UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA
Programa de Pós-Graduação em Ciências Biológicas (Bioquímica)

THIAGO LUIZ ARAUJO MILLER

sideRETRO: uma ferramenta de bioinformática dedicada à identificação de inserções polimórficas, germinativas ou somáticas, de pseudogenes processados

Versão original da Tese defendida

São Paulo

Data do Depósito na SPG:
03/05/2022

THIAGO LUIZ ARAUJO MILLER

sideRETRO: uma ferramenta de bioinformática dedicada à identificação de inserções polimórficas, germinativas ou somáticas, de pseudogenes processados

Tese apresentada ao Instituto de Química da Universidade de São Paulo para obtenção do título de Doutor em Ciências (Bioquímica**).**

Orientador: Dr. Pedro Alexandre Favoretto Galante

São Paulo

2022

Ficha Catalográfica elaborada eletronicamente pelo autor, utilizando o programa desenvolvido pela Seção Técnica de Informática do ICMC/USP e adaptado para a Divisão de Biblioteca e Documentação do Conjunto das Químicas da USP

Bibliotecária responsável pela orientação de catalogação da publicação:
Marlene Aparecida Vieira - CRB - 8/5562

M647s

Miller, Thiago Luiz Araujo
sideRETRO: uma ferramenta de bioinformática dedicada à identificação de inserções polimórficas, germinativas ou somáticas, de pseudogenes processados / Thiago Luiz Araujo Miller. - São Paulo, 2022.
208 p.

Tese (doutorado) - Instituto de Química da Universidade de São Paulo. Departamento de Bioquímica.
Orientador: Galante, Pedro Alexandre Favoretto

1. Bioinformática. 2. Pseudogenes processados. 3. Retrocópias. 4. Polimorfismo. 5. Genômica. I. T. II. Galante, Pedro Alexandre Favoretto, orientador.

FOLHA DE AVALIAÇÃO

(Esta página foi intencionalmente deixada em branco)

Dedico este trabalho aos meus queridos pais Dona Neide e Seu Jayme e a minha querida irmã Miriam.

AGRADECIMENTOS

A D’us por tudo.

Ao Dr. Pedro A. F. Galante pela orientação e por ter confiado a mim o desenvolvimento deste trabalho.

À Dra. Cibele Masotti pelos conselhos e pelas palavras de incentivo.

Ao J. Leonel L. Buzzo pela barda amizade desde os saudosos tempos de UNESP.

Ao Diogo Pessoa pela ajuda no meu processo de adaptação à USP.

À Fernanda Orpinelli pelo carinho e amizade desde o início da minha iniciação científica.

Ao Rodrigo Barreiro e à Gabriela Guardia pelo auxílio na criação do símbolo visual da ferramenta sideRETRO.

Ao clube do Agar.io – David Berl, Felipe A. C. dos Santos, Filipe F. dos Santos, Rafael Mercuri, Ramon T. do Carmo, Ricardo Piuco – pelas divertidas e inusitadas conversas científicas.

Ao técnico de informática Daniel Ohara por todo o suporte prestado.

Aos membros do Centro de Oncologia Molecular do Hospital Sírio-Libanês pelo ambiente de trabalho acolhedor.

Ao Instituto de Química da USP pela oportunidade de realização do doutorado.

Ao Instituto de Ensino e Pesquisa do Hospital Sírio-Libanês pela infraestrutura.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“Suponho que seja tentador, se a única ferramenta que você tem é um martelo, tratar tudo como se fosse um prego.”

(MASLOW, 1966, p. 15, tradução nossa).

RESUMO

MILLER, T. L. A. **sideRETRO**: uma ferramenta de bioinformática dedicada à identificação de inserções polimórficas, germinativas ou somáticas, de pseudogenes processados. 2022. Tese – Programa de Pós-Graduação em Ciências Biológicas (Bioquímica). Instituto de Química, Universidade de São Paulo, São Paulo. 2022.

Os avanços metodológicos e instrumentais decorrentes do Projeto Genoma Humano formaram o arcabouço necessário para o surgimento das tecnologias de sequenciamento de DNA de Nova Geração, as quais se caracterizam por um custo reduzido, uma baixa demanda operacional e a produção de um grande volume de dados por experimento. Concomitantemente a isso, o aumento no poder de processamento computacional permitiu o desenvolvimento de análises genéticas em larga escala, de modo que, atualmente, é possível estudar características genômicas individualizadas e, até então, pouco ou nunca exploradas. Dentre essas características, aquelas relacionadas às variações estruturais em genomas têm recebido bastante atenção. Os pseudogenes processados, ou retrocópias, são variações estruturais causadas pela duplicação de genes codificadores mediante à transposição de seu RNA mensageiro maduro pela maquinaria enzimática de LINE-1. As retrocópias podem estar fixadas, ou seja, presentes em todos os genomas de uma dada espécie, os quais são representados pela montagem modelo do genoma de referência, ou podem não estar fixadas, sendo polimórficas, germinativas ou somáticas. No entanto, o conhecimento acerca das retrocópias não fixadas ainda é limitado devido à falta de ferramentas de bioinformática dedicadas a sua identificação e anotação em dados de sequenciamento de DNA. Posto isso, este trabalho apresenta o sideRETRO – um programa computacional especializado na detecção de pseudogenes processados ausentes do genoma de referência, mas presentes em dados de sequenciamento de genoma completo e exoma de outros indivíduos. Além de apontar para a presença de retrocópias não fixadas, o sideRETRO é capaz de anotar várias outras características relacionadas a esses evento, tais como: a coordenada genômica de inserção do pseudogene processado, a qual constitui o cromossomo, o ponto de inserção e a fita de DNA (líder or retardada); o contexto genômico do evento (exônico, intrônico ou intergênico); a genotipagem (presente ou ausente) e a haplotipagem (em homozigose ou heterozigose). Para atestar a eficiência da ferramenta, o sideRETRO foi executado para dados simulados e para dados reais validados experimentalmente por um grupo independente. Portanto, em resumo, nesta tese são descritos o desenvolvimento e o uso do sideRETRO – uma ferramenta computacional robusta e eficiente, designada para identificar e anotar pseudogenes processados não fixados. Por fim, vale destacar que o sideRETRO preenche uma lacuna metodológica e possibilita novas hipóteses e investigações sistemáticas no campo de chamada de variantes estruturais.

Palavras-chave: Bioinformática. Pseudogenes processados. Retrocópias. Polimorfismo. Genômica.

ABSTRACT

MILLER, T. L. A. **sideRETRO**: a bioinformatics tool for identifying somatic and polymorphic insertions of processed pseudogenes. 2022. Tese – Programa de Pós-Graduação em Ciências Biológicas (Bioquímica). Instituto de Química, Universidade de São Paulo, São Paulo. 2022.

The methodological and instrumental advances resulting from the Human Genome Project have created the necessary framework to the emergence of Next Generation DNA sequencing technologies, which are characterized by a reduced cost, low operational demand and the generation of a large volume of data per experiment. Concomitantly with this, the increase in computational processing power has driven the development of large-scale genetic analyses, which allowed us to study individualized genomic traits little or never explored before. Among these characteristics, those related to structural variations in genomes have received much attention. Processed pseudogenes, or retrocopies, are structural variations caused by the duplication of coding genes through the transposition of their mature messenger RNA by the LINE-1 enzymatic machinery. Retrocopies can be fixed (i.e., present in all genomes of a given species and included into the assembly of the reference genome) or unfixed, being polymorphic, germinal or somatic. However, knowledge about unfixed retrocopies is still limited due to the lack of bioinformatics tools dedicated to their identification and annotation in DNA sequencing data. Therefore, this work presents sideRETRO – a computer program specialized in the detection of processed pseudogenes absent from the reference genome, but present in whole genome and exome sequencing data from other individuals. In addition to pointing out the presence of unfixed retrocopies, sideRETRO is able to annotate several other characteristics related to these events, such as: the genomic coordinate of the processed pseudogene insetion, which constitutes the chromosome, the insertion point and the DNA strand (leader or retard); the genomic context of the event (exonic, intronic or intergenic); genotyping (present or absent) and haplotyping (homozygous or heterozygous). To certify the sideRETRO efficiency, it was run on simulated data and on real data experimentally validated by an independent group. Therefore, in summary, this thesis describes the development and use of sideRETRO – a robust and efficient computational tool, designed to identify and annotate unfixed processed pseudogenes. Finally, it is worth noting that sideRETRO fills a methodological gap and allows new hypotheses and systematic investigations in the field of structural variant calling.

Keywords: Bioinformatics. Processed pseudogenes. Retrocopies. Polymorphism. Genomics.

LISTA DE FIGURAS

Figura 1 - A química Rosalind Elsie Franklin.....31
Figura 2 - A citogeneticista Barbara McClintock.....34
Figura 3 - Mecanismo de transposição para os transposons de classes I e II.....36
Figura 4 - Estrutura genética do LINE-1.....38
Figura 5 - Reação de TPRT para L1.....40
Figura 6 - Retrotransposição em *trans* de um gene codificador.....42
Figura 7 - Alinhamentos indistinguíveis.....56
Figura 8 - Leituras pareadas alinhadas em diferentes éxons.....57
Figura 9 - Exemplo de leituras pareadas que cruzam o íntron.....57
Figura 10 - Leituras pareadas alinhadas em cromossomos diferentes.....58
Figura 11 - Exemplo de leituras pareadas alinhadas em cromossomos diferentes.. .59
Figura 12 - Leituras pareadas alinhadas em regiões distantes.....60
Figura 13 - Exemplo de leituras pareadas alinhadas em regiões distantes.....60
Figura 14 - Leituras divididas.....62
Figura 15 - Exemplo de leituras pareadas com alinhamento suplementar.....63
Figura 16 - Diagrama de relacionamento de entidade para o banco de dados do sideRETRO.....70
Figura 17 - Três exemplos de sobreposição entre éxon e leitura.....79
Figura 18 - As tabelas que são preenchidas durante o subcomando *process-sample*.....80
Figura 19 - Fusão dos bancos de dados.....83
Figura 20 - Agrupamento de leituras pelo algoritmo DBSCAN.....87
Figura 21 - Preenchimento das tabelas do banco de dados do sideRETRO durante a etapa de agrupamento.....90
Figura 22 - O arquivo com as regiões genômicas proibidas é anotado na tabela *blacklist*.....92
Figura 23 - Armazenamento dos agrupamentos sobrepostos.....97
Figura 24 - Cálculo do ponto de inserção, segundo a posição, anterior ou posterior, do alinhamento suplementar.....100
Figura 25 - retroCNV e seu gene parental estão na mesma fita.....103
Figura 26 - retroCNV e seu gene parental estão em fitas opostas.....103

Figura 27 - Alinhamentos anormais são usados como evidência do alelo alternativo.....107
Figura 28 - Desempenho geral do sideRETRO durante a genotipagem dos 100 indivíduos, sequenciamentos, simulados.....142
Figura 29 - Desempenho geral do sideRETRO durante a genotipagem dos 100 indivíduos simulados, após a remoção dos eventos localizados em regiões altamente repetitivas.....152
Figura 30 - O alinhamento genômico da região da retroCNV do gene parental CACNA1B.....153
Figura 31 - Análise de retroelementos em um contexto de progressão tumoral. Eventos de retroCNVs, ao longo dos cromossomos do genoma humano, de amostras normais e tumorais.....157
Figura 32 - Análise de retroelementos em um contexto de progressão tumoral. Número de retroCNVs (únicos) aumentam com a progressão do câncer colorretal.....158

LISTA DE FLUXOGRAMAS

Fluxograma 1 - Subcomando *process-sample*.....72
Fluxograma 2 - Filtro de qualidade.....73
Fluxograma 3 - Ffiltro de alinhamentos anormais.....75
Fluxograma 4 - Filtro de sobreposição a um éxon.....77
Fluxograma 5 - Registro dos éxons codificantes no banco de dados do sideRETRO.....78
Fluxograma 6 - Subcomando *merge-call*.....82
Fluxograma 7 - Fase de agrupamento dos alinhamentos anormais.....85
Fluxograma 8 - O algoritmo abstrato de DBSCAN.....89
Fluxograma 9 - Filtros de agrupamento.....93
Fluxograma 10 - Anotação de retroCNVs.....94
Fluxograma 11 - Resolução de agrupamentos sobrepostos.....96
Fluxograma 12 - Cálculo do ponto de inserção.....101
Fluxograma 13 - O algoritmo de cálculo da orientação da retroCNV.....105
Fluxograma 14 - O algoritmo de genotipagem.....109
Fluxograma 15 - O teste do sideRETRO com dados simulados.....117
Fluxograma 16 - O teste do sideRETRO com dados reais.....120

LISTA DE TABELAS

Tabela 1 - *Flags bit* a *bit* do arquivo SAM/BAM verificadas pelo filtro de qualidade.....74
Tabela 2 - Valores padrão para o nível de qualidade *Phred* e a frequência de bases.....75
Tabela 3 - *Flag* que define um alinhamento suplementar no formato SAM/BAM.....76
Tabela 4 - Valor padrão para leituras distantes.....76
Tabela 5 - Valor mínimo para a sobreposição entre um éxon e uma das leituras do alinhamento anormal.....79
Tabela 6 - Sumário dos valores padrões para o subcomando *process-sample*.....81
Tabela 7 - Sumário das *flags bit* a *bit* do arquivo no formato SAM verificadas pelo sideRETRO.....81
Tabela 8 - Exemplo de uma fila de agrupamentos.....86
Tabela 9 - Opções para os parâmetros de DBSCAN com seus respectivos valores padrão.....88
Tabela 10 - Sumário dos *bits* dos filtros de agrupamento.....93
Tabela 11 - Opções para os parâmetros do filtro de agrupamento.....94
Tabela 12 - Sumário dos *bits* relativos ao contexto dos genes parentais dos agrupamentos sobrepostos.....98
Tabela 13 - Opção para a resolução de agrupamentos sobrepostos.....98
Tabela 14 - *Bits* usados pelo sideRETRO para identificar a forma como se calculou o ponto de inserção da retroCNV.....100
Tabela 15 - Sumário dos valores padrões para o subcomando *merge-call*.....108
Tabela 16 - Opções para a anotação do arquivo VCF.....113
Tabela 17 - As 100 retroCNVs simuladas, com seus respectivos genes parentais e posições genômicas.....132
Tabela 18 - As 79 retroCNVs anotadas pelo sideRETRO durante a simulação.....135
Tabela 19 - Detecção das retroCNVs segundo as categorias de simulação.....138
Tabela 20 - Erro na predição do ponto de inserção.....138
Tabela 21 - Erro na predição da fita.....138
Tabela 22 - Desempenho do sideRETRO durante a genotipagem dos 100 indivíduos, sequenciamentos, simulados.....138

Tabela 23 - RetroCNVs validadas experimentalmente por PCR e genotipadas por Abyzov et al. e pelo sideRETRO em indivíduos de 14 populações humanas.....143
Tabela 24 - Genotipagem das retroCNVs validadas experimentalmente por Abyzov et al. em 14 populações humanas do projeto 1000 Genomas.....143
Tabela 25 - Os 21 eventos simulados de retroCNV não encontrados pelo sideRETRO.....148
Tabela 26 - Desempenho do sideRETRO durante a genotipagem dos 100 indivíduos simulados, após a remoção dos eventos localizados em regiões altamente repetitivas.....149

LISTA DE QUADROS

Quadro 1 - Resumo dos tipos de padrões de alinhamentos anormais.....64
Quadro 2 - Cálculo do ponto de inserção por alinhamento suplementar.....99
Quadro 3 - Sumário da regra usada para se determinar a fita, na qual se inseriu a retroCNV.....104
Quadro 4 - Descrição dos campos do arquivo VCF que são adicionados pelo sideRETRO.....110
Quadro 5 - As 9 retroCVNs validadas por PCR por Abyzov e colaboradores.....118
Quadro 6 - As 14 populações do 1KGP que foram genotipadas por Abyzov e colaboradores.....119
Quadro 7 - Atributos da retroCNV detectados pelo sideRETRO.....124
Quadro 8 - Opções obrigatórias para o *process-sample*.....127
Quadro 9 - Demais opções para o *process-sample*.....127
Quadro 10 - Opções obrigatórias para o *merge-call*.....129
Quadro 11 - Demais opções para o *merge-call*.....129
Quadro 12 - Opções para o *make-vcf*.....131

LISTA DE ABREVIATURAS E SIGLAS

1KGP	*1000 Genomes Project*
Ac	Lócus Ativador
BAM	*Binary Alignment Map*
BED	*Browser Extensible Data*
CC	*Coiled-coil*
CTD	*C-Terminal Domain*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
Ds	Lócus de Dissociação
FN	Falso Negativo
FP	Falso Positivo
GDH	Glutamato Desidrogenase
GFF3	*General Feature Format version 3*
GPLv3	*General Public License version 3*
GTF	*Gene Transfer Format*
H	*Hard clipping*
HET	Heterozigoto
HGP	*The Human Genome Project*
HIV	*Human Immunodeficiency Virus*
HOA	Homozigoto Alternativo
HOR	Homozigoto Referência
ICGC	*The International Cancer Genome Consortium*
LINE	*Long Interspersed Nuclear Element*
LTR	*Long Terminal Repeats*
M	*Match*
MEDSE	*Median Squared Error*
MSE	*Mean Squared Error*
NCBI	*National Center for Biotechnology Information*
NGS	*Next Generation Sequencing*
ORF	*Open Reading Frame*
PCR	*Polymerase Chain Reaction*
PTEN	*Phosphatase and Tensin homolog*

RDBMS	*Relational Database Management System*
RNA-Seq	Sequenciamento de RNA
S	*Soft clipping*
SAM	*Sequence Alignment Map*
SINE	*Short Interspersed Nuclear Element*
SQL	*Structured Query Language*
TCGA	*The Cancer Genome Atlas*
TPRT	*Target-primed Reverse Transcription*
UTR	*Untranslated Region*
VCF	*Variant Call Format*
VP	Verdadeiro Positivo
WES	*Whole Exome Sequencing*
WGS	*Whole Genome Sequencing*
cDNA	DNA complementar
ceRNA	*Competing endogenous RNA*
mRNA	RNA mensageiro
miRNA	MicroRNA
retroCNV	*Gene Copy-Number Variant caused by retrotransposition*

SUMÁRIO

1 INTRODUÇÃO.....31
1.1 O GENOMA HUMANO E O SEQUENCIAMENTO DE DNA.....31
1.2 A DESCOBERTA DOS ELEMENTOS TRANSPONÍVEIS.....33
1.3 A CLASSIFICAÇÃO DOS TRANSPOSONS.....35
1.3.1 A classificação dos transposons de classe I.....36
1.4 LINE-1.....37
1.4.1 A estrutura do gene de LINE-1.....37
1.4.2 A proteína ORF1p.....38
1.4.3 A proteína ORF2p.....39
1.4.4 O processo de mobilização do elemento L1.....39
1.5 ELEMENTOS TRANSPONÍVEIS NÃO AUTÔNOMOS E A RETROTRANSPOSIÇÃO EM *TRANS*.....41
1.5.1 SINEs.....41
1.5.2 Pseudogenes processados.....42
1.6 RETROCÓPIAS DE GENES CODIFICADORES E O SEU PAPEL.....43
1.6.1 Exemplo de retrocópia funcional: GLUD2.....44
1.6.2 Exemplo de retrocópia funcional: PTENP1.....45
1.7 RETROCÓPIAS FIXAS E POLIMÓRFICAS NO GENOMA HUMANO.....45
1.8 ALGORITMOS PARA IDENTIFICAÇÃO DE RETROCÓPIAS NÃO FIXADAS.....46
1.9 SIDERETRO.....47
2 OBJETIVOS.....51
2.1 OBJETIVO GERAL.....51
2.2 OBJETIVOS ESPECÍFICOS.....51
3 METODOLOGIA.....55
3.1 RESOLVENDO AS DIFICULDADES INTRÍNSECAS AO ALINHAMENTO DAS SEQUÊNCIAS CONTRA O GENOMA.....55
3.1.1 Os alinhamentos indistinguíveis.....55
3.1.2 Leituras pareadas alinhadas em diferentes éxons.....56
3.1.3 Leituras pareadas alinhadas em cromossomos diferentes.....57
3.1.4 Leituras pareadas alinhadas em regiões distantes.....59

3.1.5 Leituras divididas (leituras com alinhamento suplementar).....61
3.1.6 Resumo dos alinhamentos anormais.....63
3.2 O DESENVOLVIMENTO DO SIDERETRO.....64
3.2.1 Banco de dados.....66
3.2.2 Subcomando *process-sample*.....71
3.2.2.1 Filtro de qualidade.....73
3.2.2.2 Filtro de alinhamentos anormais.....75
3.2.2.3 Filtro de sobreposição a um éxon.....77
3.2.2.4 Registro do alinhamento.....79
3.2.3 Subcomando *merge-call*.....81
3.2.3.1 Fusão dos bancos de dados.....83
3.2.3.2 Agrupamento.....84
3.2.3.2.1 Agrupamento por gene codificador.....85
3.2.3.2.2 Agrupamento por cromossomo.....86
3.2.3.2.3 Criação da fila de agrupamentos.....86
3.2.3.2.4 DBSCAN.....87
3.2.3.3 Filtro de agrupamento.....90
3.2.3.4 Anotação de retroCNVs.....94
3.2.3.4.1 Resolução de agrupamentos sobrepostos.....95
3.2.3.4.2 Cálculo do ponto de inserção.....98
3.2.3.4.3 Cálculo da orientação.....102
3.2.3.4.4 Genotipagem.....106
3.2.4 Subcomando *make-vcf*.....110
3.3 TESTANDO O SIDERETRO COM DADOS SIMULADOS.....113
3.3.1 Desenho das retroCNVs.....113
3.3.2 Desenho da coorte.....114
3.3.3 Simulação dos genomas sequenciados.....115
3.3.4 Alinhamento contra o genoma de referência.....115
3.3.5 Análise com o sideRETRO.....116
3.4 TESTANDO O SIDERETRO COM DADOS REAIS.....118
4 RESULTADOS.....123
4.1 O SIDERETRO NUMA CASCA DE NOZ.....123
4.1.1 Uma visão geral do sideRETRO.....123
4.1.2 Instalando o sideRETRO.....124

4.1.3 Usando o sideRETRO: o pré-processamento dos dados de NGS....125
4.1.4 Usando o sideRETRO: *process-sample*....126
4.1.5 Usando sideRETRO: *merge-call*....128
4.1.6 Usando sideRETRO: *make-vcf*....130
4.1.6.1 Arquivo VCF gerado pelo sideRETRO....131
4.2 RESULTADO DO SIDERETRO PARA DADOS SIMULADOS....132
4.3 RESULTADOS DO SIDERETRO PARA DADOS REAIS....142
5 DISCUSSÃO....147
5.1 ESTIMANDO O DESEMPENHO DO SIDERETRO COM DADOS SIMULADOS....147
5.2 ESTIMANDO O DESEMPENHO DO SIDERETRO EM DADOS REAIS 152
5.3 DISCUSSÃO GERAL ACERCA DO SIDERETRO....154
6 CONCLUSÃO....161
REFERÊNCIAS....165
APÊNDICE A - Artigo do sideRETRO publicado na revista *Bioinformatics*....179
APÊNDICE B - *Script make_rtc.pl*....183
APÊNDICE C - *Script make_cohort.pl*....187
APÊNDICE D - *Script compare_sim.pl*....195
APÊNDICE E - *Script confusion_analysis.pl*....201
APÊNDICE F - *Script make_simulation.sh*....205
ANEXO A – Súmula Curricular....211

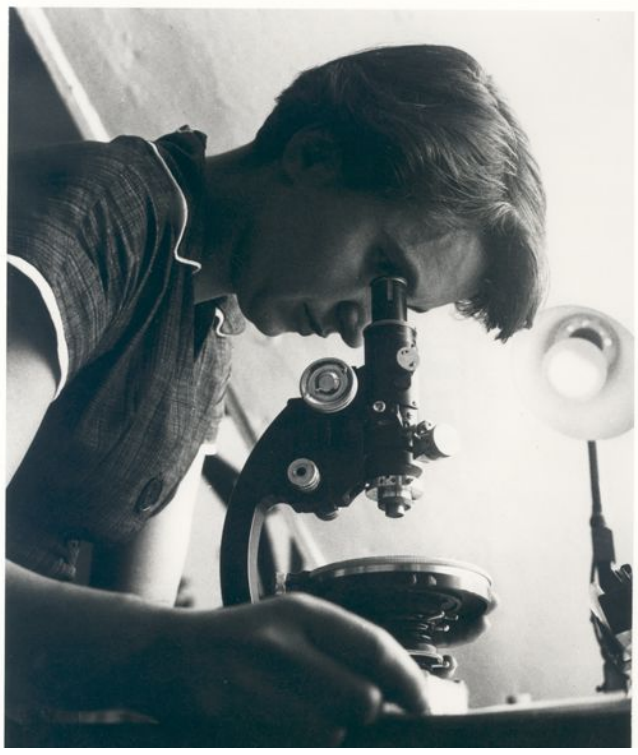
1 Introdução

1 INTRODUÇÃO

1.1 O GENOMA HUMANO E O SEQUENCIAMENTO DE DNA

Desde o descobrimento e o estabelecimento de um modelo para a estrutura de dupla hélice do DNA por Rosalind Elsie Franklin (1920-1958, Figura 1), James Dewey Watson (1928) e Francis Crick (1916-2004) (BRIDGES, 2003; CRICK, 1974; WATSON, 1970), o campo de pesquisa em genética molecular recebeu uma considerável atenção para o seu desenvolvimento teórico (CHARGAFF, 1974) e empírico (CREAGER; MORGAN, 2008). As inovações tecnológicas provenientes disso fomentaram o arcabouço necessário para o estabelecimento em 1990 do Projeto Genoma Humano[1] (HGP) (MARDIS, 2013; WATSON; JORDAN, 1989) com a ambiciosa meta de sequenciar os quase 3 bilhões de pares de bases, contidos nos 46 cromossomos nucleares das células.

Figura 1 - A química Rosalind Elsie Franklin.



Fonte: (MRC LABORATORY OF MOLECULAR BIOLOGY, 1955).

[1] *The Human Genome Project* (HGP).

O término do sequenciamento do genoma humano se deu em 2003 (PEARSON, 2003) e foi um marco, tendo encerrado um dos mais bem sucedidos projetos científicos da história (LANDER et al., 2001; VENTER et al., 2001). Dentro dos avanços da instrumentalidade do método decorrentes do projeto, três anos mais tarde, começaram a surgir as tecnologias de sequenciamento de DNA de segunda geração, também chamadas de Sequenciamento de Nova Geração[2] (NGS), diminuindo em cerca de 50 mil vezes os custos operacionais no sequenciamento de um genoma humano (GOODWIN; MCPHERSON; MCCOMBIE, 2016). Em consequência disso, sequenciar genomas individualizados têm se tornado uma prática laboratorial cada vez menos custosa e vagarosa, podendo ser realizada em poucos dias (LEVY; MYERS, 2016).

Concomitantemente com essa evolução técnica, ocorreu também a expansão dos projetos de sequenciamento em larga escala. Se nos anos 90 havia o Projeto Genoma Humano, ainda nos anos 2000 foi estabelecido, por exemplo, o Projeto 1.000 genomas[3] (1KGP) (1000 GENOMES PROJECT CONSORTIUM et al., 2010), o qual já sequenciou o genoma completo de mais de 2,5 mil indivíduos de 26 populações humanas (SUDMANT et al., 2015). Há também consórcios internacionais como o Atlas do Genoma do Câncer[4] (TCGA) (CANCER GENOME ATLAS RESEARCH NETWORK et al., 2013), e o Consórcio Internacional do Genoma do Câncer[5] (ICGC) (ZHANG et al., 2019), os quais estão sequenciando milhares de genomas de diferentes tumores. Boa parte desses dados está disponível e organizado para ser baixado no Centro Estadunidense de Informações sobre Biotecnologia[6] (NCBI), o qual também possui sequências de genomas inteiros de mais de mil espécies e linhagens celulares, contemplando todos os três principais domínios da vida (bactérias, archaea e eucariotos), assim como muitos vírus, fagos, viróides, plasmídeos e organelas (ENTREZ HELP [INTERNET]. BETHESDA (MD), 2016).

Além de todo desenvolvimento científico e metodológico, o HGP deixou um outro legado extremamente importante: a cultura do acesso livre aos dados de sequenciamento produzidos em qualquer artigo científico. Por exemplo, a sequência do genoma humano está disponível publicamente desde o início do projeto, há mais

[2] *Next Generation Sequencing* (NGS).
[3] *1000 Genomes Project* (1KGP).
[4] *The Cancer Genome Atlas* (TCGA).
[5] *The International Cancer Genome Consortium* (ICGC).
[6] *National Center for Biotechnology Information* (NCBI).

de 30 anos. O 1KGP seguiu a mesma linha e também disponibiliza publicamente todas as sequências geradas. A consequência disso é que, graças a esses projetos, há uma enorme quantidade de informação, rica e totalmente aberta para ser explorada pela comunidade científica. As únicas exigências para isso são:

a) possuir métodos e ferramentas apropriados;
b) identificar e focar no estudo de fenômenos e características biológicas relevantes, ainda não ou pouco explorados;
c) ter uma estrutura de laboratório adequada para lidar com essa miríade de dados (CANNATA; MERELLI; ALTMAN, 2005).

1.2 A DESCOBERTA DOS ELEMENTOS TRANSPONÍVEIS

No final da década de 1940, ainda antes da icônica foto 51 tirada por difração de raio X pelo orientando de Rosalind Franklin, Raymond Gosling (WITKOWSKI, 2019), a citogeneticista Barbara McClintock (1902-1992, Figura 2) desafiou o dogma da imutabilidade dos cromossomos, formulada pelo geneticista Thomas Hunt Morgan (1866-1945) (FEDOROFF, 2012). No artigo de 1910, publicado na revista *Science*, Thomas Hunt Morgan propôs, por intermédio de seu trabalho com cruzamento de moscas-das-frutas (*Drosophila melanogaster*) de olhos brancos e vermelhos, que os genes seriam as unidades fundamentais de estrutura e função e estariam dispostos de forma ordenada e estática nos cromossomos, tal qual contas num colar (MORGAN, 1910). Foi então que Barbara McClintock, pesquisando a quebra e a fusão nos cromossomos do milho (*Zea Mays* ssp.), identificou sob microscopia óptica uma quebra que sempre ocorria no mesmo lócus, localizado no braço curto do cromossomo 9 do milho e o nomeou de Lócus de Dissociação (Ds). Futuras observações feitas pela cientista acerca desse peculiar fenômeno mostraram que o Ds podia se movimentar aleatoriamente dentro do mesmo cromossomo. Experimentos adicionais mostraram que a quebra no Lócus de Dissociação dependia de um segundo lócus, o qual também era capaz de promover a própria mobilização pelo cromossomo, nomeado de Lócus Ativador (Ac). Barbara McClintock notou que os elementos móveis Ds e Ac podiam outrossim interferir na

função de genes presentes nos *loci*, os quais recebiam as inserções deles, gerando mutações instáveis, e que o deslocamento dos elementos desses *loci* mutados podia restaurar a função gênica (MCCLINTOCK, 1950).

Inicialmente os achados de Barbara McClintock não foram bem recebidos pela comunidade científica. Nas palavras de Barbara (MCGRAYNE, 2001, p. 144–174, tradução nossa): “Fiquei assustada quando descobri que eles não entendiam isso; não levavam a sério” (apud RAVINDRAN, 2012). A ideia de *loci* a se movimentar pelo cromossomo, genes que eram reversivelmente inativados, o cromossomo não como uma entidade estática, mas dinâmica; todos esses conceitos não se enquadraram nas bases genéticas da época.

Figura 2 - A citogeneticista Barbara McClintock.

Fonte: (SMITHSONIAN INSTITUTION - RESTORED BY ADAM CUERDEN, 1947).

Conforme a ciência se desenrolou desde o estabelecimento do modelo de Watson e Crick para a molécula de DNA em 1953; avanços metodológicos na área da biologia molecular permitiram compreender os passos que levam da transcrição do DNA a RNA mensageiro (mRNA), e da tradução do RNA mensageiro a sequências de aminoácidos que constituem as proteínas (CRICK, 1970). Os genes

deixaram de ser conceitos abstratos, para se tornarem indivíduos moleculares concretos, passíveis de serem manipulados em laboratório. A tecnologia do DNA recombinante possibilitou que nas décadas seguintes elementos móveis fossem também descobertos em bacteriófagos (TAYLOR, 1963), bactérias (SHAPIRO, 1969), assim como *Drosophila* (ENGELS; PRESTON, 1981). Estava, por conseguinte, evidente para os cientistas que elementos móveis não eram mais algo intrínseco apenas ao milho. Em 1983, finalmente a citogeneticista Barbara McClintock foi reconhecida e premiada com o Nobel de Fisiologia, quase 40 anos depois da descoberta dos elementos móveis (FEDOROFF, 1994).

Hoje se sabe que os elementos móveis, comumente chamados de elementos transponíveis, de transposição, ou simplesmente transposons, estão presentes praticamente em todas as espécies (ABDEL-HALEEM, 2007). Em boa parte dos mamíferos, compreendem mais da metade do genoma (BURNS, 2017). Especificamente, aproximadamente dois terços do genoma humano são produto das atividades direta e indireta dos elementos transponíveis (KONING et al., 2011).

1.3 A CLASSIFICAÇÃO DOS TRANSPOSONS

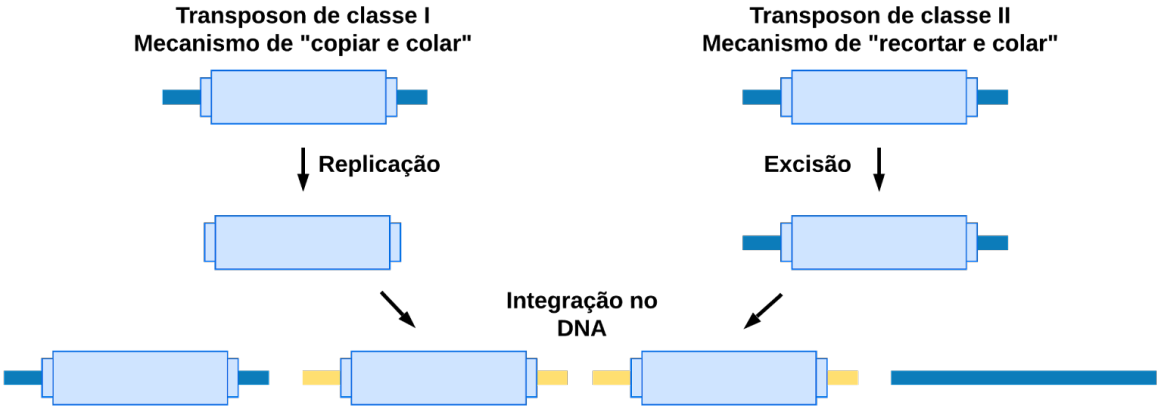
Os elementos transponíveis podem ser classificado de acordo com a molécula intermediária utilizada no processo de mobilização (Figura 3):

a) aqueles que usam uma molécula de RNA como intermediário para mobilização são chamados de transposons de classe I;
b) aqueles que se mobilizam na forma de DNA são chamados de transposons de classe II (WICKER et al., 2007).

Interessante citar que os elementos Ds e Ac que foram observados e descritos pela Barbara McClintock pertencem à classe II. Os transposons também são classificados segundo a presença, ou ausência, de genes que codifiquem para a sua própria mobilização, sendo chamados de autônomos, no caso de codificarem para os genes necessários à transposição, ou não autônomos, no caso de não

possuírem maquinaria enzimática própria e, por isso, dependerem da maquinaria enzimática de um elemento autônomo (WICKER et al., 2007).

Figura 3 - Mecanismo de transposição para os transposons de classes I e II.



Fonte: autoria própria.
Transposons de classe I usam uma molécula intermediária de RNA para se copiarem para outra posição genômica. Transposons de classe II se mobilizam, saltam, de seu lócus atual para outra posição genômica.

1.3.1 A classificação dos transposons de classe I

Os transposons de classe I, ou igualmente nomeados de retrotransposons, por sua vez, podem ser subclassificados com base na existência de regiões de sequências repetitivas de nucleotídeos flanqueando o elemento móvel, que são conhecidas por regiões de Repetição Terminal Longa[7] (LTR) (CORDAUX; BATZER, 2009). Ou seja:

a) há os elementos LTR que possuem LTR flanqueantes, como os retrovírus - o Vírus da Imunodeficiência Humana[8] (HIV), por exemplo (FOLEY et al., 2018);
b) e há os elementos não LTR que não possuem as Repetições Terminais Longas, como no caso dos autônomos Elementos Nucleares Intercalados Longos[9] (LINE) e os não autônomos Elementos Nucleares Intercalados Curtos[10] (SINE) (FESCHOTTE, 2012).

[7] *Long Terminal Repeats* (LTR).
[8] *Human Immunodeficiency Virus* (HIV).
[9] *Long Interspersed Nuclear Element* (LINE).
[10] *Short Interspersed Nuclear Element* (SINE).

1.4 LINE-1

Os LINEs são transposons de classe I, não LTR e autônomos. Foram descritos pela primeira vez em 1980 (ADAMS et al., 1980) e, desde então, têm sido ostensivamente descobertos em eucariotos (JURKA, 1998; SINGER, 1982). A família LINE-1, ou somente L1, é uma subclasse de LINE que está amplamente distribuída nos genomas de mamíferos (BECK et al., 2010). Do genoma humano, o L1 corresponde a aproximadamente 17%, o que totaliza algo em torno de 516 mil elementos (LANDER et al., 2001). A maioria desses retrotransposons perdeu a sua capacidade de mobilização e são considerados como fósseis moleculares (BECK et al., 2010), não obstante, alguns continuam ativos. Sumarizando: há mais ou menos 7 mil L1 com regiões promotoras intactas, sendo 5 mil com a sequência de nucleotídeos completa e 20 aptos a promover a retrotransposição no genoma (BECK et al., 2010; BROUHA et al., 2003; KHAN; SMIT; BOISSINOT, 2006). Observa-se que os L1 competentes em humanos apresentam uma assinatura de três nucleotídeos, “ACA”, na Região Não Traduzida[11] (UTR) a jusante, 3’UTR, o que os classifica dentro da subfamília mais jovem de L1: a subfamília L1Hs (BROUHA et al., 2003; KHAN; SMIT; BOISSINOT, 2006).

1.4.1 A estrutura do gene de LINE-1

O L1 em sua forma intacta possui cerca de 6 mil nucleotídeos (Figura 4) que compreendem:

a) uma região não traduzida a montante, 5’UTR, que tem uma atividade de promotor bidirecional, com um promotor senso para a enzima RNA polimerase II, que produz o mRNA de L1, e também um promotor antissenso, que produz RNAs quiméricos do gene de L1 (CRISCIONE et al., 2016; ISHIGURO et al., 2018; SPEEK, 2001);

[11] *Untranslated Region* (UTR).

b) duas Fases de Leitura Aberta[12] (ORF), a ORF1, que codifica para a proteína ORF1p, e a ORF2, que codifica para a proteína ORF2p;
c) uma região não traduzida a jusante, 3'UTR;
d) uma cauda poli-A integrada à região 3'UTR;
e) uma nova ORF no sentido antissenso, a ORF0, encontrada exclusivamente em primatas, que codifica para uma proteína de 70 aminoácidos e com função ainda desconhecida (DENLI et al., 2015; HANCKS; KAZAZIAN, 2016).

Figura 4 - Estrutura genética do LINE-1.

Fonte: autoria própria.
Um promotor bidirecional na região 5'UTR com atividade senso para a enzima RNA polimerase II. A ORF1 que codifica para a proteína ORF1p, uma proteína ligante de RNA, com três domínios, CC, RRM e CTD. A ORF2 que codifica para a proteína ORF2p, uma transcriptase reversa (RT) com atividade de endonuclease (EN). Uma cauda poli-A integrada à região 3'UTR.

1.4.2 A proteína ORF1p

A ORF1p é uma proteína ligante de RNA e que também possui atividade de chaperona de ácidos nucleicos (MARTIN, 2010). Tanto *in vivo* (HOHJOH; SINGER, 1996; MARTIN, 1991), quanto *in vitro* (BASAME et al., 2006), observa-se a formação de um homotrímero de ORF1p, que origina um complexo ligante de RNA. Três domínios distintos foram caracterizados na proteína; da extremidade N-terminal à extremidade C-terminal, são eles:

a) um domínio Super-hélice[13] (CC), que medeia a homotrimerização (MARTIN et al., 2003; MARTIN; LI; WEISZ, 2000);
b) um domínio para o Motivo de Reconhecimento de RNA[14] (RRM) (KHAZINA; WEICHENRIEDER, 2009);

[12] *Open Reading Frame* (ORF).
[13] *Coiled-coil* (CC).
[14] *RNA Recognition Motif* (RRM).

c) um domínio C-Terminal[15] (CTD) (JANUSZYK et al., 2007). Os domínios RRM e CTD são responsáveis pela ligação da proteína à molécula de RNA (JANUSZYK et al., 2007; MARTIN; LI; WEISZ, 2000).

1.4.3 A proteína ORF2p

A ORF2p, por sua vez, é uma proteína com atividade de endonuclease e transcriptase reversa. Ela tem dois domínios que, pelo sentido da extremidade N à C-terminal, são:

a) um domínio endonucleolítico com atividade de Endonuclease Apurínica-apimidínica[16] (APE) (SULTANA et al., 2017), que diretamente contribui com o reconhecimento do sítio genômico alvo, tipicamente a clivagem ocorre no motivo “TTTT/AA” (JURKA, 1997);
b) um domínio com atividade de transcriptase reversa, responsável por converter uma sequência de RNA noutra de DNA complementar (cDNA) (COST, 2002).

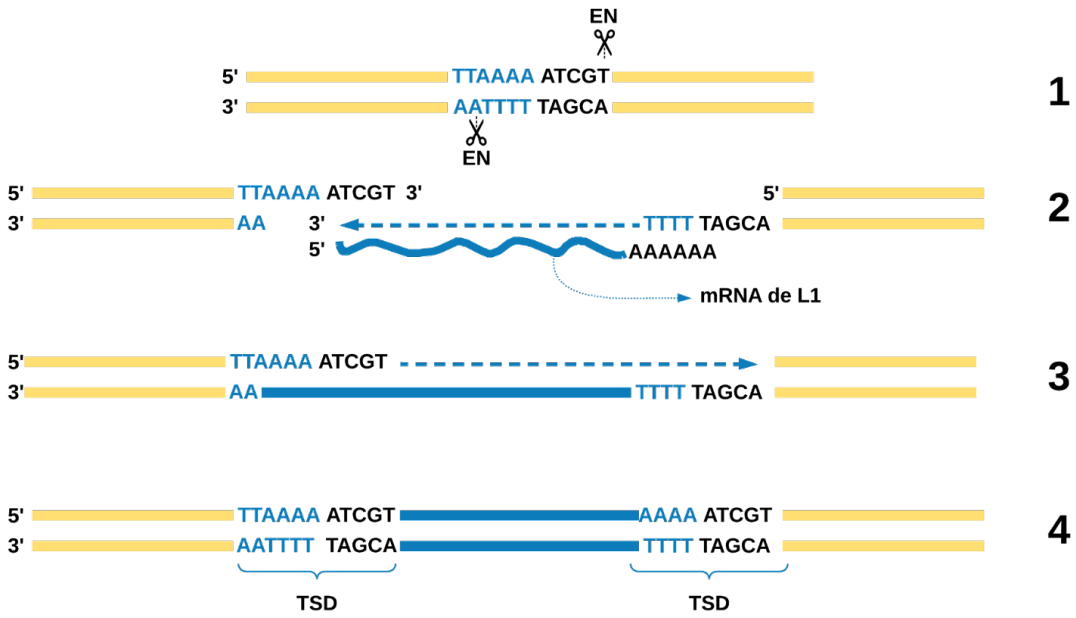
1.4.4 O processo de mobilização do elemento L1

A mobilização de L1 pelo genoma se dá, como todos os retrotransposons, através de um intermediário de RNA. Cada ciclo de transposição começa com a transcrição de uma cópia ativa de L1 em um RNA mensageiro bicistrônico. O mRNA de L1 migra para o citoplasma, onde irá servir de molde para a tradução de ambas as ORFs nas proteínas ORF1p e ORF2p. Essas proteínas se associam em *cis* ao mRNA de L1, formando uma partícula ribonucleoproteica (WEI et al., 2001). A partícula retorna ao núcleo, onde o domínio endonucleotídico APE da ORF2p reconhece o motivo “TTTT/AA” (JURKA, 1997) e o cliva, liberando uma ponta 3’OH.

[15] *C-Terminal Domain* (CTD).
[16] *Apurinic-apyrimidinic Endonuclease* (APE).

Essa ponta 3'OH liberada na extremidade da sequência poli-T se anela à cauda poli-A do mRNA de L1 e é estendida pela transcriptase reversa presente no segundo domínio da proteína ORF2p, usando o próprio mRNA de L1 como molde (FENG et al., 1996). Para essa reação é dado o nome de Transcrição Reversa com o Alvo como Iniciador (Figura 5), ou Transcrição Reversa com o Alvo como *Primer*[17] (TPRT). A integração do transposon é finalizada por uma reação análoga à TPRT, agora usando como molde a sequência de cDNA de L1 recém sintetizada (MARTIN, 2010; MARTIN et al., 2005). Também é possível que ocorra a retrotransposição de L1 por uma via menos comum e independente da endonuclease. Nessa via, lesões pré existentes na cadeia de DNA são usadas pela transcriptase reversa como iniciadores da reação de polimerização (VIOLLET; MONOT; CRISTOFARI, 2014).

Figura 5 - Reação de TPRT para L1.



Fonte: autoria própria.
1) o domínio de endonuclease da ORF2p reconhece o motivo "TTTT/AA" e o cliva, liberando uma ponta 3'OH; 2) ocorre o pareamento da cauda poli-A do mRNA de L1 com a sequência de poli-T do DNA, então a extremidade 3'OH livre é usada como iniciador pelo domínio transcriptase reversa da ORF2p. O próprio mRNA de L1 serve como molde para a reação de polimerização; 3) uma reação análoga à TPRT finaliza a integração de L1, usando a sequência de cDNA sintetizada como molde; 4) uma nova cópia de L1 está pronta. Vê-se a presença das regiões de TSD flanqueando o transposon.

Nota-se que a maioria das novas inserções de L1 são truncadas na região 5' (SZAK et al., 2002). A processividade média da transcriptase reversa da ORF2p, em

[17] *Target-primed Reverse Transcription* (TPRT).

humanos, está em aproximadamente 900 pares de bases (NAVARRO; GALANTE, 2015; SZAK et al., 2002), o que faz com que as novas inserções não tenham a região 5'UTR, possuidora do promotor interno para RNA polimerase II. Por conseguinte, assume-se *a priori* a não funcionalidade desses eventos (BECK et al., 2011).

1.5 ELEMENTOS TRANSPONÍVEIS NÃO AUTÔNOMOS E A RETROTRANSPOSIÇÃO EM *TRANS*

Não obstante a maquinaria enzimática de L1 se ligar preferencialmente em *cis* ao prórpio RNA mensagiero (WEI et al., 2001), ela também é capaz de promover a mobilização de outros elementos, ou seja em *trans*. Dentre esses elementos, podem ser citados os transposons não autônomos SINEs, assim como RNAs mensageiros de outros genes codificadores.

