



UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE BIOCÊNCIAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ISABELA PIMENTEL DE ALMEIDA

PIPELINE OLIGOY PARA DESENHO DE SONDAS ***OLIGOPAINT*** DO
CROMOSSOMO Y INCLUINDO SEQUÊNCIAS REPETITIVAS

Durante o desenvolvimento deste trabalho a autora recebeu auxílio financeiro
da FAPESP (Processo nº 2019/14878-4)

São Paulo/SP

2022

ISABELA PIMENTEL DE ALMEIDA

***PIPELINE* OLIGOY PARA DESENHO DE SONDAS *OLIGOPAINT* DO
CROMOSSOMO Y INCLUINDO SEQUÊNCIAS REPETITIVAS**

Versão Corrigida

Dissertação de Mestrado apresentada ao Programa
Interunidades de Pós-Graduação em Bioinformática da
Universidade de São Paulo para obtenção do título de
Mestre em Ciências.

Área de Concentração: Bioinformática

Orientadora: Profa. Dra. Maria Dulcetti Vibranovski

Co-Orientador: Prof. Dr. Antonio Bernardo de Carvalho

São Paulo/SP

2022

FICHA CATALOGRÁFICA

A447 Almeida, Isabela Pimentel de
Pipeline OligoY para desenho de sondas oligopaint do cromossomo Y incluindo sequências repetitivas / Isabela Pimentel de Almeida, orientadora Maria Dulcetti Vibranovski, co-orientador Antonio Bernardo de Carvalho -- São Paulo : 2022.
150 p.

Dissertação (Mestrado) -- Universidade de São Paulo
Orientadora: Profa. Dra. Maria Dulcetti Vibranovski
Co-orientador: Prof. Dr. Antonio Bernardo de Carvalho
Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática
Versão Corrigida

1. Bioinformática. 2. FISH. 3. Citogenética. 4. Heterocromatina. I. Vibranovski, Maria Dulcetti, orientadora. II. Carvalho, Antonio Bernardo de, co-orientador. III. Universidade de São Paulo. IV. Título.

CDD - 572.8



Universidade de São Paulo

ATA DE DEFESA

Aluno: 95131 - 11359241 - 1 / Página 1 de 1

Ata de defesa de Dissertação do(a) Senhor(a) Isabela Pimentel de Almeida no Programa: Bioinformática, do(a) Interunidades em Bioinformática da Universidade de São Paulo.

Aos 26 dias do mês de janeiro de 2022, no(a) realizou-se a Defesa da Dissertação do(a) Senhor(a) Isabela Pimentel de Almeida, apresentada para a obtenção do título de Mestra intitulada:

"Pipeline OligoY para desenho de sondas oligopaint do cromossomo Y incluindo sequências repetitivas"

Após declarada aberta a sessão, o(a) Sr(a) Presidente passa a palavra ao candidato para exposição e a seguir aos examinadores para as devidas arguições que se desenvolvem nos termos regimentais. Em seguida, a Comissão Julgadora proclama o resultado:

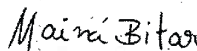
Nome dos Participantes da Banca	Função	Sigla da CPG	Resultado
Maria Dulcetti Vibranovski	Presidente	IB - USP	Não Votante
Ariane Machado Lima	Titular	EACH - USP	<u>APROVADA</u>
Mainá Bitar Lourenço	Titular	Externo	<u>APROVADA</u>
Leonardo Barbosa Koerich	Suplente	ECI-UFGM - Externo	<u>APROVADA</u>

Resultado Final: APROVADA

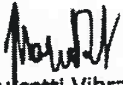
Parecer da Comissão Julgadora *

Eu, Maria Dulcetti Vibranovski, lavrei a presente ata, que assino juntamente com os(as) Senhores(as). São Paulo, aos 26 dias do mês de janeiro de 2022.


Ariane Machado Lima


Mainá Bitar Lourenço


Leonardo Barbosa Koerich


Maria Dulcetti Vibranovski
Presidente da Comissão Julgadora

* Obs: Se o candidato for reprovado por algum dos membros, o preenchimento do parecer é obrigatório.

A defesa foi homologada pela Comissão de Pós-Graduação em 16.02.2022, portanto, o(a) aluno(a) acima jus ao título de Mestra em Ciências obtido no Programa Bioinformática.


Presidente da Comissão de Pós-Graduação

Prof. Dr. André Fujita
Presidente
Comissão de Pós-Graduação
Interunidades em Bioinformática

*Às minúcias da vida
abrilhantadas por grandes mulheres pesquisadoras*

Agradecimentos

À Isabela Almeida, a mulher mais especial da minha vida, por toda a sua trajetória acadêmica, pelas longas horas dedicadas a esta pesquisa, por me olhar nos olhos através do espelho e me assegurar que, se não fosse por nossa própria capacidade, não estaríamos aqui, por acreditar em mim e por amar a vida mais que tudo, me concedendo tempos de descanso merecido.

À Bryan Khelven, pelas definições computacionais do glossário, por ter a paciência de me guiar pelo mundo das classes em Python, por me dizer para me basear na padronização dos códigos no livro de Algoritmos de Cormem para a escrita dessa dissertação, pela revisão editorial inicial e cada *no comma*, por me visitar em São Paulo me mostrando tudo de bom pela cidade, pelo colchão que me deu, por me ouvir ensaiar apresentações inúmeras vezes (mesmo antes do mestrado), pelas comidas maravilhosas, pela companhia constante durante a pandemia, por ser um ser humano consciente, por existir na minha vida e me deixar existir na sua.

À Maria Vibranovski, por responder meu primeiro e-mail em janeiro de 2019 e aceitar me entrevistar, por me dar a liberdade de tentar abordagens diferentes, por entender que eu não sou boa de bancada, pelas reuniões, por me proporcionar conhecer o Parque Nacional de Itatiaia, por ser uma pesquisadora e uma mulher feminista incrível, por todo o apoio, por ter tanta calma e sabedoria, por me dizer que iria dar tudo certo, por me oferecer tanto conforto e por acreditar em mim. Ah, por me orientar!

À José Joelson Pimentel de Almeida e Sirlene Aparecida Gomes Pimentel de Almeida, por terem me concedido a vida, me dado amor e carinho, pelas bananas que reservaram para as minhas vitaminas, por me proporcionarem conhecer o IME e a USP desde criança, pelos livros de matemática do papai e pelos livros de poesia da mamãe, pelo tanto de cor que sempre me ofereceram na vida, por estimularem que eu fosse em todas as excursões da escola, pelas viagens ao Paraná, por terem se mudado para a Paraíba e, assim, me dado de presente o Brasil inteiro, por terem me ajudado nas tarefas de casa e por terem paciência quando eu era a única criança na sala de aula que ainda não sabia ler, por me julgarem inteligente mesmo

sem entender o que é que eu faço.

À Eduardo Benedito e à Jóyce Kaynara, por ouvirem minhas lamentações, por clarearem a minha cabeça, por todas as conversas aleatórias sobre Harry Potter, a vida, o governo e a morte, pelas músicas enviadas, por perguntarem se eu estava bem, por estarem comigo desde o primeiro ano do IF e por todas as risadas loucas.

À Vitória Almeida, Eduardo Benedito, Alícia Melo e Rodrigo Ogava, por todas as discussões sobre a pós-graduação, por podermos compartilhar os dias de desânimo e as alegrias, por entenderem as dificuldades triplicadas da pós-graduação na pandemia. À Vitória por ser minha melhor lembrança da Biologia na UEPB, por todos os valhas, por amar junto comigo os gráficos e pelos congressos. À Alícia e ao Rodrigo por todos os almoços na Física (os da Química não), pelas manhãs e pelas tardes estudando na biblioteca do IME, pelos quadros completos desenhados sobre alinhamento de sequências, por ouvirem minha primeira apresentação de projeto, pelos trabalhos que fizemos juntos e por todo o apoio ao longo deste trabalho.

À Raul Gomes, pelos passos iniciais com Reconhecimento de Padrões, por estar sempre lá do outro lado da internet e por ter feito de mim a Lalá há muitos anos, quando eu ainda nem sonhava em ser Bioinformata. À Maria do Livramento, pela existência de Bryan, por ter me aceitado em sua vida de braços abertos, por me oferecer um segundo lar, verdadeiramente *home sweet home* e pelas conversas e risadas aleatórias.

Ao pessoal do Vibranovski Lab, por formarem um grupo tão diverso, bonito e inteligente. À Camila Avelino, por sentar do outro lado da minha mesa, pelas tantas conversas sobre gatos e por entender tão bem sobre citogenética e protocolos. À Carolina Mendonça, por me mostrar como repicar moscas e por todas as horas e horas de reunião sobre o OligoMiner, sondas, Beliveau, lógica de PCR, enzimas, *primers* e esquemas de amplificação de biblioteca. À Eduardo Dupim, por me ensinar como usar o BlobTools, por me explicar sobre picos hetero e homogaméticos, por me convidar para fazer análises e, quem sabe, melhorar o código do YGS2. À Henry Bonilla por toda a colaboração, por ter se dedicado com tanto afinho aos experimentos de bancada e por ter me feito perceber mil e um detalhes de biologia molecular. À Mara Pinheiro, por ser uma grande citogeneticista, pelas habilidades quase surreais de microscopia. À Amanda Luvisotto por me ensinar a dissecar o cérebro de larva de *Drosophila melanogaster*, pelas reuniões sobre amplificação de sondas e pelo empenho nos experimentos de bancada. À Gabriel Goldstein por prontamente me ajudar com o servidor e por acreditar que eu poderia de alguma forma contribuir com seu trabalho lindo de genes novos. À Ana Ferretti por sempre responder meus e-mails com tanta atenção e riqueza de detalhes. À Bernardo de Carvalho por todas as pesquisas anteriores e pela formação de pessoas essenciais para o desenvolvimento desta. À todas as moscas que derem involuntariamente suas vidas

para essa pesquisa. E à Maria Vibranosvki por dar vida a este laboratório.

À Banca examinadora deste trabalho, que em parte me acompanhou desde o Exame de Qualificação, pelas suas valiosas contribuições. Em especial à Mainá Bitar, por me abrir a porta para o mundo da Bioinformática, por me mostrar que posso ser voluntária em congressos e por ser minha eterna mentora. À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP – Processo nº 2019/14878-4), pelo apoio financeiro sem o qual esta pesquisa e sua pesquisadora não poderiam ter se eternizado nestas páginas. Aos professores da pós-graduação por todos os ensinamentos. À Universidade de São Paulo, por ter sido para mim um sonho por tantos anos e por finalmente ter me aceitado e me deixado carregar uma carteirinha com seu nome.

À Zag, Amorinha e Tutu (todos *in memoriam*), meus bebês amarelo, preto e tuti-fruti, que sempre entenderam mais ou menos que eu tinha algo muito importante para fazer lá fora, já que esse seria o único motivo para não ficar com eles, e por terem me dado o privilégio de conhecê-los em vida. Ao meu Zaguinho, nem se eu juntasse todas as palavras do mundo poderia expressar o quanto cada batida do seu coração permitiu que eu fosse hoje quem sou. Ao Eustáquio, por surgir em minha vida quando eu mais precisava lembrar que todos os gatinhos são absolutamente incríveis e por ser o bonitinho mais bonitinho da minha existência. À Mini (Balãozinho), Marcelino, Três e Eustáquio por fazerem de casa um parque de diversão e me ajudar a esquecer tudo de ruim que acontece do lado de fora durante a pandemia. À Marcelino por me amar disfarçando que me odeia. À Fofinha, Dois, Três e Quatro, por todos os *uémis* e fofurinhas, por me fazerem levantar da cadeira e ir tomar um sol. À Branquinho por ser um pai tão amoroso para Mini e Eustáquio, e à ele e Eustáquio pela companhia fiel na mesa "de escritório"/cozinha enquanto eu trabalhava. À Branquinho e ao Três, que mais me acompanharam nas práticas de Yoga – e à Adriene Mishler, que nem sabe que eu existo, e me proporcionou a prática diária de Yoga ao longo de toda a pandemia, pelas palavras serenas, por cada *lots of love in* e por cada *namastê*.

Ao Universo, por colapsar. Pelas formações rochosas que deram origem à São Paulo, Paraná, Pernambuco e Paraíba. Pela complexidade e diversidade da comunicação e das formas de se comunicar. Por no meio do seu equilíbrio desequilibrado ter permitido que todos os seres que já viveram tenham vivido e evoluído e, assim, que todos mencionados nessa seção tenham colidido suas existências. Por permitir que tenhamos nos conhecido e, aos que fazem ciência, fazer acontecer ciência e estudar a organização no meio de todo esse lindo e grandioso caos.

Com um sincero, profundo e grandioso sentimento,

Obrigada!

Isabela Almeida

*I went looking for my dreams outside of myself and discovered,
it's not what the world holds for you,
it's what you bring to it.*

Lucy Maud Montgomery (2014)

Resumo

Um dos principais obstáculos em estudos com o cromossomo Y está relacionado ao estado heterocromático e altamente repetitivo desta estrutura, levando a dificuldades de montagem de *scaffolds* e *contigs* e, assim, na falta de sequências finais montadas para o mesmo. O cromossomo Y de *Drosophila melanogaster*, organismo modelo utilizado nessa pesquisa, tem tamanho estimado em 41 Mb de sequências ricas em repetições, mas apenas 10% das mesmas estão montadas na versão mais recente do genoma. Em contrapartida, o protocolo para desenho de sondas utilizadas em experimentos de marcação fluorescente de cromossomo completo (FISH *Oligopaint*) não inclui sequências repetitivas para evitar hibridização fora da região alvo. Por esse motivo, existem atualmente menos de 1500 sondas *oligopaint* para este cromossomo Y modelo, quantidade esta ao menos dez vezes menor quando comparada com a de outros cromossomos da mesma espécie. Além disso, essa quantidade é insuficiente para realizar ensaios de FISH *Oligopaint* com eficiência. O objetivo principal desta pesquisa é desenvolver uma *pipeline* que permita desenhar sondas *oligopaint* para o cromossomo Y de qualquer espécie de interesse. A *pipeline* final inclui a utilização de ferramentas livres e existentes em Bioinformática, identificação de sequências exclusivas do cromossomo de interesse, garante ao usuário a autonomia para escolha de parâmetros e efetivamente usa sequências repetitivas exclusivas do cromossomo alvo para desenhar sondas, maximizando assim a eficiência geral dos experimentos de citogenética. Após extensos testes e validações *in silico* e *in situ*, foi constatado que a aplicação da *pipeline* desenvolvida, OligoY, permite marcar o cromossomo Y sem gerar sinal fora do alvo, apesar da utilização de sequências repetitivas para desenho das sondas *oligopaint*.

Palavras-chave: Bioinformática, FISH, citogenética, heterocromatina

Abstract

One of the main obstacles in studies with the Y chromosome is related to the heterochromatic and highly repetitive state of this structure, leading to difficulties in assembling scaffolds and contigs and, thus, in a lack of final assembled sequences for it. The Y chromosome of *Drosophila melanogaster*, the model organism used in this research, has an estimated size of 41 Mb of repeat-rich sequences, but only 10% of them are assembled in the most recent genome release. In contrast, the protocol for designing probes used in full chromosome fluorescent labeling experiments (FISH Oligopaint) does not include repetitive sequences to avoid off-target hybridization. For this reason, there are currently less than 1500 oligopaint probes for this Y chromosome model, which is a value at least ten times smaller when compared to the one observed for other chromosomes of the same species. Furthermore, this amount is insufficient to carry out FISH Oligopaint assays efficiently. The main objective of this research is to develop a pipeline that allows the design of oligopaint probes for the Y chromosome of any species of interest. The final pipeline includes the use of open-source and existing tools in Bioinformatics, identification of sequences unique to the chromosome of interest, guarantees the user the autonomy to choose parameters and effectively uses repetitive sequences unique to the target chromosome to design probes, thus maximizing overall efficiency of cytogenetic experiments. After extensive tests and validations *in silico* and *in situ*, it was verified that the application of the developed pipeline, OligoY, allows staining the Y chromosome without generating off-target signal, despite the use of repetitive sequences for oligopaint probe design.

Key-words: Bioinformatics, FISH, cytogenetics, heterochromatin

Lista de Figuras

1.1	Cromossomos mitóticos de <i>Drosophila melanogaster</i>	20
1.2	Experimentos de FISH <i>Oligopaint</i> em <i>Drosophila melanogaster</i>	22
1.3	Estrutura em <i>loop</i> do cromossomo Y de <i>Drosophila melanogaster</i>	24
2.1	Fluxograma generalizado para desenho de sondas <i>oligopaint</i> para o cromossomo Y	29
2.2	Fluxograma de execução do método YGS	33
2.3	Esquema de Histograma a partir de <i>k-mers</i> distintos do Jellyfish histo	36
2.4	Estrutura inicial e final do arquivo de saída do YGS.pl	39
2.5	Fluxograma de execução do BlobTools	40
2.6	Esquema simplificado da <i>pipeline</i> do OligoMiner	43
2.7	Estrutura de arquivo SAM produzido pelo bowtie2	46
2.8	Esquema do protocolo de amplificação com IVT e RT	50
2.9	Sexagem em <i>Drosophila melanogaster</i>	53
2.10	Amostras teciduais – testículos de adultos e cérebros de larvas de terceiro instar de <i>Drosophila melanogaster</i>	55
2.11	Esquema de experimentos de validação com FISH <i>Oligopaint</i> do cromossomo Y	56
3.1	Esquema simplificado do método YGS	59
3.2	Jellyfish histo dos <i>reads</i> de fêmea e macho utilizados	61
3.3	Sequências inferidas para o cromossomo Y e para os cromossomos autossômicos e X	63
3.4	Diferentes tamanhos de <i>k</i> aplicados e cromossomos inferidos	66
3.5	Divergências e coincidências entre inferências realizadas a partir do YGS e definições cromossomais de Chang e Larracunte (2019)	67
3.6	Comparação entre resultados do YGS e do YGS2	69
3.7	Análise do Blobtools realizada com o conjunto de sequências inferidas para o cromossomo Y a partir da montagem Illumina e <i>k</i> = 18	75

4.1	Fluxograma da abordagem clássica do OligoMiner a partir dos resultados de inferência para o cromossomo Y	78
4.2	Esquema de subgrupos de sequências inferidas para o cromossomo Y	80
4.3	Conjuntos de sondas criados através da abordagem clássica	83
4.4	Desenhando sondas candidatas com ou sem permissão da sobreposição entre as mesmas	87
4.5	Esquema da nova abordagem que permite selecionar sondas repetitivas exclusivas	92
4.6	Genoma de Referência sem as sequências inferidas para o Cromossomo Y	94
4.7	Alternativas exploradas para filtragem com <i>short-reads</i> de fêmea	100
4.8	Relatório do kmerFilter.py para filtragem com genoma ou <i>short reads</i> de fêmea	101
4.9	Resultados esperados e observados na comparação dos modos do OligoMiner	105
4.10	Estimativa da densidade de alvos/kb dos subgrupos comuns de sondas desenhadas	108
4.11	Comparação do número final de sondas <i>oligopaint</i> obtidas com diferentes abordagens	110
5.1	Importância das regiões de <i>priming</i>	112
5.2	Sondas compartilhadas entre diferentes montagens	115
5.3	Amplificação de grupos de sondas em biblioteca com até duas regiões de <i>priming forward</i> em um oligonucleotídeo	116
5.4	Reações de PCR em Tempo Real	123
5.5	Eletroforese horizontal do produto de PCR em Tempo Real	124
5.6	Eletroforese vertical do produto final de amplificação	125
5.7	Resultados de FISH <i>Oligopaint</i> na linhagem BM5	126
5.8	OligoY: Fluxograma da <i>Pipeline</i> final	130

Lista de Tabelas

2.1	Montagens genômicas utilizadas	27
2.2	Parâmetros utilizados na análise através do YGS	30
2.3	Conjuntos de parâmetros utilizados para desenho e seleção de sondas candidatas	31
3.1	Tamanhos das sequências inferidas para o cromossomo Y	64
3.2	Tempo de execução, uso de memória e processador pelo YGS	68
3.3	Análise de contaminantes a partir de diferentes conjuntos de sequências inferidas para o cromossomo Y	73
4.1	Subgrupos de sequências inferidas para o cromossomo Y	79
4.2	Quantidade de sondas candidatas não sobrepostas desenhadas através do blockParse.py	82
4.3	Quantidade de sondas sem sobreposição entre candidatas – abordagem clássica OligoMiner (UM)	84
4.4	Quantidade de sondas sem sobreposição entre candidatas – abordagem clássica OligoMiner (LDM)	85
4.5	Quantidade de sondas candidatas sobrepostas desenhadas através do blockParse.py	88
4.6	Quantidade de sondas com sobreposição entre candidatas – abordagem clássica OligoMiner (UM)	89
4.7	Quantidade de sondas com sobreposição entre candidatas – abordagem clássica OligoMiner (LDM)	90
4.8	Quantidade de sondas sem sobreposição entre candidatas – OligoMiner (ZM)	97
4.9	Quantidade de sondas com sobreposição entre candidatas – OligoMiner (ZM)	98
4.10	Quantidade de sondas filtradas contra <i>short reads</i> de fêmea – sobreposição não permitida	103
4.11	Quantidade de sondas filtradas contra <i>short reads</i> de fêmea – sobreposição permitida	104
4.12	Densidades de alvos/kb das sondas desenhadas com sobreposição e filtradas através do ZM	109
5.1	Regiões de <i>priming</i> utilizadas na Biblioteca final de sondas <i>oligopaint</i>	118
5.2	<i>Primers</i> utilizados no protocolo de amplificação e adição de fluoróforo	119

Lista de Abreviaturas

- ASCII** *American Standard Code for Information Interchange* – Código Padrão Americano para o Intercâmbio de Informação. *Glossário*: ASCII, 33, 34, 46
- CQ** *Chromosome Quotient method* – método de quociente cromossômico. 58
- FISH** *Fluorescence in situ hybridization* – Hibridização Fluorescente *in situ*. 16, 17, 20, 22, 23, 24, 26, 29, 48, 49, 51, 52, 55, 56, 57, 76, 83, 86, 87, 96, 106, 108, 111, 112, 113, 114, 126, 127, 128, 129, 131, 132, 133
- IVT** *In vitro Transcription* – Transcrição *in vitro*. 49, 113, 118, 122, 132
- LDM** *Linear Discriminant Analysis Mode* – Modo de Análise Linear Discriminante. 44, 45, 46, 77, 78, 81, 82, 83, 85, 88, 90, 102, 103, 104, 105, 106, 110, 114, 132
- NGS** *Next Generation Sequencing* – Sequenciamento de Nova Geração. 43, 95, 99, 100, 101, 102, 103, 104
- PBS** *Phosphate buffered saline* – Tampão salino de fosfato. 53, 54
- PVSCUK** *Percentage of valid single copy k-mer unmatched by the female reads* – Porcentagem de *k-mers* de cópia única validados e sem correspondência com os *reads* de fêmea. 30, 37, 38, 39, 62, 63, 66, 67
- RT** *Reverse Transcription* – Transcrição reversa. 49, 50, 51, 113, 117, 118, 119, 122, 123, 124, 125, 132
- SNP** *Single Nucleotide Polymorphisms* – Polimorfismo de nucleotídeo único. *Glossário*: SNP, 35, 60
- UM** *Unique Mode* – Modo Único. 44, 45, 77, 78, 81, 82, 83, 84, 88, 89, 93, 102, 103, 104, 105, 106, 110, 114, 132
- VSCK** *Valid single copy k-mer* – *k-mers* de cópia única validados. 30, 39, 62, 63, 66, 67
- YGS** *Y Genome Scan* – Varredura do genoma para o Y. 21, 28, 29, 30, 31, 32, 33, 35, 37, 38, 39, 58, 59, 60, 62, 63, 64, 65, 67, 68, 69, 70, 71, 73, 74, 76, 78, 79, 80, 81, 82, 83, 86, 88, 93, 94, 109, 114, 129, 131, 134
- ZM** *Zero Mode* – Modo Zero. 45, 92, 93, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 109, 110, 114, 129, 130, 132, 133

1	Introdução	16
1.1	DNA heterocromático e sequências repetitivas	18
1.2	Montagem do genoma de <i>Drosophila melanogaster</i>	19
1.3	Citogenética e sequências repetitivas	22
1.4	Objetivos	24
2	Materiais e Métodos	26
2.1	Organismo Modelo e Sequências utilizadas	26
2.2	Ferramentas em Bioinformática aplicadas	28
2.2.1	Como executar e como é executado o YGS	32
2.2.2	Análise de Contaminantes com BlobTools	40
2.2.3	Como executar e como é executado o OligoMiner	42
2.3	Protocolos de validação <i>in situ</i>	48
2.3.1	Amplificação da biblioteca sintetizada de sondas <i>oligopaint</i>	49
2.3.2	Sexagem, Dissecção e Fixação de Tecidos	52
2.3.3	Experimentos de FISH <i>Oligopaint</i>	56
3	Inferindo sequências para o cromossomo Y	58
3.1	Testando diferentes parâmetros	59
3.2	Análise de Contaminantes	71
4	Desenhando sondas <i>oligopaint</i> para o cromossomo Y	76
4.1	Explorando as abordagens clássicas do OligoMiner	77
4.2	Uma nova abordagem a partir do OligoMiner Zero Mode para selecionar sondas repetitivas exclusivas	91
4.2.1	Explorando alternativas de filtragem	99
4.3	Comparações entre as abordagens	102

5	Biblioteca final e Validação <i>in situ</i> da <i>pipeline</i>	111
5.1	Procedimentos pré-síntese da Biblioteca Final	111
5.1.1	Triagens da Biblioteca Final	113
5.1.2	Escolha e Adição de regiões de <i>priming</i> aos grupos de sondas	117
5.2	Amplificação dos Subgrupos de Oligonucleotídeos	122
5.3	Experimentos de FISH <i>Oligopaint</i> e Microscopia	125
6	Considerações finais	131
	Referências	135
	Glossário	145
	Apêndice A – Memorial descritivo da Autora	148

1

Introdução

Estudos em citogenética são relevantes para a compreensão não somente das estruturas cromossômicas, como também das interações proteicas ao longo do ciclo celular, podendo revelar aspectos importantes acerca de um organismo e inclusive seus mecanismos de evolução (FERGUSON-SMITH, 2015; VALENTE *et al.*, 2016). Técnicas de marcação cromossômica por fluorescência são essenciais nessa área, visto que permitem detectar e localizar a presença e, por correlação, até mesmo a ausência de sequências de DNA de interesse (FERGUSON-SMITH, 2015). As diferentes técnicas de marcação cromossômica por fluorescência se tratam de variações do protocolo de FISH (*Fluorescence in situ hybridization* – Hibridização Fluorescente *in situ*), desenvolvido a partir de técnicas mais simples de hibridização *in situ*, como inicialmente descrito nos trabalhos de Gall e Pardue (GALL; PARDUE, 1969; PARDUE; GALL, 1969).

É indispensável que, para a realização de qualquer um destes experimentos de FISH, previamente seja escolhida a sequência de DNA de interesse a ser marcada para que a sonda possa ser sintetizada. Por exemplo, através de segmentos purificados de DNA amplificado, seja em vetores bacterianos ou através de PCR comum, sondas podem ser geradas e utilizadas em experimentos de FISH (BELIVEAU *et al.*, 2015). Para o desenho de sondas de FISH *Oligopaint*, entretanto, são sempre utilizadas as sequências genômicas montadas (BELIVEAU *et al.*, 2015; BELIVEAU *et al.*, 2018). Isso porque esse experimento almeja marcar o cromossomo por completo e não somente uma pequena região do mesmo – ex. cromossomo X *vs* DNA centromérico. A fim de evitar hibridização fora da região alvo, sequências ricas em repetições não são utilizadas para desenhar sondas *oligopaint*, dado que provavelmente se repetem fora da sequência de interesse (BELIVEAU *et al.*, 2018).

Particularmente, cromossomos Y (ou W) costumam ter uma natureza altamente repetitiva que tanto se reflete em dificuldades relacionadas com a montagem de suas sequências como também na precária disponibilidade de sondas *oligopaint* para o mesmo (CHARLESWORTH;

LANGLEY; STEPHAN, 1986; ADAMS, 2000; BELIVEAU *et al.*, 2018). Como resultado, experimentos de FISH *Oligopaint* não são realizados para cromossomos Y, apesar de que essa técnica possa ajudar a esclarecer mais aspectos sobre sua estrutura e comportamento ao longo do ciclo celular.

Assim, essa pesquisa trata da aplicação de soluções existentes em Bioinformática para desenho de sondas *oligopaint* incluindo as sequências repetitivas do cromossomo Y. A compilação das ferramentas utilizadas permite que, ao aplicar a *pipeline* aqui desenvolvida, pesquisas em citogenética com o cromossomo Y sejam realizadas para qualquer espécie de interesse. Para tanto, usamos como modelo o cromossomo Y de *Drosophila melanogaster*, formado em suma por sequências heterocromáticas repetitivas (ADAMS, 2000). Além disso, a fácil manutenção de bibliotecas de cultura desse organismo permite que a metodologia aqui proposta tenha sua eficiência averiguada.

Ao longo deste primeiro capítulo serão discutidas algumas técnicas em Biologia Celular e conceitos biológicos essenciais para compreensão desta pesquisa. Na primeira sessão, são introduzidos quesitos importantes nas discussões acerca de DNA heterocromático e sequências repetitivas (Seção 1.1). A seguir, são elucidados aspectos quanto à montagem do genoma de *Drosophila melanogaster* e do próprio cromossomo Y (Seção 1.2). A Seção 1.3 introduz a citogenética, conflitando suas principais técnicas ao uso de sequências repetitivas. Por fim, são delineados os objetivos desta pesquisa (Seção 1.4).

No capítulo seguinte são apresentados brevemente os materiais e métodos utilizados para o desenvolvimento e validação da *pipeline* proposta (Capítulo 2). Estão contidas no terceiro capítulo todas as discussões e critérios envolvidos na execução da ferramenta escolhida para inferir sequências para o cromossomo Y, sendo essa uma etapa importante para enriquecer a disponibilidade de regiões a partir das quais podem ser desenhadas sondas de interesse (Capítulo 3). No Capítulo 4 são apresentadas as diferentes abordagens testadas para desenhar as sondas *oligopaint*, assim como todas as filtragens aplicadas a fim de evitar marcações com viés indesejado. Em seguida, é explicada em detalhes a etapa diretamente envolvida com a síntese da biblioteca das sondas finais selecionadas, incluindo a adição de regiões de *priming* correlacionadas ao protocolo de amplificação e os resultados de validação da *pipeline*, a fim de garantir sua eficiência (Capítulo 5). Por fim, são traçadas as considerações finais envolvidas em todo o processo do desenvolvimento da *pipeline* OligoY para desenho de sondas *oligopaint* para cromossomos Y (Capítulo 6). Vale ressaltar que, para demonstrar o toque de personalidade que existe por trás desta pesquisa, é apresentado de forma breve o Apêndice A – Memorial descritivo da Autora.

1.1 DNA heterocromático e sequências repetitivas

Modificações covalentes nas histonas determinam a estrutura da cromatina e, por conseguinte, dos nucleossomos, que correspondem respectivamente ao complexo formado por porções de DNA associado a proteínas e ao arranjo de várias dessas unidades (KURUMIZAKA, 2013; MURAKAMI, 2013). Sequências heterocromáticas são regiões de DNA organizadas em nucleossomos compactados, normalmente concentrados em regiões centroméricas, que permanecem em grande parte transcricionalmente silenciadas em todas as células, ou em regiões de heterocromatina facultativa, silenciadas apenas temporariamente (BRAHMACHARI; JAIN, 2013; MURAKAMI, 2013). A heterocromatina pode ser diferenciada da eucromatina tanto por composição de sequências de DNA, quanto por aspectos relacionados ao processo de replicação e ao nível de empacotamento ao longo do ciclo celular, dado que porções eucromáticas do DNA são menos condensadas, mais acessíveis e, em geral, mais facilmente transcritas (HUISINGA; BROWER-TOLAND; ELGIN, 2006; ELGIN; WORKMAN, 2002).

Além disso, regiões heterocromáticas também costumam ser compostas por muitos elementos repetitivos, assim como discutido por Charlesworth, Langley e Stephan (1986). Esses mesmos autores mostraram que regiões heterocromáticas não centroméricas que não participam do processo de recombinação, como os cromossomos sexuais Y e W, são propícias ao acúmulo de DNA repetitivo através, por exemplo, de elementos transponíveis.

Por muitos anos acreditou-se que essas regiões não contribuam com a expressão e a funcionalidade celular. Entretanto, atualmente existem diversas evidências mostrando que regiões heterocromáticas estão envolvidas justamente no controle da expressão gênica e estabilidade do genoma, contendo inclusive genes essenciais para o organismo (CHANG; LARRACUENTE, 2019; HOSKINS *et al.*, 2002; SULLIVAN; BLOWER; KARPEN, 2001). Por exemplo, sequências nucleotídicas que normalmente se condensam em heterocromatina estão envolvidas em processos biológicos relevantes como especiação (BAYES; MALIK, 2009; FERREE; BARBASH, 2009; CATTANI; PRESGRAVES, 2012), organização nuclear (CSINK; HENIKOFF, 1996), segregação e pareamento de cromossomos (DERNBURG; SEDAT; HAWLEY, 1996; MCKEE; HONG; DAS, 2000; ROŠIĆ; KÖHLER; ERHARDT, 2014).

Apesar da riqueza de informações que estudos das regiões heterocromáticas possam trazer, existe um impasse para a obtenção dessas sequências desde a etapa de sequenciamento, mesmo quando as abordagens utilizadas procuram contornar os efeitos de toxicidade de sequências repetitivas para *Escherichia coli* ou a instabilidade destas sequências na clonagem de vetores para construção das bibliotecas de sequenciamento (CARLSON; BRUTLAG, 1977; LOHE; BRUTLAG, 1987a; LOHE; BRUTLAG, 1987b; HOSKINS *et al.*, 2002; CELNIKER *et al.*, 2002; EICHLER; CLARK; SHE, 2004; HOSKINS *et al.*, 2007). Em paralelo, os algoritmos

utilizados para montar as sequências não resolvem de maneira perfeita os *reads* repetitivos advindos do sequenciamento: diferentes sequências acabam convergindo para uma mesma região – a repetição – e o algoritmo escolhe adicionar a região repetitiva em apenas um desses caminhos (MILLER; KOREN; SUTTON, 2010).

Levando isso em consideração, na próxima seção é apresentado o organismo modelo utilizado nessa pesquisa, *Drosophila melanogaster*, discussões interessantes sobre seu genoma montado e, mais particularmente, acerca das dificuldades de montagem do cromossomo Y dada sua natureza altamente heterocromática e repetitiva.

1.2 Montagem do genoma de *Drosophila melanogaster*

Alguns aspectos similares entre as regiões heterocromáticas de *Drosophila melanogaster* e outras espécies, como sua estrutura e suas modificações na cromatina, sugerem que esse seja um modelo excelente para estudar sequências de DNA ricas em repetições em outras espécies, inclusive em seres humanos (SMITH *et al.*, 2007; International Human Genome Sequencing Consortium *et al.*, 2001; VENTER *et al.*, 2001). Taxonomicamente, esse é um organismo Arthropoda da super família Ephydroidea, onde a família dos drosofilídeos está inserida, agrupando os drosofilíneos e, estes, englobando o gênero (e subgênero) das *Drosophila* Sophophoras e estas ao grupo e subgrupo das melanogaster (MARKOW; O'GRADY, 2006; O'GRADY; MARKOW, 2009). Existem ao menos outros 5 grupos e 12 subgrupos de *Drosophila* Sophophoras – e essa é uma mera demonstração da diversidade de organismos intimamente relacionados, esclarecendo que, embora *Drosophila melanogaster* seja um organismo modelo amplamente utilizado, os estudos sobre efídróides envolvem uma série muito maior de espécies diferentes. Para *Drosophila melanogaster*, seu cariótipo, nominalmente compreendido como sendo o arranjo dos cromossomos condensados, inclui os cromossomos autossômicos 2, 3 e 4, e os cromossomos sexuais X e Y (Figura 1.1).

Estima-se que pouco mais de 30% do genoma dessa espécie seja constituído por regiões heterocromáticas (HOSKINS *et al.*, 2007; SMITH *et al.*, 2007). Sua heterocromatina contém sequências simples repetidas, como as de DNAs satélites, elementos repetitivos médios, como elementos transponíveis e DNA ribossômico e algumas sequências de DNA de cópia única (ADAMS, 2000; HOSKINS *et al.*, 2002; LOHE; BRUTLAG, 1986). E, apesar de possuir um dos genomas mais bem montados, principalmente com relação à sua contiguidade (CHAKRABORTY *et al.*, 2018; CHAKRABORTY *et al.*, 2016), apenas 143 Mb (megabases) dos 180 Mb de genoma haploide estimado está presente na forma de *contigs* e *scaffolds* (HOSKINS *et al.*, 2015). Uma vez que *Drosophila melanogaster* se trata de um importante modelo de estudo animal, não somente existem as versões de sua montagem consideradas

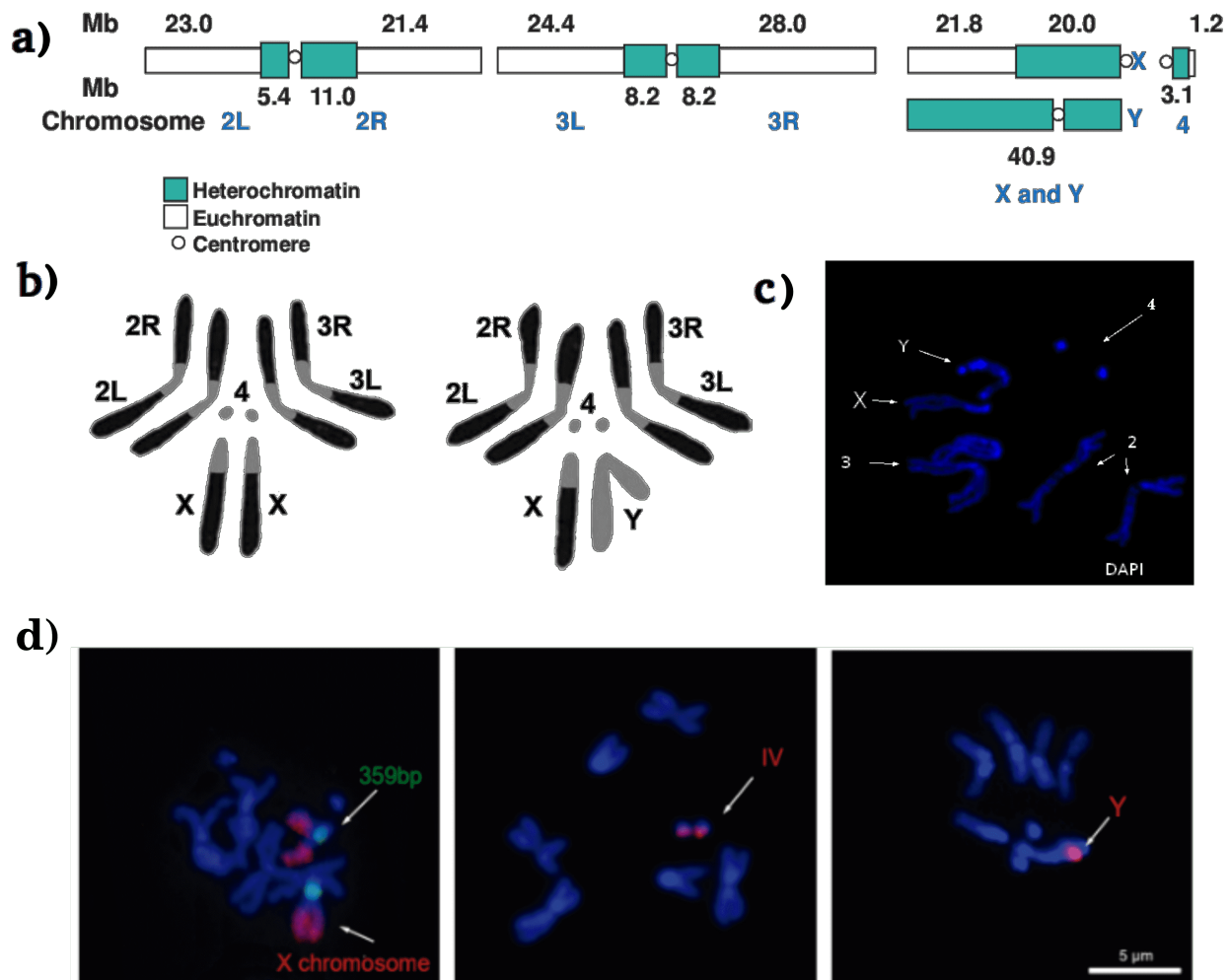


Figura 1.1: Cromossomos mitóticos de *Drosophila melanogaster*

Nota da figura: a) esquema mostrando regiões eucromáticas, heterocromáticas e centroméricas e seus tamanhos estimados em Mb (megabases; Extraído de Adams (2000)); b) cariótipo esquemático, onde as cores cinza e preta representam regiões heterocromáticas e eucromáticas, respectivamente (Extraído de Kaufman (2017)); c) cariótipo marcado por DAPI e visualizado em microscópio de fluorescência (Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa); d) ensaio de dupla hibridização FISH em célula metafásica, onde o cromossomo Y é marcado com sonda de DNA satélite Digoxigenina-(AATAC) em vermelho; cromossomo X em verde marcado com sonda para heterocromatina (359bp) com biotina e com sondas *oligopaint* desenhadas para eucromatina; todas as regiões de DNA estão marcadas com DAPI em azul (Produzido por Mara Pinheiro, Carolina Mendonça, Camila Avelino e Octavio Palacios – Laboratório da Professora Dra. Maria Vibranovski).

oficiais pelo consórcio do FlyBase (LARKIN *et al.*, 2021), indo desde a primeira versão nos anos 2000 (ADAMS, 2000) até a mais recente (HOSKINS *et al.*, 2015), como também existem diversas outras versões e abordagens exploradas na tentativa de resolver completamente sua montagem. Trabalhos como esses envolvem, por exemplo, curagens manuais buscando

aumentar a qualidade e podem usar *reads* providos de diferentes tecnologias de sequenciamento (CHANG; LARRACUENTE, 2019).

Com relação ao cromossomo Y, um dos principais obstáculos na identificação de genes e demais estudos com o mesmo está relacionado justamente ao estado heterocromático desta estrutura, formada por cerca de 40 Mb de sequências altamente repetitivas que o compõem quase que inteiramente, como esquematizado na Figura 1.1a (CELNIKER *et al.*, 2002; CARVALHO; LAZZARO; CLARK, 2000; ADAMS, 2000). Vale ressaltar que 80% dessas repetições são em *tandem*, como mostrado por Lohe em diferentes trabalhos (LOHE; HILLIKER; ROBERTS, 1993; BONACCORSI; LOHE, 1991). Como resultado, pelas dificuldades envolvidas com o sequenciamento e montagem de regiões repetitivas, ainda existe uma grande porção do cromossomo Y de *Drosophila melanogaster* que permanece desconhecida, com blocos altamente repetitivos e apenas 10% de seu tamanho estimado montado na versão mais recente do genoma (HOSKINS *et al.*, 2015; CHANG; LARRACUENTE, 2019). Mesmo a identificação dos genes desse cromossomo, viabilizada pelos estudos realizados por Bridges (1916), é dificultada, uma vez que existem regiões mínimas de eucromatina flanqueadas por regiões muito mais extensas de heterocromatina (CARVALHO *et al.*, 2001; HOSKINS *et al.*, 2015).

Em paralelo, existem abordagens que permitem recuperar mais sequências exclusivas do cromossomo Y, como o método YGS (*Y Genome Scan* – Varredura do genoma para o [cromossomo] Y) proposto inicialmente por Carvalho e Clark (2013). Através desse método é possível identificar *contigs*, por exemplo, que apesar de serem sequências montadas, não foram mapeados para um cromossomo específico. Nas montagens oficiais de *Drosophila melanogaster* esses *contigs* são identificados como *scaffolds* não mapeados (*unmapped scaffolds*), compondo as sequências do armU (CARVALHO *et al.*, 2001). A utilização dessa ferramenta resulta no aumento da riqueza de regiões conhecidas para o cromossomo Y, ainda que os *contigs* não estejam propriamente montados em uma sequência ordenada. Em *Drosophila melanogaster*, justamente a partir de *contigs* do armU alguns dos poucos genes do cromossomo Y foram identificados, e, assim, tais *contigs* se tornaram parte deste cromossomo (GEPNER; HAYS, 1993; CARVALHO; LAZZARO; CLARK, 2000; CARVALHO *et al.*, 2001; VIBRANOVSKI; KOERICH; CARVALHO, 2008; CARVALHO *et al.*, 2015).

A seguir são discutidas questões importantes para realização de experimentos em citogenética, contrastados com a não utilização de sequências repetitivas.

1.3 Citogenética e sequências repetitivas

Conforme já mencionado, os protocolos utilizados para desenhar sondas *oligopaint* não costumam incluir sequências repetitivas (BELIVEAU; APOSTOLOPOULOS; WU, 2014). Por esse motivo, é indicado que as repetições sejam mascaradas de maneira a evitar selecionar sondas que hibridizem fora da região alvo (BELIVEAU *et al.*, 2018), ou seja, que as bases repetitivas sejam “escondidas”, por exemplo, trocando-as pela quantidade respectivas de bases em múltiplos caracteres N. Por outro lado, a quantidade de sondas utilizadas e a densidade das mesmas por área cromossômica são bem maiores em experimentos de FISH *Oligopaint* do que nas variações mais comuns dessa técnica (BELIVEAU *et al.*, 2015). Isso assegura que regiões muito maiores, com dimensões cromossômicas, possam ser identificadas através de microscopia de fluorescência, assim como mostraram os trabalhos de Rosin, Nguyen e Joyce (2018) e de Beliveau *et al.* (2015) (Figura 1.2). Dessa forma, diferentes perguntas quanto às relações da região marcada com proteínas de interesse, como RNA polimerase, assim como o estudo de mecanismos celulares que a envolvam e investigações sobre momentos específicos do ciclo celular podem ser respondidas (BELIVEAU; APOSTOLOPOULOS; WU, 2014).

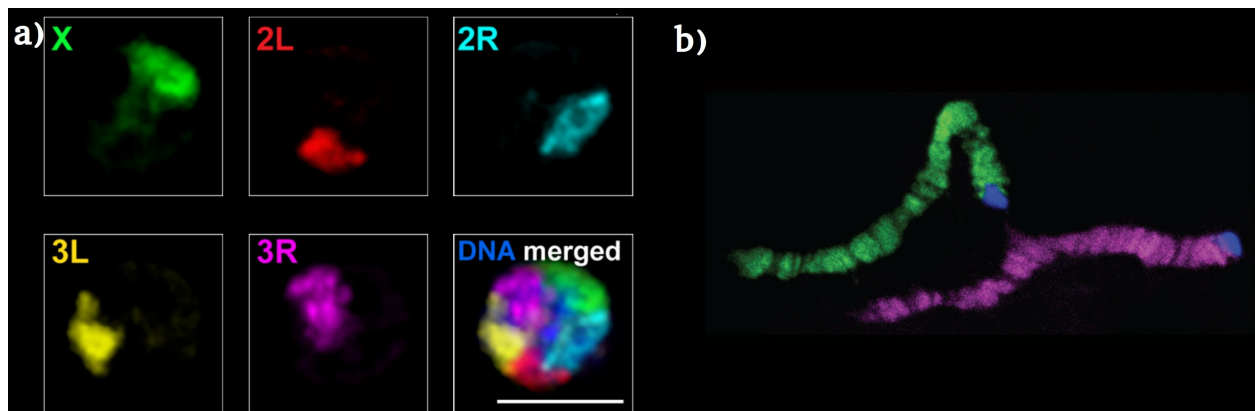


Figura 1.2: Experimentos de FISH *Oligopaint* em *Drosophila melanogaster*

Nota da figura: São apresentados dois experimentos com sondas de FISH *Oligopaint*, sendo: a) sinais de alta resolução obtidos a partir de sondas que incluíam apenas regiões não repetitivas dos cromossomos X, 2 e 3 (Extraído de Rosin, Nguyen e Joyce (2018)); b) sondas desenhadas a partir de regiões específicas para cada cromossomo politênico parental obtido do núcleo de glândulas salivares (Extraído de Beliveau *et al.* (2015)).

Dada a natureza altamente repetitiva de cromossomos Y, existe um impasse para realização de experimentos de FISH *Oligopaint* com os mesmos, já que a maior parte de suas sequências não pode ser utilizada no desenho das sondas. Como resultado, as sondas desenhadas apresentariam baixa densidade de marcações distribuídas ao longo do cromossomo,

provavelmente resultando em um experimento de FISH *Oligopaint* com pouco ou mesmo nenhum sinal observável pela microscopia de fluorescência. Atualmente, para o cromossomo Y de *Drosophila melanogaster*, são conhecidas em média menos de 1500 sondas, desenhadas a partir da versão oficial mais recente do genoma dessa espécie (BELIVEAU *et al.*, 2018). Em contraste, para todos seus outros cromossomos, a quantidade de sondas desenhadas é muito maior: para o cromossomo X, que tem tamanho muito similar ao Y, as sondas desenhadas por Beliveau *et al.* (2018) chegam a ultrapassar o valor de 300 mil, muito embora metade do cromossomo X também seja heterocromático, como pode ser observado na Figura 1.1 (ADAMS, 2000). Até mesmo para o menor cromossomo de *Drosophila melanogaster*, (cromossomo 4), para o qual é conhecida uma pequena região de eucromatina (1.2 Mb), o número de sondas desenhadas chega a ser em média dez vezes maior do que as encontradas para o cromossomo Y (BELIVEAU *et al.*, 2018; ADAMS, 2000). Essas comparações demonstram que a maior dificuldade enfrentada no desenho de sondas *oligopaint* para o cromossomo Y está realmente relacionada à natureza altamente repetitiva do mesmo.

Consequentemente, estudos de citogenética do cromossomo X bem como de cromossomos autossômicos são feitos comumente (Figura 1.2), entretanto, esses trabalhos não costumam incluir análises relacionadas ao cromossomo Y. Um dos interesses para realizar experimentos de FISH *Oligopaint* neste cromossomo parte da premissa de que assim seria possível observar em mais detalhes a estrutura em forma de *loop* que ele adquire durante a meiose, como mostrado por Fingerhut, Moran e Yamashita (2019) (Figura 1.3b). Essa estrutura é assim chamada por se parecer com um laço formado pelo cromossomo descondensado nesse momento da divisão celular (Figura 1.3a, em vermelho), se espalhando por toda a célula – diferente do observado para os demais cromossomos, que ocupam apenas uma região da mesma (MAHADEVARAJU *et al.*, 2021).

No trabalho mencionado, as autoras utilizaram marcações com sondas de DNA satélite (Figura 1.3b) que mostram diferentes regiões do *loop*. Essa estrutura pode estar relacionada com o controle da expressão gênica ao longo do genoma, justamente por se estender por todas as regiões da célula, potencialmente influenciando a expressão de outros cromossomos (CHANG; LARRACUENTE, 2019; FINGERHUT; MORAN; YAMASHITA, 2019). Assim, dispor de mais sondas *oligopaint* que melhor se distribuam ao longo do cromossomo Y, poderia ajudar a investigar o papel deste durante a meiose e revelar aspectos importantes sobre sua evolução e relações com outras partes do genoma.

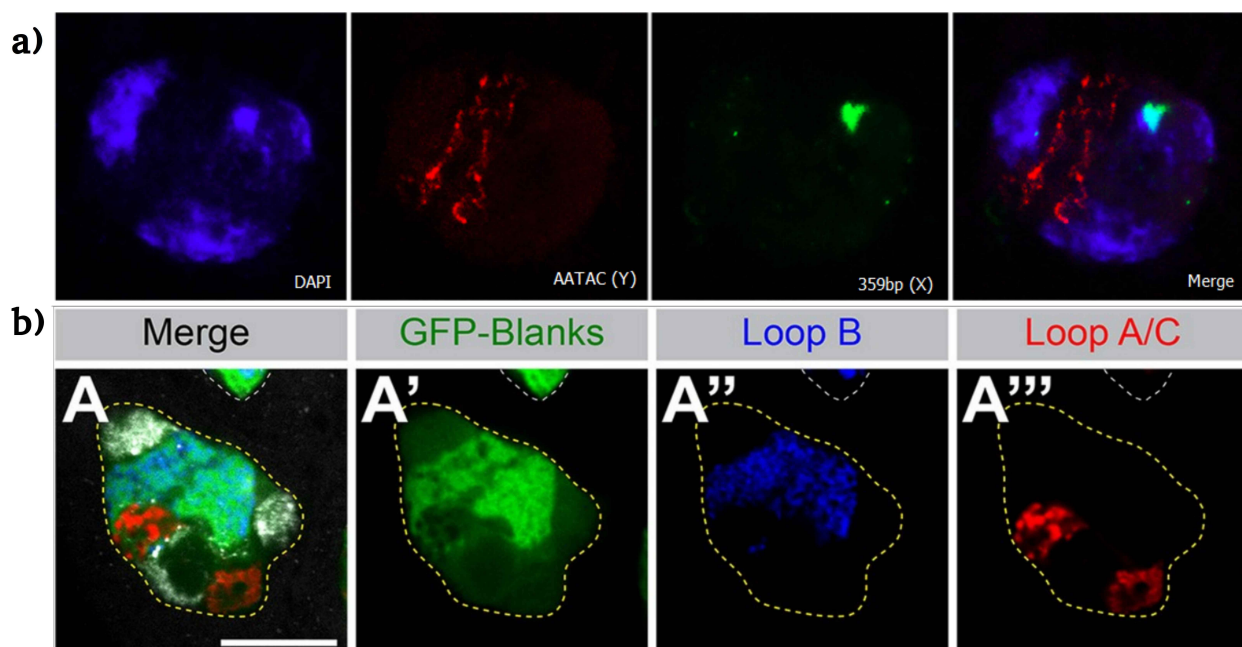


Figura 1.3: Estrutura em *loop* do cromossomo Y de *Drosophila melanogaster*

Nota da figura: A sobreposição das imagens permite observar que o cromossomo Y se espalha por toda a célula durante a meiose, sendo: a) experimento comum de FISH realizado a partir de espermatócito tardio marcando o cromossomo Y com sonda de DNA satélite Digoxigenina-(AATAC) em vermelho; cromossomo X em verde marcado com sonda para heterocromatina (359b) com biotina e todas as regiões de DNA marcadas com DAPI em azul (Produzido por Mara Pinheiro, Carolina Mendonça, Camila Avelino e Octavio Palacios – Laboratório da Professora Dra. Maria Vibranovski); b) FISH de RNA com expressão de GFP-Blanks para marcação de diferentes regiões conhecidas do *loop* com sondas de DNA satélite, onde os *loops* A e C são marcados com sonda Cy3-(AAGAC) em vermelho e o *loop* B é marcado com sonda Cy5-(AATAT) em azul (Extraído de Fingerhut, Moran e Yamashita (2019)).

