabela 12 – Tamanhos dos pares de transcritos (TQ,ET) em Bovino.*
*Tamanho dos pares de transcritos (TQ,ET) envolvidos na formação das sequências quiméricas candidatas em bovino.*

| Transcrito candidato | Tamanho TQ (pb) | Tamanho ET (pb) |
|---|---|---|
| NM_203358 | 100934 | 7246 |
| NM_174719 | 12932 | 12932 |
| NM_001015598 | 6820 | 17198 |
| NM_001080267 | 7886 | 12000 |
| BC113235 | 646 | 7020 |
| BC133483 | 25624 | 15392 |
| BC134601 | 2475 | 2469 |
| BC148157 | 2963 | 22872 |
| BC112810 | 22771 | 17788 |
| BC105254 | 46564 | 9320 |
| NM_174055 | 96192 | 2329 |

*Fonte: da própria pesquisa.*

Todos os pares são formados por sequências cujo tamanho da primeira é equivalente e TQ, e o da segunda equivalente a ET. Por meio deles, é possível comparar, conforme discutido no corpo do trabalho, a quantidade de SR do par (TQ,ET) com relação aos pares

(GR,GR), (TR,GR) e (TR,R). Os tipos de SR considerados pelos gráficos são dos tipos: repetição direta (RD), repetição reversa (RR), repetição complementar (RC) e repetição reversa complementar (RRC).

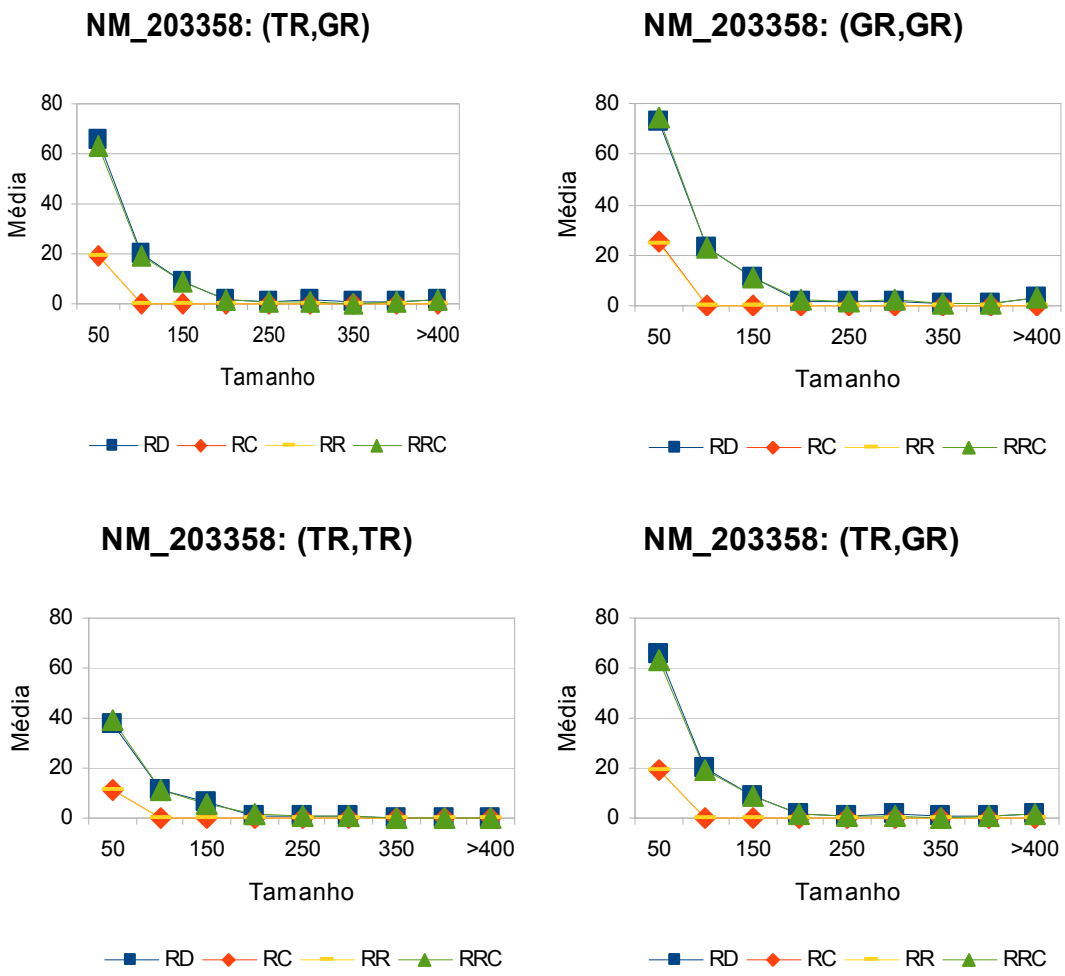


*Figura 65 – SR entre sequências relacionadas com o transcrito [RefSeq:NM_203358].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [RefSeq:NM_203358].*
*Fonte: da própria pesquisa.*