1.5.1 SINEs

Os SINEs representam 13,5% de todo genoma humano (MANDAL; KAZAZIAN, 2008), sendo *Alu* o elemento mais frequente dessa classe de transposons. Sozinho, o *Alu* ocupa cerca de 11% do genoma humano e é encontrado especificamente em primatas (DEININGER, 2011). Esses retrotransposons são desprovidos de transposase própria para a sua mobilização (não autônomos) e, por isso, parasitam a maquinaria enzimática de L1 de modo a manter o ciclo reprodutivo. Por exemplo, o elemento *Alu* tem por volta de 300 nucleotídeos de comprimento (DEININGER, 2011), distribuídos dentro do seguinte modelo:

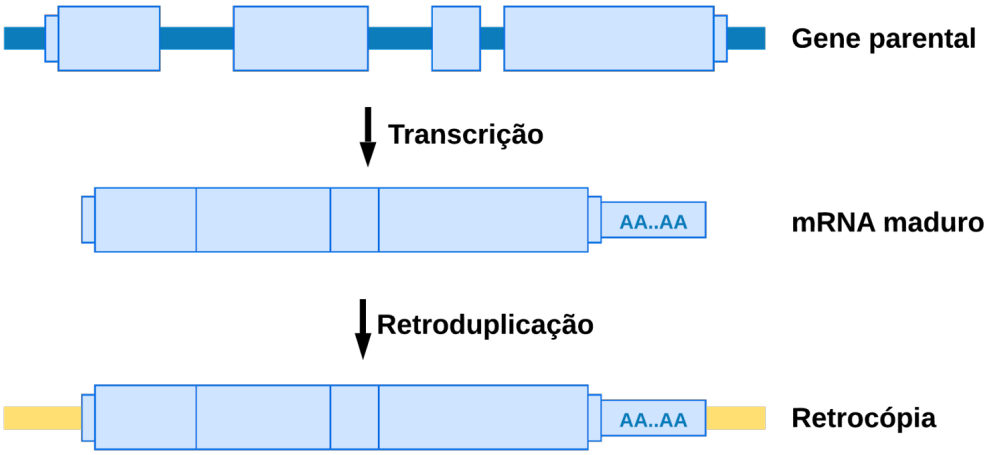
a) uma região 5', na qual há marcas internas para o reconhecimento da enzima RNA polimerase III;

b) mais a jusante, há duas regiões de sequências similares, as quais são conhecidas como partes A e B, ou braços esquerdo e direito, separadas por uma cadeia rica em adenina;
c) por fim, na região 3', uma cauda poli-A que vai atuar como alvo para a transcrição reversa (AGUILAR, 2019, p. 135–139; BAKSHI et al., 2016).

No citoplasma, os RNAs de *Alu* conseguem cooptar as proteínas ORF1p e ORF2p de L1, que irão promover a retrotransposição em *trans* desses elemento, seguindo a mesma estratégia de inserção por TPRT, que é usada para os próprios RNAs mensageiros de L1 (BENNETT et al., 2008).

1.5.2 Pseudogenes processados

Figura 6 - Retrotransposição em *trans* de um gene codificador.



Fonte: autoria própria.
O mRNA maduro do gene é mobilizado pelas enzimas de L1, de modo que apenas as regiões exônicas do gene são copiadas. Na retrocópia, há a presença de cauda poli-A.

Assim como os retrotransposons não autônomos se mobilizam por intermédio da transposase de L1, pode acontecer de RNAs mensageiros de genes codificadores serem capturados pelas ORF1p e ORF2p e, dessa forma, também serem mobilizados pelo genoma (KAESSMANN; VINCKENBOSCH; LONG, 2009). Esse processo gera duplicações gênicas (Figura 6), nas quais apenas os éxons presentes no mRNA maduro são copiados para outra região cromossômica (VANIN,

1985). As duplicações gênicas mediadas pelas enzimas de L1 são chamadas de pseudogenes processados, ou retrocópias, enquanto que os genes que foram retrocopiados são conhecidos como genes parentais.

1.6 RETROCÓPIAS DE GENES CODIFICADORES E O SEU PAPEL

As retrocópias são o resultado da duplicação gênica promovida pela maquinaria enzimática de L1. Foi na década de 80 que os primeiros grupos científicos começaram a apontar para a sua existência. Em 1982, a sequência completa de nucleotídeos do pseudogene processado da alfa globina humana foi publicada na revista *Cell* por Proudfoot e Maniatis (PROUDFOOT; GIL; MANIATIS, 1982), no entanto, os mecanismos pelos quais se deu a duplicação gênica ainda eram desconhecidos. Trabalhos subsequentes de outros grupos de pesquisa continuaram a apresentar à comunidade mais casos de pseudogenes processados e, em 1988, Begg e colaboradores elucidaram a conexão entre retrocópias e L1 (BEGG; DELIUS; LEADER, 1988).

Atualmente, sabe-se que as retrocópias são frequentes em diversas espécies (BAERTSCH et al., 2008; NAVARRO; GALANTE, 2013). No genoma humano, a estimativa do número de retrocópias varia de acordo com o tipo de metodologia empregada:

a) metodologias baseadas no alinhamento de sequências de RNAs mensageiros identificaram entre 7 mil e 13 mil retrocópias fixadas (BAERTSCH et al., 2008; NAVARRO; GALANTE, 2013, 2015; PEI et al., 2012; SAKAI et al., 2007);
b) enquanto que aquelas baseadas no alinhamento de sequências de proteínas postularam entre 3 mil e 6 mil dessas retrocópias (MARQUES et al., 2005; VINCKENBOSCH; DUPANLOUP; KAESSMANN, 2006).

Segundo os métodos desenvolvidos e publicados por Pedro A. F. Galante e colaboradores do laboratório de bioinformática do Hospital Sìrio-Libanês em São Paulo, há aproximadamente 8 mil retrocópias fixadas no genoma humano (NAVARRO; GALANTE, 2013, 2015).

Inicialmente, acreditava-se que as retrocópias eram “natimortas”, pois, devido a algumas de suas características, tais como ausência de regiões reguladoras (KAESSMANN; VINCKENBOSCH; LONG, 2009), acúmulo de mutações e sequências truncadas a montante (à semelhança das duplicatas de L1) (PISKAREVA; SCHMATCHENKO, 2006), eram consideradas não funcionais. De fato, há um considerável montante de retrocópias que não apresentam funcionalidade (ZHANG; CARRIERO; GERSTEIN, 2004), mas também há aquelas que se inserem próximas a regiões promotoras, ou em íntrons de genes codificadores (NAVARRO; GALANTE, 2015), e, em vista disso, acabam sendo transcritas e potencialmente funcionais (CARELLI et al., 2016).

1.6.1 Exemplo de retrocópia funcional: GLUD2

Graças a trabalhos pontuais, são conhecidas algumas retrocópias funcionais. Burki e Kaessmann (2004) demonstraram que o gene GLUD2, específico de alguns primatas (incluindo humanos), se originou por retrotransposição do gene GLUD1. Ambos os genes GLUD1 e GLUD2 codificam para a enzima Glutamato Desidrogenase (GDH), a qual atua nas mitocôndrias das células eucarióticas, conduzindo a reação anaplerótica de desaminação oxidativa reversível do glutamato em α-cetoglutarato. O GLUD1 é um gene localizado no cromossomo autossômico humano 10 e possui 13 éxons, sendo amplamente expresso em diversos tecidos, ao passo que o gene GLUD2 está localizado no cromossomo sexual humano X e possui apenas 1 éxon, sendo expresso de forma específica nos neurônios. O fato de o gene GLUD2 não possuir íntrons sugere que este seja uma retrocópia funcional de GLUD1. Como no cérebro a enzima GDH catalisa a reciclagem do glutamato, que é o principal neurotransmissor excitatório (MASTORODEMOS et al., 2009; MCKENNA; FERREIRA, 2016), os autores sugerem que a participação de GLUD2 no metabolismo neuronal do glutamato parece ter sido responsável por uma melhora na função cognitiva de hominídeos.

1.6.2 Exemplo de retrocópia funcional: PTENP1

Um outro trabalho, agora envolvendo um exemplo dentro da patologia, é o papel da retrocópia PTENP1 na tumorigênese. O gene que codifica para a Fosfatase Homóloga à Tensina[18] (PTEN), é um gene supressor tumoral que possui uma retrocópia conhecida, a PTENP1. Essa retrocópia não codifica para proteína, mas possui uma região 3'UTR cuja similaridade com a do seu gene parental PTEN permite que seja alvo dos mesmos microRNAs (miRNAs). A ligação do miRNA ao RNA mensageiro inibe a tradução, ou promove processos degradativos (HUNTZINGER; IZAURRALDE, 2011). Assim, a retrocópia PTENP1 atua como um RNA endógeno concorrente[19] (ceRNA), ou seja, em condições fisiológicas, a expressão do gene PTEN aumenta quando há a co-expressão da retrocópia PTENP1. Poliseno e colaboradores (2010) observaram que, em alguns tumores, a retrocópia PTENP1 está silenciada, o que faz com que todos os miRNAs sejam direcionados ao gene parental PTEN. Como resultado, a expressão do gene PTEN é reduzida, gerando um cenário favorável à tumorigênese.

1.7 RETROCÓPIAS FIXAS E POLIMÓRFICAS NO GENOMA HUMANO

Até aqui foram apresentados exemplos pontuais de 2 retrocópias dentre as quase 8 mil presentes no genoma humano. Todos esses eventos de retrotransposição detectados e anotados no genoma são classificados como eventos "fixados", ao passo que aqueles que estão ausentes recebem o termo "não fixados". As retrocópias não fixadas são nomeadas de Variantes de Número de Cópias do gene causadas por retrotransposição[20] (retroCNVs) (SCHRIDER et al., 2013). Dos eventos não fixados, muitos podem se enquadrar como variações raras (1 em 1.000 indivíduos) ou polimórficas (1 em 100 indivíduos) e serem de origem germinativa, ou somática.

[18] *Phosphatase and Tensin homolog* (PTEN).
[19] *Competing endogenous RNA* (ceRNA).
[20] *Gene Copy-Number Variants caused by retrotransposition* (retroCNVs).

Constata-se que ao passo que as retrocópias fixadas têm sido continuamente investigadas (KABZA; CIOMBOROWSKA; MAKAŁOWSKA, 2014; NAVARRO; GALANTE, 2013), as suas contrapartes, não fixadas, encontram-se sub-exploradas. Essa negligência para com as retroCNVs não ocorre, por exemplo, com os eventos não fixados de LINEs e SINEs, os quais têm recebido a atenção da comunidade científica, não somente com relação a estudos evolutivos (KAZAZIAN, 2004), mas também com relação a estudos patológicos (BURNS, 2017; HANCKS; KAZAZIAN, 2016). Um trabalho interessante a citar na área da patologia foi publicado na revista *Science* por Lee e colaboradores (2012). Nele, os pesquisadores descreveram a ocorrência de eventos de retrotransposição somática de L1 em genomas tumorais. É sabido que o desenvolvimento e a progressão de algumas doenças, como o câncer, estão intimamente associados à alterações genômicas, tais como mutações pontuais e variações estruturais. Averiguou-se, pois, que tumores de cólon apresentam um número elevado de eventos não fixados de L1, os quais causam alterações genômicas estruturais no local onde são inseridos, podendo alterar a expressão de vários genes de maneira direta ou indireta. Contudo, nunca se estudou sistematicamente a retrotransposição somática de outros RNAs mensageiros em tumores, o que parece ser algo lógico de acontecer.

1.8 ALGORITMOS PARA IDENTIFICAÇÃO DE RETROCÓPIAS NÃO FIXADAS

Um dos motivos pelos quais o estudo sistemático de retrocópias não fixadas não tem recebido a devida atenção por parte da comunidade científica é devido à falta de algoritmos dedicados à identificação e anotação de retroCNVs em dados de NGS. Diferentes pesquisadores já despenderam esforços na análise de retroCNVs na população humana (ABYZOV et al., 2013; EWING et al., 2013; SCHRIDER et al., 2013; ZHANG et al., 2017). No entanto, esses trabalhos não tiveram como objetivo a descrição de uma metodologia computacional para detecção de eventos não fixados no genoma, tampouco a publicação de um programa de bioinformática, o qual automatizasse as etapas analíticas necessárias.

1.9 SIDERETRO

Nós acreditamos na importância de se estudar a presença de retroCNVs e no seu papel funcional, tanto do ponto de vista evolutivo, quanto do ponto de vista patológico, com destaque para os eventos relacionados a câncer. Tendo isso em mente, desafiamo-nos a desenvolver neste projeto uma ferramenta de bioinformática voltada a facilitar, estimular e acelerar as pesquisas de retrocópias não fixadas. Então surgiu o sideRETRO, um algoritmo dedicado a identificação de inserções de retrocópias somáticas e polimórficas, que provê informações-chave para o entendimento do fenômeno, tais como:

a) o gene parental;
b) a posição, coordenada, genômica na qual se inseriu a retroCNV;
c) a orientação da inserção do evento – se na fita líder (+), ou na fita retardada (-);
d) o contexto genômico da inserção – se a retrotransposição ocorreu numa região intergênica ou intragênica;
e) a genotipagem – quando analisando múltiplos indivíduos, em quais deles foi identificada a retroCNV;
f) haplotipagem – se o evento é homozigótico alternativo ou heterozigótico.

Por fim, nós humildemente desejamos que o sideRETRO possa impulsionar mais estudos de retroCNVs no contexto de pesquisa básica e translacional e que isso possa ser usado tanto para o enriquecimento do nosso conhecimento acerca dos elementos móveis, como para o bem da humanidade.

2 Objetivos

2 OBJETIVOS

2.1 OBJETIVO GERAL

Construir uma ferramenta computacional eficiente e de fácil uso para identificar e analisar eventos somáticos, germinativos e polimórficos de retroCNVs a partir de dados de sequenciamento de genomas completos e dados de exomas.

2.2 OBJETIVOS ESPECÍFICOS

a) investigar de maneira sistemática os padrões de alinhamento de leituras pareadas contra o genoma de referência e definir aqueles que indicam eventos não fixados de retroCNVs;
b) desenvolver e implementar um algoritmo computacional (nomeado de sideRETRO) para analisar dados de sequenciamento de genomas completos e exomas e, a partir dos padrões de alinhamento das sequências contra o genoma de referência, identificar possíveis eventos somáticos, germinativos e polimórficos de retroCNVs;
c) atribuir informações adicionais, tais como o ponto de inserção, a orientação da fita, o contexto genômico, a genotipagem e a zigosidade, para os eventos de retroCNVs identificados;
d) definir, através de dados simulados e experimentais, a eficiência de identificação de retroCNVs pelo sideRETRO;
e) tornar o sideRETRO uma ferramenta computacional de fácil instalação e execução, eficiente do ponto de vista computacional e capaz de produzir resultados completos e bem organizados;
f) criar uma documentação organizada, detalhada e acessível para descrever e ilustrar o uso da ferramenta sideRETRO.

3 Metodologia

3 METODOLOGIA

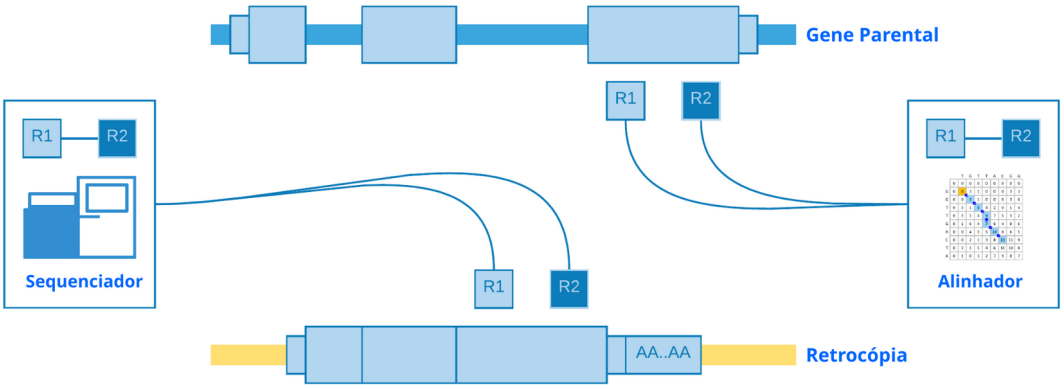
3.1 RESOLVENDO AS DIFICULDADES INTRÍNSECAS AO ALINHAMENTO DAS SEQUÊNCIAS CONTRA O GENOMA

Para realizar o desenvolvimento desta ferramenta de detecção de eventos polimórficos e somáticos de inserção de retroCNVs no genoma, foi seguida uma metodologia composta por diversos passos. O ponto de partida para a análise de tais eventos é o alinhamento das sequências de interesse contra o genoma de referência. No entanto, para identificar as retroCNVs, existem algumas dificuldades intrínsecas ao alinhamento e também às metodologias de sequenciamento (de segunda geração), sobretudo oriundas da alta similaridade entre as sequências genômicas de uma retrocópia e de seu gene parental. Portanto, para realizar esse objetivo, primeiramente foi investigado minuciosamente os diferentes padrões encontrados quando se alinha a sequência de interesse contra o genoma de referência.

3.1.1 Os alinhamentos indistinguíveis

Quando há um evento de inserção de retroCNV em um indivíduo e seu genoma é sequenciado com uma tecnologia de NGS (isto é, com sequências/leituras curtas), pode-se esperar que o alinhador (algoritmo de alinhamento) "se confunda" quanto à origem de certas leituras, gerando um alinhamento incorreto, principalmente em éxons do gene parental desse evento de retroduplicação. A Figura 7 faz uma ilustração resumida deste padrão e mostra que esse tipo de alinhamento é indistinguível, pois ele não dá nenhuma pista sobre a presença da retroCNV.

Figura 7 - Alinhamentos indistinguíveis.



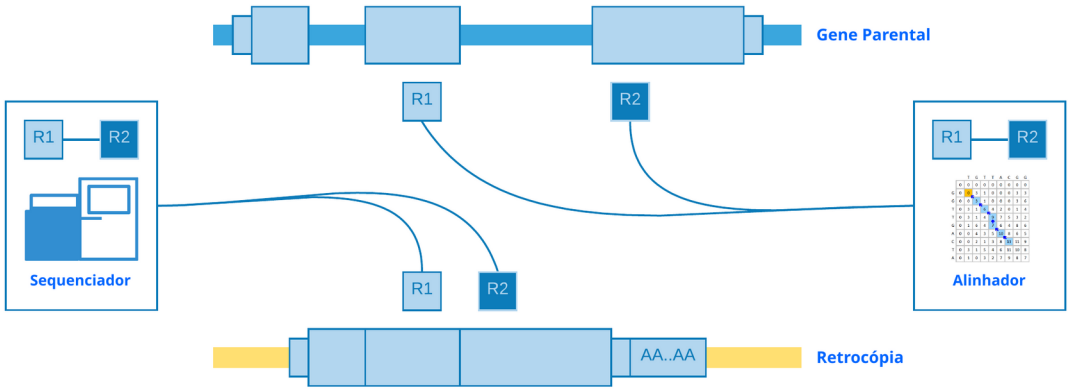
Fonte: autoria própria.
Leituras provindas de uma retroCNV podem ser erroneamente alinhadas a um éxon do gene parental. Alinhamentos como esses são indistinguíveis e não fornecem evidências da presença de um evento de retroCNV.

No entanto, existem alinhamentos que fornecem evidências de um evento oriundo de retrocópia, os alinhamentos anormais, os quais serão descritos a seguir.

3.1.2 Leituras pareadas alinhadas em diferentes éxons

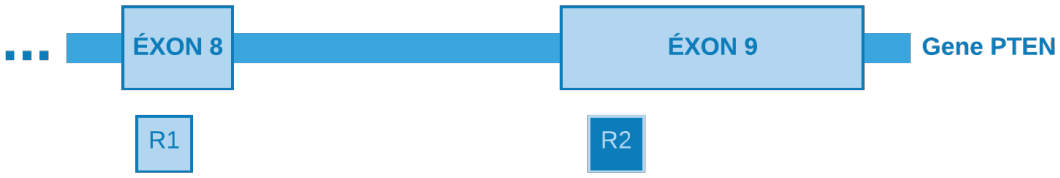
A característica predominante de uma retrocópia é a sua ausência de íntrons. Portanto, é lógico imaginar que essa característica deva refletir em certos padrões de alinhamentos. Para as metodologias de sequenciamento baseadas em pares de leituras, cada sequência é alinhada e, subsequentemente, esses elementos são pareados. Quando há um padrão de alinhamento em que se observa cada leitura em éxons contíguos do possível gene parental, há um indício de que tais sequências são de uma região ausente do genoma de referência, e, claro, desprovida de íntrons (Figuras 8 e 9), isto é, uma nova retroCNV. Vale destacar que este tipo de alinhamento ocorre em dados de sequenciamento de RNA (RNA-Seq), pelo fato do sequenciamento usar como matéria prima mRNAs maduros, desprovidos de regiões intrônicas, mas não ocorre em dados de sequenciamento de DNA genômico.

Figura 8 - Leituras pareadas alinhadas em diferentes éxons.



Fonte: autoria própria.
Representação de alinhamento de leituras pareadas e provindas de uma retrocópia. Nesta representação, supõe-se que os dados são de sequenciamento de DNA genômico e que esse padrão de leituras alinhadas em éxons contíguos (no gene parental) indique a existência de uma retroCNV na amostra que fora sequenciada.

Figura 9 - Exemplo de leituras pareadas que cruzam o íntron.



QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN
R1	99	chr10	87.960.894	30	101M	chr10	87.965.398	4605
R2	147	chr10	87.965.398	30	101M	chr10	87.960.894	-4605

* Saída reduzida de um arquivo SAM.

Fonte: autoria própria.
Alinhamento relativo ao gene PTEN. A leitura R1 está alinhada no éxon 8, ao passo que seu par R2 está alinhado no éxon 9. Ambas as leituras têm 101 bases de comprimento e estão a uma distância de 4.605 bases.

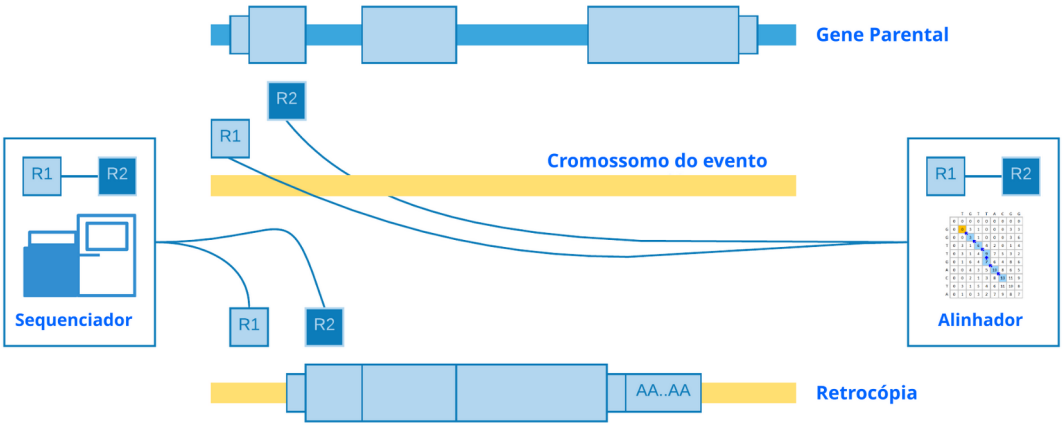
3.1.3 Leituras pareadas alinhadas em cromossomos diferentes

Uma retroCNV pode se inserir num cromossomo diferente do cromossomo de seu gene parental (SCHRIDER et al., 2013). Esse evento, ao ser sequenciado com uma tecnologia NGS e alinhado contra o genoma de referência, poderá apresentar o padrão, no qual leituras pareadas são mapeadas em cromossomos diferentes, isto é, parte das leituras se alinham na região do gene parental e parte das leituras na

região próxima ao ponto de inserção da retrocópia (em um cromossomo diferente daquele em que o gene parental se localiza, vide Figura 10). Outra característica deste alinhamento anormal é que, pelo menos, uma das leituras pareadas é alinhada num éxon do gene parental (geralmente o éxon mais a jusante).

Este padrão emerge quando as leituras pareadas advêm de uma região limítrofe da retroCNV - sendo um dos pares internamente oriundo da retrocópia e o seguinte oriundo externamente a ela (isto é, na região genômica imediatamente próxima ao ponto de inserção). Como essa retroCNV não existe no genoma de referência, o alinhador irá mapear a leitura de origem interna à retrocópia no respectivo éxon do gene parental, enquanto a leitura par de origem externa à retrocópia será mapeada noutra região genômica - neste caso, no outro cromossomo, onde se deu o evento de inserção de retroCNV (Figura 10).

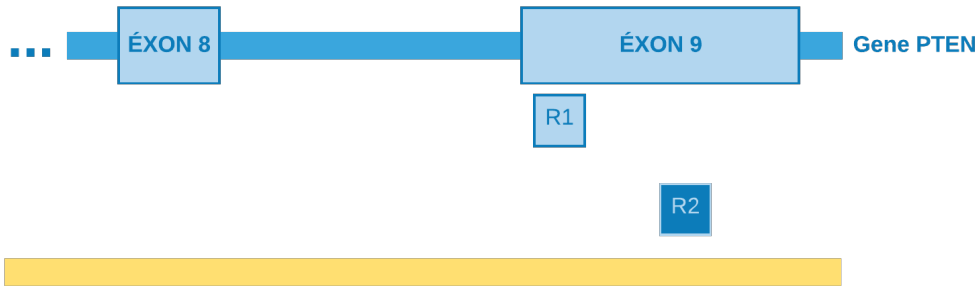
Figura 10 - Leituras pareadas alinhadas em cromossomos diferentes.



Fonte: autoria própria.
Representação esquemática de alinhamentos de leituras pareadas oriundas de uma retroCNV. Neste exemplo, na parte superior, há uma representação da leitura R2 alinhada no gene parental e a leitura R1 em um outro cromossomo do genoma de referência. Na parte inferior, é mostrado como, de fato, as leituras se alinham no genoma que contém o evento (isto é, da amostra que fora sequenciada).

Uma vantagem deste tipo de padrão de alinhamento anormal é que ele fornece evidências da posição genômica da janela (mas não o ponto exato), em que se sucedeu a integração da retroCNV no genoma investigado/sequenciado. A Figura 11 mostra um exemplo disso.

Figura 11 - Exemplo de leituras pareadas alinhadas em cromossomos diferentes.



QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN
R1	99	chr10	87.965.404	30	101M	chr1	85.186	0
R2	145	chr1	85.186	30	101M	chr10	87.965.404	0

* Saída reduzida de um arquivo SAM.

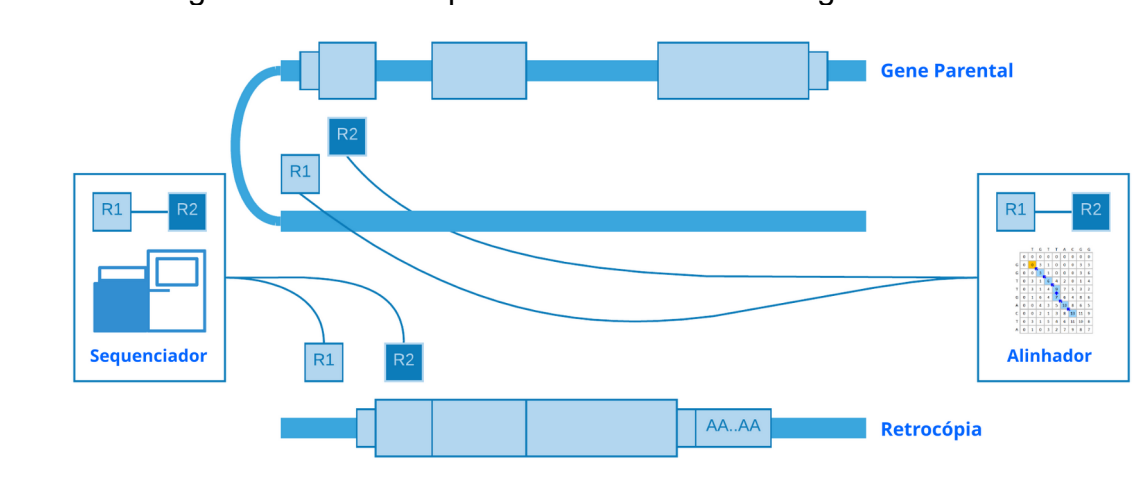
Fonte: autoria própria
Alinhamento relativo ao gene PTEN. A leitura R1 é alinhada ao éxon 9 do gene PTEN, localizado no cromossomo 10, ao passo que a leitura R2 está alinhada no cromossomo 1, indicando um possível retroCNV deste gene no cromossomo 1. Ambas as leituras têm 101 bases de comprimento.

3.1.4 Leituras pareadas alinhadas em regiões distantes

Apesar de possuir uma probabilidade menor (P = 1 / [número de cromossomos do organismo]), uma retroCNV também pode se inserir no mesmo cromossomo de seu gene parental. Este evento, ao ser sequenciado com uma tecnologia NGS e alinhado contra o genoma de referência, poderá manifestar o padrão, no qual leituras pareadas são mapeadas em regiões mais distantes do que o esperado (Figura 12). Normalmente a distância esperada para o mapeamento de leituras pareadas varia de acordo com o comprimento do fragmento de DNA de origem dessas leituras, sendo que os fragmentos de DNA são gerados (física ou mecanicamente) ao longo de processos experimentais de biologia molecular, necessários para a produção da biblioteca de sequenciamento (HESS et al., 2020), os quais precedem o sequenciamento (GOODWIN; MCPHERSON; MCCOMBIE, 2016). A distância esperada para o mapeamento de leituras pareadas varia de acordo com uma curva normal, com média em torno de 300 bases de comprimento, para sequenciamentos em plataformas Illumina, e um desvio padrão que dificilmente chega a 100 bases. Portanto, distâncias maiores que 600 bases (Figura 13) já são

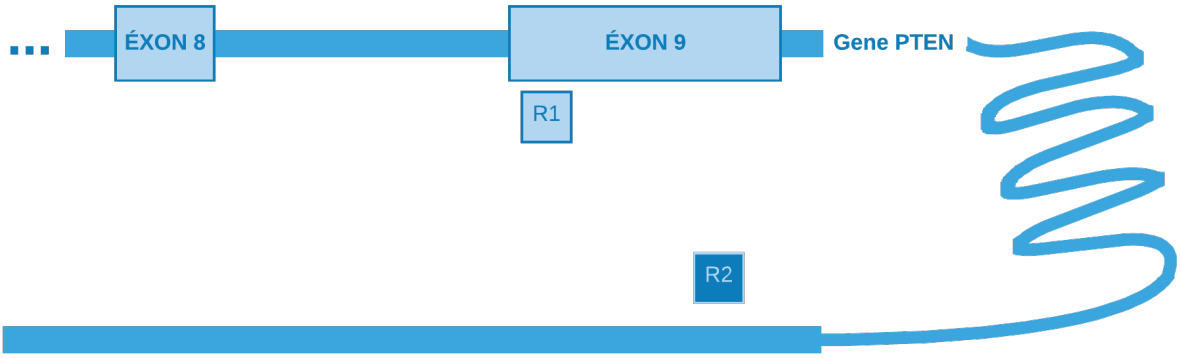
vistas como suspeitas de estarem relatando um evento de alinhamento anormal (não necessariamente por consequência de um evento de retroCNV).

Figura 12 - Leituras pareadas alinhadas em regiões distantes.



Fonte: autoria própria.
Leituras pareadas oriundas de uma retroCNV podem alinhar no mesmo cromossomo do genoma de referência em regiões distantes.

Figura 13 - Exemplo de leituras pareadas alinhadas em regiões distantes.



QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN
R1	97	chr10	87.965.404	30	101M	chr10	133.688.018	45.722.715
R2	145	chr10	133.688.018	30	101M	chr10	87.965.404	-45.722.715

* Saída reduzida de um arquivo SAM.

Fonte: autoria própria.
Alinhamento relativo ao gene PTEN. A leitura R1 é alinhada ao éxon 9 do gene PTEN e localizada na posição genômica de 87.965.404 no cromossomo 10. A leitura pareada R2 está localizada na posição 133.688.018 do mesmo cromossomo 10, à uma distância de 45.722.715 - uma valor não condizente com o comprimento de um fragmento de NGS (aproximadamente 300 bases) de leituras de 101 bases.

Para um possível evento de retroCNV, primeiro é observado que pelo menos uma das leituras, pareadas e alinhadas em regiões distantes, é mapeada dentro de um éxon (gene parental). Este padrão sucede, quando o par de leituras se origina de

uma das pontas da retroCNV, sendo uma das leituras interna à retrocópia e seu par externo a ela. Dado que tal retrocópia não existe no genoma de referência, o alinhador irá mapear a leitura interna à retroCNV no respectivo gene parental, enquanto que a leitura externa à retroCNV será mapeada noutra região genômica no mesmo cromossomo - o que neste caso implica a região onde ocorreu a retroduplicação (Figura 12).

3.1.5 Leituras divididas (leituras com alinhamento suplementar)

Leituras divididas se originam a partir de um tipo de alinhamento chamado de alinhamento quimérico. Segundo o manual de especificações do formato de Mapa de Alinhamento de Sequência[21] (SAM) e sua versão compactada Mapa de Alinhamento Binário[22] (BAM) (2021a, p. 2, tradução nossa):

> Alinhamento quimérico é um alinhamento de uma leitura que não pode ser representado como um alinhamento linear. Um alinhamento quimérico é representado como um conjunto de alinhamentos lineares que não têm grandes sobreposições. Normalmente, um dos alinhamentos lineares em um alinhamento quimérico é considerado o alinhamento "representativo", e os outros são chamados de “suplementares” e são diferenciados pela *flag* de alinhamento suplementar.

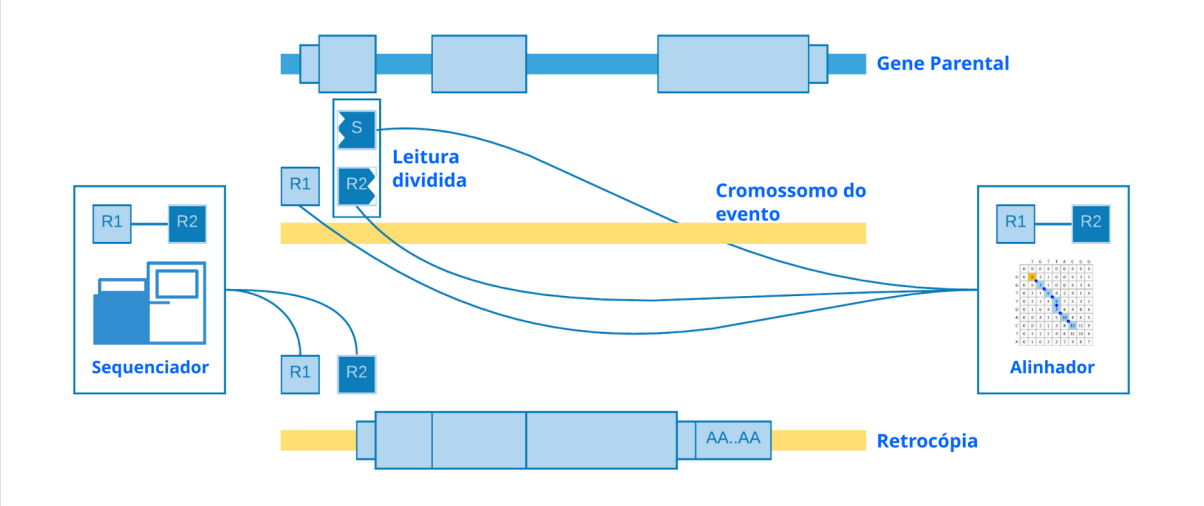
Por conseguinte, a leitura dividida é o modo como o alinhamento quimérico é anotado, sendo uma das partes da leitura o alinhamento representativo, e as demais partes os alinhamentos suplementares. Ainda segundo o manual de especificações do formato SAM/BAM, os alinhamentos quiméricos são causados principalmente por variações estruturais, fusões gênicas, montagens incorretas, RNA-Seq (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021a).

Do ponto de vista de uma retroCNV, uma leitura dividida pode ocorrer quando a leitura cobre uma das extremidades (5’ ou 3’) da retrocópia, de modo que parte da leitura advenha da região em que ocorreu o evento de retroduplicação e parte

[21] *Sequence Alignment Map* (SAM).
[22] *Binary Alignment Map* (BAM).

advenha da própria retrocópia (Figura 14). Esta leitura ao ser alinhada contra o genoma de referência gerará um alinhamento quimérico, sendo representado por leituras divididas - cada qual alinhada a sua posição de origem, estando, portanto, uma mapeada num dos éxons do gene parental e a outra mapeada na região, onde se sucedeu a inserção da retroCNV.

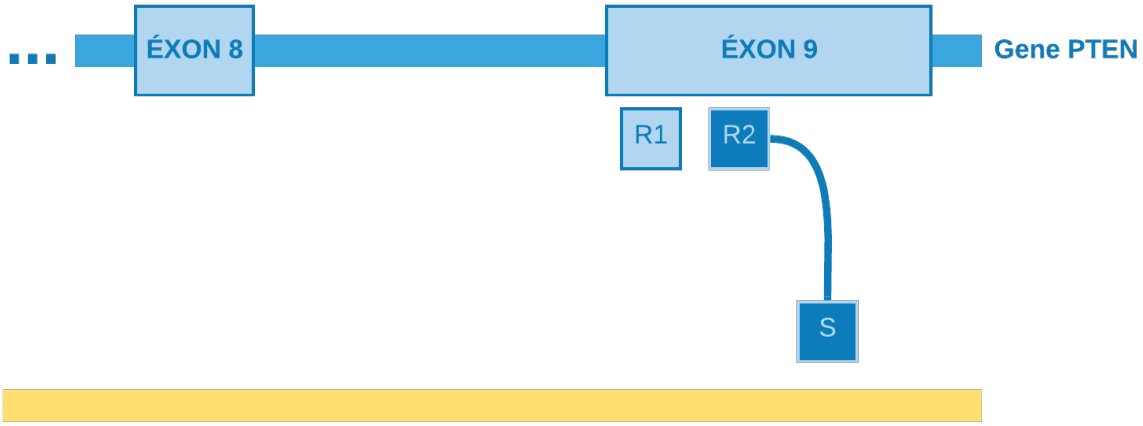
Figura 14 - Leituras divididas.



Fonte: autoria própria.
Isso ocorre quando a leitura cobre uma das extremidades da retrocópia, então o alinhador precisa gerar mais de um alinhamento para representar a mesma leitura. Um dos alinhamentos é tido como o representativo e os demais como suplementares.

O alinhamento dividido de uma leitura, dada a sua natureza quimérica, pode ser útil para a determinação do ponto de inserção da retroCNV. Ele fornece as pistas necessárias ao se observar o ponto de quebra entre o alinhamento representativo e seu respectivo alinhamento suplementar (Figura 15). Um deles estará mapeado num éxon do gene parental e o seguinte estará mapeado na provável região de retrotransposição, portanto, este segundo dará evidências da posição em que ocorreu o evento.

Figura 15 - Exemplo de leituras pareadas com alinhamento suplementar.



QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN
R1	97	chr10	87.965.404	30	101M	chr10	87.965.704	401
R2	147	chr10	87.965.704	30	55M46S	chr10	87.965.404	-401
S	2147	chr1	85.186	30	55H46M	chr10	87.965.404	0

* Saída reduzida de um arquivo SAM.

Fonte: autoria própria.
Neste exemplo, a leitura R2 tem um alinhamento suplementar S alinhado em outro cromossomo. Alinhamento relativo ao gene PTEN. A leitura R1 é alinhada ao éxon 9 do gene PTEN e localizada na posição genômica de 87.965.404 no cromossomo 10. A leitura pareada R2 está localizada na posição 87.965.704 do mesmo cromossomo 10 e possui também um alinhamento suplementar S, que está mapeado no cromossomo 1, na posição 85.186. Observando-se a coluna CIGAR, nota-se que a leitura R2 foi dividida, tendo, das 101 bases de comprimento, 55 bases mapeadas no éxon 9 do gene PTEN (como visto no alinhamento R2) e, suplementarmente, 46 bases mapeadas no cromossomo 1 (como visto no alinhamento S). A posição do alinhamento S é útil para determinar o ponto de inserção da retroCNV no cromossomo 1.

3.1.6 Resumo dos alinhamentos anormais

Eis, então, que foram vistos, em mais detalhes, os alinhamentos anormais e suas respectivas características (Quadro 1). Entender os diferentes padrões advindos de um evento de retroduplicação num indivíduo - o qual teve o genoma sequenciado por uma tecnologia de NGS e alinhado contra o genoma de referência - é essencial para a construção de uma ferramenta computacional, que possa fazer uso desses padrões e, por conseguinte, identificar as retroCNVs.

Quadro 1 - Resumo dos tipos de padrões de alinhamentos anormais.

Alinhamentos anormais	Detecção	
	Da coordenada	Do ponto de Inserção
Em éxons diferentes	NÃO	NÃO
Em cromossomos diferentes	SIM	NÃO
Em regiões distantes	SIM	NÃO
Leituras divididas	SIM	SIM

Fonte: autoria própria.

E para este fim, foi desenvolvida a ferramenta sideRETRO, que faz uso dos alinhamentos anormais para identificar retroCNVs em dados de sequenciamento de genomas completos e exomas de indivíduos de quaisquer espécies que tenham seus genomas de referência já sequenciados e bem montados. A seguir, será apresentado o sideRETRO e o seu modo de funcionamento.

3.2 O DESENVOLVIMENTO DO SIDERETRO

O sideRETRO foi desenvolvido *ab initio* na linguagem de programação C (KERNIGHAN; RITCHIE, 1988) e usa as bibliotecas HTSlib (BONFIELD et al., 2021) e SQLite3 (HIPP, 2019) para gerenciar a leitura e a análise dos dados.

O fluxo de informação, de modo resumido, pela ferramenta começa com a entrada do arquivo de anotação dos genes conhecidos para a espécie que será investigada, no Formato ou de Transferência de Genes[23] (GTF) ou de Funcionalidade Geral versão 3[24] (GFF3) (EMBL-EBI, 2021), e dos dados de NGS previamente alinhados contra o genoma de referência, podendo estes estarem no formato SAM ou nos respectivos formatos de compressão BAM e CRAM (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021c). Daí, ocorre a captura dos alinhamentos anormais presentes nos arquivos de alinhamento, os quais são registrados num banco de dados. Então, usando-se o algoritmo de aprendizado de máquina não supervisionado chamado Agrupamento Espacial

[23] *Gene Transfer Format* (GTF).
[24] *General Feature Format version 3* (GFF3).

Baseado em Densidade de Aplicações com Ruído[25] (DBSCAN), são agrupadas as leituras anormais, reconstruindo, assim, a janela de inserção da retroCNV. Por fim, os eventos de retroduplicação são anotados com relação ao gene parental, a posição genômica de integração, a polaridade da fita, o contexto genômico, a genotipagem e a haplotipagem. A saída do sideRETRO é um arquivo no Formato de Chamada de Variante[26] (VCF), contendo essas informações detalhadas das retroCNVs.

Portanto, toda a análise pode ser dividida em três partes: a primeira corresponde ao registro dos alinhamentos anormais no banco de dados, a segunda ao agrupamento dos alinhamentos anormais e anotação das retroCNVs e a terceira parte é a geração do arquivo VCF. Seguindo essa linha de raciocínio, o sideRETRO funciona por intermédio de um programa computacional chamado *sider*, o qual possui três subcomandos, sendo cada qual desenvolvido para rodar uma dessas etapas:

a) subcomando *process-sample* – primeira etapa de análise. Aqui, o sideRETRO indexa a anotação do genoma no formato GTF ou GFF3 e usa a biblioteca HTSlib para ler os arquivos de alinhamento no formato SAM, BAM ou CRAM. Os alinhamentos anormais identificados são registrados no banco de dados gerenciado pela biblioteca SQLite3;
b) subcomando *merge-call* – no passo seguinte ao *process-sample*, os alinhamentos anormais registrados no banco de dados são agrupados segundo o algoritmo de aprendizado de máquina não supervisionado DBSCAN. Os agrupamentos encontrados são registrados e anotados como retroCNVs no banco de dados;
c) subcomando *make-vcf* – nesta última etapa, as retrocópias presentes no banco de dados são escritas no arquivo de formato VCF.

Antes de falar especificamente da primeira etapa de processamento do *sider*, cabe uma explicação sobre o banco de dados em SQLite3 que é usado para organizar todos os dados intermediários de análise.

[25] *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN).
[26] *Variant Call Format* (VCF).

3.2.1 Banco de dados

O banco de dados usado pelo sideRETRO foi escrito, usando a biblioteca SQLite3. SQLite3 é um Sistema de Gerenciamento de Banco de Dados Relacional[27] (RDBMS) (CODD, 2002) que usa a Linguagem de Busca Estruturada[28] (SQL) (DATE, 1989) para a administração dos dados.

A escolha de se usar um banco de dados em SQL, deve-se sobretudo a dois fatores: o primeiro é manter toda a quantidade de dados intermediários - que são gerados pelas diferentes etapas de análise - organizados num único lugar e facilmente acessíveis e o segundo fator diz respeito à própria possibilidade de se usar a linguagem SQL para auxiliar nas análises dos dados: tendo todas as informações, concernentes a genes, éxons, alinhamentos anormais etc, divididas ordenadamente em tabelas, a busca correlacionando diferentes atributos e características torna-se mais simples.