1.4 Objetivos

Diante do exposto, esta pesquisa tem por objetivo desenvolver uma *pipeline* que permita desenhar sondas *oligopaint* para o cromossomo Y de qualquer espécie de interesse. Visando maximizar o potencial sinal obtido nos experimentos de fluorescência, são objetivos específicos:

- criar uma *pipeline* em que seja possível:
 - encontrar sequências exclusivas do cromossomo Y através de métodos existentes, aumentando assim a disponibilidade de regiões para o desenho de sondas;
 - permitir, através de métodos existentes, que as sondas sejam desenhadas conforme critérios estabelecidos pelo usuário da mesma;

- incluir as regiões repetitivas no desenho das sondas;
 - aplicar diferentes filtragens para evitar potenciais falhas no experimento, tanto de hibridização fora da região alvo quanto de problemas inerentes à estrutura da própria sonda;
 - garantir que a *pipeline* retorne a maior quantidade de sondas *oligopaint* possível.
- orientar sobre a transição *in silico-in situ*, com a adição das regiões de *priming*, importantes para os experimentos de fluorescência;
- realizar diferentes testes com a *pipeline* desenvolvida, incluindo:
 - submeter diferentes arquivos de entrada;
 - alterar os parâmetros dos métodos utilizados;
 - analisar os resultados advindos das combinações dos dois testes anteriores;
 - validar a *pipeline* a partir de experimentos com o cromossomo Y de *Drosophila melanogaster*.
- traçar possíveis aplicações da *pipeline*.

2

Materiais e Métodos

O desenho de sondas *oligopaint* para uma região altamente repetitiva apresenta uma série de desafios em Bioinformática. Isso porque as sondas desenhadas não devem marcar regiões fora do cromossomo alvo, o que diminuiria a eficiência dos ensaios realizados para a visualização *in situ* das marcações fluorescentes, ao passo que devem maximizar a distribuição dos sinais ao longo da região de interesse, principalmente em se tratando de experimentos com FISH *Oligopaint*. Nos próximos capítulos dessa dissertação de mestrado serão melhor esclarecidos os desafios e as soluções propostas para o desenvolvimento de uma *pipeline* que permita aos diferentes usuários desenhar sondas *oligopaint* para o cromossomo Y de qualquer espécie de interesse, independente da disponibilidade de um *scaffold* completamente montado para o mesmo e utilizando suas regiões repetitivas.

Neste capítulo são apresentados os materiais e métodos empregados para o desenvolvimento e validação da *pipeline* OligoY. Na primeira seção são apresentadas as sequências utilizadas para o cromossomo modelo (Seção 2.1). Em seguida, são descritas as ferramentas em Bioinformática aplicadas (Seção 2.2). Por fim, na Seção 2.3, estão postos os materiais e protocolos envolvidos nos experimentos de FISH *Oligopaint* a fim de validar a *pipeline* e traçar conclusões pertinentes sobre a mesma.

2.1 Organismo Modelo e Sequências utilizadas

O modelo utilizado nessa pesquisa é o cromossomo Y de *Drosophila melanogaster*. Como discutido anteriormente (Capítulo 1), pela alta repetitividade de suas sequências, apenas 10% de seu tamanho – estimado em 40.9 Mb – está montado na versão mais recente do genoma (ADAMS, 2000; HOSKINS *et al.*, 2015). Arelado aos *gaps* existentes e falta de sequências conhecidas para esse cromossomo, a escolha desse modelo para realização de testes, assegura que a *pipeline* proposta possa ser aplicada para outras espécies nas quais a montagem do

cromossomo Y seja tão precária quanto a do mesmo. Arelado a isto, a manutenção facilitada de bibliotecas de cultura deste organismo em laboratório permite a investigação da eficiência desta *pipeline* para validação da mesma.

Outra vantagem de utilizar este cromossomo modelo é a alta disponibilidade de sequências montadas a partir de diferentes abordagens e tecnologias de sequenciamento. Dessa forma, cinco genomas de *Drosophila melanogaster* são utilizados ao longo desta pesquisa (Tabela 2.1). Estas são montagens não finalizadas, incluindo a última versão oficial do genoma, *Release 6* (HOSKINS *et al.*, 2015), onde nem todos os cromossomos estão complementemente montados e existem muitos *contigs* não mapeados para um cromossomo específico, similarmente aos *contigs* do armU da *Release 6* (CARVALHO *et al.*, 2001; HOSKINS *et al.*, 2015). Em especial, a montagem mais recente é a de Chang e Larracunte (2019), onde os autores usaram uma abordagem de sequenciamento de molécula única e *long-reads* pela Pacific Biosciences, resultando em uma montagem genômica com menos *contigs*, mais *scaffolds* e menos *gaps*. Para o cromossomo Y, esses autores conseguiram elevar seu conteúdo nucleotídico em cerca de 10 Mb (14.5 Mb no total). Como eles procuravam gerar um genoma com mais contiguidade, diferentes abordagens de montagem foram aplicadas, incluindo o uso do montador Falcon para a montagem *de novo*, de maneira a aproveitar grande parte da informação sequenciada a partir das regiões heterocromáticas.

Assemblies	contigs	bP	Mb	Source
Chang & Larracunte	171	155613520	155.61	(CHANG; LARRACUENTE, 2019)
Illumina	144366	132243296	132.24	Carvalho with Gutzwiller <i>et al.</i> (2015) <i>reads</i>
Falcon	234	144185375	144.19	Carvalho with Kim <i>et al.</i> (2014) <i>reads</i>
Sanger	37814	163632981	163.63	(HOSKINS <i>et al.</i> , 2002)
Release 6	1870	143726386	143.73	(HOSKINS <i>et al.</i> , 2015)

Tabela 2.1: Montagens genômicas utilizadas

Nota da tabela: Para cada montagem (*Assembly*) são apresentadas as quantidade de *contigs* disponíveis na mesma, seu tamanho total em pares de bases (bP) e em megabases (Mb), assim como a fonte para obtenção dos mesmo (*Source*). Os nomes das montagens aparecem de acordo com a tecnologia de sequenciamento ou montador utilizado na mesma e são citados dessa forma ao longo de todo o trabalho. Genomas Falcon e Illumina foram montados pelo Professor Dr. Antonio Bernardo de Carvalho (UFRJ) usando *reads* dos trabalhos citados.

Para a última versão oficial do genoma, *Release 6* (HOSKINS *et al.*, 2015), existem algumas sequências montadas para o cromossomo Y: sendo 1 *scaffold* (3,67 Mb) e 199 *contigs* (0,86 Mb), totalizando 4.53 Mb de sequências conhecidas para este cromossomo. Nessa mesma versão, existem 978 *contigs* do armU, ou seja, pouco mais da metade dos *contigs* da *Release 6*

não foram mapeados em nenhum cromossomo. No entanto, as sequências do armU somam apenas 3.05 Mb, já que os demais cromossomos de *Drosophila melanogaster* estão muito bem montados em *scaffolds* (HOSKINS *et al.*, 2015).

Além de sequência genômica, a *pipeline* proposta também necessita de *short-reads* de macho e fêmea da espécie de interesse. Aqui, os arquivos utilizados foram ambos obtidos a partir do sequenciamento da linhagem ISO1 de *Drosophila melanogaster*: os *short-reads* de fêmea, advindos de sequenciamento Illumina, foram cedidos pelo Professor Dr. Antonio Bernardo de Carvalho (UFRJ), com curagem realizada anteriormente por seu grupo (GOLDSTEIN, 2016); já os *short-reads* de macho, advindos de sequenciamento com tecnologia da Pacific Biosciences, foram obtidos através do trabalho de Kim *et al.* (2014). Os *short-reads* de fêmea resultaram de um sequenciamento com aproximadamente 5 GB de sequências Illumina e 17X de cobertura (baseado em genoma de 142 Mb do Drosophila 12 Genomes Consortium, Project Leaders e Clark (2007)) (GOLDSTEIN, 2016). Kim *et al.* (2014) detalham uma série de avaliações de qualidade do material sequenciado utilizado, incluindo, por exemplo, o tamanho N50. Eles avaliaram a cobertura do sequenciamento em 95X, baseado na terceira versão do genoma de *Drosophila melanogaster* (160 Mb – Celniker *et al.* (2002)).

2.2 Ferramentas em Bioinformática aplicadas

Para desenvolvimento da *pipeline* proposta, foram utilizadas diferentes ferramentas, todas disponíveis gratuitamente. As principais delas são o método YGS (CARVALHO; CLARK, 2013) e a suíte de programas em Python, OligoMiner (BELIVEAU *et al.*, 2018). Essas duas ferramentas foram utilizadas para atender aos objetivos desta pesquisa, como será possível observar em mais detalhes nos próximos capítulos. Diferentes abordagens dentro dessas duas ferramentas foram testadas culminando na *pipeline* final, OligoY. Entretanto, todas as alternativas exploradas seguem a mesma ordem lógica apresentada na Figura 2.1.

O método YGS é fundamental para a inferência de *contigs* exclusivos do cromossomo Y, tornando possível explorar ao máximo montagens já disponíveis para enriquecer o conteúdo de bases nucleotídicas conhecidas para esse cromossomo (CARVALHO; CLARK, 2013). A importância da utilização desse método, bem como uma explicação mais detalhada de como executá-lo, são melhor discutidos a seguir (Subseção 2.2.1). O Capítulo 3 inclui os resultados das análises com YGS e também um tópico sobre a remoção de contaminantes através do BlobTools (LAETSCH; BLAXTER, 2017; LAETSCH *et al.*, 2017), de maneira a assegurar que as sequências inferidas para o cromossomo Y são de fato exclusivas e pertencentes ao mesmo. As orientações gerais de execução do BlobTools estão contidas na Subseção 2.2.2.

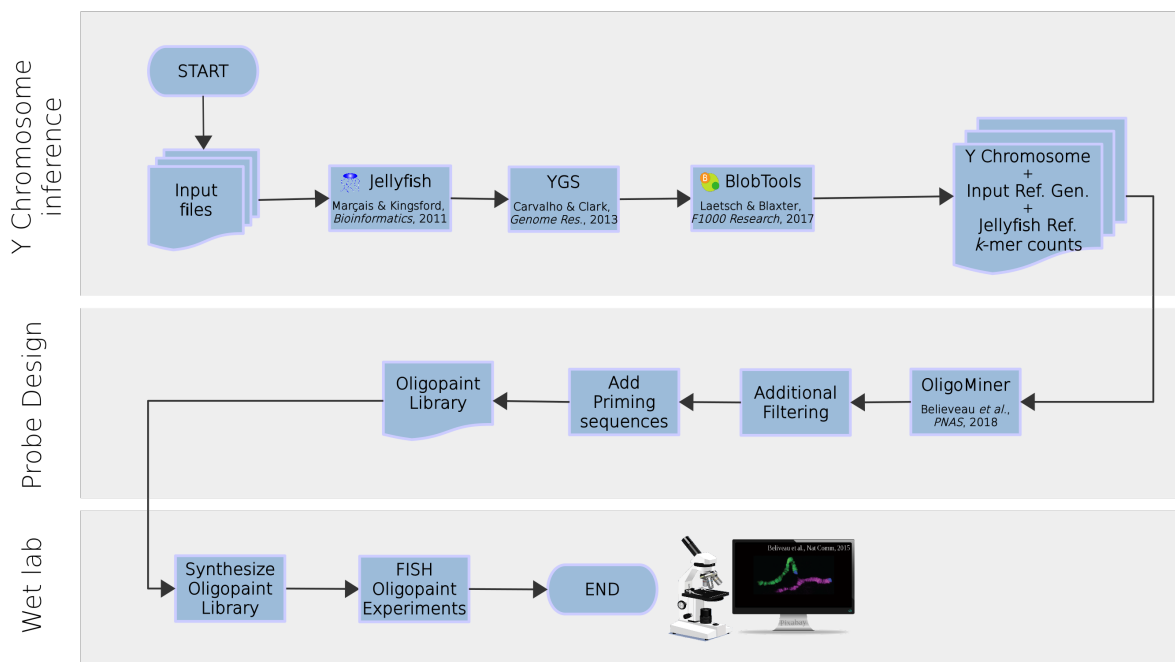


Figura 2.1: Fluxograma generalizado para desenho de sondas *oligopaint* para o cromossomo Y

Nota da figura: De cima para baixo, a *pipeline* pode ser dividida em três etapas: (i) Inferências de sequências para o cromossomo Y (*Y Chromosome inference*); (ii) Desenho de sondas (*Probe Design*); e (iii) Laboratório Molhado (*Wet Lab*). Para começar, os arquivos de entrada fornecidos são o genoma de referência e os *short-reads* de macho e fêmea. Na primeira etapa (i) esses arquivos são processados através do Jellyfish (MARÇAIS; KINGSFORD, 2011) e as tabelas *hash* geradas são usados para inferir *contigs* para o cromossomo Y através do YGS (CARVALHO; CLARK, 2013). Ainda nessa etapa, são removidas sequências contaminadas através do Blobtools (LAETSCH; BLAXTER, 2017; LAETSCH *et al.*, 2017). Na segunda etapa (ii), as sequências puras inferidas para o cromossomo Y são utilizadas para desenhar e filtrar sondas candidatas através do OligoMiner (BELIVEAU *et al.*, 2018). Filtragens adicionais também são empregadas e as sondas são processadas para adição de sequências de *priming*. A Biblioteca de Oligopaints finalizada marca a transição *in silico* – *in situ* (iii), com a síntese da mesma e, por fim, a realização dos experimentos de FISH *Oligopaint*. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

Sucintamente, para a inferência de sequências para o cromossomo Y através do método YGS foram utilizados os parâmetros descritos na Tabela 2.2. Além do intuito de realizar diferentes análises, a escolha destes valores seguem as orientações de Carvalho e Clark (2013) para evitar a inferência de falsos-positivos para o cromossomo Y. Mais orientações sobre os mesmos podem ser encontradas na Subseção 2.2.1.

Após a inferência de sequências para o cromossomo Y, é possível dar início ao desenho e seleção de sondas *oligopaint*. O OligoMiner é a ferramenta aplicada para selecionar sondas candidatas no cromossomo de interesse a partir de diferentes critérios fornecidos pelo usuário

Parameters	Value(s)
<i>k-mer</i>	$k = 15$ e $k = 18$
pontuação <i>Phred</i>	“5”
<i>lower-count/-L</i> para <i>short-reads</i> de macho	5
<i>lower-count/-L</i> para <i>short-reads</i> de fêmea	3
tamanho do <i>hash</i>	tamanho de cada arquivo de entrada
VSCK	maior ou igual à 20
PVSCUK	menor ou igual à 75%.

Tabela 2.2: Parâmetros utilizados na análise através do YGS

Nota da tabela: *k-mer* corresponde ao tamanho para fragmentação das sequências de entrada; pontuação *Phred* corresponde à qualidade mínima dos *short-reads* fornecidas; *lower-count/-L* é um valor numérico para remoção de *k-mers* que apareçam poucas vezes em relação à cobertura do genoma; tamanho do *hash* é determinado para que o mínimo de arquivos intermediários seja gerado sem comprometer o uso de memória; VSCK (Número de *k-mers* de cópia única validados) e PVSCUK (Porcentagem de *k-mers* de cópia única validados e sem correspondência com os *reads* de fêmea) servem para processar o arquivo de saída do YGS e efetivamente inferir sequências para o cromossomo Y.

(BELIVEAU *et al.*, 2018). Através dessa ferramenta, os parâmetros utilizados para desenhar sondas podem variar conforme as necessidades de futuros usuários da *pipeline*. Além disso, o OligoMiner também oferece diversas etapas de filtragem para construir a biblioteca final de sondas *oligopaint*. Todas as peculiaridades sobre execução do OligoMiner são melhor explicadas na Subseção 2.2.3. As diferentes abordagens exploradas através desta ferramenta e seus resultados são apresentados no Capítulo 4.

Em suma, além dos testes de desenho e seleção de sondas *oligopaint* feitos a partir dos diferentes genomas de referência fornecidos (Tabela 2.1) e respectivos conjuntos de sequências inferidas para o cromossomo Y através do YGS, também foram testados três conjuntos de parâmetros (Tabela 2.3). A determinação dos valores para cada um destes conjuntos de parâmetros foi baseada em Beliveau *et al.* (2018), autores do OligoMiner.

Outros programas são também utilizados ao longo da *pipeline*, seguindo as orientações de seus respectivos autores, incluindo bowtie2 (LANGMEAD; SALZBERG, 2012) e Jellyfish (MARÇAIS; KINGSFORD, 2011). Além disso, outras ferramentas são utilizadas a fim de analisar os dados gerados, como alinhamento local com BLASTn 2.2.31+ (ALTSCHUL *et al.*, 1990), seqtk (Li, <https://github.com/lh3/seqtk>) e comandos básicos em UNIX, como grep (KERNIGHAN; PIKE, 1984; RUALTHANZAUVA, 2014), SED (MCMAHON, 1978), trade (MCILROY, 1987) e AWK (AHO; KERNIGHAN; WEINBERGER, 1978). Todos os gráficos foram gerados em Python 2.7+. Diferentes bibliotecas de Python também são necessárias para execução dos programas que compõem a suíte de programas em Python OligoMiner,

Parameters	stringent	balance	coverage
Probe length	40 – 46	36 – 41	30 – 35
Melting Temperature (°C)	47 – 52	42 – 47	37 – 42
Hybridization Temperature (°C)	47	42	37
<i>k-mer</i> size	18	18	16
<i>k-mer</i> threshold	5	5	5

Tabela 2.3: Conjuntos de parâmetros utilizados para desenho e seleção de sondas candidatas

Nota da tabela: Do primeiro para o quinto parâmetro são apresentados, respectivamente, os comprimentos mínimo e máximo permitido para as sondas, a faixa de temperatura de desnaturação, a temperatura de hibridização, o tamanho do *k-mer* e seu *threshold* (limite). Da esquerda para a direita, são apresentados os conjuntos de parâmetros mais restritivos, balanceados e permissivos. Os valores de *stringente* e *balance* seguem a determinação feita por Beliveau *et al.* (2018) e, em especial para o conjunto de parâmetros *coverage*, que são os mais permissivos, foi alterado o comprimento das sondas sugerido por esses autores de 26-32 nt para 30-35 nt. O conjunto de parâmetros *balance* corresponde aos valores padronizados do OligoMiner para esses critérios. Para todos os critérios não mencionados, foram utilizados os valores padronizados pela suíte de programas do OligoMiner.

incluindo NUPACK (DIRKS; PIERCE, 2003; DIRKS; PIERCE, 2004; DIRKS *et al.*, 2007) e scikit-learn 0.17+ (PEDREGOSA *et al.*, 2011). Em casos específicos, Python 3+ também foi necessário.

O sistema operacional utilizado nessa pesquisa foi Ubuntu 16.04.7 LTS (GNU/Linux 4.15.0-041500-generic x86_64) em uma máquina com hardware de alto desempenho (HD: 17 T; RAM: 300 G; 40 CPUs; Hertz máximo da memória: 2133 MHz; Hertz máximo da CPU: 3100 MHz; Modelo do processador: Intel® Xeon® CPU E5-2630 v4 @ 2.20 GHz). Uma máquina de uso pessoal foi utilizada em alguns momentos para gerar alguns dos gráficos e testes com YGS, sendo esta com sistema operacional Ubuntu 16.04.6 LTS (GNU/4.4.0-206-generic x86_64) com desempenho comum (HD: 87G B; RAM: 15G; 8 CPUs; Hertz máximo da memória: 2400 MHz; Hertz máximo da CPU: 4000 MHz; Modelo do processador: Intel® Core™ i7-8550U CPU @ 1.80 GHz).

Os conteúdos de bases nucleotídicas em pares de bases (*base pairs* – bP) apresentados neste trabalho foram obtidos a partir do código em AWK (AHO; KERNIGHAN; WEINBERGER, 1978) e do comando grep do UNIX (KERNIGHAN; PIKE, 1984; RUALTHANZAUVA, 2014) a seguir. A conversão para Mb foi obtida dividindo bP/1000000, já que 1 milhão de bases formam 1 Mb.

```

1 ## Entre sequencias comuns a dois diferentes arquivos
2 awk 'NR==FNR{seen[$0]=1; next} seen[$0]' file1.fasta file2.fasta | grep <=>
    -v ">" | awk '/^>/ {if (seqlen){print seqlen}; print <=>
    ;seqlen=0;next; } { seqlen += length($0)}END{print seqlen}'

```

```

3 ## Diretamente atraves de um arquivo
4 grep -v ">" file1.fasta | awk '/^>/ {if (seqlen){print seqlen}; print ↵)
    ;seqlen=0;next; } { seqlen += length($0)}END{print seqlen}'

```

Ao longo da dissertação são apresentados e explicados todos os comando e códigos envolvidos na aplicação da *pipeline*, a maior parte dos mesmos com trechos de arquivos de texto em Bash que poderiam compor um *shell script*. Complementarmente, no repositório OligoY do GitHub estão disponíveis arquivos executáveis .sh (Bash *script*), sendo possível determinar caminhos para os arquivos de entrada e definir todos os parâmetros desejados.

Na subseção a seguir é explicada a lógica de execução e os passos para se executar o YGS, incluindo as opções e parâmetros diretamente relacionados com a inferência de *contigs* para o cromossomo Y e com a prévia triagem de qualidade dos *short-reads* utilizados no processo (Subseção 2.2.1). Em seguida, são descritos os passos gerais para executar a ferramenta BlobTools, fundamental para realização da análise de contaminantes (Subseção 2.2.2). Posteriormente, são apresentadas as opções e parâmetros envolvidos no desenho de sondas candidatas e nas filtrações que podem ser realizadas através da execução da suíte de programas em Python do OligoMiner (Subseção 2.2.3).

2.2.1 Como executar e como é executado o YGS

Essa seção esclarece com detalhe cada um dos passos que devem ser executados para aplicação do método YGS (CARVALHO; CLARK, 2013) assim como os parâmetros necessários e orientações sobre a determinação de seus valores. De maneira sintetizada, a Figura 2.2 apresenta todas as etapas do método, incluindo o processamento dos arquivos de entrada, que são: um arquivo FASTA do genoma montado e arquivos FASTQ dos *short-reads* do DNA da fêmea e do macho.

O YGS.pl é a implementação do método YGS construída na Linguagem de Programação Perl por Carvalho e Clark (2013). Antes de utilizar esse programa é necessário criar tabelas *hash* com *k-mers* e respectiva contagem e, conforme as recomendações dos autores, realizar triagem de qualidade para cada um dos arquivos de entrada através do Jellyfish (MARÇAIS; KINGSFORD, 2011). Para gerar as tabelas *hash* de cada um dos arquivos de entrada, contendo os *k-mers* que passaram pela triagem de qualidade, e submeter estes arquivos ao YGS.pl, o usuário precisa necessariamente estabelecer os seguintes valores:

- tamanho da subsequência de nucleotídeos (valor de *k*) para fragmentação das sequências de entrada, que deve ser o mesmo tanto para o Jellyfish (*-m*), quanto para o YGS.pl (*m_jelly* e *kmer_size*);

- pontuação *Phred* de qualidade mínima, que deve ser convertida em sinal gráfico correspondente na Tabela ASCII (ASA, 1963), ou seja, a partir dos sinais imprimíveis da codificação ASCII, utilizar o sinal que corresponder à pontuação *Phred* mínima de qualidade desejada ($-Q$);
- valor numérico para remoção de k -mers que apareçam poucas vezes em relação à cobertura do genoma (*lower-count* e $-L$);
- tamanho do *hash* ($-s$) para que o mínimo de arquivos intermediários seja gerado sem comprometer o uso de memória – conforme o manual do Jellyfish (MARÇAIS; KINGSFORD, 2011), se o arquivo de entrada, com tamanho G , for um genoma montado, $(-s) = G$, e se for um arquivo de *short-reads*, $(-s) = (G + k * n)/0.8$, onde k corresponde ao tamanho do k -mer e n à quantidade de *reads*. A unidade de medida é determinada como no exemplo: $10M = 10$ Milhões ou $50G = 50$ bilhões.

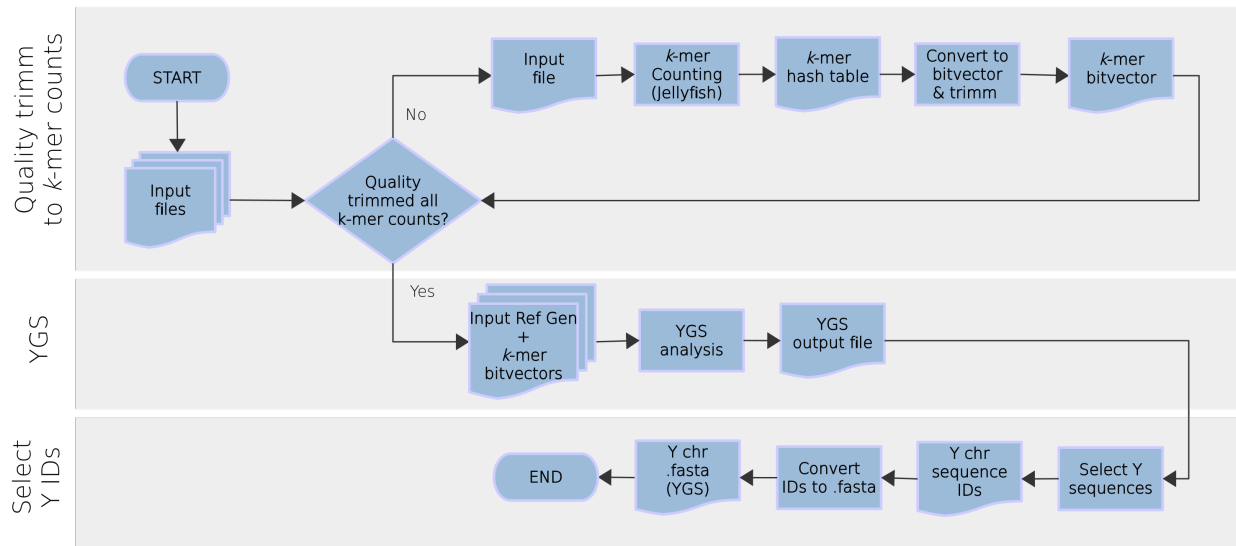


Figura 2.2: Fluxograma de execução do método YGS

Nota da figura: São necessários três arquivos de entrada: um arquivo .fasta do genoma montado e arquivos .fastq dos *short-reads* do DNA da fêmea e do macho. A primeira etapa utiliza jellyfish count (MARÇAIS; KINGSFORD, 2011) para remover *reads* com baixa qualidade e gerar tabelas *hash* com k -mers e respectiva contagem. Em seguida, essas tabelas são transformados em *bit-arrays*, removendo também k -mers com frequência baixa em relação ao genoma. Por fim, o programa do YGS pode ser executado e análises podem ser traçadas a partir de seus resultados. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

Além disso, é preciso também determinar a orientação das sequências, indicando que apenas a representação canônica de um k -mer que primeiro aparecer lexicograficamente (o próprio k -mer ou seu complemento reverso) deve ser mantida ($-C$).

O pseudocódigo abaixo é uma demonstração de arquivo de texto em Bash (*shell script*) que especifica como proceder para iniciar a execução da primeira etapa com o jellyfish count (MARÇAIS; KINGSFORD, 2011), que consiste na criação das tabelas *hash* com *k-mers* presentes em cada um dos arquivos de entrada e respectivas contagens. Essa etapa é anterior ao uso do próprio YGS.pl.

```

1 ## I - Gerar tabelas hash de k-mers com jellyfish count
2 # -t --threads          -C --both strands          -o --output
3 # -s --hash size        -m --Length of mer         -Q --min quality
4
5 gunzip -c FemaleReads*.fastq.gz | jellyfish count -t 6 -C -s ↔)
        $femaleReadsSize -m $KmerSize -o femaleReads.jelly -Q "${quality}" ↔)
        /dev/fd/0
6 gunzip -c MaleReads*.fastq.gz | jellyfish count -t 6 -C -s ↔)
        $maleReadsSize -m $KmerSize -o maleReads.jelly -Q "${quality}" ↔)
        /dev/fd/0
7 jellyfish count -t 6 -C -s $genomeSize -m $KmerSize -o genome.jelly ↔)
        Genome.fasta

```

Paralelamente à criação de tabelas *hash*, o jellyfish count também remove bases nucleotídicas de baixa qualidade, sendo esta remoção uma opção necessária apenas para os arquivos *.fastq* contendo os *short-reads* do DNA da fêmea e do macho, uma vez que alguns destes *reads* podem representar erros de sequenciamento. Além disso, *short-reads* estão normalmente compreendidos em mais de um arquivo *.fastq* – as linhas 5 e 6 do código acima exemplificam como descomprimir uma série de arquivos *.fastq.gz* e repassar essa informação para o jellyfish count. Em suma, além dos valores obrigatório de *k* (*-m*), qualidade mínima (*-Q* “*Phred* mínimo na Tabela ASCII”), tamanho do *hash* (*-s*) e orientação das sequências (*-C*), o usuário pode opcionalmente informar quantos *threads* deseja alocar para o processo (*-t*), o nome do arquivo de saída (*-o*) e utilizar também outras opções presentes no manual do Jellyfish (MARÇAIS; KINGSFORD, 2011). Após a execução dos comandos descritos anteriormente, o jellyfish count retorna o arquivo com uma tabela *hash* contendo as contagens (quantidade/*value*) para cada um dos *k-mers* (chave/*key*) presentes nos respectivos arquivos de entrada, tendo mascarado as bases nucleotídicas com qualidade inferior a *-Q* para os *short-reads* do DNA da fêmea e do macho.

Em posse destas tabelas, é possível prosseguir para a remoção de *k-mers* que são muito raros em comparação com a cobertura do sequenciamento (seguindo o valor numérico de *lower-count* fornecido pelo usuário). Por uma questão prática, esses *k-mers* devem ser

removidos, pois, são muitos e, apesar de não adicionarem nenhuma informação para a análise, consomem muita memória de processamento. Além disso, podem prejudicar a análise ao acusar *k-mers* únicos falsamente, se, pela inclusão do erro, estiveram também presentes na montagem genômica. Os valores de corte podem mudar de um arquivo para outro (*short-reads* do macho e da fêmea), já que a qualidade do sequenciamento também pode variar. Uma maneira de avaliar isso é observando os picos em um histograma gerado a partir dos *k-mers* presentes no arquivo .jelly – a seguir é apresentando o comando.

```
1 ## Gerar arquivo com histograma do .jelly
2 jellyfish histo femaleReads.jelly > femaleReads.histo
3 jellyfish histo maleReads.jelly > maleReads.histo
4 jellyfish histo genome.jelly > genome.histo
```

Biologicamente, a interpretação dos picos está relacionada com a proporção de *k-mers* hétero e homogaméticos: considerando organismos com sistema XY, o primeiro pico corresponde aos *k-mers* heterozigóticos, enquanto o segundo pico corresponde aos *k-mers* homozigóticos. Para os *short-reads* do DNA da fêmea (ou indivíduo homozigoto), se houver dois picos, o primeiro corresponderá à *k-mers* heterozigóticos que são produtos de SNPs (*Single Nucleotide Polymorphisms* – Polimorfismo de nucleotídeo único) e diferenças entre alelos; e o segundo pico conterá os *k-mers* homozigóticos advindos de sequências dos cromossomos autossômicos e X. Para os *short-reads* do DNA do macho (ou indivíduo heterozigoto), será possível visualizar dois picos iniciais, o primeiro deles incluindo, além de *k-mers* heterozigóticos que são produtos de SNPs e diferenças entre alelos, aqueles pertencentes às regiões dos cromossomos X e Y. Em machos, o segundo pico incluirá *k-mers* de sequências únicas homozigotas, ou seja, de cromossomos autossômicos. Em linhagens com baixa heterozigosidade, como *Drosophila melanogaster*, haverá apenas o pico homozigótico em fêmeas e dois picos em machos, o primeiro referente aos cromossomos X e Y, e o segundo aos autossomos. O corte de *k-mers* que são muito raros em comparação com a cobertura do sequenciamento pode ser realizado pouco antes do primeiro pico para machos e, semelhantemente, pouco antes do pico homozigótico para fêmeas (Figura 2.3). Esse corte em machos é obrigatório, pois está diretamente atrelada com a validação de *k-mers* realizada pelo YGS. Em fêmeas, os *k-mers* raros devem ser removidos para reduzir o tamanho do arquivo e respectivo uso de memória, como também para evitar erros nas comparações de *k-mers* exclusivos de machos. Com relação ao genoma montado, devem ser selecionados apenas *k-mers* que aparecem repetidas vezes (*lower-count* > 1), como requisitado para as análises do YGS.

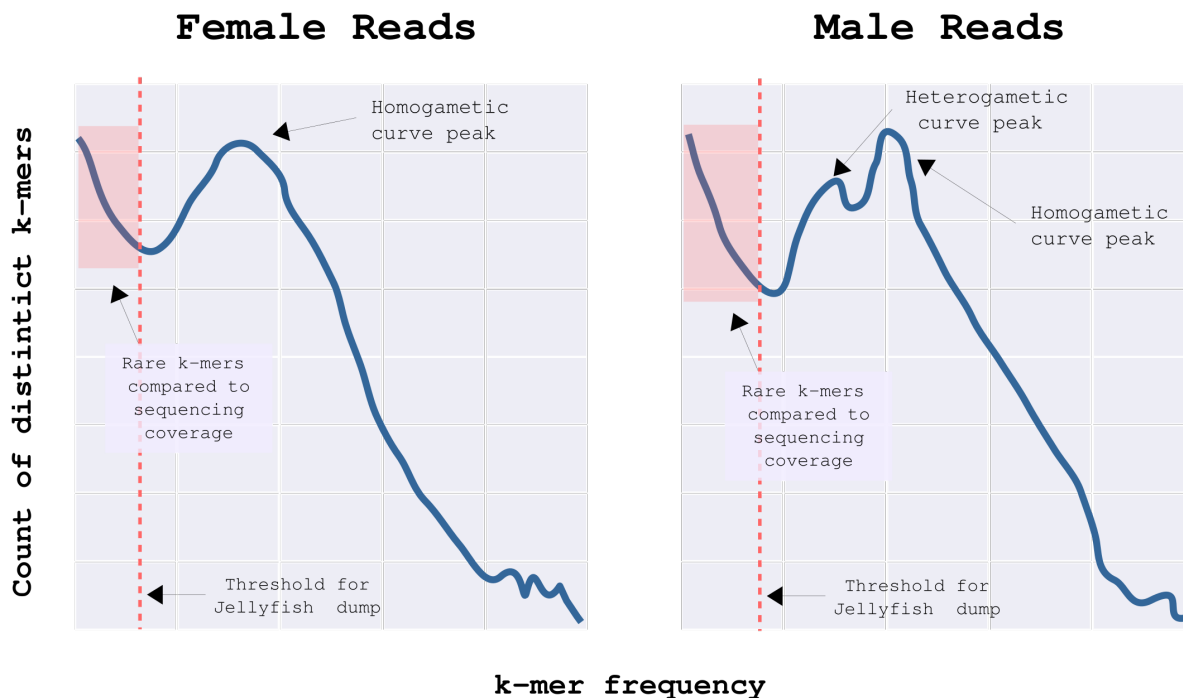


Figura 2.3: Esquema de Histograma a partir de *k-mers* distintos do Jellyfish histo

Nota da figura: O esquema mostra os picos heterogaméticos e homogaméticos para os *reads* de fêmea e de macho (à esquerda e à direita, respectivamente) em uma linhagem com baixa heterozigosidade (apenas pico homozigótico aparece em fêmeas). O tracejado em vermelho representa o *threshold* para corte, através do Jellyfish dump, de *k-mers* raros em relação à cobertura de sequenciamento (realçados na região avermelhada). Esse *threshold* deve ser determinado pelo usuário após observação dos gráficos gerados. A ordenada corresponde às contagens de *k-mers* distintos e a abscissa à frequência de *k-mers*. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

Em seguida é necessário converter as tabelas *hash* geradas em *bit-arrays*, removendo também *k-mers* com frequência baixa/raros em relação à cobertura do sequenciamento. Para gerar os *bit-arrays* a partir das tabelas *hash*, Carvalho e Clark (2013) desenvolveram um *script* em Perl que já chama o jellyfish dump (MARÇAIS; KINGSFORD, 2011) internamente para remoção de *k-mers* com frequência baixa em relação à cobertura do sequenciamento. Como essa conversão deve ser feita individualmente para cada arquivo de tabela *hash*, é retornado para o usuário arquivos individuais com cada *bit-array*. Nestes, cada *k-mer* aparecerá com correspondente presença codificada como “1” ou ausência como “0”. Posteriormente estes *bit-arrays* serão submetidos como entrada ao YGS.pl para realização das devidas comparações. A seguir, o pseudocódigo em Bash descreve os comandos e variáveis pertinentes a essa etapa.

```

1 ## II - Remover contagens baixas com jellyfish dump e gerar bit-arrays
2 # m_jelly --KmerSize          kmer_size --KmerSize
3 # jelly_file --input          lower-count --min k-mer count
4
5 perl jelly_2_bitvector_mxky_1eg.pl m_jelly=$KmerSize ↔)
6     kmer_size=$KmerSize jelly_file=FemaleReads.jelly ↔)
7     lower-count=$femaleLowerCount
8
9 perl jelly_2_bitvector_mxky_1eg.pl m_jelly=$KmerSize ↔)
10    kmer_size=$KmerSize jelly_file=MaleReads.jelly ↔)
11    lower-count=$maleLowerCount
12
13 perl jelly_2_bitvector_mxky_1eg.pl m_jelly=$KmerSize ↔)
14    kmer_size=$KmerSize jelly_file=Genome.jelly ↔)
15    lower-count=$genomeLowerCount

```

Por fim, o YGS.pl no modo `final_run` pode ser executado utilizando como entrada o genoma montado e os *bit-arrays* gerados (vide pseudocódigo a seguir). Nesta etapa o programa gera internamente outro *bit-array* para o genoma montado, desta vez a partir de todos os *k-mers* de *scaffolds/contigs* presentes no mesmo. A lógica do método YGS envolve a simples comparação desses vetores de *bits*, criando novos *bit-arrays* para as diferentes combinações. Por exemplo, será criado um *bit-array* interno com *k-mers* de cópia única: aqueles *k-mers* presentes no genoma montado, porém não repetitivos, ou seja, que estavam no *bit-array* gerado a partir de todas as sequências do genoma, mas não no *bit-array* gerado na etapa anterior (com filtragem de frequência através do jellyfish dump). Esse *bit-array* com *k-mers* de cópia única é comparado com o *bit-array* gerado a partir dos *short-reads* do DNA da fêmea na etapa anterior, resultando em uma coleção de *k-mers* de cópia única que não estão presentes entre os *k-mers* da fêmea. Dessa forma, para cada sequência, o YGS traça uma série de métricas, incluindo a proporção de *k-mers* de cópia única sem correspondente com os *short-reads* de fêmea (PVSCUK) – sendo essa uma das métricas importantes para inferência de *contigs* para o cromossomo Y.

```

1 ## III - Executar YGS no modo final_run
2 # kmer_size --KmerSize          mode --final_run, trace or contig
3 # contig --assembled genome     gen_rep --genome bit-array
4 # trace --female bit-array      male_trace --male bit-array

```

```

5 perl YGS.pl kmer_size=KmerSize mode=final_run contig=Genome.fasta ↔)
    gen_rep=GenomeVector.gz trace=FemaleReadsVector.gz ↔)
    male_trace=MaleReadsVector.gz

```

Além de proporções e contagens, o YGS também faz validação dos *k-mers* para desconsiderar sequências contaminadas no genoma montado. Para tanto, é utilizado o *bit-array* gerado a partir dos *k-mers* dos *short-reads* do DNA do macho. Essa validação é necessária por conta de que os genomas montados podem conter sequências contaminadas e, respectivamente, *k-mers* que normalmente não estariam presentes nem no genoma montado, nem entre os *short-reads* sequenciados a partir do DNA da fêmea. Assim, a alta proporção de tais *k-mers* em uma sequência contaminada implicaria em sua inferência como parte do cromossomo Y, uma vez que os mesmos não teriam correspondente contra os *short-reads* sequenciados a partir do DNA da fêmea. Para evitar isso, Carvalho e Clark (2013) usaram o raciocínio de comparar os *k-mers* obtidos a partir dos *short-reads* do DNA do macho com os do genoma montado, assumindo que o contaminante estaria ausente entre os *short-reads* de macho. Por isso mesmo, PVSCUK, que é a porcentagem de *k-mers* de cópia única sem correspondência com os *reads* de fêmea e que foram **validados**, portanto estão também presentes entre os *short-reads* de macho, é a métrica que deve ser utilizada para inferência de sequências para o cromossomo Y.

O programa retorna ao usuário um arquivo com um relatório curto da execução e, para cada sequência montada (*contigs* e *scaffolds*), uma série de informações que podem ser usadas na inferência de sequências para o cromossomo Y (Figura 2.4). Dessa forma, o usuário pode definir parâmetros para selecionar *contigs*, por exemplo usando o número de *k-mers* de cópia única validados e a porcentagem destes relacionada com os que não têm correspondente entre os *reads* de fêmea, seguindo as recomendações de Carvalho e Clark (2013).

Um importante adendo concerne ao tempo de execução e uso de memória, que devem ser considerados pelo usuário antes da inicialização do processo. O programa desenvolvido por Carvalho e Clark (2013) é de ordem 4^k , sendo 4 o número de *bit-arrays* comparados e k o tamanho do *k-mer* fornecido pelo usuário. Desta forma, apesar de que o valor do *k-mer* tenha que ser grande o suficiente de maneira a evitar a correspondência entre *k-mers* não similares, de acordo com Carvalho e Clark (2013), usar um valor superior a 18 requer aproximadamente 160 Gb de memória RAM, variando ainda conforme a fragmentação do genoma inicialmente disponibilizado. Normalmente essas especificações de hardware não estão disponíveis aos usuários comuns, a menos que estes tenham acesso a um *cluster* de alto desempenho.

Por este motivo, foi desenvolvida a segunda versão desse método, YGS2 (DUPIM; ALMEIDA; CARVALHO, versão não publicada), de maneira a sanar os problemas de desem-

```

Wed Aug 26 08:46:43 2020

Input
contig file (fasta): genome.fasta
trace file (vector): femaleReads.vector_rep3.gz
genome repetitive file (vector): genome.vector_rep2.gz
male trace file (vector): maleReads.vector_rep5.gz

Output
final results (text): genome_femaleReads_maleReads.final_result

program: /data1/programas/YGS/YGS.pl version: v_11b 8 Oct 2012 10AM PID: 15073
program mode: final_run
kmer_size: 15

command line:

perl /data1/programas/YGS/YGS.pl kmer_size=15 mode=final_run contig=genome.fasta
gen_rep=genome.vector_rep2.gz trace=femaleReads.vector_rep3.gz male_trace=maleReads.vector_rep5.gz

load_vector finished. 19378995 kmers loaded from file genome.vector_rep2.gz
load_vector finished. 90554159 kmers loaded from file femaleReads.vector_rep3.gz
load_vector finished. 90620231 kmers loaded from file maleReads.vector_rep5.gz

GI      NUM      MAX_K  K      UK      SC_K  SC_UK  P_SC_UK  VSC_K  VSC_UK  P_VSC_UK
>gi|1    1      17890  15297  2611  494   470    95.1    416    392    94.2
>gi|2    2      24627  19144  3447  1361  1337   98.2    1164   1142   98.1
>gi|3    3      13143  10819  21    50    20     40.0    31     1     3.2
>gi|4    4      1129   867    13    91    13     14.3    78     0     0.0
>gi|5    5      3769   3769   0      0      0      .       0      0      .

[all lines including >gi|6 until >gi|1865]

>gi|1866 1866 2908 2982 1144 680 667 98.1 657 644 98.0
>gi|1867 1867 5511 5235 961 787 740 94.0 739 693 93.8
>gi|1868 1868 5516 4185 1160 483 399 99.0 373 369 98.9
>gi|1869 1869 6449 5418 2367 46 46 100.0 1 1 100.0
>gi|1870 1870 60431 31793 122 1911 98 5.1 1832 22 1.2

processing contig finished
contig_wide analysis: (Note: kmers_found is the sum of kmers found in each contig, and hence
contains between-contig repeats)

1870 contigs. 142539220 max kmers 114878235 kmers found

within contigs repeats: 27660985 kmers

genome_wide analysis:

total kmers in the genome: 89301150
scp kmers in the genome: 69922155
rep kmers in the genome: 19378995 ( 21.7 % of the total )

unmatched total kmers in the genome: 790540 ( 0.9 % of the total genome kmers )
unmatched scp kmers in the genome: 529456 ( 0.8 % of the scp genome kmers )
unmatched rep kmers in the genome: 261084 ( 1.3 % of the rep genome kmers )

total kmers in the traces: 90554159
unmatched kmers in the traces: 2043549 ( 2.3 % of the total )

program finished at: Wed Aug 26 09:33:37 2020

```

Figura 2.4: Estrutura inicial e final do arquivo de saída do YGS.pl

Nota da figura: Estrutura inicial à esquerda, seguida da estrutura final do arquivo à direita. Descrição das colunas impressas no arquivo: GI – nome da sequência, assim como aparece no genoma montado; NUM – index numérico da sequência; MAX_K – número máximo de k -mers (se todos fossem únicos, ou seja, MAX_K = tamanho da sequência - tamanho do k -mer + 1); K – número de k -mers encontrados na sequência; UK – número de k -mers sem correspondente nos *reads* de fêmea; SC_K – número de k -mers de cópia única encontrado na sequência em relação ao genoma e sem validação; SC_UK – número de k -mers de cópia única sem correspondente nos *reads* de fêmea; P_SC_UK – porcentagem de SC_UK em relação à SC_K; VSC_K – número de k -mers de cópia única validados; VSC_UK – número de k -mers de cópia única validados e sem correspondente nos *reads* de fêmea; PVSCUK – porcentagem de VSC_UK em relação a VSC_K. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

penho e facilitando a usabilidade da ferramenta. Agora as tabelas *hash* geradas pelo Jellyfish (MARÇAIS; KINGSFORD, 2011) não precisam mais ser convertidas em *bit-arrays* e, em particular para a tabela *hash* formada a partir do genoma montado, apenas k -mers únicos são mantidos. Isso leva a algumas alterações na lógica de comparações realizadas pelo programa. Assim, por exemplo, não é mais necessário gerar internamente um quarto *bit-array* com os todos os k -mers presentes no genoma montado (únicos e repetitivos). Alterações como essa tornaram o processo mais eficiente em termos de uso de memória e tempo de execução. Vale ressaltar que este trabalho está sendo desenvolvido paralelamente à pesquisa aqui apresentada, sendo fruto de uma colaboração entre a autora do presente e o grupo de seu co-orientador, o Professor Dr. Antonio Bernardo de Carvalho, com autoria principal de Eduardo Dupim. Além de sumariamente importante para obtenção de alguns resultados que serão apresentados a seguir (Capítulo 3, Seção 3.1), com a publicação do mesmo e disponibilização da ferramenta, se espera que o método YGS passe a ser amplamente utilizado.

2.2.2 Análise de Contaminantes com BlobTools

As análises de contaminantes foram feitas através do BlobTools (LAETSCH; BLAXTER, 2017; LAETSCH *et al.*, 2017). De maneira sintetizada, a Figura 2.5 apresenta todas as etapas do método, incluindo o processamento dos arquivos gerados.

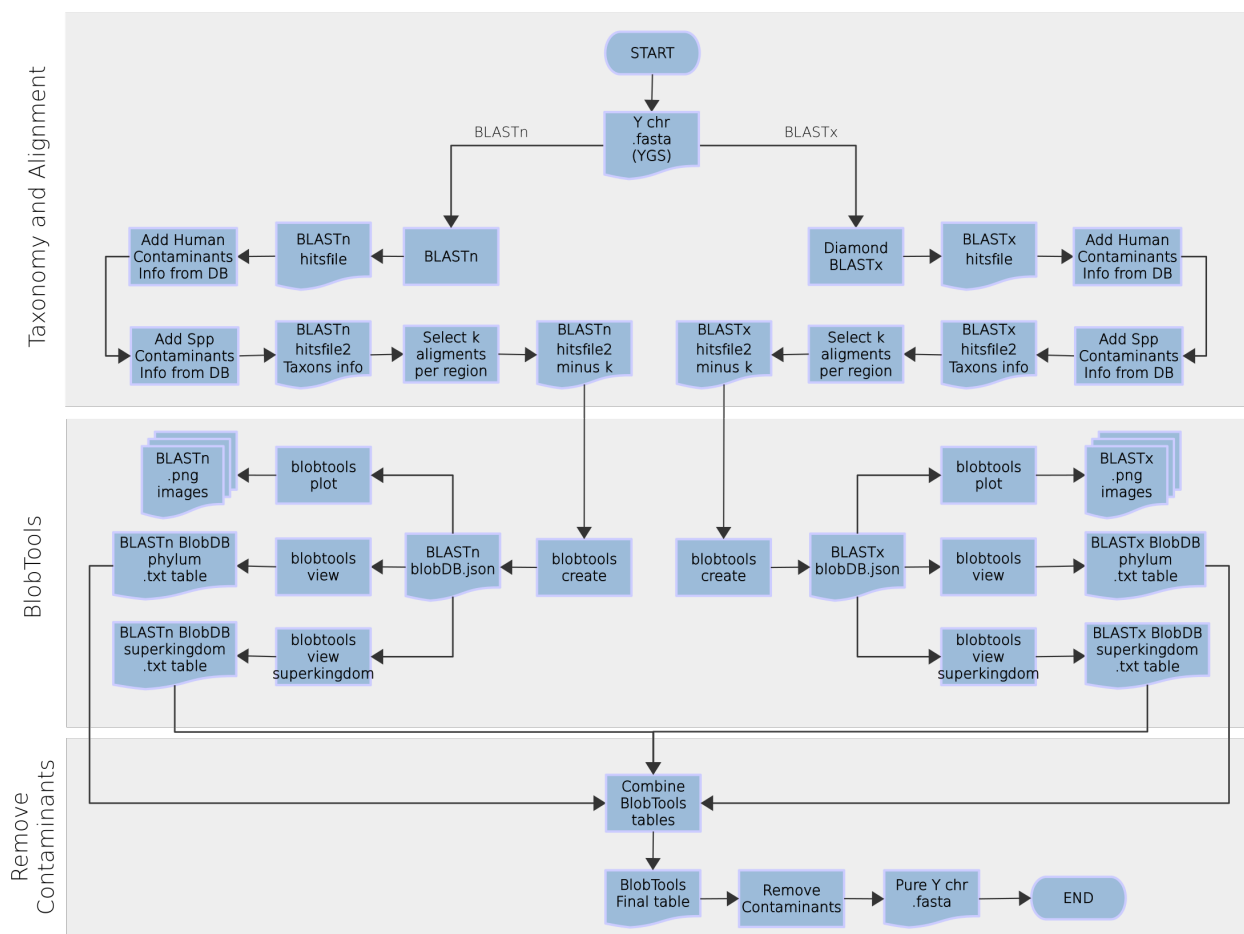


Figura 2.5: Fluxograma de execução do BlobTools

Nota da figura: A partir do arquivo FASTA para o qual se deseja realizar análise de contaminantes e do respectivo arquivo de cobertura é possível executar o BlobTools. A primeira etapa (*Taxonomy and Alignment*) é realizar alinhamentos e processamento das informações taxonômicas – aqui exemplificada com BLASTx e BLASTn (ALTSCHUL *et al.*, 1990; BUCHFINK; XIE; HUSON, 2014). Em seguida, o BlobTools em si é executado, gerando imagens e tabelas. Por fim, as informações dessas tabelas podem ser combinadas e os contaminantes encontrados podem ser removidos do arquivo FASTA de entrada, gerando um arquivo final puro – que nesse caso é o cromossomo Y. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

Esse programa analisa o conteúdo das bases nucleotídicas guanina e citosina presentes em cada uma das sequências do arquivo .fasta, assim como a cobertura da biblioteca de sequenciamento (quando disponível) e a taxonomia da sequência de acordo com os resultados

de alinhamento, os quais também devem ser fornecidos como arquivos de entrada. A proporção de guanina e citosina varia substancialmente entre genomas (LAETSCH; BLAXTER, 2017; LAETSCH *et al.*, 2017) e, combinando com as informações de taxonomia e cobertura, o BlobTools consegue distinguir as sequências contaminantes presentes.