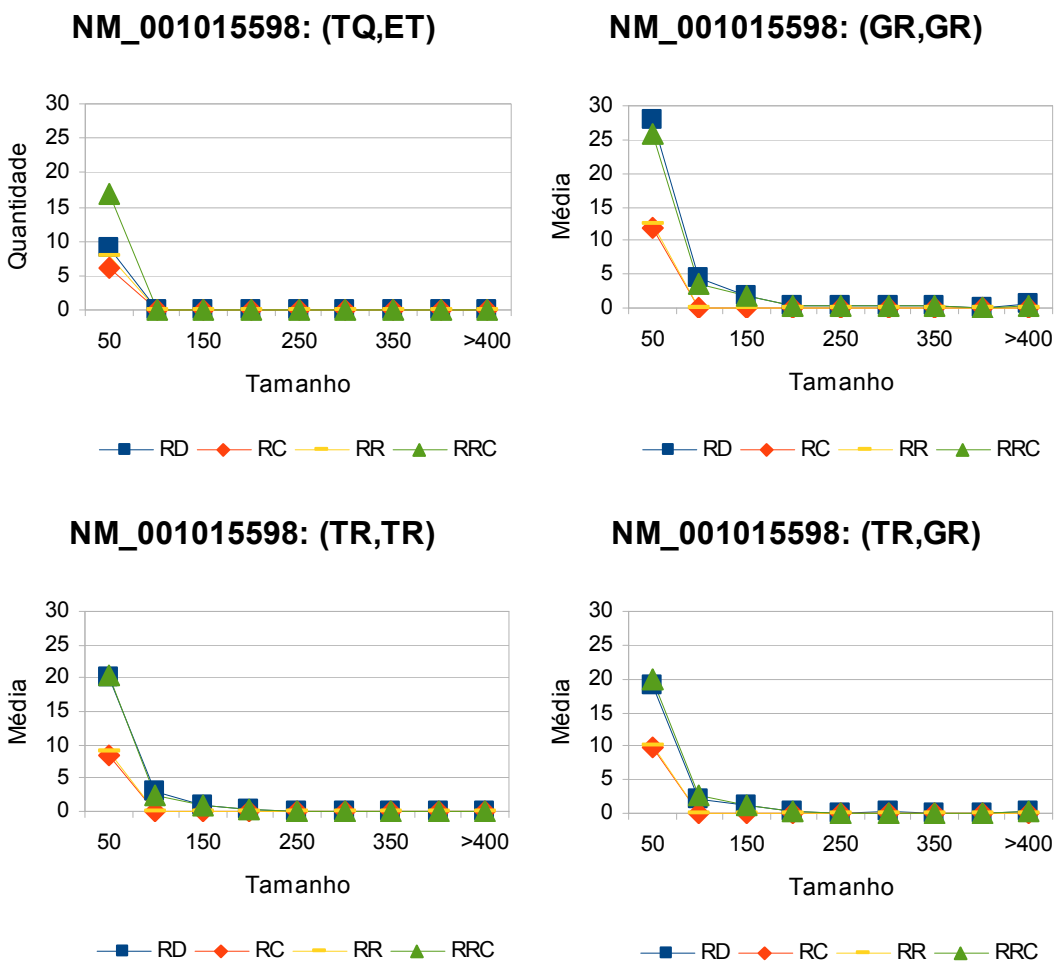


*Figura 66 - SR entre sequências relacionadas com o transcrito [RefSeq:NM_001015598].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [RefSeq:NM_001015598].*
*Fonte: da própria pesquisa.*