Sendo assim, a base de dados está dividida em 13 tabelas (Figura 16):

a) *schema* – onde se encontra o versionamento do banco de dados. Como o banco de dados sofreu alterações no decorrer do seu desenvolvimento, esta é uma forma de garantir que apenas a versão adequada do sideRETRO vá poder alterar os dados das tabelas;
b) *batch* – entidade forte que se relaciona com a tabela *source* numa cardinalidade 1:N. Esta tabela serve para guardar informações das diferentes bateladas: uma batelada pode ser interpretada como os arquivos de alinhamento no formato SAM, BAM ou CRAM que foram processados juntos. Como o sideRETRO permite que vários banco de dados - gerados cada qual para um determinado propósito - possam ser unidos durante uma análise, assim pode-se distinguir de qual grupo, batelada, um determinado arquivo de alinhamento veio;
c) *source* – entidade forte que se relaciona com as tabelas *alignment* e *genotype* com cardinalidade 1:N. Nesta tabela são guardadas as informações individuais de cada arquivo de alinhamento SAM, BAM ou CRAM - como a

[27] *Relational Database Management System* (RDBMS).
[28] *Structured Query Language* (SQL).

batelada a que pertence e o caminho no disco rígido até o arquivo processado. Desta forma, é possível saber de qual arquivo um dado alinhamento anormal vem, assim como atribuir uma retrocópia detectada durante a genotipagem ao seu respectivo indivíduo;

d) *exon* – entidade forte que se relaciona com a tabela *overlapping* numa cardinalidade 1:N. Esta tabela armazena os dados dos arquivos de anotação do genoma, nos formatos GTF ou GFF3, referentes aos éxons dos genes codificadores de proteína. As informações de interesse são a posição do éxon (cromossomo, posição de início e fim, fita) e o gene, ao qual pertence. O sideRETRO usa estas informaçõe para verificar se alguma das leituras pareadas do alinhamento anormal se sobrepõe a um éxon de um gene codificador de proteína;

e) *alignment* – entidade forte que se relaciona com a tabela *overlapping* com cardinalidade 1:N e com a tabela clustering N:M. Nesta tabela, os dados provindos dos alinhamentos anormais são armazenados, tais como: posição genômica (cromossomo, posição de início e fim), qualidade (nível de qualidade *Phred*), mapeamento do alinhador (*flag* de alinhamento, CIGAR *string*), o nome que identifica as leituras pareadas dos alinhamentos anormais (*qname*), o arquivo de alinhamento de origem (tabela *source*) e o tipo de alinhamento anormal (em cromossomos diferentes, em regiões distantes, leituras divididas com alinhamento suplementar);

f) *overlapping* – entidade fraca que conecta as tabelas *exon* e *alignment* numa cardinalidade N:M. Esta tabela anota a sobreposição entre alinhamentos anormais e éxons de genes codificadores. Também são anotados a posição em que começa a sobreposição e o comprimento da sobreposição em bases;

g) *cluster* – entidade forte que se relaciona com as tabelas *clustering* e *overlapping_blacklist* com cardinalidade 1:N e *cluster_merging* com cardinalidade 1:1. Esta tabela contém as informações concernentes à etapa de agrupamento. É anotada a posição genômica do grupo (cromossomo, início e fim), o seu gene parental (das leituras pareadas, uma alinhou num éxon do gene parental e a seguinte alinhou dentro do agrupamento) e por quantos filtros de qualidade o agrupamento passou. Os filtros usados são: de cromossomo (pode-se excluir agrupamentos de um determinado cromossomo, como por exemplo o mitocondrial), de distância (para excluir os

agrupamentos que caem próximos do gene parental - que devem advir de erro do alinhador), de região (as regiões já previamente anotadas como inserção de retrocópia - vindas de um arquivo de anotação do genoma GTF ou GFF3) e, por fim, de suporte (se o agrupamento possui um valor mínimo de cobertura de alinhamentos anormais). Todos estes filtros possuem valores padrão, mas é oferecido ao usuário a possibilidade de modificá-los - sobre isso, será falado com mais detalhes no transcorrer desta metodologia;

h) *clustering* – entidade fraca que conecta as tabelas *cluster* e *alignment* numa cardinalidade N:M. Aqui são organizadas as informações geradas pelo algoritmo de aprendizado de máquina não supervisionado DBSCAN, tais como: o alinhamento anormal, o número de alinhamentos anormais que são vizinhos dele, um rótulo, *label*, que identifica esse alinhamento dentro do grupo (podendo ser um membro núcleo, alcançável ou ruído) e o identificador do agrupamento. Com esta tabela é possível saber quais alinhamentos estão presentes num dado grupo da tabela *cluster*;

i) *blacklist* – entidade forte que se relaciona com tabela *overlapping_blacklist* com cardinalidade 1:N. Nesta tabela são registradas todas as regiões do genoma (cromossomo, início e fim), onde não pode haver um agrupamento de alinhamentos anormais. Estas regiões são retiradas de um arquivo de anotação do genoma GTF ou GFF3, que previamente já as publicou como inserção de retrocópia. Essas informações são usadas no filtro de qualidade da tabela *cluster*, especificamente para o filtro de região. Desta forma, eliminam-se das análises as retrocópias anotadas no genoma de referência, que, portanto, são fixadas;

j) *overlapping_blacklist* – entidade fraca que relaciona as tabelas *cluster* e *blacklist* com cardinalidade N:M. Esta tabela registra quais grupos da tabela *cluster* se sobrepõem a uma região de agrupamento proibido - região vinda da tabela *blacklist*. Também são anotados a posição que começa a sobreposição e o comprimento em bases dela;

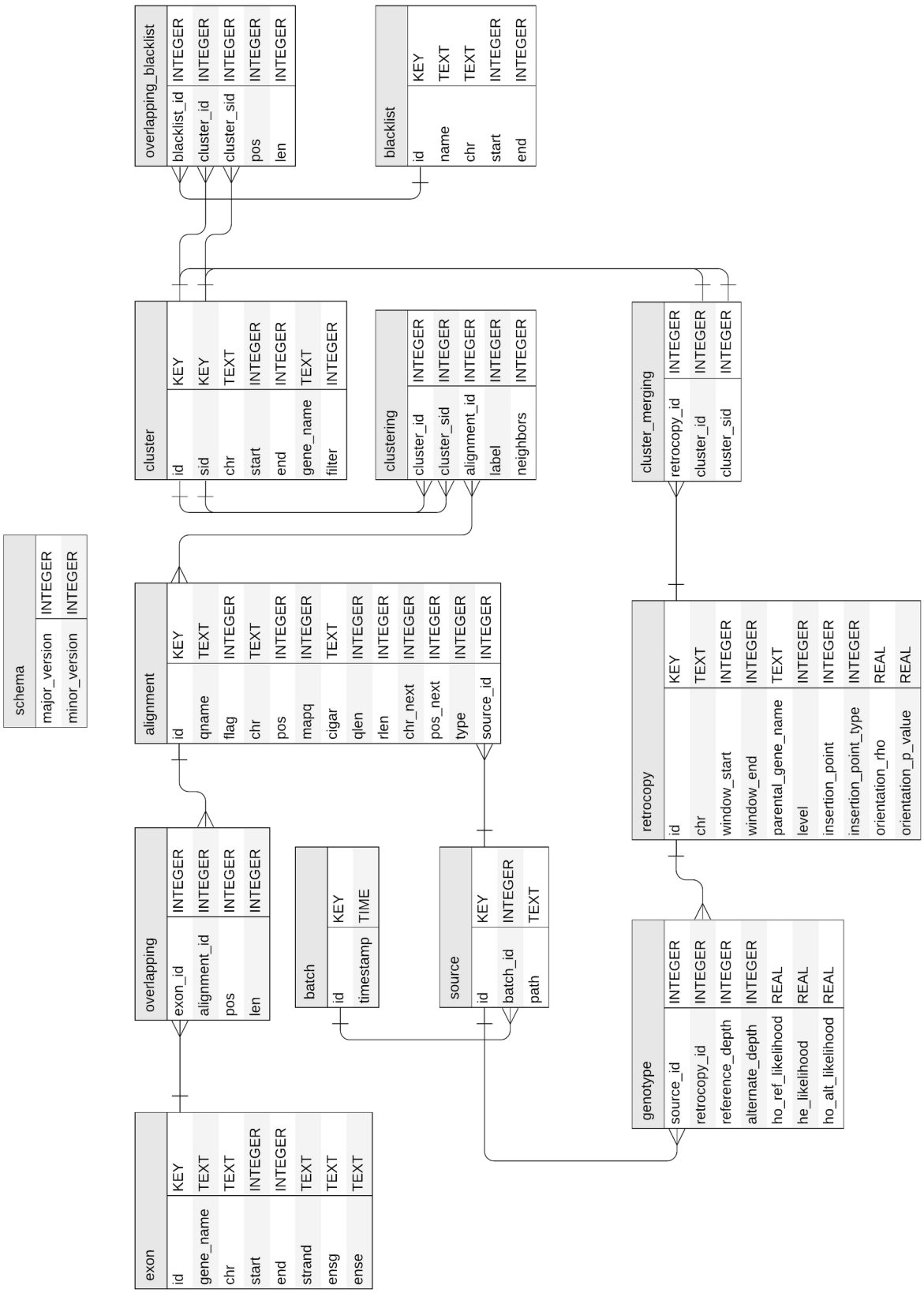
k) *retrocopy* – entidade forte que se relaciona com as tabelas *cluster_merging* e *genotype* com cardinalidade 1:N. Aqui são registradas as retroCNVs e os seguintes achados com relação a elas: posição genômica, onde ocorreu a janela do evento de retroduplucação (cromossomo, início da janela de inserção, fim da janela de inserção) - informação advinda dos dados de

agrupamento na tabela *cluster* - o ponto de inserção da retroCNV e o modo como este foi calculado (ou o ponto de inserção foi calculado a partir dos alinhamentos suplementares, ou foi feita uma média aritmética dos pontos de início e fim da janela de inserção), o gene parental e a orientação da retroCNV em relação ao seu gene parental (se está na mesma fita, ou em fitas opostas) - neste caso é calculado o coeficiente *rho* de correlação de postos de *Spearman* - entre as posições dos alinhamentos no gene parental e as posições de seus respectivos pares alinhados na região de agrupamento - e também o valor-p dessa análise estatística;

l) *cluster_merging* – entidade fraca que relaciona as tabelas *retrocopy* e *cluster* com cardinalidade 1:N. Nesta tabela, são indicados quais agrupamentos da tabela *cluster* deram origem a retrocópia na tabela *retrocopy*;

m) *genotype* – entidade fraca que relaciona as tabelas *retrocopy* e *source* com cardinalidade N:M. Nesta tabela, são anotadas as retroCNVs da tabela *retrocopy* que estão presentes num dado arquivo de alinhamento SAM, BAM ou CRAM, o qual consta na tabela *source*. São também registrados os detalhes concernentes a genotipagem, tais como: a cobertura de alinhamentos anormais que evidenciam a inserção de uma retroCNV (o alelo alternativo) e a cobertura de alinhamentos normais que evidenciam a ausência de um inserção de retroCNV na dada região de agrupamentos dos alinhamentos anormais (alelo referência). São calculadas as probabilidades de genotipagem para homozigoto referência, heterozigoto e homozigoto alternativo.

Figura 16 - Diagrama de relacionamento de entidade para o banco de dados do sideRETRO.



Fonte: sideRETRO versão 1.0.0.

O banco de dados é usado no transcorrer de toda a análise feita pelo sideRETRO, por cada subcomando do executável *sider*:

a) o *process-sample* gera o banco de dados e o popula com os dados de anotação do genoma e dos alinhamentos anormais capturados dos arquivos de alinhamento;
b) o *merge-call* processa as informações já presentes no banco de dados e anota as retroCNVs, entre outras informações concernentes a elas;
c) então, o subcomando *make-vcf* lê as anotações de retroCNV presentes no banco de dados e as escreve num arquivo de formato VCF.

Tendo, portanto, o banco de dados sido descrito, passar-se-á à metodologia usada por cada subcomando do *sider*.

3.2.2 Subcomando *process-sample*

A primeira das três partes da análise conduzida pelo sideRETRO é chamada de *process-sample*. Esta etapa consiste na captura de alinhamentos anormais em dados de NGS, já previamente alinhados contra o genoma de referência, e no registro deles no banco de dados em SQL - visando a próxima análise a ser feita pelo subcomando *merge-call*.

No processo de capturar alinhamentos, ditos anormais, estes são submetidos a uma série de filtros que visam garantir a qualidade do alinhamento e as características que determinam o fato de eles serem considerados anormais, tais como:

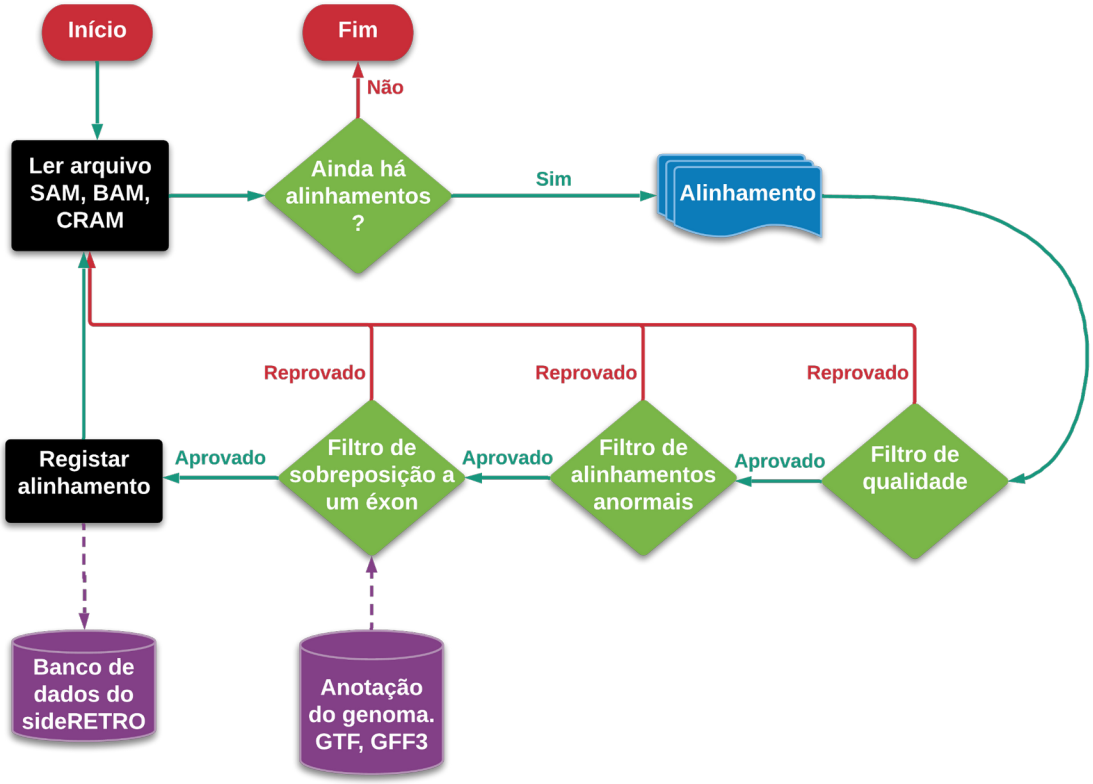
a) leituras pareadas em que os elementos do par se alinham em cromossomos diferentes;
b) leituras pareadas em que os elementos do par se alinham a uma distância maior que o esperado;
c) alinhamentos suplementares.

Outro fator a ser levado em conta é se pelo menos uma das leituras pertinentes ao mesmo fragmento alinha num éxon de algum gene codificador de proteína. Esse é um filtro importante para com a anotação de retroCNVs: como elas são cópias truncadas de genes codificantes - geradas pela maquinaria de L1 - o alinhamento exônico, de um dos pares de leituras anormais, indica o possível gene parental, do qual se originou a retrocópia.

Sendo assim, o subcomando *process-sample* pode ser dividido em quatros etapas (Fluxograma 1): três filtros consecutivos, seguidos de uma última etapa de registro no banco de dados:

a) filtro de qualidade;
b) filtro de alinhamentos anormais;
c) filtro de sobreposição a um éxon;
d) registro do alinhamento.

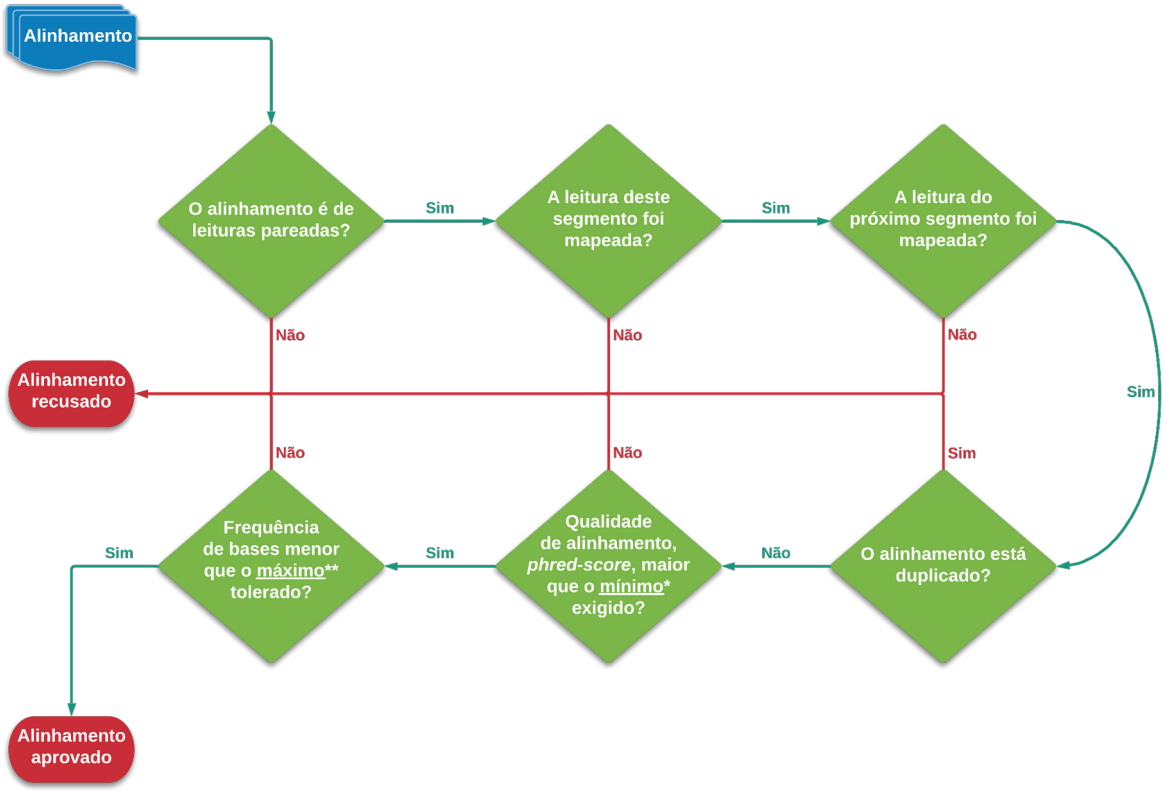
Fluxograma 1 - Subcomando *process-sample*.



Fonte: autoria própria.
A visão geral do algoritmo usado pelo subcomando do sider *process-sample*. Nele, são destacadas as etapas de filtragem e registro dos alinhamentos no banco de dados do programa.

3.2.2.1 Filtro de qualidade

Fluxograma 2 - Filtro de qualidade.



Fonte: autoria própria.
O alinhamento precisa passar por 6 subfiltros, de modo a ser aprovado para as próximas etapas do process-sample.
* Valores padrão que podem ser alterados pelo usuário: nível mínimo de qualidade *Phred* requerido; frequência máxima de bases tolerada.

O filtro de qualidade visa garantir a consistência dos dados de alinhamento. Para isso, são selecionados aqueles que apresentem as seguintes seis características (Fluxograma 2):

a) são oriundos de um sequenciamento de leituras pareadas - pois, esta metodologia é exclusiva para essa tecnologia de sequenciamento;
b) o alinhador conseguiu mapear a leitura do segmento no genoma de referência;
c) o alinhador conseguiu mapear a leitura do próximo segmento no genoma de referência. Do próximo segmento é a leitura pareada advinda do mesmo fragmento;

d) a alinhamento não foi marcado como uma duplicação oriunda da Reação em Cadeia da Polimerase[29] (PCR) ou uma duplicação óptica;
e) o nível de qualidade *Phred* é maior que o mínimo requerido;
f) a frequência de bases da leitura é menor que o máximo tolerado.

Com exceção das alíneas e) e f), todos as demais são verificados pela presença ou ausência de *flags bit* a *bit* - que estão combinadas num único valor numérico presente na segunda coluna do formato SAM (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021a, p. 6). Cada *bit* representa um atributo assinalado pelo alinhador e deve estar de acordo com as especificações do formato SAM/BAM (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021a, p. 7). Os *bits* de interesse para o filtro de qualidade encontram-se na Tabela 1.

Tabela 1 - *Flags bit* a *bit* do arquivo SAM/BAM verificadas pelo filtro de qualidade.

Descrição	*Bit*	
	Hexadecimal	Decimal
Alinhamentos pareados	0x1	1
Seguimento não mapeado	0x4	4
Próximo segmento não mapeado	0x8	8
Duplicação de PCR ou ótica	0x400	1024

Fonte: (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021a, p. 7).

Portanto, o filtro de qualidade verifica se está presente a *flag* 0x1 - alinhamento pareado - e se estão ausentes as *flags* 0x4, 0x8 e 0x400 - o alinhamento precisa ter sido mapeado pelo alinhador, assim como o seu par, e não ser uma duplicação de PCR ou óptica.

As alíneas e) e f) são verificadas por dois valores padrão (Tabela 2), que podem ser alterados pelo usuário da ferramenta. Os valores padrão para o nível de qualidade *Phred* e a frequência de bases foram definidos heurísticamente. O nível de qualidade *Phred* tem um valor padrão de 8, que equivale aproximadamente à probabilidade de 16 de 100 bases estarem incorretas - uma precisão de aproximadamente 84% (EWING; GREEN, 1998). O filtro de frequência de bases visa evitar leituras com um alto teor repetitivo de bases, que, por conseguinte, alinham em regiões repetitivas do genoma (LI; FREUDENBERG, 2014). Por padrão, o valor

[29] *Polymerase Chain Reaction* (PCR).

máximo tolerado para a frequência de bases é de 75% - 75 de 100 bases podem se repetir em uma leitura.

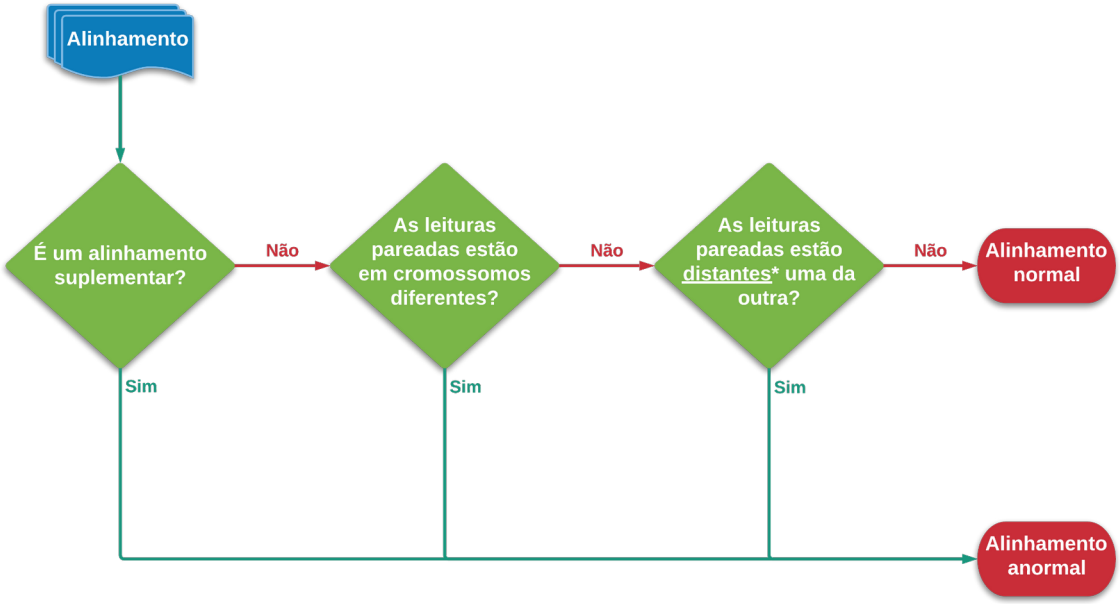
Tabela 2 - Valores padrão para o nível de qualidade *Phred* e a frequência de bases.

Filtro	Valor padrão
Nível mínimo de qualidade Phred requerido	8
Frequência máxima de bases tolerada	75%

Fonte: sideRETRO versão 1.0.0.

3.2.2.2 Filtro de alinhamentos anormais

Fluxograma 3 - Ffiltro de alinhamentos anormais.



.

Fonte: autoria própria.
O algoritmo classifica o alinhamento em normal ou anormal. Os alinhamentos normais são descartados e os anormais são encaminhados para o filtro de sobreposição a um éxon.
* Valor padrão que pode ser alterado pelo usuário: leituras pareadas alinhadas em regiões distantes.

Os alinhamentos que atenderem a todos os critérios do filtro de qualidade passam, então, para esta etapa, na qual serão classificados em alinhamento normal ou anormal (Fluxograma 3). Os alinhamentos que forem julgados anormais seguem para a etapa seguinte - o filtro de sobreposição a um éxon.

O classificador testa para três tipos de alinhamentos anormais:

a) leituras divididas (leituras com alinhamento suplementar);
b) leituras pareadas em que os membros do par se alinham em cromossomos diferentes;
c) leituras pareadas em que os elementos do par se alinham a uma distância maior que o esperado.

O alinhamento suplementar pode ser identificado pela presença do *bit* 0x800 (Tabela 3) na *flag* do alinhamento no formato SAM ou BAM. Caso o *bit* de alinhamento suplementar não esteja presente, o classificador, então, verifica se o alinhamento para a leitura e o alinhamento para o par dela estão em cromossomos diferentes. Isso pode ser feito, comparando-se as colunas referentes aos cromossomos, mapeados pelo alinhador para o par de leituras: terceira e sétima colunas (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021a, p. 6), segundo o formato SAM/BAM. Por último, se o alinhamento não é suplementar e ambos os pares de leitura foram mapeados no mesmo cromossomo, testa-se se os pares de leitura foram mapeados em posições distantes do cromossomo. O valor padrão para considerar as leituras pareadas distantes é de 10.000 bases (Tabela 4), definido heurísticamente. Esse valor, assim como todos os valores padrão, pode ser alterado pelo usuário, de modo a se adequar às características dos dados.

Tabela 3 - *Flag* que define um alinhamento suplementar no formato SAM/BAM.

Descrição	*Bit*	
	Hexadecimal	Decimal
Alinhamento suplementar	0x800	2048

Fonte: (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021a, p. 7).

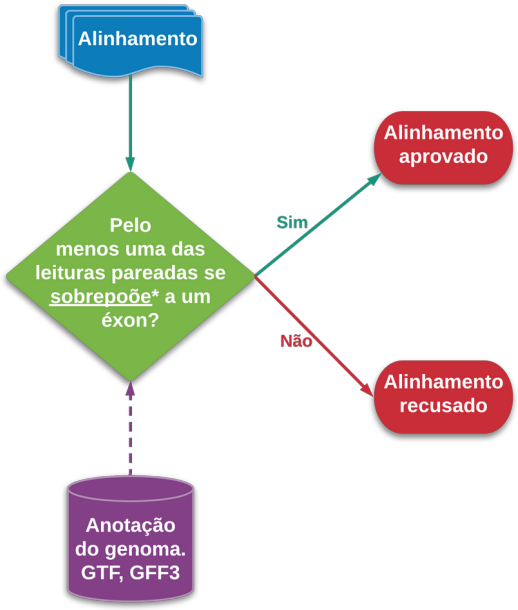
Tabela 4 - Valor padrão para leituras distantes.

Filtro	Valor padrão
Leituras pareadas alinhadas em regiões distantes	10.000 bases

Fonte: sideRETRO versão 1.0.0.

3.2.2.3 Filtro de sobreposição a um éxon

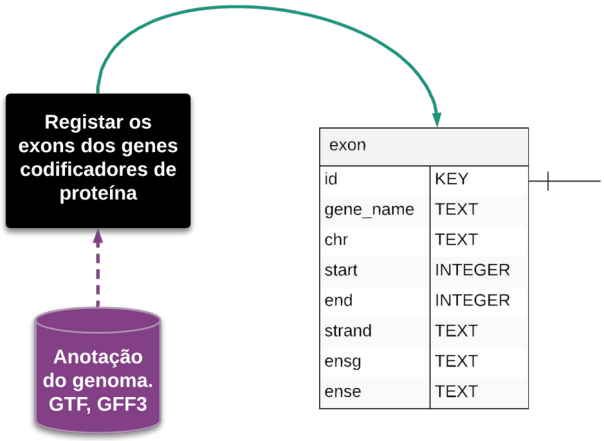
Fluxograma 4 - Filtro de sobreposição a um éxon.



Fonte: autoria própria.
Caso uma das leituras pareadas do alinhamento que foi classificado em anormal sobreponha-se a um éxon de um gene codificador, o alinhamento é aprovado e enviado para armazenamento no banco de dados do sideRETRO.
* Valor padrão que pode ser alterado pelo usuário: sobreposição mínima entre uma das leituras do alinhamento anormal e um éxon.

Os alinhamentos que foram classificados em alinhamentos anormais são, então, testados quanto a sobreposição deles a algum éxon de um gene codificador (Fluxograma 4). Esta etapa depende do arquivo de anotação do genoma de referência (no formato GTF ou GFF3) para determinar quais genes foram anotados como genes codificadores de proteínas e quais as respectivas posições genômicas de seus éxons. Então, primeiramente, o sideRETRO anota os éxons do arquivo GTF ou GFF3 no banco de dados em SQL (Fluxograma 5), o que vai garantir o registro e o rápido acesso a essas posições genômicas, tanto para esta análise, quanto nas análises subsequentes.

Fluxograma 5 - Registro dos éxons codificantes no banco de dados do sideRETRO.



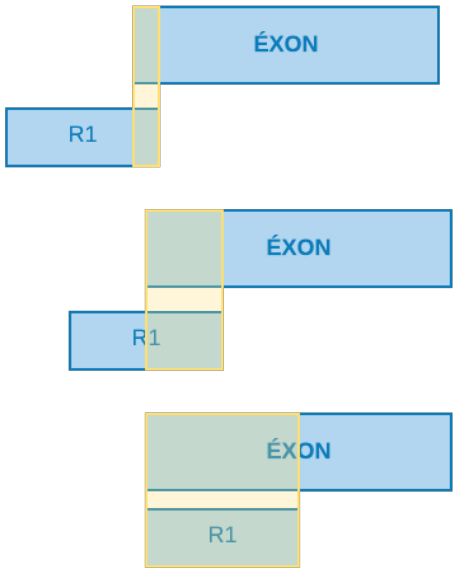
Fonte: autoria própria.

Tendo isso em mãos, a ferramenta compara as posições em que foram mapeadas as leituras pareadas do alinhamento anormal com as posições exônicas - já inscritas no banco de dados. Se houver sobreposição entre as posições, então o alinhamento é aprovado e enviado para armazenamento no banco de dados do sideRETRO.

Para considerar que houve uma sobreposição entre uma das leituras do alinhamento anormal e algum dos éxons de um gene codificador, precisa haver, por padrão, a justaposição de pelo menos uma base entre as sequências. Esse comportamento também pode ser alterado pelo usuário (Figura 17 e Tabela 5), segundo as características dos dados. Portanto, é possível determinar uma sobreposição, levando-se em consideração:

a) a fração do alinhamento que foi coberta pelo éxon;
b) a fração do éxon que foi coberta pelo alinhamento;
c) ambas as frações sendo cobertas;
d) pelo menos uma das frações sendo coberta.

Figura 17 - Três exemplos de sobreposição entre éxon e leitura.



Fonte: autoria própria.

Tabela 5 - Valor mínimo para a sobreposição entre um éxon e uma das leituras do alinhamento anormal.

Filtro	Valor padrão
Sobreposição mínima entre uma das leituras do alinhamento anormal e um éxon	1 base

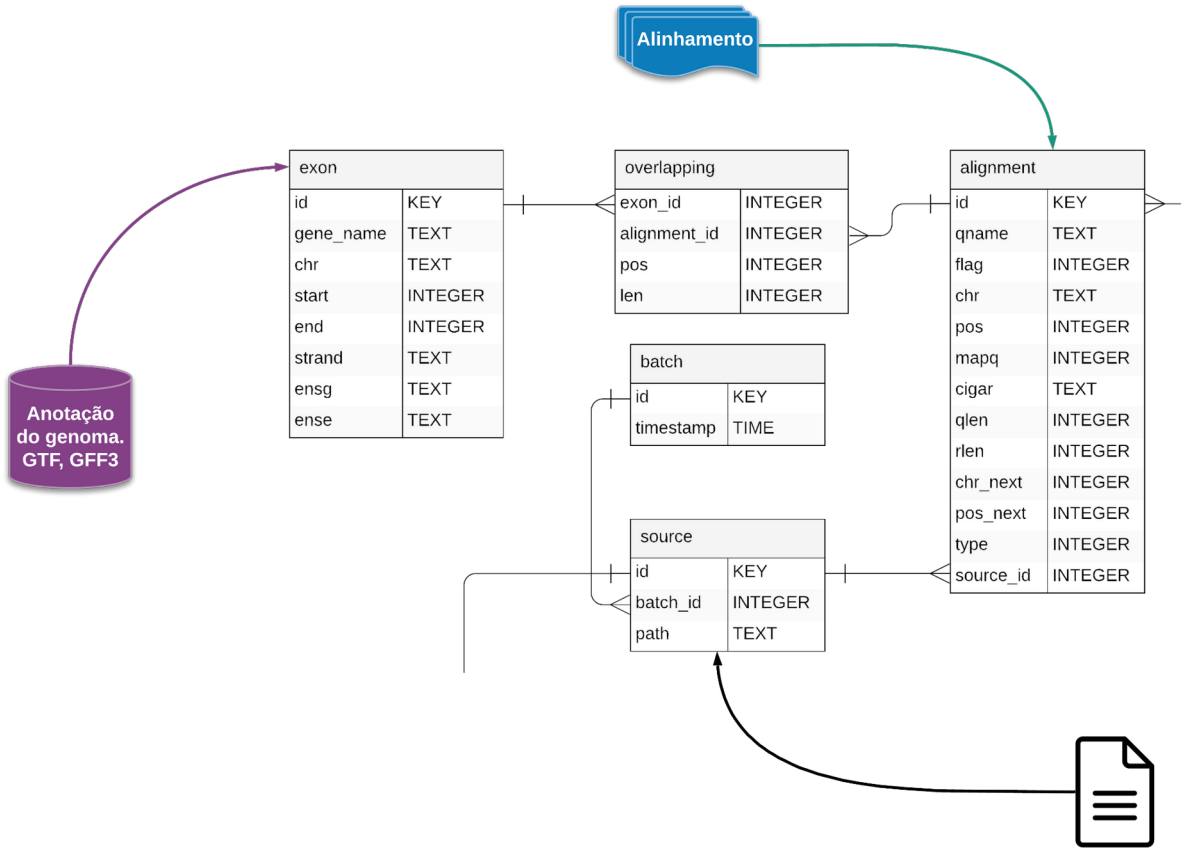
Fonte: sideRETRO versão 1.0.0.

3.2.2.4 Registro do alinhamento

O alinhamento que, por fim, passou por todos os filtros de qualidade, que foi classificado como anormal, cuja uma das leituras está sobreposta a um éxon codificante, é registrado no banco de dados do sideRETRO, na tabela *alignment* - assim como o tipo de alinhamento anormal que ele é. A posição genômica e o comprimento em bases da sobreposição com o éxon são inscritas na tabela *overlapping* (Figura 18).

Informações sobre o arquivo de alinhamento que está sendo processado também são inseridas no banco de dados - nas tabelas *batch* e *source* (Figura 18). Dessa forma, pode-se saber de qual arquivo veio o alinhamento anormal, e essa será uma informação importante para a etapa de genotipagem do subcomando *merge-call*.

Figura 18 - As tabelas que são preenchidas durante o subcomando *process-sample*.



Fonte: autoria própria.
Os éxons do genes codificadores de proteínas são registrados na tabela *exon*, o alinhamento anormal que passou por todos os filtros é registrado na tabela *alignment*. Os detalhes da sobreposição entre o alinhamento anormal e o éxon codificante são anotados na tabela *overlapping*. As informações sobre o arquivo que está sendo analisado vão para as tabelas *batch* e *source*.

O subcomando *process-sample* termina quando os arquivos de alinhamento forem todos processados. Como produto desta análise, fica o banco de dados em SQL - com as tabelas *alignment*, *exon*, *overlapping*, *batch* e *source* preenchidas. As Tabelas 6 e 7 sumarizam os valores padrão usados nesta etapa de análise e as *flags bit* a *bit* do formato SAM verificadas pelo sideRETRO, respectivamente.

O próximo passo é o subcomando *merge-call*, que irá analisar os dados gerados durante o *process-sample* e usá-los para descobrir, anotar e genotipar retroCNVs.

Tabela 6 - Sumário dos valores padrões para o subcomando *process-sample*.

Filtro	Valor padrão
Nível de qualidade *Phred* mínimo requerido	8
Frequência máxima de bases tolerada	75%
Leituras pareadas alinhadas em regiões distantes	10.000 bases
Sobreposição mínima entre uma das leituras do alinhamento anormal e um éxon	1 base

Fonte: sideRETRO versão 1.0.0.

Tabela 7 - Sumário das *flags bit* a *bit* do arquivo no formato SAM verificadas pelo sideRETRO.

Descrição	*Bit*	
	Hexadecimal	Decimal
Alinhamentos pareados	0x1	1
Seguimento não mapeado	0x4	4
Próximo segmento não mapeado	0x8	8
Duplicação de PCR ou ótica	0x400	1024
Alinhamento suplementar	0x800	2048

Fonte: (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021a, p. 7).

3.2.3 Subcomando *merge-call*

A etapa subsequente à captura dos alinhamentos anormais, *process-sample*, é chamada de *merge-call*. Durante o *merge-call* (Fluxograma 6), os alinhamentos anormais registrados no banco de dados do sideRETRO são agrupados, usando-se o algoritmo de aprendizado de máquina não supervisionado DBSCAN, e passam por uma filtragem, visando a consistência e a significância dos agrupamentos formados. Os agrupamentos que, portanto, forem aprovados são anotados e genotipados como retroCNVs.

O *merge-call* permite que mais de um banco de dados seja analisado conjuntamente, por conseguinte, as análises são precedidas por um ciclo de fusão dos diferentes bancos de dados. Dessa forma, um número grande de arquivos de alinhamento pode ser processado pelo *process-sample* separadamente, por exemplo em diferentes máquinas, e, durante o *merge-call*, todos os bancos de

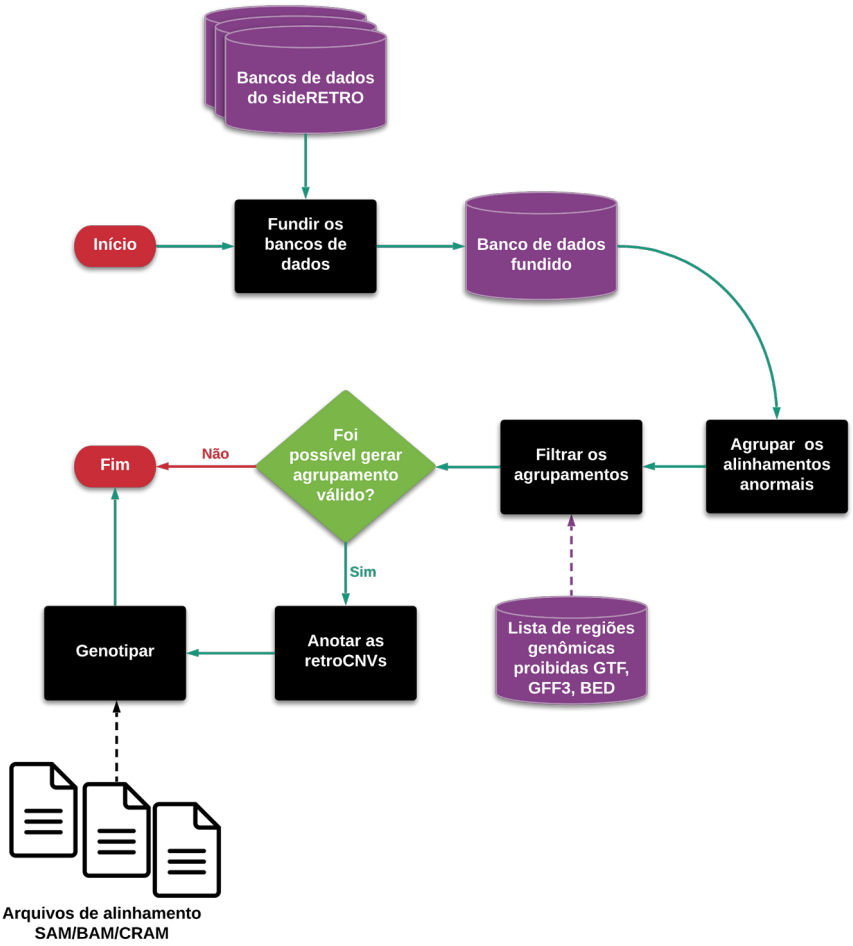
dados gerados são fundidos num único arquivo - que será enviado, sucessivamente, para a fase de agrupamento.

Pontuando as etapas pertinentes ao *merge-call*:

a) fusão dos bancos de dados;
b) agrupamento;
c) filtro de agrupamento;
d) anotação de retroCNVs;
e) genotipagem.

Cada uma dessas etapas será vista com mais detalhes a seguir.

Fluxograma 6 - Subcomando *merge-call*.

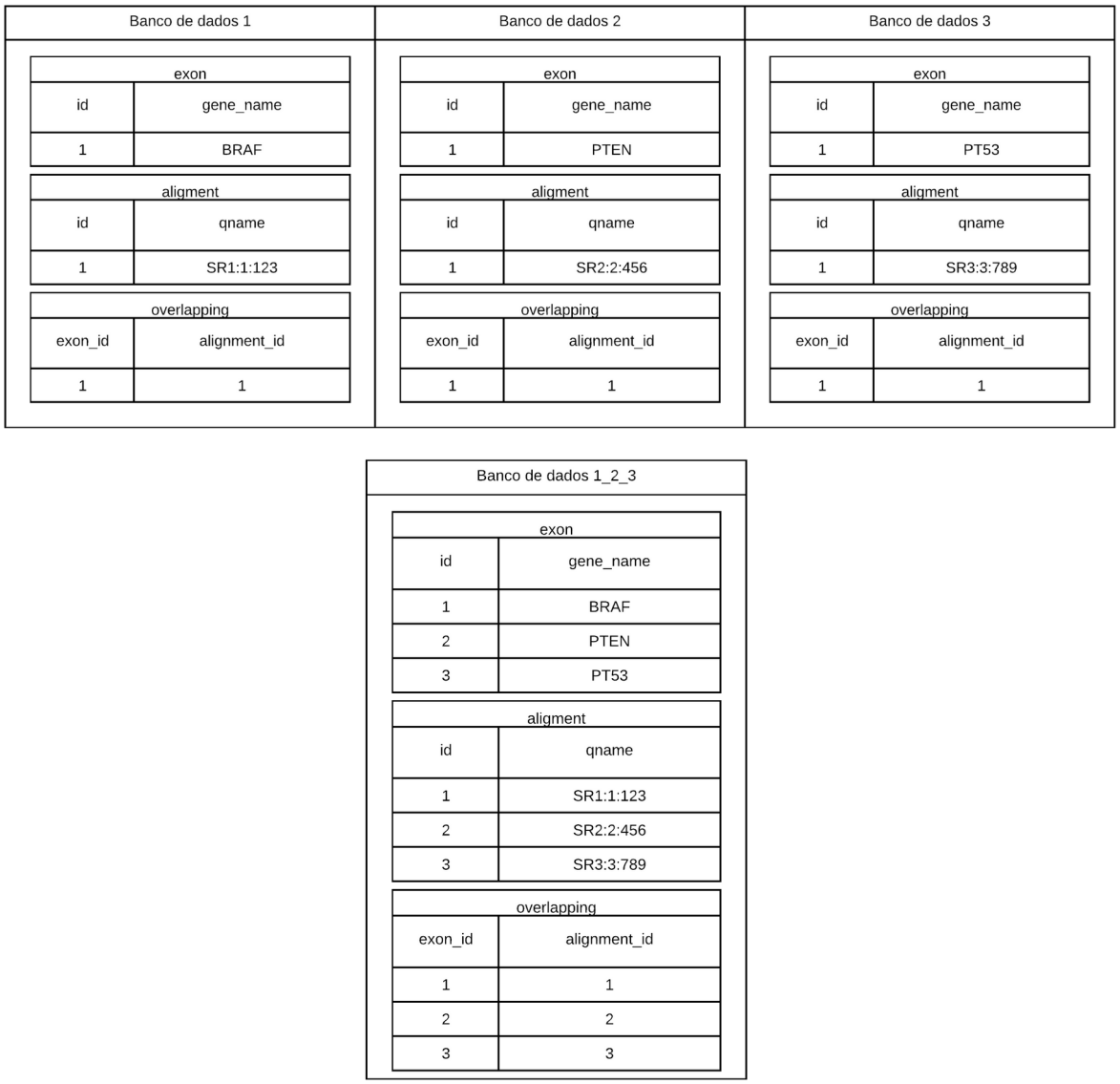


Fonte: autoria própria.

A visão geral do algoritmo usado pelo subcomando do sider *merge-call*. Em destaque estão as etapas de fusão dos banco de dados, agrupamento, filtragem dos agrupamentos, anotação das retroCNVs e genotipagem.

3.2.3.1 Fusão dos bancos de dados

Figura 19 - Fusão dos bancos de dados.



Fonte: autoria própria.
Exemplo resumido, para dois atributos apenas, da fusão das tabelas de diferentes bancos de dados gerados durante o subcomando do *sider* - *process-sample*. São vistas três tabelas: *exon*, *alignment* e *overlapping*.

O subcomando *merge-call* inicia os seus trabalhos com a fusão dos diferentes bancos de dados (Figura 19) - que foram gerados durante o *process-sample*. As tabelas que se esperam estar preenchidas até aqui são cinco, quatro entidades fortes e uma entidade fraca:

a) *batch* – relaciona os arquivos de alinhamento que rodaram conjuntamente durante o *process-sample*;
b) *source* – inclui os caminhos para os arquivos de alinhamento;
c) *exon* – éxons dos genes codificadores de proteína;
d) *alignment* – alinhamentos anormais que foram selecionados;
e) *overlapping* – única entidade fraca dentre as cinco tabelas, anota as sobreposições entre alinhamentos anormais e éxons.

O algoritmo, então, promove a união das tabelas análogas, atualizando tão somente o identificador das diferentes entradas - como pode ser visto na Figura 19. Aqui é importante salientar que a relação entre as entidades é mantida, de maneira que uma dada sobreposição entre um éxon e um alinhamento para um dado arquivo de alinhamento (SAM, BAM ou CRAM) vai continuar inalterada.

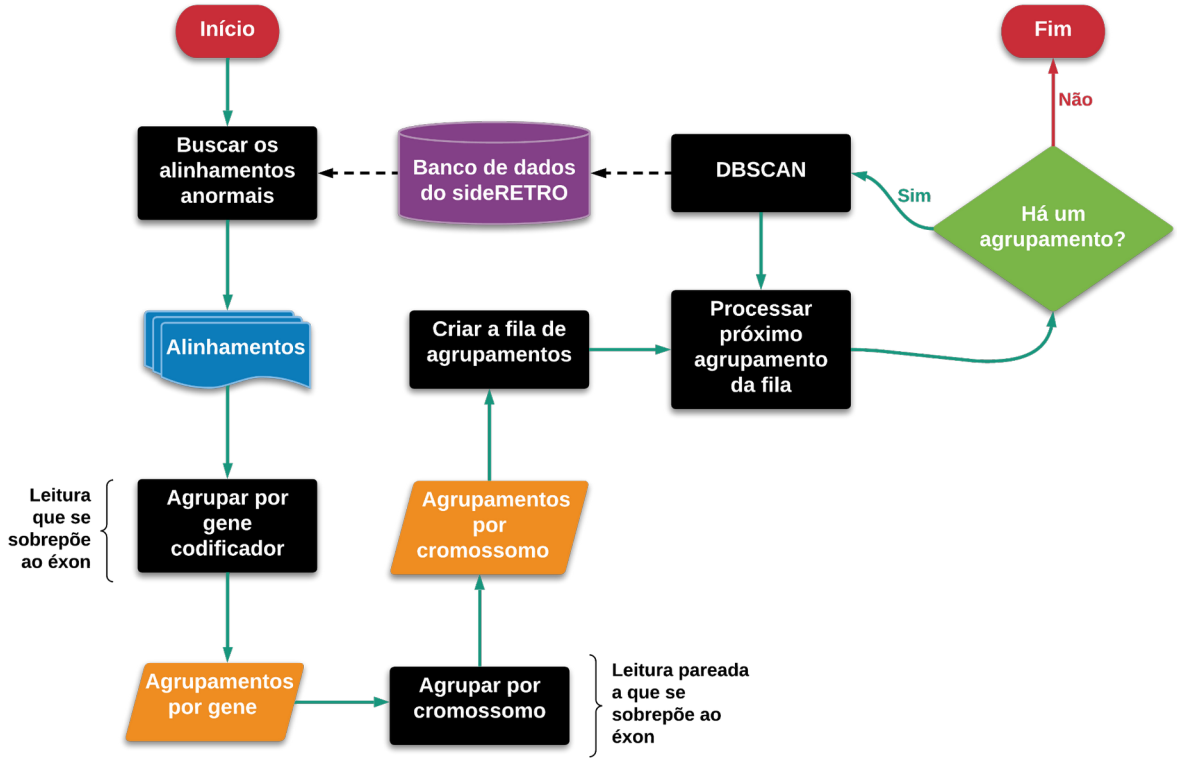
A saída da fusão dos bancos de dados é um novo banco de dados, contendo a união de todos os outros gerados na etapa de *process-sample* e passados como parâmetro para o *merge-call*.