De acordo com as orientações dos autores dessa ferramenta, o arquivo de *hits* (alinhamentos) deve ser tabular e conter, da primeira para a terceira coluna, respectivamente: a identificação da sequência; a identificação do táxon de acordo com o NCBI; e a pontuação do alinhamento. O usuário pode realizar o alinhamento usando o banco de dados de sua preferência, sendo que análises individuais do BlobTools poderão ser executadas a partir dos alinhamentos realizados contra o banco de dados de nucleotídeos e de proteínas. Aqui os banco de dados utilizados foram gerados pela equipe do Professor Dr. Antonio Bernardo de Carvalho (VANDERLINDE *et al.*, 2018) a partir de banco de dados de proteínas RefSeq e considerando Breitwieser *et al.* (2019). Os arquivos resultantes podem ser processados com AWK (AHO; KERNIGHAN; WEINBERGER, 1978) para adicionar as informações de táxons de acordo com contaminantes humanos e de outras espécies (VANDERLINDE *et al.*, 2018). Em seguida, o BlobTools pode ser executado a partir dos arquivos fornecidos, resultando em um banco de dados que deve ser processado para gerar gráficos (blobtools plot) e arquivos tabulares (blobtools view) que podem ser analisados pelo usuário. A seguir, são apresentados os comandos executados para realização da análise de contaminantes usando o programa descrito.

```

1 ## I - Taxonomia e Alinhamento com Alinhador de preferencia do usuario
2
3 ## II - Processamento dos alinhamentos gerados
4 # Arquivos de entrada que variam para alinhamento contra banco de dados ↔
   de proteínas ou contra nucleotídeos:
5 #       Breitwieser.txt & chrY.hitsfile
6 # O script minusK_blob.awk -- seleciona apenas k alinhamentos por local
7 awk 'BEGIN {OFS="\t"} FNR==1{ f++ } f==1 { a[$1]=$1; next } { if($4 in ↔)
   a); else print $0 }' Breitwieser.txt chrY.hitsfile > ↔)
   humanContam_chrY.txt
8 awk 'BEGIN {OFS="\t"} FNR==1{ f++ } f==1 { a[$1]=$1; next } { if($2 in ↔)
   a); else print $0 }' SppContaminants.txt humanContam_chrY.txt > ↔)
   contam_chrY.hitsfile2
9
10 minusK_blob.awk k=3 contam_chrY.hitsfile2 > contam_chrY.hitsfile2_k3

```



```

11 ## III - BlobTools
12 # -i --FASTA/blobDB.json -c --coverage file -o --BlobDB output prefix
13 # -t --Hits file in format (qseqid\ttaxid\tbitscore)
14 blobtools create -i chrY.fasta -c chrY.coverage -t ↔)
    contam_chrY.hitsfile2_k3 -o blob_chrY_k3
15 blobtools plot -i blob_chrY_k3.blobDB.json -p 10 --noreads --colours ↔)
    blobcolors_24ago19.txt --sort_first "no-hit" --multiplot
16 blobtools view -i blob_chrY_k3.blobDB.json
17 blobtools view -i blob_chrY_k3.blobDB.json -r superkingdom -o superkingdom

```

2.2.3 Como executar e como é executado o OligoMiner

Considerando que sondas *oligopaint* são desenhadas a partir de sequências genômicas, o arquivo de entrada que o usuário precisa fornecer para o OligoMiner desenhar sondas candidatas consiste exatamente na região sequenciada de interesse em formato FASTA (Figura 2.6). Em específico, a versão mais recente do OligoMiner (1.7) suporta apenas arquivos FASTA com uma sequência (*single-entry* FASTA). Além disso, é necessário um genoma de referência contra o qual essas sondas candidatas deverão ser alinhadas. O caminho essencial que o usuário deve percorrer para identificar sondas candidatas e prever se as mesmas possuem mais de um alvo de hibridização, a fim de filtrá-las e obter as sondas finais, envolve, respectivamente, os programas `blockParse.py` e `outputClean.py`, que fazem parte da suíte de programas em Python do OligoMiner (Figura 2.6a). Todos os arquivos BED gerados pelo OligoMiner estão estruturados como mostra a Figura 2.6b.

Varrendo a sequência de interesse, o `blockParse.py` procura por oligonucleotídeos que satisfaçam os parâmetros padronizados pelo programa e/ou que foram fornecidos pelo usuário (BELIVEAU *et al.*, 2018). Essa é uma das prerrogativas para a escolha do OligoMiner, que oferece ao usuário a possibilidade de desenhar sondas a partir de uma série de critérios que podem ser alterados conforme as peculiaridades envolvidas no experimento. São considerados, por exemplo, os tamanhos mínimo (*-l*) e máximo (*-L*) permitidos para a sonda, a temperatura de desnaturação mínima (*-t*) e máxima (*-T*), a porcentagem GC mínima (*-g*) e máxima (*-G*) permitida, assim como sequências que não devem aparecer nas sondas (*-X*), concentração de sódio (*-s*), porcentagem de formamida (*-F*), permissão ou não da sobreposição entre sondas candidatas (*-O*) e espaço mínimo entre sondas adjacentes (*-S*).

É apresentado a seguir o código para a execução desse primeiro programa, supondo que o usuário esteja fornecendo os comprimentos permitidos para as sondas e as temperaturas envolvidas no processo de desnaturação. Isso implica que, para todos os outros critérios

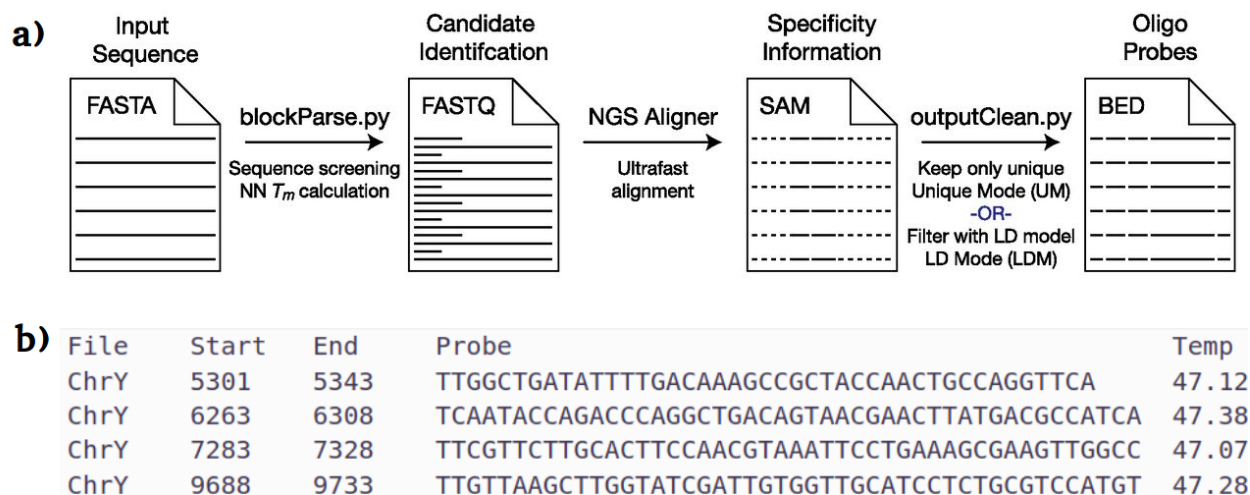


Figura 2.6: Esquema simplificado da *pipeline* do OligoMiner

Nota da figura: a) *Input Sequence* – Sequência da região para a qual se deseja desenhar sondas *oligopaint*; *Candidate Identification* – Identificação de Candidatas a partir do `blockParse.py`; *Specificity Information* – Informações de especificidade são obtidas com alinhamento NGS; *Oligo Probes* – Sondas de oligonucleotídeos são filtradas através do `outputClean.py` a partir de um dos modos que esse programa oferece (variando conforme escolha do usuário) usando como bases as informações obtidas com alinhamento NGS. O formato dos arquivos em cada uma dessas etapas é apresentado na figura (Extraído de Beliveau *et al.* (2018)); b) exemplo de arquivo BED gerado pelo OligoMiner – o cabeçalho foi inserido apenas a título de informação. Da primeira para a quinta coluna estão presentes: nome do arquivo a partir do qual a sonda foi gerada; posição de início da sonda nesse arquivo FASTA; posição final da sonda no arquivo FASTA; sequência da sonda; temperatura de hibridização da sonda. Figura b produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

analisados pelo `blockParse.py`, serão consideradas seus valores padronizadas. O arquivo padrão de saída contendo as sondas candidatas é em formato FASTQ, ideal para realização de alinhamento NGS (*Next Generation Sequencing* – Sequenciamento de Nova Geração). O usuário também pode solicitar que o mesmo seja salvo em formato BED (-b).

```
1 python2.7 blockParse.py -l $minLength -L $maxLength -t $minTempMelting ↔)
   -T $maxTempMelting -f chrY.fasta -o candidateProbesChrY
```

As sondas encontradas nesse primeiro passo são ditas candidatas, uma vez que ainda não foram submetidas a nenhuma filtragem. Para que o OligoMiner, através do programa `outputClean.py`, possa filtrar tais sondas de acordo com o número predito de seus alvos, as mesmas devem ser alinhadas contra o genoma de referência usando um alinhador NGS. De acordo com Beliveau *et al.* (2018), esse alinhamento pode ser realizado a partir do `bowtie2`, que é um alinhador de *reads* rápido e sensível (LANGMEAD; SALZBERG, 2012). Vale

explicar que existem opções predefinidas do bowtie2 e que foram projetadas de maneira a reduzir ao máximo a troca entre velocidade, sensibilidade e precisão.

Para realizar alinhamentos com essa ferramenta, é necessário gerar um *index* do genoma de referência com bowtie2-build. Tendo em conta que o OligoMiner pode filtrar sondas de diferentes modos, a sensibilidade do alinhamento pode variar de acordo com o modo que o usuário deseja utilizar: para sondas que serão selecionadas através do UM (*Unique Mode* - Modo Único), o alinhamento local precisa ser muito sensitivo, por exemplo não sendo admitidos *mismatches*, de maneira a garantir que alinhamentos múltiplos acontecerão apenas para sondas que aparecem idênticas em várias partes do genoma – por essa razão, a opção predefinida *-very-sensitive-local* é utilizada. Já para sondas que terão seu número de alvos predito a partir do LDM (*Linear Discriminant Analysis Mode* - Modo de Análise Linear Discriminante), um alinhamento local que aceita *mismatches* é suficiente. O motivo para tanto será explicado com mais detalhes em breve e está relacionado com o fato de que o LDM faz uma simulação da hibridização desses alinhamentos para filtrar as sondas candidatas, portanto, *mismatches* se tornam toleráveis e até mesmo pertinentes. De maneira contrária, o UM apenas seleciona as sondas com alinhamento único – assim, aceitar *mismatches* influenciaria na quantidade de alinhamentos para cada sonda e em sua seleção. Essas informações seguem as recomendações de Beliveau *et al.* (2018), e os comandos apresentados a seguir são uma reprodução das mesmas.

```
1 ## Gerar index do Genoma de Referencia
2 bowtie2-build Genome.fasta GenomeIndex
3
4 ## Alinhar sondas candidatas contra Index do Genoma
5 ## Para o UM
6 bowtie2 -x GenomeIndex -U candidateProbesChrY.fastq --no-hd -t -k 2 ↵)
   --very-sensitive-local -S aligned_candidateProbesChrY.sam
7
8 ## Para o LDM
9 bowtie2 -x GenomeIndex -U candidateProbesChrY.fastq --no-hd -t -k 2 ↵)
   --local -D 20 -R 3 -N 1 -L 20 -i C,4 --score-min G,1,4 -S ↵)
   aligned_candidateProbesChrY.sam
```

Como mencionado, o OligoMiner filtra sondas de acordo com o número predito de seus alvos e essa predição pode variar meramente com os resultados do alinhamento ou juntamente com os de uma análise linear discriminante. No primeiro caso, existem dois modos que podem

ser escolhidos pelo usuário para selecionar sondas usando apenas a quantidade de vezes que as mesmas alinharam no genoma de referência:

- UM, filtrando sondas que alinharam apenas uma vez no genoma de referência de acordo com arquivo de entrada SAM (-f) – para tanto, a opção -u deve ser informada, como mostrado no comando abaixo:

```
1 | python2.7 outputClean.py -u -f aligned_candidateProbesChrY.sam
```

- ZM, filtrando as sondas que alinharam zero vezes no genoma de referência de acordo com arquivo de entrada SAM (-f); esse modo pode ser útil no desenho de sondas para regiões exógenas ou transgenes – para tanto, a opção -0 deve ser informada, como mostrado no comando abaixo:

```
1 | python2.7 outputClean.py -0 -f aligned_candidateProbesChrY.sam
```

Para entender como o outputClean.py filtra as sondas, é necessário entender a estrutura básica do arquivo SAM retornado pelo bowtie2 (Figura 2.7). Para o OligoMiner, importam apenas as *flags* (segunda coluna) e as pontuações adicionais fornecidas para cada alinhamento (décima segunda coluna em diante – *Optional Fields*). Em arquivos SAM, *flags* indicam essencialmente a quantidade e a natureza de alinhamentos que não sejam únicos e/ou que provenham de *reads* pareados, aparecendo valor 0 para *reads* não pareados que alinharam uma única vez. Mais informações a respeito das *flags* reportadas por um alinhamento com bowtie2 podem ser encontradas no manual desse programa (LANGMEAD; SALZBERG, 2012) ou nas especificações do formato SAM (LI *et al.*, 2009).

Para o UM, serão selecionadas aquelas sondas que apresentarem *flag* 0 e apenas um alinhamento (campo XS não aparecer entre as pontuações adicionais). Essencialmente, isso permite que esse modo selecione sondas que tiveram apenas um alinhamento no genoma de referência. Para o ZM, serão selecionadas as sondas com *flag* 4 ou 16, respectivamente indicando nenhum alinhamento encontrado ou alinhamento na sequência reversa. Considerando que as sondas que serão sintetizadas são fita simples e terão fluoróforo adicionado na sequência que corresponde exatamente à da sonda desenhada, qualquer um dos valores 4 ou 16 corresponderá a uma sonda que, em tese e em prática, não hibridizará no genoma de referência fornecido.

De forma padronizada, o outputClean.py executa um terceiro modo, LDM, que faz análise linear discriminante em todas as sondas usando tanto as informações do alinhamento – mais precisamente seguindo os mesmos parâmetros do UM –, quanto simulando o processo de hibridização na fita complementar. É utilizada a temperatura de hibridização fornecida pelo usuário ou seu valor padrão (-T 42) para encontrar a quantidade de alvos termodinamicamente

1 Query	2 Flag	3 Reference	4 Position	5 MAPQ	6 CIGAR	7 RNEXT	8 PNEXT	9 TLEN	10 Sequence
ChrY:569-612	0	gi 175	27407400	1	44M	*	0	0	TGGGTAATTGCCCAATGAATTGCCGCTAGTTCTTGTTCGGTTGT
ChrY:569-612	256	gi 29	22397856	1	44M	*	0	0	TGGGTAATTGCCCAATGAATTGCCGCTAGTTCTTGTTCGGTTGT
ChrY:639-683	16	gi 176	2168420	1	45M	*	0	0	ACCGTAACGGGATTCAGCTCCCAATAGCTTATGCATCAGTTTCAT
ChrY:639-683	272	gi 30	3653833	1	45M	*	0	0	ACCGTAACGGGATTCAGCTCCCAATAGCTTATGCATCAGTTTCAT
ChrY:640-683	16	gi 30	3653833	1	44M	*	0	0	ACCGTAACGGGATTCAGCTCCCAATAGCTTATGCATCAGTTTCAT
ChrY:640-683	272	gi 176	532680	1	44M	*	0	0	ACCGTAACGGGATTCAGCTCCCAATAGCTTATGCATCAGTTTCAT
ChrY:641-686	0	gi 1157	23516865	1	46M	*	0	0	GAACGTGATGCATAAGCTATTGGGAGCTGAATCCCGTTACGGTTCT
ChrY:641-686	256	gi 1157	23478207	1	46M	*	0	0	GAACGTGATGCATAAGCTATTGGGAGCTGAATCCCGTTACGGTTCT
11 ASCII qualities				12- Optional Fields					
~!@#\$%^&*()-_+{} :;'"<.,>/?\`~				AS:i:88 XS:i:88 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:44 YT:Z:UU					
~!@#\$%^&*()-_+{} :;'"<.,>/?\`~				AS:i:88 XS:i:88 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:44 YT:Z:UU					
~!@#\$%^&*()-_+{} :;'"<.,>/?\`~				AS:i:90 XS:i:90 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:45 YT:Z:UU					
~!@#\$%^&*()-_+{} :;'"<.,>/?\`~				AS:i:90 XS:i:90 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:45 YT:Z:UU					
~!@#\$%^&*()-_+{} :;'"<.,>/?\`~				AS:i:88 XS:i:88 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:44 YT:Z:UU					
~!@#\$%^&*()-_+{} :;'"<.,>/?\`~				AS:i:88 XS:i:88 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:44 YT:Z:UU					
~!@#\$%^&*()-_+{} :;'"<.,>/?\`~				AS:i:92 XS:i:92 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:46 YT:Z:UU					
~!@#\$%^&*()-_+{} :;'"<.,>/?\`~				AS:i:92 XS:i:92 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:46 YT:Z:UU					

Figura 2.7: Estrutura de arquivo SAM produzido pelo bowtie2

Nota da figura: Da primeira para a última coluna estão presentes: 1 *Query* – nome do *read* que alinhou; 2 *Flag* – natureza do alinhamento; 3 *Reference* – nome da referência onde o alinhamento ocorreu; 4 *Position* – posição onde inicia o alinhamento; 5 *MAPQ* – qualidade do alinhamento; 6 *CIGAR* – operações *CIGAR* realizadas; 7 *RNEXT* – nome da referência onde alinhamento pareado ocorreu (* se nenhum *read* pareado alinhou); 8 *PNEXT* – posição onde se inicia o alinhamento no *read* pareado; 9 *TLEN* – tamanho do alinhamento no *read* pareado; 10 *Sequence* – sequência que foi alinhada; 11 *ASCII qualities* – qualidade da sequência na Tabela ASCII; 12 *Optional Fields* – campos adicionais da coluna 12 em diante com pontuações extras para cada alinhamento. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

preditos a partir desse modelo de aprendizagem de máquina. Isso permite selecionar algumas sondas que apresentaram mais de um resultado de alinhamento, bem como remover aquelas que, apesar de apresentarem um único alinhamento, não passaram na simulação de hibridização. Esse modo requer scikit-learn 0.17+ (PEDREGOSA *et al.*, 2011). O comando abaixo é uma demonstração de como executar o outputClean.py usando o LDM. Além da temperatura de hibridização, o usuário pode também alterar a probabilidade *-p* a ser usada para classificar as sondas como tendo ou não hibridização fora da sequência alvo, a concentração de sal *-s* e a porcentagem de formamida *-F* que devem ser utilizadas para o cálculo da temperatura de desnaturação.

```
1 python2.7 outputClean.py -l -T $hybridTemp -f ↔)
   aligned_candidateProbesChrY.sam
```

O arquivo retornado pelo outputClean.py é em formato BED contendo as sondas filtradas na quarta coluna, nomeado de forma padrão com o mesmo nome do arquivo de entrada + __probes.bed. Uma vez que as sondas foram filtradas segundo a predição de seus alvos, o usuário pode ou não realizar outras etapas de filtragem. Na suíte de programas do

OligoMiner, existem algumas funcionalidades que são úteis justamente para isso, como o `kmerFilter.py` e `structureCheck.py`.

O filtro de *k-mers* é realizado para remover aquelas sondas que, mesmo após filtradas pelo `outputClean.py`, ainda apresentem *k-mers* que sejam abundantes no genoma de referência. Para tanto, o `kmerFilter.py` requer um arquivo BED (*-f*) contendo sequências de sondas na quarta coluna e ainda que o Jellyfish esteja instalado e no mesmo PATH (MARÇAIS; KINGSFORD, 2011). Isso porque esse programa usa a tabela *hash* de *k-mers* de tamanho *-m* elaborado a partir do genoma de referência com `jellyfish count` para procurar e remover sondas com *k-mers* abundantes na referência. Essa remoção é realizada quando o *threshold* (*-k*) escolhido pelo usuário é atingido. Dessa forma, são retornadas em um arquivo BED (*-o*) apenas as sondas que passaram o *threshold* e esse arquivo é nomeado conforme as preferências do usuário ou é automaticamente adicionado, após o nome do arquivo de entrada, o tamanho dos *k-mer* presentes na tabela *hash* e o *threshold* utilizado. Os comandos para gerar a tabelas *hash* de *k-mers* usando o `jellyfish count` e para executar a filtragem de *k-mers* abundantes através do `kmerFilter.py` são apresentados a seguir.

```

1 ## Tabela hash de k-mers do Genoma de Referencia
2 # -s --hash size      -m --Length of mer      -o --output
3 # -c --out-counter-len Reduz tamanho do arquivo de saida
4 # -L --lower-count Nao imprime k-mer menor que -L
5 jellyfish count -s $genomeSize -m $kmerSize -o Genome.jf -c 1 -L 2 ↔)
    Genome.fasta
6
7 ## Filtrar k-mers abundantes no Genoma de Referencia
8 python2.7 kmerFilter.py -f aligned_candidateProbesChrY.bed -m $kmerSize ↔)
    -j Genome.jf -k $kmerThreshold -o ↔)
    aligned_candidateProbesChrY_probes_${kmerSize}_${kmerThreshold}.bed

```

Outra funcionalidade oferecida pelo OligoMiner é a filtragem de sondas que, após sua síntese, possam apresentar estruturas secundárias. A formação de estruturas secundárias não é desejada, uma vez que implica na indisponibilidade da sonda *oligopaint* para hibridização na sequência alvo. Para executar essa filtragem, o usuário precisa fornecer ao `structureCheck.py` um arquivo BED (*-f*) contendo as sequências de sondas na quarta coluna. Esse programa requer que o NUPACK esteja instalado e no mesmo PATH (DIRKS; PIERCE, 2003; DIRKS; PIERCE, 2004; DIRKS *et al.*, 2007). É preciso fornecer também a temperatura de hibridização *-T* a partir da qual será calculado se a sonda pode ou não apresentar formação de estrutura secundária. Além deste parâmetro, podem ser alterados o limite de probabilidade *-t* a ser

usado para considerar que as sondas apresentem estrutura linear, a concentração de sal $-s$ e a porcentagem de formamida $-F$ que devem ser utilizadas para o calcular, junto à temperatura de hibridização, a probabilidade de formação de estrutura secundária. O comando para executar essa filtragem é o seguinte:

```
python3.6 structureCheck.py -T $hybridTemp -f  $\leftrightarrow$ )  
aligned_candidateProbesChrY_probes_${kmerSize}_${kmerThreshold}.bed
```

Todos os programas do OligoMiner aqui descritos contam ainda com a opção de escrever um relatório ($-R$) ou imprimir essas informações diretamente no terminal ($-D$), podendo também escrever um arquivo com meta informações ($-M$). Ao ativar uma dessas opções, entretanto, a velocidade de execução dos programas cai drasticamente, assim como apontado por Beliveau *et al.* (2018), indo de poucos minutos para até mesmo mais de 24 horas. A versão 1.7 desses programas foi utilizada nas análises apresentadas nas próximas seções. Essas versões foram desenvolvidos para serem executados em Python 2. Em especial, o structureCheck.py teve uma de suas bibliotecas descontinuada e, por isso, o programa foi convertido para Python 3+ usando 2to3 para que pudesse ser aqui executado.

Em posse das sondas finais, o usuário pode fazer sua síntese *in house* ou através de um comerciante, desde que siga as especificações mencionadas no Capítulo 5 – as quais são importantes para a amplificação e seleção de sondas em uma biblioteca complexa de oligonucleotídeos, bem como para adição do fluoróforo que, efetivamente, permitirá observar a marcação. Neste trabalho, as sondas selecionadas foram sintetizadas através do comerciante Genscript (Mais detalhes no Capítulo 5, Seção 5.1).

2.3 Protocolos de validação *in situ*

O propósito da *pipeline in silico* desenvolvida ao longo desta pesquisa é desenhar sondas *oligopaint* para o cromossomo Y de qualquer espécie de interesse de maneira que tais sondas possam ser utilizadas em experimentos *in situ* de fluorescência. Assim, além dos testes de validação *in silico* aplicados ao longo do desenvolvimento da referida *pipeline*, é de suma importância a realização de ensaios FISH *Oligopaint* com as sondas selecionadas e sintetizadas através do comerciante Genscript, permitindo averiguar por completo a eficiência da *pipeline* desenvolvida no organismo modelo utilizado (Seção 2.1).

Na Subseção 2.3.1 são apresentados o protocolo e modificações realizadas para amplificar a biblioteca sintetiza de sondas *oligopaint*. Posteriormente são descritos os materiais e métodos envolvidos na seleção e fixação das amostras utilizadas nos ensaios de FISH *Oligopaint*

(Subseção 2.3.2). Por fim, na Subseção 2.3.3 estão descritos o protocolo e microscópios utilizado para realização do ensaio de FISH *Oligopaint*, incluindo também o desenho experimental.

2.3.1 Amplificação da biblioteca sintetizada de sondas *oligopaint*

Após desenhar e filtrar sondas *oligopaint*, adicionar regiões de *priming* e sintetizá-las, as sondas precisam passar por um protocolo de amplificação para que possam efetivamente ser utilizadas nos mais diversos ensaios de FISH *Oligopaint*. Existem diferentes protocolos de amplificação, que devem ser levados em consideração durante a adição de regiões de *priming*. Aqui, dada a complexidade da biblioteca sintetizada, as sondas foram preparadas para que fossem amplificadas através do protocolo IVT-RT (*In vitro Transcription – Reverse Transcription*; Transcrição *in vitro* – Transcrição Reversa) proposto por Beliveau *et al.* (2017). Esse protocolo conta com PCR em Tempo Real seguida de Transcrição *in vitro* e Transcrição Reversa para selecionar apenas sondas de interesse. Além de produzir concentrações maiores de sondas marcadas, em relação à protocolos mais simples (BELIVEAU *et al.*, 2015; MURGHA; ROUILLARD; GULARI, 2014), o protocolo IVT-RT permite maior complexidade no desenho da biblioteca, graças às duas regiões de *priming* diferentes – uma para a amplificação inicial com PCR em Tempo Real e outra para a Transcrição Reversa. Assim, diferentes combinações dessas duas regiões possibilitam a seleção de diferentes subgrupos de sondas (BELIVEAU *et al.*, 2017; MURGHA; ROUILLARD; GULARI, 2014).

A biblioteca desenhada para amplificação através desse protocolo contém oligonucleotídeos com polaridade 5'-3'. Essa sequência tem a estrutura apresentada a seguir (Figura 2.8 - *Oligo pool*). Através de PCR em Tempo Real são selecionadas apenas as sondas do subgrupo de interesse ao utilizar *primer forward* próprio do mesmo. Também durante essa reação são gerados os sítios de reconhecimento para a enzima T7 RNA Polimerase nas fitas duplas de DNA que são produzidas. Isso é possível por o *primer* reverso sintetizado, além de ser a sequência complementar reversa da região de *priming* reverso presente na biblioteca de oligonucleotídeos, carregar também o promotor da T7 RNA Polimerase – TAATACGACT-CACTATAGGG.

Na Figura 2.8, algumas regiões do promotor aparecem em negrito e são elas que correspondem aos sítios de reconhecimento da T7 RNA Polimerase distribuídos na dupla fita (MULLER; MARTIN; COLEMAN, 1989; PADMANABHAN; SARCAR; MILLER, 2019). Assim, na etapa de IVT, são transcritas em RNA as fitas duplas produzidas na PCR em Tempo Real que apresentam o sítio de reconhecimento da T7 RNA Polimerase. É importante ressaltar que a fita 3'-5' começando pelo sítio de reconhecimento serve como modelo para codificação da fita de RNA que terá a sequência correspondente à da outra fita de DNA (5'-3').

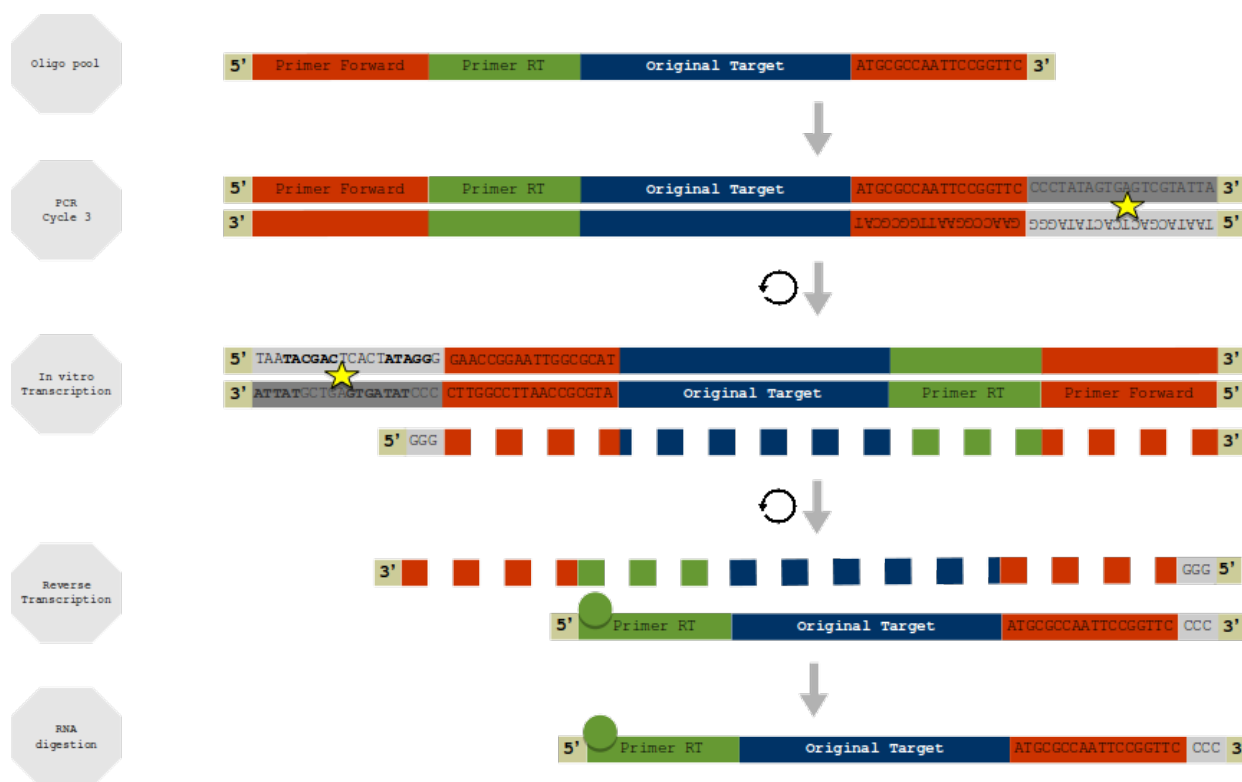


Figura 2.8: Esquema do protocolo de amplificação com IVT e RT

Nota da figura: A biblioteca de oligonucleotídeos (*Oligo pool*) é constituída de sondas com fita simples (5'-3') que começam com a região de *primer forward*, seguida da região de *primer* da Transcrição Reversa (RT), da sequência alvo originalmente sintetizada (*Original Target*) e, por fim, da região de *primer* reverso, que na figura conta com uma sequência de exemplo para melhor ilustrar o esquema. Essa biblioteca de sondas sintetizadas é utilizada para seleção inicial com PCR em Tempo Real. Os ciclos 2 e 3 de PCR em Tempo Real produzem fitas duplas de DNA com sítios de reconhecimento da enzima T7 RNA Polimerase (regiões da fita em tons de cinza e marcadas com estrela); além disso algumas fitas simples, que não aparecem na figura, também são geradas, tendo um ou nenhum sítio de reconhecimento. Em seguida, as fitas duplas de DNA com sítio de reconhecimento da T7 RNA Polimerase são utilizadas como molde para a Transcrição *in vitro* de DNA para RNA (linha pontilhada) – a imagem está invertida pois é nessa direção que as fitas entram na enzima, facilitando assim o reconhecimento das fitas codificante (3'-5') e não codificante (5'-3') com polaridade determinada do lado que começa o sítio de reconhecimento da enzima. O RNA é produzido a partir das últimas 3 citosinas presentes no promotor da T7 RNA Polimerase. Nele, o *primer* da RT acoplado ao fluoróforo irá anelar e permitir a Transcrição Reversa para DNA fluorescentemente marcado. Por fim, é recomendado fazer a digestão do RNA. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa, com auxílio de Carolina Mendonça.

Por esse motivo é seguro concluir que, ao adicionar uma mesma região de *priming* reverso para todas os conjuntos de sondas presentes na mesma biblioteca, não é gerada a possibilidade de que a seleção seja anulada pela formação de RNA de todos os oligonucleotídeos. Isso porque a T7 RNA Polimerase não conseguiria sintetizar a fita de RNA a partir de uma fita simples

de DNA não codificante que tenha polaridade 5'-3' (começando pelo sítio de reconhecimento da enzima). Ou seja, mesmo que o primeiro ciclo da PCR em Tempo Real adicione um dos promotores da T7 RNA Polimerase em todos os oligonucleotídeos, formando fitas não codificantes, a fita codificante só é gerada a partir da adição de *primer forward* específico.

Por fim, a fita de RNA sintetizada é transcrita reversamente à DNA a partir de *primer* acoplado ao fluoróforo desejado. Dessa forma, a fita de DNA produzida é fluorescentemente marcada e, após digestão do material de RNA presente, as sondas estarão prontas para serem utilizadas nos mais diversos experimentos de FISH *Oligopaint*.

Note que diferentes combinações entre as regiões de *priming* da PCR em Tempo Real e do processo de RT permitem selecionar e adicionar fluoróforo apenas no grupo de sondas de interesse, assim como discutido no Capítulo 5 (Seção 5.1).

Vale ressaltar que algumas modificações foram realizadas em relação ao protocolo original de Beliveau *et al.* (2017). A primeira delas concerne substituição de dois reagentes, a enzima Taq DNA Polimerase de alta fidelidade e o corante intercalante fluorescente para PCR em Tempo Real: *Phusion Hot-start Master Mix* (New England BioLabs), que contém enzima Taq DNA Polimerase com fidelidade 50 vezes maior em relação à enzima comum (Taxa de erro: 4.4×10^{-7}) e corante Eva Green foram substituídos por *Maxima SYBR Green/ROX qPCR Master Mix (2X)* (ThermoFisher Scientific/Invitrogen), que contém enzima Taq DNA Polimerase com fidelidade similar à enzima comum (Taxa de erro: 2×10^{-4}) e corante SYBR Green (FILGES *et al.*, 2019; WITTE *et al.*, 2018). Essa substituição foi realizada em razão da disponibilidade desse *kit* em laboratório.

Além disso, de maneira generalizada Beliveau *et al.* (2017) sugerem um programa de PCR em Tempo Real com incubação inicial a 98 °C por 3 minutos, seguido de incubação a 98 °C por 5 segundos e, por fim, incubação a 72 °C por 20 segundos. Esse programa de PCR em Tempo Real foi alterado conforme cada um dos *primers* utilizados. A reação de PCR em Tempo Real, entretanto, foi interrompida antes que a saturação fosse atingida, conforme recomendado pelos autores. Sendo este um aspecto muito importante para evitar o desequilíbrio da reação.

Os produtos das reações foram testados conforme recomendado por Beliveau *et al.* (2017), mas os ensaios de eletroforese realizados foram ligeiramente alterados para se adequar a disponibilidade de materiais. Para o produto da PCR em Tempo Real foi realizada uma eletroforese horizontal com gel de agarose 4% em solução tampão TBE 1X (Tris, ácido bórico, EDTA); essa corrida de eletroforese foi feita usando solução tampão TBE 1X com voltagem constante em 120 V por cerca de 50 minutos. Para o produto da reação de amplificação em sua totalidade foi empregado uma eletroforese vertical com gel de poliacrilamida 15% desnaturante (Acrilamida, bis-acrilamida e ureia); nesse caso, antes de adicionadas ao gel, as

amostras foram misturadas na proporção 1:1 com solução tampão desnaturante (*Gel Loading Buffer II* – Invitrogen) e incubadas a 90 °C por 4 minutos; a corrida de eletroforese foi feita usando solução tampão TBE 1X com voltagem constante em 100 V por aproximadamente 50 minutos.

2.3.2 Sexagem, Dissecção e Fixação de Tecidos

Para os ensaios de FISH *Oligopaint*, além da linhagem utilizada para obtenção das sequências genômicas e dos *short-reads* de sequenciamento (*Drosophila melanogaster* ISO1), outras duas linhagens também foram empregadas no intuito de averiguar se existe alguma flexibilidade na utilização de sequências obtidas através de linhagem diferente da sequenciada. Para tanto, a linhagem selvagem de *Drosophila melanogaster*, Tempe-T (CLARK *et al.*, 2006), foi empregada em alguns destes ensaios. Essa linhagem foi originalmente identificada na cidade de Tempe (Arizona, EUA) na presença de tetraciclina, um antibiótico que ataca bactérias do gênero *Wolbachia* (MARKOW; RICKER, 1991). Além de Tempe-T, também foram realizados ensaios com a linhagem BM5, originalmente obtida por colaboração com os Professores Oliver e Kenissson do NIH (National Institutes of Health, Estados Unidos). Essa linhagem possui uma translocação entre o cromossomo X e 2 que torna os machos translocados inférteis. Dessa forma, apenas machos não translocados, que devem ser mantidos na biblioteca para manutenção da linhagem, foram utilizados nos ensaios de FISH *Oligopaint* dessa pesquisa. As bibliotecas de todas essas linhagens são mantidas em câmara quente a 22 °C para fins de manutenção regular do metabolismo. O meio de cultura utilizado contém alimentos gelatinosos, tipicamente feitos de uma mistura de água, fubá, fermento, farinha de soja, extrato de malte, xarope de milho e ágar, resistente o suficiente para que as moscas não fiquem presas no meio, mas macios para poderem se alimentar (STOCKER; GALLANT, 2008).

Os ensaios de FISH devem ser realizados com lâminas nas quais a amostra tecidual desejada tenha sido fixa, considerando que tecidos diferentes permitem melhor observar o ciclo celular em distintos momentos: A realização de ensaios a partir de amostras obtidas da dissecção do cérebro de larvas de terceiro instar de *Drosophila melanogaster* permite visualizar a célula durante a metáfase, enquanto ensaios realizados a partir de amostras de dissecção do testículo de adultos permitem visualizar a célula durante a meiose. Além disso, como o objetivo desses experimentos de validação *in situ* é a observação da marcação no cromossomo Y, seja na mitose ou na meiose, devem ser dissecados majoritariamente indivíduos machos, tanto em estágio larval como adulto. Para tanto, é essencial distinguir o sexo dos espécimes (Figura 2.9) antes da dissecção dos tecidos amostrais e posterior fixação dos mesmos em lâminas.

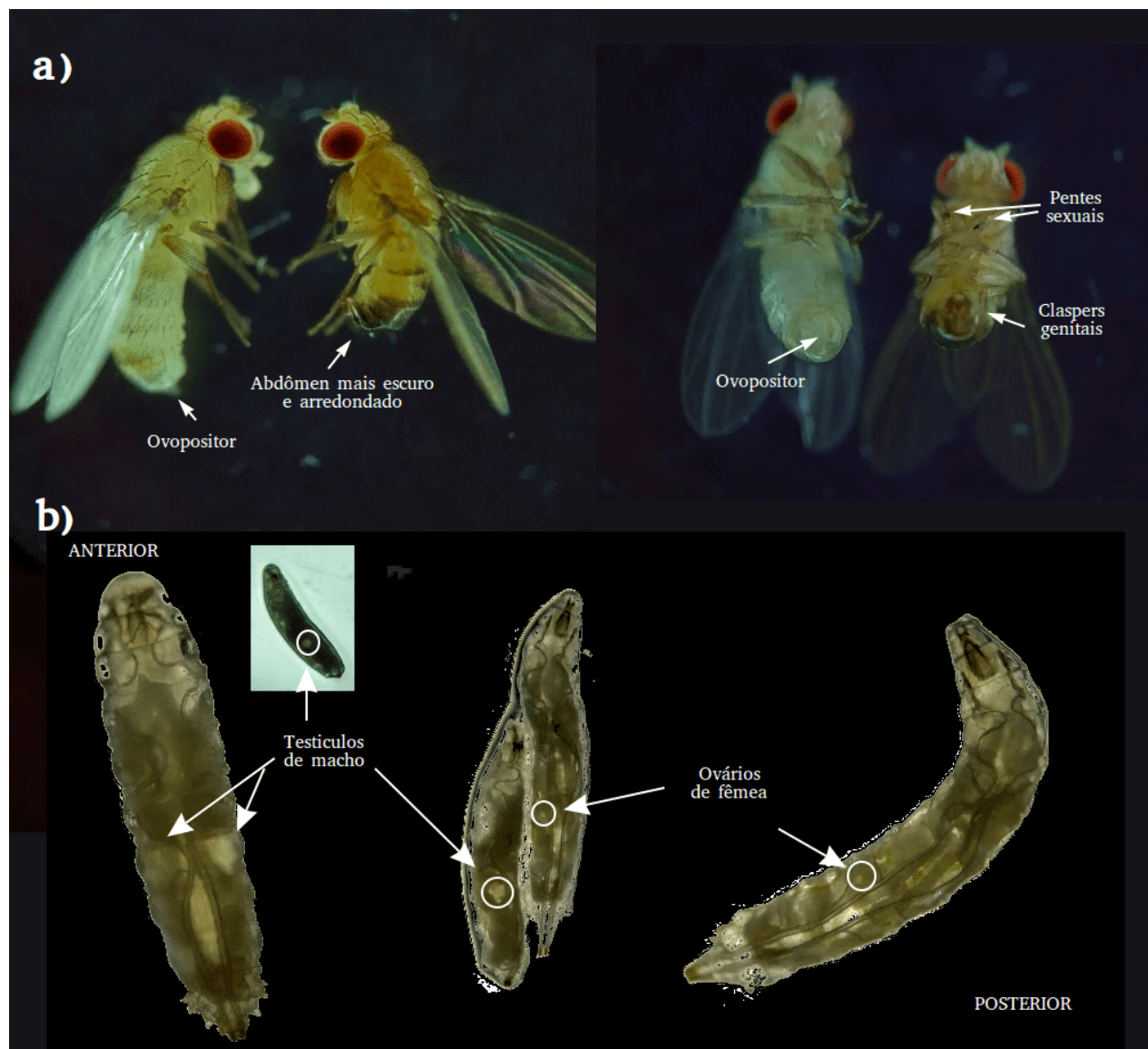


Figura 2.9: Sexagem em *Drosophila melanogaster*

Nota da figura: Distinção entre machos e fêmeas em *Drosophila melanogaster* – a) sexagem em espécimes adultos; b) sexagem em espécimes de larva de terceiro instar. Figuras produzidas por Isabela Almeida durante o desenvolvimento desta pesquisa.

Os instrumentos e reagentes necessários para a diferenciação do sexo nesses dois momentos do ciclo de *Drosophila melanogaster*, assim como os reagentes requeridos nos protocolos para dissecação e fixação dos tecidos almeçados incluem:

- Instrumentos: pinças, lâminas, alfinetes de insetos, pincel, assim como estereomicroscópio;
- Reagentes: Trietilamina (Anestesiante FlyNap) e PBS (*Phosphate buffered saline* –

Tampão salino de fosfato) são os mais frequentemente utilizados, além dos reagentes descritos nos protocolos publicados citados.

O protocolo de sexagem varia entre as espécies e seus diferentes momentos do ciclo de vida. Em *Drosophila melanogaster* a distinção entre machos e fêmeas adultas (Figura 2.9a) começa com a anestesia dos indivíduos com Trietilamina (FlyNap). Usando estereomicroscópio é possível notar que, além de menores, indivíduos machos adultos apresentam pentes sexuais em ambos os tarsos da perna anterior e um abdômen mais arredondado e escuro, formado por uma série de listras escurecidas (ZAMORE; MA, 2011). Em paralelo, as fêmeas adultas possuem ovipositores pontudos e um abdômen mais claro.

A distinção do sexo em estágio larval em *Drosophila melanogaster* requer muita habilidade de coordenação motora e visualização atenta, uma vez que o corpo dos espécimes é todo translúcido e suas gônadas, também translúcidas, devem ser identificadas. O protocolo para tal inclui o uso de pincéis para manipular as larvas cuidadosamente e buscar pela presença de grandes gônadas translúcidas na região dorsal de seus corpos – nos machos, as gônadas são grandes discos ovais embutidos no terço posterior do corpo, e, quando a larva rasteja, esses discos se movem ao longo da parede do corpo, entre os segmentos abdominais inferiores da parede corporal, apesar de permanecerem em sua posição relativa (Figura 2.9b) (MAIMON; GILBOA, 2011; PARK; GODT; KALDERON, 2018). Nas fêmeas, que, em geral, também são maiores que as larvas de machos, esses grandes discos translúcidos não estão presentes, mas sim esferas pequenas, arredondadas e mais claras – os ovários (Figura 2.9b).

Para dissecação das amostras teciduais, PBS é utilizado para mantê-las hidratadas. A isolamento dos testículos de espécimes adultos é feita mantendo o corpo do indivíduo fixo com uma pinça e, com o auxílio de outra pinça, seu abdômen é puxado do tórax, expondo o testículo anexado à vesícula seminal em meio a outros tecidos (Figura 2.10a–c). Usando alfinetes de insetos os testículos podem ser dissecados por completo (ZAMORE; MA, 2011). A dissecação do cérebro de larvas de terceiro instar também é realizada com o auxílio de pinças: a ponta dos ganchos bucais é puxada com um par de pinças enquanto o corpo é mantido fixo com outra pinça; a dissecação pode ser finalizada com um par de alfinetes de inseto (Figura 2.10d) (CAI *et al.*, 2010).

Uma vez que as amostras teciduais foram dissecadas, as mesmas podem ser fixas em lâminas, variando o protocolo para tanto conforme a amostra. O protocolo de fixação de cérebros de larva de terceiro instar começa com a dissecação de 10 à 12 destes, conforme descrito anteriormente. Com o auxílio de instrumentos comuns de bancada, estes tecidos são lavados duas vezes com 100 μ l de 1X PBS e incubados em 100 μ l de KCL (0.075 mM) por 5 minutos. Posteriormente, são retirados 30 μ l de KCl e adicionados 30 μ l de Solução de Fixação de Carnoy (3:1 etanol absoluto: ácido acético) por 5 minutos. Depois, toda a solução

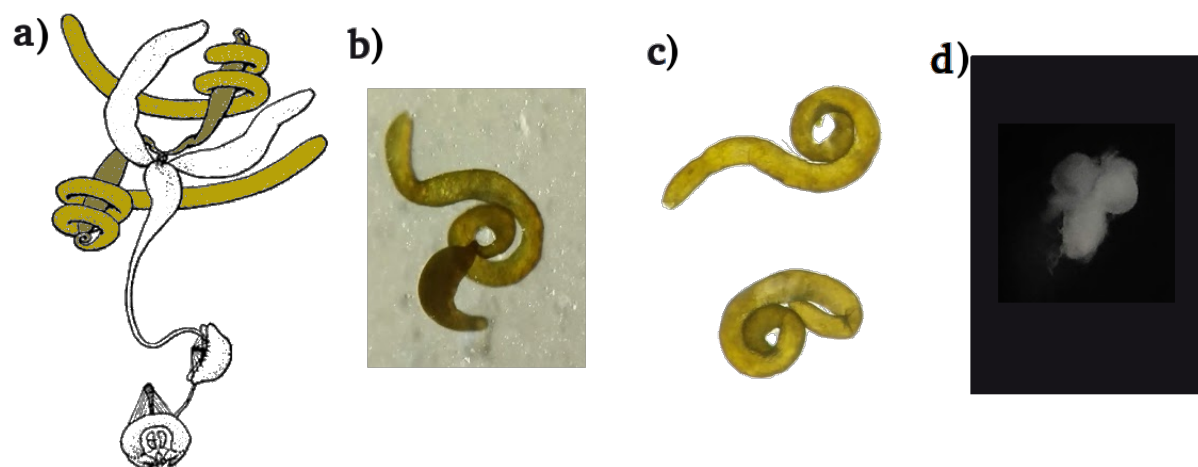


Figura 2.10: Amostras teciduais – testículos de adultos e cérebros de larvas de terceiro instar de *Drosophila melanogaster*

Nota da figura: Amostras teciduais dissecadas – a) esquema do sistema reprodutor de machos em *Drosophila melanogaster*, destacando os testículos espiralados (amarelo claro) anexados à vesícula seminal (amarelo escuro) – Figura adaptada de Miller (2008) e utilizada conforme permissão da Cold Spring Harbor Laboratory (CSHL) Press (Ver <<http://www.cshlpress.com>>); b) imagem de testículo anexado à vesícula seminal; c) par de testículos dissecados por completo; d) cérebro de larva de terceiro instar dissecada por completo. Figuras b, c e d produzidas por Isabela Almeida durante o desenvolvimento desta pesquisa.

é retirada e se adiciona 100 μl de Solução de Fixação de Carnoy por 5 minutos. Logo após, os cérebros são colocados em um microtubo e fixados com 75 μl de Solução de Fixação de ácido acético (60%). Finalmente, cerca de 20 μl da solução de cérebros homogeneizados são adicionados e espalhados em uma lâmina pré-aquecida a 40 °C. Para fins de armazenamento, as lâminas devem ser mantidas à -20 °C até o momento do uso.

Para observação do cromossomo Y durante a meiose é necessário isolar os diferentes estágios do testículo: região apical contém células enriquecidas para a fase mitótica; região proximal contém células meióticas; e região distal contém cistos de espermatócitos e baixos níveis de células pós-meioticas, normalmente enriquecidas com cistos de espermatídes em processo de alongamento e grupos de espermatídes e espermatozoides enrolados (VIBRANOVSKI; KOERICH; CARVALHO, 2008). Essas diferentes regiões podem ser isoladas com o uso de alfinetes de inseto, seguindo as recomendações de Vibranovski, Koerich e Carvalho (2008). E o protocolo para fixação de amostra tecidual de testículos segue os passos descritos por Mahadevaraju *et al.* (2021) na seção de métodos para FISH utilizando *oligopaints*.

2.3.3 Experimentos de FISH *Oligopaint*

Para validar a *pipeline* desenvolvida foi planejada a realização dos experimentos de FISH *Oligopaint* descritos na Figura 2.11.