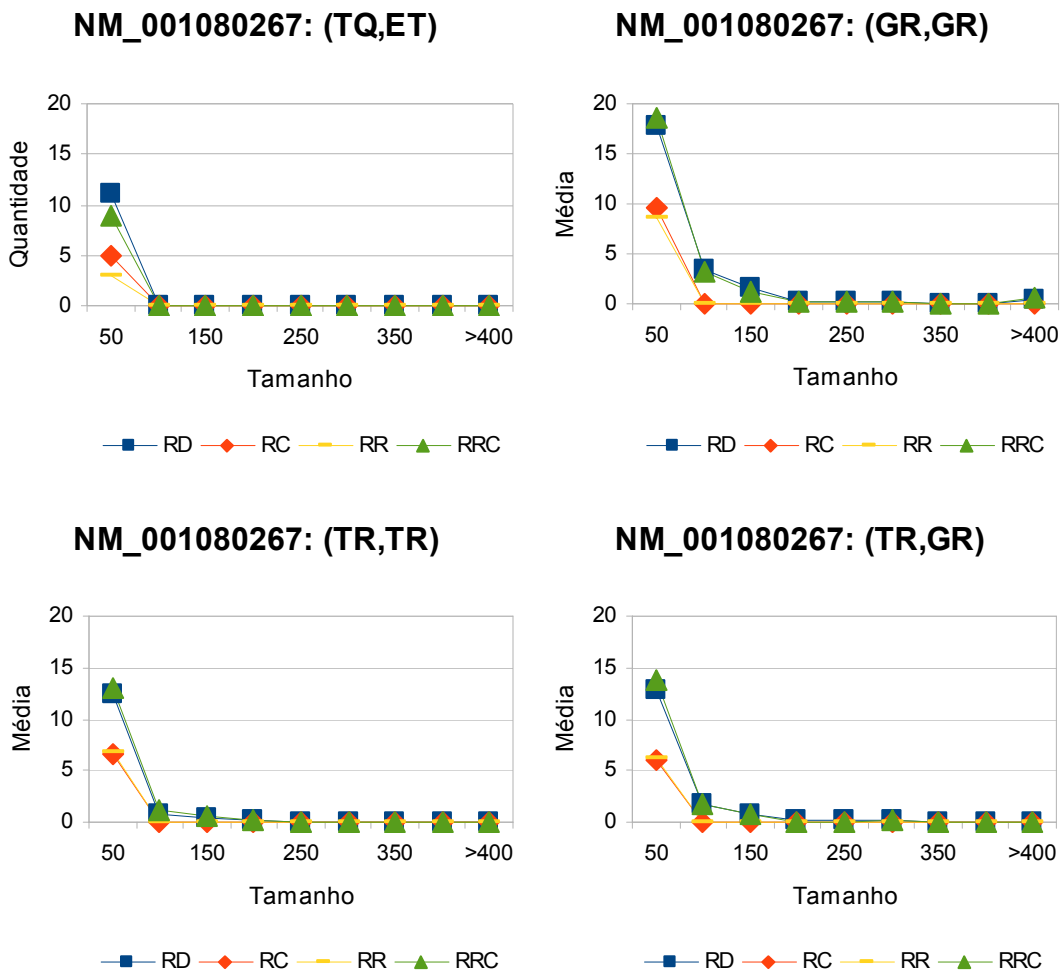


*Figura 67 - SR entre sequências relacionadas com o transcrito [RefSeq:NM_001080267].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [RefSeq:NM_001080267].*
*Fonte: da própria pesquisa.*

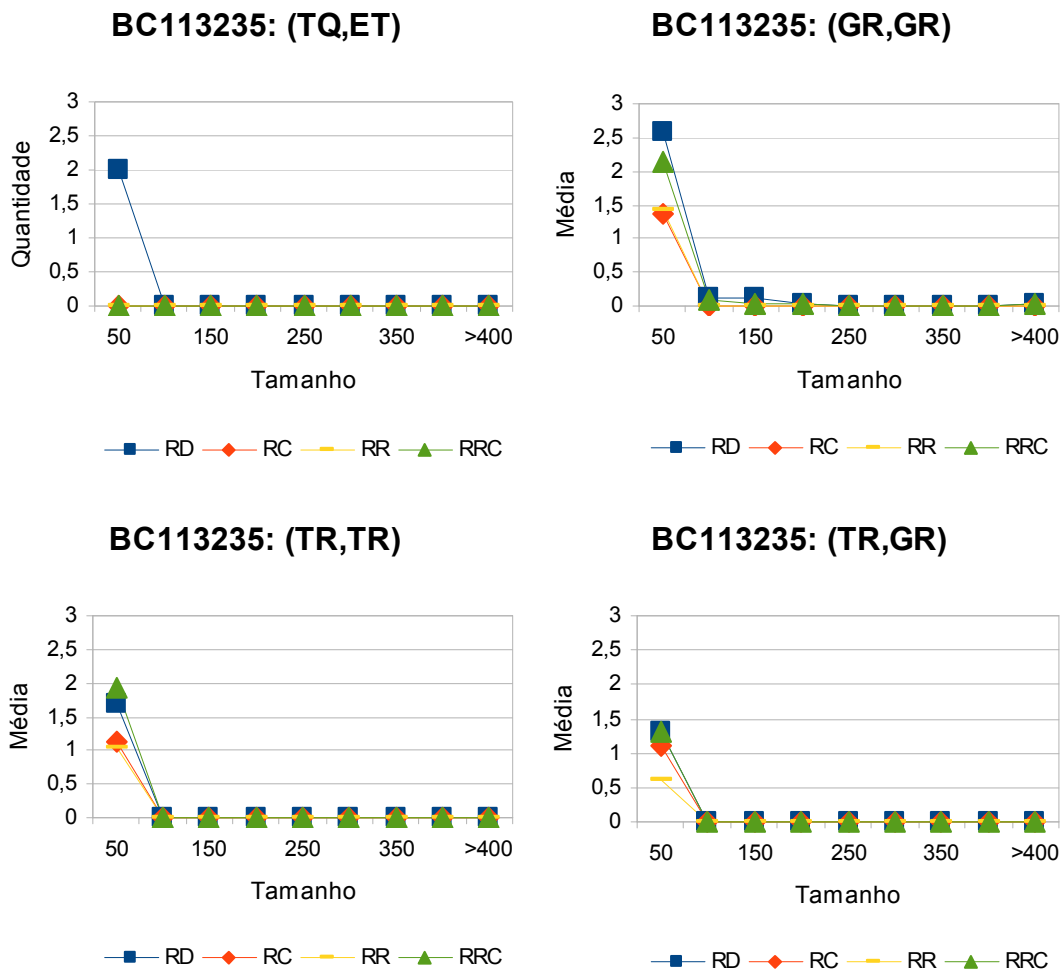


*Figura 68 - SR entre sequências relacionadas com o transcrito [MGC:BC113235].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [MGC:BC113235].*
*Fonte: da própria pesquisa.*