3.2.3.2 Agrupamento

Após os bancos de dados do sideRETRO terem sido fundidos, os dados resultantes são organizados em um novo banco de dados, o qual é passado, então, para a fase de agrupamento (Fluxograma 7). Nesta fase, os alinhamentos anormais presentes serão agrupados e processados pelo algoritmo DBSCAN, visando a descoberta de retroCNVs. O agrupamento pode ser subdividido em quatro partes procedimentais:

a) agrupamento por gene codificador;
b) agrupamento por cromossomo;
c) criação da fila de agrupamentos;
d) DBSCAN.

Fluxograma 7 - Fase de agrupamento dos alinhamentos anormais.



Fonte: autoria própria.

Os alinhamentos anormais são agrupados segundo o gene codificador, ao qual um dos pares de leitura se sobrepõe, e em seguida esses agrupamentos são subagrupados por cromossomo - aqui o cromossomo em que se alinhou a leitura pareada àquela sobreposta ao gene codificador. Em seguida é criada uma fila de processamento dos agrupamentos, que irá alimentar o algoritmo DBSCAN - com o intuito de verificar se os alinhamentos anormais, de fato, formaram agrupamentos válidos.

3.2.3.2.1 Agrupamento por gene codificador

Todos os alinhamentos anotados como anormais - com leituras pareadas alinhadas em cromossomos diferentes, a uma distância maior que o esperado e alinhamentos suplementares - têm em comum o fato de que pelos menos uma das leituras do par se sobrepõe a algum éxon de um gene codificador. Portanto, primeiramente, os pares de leituras são ordenados de acordo com o gene codificador, ao qual uma das leituras do par se sobrepõe. Assim são formados agrupamentos de alinhamentos anormais por gene codificador.

3.2.3.2.2 Agrupamento por cromossomo

Os agrupamentos de alinhamentos anormais por gene codificador são, por sua vez, ordenados pelo cromossomo. Aqui a ideia é separar as leituras pareadas entre aquelas que se sobrepõem ao gene codificador em questão e as demais e subagrupar estas pelo cromossomo ao qual alinharam. Sendo assim, cada gene codificador terá uma lista de alinhamentos anormais agrupados por cromossomo.

3.2.3.2.3 Criação da fila de agrupamentos

De modo a preparar os devidos alinhamentos anormais para alimentar o algoritmo DBSCAN, é construída uma fila de agrupamentos. Nesta fila, é organizada a ordem de processamento dos agrupamentos pertinentes a cada gene codificador - segundo os cromossomos, nos quais houve alinhamentos anormais (Tabela 8). Cada agrupamento é passado, então, ao algoritmo DBSCAN, a fim de se avaliar se se trata de uma inserção de retroCNV.

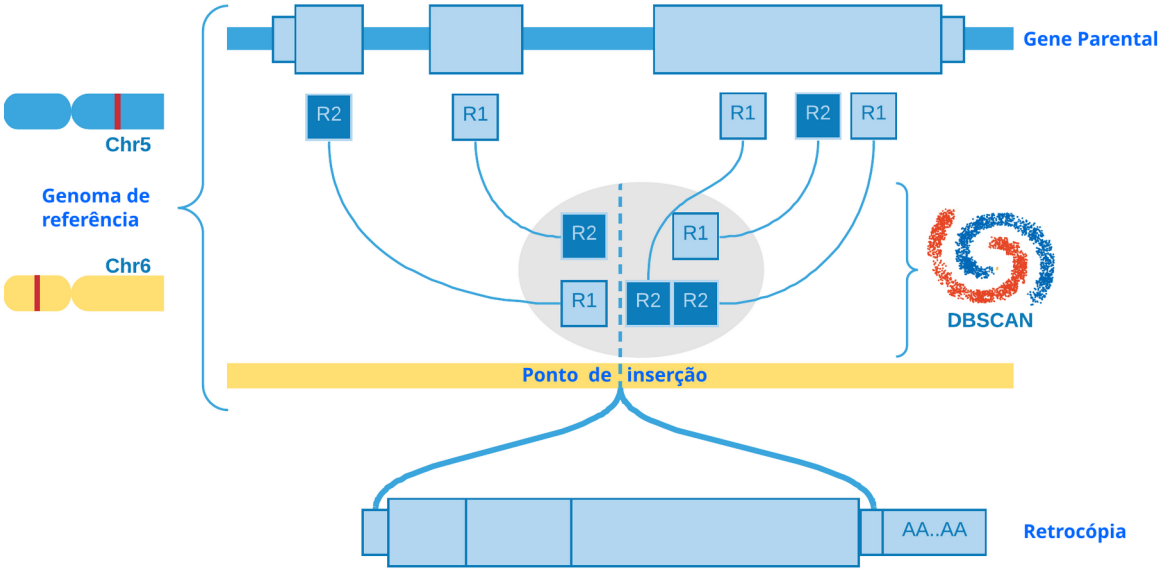
Tabela 8 - Exemplo de uma fila de agrupamentos.

Fila	Gene codificador	Alinhamento	
		Cromossomo	Posição
1	PTEN	chr1	85.186
	PTEN	chr1	85.121
	PTEN	chr1	85.127
2	BRAF	chr7	130.980
	BRAF	chr7	131.000
3	BRAF	chrX	15.450.376
	BRAF	chrX	15.450.501

Fonte: autoria própria.
Aqui se vêem três agrupamentos para dois genes codificadores, um para o gene PTEN e dois para o gene BRAF. O gene PTEN, localizado no cromossomo 10, neste caso possui um agrupamento no cromossomo 1, enquanto o gene BRAF possui um agrupamento no mesmo cromossomo 7 em que está localizado e outro agrupamento no cromossomo X. Esses três agrupamentos são passados ao algoritmo DBSCAN, o qual poderá apontá-los como inserções de retroCNVs.

3.2.3.2.4 DBSCAN

Figura 20 - Agrupamento de leituras pelo algoritmo DBSCAN.



Fonte: autoria própria.
Exemplo de agrupamento que representa uma retroCNV no cromossomo 6 provinda de um gene codificador presente no cromossomo 5.

DBSCAN é um algoritmo de aprendizado de máquina não supervisionado para o agrupamento espacial, baseado na densidade de pontos presentes na vizinhança um dos outros (ESTER et al., 1996). Para este projeto em particular, o espaço é compreendido como a extensão do cromossomo e um ponto como o intervalo entre o início e o fim de uma leitura (Figura 20). O algoritmo DBSCAN necessita de dois parâmetros para ser executado: a distância ε e o número mínimo de pontos.

a) a distância, ou raio, ε é a distância máxima entre dois pontos para que estes sejam considerados vizinhos. Sendo os pontos as leituras alinhadas, a vizinhança é constituída pela união das leituras que estão a uma distância ε a montante e a jusante para uma dada leitura em específico - assim como sobrepostas a ela;
b) o número mínimo de pontos é o número mínimo de leituras presentes na vizinhança para que a região formada seja considerada uma região densa.

Esses dois parâmetros precisam ser determinados de forma heurística, de modo a se adaptar aos dados. O sideRETRO inclui valores padrão para esses parâmetros (Tabela 9), assim como dá a possibilidade do usuário escolher os valores que mais poderão se adequar aos seus dados.

Tabela 9 - Opções para os parâmetros de DBSCAN com seus respectivos valores padrão.

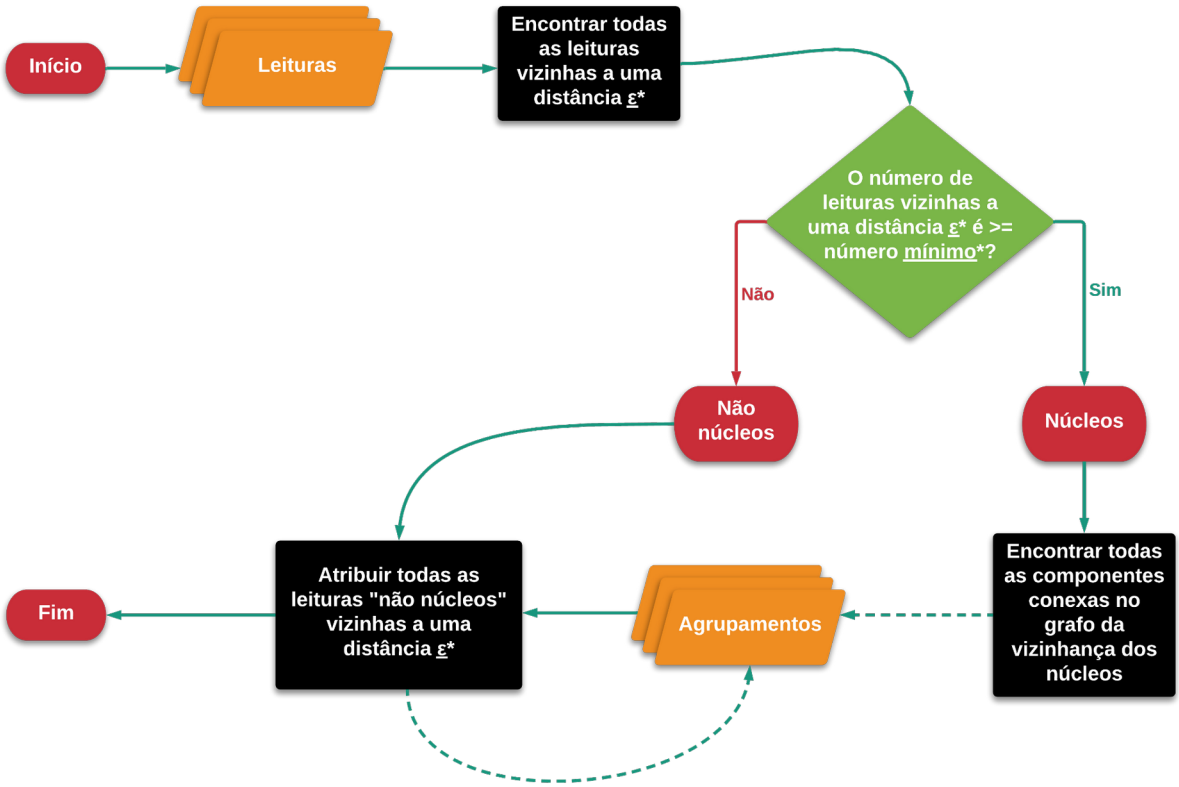
Parâmetros para DBSCAN	Valor padrão
Distância épsilon ε	300 bases
Número mínimo de pontos	10 alinhamentos

Fonte: sideRETRO versão 1.0.0.

O algoritmo abstrato de DBSCAN (SCHUBERT et al., 2017), aplicado no contexto de descoberta de retroCNVs, consiste em agrupar todas as leituras provindas de alinhamentos anormais - que possuem em comum o fato de que estão alinhadas contra o mesmo cromossomo, enquanto as leituras pareadas a elas se sobrepõem a algum éxon do mesmo gene codificador.

Primeiramente, é calculada a vizinhança para cada leitura, sendo a vizinhança constituída pelas leituras que estão a uma distância ε. As leituras que tiverem em sua vizinhança um número maior que o número mínimo de pontos aceito são classificadas como núcleos. É, então, formado um grafo com as vizinhanças das leituras classificadas como núcleos e, conforme a presença dos núcleos na vizinhança um dos outros, estabelecem-se componentes conexas. Essas componentes conexas formam regiões densas, podendo ser chamadas agora de agrupamentos. As leituras que não foram classificadas como núcleo são atribuídas a algum agrupamento formado, caso estejam a uma distância ε dele (Fluxograma 8). Assim, como resultado, são obtidos grupos que formam regiões densas e são candidatos a serem contados como retroCNVs nas análises subsequentes.

Fluxograma 8 - O algoritmo abstrato de DBSCAN.



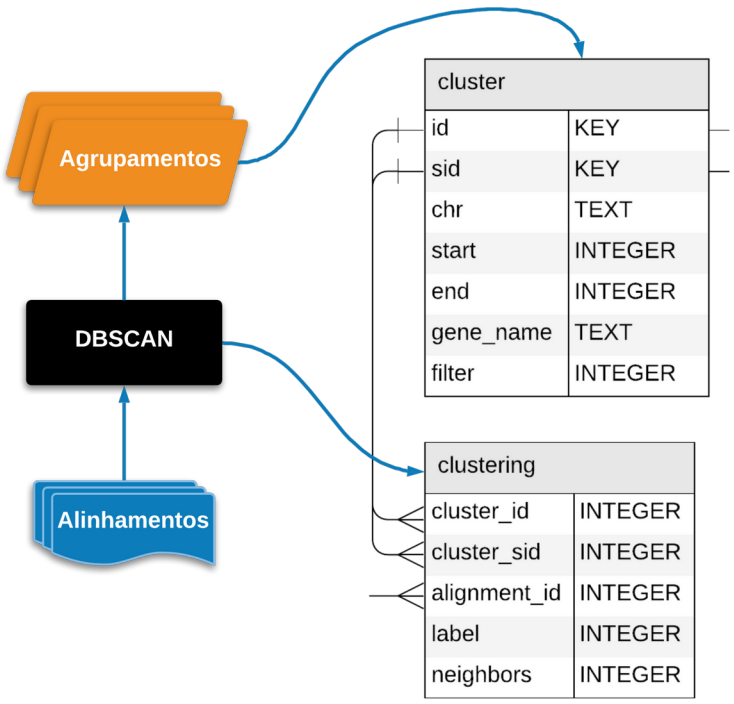
Fonte: autoria própria.
Para cada leitura, calcula-se o número de leituras vizinhas a uma distância ε. Se esse número de leituras vizinhas for maior ou igual a um número mínimo, então a dada leitura é classificada como núcleo, se não, como não núcleo. As leituras núcleos são agrupadas conforme a presença na vizinhança uma das outras (como componentes conexas), formando regiões densas. As leituras classificadas como não núcleos se estiverem a uma distância ε de alguma região densa são, por conseguinte, incorporadas a ela. Essas regiões densas são os agrupamentos formados pelo DBSCAN.
* Valores padrão que podem ser alterados pelo usuário: distância ε; número mínimo de pontos.

Os resultados obtidos durante o processo de agrupamento são registrados no banco de dados do sideRETRO (Figura 21). Especificamente, as tabelas preenchidas são: *cluster* e *clustering*.

a) tabela *cluster* – contém as informações de mais alto nível concernentes aos agrupamentos formados pelo algoritmo DBSCAN. Isso inclui o cromossomo, o intervalo compreendido entre o início e o fim do agrupamento, o gene codificador - cujos éxons são sobrepostas pelas leituras pareadas às que formam o grupo - e os filtros que validam como uma retroCNV, que serão computados na próxima etapa - a de filtragem dos agrupamentos;
b) tabela clustering – contém as informações de mais baixo nível concernentes aos agrupamentos. São inseridas as leituras do grupo, assim como os dois

dados gerados pelo DBSCAN: *label*, que identifica se a leitura foi classificada como núcleo, e *neighbors*, que é o número de leituras presentes na vizinhança.

Figura 21 - Preenchimento das tabelas do banco de dados do sideRETRO durante a etapa de agrupamento.



Fonte: autoria própria.

3.2.3.3 Filtro de agrupamento

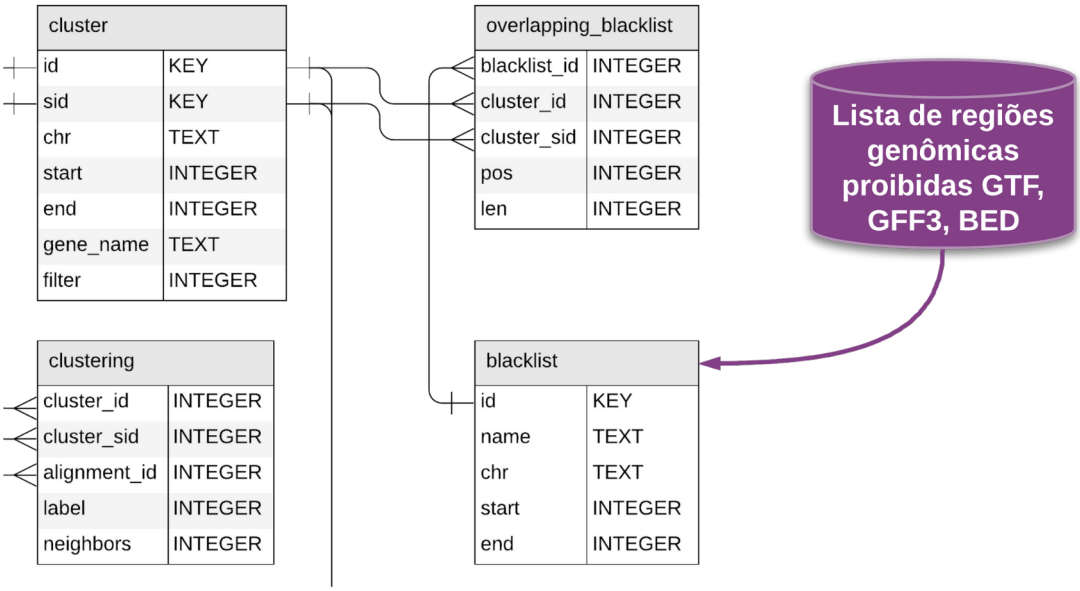
Os agrupamentos formados são validados a partir de uma série de filtros (Tabela 11), que visam remover os grupos que estão presentes em regiões que foram classificadas como proibidas, ou que estão próximos de seus genes parentais (gene parental é o gene codificador, cujos éxons são sobrepostos pelas leituras pareadas às leituras mapeadas no grupo), ou que foram formados por indivíduos (arquivos de alinhamento) com baixa cobertura de leituras dentro do agrupamento. Então há quatro filtros (Fluxograma 9):

a) filtro de cromossomo – visa remover agrupamentos que estão presentes num dado cromossomo desinteressante para as análises, ou cujo gene parental esteja presente neste cromossomo. Um exemplo de possível uso deste filtro é a remoção de agrupamentos formados no cromossomo mitocondrial, ou com o gene parental vindo desse cromossomo;
b) filtro de região – aqui o usuário pode passar um arquivo com regiões vedadas para agrupamento. Este arquivo pode estar nos formatos GTF, GFF3 ou no formato de Dados Extensíveis do Navegador[30] (BED) (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2022) (Figura 22). A ideia deste filtro é evitar a anotação de retroCNVs em regiões sabidamente de baixa qualidade na montagem do genoma de referência, ou também regiões que já foram anotadas, contendo pseudogenes. O arquivo com as regiões é indexado internamente no banco de dados do sideRETRO na tabela *blacklist*. Quando houver a justaposição de um agrupamento e de uma região proibida, a sobreposição é registrada na tabela *overlapping_blacklist*;
c) filtro de distância – para agrupamentos formados no mesmo cromossomo de seu gene parental. Este filtro remove os grupos que se encontram a uma distância inferior à distância mínima permitida em relação a esse gene. Portanto, este filtro visa evitar a anotação de eventos falsos positivos. Os grupos que falharem neste filtro passam por um processo de reagrupamento com DBSCAN. Primeiramente, são removidas do grupo as leituras consideradas próximas do gene e as restantes seguem para validação com DBSCAN. Caso o algoritmo possa agrupar as leituras filtradas, então o novo grupo criado segue para as próximas análises;
d) filtro de cobertura – a cobertura mínima por arquivo de alinhamento. Como a ferramenta *sider* foi desenvolvida para analisar um vasto número de indivíduos (arquivos de alinhamento) conjuntamente, podem-se validar os agrupamentos segundo uma quantidade mínima de leituras participantes provindas de cada arquivo. Essa filtragem visa remover aqueles grupos de baixa cobertura por indivíduo - que podem significar apenas um artefato. Os grupos que falharem neste filtro passam por um processo de reagrupamento com DBSCAN, no qual, primeiramente, são removidas as leituras pertinentes aos indivíduos com menos leituras que o mínimo requerido. As leituras que

[30] *Browser Extensible Data* (BED).

sobrarem são reagrupadas novamente e, se validadas pelo DBSCAN como grupos válidos, são consideradas aprovadas.

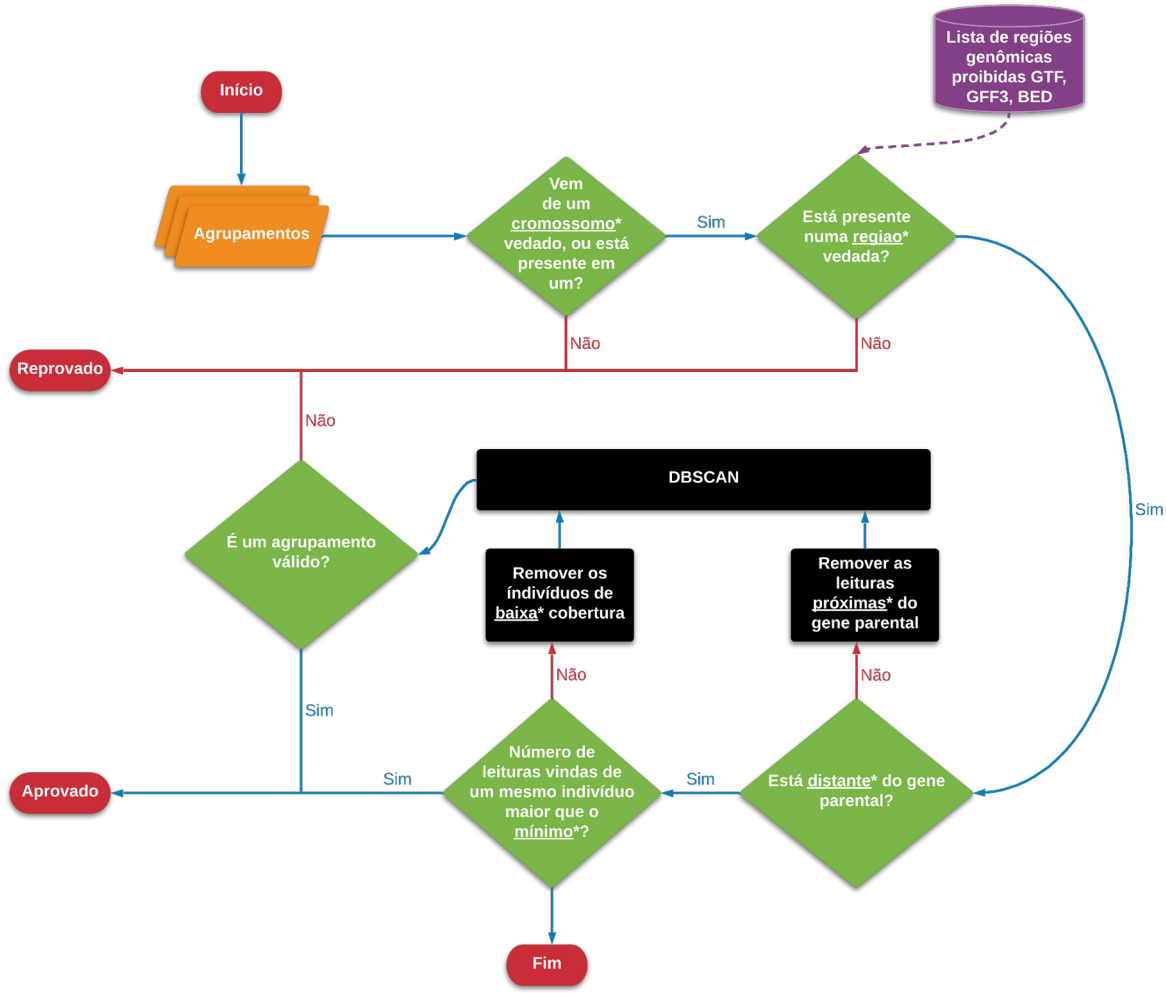
Figura 22 - O arquivo com as regiões genômicas proibidas é anotado na tabela *blacklist*.



Fonte: autoria própria.
A tabela *overlapping_blacklist* registra as sobreposições entre as regiões proibidas e os agrupamentos.

O campo *filter* da tabela *cluster* é por fim atualizado com os valores obtidos. Cada filtro é representado por um *bit* (Tabela 10). A presença de um *bit* específico indica que o grupo foi aprovado nesse critério. Os agrupamentos que possuírem todos os *bits* seguirão para os próximos passos.

Fluxograma 9 - Filtros de agrupamento.



Fonte: autoria própria.

São quatro os filtros desta etapa: 1) filtro de cromossomo - remove agrupamentos presentes, ou com o gene parental localizado, em determinados cromossomos; 2) filtro de região - remove agrupamentos em regiões genômicas proibidas - determinadas através de um arquivo GTF, GFF3 ou BED; 3) filtro de distância - distância mínima que precisa haver entre o gene parental e seu agrupamento, quando este está no cromossomo de origem daquele; 4) filtro de cobertura - o número mínimo de leituras por indivíduo para um dado agrupamento.

Se a filtragem falhar nos filtros 3 e 4, ocorre um reagrupamento com DBSCAN.

* Valores padrão que podem ser alterados pelo usuário: lista de cromossomos vedados; regiões genômicas proibidas; distância do gene parental e cobertura mínima por indivíduo.

Tabela 10 - Sumário dos *bits* dos filtros de agrupamento.

Filtro	Descrição	*Bit*	
		Hexadecimal	Decimal
-	Nenhum	0x1	1
Cromossomo	Cromossomo vedado	0x2	2
Distância	Distância do gene parental	0x4	4
Região	Regiões genômicas proibidas	0x8	8
Cobertura	Cobertura mínima por indivíduo	0x10	16

Fonte: sideRETRO versão 1.0.0.

Tabela 11 - Opções para os parâmetros do filtro de agrupamento.

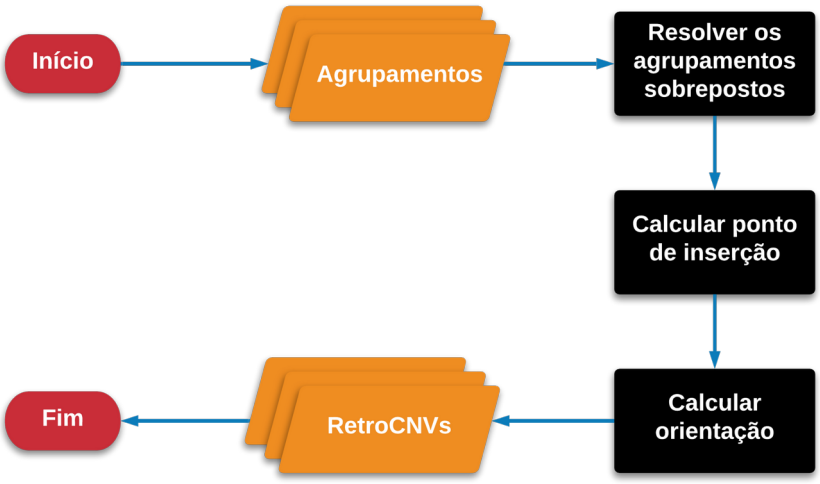
Filtro	Valor padrão
Lista de cromossomos vedados	Cromossomo mitocondrial
Regiões genômicas proibidas	GTF, GFF3 e BED
Distância do gene parental	1.000.000 bases
Cobertura mínima por indivíduo	1 leitura

Fonte: sideRETRO versão 1.0.0.

3.2.3.4 Anotação de retroCNVs

Os agrupamentos encontrados pelo algoritmo DBSCAN e que passaram por todos os crivos de filtragem seguem, por conseguinte, para esta etapa, na qual serão anotados como retroCNVs (Fluxograma 10).

Fluxograma 10 - Anotação de retroCNVs.



Fonte: autoria própria.
Os agrupamentos advindos das etapas de agrupamento e filtragem são anotados como retroCNVs. Primeiramente são resolvidas as sobreposições entre os grupos, então é calculado o ponto de inserção do evento. Por fim, calcula-se a orientação da inserção.

A anotação pode ser dividida em três partes:

a) resolução de agrupamentos sobrepostos;
b) cálculo do ponto de inserção;

c) cálculo da orientação.

3.2.3.4.1 Resolução de agrupamentos sobrepostos

Pode ocorrer até aqui de que haja agrupamentos que se sobreponham no mesmo cromossomo. No advento deste fenômeno, há que se analisar o contexto - de que forma se comportam os genes parentais dos grupos justapostos (Tabela 12 e Fluxograma 11):

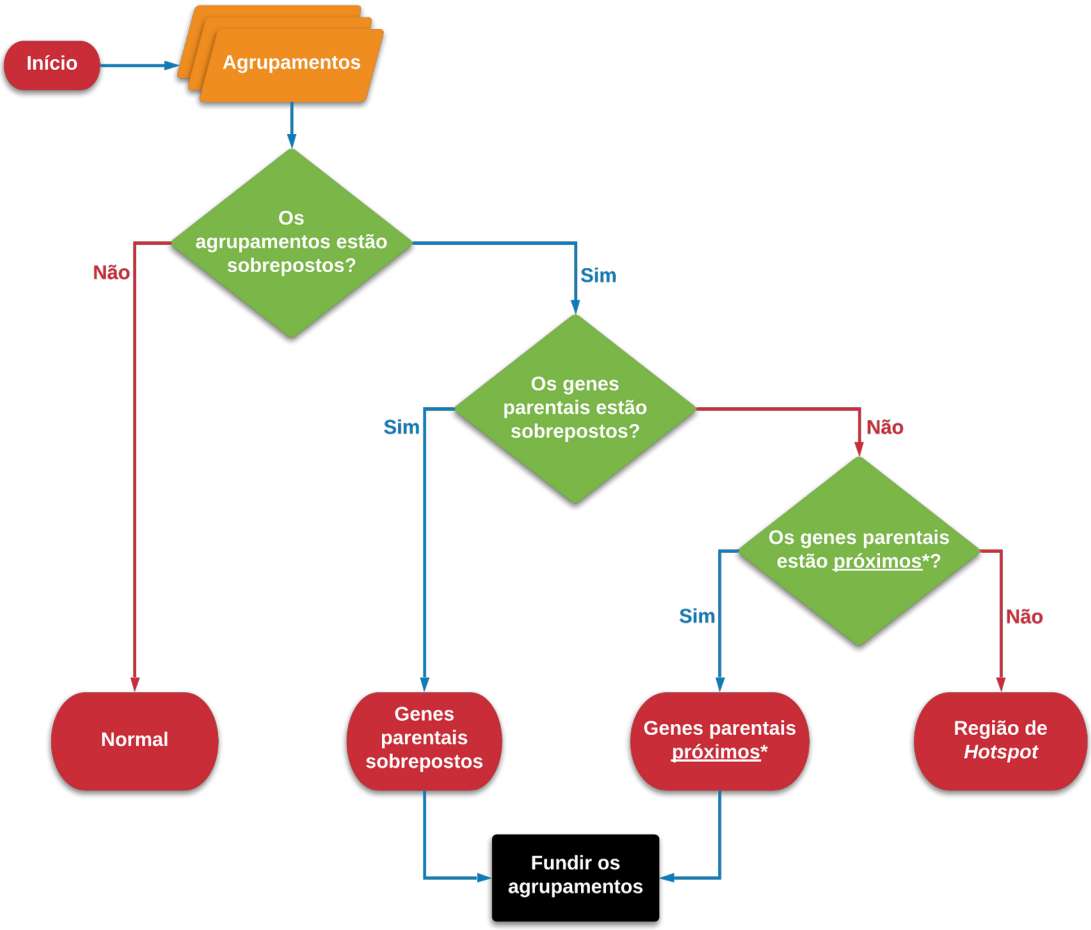
a) eles podem também estar sobrepostos;
b) eles podem estar próximos;
c) para os demais cenários, os agrupamentos devem representar uma região de *hotspot* (SULTANA et al., 2019).

Genes que se sobrepõem no genoma (genes senso-antisenso, por exemplo) compartilham entre si uma considerável quantidade de informação gênica, por isso, pode acontecer de o alinhador atribuir erroneamente leituras de um para o outro (FIRTH; BROWN, 2005; GALANTE et al., 2007). Agrupamentos justapostos podem, então, surgir exatamente por conta disso - dos seus genes parentais estarem sobrepostos, o que pode ser um fator confundidor para o alinhador. Nesse tipo de caso, não é possível determinar qual dos genes é o parental, ou ainda se cada um deles de fato foi retroduplicado para a mesma região gênica. Então os agrupamentos são fundidos numa única retroCNV e o gene parental dela se torna a junção dos nomes dos parentais anteriores. Um *bit* é usado no banco de dados do sideRETRO para marcar a retroCNV como oriunda de genes parentais sobrepostos (Tabela 12).

Se os genes parentais dos agrupamentos justapostos não se sobrepõem, mas estão próximos, então os grupos, assim como no caso de sobreposição dos parentais, são fundidos numa única retroCNV, cujo gene parental é a concatenação dos nomes dos genes próximos, e recebe o *bit* que a classifica como sendo advinda desses genes em questão. Aqui é o termo “próximos” se refere próximo em relação à uma distância ranqueada, e não em bases (Tabela 13). Dado um cromossomo, se

se classificar o primeiro gene codificador a montante como 1, o seguinte como 2 e assim sucessivamente até o último, ter-se-á o ranqueamento desses genes e assim é possível calcular a distância entre eles segundo a posição que ocupam. O critério de próximos, segundo uma distância ranqueada, é usado por haver no genoma genes duplicados em tandem, de modo que o alinhador pode equivocar-se em relação à origem de uma leitura provinda de um desses genes (PAN; ZHANG, 2008). Portanto, não há como garantir qual dos genes codificadores próximos, segundo uma distância ranqueada, é o parental da retroCNV em questão.

Fluxograma 11 - Resolução de agrupamentos sobrepostos.



Fonte: autoria própria.

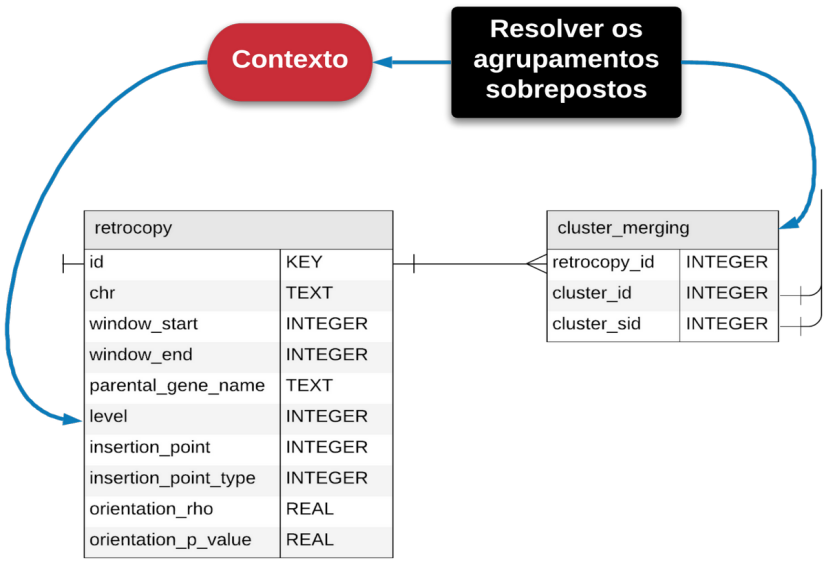
Se houver grupos sobrepostos, verificam-se se os genes parentais também se sobrepõem, ou se estão próximos - segundo uma distância genômica ranqueada. Se sim, os agrupamentos são fundidos numa única retroCNV. Um bit associado ao contexto dos genes parentais é atribuído à retroCNV. Os contextos verificados são: normal (não há agrupamentos sobrepostos); genes parentais sobrepostos; genes parentais próximos; e região de hotspot (genes parentais não se sobrepõem e tampouco estão próximos).

* Valor padrão que pode ser alterado pelo usuário: genes parentais próximos segundo uma distância ranqueada.

Os agrupamentos justapostos que não possuem genes parentais sobrepostos ou próximos não são fundidos, portanto são considerados cada qual uma retroCNVs distinta que foi inserida numa possível região de *hotspot* - similar ao que ocorre com as regiões de *hotspot* de L1 (SULTANA et al., 2019). Todas as retroCNVs sobrepostas recebem um *bit* (Tabela 12) para identificá-las como provenientes de uma região de *hotspot*.

Os dados gerados pela resolução de agrupamentos sobrepostos são computados no banco de dados do sideRETRO (Figura 23). As tabelas preenchidas são *retrocopy* e *cluster_merging* - a tabela *cluster_merging* conecta os agrupamentos presentes na tabela *cluster* com a *retrocopy*, que guarda as anotações concernentes às retroCNVs. Até aqui, na tabela *retrocopy*, são populadas as informações de posição genômica e de contexto de resolução dos grupos justapostos - quando houver sobreposição de agrupamentos, o início e o fim da janela de inserção da retroCNV passarão a ser o ínicio do elemento mais a montante e o fim do elemento mais a jusante. Ademais, as informações que faltam completar são as de ponto de inserção e orientação, que serão calculadas nas próximas etapas.

Figura 23 - Armazenamento dos agrupamentos sobrepostos.



Fonte: autoria própria.
Os resultados obtidos na resolução de agrupamentos sobrepostos são armazenados nas tabelas *cluster_merging* e *retrocopy*. A primeira contém a relação entre as tabelas *cluster* e *retrocopy*, enquanto a segunda contém as anotações das retroCNVs.

Tabela 12 - Sumário dos *bits* relativos ao contexto dos genes parentais dos agrupamentos sobrepostos.

Contexto	Descrição	Os agrupamentos são fundidos?	*Bit*	
			Hexadecimal	Decimal
Normal	Sem sobreposição entre os agrupamentos que formam as retroCNVs	Não	0x1	1
Genes parentais sobrepostos	Os genes parentais dos agrupamentos justapostos estão sobrepostos	Sim	0x2	2
Genes parentais próximos	Os genes parentais dos agrupamentos justapostos estão próximos, segundo uma distância ranqueada	Sim	0x4	4
Região de hotspot	Cada agrupamento representa uma retroCNV distinta que foi inserida na mesma região genômica	Não	0x8	8

Fonte: sideRETRO versão 1.0.0.

Tabela 13 - Opção para a resolução de agrupamentos sobrepostos.

Contexto	Valor padrão
Genes parentais próximos segundo uma distância ranqueada	3

Fonte: sideRETRO versão 1.0.0.

3.2.3.4.2 Cálculo do ponto de inserção

As retroCNVs começaram a ser anotadas no passo anterior, o passo de resolução de agrupamentos sobrepostos. Têm-se, até então, as informações concernentes ao gene parental e à janela de inserção da retroduplicação, com o cromossomo e as posições de início e fim. Para se determinar, dentro desta janela, o ponto exato de inserção do evento, há que se remeter às leituras dos alinhamentos suplementares presentes no agrupamento - caso haja alguma. O alinhamento suplementar tem como característica o fato de ser advindo de uma leitura dividida, a qual parte foi mapeada numa região genômica e parte noutra - recebendo uma delas do alinhador o *bit* 0x800, que a classifica como sendo a parte suplementar (como explicado anteriormente). Para alinhamentos anormais, isso significa que uma fração

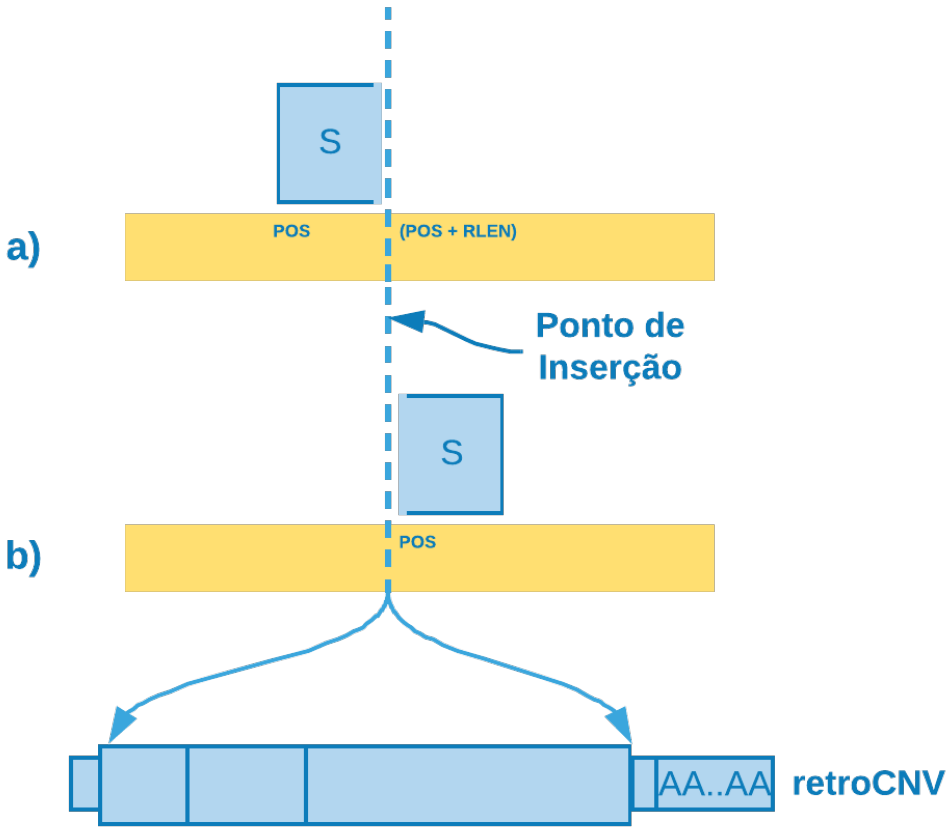
da leitura foi alinhada dentro da janela de inserção, enquanto a outra foi alinhada num éxon do gene parental. Esse tipo de evento ocorre por conta de parte do fragmento sequenciado cobrir tanto a retrocópia, quanto o cromossomo, onde se inseriu, e como essa retrocópia não está presente no genoma de referência, o alinhador acaba por atribuir a sequência localizada no elemento móvel ao éxon do gene, ficando a sequência do cromossomo mapeada na região do evento. A parte pertinente a região do evento é um importante padrão, pois a posição onde ocorreu a quebra da leitura coincide com o ponto de inserção. Sendo assim, o método para se calcular a posição de fixação da retroCNV leva em consideração o quarto e o sexto campos do arquivo de alinhamento no formato SAM - posição de mapeamento da primeira operação CIGAR e as operações CIGAR respectivamente (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021a, p. 6–8) - para os alinhamentos suplementares dentro da janela. As operações CIGAR para o alinhamento suplementar devem mostrar um padrão de *soft clipping* (S), ou *hard clipping* (H), sucedendo ou precedendo o *match* (M) (Quadro 2 e Figura 24). Se o M preceder o *clipping*, então o ponto de inserção deve estar localizado na posição após o consumo da operação de *match*. Do contrário, se a operação M suceder alguma operação de *clipping*, então o ponto de inserção deve coincidir com a posição da primeira operação CIGAR. Em suma, se nas operações CIGAR houver um padrão de expressão regular (THOMPSON, 1968) do tipo *[0-9]+M[0-9]+[SH]*, então o ponto de inserção é calculado pela soma da posição da primeira operação CIGAR, quarta coluna do arquivo SAM, e do número de bases mapeadas na operação de *match*; e se o padrão for *[0-9]+[SH][0-9]+M*, então usa-se o próprio valor da quarta coluna (Fluxograma 12).

Quadro 2 - Cálculo do ponto de inserção por alinhamento suplementar.

Expressão regular das operações CIGAR	Cálculo do ponto de inserção
*[0-9]+**M**[0-9]+[**SH**]*	POS* + RLEN**
*[0-9]+[**SH**][0-9]+**M***	POS*

Fonte: autoria própria.
Cálculo usando expressão regular nas operações CIGAR.
* Posição da primeira operação CIGAR;
** Número de bases mapeadas na operação CIGAR de *match*.

Figura 24 - Cálculo do ponto de inserção, segundo a posição, anterior ou posterior, do alinhamento suplementar.



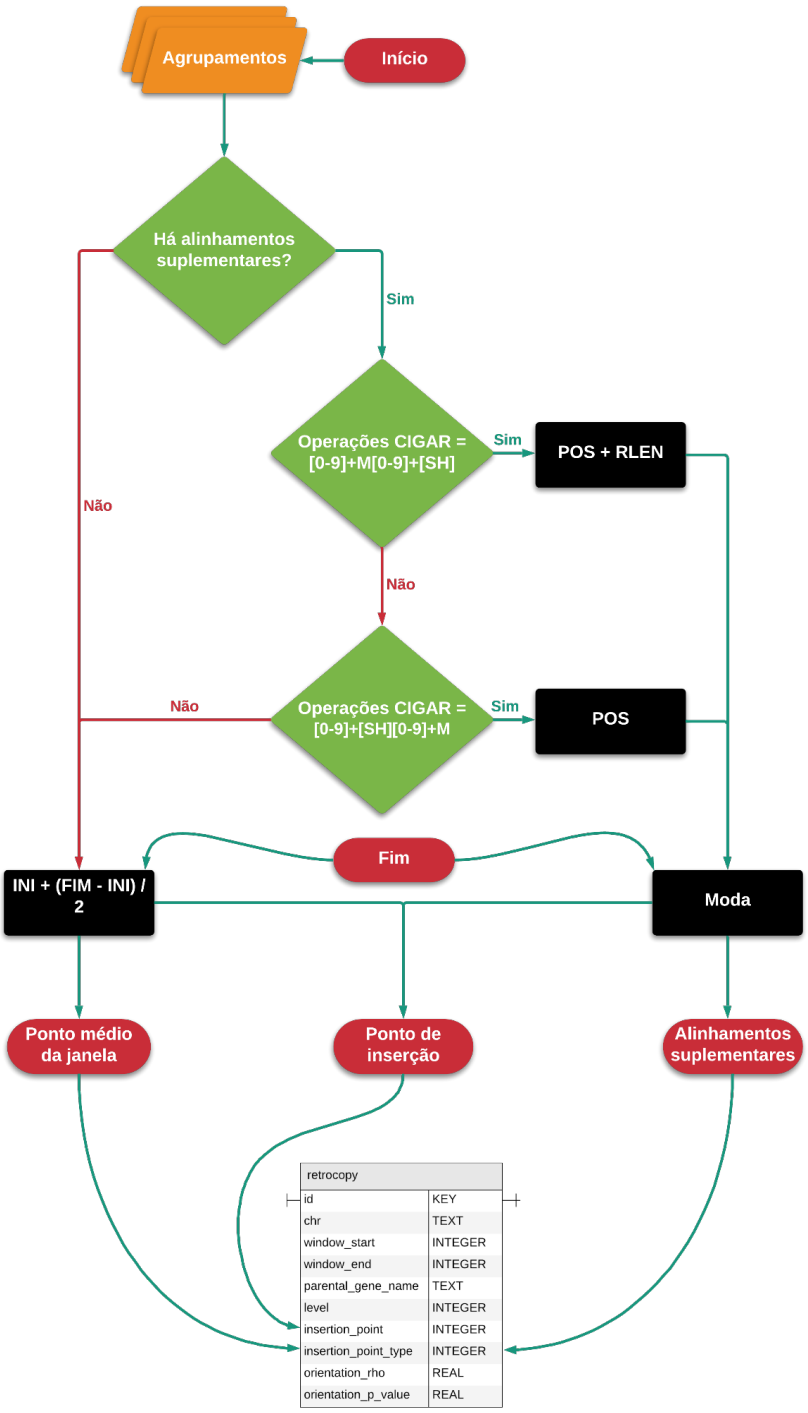
Fonte: autoria própria.
a) alinhamento anterior – soma-se a posição da primeira operação CIGAR ao número bases mapeadas na operação CIGAR de *match*; b) alinhamento posterior – a posição da primeira operação CIGAR coincide com o ponto de inserção.
POS = posição da primeira operação CIGAR; RLEN = número de bases mapeadas na operação CIGAR de *match*; S = alinhamento suplementar.

Tabela 14 - *Bits* usados pelo sideRETRO para identificar a forma como se calculou o ponto de inserção da retroCNV.

Descrição	*Bit*	
	Hexadecimal	Decimal
Ponto médio da janela	0x1	1
Alinhamentos suplementares	0x2	2

Fonte: sideRETRO versão 1.0.0.