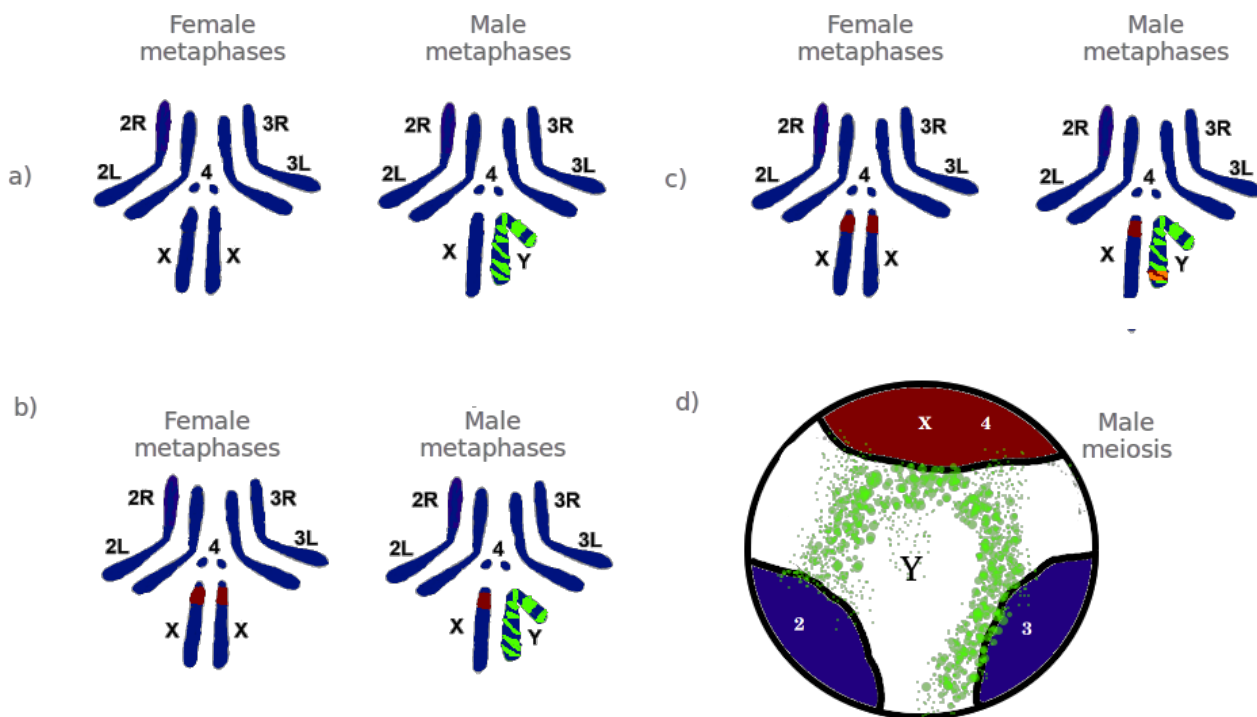


Figura 2.11: Esquema de experimentos de validação com FISH *Oligopaint* do cromossomo Y

Nota da figura: Estão representados de maneira generalizada os diferentes experimentos de validação que podem ser empregados para observar metáfases (a,b,c) e meioses (d). Pela alta fragmentação das sequências montadas para o cromossomo Y não se espera que o mesmo seja marcado por completo. Os esquemas compreendem: a) experimentos em lâminas com amostra fixada de cérebro de larva utilizando apenas as sondas *oligopaint* desenhadas para o cromossomo Y (fluoróforo Cy3 – marcação verde) e DAPI (marcando todo DNA em azul); b) experimento controle em lâminas com amostra fixada de cérebro de larva utilizando o melhor conjunto de sondas do Y (fluoróforo Cy3 – marcação verde), DAPI (marcando todo DNA em azul) e sonda para heterocromatina do cromossomo X (359 bp marcado com digoxigenina – vermelho) e, opcionalmente c) com sonda AATAC de DNA satélite do Y marcada com digoxigenina (vermelho), nesse caso a sobreposição de marcações em verde e vermelho aparece em laranja; d) experimento controle em lâminas com amostra fixada de testículo de macho adulto utilizando o melhor conjunto de sondas do Y (fluoróforo Cy3 – marcação verde), DAPI (marcando todo DNA em azul), sonda para heterocromatina do cromossomo X (359 bp) e/ou sonda ribossomal do cromossomo X (ambas marcadas com digoxigenina – vermelho). Nota: O esquema a) deverá ser reproduzido individualmente para cada um dos subgrupos de sondas sintetizadas, podendo também ser reproduzido com dois subgrupos de sondas, cada um marcado com uma cor (Cy3 ou Cy5), para observar a sobreposição ou não entre os mesmos. Os esquemas metafásicos foram adaptados a partir de Kaufman (2017) e o esquema da célula meiótica baseado em Mahadevaraju *et al.* (2021). Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

Sucintamente, esses experimentos permitem avaliar qual ou quais conjuntos de sondas *oligopaint* marcam o cromossomo Y com mais eficiência, ou seja, com maior sinal e em maior extensão. Além de demonstrar a eficiência da marcação no cromossomo Y isoladamente, é também necessário comparar essa marcação em experimentos usando sondas controle: seja usando sondas que já existem para a região satélite do cromossomo Y (AATAC) ou mesmo para sondas de outros cromossomos, como o X, de maneira a confirmar que apenas o Y está sendo marcado com as sondas desenhadas através da *pipeline* desenvolvida.

O protocolo de FISH *Oligopaint* utilizado é o de Nguyen e Joyce (2019). As lâminas produzidas foram analisadas em um Microscópio Zeiss Axiophot 2 equipado com campo claro e óptica de epifluorescência. Lâminas produzidas a partir de amostras do tecido de testículos de machos adultos puderam ser analisadas em mais detalhes através do Microscópio Confocal de Varredura a *laser* Zeiss LSM 800 com objetiva de imersão em óleo 63X / 1,4 NA.

3

Inferindo sequências para o cromossomo Y

Mesmo desprovido de um genoma completamente montado e, em particular, sem um *scaffold* de sequências bem estabelecidas para o cromossomo sexual Y/W, existem métodos que permitem inferir mais *contigs* para estes – uma vez que, em espécies com cromossomos sexuais heteromórficos, eles estão presentes somente em um dos sexos. A aplicação de um método como esse leva ao aumento da riqueza de informações e regiões conhecidas para esses cromossomos, ainda que estes *contigs* não possam ser propriamente montados em uma sequência ordenada. No contexto desta pesquisa, este aumento de sequências disponíveis é extremamente valioso por permitir desenhar sondas *oligopaint* que marcam o cromossomo Y em maior extensão e, em um cenário mais genérico, facilitar a reprodução da *pipeline* proposta independentemente da disponibilidade de genomas finalizados para a espécie sendo estudada e das diferentes circunstâncias em que a pesquisa possa ser realizada.

Com a finalidade de detectar sequências ligadas ao cromossomo Y, o método YGS foi descrito e implementado por Carvalho e Clark (2013) e parte da premissa de que o genoma de macho não finalizado pode ser varrido à procura de *k-mers* que estejam ausentes entre os *short-reads* do DNA da fêmea e que foram validados a partir dos *short-reads* do DNA do macho (Figura 3.1). Isso permite separar os *contigs* em dois grupos – aqueles que são do cromossomo Y e aqueles que não são, pertencendo assim aos cromossomos autossômicos e X.

Existem dois outros métodos que se propõem a encontrar sequências gênicas para o cromossomo Y: O CQ, desenvolvido por Hall *et al.* (2013), e o DiscoverY, desenvolvido por Rangavittal *et al.* (2019). O primeiro deles, CQ, usa DNA de macho e de fêmea e analisa, para cada sequência, a razão entre alinhamentos de fêmea e alinhamentos de macho obtidos (*chromosome quotient* – quociente cromossômico) para determinar se a sequência é ou não ligada ao cromossomo Y. O segundo método, DiscoverY, parte do pressuposto de combinar as metodologias do YGS e CQ, analisando a proporção de *k-mers* compartilhada com genoma de fêmea e a profundidade de cobertura de *reads* de macho para classificar os *contigs*; no entanto,

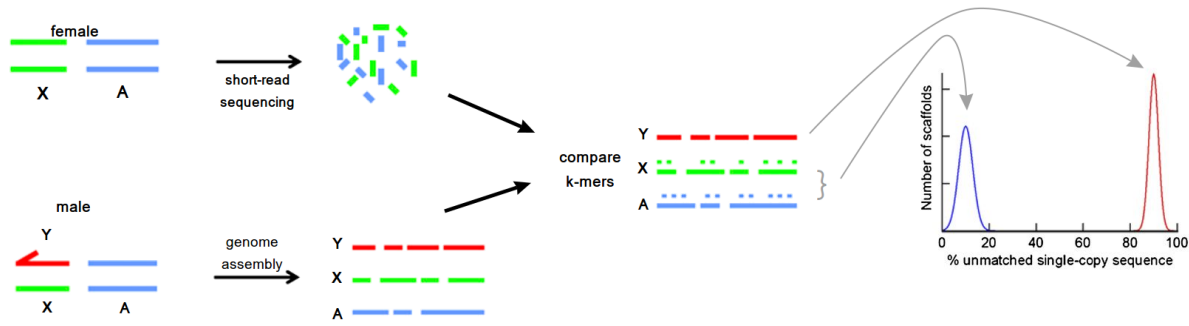


Figura 3.1: Esquema simplificado do método YGS

Nota da figura: Os *k-mers* gerados a partir dos *short-reads* da fêmea e do genoma montado são comparados – aqueles que são exclusivos do cromossomo Y só estão presentes no genoma montado de macho e, desta forma é possível fazer a distinção entre estes e os dos cromossomos autossômicos e X, como ilustrado no gráfico (Figura adaptada de Carvalho e Clark (2013)).

Rangavittal *et al.* (2019) não se baseiam de fato em toda a lógica envolvida no método YGS, por exemplo, não validando *k-mers*. Assim, estes métodos foram mais profundamente analisados por Dupim, Almeida e Carvalho (versão não publicada), trabalho ainda não publicado que apresenta a segunda versão do YGS (YGS2). Como discutido anteriormente, que este trabalho está sendo desenvolvido paralelamente à pesquisa aqui apresentada, por uma colaboração entre a autora do presente e o grupo de seu co-orientador, o Professor Dr. Antonio Bernardo de Carvalho, com autoria principal de Eduardo Dupim. O seu desenvolvimento foi sumariamente importante na obtenção de alguns dos resultados apresentados na Seção 3.1.

Na próxima seção são apresentadas as análises dos resultados e discussões pertinentes ao YGS (Seção 3.1). Em seguida, a partir dos *contigs* inferidos para o cromossomo Y, é descrita a análise de contaminantes aplicada e seus resultados (Seção 3.2). Para fins de melhor compreensão, o genoma de macho não finalizado é mencionado no texto como *genoma montado* e quaisquer exceções que figurem um genoma completamente finalizado e/ou que não seja um genoma de macho são assim esclarecidas.

3.1 Testando diferentes parâmetros

Conforme mencionado no capítulo anterior (Seção 2.2), dois valores de *k-mer* foram utilizados para executar o YGS a fim de realizar a inferência de *contigs* para o cromossomo Y: $k = 15$ e $k = 18$ (Tabela 2.2). A escolha desses dois valores é baseada em Carvalho e Clark (2013), quando explicaram que, apesar de o tamanho do *k-mer* ter de ser grande o suficiente de maneira a evitar que *k-mers* idênticos sejam alinhados, deve-se balancear também o uso de

memória, tempo de processamento e poder da máquina utilizada (Para detalhes, ver Capítulo 2, Subseção 2.2.1). Além disso, esses autores realizaram testes com diferentes tamanhos de k , mostrando que $k = 15$ era um valor ideal para *Drosophila melanogaster* e organismos com genomas similares aos de outros insetos. Para essa mesma espécie, eles mostraram que, ao utilizar $k = 17$, houve pouco ganho de resolução. Dessa forma, dado o poder de processamento disponível na máquina aqui utilizada (Capítulo 2 – Seção 2.2), testamos também $k = 18$ de maneira a averiguar se, ao aumentar um pouco o valor de k em relação aos testes realizados por Carvalho e Clark (2013), as inferências realizadas demonstrariam maior ganho de resolução e, talvez, resultariam em mais sondas finais selecionadas.

Ao submeter individualmente os *short-reads* de fêmea e macho ao Jellyfish (MARÇAIS; KINGSFORD, 2011) foram gerados os gráficos com a distribuição de frequência dos k -mers para cada um dos valores de k utilizados (Figura 3.2). Com os *short-reads* de fêmea, é observado a formação de apenas um pico (homogamético) para ambos os valores de k . Com os *short-reads* de macho são formados vários picos, dada a alta profundidade do sequenciamento (95X), sendo o primeiro deles o pico heterogamético, que em parte representa os k -mers do cromossomo Y e em parte representa os k -mers do cromossomo X, por serem heterozigóticos; e o segundo pico corresponde aos k -mers homozigóticos (dos cromossomos autossômicos) presentes em ambos os sexos. No pico heterogamético dos machos se espera que a maior parte dos k -mers sejam do cromossomo X, já que o cromossomo Y possui poucas sequências de cópia única.

De maneira a uniformizar as análises aqui realizada, a partir da observação dos histogramas gerados (Figura 3.2), o corte de k -mers raros em relação à cobertura do sequenciamento foi definido como *lower-count*=3 para os *short-reads* de fêmeas e *lower-count*=5 para os *short-reads* de macho, independente da análise ser com $k = 15$ ou $k = 18$ (Tabela 2.2). Vale ressaltar que a proporção entre os picos hétero e homogamético variam conforme o valor de k escolhido: quanto maior o valor de k , maior será o pico de k -mers heterozigóticos e menor o pico homogamético. Supondo que exista apenas um SNP a cada 100 nt, com $k = 15$ existiriam até 86 k -mers dos quais apenas 15 seriam heterozigóticos, enquanto com $k = 18$ existiriam até 83 k -mers dos quais 18 seriam heterozigóticos, aumentando assim o tamanho do primeiro pico em relação ao do segundo.

Dispondo de genomas de *Drosophila melanogaster* obtidos através de diferentes tecnologias de sequenciamento e montagem, foram necessárias distintas execuções das etapas descritas no capítulo anterior (Subseção 2.2.1), uma para cada uma das montagens e, para cada uma destas, uma execução para cada valor de k .

Outros dois parâmetros singularmente relevantes para a inferência de sequências para o cromossomo Y a partir do arquivo retornado pelo YGS (Figura 2.4) são o número

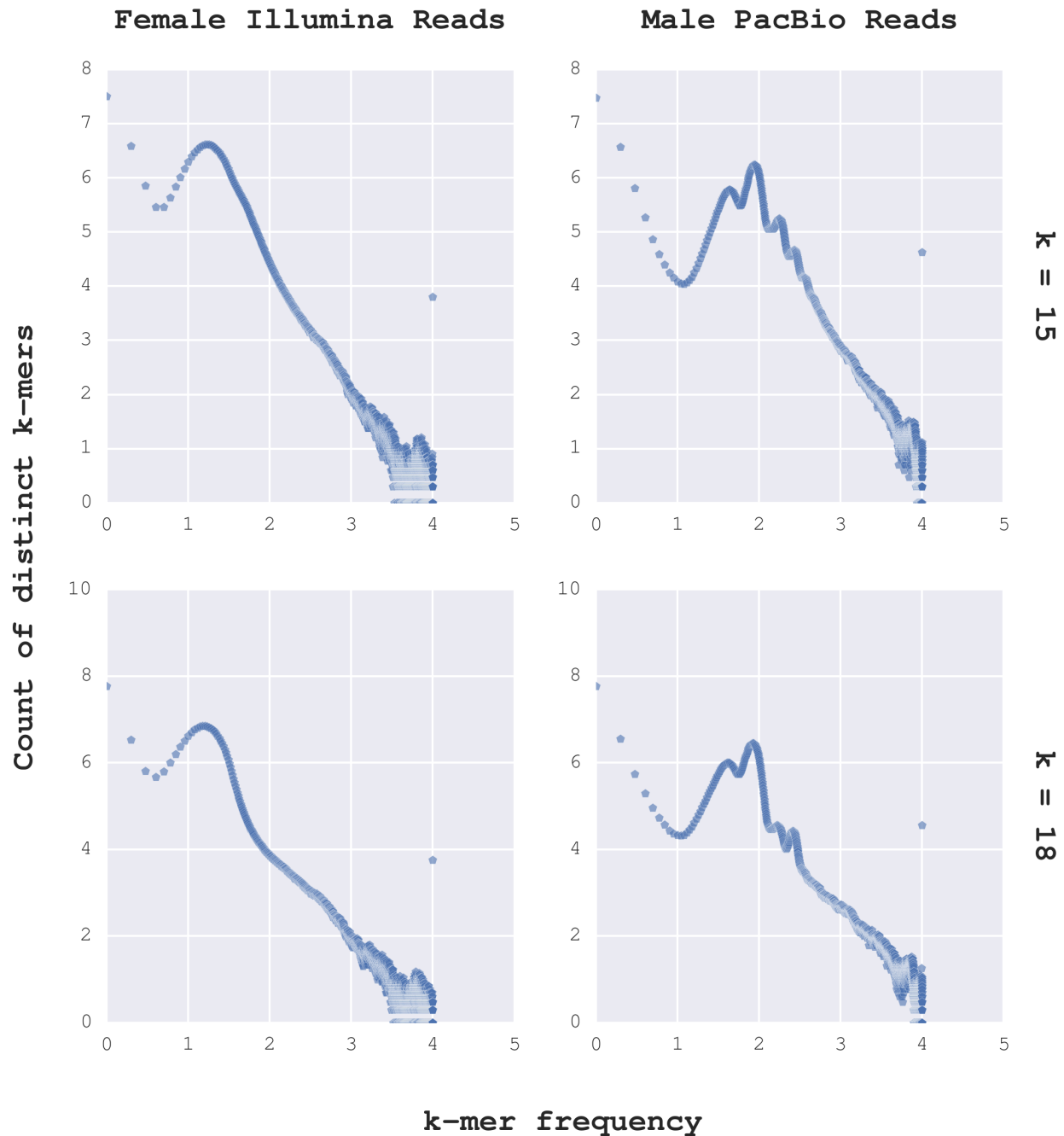


Figura 3.2: Jellyfish histo dos *reads* de fêmea e macho utilizados

Nota da figura: São apresentadas as contagens de *k-mers* distintos em relação à sua frequência para *reads* de fêmea e de macho (à esquerda e à direita, respectivamente); com $k = 15$ acima e $k = 18$ abaixo. Para os *reads* de fêmea apenas o primeiro pico é importante (pico homogamético) e para os de macho os dois primeiros picos (heterogamético e homogamético, respectivamente). Os dados estão em escala logarítmica de base 10. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

de *k-mers* de cópia única validados e a porcentagem de *k-mers* de cópia única validados e sem correspondência com os *reads* de fêmea, respectivamente VSCK e PVSCUK. Os valores escolhidos para selecionar as sequências a partir destes dois parâmetros foram de que ao menos 20 *k-mers* de cópia única validados deveriam ter sido detectados pelo YGS naquele mesmo *contig* e ao menos 75% destes não deveriam ter correspondência com os *reads* de fêmea (Tabela 2.2). Como mencionado no capítulo anterior (Seção 2.2), além de baseada em Carvalho e Clark (2013), a escolha desses parâmetros foi feita de maneira a evitar falsos-positivos, garantindo assim que sequências conhecidas para os cromossomos autossômicos ou para o cromossomo X não passassem esse *threshold* mínimo (limite mínimo) e fossem erroneamente inferidas para o cromossomo Y.

Para tanto, em um arquivo com apenas os valores gerados pelo YGS para cada uma das sequências, foi executada a linha de comando descrita abaixo, em que apenas as colunas de interesse (Coluna 1 – GI/identificadores dos *contigs*; Coluna 9 – VSCK; Coluna 11 – PVSCUK) são passadas para um comando em AWK (AHO; KERNIGHAN; WEINBERGER, 1978), empregado para selecionar os identificadores dos *contigs* que satisfaçam os critérios determinados.

```
1 cut -f1,9,11 genome_femaleReads_maleReads_onlyGIs.final_result | awk -v ↵)
    vsck="20" -v pvscuk="75" -F'\t' '{if($2>=vsck && ↵)
    $3>=pvscuk)print$1}' > genome_femaleReads_maleReads_finalresult_YGIs
```

Após submeter individualmente os diferentes genomas não finalizados ao Jellyfish e YGS (MARÇAIS; KINGSFORD, 2011; CARVALHO; CLARK, 2013) uma vez para cada valor de *k-mer*, usando como entrada os mesmos *short-reads* de fêmea e macho já descritos e analisados, foram processados (usando o código mostrado acima) cada um dos arquivos gerados pelo YGS de maneira a inferir quais *contigs* de cada montagem são do cromossomo Y. São apresentados a seguir os gráficos resultantes, mostrando o contraste entre as inferências feitas para as sequências: em azul, as inferidas para o cromossomo Y; em preto, as restantes – por correlação, inferidas para os cromossomos autossômicos ou X (Figura 3.3). O corte das sequências que apresentaram menos de 20 *k-mers* de cópia única validados aparece em vermelho. É notável que as quantidades de sequências (diferentes GIs) inferidas para o cromossomo Y variam entre uma montagem e outra, sendo visível que, para as montagens realizadas a partir de tecnologia Illumina, Sanger e, até mesmo, para a versão mais recente do genoma de *Drosophila melanogaster* (Release 6), essas quantidades são muito maiores. Contudo, o tamanho particular destes *contigs* varia exatamente com a tecnologia de sequenciamento empregada e, assim, grandes quantidades de *contigs* não necessariamente equivalem à extensas sequências de nucleotídeos. Para a Release 6, por exemplo, vemos mais *contigs* inferidos

para o cromossomo Y do que para os outros cromossomos, devido às sequências dos outros cromossomos estarem montadas em *scaffolds*, o que resulta em poucos pontos no gráfico e a restarem poucos *contigs* que não foram inferidos para o cromossomo Y ou que ficaram abaixo do corte de *k-mers* de cópia única validados.

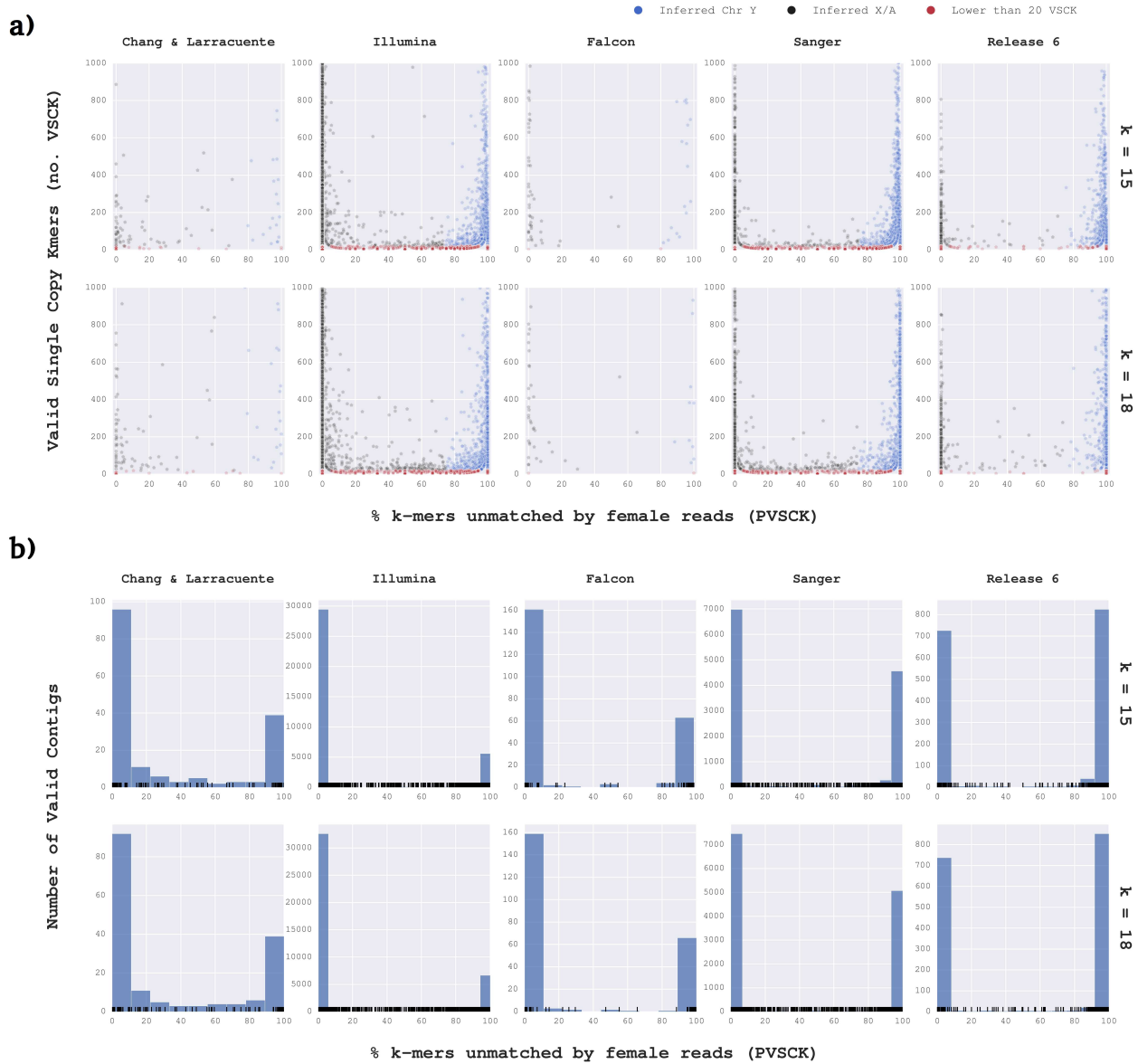


Figura 3.3: Sequências inferidas para o cromossomo Y e para os cromossomos autossômicos e X

Nota da figura: a) Contraste entre sequências inferidas para os cromossomos autossômicos e X (X/A), como resultado da análise do YGS para cada um dos arquivos de entrada e parâmetros utilizados ($k = 15$ ou $k = 18$); a ordenada corresponde ao número de *k-mers* de cópia única validados (VSCK) e a abscissa à porcentagem de *k-mers* de cópia única validados e sem correspondência com os *reads* de fêmea (PVSCUK). b) Histograma do número de *contigs* com PVSCUK. Figuras produzidas por Isabela Almeida durante o desenvolvimento desta pesquisa.

Para a montagem do genoma de *Drosophila melanogaster* feita por Chang e Larracuenta (2019), a qual possui menos *contigs* e, assim, menos *gaps*, embora existam poucas sequências, representadas por poucos pontos no gráfico (Figura 3.3a), grande quantidade de heterocromatina e informação nucleotídica está presente nas mesmas. Como nesse trabalho os autores procuravam gerar um genoma mais contíguo, diferentes abordagens de montagem foram aplicadas, incluindo o uso do montador Falcon para a montagem *de novo*, de maneira a aproveitar grande parte da informação sequenciada a partir das regiões heterocromáticas. Outra das montagens não finalizadas que aqui foram submetidas ao método YGS contou exatamente com a montagem Falcon a partir de *short-reads* da Pacific Biosciences, apesar de as abordagens de montagem empregadas serem diferentes. Repara-se que ambas as montagens contam com poucos *contigs* e, no entanto, os diferentes conjuntos de sequências inferidas para o cromossomo Y a partir destas contam com os dois maiores conteúdos de bases nucleotídicas (Tabela 3.1), sendo a montagem a partir de sequenciamento Illumina a com menor conteúdo entre todas as outras.

Assembly	Common		$k = 15$		$k = 18$	
	contigs	Mb	contigs	Mb	contigs	Mb
Chang & Larracuenta	40	13.51	40	13.51	44	13.79
Illumina	1887	1.15	1897	1.15	2964	1.30
Falcon	66	8.80	66	8.80	66	8.80
Sanger	2654	5.50	2659	5.51	3449	6.19
Release 6	768	6.57	769	6.57	821	6.65

Tabela 3.1: Tamanhos das sequências inferidas para o cromossomo Y

Nota da tabela: Números de *contigs* e pares de bases em Mb são fornecidos para as sequências que foram inferidas para o cromossomo Y a partir das diferentes montagens (*Assembly*). Os dados correspondem ao número de *contigs* e pares de bases obtidos com ambos os valores de k utilizados (*common*), seguido do total de informações para as sequências inferidas com cada valor de k .

Diferentes escolhas de valores de k figuram diferentes resultados, uma vez que o tamanho do k -mer influencia diretamente nas comparações que são feitas pelo YGS e, até mesmo, nos prévios cortes de frequência e qualidade. Entretanto, ao testar diferentes tamanhos de k -mer, apesar das análises do YGS de fato retornarem proporções que variam entre um e outro valor de k , é possível que ainda assim algumas das sequências, senão todas, passem os critérios de inferência da mesma forma. Esse aspecto foi notado nesta pesquisa: para todos os genomas submetidos à análise do YGS foram obtidas sequências para o cromossomo Y que compartilhavam *contigs* entre os dois valores de k testados (Figura 3.4). Para o genoma montado com Falcon, por exemplo, ambos os k -mers testados levaram à inferência não somente da mesma quantidade de *contigs*, mas dos mesmos 66 *contigs*. Já para a montagem

feita por Chang e Larracuent (2019) todos os *contigs* inferidos para o cromossomo Y a partir de $k = 15$ também foram inferidos com $k = 18$, apesar de que para esse último valor outros 4 *contigs* foram adicionalmente inferidos, elevando o conteúdo de bases nucleotídicas de 13.51 Mb para 13.79 Mb. A partir de cada um dos outros três genomas não finalizados que foram processados, além de majoritariamente os *contigs* inferidos para o cromossomo Y serem independentes do valor de k utilizado, individualmente ambas as análises resultaram em um acréscimo de sequências, ou seja, para a *Release 6*, Illumina e Sanger, ao menos um *contig* foi exclusivamente inferido para o cromossomo Y com $k = 15$ e também para $k = 18$.

Em especial para o conjunto de sequências inferidas para cromossomo Y a partir da montagem realizada por Chang e Larracuent (2019), quando comparado com as sequências que esses autores determinaram para o cromossomo Y, fica evidenciado que o uso do YGS permite resgatar grande parte das mesmas, independente do k -mer escolhido (Figura 3.5): das 55 sequências determinadas para o cromossomo Y na referida montagem, 42 são também inferidas para esse cromossomo pelo YGS com $k = 18$ e 38 com $k = 15$. Em conteúdo nucleotídico, menos de 2 Mb são reduzidos do cromossomo Y montado por Chang e Larracuent (2019) em relação ao conjunto de *contigs* inferidos na presente análise, havendo pouca variação entre os valores de k utilizados (Tabela 3.1).

Com um mínimo de 97.3% de k -mers válidos que não estavam presentes entre os *short-reads* de fêmea, apenas duas sequências que os autores mencionados originalmente definiram para outra região genômica (Região centromérica do cromossomo 3) foram inferidas para o cromossomo Y de acordo com as análises do YGS. Além disso, algumas sequências do cromossomo Y do trabalho mencionado não apresentam o número mínimo necessário de k -mers de cópia única válidos para que sejam inferidos para este mesmo cromossomo a partir da análise com o YGS. Isso implica mostrar que, se por um lado Chang e Larracuent (2019) aplicaram uma abordagem extremamente minuciosa, que contou com uma série de curagens manuais para enriquecer as sequências conhecidas para o cromossomo Y, o YGS propicia alcançar resultados muito semelhantes, sem demandar tanto empenho e meticulosidade nas análises efetuadas. Dessa forma, é reforçado que, ao designar o método YGS (CARVALHO; CLARK, 2013; DUPIM; ALMEIDA; CARVALHO, versão não publicada) como parte da *pipeline* para desenho de sondas *oligopaint*, a reprodução da mesma é facilitada, permitindo sua execução independente da disponibilidade de genomas finalizados para a espécie sendo estudada.

Outro aspecto importante que deve ser levado em consideração antes da reprodução dessa *pipeline* é se a máquina utilizada possui memória e processamento suficientes para tanto. Testando diferentes arquivos de entrada com $k = 15$, foi observado um máximo de 43 horas de execução (montagem a partir de sequenciamento Illumina) e mínimo de 15 minutos (monta-

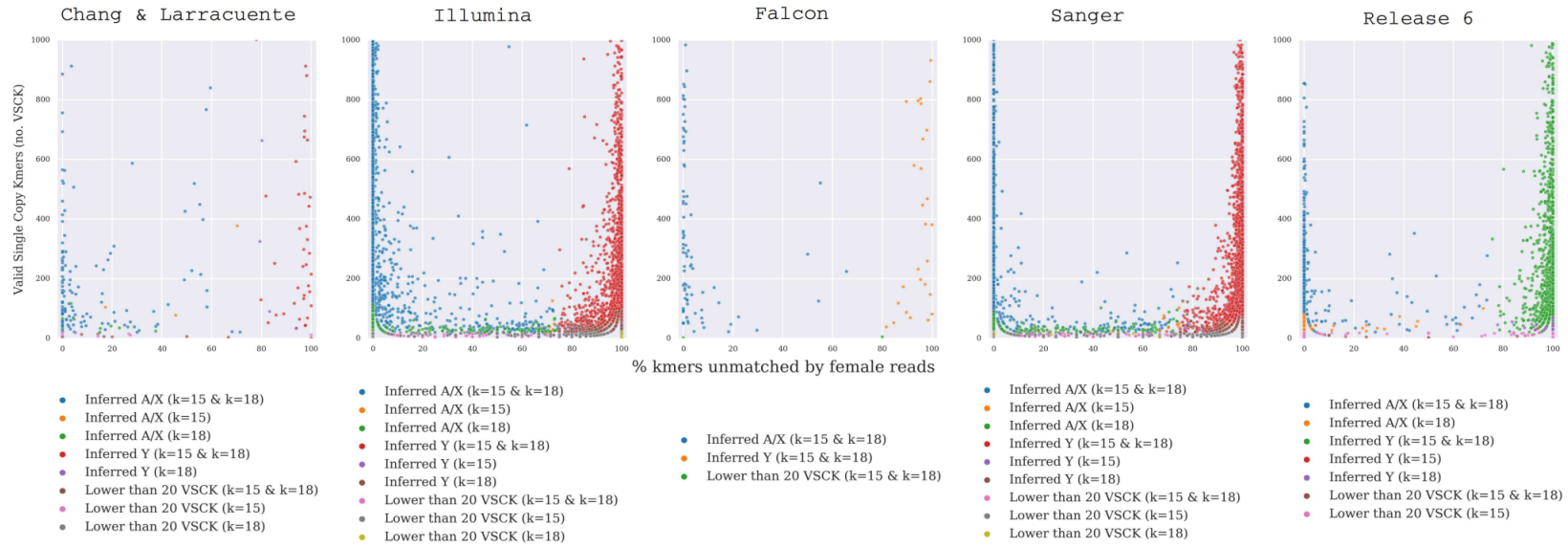


Figura 3.4: Diferentes tamanhos de k aplicados e cromossomos inferidos

Nota da figura: X/A representam as seqüências que não foram inferidas para o cromossomo Y, sendo então assumidas para o cromossomo X ou para os cromossomos autossômicos. Inferências obtidas para o cromossomo Y com ambos valores de k aparecem duplicadas no gráfico, com as coordenadas obtidas a partir de cada uma das análises individuais. A ordenada corresponde ao número de k -mers de cópia única validados (VSCK de 0 à 1000) e a abscissa à porcentagem de k -mers de cópia única validados e sem correspondência com os *reads* de fêmea (PVSCUK - 0 à 100%). Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

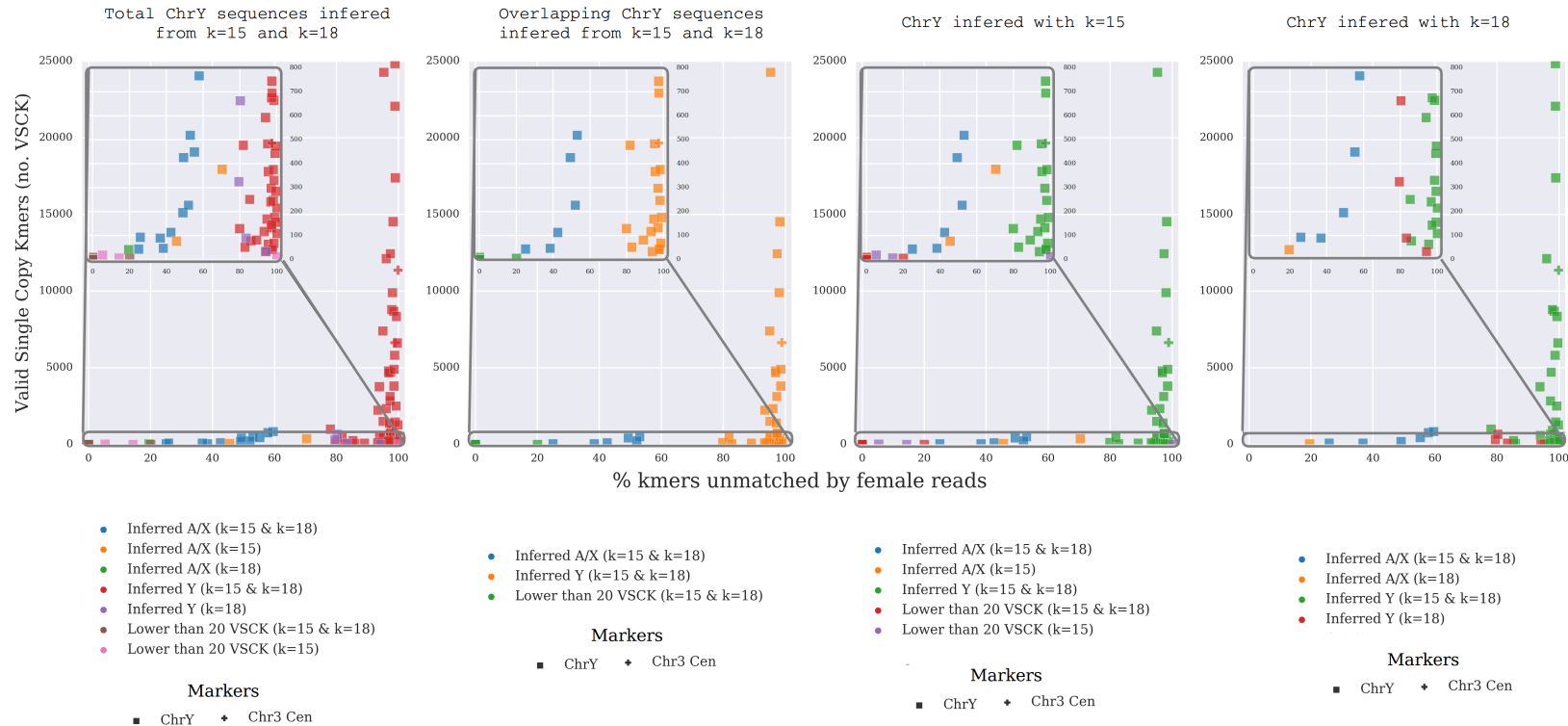


Figura 3.5: Divergências e coincidências entre inferências realizadas a partir do YGS e definições cromossomais de Chang e Larracuent (2019)

Nota da figura: Estão plotadas todas as 55 seqüências definidas pelo trabalho citado como parte do cromossomo Y em cada um dos gráficos, sendo a coloração determinada pelas inferências realizadas a partir das análises do YGS e os marcadores pelas definições dos referidos autores, mostrando divergências e coincidências entre as duas análises. Da esquerda para a direita são apresentadas todas as inferências a partir do YGS usando, respectivamente: $k = 15$ e $k = 18$, com e sem sobreposições de inferências; apenas as sobreposições entre as inferências com $k = 15$ e $k = 18$; apenas as inferências a partir de $k = 15$; e apenas as inferências a partir de $k = 18$. X/A representam as seqüências que não foram inferidas para o cromossomo Y pelo YGS, sendo então assumidas para o cromossomo X ou para os cromossomos autossômicos. Seqüências que apresentam a mesma inferência final tanto com $k = 15$ quanto com $k = 18$ são plotadas apenas com as coordenadas obtidas com a análise que usou $k = 15$, para fins de melhor visualização. A ordenada corresponde ao número de k -mers de cópia única validados (VSCK de 0 à 25000 no gráfico maior e região ampliada vai de 0 à 800) e a abscissa à porcentagem de k -mers de cópia única validados e sem correspondência com os *reads* de fêmea (PVSCUK de 0 à 100%). Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

gens menos fragmentadas - Chang e Larracuent (2019) e Falcon), variando conforme a fragmentação do genoma, em uma máquina com hardware de alto desempenho (Tabela 3.2, detalhes em Capítulo 2, Seção 2.2). Para $k = 18$, o YGS foi executado entre 4 e 5 horas para as montagens menos fragmentadas, atingindo até cerca de 751 horas (1 mês) ao ser executado com a montagem de genoma através de sequenciamento Sanger e chegando a ultrapassar 3 meses de execução com o genoma obtido através de sequenciamento Illumina. O máximo de memória residente alcançado foi de 934 MB para $k = 15$ e sempre pouco mais de 48 GB para $k = 18$, embora esses valores não variem muito entre as análises realizadas com o mesmo valor de k entre os diferentes genomas analisados.

Assembly ($k = 15$)	Time (h)	Mres (MB)	CPUs Kernel (s)	CPUs User (s)
Chang & Larracuent	0.25	856.42	66.07	838.89
Illumina	43.89	760.13	52431.88	105552.97
Falcon	0.25	847.19	86.13	821.76
Sanger	11.77	860.78	13997.09	28373.07
Release 6	0.78	934.24	724.22	2089.30

Assembly ($k = 18$)	Time (h)	Mres (MB)	CPUs Kernel (s)	CPUs User (s)
Chang & Larracuent	3.68	48093.56	4148.84	9114.12
Illumina	2100+	Failed	Failed	Failed
Falcon	4.84	48093.74	5552.75	11869.75
Sanger	751.24	48110.75	940519	1763794
Release 6	36.98	48184.29	44255.16	88872.77

Tabela 3.2: Tempo de execução, uso de memória e processador pelo YGS

Nota da tabela: Informações obtidas através do `/usr/bin/time` para as execuções individuais do `YGS.pl mode=final_mode`, sendo o primeiro bloco da tabela para as execuções com $k = 15$ e o segundo para as execuções com $k = 18$. As células preenchidas com “*Failed*” correspondem às execuções do programa onde houve falha, sendo mostrado o tempo aproximado de processamento até o momento da falha. Descrição das colunas da tabela: *Assembly* (valor de k) – genoma utilizado como entrada e tamanho de k ; *Time* – Tempo em horas; *Mres* – Máximo de memória residente utilizada durante o processo; *CPUs Kernel* – CPU-seconds (em segundos) usado pelo sistema durante o processo pelo modo kernel; *CPUs User* – CPU-seconds (em segundos) diretamente usado pelo processo no modo do usuário.

Como discutido no capítulo anterior (Subseção 2.2.1), ao aumentar o valor de k , o YGS diminui sua eficiência, levando mais tempo e utilizando mais memória para processar as informações. Essa perda de eficiência é ainda mais notável em genomas muito fragmentados, já que cada uma das sequências será igualmente processada com ordem 4^k . Por essa razão, ao tentar executar o YGS utilizando $k = 18$ com o genoma não finalizado obtido a partir de

sequenciamento Illumina, que contém quase 145 mil *contigs*, o programa falhou por falta de memória disponível, apesar da máquina utilizada ser um *cluster* de alto desempenho (Capítulo 2 – Seção 2.2). Essa falha ocorreu após três meses de execução contínua do programa, momento onde cerca de 50 mil *contigs* ainda estavam pendentes de processamento.

Em contrapartida, com a versão mais recente deste mesmo método, YGS2 (DUPIM; ALMEIDA; CARVALHO, versão não publicada), foi possível executar o programa para as cinco diferentes montagens, usando dois valores de *k-mer* e, portanto, duas execuções para cada montagem, em um tempo total de menos de uma semana, tanto na máquina de alto desempenho quanto em uma de uso pessoal (especificações citadas no capítulo anterior (Seção 2.2). Esse tempo total de execução incluiu também as filtrações de qualidade e frequência realizadas com o Jellyfish (MARÇAIS; KINGSFORD, 2011).

Para as montagens e escolhas de *k* em que a primeira versão do YGS (CARVALHO; CLARK, 2013) foi executada sem falhas, foram comparados seus resultados com os obtidos na execução do YGS2 (DUPIM; ALMEIDA; CARVALHO, versão não publicada) (Figura 3.6). Como ambas as versões, independente do valor de *k-mer* escolhido, geram resultados idênticos, as inferências realizadas a partir do genoma montado adivindo de sequenciamento Illumina e *k* = 18 apresentadas nesse capítulo foram as obtidas a partir do YGS2.

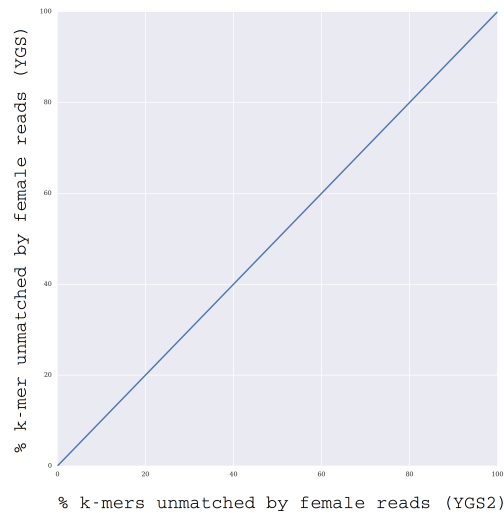


Figura 3.6: Comparação entre resultados do YGS e do YGS2

Nota da figura: Para todas as montagens e escolhas de *k* em que a primeira versão do YGS (CARVALHO; CLARK, 2013) foi executada sem falhas (Chang & Larracunte (*k* = 15 e *k* = 18), Falcon (*k* = 15 e *k* = 18), *Release 6* (*k* = 15 e *k* = 18), Sanger (*k* = 15 e *k* = 18) e Illumina (*k* = 15)) o mesmo gráfico foi obtido, com total correspondência dos resultados entre as análises do YGS e YGS2. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

Para executar o YGS2 também é realizada a triagem de qualidade descrita no capítulo anterior (Subseção 2.2.1). Entretanto, não é necessária a transformação das informações para *bit-arrays*, permitindo que a remoção de *k-mers* que são muito raros em comparação com a cobertura do genoma possa ser realizada diretamente através do jellyfish dump (MARÇAIS; KINGSFORD, 2011), como exposto no código a seguir. A opção *-L* corresponde à opção *lower-count* do programa em Perl desenvolvido por Carvalho e Clark (2013) que chama o jellyfish dump (MARÇAIS; KINGSFORD, 2011) internamente. Para a tabela *hash* com as frequências de *k-mers* geradas a partir do genoma de referência utilizado, são selecionados apenas os *k-mers* de cópia única com o parâmetro *-U 1*. Ou seja, qualquer *k-mer* com frequência acima de 1, portanto *k-mer* repetitivo, será removido. Essa alteração se relaciona apenas à uma lógica revertida das execuções internas de cada versão do programa.

```
1 ## Remover contagens baixas com jellyfish dump
2 ## a partir de arquivos gerado por jellyfish count (etapa I)
3 # -c --out-counter-len Reduz tamanho do arquivo de saída
4 # -L --lower-count Nao imprime k-mer menor que -L
5
6 jellyfish dump -c -L $CountFemale FemaleReads.jelly > FemaleReads.dump
7 jellyfish dump -c -L $CountMale MaleReads.jelly > MaleReads.dump
8 jellyfish dump -c -U 1 Genome.jelly > Genome.dump
```

A execução do YGS2 em si, exemplificada a seguir, também requer (i) o tamanho de *k*; (ii) os arquivos de entrada – os gerados pelo jellyfish dump e o genoma de referência; (iii) o nome do arquivo de saída. O arquivo gerado terá o mesmo formato do arquivo produzido pela primeira versão do YGS (CARVALHO; CLARK, 2013), conforme ilustrado na Figura 2.4. Nesta versão, houve um ganho significativo, diminuindo consideravelmente o consumo de memória RAM para o processamento de um mesmo arquivo, em relação à primeira versão do YGS.

```
1 ## Executar YGS2 com k sendo k-mer size
2 perl YGS2.p $KmerSize MaleReads.dump FemaleReads.dump Genome.dump ↔
   Genome.fasta > out
```

3.2 Análise de Contaminantes

A validação de *k-mers* realizada pelo método YGS usando os *short-reads* do DNA do macho serve para remover contaminantes que poderiam estar presentes no genoma montado. Entretanto, contaminações nos *short-reads* não são consideradas pelo programa, tanto nos de macho, usados para a validação no próprio software, quanto nos de fêmea, com papel essencial da determinação de *contigs* para o cromossomo Y. Essas contaminações podem, potencialmente, ser exclusivas entre os *bit-arrays* gerados e, assim, acabar entre o conjunto de sequências inferidas para o cromossomo Y. Levando esses aspectos em consideração, para cada um dos arquivos com as sequências inferidas para o cromossomo Y gerado a partir dos testes descritos na seção anterior (Seção 3.1), foram realizadas análises de contaminantes através do BlobTools (LAETSCH; BLAXTER, 2017; LAETSCH *et al.*, 2017) seguindo as orientações descritas no capítulo anterior (Subseção 2.2.2, Figura 2.5).

Para esta pesquisa, as sequências inferidas para o cromossomo Y, que são as sequências de interesse, foram submetidas como *query* para alinhamento contra banco de dados de proteínas e de nucleotídeos a partir do BLAST/Diamond (ALTSCHUL *et al.*, 1990; BUCHFINK; XIE; HUSON, 2014). Os comandos abaixo apresentam todos os parâmetros utilizados para executar essa etapa. Em seguida, os resultados dos alinhamentos gerados foram processados com os comandos em AWK (AHO; KERNIGHAN; WEINBERGER, 1978) descritos no Capítulo 2 (Subseção 2.2.2), para adicionar as informações de táxons de acordo com contaminantes humanos e de outras espécies – aqui utilizamos um banco de dados montado pela equipe do Professor Dr. Antonio Bernardo de Carvalho (VANDERLINDE *et al.*, 2018) (oakdb corresponde à *Banco de dados Carvalho*).

```
1 ## I - Taxonomia e Alinhamento
2 diamond blastx --query chrY.fasta --db oakdb_prot_01ago19.dmnd <↔>
    --block-size 2 --max-target-seqs 250 --sensitive --index-chunks 1 <↔>
    --evaluate 1e-25 --threads 64 --outfmt 6 qseqid staxids bitscore <↔>
    sseqid sstart send qstart qend pident --out blastx_chrY.hitsfile > <↔>
    blastx_chrY.out
3 blastn -db oakdb_genomic_01ago19 -query chrY -outfmt '6 qseqid staxids <↔>
    bitscore sseqid sstart send qstart qend pident' -evaluate 1e-10 <↔>
    -num_threads 64 > blastn_chrY.hitsfile
```

As informações obtidas a partir da execução do BlobTools com os resultados dos alinhamentos contra o banco de dados de proteínas e de nucleotídeos foram então unidas,

permitindo remover em seguida todos os *contigs* que fossem de um filo diferente de Arthropoda e/ou que não fossem de Eukaryota. Além disso, foram também mantidas aquelas sequências que não apresentaram nenhum alinhamento (*no-hit*) contra o banco de dados fornecido. Todas essas etapas aparecem no fluxograma anteriormente apresentado para o BlobTools (Figura 2.5). Abaixo, são apresentados os comandos executados para unir os resultados finais do BlobTools e identificar as sequências contaminadas.

```

1 ## Unir resultados do BlobTools
2 grep -v "^#" superkingdom.blastx_blob_chrY_k3.blobDB.table.txt | cut -f6- > temporary_SuperKingBLASTxColumns.txt
3 grep -v "^#" superkingdom.blastn_blob_chrY_k3.blobDB.table.txt | cut -f6- > temporary_SuperKingBLASTnColumns.txt
4 grep -v "^#" blastn_blob_chrY_k3.blobDB.table.txt | cut -f6- > temporary_PhylumBLASTnColumns.txt
5 grep -v "^#" blastx_blob_chrY_k3.blobDB.table.txt > temporary_PhylumBLASTxColumns.txt
6 paste temporary_PhylumBLASTxColumns.txt temporary_SuperKingBLASTxColumns.txt temporary_PhylumBLASTnColumns.txt temporary_SuperKingBLASTnColumns.txt > BlobtoolsFinalTable_chrY_blastx_blastn_k3.txt
7
8 ## Identificar contaminacoes
9 cut -f6,9,12,15 BlobtoolsFinalTable_chrY_blastx_blastn_k3.txt | tr '\t' '\n' | sort | uniq | grep -v -e "$phylum" -v -e "$superkingdom" -v -e "no-hit" | while read contam ; do grep "$contam" BlobtoolsFinalTable_chrY_blastx_blastn_k3.txt ; done | sort | uniq | cut -f1 | cut -d"|" -f2 > BlobtoolsResults_chrY_contaminatedGIs

```

Foram realizadas análises de contaminantes para cada um dos conjuntos de sequências inferidas para o cromossomo Y (Seção 3.1) e para os *contigs* que Chang e Larracunte (2019) originalmente determinaram como parte do cromossomo Y de *Drosophila melanogaster*. Para a maioria das sequências nenhum contaminante foi detectado (Tabela 3.3) e, aquelas que apresentaram um alinhamento contra um táxon diferente de Arthropoda e/ou que não fossem de Eukaryota, foram removidas do respectivo arquivo .fasta contendo os *contigs* desse cromossomo. Desta maneira, é seguro afirmar que as sequências inferidas através das análises descritas nesse capítulo não são uma mera sequela de contaminações presentes nos arquivos

inicialmente submetidos (genoma montado, *short-reads* de macho e de fêmea), removendo a possibilidade de desenhar sondas *oligopaint* que não iriam hibridizar na amostra fixada de interesse, seja no cromossomo Y ou fora da sequência alvo. Assim, as análises posteriormente apresentadas (Capítulos 4 e 5) sobre a densidade de sondas/kb gerada e sintetizada e os resultados dos experimentos de validação não são enviesadas por uma quantidade de sondas *oligopaint* que efetivamente não poderiam marcar o cromossomo Y.

ChrY (Original)	Contigs	Mb	Pure Contigs	Pure Mb	Contaminations
Chang & Larracuent	55	14.60	54	14.55	1

ChrY ($k = 15$)	Contigs	Mb	Pure Contigs	Pure Mb	Contaminations
Chang & Larracuent	40	13.51	40	13.51	0
Illumina	1897	1.16	1783	1.08	114
Falcon	66	8.81	66	8.81	0
Sanger	2659	5.51	2658	5.51	1
Release 6	769	6.57	769	6.57	0

ChrY ($k = 18$)	Contigs	Mb	Pure Contigs	Pure Mb	Contaminations
Chang & Larracuent	44	13.79	44	13.79	0
Illumina	2964	1.30	2839	1.23	125
Falcon	66	8.81	66	8.81	0
Sanger	3449	6.19	3440	6.18	9
Release 6	821	6.65	821	6.65	0

Tabela 3.3: Análise de contaminantes a partir de diferentes conjuntos de sequências inferidas para o cromossomo Y

Nota da tabela: Números de *contigs* e pares de bases em Mb são fornecidos para as sequências que foram inferidas para o cromossomo Y através do YGS/YGS2 com ambos os valores de k utilizados e para o conjunto de sequências originalmente montadas para o cromossomo Y por Chang e Larracuent (2019), seguido do número de *contigs* não contaminados e seu tamanho em Mb e do número de contaminações encontradas para cada um dos conjuntos de sequências.