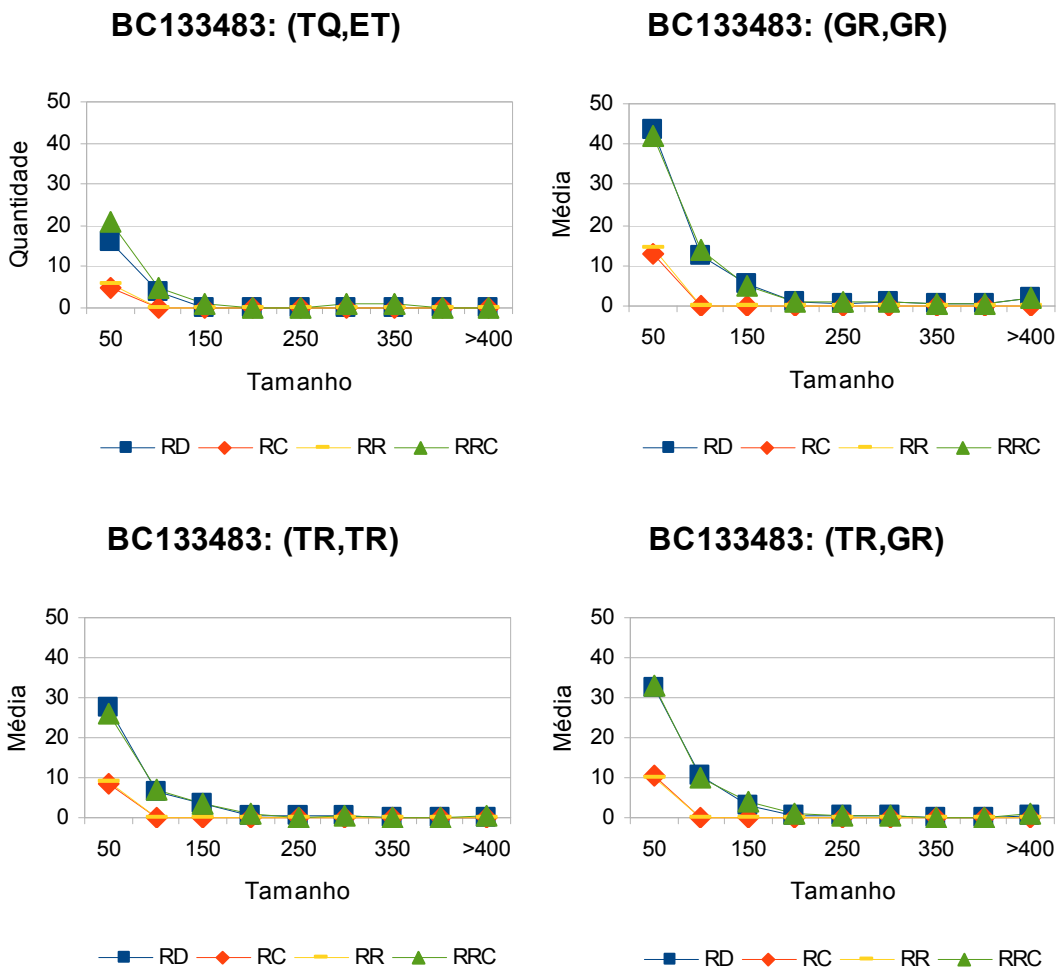


*Figura 69 - SR entre sequências relacionadas com o transcrito [MGC:BC133483].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [MGC:BC133483].*
*Fonte: da própria pesquisa.*

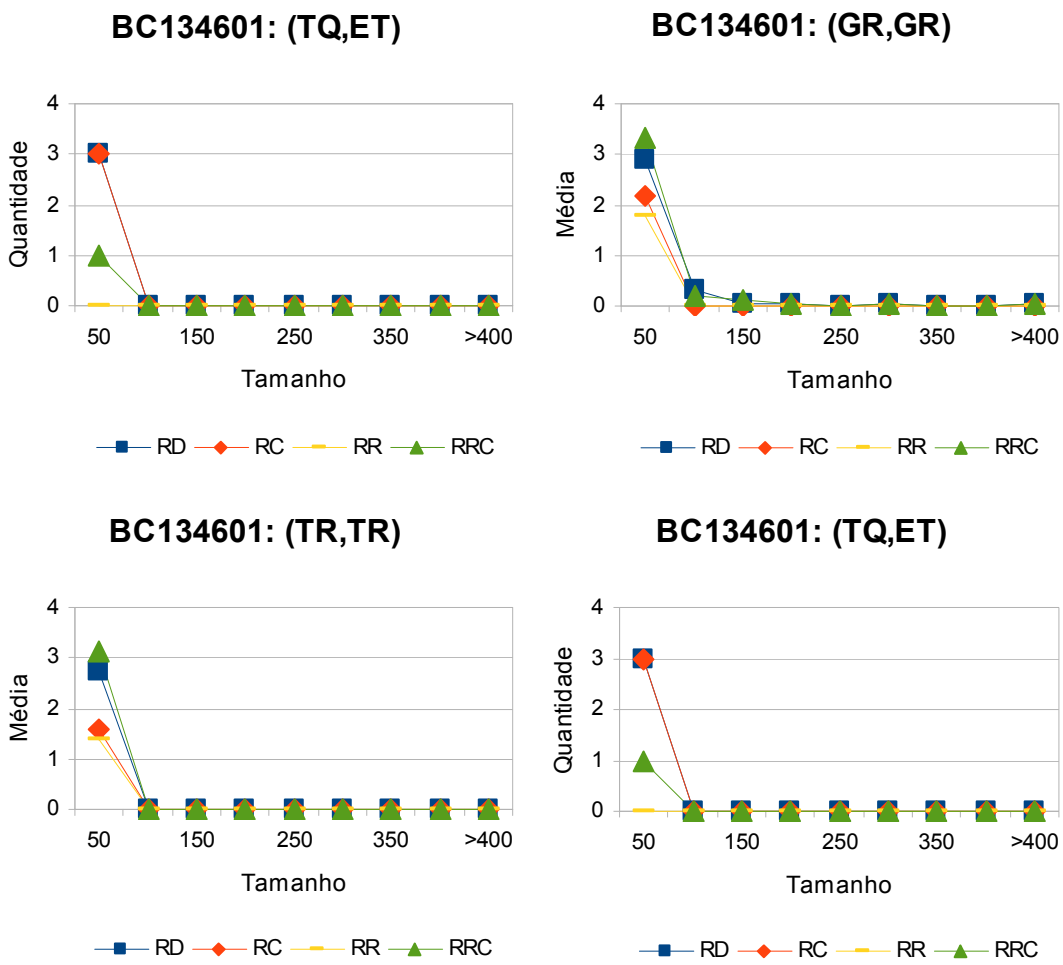


*Figura 70 - SR entre sequências relacionadas com o transcrito [MGC:BC134601].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [MGC:BC134601].*
*Fonte: da própria pesquisa.*