Fluxograma 12 - Cálculo do ponto de inserção.



Fonte: autoria própria.

Se para um dado agrupamento houver alinhamento suplementar, então é possível calcular o ponto de inserção, segundo dois padrões de operações CIGAR. Se mais de um alinhamento suplementar estiver no agrupamento, então a moda dos pontos de inserção é usada. No caso de não haver alinhamentos suplementares, então o ponto médio da janela de inserção é usado. O ponto de inserção e o modo como este foi calculado são anotados na tabela *retrocopy* do banco de dados do sideRETRO.

POS = posição da primeira operação CIGAR; RLEN = número bases mapeadas na operação CIGAR de match; INI = posição de início da janela de inserção da retroCNV; FIM = posição final da janela de inserção da retroCNV.

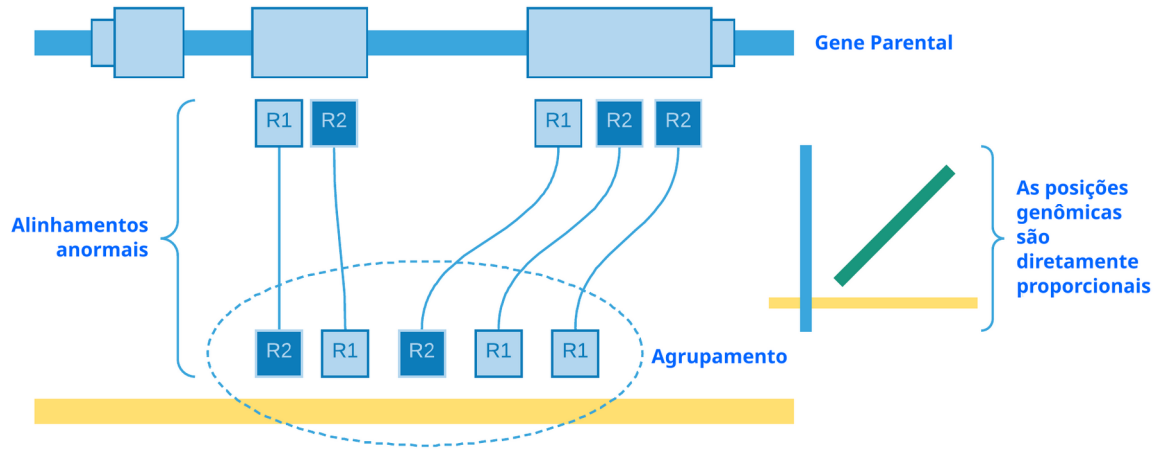
Para cada alinhamento suplementar dentro da janela, calcula-se o ponto de inserção, gerando um vetor de resultados. A partir desse vetor, a posição que tiver a maior moda representará o ponto de integração da retroCNV. Um outro cenário possível é que não haja alinhamentos suplementares mapeados na janela. Quando isso ocorre, não é possível calcular com precisão o ponto de inserção, então é usado o ponto médio do intervalo da janela.

O ponto de inserção é anotado na tabela *retrocopy* no campo *insertion_point*, assim como é atribuído um *bit* ao campo *insertion_point_type*, de maneira a identificar o modo de cálculo dessa posição. Por conseguinte, há duas possibilidades (Tabela 14), uma para alinhamentos suplementares e outra para o ponto médio da janela.

3.2.3.4.3 Cálculo da orientação

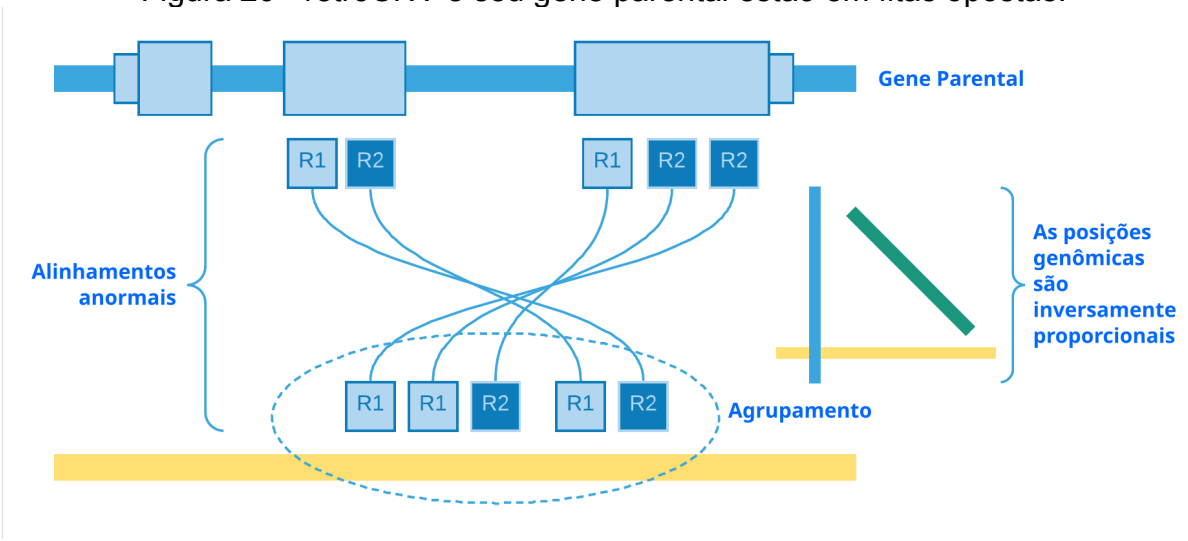
Outra informação importante, que pode ser obtida a partir dos dados, é a orientação da retroCNV em relação ao seu gene parental. As leituras de alinhamento anormal dão a pista para resolver esse problema. Foram capturadas as leituras pareadas quando uma das leituras se alinhou com um éxon de um gene codificador e seu par se alinhou com alguma outra região genômica. Então, podem-se ordenar em ordem crescente as leituras do sítio exônico - segundo as suas posições genômicas no cromossomo - e verificar, após isso, se seus pares estarão ordenados em ordem crescente ou decrescente na outra região genômica (Figuras 25 e 26). Como resultado, se se observar que eles são diretamente proporcionais, pode-se assumir que a retroCNV está na mesma fita do gene parental, caso contrário, eles estão em fitas opostas.

Figura 25 - retroCNV e seu gene parental estão na mesma fita.



Fonte: autoria própria.

Figura 26 - retroCNV e seu gene parental estão em fitas opostas.



Fonte: autoria própria.

Portanto, é usado o coeficiente de correlação de postos de *Spearman* (FIELLER; HARTLEY; PEARSON, 1957) para ter uma medida da relação entre as leituras do éxon e seus pares no sítio do agrupamento. Como os dados não são paramétricos, o *rho* de *Spearman* pode avaliar a relação monotônica, ou seja, pode dizer se a posição genômica das leituras do éxon aumenta quando a posição genômica de seus pares no local do agrupamento aumenta (*rho* positivo) - ou o contrário (*rho* negativo). Portanto, estando o gene parental na fita líder (+), um *rho* positivo indicaria uma retroCNV na fita positiva e um *rho* negativo indicaria na fita negativa. Seguindo esse mesmo raciocínio, estando o gene parental na fita

retardada (-), um *rho* positivo indicaria uma retroCNV na fita negativa, enquanto um *rho* negativo indicaria uma retroCNV na fita positiva (Quadro 3 e Fluxograma 13).

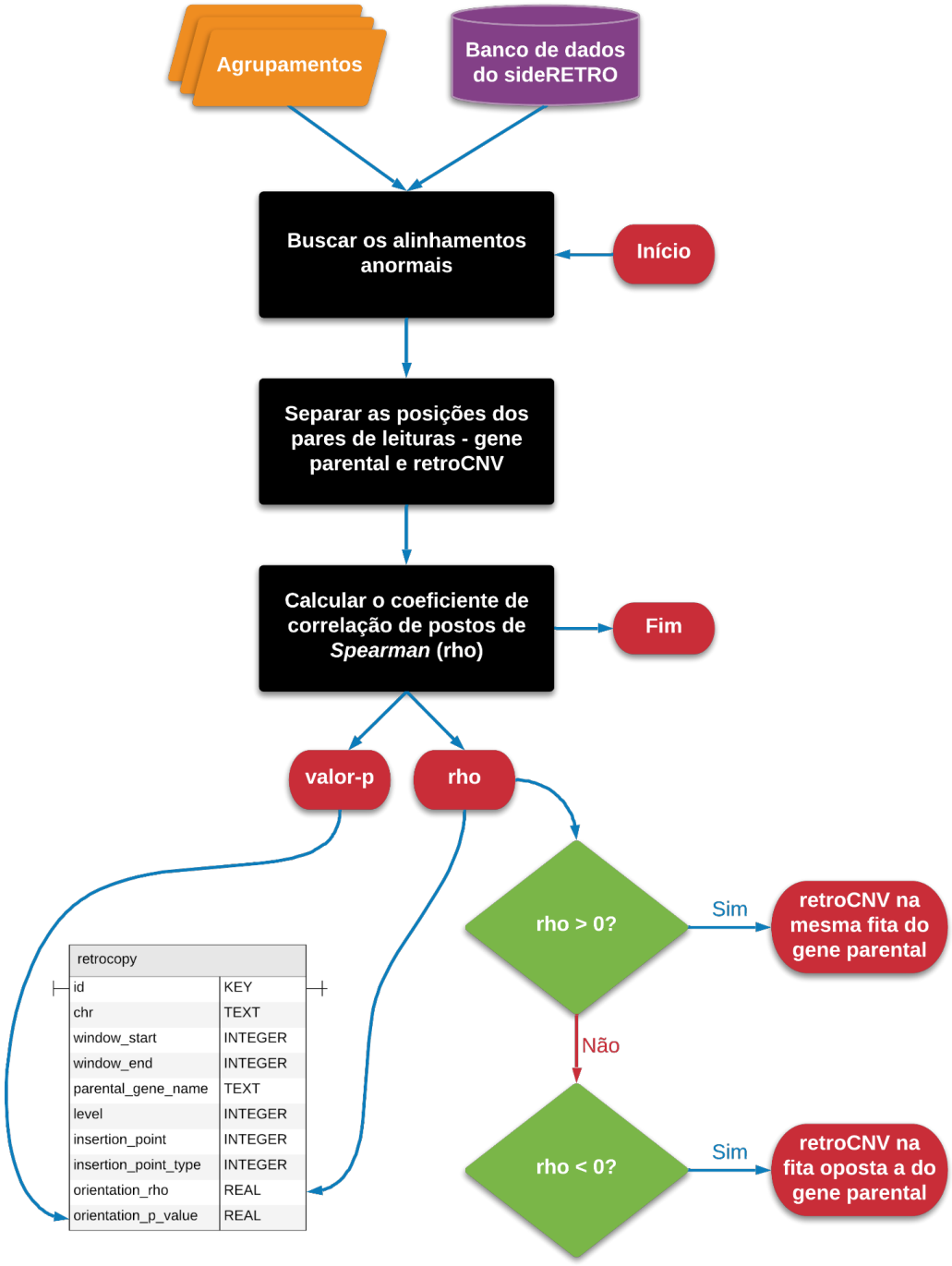
Quadro 3 - Sumário da regra usada para se determinar a fita, na qual se inseriu a retroCNV.

Fita do gene parental	Fita da retroCNV	
	rho > 0	*rho* < 0
+	+	-
-	-	+

Fonte: autoria própria.

Por conseguinte, o algoritmo do sideRETRO busca os alinhamentos anormais que formam o agrupamento da retroCNV. As posições genômicas das leituras mapeadas no gene parental e das leituras mapeadas na retroCNV são inseridas cada qual num vetor - respeitando a ordem, de modo que a posição de uma leitura no índice *i* de um vetor, terá sua leitura pareada no índice *i* do outro vetor. Então ambos os vetores são usados para se calcular o *rho* de *Spearman* e o valor-p, que são anotados no banco de dados do sideRETRO, na tabela *retrocopy* - campos *orientation_rho* e *orientation_p_value*, respectivamente (Fluxograma 13).

Fluxograma 13 - O algoritmo de cálculo da orientação da retroCNV.



Fonte: autoria própria.

Os alinhamentos anormais pertinentes a retroCNV são separados em dois vetores, contendo as posições das leituras do gene parental num, e as posições das leituras da retroCNV noutro. Os vetores são aplicados na cálculo do coeficiente de correlação de postos de *Spearman*; o *rho* e seu valor-p são anotados na tabela *retrocopy*, nos campos *orientation_rho* e *orientation_p_value*.

3.2.3.4.4 Genotipagem

A última parte do subcomando *merge-call* é a genotipagem das retroCNVs previamente anotadas no banco de dados. Aqui são verificados quais indivíduos - arquivos de alinhamento (SAM, BAM ou CRAM) - anotados na tabela *source*, possuem a variação estrutural e com qual zigosidade (em heterozigose ou em homozigose).

Há três possibilidades para sítios bialélicos (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021b, p. 6): se A é o alelo de referência e B é o alelo alternativo, a ordem dos genótipos para as probabilidades é AA, AB, BB. As probabilidades, por sua vez, são calculadas com base no artigo de Heng Li (2011): supondo-se que, em um determinado ponto de inserção de retrotransposição, haja *k* leituras. Sejam as primeiras *l* leituras idênticas ao genoma de referência e o restante seja diferente. A probabilidade de erro de alinhamento da *j*-ésima leitura é ϵ_j. Presumindo independência de erro, pode-se deduzir que:

$$\delta(g)=\frac{1}{m^k}\prod_{j=1}^{l}\left[(m-g)\epsilon_j+g(1-\epsilon_j)\right]\prod_{j=l+1}^{k}\left[(m-g)(1-\epsilon_j)+g\epsilon_j\right] \quad (1)$$

onde:

$\delta(g)$	Probabilidade para um dado evento de retroduplicação;
m	Ploidia;
g	Genótipo (o número de alelos de referência).

Portanto, pode-se resumir a fórmula (1) para homozigoto referência (HOR) (2), heterozigoto (HET) (3) e homozigoto alternativo (HOA) (4):

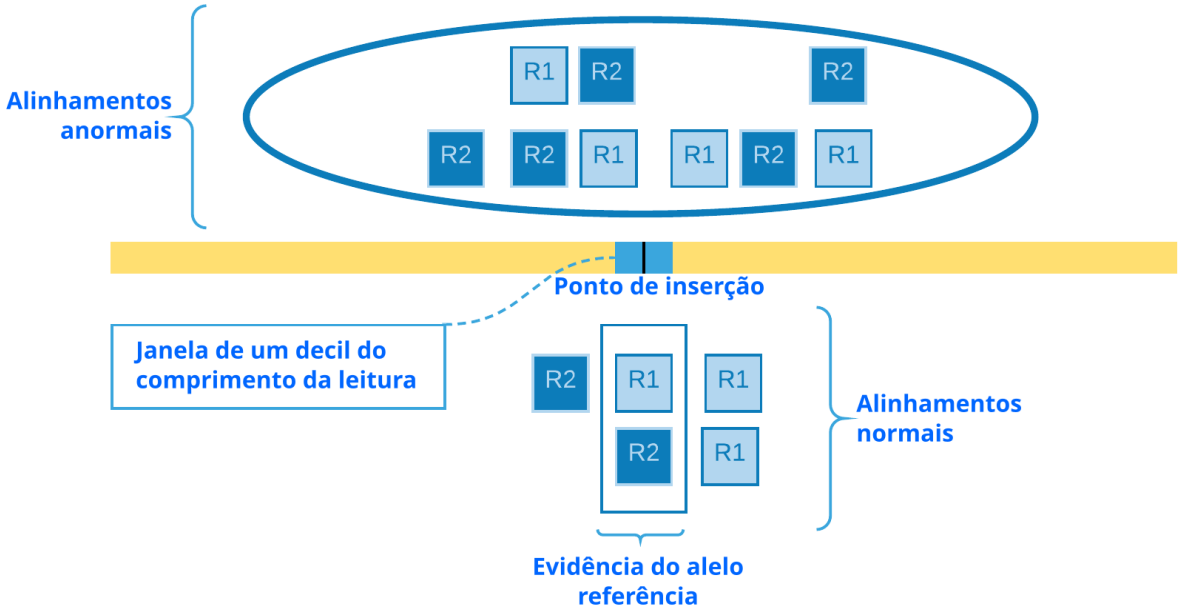
$$\delta(HOR)=\frac{1}{2^k}\prod_{j=1}^{l}2(1-\epsilon_j)\prod_{j=l+1}^{k}2\epsilon_j \quad (2)$$

$$\delta(HET)=\frac{1}{2^k} \quad (3)$$

$$\delta(HOA)=\frac{1}{2^k}\prod_{j=1}^{l}2\epsilon_j\prod_{j=l+1}^{k}2(1-\epsilon_j) \quad (4)$$

As leituras anormais pertencentes ao agrupamento que identifica a retroCNV serão usadas como o $k-l$ restante das leituras que diferem do genoma de referência (alelo alternativo) (Figura 27). Para verificar se há evidência do alelo de referência, precisa-se voltar ao arquivo de alinhamento (SAM, BAM ou CRAM) e verificar a presença de leituras normais que cruzem o ponto de inserção. De modo a mitigar o erro de alinhamento - que de outra forma superestimaria o número de leituras de alelos de referência - selecionam-se as leituras que cubram pelo menos uma janela de um decil do comprimento da leitura sobre o ponto de inserção, então chega-se às leituras idênticas ao genoma de referência (Figura 27). Aqui o usuário pode definir um nível de qualidade *Phred*, para remover alinhamentos normais de baixa qualidade.

Figura 27 - Alinhamentos anormais são usados como evidência do alelo alternativo.



Fonte: autoria própria.
Alinhamentos anormais do agrupamento, que evidencia a inserção da retroCNV, são usados como evidência do alelo alternativo. Os alinhamentos normais que cubram uma janela de um decil do tamanho da leitura em torno do ponto de inserção são usados como evidência do alelo referência.

Das leituras, é extraído o nível de qualidade *Phred*, quinta coluna do formato SAM (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021a, p. 6) e a partir dele, pode-se calcular a probabilidade de erro ϵ:

$$Q = -10 \log_{10} \epsilon \quad (5)$$

$$\epsilon = 10^{\frac{-Q}{10}} \quad (6)$$

onde:
Q Nível de qualidade *Phred*;
ϵ Probabilidade de erro de alinhamento.

As probabilidades de erro de alinhamento (6) para o alelo alternativo e o alelo referência são usadas para calcular as probabilidades do genótipo, segundo as equações de Heng Li supracitadas. Os resultados, para HOR, HET e HOA, são anotados na tabela *genotype*, do banco de dados do sideRETRO, nos campos *ho_ref_likelihood*, *he_likelihood* e *ho_alt_likelihood* - respectivamente. Também são anotados os números das leituras para o alelo referência (campo *reference_depth*) e o alelo alternativo (campo *alternate_depth*) (Fluxograma 14).

Enfim, o subcomando *merge-call* termina, tendo descoberto, anotado e genotipado retroCNVs. Essa quantidade de informação fica registrada no banco de dados SQL do sideRETRO, tendo sido preenchidas as tabelas *cluster, clustering*, *blacklist, overlapping_blacklist*, *cluster_merging*, *retrocopy* e *genotype*. A Tabela 15 sumariza os diferentes valores padrão usados no transcorrer das análises pelo *merge-call*.

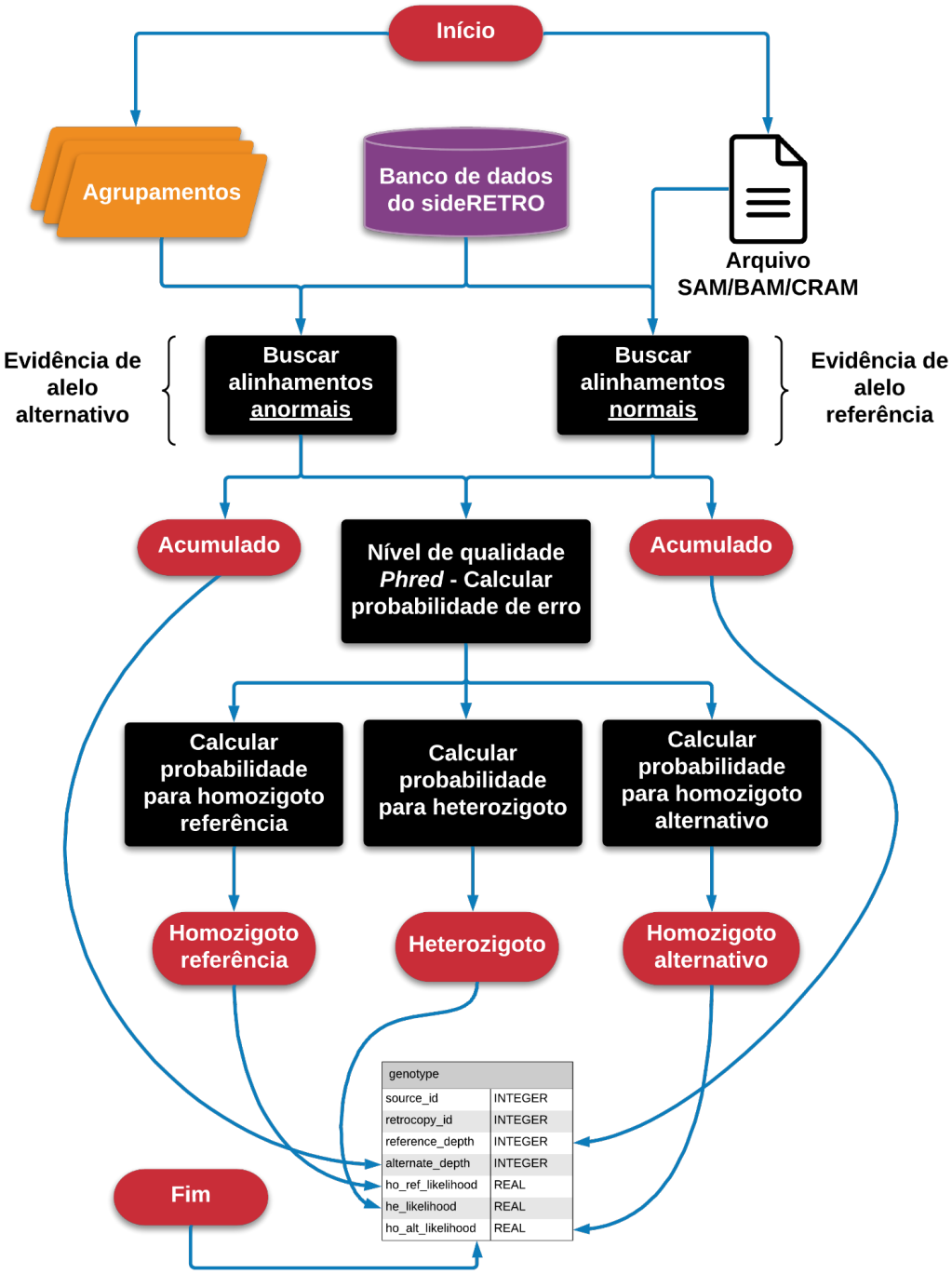
O próximo, e último passo do *sider*, é o subcomando *make-vcf*. Nele, serão organizadas as informações contidas no banco de dados num arquivo de formato VCF.

Tabela 15 - Sumário dos valores padrões para o subcomando *merge-call*.

Opção	Valor padrão
Distância épsilon ε	300 bases
Número mínimo de pontos	10 alinhamentos
Regiões genômicas proibidas	GTF, GFF3 e BED
Distância do gene parental	1.000.000 bases
Cobertura mínima por indivíduo	1 leitura
Genes parentais próximos	3
Lista de cromossomos vedados	Cromossomo mitocondrial
Nível de qualidade *Phred* mínimo requerido (alinhamentos normais)	8

Fonte: sideRETRO versão 1.0.0.

Fluxograma 14 - O algoritmo de genotipagem.



Fonte: autoria própria.

Os alinhamentos anormais da retroCNV e os alinhamentos normais que cubram o ponto de inserção da retroCNV são usados para a genotipagem de um dado indivíduo, arquivo de alinhamento. Dos alinhamentos é extraído o nível de qualidade *Phred* e com ele se calcula a probabilidade de erro de alinhamento. Então, aplicam-se as fórmulas para sítios bialélicos segundo o artigo de Heng Li, chegando-se às probabilidades de homozigoto referência, heterozigoto e homozigoto alternativo. Os valores obtidos são registrados no banco de dados do sideRETRO.

3.2.4 Subcomando *make-vcf*

O subcomando *make-vcf* é a última etapa de análise do sideRETRO. Após os alinhamentos anormais terem sido registrados pelo *process-sample* e as retroCNVs descobertas, anotadas e genotipadas durante o *merge-call*, é possível gerar, como resultado, um arquivo no formato VCF, contendo as informações da análise de forma organizada.

O arquivo VCF é escrito, seguindo as especificações da versão 4,2 (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021b). São adicionados campos específicos para eventos de retroCNV no cabeçalho do arquivo (Quadro 4), de modo que o usuário obtenha a conclusão do experimento com todos os detalhes gerados.

Quadro 4 - Descrição dos campos do arquivo VCF que são adicionados pelo sideRETRO.

(continua)

ID	Coluna	Descrição
INS:ME:RTC	*ALT*	Identifica o alelo alternativo como uma inserção de retroCNV.
SVTYPE	*INFO*	Tipo de variação estrutural. O valor usado é sempre *INS*, por se tratar de uma inserção de retroCNV.
IMPRECISE	*INFO*	Quando a *POS* não foi calculada através de leituras divididas, então usa-se o ponto médio da janela de inserção.
CIPOS	*INFO*	Intervalo de confiança em torno da *POS* para variações *IMPRECISE*, são 2 valores separados por vírgula: início da janela de inserção subtraída a *POS*, fim da janela de inserção subtraída a *POS*.
DP	*INFO*	Número total de leituras que cobrem o alelo alternativo.
SR	*INFO*	Número total de leituras divididas que estimam a *POS*.
ORHO	*INFO*	Coeficiente de correlação de postos de *Spearman* (*rho*) usado para estimar a polaridade da retroCNV.
POLARITY	*INFO*	Polaridade (fita) da retroCNV de acordo com o *rho*.
EXONIC	*INFO*	retroCNV exônica. IDs dos éxons separados por ‘/’.
INTRONIC	*INFO*	retroCNV intrônica: IDs dos íntrons separados por ‘/’
NEAR	*INFO*	retroCNV intragênica: IDs dos genes próximos da inserção.
PG	*INFO*	Nome do gene parental.

Quadro 4 - Descrição dos campos do arquivo VCF que são adicionados pelo sideRETRO.

(conclusão)

ID	Coluna	Descrição
PGTYPE	*INFO*	Tipo do gene parental: a) normal; b) genes parentais sobrepostos; c) genes parentais próximos; d) região de *hotspot*.
DP2	*FORMAT*	Número de leituras que cobrem o alelo referência e o alelo alternativo. São 2 valores separados por vírgula.
GT	*FORMAT*	Genotipagem: presença de 1 alelo alternativo (heterozigoto) ‘0/1’, de 2 alelos alternativos (homozigoto alternativo) ‘1/1’ e de nenhum alelo alternativo ‘0/0’ (homozigoto referência).
GL	*FORMAT*	Probabilidade da genotipagem, *genotype likelihood*. São 3 valores separados por vírgula: probabilidade para homozigoto referência, heterozigoto e homozigoto alternativo.

Fonte: (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021b, p. 4–6; sideRETRO versão 1.0.0).
POS = segunda coluna; *ALT* = quinta coluna; *INFO* = oitava coluna; *FORMAT* = nona coluna e adiante.

O sideRETRO, de modo a criar o arquivo VCF de saída, volta ao banco de dados já preenchido com a anotação das retroCNVs. As informações concernentes à posição genômica e ao gene parental são extraídas da tabela *retrocopy*:

a) *CHROM* (primeira coluna) – o cromossomo;
b) *POS* – o ponto de inserção da retroCNV;
c) na coluna *INFO*:
   - *PG* – o gene parental;
   - *PGTYPE* – o contexto relativo ao gene parental, podendo ser: normal (1), genes parentais sobrepostos (2), genes parentais próximos (4) ou região de hotspot (8);
   - se o ponto de inserção foi calculado pela média da janela de inserção, então inclui-se *IMPRECISE*;
   - para *IMPRECISE*, calcula-se o intervalo de confiança *CIPOS* com o início e fim da janela de inserção (ínicio da janela - *POS*, fim da janela - *POS*);
   - *ORHO* – se o valor-p do coeficiente de correlação de postos de *Spearman* for menor que o erro máximo tolerado, então se inclui o valor de *rho* (Tabela 16);

- *POLARITY* – se o *rho* for positivo, então a retoCNV está na mesma fita do gene parental, se for negativo, então está na fita oposta.

As informações de genotipagem são encontradas na tabela *genotype*:

a) a partir da coluna *FORMAT*, seguindo por todos os arquivos de alinhamento:
   - *DP2* – o número de leituras que cobrem o alelo referência e o alelo alternativo vem dos atributos da tabela *reference_depth* e *alternate_depth*, respectivamente;
   - *GL* – as probabilidades para homozigoto referência, heterozigoto e homozigoto alternativo estão, pela ordem, nos campos *ho_ref_likelihood*, *he_likelihood*, *ho_alt_likelihood*;
   - *GT* – a genotipagem. Se a maior probabilidade em *GL* for para homozigoto referência, então se adiciona o valor 0/0, se for para heterozigoto, então se adiciona 0/1, ou se for para homozigoto alternativo, 1/1.
b) na coluna *INFO*:
   - *DP* – o número total de leituras que cobrem o alelo alternativo. É a somatória de todos os valores para o atributo da tabela *alternate_depth*.

O valor para o número de leituras divididas, *SR*, na coluna *INFO* é obtido pela somatória de todos os alinhamentos anormais, na tabela alignment, pertencentes ao agrupamento da retroCNV e que possuem o *bit* de alinhamento suplementar e um dos padrões de expressão regular para as operações CIGAR - *[0-9]+M[0-9]+[SH]* ou *[0-9]+[SH][0-9]+M*.

O contexto de inserção da retroCNV, se é uma retrocópia intragênica (*EXONIC*, *INTRONIC*) ou intergênica próxima (Tabela 16) de algum gene (*NEAR*) na coluna *INFO*, é obtido a partir do cruzamento das posições das retroCNVs na tabela *retrocopy* e das posições dos éxons dos genes codificadores na tabela *exon*. Os valores *EXONIC*, *INTRONIC* e *NEAR* representam listas separadas por ‘/’, contendo os nomes dos genes, aos quais a retroCNV se relaciona. Um mesmo evento pode ser exônico (*EXONIC*) para certos genes e intrônico (*INTRONIC*) para outros. Os genes próximos à retroCNV só são adicionados no campo *NEAR*, se esta for intergênica.

Opcionalmente, se o genoma de referência - no formato FASTA (NATIONAL LIBRARY OF MEDICINE, [s.d.]) - tiver sido passado, a quarta coluna *REF*, com a

base nitrogenada do alelo referência do cromossomo *CHROM* e posição *POS*, é preenchida, assim como os campos *contig* do cabeçalho - com os nomes dos cromossomos e os seus comprimentos em número de bases.

Tabela 16 - Opções para a anotação do arquivo VCF.

Opções	Valor padrão
Erro máximo para valor-p do coeficiente de correlação de postos de *Spearman*	0,05
Distância próxima de algum gene	10.000 bases

Fonte: sideRETRO versão 1.0.0.

3.3 TESTANDO O SIDERETRO COM DADOS SIMULADOS

Com o intuito de averiguar o desempenho do sideRETRO num cenário cujas variáveis pudessem ser controladas, foi desenvolvida uma simulação computacional, na qual genomas com inserções de retroCNVs foram desenhados previamente e analisados posteriormente com o *sider*. A simulação pode, então, ser dividida em cinco etapas (Fluxograma 15):

a) desenho das retroCNVs;
b) desenho da coorte;
c) simulação dos genomas sequenciados;
d) alinhamento contra o genoma de referência;
e) análise com o sideRETRO.

3.3.1 Desenho das retroCNVs

Foram desenhadas 100 retroCNVs distintas, a partir do arquivo FASTA com as sequências dos transcritos maduros (isentos de regiões intrônicas) dos genes codificadores de proteínas, distribuído pelo GENCODE versão 32 (FRANKISH et al., 2019). Sortearam-se 100 genes parentais e, para cada gene selecionado, extraíram-

se as últimas (mais a jusante) 1000 bases do transcrito mais comprido para serem usadas como sequência da retrocópia.

O programa usado para desenhar as 100 retroCNVs, *make_rtc.pl* (Apêndice B), foi executado com os parâmetros: *--seed*=17, *--rtc_num*=100 e *–length*=1000.

3.3.2 Desenho da coorte

Para as 100 retroCNVs desenhadas, foram sorteadas as regiões genômicas de inserção - o cromossomo, a fita e o ponto de inserção - e a zigosidade - heterozigoto ou homozigoto alternativo. Os eventos foram aleatoriamente divididos em 3 grupos:

a) retroCNVs fixadas – 25 eventos;
b) retroCNVs polimórficas – 50 eventos;
c) retroCNVs somáticas – 25 eventos.

A coorte de teste que recebeu as retroCNVs projetadas foi composta de 100 indivíduos. O modo de distribuição estocástico dos eventos por indivíduo seguiu a seguinte regra:

a) as retroCNVs fixadas foram inseridas em todos os indivíduos da coorte;
b) pelo menos dois indivíduos receberam a mesma retroCNV polimórfica;
c) retroCNVs somáticas foram inseridas cada uma num único indivíduo da coorte.

O desenho da coorte se deu pelo programa *make_cohort.pl* (Apêndice C). Ele foi rodado com os parâmetros: *--cohort*=100 e *–seed*=17.

3.3.3 Simulação dos genomas sequenciados

Para cada indivíduo da coorte, foi simulado um sequenciamento genômico de acordo com as inserções de retroCNVs, projetadas nos passos anteriores. As simulações foram feitas com uma baixa cobertura de sequenciamento de 20 vezes, ou seja, eventos heterozigóticos tiveram apenas 10 vezes de cobertura. Essa estratégia permitiu verificar a capacidade do sideRETRO em identificar eventos mesmo em um “cenário não ideal” de baixa cobertura de sequenciamento.

A ferramenta usada para simular os sequenciamentos foi a *SANDY* versão 0,23 (MILLER, 2019). A *SANDY* indexou as devidas variações estruturais por indivíduo e gerou as simulações dos sequenciamentos no formato FASTQ (COCK et al., 2010). Ao final, obtiveram-se 100 sequenciamentos genômicos pareados, equivalentes a experimentos de WGS.

A *SANDY* foi rodada com os seguintes parâmetro:

a) etapa de indexação das variações estruturais – *sandy variation add:* --*structural-variation*=id-do-indivíduo;
b) simulação – sandy genome: --*id*='%i.%U_%c:%S-%E_%v', --*structural-variation*=id-do-indivíduo, --*jobs*=20, --*seed*=1, --*quality-profile*='hiseq_101' e –*coverage*=20.

3.3.4 Alinhamento contra o genoma de referência

Os sequenciamentos simulados foram alinhados contra o genoma de referência, GENCODE versão hg38 (FRANKISH et al., 2019). O alinhador escolhido foi o *BWA* versão 0.7.9 (LI; DURBIN, 2009) - por ser capaz de lidar com leituras divididas. O *BWA* rodou com o algoritmo *mem* e os parâmetros padrão - *bwa mem*, gerando arquivos de alinhamento no formato SAM.

3.3.5 Análise com o sideRETRO

Finalmente, rodou-se o sideRETRO, versão 0.14.1, sobre os sequenciamentos simulados alinhados. Os parâmetros usados para os subcomando *process-sample* e *merge-call* foram[31]:

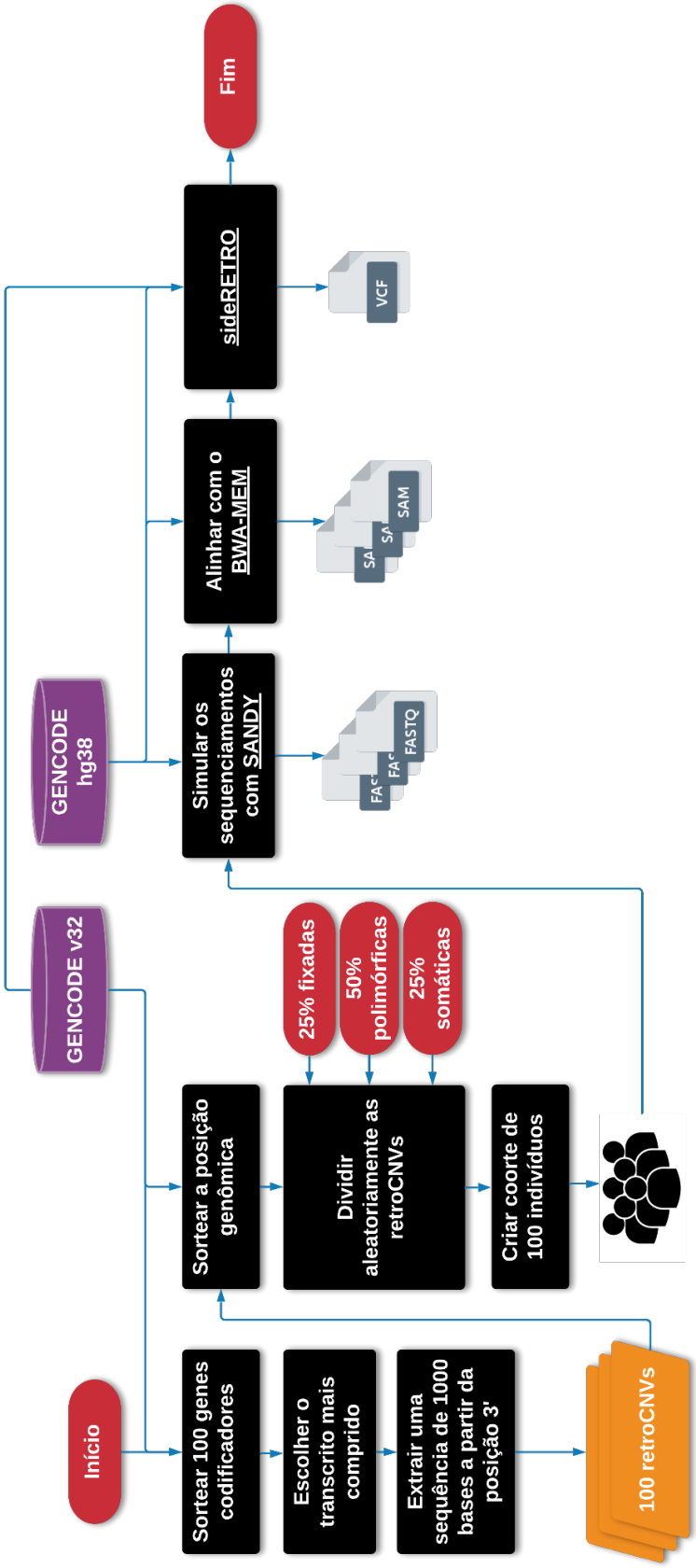
a) *sider process-sample: --alignment-frac*=0.9 e *--phred-quality*=20;
b) *sider merge-call*: *--epsilon*=500, *--min-pts*=10 e *–phred-quality*=20.

O genoma de referência usado foi o do GENCODE hg38 e a anotação do genoma também do projeto GENCODE versão 32.

Os *scripts* exemplificando todo o processo de desenho das retroCNVs, simulação dos genomas sequenciados e análise com o sideRETRO constam nos apêndices deste trabalho (Apêndice F).

[31] Ponto como separador decimal (para *LC_NUMERIC=en_US.UTF-8*).

Fluxograma 15 - O teste do sideRETRO com dados simulados.



Fonte: autoria própria.

3.4 TESTANDO O SIDERETRO COM DADOS REAIS

Para avaliar o desempenho do sideRETRO contra dados reais, foi selecionado o trabalho de Abyzov e colaboradores (2013). O método computacional desenvolvido e utilizado pelos autores conta com a utilização de leituras de junções exon-exon para a identificação de retroCNVs. Para aumentar a confiabilidade de seus candidatos, os autores realizaram validações experimentais (Quadro 5) por PCR para nove retroCNVs e, para seis delas, encontraram seus pontos de inserção com as coordenadas no genoma de referência.

Quadro 5 - As 9 retroCVNs validadas por PCR por Abyzov e colaboradores.

Gene parental	Ponto de inserção
LAPTM4B	Sim
TMEM66	Sim
BOD1	Não
SKA3	Sim
AP3S1	Não
CACNA1B	Sim
TDG	Sim
CBX3	Sim
MTCH2	Não

Fonte: dados extraídos da Tabela X (ABYZOV et al., 2013).

Essas retroCNVs foram genotipadas por Abyzov e colaboradores em 974 indivíduos de 14 populações do 1KGP (Quadro 6). É possível obter do 1KGP, os arquivos de alinhamento no formato CRAM para todos os indivíduos de todas as populações contempladas pelo projeto (EMBL-EBI, [s.d.]). Logo, foi rodado o sideRETRO a fim de se verificar se este seria capaz de detectar as seis retroCNVs, com o ponto de inserção, validadas por Abyzov et al. e genotipá-las nas mesmas 14 populações (Fluxograma 16).

Quadro 6 - As 14 populações do 1KGP que foram genotipadas por Abyzov e colaboradores.

Código da população	Número de indivíduos	Detalhe
ASW	50	Ancestralidade Africana no Sudoeste dos EUA
CEU	91	Residentes de Utah (CEPH) com ascendência do norte e da Europa Ocidental
CHB	81	Etnia Han em Pequim, China
CHS	92	Etnia Han no sul da China
CLM	52	Colombianos em Medellín, Colômbia
FIN	77	Finlandeses na Finlândia
GBR	72	Britânicos na Inglaterra e Escócia
IBS	6	Populações ibéricas na Espanha
JPT	80	Japoneses em Tóquio, Japão
LWK	83	Luhya em Webuye, Quênia
MXL	54	Ascendência mexicana em Los Angeles, Califórnia
PUR	53	Porto-riquenhos em Porto Rico
TSI	100	Toscani na Itália
YRI	83	Yoruba em Ibadan, Nigéria

Fonte: dados extraídos da Tabela S1. Material suplementar (ABYZOV et al., 2013).

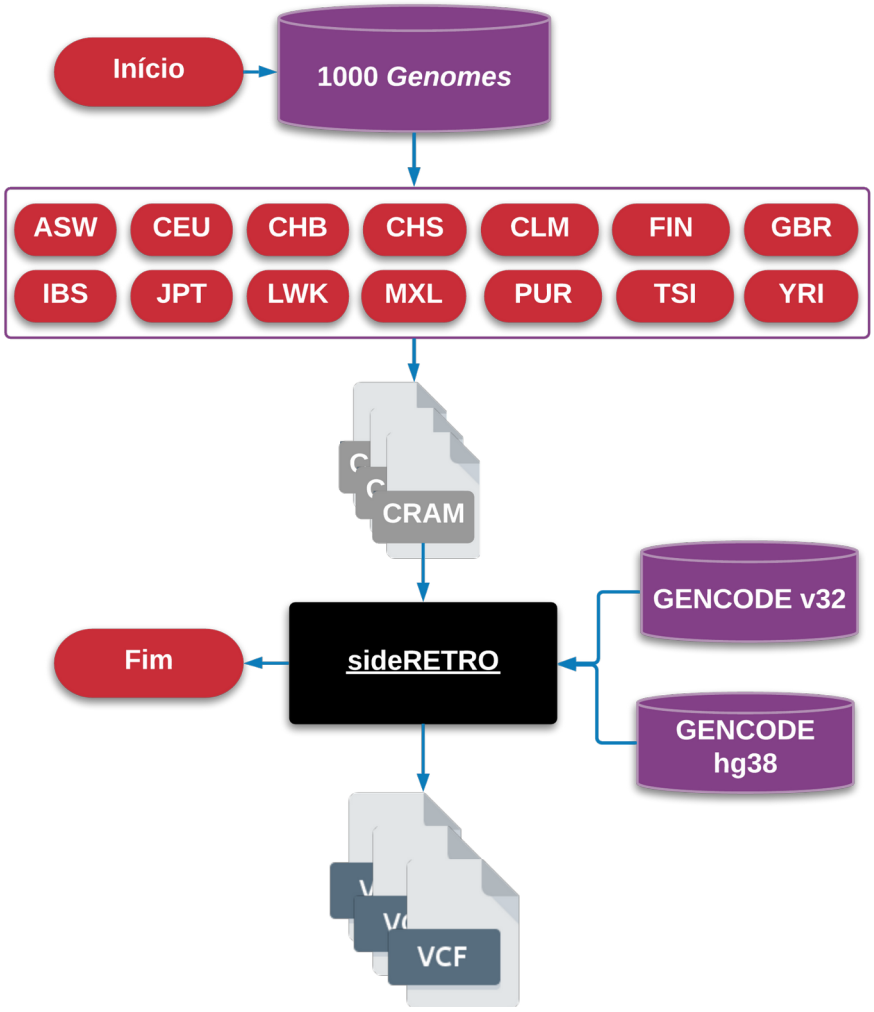
Os parâmetros usados pelo executável foram[32]:

a) subcomando *process-sample*: *--max-distance*=15000, *--alignment-frac*=0.9 e *--phred-quality*=20;
b) subcomando *merge-call*: *--epsilon*=500, *--min-pts*=20, *--genotype-support*=5, *--near-gene-rank*=3 e *–phred-quality*=20.

O genoma de referência e a anotação do genoma usados foram do GENCODE versão 32 (genoma de referência hg38).

[32] Ponto como separador decimal (para *LC_NUMERIC=en_US.UTF-8*).

Fluxograma 16 - O teste do sideRETRO com dados reais.



Fonte: autoria própria.
Análise para 14 populações do 1KGP a partir dos arquivos de alinhamento no formato CRAM disponibilizados pelo projeto.

4 Resultados

4 RESULTADOS

4.1 O SIDERETRO NUMA CASCA DE NOZ

4.1.1 Uma visão geral do sideRETRO

O sideRETRO é uma ferramenta de bioinformática dedicada à detecção de inserção polimórficas, germinativas ou somáticas de retrocópias (aqui chamadas de retroCNVs) em dados de Sequenciamento Completo de Genomas[33] (WGS) e também de Exoma[34] (WES). O programa foi escrito *ab initio* na linguagem de programação C e com o intuito de ser rodado em sistemas operacionais baseados em Linux (TORVALDS, 1997). O código-fonte é distribuído sob a Licença Pública Geral GNU versão 3[35] (GPLv3) (“GNU General Public License, version 3”, 2007). Além de detectar a mobilização de retrocópias, o sideRETRO anota várias outras características relacionadas ao evento de retroCNV (Quadro 7).

O sideRETRO usa dados NGS para identificar retrocópias não fixadas - polimórficas, germinativas ou somáticas - ausentes no genoma de referência, mas presentes no genoma sequenciado. A metodologia consiste em detectar alinhamentos anormais nos arquivos SAM, BAM ou CRAM e, com o algoritmo de aprendizado de máquina não supervisionado DBSCAN, agrupar essas leituras e descobrir inserções de retrocópias.

O código-fonte do sideRETRO compila para um executável chamado *sider*, o qual possui três subcomandos, cada qual para uma etapa consecutiva de análise:

a) *process-sample* – primeira etapa de análise na qual o sideRETRO identifica os alinhamentos anormais nos arquivos de alinhamento SAM, BAM ou CRAM. Os alinhamentos anormais são registrados no banco de dados;
b) *merge-call* – na etapa seguinte ao *process-sample*, os alinhamentos anormais registrados no banco de dados são agrupados segundo o algoritmo de

[33] *Whole Genome Sequencing* (WGS).
[34] *Whole Exome Sequencing* (WES).
[35] *General Public License version 3* (GPLv3).

aprendizado de máquina não supervisionado DBSCAN. Os agrupamentos encontrados são genotipados e registrados como retrocópias no banco de dados;

c) *make-vcf* – nesta última etapa, as retrocópias presentes no banco de dados são anotadas num arquivo de formato VCF.