Foram detectadas contaminações entre os conjuntos de sequências inferidas para o cromossomo Y a partir das montagens de Illumina e Sanger e para o cromossomo Y originalmente montado por Chang e Larracuent (2019). Para o Y dos autores citados, apenas através de alinhamento com BLASTx/Diamond (ALTSCHUL *et al.*, 1990; BUCHFINK; XIE; HUSON, 2014) foi detectada uma sequência que não era pertencente ao táxon Arthropoda, mas sim Basidiomycota. Para o conjunto de sequências inferidas para cromossomo Y utilizando $k = 15$ a partir da montagem Sanger ambas as análises de BLASTn e BLASTx/Diamond mostraram uma contaminação do filo Chordata. Para o conjunto de sequências inferidas para cromossomo Y utilizando $k = 18$ a partir dessa mesma montagem foram detectadas

9 contaminações advindas dos filos Chrysiogenetes, Chordata e Proteobacteria (BLASTn) e Chordata, Tenericutes e Proteobacteria (BLASTx/Diamond). Vale ressaltar que esses *contigs* contaminados, em todos os três casos mencionados, não chegaram a totalizar 1 Mb de sequência nucleotídica.

Para o conjunto de sequências inferidas para cromossomo Y utilizando $k = 15$ a partir da montagem Illumina foram detectadas 114 sequências contaminadas (Totalizando menos de 1 Mb de sequência nucleotídica) – através do alinhamento com BLASTn, o BlobTools atrelou algumas dessas sequências aos táxons de Ascomycota, Spirochaetes, Firmicutes, Proteobacteria, Eukaryota-undef ou Chordata; já para alinhamento contra banco de dados de proteínas, foram detectados os filos Ascomycota, Basidiomycota, Firmicutes, Viruses-undef e Chordata. Para o conjunto de sequências inferidas para cromossomo Y a partir dessa mesma montagem, mas com $k = 18$, foram detectadas 125 contaminações, também somando menos de 1 Mb de sequência nucleotídica contaminada, advindas dos filos Ascomycota, Spirochaetes, Firmicutes, Proteobacteria, Eukaryota-undef e Chordata para análise com BLASTn e Ascomycota, Basidiomycota, Viruses-undef, Firmicutes e Chordata para análises com BLASTx/Diamond. Para fins de visualização, é apresentada apenas a figura resultante da análise de BLASTx/Diamond feita com o conjunto de sequências inferidas para o cromossomo Y utilizando $k = 18$ a partir da montagem Illumina, que conta com o maior número de contaminações e, assim, permite facilmente contrastar as sequências de Arthropoda contra as alinhadas para outros táxons (Figura 3.7).

Através das análises realizadas, o maior conteúdo nucleotídico inferido para o cromossomo Y foi obtido usando $k = 18$ a partir da montagem realizada por Chang e Larracuent (2019) – 13.79 Mb, que muito se aproxima dos originais 14.60 Mb (14.55 Mb após remoção de contaminantes) obtidos através dessa mesma montagem para o cromossomo Y pelos autores citados. Como discutido anteriormente, a abordagem utilizada por esses autores foi extremamente minuciosa, contando com uma série de curagens manuais para enriquecer as sequências conhecidas para o cromossomo Y e também com o sequenciamento de molécula única e *long-reads* da Pacific Biosciences. O resultado mais próximo a esse conteúdo nucleotídico foi alcançado a partir de um genoma montado com Falcon, chegando a 8.81 Mb de sequências inferidas para o cromossomo Y. Além disso, ao aumentar o valor de k de 15 para 18 não houve muita variação entre as sequências inferidas (Tabela 3.1) e o mesmo foi observado após a análise de contaminantes (Tabela 3.3). Esses resultados corroboram o método YGS (CARVALHO; CLARK, 2013; DUPIM; ALMEIDA; CARVALHO, versão não publicada) como uma escolha sensata para a inferência de sequências para o cromossomo Y, realçando a reprodução facilitada da *pipeline* proposta nessa pesquisa.

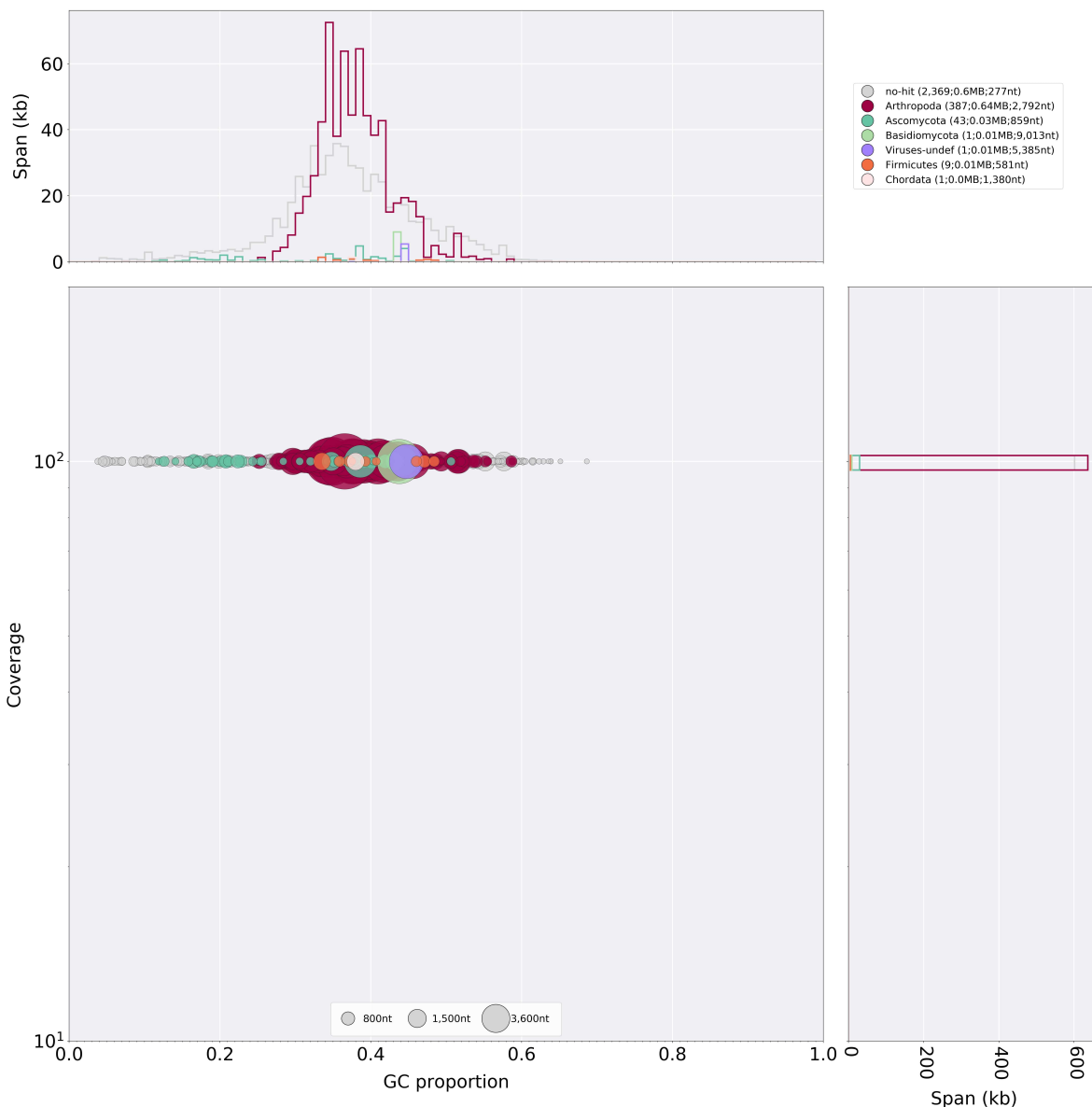


Figura 3.7: Análise do Blobtools realizada com o conjunto de sequências inferidas para o cromossomo Y a partir da montagem Illumina e $k = 18$

Nota da figura: Essa análise com BLASTx/Diamond mostrou que as contaminações encontradas entre os *contigs* inferidos para o cromossomo Y representavam sequências pertencentes aos táxons Ascomycota, Basidiomycota, Viruses-undef, Firmicutes e Chordata. Todas as sequências ditas puras também aparecem na figura como parte do filo Arthropoda. Essa é uma figura gerada pelo próprio BlobTools (blobtools plot) (LAETSCH; BLAXTER, 2017; LAETSCH *et al.*, 2017), e mostra na ordenada a cobertura de sequenciamento e na abscissa a proporção de bases CG – a proporção da circunferência vai de 800 nt, passando por 1500 nt, até 3600 nt, variando justamente com a proporção de nucleotídeos contaminados. Na legenda da figura no canto superior direito, lê-se: no-hit, Arthropoda, Ascomycota, Basidiomycota, Viruses-undef, Firmicutes, Chordata.

4

Desenhando sondas *oligopaint* para o cromossomo Y

Uma das variações da técnica de FISH usa sondas *oligopaint*. Desenhadas a partir de sequências genômicas, essas sondas constituem uma biblioteca sintética de oligonucleotídeos de DNA de filamento simples com polaridade 5'-3' (BELIVEAU; APOSTOLOPOULOS; WU, 2014; BELIVEAU *et al.*, 2015). Esse caráter sintético das sondas *oligopaint* permite que uma série de parâmetros sejam considerados para o desenho das mesmas (BELIVEAU *et al.*, 2018). Alguns desses parâmetros estão diretamente envolvidos com o experimento de FISH que será realizado com as sondas desenhadas: por exemplo, a escolha da temperatura de hibridização, que deve ser a mesma tanto para a seleção de sondas, quanto para que, *in situ*, as sondas sintetizadas hibridizem na amostra fixada. Além disso, a quantidade de sondas desenhadas, assim como a densidade de sondas por área cromossômica, costumam ser muito superiores aos experimentos comuns de FISH, assegurando que a marcação aconteça em uma região muito mais extensa. Uma vez que todas as sondas são marcadas com o mesmo fluoróforo, a extensão dessa marcação permite a visualização do cromossomo em sua totalidade (BELIVEAU *et al.*, 2015; BELIVEAU *et al.*, 2018).

Sondas desenhadas a partir de sequências repetitivas podem hibridizar fora da região alvo e, por esse motivo, sequências ricas em repetições costumam ser mascaradas para minimizar o viés de marcação indesejada (BELIVEAU *et al.*, 2018).

Um dos grandes desafios na implementação da *pipeline* desenvolvida é justamente desenhar sondas para cromossomos Y (ou W), em especial se for considerado que, além de apresentarem uma natureza altamente repetitiva, também há dificuldades relacionadas à montagem de suas sequências. Por esse motivo, usamos as sequências (sem serem mascaradas) inferidas a partir do método YGS (Capítulo 3) como base para o desenho de sondas *oligopaint* através do OligoMiner – um método desenvolvido especialmente para encontrar e selecionar sondas *oligopaint* (BELIVEAU *et al.*, 2018).

A escolha desse método em detrimento de outros está relacionada à facilidade de sua

utilização e também à autonomia concedida ao usuário na escolha dos parâmetros envolvidos no desenho e seleção de sondas. Existem outras alternativas, como o programa OligoArray (ROUILLARD; ZUKER; GULARI, 2003) ou mesmo através da simples fragmentação da região de interesse em sequências do tamanho desejado para as sondas, abordagem realizada por Albert *et al.* (2019). No entanto, essas alternativas ou não permitem a flexibilização dos parâmetros envolvidos (ROUILLARD; ZUKER; GULARI, 2003) ou exigem muito mais habilidades do usuário para aplicação das mesmas (ALBERT *et al.*, 2019), inclusive sendo mais restrita a janela de parâmetros que podem ser manipulados: geralmente incluindo apenas o tamanho das sondas, similaridade a outras regiões do genoma e cálculo da temperatura de hibridização. Mesmo a publicação mais recente do OligoMiner, o OligoMinerApp, desenvolvido por Passaro *et al.* (2020), não garante ao usuário a mesma flexibilidade na escolha de parâmetros – visto que, para essa interface gráfica ainda não foram implementadas todas as funcionalidades da publicação original.

Na próxima seção são exploradas as abordagens clássicas do OligoMiner, como demonstrado por Beliveau *et al.* (2018), esclarecendo os parâmetros aqui utilizados ao longo de todo o processo e os resultados encontrados (Seção 4.1). Na segunda seção deste capítulo é revelada uma nova abordagem para desenho de sondas a partir de regiões tão repetitivas quanto as de cromossomos Y. Essa abordagem é executada a partir de uma alternativa encontrada no próprio OligoMiner, contando também com filtragens adicionais – todos esses aspectos, assim como os resultados alcançados, são melhor esclarecidos na Seção 4.2. Subsequentemente, são comparados os resultados obtidos a partir das duas abordagens, incluindo também discussões acerca das densidades de alvos/kb obtidas para o cromossomo Y (Seção 4.3).

4.1 Explorando as abordagens clássicas do OligoMiner

Aqui é considerada uma abordagem clássica do OligoMiner a utilizada e discutida por Beliveau *et al.* (2018) (Figura 4.1). A abordagem clássica inclui: o desenho de sondas através do UM – o qual mantém apenas as sondas com um único alinhamento no genoma de referência – e o desenho de sondas através do LDM – modo que usa os resultados de alinhamento para selecionar as sondas que apresentarem apenas um alvo termodinamicamente predito de acordo com modelo de análise linear discriminante. Além destas duas etapas principais, foram também executadas filtragens adicionais oferecidas pelo OligoMiner: processamento das sondas para remoção daquelas que apresentassem *k-mers* abundantes no genoma ou formação de estrutura secundária. Foram seguidas as recomendações esclarecidas por Beliveau *et al.* (2018) e apresentadas anteriormente (Capítulo 2, Subseção 2.2.3).

Alguns dos critérios para seleção de sondas candidatas e filtragens realizadas foram

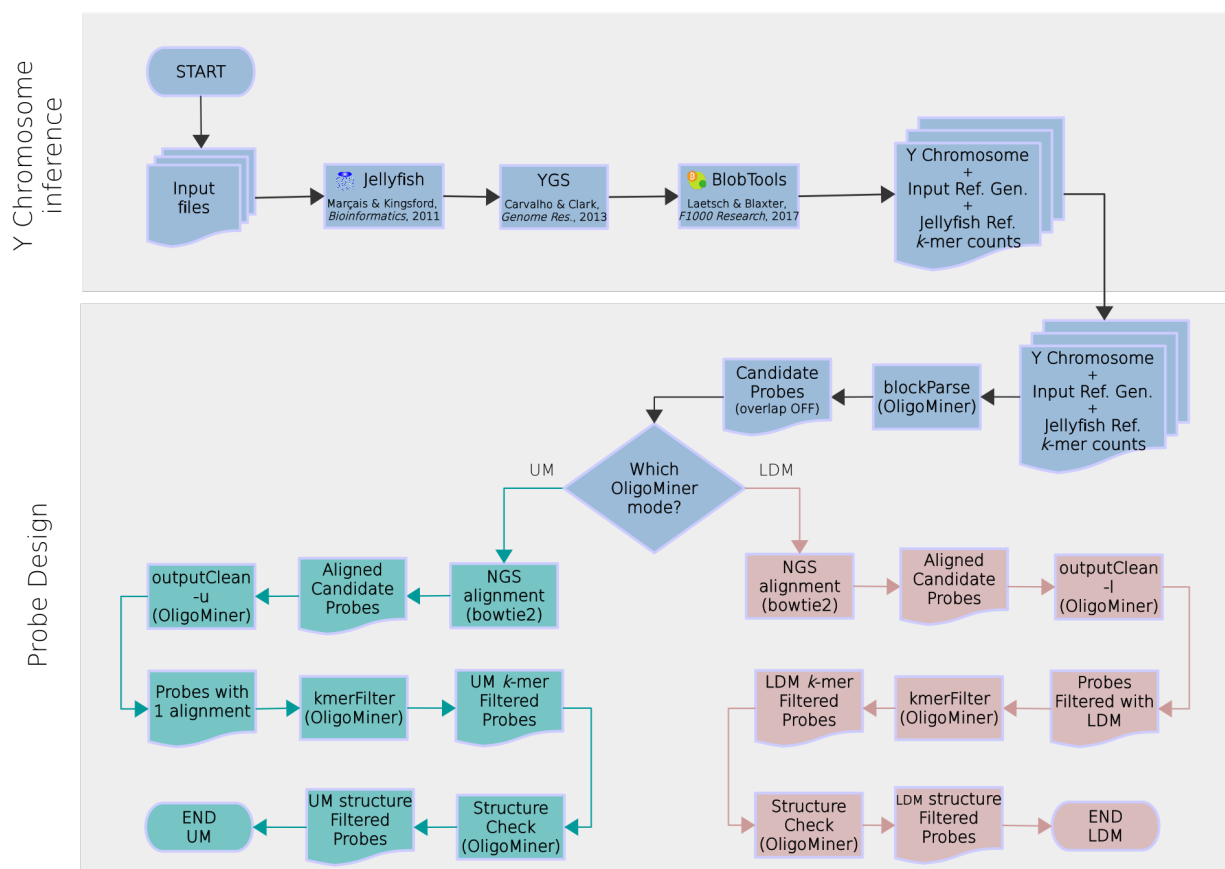


Figura 4.1: Fluxograma da abordagem clássica do OligoMiner a partir dos resultados de inferência para o cromossomo Y

Nota da figura: Através do `blockParse.py` são desenhadas sondas candidatas a partir dos conjuntos de sequências inferidas para o cromossomo Y com as análises do YGS e BlobTools; em seguida as mesmas são alinhadas contra o genoma de referência com `bowtie2`; o resultado do alinhamento é processado pelo programa `outputClean.py` (varia conforme o modo escolhido – UM, em esverdeado, ou LDM, em rosado); por fim são removidos *k-mers* abundantes no genoma de referência com `kmerFilter.py` e sondas com probabilidade de formar estruturas secundárias com `structureCheck.py`. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

modificados, de modo que três conjuntos de parâmetros foram utilizados para o desenho de sondas através das abordagens clássicas do OligoMiner: um conjunto com parâmetros mais restritivos, outro com parâmetros mais permissivos e um terceiro com todos os critérios padronizados pelo próprio OligoMiner. Esses conjuntos de parâmetros são referidos neste trabalho como *stringent*, *coverage* e *balance*, respectivamente. Seus valores foram apresentados anteriormente (Tabela 2.3) e dizem respeito ao comprimento das sondas, temperaturas de desnaturação e hibridização permitidas, tamanho do *k-mer* para filtragem de *k-mers* abundantes e seu respectivo *threshold*. A determinação dos valores para *stringent* e *balance* foi feita por Beliveau *et al.* (2018) e, em especial para os parâmetros *coverage*, mais permissivos,

foi alterado o comprimento das sondas sugerido por esses autores de 26 – 32 nt para 30 – 35 nt.

Dessa forma, cada conjunto de sequências obtido para o cromossomo Y através das análises de inferência de sequências e remoção de contaminantes (Capítulo 3, Tabela 3.3) foi utilizado para desenho de sondas *oligopaint*. Considerando que a inferência de *contigs* para o cromossomo Y a partir de dois valores de k resultou em grande parte dessas sequências serem obtidas por ambas as análises (Tabela 3.1 e suas discussões), para cada montagem, o desenho das sondas foi feito a partir dos diferentes subgrupos de sequências em comum entre $k = 15$ e $k = 18$ ou exclusivas de cada uma dessas análises após remoção de contaminantes (Tabela 4.1 e Figura 4.2).

Assembly	Common		$k = 15$		$k = 18$	
	contigs	Mb	contigs	Mb	contigs	Mb
Chang & Larracuent	40	13.51	None	None	4	0.28
Illumina	1778	1.08	5	0.001	1061	0.14
Falcon	66	8.80	None	None	None	None
Sanger	2653	5.50	5	0.004	787	0.67
Release 6	768	6.57	1	0.001	53	0.08

Tabela 4.1: Subgrupos de sequências inferidas para o cromossomo Y

Nota da tabela: Números de *contigs* e pares de bases em Mb são fornecidos para os subgrupos de cromossomos Y utilizados como arquivo de entrada para execução do blockParse.py, sendo: *Common* – sequências obtidas em comum entre $k = 15$ e $k = 18$ para cada montagem; $k = 15$ – sequências exclusivamente obtidas com $k = 15$; $k = 18$ – sequências exclusivamente obtidas com $k = 18$. Células que aparecem com *None* indicam um subgrupo em que não existia nenhuma sequência exclusiva e, assim, nenhum conteúdo de bases nucleotídicas. Em especial para Chang e Larracuent (2019), um subgrupo com os 16 *contigs* exclusivos desses autores em relação aos 40 em comum às análises do YGS também foi utilizado, tendo este um total de 1.15 Mb.

Em especial para o cromossomo Y montado por Chang e Larracuent (2019), como alguns de seus *contigs* originais não foram inferidos pelo YGS com nenhum dos dois valores de k utilizados, outro subgrupo foi utilizado para desenho das sondas candidatas contando com os *contigs* originais exclusivos quando comparado aos *contigs* que tanto as análises com $k = 15$ quanto $k = 18$ recuperaram. Nesse caso, considerando que a sequência contaminada encontrada também foi removida (Tabela 3.3) e que o subgrupo das sequências em comum entre $k = 15$ e $k = 18$ conta com dois *contigs* que esses autores não determinaram originalmente para o cromossomo Y (Tabela 4.1), foram utilizados 16 *contigs* com total de 1.15 Mb.

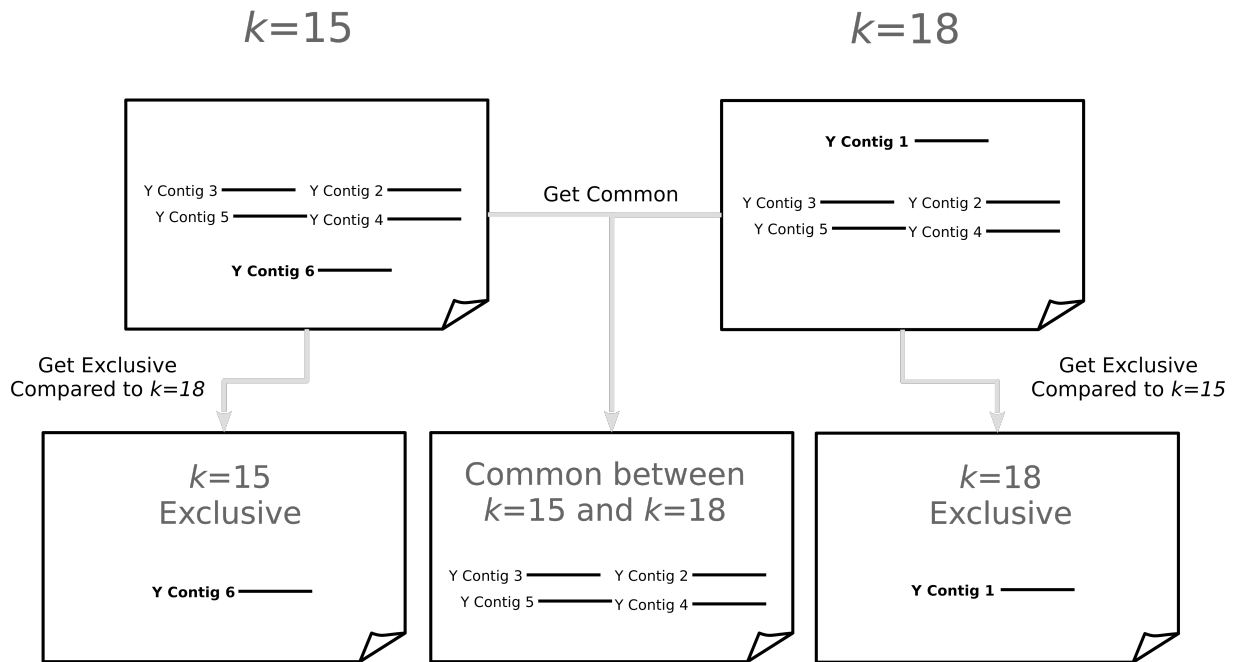


Figura 4.2: Esquema de subgrupos de sequências inferidas para o cromossomo Y

Nota da figura: As sondas candidatas foram desenhadas pelo programa `blockParse.py` a partir dos subgrupos gerados conforme esquematizado e apontado na Tabela 4.1.

A seleção das sequências de cada um desses subgrupos foi dada a partir dos comandos em AWK (AHO; KERNIGHAN; WEINBERGER, 1978) descritos a seguir.

```

1 ## Selecionar contigs em comum entre a.fasta e b.fasta
2 awk 'NR==FNR{seen[$0]=1; next} seen[$0]' a.fasta b.fasta > ab_common.fasta
3
4 ## Selecionar contigs exclusivos de a.fasta comparado a b.fasta
5 awk 'FNR==NR{seen[$0]++; next} !($0 in seen)' b.fasta a.fasta > (↔)
   a_exclusive.fasta

```

As sondas foram desenhadas dessa maneira uma vez que aquelas obtidas a partir de sequências inferidas com dois valores de k têm maior confiabilidade. A discriminação dos resultados para esses subgrupos de *contigs* ao longo de cada etapa do desenho de sondas é apresentada apenas para aqueles encontrados em comum às diferentes análises do YGS. Para os subgrupos de *contigs* exclusivos são apresentados apenas os valores finais de sondas desenhadas após aplicação de todos os filtros, tanto individualmente quanto somado às sondas comuns, totalizando as sondas desenhadas para cada conjunto individual de sequências inferidas para o cromossomo Y quando subgrupos não são considerados. Por exemplo, para o conjunto

de sequências inferidas para o cromossomo Y a partir da montagem com sequenciamento Illumina são apresentados os valores de sondas obtidas ao longo de todo o processo apenas para os *contigs* que foram obtidos em comum entre as análises do YGS com $k = 15$ e $k = 18$ e, para os *contigs* exclusivos de um desses valores de k são apresentados apenas os resultados finais individuais e o valor total quando somados às sondas desenhadas em comum para ambas as sequências.

Como o programa `blockParse.py` aceita apenas arquivos com uma única entrada FASTA, cada subgrupo contendo os *contigs* de interesse (comuns, exclusivos de $k = 15$, exclusivos de $k = 18$ e exclusivos do cromossomo montado por Chang e Larracuent (2019)) foi, individualmente, unido em uma única sequência FASTA com blocos de 200 caracteres N separando-os. Isso foi realizado para evitar que sondas fossem desenhadas entre a região final de um *contig* e a inicial de outro e, simultaneamente, permitir que pudessem ser obtidas, com apenas uma execução, todas as sondas candidatas para a sequência de interesse. Para tanto, foram executados comandos em SED e tr (MCMAHON, 1978; MCILROY, 1987), como os descritos abaixo, onde o bloco “NNNNN” era substituído por um bloco de 200 N.

```

1 sed -i '/^>/c\NNNNN' multipleContigs_chrY.fasta
2 tr -d '\n' < multipleContigs_chrY.fasta | sed '$s/ $/\n/' > <=)
   singleEntry_chrY.fasta
3 sed -i "1 i>singleEntry_ChrY" singleEntry_chrY.fasta
4 echo >> singleEntry_chrY.fasta

```

A relação de sondas candidatas desenhadas para os *contigs* comuns está disponível na Tabela 4.2. Já que essas sondas representam todas as sequências presentes no arquivo de interesse que satisfazem os parâmetros pré-determinados, não é necessário executar o `blockParse.py` mais de uma vez para filtrá-las a partir dos diferentes modos do `outputClean.py`, uma vez que essa filtragem é posterior e depende somente dos resultados do alinhamento das sondas candidatas. Um importante adendo diz respeito à sobreposição entre as sondas candidatas encontradas – que não foi permitida nesse caso – ou seja, foi utilizada a opção padronizada do `blockParse.py` que só procura por uma nova sonda a partir da posição *1 + posição final da sonda candidata anterior*.

Em seguida foi realizado o alinhamento das sondas candidatas segundo as orientações descritas no capítulo anterior (Subseção 2.2.3) para poderem ser filtradas a partir dos modos UM e LDM. Os genomas de referência utilizados variavam exatamente conforme a origem do conjunto de sequências inferidas para o cromossomo Y e, além disso, todas as sequências contaminadas encontradas (Tabela 3.3) foram também removidas das sequências genômicas para impedir que sondas fossem descartadas desnecessariamente. Assim, por exemplo, para o

Assembly (Y chr seq)	Stringent Candidates	Balance Candidates	Coverage Candidates
Chang & Larracunte	113217	189995	270729
Illumina	7713	14320	20893
Falcon	77554	128988	179073
Sanger	37630	63405	88759
Release 6	51454	86345	120667

Tabela 4.2: Quantidade de sondas candidatas não sobrepostas desenhadas através do blockParse.py

Nota da tabela: Para cada subgrupo de sequências encontradas em comum a partir da mesma montagem genômica (*Assembly – Y chromosome sequences*) e com ambas as análises do YGS com $k = 15$ e $k = 18$ são apresentadas as quantidades de sondas candidatas desenhadas a partir de três diferentes conjuntos de parâmetros: *stringent* – mais restritivos; *balance* – valores padrão do OligoMiner; *coverage* – parâmetros mais permissivos. Os valores desses parâmetros foram apresentados na Tabela 2.3.

conjunto de sequências inferidas para o cromossomo Y a partir da montagem Illumina, esse genoma de referência sem os contaminantes detectados foi a utilizado. Posteriormente à seleção de sondas com UM ou LDM, as mesmas foram filtradas com os programas kmerFilter.py e structureCheck.py, nessa ordem.

Em sua totalidade, foram obtidos 96 conjuntos de sondas a partir da abordagem clássica do OligoMiner (Figura 4.3), considerando os subgrupos de sequências do cromossomo Y fornecidas como entrada para o OligoMiner (Figura 4.2). Para cada arquivo de entrada, as sondas candidatas foram filtradas primariamente através do UM e LDM (análises individuais), além de filtragens adicionais anteriormente explicadas (Figura 4.1). Dessa forma, a partir de cada montagem genômica, foram desenhados 18 conjuntos de sondas – metade a partir do UM e a outra metade a partir do LDM. Em especial para as sequências do cromossomo Y montado por Chang e Larracunte (2019), o subgrupo de *contigs* originais exclusivos também foi utilizado para desenhar sondas, gerando 6 conjuntos de sondas finais a mais a partir dessa montagem (totalizando 24 conjuntos de sondas).

Para as sondas filtradas com o UM (Tabela 4.3), independente da origem das sequências do cromossomo Y utilizadas, foram retornadas menos de 12.5% das sondas candidatas ao final das filtragens. O conjunto de sequências do cromossomo Y obtidas a partir de sequenciamento Illumina, o menos repetitivo entre os diferentes conjuntos de sequências do cromossomos Y utilizados, apesar de apresentar os menores valores de sondas candidatas, mantém de 33% a 43% das mesmas após filtragem inicial com outputClean.py, se mantendo nessa faixa de 30–40% após execução das etapas seguintes. Para os outros conjuntos de sequências inferidos para o cromossomos Y e que foram utilizados, são mantidas menos de 10% das sondas candidatas entre as finais e a maior parte das candidatas são removidas justamente

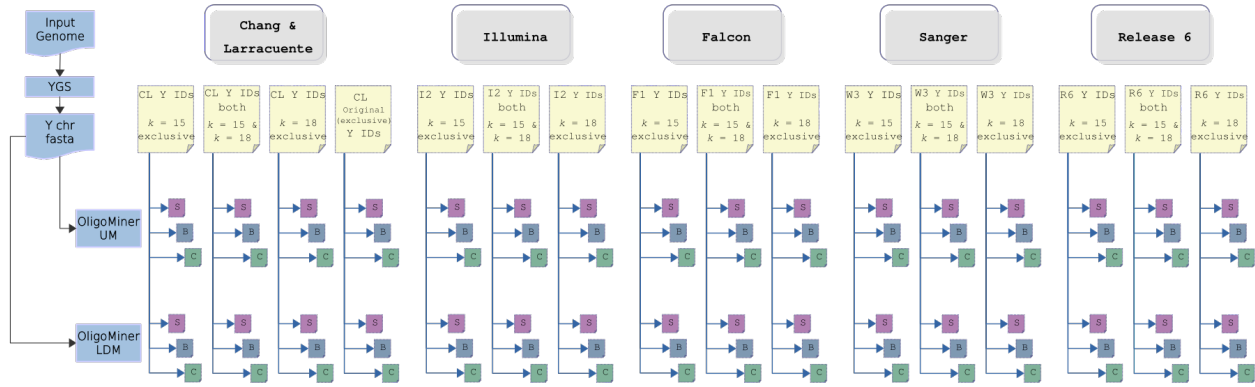


Figura 4.3: Conjuntos de sondas criados através da abordagem clássica

Nota da figura: A partir dos arquivos de sondas gerados com YGS e processados conforme exemplificado na Figura 4.2 (Detalhes na Tabela 4.1), as sondas candidatas foram desenhadas e filtradas pelo OligoMiner a partir de três conjuntos de parâmetros: s – stringent; b – balance; c – coverage; conforme detalhado na Tabela 2.3.

na predição de alvos através do alinhamento. Isso indica que grande parte dessas sondas são filtradas devido à alta repetitividade das sequências do cromossomo Y, já que muitas das vezes essas sequências também aparecem em outros cromossomos, portanto, não poderiam ser mantidas entre as sondas finais, posto que hibridizariam fora da região de interesse, reduzindo assim a eficiência da *pipeline* proposta. As demais filtrações, embora não removam muitas sondas, permitem tornar a biblioteca de oligonucleotídeos em uma que seja mais confiável, reduzindo assim a possibilidade de hibridização devido à presença de *k-mer* abundante ou a inutilização da sonda por formação de estrutura secundária.

Em relação às sondas candidatas submetidas à filtragem através do modelo de análise linear discriminante do OligoMiner (LDM), é observado o mesmo padrão de remoção da grande maioria das sondas (Tabela 4.4). Dado que esse modo faz a predição de alvos usando tanto as informações inerentes do alinhamento contra o genoma de referência quanto o modelo que simula o processo de hibridização segundo a temperatura fornecida, ele inicialmente filtra mais sondas quando comparado diretamente com o UM, como pode ser observado na Tabela 4.4 em comparação à Tabela 4.3. Exatamente por reavaliar a possibilidade de hibridização fora da região alvo para aquelas sondas que apresentaram mais de um alinhamento, uma vez que alinhamentos podem conter *gaps* e *mismatches*, é possível manter algumas destas sondas através do LDM, já que nos experimentos de *FISH* tais *gaps* e *mismatches* podem culminar na não hibridização da mesma fora da região desejada. Além disso, após a filtragem de *k-mers* abundantes, a maior parte dessas sondas a mais que são mantidas pelo LDM é removida, levando a uma menor quantidade geral de sondas finais quando comparado ao UM (Tabela

STRINGENT PARAMETERS					
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Structure Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	2485	2063	2044	None 6 37	2044 2050 2081
Illumina	2591	2431	2414	0 21	2414 2435
Falcon	1883	1540	1531	None None	1531 1531
Sanger	2772	2316	2301	0 6	2301 2307
Release 6	2330	1967	1947	0 0	1947 1947

BALANCE PARAMETERS					
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Structure Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	5665	4837	4795	None 14 76	4795 4809 4871
Illumina	5566	5254	5211	0 98	5211 5309
Falcon	4307	3536	3500	None None	3500 3500
Sanger	6117	5290	5247	0 21	5247 5268
Release 6	5198	4528	4489	0 3	4489 4492

COVERAGE PARAMETERS					
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Structure Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	10293	7392	7313	None 15 102	7313 7328 7415
Illumina	8948	7979	7909	1 108	7910 8017
Falcon	8456	5591	5524	None None	5524 5524
Sanger	10683	7985	7900	1 50	7901 7951
Release 6	9104	7017	6941	0 4	6941 6945

Tabela 4.3: Quantidade de sondas sem sobreposição entre candidatas – abordagem clássica OligoMiner (UM)

Nota da tabela: Para cada subgrupo de sequências do cromossomo Y são apresentadas as quantidades de sondas obtidas ao longo das filtragens através do OligoMiner, sendo: (1) outputClean.py – retorna as sondas filtradas de acordo com o UM (*Mode Filtered*) a partir dos resultados de alinhamento com bowtie2; (2) kmerFilter.py – retorna as sondas que não apresentam *k-mers* abundantes no genoma de referência; (3) structureCheck.py – retorna as sondas que não apresentam formação de estrutura secundária, sendo essas as sondas finais nessa análise; (4) resultados do structureCheck.py a partir de subgrupos de sequências exclusivas, como descrito na Tabela 4.1; (5) valor total de sondas para cada um dos referidos subgrupos. Para o cromossomo Y montado por Chang e Larracuent (2019) os valores exclusivos incluem, por último, os de seu subgrupo exclusivo. Os blocos na tabela correspondem aos resultados a partir de diferentes parâmetros, detalhados na Tabela 2.3.

STRINGENT PARAMETERS					
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Structure Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	3696	1912	1893	None 6 38	1893 1899 1931
Illumina	2974	2390	2367	0 41	2367 2408
Falcon	3191	1434	1423	None None	1423 1423
Sanger	3760	2131	2105	0 12	2105 2017
Release 6	3151	1822	1799	0 1	1799 1800

BALANCE PARAMETERS					
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Structure Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	6170	3876	3842	None 13 73	3842 3855 3915
Illumina	5122	4403	4364	0 84	4364 4448
Falcon	4789	2712	2683	None None	2683 2683
Sanger	6024	4063	4025	0 16	4025 4041
Release 6	5084	3527	3494	0 0	3494 3494

COVERAGE PARAMETERS					
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Structure Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	6560	4917	4863	None 14 94	4863 4877 4957
Illumina	5929	5419	5377	1 58	5378 5435
Falcon	5187	3454	3406	None None	3406 3406
Sanger	6479	5093	5034	0 29	5034 5063
Release 6	5583	4527	4475	0 3	4475 4478

Tabela 4.4: Quantidade de sondas sem sobreposição entre candidatas – abordagem clássica OligoMiner (LDM)

Nota da tabela: Para cada subgrupo de sequências do cromossomo Y são apresentadas as quantidades de sondas obtidas ao longo das filtragens através do OligoMiner, sendo: (1) outputClean.py – retorna as sondas filtradas de acordo com o LDM (*Mode Filtered*) a partir dos resultados de alinhamento com bowtie2; (2) kmerFilter.py – retorna as sondas que não apresentam *k-mers* abundantes no genoma de referência; (3) structureCheck.py – retorna as sondas que não apresentam formação de estrutura secundária, sendo essas as sondas finais nessa análise; (4) resultados do structureCheck.py a partir de subgrupos de sequências exclusivas, como descrito na Tabela 4.1; (5) valor total de sondas para cada um dos referidos subgrupos. Para o cromossomo Y montado por Chang e Larracuent (2019) os valores exclusivos incluem, por último, os de seu subgrupo exclusivo. Os blocos na tabela correspondem aos resultados a partir de diferentes parâmetros, detalhados na Tabela 2.3.

4.3). Em geral, isso acontece, pois, ao remover *k-mers* abundantes, as sondas que tiveram seus múltiplos resultados de alinhamento reavaliados também será removida.

Essas conclusões são também independentes da origem das sequências inferidas para o cromossomo Y, por esse ser um problema pertencente à natureza repetitiva dessa estrutura. Pensando que grande porcentagem das sondas candidatas é descartada por apresentar mais de um alvo predito, uma solução seria aumentar a quantidade inicial das mesmas e, assim, manter maior quantidade de sondas finais. Isso garantiria ao experimento de FISH mais fontes de marcação do cromossomo Y, resultando em sinal mais intenso de fluorescência que talvez seja também mais contíguo ao longo dele.

Uma das maneiras de desenhar mais sondas candidatas é através da disponibilização de uma sequência de entrada que seja mais rica, por exemplo, a montagem de um cromossomo completamente finalizado. Por esse motivo, o método YGS é empregado para inferir mais sequências para o cromossomo Y a fim de aumentar a disponibilidade de regiões para as quais sondas *oligopaint* podem ser desenhadas (Capítulo 3). Além disso, dados os resultados apresentados até o momento, foi explorada uma opção no desenho de sondas candidatas através do `blockParse.py` que diz respeito à permissão de sobreposição ou não entre as mesmas. Ou seja, uma vez que o programa encontra uma sonda candidata que satisfaça todos os parâmetros (padronizados e/ou fornecidos pelo usuário), a procura pela próxima região que também os satisfaça deve se dar: após a posição de fim da última sonda candidata (quando a sobreposição não é permitida), sendo essa a opção padronizada no `blockParse.py`; ou na posição imediatamente posterior à de início da última sonda candidata encontrada quando a sobreposição é permitida (*-O*) (Figura 4.4).

Permitindo a sobreposição entre sondas candidatas, é possível que surjam algumas regiões que apresentem alvos únicos, assim como que não tenham *k-mers* abundantes ou probabilidade de formar estrutura secundária (Figura 4.4), aumentando então a quantidade de sondas finais após execução das devidas filtrações.

Não é interessante, entretanto, que o conjunto de sondas finais apresente sobreposição entre algumas delas, uma vez que o importante é marcar mais regiões distintas e não a mesma região várias vezes. Por isso, após execução das filtrações por predição do número de alvos, remoção de *k-mers* abundantes e de sondas que possam formar estrutura secundária, foram também removidas aquelas sondas que ainda se sobrepunham. O programa em Python abaixo foi escrito para, a partir de um arquivo BED contendo a posição inicial de sondas na segunda coluna, remover todas as sondas que apresentaram sobreposição por *-n* posições. Alterações nesse *script* podem também ser úteis para a redução da densidade de sondas.

```
1 # usage: noOverlap.py [-h] [-n NOOVERLAP] [input]
2 import argparse, sys
```

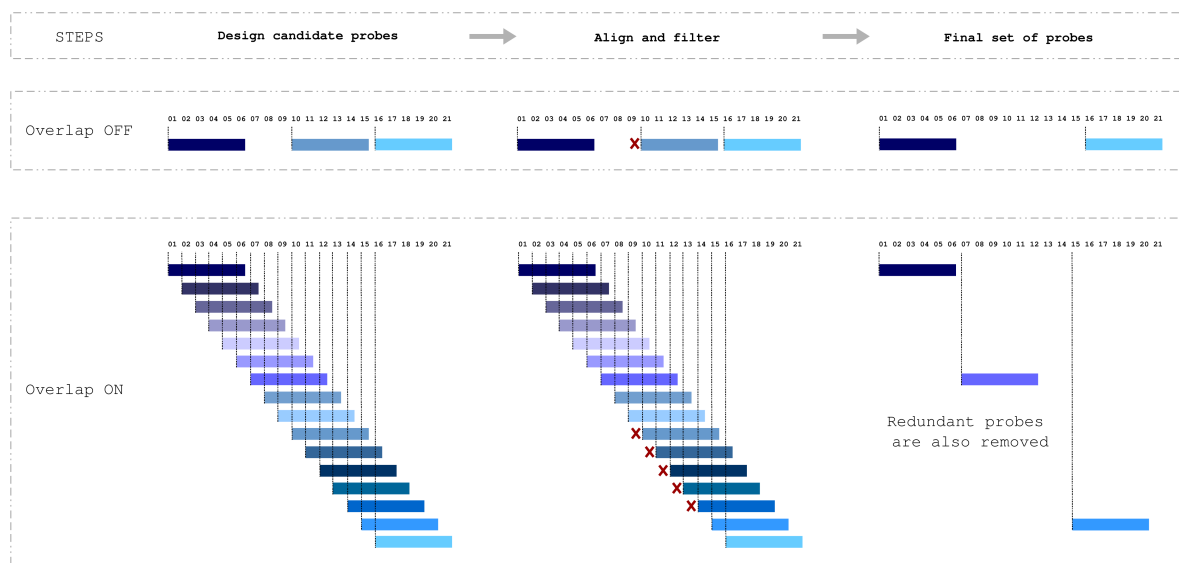


Figura 4.4: Desenhando sondas candidatas com ou sem permissão da sobreposição entre as mesmas

Nota da figura: São apresentados os passos generalizados que vão desde o desenho de sondas candidatas, passando pelo alinhamento e filtragem das mesmas, por fim resultando no conjunto final de sondas. O primeiro esquema exemplifica a não sobreposição de sondas candidatas (*Overlap OFF*), enquanto o segundo mostra a permissão da sobreposição (*Overlap ON*). Nesse último caso, a sobreposição entre sondas candidatas iniciais permite selecionar sondas que alinham em mais regiões diferentes quando comparado com o *Overlap OFF*, o que leva à seleção de mais sondas finais. Entretanto, seria prejudicial para o experimento de FISH se sondas de DNA sobrepostas competissem para hibridizar em uma mesma região. Dessa forma, após todas as filtrações devem também ser removidas as sondas redundantes, mantendo assim sondas que marcam a maior quantidade possível de regiões diferentes, mas sem sobreposição. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

```

3 ## Get command-line args
4 parser = argparse.ArgumentParser(description="Takes a probes .bed file")
5 parser.add_argument("input", help="a probes.bed file", ↵)
    default=sys.stdin, type=argparse.FileType('r'), nargs='?')
6 parser.add_argument("-n", "--noOverlap", help="positions that must not ↵)
    have probe overlap", type=int, default=20)
7 args = parser.parse_args()
8
9 ## Open data and create output file
10 data = args.input.readlines()
11 inputfile = sys.argv[3]
12 outputfile = inputfile.split(".")[0] + "_n0.bed"

```

```

13 output = open(outputfile, "w+" )
14 args.input.close()
15
16 ## Get probes that do not overlap for up to n positions
17 output.write(data[0])
18 first_probe=data[0].split("\t")
19 previous=int(first_probe[1])
20 for probe in range (1, len(data)):
21     line = data[probe].split("\t")
22     starts = int(line[1])
23     for i in range (0,args.noOverlap+1):
24         if (previous+i) == (starts): break
25     if (previous+i) != (starts):
26         previous = starts
27     output.write(data[probe])

```

Como previsto, ao desenhar sondas permitindo que haja sobreposição entre as regiões candidatas (Tabela 4.5), foram encontrados valores iniciais muito superiores aos anteriores (Tabela 4.2) independente do conjunto de parâmetros utilizado. Essas sondas candidatas foram então submetidas às filtragens com UM e LDM e, posteriormente, à remoção daquelas que apresentassem *k-mers* abundantes ou probabilidade de formar estrutura secundária. Por fim, foram mantidas apenas as sondas que não eram sobrepostas, usando o programa apresentado acima.

Assembly (Y chr seq)	Stringent Candidates	Balance Candidates	Coverage Candidates
Chang & Larracunte	2919142	4482307	5614166
Illumina	178928	320212	420232
Falcon	1997727	3029494	3654180
Sanger	953417	1473599	1802099
Release 6	1319071	2025137	2467195

Tabela 4.5: Quantidade de sondas candidatas sobrepostas desenhadas através do blockParse.py

Nota da tabela: Para cada subgrupo de sequências encontradas em comum a partir da mesma montagem genômica (*Assembly - Y chromosome sequences*) e com ambas as análises do YGS com $k = 15$ e $k = 18$, são apresentadas as quantidades de sondas candidatas desenhadas a partir de três diferentes conjuntos de parâmetros: *stringent* – mais restritivos; *balance* – valores padrão do OligoMiner; *coverage* – parâmetros mais permissivos. Os valores desses parâmetros foram apresentados na Tabela 2.3.

Agora, as menores quantidades de sondas finais encontradas através da filtragem com UM é de 2945, chegando a pouco mais de 13 mil sondas (Tabela 4.6). Quando a sobreposição de candidatas não era permitida esses valores estão entre 1531 e 8017 sondas, variando conforme a origem das sequências inferidas para o cromossomo Y (Tabela 4.3).

STRINGENT PARAMETERS

Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Overlapping Common	Remove Overlapping <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	60184	49753	49358	3611	– 19 76	3611 3630 3687
Illumina	62118	58546	58142	3540	0 87	3540 3627
Falcon	47777	38827	38435	2945	– –	2945 2945
Sanger	66708	56067	55601	4049	0 32	4049 4081
Release 6	56970	48057	47641	3448	0 4	3448 3452

BALANCE PARAMETERS

Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Overlapping Common	Remove Overlapping <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	136886	116022	114944	8223	– 25 143	8223 8248 8366
Illumina	131077	124611	123618	7297	2 254	7299 7551
Falcon	105257	86161	85367	6925	– –	6925 6925
Sanger	145885	125764	124589	9035	1 125	9036 9160
Release 6	125730	109322	108242	7714	0 10	7714 7724

COVERAGE PARAMETERS

Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Overlapping Common	Remove Overlapping <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	218761	160329	158967	12160	– 54 248	12160 12214 12408
Illumina	189415	170436	169183	10452	6 376	10458 10828
Falcon	178737	120170	119055	10554	– –	10554 10554
Sanger	225452	172356	170722	13074	4 245	13078 13319
Release 6	192869	150548	149170	11059	0 22	11059 11081

Tabela 4.6: Quantidade de sondas com sobreposição entre candidatas – abordagem clássica OligoMiner (UM)

Nota da tabela: Para cada subgrupo de sequências do cromossomo Y são apresentadas as quantidades de sondas obtidas ao longo das filtrações através do OligoMiner, sendo: (1) `outputClean.py` – retorna as sondas filtradas de acordo com o UM (*Mode Filtered*) a partir dos resultados de alinhamento com `bowtie2`; (2) `kmerFilter.py` – retorna as sondas que não apresentam *k-mers* abundantes no genoma de referência; (3) `structureCheck.py` – retorna as sondas que não apresentam formação de estrutura secundária, sendo essas as sondas finais nessa análise; (4) Remoção de sondas que se sobrepõem mesmo após as filtrações anteriores (*Remove Overlapping*); (5) resultados da remoção de sondas sobrepostas para cada um dos subgrupos de sequências exclusivas, como descrito na Tabela 4.1; (6) valor total de sondas para cada um dos referidos subgrupos. Para o cromossomo Y montado por Chang e Larracuent (2019) os valores exclusivos incluem, por último, os de seu subgrupo exclusivo. Os blocos na tabela correspondem aos resultados a partir de diferentes parâmetros, detalhados na Tabela 2.3.

Para a filtragem através do LDM, o mínimo de sondas finais é 3295, atingindo até pouco mais de 10 mil sondas (Tabela 4.7). Enquanto antes, através desse mesmo modo, os valores estão entre 1423 e 5435 sondas (Tabela 4.4).

STRINGENT PARAMETERS						
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Overlapping Common	Remove Overlapping <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	83307	43801	43459	3766	– 19 81	3766 3785 3847
Illumina	67160	54995	54540	3911	0 163	3911 4074
Falcon	75246	34863	34511	3295	– –	3295 3295
Sanger	84211	48695	48245	4190	0 61	4190 4251
Release 6	70501	42052	41697	3674	0 6	3674 3680

BALANCE PARAMETERS						
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Overlapping Common	Remove Overlapping <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	135284	89925	89009	7734	– 25 139	7734 7759 7873
Illumina	113922	100935	100086	7242	1 296	7243 7538
Falcon	108180	64706	64043	6535	– –	6535 6535
Sanger	133101	93521	92487	8346	2 124	8348 8470
Release 6	114136	82831	81895	7231	0 11	7231 7242

COVERAGE PARAMETERS						
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Overlapping Common	Remove Overlapping <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	133940	105914	105000	9900	– 53 229	9953 10129
Illumina	121426	113885	113030	8642	2 305	8644 8948
Falcon	104413	73938	73277	8494	– –	8494 8494
Sanger	130684	107830	106709	10353	3 193	10356 10546
Release 6	114674	96576	95649	8870	0 16	8870 8886

Tabela 4.7: Quantidade de sondas com sobreposição entre candidatas – abordagem clássica OligoMiner (LDM)

Nota da tabela: Para cada subgrupo de sequências do cromossomo Y são apresentadas as quantidades de sondas obtidas ao longo das filtrações através do OligoMiner, sendo: (1) outputClean.py – retorna as sondas filtradas de acordo com o LDM (*Mode Filtered*) a partir dos resultados de alinhamento com bowtie2; (2) kmerFilter.py – retorna as sondas que não apresentam *k-mers* abundantes no genoma de referência; (3) structureCheck.py – retorna as sondas que não apresentam formação de estrutura secundária, sendo essas as sondas finais nessa análise; (4) Remoção de sondas que se sobrepõem mesmo após as filtrações anteriores (*Remove Overlapping*); (5) resultados da remoção de sondas sobrepostas para cada um dos subgrupos de sequências exclusivas, como descrito na Tabela 4.1; (6) valor total de sondas para cada um dos referidos subgrupos. Para o cromossomo Y montado por Chang e Larracuent (2019) os valores exclusivos incluem, por último, os de seu subgrupo exclusivo. Os blocos na tabela correspondem aos resultados a partir de diferentes parâmetros, detalhados na Tabela 2.3.