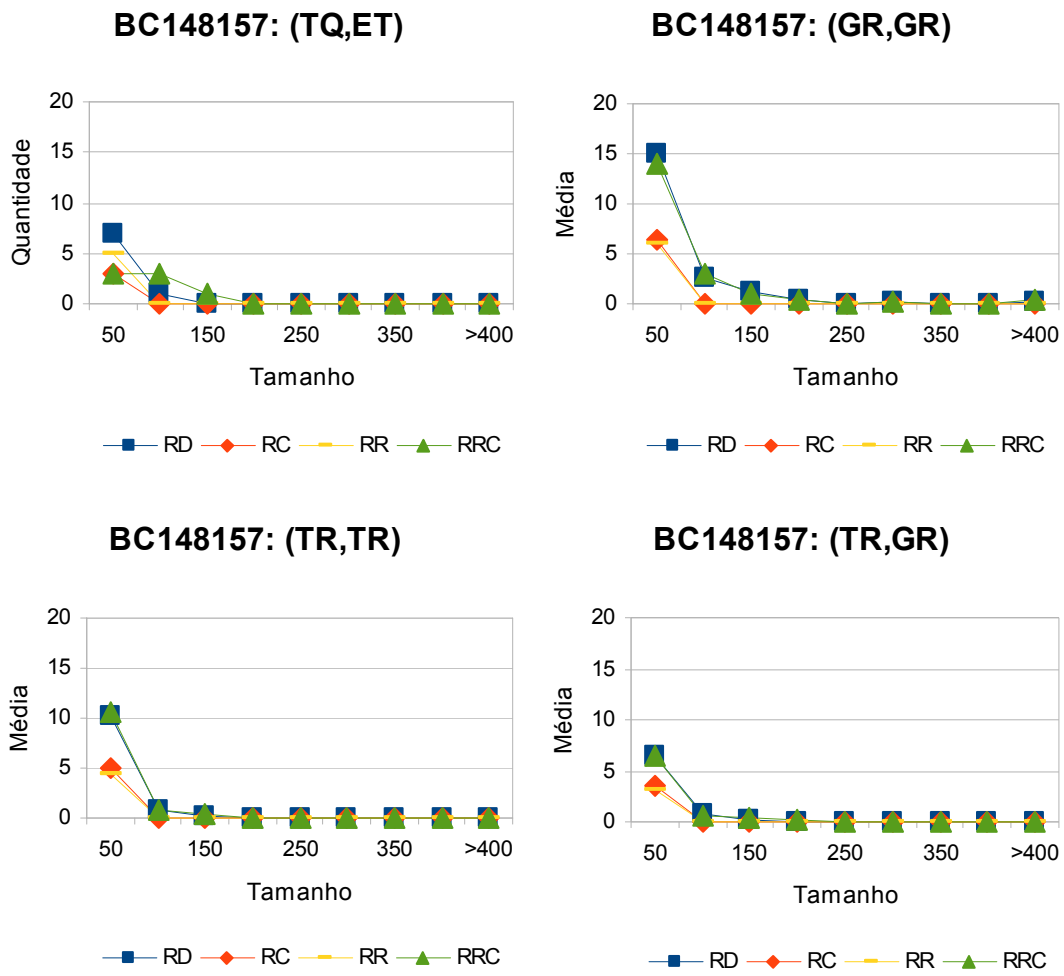


*Figura 71 - SR entre sequências relacionadas com o transcrito [MGC:BC148157].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [MGC:BC148157].*
*Fonte: da própria pesquisa.*