Quadro 7 - Atributos da retroCNV detectados pelo sideRETRO.

Gene parental	Gene que passou pelo processo de retrotransposição, dando origem à retrocópia.
Posição genômica	A coordenada do genoma onde ocorreu a integração da retrocópia (cromossomo: início-fim). Inclui o ponto de inserção.
Fita	A orientação da inserção: se ela ocorreu na fita líder (+), ou na fita retardada (-).
Contexto genômico	O contexto do sítio de integração da retrócopia: se o evento de retrotransposição ocorreu em uma região intergênica ou intragênica - a última pode ser dividida em exônica e intrônica de acordo com o gene hospedeiro.
Genótipo	Quando vários indivíduos são analisados, são anotados os eventos para cada um deles. Dessa forma, é possível distinguir se um evento é exclusivo ou compartilhado entre a coorte.
Haplótipo	A ferramenta fornece informações sobre a ploidia do evento, ou seja, se ele ocorreu em um ou ambos os cromossomos homólogos (homozigoto ou heterozigoto).

Fonte: autoria própria.

4.1.2 Instalando o sideRETRO

O sideRETRO armazena seu código-fonte de forma pública no *github* (GITHUB, 2022) e usa o sistema de compilação *Meson* (“The Meson Build system”, [s.d.]) para gerenciar o processo de configuração e compilação nos sistemas operacionais baseados em Linux. O sistema de compilação *Meson* necessita da linguagem de programação interpretada *Python* (VAN ROSSUM; DRAKE, 2009), de um compilador de linguagem C (gcc (GCC TEAM, 2021), por exemplo) e do sistema de compilação *Ninja* (“The Ninja build system”, 2020) para funcionar.

O código-fonte da ferramenta pode ser obtido diretamente no sítio <https://github.com/galantelab/sideRETRO.git>, ou através do programa de versionamento *git* (CHACON; STRAUB, 2014):

git clone https://github.com/galantelab/sideRETRO.git *(1)*

Tendo sido obtidos os *scripts* do programa, pode-se seguir, então, ao processo de compilação e de instalação:

cd sideRETRO *(2)*
meson build *(3)*
ninja -C build *(4)*

O programa *git* irá baixar o código-fonte (1) para o diretório sideRETRO. Então, entra-se no diretório criado (2) e executa-se o sistema *Meson* (3). Será criado o diretório *build* (3) com as configurações necessárias do processo de compilação, os quais, por sua vez, serão usados pelo sistema *Ninja*. Logo, roda-se o comando *ninja* (4) sobre o diretório *build*. Isso dará início ao processo de compilação do código do sideRETRO e será criado o executável *sider*.

O executável *sider* poderá ser encontrado no diretório *build/src/*. Caso se queira instalar o *sider* num diretório padrão do sistema, pode-se executar o comando:

ninja -C build install *(5)*

Para instalar o sider (5) no diretório padrão será necessário ter permissões de superusuário.

4.1.3 Usando o sideRETRO: o pré-processamento dos dados de NGS

O *sider* analisa dados que já tenham sido previamente alinhados contra o genoma de referência. Os arquivos de sequenciamento gerados pelas tecnologias

NGS, FASTQs (COCK et al., 2010), precisam ser alinhados contra o genoma de referência para, então, poderem ser analisados pela primeira etapa do sideRETRO, o *process-sample*. A escolha do alinhador fica a critério do usuário, no entanto, alinhadores que sejam capazes de detectar as leituras divididas fornecerão mais informações à ferramenta e, por conseguinte, obterão resultados mais precisos - sobretudo com relação ao ponto de inserção da retroCNV.

Alguns exemplos de alinhadores que sejam capazes de detectar as leituras divididas:

a) *Bowtie* (LANGMEAD et al., 2009);
b) *BWA* - subcomando *mem* (LI; DURBIN, 2009);
c) *bwa-mem2* (VASIMUDDIN et al., 2019).

Tendo os dados de NGS sido alinhados contra o genoma de referência e, finalmente, os arquivos no formato SAM, BAM ou CRAM sido gerados, é chegada a hora de rodar o *sider* e descobrir novas inserções de retroCNV.

4.1.4 Usando o sideRETRO: *process-sample*

O passo seguinte ao pré-processamento é a execução do subcomando do *sider* - *process-sample*. O subcomando *process-sample* cria um banco de dados de leituras anormais de um conjunto de arquivos SAM, BAM ou CRAM. Para isso, existem algumas opções obrigatórias que o usuário deve fornecer para fazer uma busca correta. Chamar o comando *process-sample* sem nenhum argumento (1) dará uma ajuda específica na qual o usuário poderá conhecer todas as opções obrigatórias para este comando:

sider process-sample *(1)*

Obrigatoriamente, é necessário que o usuário indique um arquivo de anotação no formato GTF ou GFF3 e uma lista de arquivos de alinhamento SAM, BAM ou CRAM (2).

sider process-sample -a gencode.gtf ind1.bam ind2.bam ind3.bam *(2)*

Seguem as opções disponíveis (Quadros 8 e 9) do subcomando *process-sample*:

Quadro 8 - Opções obrigatórias para o *process-sample*.

Opção	Descrição
-a, --annotation-file	Anotação dos genes no genoma de referência no formato GTF ou GFF3. O sider procurará por éxon com o atributo *transcript_type=protein_coding*. Os atributos *gene_name*, *gene_id* e *exon_id* também são necessários.
-i, --input-file	Arquivo contendo uma lista separada por nova linha de arquivos de alinhamento no formato SAM, BAM ou CRAM. Esta opção não é obrigatória se um ou mais arquivos SAM, BAM ou CRAM forem passados como argumento. Se *input-file* e argumentos forem definidos concomitantemente, a união de todos os arquivos de alinhamento será usada.

Fonte: sideRETRO versão 1.0.0.

Quadro 9 - Demais opções para o *process-sample*.

(continua)

Opção	Descrição*
-h, --help	Mostra as opções de ajuda.
-q, --quiet	Diminui a verbosidade apenas para mensagens de erro ou suprime as saídas do terminal se *log-file* for passado.
--silent	O mesmo que *–quiet*.
-d, --debug	Aumente a verbosidade para o nível de depuração.
-l, --log-file	Imprime mensagens de *log* em um arquivo.
-o, --output-dir	Diretório de saída. Cria o diretório caso este não exista.
-p, --prefix	Prefixo para os arquivos de saída. Padrão: *out*.
-c, --cache-size	Define o tamanho do *cache* para o SQLite3 em KiB. Padrão: 200000.
-Q, --phred-quality	Qualidade mínima de mapeamento das leituras. Padrão: 8.
-M, --max-base-freq	Fração máxima permitida de frequência de bases. Padrão : 0.75.
-D, --deduplicate	Remove as leituras duplicadas. As leituras são consideradas duplicadas quando compartilham as posições 5’ de ambas as leituras e pares de leituras.
-s, --sorted	Assume que os arquivos de alinhamento estão ordenados por *queryname*.

Quadro 9 - Demais opções para o *process-sample*.

(conclusão)

Opção	Descrição*
-t, --threads	Número de *threads*. Padrão: 1.
-m, --max-distance	Distância máxima entre os pares de leituras antes de considerá-las mais distantes que o esperado. Padrão: 10000.
-f, --exon-frac	Sobreposição mínima necessária como uma fração de éxon. Padrão: 1 base.
-F, --alignment-frac	Sobreposição mínima necessária como uma fração de alinhamento. Padrão: 1 base.
-e, --either	A fração mínima que deve ser satisfeita para pelo menos o alinhamento OU o éxon. Sem *-e*, ambas as frações teriam que ser satisfeitas.
-r, --reciprocal	A sobreposição de fração deve ser recíproca para éxon e alinhamento. Se *-f* for 0.5, então *-F* também será definido como 0.5.

Fonte: sideRETRO versão 1.0.0.
* Ponto como separador decimal (para *LC_NUMERIC=en_US.UTF-8*).

4.1.5 Usando sideRETRO: *merge-call*

O segundo subcomando do sideRETRO é o *merge-call*. O objetivo dele é pegar o banco de dados criado pela etapa de *process-sample*, como entrada, e preencher as tabelas nele restantes com informações levantadas do processo de agrupamento das leituras anormais e a da anotação de eventos de retroCNV. Chamar o comando *merge-call* sem nenhum argumento (1), assim como para *process-sample*, dará uma ajuda específica na qual o usuário poderá conhecer todas as opções obrigatórias para este comando:

sider merge-call *(1)*

É necessário que o usuário indique um banco de dados, gerado na etapa de *process-sample*, ou uma lista de bancos de dados (2).

sider merge-call out1.db out2.db out3.db *(2)*

Seguem as opções disponíveis (Quadros 10 e 11) do subcomando *merge-call*:

Quadro 10 - Opções obrigatórias para o *merge-call*.

Opção	Descrição
-i, --input-file	Arquivo contendo uma lista separada por nova linha de bancos de dados SQLite3 a serem processados. Esta opção não é obrigatória se um ou mais bancos de dados SQLite3 forem passados como argumento. Se input-file e argumentos forem definidos concomitantemente, a união de todos os arquivos será usada.

Fonte: sideRETRO versão 1.0.0.

Quadro 11 - Demais opções para o *merge-call*.

(continua)

Opção	Descrição
-h, --help	Mostra as opções de ajuda.
-q, --quiet	Diminui a verbosidade apenas para mensagens de erro ou suprime as saídas do terminal se *log-file* for passado.
--silent	O mesmo que *–quiet*.
-d, --debug	Aumente a verbosidade para o nível de depuração.
-l, --log-file	Imprime mensagens de *log* em um arquivo.
-o, --output-dir	Diretório de saída. Cria o diretório caso este não exista.
-p, --prefix	Prefixo para os arquivos de saída. Padrão: *out*.
-I, --in-place	Funde todos os bancos de dados da lista num único banco de dados, em vez de criar um novo.
-c, --cache-size	Define o tamanho do *cache* para o SQLite3 em KiB. Padrão: 200000.
-e, --epsilon	DBSCAN - Distância máxima entre dois alinhamentos para considerá-los nas vizinhanças um do outro. Padrão: 300 bases.
-m, --min-pts	DBSCAN - Número mínimo de alinhamentos num agrupamento para serem considerados uma região densa. Padrão: 10 alinhamentos.
-b, --blacklist-chr	Evita agrupar alinhamentos anormais neste cromossomo ou cujo gene parental esteja neste cromossomo. Padrão: cromossomo mitocondrial.
-B, --blacklist-region	Evita agrupar alinhamentos anormais nas regiões presentes no arquivo. São aceitos os formatos GTF/GFF3/BED. Se o formato for GTF/GFF3. pode-se filtrar as regiões pela característica (terceira coluna) e pelos atributos (nona coluna).
-P, --blacklist-padding	Expande as janelas das regiões proibidas para grupamento.

Quadro 11 - Demais opções para o *merge-call*.

(conclusão)

Opção	Descrição
-T, --gff-feature	O valor da característica (terceira coluna), no caso de um arquivo GTF/GFF3 ter sido passado para a opção *--blacklist-region*. Padrão: *gene*.
-H, --gff-hard-attribute	O valor do atributo (nona coluna), no caso de um arquivo GTF/GFF3 ter sido passado para a opção *–blacklist-region*. Esta opção deve ser escrita na formato chave=valor. É possível passar *regex*, então *gene_type=pseudogene* irá capturar *IG_C_pseudogene* e *IG_V_pseudogene* - por exemplo. Pode-se passar esta opção múltiplas vezes e só será capturada a entrada que passar por todos os filtros.
-S, --gff-soft-attribute	Funciona como *–gff-hard-attribute*. A diferença é que se esta opção for passada múltiplas vezes, apenas um filtro passando basta para a entrada ser capturada. Padrão: *gene_type=processed_pseudogene* e *tag=retrogene*.
-x, --parental-distance	Distância mínima permitida entre um agrupamento e seu gene parental. Padrão: 1000000.
-g, --genotype-support	Número mínimo de alinhamentos vindos de uma única fonte, arquivo de alinhamento SAM/BAM/CRAM, dentro de um agrupamento. Padrão: 1.
-n, --near-gene-rank	Distância ranqueada mínima entre dois genes para considerá-los próximos. Padrão: 3.
-t, --threads	Número de *threads*. Padrão: 1.
-Q, --phred-quality	Qualidade mínima de mapeamento das leituras. Padrão: 8.

Fonte: sideRETRO versão 1.0.0.

4.1.6 Usando sideRETRO: *make-vcf*

O terceiro e último subcomando do sider é o *make-vcf*. Este subcomando exige um banco de dados gerado nas etapas anteriores de *process-sample* e *merge-call*. Como resultado, é gerado um arquivo de anotação no formato VCF.

Chamar o comando *make-vcf* sem nenhum argumento (1), assim como para *process-sample* e *merge-call*, dará uma ajuda específica onde o usuário poderá conhecer todas as opções para este comando:

sider make-vcf *(1)*

Um exemplo de uso (2):

sider vcf -r hg38.fa -o result in.db *(2)*

Seguem as opções disponíveis (Quadro 12) no subcomando *make-vcf*.

Quadro 12 - Opções para o *make-vcf*.

Opção	Descrição*
-h, --help	Mostra as opções de ajuda.
-q, --quiet	Diminui a verbosidade apenas para mensagens de erro ou suprime as saídas do terminal se *log-file* for passado.
--silent	O mesmo que *–quiet*.
-d, --debug	Aumente a verbosidade para o nível de depuração.
-l, --log-file	Imprime mensagens de *log* em um arquivo.
-o, --output-dir	Diretório de saída. Cria o diretório caso este não exista.
-p, --prefix	Prefixo para os arquivos de saída. Padrão: *out*.
-n, --near-gene-dist	Distância mínima entre genes para que estes sejam considerados próximos. Padrão: 10000.
-e, --orientation-error	Erro tipo-I para o teste de correlação de postos de *Spearman* (*rho*) durante a detecção da orientação da fita. Padrão: 0.05.
-r, --reference-file	Arquivo FASTA para o genoma de referência.

Fonte: sideRETRO versão 1.0.0.
* Ponto como separador decimal (para *LC_NUMERIC=en_US.UTF-8*).

4.1.6.1 Arquivo VCF gerado pelo sideRETRO

O arquivo VCF de saída é escrito, seguindo as especificações da versão 4,2 (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021b). São adicionados campos específicos para eventos de retroCNV no cabeçalho e corpo do arquivo, de modo que o usuário obtenha a conclusão do experimento com todos os detalhes gerados. O Quadro 4 que consta na metodologia contém detalhadamente essas informações e o modo de interpretá-las.

4.2 RESULTADO DO SIDERETRO PARA DADOS SIMULADOS

Resultados para a simulação computacional de 100 eventos de retroCNVs, desenhados e inseridos aleatoriamente numa população de 100 indivíduos (Tabela 17). Foram anotados pelo sideRETRO 79 eventos (Tabela 18), destes 19 são retroCNVs fixadas (de um total de 25), 42 são polimórficas (de um total de 50) e 18 são somáticas (de um total de 25) (Tabela 19). Quanto a sua posição genômica, as retroCNVs detectadas tiveram um Erro Quadrático Médio[36] (MSE) de bases de 128 e um Erro Quadŕatico Mediano[37] (MEDSE) de bases de 1 (Tabela 20), em relação ao que foi simulado e anotado. A fita (polaridade) foi identificada em 78 de 79 casos (Tabela 21) e, em todos os casos, de forma correta.

Os indivíduos tiveram uma inserção média de 30 retroCNVs (Tabela 22). O desempenho do sideRETRO quanto a genotipagem dos 100 indivíduos foi mensurado pelo F1-*score* (SASAKI; OTHERS, 2007) - que é a média harmônica entre a precisão e a sensibilidade. Os acertos, Verdadeiro Positivo (VP), e os erros, Falso Positivo (FP) e Falso Negativo (FN), foram computados segundo a detecção do alelo alternativo, então eventos heterozigóticos contaram 1 ponto e eventos homozigóticos 2 pontos. A detecção de um evento heterozigótico somou 1 ponto para VP, a não detecção somou 1 ponto para FN e a detecção do alelo não simulado somou 1 ponto para FP. Para eventos homozigóticos, a detecção somou 2 pontos para VP, a não detecção 2 pontos para FN, a detecção como evento heterozigótico somou 1 ponto para VP e 1 ponto para FN; por fim, a detecção de alelos não simulados contou 2 pontos para FP. Sendo assim, o sideRETRO obteve uma precisão média de 96%, uma sensibilidade média de 78% e uma acuŕacia média, segundo o F1-*score*, de 86% (Tabela 22 e Figura 28).

Tabela 17 - As 100 retroCNVs simuladas, com seus respectivos genes parentais e posições genômicas.

(continua)

Gene parental	Cromossomo	Posição	Fita
AC002310.4	chr9	94.545.202	-
AC135178.3	chr7	74.794.901	-

[36] *Mean Squared Error* (MSE).
[37] *Median Squared Error* (MEDSE).

Tabela 17 - As 100 retroCNVs simuladas, com seus respectivos genes parentais e posições genômicas.

(continuação)

Gene parental	Cromossomo	Posição	Fita
ACSBG2	chr21	43.058.887	-
ADD2	chr3	9.759.497	+
AL645922.1	chr6	38.626.680	-
ALG2	chr10	30.778.982	-
ARMC2	chr5	52.723.637	-
ATG2B	chr5	177.026.995	-
BTF3	chr7	146.774.631	-
C21orf91	chr14	54.886.570	-
C2orf92	chr6	112.158.328	-
C8orf76	chr9	94.927.085	-
C9orf64	chr17	40.139.106	+
CABP7	chr5	153.788.597	+
CARD8	chrX	99.922.659	+
CASTOR3	chr3	189.081.695	-
CDH22	chr9	113.306.486	-
CERS1	chr20	41.341.204	+
CFAP69	chr11	10.733.916	-
COL4A3	chr16	46.427.444	+
COPS2	chr1	38.773.310	-
CPNE7	chr9	42.228.417	+
CWC25	chr13	39.475.646	-
DENND2D	chr18	37.314.709	+
DHRSX	chr5	166.496.220	-
DNAJC27	chr12	60.940.050	-
EPC2	chr13	94.468.157	-
EPS8	chr21	26.428.011	+
ERCC4	chr6	93.262.920	+
FAAP20	chr9	77.384.901	-
FAM177B	chr12	130.498.191	+
FAM71E2	chr2	225.319.689	+
HAO2	chr14	69.901.152	+
HEG1	chr3	15.517.386	-
HIP1	chr8	75.177.754	+
IL1R1	chr8	30.386.429	-
IQGAP3	chr6	124.358.143	+
KIF7	chrX	89.251.626	-
LAMP1	chr13	87.908.197	-

Tabela 17 - As 100 retroCNVs simuladas, com seus respectivos genes parentais e posições genômicas.

(continuação)

Gene parental	Cromossomo	Posição	Fita
LARS	chr9	64.069.435	+
LETM1	chrY	24.793.930	-
LRRC6	chr4	180.728.002	-
MACROD2	chr20	18.178.487	+
MALL	chr7	110.598.366	+
MRPS7	chr2	1.490.696	+
MTNR1A	chr8	86.938.090	-
MYH10	chr4	186.290.075	+
MYH7B	chr13	104.241.206	+
MYO7A	chr11	14.072.547	+
NAE1	chr18	74.528.384	+
NDUFA6	chr10	38.060.463	+
OR14A16	chr1	52.758.590	+
OR51M1	chr2	37.409.208	-
OSER1	chr5	53.846.631	-
PAFAH1B1	chr15	86.208.543	+
PDGFB	chr8	133.462.380	-
PFKFB2	chr5	36.822.019	-
PLAC8	chr9	39.225.441	+
PLCB1	chr9	25.165.703	+
PNRC1	chr15	48.607.415	+
PRMT2	chr8	50.511.539	-
PRPF18	chr20	51.460.729	+
PRSS45P	chr19	5.420.707	-
PTCHD4	chr15	31.035.142	-
PTPRF	chr19	7.227.546	+
RAB18	chr4	10.281.361	-
RAB5B	chr6	46.561.322	+
RADX	chr12	117.277.769	+
RASGEF1C	chr5	115.992.817	+
RBM4	chr7	101.199.285	+
RMDN3	chr3	28.655.572	-
RNF6	chr4	39.797.761	-
SART1	chr2	109.317.943	+
SDHA	chr4	179.658.356	+
SEZ6L	chr18	560.651	-
SKP2	chr5	88.746.051	-

Tabela 17 - As 100 retroCNVs simuladas, com seus respectivos genes parentais e posições genômicas.

(conclusão)

Gene parental	Cromossomo	Posição	Fita
SLC44A4	chrY	4.417.954	+
SLC9A3	chr4	140.369.141	-
SMTNL2	chr3	144.112.843	-
SNRNP27	chrX	13.251.389	-
STK17B	chrX	36.995.058	-
STON2	chrX	468.106	+
TACO1	chrY	12.987.416	+
TAF7	chr22	22.384.919	-
TBC1D3F	chr16	65.760.883	+
TMEM63C	chr17	49.131.966	+
TMEM95	chr2	234.301.985	-
TRIM40	chr5	45.713.519	+
TSFM	chr12	80.384.739	-
TUBGCP2	chr1	197.233.691	+
VIPAS39	chr12	54.021.508	-
WDR74	chr11	112.552.782	-
WDR75	chr6	132.636.317	+
ZNF136	chr16	59.509.103	+
ZNF326	chr8	29.273.486	-
ZNF385A	chr12	92.752.469	-
ZNF431	chr16	88.101.015	-
ZNF585A	chr18	78.888.223	-
ZNF738	chr6	139.608.184	-
ZNF793	chr9	120.420.222	+

Fonte: autoria própria.

Tabela 18 - As 79 retroCNVs anotadas pelo sideRETRO durante a simulação.

(continua)

Gene parental	Posição simulada			Posição predita pelo sideRETRO		
	Cromossomo	Posição	Fita	Cromossomo	Posição	Fita
ALG2	chr10	30.778.982	-	chr10	30.778.981	-
ARMC2	chr5	52.723.637	-	chr5	52.723.638	-
ATG2B	chr5	177.026.995	-	chr5	177.026.990	-
BTF3	chr7	146.774.631	-	chr7	146.774.629	-
C2orf92	chr6	112.158.328	-	chr6	112.158.327	-
C8orf76	chr9	94.927.085	-	chr9	94.927.084	-
C9orf64	chr17	40.139.106	+	chr17	40.139.104	+

Tabela 18 - As 79 retroCNVs anotadas pelo sideRETRO durante a simulação. (continuação)

Gene parental	Posição simulada			Posição predita pelo sideRETRO		
	Cromossomo	Posição	Fita	Cromossomo	Posição	Fita
CABP7	chr5	153.788.597	+	chr5	153.788.596	+
CARD8	chrX	99.922.659	+	chrX	99.922.658	+
CASTOR3	chr3	189.081.695	-	chr3	189.081.692	-
CDH22	chr9	113.306.486	-	chr9	113.306.485	-
CFAP69	chr11	10.733.916	-	chr11	10.733.915	-
COL4A3	chr16	46.427.444	+	chr16	46.427.444	+
COPS2	chr1	38.773.310	-	chr1	38.773.309	-
CPNE7	chr9	42.228.417	+	chr9	42.228.469	∅
DENND2D	chr18	37.314.709	+	chr18	37.314.708	+
DNAJC27	chr12	60.940.050	-	chr12	60.940.049	-
EPC2	chr13	94.468.157	-	chr13	94.468.156	-
EPS8	chr21	26.428.011	+	chr21	26.428.011	+
ERCC4	chr6	93.262.920	+	chr6	93.262.919	+
FAAP20	chr9	77.384.901	-	chr9	77.384.898	-
FAM177B	chr12	130.498.191	+	chr12	130.498.188	+
FAM71E2	chr2	225.319.689	+	chr2	225.319.688	+
HAO2	chr14	69.901.152	+	chr14	69.901.150	+
HEG1	chr3	15.517.386	-	chr3	15.517.382	-
HIP1	chr8	75.177.754	+	chr8	75.177.754	+
IL1R1	chr8	30.386.429	-	chr8	30.386.427	-
IQGAP3	chr6	124.358.143	+	chr6	124.358.101	+
KIF7	chrX	89.251.626	-	chrX	89.251.603	-
LAMP1	chr13	87.908.197	-	chr13	87.908.197	-
LARS	chr9	64.069.435	+	chr9	64.069.377	+
LRRC6	chr4	180.728.002	-	chr4	180.728.002	-
MACROD2	chr20	18.178.487	+	chr20	18.178.486	+
MYH10	chr4	186.290.075	+	chr4	186.290.074	+
MYH7B	chr13	104.241.206	+	chr13	104.241.205	+
MYO7A	chr11	14.072.547	+	chr11	14.072.546	+
NAE1	chr18	74.528.384	+	chr18	74.528.383	+
OR14A16	chr1	52.758.590	+	chr1	52.758.589	+
OR51M1	chr2	37.409.208	-	chr2	37.409.207	-
OSER1	chr5	53.846.631	-	chr5	53.846.596	-
PAFAH1B1	chr15	86.208.543	+	chr15	86.208.562	+
PDGFB	chr8	133.462.380	-	chr8	133.462.379	-
PFKFB2	chr5	36.822.019	-	chr5	36.822.019	-

Tabela 18 - As 79 retroCNVs anotadas pelo sideRETRO durante a simulação. (conclusão)

Gene parental	Posição simulada			Posição predita pelo sideRETRO		
	Cromossomo	Posição	Fita	Cromossomo	Posição	Fita
PLCB1	chr9	25.165.703	+	chr9	25.165.702	+
PNRC1	chr15	48.607.415	+	chr15	48.607.414	+
PRMT2	chr8	50.511.539	-	chr8	50.511.540	-
PRPF18	chr20	51.460.729	+	chr20	51.460.728	+
PRSS45P	chr19	5.420.707	-	chr19	5.420.706	-
PTPRF	chr19	7.227.546	+	chr19	7.227.546	+
RAB18	chr4	10.281.361	-	chr4	10.281.361	-
RAB5B	chr6	46.561.322	+	chr6	46.561.322	+
RADX	chr12	117.277.769	+	chr12	117.277.768	+
RASGEF1C	chr5	115.992.817	+	chr5	115.992.816	+
RBM4	chr7	101.199.285	+	chr7	101.199.284	+
RMDN3	chr3	28.655.572	-	chr3	28.655.571	-
RNF6	chr4	39.797.761	-	chr4	39.797.759	-
SART1	chr2	109.317.943	+	chr2	109.317.942	+
SDHA	chr4	179.658.356	+	chr4	179.658.355	+
SEZ6L	chr18	560.651	-	chr18	560.650	-
SKP2	chr5	88.746.051	-	chr5	88.746.050	-
SLC9A3	chr4	140.369.141	-	chr4	140.369.139	-
SMTNL2	chr3	144.112.843	-	chr3	144.112.842	-
SNRNP27	chrX	13.251.389	-	chrX	13.251.387	-
STK17B	chrX	36.995.058	-	chrX	36.995.057	-
TACO1	chrY	12.987.416	+	chrY	12.987.415	+
TMEM63C	chr17	49.131.966	+	chr17	49.131.965	+
TMEM95	chr2	234.301.985	-	chr2	234.301.984	-
TSFM	chr12	80.384.739	-	chr12	80.384.736	-
TUBGCP2	chr1	197.233.691	+	chr1	197.233.690	+
VIPAS39	chr12	54.021.508	-	chr12	54.021.507	-
WDR74	chr11	112.552.782	-	chr11	112.552.781	-
WDR75	chr6	132.636.317	+	chr6	132.636.316	+
ZNF136	chr16	59.509.103	+	chr16	59.509.104	+
ZNF326	chr8	29.273.486	-	chr8	29.273.482	-
ZNF385A	chr12	92.752.469	-	chr12	92.752.468	-
ZNF431	chr16	88.101.015	-	chr16	88.101.015	-
ZNF585A	chr18	78.888.223	-	chr18	78.888.222	-
ZNF738	chr6	139.608.184	-	chr6	139.608.183	-
ZNF793	chr9	120.420.222	+	chr9	120.420.223	+

Fonte: autoria própria.

Tabela 19 - Detecção das retroCNVs segundo as categorias de simulação.

retroCNVs	Eventos simulados	Eventos anotados pelo sideRETRO	Acertos
Fixadas	25	19	0,76
Polimórficas	0	42	0,84
Somáticas	25	18	0,72

Fonte: autoria própria.
Fixadas (inseridas em todos os indivíduos), polimórficas (inseridas em pelo menos 2 indivíduos) e somáticas (inseridas em um único indivíduo).

Tabela 20 - Erro na predição do ponto de inserção.

Erro	Valor
MSE	128
MEDSE	1

Fonte: autoria própria.
As posições (pontos de inserção) preditas pelo sideRETRO para as 79 retroCNVs anotadas das 100 simuladas.

Tabela 21 - Erro na predição da fita.

retroCNVs detectadas	Fitas detectadas	Erro	Erro médio
79	78	1	0,013

Fonte: autoria própria.
Erro para as 79 retroCNVs anotadas dos 100 eventos simulados. A não predição de fita foi contada como erro, juntamente com a predição errônea de fita.

Tabela 22 - Desempenho do sideRETRO durante a genotipagem dos 100 indivíduos, sequenciamentos, simulados.

(continua)

Indivíduo	VP	FP	FN	Precisão	Sensibilidade	F1-*score*
0	38	0	9	1,00	0,81	0,89
1	36	2	11	0,95	0,77	0,85
2	33	1	10	0,97	0,77	0,86
3	35	1	12	0,97	0,74	0,84
4	29	1	9	0,97	0,76	0,85
5	37	4	12	0,90	0,76	0,82
6	45	0	10	1,00	0,82	0,90
7	37	2	11	0,95	0,77	0,85
8	32	2	11	0,94	0,74	0,83
9	33	3	11	0,92	0,75	0,83
10	34	1	9	0,97	0,79	0,87
11	37	2	12	0,95	0,76	0,84
12	30	1	10	0,97	0,75	0,85
13	43	3	11	0,93	0,80	0,86
14	38	0	10	1,00	0,79	0,88

Tabela 22 - Desempenho do sideRETRO durante a genotipagem dos 100 indivíduos, sequenciamentos, simulados.

(continuação)

Indivíduo	VP	FP	FN	Precisão	Sensibilidade	F1-*score*
15	31	1	8	0,97	0,79	0,87
16	30	4	13	0,88	0,70	0,78
17	39	1	9	0,98	0,81	0,89
18	37	0	10	1,00	0,79	0,88
19	39	1	10	0,98	0,80	0,88
20	39	2	12	0,95	0,76	0,85
21	42	3	12	0,93	0,78	0,85
22	39	0	10	1,00	0,80	0,89
23	41	2	10	0,95	0,80	0,87
24	43	1	8	0,98	0,84	0,91
25	41	0	9	1,00	0,82	0,90
26	43	0	10	1,00	0,81	0,90
27	34	0	10	1,00	0,77	0,87
28	38	4	14	0,90	0,73	0,81
29	36	1	11	0,97	0,77	0,86
30	47	3	11	0,94	0,81	0,87
31	43	3	12	0,93	0,78	0,85
32	38	0	11	1,00	0,78	0,87
33	34	1	12	0,97	0,74	0,84
34	35	4	12	0,90	0,74	0,81
35	43	2	10	0,96	0,81	0,88
36	41	2	11	0,95	0,79	0,86
37	38	1	11	0,97	0,78	0,86
38	34	1	9	0,97	0,79	0,87
39	39	0	8	1,00	0,83	0,91
40	35	1	9	0,97	0,80	0,88
41	33	1	9	0,97	0,79	0,87
42	39	1	11	0,98	0,78	0,87
43	37	4	13	0,90	0,74	0,81
44	39	4	13	0,91	0,75	0,82
45	35	3	11	0,92	0,76	0,83
46	31	0	9	1,00	0,78	0,87
47	36	0	10	1,00	0,78	0,88
48	40	3	11	0,93	0,78	0,85
49	34	1	10	0,97	0,77	0,86
50	41	4	13	0,91	0,76	0,83
51	34	0	9	1,00	0,79	0,88

Tabela 22 - Desempenho do sideRETRO durante a genotipagem dos 100 indivíduos, sequenciamentos, simulados.

(continuação)

Indivíduo	VP	FP	FN	Precisão	Sensibilidade	F1-*score*
52	36	3	12	0,92	0,75	0,83
53	39	2	11	0,95	0,78	0,86
54	47	0	10	1,00	0,82	0,90
55	36	1	12	0,97	0,75	0,85
56	40	2	12	0,95	0,77	0,85
57	41	1	9	0,98	0,82	0,89
58	40	0	10	1,00	0,80	0,89
59	34	3	11	0,92	0,76	0,83
60	35	2	10	0,95	0,78	0,85
61	38	1	9	0,97	0,81	0,88
62	30	1	8	0,97	0,79	0,87
63	38	4	13	0,90	0,75	0,82
64	43	2	10	0,96	0,81	0,88
65	46	1	10	0,98	0,82	0,89
66	41	1	10	0,98	0,80	0,88
67	37	2	9	0,95	0,80	0,87
68	44	5	13	0,90	0,77	0,83
69	36	0	9	1,00	0,80	0,89
70	42	4	14	0,91	0,75	0,82
71	44	3	14	0,94	0,76	0,84
72	41	3	13	0,93	0,76	0,84
73	34	1	9	0,97	0,79	0,87
74	42	1	10	0,98	0,81	0,88
75	37	3	11	0,93	0,77	0,84
76	34	2	9	0,94	0,79	0,86
77	37	3	10	0,93	0,79	0,85
78	38	0	8	1,00	0,83	0,90
79	40	2	9	0,95	0,82	0,88
80	35	0	9	1,00	0,80	0,89
81	40	1	10	0,98	0,80	0,88
82	41	2	11	0,95	0,79	0,86
83	39	2	11	0,95	0,78	0,86
84	40	3	10	0,93	0,80	0,86
85	36	4	12	0,90	0,75	0,82
86	37	4	13	0,90	0,74	0,81
87	32	2	11	0,94	0,74	0,83
88	42	2	12	0,95	0,78	0,86

Tabela 22 - Desempenho do sideRETRO durante a genotipagem dos 100 indivíduos, sequenciamentos, simulados.

(conclusão)

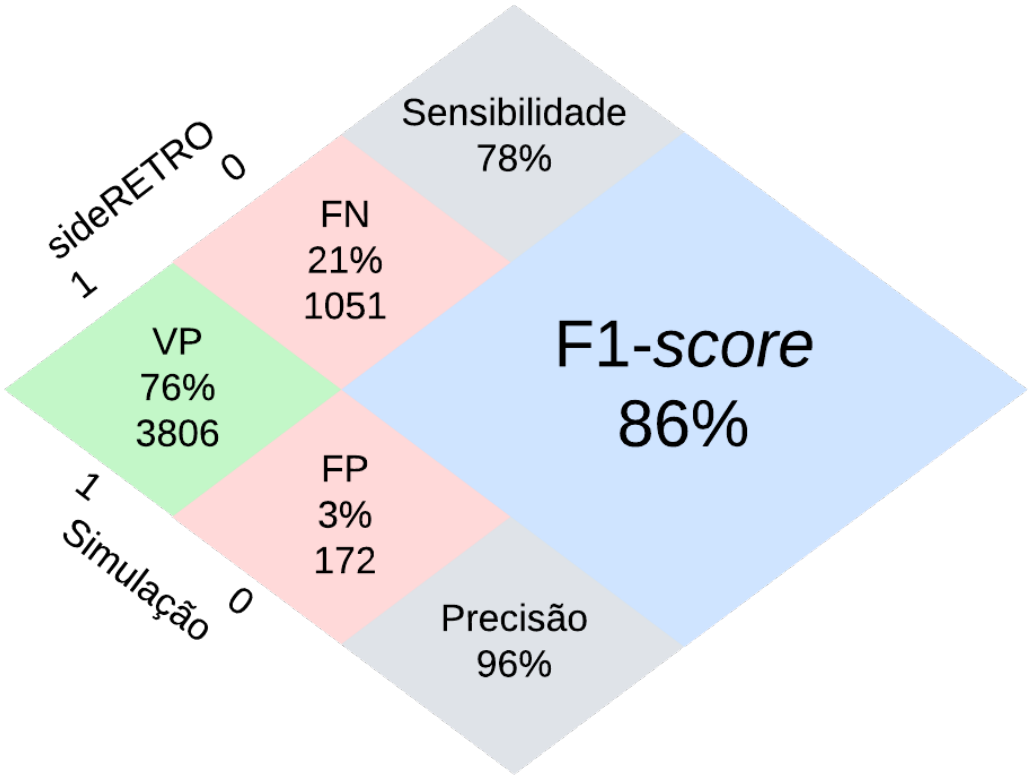
Indivíduo	VP	FP	FN	Precisão	Sensibilidade	F1-*score*
89	34	1	9	0,97	0,79	0,87
90	41	2	10	0,95	0,80	0,87
91	45	0	9	1,00	0,83	0,91
92	39	2	8	0,95	0,83	0,89
93	39	2	11	0,95	0,78	0,86
94	34	3	12	0,92	0,74	0,82
95	44	4	11	0,92	0,80	0,85
96	36	1	9	0,97	0,80	0,88
97	39	2	10	0,95	0,80	0,87
98	48	0	9	1,00	0,84	0,91
99	40	0	10	1,00	0,80	0,89
Total	3.806	172	1051	0,96	0,78	0,86

Fonte: autoria própria.

Cada indivíduo teve uma inserção média de 30 das 100 retroCNVs simuladas. Foram contados os acertos com relação à detecção do alelo alternativo, então eventos heterozigóticos contaram 1 ponto e eventos homozigóticos contaram 2 pontos. Para eventos heterozigóticos: detecção do alelo, VP = VP + 1; a não detecção do alelo, FN = FN + 1; a detecção de um evento não simulado FP = FP + 1. Para eventos homozigóticos: detecção dos dois alelos, VP = VP + 2; a não detecção do evento, FN = FN + 2; um alelo detectado, VP = VP + 1 e FN = FN +1; a detecção de um evento não simulado FP = FP + 2.

VP = Verdadeiro Positivo; FP = Falso Positivo; FN = Falso Negativo.

Figura 28 - Desempenho geral do sideRETRO durante a genotipagem dos 100 indivíduos, sequenciamentos, simulados.



Fonte: autoria própria.
VP = Verdadeiro Positivo; FP = Falso Positivo; FN = Falso Negativo; 1 = Presença de retroCNV; 0 = Ausência de retroCNV.

4.3 RESULTADOS DO SIDERETRO PARA DADOS REAIS

Resultados do sideRETRO para os dados validados experimentalmente por Abyzov e colaboradores (ABYZOV et al., 2013). Das seis retroCNVs validadas por PCR, o sideRETRO anotou cinco: dos genes parentais CBX3, LAPTM4B, TMEM66/SARAF, SKA3 e TDG. Só não foi detectado o evento validado para o gene parental CACNA1B (Tabela 23).

A genotipagem para as 14 populações do 1KGP pelo sideRETRO obteve os mesmos resultados que os obtidos por Abyzov et al. para as cinco retroCNVs identificadas pela ferramenta (Tabela 24).

Tabela 23 - RetroCNVs validadas experimentalmente por PCR e genotipadas por Abyzov et al. e pelo sideRETRO em indivíduos de 14 populações humanas.

Gene parental	Abyzov et al.			sideRETRO	
	Cromossomo	Início	Fim	Cromossomo	Ponto de inserção
CBX3	chr15	40.561.954	40.561.998	chr15	40.561.980
LAPTM4B	chr6	166.920.412	166.920.482	chr6	166.920.475
TMEM66*	chr1	191.829.533	191.829.591	chr1	191.829.594
SKA3	chr11	108.714.998	108.715.054	chr11	108.715.020
TDG	chr12	125.316.536	125.316.676	chr12	125.316.601
CACNA1B	chr1	148.027.670	148.027.843	-	-

Fonte: autoria própria.
TMEM66 (usado em Abyzov et al.): agora seu nome oficial é SARAF.

Tabela 24 - Genotipagem das retroCNVs validadas experimentalmente por Abyzov et al. em 14 populações humanas do projeto 1000 Genomas.

População	Gene parental da retroCNV					
	CBX3	LAPTM4B	TMEM66*	SKA3	TDG	CACNA1B
ASW	1/1	0/0	0/0	1/1	1/1	1/0
CEU	1/1	1/1	1/1	1/1	1/1	1/0
CHB	1/1	0/0	0/0	1/1	1/1	1/0
CHS	1/1	0/0	0/0	1/1	1/1	1/0
CLM	1/1	1/1	0/0	1/1	1/1	1/0
FIN	1/1	1/1	1/1	1/1	1/1	1/0
GBR	1/1	1/1	1/1	1/1	1/1	1/0
IBS	1/1	0/0	0/0	1/1	1/1	1/0
JPT	1/1	0/0	0/0	1/1	1/1	1/0
LWK	1/1	0/0	0/0	1/1	1/1	1/0
MXL	1/1	0/0	0/0	1/1	1/1	1/0
PUR	1/1	1/1	1/1	1/1	1/1	1/0
TSI	1/1	1/1	1/1	1/1	1/1	1/0
YRI	1/1	0/0	0/0	1/1	1/1	1/0

Fonte: autoria própria.
Genotipagem (“retroCNV anotada por Abyzov et al.” / “retroCNV anotado pelo sideRETRO”). Evento anotado por Abyzov et al. e sideRETRO - 1/1; Eventos anotados apenas por Abyzov et al. - 1/0; Eventos ausentes na população - 0/0.
TMEM66 (usado em Abyzov et al.): agora seu nome oficial é SARAF.

5 Discussão

5 DISCUSSÃO

5.1 ESTIMANDO O DESEMPENHO DO SIDERETRO COM DADOS SIMULADOS

Das retroCNVs desenhadas, 79% foram anotadas pelo sideRETRO com uma precisão média de 96% na detecção do alelo alternativo durante a genotipagem. Para os eventos anotados, foi computado o ponto de inserção com um erro quadrático mediano de 1, enquanto que a fita foi detectada, e corretamente, para 78 de 79 casos - sendo que o único caso não detectado não adveio de um erro algorítmico, e sim do erro tipo I no teste de correlação de postos de *Spearman*.

Nota-se que não obstante uma cobertura de sequenciamento simulado de 20 vezes (10 vezes de cobertura para um evento heterozigótico), eventos polimórficos (inseridos em pelo menos dois indivíduos) e somáticos (inseridos em um indivíduo) obtiveram um percentual de anotação similar aos eventos fixados (presentes em todos os indivíduos): 84%, 72% e 76%, respectivamente - sendo ainda que os polimórficos e somáticos correspondem juntos a 75% de todos os eventos. Ou seja, o sideRETRO demonstrou uma capacidade de detecção não apenas de retroCNVs presentes numa dada população, como um todo, mas também um poder de detecção de eventos raros, talvez presentes num único indivíduo - o que seria útil para estudos relacionados aos efeitos das variações estruturais do genoma: num contexto tumorigênico, por exemplo.

Não foram anotadas pelo sideRETRO 21% das retroCNVs, o que resultou, durante a genotipagem, no aumento de falsos negativos e, consequentemente, na queda do desempenho quanto a sensibilidade média - que obteve um valor de 78%. Comparativamente, observou-se um valor baixo para o número de falsos positivos, e por conseguinte uma precisão de 96%. Vide esse fato, averiguou-se o porquê de 21 eventos não serem encontrados em nenhum indivíduo. A abordagem consistiu em cruzar as posições genômicas simuladas das 21 retroCNVs com um banco de dados contendo regiões consideradas ambíguas - por estarem duplicadas no genoma (AMEMIYA; KUNDAJE; BOYLE, 2019). Com isso, foi verificado que 14 das 21 retrocópias haviam sido inseridas em regiões ambíguas e de difícil acesso metodológico computacional (Tabela 25). Por esta razão, o resultado foi

reinterpretado de modo a excluir esses eventos das análises e assim obteve-se o seguinte desempenho, agora para 86 retroCNVs e não 100: na genotipagem, o número total de FN foi de 1.051 alelos não detectados para 551, o que representou uma queda de aproximadamente 48%. A sensibilidade, então, aumentou de 78% para 87% e o F1-*score* de 86% para 91% (Tabela 26 e Figura 29).

Tabela 25 - Os 21 eventos simulados de retroCNV não encontrados pelo sideRETRO.

RetroCNV não anotada pelo sideRETRO			Região genômica* ambígua (duplicada)
Gene parental	Cromossomo	Posição	
AC002310.4	chr9	94.545.202	chr8:115.819.078-115.819.180
			chr7:75.151.009-75.151.108
AC135178.3	chr7	74.794.901	chr7:74.794.851-74.794.950
			chr7:73.241.459-73.241.558
ACSBG2	chr21	43.058.887	chr21:6.450.515-6.450.614
			chr21:43.058.837-43.058.936
ADD2	chr3	9.759.497	Não
AL645922.1	chr6	38.626.680	Não
C21orf91	chr14	54.886.570	Duplicações em genoma 7x
CERS1	chr20	41.341.204	Não
CWC25	chr13	39.475.646	Não
DHRSX	chr5	166.496.220	Região altamente repetitiva (8x > 90%)
LETM1	chrY	24.793.930	Duplicação em 8 regiões idênticas no chrY
MALL	chr7	110.598.366	Não
MRPS7	chr2	1.490.696	chr2_KI270774v1_alt
MTNR1A	chr8	86.938.090	chrX, chr4
NDUFA6	chr10	38.060.463	chr10:42.588.649-42.588.750
PLAC8	chr9	39.225.441	chr9:61.393.599-61.393.698
PTCHD4	chr15	31.035.142	chr15_KI270905v1_alt
SLC44A4	chrY	4.417.954	chrX:90.835.484-90.835.583
STON2	chrX	468.106	chrY:468.056-468.155
TAF7	chr22	22.384.919	chr22_KI270875v1_alt
TBC1D3F	chr16	65.760.883	Não
TRIM40	chr5	45.713.519	Não

Fonte: (AMEMIYA; KUNDAJE; BOYLE, 2019).
14 estão localizados em regiões genômicas ambíguas.
* região de 100 bases ao redor do ponto de inserção.

Tabela 26 - Desempenho do sideRETRO durante a genotipagem dos 100 indivíduos simulados, após a remoção dos eventos localizados em regiões altamente repetitivas.

(continua)

Indivíduo	VP	FP	FN	Precisão	Sensibilidade	F1-*score*
0	38	0	5	1,00	0,88	0,94
1	36	2	7	0,95	0,84	0,89
2	33	1	6	0,97	0,85	0,90
3	35	1	5	0,97	0,88	0,92
4	29	1	5	0,97	0,85	0,91
5	37	4	5	0,90	0,88	0,89
6	45	0	6	1,00	0,88	0,94
7	37	2	5	0,95	0,88	0,91
8	32	2	5	0,94	0,86	0,90
9	33	3	5	0,92	0,87	0,89
10	34	1	5	0,97	0,87	0,92
11	37	2	5	0,95	0,88	0,91
12	30	1	5	0,97	0,86	0,91
13	43	3	5	0,93	0,90	0,91
14	38	0	6	1,00	0,86	0,93
15	31	1	5	0,97	0,86	0,91
16	30	4	6	0,88	0,83	0,86
17	39	1	5	0,98	0,89	0,93
18	37	0	5	1,00	0,88	0,94
19	39	1	6	0,98	0,87	0,92
20	39	2	6	0,95	0,87	0,91
21	42	3	5	0,93	0,89	0,91
22	39	0	6	1,00	0,87	0,93
23	41	2	5	0,95	0,89	0,92
24	43	1	5	0,98	0,90	0,93
25	41	0	6	1,00	0,87	0,93
26	43	0	6	1,00	0,88	0,93
27	34	0	5	1,00	0,87	0,93
28	38	4	7	0,90	0,84	0,87
29	36	1	6	0,97	0,86	0,91
30	47	3	5	0,94	0,90	0,92
31	43	3	5	0,93	0,90	0,91
32	38	0	5	1,00	0,88	0,94
33	34	1	6	0,97	0,85	0,91
34	35	4	6	0,90	0,85	0,88
35	43	2	6	0,96	0,88	0,91

Tabela 26 - Desempenho do sideRETRO durante a genotipagem dos 100 indivíduos simulados, após a remoção dos eventos localizados em regiões altamente repetitivas.