Após exploradas todas essas opções usando as abordagens clássicas do OligoMiner, ainda é observada, no entanto, uma porcentagem semelhante de sondas sendo removidas por

apresentar mais de um alvo predito e, embora muitas das sequências do cromossomo Y de fato se repitam fora do mesmo, algumas das vezes essa repetição ocorre dentro dele próprio. Assim, sondas desenhadas para essas regiões não precisariam ser removidas, dado que não estariam hibridizando fora do cromossomo de interesse. Explorando mais esses aspectos, foi encontrada uma nova abordagem – que ainda usa o OligoMiner e que mantém aquelas sondas que se repetem apenas no próprio cromossomo de interesse (Seção 4.2).

4.2 Uma nova abordagem a partir do OligoMiner Zero

Mode para selecionar sondas repetitivas exclusivas

Ambos os modos do `outputClean.py` utilizados na abordagem clássica (Seção 4.1) partem da premissa de que o alvo com predição única encontrada alinhou uma única vez na própria sequência de interesse fornecida como arquivo de entrada para o `blockParse.py`, já que essa mesma sequência também deve estar presente no genoma de referência. Lembrando que os autores do OligoMiner recomendam mascarar o arquivo FASTA de entrada (Capítulo 2, Subseção 2.2.3) para evitar até mesmo que sondas candidatas sejam desenhadas para regiões repetitivas (BELIVEAU *et al.*, 2018). Deste modo, faz sentido que esse alinhamento único seja utilizado, evitando com mais segurança hibridização fora do cromossomo desejado.

Para esta pesquisa, entretanto, muitas das sequências repetitivas do cromossomo Y se repetem exclusivamente no próprio cromossomo. Assim, a remoção de sondas desenhadas para essas regiões não é necessária, já que estas não apresentarão complementariedade com outros cromossomos. Dessa forma, apenas é necessário remover sondas que alinhem fora do cromossomo Y ou mesmo que apresentem *k-mers* que sejam abundantes fora da região de interesse.

Uma alternativa encontrada para satisfazer esse critério está no próprio OligoMiner quando o usuário escolhe a opção `-0` para que sejam filtradas as sondas que tiveram nenhum alinhamento no genoma de referência (Capítulo 2, Subseção 2.2.3). Beliveau *et al.* (2018) descrevem que esse modo pode ser útil quando a região alvo envolve transgenes ou mesmo sequências exógenas, uma vez que as sondas desenhadas para essas regiões não iriam alinhar no genoma de referência do organismo onde essas sequências genéticas teriam sido inseridas.

Como esse modo mantém as sondas com zero alinhamentos, basta então que o genoma de referência fornecido para alinhar as sondas candidatas não contenha a sequência de entrada fornecida para o `blockParse.py` (Figura 4.5). Dessa maneira, todas as sondas desenhadas para o cromossomo Y que se repitam fora do mesmo apresentarão um ou mais resultados de alinhamento, enquanto aquelas sondas que são exclusivas desse cromossomo resultarão em

zero alinhamentos.

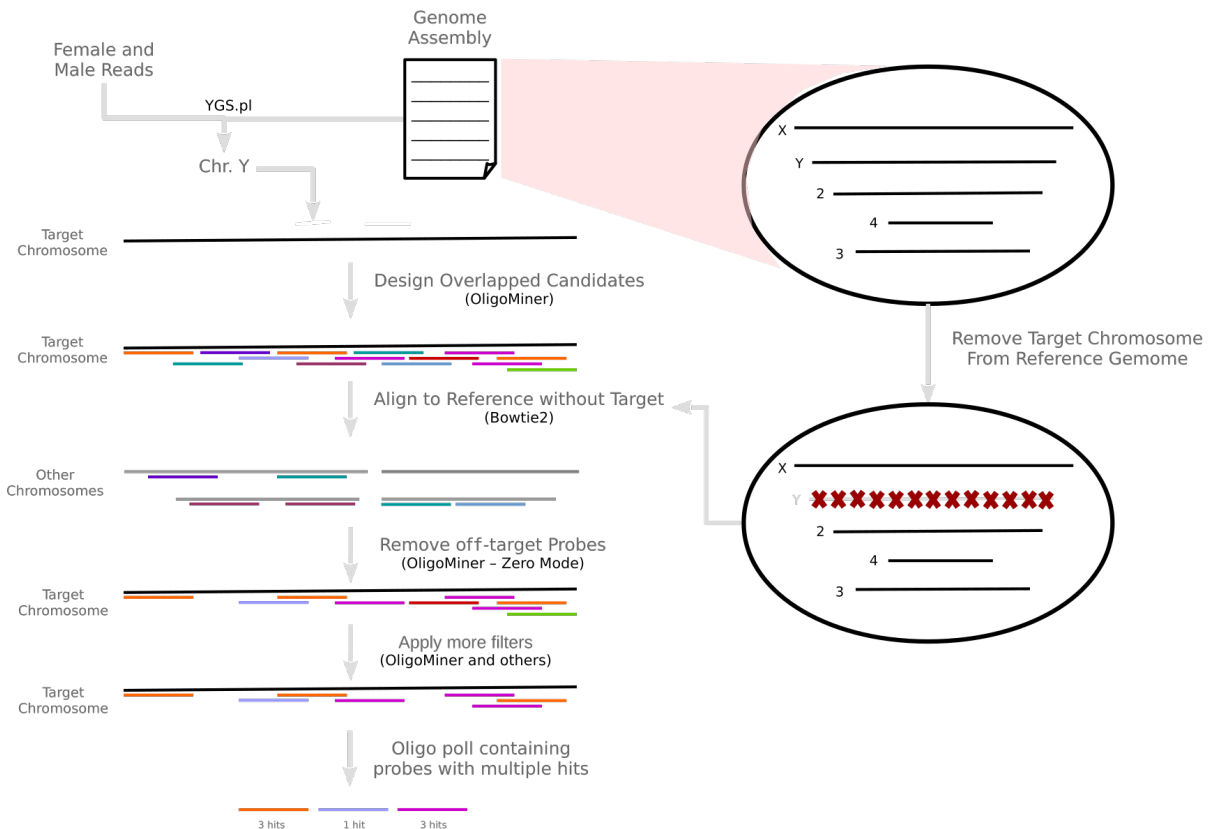


Figura 4.5: Esquema da nova abordagem que permite selecionar sondas repetitivas exclusivas

Nota da figura: As sondas desenhadas com o programa `blockParse.py` a partir do conjunto de sequências inferidas para o cromossomo Y com as análises do YGS e `BlobTools` (*Target Chromosome* - cromossomo alvo) são alinhadas contra o genoma de referência sem esse mesmo conjunto de sequências do cromossomo alvo. Isso permite que o `outputClean.py` com argumento `-0` (*Zero Mode* - ZM) remova todas as sondas que alinham uma ou múltiplas vezes em outros cromossomos do genoma que não o alvo (*Other Chromosomes*). Ou seja, são mantidas as sondas exclusivas do cromossomo Y, tanto as que aparecem apenas uma vez nesse cromossomo como as que se repetem apenas nele. Em seguida são aplicadas as outras filtragens e por fim a biblioteca de sondas conterá sondas com vários alvos no cromossomo de interesse. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

Para selecionar os identificadores das sequências que não foram inferidos para o cromossomo Y, ou seja, os identificadores das sequências que eram de outros cromossomos e que, portanto, iriam compor o genoma de referência sem o Y, foram utilizados os comandos em UNIX abaixo, como `SED` e `grep` (MCMAHON, 1978; KERNIGHAN; PIKE, 1984; RUALTHANZAUVA, 2014). De acordo com a origem do conjunto de sequências inferidas para o cromossomo Y utilizado como arquivo de entrada para desenho de sondas candidatas, o mesmo foi também removido de seu respectivo genoma de referência para que, contra esse,

essas sondas pudessem ser alinhadas e filtradas com ZM.

```
1 ## Gerar genoma sem os contigs inferidos para o cromossomo Y
2 awk 'FNR==NR{seen[$0]++; next} !($0 in seen)' crhY.fasta genome.fasta <->
   > noYgenome.fasta
```

A criação dos *index* e tabela *hash* de *k-mers* e respectivas contagens, necessários para as etapas de filtragem do `outputClean.py` e `kmerFilter.py`, respectivamente, foi realizada através dos programas `bowtie2-build` e `jellyfish count`, seguindo as mesmas orientações descritas no capítulo anterior (Subseção 2.2.3) (LANGMEAD; SALZBERG, 2012; MARÇAIS; KINGSFORD, 2011). Em específico, como os *contigs* inferidos para o cromossomo Y variam com o valor de *k* utilizado na análise do YGS, foram gerados genomas sem cada um dos diferentes conjuntos de sequências inferidas para o cromossomo Y e estes genomas foram usados adequadamente (Figura 4.6). Para as sondas desenhadas para o subgrupo de *contigs* encontrados por ambos os tamanhos de *k-mer* (Tabela 4.1), o genoma sem o conjunto de sequências inferido para o cromossomo Y com $k = 15$ foi utilizado, já que todos os *contigs* em comum necessariamente estariam removidos dessa referência.

Para a etapa de alinhamento foram empregados os mesmos parâmetros mais sensíveis recomendados para filtrar as sondas através do UM por Beliveau *et al.* (2018) (Capítulo 2, Subseção 2.2.3). Essa escolha foi tomada uma vez que o resultado do alinhamento é o único indicativo utilizado para predição do número de alvos de cada uma das sondas. Ou seja, para filtrar sondas na nova abordagem com ZM, o alinhamento deve ser tão sensível quanto o realizado para filtrar sondas através do UM.

Pela natureza do desenho de sondas através do ZM é possível que algumas sequências idênticas tenham sido desenhadas a partir de diferentes regiões do arquivo de entrada compoñham o conjunto final de sondas após remoção por predição do número de alvos, presença de *k-mers* abundantes e formação de estrutura secundária. Em especial para a predição do número de alvos, sondas idênticas desenhadas para mais de uma região no cromossomo Y serão mantidas sempre que tiverem nenhum alinhamento no genoma de referência sem o Y e isso não somente permite aumentar a quantidade de sondas finais, como também que uma quantidade menor de sequências efetivamente apresente mais alvos, como esquematizado na Figura 4.5. Por exemplo, tendo 500 sondas finais das quais 300 alinham apenas uma vez no alvo e as outras 200 alinham, cada uma, duas vezes no alvo, implica dizer que, na verdade, o conjunto final de sondas era composto por 700 sondas, quando efetivamente apenas 500 precisam ser sintetizadas para que essas 700 diferentes regiões sejam fluorescentemente marcadas. Na prática, essas sondas idênticas desenhadas para o cromossomo Y e que não forem removidas nas etapas de filtragem podem aparecer mais que apenas duas vezes e isso

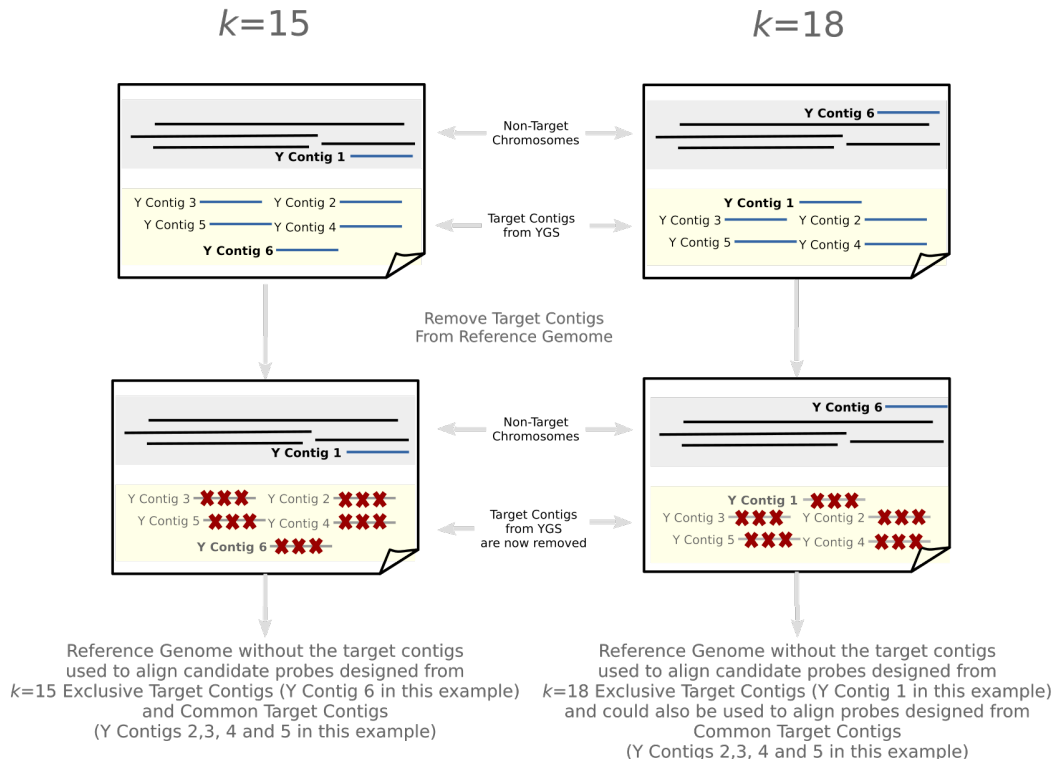


Figura 4.6: Genoma de Referência sem as sequências inferidas para o Cromossomo Y

Nota da figura: As sondas candidatas serão alinhadas e filtradas usando o genoma de referência sem os *contigs* do cromossomo Y alvo que foi utilizado para desenhar as mesmas. Detalhadamente, para o alinhamento e remoção de k -mers abundantes serão usados: (i) para os *contigs* inferidos exclusivamente com $k = 15$ será usado o genoma de referência sem todos os *contigs* inferidos com $k = 15$; (ii) para os *contigs* inferidos exclusivamente com $k = 18$ será usado o genoma de referência sem todos os *contigs* inferidos com $k = 18$; (iii) para os *contigs* inferidos em ambas as análises do YGS (tanto $k = 15$ quanto $k = 18$) será usado o genoma de referência sem todos os *contigs* inferidos com $k = 15$ – essa foi apenas uma escolha, poderia ter sido também $k = 18$. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

implica tanto uma economia para síntese comercial da biblioteca de oligonucleotídeos, como também no enriquecimento do sinal ao longo de diferentes regiões do cromossomo Y.

Como apontado, entretanto, não é necessário que em uma mesma biblioteca de oligonucleotídeos existam sondas com mais de uma cópia, já que cada cópia única sintetizada será amplificada, o que permitirá sua hibridização em vários pontos da amostra fixada. Dessa forma, é preciso que cópias extras de sondas sejam removidas, para que cada uma destas apareça apenas uma vez no conjunto final de sondas a ser sintetizado. Para executar essa etapa foi escrito o programa em Python abaixo, que recebe um arquivo BED contendo sondas na quarta coluna e retorna apenas a primeira exibição de cada uma delas para um arquivo

“_nR.bed”. Essa etapa é esquematizada no último passo da Figura 4.5 – após aplicar diversas filtragens ainda estão presentes sondas que se sobrepõe (sendo removidas conforme explicado anteriormente) e outras que aparecem várias vezes. Assim, na biblioteca final de sondas (*Oligo pool*) é mantida apenas uma cópia de cada sonda.

```

1 # usage: nonRedundant.py [-h] [input]
2 import argparse, sys
3 ## Open data and create output file
4 data = args.input.readlines()
5 inputfile = sys.argv[1]
6 outputfile = inputfile.split(".")[0] + "_nR.bed"
7 output = open(outputfile, "w+" )
8 args.input.close()
9 ## Get the first report of each probe
10 out_probes=[]
11 times_probes=[]
12 for probe in range (0, len(data)):
13     thisLine = data[probe].split("\t")
14     thisProbe = thisLine[3]
15     unique=True
16     for eachProbe in range (0,len(out_probes)):
17         outLine = out_probes[eachProbe].split("\t")
18         outProbe = outLine[3]
19         if thisProbe == outProbe:
20             times_probes[eachProbe] += 1
21             unique=False
22             break
23     if unique == True:
24         out_probes.append(data[probe])
25         times_probes.append(1)
26         output.write(data[probe])

```

Dito isto, todas as mesmas sondas candidatas desenhadas com os critérios previamente citados ao permitir ou não permitir a sobreposição entre estas no `blockParse.py` foram utilizadas para alinhamento NGS local muito sensitivo (Tabela 4.5 e Tabela 4.2, respectivamente). Esses resultados de alinhamento foram submetidos ao programa `outputClean.py` no modo que filtra apenas sondas com zero alinhamentos (ZM) e, em seguida, as sondas foram filtradas

com os programas `kmerFilter.py` e `structureCheck.py`, foram então removidas as sondas que apresentaram sobreposição (para a análise em que o desenho de sondas candidatas as permitia), sendo por fim mantida apenas uma cópia de cada sequência de sondas, seguindo a ordem aqui mencionada.

As quantidades de sondas desenhadas após a aplicação dessa análise mostram que, de fato, a utilização do ZM é apropriada para manter sondas do cromossomo Y que, embora se repitam, não hibridizam fora da região alvo (Tabela 4.8 e Tabela 4.9). Mesmo após todas as filtragens aplicadas, discutidas em mais detalhe a seguir (Subseção 4.2.1), o mínimo de sondas desenhadas com a abordagem aqui proposta através do ZM é de 2634 sondas que apresentam mais de 2800 alvos, dado que algumas delas hibridizam mais de uma vez no cromossomo Y. Ao usar as sondas que não se sobrepõem, foram selecionadas mais de 18,500 sondas com mais de 34 mil alvos, a partir dos parâmetros mais restritivos e usando o conjunto de sequências inferidas para o cromossomo Y com $k = 18$ e com a montagem de Chang e Larracuent (2019).

Para as sondas desenhadas com sobreposição entre candidatos, um mínimo de 3500 sondas com quase 4 mil alvos compõe parte do conjunto final de sequências filtradas a partir dos parâmetros mais restritivos. As quantidades máximas de sondas desenhadas para o cromossomo Y usando o ZM com sobreposição entre candidatos passa das 26 mil sondas e estas apresentam cerca de 50 mil alvos diferentes nessa região de interesse.

De maneira generalizada, enquanto a partir da montagem de Chang e Larracuent (2019) são geradas as quantidades de sondas mais promissoras e que a montagem Falcon em muito dela se aproxima, o conjunto de sequências inferidas para o cromossomo Y a partir da montagem Illumina, independente do valor de k utilizado, sempre produz a menor quantidade de sondas e alvos. Entretanto, vale ressaltar que todas as discussões apresentadas neste capítulo a respeito de quantidade de sondas finais consideram apenas o potencial de marcação com sinal mais forte e extensão esperada. Ou seja, quanto maior a quantidade de sondas e mais esparsas as mesmas estiverem ao longo do cromossomo de interesse, mais eficiente será o experimento de FISH. Para outros cromossomos que não o Y e que, portanto, são melhor montados, nem todas as sondas desenhadas precisam ser sintetizadas, dado que o objetivo é simplesmente manter sinal consistente ao longo de todo o cromossomo. Mas para o cromossomo Y, que não tem montagem completa, e que muitas vezes não se conhece a ordem de seus *contigs* definidos e o tamanho do espaçamento entre os mesmos, não é possível pré-determinar se as sondas desenhadas se distribuem consistentemente ao longo de todo ele. Assim, maiores quantidades de sondas acabam sendo associadas ao provável sucesso do experimento de FISH. A validação dessa *pipeline*, com a devida discussão da marcação observada através do experimento de FISH *Oligopaint* é apresentada no Capítulo 5.

Note que as tabelas com os resultados do ZM contam também com uma filtragem

STRINGENT PARAMETERS							
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Redundant Common	Fem. Reads Filtered Common	Fem. Reads Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	12620	11174	11082	6172	5760	– 113 58	5760 5873 5818
Illumina	2965	2836	2816	2805	2634	0 52	2634 2686
Falcon	11705	10656	10591	6182	5730	– –	5730 5730
Sanger	3970	3396	3362	3274	3106	0 65	3106 3171
Release 6	6368	5692	5635	4620	4335	0 46	4335 4381

BALANCE PARAMETERS							
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Redundant Common	Fem. Reads Filtered Common	Fem. Reads Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	25587	23295	23033	13022	12231	– 177 105	12231 12408 12336
Illumina	6217	5988	5941	5921	5594	0 155	5594 5749
Falcon	22113	20382	20119	11751	10980	– –	10980 10980
Sanger	8674	7631	7567	7318	6972	0 176	6972 7148
Release 6	12887	11710	11607	9458	8898	0 111	8898 9009

COVERAGE PARAMETERS							
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Redundant Common	Fem. Reads Filtered Common	Fem. Reads Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	41363	34369	34041	19211	18316	– 226 135	18316 18542 18451
Illumina	9842	9098	9020	8980	8576	0 181	8576 8757
Falcon	35403	29702	29406	16891	16071	– –	16071 16071
Sanger	14742	11586	11458	11052	10633	1 300	10634 10933
Release 6	21084	17602	17420	14127	13443	0 165	13443 13608

Tabela 4.8: Quantidade de sondas sem sobreposição entre candidatas – OligoMiner (ZM)

Nota da tabela: Para cada subgrupo de sequências do cromossomo Y são apresentadas as quantidades de sondas obtidas ao longo das filtragens através do OligoMiner, sendo: (1) outputClean.py – retorna as sondas filtradas de acordo com o ZM (*Mode Filtered*) a partir dos resultados de alinhamento com bowtie2; (2) kmerFilter.py – retorna as sondas que não apresentam *k-mers* abundantes no genoma de referência; (3) structureCheck.py – retorna as sondas que não apresentam formação de estrutura secundária, sendo essas as sondas finais nessa análise; (4) Remoção das sondas redundantes, mantendo apenas uma cópia de cada sequência de oligonucleotídeos (*Remove Redundant*); (5) Resultados de Alinhamento e Filtragem (outputClean.py -0) contra os *short-reads* de fêmea (*Fem. Reads Filtered*); (6) Resultados de Alinhamento e Filtragem (outputClean.py -0) contra os *short-reads* de fêmea para cada um dos subgrupos de sequências exclusivas, como descrito na Tabela 4.1; (7) valor total de sondas para cada um dos referidos subgrupos. Para o cromossomo Y montado por Chang e Larracuent (2019) os valores exclusivos incluem, por último, os de seu subgrupo exclusivo. Os blocos na tabela correspondem aos resultados a partir de diferentes parâmetros, detalhados na Tabela 2.3.

STRINGENT PARAMETERS								
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Overlapping Common	Remove Redundant Common	Fem. Reads Filtered Common	Fem. Reads Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	320430	287416	284907	18631	9887	8757	– 164 90	8757 8921 8847
Illumina	70848	68013	67561	3925	3905	3500	0 127	3500 3627
Falcon	308159	281996	279800	17077	9464	8388	– –	8388 8388
Sanger	97195	83678	83003	5705	5484	4956	0 134	4956 5090
Release 6	160537	144382	143222	9229	7289	6428	0 86	6428 6514

BALANCE PARAMETERS								
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Overlapping Common	Remove Redundant Common	Fem. Reads Filtered Common	Fem. Reads Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	611405	557938	553239	37343	20165	18185	– 244 150	18185 18429 18335
Illumina	145364	140558	139472	7964	7924	7206	0 291	7206 7497
Falcon	531251	490172	485222	32589	18210	16383	– –	16383 16383
Sanger	205501	180919	179174	12574	12025	11063	1 353	11064 11416
Release 6	306021	279116	276746	18545	14796	13316	0 175	13316 13491

COVERAGE PARAMETERS								
Assembly (Y chr seq)	Mode Filtered Common	k-mer Filtered Common	Structure Filtered Common	Remove Overlapping Common	Remove Redundant Common	Fem. Reads Filtered Common	Fem. Reads Filtered <i>k</i> exclusive (<i>k</i> = 15 <i>k</i> = 18)	Total (<i>k</i> = 15 <i>k</i> = 18)
Chang & Larracuent	875399	731079	724796	51808	28052	25991	– 332 257	25991 26323 26248
Illumina	207558	192696	191279	11503	11436	10608	5 454	10613 11062
Falcon	728030	615630	610215	44599	24781	22883	– –	22883 22883
Sanger	309308	247688	245372	17975	17152	16132	4 627	16136 16759
Release 6	439212	371261	368126	25802	20522	19007	0 257	19007 19264

Tabela 4.9: Quantidade de sondas com sobreposição entre candidatas – OligoMiner (ZM)

Nota da tabela: Para cada subgrupo de sequências do cromossomo Y são apresentadas as quantidades de sondas obtidas ao longo das filtragens através do OligoMiner, sendo: (1) outputClean.py – retorna as sondas filtradas de acordo com o ZM (*Mode Filtered*) a partir dos resultados de alinhamento com bowtie2; (2) kmerFilter.py – retorna as sondas que não apresentam *k-mers* abundantes no genoma de referência; (3) structureCheck.py – retorna as sondas que não apresentam formação de estrutura secundária, sendo essas as sondas finais nessa análise; (4) Remoção de sondas que se sobrepõem mesmo após as filtragens anteriores (*Remove Overlapping*); (5) Remoção das sondas redundantes, mantendo apenas uma cópia de cada sequência de oligonucleotídeos (*Remove Redundant*); (6) Resultados de Alinhamento e Filtragem (outputClean.py -0) contra os *short-reads* de fêmea (*Fem. Reads Filtered*); (7) Resultados de Alinhamento e Filtragem (outputClean.py -0) contra os *short-reads* de fêmea para cada um dos subgrupos de sequências exclusivas, como descrito na Tabela 4.1; (8) valor total de sondas para cada um dos referidos subgrupos. Para o cromossomo Y montado por Chang e Larracuent (2019) os valores exclusivos incluem, por último, os de seu subgrupo exclusivo. Os blocos na tabela correspondem aos resultados a partir de diferentes parâmetros, detalhados na Tabela 2.3.

contra os *short-reads* de fêmea (Tabela 4.8 e Tabela 4.9). Essa filtragem foi empregada como parte da *pipeline* proposta após uma série de testes que objetivam garantir a eliminação de quaisquer sondas que pudessem ter hibridização fora da região alvo. Sendo essa uma preocupação inerente à alta repetitividade do cromossomo Y e à proposição de uma nova abordagem a partir da opção ZM do OligoMiner (BELIVEAU *et al.*, 2018). As discussões pertinentes a estes testes são brevemente citadas a seguir (Subseção 4.2.1).

4.2.1 Explorando alternativas de filtragem

Como essa é uma abordagem diferente da originalmente proposta pelos autores do OligoMiner para o ZM (BELIVEAU *et al.*, 2018), algumas alternativas de filtragem extras foram exploradas de maneira que quaisquer possibilidades de hibridização fora da sequência alvo pudessem ser anuladas (Figura 4.7). Dispondo dos *short-reads* de fêmea, necessários para a inferência de *contigs* para o cromossomo Y (Capítulo 3), foram feitos testes breves para incluir o uso destes *reads* na filtragem das sondas desenhadas. Isso porque algumas vezes os *short-reads* de fêmea podem conter *k-mers* que sejam abundantes em fêmeas, mas que por algum motivo foram descartados pelos algoritmos de montagem. Esses testes foram realizados apenas através da versão oficial mais recente do genoma de *Drosophila melanogaster* (Release 6). Uma vez que muitos dos *scaffolds* e *contigs* estão montados para algum cromossomo, conferir se determinada alternativa de filtragem está filtrando as sondas inadequadamente se torna um processo mais simples.

Foi brevemente examinado se para a etapa de alinhamento NGS seria possível substituir o *index* criado a partir do genoma sem o Y por um criado a partir dos *short-reads* de fêmea. A vantagem que essa substituição traria está na abundância dos *k-mers* em fêmea que acabam não estando montados no genoma de referência e que, dessa forma, poderiam ser considerados. Entretanto, foi constatado que muitas das sondas candidatas que não estavam sendo removidas, presumidamente por não alinharem em nenhum cromossomo que não fosse o Y, em verdade alinhavam em várias outras regiões do genoma (2R, 2cen, 2L, 3R, 3cen, 3L, 4, X, rDNA). Essa conferência foi realizada através do comando abaixo que procura por sequências idênticas às das sondas em cada um dos cromossomos – onde armX foi substituído pelo nome de cada um dos cromossomos citados no genoma de referência (Release 6).

```
1 cut -f4 filteredProbesWithFemaleReadsIndex.bed | while read probe; do ↵
    if grep -q "$probe" genome.fasta; then grep -n -e " armX " -e ↵
        "$probe" genome.fasta; fi; done
```

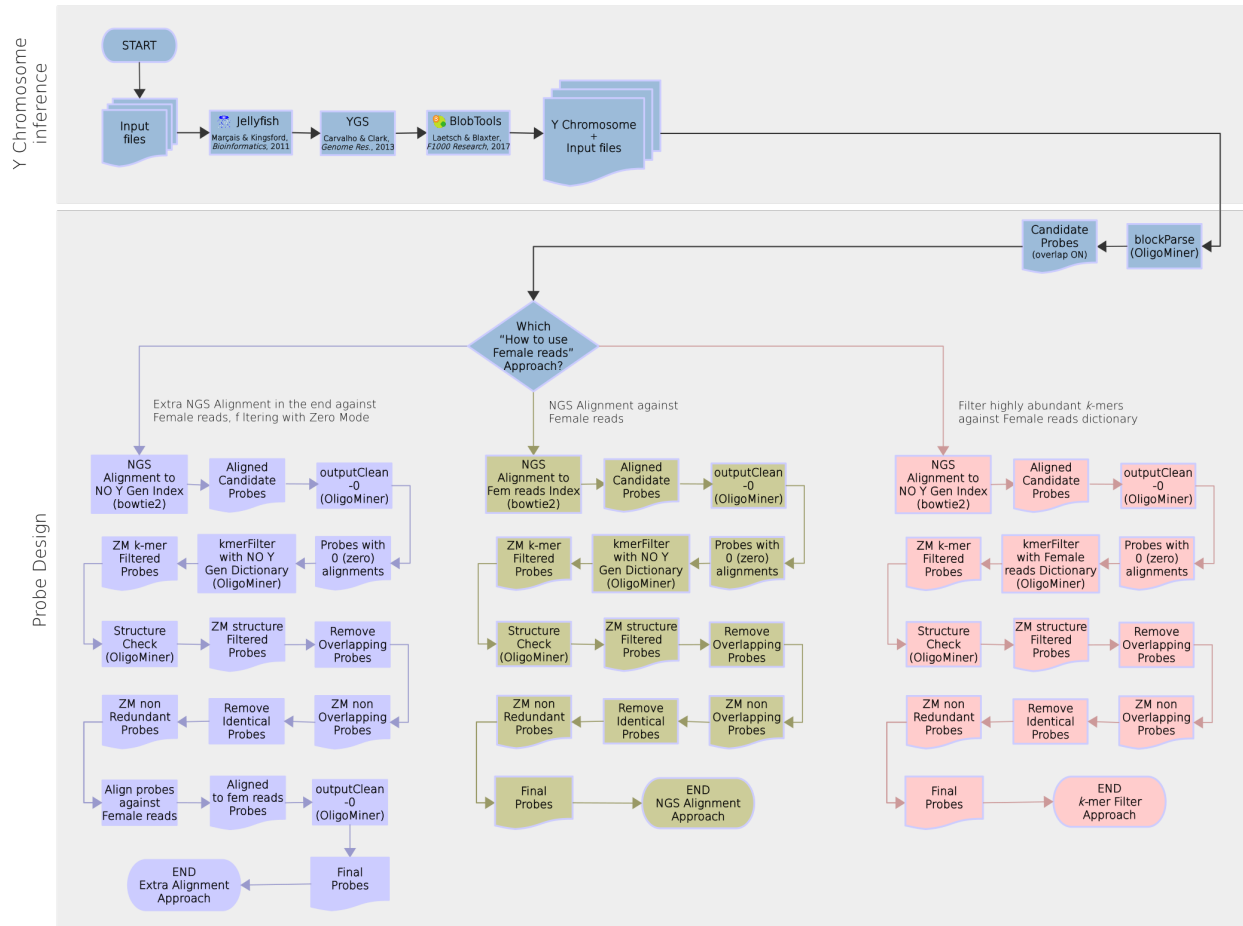


Figura 4.7: Alternativas exploradas para filtragem com *short-reads* de fêmea

Nota da figura: A primeira alternativa explorada, que aparece à esquerda no esquema em tons azulados, utiliza um *index* gerado a partir dos *short-reads* de fêmea em um alinhamento extra no final da *pipeline*, seguido de filtragem através do `outputClean.py -0` (ZM). A segunda alternativa, centralizada na imagem em tons esverdeados, substitui o *index* gerado a partir do genoma sem o Y por um gerado a partir dos próprios *short-reads* de fêmea para o primeiro alinhamento NGS. A terceira alternativa explorada, à direita do esquema em tons rosados, substitui a tabela *hash* de *k-mers* e respectivas contagens (*Dictionary*) gerada a partir do genoma sem o Y por uma tabela *hash* gerada a partir dos *short-reads* de fêmea. Todos os esquemas partem de sondas com sobreposição entre candidatas e, para filtragem em que essa opção não foi utilizada, basta desconsiderar a etapa de remoção de sondas sobrepostas. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

Também foi verificado se os *short-reads* de fêmea não poderiam ser usados em substituição ao genoma para criação da tabela *hash* de *k-mer* e respectivas contagens através do jellyfish count (MARÇAIS; KINGSFORD, 2011), a fim de remover *k-mers* abundantes em fêmea e não montados. No entanto, a abundância de tais *k-mers* é afetada justamente por proverem de *reads* não montados. Esta conclusão foi realizada a partir de análise do

relatório (-R) gerado pelo programa `kmerFilter.py` (Figura 4.8), que mostra um aumento na porcentagem de sondas sendo removidas por apresentar *k-mers* que apareçam mais de 20, 50 e 100 vezes, independente das sondas terem sido desenhadas com ou sem sobreposição de seus candidatos.

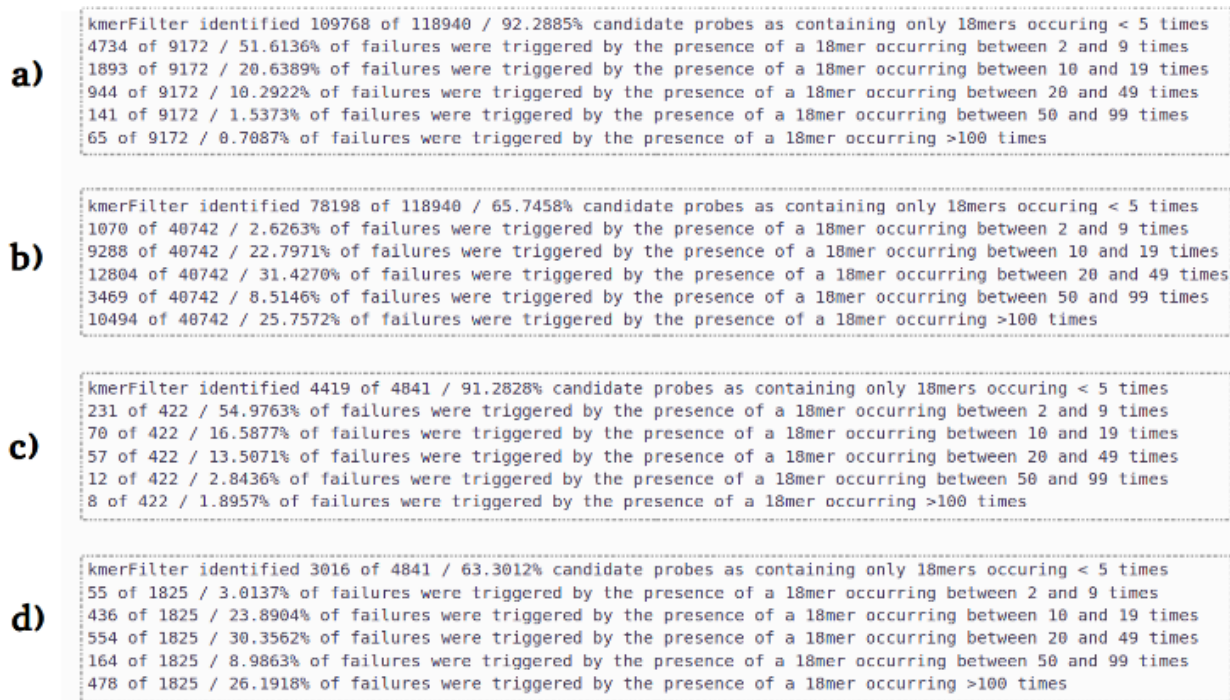


Figura 4.8: Relatório do `kmerFilter.py` para filtragem com genoma ou *short reads* de fêmea

Nota da figura: Cabeçalho do relatório fornecido pelo programa `kmerFilter.py` para análises com sondas que foram geradas permitindo sobreposição entre candidatas (a;b) ou não (c;d). Tais sondas foram filtradas contra a tabela *hash* feita a partir de (a;c) genoma sem o cromossomo Y ou (b;d) dos *short-reads* de fêmeas. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

A alternativa encontrada para utilizar esses *reads* foi adicionar uma última filtragem que também utiliza o programa `outputClean.py` através do ZM para processar os resultados do alinhamento NGS feito contra os mesmos. Essa filtragem é realizada após todas as outras, incluindo a remoção de sondas que sobrepõem – quando a sobreposição entre candidatos é permitida – e a remoção de cópias extras de uma mesma sonda. Para a realização dessa etapa, os *short-reads* de fêmea são transformados em sequência FASTA usando `seqtk seq` (Disponível em <https://github.com/lh3/seqtk>) como exemplificado no comando a seguir.

```
1 seqtk seq -a FemaleReads *.fastq.gz > FemaleReads.fasta
```

Posteriormente, o *index* desse arquivo é gerado através de bowtie2-build (LANGMEAD; SALZBERG, 2012) conforme as orientações descritas no capítulo anterior (Subseção 2.2.3). Para que o alinhamento NGS seja realizado, entretanto, é necessário também que as sondas presentes no arquivo BED sejam convertidas para FASTQ. Para tanto, o programa `bedToFastq.py` presente na suíte de programas do OligoMiner pode ser utilizado (BELIVEAU *et al.*, 2018). Os parâmetros usados nesse alinhamento NGS final contra *index* dos *short-reads* de fêmea são os mesmos usados tanto para filtragem inicial com UM quanto ZM.

4.3 Comparações entre as abordagens

Para fins de melhor comparar os três diferentes modos do OligoMiner (UM, LDM e ZM), foi realizada a filtragem contra os *short-reads* de fêmea para os resultados obtidos dos modos clássicos UM e LDM, fazendo a devida substituição do Genoma sem o Y pelo Genoma de Referência completo, assim como apenas removendo sondas que sobrepunham quando tal sobreposição foi permitida no desenho de sondas candidatas (Tabela 4.10 e Tabela 4.11). Isso garante mais confiabilidade aos resultados finais obtidos através da abordagem clássica, já que a própria sequência utilizada como arquivo de interesse foi inferida e não propriamente montada e curada.

A análise através do ZM sempre retorna o maior número de sondas finais, em média duas vezes maior quando comparado aos outros dois modos, independente da análise que permite ou não a sobreposição entre as sondas desenhadas pelo programa `blockParse.py`. Ademais, comparações mais diretas entre os diferentes modos e parâmetros explorados foram realizadas.

A primeira conferência realizada foi justamente entre as sondas com e sem sobreposição, a fim de observar se, quando a sobreposição é permitida, todas as sondas desenhadas sem sobreposição entre candidatos não estariam presentes em ambas as análises. Isso implicaria que de fato o desenho de sondas com sobreposição entre candidatos é uma opção ideal, já que além de resultar em mais sondas, não perde nenhuma daquelas que teriam sido desenhadas da outra forma. Como essas sondas vão sendo filtradas, a comparação foi feita entre as sondas finais que originalmente não tinham sobreposição entre seus candidatos e, imediatamente após a filtragem com `outputClean.py` par a par para cada análise (modo – UM, LDM ou ZM; subgrupo utilizado; conjunto de parâmetros), para o grupo de sondas que permitia tal sobreposição. Isso porque, se as sondas forem selecionadas pelo número predito de alvos, para aquelas com sobreposição entre candidatos será possível que essa mesma sonda não seja selecionada em detrimento de outra na etapa de remoção de sondas sobrepostas redundantes.

STRINGENT PARAMETERS						
UM			LDM			ZM
Assembly (Y chr seq)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)
Chang & Larracuento	1964	1964 1970 2000	1594	1594 1600 1631	5760	5760 5873 5818
Illumina	2287	2287 2305	1915	1915 1932	2634	2634 2686
Falcon	1455	1455 1455	1143	1143 1143	5730	5730 5730
Sanger	2207	2207 2211	1757	1757 1764	3106	3106 3171
Release 6	1857	1857 1857	1514	1514 155	4335	4335 4381

BALANCE PARAMETERS						
UM			LDM			ZM
Assembly (Y chr seq)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)
Chang & Larracuento	4602	4602 4615 4672	3424	3424 3437 3493	12231	12231 12408 12336
Illumina	4962	4962 5051	3788	3788 3849	5594	5594 5749
Falcon	3329	3329 3329	2325	2325 2325	10980	10980 10980
Sanger	5039	5039 5057	3538	3538 3549	6972	6972 7148
Release 6	4282	4282 4284	3078	3078 3078	8898	8898 9009

COVERAGE PARAMETERS						
UM			LDM			ZM
Assembly (Y chr seq)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)
Chang & Larracuento	7074	7074 7088 7171	4765	4765 4779 4857	18316	18316 18542 18451
Illumina	7590	7590 7688	5213	5213 5263	8576	8576 8757
Falcon	5332	5332 5332	3312	3312 3312	16071	16071 16071
Sanger	7636	7637 7682	4902	4902 4929	10633	10634 10933
Release 6	6663	6663 6667	4365	4365 4368	13443	13443 13608

Tabela 4.10: Quantidade de sondas filtradas contra *short reads* de fêmea – sobreposição não permitida

Nota da tabela: Para cada subgrupo de sequências inferidas para o cromossomo Y são apresentadas as quantidades de sondas obtidas após a filtragem feita a partir dos resultados de alinhamento NGS contra os *short-reads* de fêmea com bowtie2 (*FR Filtered*). Essa filtragem foi feita a partir do programa `outputClean.py -0`. Para cada modo (UM, LDM e ZM, respectivamente) são apresentados os valores de sondas para os subgrupos de *contigs* em comum do cromossomo Y, seguido do total para cada conjunto de sequências inferidas para o cromossomo Y a partir de um valor de k diferente. Para o cromossomo Y montado por Chang e Larracuento (2019) os valores exclusivos incluem, por último, os de seu subgrupo exclusivo. Os blocos na tabela correspondem aos resultados a partir de diferentes parâmetros, detalhados na Tabela 2.3.

Assim, o comando abaixo foi executado, retornando em um arquivo aquelas que seriam as sondas exclusivas da análise que não permite a sobreposição de candidatos. Como esperado, a não sobreposição de sondas candidatas é uma análise que nunca retorna sondas exclusivas.

```
1 awk 'FNR==NR{a[$4];next};!($4in a)' oCprobesWithOverlapON.bed <->
    finalProbesWithOverlapOFF.bed > exclusiveFinalProbesOverlapOFF.bed
```

Pela lógica da abordagem clássica, comparada à da nova abordagem proposta com

STRINGENT PARAMETERS								
UM			LDM			ZM		
Assembly (Y chr seq)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)
Chang & Larracuent	3318	3318 3334 3382	2681	2681 2698 2748	8757	8757 8921 8847		
Illumina	3219	3219 3291	2630	2630 2708	3500	3500 3627		
Falcon	2684	2684 2684	2215	2215 2215	8388	8388 8388		
Sanger	3702	3702 3728	2899	2899 2939	4956	4956 5090		
Release 6	3122	3122 3125	2533	2533 2537	6428	6428 6514		

BALANCE PARAMETERS								
UM			LDM			ZM		
Assembly (Y chr seq)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)
Chang & Larracuent	7675	7675 7697 7795	6090	6090 6112 6206	18185	18185 18429 18335		
Illumina	6734	6734 6947	5357	5357 5521	7206	7206 7497		
Falcon	6426	6426 6426	4941	4941 4941	16383	16383 16383		
Sanger	8422	8423 8530	6436	6437 6529	11063	11064 11416		
Release 6	7101	7101 7109	5561	5561 5568	13316	13316 13491		

COVERAGE PARAMETERS								
UM			LDM			ZM		
Assembly (Y chr seq)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)	FR Filtered Common	Total ($k = 15 \mid k = 18$)
Chang & Larracuent	11542	11542 11593 11769	9372	9372 9423 9588	25991	25991 26323 26248		
Illumina	9800	9805 10132	8068	8069 8331	10608	10613 11062		
Falcon	10007	10007 10007	8022	8022 8022	22883	22883 22883		
Sanger	12437	12441 12669	9788	9781 9969	16132	16136 16759		
Release 6	10415	10415 10436	8337	8337 8352	19007	19007 19264		

Tabela 4.11: Quantidade de sondas filtradas contra *short reads* de fêmea – sobreposição permitida

Nota da tabela: Para cada subgrupo de sequências inferidas para o cromossomo Y são apresentadas as quantidades de sondas obtidas após a filtragem feita a partir dos resultados de alinhamento NGS contra os *short-reads* de fêmea com bowtie2 (*FR Filtered*). Essa filtragem foi feita a partir do programa `outputClean.py -0`. Para cada modo (UM, LDM e ZM, respectivamente) são apresentados os valores de sondas para os subgrupos de *contigs* em comum do cromossomo Y, seguido do total para cada conjunto de sequências inferidas para o cromossomo Y a partir de um valor de k diferente. Para o cromossomo Y montado por Chang e Larracuent (2019) os valores exclusivos incluem, por último, os de seu subgrupo exclusivo. Os blocos na tabela correspondem aos resultados a partir de diferentes parâmetros, detalhados na Tabela 2.3.

ZM, as sondas filtradas a partir do UM deveriam estar todas presentes entre as sondas do ZM e majoritariamente para aquelas filtradas através do LDM (Figura 4.9) – já que o ZM mantém sondas exclusivas do cromossomo alvo, seja com um único alinhamento (portanto, todas do UM) ou múltiplos (exclusivamente selecionadas com ZM). Ou seja, apenas as análises do LDM deveriam retornar algumas sondas exclusivas com a execução dos comandos a seguir, dado que o ZM não conta com análise linear discriminante para reavaliar a predição do número de alvos. O arquivo com as sondas do ZM utilizado é aquele gerado pelo `outputClean.py`, enquanto o arquivo final de sondas para os outros dois modos é utilizado nessa análise – já

que a abordagem que mantém as sondas que tiveram zero alinhamentos passa por filtragens a mais e, assim, se uma sonda tiver sido filtrada nesta etapa, mas removida posteriormente, ainda deve ser considerado que essa abordagem a selecionou em um primeiro momento.

```

1 ## Sondas exclusivas do UM
2 awk 'FILENAME != ARGV[3] { m[$4,$4] = 1; next}; !(($4,$4) in m)' <↔>
    ZM_oCprobesWithOverlapON.bed LDM_finalProbesWithOverlapON.bed <↔>
    UM_finalProbesWithOverlapON.bed > UM_exclusiveFinalProbesOverlapON.bed
3 ## Sondas exclusivas do LDM
4 awk 'FILENAME != ARGV[3] { m[$4,$4] = 1; next}; !(($4,$4) in m)' <↔>
    ZM_oCprobesWithOverlapON.bed UM_finalProbesWithOverlapON.bed <↔>
    LDM_finalProbesWithOverlapON.bed > <↔>
    LDM_exclusivefinalProbesOverlapON.bed

```

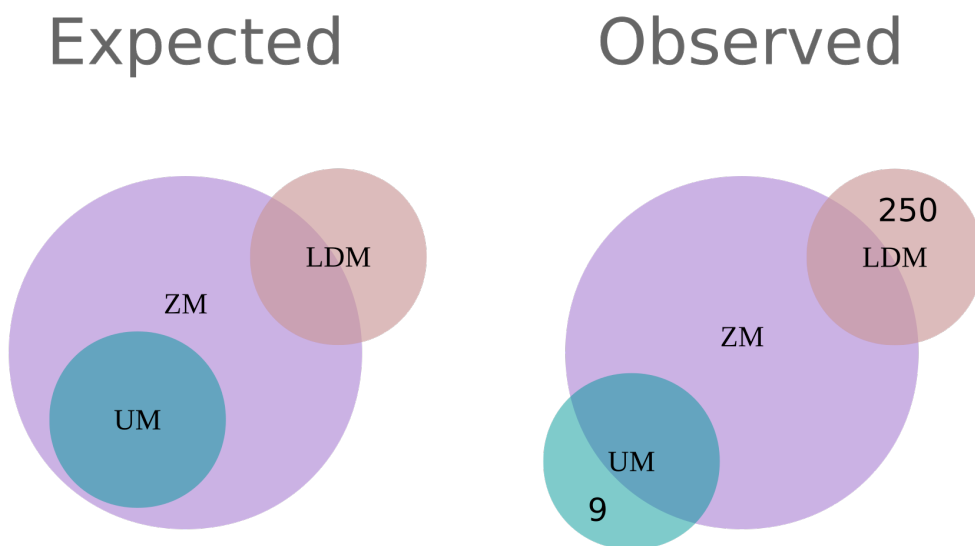


Figura 4.9: Resultados esperados e observados na comparação dos modos do OligoMiner

Nota da figura: São apresentados à esquerda os resultados esperados (*Expected*) e à direita (*Observed*). Era esperado que as sondas filtradas através do ZM englobassem todas as filtradas com UM e parte dos resultados do LDM; entretanto, algumas sondas foram unicamente selecionadas através do UM. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

Como resultado dessa análise, foram obtidas para o UM, somando as sondas exclusivas a partir dos três conjuntos de parâmetros e subgrupos do cromossomo Y (Tabela 4.1), nove sondas que não são filtradas com o ZM. Esse resultado difere do inicialmente esperado, já

que sondas com um único alinhamento no cromossomo em análise não deveriam apresentar nenhum alinhamento no genoma sem essa sequência de entrada. Ao buscar a fonte de tais sondas serem exclusivas do UM, foi observado que quando essas mesmas nove sequências são alinhadas contra um *index* do genoma sem Y, provavelmente por a sequência não estar presente de forma alguma neste *index*, o bowtie2 (LANGMEAD; SALZBERG, 2012) faz o alinhamento desta contra o complemento reverso do genoma e, por ser encontrado um alinhamento, estas sondas não são filtradas com o ZM. Apesar do oligonucleotídeo sintetizado que tomará parte no experimento de FISH *Oligopaint* ter sequência idêntica à da sonda desenhada, em teoria essas nove sondas não precisariam ser descartadas pelo ZM, podendo ser resgatadas posteriormente e adicionadas ao conjunto de sondas selecionadas. Entretanto, cabe ao próprio usuário definir qual a fita alvo: se é a complementar à região montada ou a própria região montada. Por essa razão, não vemos problemas em continuar descartando tais sondas e mantendo a simplicidade da análise.

Para o LDM também aparecem algumas sondas exclusivas (sempre menos de 250 sondas) e elas não se encontram em todos os subgrupos analisados. Dada a natureza da seleção por análise linear discriminante com simulação do número previsto de alvos, já era esperado que os resultados desse modo apresentassem algumas exclusividades. No entanto, pela alta repetitividade do cromossomo Y e pelos promissores resultados da nova abordagem proposta, essas sondas não foram escolhidas para o experimento de validação. Embora idealmente tal experimento contaria com todas as diferentes análises aqui realizadas, em prática determinados cortes foram realizados devido ao custo de síntese de oligonucleotídeos e orçamento reservado para tanto. Essas questões são melhor descritas no Capítulo 5.

Vale ressaltar que, embora seja possível comparar as sondas obtidas através dos diferentes modos, tal comparação não seria possível com os diferentes parâmetros – muito embora trechos de sondas menores em tamanho de nucleotídeos estejam contidos em sondas mais compridas. Entretanto, essa redundância não é compatível devido às temperaturas envolvidas nos experimentos de FISH para desnaturação da fita de DNA fixada e hibridização do oligonucleotídeo fluorescentemente marcado. E esses valores foram utilizados justamente para desenhar e selecionar as sondas.

Para finalizar as discussões pertinentes ao desenho de sondas *oligopaint* para o cromossomo Y, é interesse discutir a densidade das sondas finais, uma medida considerada importante para os experimentos de FISH *Oligopaint* justamente porque quanto maior a densidade, maior a cobertura de sondas por região do cromossomo. Os valores ideais de densidade de alvos/kb, entretanto, variam principalmente com o tamanho da área a ser marcada – se o cromossomo tem mais de 100 Mb, a densidade média de 0.25 alvos/kb foi apontada como suficiente para cobrir bem toda a área cromossômica (ALBERT *et al.*, 2019).

Beliveau *et al.* (2018) mostraram, entretanto, que quanto menor é a região alvo, maior precisa ser a densidade de alvos/kb, chegando à 20 alvos/kb para uma região curta de 5 kb, embora para uma região de 4 Mb 1.5 alvos/kb já sejam suficientes para marcar a mesma eficientemente.