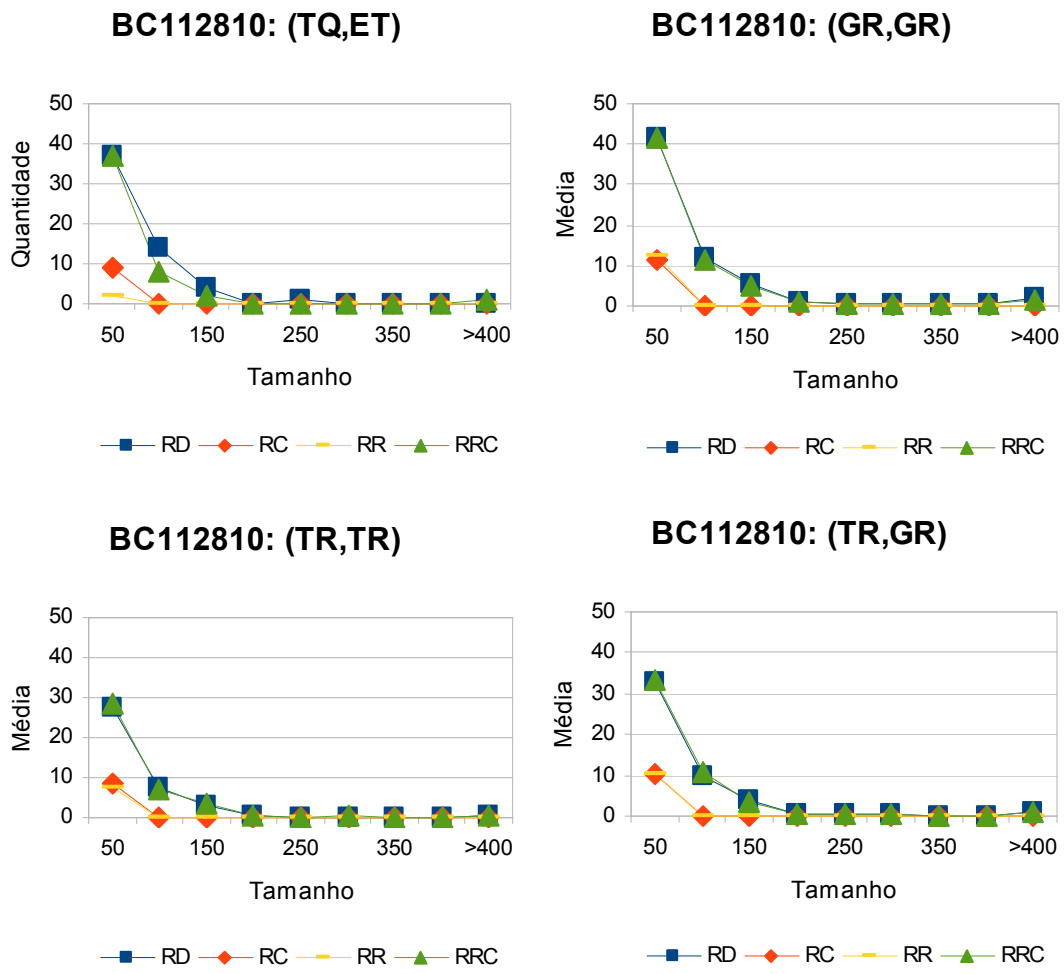


*Figura 72 - SR entre sequências relacionadas com o transcrito [MGC:BC112810].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [MGC:BC112810].*
*Fonte: da própria pesquisa.*

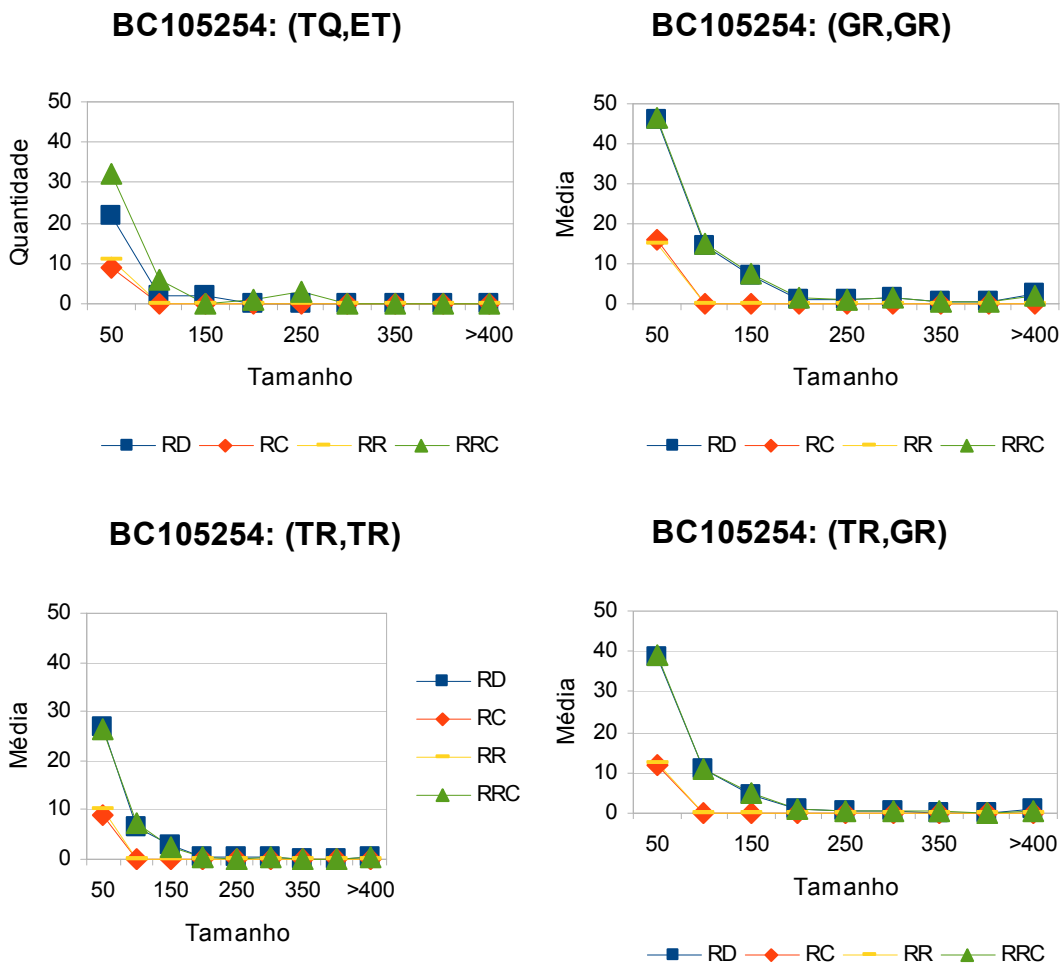


*Figura 73 - SR entre sequências relacionadas com o transcrito [MGC:BC105254].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [MGC:BC105254].*
*Fonte: da própria pesquisa.*