(continuação)

Indivíduo	VP	FP	FN	Precisão	Sensibilidade	F1-*score*
36	41	2	6	0,95	0,87	0,91
37	38	1	6	0,97	0,86	0,92
38	34	1	5	0,97	0,87	0,92
39	39	0	5	1,00	0,89	0,94
40	35	1	5	0,97	0,88	0,92
41	33	1	5	0,97	0,87	0,92
42	39	1	7	0,98	0,85	0,91
43	37	4	7	0,90	0,84	0,87
44	39	4	6	0,91	0,87	0,89
45	35	3	6	0,92	0,85	0,89
46	31	0	5	1,00	0,86	0,93
47	36	0	5	1,00	0,88	0,94
48	40	3	6	0,93	0,87	0,90
49	34	1	5	0,97	0,87	0,92
50	41	4	6	0,91	0,87	0,89
51	34	0	5	1,00	0,87	0,93
52	36	3	5	0,92	0,88	0,90
53	39	2	5	0,95	0,89	0,92
54	47	0	6	1,00	0,89	0,94
55	36	1	5	0,97	0,88	0,92
56	40	2	6	0,95	0,87	0,91
57	41	1	5	0,98	0,89	0,93
58	40	0	5	1,00	0,89	0,94
59	34	3	6	0,92	0,85	0,88
60	35	2	5	0,95	0,88	0,91
61	38	1	5	0,97	0,88	0,93
62	30	1	5	0,97	0,86	0,91
63	38	4	6	0,90	0,86	0,88
64	43	2	5	0,96	0,90	0,92
65	46	1	6	0,98	0,88	0,93
66	41	1	6	0,98	0,87	0,92
67	37	2	5	0,95	0,88	0,91
68	44	5	6	0,90	0,88	0,89
69	36	0	5	1,00	0,88	0,94
70	42	4	7	0,91	0,86	0,88
71	44	3	7	0,94	0,86	0,90

Tabela 26 - Desempenho do sideRETRO durante a genotipagem dos 100 indivíduos simulados, após a remoção dos eventos localizados em regiões altamente repetitivas.

(conclusão)

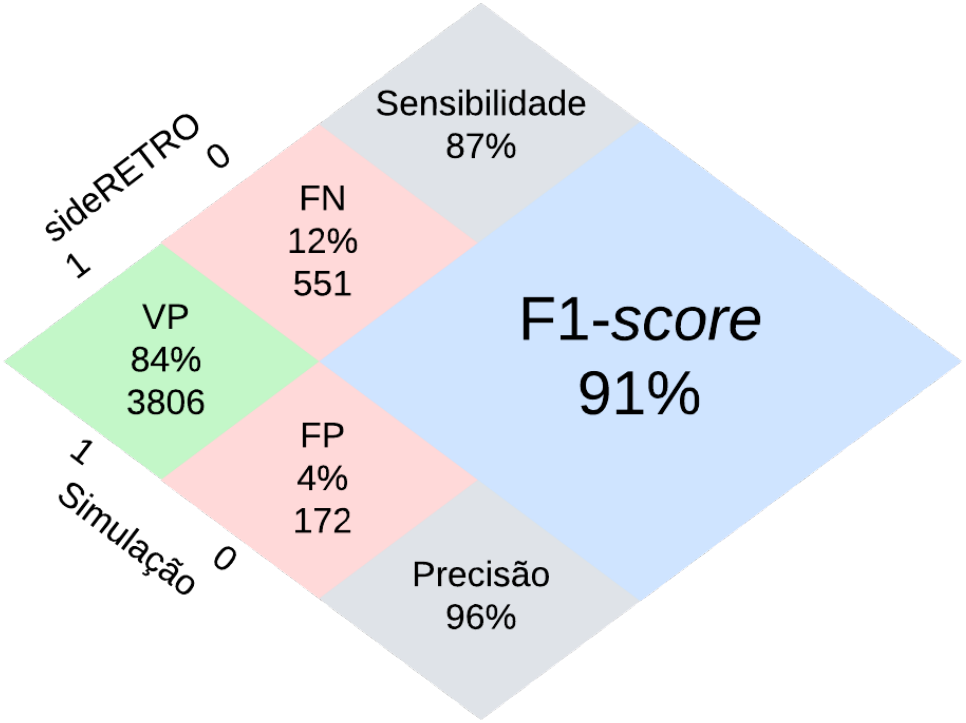
Indivíduo	VP	FP	FN	Precisão	Sensibilidade	F1-*score*
72	41	3	6	0,93	0,87	0,90
73	34	1	5	0,97	0,87	0,92
74	42	1	5	0,98	0,89	0,93
75	37	3	5	0,93	0,88	0,90
76	34	2	5	0,94	0,87	0,91
77	37	3	5	0,93	0,88	0,90
78	38	0	5	1,00	0,88	0,94
79	40	2	5	0,95	0,89	0,92
80	35	0	5	1,00	0,88	0,93
81	40	1	6	0,98	0,87	0,92
82	41	2	7	0,95	0,85	0,90
83	39	2	6	0,95	0,87	0,91
84	40	3	6	0,93	0,87	0,90
85	36	4	5	0,90	0,88	0,89
86	37	4	6	0,90	0,86	0,88
87	32	2	5	0,94	0,86	0,90
88	42	2	7	0,95	0,86	0,90
89	34	1	5	0,97	0,87	0,92
90	41	2	5	0,95	0,89	0,92
91	45	0	6	1,00	0,88	0,94
92	39	2	5	0,95	0,89	0,92
93	39	2	6	0,95	0,87	0,91
94	34	3	5	0,92	0,87	0,89
95	44	4	5	0,92	0,90	0,91
96	36	1	5	0,97	0,88	0,92
97	39	2	5	0,95	0,89	0,92
98	48	0	6	1,00	0,89	0,94
99	40	0	6	1,00	0,87	0,93
Total	3.806	172	551	0,96	0,87	0,91

Fonte: autoria própria.

Foram contados os acertos com relação à detecção do alelo alternativo, então eventos heterozigóticos contam 1 ponto e eventos homozigóticos contam 2 pontos. Para eventos heterozigóticos: detecção do alelo, VP = VP + 1; a não detecção do alelo, FN = FN + 1; a detecção de um evento não simulado FP = FP + 1. Para eventos homozigóticos: detecção dos dois alelos, VP = VP + 2; a não detecção do evento, FN = FN + 2; um alelo detectado, VP = VP + 1 e FN = FN +1; a detecção de um evento não simulado FP = FP + 2.

VP = Verdadeiro Positivo; FP = Falso Positivo; FN = Falso Negativo.

Figura 29 - Desempenho geral do sideRETRO durante a genotipagem dos 100 indivíduos simulados, após a remoção dos eventos localizados em regiões altamente repetitivas.



Fonte: autoria própria.
VP = Verdadeiro Positivo; FP = Falso Positivo; FN = Falso Negativo; 1 = Presença de retroCNV; 0 = Ausência de retroCNV.

5.2 ESTIMANDO O DESEMPENHO DO SIDERETRO EM DADOS REAIS

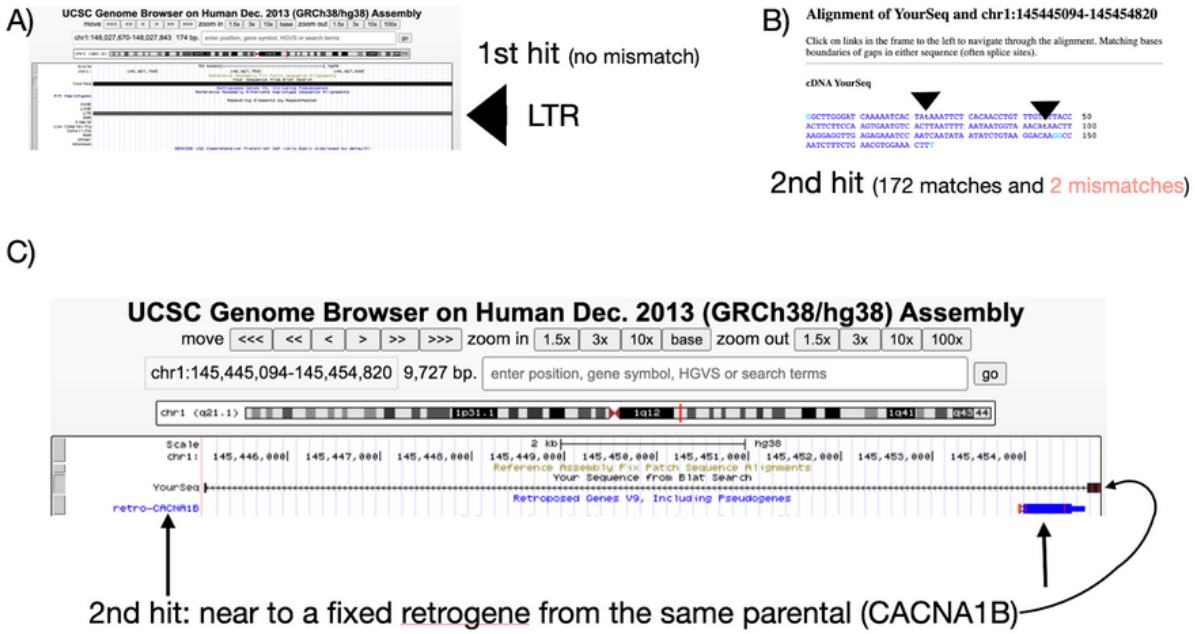
Das seis retroCNVs validadas experimentalmente por PCR por Abyzov e colaboradores (ABYZOV et al., 2013), o sideRETRO anotou cinco. Para esses eventos, a genotipagem conduzida pela ferramenta coincidiu com a mesma conduzida por Abyzov et al. dentro das 14 populações do projeto 1KGP contempladas no seu trabalho.

O evento não anotado pelo sideRETRO corresponde a retroCNV do gene parental CACNA1B, inserido o cromossomo 1, na região entre as posições 148.027.670 e 148.027.843. Curiosamente, Abyzov et al. não encontrou um bom suporte de cobertura de leituras para essa retrocópia, como consta no manuscrito do artigo publicado por ele (ABYZOV et al., 2013). Investigando um pouco mais a fundo, foi observado que a posição genômica chr1:148.027.670-148.027.843 condiz

com uma região de LTR (Figura 30A). Também verificou-se que essa região alinha quase que perfeitamente (duas bases discordantes) noutra região do mesmo cromossomo - chr1:145.445.094-145.454.820 (Figura 30B). Além disso, esse segundo alinhamento está próximo a uma retrocópia fixa do mesmo gene parental, CACNA1B (Figura 30C).

Dadas todas essas observações concernentes ao evento não detectado, pode-se hipotetizar que ele surgiu em decorrência de alinhamentos falsos positivos - que sucedem das dificuldades metodológicas que implicam na distinção de leituras advindas dos éxons do gene parental, de leituras pertinentes à retrocópia fixa e mesmo de regiões repetitivas no genoma (AMEMIYA; KUNDAJE; BOYLE, 2019). No entanto, salienta-se que apenas uma validação experimental adicional poderia confirmar essa hipótese.

Figura 30 - O alinhamento genômico da região da retroCNV do gene parental CACNA1B.



Fonte: <http://genome.ucsc.edu> (KENT et al., 2002).
A) o alinhamento genômico do intervalo definido como ponto de inserção da retroCNV do gene parental CACNA1B se encontra dentro de uma região de LTR; B) o segundo alinhamento da sequência no genoma (apenas 2 bases discordantes em 174 bases totais); C) o segundo alinhamento no genoma está próximo de uma retrocópia fixa do mesmo gene parental, CACNA1B.
A versão 38 do genoma de referência foi usada para as análises.

Desta forma, o sideRETRO apresentou uma correspondência na anotação dos eventos de retroCNV variando de cinco sextos (considerando todos os eventos) a cinco quintos (excluindo o candidato suspeito do gene parental CACNA1B) contra

um conjunto de dados experimentais de um grupo independente. Vale destacar também que houve uma concordância total entre a genotipagem pelo sideRETRO e por Abyzov et al. desses cinco eventos nas 14 populações do projeto 1KGP.

5.3 DISCUSSÃO GERAL ACERCA DO SIDERETRO

Os eventos somáticos e polimórficos de retroduplicação de mRNAs maduros de genes codificadores têm sido subaproveitados nas análises populacionais evolutivas, assim como subexplorados em estudos patológicos - por exemplo, naqueles que gerem sequenciamento tumoral exônico e genômico (CASOLA; BETRÁN, 2017). Muito disso se deveu à falta de metodologias padronizadas de detecção de retroCNVs, tornando-se o fator limitante no entendimento do papel funcional dos pseudogenes processados não fixados na pesquisa básica e translacional. Com o intuito de preencher essa lacuna metodológica, foi desenvolvido o programa computacional sideRETRO.

O sideRETRO foi pensado desde de sua concepção para ser uma ferramenta simples de instalar e de usar. O código-fonte do programa encontra-se publicamente disponível no sítio do *github* (GITHUB, 2022) sob a licença livre GPLv3. Qualquer grupo de pesquisa está autorizado a baixar, modificar (se assim o desejar), compilar, instalar e fazer uso do executável *sider* em suas pesquisas. O *sider* requer, como entrada, arquivos de alinhamento genômico e anotação do genoma de referência nos formatos já conhecidos e amplamente usados pela comunidade científica: os formatos SAM, BAM ou CRAM para os alinhamentos e os formatos GTF ou GFF3 para a anotação do transcriptoma. A saída, após o término das sequências analíticas do *sider*, é um arquivo no formato VCF, outrossim utilizado extensivamente pelos pesquisadores para disponibilizar variações estruturais do genoma. Dessa maneira, foi atingido o intuito primordial de se fornecer um programa computacional fácil de ser manejado e que esteja dentro das boas práticas já preestabelecidas na área da bioinformática.

Outro ponto importante concerne o desempenho e a confiabilidade da ferramenta. Em vista disso, o sideRETRO foi testado contra dados simulados quanto a uma série de variáveis controladas, como a detecção das retroCNVs, a posição

genômica e a orientação (fita) de integração dos eventos e a identificação do alelo alternativo. Assim, obtiveram-se os seguintes resultados para a simulação de sequenciamento de 100 genomas com cobertura de 20 vezes:

a) de 86 retroCNVs curadas[38], 79 foram identificadas na coorte de 100 indivíduos (genomas simulados), aproximadamente 92% dos eventos detectados;
b) quanto à identificação do alelo alternativo[39], obteve-se uma sensibilidade de 87%, com 12% dos alelos não identificados (falsos negativos). A precisão foi de 96%, com um número baixo de falsos positivos, 4%, os quais não correspondem, neste caso, a retroCNVs não simuladas, mas sim a eventos heterozigóticos sendo qualificados como homozigóticos;
c) um erro quadrático mediano de uma base para a detecção do ponto de inserção das retroCNVs;
d) só não foi detectada a orientação da retroCNV em 1 caso dentre os 79 possíveis. Para os demais 78, a orientação de inserção do evento foi corretamente assinalada;
e) com relação à frequência de eventos, o sideRETRO identificou tão bem eventos fixados (84% detectados), quanto aqueles menos frequentes, como os polimórficos e somáticos (72% e 76% detectados respectivamente).

Testou-se também o algoritmo contra dados reais gerados por um grupo independente. Para isso, foi escolhido o trabalho de Abyzov e colaboradores (2013), no qual os pesquisadores validaram experimentalmente nove retroCNVs por PCR e, para seis delas, encontraram o ponto de inserção. Esses eventos foram, então, genotipados em 974 indivíduos de 14 populações do 1KGP. O sideRETRO detectou cinco das seis retroCNVs validadas, sendo que o único evento perdido possui algumas características suspeitas (Figura 30) que poderiam evidenciar um caso de falso positivo por parte das análises de Abyzov et al. Para as demais cinco retrocópias não fixadas, houve uma correspondência de 100% na genotipagem feita pelo grupo de Abyzov e pelo sideRETRO.

[38] Das 21 retroCNVs não identificadas, 14 foram removidas por estarem em regiões altamente repetitivas do genoma (Tabela 25).
[39] Das 21 retroCNVs não identificadas, 14 foram removidas por estarem em regiões altamente repetitivas do genoma (Tabela 26).

Dadas as evidências de fidedignidade apresentadas pelos testes da metodologia, o sideRETRO passou a ser adotado em diferentes frentes de trabalho dentro do laboratório de bioinformática do Hospital Sírio-Libanês em São Paulo. Um caso translacional a ser citado foi a Análise de retroelementos em um contexto de progressão tumoral (REGO, 2021). Parte do estudo consistiu no sequenciamento genômico de amostras pareadas (tecido normal, tecido tumoral) de cinco pacientes de câncer colorretal. Foi rodado o sideRETRO sobre os dados de NGS, já previamente alinhados contra o genoma de referência, e, subsequentemente, foram identificados eventos somáticos para todos os indivíduos. Observou-se que o número de retroCNVs aumenta de acordo com a progressão tumoral (Figura 31), então há mais retroduplicações em metástase do que em tumor primário, que, por sua vez, tem mais do que em tecido normal (Figura 32). Outro achado importante é a inserção de eventos somáticos (intragênicos) em genes já descritos na literatura por estarem relacionados ao controle e à progressão do câncer colorretal, como: ALK, EPS15, FHIT, FOXP1, MSH6, NCOR2, PTPRT, ZNF521 (TATE et al., 2019). Fato esse que poderia estar relacionado à piora no prognóstico desse tipo tumoral.

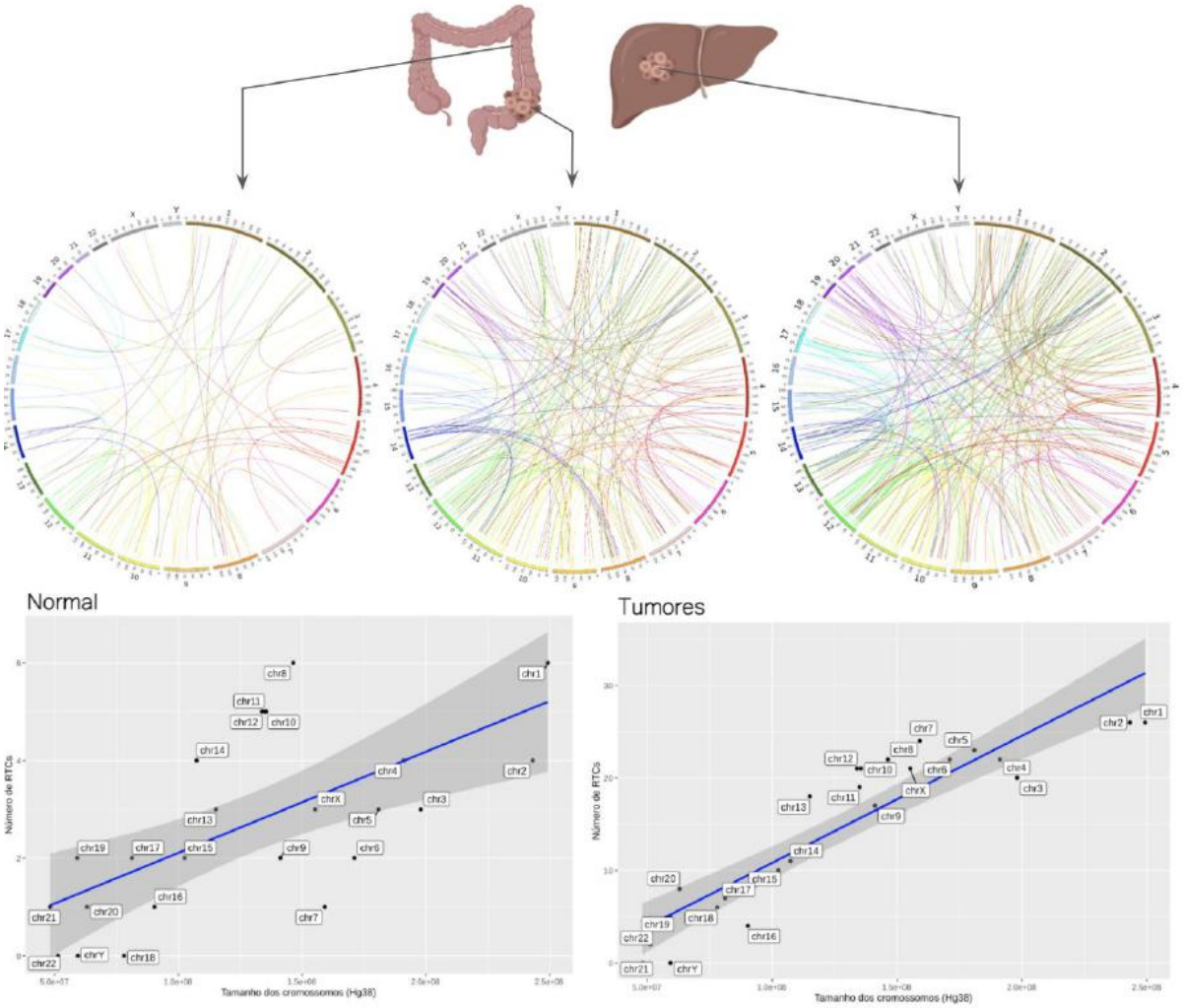
O sideRETRO foi o primeiro programa de bioinformática inteiramente dedicado à descoberta e à caracterização de retrocópias somáticas e polimórficas, o qual teve a sua metodologia publicada na revista científica britânica *Bioinformatics* (MILLER et al., 2021) - vide Apêndice A. Em 2021, mais duas ferramentas para anotação de retroCNVs foram publicadas: *RetroScan* (WEI et al., 2021) e *PΨFinder* (ABRAHAMSSON et al., 2022).

O *RetroScan* é um programa que usa alinhamento de sequências de proteínas com sequências de genoma para identificação e visualização de retrocópias. Os autores citam o sideRETRO durante a discussão de seu artigo, salientando as vantagens de se utilizá-lo para análises de eventos somáticos e polimórficos nas populações, enquanto que o *RetroScan* contribuiria para a pesquisa de retrocópias em organismos individuais.

O *PΨFinder* também é um programa de identificação e visualização de retrocópias que, assim como o sideRETRO, usa alinhamentos anormais. Os autores testaram o *PΨFinder* contra o sideRETRO e obtiveram resultados similares, concernentes ao desempenho e à determinação da posição genômica de inserção dos eventos. No entanto, o *PΨFinder* não anota outras informações úteis acerca da retroduplicação, como a orientação da retroCNV, o contexto de inserção e a

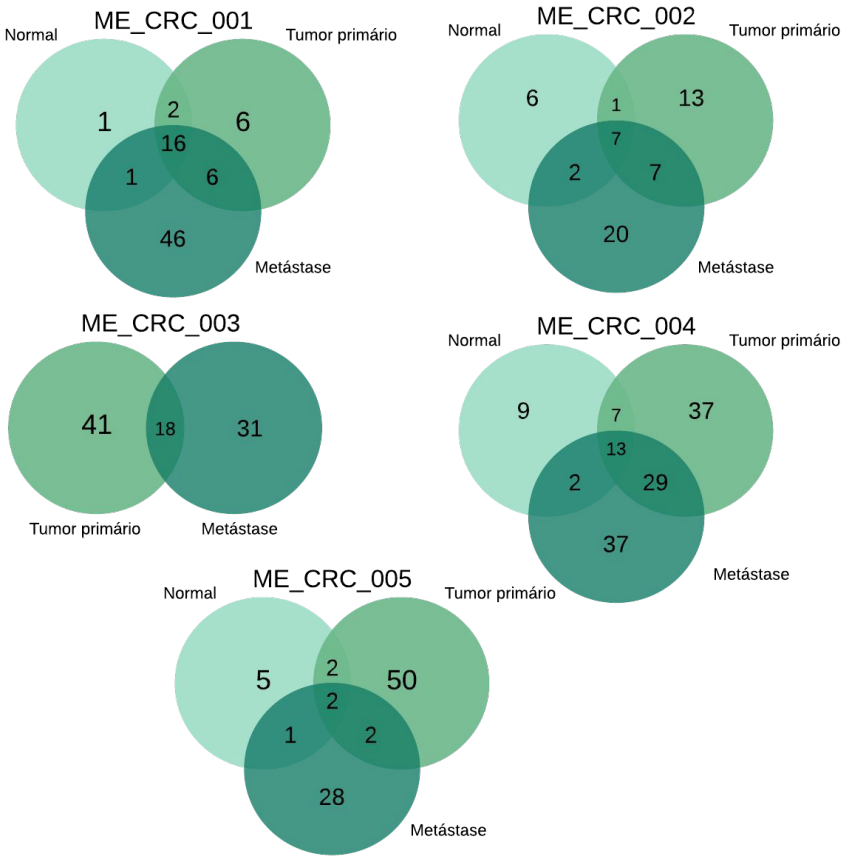
genotipagem com a haplotipagem (características anotadas de forma automática pelo sideRETRO). Outro ponto importante levantado pelos próprios autores da ferramenta é que o *PΨFinder* não usa um formato padrão para disponibilizar suas análises, focando nas facilidades do resultado visual que gera, ao passo que o sideRETRO reforça as boas práticas da área da bioinformática, ao fornecer como saída um arquivo no formato já conhecido, VCF.

Figura 31 - Análise de retroelementos em um contexto de progressão tumoral. Eventos de retroCNVs, ao longo dos cromossomos do genoma humano, de amostras normais e tumorais.



Fonte: (REGO, 2021).
(A) *circosplot* com os eventos por tipo de amostras (normais, tumores primários, metástases); (B) correlação entre o número de inserções de retroCNVs e o tamanho dos cromossomos do genoma humano (hg38) para tecidos normais e tumorais.

Figura 32 - Análise de retroelementos em um contexto de progressão tumoral. Número de retroCNVs (únicos) aumentam com a progressão do câncer colorretal.



Fonte: (REGO, 2021).
Número de eventos únicos e compartilhados entre as amostras normais (verde mais claro), as amostras de tecido tumoral primário (verde) e as amostras metastáticas (verde escuro). Para todos os pacientes analisados as amostras normais apresentaram um número menos que as amostras de tumores primários que, por sua vez, apresentaram um número menor de eventos de retroCNVs que as amostras de tumores secundários (metástases).

Ainda em 2021, os pesquisadores Stepanka Zverinova e Victor Guryev publicaram uma lista curada com os programas padrão ouro para a detecção de variações estruturais no DNA. O sideRETRO foi a única ferramenta adicionada por eles para detecção e genotipagem de retroCNVs (ZVERINOVA; GURYEV, 2021). Tudo isso mostra que, não obstante a sua recente publicação, o sideRETRO tem sido útil e recebido reconhecimento da comunidade científica, inclusive de grupos que poderiam ser denominados concorrentes.

6 Conclusão

6 CONCLUSÃO

Os pseudogenes processados polimórficos, somáticos ou germinativos (também conhecidos por retroCNVs), quer em ciência básica, quer em ciência aplicada, encontram-se subexplorados, sendo, por conseguinte, desconsiderados em análises evolutivas e translacionais. Isso se dá, parcialmente, em consequência da falta de metodologias computacionais dedicadas a sua identificação e anotação em dados de sequenciamento de DNA genômico. De modo a contribuir para a mudança desse *status quo*, foi desenvolvido o sideRETRO.

O sideRETRO é uma ferramenta de bioinformática especializada na detecção de pseudogenes processados não fixados no genoma de referência, mas presentes em indivíduos sequenciados por WGS e WES. O sideRETRO é capaz de determinar o gene parental do evento, a sua coordenada genômica de integração (com o cromossomo, o ponto de inserção e a orientação da fita), o contexto genômico do sítio onde se sucedeu a mobilização (evento intragênico ou intergênico), a genotipagem e a haplotipagem (se o evento está em homozigose ou heterozigose).

De modo a testar a confiabilidade da metodologia ora apresentada, foi rodado o sideRETRO com dados simulados. Foram detectados 92% dos eventos de retrotransposição simulados e curados com uma precisão e uma sensibilidade de 96% e 87%, respectivamente, para a identificação do alelo alternativo. Aferiu-se também o desempenho do algoritmo, agora, para dados reais de um grupo independente. A análise contemplou seis retroCNVs validadas experimentalmente e genotipadas em 974 indivíduos de 14 populações do 1KGP. O sideRETRO encontrou cinco das seis retroCNVs, com uma correspondência de 100% concernente à genotipagem.

Não obstante alguns grupos de pesquisa já terem publicado estudos que abrangessem pseudogenes processados não fixados, salienta-se o fato de o sideRETRO ter sido o primeiro programa de bioinformática dedicado à descoberta e anotação de retroCNVs, o qual permite que quaisquer pesquisadores, independentemente de seu grau de especialização em variações estruturais do genoma, possam analisar os seus dados de NGS em busca de retrocópias polimórficas de genes codificadores.

Um fato interessante a ser mencionado é que a metodologia do sideRETRO foi desenvolvida para ser um protótipo genérico de identificação de elementos que se mobilizam no DNA através de um intermediário de RNA. Portanto, outros retrotransposons, como L1Hs, *AluY*, HERVK, SVA (para citar os principais em humanos), poderiam, outrossim, serem detectados e anotados pelo sideRETRO, desde que se fizessem as devidas alterações algorítmicas para que o programa procurasse por esses transposons nas anotações do transcriptoma, tal qual procura por éxons de genes codificadores durante as análises de retroCNV. No entanto, é necessário frisar que este programa foi ostensivamente testado e validado em contextos de inserção de retrocópias de genes codificadores, de modo que o uso do sideRETRO com relação a outros retrotransposons carece de ser corroborado.

Com o crescente aprimoramento e a popularização dos métodos de sequenciamento de terceira geração, uma futura atualização que vise aperfeiçoar a metodologia deste projeto deve levar em consideração o uso de leituras longas. Uma sugestão é um algoritmo que use de modo híbrido leituras curtas e leituras longas para a montagem *de novo* do sítio contendo a retrocópia polimórfica.

É chegado, pois, o término desta tese. O sideRETRO mostrou-se uma ferramenta capaz de identificar e anotar pseudogenes processados não fixados em dados de NGS, preenchendo, dessa forma, uma lacuna metodológica no campo de chamada de variante. O seu uso tem o potencial de auxiliar os cientistas na elucidação do papel das retroCNVs nos contextos evolutivo e translacional. Bom uso e sideRETRO!

Referências

REFERÊNCIAS[40]

1000 GENOMES PROJECT CONSORTIUM et al. A map of human genome variation from population-scale sequencing. **Nature**, v. 467, n. 7319, p. 1061–1073, 28 out. 2010.

ABDEL-HALEEM, H. The Origins of Genome Architecture. **The Journal of heredity**, v. 98, n. 6, p. 633–634, 28 ago. 2007.

ABRAHAMSSON, S. et al. PΨFinder: a practical tool for the identification and visualization of novel pseudogenes in DNA sequencing data. **BMC bioinformatics**, v. 23, n. 1, p. 59, 3 fev. 2022.

ABYZOV, A. et al. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. **Genome Research**, 2013. Disponível em: <http://dx.doi.org/10.1101/gr.154625.113>

ADAMS, J. W. et al. A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. **Nucleic acids research**, v. 8, n. 24, p. 6113–6128, 20 dez. 1980.

AGUILAR, L. **Genes, Genomes, Genetics and Chromosomes**. [s.l.] Scientific e-Resources, 2019.

AMEMIYA, H. M.; KUNDAJE, A.; BOYLE, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. **Sci Rep**, v. 9, n. 1, 2019. Disponível em: <https://doi.org/10.1038/s41598-019-45839-z>

BAERTSCH, R. et al. Retrocopy contributions to the evolution of the human genome. **BMC Genomics**, 2008. Disponível em: <http://dx.doi.org/10.1186/1471-2164-9-466>

BAKSHI, A. et al. DNA methylation variation of human-specific Alu repeats. **Epigenetics:** official journal of the DNA Methylation Society, v. 11, n. 2, p. 163–173, 18 fev. 2016.

BASAME, S. et al. Spatial Assembly and RNA Binding Stoichiometry of a LINE-1 Protein Essential for Retrotransposition. **Journal of Molecular Biology**, 2006. Disponível em: <http://dx.doi.org/10.1016/j.jmb.2005.12.063>

BECK, C. R. et al. LINE-1 retrotransposition activity in human genomes. **Cell**, v. 141, n. 7, p. 1159–1170, 25 jun. 2010.

BECK, C. R. et al. LINE-1 elements in structural variation and disease. **Annual review of genomics and human genetics**, v. 12, p. 187–215, 2011.

[40] De acordo com a Associação Brasileira de Normas Técnicas (ABNT NBR 6023).

BEGG, C. E.; DELIUS, H.; LEADER, D. P. Duplicated region of the mouse genome containing a cytoplasmic gamma-actin processed pseudogene associated with long interspersed repetitive elements. **Journal of molecular biology**, v. 203, n. 3, p. 677–687, 5 out. 1988.

BENNETT, E. A. et al. Active Alu retrotransposons in the human genome. **Genome Research**, 2008. Disponível em: <http://dx.doi.org/10.1101/gr.081737.108>

BONFIELD, J. K. et al. HTSlib: C library for reading/writing high-throughput sequencing data. **GigaScience**, v. 10, n. 2, 16 fev. 2021.

BRIDGES, B. Rosalind Franklin. The Dark Lady of DNA: Brenda Maddox, HarperCollins**,** ISBN 0-00-257149-8. **DNA Repair**, v. 2, n. 3, p. 359–360, 2003. Disponível em: <http://dx.doi.org/10.1016/s1568-7864(02)00244-6>

BROUHA, B. et al. Hot L1s account for the bulk of retrotransposition in the human population. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 9, p. 5280–5285, 29 abr. 2003.

BURKI, F.; KAESSMANN, H. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. **Nature Genetics**, 2004. Disponível em: <http://dx.doi.org/10.1038/ng1431>

BURNS, K. H. Transposable elements in cancer. **Nature reviews. Cancer**, v. 17, n. 7, p. 415–424, jul. 2017.

CANCER GENOME ATLAS RESEARCH NETWORK et al. The Cancer Genome Atlas Pan-Cancer analysis project. **Nature genetics**, v. 45, n. 10, p. 1113–1120, out. 2013.

CANNATA, N.; MERELLI, E.; ALTMAN, R. B. Time to Organize the Bioinformatics Resourceome. **PLoS Computational Biology**, 2005. Disponível em: <http://dx.doi.org/10.1371/journal.pcbi.0010076>

CARELLI, F. N. et al. The life history of retrocopies illuminates the evolution of new mammalian genes. **Genome research**, v. 26, n. 3, p. 301–314, mar. 2016.

CASOLA, C.; BETRÁN, E. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? **Genome biology and evolution**, v. 9, n. 6, p. 1351–1373, 1 jun. 2017.

CHACON, S.; STRAUB, B. **Pro git**. [s.l.] Springer Nature, 2014.

CHARGAFF, E. Building the Tower of Babble. **Nature**, v. 248, n. 5451, p. 776–779, 1974. Disponível em: <http://dx.doi.org/10.1038/248776a0>

COCK, P. J. A. et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. **Nucleic acids research**, v. 38, n. 6, p. 1767–1771, abr. 2010.

CODD, E. F. A Relational Model of Data for Large Shared Data Banks. Em: BROY, M.; DENERT, E. (Eds.). **Software Pioneers:** Contributions to Software Engineering. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. p. 263–294.

CORDAUX, R.; BATZER, M. A. The impact of retrotransposons on human genome evolution. **Nature reviews. Genetics**, v. 10, n. 10, p. 691–703, out. 2009.

COST, G. J. Human L1 element target-primed reverse transcription in vitro. **The EMBO Journal**, 2002. Disponível em: <http://dx.doi.org/10.1093/emboj/cdf592>

CREAGER, A. N. H.; MORGAN, G. J. **After the Double HelixIsis**, 2008. Disponível em: <http://dx.doi.org/10.1086/588626>

CRICK, F. Central dogma of molecular biology. **Nature**, v. 227, n. 5258, p. 561–563, 8 ago. 1970.

CRICK, F. The double helix: a personal view. **Nature**, v. 248, n. 5451, p. 766–769, 26 abr. 1974.

CRISCIONE, S. W. et al. Genome-wide characterization of human L1 antisense promoter-driven transcripts. **BMC genomics**, v. 17, p. 463, 14 jun. 2016.

DATE, C. J. **A guide to the SQL standard (2nd ed.)**. USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

DEININGER, P. Alu elements: know the SINEs. **Genome biology**, v. 12, n. 12, p. 236, 28 dez. 2011.

DENLI, A. M. et al. Primate-specific ORF0 contributes to retrotransposon-mediated diversity. **Cell**, v. 163, n. 3, p. 583–593, 22 out. 2015.

EMBL-EBI. **GFF/GTF File Format - Definition and supported options**. Disponível em: <http://dec2021.archive.ensembl.org/info/website/upload/gff.html>. Acesso em: 1 de jan. de 2022.

EMBL-EBI. **Using data from IGSR**. Disponível em: <https://www.internationalgenome.org/data>. Acesso em: 5 de jan. de 2022.

ENGELS, W. R.; PRESTON, C. R. Identifying P factors in Drosophila by means of chromosome breakage hotspots. **Cell**, v. 26, n. 3 Pt 1, p. 421–428, nov. 1981.

ENTREZ HELP [INTERNET]. BETHESDA (MD). **Entrez Help**. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK3837/>. Acesso em: 1 de fev. de 2022.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. **KDD'96**: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.

EWING, A. D. et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. **Genome Biology**, 2013. Disponível em: <http://dx.doi.org/10.1186/gb-2013-14-3-r22>

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. **Genome research**, v. 8, n. 3, p. 186–194, mar. 1998.

FEDOROFF, N. V. Barbara McClintock, 16 June 1902 - 2 September 1992. **Biographical memoirs of fellows of the Royal Society. Royal Society** , v. 40, p. 265–280, 1 nov. 1994.

FEDOROFF, N. V. McClintock's challenge in the 21st century. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 50, p. 20200–20203, 11 dez. 2012.

FENG, Q. et al. Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition. **Cell**, 1996. Disponível em: <http://dx.doi.org/10.1016/s0092-8674(00)81997-2>

FESCHOTTE, C. Review of Mobile DNA - finding treasure in junk by Haig H Kazazian. **Mobile DNA**, 2012. Disponível em: <http://dx.doi.org/10.1186/1759-8753-3-16>

FIELLER, E. C.; HARTLEY, H. O.; PEARSON, E. S. Tests for Rank Correlation Coefficients. I. **Biometrika**, v. 44, n. 3/4, p. 470–481, 1957.

FIRTH, A. E.; BROWN, C. M. Detecting overlapping coding sequences with pairwise alignments. **Bioinformatics** , v. 21, n. 3, p. 282–292, 1 fev. 2005.

FOLEY, B. T. et al. **HIV Sequence Compendium 2018**. [s.l.] Los Alamos National Laboratory (LANL), 27 jun. 2018. Disponível em: <http://www.osti.gov/servlets/purl/1458915/>

FRANKISH, A. et al. GENCODE reference annotation for the human and mouse genomes. **Nucleic acids research**, v. 47, n. D1, p. D766–D773, 8 jan. 2019.

GALANTE, P. A. F. et al. Sense-antisense pairs in mammals: functional and evolutionary considerations. **Genome biology**, v. 8, n. 3, p. R40, 2007.

GCC TEAM. **GCC**. Disponível em: <https://gcc.gnu.org/onlinedocs/>. Acesso em: 5 de jan. de 2022.

GITHUB. **GitHub**, 2022. Disponível em: <https://github.com/>

GNU General Public License, version 3. Disponível em: <http://www.gnu.org/licenses/gpl.html>

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. **Nature reviews. Genetics**, v. 17, n. 6, p. 333–351, 17 maio 2016.

HANCKS, D. C.; KAZAZIAN, H. H., Jr. Roles for retrotransposon insertions in human disease. **Mobile DNA**, v. 7, p. 9, 6 maio 2016.

HESS, J. F. et al. Library preparation for next generation sequencing: A review of automation strategies. **Biotechnology advances**, v. 41, p. 107537, jul. 2020.

HIPP, R. D. **SQLite**. [s.l: s.n.]. Diponível em: <https://www.sqlite.org>. Acesso em: 5 de jan. de 2022.

HOHJOH, H.; SINGER, M. F. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. **The EMBO journal**, v. 15, n. 3, p. 630–639, 1 fev. 1996.

HUNTZINGER, E.; IZAURRALDE, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. **Nature reviews. Genetics**, v. 12, n. 2, p. 99–110, fev. 2011.

ISHIGURO, K. et al. Establishment of a genome-wide and quantitative protocol for assessment of transcriptional activity at human retrotransposon L1 antisense promoters. **Genes & genetic systems**, v. 92, n. 5, p. 243–249, 10 abr. 2018.

JANUSZYK, K. et al. Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. **The Journal of biological chemistry**, v. 282, n. 34, p. 24893–24904, 24 ago. 2007.

JURKA, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. **Proceedings of the National Academy of Sciences**, 1997. Disponível em: <http://dx.doi.org/10.1073/pnas.94.5.1872>

JURKA, J. Repeats in genomic DNA: mining and meaning. **Current opinion in structural biology**, v. 8, n. 3, p. 333–337, jun. 1998.

KABZA, M.; CIOMBOROWSKA, J.; MAKAŁOWSKA, I. RetrogeneDB—A Database of Animal Retrogenes. **Molecular Biology and Evolution**, 2014. Disponível em: <http://dx.doi.org/10.1093/molbev/msu139>

KAESSMANN, H.; VINCKENBOSCH, N.; LONG, M. RNA-based gene duplication: mechanistic and evolutionary insights. **Nature reviews. Genetics**, v. 10, n. 1, p. 19–31, jan. 2009.

KAZAZIAN, H. H., Jr. Mobile elements: drivers of genome evolution. **Science**, v. 303, n. 5664, p. 1626–1632, 12 mar. 2004.

KENT, W. J. et al. The human genome browser at UCSC. **Genome research**, v. 12, n. 6, p. 996–1006, jun. 2002.

KERNIGHAN, B. W.; RITCHIE, D. M. **The C Programming Language**. [s.l.] Pearson Educación, 1988.

KHAN, H.; SMIT, A.; BOISSINOT, S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. **Genome research**, v. 16, n. 1, p. 78–87, jan. 2006.

KHAZINA, E.; WEICHENRIEDER, O. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. **Proceedings of the National Academy of Sciences**, 2009. Disponível em: <http://dx.doi.org/10.1073/pnas.0809964106>

KONING, A. P. J. DE et al. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. **PLoS Genetics**, 2011. Disponível em: <http://dx.doi.org/10.1371/journal.pgen.1002384>

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 15 fev. 2001.

LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome biology**, v. 10, n. 3, p. R25, 4 mar. 2009.

LEE, E. et al. Landscape of somatic retrotransposition in human cancers. **Science**, v. 337, n. 6097, p. 967–971, 24 ago. 2012.

LEVY, S. E.; MYERS, R. M. Advancements in Next-Generation Sequencing. **Annual review of genomics and human genetics**, v. 17, p. 95–115, 31 ago. 2016.

LI, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. **Bioinformatics** , 2011.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows–Wheeler transform. **Bioinformatics** , v. 25, n. 14, p. 1754–1760, 18 maio 2009.

LI, W.; FREUDENBERG, J. Mappability and read length. **Frontiers in genetics**, v. 5, p. 381, 10 nov. 2014.

MANDAL, P. K.; KAZAZIAN, H. H. SnapShot: Vertebrate Transposons. **Cell**, 2008. Disponível em: <http://dx.doi.org/10.1016/j.cell.2008.09.028>

MARDIS, E. R. Next-generation sequencing platforms. **Annual review of analytical chemistry** , v. 6, p. 287–303, 2013.

MARQUES, A. C. et al. Emergence of young human genes after a burst of retroposition in primates. **PLoS biology**, v. 3, n. 11, p. e357, nov. 2005.

MARTIN, S. L. Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. **Molecular and cellular biology**, v. 11, n. 9, p. 4804–4807, set. 1991.

MARTIN, S. L. et al. Trimeric structure for an essential protein in L1 retrotransposition. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 24, p. 13815–13820, 25 nov. 2003.

MARTIN, S. L. et al. The structures of mouse and human L1 elements reflect their insertion mechanism. **Cytogenetic and Genome Research**, 2005. Disponível em: <http://dx.doi.org/10.1159/000084956>

MARTIN, S. L. Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. **RNA biology**, v. 7, n. 6, p. 706–711, nov. 2010.

MARTIN, S. L.; LI, J.; WEISZ, J. A. Deletion analysis defines distinct functional domains for protein-protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1. **Journal of molecular biology**, v. 304, n. 1, p. 11–20, 17 nov. 2000.

MASLOW, A. H. **The Psychology of Science a Reconnaissance**. [s.l.] Harper & Row, 1966.

MASTORODEMOS, V. et al. Human GLUD1 and GLUD2 glutamate dehydrogenase localize to mitochondria and endoplasmic reticulum. **Biochemistry and cell biology**, v. 87, n. 3, p. 505–516, jun. 2009.

MCCLINTOCK, B. The origin and behavior of mutable loci in maize. **Proceedings of the National Academy of Sciences**, 1950. Disponível em: <http://dx.doi.org/10.1073/pnas.36.6.344>

MCGRAYNE, S. B. **Nobel Prize women in science:** their lives, struggles, and momentous discoveries. Secaucus, N.J.: Carol Publishing Group, 2001.

MCKENNA, M. C.; FERREIRA, G. C. Enzyme Complexes Important for the Glutamate–Glutamine Cycle. **Advances in Neurobiology**, 2016. Disponível em: <http://dx.doi.org/10.1007/978-3-319-45096-4_4>

MILLER, T. L. A. **galantelab/sandy:** Release v0.23. Zenodo, 2019. Disponível em: <https://doi.org/10.5281/zenodo.2589575>

MILLER, T. L. A. et al. sideRETRO: a pipeline for identifying somatic and polymorphic insertions of processed pseudogenes or retrocopies. **Bioinformatics** , v. 37, n. 3, p. 419–421, 20 abr. 2021.

MORGAN, T. H. Sex limited inheritance in Drosophila. **Science**, v. 32, n. 812, p. 120–122, 22 jul. 1910.

MRC LABORATORY OF MOLECULAR BIOLOGY. **Rosalind Franklin**. From the personal collection of Jenifer Glynn, 1955. Disponível em: <https://commons.wikimedia.org/wiki/File:Rosalind_Franklin.jpg>

NATIONAL LIBRARY OF MEDICINE. **BLAST topics**. Disponível em: <https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp>. Acesso em: 5 de jan. de 2022.

NAVARRO, F. C. P.; GALANTE, P. A. F. RCPedia: a database of retrocopied genes. **Bioinformatics**, 2013. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btt104>

NAVARRO, F. C. P.; GALANTE, P. A. F. A Genome-Wide Landscape of Retrocopies in Primate Genomes. **Genome biology and evolution**, v. 7, n. 8, p. 2265–2275, 29 jul. 2015.

PAN, D.; ZHANG, L. Tandemly arrayed genes in vertebrate genomes. **Comparative and functional genomics**, p. 545269, 2008.

PEARSON, H. Human genome done and dusted. **Nature**, 14 abr. 2003.