Para o cromossomo Y, no entanto, por sua montagem ser altamente fragmentada, as medidas de densidade não poderiam ser precisas: em prática, as sondas desenhadas deverão hibridizar em uma região de tamanho estimado em 40.9 Mb (ADAMS, 2000), embora as sequências montadas utilizadas nunca passem de 14 Mb, sendo o tamanho mínimo de 1.08 Mb (Tabela 4.1). Além disso, os valores de sondas utilizados para obter as densidades apresentadas (Figura 4.10 e Tabela 4.12) correspondem às sondas após remoção de sobreposição, já que existem algumas sondas que alinham mais de uma vez no próprio cromossomo Y. Para obter valores que fossem condizentes com a filtragem que remove aquelas sondas que alinham contra o *index* dos *short-reads* de fêmea, essa última etapa de filtragem foi executada para cada um dos diferentes subgrupos a partir do arquivo gerado após remoção de sondas que tivessem sobreposição. Esses valores foram então utilizados para calcular o número de alvos/kb, sendo apresentada tanto a densidade para o valor estimado do cromossomo Y (40.9 Mb) quanto para o correspondente tamanho do conjunto de sequências inferidas que foi efetivamente utilizado.

Por fim, é possível observar que a densidade varia extremamente ao utilizar os dois diferentes tamanhos de cromossomo (estimado ou inferido), entre 0.09 e 1.19 alvos/40900 kb e entre 0.86 e 9.89 alvos/tamanho específico do cromossomo em questão em kb (Tabela 4.12). Nesse último caso é bom ressaltar que existe um viés criado pelo cromossomo inferido a partir da montagem com sequenciamento Illumina: a relação, principalmente a partir dos parâmetros mais permissivos, entre o menor conteúdo nucleotídico altamente repetitivo (1.08 Mb) e o menor número de sondas geradas, resulta em uma densidade aparentemente tão superior às dos demais conjuntos de sondas/tamanho do alvo montado. Quando a densidade desse cromossomo é obtida em relação ao tamanho estimado de 40.9 Mb é observado que de fato o conjunto de sequências inferidas para o cromossomo Y a partir de Illumina resulta nas menores densidades, com mínimo de 0.09 alvos/40900 kb.

Além de Illumina, a montagem com as maiores densidades em relação ao tamanho da sequência utilizada para desenho das sondas foi obtida a partir do conjunto de sequências inferidas para o cromossomo Y com a montagem Falcon, chegando à 4.63 alvos/8804 kb – para o cromossomo estimado em 40.9 Mb, entretanto, a densidade cai para 1 alvo/40900 kb, se aproximando mais dos demais valores.

A densidade aumenta conforme a permissividade dos parâmetros, mas mesmo para os mais restritivos ela se aproxima, em média, de 0.24 alvos/40900 kb. Essa é uma boa estimativa da provável eficiência dos experimentos de validação, dado que esse é o tamanho estimado do cromossomo Y – justamente a região em que as sondas irão hibridizar, muito

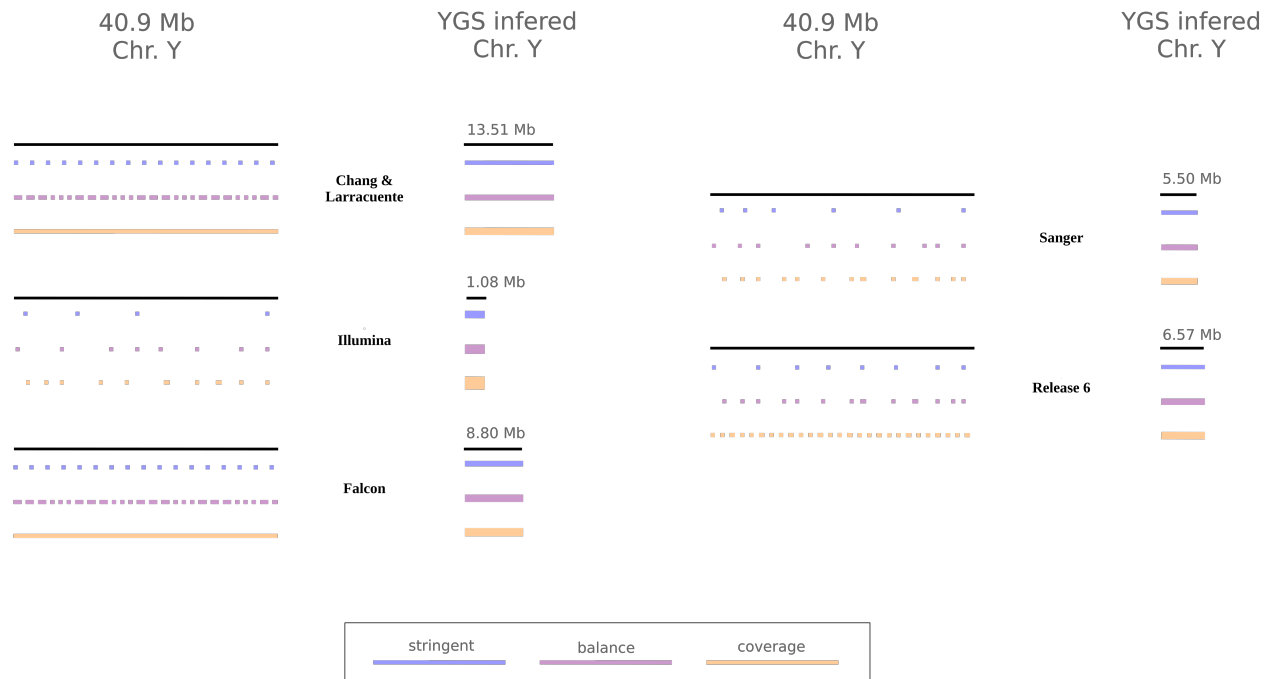


Figura 4.10: Estimativa da densidade de alvos/kb dos subgrupos comuns de sondas desenhadas

Nota da figura: Representando os experimentos de fluorescência, o esquema demonstra as densidades das sondas esperadas no tamanho estimado do cromossomo Y (40900 kb) e respectivo tamanho do cromossomo de acordo com o subgrupo para o qual as sondas foram desenhadas Tabela 4.1). Notas: (i) Apenas sondas desenhadas para os *contigs* selecionados por ambos os valores de k ($k = 15$ e $k = 18$) estão esquematizadas; (ii) Devido ao tamanho montado do cromossomo inferido a partir do sequenciamento Illumina ser muito menor em relação aos demais, sua escala está três vezes maior para fins de melhor visualização. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

embora tenham sido desenhadas a partir de uma sequência montada pelo menos três vezes menor. Vale ressaltar que todos valores de densidade apresentados são apenas estimativas, podendo ser inclusive muito maiores, dada a alta repetitividade de sequências do cromossomo Y e, assim, ao possível maior número de alvos.

Em síntese, a *pipeline* aqui desenvolvida permitiu desenhar um total de sondas *oligopaint* que, grande parte das vezes, simboliza boas estimativas do número de alvos/kb, sendo essa uma provável vantagem para a eficiência dos experimentos de FISH. Outra grande vantagem toca justamente no fato de que as sondas finais sintetizadas possam hibridizar mais de uma vez no cromossomo, o que foi considerado nos cálculos de densidade/kb cromossômico anteriormente apresentados. Assim, além de aumentar a expectativa do número de alvos, pela alta repetitividade das sequências do cromossomo Y, o número real de alvos que as sondas terão pode ser ainda maior. Ademais, o valor comercial para síntese da biblioteca de oligonucleotídeos é reduzido, dado que é desnecessário sintetizar uma mesma sonda múltiplas

STRINGENT PARAMETERS									
Assembly (Y chr seq)	Hits	Common		$k = 15$			$k = 18$		
		40.9 Mb	Chr. size	Hits	40.9 Mb	Chr. size	Hits	40.9 Mb	Chr. size
Chang & Larracunte	16426	0.40	1.22	16426	0.40	1.22	16616	0.41	1.21
Illumina	3519	0.09	3.26	3519	0.09	3.26	3647	0.09	2.98
Falcon	14889	0.36	1.69	14889	0.36	1.69	14889	0.36	1.69
Sanger	5153	0.13	0.94	5153	0.13	0.94	5289	0.13	0.86
Release 6	8123	0.20	1.24	8123	0.20	1.24	8210	0.20	1.23

BALANCE PARAMETERS									
Assembly (Y chr seq)	Hits	Common		$k = 15$			$k = 18$		
		40.9 Mb	Chr. size	Hits	40.9 Mb	Chr. size	Hits	40.9 Mb	Chr. size
Chang & Larracunte	33676	0.82	2.49	33676	0.82	2.49	33975	0.83	2.46
Illumina	7241	0.18	6.71	7241	0.18	6.71	7532	0.18	6.16
Falcon	28955	0.71	3.29	28955	0.71	3.29	28955	0.71	3.29
Sanger	11573	0.28	2.10	11574	0.28	2.10	11933	0.29	1.93
Release 6	16592	0.41	2.53	16592	0.41	2.53	16768	0.41	2.52

COVERAGE PARAMETERS									
Assembly (Y chr seq)	Hits	Common		$k = 15$			$k = 18$		
		40.9 Mb	Chr. size	Hits	40.9 Mb	Chr. size	Hits	40.9 Mb	Chr. size
Chang & Larracunte	48085	1.18	3.56	48085	1.18	3.56	48497	1.19	3.52
Illumina	10666	0.26	9.89	10671	0.26	9.89	11123	0.27	9.10
Falcon	40727	1.00	4.63	40727	1.00	4.63	40727	1.00	4.63
Sanger	16897	0.41	3.07	16901	0.41	3.07	17538	0.43	2.84
Release 6	23816	0.58	3.63	23816	0.58	3.63	24075	0.59	3.62

Tabela 4.12: Densidades de alvos/kb das sondas desenhadas com sobreposição e filtradas através do ZM

Nota da tabela: Valores obtidos com o número de alvos (*Hits* - obtido após remoção de sobreposição e filtro contra *short-reads* de fêmea) dividido pelo tamanho do cromossomo em kb, sendo: 40900 kb para o tamanho estimado do cromossomo Y e respectivo tamanho do cromossomo de acordo com o subgrupo para o qual as sondas foram desenhadas (*Chr. size*; Tabela 4.1). São apresentados as densidades para os subgrupos “comuns” às análises do YGS (*Common*) e as densidades com o total de sondas finais para as análises com $k = 15$ e $k = 18$, usando então o tamanho total desses cromossomos específicos (Tabela 3.3). Em especial para Chang e Larracunte (2019), o subgrupo original, que não aparece na tabela, resultou em: (0.40 alvos/kb, 1.13 alvos/kb), (0.83 alvos/kb, 2.31 alvos/kb) e (1.18 alvos/kb e 3.30 alvos/kb) indo de *stringent* para *coverage* com cromossomo de 40.9 Mb e 14.65 Mb, respectivamente.

vezes, e que sondas com múltiplos alvos marcarão o cromossomo de interesse em várias regiões. Além disso, quando comparado com a quantidade total de sondas que já existiam para o cromossomo Y de *Drosophila melanogaster* antes desta pesquisa (BELIVEAU *et al.*, 2018) (Figura 4.11), se observa que todas as abordagens exploradas proporcionaram um enorme ganho de sondas – o ZM em especial, e suas estimativas de número total de alvos, tendo sido um destaque.

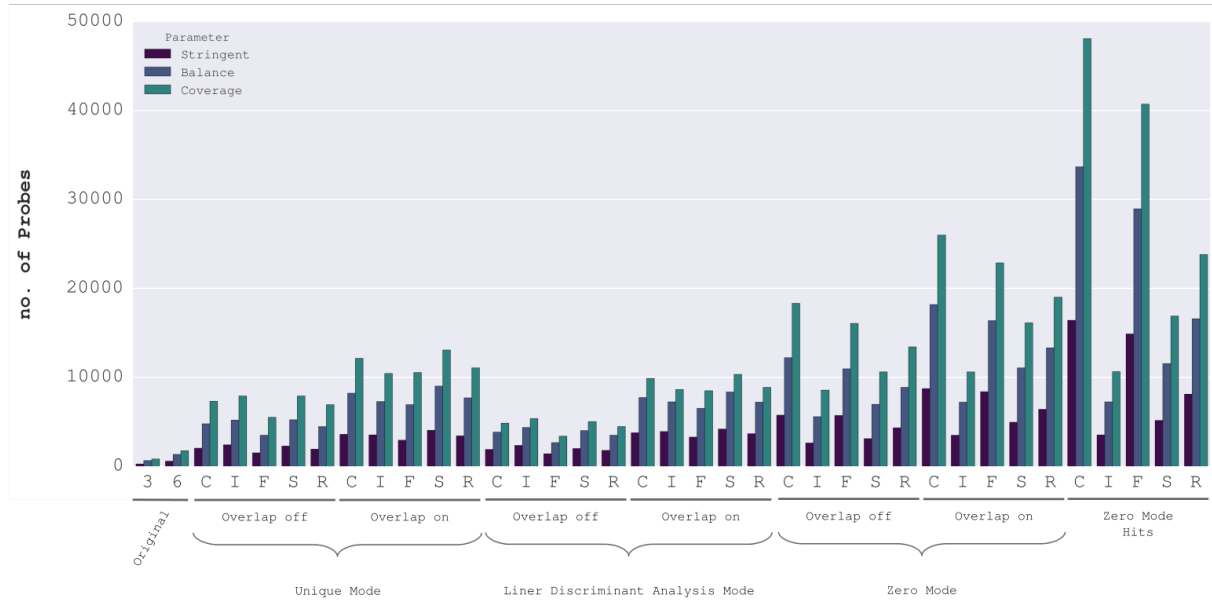


Figura 4.11: Comparação do número final de sondas *oligopaint* obtidas com diferentes abordagens

Nota da figura: São apresentadas as quantidades de sondas (ordenada) finais com os parâmetros *stringent*, *balance* e *coverage*. Da esquerda para a direita: *original* representam as sondas desenhadas por Beliveau *et al.* (2018), sendo 3 e 6 respectivamente as versões dm3 e dm6 do genoma de *Drosophila melanogaster*; sondas filtradas através da abordagem clássica do OligoMiner com UM sem (*Overlap off*) e com sobreposição (*Overlap on*) de candidatas; sondas filtradas através da abordagem clássica do OligoMiner com LDM sem (*Overlap off*) e com sobreposição (*Overlap on*) de candidatas; sondas filtradas através da nova abordagem do OligoMiner com ZM sem (*Overlap off*) e com sobreposição (*Overlap on*) de candidatas; e, por fim, o número estimado de alvos (*hits*) a partir das sondas do ZM desenhadas com sobreposição de sondas candidatas. C I F S R representam, respectivamente, as montagens *Chang & Larracuenta*, *Ilumina*, *Falcon*, *Sanger*, *Release 6*. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

5

Biblioteca final e Validação *in situ* da *pipeline*

Em posse do conjunto de sondas finais as sequências podem ser preparadas para constituir uma biblioteca que será sintetizada, seja *in house* ou através de um comerciante. Esse preparo é fundamental para a transição *in silico* – *in situ*, permitindo que sondas sintetizadas passem pelo processo de amplificação e sejam propriamente utilizadas em ensaios de FISH *Oligopaint*.

Sucintamente, esse capítulo é dividido em três seções, sendo a primeira delas dedicada à adição de regiões de *priming*, um processo feito *in silico* (Seção 5.1). Posteriormente, são detalhados os controles de qualidade empregados no protocolo de amplificação da biblioteca sintetizada (Seção 5.2), seguido dos resultados obtidos a partir dos ensaios de validação com FISH *Oligopaint* (Seção 5.3). É importante ressaltar que os resultados obtidos nas duas últimas seções tiveram participação crucial de Henry Bonilla, Amanda Luvisotto e Mara Pinheiro, sendo os dois primeiros alunos sob orientação da Professora Dra. Maria Vibranovski, seguidos pela Técnica em Citogenética do Laboratório. Bonilla foi responsável pela padronização do protocolo de amplificação, enquanto Bonilla, Pinheiro e Luvisotto conduziram os experimentos de FISH *Oligopaint* e microscopia.

5.1 Procedimentos pré-síntese da Biblioteca Final

O preparo das sequências alvo selecionadas, até então mencionadas como *sondas* ou *sondas oligopaint*, é caracterizado pela adição de regiões de *priming*. Tais regiões são importantes para a amplificação e seleção de subgrupos de oligonucleotídeos (alvo + regiões de *priming*) em uma biblioteca complexa. Além disso, as regiões de *priming* proporcionam adicionar o fluoróforo que permitirá observar a marcação cromossômica em ensaios de citogenética. Dado que existem diferentes protocolos de amplificação e necessidades particulares de cada experimento, todos esses quesitos têm de ser considerados antes da síntese da biblioteca final.

Comercialmente, são sintetizadas apenas uma cópia de cada fita de oligonucleotídeo e mais de um grupo de sondas pode compor a mesma biblioteca (Figura 5.1a). Vale frisar que cada biblioteca, independente de possuir um ou mais grupos de sondas, será sintetizada em um único tubo de reagente. Por esse motivo, a região de *priming* permitirá a diferenciação dos mesmos, proporcionando amplificar e empregar cada um separadamente em variados ensaios de FISH *Oligopaint* (Figura 5.1b_i).

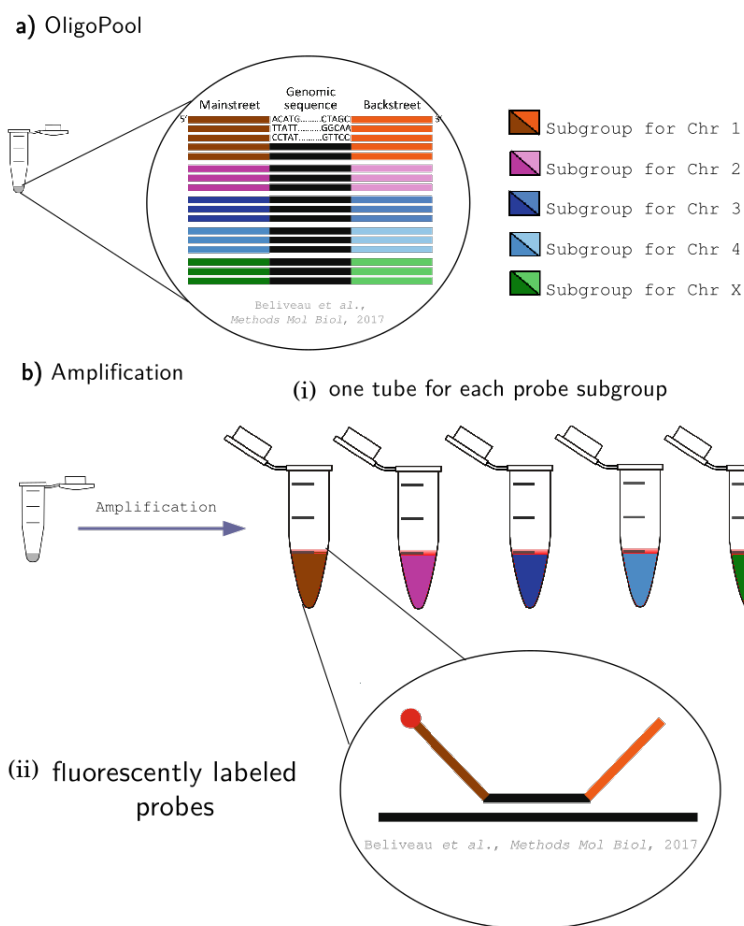


Figura 5.1: Importância das regiões de *priming*

Nota da figura: a) Em uma biblioteca complexa de sondas *oligopaint* (*Oligo Pool*), as regiões de *priming* são importantes durante b) o processo de amplificação das mesmas, permitindo (i) amplificar separadamente cada um dos subgrupos de sondas presentes na biblioteca, aqui esquematizado com diferentes cores para *Mainstreet* (*priming forward*) e *Backstreet* (*priming reverso*) para subgrupos de sondas desenhadas para cromossomos 1, 2, 3, 4 e X ; e (ii) para acoplar fluoróforo nas sondas amplificadas, permitindo identificar fluorescentemente as regiões hibridizadas nos experimentos de FISH *Oligopaint*. Vale ressaltar que a intensidade do sinal observado pode ter diferentes níveis, variando, por exemplo, se obtido somente a partir de anticorpo secundário, com dupla coloração ou dupla marcação ou mesmo mais fluoróforos de forma a personalizar ainda mais o sinal observado. (Adaptado de Beliveau *et al.* (2017)).

A segunda importância da região de *priming* é exatamente a adição de fluoróforo à sonda de interesse (Figura 5.1b_{ii}). O fluoróforo é a molécula química responsável por criar uma emissão fluorescente em um espectro de luz visível, efetivamente permitindo observar a marcação em experimentos de FISH *Oligopaint*. As diferentes maneiras de adicionar essa molécula às sondas podem levar ao aumento da intensidade do sinal observado e influenciar na escolha do protocolo de amplificação utilizado.

Conforme mencionado no Capítulo 2 (Seção 2.3, Subseção 2.3.1), dada a complexidade da biblioteca sintetizada, nesta pesquisa as sondas foram preparadas para serem amplificadas através do protocolo IVT-RT descrito por Beliveau *et al.* (2017), ou seja, os oligonucleotídeos devem conter duas regiões de *priming*. A seguir são descritas as polaridades e peculiaridades de cada uma dessas regiões e de seus respectivos *primers*, os reagentes sintetizados à parte para serem utilizados no protocolo de amplificação:

- *priming forward* – sequência com polaridade 5'-3', sendo o *primer forward* um oligonucleotídeo sintetizado exatamente a partir da mesma sequência, portanto, que terá a mesma polaridade;
- *priming* reverso – sequência com polaridade 5'-3', sendo o *primer* reverso um oligonucleotídeo sintetizado a partir de sua sequência complementar reversa, começando com promotor T7 (T7 + complemento reverso da região de *priming* reverso) também com polaridade 5'-3';
- *priming* da RT – sequência com polaridade 5'-3', sendo o *primer* da RT um oligonucleotídeo sintetizado exatamente a partir da mesma sequência, portanto, que terá a mesma polaridade;

5.1.1 Triagens da Biblioteca Final

Para realização de tais experimentos *in situ* de modo a validar a *pipeline in silico* proposta, seriam idealmente sintetizadas e empregadas nos diferentes experimentos de FISH as sondas que compõe todas as análises realizadas por este trabalho (Capítulo 4). Isso possibilitaria observar visualmente, por exemplo, quais as regiões extras que as sondas desenhadas exclusivamente a partir de um valor de k marcam, inclusive avaliando se a escolha de $k = 18$ para inferir sequências para o cromossomo Y leva de fato a marcação mais eficiente deste em detrimento de $k = 15$; ou mesmo se as sequências em comum (Tabela 4.1) já são suficientes para a marcação eficaz. Além disso, esses resultados poderiam apontar diferenças observadas entre as tecnologias de sequenciamento utilizadas para gerar os diferentes genomas de entrada (Tabela 2.1). Em suma, isso mostraria tanto formas de aprimorar a *pipeline*

conforme a tecnologia de sequenciamento empregada para gerar o genoma disponível, como também quais dessas resultariam nas marcações mais e menos eficientes do cromossomo Y.

Na prática, porém, as sondas finais não poderiam ultrapassar o total de 91,766 oligonucleotídeos, correspondendo ao maior *chip* comercial com melhor custo benefício através do comerciante Genscript, tendo em consideração o orçamento reservado para os experimentos de validação *in situ*. Nessa empresa, existe ainda uma variação no preço do *chip* mencionado conforme o tamanho do maior oligonucleotídeo presente na biblioteca, indo de USD \$4K à USD \$6K (Para mais detalhes: <<https://www.genscript.com/precise-synthetic-oligo-pools.html>>).

Dessa maneira, houve a necessidade de escolher alguns conjuntos de sondas em detrimento de outros. A primeira seleção concerne as sondas desenhadas permitindo ou não a sobreposição entre candidatos: dado que os resultados com sobreposição foram mais promissores e englobaram todas as sondas desenhadas sem ativar essa opção, os experimentos de validação com esse último grupo não trariam resultados diferentes dos já observados.

Em seguida, foi realizada uma seleção entre os conjuntos de parâmetros usados para desenhar as sondas através do OligoMiner: *stringent*, *balance* e *coverage* (Tabela 2.3). Como os dois primeiros possuem temperaturas e tamanhos comumente empregados em ensaios de FISH, os grupos de sondas desenhados a partir dos parâmetros mais permissivos (*coverage*) não foram sintetizados.

A terceira seleção de sondas para compor a biblioteca final foi com relação aos diferentes modos de predição do número de alvos do OligoMiner. Considerando que a maior parte das sondas desenhadas através dos modos UM e LDM foram também selecionadas através do ZM, foi decidido que não seriam realizados experimentos de validação separadamente para cada um dos modos de filtragem do OligoMiner. Visto que a abordagem com o ZM permitiu elevar em muito as quantidades de sondas finais em relação aos outros dois modos, não se acredita haver prejuízo nessa escolha (Tabela 4.11 e Tabela 4.12).

Com relação aos tamanhos de *k-mers* utilizados para inferir sequências para o cromossomo Y através do YGS, $k = 15$ e $k = 18$, as sondas desenhadas a partir dos *contigs* inferidos com ambos os valores de k (Subgrupo *Common*) foram selecionadas para compor a biblioteca final. Vale ressaltar que a quantidade total de sondas exclusivas de cada um dos valores de k para as diferentes montagens não chega a 4K, indicando que experimentos com as mesmas provavelmente não iriam mostrar contrastes sólidos entre estes dois valores.

Assim, um total de 98,182 sondas foram selecionadas para compor a biblioteca final. Contudo, o *chip* comercial da Genscript com melhor custo benefício tem espaço para apenas 91,766 oligonucleotídeos. Em paralelo, muitas sondas foram desenhadas de maneira idêntica para mais de um dos conjuntos de sequências inferidas para o cromossomo Y utilizado. Por exemplo, existem sondas finais selecionadas para ambos os conjuntos de sequências do

cromossomo Y inferidos com as montagens Falcon e Sanger. O diagrama de Venn a seguir demonstra as quantidades finais de sondas compartilhadas entre as diferentes montagens (Figura 5.2).

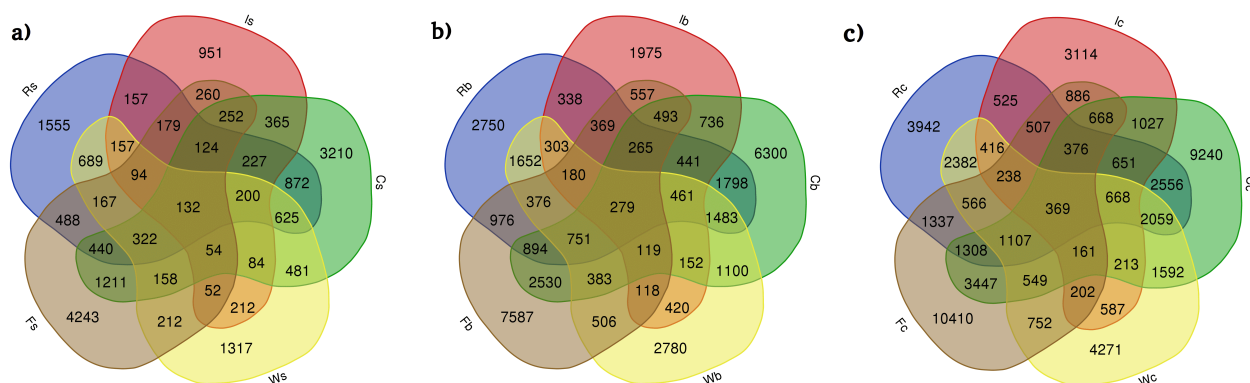


Figura 5.2: Sondas compartilhadas entre diferentes montagens

Nota da figura: São apresentadas as quantidades de sondas finais compartilhadas entre as diferentes montagens, sendo: *I* – Illumina; *C* – Chang & Larracuente; *W* – Sanger; *R* – Falcon; *R* – Release 6; s/b/c corresponde, respectivamente, às sondas desenhadas e filtradas com os parâmetros a) *stringent*; b) *balance* e c) *coverage* (Tabela 2.3). Esses diagramas de Venn foram produzidos através da ferramenta desenvolvida pelo grupo do prof. Van de Peer na (Disponível em <<http://bioinformatics.psb.ugent.be/webtools/Venn/>>), por Isabela Almeida durante o desenvolvimento desta pesquisa.

Uma maneira de aproveitar esses compartilhamentos é usar uma região de *priming* para cada uma das “áreas do diagrama de Venn”, o que é potencialmente problemático conforme aumenta a complexidade da biblioteca – nesse caso, seriam necessários 31 diferentes sequências somente para a região de *priming forward*. Outra forma é adicionar, em uma mesma região de *priming*, o correspondente a duas: assim, para sondas compartilhadas entre dois grupos de sondas, como o exemplo de Falcon e Sanger, a região de *priming* deve conter ambas as sequências de *priming* escolhidas para amplificar individualmente cada um desses grupos.

Todavia, a adição das cinco regiões de *priming* nas sondas compartilhadas por todos os grupos (centro dos diagramas de Venn na Figura 5.2) levaria à variação de mais de 10% em tamanho na mesma biblioteca, já que a maior parte das sondas teria menos regiões de *priming*. Essa variação não é permitida pela Genscript de maneira a garantir a eficiência da síntese dos oligonucleotídeos.

Além disso, o posicionamento das regiões de *priming* determina a eficiência da amplificação por PCR em Tempo Real: oligonucleotídeos que fossem exclusivos de cada montagem, teriam apenas uma região de *priming* e seriam muito mais curtos do que os

compartilhados por cinco grupos. Isso determina que esses grupos de oligonucleotídeos menores amplificariam muito mais, causando desequilíbrio dos produtos finais da PCR em Tempo Real. Mesmo com a normalização do tamanho dos oligonucleotídeos pela adição de bases nucleotídicas aleatórias no terminal 3', a posição relativa da região de *priming* que irá formar os menores oligonucleotídeos levaria ao desequilíbrio.

Considerando esses aspectos, foram aproveitados os compartilhamentos apenas entre duas montagens – de maneira que os conjuntos de sondas sintetizados foram preparados contendo de uma a duas regiões de *priming forward* (Figura 5.3). Ainda assim, os tamanhos mínimo e máximo de oligonucleotídeos ainda variavam em mais de 10%, entre 102 e 130 nt: mas, por não corresponder a um espectro tão amplo, a normalização desse tamanho não é problemática. Essa normalização, realizada pela empresa Genscript, é conhecida como *padding the sequence* e corresponde à adição de bases nucleotídicas aleatórias no terminal 3'.

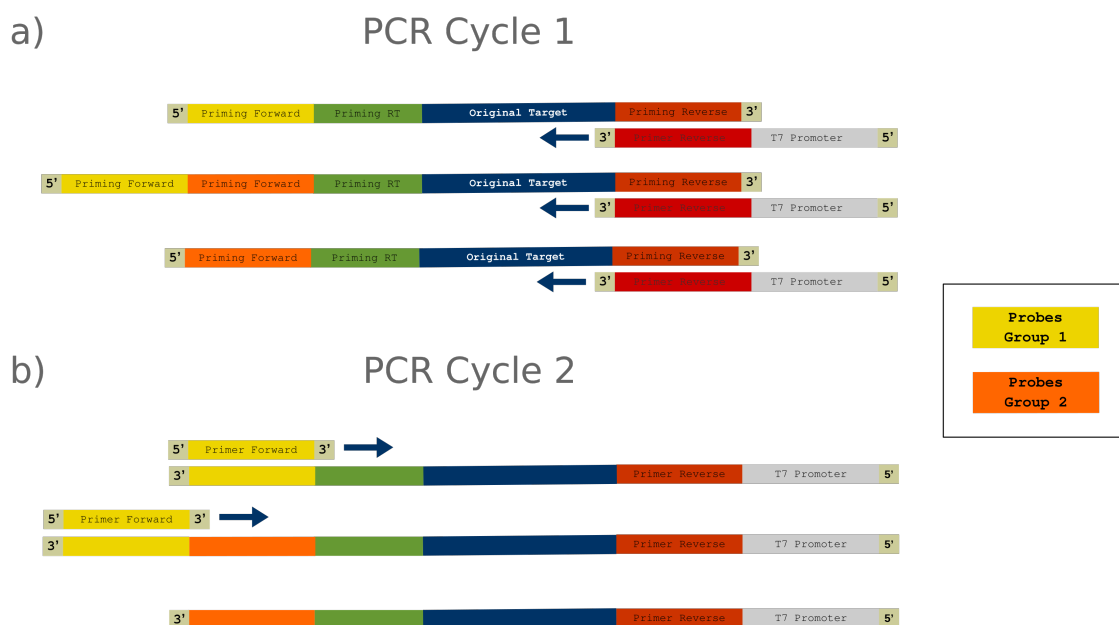


Figura 5.3: Amplificação de grupos de sondas em biblioteca com até duas regiões de *priming forward* em um oligonucleotídeo

Nota da figura: Nesse esquema simplificado, é apresentado o momento no protocolo de amplificação em que ocorre a seleção inicial de grupos de sondas em uma mesma biblioteca de oligonucleotídeos. Seguindo a lógica de adição de regiões de *priming* é possível observar em a) o ciclo de PCR em Tempo Real 1 amplificando todos os grupos de sondas com o anelamento do *primer forward* + promotor da Enzima T7 RNA Polimerase; em b) são amplificadas efetivamente apenas as sondas que possuem a região de *priming forward* do grupo de interesse (em amarelo – *Probes group 1*), independente de o grupo de sondas possuir apenas uma região de *priming* ou duas. Esse esquema segue a lógica apresentada na Figura 2.8. Figura produzida por Isabela Almeida durante o desenvolvimento desta pesquisa.

5.1.2 Escolha e Adição de regiões de *priming* aos grupos de sondas

Algumas precauções devem ser tomadas ao escolher as regiões de *priming* que irão compor os oligonucleotídeos sintetizados. As recomendações e valores de similaridade a seguir seguem as orientações de diferentes trabalhos (BOETTIGER *et al.*, 2016; CHEN *et al.*, 2015; MOFFITT; ZHUANG, 2016), sendo elas:

- alinhar as regiões de *priming* umas contra as outras, evitando que apresentem similaridade maior ou igual a 10 bP;
- alinhar as regiões de *priming* contra o terminal 3' do promotor da T7, eliminando aquelas que apresentem similaridade maior ou igual a 5 bP;
- remover a guanina do terminal 5' da região de *priming* reverso, evitando a criação de um quádruplo G com terminal GGG do promotor T7 no *primer* reverso sintetizado;
- alinhar as regiões de *priming* contra as sondas selecionadas, removendo qualquer uma que apresente similaridade maior ou igual a 12 bP para as regiões de *priming* da PCR em Tempo Real e aquelas com similaridade maior ou igual a 20 bP para as da RT;
- após atentar para os passos anteriores que funcionam como uma seleção para encontrar regiões de *priming*, adicionar as mesmas às sondas
- em posse das *sondas com as regiões de priming*, alinhá-las contra as regiões individuais de *priming*, removendo qualquer uma que apresente um novo sítio de similaridade maior ou igual a 12 bP (PCR em Tempo Real) e a 20 bP (RT) – nesse caso, o processo deve ser feito novamente com uma região de *priming* alternativa.

A escolha das sequências que iriam compor as regiões de *priming* foi baseada em trabalhos anteriores (BELIVEAU *et al.*, 2015; CHEN *et al.*, 2015) e nos alinhamentos locais realizados através de BLASTn (ALTSCHUL *et al.*, 1990) seguindo a linha de comando descrita a seguir.

```
1 blastn -query seqsQuery.fasta -subject seqSubject.fasta -outfmt '6 ↔)
    qseqid sseqid pident length mismatch gapopen qstart qend sstart send ↔)
    evalue bitscore' -word_size $similarity > blast
```

Em especial para a sequência de *priming* reverso, inicialmente não se havia atentado para a remoção da base nucleotídica guanina no terminal 3' da mesma – de forma a evitar a formação de um quádruplo G entre o terminal 3' GGG do promotor T7 e o início da região complementar reversa do *priming* reverso – como orientado por Moffitt e Zhuang (2016). Por esse motivo, embora as regiões de *priming* reverso de todas as sondas sejam compostas pela sequência ATGCGCCAATTCCGGTTC (Tabela 5.1), o *primer* sintetizado para anelar nessa região durante a PCR em Tempo Real não possui a guanina correspondente no terminal 3'. Dessa maneira, o anelamento só acontece a partir da sequência T terminal 3' e não C terminal 3' (Tabela 5.2). Além de não interferir no anelamento do *primer* reverso à região de *priming* reverso, essa precaução pode aumentar a eficiência do experimento de IVT (MOFFITT; ZHUANG, 2016), embora Beliveau *et al.* (2018) utilizaram *primer* reverso com formação de quádruplo G entre o terminal 3' GGG do promotor T7 e o início da região complementar reversa à de *priming* reverso.

STRINGENT PARAMETERS				
Library subset	priming forward (1)	priming forward (2)	RT priming	priming reverse
Chang & Larracuente (non YGS)	ACAAATCCGACCATCG		CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Chang & Larracuente exclusive	CAGTGCTCGTGTGAGAAG		CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Chang & Larracuente ∩ Falcon	CAGTGCTCGTGTGAGAAG	CATCGGCCACGGTCCCGT	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Chang & Larracuente ∩ Sanger	CAGTGCTCGTGTGAGAAG	GACTGGTACTCGCGTGAC	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Illumina exclusive	GCCCGTATTCCCGCTTGC		CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Illumina exclusive ∩ Chang & Larracuente	GCCCGTATTCCCGCTTGC	CAGTGCTCGTGTGAGAAG	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Illumina exclusive ∩ Falcon	GCCCGTATTCCCGCTTGC	CATCGGCCACGGTCCCGT	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Illumina exclusive ∩ Sanger	GCCCGTATTCCCGCTTGC	GACTGGTACTCGCGTGAC	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Falcon exclusive	CATCGGCCACGGTCCCGT		CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Sanger exclusive	GACTGGTACTCGCGTGAC		CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Sanger ∩ Falcon	GACTGGTACTCGCGTGAC	CATCGGCCACGGTCCCGT	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Release 6 exclusive	CGCTCGGTCTCCGTTTGGT		CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Release 6 ∩ Chang & Larracuente	CGCTCGGTCTCCGTTTGGT	CAGTGCTCGTGTGAGAAG	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Release 6 ∩ Illumina	CGCTCGGTCTCCGTTTGGT	GCCCGTATTCCCGCTTGC	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Release 6 ∩ Falcon	CGCTCGGTCTCCGTTTGGT	CATCGGCCACGGTCCCGT	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC
Release 6 ∩ Sanger	CGCTCGGTCTCCGTTTGGT	GACTGGTACTCGCGTGAC	CGCAACGCTTGGGACGGTTCCAATCGGATC	ATGGGCCAATTCGGTTC

BALANCE PARAMETERS				
Library subset	priming forward (1)	priming forward (2)	RT priming	priming reverse
Chang & Larracuente (non YGS)	ACAAATCCGACCATCG		CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Chang & Larracuente exclusive	CAGTGCTCGTGTGAGAAG		CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Chang & Larracuente ∩ Falcon	CAGTGCTCGTGTGAGAAG	CATCGGCCACGGTCCCGT	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Chang & Larracuente ∩ Sanger	CAGTGCTCGTGTGAGAAG	GACTGGTACTCGCGTGAC	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Illumina exclusive	GCCCGTATTCCCGCTTGC		CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Illumina exclusive ∩ Chang & Larracuente	GCCCGTATTCCCGCTTGC	CAGTGCTCGTGTGAGAAG	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Illumina exclusive ∩ Falcon	GCCCGTATTCCCGCTTGC	CATCGGCCACGGTCCCGT	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Illumina exclusive ∩ Sanger	GCCCGTATTCCCGCTTGC	GACTGGTACTCGCGTGAC	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Falcon exclusive	CATCGGCCACGGTCCCGT		CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Sanger exclusive	GACTGGTACTCGCGTGAC		CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Sanger ∩ Falcon	GACTGGTACTCGCGTGAC	CATCGGCCACGGTCCCGT	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Release 6 exclusive	CGCTCGGTCTCCGTTTGGT		CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Release 6 ∩ Chang & Larracuente	CGCTCGGTCTCCGTTTGGT	CAGTGCTCGTGTGAGAAG	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Release 6 ∩ Illumina	CGCTCGGTCTCCGTTTGGT	CATCGGCCACGGTCCCGT	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Release 6 ∩ Falcon	CGCTCGGTCTCCGTTTGGT	GACTGGTACTCGCGTGAC	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC
Release 6 ∩ Sanger	CGCTCGGTCTCCGTTTGGT	GACTGGTACTCGCGTGAC	CGCTCGTCTCCGGTCCACCGTTGCGGTAC	ATGGGCCAATTCGGTTC

Tabela 5.1: Regiões de *priming* utilizadas na Biblioteca final de sondas *oligopaint*

Nota da tabela: A estrutura de todas as sondas na biblioteca segue o seguinte padrão com polaridade 5'-3': *priming forward* + *priming* da RT + sequência da sonda selecionada + *priming* reverso, variando conforme o subgrupo dentro da mesma biblioteca de oligonucleotídeos (*Library subset*). Células vazias implicam em apenas subgrupos exclusivos de sondas (*exclusive*), ou seja, que não são compartilhados com outras linhagens.

A região de *priming* da RT foi utilizada para separar as sondas desenhadas a partir dos parâmetros *stringent* e *balance*. Ou seja, todas as sondas desenhadas com os parâmetros mais restritivos, independente da montagem, têm a mesma região de *priming* da RT e o mesmo vale para as sondas desenhadas com os parâmetros *balance*. Como as etapas de PCR em Tempo Real e IVT iniciais selecionam apenas sondas de uma montagem específica, a Transcrição Reversa irá adicionar o fluoróforo apenas às sondas desenhadas a partir de um

STRINGENT PARAMETERS			
Library subset	priming forward	RT priming	priming reverse
Chang & Larracuent (non YGS)	ACAAATCCGACGATCG	CGCAACGTTGGGACGTTCCAATCGGATC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Chang & Larracuent	CAGTGTCTCGTGAAG	CGCAACGTTGGGACGTTCCAATCGGATC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Illumina	GCCCGTATTCCCGCTTGC	CGCAACGTTGGGACGTTCCAATCGGATC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Falcon	CATCGGCGACGGTCCCGT	CGCAACGTTGGGACGTTCCAATCGGATC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Sanger	GACTGGTACTCGGTGAC	CGCAACGTTGGGACGTTCCAATCGGATC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Release 6	CGTCGGTCTCCGTTGCT	CGCAACGTTGGGACGTTCCAATCGGATC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT

BALANCE PARAMETERS			
Library subset	priming forward	RT priming	priming reverse
Chang & Larracuent (non YGS)	ACAAATCCGACGATCG	CGGTGCTCTCCGTTCCACCGTTGCGCTTAC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Chang & Larracuent	CAGTGTCTCGTGAAG	CGGTGCTCTCCGTTCCACCGTTGCGCTTAC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Illumina	GCCCGTATTCCCGCTTGC	CGGTGCTCTCCGTTCCACCGTTGCGCTTAC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Falcon	CATCGGCGACGGTCCCGT	CGGTGCTCTCCGTTCCACCGTTGCGCTTAC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Sanger	GACTGGTACTCGGTGAC	CGGTGCTCTCCGTTCCACCGTTGCGCTTAC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT
Release 6	CGTCGGTCTCCGTTGCT	CGGTGCTCTCCGTTCCACCGTTGCGCTTAC	TAATACGACTCATTATAGGGAACCGGAATTGGCGCAT

Tabela 5.2: *Primers* utilizados no protocolo de amplificação e adição de fluoróforo

Nota da tabela: Relação de *primers* utilizados para amplificar cada subgrupo dentro da biblioteca (*Library subset*). *Primers* da RT foram sintetizados duas vezes – uma vez acoplado ao marcador Cy3 e outra ao Cy5 – para que diferentes combinações das duas cores pudessem ser utilizadas para análises entre uma e outra montagem.

parâmetro específico. Além disso, dada a complexidade deste protocolo paralela à biblioteca desenhada, foi utilizada a mesma região de *priming* reverso para todos os subgrupos de sondas.

Por fim, para amplificar e adicionar fluoróforo em um subgrupo desejado de oligonucleotídeos é necessário simplesmente utilizar três *primers* como reagentes: aquele específico da montagem (*primer forward*), o *primer* da RT correspondente a um conjunto específico de parâmetros e o mesmo *primer* reverso acoplado ao promotor da T7 (Tabela 5.2). Tais *primers* como reagentes foram sintetizados através do comerciante Sigma Aldrich/Merk.

Para propriamente adicionar as regiões de *priming* às sondas foi necessário criar os arquivos com os diferentes subgrupos de sondas, incluindo as intersecções entre duas diferentes montagens. Para tanto, foi executado o comando abaixo que, a partir de um arquivo TXT produzido através da ferramenta InteractVenn (HEBERLE *et al.*, 2015), gerou cada um dos grupos de sondas, mantendo a maior parte das intersecções como exclusivas do grupo desejado e apenas o compartilhamento par a par entre as duas montagens.

```

1 ## ex: Is (Illumina stringent) exclusivos
2 cat InteractiVenn | sed 's/and//g' | sed -e 's/\[\([^\]]*\)\]/\_\\1/g' | ↔)
   sed s/___*/_/g | grep -E "^_Is_:|^_Is_ _Cs_ _Fs_:|^_Is_ _Cs_ _Ws_ ↔)
   _Fs_:|^_Is_ _Cs_ _Ws_:|^_Rs_ _Is_ _Cs_ _Ws_:|^_Rs_ _Is_ _Cs_ ↔)
   _Ws_ _Fs_:|^_Rs_ _Is_ _Cs_:|^_Rs_ _Is_ _Cs_ _Fs_:|^_Rs_ _Is_ ↔)
   _Fs_:|^_Rs_ _Is_ _Ws_ _Fs_:|^_Is_ _Ws_ _Fs_:|^_Rs_ _Is_ ↔)
   _Ws_:" | cut -d":" -f2 | cut -d" " -f2- | tr '\n' ',' | sed ↔)
   's/~/[Is]: /' | sed 's/.$//'' | sed -e '$a\' >> groups_InteractiVenn
3 ## ex: Is (Illumina stringent) compartilhado com Fs (Falcon stringent)
4 cat InteractiVenn | sed 's/and//g' | sed -e 's/\[\([^\]]*\)\]/\_\\1/g' | ↔)
   sed s/___*/_/g | grep "^_Is_ _Fs_:" | cut -d":" -f2 | cut -d" " -f2- ↔)
   | sed 's/~/[Is] and [Fs]: /' | sed 's/.$//'' | sed -e '$a\' >> ↔)

```

groups_InteractiveVenn

Uma vez gerado o arquivo com apenas as intersecções desejadas, contendo todas as sondas finais e dispondo das regiões de *priming* que passaram nos critérios de alinhamento mencionados, as mesmas foram adicionadas às sondas de acordo com o respectivo grupo.

```

1 ## Lista de arquivos Venn, ex:
2 vennFiles=("groups_InteractiveVenn_stringent" ↔)
   "groups_InteractiveVenn_balance")
3 ## Arquivo com as regioes completas de priming forward (1 ou 2) em ↔)
   ordem alfabetica do Diagrama de Venn
4 GroupPriming=GroupPrimingMax2.fasta
5 ## Arquivo com um primer RT para cada InteractiveVenn
6 RTPriming=RTPriming.fasta
7 ## Arquivo com a sequencia priming reverse
8 ReversePriming=PrimingReverse.fasta
9 ## Arquivos finais
10 vennOutputFiles=("stringent_PrimingTarget.oligos" ↔)
   "balance_PrimingTarget.oligos")
11 finalFile=("finalProbes_Max2Groups_synthesize.oligos")
12
13 ## Adicionar regioes de priming
14 runs="$(( ${#vennFiles[@]} - 1 ))"
15 for venn in $(seq 0 $runs)
16 do
17     vennFile=${vennFiles[${venn}]}
18     # Criar arquivos para cada grupo de sondas em groups_InteractiveVenn
19     cat ${vennFile} | while read line
20     do
21         group='echo $line | cut -d":" -f1 | tr -d ' ' | sed 's/and//g' \
22         | sed -e 's/\[([~])*\]\]/\_1_/g' | sed s/__$/_/g'
23         echo $line | cut -d":" -f2 | cut -d" " -f2- \
24         | sed 's/,/\n/g' >> Group${group}probes.oligos
25     done
26     groupFiles=( $(ls Group*probes.oligos) )
27     primersForward=( $(grep -v ">" ${GroupPriming}) )
28     primersRT=( $(grep -v ">" ${RTPriming}) )

```

```

29 pr=( $(grep -v ">" ${ReversePriming}) )
30 rt=${primersRT[venn]}
31 groups="$(( ${#groupFiles[@]} - 1 ))"
32 for i in $(seq 0 $groups)
33 do
34     # Adicionar regioao priming conforme grupo
35     pf=${primersForward[i]}
36     vennGroupName='echo ${groupFiles[i]} | cut -d"/" -f5 | cut ↵
-d"." -f1'
37     vennGroup=${groupFiles[i]}
38     probeNumber=1
39     cat ${vennGroup} | while read target
40     do
41         echo -e "${pf} \t ${rt} \t ${target} \t ${pr} \t ↵
${probeNumber}_${vennGroupName} \t ${pf}${rt}${target}${pr}" >> ↵
${vennOutputFiles[venn]}
42         echo -e "${pf} \t ${rt} \t ${target} \t ${pr} \t ↵
${probeNumber}_${vennGroupName} \t ${pf}${rt}${target}${pr}" >> ↵
${vennGroupName}_ProbesAndPrimers.oligos
43         ((probeNumber=probeNumber+1))
44     done
45
46     # Alinhar main primings contra sondas com regioes priming
47     cat ${vennGroupName}_ProbesAndPrimers.oligos | while read line
48     do
49         id='echo $line | cut -d" " -f5'
50         seq='echo $line | cut -d" " -f6'
51         echo ">${id}" >> ${vennGroupName}_ProbesAndPrimers.fasta
52         echo "${seq}" >> ${vennGroupName}_ProbesAndPrimers.fasta
53     done
54     blastn -query main_priming -subject ↵
${vennGroupName}_ProbesAndPrimers.fasta -outfmt '6 qseqid sseqid ↵
pident length mismatch gapopen qstart qend sstart send eval ↵
bitscore' -word_size 12 > ↵
blast_MainPrimers_${vennGroupName}_ProbesAndPrimers
55     blastn -query RT_priming -subject ↵

```

```
    ${vennGroupName}_ProbesAndPrimers.fasta -outfmt '6 qseqid sseqid ↔)
    pident length mismatch gapopen qstart qend sstart send evaluate ↔)
    bitscore' -word_size 20 >> ↔)

    blast_MainPrimers_${vennGroupName}_ProbesAndPrimers
56     done
57     finalFiles+=("${vennOutputFiles[venn]}")
58 done
59
60 ## Criar arquivo final da Biblioteca completa de oligonucleotídeos
61 ls ${finalFiles[@]} | while read file; do cut -f6 $file >> ${finalFile} ↔)
    ; done
```

5.2 Amplificação dos Subgrupos de Oligonucleotídeos

O protocolo IVT-RT de Beliveau *et al.* (2017) foi realizado, seguindo as alterações descritas anteriormente (Capítulo 2, Seção 2.3, Subseção 2.3.1).

As reações de PCR em Tempo Real foram interrompidos antes de atingir saturação (Figura 5.4), conforme recomendado por Beliveau *et al.* (2017), para evitar o desequilíbrio da reação. Dado que essa reação conta com diversos fragmentos diferentes de DNA, também com tamanhos variados, a amplificação das sondas menores pode saturar a enzima Taq DNA Polimerase, suprimindo a amplificação das maiores sondas e causando um desequilíbrio não desejado nessa reação. Além disso, os próprios reagentes utilizados na PCR em Tempo Real começam a ser amplificados. Isso é bastante problemático, pois, embora essa reação inicial sirva apenas para realizar uma seleção primária no grupo de sondas desejado, o restante do protocolo depende de uma PCR em Tempo Real bem executada.

Os produtos das reações de PCR em Tempo Real foram testados através de eletroforese horizontal com gel de agarose 4% em solução tampão TBE 1X (Tris, ácido bórico, EDTA) (Figura 5.5). Os controles com água não apresentam nenhuma banda no gel, demonstrando que não houve contaminação na reação. Além disso, o tamanho da banda apresentada para as amostras de sondas *oligopaint* é consistente com o tamanho esperado após a adição da sequência nucleotídica promotora para a enzima T7 RNA Polimerase, com 20 nt: originalmente, o tamanho das sondas a serem sintetizadas variava de 102 a 130 bp, assim, após a reação de PCR em Tempo Real, o tamanho das sondas deveria variar de 122 a 150 nt.