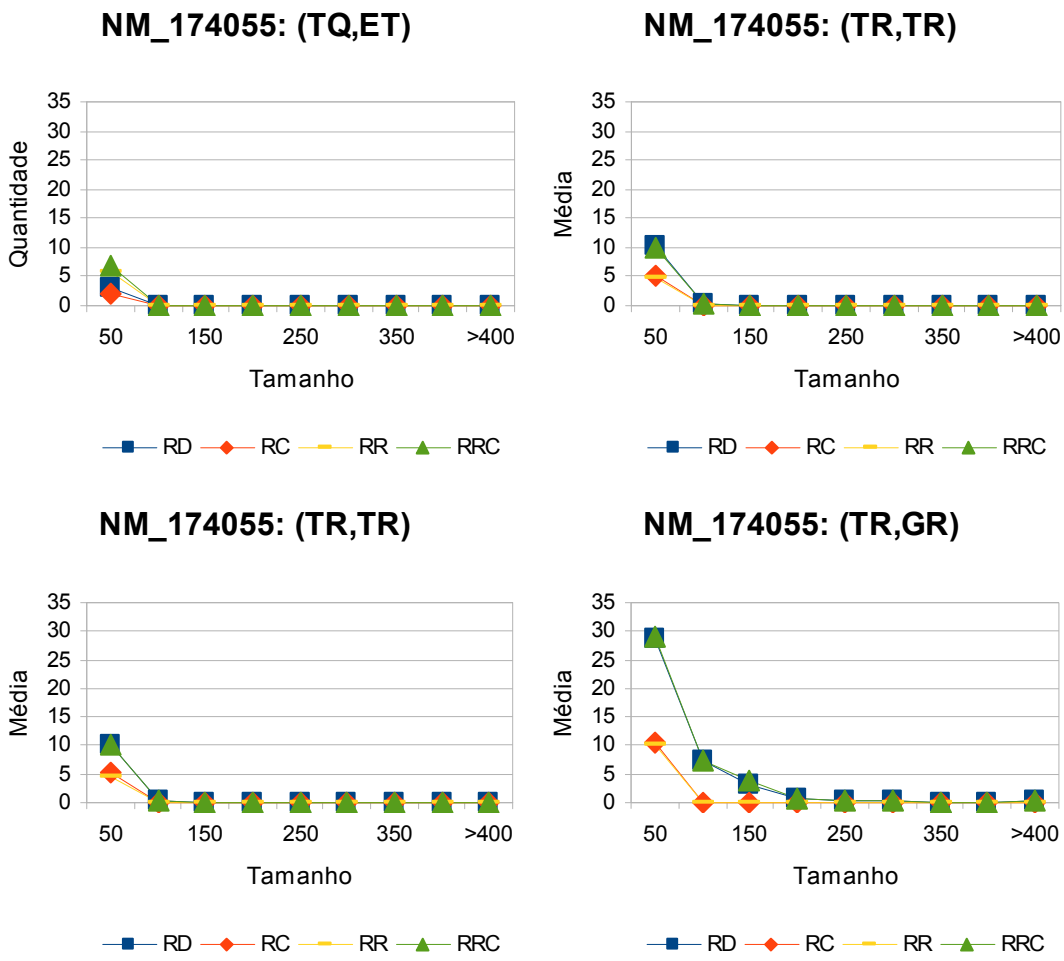


*Figura 74 - SR entre sequências relacionadas com o transcrito [RefSeq:NM_174055].*
*Quantidade média de SR entre diferentes tipos de pares de transcrito com relação às quantidades detectadas no candidato quimérico bovino, [RefSeq:NM_174055].*
*Fonte: da própria pesquisa.*

## Apêndice C

Um transcrito quimérico candidato de bovino, o [MGC:BC133483], é formado pelos transcritos [RefSeq:NM_001163190] e pela evidência de transcrição da TSR, o transcrito [RefSeq:NM_001163188]. Analisando as sequências de aminoácidos de cada uma delas, pode-se verificar que a proteína putativa gerada pelo transcrito quimérico, a AAI33484 (Tabela 13), possui uma parte idêntica à proteína NP_001156662 (em azul na Tabela 14) e o restante é idêntico ao da proteína NP_001156660 (em vermelho na Tabela 15).

*Tabela 13 – Sequência de aminoácidos da proteína AAI33484.*
*Sequência de aminoácidos da proteína AAI33484 gerada pelo transcrito quimérico candidato [RefSeq:BC133483].*

```
> AAI33484 | BC133483
MADTDLFMECEEEELEPWQKISDVIEDSVVEDYNSVDKTTTAGN
SLVQQGGQPLILTQNPAPGLGTMVTQPVLRPVQVMQNANHVTSSPVASQPIFITTQGF
PVRNVRPVQNAMNQVGIVLNVQQGQTVRPITLVPAPGTQFVKPTVGVPQVFSQMTPVR
PGSTMPVRPTTNTFTTVIPATLTIRSTVPQSQSQSTKSTPSTSTTPTATQPTSLGHLA
VQPPGQSSQTQNPKLVTSPIPVFDLQDGGRKICPQCNAHFRVTEALRGHMCYCCPEMV
EYQKKGKSLDSEPSVPSAAKPPSPEKTAPVASTPSSTPIPALSPPTKVPEPNENVGDA
VQTKLIMLVDDFYYGRDGGTVAQLTNFPKVATSFRCPHCTKRLKNNIRFMNHMKHHVE
LDQQNGEVDGHTICQHCYRQFSTPFQLQCHLENVHSPYESTTKCKICEWAFESEPLFL
QHMKDTHKPGEMPYVCQVCQYRSSLYSEVDVHFRMIHEDTRHLLCPYCLKVFKNGSAF
QQHYMRHQKRNVYHCNKCRLQFLFAKDKIEHKLQHHKTFRKPKQLEGLKPGTKVTIRA
SRGQPRTAPVPSSDGPPGSLQEAAPLASSADPLPVFLYPPVQRNVQKRAVRKIQVPQE
ALRRTLFPLGELTKDFVKKIAAENRLHHVLQKKESMGICFIGKRNFENFILEYLQPRP
GRFISIEDNKVLGTHKGWFLYTLGQRANIGGLREPWYVVDKDGAKGDVLVAPRTDHPA
LYRDLLRTGRVHWIAEEPPAALVRDKMMECHFRFRHQMALVPCVLTLNQDGTVWVTAV
KAVRALAPGQFAVFYKGDECLGSGKILRLGPSAYTLQKGQRTSSVAKEGPSDSPGLGP
AP
```