PEI, B. et al. The GENCODE pseudogene resource. **Genome biology**, v. 13, n. 9, p. R51, 26 set. 2012.

PISKAREVA, O.; SCHMATCHENKO, V. DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template in vitro. **FEBS letters**, v. 580, n. 2, p. 661–668, 23 jan. 2006.

POLISENO, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. **Nature**, v. 465, n. 7301, p. 1033–1038, 24 jun. 2010.

PROUDFOOT, N. J.; GIL, A.; MANIATIS, T. The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene. **Cell**, v. 31, n. 3 Pt 2, p. 553–563, dez. 1982.

RAVINDRAN, S. Barbara McClintock and the discovery of jumping genes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 50, p. 20198–20199, 11 dez. 2012.

REGO, F. O. R. **Análise de retroelementos em um contexto de progressão tumoral**. Phd—[s.l.] Instituto Sírio-Libanês de Ensino e Pesquisa, 4 mar. 2021.

SAKAI, H. et al. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. **Gene**, v. 389, n. 2, p. 196–203, 2007. Disponível em: <http://dx.doi.org/10.1016/j.gene.2006.11.007>

SASAKI, Y. The truth of the f-measure. **Teach Tutor Mater**, 2007. Disponível em: <https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf>. Acesso em: 26 de maio de 2021.

SCHRIDER, D. R. et al. Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans. **PLoS Genetics**, 2013. Disponível em: <http://dx.doi.org/10.1371/journal.pgen.1003242>

SCHUBERT, E. et al. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. **ACM Trans. Database Syst.**, v. 42, n. 3, p. 1–21, 31 jul. 2017.

SHAPIRO, J. A. Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli. **Journal of molecular biology**, v. 40, n. 1, p. 93–105, 28 fev. 1969.

SINGER, M. F. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. **Cell**, v. 28, n. 3, p. 433–434, mar. 1982.

SMITHSONIAN INSTITUTION - RESTORED BY ADAM CUERDEN. **Barbara McClintock (1902-1992) shown in her laboratory in 1947.** Smithsonian Institution Archives,1947. Disponível em: <https://commons.wikimedia.org/wiki/File:Barbara_McClintock_(1902-1992)_shown_in_her_laboratory_in_1947.jpg>

SPEEK, M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. **Molecular and cellular biology**, v. 21, n. 6, p. 1973–1985, mar. 2001.

SUDMANT, P. H. et al. An integrated map of structural variation in 2,504 human genomes. **Nature**, v. 526, n. 7571, p. 75–81, 1 out. 2015.

SULTANA, T. et al. Integration site selection by retroviruses and transposable elements in eukaryotes. **Nature reviews. Genetics**, v. 18, n. 5, p. 292–308, maio 2017.

SULTANA, T. et al. The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. **Molecular cell**, v. 74, n. 3, p. 555–570.e7, 2 maio 2019.

SZAK, S. T. et al. Molecular archeology of L1 insertions in the human genome. **Genome biology**, v. 3, n. 10, p. research0052, 19 set. 2002.

TATE, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. **Nucleic acids research**, v. 47, n. D1, p. D941–D947, 8 jan. 2019.

TAYLOR, A. L. Bacteriophage-induced mutation in escherichia coli. **Proceedings of the National Academy of Sciences of the United States of America**, v. 50, p. 1043–1051, dez. 1963.

The Meson Build system. Disponível em: <https://mesonbuild.com>. Acesso em: 5 de jan. de 2022.

The Ninja build system. Disponível em: <https://ninja-build.org/manual.html>. Acesso em: 5 de jan. de 2022.

THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP. **Sequence Alignment/Map Format Specification**. Toronto: Global Alliance for Genomics & Health (GA4GH), 3 jun. 2021a. Disponível em: <https://github.com/samtools/hts-specs>

THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP. **The Variant Call Format (VCF) Version 4.2 Specification**. Toronto: Global Alliance for Genomics & Health (GA4GH), 27 jul. 2021b. Disponível em: <https://github.com/samtools/hts-specs>

THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP. **CRAM format specification**. Toronto: Global Alliance for Genomics and Health (GA4GH), 15 out. 2021c. Disponível em: <https://github.com/samtools/hts-specs>

THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP. **The Browser Extensible Data (BED) format**. Toronto: Global Alliance for Genomics and Health (GA4GH), 5 jan. 2022. Disponível em: <https://github.com/samtools/hts-specs>

THOMPSON, K. Programming Techniques: Regular expression search algorithm. **Communications of the ACM**, v. 11, n. 6, p. 419–422, 1 jun. 1968.

TORVALDS, L. **Linux**: a portable operating system. University of Helsinki Department of Computer Science,1997.

VANIN, E. F. Processed pseudogenes: characteristics and evolution. **Annual Review of Genetics**, 1985. Disponível em: <http://dx.doi.org/10.1146/annurev.ge.19.120185.001345>

VAN ROSSUM, G.; DRAKE, F. L. **Python 3 Reference Manual**. [s.l.] CreateSpace Independent Publishing Platform, 2009.

VASIMUDDIN, M. et al. **Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems**. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), p. 314–324, 2019.

VENTER, J. C. et al. The sequence of the human genome. **Science**, v. 291, n. 5507, p. 1304–1351, 16 fev. 2001.

VINCKENBOSCH, N.; DUPANLOUP, I.; KAESSMANN, H. Evolutionary fate of retroposed gene copies in the human genome. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 9, p. 3220–3225, 28 fev. 2006.

VIOLLET, S.; MONOT, C.; CRISTOFARI, G. L1 retrotransposition: The snap-velcro model and its consequences. **Mobile genetic elements**, v. 4, n. 1, p. e28907, 1 jan. 2014.

WATSON, J. D. **The double helix**. London: Weidenfeld & Nicolson, v. 968, 1970.

WATSON, J. D.; JORDAN, E. The Human Genome Program at the National Institutes of Health. **Genomics**, v. 5, n. 3, p. 654–656, out. 1989.

WEI et al. Human L1 Retrotransposition: *cis* Preference versus *trans* Complementation. **Molecular and Cellular Biology**, 2001. Disponível em: <http://dx.doi.org/10.1128/mcb.21.4.1429-1439.2001>

WEI, Z. et al. RetroScan: An Easy-to-Use Pipeline for Retrocopy Annotation and Visualization. **Frontiers in genetics**, v. 12, p. 719204, 16 ago. 2021.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature reviews. Genetics**, v. 8, n. 12, p. 973–982, dez. 2007.

WITKOWSKI, J. The forgotten scientists who paved the way to the double helix. **Nature**, v. 568, n. 7752, p. 308–309, 16 abr. 2019.

ZHANG, J. et al. The International Cancer Genome Consortium Data Portal. **Nature biotechnology**, v. 37, n. 4, p. 367–369, abr. 2019.

ZHANG, Y. et al. Landscape and variation of novel retroduplications in 26 human populations. **PLOS Computational Biology**, 2017. Disponível em: <http://dx.doi.org/10.1371/journal.pcbi.1005567>

ZHANG, Z.; CARRIERO, N.; GERSTEIN, M. Comparative analysis of processed pseudogenes in the mouse and human genomes. **Trends in genetics: TIG**, v. 20, n. 2, p. 62–67, fev. 2004.

ZVERINOVA, S.; GURYEV, V. Variant calling: Considerations, practices, and developments. **Human mutation**, 9 dez. 2021.

Apêndices

APÊNDICE A - Artigo do sideRETRO publicado na revista *Bioinformatics*

Bioinformatics, 37(3), 2021, 419–421
doi: 10.1093/bioinformatics/btaa689
Advance Access Publication Date: 27 July 2020
Applications Note

Genome analysis

sideRETRO: a pipeline for identifying somatic and polymorphic insertions of processed pseudogenes or retrocopies

Thiago L. A. Miller[1,2], Fernanda Orpinelli Rego [1], José Leonel L. Buzzo[1,2] and Pedro A. F. Galante [1,*]

[1]Centro de Oncologia Molecular, Hospital Sírio-Libanês, São Paulo 01308-060, Brazil and [2]Departmento de Bioquímica, Universidade de São Paulo, São Paulo 05508-000, Brazil

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson
Received on January 22, 2020; revised on June 29, 2020; editorial decision on July 21, 2020; accepted on July 23, 2020

Abstract

Motivation: Retrocopies or processed pseudogenes are gene copies resulting from mRNA retrotransposition. These gene duplicates can be fixed, somatically inserted or polymorphic in the genome. However, knowledge regarding unfixed retrocopies (retroCNVs) is still limited, and the development of computational tools for effectively identifying and genotyping them is an urgent need.

Results: Here, we present sideRETRO, a pipeline dedicated not only to detecting retroCNVs in whole-genome or whole-exome sequencing data but also to revealing their insertion sites, zygosity and genomic context and classifying them as somatic or polymorphic events. We show that sideRETRO can identify novel retroCNVs and genotype them, in addition to finding polymorphic retroCNVs in whole-genome and whole-exome data. Therefore, sideRETRO fills a gap in the literature and presents an efficient and straightforward algorithm to accelerate the study of bona fide retroCNVs.

Availability and implementation: sideRETRO is available at https://github.com/galantelab/sideRETRO

Contact: pgalante@mochsl.org.br

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Studies of genetic variability caused by the insertion of long (LINE) and short (SINE) interspersed mobile elements (MEs) have been increasingly recognized as important not only for the study of evolution (Kazazian, 2004) but also that of pathologies (Burns, 2017; Hancks and Kazazian, 2016). In addition to the insertion of MEs, it has been reported that cellular mRNAs are retrocopied as a byproduct of LINE retrotransposition. Humans, other primates, and mice exhibit a similar number of fixed (Navarro and Galante, 2013, 2015; Zhang *et al.*, 2004) and an unclear number of unfixed retrocopies (Abyzov *et al.*, 2013; Ewing *et al.*, 2013; Schrider *et al.*, 2013; Zhang *et al.*, 2017). While the former have been well studied (Kabza *et al.*, 2014; Navarro and Galante, 2013), the latter [usually referred to as retroCNVs (Schrider *et al.*, 2013)] are still underexploited, especially because of the lack of a well-established algorithm for their identification.

Here, we present sideRETRO, a pipeline dedicated to detecting retroCNVs. sideRETRO is a mapping-based algorithm that uses whole-genome sequencing (WGS) or whole-exome sequencing (WXS) data to identify somatic or polymorphic retroCNVs and provides their genomic insertion sites, zygosity, genomic context and parental genes.

2 sideRETRO: description, main features and availability

sideRETRO detects somatic (*de novo*) and polymorphic insertions of retrocopies that are absent in the reference genome (referred to as retroCNVs herein). sideRETRO is written in the C programming language and distributed under the GNU General Public License. It is easy to install via the command line and is also available as a Docker image (see Supplementary Data).

sideRETRO is straightforward to use. It requires only an aligned BAM, SAM or CRAM file (WGS or WXS), a reference for the genome and a reference for the transcriptome to infer the presence of a retroCNV event. In summary, the sideRETRO algorithm follows a sequence of steps. First, it selects two classes of anomalously mapped reads (Fig. 1A): the discordantly aligned paired-end reads (DPE; read pairs aligned out of the expected distance or into distinct chromosomes) and split reads [SR; reads presenting non-continuous

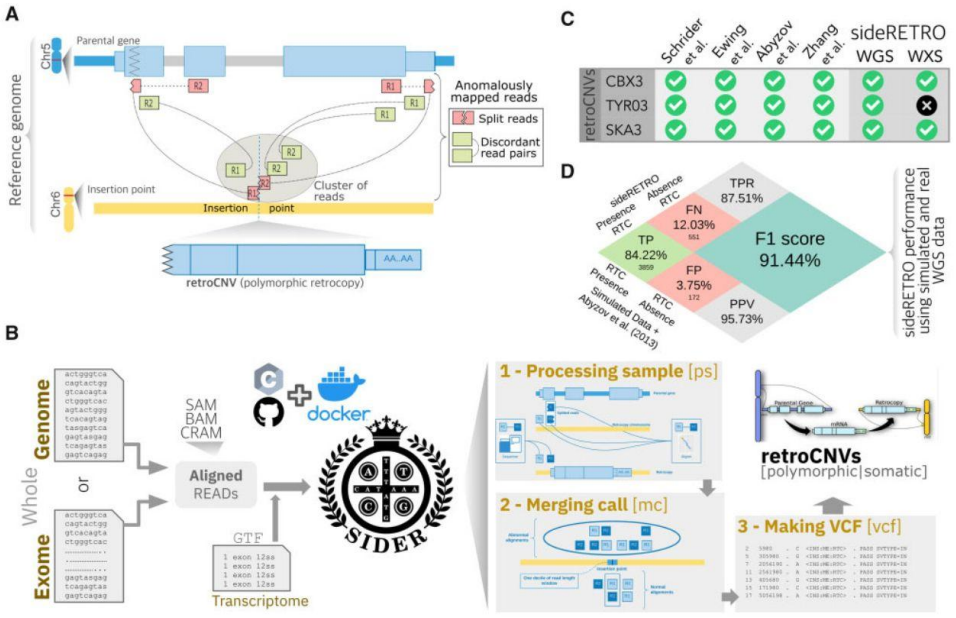


Fig. 1. The sideRETRO pipeline is straightforward to use and accurately identifies retroCNVs. (**A**) sideRETRO uses multiple sources of anomalously mapped reads, such split reads, discordant read pairs and grouping of reads presenting a concordant mapping patter to infer the occurrence of retroCNV events. (**B**) sideRETRO requires as input a (SAM/BAM/CRAM) file containing aligned reads and a reference genome and a reference transcriptome. Next, sideRETRO performs three major steps to identify retroCNVs and several of their characteristics. (**C**) Consensus list of retroCNVs available in the literature (Abyzov *et al.*, 2013; Ewing *et al.*, 2013; Schrider *et al.*, 2013; Zhang *et al.*, 2017) for the individuals used here (NA12778, NA12878, NA18559, NA20754, NA29759). Events identified by sideRETRO using WGS and WXS data are highlighted in green. (**D**) Imbalanced confusion matrix showing the performance of sideRETRO for the simulated and real (whole-genome sequencing) data (retroCNV genotyping in 974 individuos also analyzed in Abyzov *et al.*, 2013). TP, true positive; FP, false positive; FN, false negative; TPR, true positive rate or sensitivity; PPV, positive predictive value or precision; F1-score, harmonic mean of the precision and recall.

(split) alignments]. Second, sideRETRO sub-selects those DPE mapping into an exonic region (parental gene) and its corresponding mate pair mapping elsewhere in the genome (putative insertion point). Third, our algorithm sub-selects SR alignments mapping into an exonic region (parental gene) and its remaining bases mapping into another genomic location (putative insertion point). Finally, sideRETRO, using a Density Based Spatial Clustering of Applications with Noise algorithm [DBSCAN; (Ram *et al.*, 2010)], groups those selected DPE and SR aiming to find a parental gene and an accordant insertion point region characterizing a retroCNV event, Figure 1A. For each detected retroCNV, sideRETRO provides its: (i) parental gene: the gene that underwent the retrotransposition process; (ii) putative genomic insertion site: the genome coordinates at which retrocopy integration occurred (chromosome: start-end); (iii) strandedness: the orientation of retroCNV insertion, on the same (+) or opposite (−) DNA strand compared with its parental gene; (iv) genotype: when multiple genomes (individuals) are analyzed, we annotate the events occurring in each one; and (v) haplotype: whether the event occurred on one (heterozygous) or both (homozygous) homologous chromosomes.

sideRETRO has three subcommands: process-sample, merge-call and make-vcf (Fig. 1B). The process-sample subcommand reads a BAM (or SAM) file and captures abnormal reads that must be related to a retrocopy event. All of these data are saved into a SQLite3 database. In the second step, merge-call, the database is processed to detect and annotate all putative retroCNVs. Finally, the make-vcf step joins information about the identified retroCNVs and produces the sideRETRO output in VCF format (Fig. 1B; see Supplementary Data for further details).

3 sideRETRO: application

To demonstrate how sideRETRO works and its performance, we used both simulated and real datasets. First, we generated a simulated dataset containing 100 human WGS (20× coverage) with 31–45 randomly distributed retroCNVs per genome. In total, we used a list of 100 retroCNVs (Supplementary Table S1). In order to have a trustworthy simulated data to be analyze by sideRETRO, we allowed sequencing error, event insertions even in repetitive sequences (e.g. mobile elements or paralog regions), as well as common retroCNVs (events present in all of the simulated genomes), polymorphic (present in two or more genomes) and somatic retroCNVs (present in only one genome). On average, each simulated retroCNV had at least two exons from the parental gene and was ∼1000 nt in length, which are characteristics that mimic a real retroCNV event (Navarro and Galante, 2015). Considering all simulated retroCNVs: (i) 14 events were randomly inserted into highly repetitive genomic region (Supplementary Table S1); (ii) 33 events were inserted into a LINE or SINE region (Supplementary Table S1); (iii) and the remaining 53 events were inserted in non-repetitive genomic sequences. As expected, we were not able to identify any of the 14 retroCNVs inserted into the highly repetitive genomic regions, but sideRETRO correctly identified 91.9% (79/86) of the remaining events, including 32 retroCNVs inserted into LINE/SINE regions (Supplementary Table S1). No significant identification bias (P-value = 0.908; $\chi^2 = 0.192$. $df = 2$; Supplementary Table S2) was found among the three set of simulated retroCNVs (common, 76% identified; polymorphic, 84% identified; somatic events, 72% identified), indicating that sideRETRO is able to well-identify not only

those known (here represented by common and polymorphic events) but also novel (somatic) retroCNV events. In summary, sideRETRO achieved a good performance using all simulated events (F1-score = 86.16; true positive rate or recall = 78.36%, False positive = 3.42%; Supplementary Fig. S1 and Table S3). By excluding those 14 retroCNVs inserted in highly repetitive regions (unmappable regions), sideRETRO obtained an event better performance (e.g. F1-score = 91.33; true positive rate or recall = 87.35%, False positive = 3.80%, and false negative = 12.18%; Supplementary Fig. S2 and Table S3).

Next, we used sideRETRO to search for retroCNVs in two sets of individuals from 1000 Genomes Projects (1000 Genomes Project Consortium *et al.*, 2015). First, five individuals with WGS and WXS data available were used. In the WGS data, we identified 20 retroCNVs in total (Supplementary Table S4). As expected, due to the sequencing of a target region, in the WXS data from the same individuals, we identified fewer candidates: six retroCNVs (Supplementary Table S4). We also matched these findings to retroCNVs available in the literature (Abyzov *et al.*, 2013; Ewing *et al.*, 2013; Schrider *et al.*, 2013; Zhang *et al.*, 2017), which are three consensus and experimentally validated events for these five individuals. sideRETRO successfully identified all of the retroCNVs with WGS data and missed only one candidate when exome data were used (Fig. 1C). Second, we checked sideRETRO performance in identifying and genotyping retroCNVs in a large cohort (∼1000 individuals) already validated in the literature. Abyzov *et al.* (2013) performed PCR validation and found an insertion region for six retroCNVs, Supplementary Table S5. Using the same 974 individuals (1000 Genomes Project Consortium *et al.*, 2015), we identified 83.3% (*5*/*6*) of their retroCNVs. Next, similarly to these authors (Abyzov *et al.* 2013), we grouped individuals into their 14 populations and performed the retroCNV genotyping. Three retroCNVs (from CBX3, SKA3 and TDG genes, Supplementary Table S6) were confirmed in individuals from the 14 populations. Two retroCNVs (from SARAF and LAPTM46 genes) were confirmed in individuals from five and six populations, respectively (Supplementary Table S6). Remarkably, we obtained 100% concordance with Abyzov *et al.* (Supplementary Table S6). In summary, sideRETRO obtained an overall good performance (e.g. F1-score = 91.44, true positive rate or recall = 87.51% and true negative = 3.75%) in identifying retroCNVs from simulated and real WGS data (Fig. 1D).

4 Discussion

sideRETRO is a method dedicated to identifying retroCNVs in WGS or WXS data, which provides several characteristics of these gene duplicates. By using sideRETRO to identify retroCNVs in individuals from 1000 Genomes Project and from simulated data, we confirm that our algorithm has a high sensibility to detect and genotype bona fide retroCNVs. Furthermore, sideRETRO is user friendly, easy to install and publicly available.

We (Schrider *et al.*, 2013) and others (Abyzov *et al.*, 2013; Ewing *et al.*, 2013; Zhang *et al.*, 2017) have already analyzed retroCNVs in the human population. However, these studies were not devoted to describing methodologies, and their pipelines were therefore limited not only in terms of their documentation, installation and use but also in terms of providing additional characteristics of retroCNV events. To overcome these limitations, we created sideRETRO, an algorithm that identifies and provides key information about retroCNV events, such as their parental genes, genomic insertion sites, zygosity and genomic context (within or near a gene), and further information to classify each event as somatic or polymorphic. sideRETRO has a high sensibility to identify retroCNVs because it uses multiple sources of evidence (e.g. SRs, discordant read pairs and a grouping of reads presenting a concord mapping patter) to infer the occurrence of a retroCNVs event. sideRETRO search for retroCNV not only in WGS but also in WXS data.

In summary, sideRETRO is the first algorithm available to identify somatic and polymorphic insertion of processed pseudogenes or retrocopies. We expect that sideRETRO will shed light on and drive further studies of retroCNVs in health and disease genomes.

Acknowledgements

We thank you G. Guardia for comments on the manuscript.

Funding

This work was partially supported by Fundação de Amparo à Pesquisa do Estado de São Paulo [FAPESP; 2012/24731-1 and 2018/15579-8] and Instituto Serrapilheira. FORR [2015/25020-0], T.L.A.M. and J.L.L.B. are supported by fellowships from FAPESP, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and Conselho Nacional de Desenvolvimento Cientifico e Tecnológico, respectively.

Conflict of Interest: none declared.

References

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

Abyzov,A. *et al.*; The 1000 Genomes Project Consortium. (2013) Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res.*, **23**, 2042–2052.

Burns,K.H. (2017) Transposable elements in cancer. *Nat. Rev. Cancer*, **17**, 415–424.

Ewing,A.D. *et al.*; Broad Institute Genome Sequencing and Analysis Program and Platform. (2013) Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.*, **14**, R22.

Hancks,D.C. and Kazazian,H.H. Jr. (2016) Roles for retrotransposon insertions in human disease. *Mob. DNA*, **7**, 9.

Kabza,M. *et al.* (2014) RetrogeneDB—a database of animal retrogenes. *Mol. Biol. Evol.*, **31**, 1646–1648.

Kazazian,H.H. Jr. (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.

Navarro,F.C.P. and Galante,P.A.F. (2015) A genome-wide landscape of retrocopies in primate genomes. *Genome Biol. Evol.*, **7**, 2265–2275.

Navarro,F.C.P. and Galante,P.A.F. (2013) RCPedia: a database of retrocopied genes. *Bioinformatics*, **29**, 1235–1237.

Ram,A. *et al.* (2010) A density based algorithm for discovering density varied clusters in large spatial databases. *Int. J. Comput. Appl.*, **3**, 1–4.

Schrider,D.R. *et al.* (2013) Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet.*, **9**, e1003242.

Zhang,Y. *et al.* (2017) Landscape and variation of novel retroduplications in 26 human populations. *PLoS Comput. Biol.*, **13**, e1005567.

Zhang,Z. *et al.* (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.*, **20**, 62–67.

APÊNDICE B - *Script make_rtc.pl*

```
#!/usr/bin/env perl

use strict;
use warnings;
use autodie;
use Data::Dumper;
use Getopt::Long;
use List::Util 'shuffle';

use constant {
      SEED     => 666,
      RTC_NUM  => 100,
      LENGTH   => 1000
};

if (@ARGV < 1) {
      usage();
      exit 0;
}

my $seed = SEED;
my $num = RTC_NUM;
my $length = LENGTH;

GetOptions(
      "seed=i"   => \$seed,
      "rtc-num=i" => \$num,
      "length=i" => \$length
) or die "Error in command line arguments\n";

my $fasta_file = shift;
unless ($fasta_file) {
      usage();
      exit 1;
}

srand $seed;

my $idx_h = index_fasta($fasta_file);
my %idx_large;
```

```
for my $gene (keys %$idx_h) {
	my $seq = (
		sort { length($b) <=> length($a) }
		@{ $idx_h->{$gene} }
	)[0];

	if (length($seq) >= $length) {
		$idx_large{$gene} = $seq;
	}
}

my @genes = shuffle sort keys %idx_large;

for my $i (1..$num) {
	my $seq = substr $idx_large{$genes[$i]}, - $length;
	print "$genes[$i]\t$seq\n";
}

sub usage {
	print "Usage: $0 [--seed=INT] [--rtc-num=INT] <FASTA>\n";
	print "FASTA: transcripts of protein-coding genes\n";
}

sub index_fasta {
	my $genome = shift;
	open my $fh, "-|" => "zcat $genome";

	my %idx;
	my ($id, $i);

	while (<$fh>) {
		chomp;
		next if /^;/;

		if (/^>/) {
			$i = 0;
			my @f = split /\|/;

			$id = $f[5];
			$id =~ s/^\s+|\s+$//g;

			if (exists $idx{$id}) {
				$i = scalar @{ $idx{$id} };
```

```
                }
            } else {
                die "Error reading fasta file '$genome': Not defined id"
                    unless defined $id;
                $idx{$id}[$i] .= $_;
            }
        }

        close $fh;
        return \%idx;
}
```

APÊNDICE C - *Script make_cohort.pl*

```
#!/usr/bin/env perl

use strict;
use warnings;
use autodie;
use Getopt::Long;
use List::Util qw/shuffle/;

use constant {
      SEED             => 666,
      COHORT           => 100,
      WINDOW           => 1000,
      MAX_NNN_PERC     => 0.25,
      FIXED_PERC       => 0.25,
      POLYMORPHIC_PERC => 0.50,
      SOMATIC_PERC     => 0.25,
      OUTPUT_DIR       => 'out'
};

if (@ARGV < 1) {
      usage();
      exit 0;
}

my $seed = SEED;
my $cohort = COHORT;
my $dir = OUTPUT_DIR;

GetOptions(
      "seed=i"       => \$seed,
      "cohort=i"     => \$cohort,
      "output-dir=s" => \$dir
) or die "Error in command line arguments\n";

my ($genome, $rtc_file) = @ARGV;

unless ($genome && $rtc_file) {
      usage();
      exit 1;
}
```

```
srand $seed;

mkdir $dir unless -d $dir;

my $chr_h = index_fasta($genome);
my $rtc_h = index_rtc($rtc_file);
my $pos_h = build_pos($rtc_h, $chr_h);
my $cohort_a = build_cohort($rtc_h, $cohort);

dump_rtc_pos($pos_h, "$dir/rtc_pos.tsv");
dump_rtc_ind($rtc_h, $pos_h, $cohort_a, $dir);

sub usage {
      print "Usage: $0 [--cohort=INT] [--seed=INT] [--output-dir=DIR]",
            "<GENOME> <RTC_FILE>\n";
}

sub dump_rtc_pos {
      my ($pos_h, $file) = @_;
      open my $fh, ">", $file;
      print $fh "#chromosome\tposition\tplus_strand\tparental\thomozygous\
n";
      for my $id (sort keys %$pos_h) {
            my $d_h = $pos_h->{$id};
            print $fh "$d_h->{chr}\t$d_h->{pos}\t$d_h->{plus_strand}\t",
                  "id\t$d_h->{homozygous}\n";
      }
      close $fh;
}

sub dump_rtc_ind {
      my ($rtc_h, $pos_h, $cohort_a, $dir) = @_;

      for my $i (0..$#$cohort_a) {
            my $ind = $cohort_a->[$i];

            open my $fh, ">", "$dir/ind${i}.sandy";
            print $fh "#seqid\tposition\tid\treference\t",
                  "alteration\tgenotype\n";

            for my $parental (@$ind) {
                  my $d_h = $pos_h->{$parental};
```

```
                my $seq = $d_h->{plus_strand}
                      ? $rtc_h->{$parental}
                      : _reverse_complement($rtc_h->{$parental});

                my $zygosity = $d_h->{homozygous} ? "HO" : "HE";

                my $id = $parental;
                $id .= $d_h->{plus_strand} ? "_p" : "_m";
                $id .= "_${zygosity}";

                print $fh "$d_h->{chr}\t$d_h->{pos}\t$id\t-\t",
                      "$seq\t$zygosity\n";
          }

          close $fh;
    }
}

sub _reverse_complement {
    my $seq = shift;
    my $rev_seq = reverse $seq;
    $rev_seq =~ tr/ATCGatcg/TAGCtagc/;
    return $rev_seq;
}

sub index_fasta {
    my $genome = shift;
    open my $fh, "-|" => "zcat $genome";
    my %chr;
    my $id;
    while (<$fh>) {
          chomp;
          next if /^;/;
          if (/^>/) {
                my @f = split /\|/;
                $id = $f[0];
                $id =~ s/^>//;
                $id =~ s/^\s+|\s+$//g;
          } else {
                die "Error reading fasta file '$genome': Not defined id"
                      unless defined $id;
                $chr{$id} .= $_;
          }
    }
```

```
	close $fh;
	return \%chr;
}

sub index_rtc {
	my $rtc_file = shift;
	open my $fh, "<" => $rtc_file;
	my %rtc;
	while (<$fh>) {
		chomp;
		my @f = split;
		$rtc{$f[0]} = $f[-1];
	}
	close $fh;
	return \%rtc;
}

sub build_pos {
	my ($rtc_h, $chr_h) = @_;

	my @chrs = sort keys %$chr_h;
	my @chrs_len = map {length $chr_h->{$_}} @chrs;
	my $weights_a = _build_weights(\@chrs_len);

	my %pos;
	for my $id (sort keys %$rtc_h) {
		my $chr_i = _bsearch(
			int(rand($weights_a->[-1]{up} + 1)),
			$weights_a,
			0,
			$#$weights_a
		);
		die "No index for chr" unless defined $chr_i;
		my $chr = $chrs[$chr_i];

		my $seq_s = \$chr_h->{$chr};
		my $pos;
		do {{
			$pos = int(rand(length $$seq_s));
		}} while(_is_pos_inside_NNN($seq_s, $pos));

		my $plus_strand = int(rand(2));
		my $homozygous = int(rand(4)) == 3 ? 1 : 0;
```

```
            $pos{$id} = {
                  chr         => $chr,
                  pos         => $pos,
                  plus_strand => $plus_strand,
                  homozygous  => $homozygous
            };
      }

      return \%pos;
}

sub _build_weights {
      my $w_a = shift;
      my @offset;
      my $left = 0;
      for (my $i = 0; $i < @$w_a; $i++) {
            my %w = (
                  down => $left,
                  up   => $left + $w_a->[$i] - 1
            );
            $left += $w_a->[$i];
            push @offset => \%w;
      }
      return \@offset;
}

sub _bsearch {
      my ($key1, $base, $start, $end) = @_;
      if ($start > $end) {
            # Not found!
            return;
      }
      my $index = int(($start + $end) / 2);
      my $key2 = $base->[$index];
      # $key1 <=> $key2
      my $rc = _cmp($key1, $key2);
      if ($rc > 0) {
            return _bsearch($key1, $base, $index + 1, $end);
      } elsif ($rc < 0) {
            return _bsearch($key1, $base, $start, $index - 1);
      } else {
            return $index;
      }
}
```

```
sub _cmp {
      # State the function to compare at _bsearch
      my ($r, $w) = @_;
      if ($r  >= $w->{down} && $r  <= $w->{up}) {
            return 0;
      }
      elsif ($r > $w->{down}) {
            return 1;
      } else {
            return -1;
      }
}

sub _is_pos_inside_NNN {
      my ($seq_s, $pos) = @_;
      my $seq_len = length $$seq_s;
      my $start_pos = $pos - int(WINDOW / 2);
      if ($start_pos < 0) {
            $start_pos = 0;
      } elsif (($start_pos + WINDOW) > $seq_len) {
            $start_pos = $seq_len - WINDOW;
      }
      my $win = substr $$seq_s, $start_pos, WINDOW;
      my $NNN_acm = $win =~ tr/Nn/Nn/;
      return $NNN_acm > int(WINDOW * MAX_NNN_PERC)
            ? 1
            : 0;
}

sub build_cohort {
      my ($rtc_h, $cohort) = @_;

      my @parental = shuffle sort keys %$rtc_h;
      my $parental_size = scalar @parental;

      my @fixed = splice @parental, 0,
            int(FIXED_PERC * $parental_size);
      my @polymorphic = splice @parental, 0,
            int(POLYMORPHIC_PERC * $parental_size);
      my @somatic = splice @parental, 0,
            int(SOMATIC_PERC * $parental_size);

      # FIXED
```

```
	my @cohort;
	push @cohort => [@fixed] for 1..$cohort;

	# POLYMORPHIC
	for my $ind (@cohort) {
		for my $gene (@polymorphic) {
			if (int(rand(4)) == 3) {
				push @$ind => $gene;
			}
		}
	}

	# SOMATIC
	my @putative_somatic = shuffle 0..$cohort - 1;
	my @somatic_inds = splice @putative_somatic, 0, scalar(@somatic);
	for my $i (0..$#somatic_inds) {
		my $ind = $cohort[$somatic_inds[$i]];
		push @$ind => $somatic[$i];
	}

	return \@cohort;
}
```

APÊNDICE D - *Script compare_sim.pl*

```
#!/usr/bin/env perl

use strict;
use warnings;
use autodie;
use File::Basename;
use Data::Dumper;

use constant DIR => 'analysis';

if (@ARGV < 2) {
      print "Usage: $0 <VCF> <DIR> [<BLACKLIST>]\n";
      print "DIR: directory with sandy files\n";
      print "BLACKLIST: file with retrocopies\n";
      exit 0;
}

my ($vcf_file, $sandy_dir, $black_file) = @ARGV;

my ($rtcs_a, $genotype_h) = index_vcf($vcf_file);
my @black_list;
my %b;

if (defined $black_file) {
      open my $fh, "<", $black_file;
      %b = map { chomp; $_ => 1 } <$fh>;
      for my $rtc (@$rtcs_a) {
            push @black_list, 0;
            for (keys %b) {
                  if (/$rtc->{pg}/) {
                        $black_list[-1] = 1;
                        last;
                  }
            }
      }
      close $fh;
}

my @sandy_files = glob "$sandy_dir/*.sandy";
```

```
mkdir DIR unless -d DIR;

for my $sandy_file (@sandy_files) {
      my $sandy = basename $sandy_file, '.sandy';
      my $genotype = $genotype_h->{$sandy};
      if (defined $genotype) {
            compare($rtcs_a, $genotype, $sandy_file, $sandy, \
@black_list);
      } else {
            warn "$sandy not found at \$genotype_h hash";
      }
}

sub compare {
      my ($rtc_a, $genotype_a, $sandy_file, $sandy, $black_list_a) = @_;

      my $sandy_rtc_h = index_sandy($sandy_file);
      open my $fh, ">" => DIR . "/$sandy.tsv";

      for my $i (0..$#$rtc_a) {
            next if $genotype_a->[$i] eq 'HOR';
            next if $black_list_a->[$i];

            my $rtc_h = $rtc_a->[$i];
            my $srtc_h;

            my $pg = '';
            for (split /\//, $rtc_h->{pg}) {
                  if ($sandy_rtc_h->{$_}) {
                        $srtc_h = delete $sandy_rtc_h->{$_};
                        $pg = $_;
                        last;
                  }
            }

            next if $b{$pg};

            if (defined $srtc_h) {
                  print $fh "$srtc_h->{chr}\t$srtc_h->{pos}\t",
                        "$pg\t$srtc_h->{strand}\t$srtc_h->{genotype}";
            } else {
                  print $fh "-\t-\t-\t-";
            }
```

```
            print $fh "\t$rtc_h->{chr}\t$rtc_h->{pos}\t$rtc_h->{pg}\t",
                  "$rtc_h->{strand}\t$genotype_a->[$i]\n";
      }

      if (%$sandy_rtc_h) {
            for my $pg (keys %$sandy_rtc_h) {
                  next if $b{$pg};
                  my $srtc_h = $sandy_rtc_h->{$pg};
                  print $fh "$srtc_h->{chr}\t$srtc_h->{pos}\t",
                        "$pg\t$srtc_h->{strand}\t$srtc_h->{genotype}";
                  print $fh "\t-\t-\t-\t-\t-\n";
            }
      }

      close $fh;
}

sub index_sandy {
      my $sandy_file = shift;
      open my $fh, "<" => $sandy_file;

      my %rtc;

      while (<$fh>) {
            chomp;
            next if /^#/;

            my @f = split /\t/;
            my ($pg, $strand, $genotype) = split /_/, $f[2];

            $rtc{$pg} = {
                  chr       => $f[0],
                  pos       => $f[1],
                  strand    => $strand,
                  genotype  => $genotype
            };
      }

      close $fh;
      return \%rtc;
}

sub index_vcf {
      my $file = shift;
```

```
open my $fh, "<" => $file;

my (@rtcs, %genotype, @genotype_idx);

while (<$fh>) {
	chomp;
	next if /^##/;

	my @f = split /\t/;

	my @genotypes = splice @f, 9;

	if (/^#/) {
		@genotype_idx = @genotypes;
		next;
	}

	for my $i (0..$#genotypes) {
		my $genotype = $genotypes[$i];
		my $haplo = 'HOR';
		if ($genotype =~ /(0\/1|1\/0)/) {
			$haplo = 'HET';
		} elsif ($genotype =~ /1\/1/) {
			$haplo = 'HOA';
		}
		push @{ $genotype{$genotype_idx[$i]} } => $haplo;
	}

	my %rtc = (
		chr       => $f[0],
		pos       => $f[1],
		imprecise => 0,
		strand    => '.'
	);

	my @infos = split /;/, $f[7];

	for my $info (@infos) {
		my ($key, $value) = split /=/, $info;
		if ($key eq 'PG') {
			$rtc{pg} = $value;
		} elsif ($key eq 'POLARITY') {
			$rtc{strand} = $value eq '+' ? 'p' : 'm';
		} elsif ($key eq 'IMPRECISE') {
```

```
                    $rtc{imprecise} = 1;
                }
            }

            push @rtcs => \%rtc;
        }

        close $fh;
        return (\@rtcs, \%genotype);
}
```

APÊNDICE E - *Script confusion_analysis.pl*

```
#!/usr/bin/env perl

use strict;
use warnings;
use autodie;
use File::Basename;

my $dir = shift;

unless ($dir) {
	print "Usage: $0 <DIR>\n";
	print "DIR: directory for analysis\n";
	exit 0;
}

my @analysis_files = glob "$dir/*.tsv";

my %res_all = (
	TP => 0,
	FP => 0,
	FN => 0
);

print "IND\tTP\tFP\tFN\tPPV/Precision\t",
	"TPR/Recall\tF1-score\n";

for my $analysis_file (sort comp @analysis_files) {
	my %res = (
		TP => 0,
		FP => 0,
		FN => 0
	);

	open my $fh, "<", $analysis_file;

	while (<$fh>) {
		chomp;
		my @f = split;
		if ($f[0] eq $f[5]) {
			$res{TP} ++;
```

```
                if ($f[4] =~ /HO/ && $f[9] =~ /HO/) {
                        $res{TP} ++;
                } elsif ($f[4] =~ /HO/ && $f[9] =~ /HE/) {
                        $res{FN} ++;
                } elsif ($f[4] =~ /HE/ && $f[9] =~ /H0/) {
                        $res{FP} ++;
                }
        } elsif ($f[0] eq '-') {
                $res{FP} ++;
                if ($f[9] =~ /HO/) {
                        $res{FP} ++;
                }
        } elsif ($f[5] eq '-') {
                $res{FN} ++;
                if ($f[4] =~ /HO/) {
                        $res{FN} ++;
                }
        }
    }

    for (keys %res) {
        $res_all{$_} += $res{$_};
    }

    my $precision = $res{TP} / ($res{TP} + $res{FP});
    my $recall = $res{TP} / ($res{TP} + $res{FN});
    my $f1 = 2 * ($precision * $recall) / ($precision + $recall);

    printf "$analysis_file\t$res{TP}\t$res{FP}\t",
        "$res{FN}\t%.6f\t%.6f\t%.6f\n"
        => $precision
        => $recall
        => $f1;

    close $fh;
}

my $precision = $res_all{TP} / ($res_all{TP} + $res_all{FP});
my $recall = $res_all{TP} / ($res_all{TP} + $res_all{FN});
my $f1 = 2 * ($precision * $recall) / ($precision + $recall);

printf "ALL\t$res_all{TP}\t$res_all{FP}\t",
    "$res_all{FN}\t%.6f\t%.6f\t%.6f\n"
    => $precision
```

```
      => $recall
      => $f1;

sub comp {
      my $x = basename $a, '.tsv';
      my $y = basename $b, '.tsv';
      $x =~ s/ind//;
      $y =~ s/ind//;
      return $x <=> $y;
}
```

APÊNDICE F - *Script make_simulation.sh*

```
#!/usr/bin/env bash

# GENCODE reference genome
REF_FASTA=/assets/hg38.fa

# GENCODE protein-coding transcripts
PC_FASTA=/assets/gencode.v32.pc_transcripts.fa

# GENCODE annotation v32
ANNOTATION=/assets/gencode.v32.annotation.gff3

# Options for simulation
COHORT=100
RTC_NUM=100
LEN=1000
DEPTH=20
SANDY_SEED=1
SEED=17

#
# SIMULATION
#

# Generate sequences
scripts/make_rtc.pl \
  --seed=$SEED \
  --rtc_num=$RTC_NUM \
  --length=$LEN \
  "$PC_FASTA" > rtc_100.tsv

# Build our cohort
scripts/make_cohort.pl \
  --cohort=$COHORT \
  --seed=$SEED \
  --output-dir=build \
  "$REF_FASTA" \
  rtc_100.tsv

# Retrocopies by individual
mapfile -t IND < <(ls build/*.sandy)
```

```
# Load build values to SANDY
for ind in "${IND[@]}"; do
  sandy variation add \
    --structural-variation="$(basename "$ind" '.sandy')" \
    "$ind"
done

mkdir -p sim

# Simulate all genomes
for ind in "${IND[@]}"; do
  sandy_index="$(basename "$ind" '.sandy')"
  sandy genome \
    --id='%i.%U_%c:%S-%E_%v' \
    --structural-variation="$sandy_index" \
    --output-dir="sim/$sandy_index" \
    --jobs=20 \
    --seed=$SANDY_SEED \
    --quality-profile='hiseq_101' \
    --coverage=$DEPTH \
    --verbose \
    $REF_FASTA
done

#
# sideRETRO PRE-PROCESSING - alignment
#

# Individual directories with the
# simulated data
mapfile -t IND_DIR < <(ls -d sim/*)

# Options for BWA
BWA_THREADS=10

# Index reference genome
bwa index $REF_FASTA

mkdir -p align

# Alignment
for ind in "${IND_DIR[@]}"; do
  id="$(basename "$ind")"
```

```
  bwa mem \
    -t $BWA_THREADS \
    $REF_FASTA \
    "$ind/out_R1_001.fastq.gz" \
    "$ind/out_R2_001.fastq.gz" > "align/$id.sam"
done

#
# RUN sideRETRO
#

# Our simulated SAM files list
mapfile -t LIST < <(ls align/*.sam)

# Options for sideRETRO
CACHE_SIZE=20000000
SIDER_THREADS=20
SIDER_ALIGN_FRAC=0.9
SIDER_QUAL=20
SIDER_EPSILON=500
SIDER_MIN_PTR=10

# Run process-sample step
sider process-sample \
  --prefix=sim \
  --cache-size=$CACHE_SIZE \
  --output-dir=sider \
  --threads=$SIDER_THREADS \
  --alignment-frac=$SIDER_ALIGN_FRAC \
  --phred-quality=$SIDER_QUAL \
  --sorted \
  --log-file=ps.log \
  --annotation-file=$ANNOTATION \
  "${LIST[@]}"

# Run merge-call step
sider merge-call \
  --cache-size=$CACHE_SIZE \
  --epsilon=$SIDER_EPSILON \
  --min-pts=$SIDER_MIN_PTR \
  --log-file=mc.log \
  --threads=$SIDER_THREADS \
  --phred-quality=$SIDER_QUAL \
  --in-place \
```

```
  sider/sim.db

# Finally run make-vcf
sider make-vcf \
  --log-file=vcf.log \
  --reference-file=$REF_FASTA \
  --prefix=sim \
  --output-dir=sider \
  sider/sim.db

#
# ANALYSIS OF PERFORMANCE
#

# Generate comparations for analysis
scripts/compare_sim.pl sider/sim.vcf build

# Confusion analysis
scripts/confusion_analysis.pl analysis > confusion.tsv
```

Anexo

ANEXO A – Súmula Curricular

DADOS PESSOAIS

Nome: Thiago Luiz Araujo Miller

E-mail: thiago_leisrael@hotmail.com

Local e data de nascimento: São Paulo, SP, 24 de dezembro de 1984

EDUCAÇÃO

Ano	Graduação	Instituição
2016-2022	Doutorado direto em Bioquímica	Instituto de Química, Universidade de São Paulo, USP, São Paulo, SP
2011-2015	Bacharelado em Engenharia Biotecnológica	Universidade Estadual Paulista, Júlio de Mesquita Filho, UNESP, Assis, SP
2004-2005	Ensino Médio	Instituto Universal Brasileiro, São Paulo, SP

OCUPAÇÃO

Ano	Atividade	Instituição
2020-2022	Pesquisador	Centro de Oncologia Molecular. Hospital Sírio-Libanês, São Paulo, SP
2016-2020	Bolsista de doutorado CAPES (PROEX)	Instituto de Química, Universidade de São Paulo, USP, São Paulo, SP
2015-2016	Bolsista PIBIC	Centro de Oncologia Molecular. Hospital Sírio-Libanês, São Paulo, SP
2013-2014	Iniciação científica	Laboratório de Bioinformática. Universidade Estadual Paulista, Júlio de Mesquita Filho, UNESP, Assis, SP
2012-2013	Bolsista FAPESP (TT-I)	Laboratório de Biotecnologia Industrial. Universidade Estadual Paulista, Júlio de Mesquita Filho, UNESP, Assis, SP
2011-2012	Iniciação científica	Laboratório de Matemática Aplicada. Universidade Estadual Paulista, Júlio de Mesquita Filho, UNESP, Assis, SP

PUBLICAÇÃO

Ano	Artigo
2022	Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. Nat Commun 13, 1004 (2022). <https://doi.org/10.1038/s41467-022-28648-3>
2021	sideRETRO: a pipeline for identifying somatic and polymorphic insertions of processed pseudogenes or retrocopies. Bioinformatics 37, 3 (2021). <https://doi.org/10.1093/bioinformatics/btaa689>
2019	Sandy - A straightforward and complete next-generation sequencing read simulator. <http://doi.org/10.5281/zenodo.2589575>
2019	Transposon insertion profiling by sequencing (TIPseq) for mapping LINE-1 insertions in the human genome. Mobile DNA 10, 8 (2019). <https://doi.org/10.1186/s13100-019-0148-5>
2018	CGNT: A database of cancer genes explored in a nontumoral context. Clin Cancer Res 24, 1_Supplement (2018). <https://doi.org/10.1158/1557-3265.TCM17-B10>
2017	Measuring plasma levels of three microRNAs can improve the accuracy for identification of malignant breast lesions in women with BI-RADS 4 mammography. Oncotarget 11, 8 (2017). <https://doi.org/10.18632/oncotarget.20806>

APRESENTAÇÃO DE TRABALHO (PÔSTER)

Ano	Trabalho	Evento
2018	SANDY - A straightforward and complete next-generation sequencing read simulator	SBG, Genética
2018	SANDY - A straightforward and complete next-generation sequencing read simulator	X-Meeting
2017	Circulating miR-15a, miR-101, and miR-144 to distinguish between benig and malignant breast lesions in women with BI-RADS 4 mammography	AACR International Conference on Translational Cancer Medicine, held in cooperation with the Latin American Cooperative Oncology Group (LACOG)