*Fonte: portal UniProt - http://www.uniprot.org/*

*Tabela 14 – Sequência de aminoácidos da proteína NP_001156662.*
*Sequência de aminoácidos da proteína NP_001156662 gerada pelo transcrito [RefSeq:NM_001163190].*

```
> NP_001156662 | NM_001163190
MADTDLFMECEEEELEPWQKISDVIEDSVVEDYNSVDKTTTAGN
SLVQQGGQPLILTQNPAPGLGTMVTQPVLRPVQVMQNANHVTSSPVASQPIFITTQGF
PVRNVRPVQNAMNQVGIVLNVQQGQTVRPITLVPAPGTQFVKPTVGVPQVFSQMTPVR
PGSTMPVRPTTNTFTTVIPATLTIRSTVPQSQSQSTKSTPSTSTTPTATQPTSLGHLA
VQPPGQSSQTQNPKLVTSPIPVFDLQDGGRKICPQCNAHFRVTEALRGHMCYCCPEMV
EYQKKGKSLDSEPSVPSAAKPPSPEKTAPVASTPSSTPIPALSPPTKVPEPNENVGDA
VQTKLIMLVDDFYYGRDGGTVAQLTNFPKVATSFRCPHCTKRLKNNIRFMNHMKHHVE
LDQQNGEVDGHTICQHCYRQFSTPFQLQCHLENVHSPYESTTKCKICEWAFESEPLFL
QHMKDTHKPGEMPYVCQVCQYRSSLYSEVDVHFRMIHEDTRHLLCPYCLKVFKNGSAF
QQHYMRHQKRNVYHCNKCRLQFLFAKDKIEHKLQHHKTFRKPKQLEGLKPGTKVTIRA
SRGQPRTAPVPSSDGPPGSLQEAAPLASSADPLPVFLYPPVQRNVQKRAVRKMSVMGR
QTCLECSFEIPDFPNHFPTYVHCSLCRYSTCCSRAYANHMINNHVPRKSPKYLALFKN
SVSGSKLACTSCTFVTSVGDAMAKHLVFNPSHRSSSILPRGLTWISHSRHGQTRDRAH
DRNLKNLYPPPSFSSNKAATVKSAGVTPAEPEELPAPVAQALPSPASTATPPPTPTHL
QTLALPPLAAEEAECLNVDDQDEGSPVTQEPEPASGGGSSSGIGKKEQLSVKKLRVVL
FALCCNTEQAAEHFRNPQRRIRRWLRRFQASQGESLEGKYLSLEAEEKLAEWVLTQRE
QQLPVNEETLFQKATKIGRSLEGGFKISYEWAVRFMLRHHLTPHARRAVAHTLPKDVA
ENAGLFIEFVQRQIHNQDLSLSMIVAIDEVSLFLDTEVLSSEDRKENALQTVGTGEPW
CDVVLAILADGTVLPTLVFYRGQVDQPANVPDSILLEAKESGYSDDEIMELWSTRVWQ
KHTACQRSKGMLVMDCHRTHLSEEVLAMLSASSTLPAVIPAGCSSKIQPLDVCIKRTV
KNFLHKKWKEQAREMADTACDSDVLLQLVLVWLAEVLGVIGDCPELVQRSFLVASVLP
GPDGNMNSPTRNADMQEELIASLEEQLKLSGEQSEEPSASTPRPRSSPEETIEPESLH
QLFEGESETESFYGFEEADLDLMEI
```

*Fonte: portal UniProt - http://www.uniprot.org/*

*Tabela 15 – Sequência de aminoácidos da proteína NP_001156660.*
*Sequência de aminoácidos da proteína NP_001156660 gerada pelo transcrito [RefSeq:NM_001163188].*

```
> NP_001156660 | NM_001163188
MQVARHVVCAVSGGVDSAVAALLLRRRGYQVTGVFMKNWDSLDE
HGVCTADRDCEDAYRVCRILDIPFRQVSYAKEYWNDVFSDFLNEYEKGRTPNPDIVCN
KHIKFRCFFNYAVDNLGADAVATGHYARTSLEDEEVFQQKHIKRPEGLFRNRFEVRNA
VKLLQAADSFKDQTFFLSQVPQEALRRTLFPLGELTKDFVKKIAAENRLHHVLQKKES
MGICFIGKRNFENFILEYLQPRPGRFISIEDNKVLGTHKGWFLYTLGQRANIGGLREP
WYVVDKDGAKGDVLVAPRTDHPALYRDLLRTGRVHWIAEEPPAALVRDKMMECHFRFR
HQMALVPCVLTLNQDGTVWVTAVKAVRALAPGQFAVFYKGDECLGSGKILRLGPSAYT
LQKGQRTSSVAKEGPSDSPGLGPAP
```

*Fonte: portal UniProt - http://www.uniprot.org/*