Após o protocolo de IVT, o tamanho ribonucleotídico total das diferentes sondas deve diminuir em 17 nt, pois segundo o fabricante apenas os três últimos nucleotídeos do

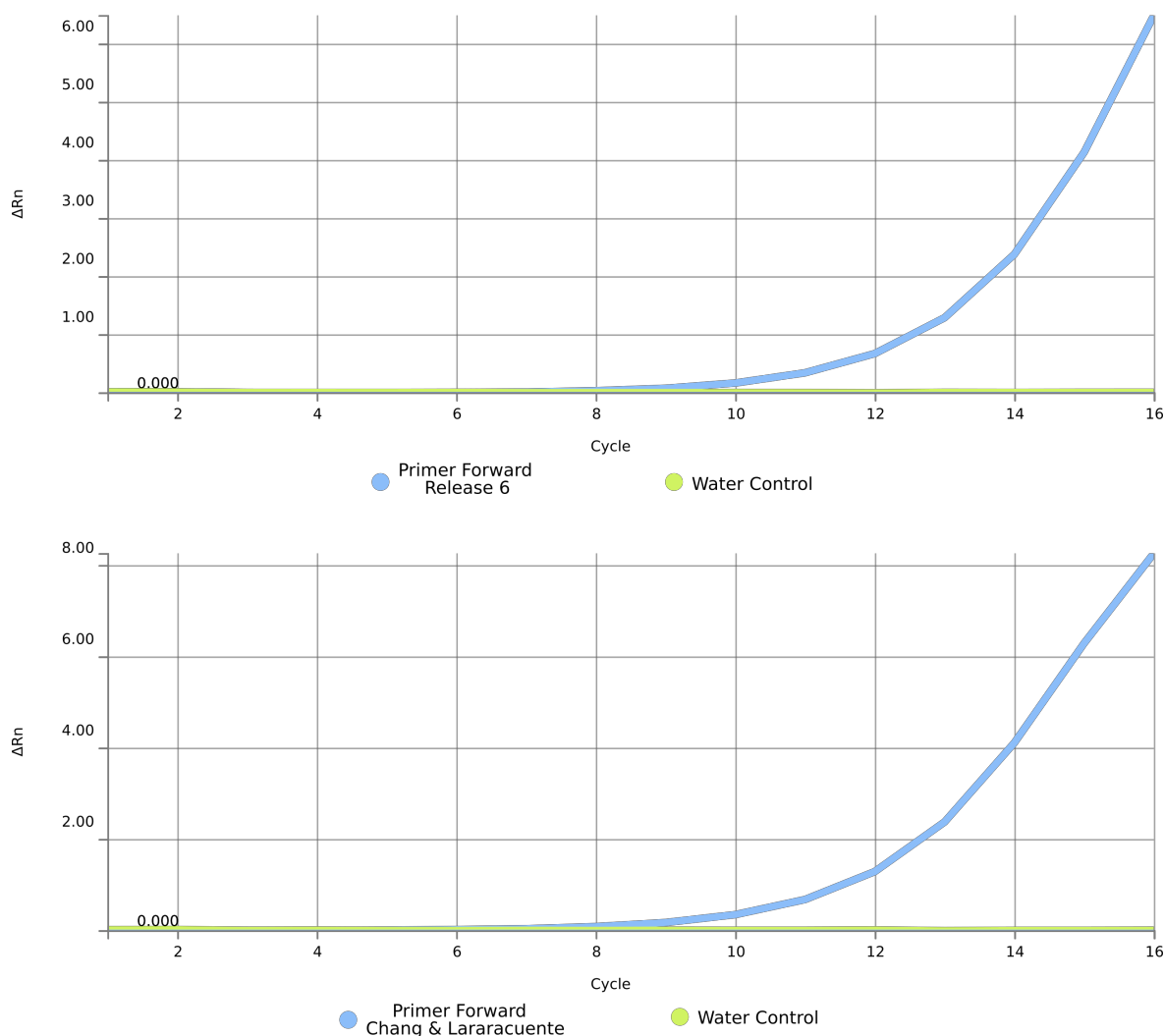


Figura 5.4: Reações de PCR em Tempo Real

Nota da figura: Cada um dos gráficos representa a reação de PCR em Tempo Real utilizando um dos diferentes *primers forward* para seleção inicial de um dos subgrupos de sondas *oligopaint*. São apresentados apenas os gráficos para os subgrupos de sondas desenhadas a partir das montagens de Chang e Larracuente (2019) e *Release 6*. Todas as reações foram finalizadas na fase exponencial, antes de atingirem saturação. Figura produzida por Henry Bonilla.

promotor da enzima T7 RNA Polimerase são transcritos à RNA – assim, após essa etapa as sondas devem variar de 105 a 133 ribonucleotídeos. Durante a RT os 18 nt que correspondem à sequência de *priming forward* não são transcritos reversamente à DNA, servindo apenas para seleção das sondas e, dessa forma, não estão presentes na fita fluorescentemente marcada de DNA. Dessa forma, após a finalização completa do protocolo de amplificação, as sondas devem variar de 87 a 115 nt.

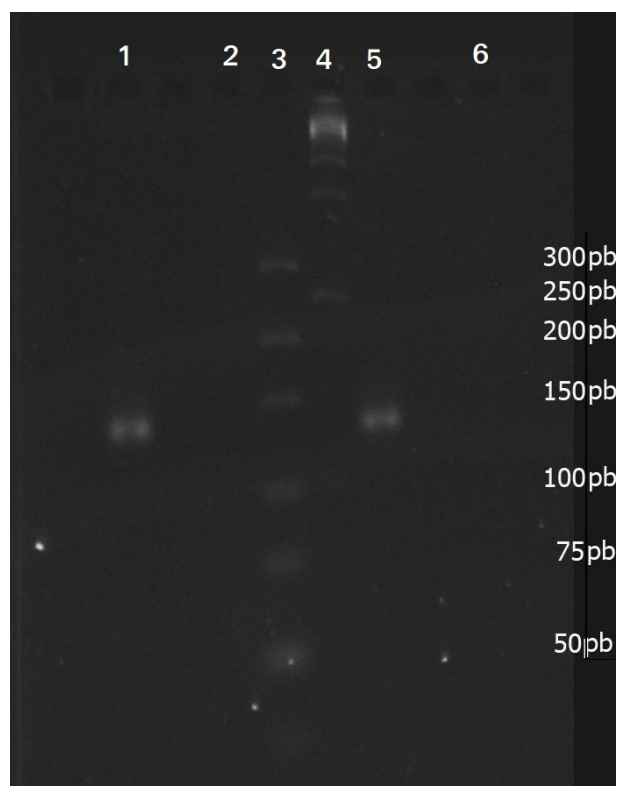


Figura 5.5: Eletroforese horizontal do produto de PCR em Tempo Real

Nota da figura: É apresentado o resultado do experimento apenas para os produtos das reações de PCR em Tempo Real com os *primers forward* dos subgrupos de sondas *Release 6* e Chang & Larracuent. Os poços marcados na figura correspondem à: (1) subgrupos de sondas Chang & Larracuent; (2) controle branco da amostra 1; (3) Marcador molecular 50 pb à 300 pb; (4) Marcador molecular 1 kb; (5) subgrupos de sondas *Release 6*; (6) controle branco da amostra 5. Figura produzida por Henry Bonilla.

Levando esses tamanhos esperados em consideração, os produtos das reações de amplificação em sua totalidade foram testados através de eletroforese vertical com gel de poliacrilamida 15% desnaturante (Acrilamida, bis-acrilamida e ureia) (Figura 5.6). É possível observar que tanto antes quanto após a purificação final das sondas amplificadas as três amostras utilizadas apresentaram uma banda forte típica da separação dos DNTPs utilizados, correspondendo a uma banda de 10 pb. Após purificação, é fácil distinguir uma segunda banda que deve corresponder aos *primers* utilizados na reação de RT, com tamanho de 30 nt. A banda que corresponde às sondas amplificadas está justamente entre 75 e 100 pb e, em razão dessas amostras advirem de reação de RT, não era esperado observar bandas fortes na eletroforese, dado o rendimento desta reação ser pequeno. Ainda assim, uma banda fraca do tamanho esperado é observada, constatando que o protocolo de amplificação funcionou.

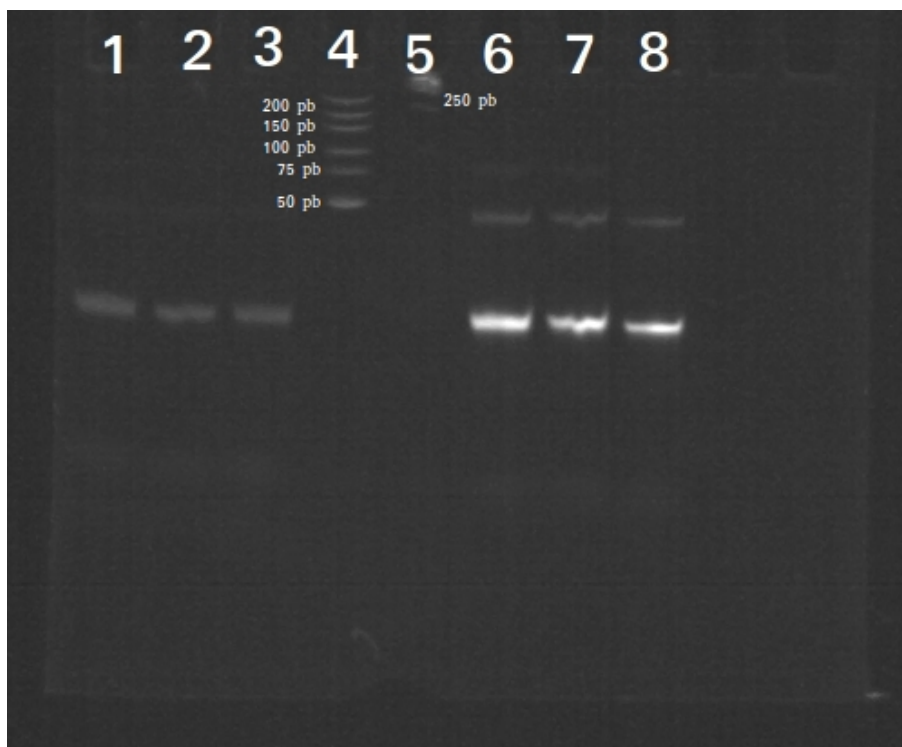


Figura 5.6: Eletroforese vertical do produto final de amplificação

Nota da figura: É apresentado o resultado do experimento apenas para os produtos das reações de PCR em Tempo Real com os *primers forward* dos subgrupos de sondas *Release 6* e Chang & Larracuenté. Os poços marcados na figura correspondem à: (1) subgrupo de sondas Chang & Larracuenté antes da etapa de purificação final; (2) subgrupo de sondas *Release 6* antes da etapa de purificação final; (3) controle branco da reação RT antes da etapa de purificação final; (4) Marcador molecular 50 pb à 300 pb; (5) Marcador molecular 1 kb; (6) subgrupo de sondas Chang & Larracuenté após etapa de purificação final; (7) subgrupo de sondas *Release 6* após etapa de purificação final; (8) controle branco da reação RT após etapa de purificação final. Figura produzida por Henry Bonilla.

5.3 Experimentos de FISH *Oligopaint* e Microscopia

Embora tenham sido sintetizadas sondas *oligopaint* desenhadas a partir dos parâmetros *stringent* e *balance* no intuito de observar a marcação do cromossomo Y metafásico e meiótico, a situação instaurada de pandemia causada pelo vírus SARS-CoV-2 desde março de 2020 implicou diretamente no acesso regular aos laboratórios. Assim, houve tanto limitação no tempo disponível para realização dos experimentos, como também limitações relacionadas às medidas de segurança impostas. Dessa forma, os ensaios para observação de células meióticas não foi possível, uma vez que depende de acesso à sala muito pequena e extremamente fechada em que fica o Microscópio Confocal de Varredura a *laser*, este só podendo ser utilizado sob supervisão do técnico responsável. Além disso, a complexidade do protocolo de amplificação

das sondas exigiu um tempo de cerca de 4 meses para completa padronização. Como resultado, essa seção concerne apenas os ensaios de FISH *Oligopaint* com sondas desenhadas a partir dos parâmetros mais restritivos (*stringent*) marcando amostra tecidual cerebral de larva de terceiro instar de *Drosophila melanogaster*, de modo a observar cromossomos metafásicos. Ademais, apenas sondas desenhadas a partir das montagens de Chang e Larracuent (2019) e *Release 6* foram empregadas nos ensaios apresentados neste capítulo.

As figuras a seguir demonstram que a *pipeline* desenvolvida foi eficiente para desenhar sondas *oligopaint* que marcam apenas o cromossomo Y (Figura 5.7). Não é observável nenhum sinal fora desse cromossomo alvo, sendo este aqui identificado pelo cariótipo característico de *Drosophila melanogaster* (Figura 1.1b,c). Vale ressaltar que durante os ensaios iniciais de FISH com as sondas desenhadas com os parâmetros mais restritivos (*stringent*), a temperatura de hibridização de 47 °C levou à deformação dos cromossomos metafásicos fixos nas lâminas. Apesar de esse valor seguir a mesma parametrização realizada por Beliveau *et al.* (2018) e ter sido utilizado para filtrar sondas através do OligoMiner, foi constatado que ao diminuir a temperatura de hibridização para 42 °C o experimento funciona, permitindo que as sondas hibridizem.

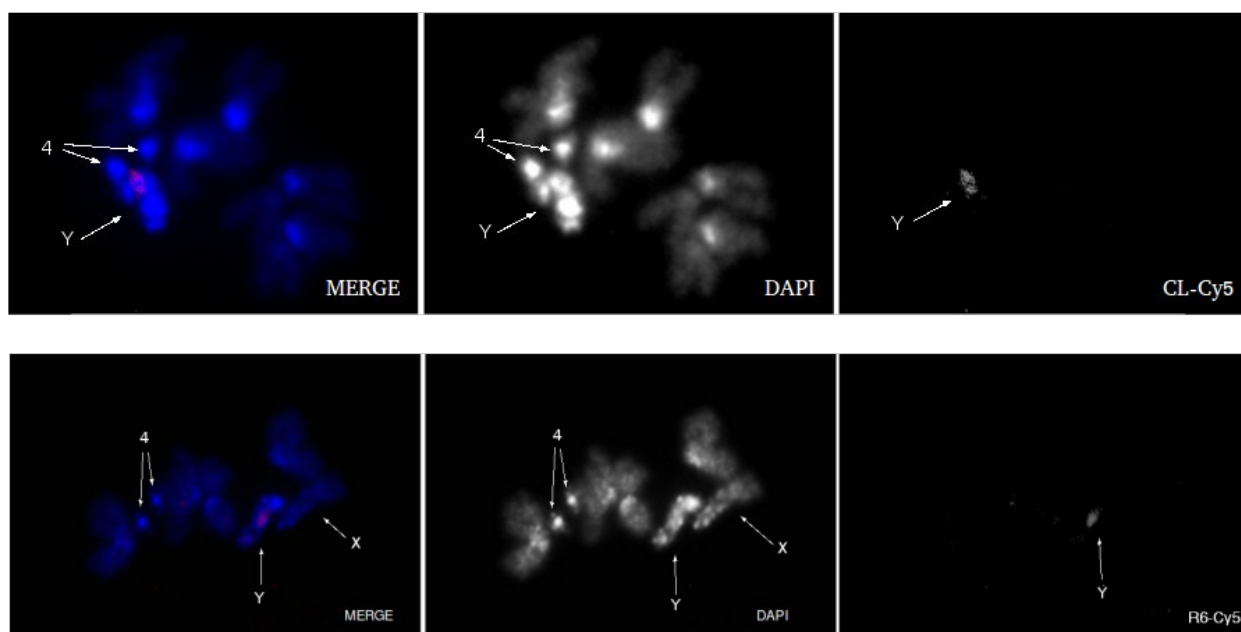


Figura 5.7: Resultados de FISH *Oligopaint* na linhagem BM5

Nota da figura: Ensaios realizados com cromossomos metafásicos fixados a partir de amostra tecidual de cérebro de larva de terceiro instar de BM5. São apresentados os resultados dos ensaios para os subgrupos de sondas obtidas a partir de *Release 6* e Chang e Larracuent (2019) com os parâmetros *stringent*. Essas sondas hibridizaram na lâmina com uma temperatura de 42 °C. Figura produzida por Henry Bonilla.

Os resultados de FISH *Oligopaint* apresentados foram realizados com amostras fixadas a partir de BM5, uma linhagem de *Drosophila melanogaster* diferente da utilizada para gerar as sondas *oligopaint* empregadas – ISO1 (Tabela 2.1). Ensaios como esse permitem avaliar a eficiência da *pipeline* ao usar, por exemplo, genoma de referência da espécie de interesse para desenhar sondas que serão aplicadas em uma linhagem diferente da sequenciada. Levando isso em consideração, ainda resta conduzir ensaios com as linhagens Tempe-T e ISO1 com as mesmas sondas (*Stringent* de *Release 6* e de Chang e Larracuent (2019)), de modo a averiguar se o padrão de marcação em outra linhagem que não ISO1 difere do observado em BM5 e se diferenças são observadas entre essas duas primeiras linhagens e ISO1 em si. Vale ressaltar que, embora tenham sido realizados ensaios com todas as três linhagens, todas as lâminas produzidas a partir de amostras teciduais de Tempe-T e ISO1 não tinham muitos cromossomos metafásicos e que, em geral, os ensaios de FISH *Oligopaint* não marcaram bem as lâminas – essa é uma particularidade inerente de experimentos com a técnica de FISH, que requer tanto destreza para realização dos ensaios, quanto tempo para vastas tentativas.

De toda forma, os ensaios realizados até então demonstraram que a *pipeline* desenvolvida pode ser eficiente para desenhar sondas utilizando genomas montados e *short-reads* disponíveis mesmo que a linhagem a ser utilizada nos ensaios de FISH *Oligopaint* difira da sequenciada. Isso é muito interessante, pois melhor comporta os diferentes cenários onde a pesquisa com o cromossomo Y se insere, dando mais flexibilidade aos futuros usuários da *pipeline* para escolher as sequências de entrada.

Os subgrupos de sondas utilizadas não marcam esse cromossomo em maior extensão quando comparadas com os resultados obtidos a partir da sonda de DNA satélite AATAC marcada com digoxigenina (Figura 1.1d). A referida sonda é amplamente utilizada para o cromossomo Y e de acordo com Tsai, Yan e McKee (2011) e Bonaccorsi e Lohe (1991) essa sonda marca apenas um pequeno trecho do cromossomo Y, a região de heterocromatina *h6*. Idealmente, seria realizado ensaio controle com a sonda AATAC para garantir que as sondas desenhadas de fato marcam o cromossomo Y, identificado até então apenas por seu cariótipo distinto. Ademais, esse ensaio permitiria constatar as diferentes regiões marcadas pelas sondas desenhadas através da *pipeline* desenvolvida e pela referida sonda AATAC, que provavelmente não são as mesmas pela alta repetitividade de sondas de DNA satélite.

Apesar de ser esperada a marcação de uma região mais ampla através da utilização da *pipeline* desenvolvida, têm que ser considerados alguns aspectos importantes sobre os experimentos de validação realizados. Por exemplo, ao alterar a enzima Taq DNA Polimerase de alta fidelidade (Taxa de erro: 4.4×10^{-7}) por *kit Maxima SYBR Green/ROX qPCR Master Mix (2X)* (ThermoFisher Scientific/Invitrogen) que contém enzima Taq DNA Polimerase com fidelidade similar à enzima comum (Taxa de erro: 2×10^{-4}), a eficiência do protocolo de

amplificação pode ter sido prejudicada – sondas podem ter sido amplificadas com muito mais erros, afetando assim a hibridização dessas sondas em cromossomos fixados e diminuindo a extensão de marcação observada. Um segundo ponto é que apenas dois subconjuntos de sondas *oligopaint* desenhadas com os parâmetros *stringent* foram testadas até então (*Release 6* e de Chang e Larracuent (2019)), restando averiguar a eficiência de sondas desenhadas a partir das sequências inferidas para o cromossomo Y com as montagens Falcon, Sanger e Illumina. Além disso, já era esperado que a região marcada com sondas *stringent* seria menor em comparação com as marcações obtidas a partir das sondas *balance*, por isso a realização de tais experimentos é crucial para melhor entender a eficiência da *pipeline* desenvolvida.

É importante ainda considerar que as sondas foram desenhadas para *Drosophila melanogaster*, um dos organismos mais estudados atualmente. Se por um lado sua utilização facilita os experimentos de validação pela alta disponibilidade de sequências, sondas controle existentes e da fácil manutenção desse organismo em biblioteca de cultura para realização de ensaios de FISH *Oligopaint*, para organismos não-modelo muitas vezes não existem sondas conhecidas para o cromossomo Y, até mesmo pela dificuldade de montagem inerente do cromossomo Y. Assim, o que pode parecer até então como uma região marcada com sondas *oligopaint* em extensão inferior à da marcada por sondas de DNA satélite, pode oferecer uma ferramenta poderosa para pesquisas com outros organismos. A facilidade de obtenção das ferramentas, alteração de parâmetros e, até mesmo, de utilização de sequências obtidas a partir de linhagem diferente daquela na qual as sondas sintetizadas hibridizarão, ainda oferece muitas vantagens para pesquisas com o cromossomo Y.

Além desses primeiros resultados, estão pendentes os ensaios de FISH *Oligopaint* utilizando sonda para heterocromatina do cromossomo X como controle. Estes, além de permitir observar a ausência de sinal fora do cromossomo Y, garantem também a distinção entre machos e fêmeas – supondo que as sondas desenhadas para o cromossomo Y não hibridizassem no mesmo, a marcação controle ainda assim apareceria, eliminando a possibilidade de que o experimento não tivesse funcionado por má execução do protocolo. Apesar disso, uma vez que nos ensaios já realizados é observada hibridização de sonda no cromossomo Y, é esperado que este ensaio resulte no mesmo padrão de marcação no cromossomo Y para cada um dos diferentes subgrupos utilizados.

Para publicação desses resultados em revista científica, é pretendido finalizar todos os ensaios de FISH *Oligopaint* a partir das sondas sintetizadas, incluindo o restante daquelas desenhadas a partir de outras montagens que não *Release 6* e Chang e Larracuent (2019), bem como realização de ensaios com as sondas obtidas usando os parâmetros padronizados (*balance*). Esses resultados permitirão constatar qual o melhor conjunto de parâmetros para desenho de sondas *oligopaint* e mostrar qual tecnologia de sequenciamento permite

resgatar, através do YGS, os melhores conjuntos de sequências exclusivas do cromossomo Y para aplicação em ensaios de FISH *Oligopaint*. Além disso, os ensaios para observação das células meióticas fixadas a partir da amostra tecidual do testículo de machos adultos de *Drosophila melanogaster* contrastados com marcação de sonda de DNA satélite AATAC devem ser concluídos para ser possível averiguar se durante a formação da estrutura em *loop* as regiões marcadas pelas sondas selecionadas com a *pipeline* desenvolvida estendem a região marcada, dado que as diferentes regiões de heterocromatina do cromossomo Y podem apresentar comportamentos até então desconhecidos.

Postos os resultados de validação *in situ* obtidos, é seguro determinar que a *pipeline* desenvolvida (Figura 5.8) permitiu marcar o cromossomo Y de *Drosophila melanogaster* com eficiência. Esses dados apontam que a nova abordagem do ZM da suíte de programas OligoMiner (BELIVEAU *et al.*, 2018) de fato permite a utilização de sondas que hibridizam mais de uma vez no mesmo cromossomo alvo, levando tanto à economia na síntese de sondas como permitindo melhor aproveitamento das sequências de entrada em cromossomos altamente repetitivos. Tendo isso em mente, a *pipeline* desenvolvida pode ser parcialmente aplicada para outros cromossomos alvo que não o Y/W na tentativa de solucionar a falta de sondas *oligopaint* para regiões destes que sejam ricas em sequências repetitivas, como foi feito em colaboração com Henry Bonilla em sua pesquisa de mestrado para tentar solucionar regiões extensas sem sondas nos cromossomos de *Drosophila miranda* (Processo FAPESP: 2019/10559-1 sob supervisão da Professora Dra. Maria Vibranovski). Para tanto, basta dispensar a utilização do YGS (CARVALHO; CLARK, 2013; DUPIM; ALMEIDA; CARVALHO, versão não publicada), dado que o mesmo foi desenvolvido para inferir sequências apenas para o cromossomo Y/W. Além disso, opcionalmente pode ainda se realizar análise de contaminantes através do BlobTools (LAETSCH; BLAXTER, 2017; LAETSCH *et al.*, 2017), garantindo que não sejam desenhadas sondas para regiões contaminadas. Nessa possível aplicação, a maior parte da *pipeline* segue idêntica, removendo apenas a última filtragem com os *short-reads* de fêmea.

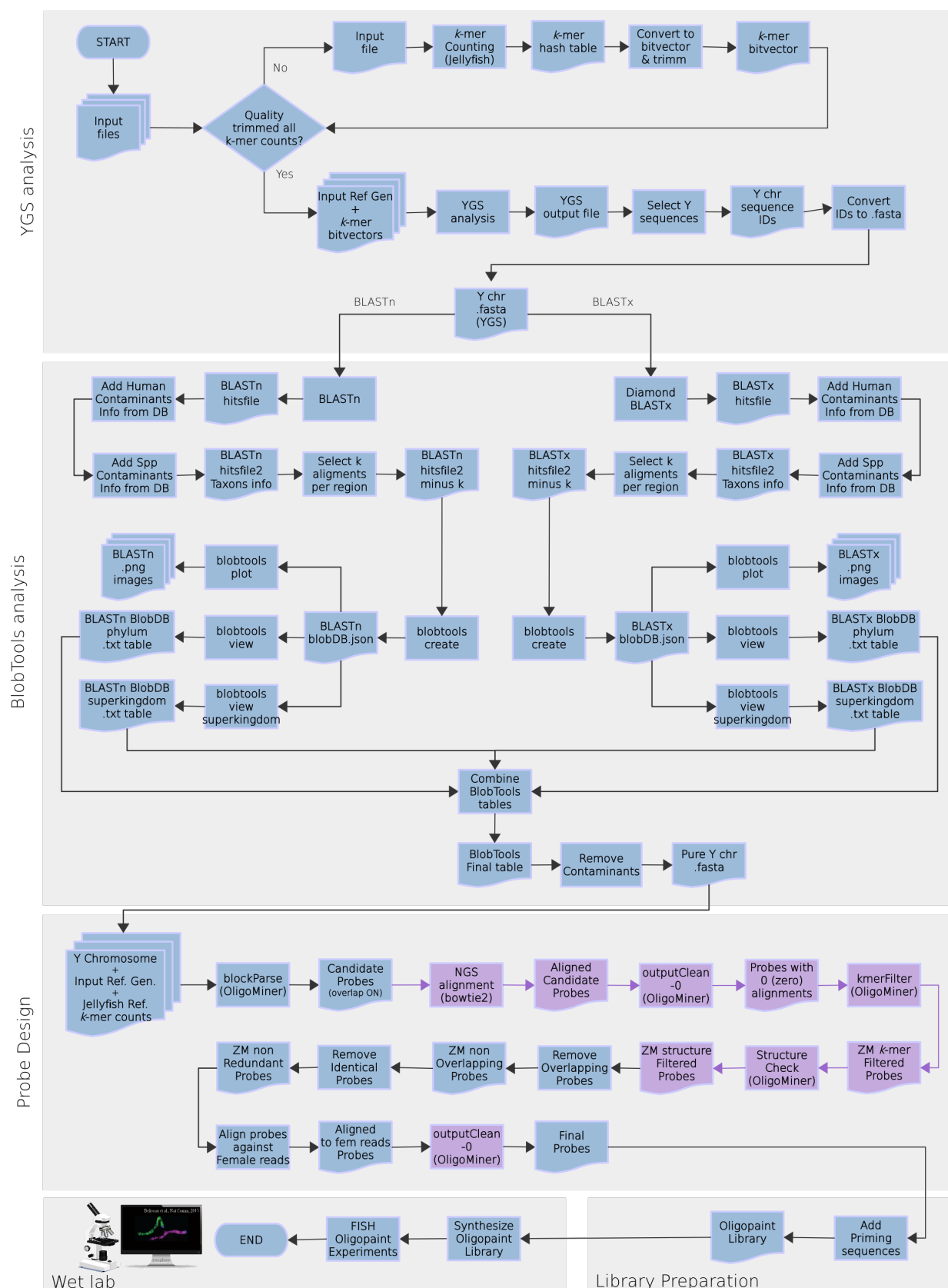


Figura 5.8: OligoY: Fluxograma da *Pipeline* final

Nota da figura: *Pipeline* final para desenho de sondas *oligopaint* para o cromossomo Y/W. Em lilás estão destacadas as partes do fluxograma diretamente ligadas ao uso do ZM.

6

Considerações finais

Em virtude dos experimentos conduzidos no desenvolvimento de uma *pipeline* para desenhar sondas *oligopaint* para o cromossomo Y, os resultados obtidos e as análises realizadas apontam que a utilização ressignificada de ferramentas existentes em Bioinformática demonstrou êxito em experimentos de FISH *Oligopaint*.

O desenvolvimento da *pipeline* foi conquistado não somente pela exploração das ferramentas disponíveis e de sua combinação, mas também através de execuções com diferentes conjuntos de parâmetros, a fim de poder melhor orientar os futuros usuários da mesma.

Dessa forma, ao submeter cinco diferentes montagens genômicas ao YGS (CARVALHO; CLARK, 2013) com dois diferentes *k-mers*, $k = 15$ e $k = 18$, foram obtidos resultados muito semelhantes para cada um dos arquivos de entrada em relação à extensão em Mb das sequências nucleotídicas inferidas para o cromossomo Y (Capítulo 3, Tabela 3.1). Isso demonstra que a escolha do tamanho de *k-mer* pode seguir as recomendações de Carvalho e Clark (2013), correspondendo ao menos a $k = 15$ para genomas de invertebrados e a $k = 18$ para genomas de vertebrados.

Seguindo a *pipeline*, os diferentes cromossomos inferidos foram submetidos à análise de contaminantes através do BlobTools (LAETSCH; BLAXTER, 2017; LAETSCH *et al.*, 2017). Nos testes realizados, apenas as montagens realizadas através de sequenciamento Illumina e Sanger apresentaram algumas sequências contaminadas (Capítulo 3, Tabela 3.3). Apesar de parecer uma etapa dispendiosa na maior parte dos testes efetuados, o uso dessa ferramenta é uma das garantias para que sondas *oligopaint* não sejam desenhadas para regiões fora do alvo, sendo este um aspecto importante tanto para custo de síntese das mesmas quanto por não ludibriar os resultados esperados a partir das densidades finais por kb cromossômico.

Todas as abordagens exploradas através do OligoMiner (BELIVEAU *et al.*, 2018), proporcionaram um enorme ganho de sondas quando comparado com a quantidade total de sondas que já existiam para o cromossomo Y de *Drosophila melanogaster* antes desta pesquisa

(BELIVEAU *et al.*, 2018) (Figura 4.11) – o ZM em especial, e suas estimativas de número total de alvos, tendo sido um destaque. Foi demonstrado que o modo ZM proporciona manter sondas que se repetem exclusivamente no cromossomo alvo fornecido (Capítulo 4, Seção 4.2). Além disso, independente do modo de filtrar as sondas pela predição do número de alvos, UM, LDM ou ZM, ao desenhá-las permitindo a sobreposição entre candidatas, maiores quantidades de sondas são selecionadas ao final de todas as filtragens (Tabelas 4.10 e 4.11) indicando potencialmente melhores densidades de alvos por área cromossômica (Tabela 4.12) e, assim, resultados mais eficientes nos experimentos de FISH *Oligopaint*.

É importante ressaltar que tanto a inserção de uma etapa para remoção de sequências contaminadas quanto a nova abordagem através do *Zero Mode* proporcionam economia na síntese de sondas. Ao remover contaminantes, sondas não são erroneamente desenhadas e selecionadas para regiões que não vão marcar cromossomos na espécie de interesse, nem o cromossomo alvo, nem cromossomos fora deste. Assim, o custo de síntese das sondas pode ser reduzido com a adição desse passo na *pipeline*. Além disso, a nova abordagem através do ZM permite selecionar sondas com mais de um alvo e, como apenas uma cópia de cada sonda *oligopaint* distinta precisa ser sintetizada, isso resulta em um número muito maior de regiões marcadas a partir de uma quantidade menor de sondas efetivamente sintetizadas. Essa também é uma economia, pois se menos sondas são sintetizadas, o custo de produção pode cair. Essa questão financeira é relevante, pois cada *chip* com cerca de 90 mil sondas custa de USD \$4K à USD \$6K através do comerciante Genscript (Para mais detalhes: <<https://www.genscript.com/precise-synthetic-oligo-pools.html>>).

Dada a complexidade da biblioteca sintetizada, mesmo após reduzir os conjuntos de sondas testadas *in situ*, foi escolhido amplificar a mesma através do protocolo IVT-RT de Beliveau *et al.* (2017) (Capítulo 5). E, apesar de que este protocolo proporcione concentração maior das sondas amplificadas, cabe ao usuário da *pipeline* escolher aquele que melhor se encaixe com suas necessidades, disponibilidade de reagentes e de recursos para realização dos experimentos. De toda maneira, vale ressaltar que a utilização inteligente das regiões de *priming* nos conjuntos de sondas, encontrada para suprir a demanda do alto número de sondas selecionadas para mais de um conjunto na mesma biblioteca, não apresentou nenhum problema, sendo uma solução plausível para àqueles que enfrentem situações de complexidade semelhantes.

Os ensaios de FISH *Oligopaint* realizados com as sondas desenhadas a partir dos parâmetros mais restritivos, *stringent*, não apresentaram hibridização fora da região alvo. Uma vez que as sondas testadas foram todas selecionadas através do ZM, esse resultado confirma ser possível utilizar sondas *oligopaint* que apresentem mais de um alvo no mesmo cromossomo, contanto que não exista nenhuma região de hibridização em outro. E a filtragem

contra os *short-reads* de fêmea, que faz parte da *pipeline* final desenvolvida, tem um papel muito importante em assegurar que marcações fora do cromossomo alvo não aconteçam: essa filtragem se baseia na abundância dos *k-mers* em fêmea que acabam não estando montados no genoma de referência e, dessa forma, são considerados para remoção de sondas que potencialmente hibridizariam fora do alvo. Dessa forma, os resultados de validação *in situ* encontrados, são extremamente positivos, pois asseguram que a *pipeline* OligoY aumenta a densidade de alvo por área cromossômica a partir de um número menor de sondas, o que também está atrelado a uma economia para sintetizar a biblioteca final de sondas, como mencionado anteriormente.

Apesar de a extensão da marcação no cromossomo metafásico de *Drosophila melanogaster* não parecer ser superior à da marcação com a sonda de DNA satélite AATAC, ainda é necessário observar se o mesmo padrão permanece durante a meiose, quando o cromossomo Y se encontra completamente disperso pela célula – as sondas *oligopaint* selecionadas podem marcar uma região cromossômica que durante a metáfase está mais condensada do que a região de heterocromatina *h6* marcada pela sonda de DNA satélite AATAC (TSAI; YAN; MCKEE, 2011; BONACCORSI; LOHE, 1991), mas que, durante a meiose, pode estar mais dispersa ou mesmo se dispersar por regiões até então desconhecidas para o território do *loop* do cromossomo Y. Além disso, é preciso verificar se todas as diferentes montagens e conjuntos de parâmetros retornam resultados semelhantes ou não. E, como mencionado, apesar de uma pequena área ser marcada pelas sondas *oligopaint* desenhadas, a *pipeline* desenvolvida pode trazer resultados mais satisfatórios para pesquisadores trabalhando com espécies menos estudadas para as quais diferentes categorias de sondas ainda não sejam conhecidas.

Ademais, além de trazer luz para a citogenética com o cromossomo Y/W, essa abordagem encontrada através do ZM pode representar uma série de perspectivas promissoras para ensaios de FISH *Oligopaint* com quaisquer regiões alvo caracteristicamente repetitivas, como apontado no final do Capítulo 5 (Seção 5.3). Nesse caso, a *pipeline* desenvolvida poderia ser parcialmente aplicada para outros cromossomos alvo que não sejam Y/W (Figura 5.8).

As ferramentas utilizadas nesta *pipeline*, todas gratuitas e de fácil usabilidade, proporcionam extrema versatilidade, garantindo aos futuros usuários autonomia na escolha de diversos parâmetros que influenciam diretamente os resultados de ensaios de fluorescência. Vale ressaltar que os ensaios de FISH *Oligopaint* conduzidos até o momento e as diversas análises realizadas e discutidas ao longo dessa dissertação, apontam que os parâmetros utilizados foram suficientes para maximizar a distribuição do sinal de marcação fluorescente ao longo de um cromossomo Y precariamente montado.

É interessante observar que os experimentos de FISH *Oligopaint* não proporcionam

meramente uma ferramenta para realizar especulações da estrutura dos cromossomos e de suas interações ao longo do ciclo celular – esse protocolo pode servir como um mecanismo de visualização de trechos de DNA. Assim, outra perspectiva promissora da *pipeline* desenvolvida está na montagem de cromossomos complexos como o Y de *Drosophila melanogaster*: ao desenhar sondas *oligopaint* para cada um dos *contigs* inferidos a partir do YGS separadamente, sintetizando-os com regiões de *priming* particulares, é possível observar a ordem dos mesmos por microscopia de fluorescência e, assim, ter uma orientação mais precisa da contiguidade do cromossomo.

Referências

ADAMS, M. D. The genome sequence of *Drosophila melanogaster*. **Science**, American Association for the Advancement of Science (AAAS), v. 287, n. 5461, p. 2185–2195, Mar 2000. Disponível em: <<https://doi.org/10.1126/science.287.5461.2185>>.

AHO, A. V.; KERNIGHAN, B. W.; WEINBERGER, P. J. **Awk — A Pattern Scanning and Processing Language**. New Jersey: Bell Laboratories, 1978. Disponível em: <<https://wolfram.schneider.org/bsd/7thEdManVol2/awk/awk.pdf>>.

ALBERT, P. S. *et al.* Whole-chromosome paints in maize reveal rearrangements, nuclear domains, and chromosomal relationships. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 116, n. 5, p. 1679–1685, Jan 2019. Disponível em: <<https://doi.org/10.1073/pnas.1813957116>>.

ALTSCHUL, S. F. *et al.* Basic local alignment search tool. **Journal of Molecular Biology**, Elsevier BV, v. 215, n. 3, p. 403–410, Out 1990. Disponível em: <[https://doi.org/10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2)>.

ASA. American Standard Code for Information Interchange. **American Standards Association**, Jun 1963. Disponível em: <<https://web.archive.org/web/20160617012149/http://worldpowersystems.com/J/codes/X3.4-1963/>>.

BAYES, J. J.; MALIK, H. S. Altered heterochromatin binding by a hybrid sterility protein in drosophila sibling species. **Science**, American Association for the Advancement of Science (AAAS), v. 326, n. 5959, p. 1538–1541, Out 2009. Disponível em: <<https://doi.org/10.1126/science.1181756>>.

BELIVEAU, B. J.; APOSTOLOPOULOS, N.; WU, C. Visualizing genomes with oligopaint FISH probes. **Current Protocols in Molecular Biology**, Wiley, v. 105, n. 1, Jan 2014. Disponível em: <<https://doi.org/10.1002/0471142727.mb1423s105>>.

BELIVEAU, B. J. *et al.* Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using oligopaint FISH probes. **Nature Communications**, Springer Science and Business Media LLC, v. 6, n. 1, Mai 2015. Disponível em: <<https://doi.org/10.1038/ncomms8147>>.

BELIVEAU, B. J. *et al.* In Situ Super-Resolution Imaging of Genomic DNA with OligoSTORM and OligoDNA-PAINT. In: **Methods in Molecular Biology**. Springer New York, 2017. p. 231–252. Disponível em: <https://doi.org/10.1007/978-1-4939-7265-4_19>.

BELIVEAU, B. J. *et al.* OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 115, n. 10, p. E2183–E2192, Fev 2018. Disponível em: <<https://doi.org/10.1073/pnas.1714530115>>.

BOETTIGER, A. N. *et al.* Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. **Nature**, Springer Science and Business Media LLC, v. 529, n. 7586, p. 418–422, Jan 2016. Disponível em: <<https://doi.org/10.1038/nature16496>>.

BONACCORSI, S.; LOHE, A. R. Fine Mapping of Satellite DNA Sequences along the Y Chromosome of *Drosophila melanogaster*: Relationships between Satellite Sequences and Fertility Factors. **Genetics**, Oxford University Press (OUP), v. 129, n. 1, p. 177–189, Set 1991.

BRAHMACHARI, V.; JAIN, S. Heterochromatin. In: **Encyclopedia of Systems Biology**. Springer New York, 2013. p. 880–880. Disponível em: <https://doi.org/10.1007/978-1-4419-9863-7_849>.

BREITWIESER, F. P. *et al.* Human contamination in bacterial genomes has created thousands of spurious proteins. **Genome Research**, Cold Spring Harbor Laboratory, v. 29, n. 6, p. 954–960, Mai 2019. Disponível em: <<https://doi.org/10.1101/gr.245373.118>>.

BRIDGES, C. B. Non-Disjunction as Proof of the Chromosome Theory of Heredity. **Genetics**, Oxford University Press (OUP), v. 1, n. 1, p. 1–52, Jan 1916. Disponível em: <<https://doi.org/10.1093/genetics/1.1.1>>.

BUCHFINK, B.; XIE, C.; HUSON, D. H. Fast and sensitive protein alignment using DIAMOND. **Nature Methods**, Springer Science and Business Media LLC, v. 12, n. 1, p. 59–60, Nov 2014. Disponível em: <<https://doi.org/10.1038/nmeth.3176>>.

CAI, W. *et al.* Preparation of drosophila polytene chromosome squashes for antibody labeling. **Journal of Visualized Experiments**, MyJove Corporation, n. 36, Fev 2010. Disponível em: <<https://doi.org/10.3791/1748>>.

CARLSON, M.; BRUTLAG, D. Cloning and characterization of a complex satellite DNA from *Drosophila melanogaster*. **Cell**, Elsevier BV, v. 11, n. 2, p. 371–381, Jun 1977. Disponível em: <[https://doi.org/10.1016/0092-8674\(77\)90054-x](https://doi.org/10.1016/0092-8674(77)90054-x)>.

CARVALHO, A. B.; CLARK, A. G. Efficient identification of Y chromosome sequences in the human and drosophila genomes. **Genome Research**, Cold Spring Harbor Laboratory, v. 23, n. 11, p. 1894–1907, Ago 2013. Disponível em: <<https://doi.org/10.1101/gr.156034.113>>.

CARVALHO, A. B. *et al.* Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 98, n. 23, p. 13225–13230, Out 2001. Disponível em: <<https://doi.org/10.1073/pnas.231484998>>.

CARVALHO, A. B.; LAZZARO, B. P.; CLARK, A. G. Y chromosomal fertility factors kl-2 and kl-3 of *Drosophila melanogaster* encode dynein heavy chain polypeptides. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 97, n. 24, p. 13239–13244, Nov 2000. Disponível em: <<https://doi.org/10.1073/pnas.230438397>>.

CARVALHO, A. B. *et al.* Birth of a new gene on the Y chromosome of *Drosophila melanogaster*. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 112, n. 40, p. 12450–12455, Set 2015. Disponível em: <<https://doi.org/10.1073/pnas.1516543112>>.

CATTANI, M. V.; PRESGRAVES, D. C. Incompatibility Between X Chromosome Factor and Pericentric Heterochromatic Region Causes Lethality in Hybrids Between *Drosophila melanogaster* and Its Sibling Species. **Genetics**, Oxford University Press (OUP), v. 191, n. 2, p. 549–559, Jun 2012. Disponível em: <<https://doi.org/10.1534/genetics.112.139683>>.

CELNIKER, S. E. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. **Genome Biology**, Springer Science and Business Media LLC, v. 3, n. 12, p. research0079.1, 2002. Disponível em: <<https://doi.org/10.1186/gb-2002-3-12-research0079>>.

CHAKRABORTY, M. *et al.* Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. **Nucleic Acids Research**, Oxford University Press (OUP), p. gkw654, Jul 2016. Disponível em: <<https://doi.org/10.1093/nar/gkw654>>.

CHAKRABORTY, M. *et al.* Hidden genetic variation shapes the structure of functional elements in *Drosophila*. **Nature Genetics**, Springer Science and Business Media LLC, v. 50, n. 1, p. 20–25, Dez 2018. Disponível em: <<https://doi.org/10.1038/s41588-017-0010-y>>.

CHANG, C.-H.; LARRACUENTE, A. M. Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the *Drosophila melanogaster* Y Chromosome. **Genetics**, Genetics, v. 211, n. 1, p. 333–348, Jan 2019. ISSN 0016-6731. Disponível em: <<https://www.genetics.org/content/211/1/333>>.

CHARLESWORTH, B.; LANGLEY, C. H.; STEPHAN, W. The evolution of restricted recombination and the accumulation of repeated DNA sequences. **Genetics**, Genetics, v. 112, n. 4, p. 947–962, Abr 1986. Disponível em: <<https://pubmed.ncbi.nlm.nih.gov/3957013/>>.

CHEN, K. H. *et al.* Spatially resolved, highly multiplexed RNA profiling in single cells. **Science**, American Association for the Advancement of Science (AAAS), v. 348, n. 6233, p. aaa6090–aaa6090, Abr 2015. Disponível em: <<https://doi.org/10.1126/science.aaa6090>>.

CLARK, M. E. *et al.* Induced Paternal Effects Mimic Cytoplasmic Incompatibility in *Drosophila*. **Genetics**, Oxford University Press (OUP), v. 173, n. 2, p. 727–734, Jun 2006. Disponível em: <<https://doi.org/10.1534/genetics.105.052431>>.

CSINK, A. K.; HENIKOFF, S. Genetic modification of heterochromatic association and nuclear organization in *Drosophila*. **Nature**, Springer Science and Business Media LLC, v. 381, n. 6582, p. 529–531, Jun 1996. Disponível em: <<https://doi.org/10.1038/381529a0>>.

DERNBURG, A. F.; SEDAT, J. W.; HAWLEY, R. Direct evidence of a role for heterochromatin in meiotic chromosome segregation. **Cell**, Elsevier BV, v. 86, n. 1, p. 135–146, Jul 1996. Disponível em: <[https://doi.org/10.1016/s0092-8674\(00\)80084-7](https://doi.org/10.1016/s0092-8674(00)80084-7)>.

DIRKS, R. M. *et al.* Thermodynamic analysis of interacting nucleic acid strands. **SIAM Review**, Society for Industrial & Applied Mathematics (SIAM), v. 49, n. 1, p. 65–88, Jan 2007. Disponível em: <<https://doi.org/10.1137/060651100>>.

DIRKS, R. M.; PIERCE, N. A. A partition function algorithm for nucleic acid secondary structure including pseudoknots. **Journal of Computational Chemistry**, Wiley, v. 24, n. 13, p. 1664–1677, Out 2003. Disponível em: <<https://doi.org/10.1002/jcc.10296>>.

DIRKS, R. M.; PIERCE, N. A. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. **Journal of Computational Chemistry**, Wiley, v. 25, n. 10, p. 1295–1304, 2004. Disponível em: <<https://doi.org/10.1002/jcc.20057>>.

Drosophila 12 Genomes Consortium; Project Leaders; CLARK, A. G. Evolution of genes and genomes on the drosophila phylogeny. **Nature**, Springer Science and Business Media LLC, v. 450, n. 7167, p. 203–218, Nov 2007. Disponível em: <<https://doi.org/10.1038/nature06341>>.

DUPIM, E. G.; ALMEIDA, I. P. de; CARVALHO, A. B. Improvements in performance of YGS using hash. **Laboratório Prof. Carvalho**, versão não publicada.

EICHLER, E. E.; CLARK, R. A.; SHE, X. An assessment of the sequence gaps: Unfinished business in a finished human genome. **Nature Reviews Genetics**, Springer Science and Business Media LLC, v. 5, n. 5, p. 345–354, Mai 2004. Disponível em: <<https://doi.org/10.1038/nrg1322>>.

ELGIN, S. C.; WORKMAN, J. L. A year dominated by histone modifications, transitory and remembered. **Current Opinion in Genetics & Development**, Elsevier BV, v. 12, n. 2, p. 127–129, Abr 2002. Disponível em: <[https://doi.org/10.1016/s0959-437x\(02\)00276-9](https://doi.org/10.1016/s0959-437x(02)00276-9)>.

FERGUSON-SMITH, M. A. History and evolution of cytogenetics. **Molecular Cytogenetics**, Springer Science and Business Media LLC, v. 8, n. 1, Mar 2015. Disponível em: <<https://doi.org/10.1186/s13039-015-0125-8>>.

FERREE, P. M.; BARBASH, D. A. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. **PLoS Biology**, Public Library of Science (PLOS), v. 7, n. 10, p. e1000234, Out 2009. Disponível em: <<https://doi.org/10.1371/journal.pbio.1000234>>.

FILGES, S. *et al.* Impact of polymerase fidelity on background error rates in next-generation sequencing with unique molecular identifiers/barcodes. **Scientific Reports**, Springer Science and Business Media LLC, v. 9, n. 1, Mar 2019. Disponível em: <<https://doi.org/10.1038/s41598-019-39762-6>>.

FINGERHUT, J. M.; MORAN, J. V.; YAMASHITA, Y. M. Satellite DNA-containing gigantic introns in a unique gene expression program during drosophila spermatogenesis. **PLoS Genetics**, Public Library of Science (PLoS), v. 15, n. 5, p. e1008028, Mai 2019. Disponível em: <<https://doi.org/10.1371/journal.pgen.1008028>>.

GALL, J. G.; PARDUE, M. L. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 63, n. 2, p. 378–383, Jun 1969. Disponível em: <<https://doi.org/10.1073/pnas.63.2.378>>.

GEPNER, J.; HAYS, T. S. A fertility region on the y chromosome of *Drosophila melanogaster* encodes a dynein microtubule motor. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 90, n. 23, p. 11132–11136, Dez 1993. Disponível em: <<https://doi.org/10.1073/pnas.90.23.11132>>.

GOLDSTEIN, G. N. R. **Identificação de sequências ligadas ao cromossomo Y no gênero *Drosophila***. Dissertação (Mestrado) — UFRJ, Mai 2016.

GUTZWILLER, F. *et al.* Dynamics of *Wolbachia pipientis* Gene Expression Across the *Drosophila melanogaster* Life Cycle. **G3 Genes|Genomes|Genetics**, Oxford University Press (OUP), v. 5, n. 12, p. 2843–2856, Out 2015. Disponível em: <<https://doi.org/10.1534/g3.115.021931>>.

HALL, A. B. *et al.* Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. **BMC Genomics**, Springer Science and Business Media LLC, v. 14, n. 1, p. 273, Abr 2013. Disponível em: <<https://doi.org/10.1186/1471-2164-14-273>>.

HEBERLE, H. *et al.* InteractiVenn: a web-based tool for the analysis of sets through venn diagrams. **BMC Bioinformatics**, Springer Science and Business Media LLC, v. 16, n. 1, Mai 2015. Disponível em: <<https://doi.org/10.1186/s12859-015-0611-3>>.

HOSKINS, R. A. *et al.* Sequence Finishing and Mapping of *Drosophila melanogaster* Heterochromatin. **Science**, American Association for the Advancement of Science (AAAS), v. 316, n. 5831, p. 1625–1628, Jun 2007. Disponível em: <<https://doi.org/10.1126/science.1139816>>.

HOSKINS, R. A. *et al.* The Release 6 reference sequence of the *Drosophila melanogaster* genome. **Genome Research**, Cold Spring Harbor Laboratory, v. 25, n. 3, p. 445–458, Jan 2015. Disponível em: <<https://doi.org/10.1101/gr.185579.114>>.

HOSKINS, R. A. *et al.* Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. **Genome Biology**, Springer Science and Business Media LLC, v. 3, n. 12, p. research0085.1, Dez 2002. Disponível em: <<https://doi.org/10.1186/gb-2002-3-12-research0085>>.

HUISINGA, K. L.; BROWER-TOLAND, B.; ELGIN, S. C. R. The contradictory definitions of heterochromatin: transcription and silencing. **Chromosoma**, Springer

Science and Business Media LLC, v. 115, n. 2, p. 110–122, Fev 2006. Disponível em: <<https://doi.org/10.1007/s00412-006-0052-x>>.

International Human Genome Sequencing Consortium *et al.* Initial sequencing and analysis of the human genome. **Nature**, Springer Science and Business Media LLC, v. 409, n. 6822, p. 860–921, Fev 2001. Disponível em: <<https://doi.org/10.1038/35057062>>.

KAUFMAN, T. C. A Short History and Description of *Drosophila melanogaster* Classical Genetics: Chromosome Aberrations, Forward Genetic Screens, and the Nature of Mutations. **Genetics**, Oxford University Press (OUP), v. 206, n. 2, p. 665–689, Jun 2017. Disponível em: <<https://doi.org/10.1534/genetics.117.199950>>.

KERNIGHAN, B. W.; PIKE, R. **The Unix Programming Environment**. New Jersey: Englewood Cliffs, N.J. : Prentice-Hall, 1984. ISBN 0-13-937681-X. Disponível em: <<https://archive.org/details/unixprogramminge0000kern/page/n3/mode/2up>>.

KIM, K. E. *et al.* Long-read, whole-genome shotgun sequence data for five model organisms. **Scientific Data**, Springer Science and Business Media LLC, v. 1, n. 1, Nov 2014. Disponível em: <<https://doi.org/10.1038/sdata.2014.45>>.

KURUMIZAKA, H. Nucleosome structure. In: **Encyclopedia of Systems Biology**. Springer New York, 2013. p. 1552–1556. Disponível em: <https://doi.org/10.1007/978-1-4419-9863-7_1410>.

LAETSCH, D. R.; BLAXTER, M. L. BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. **F1000Research**, F1000 Research Ltd, v. 6, p. 1287, Jul 2017. Disponível em: <<https://doi.org/10.12688/f1000research.12232.1>>.

LAETSCH, D. R. *et al.* Drl/blobtools: Blobtools v1.0.1. Zenodo, Ago 2017. Disponível em: <<https://doi.org/10.5281/zenodo.845347>>.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with bowtie 2. **Nature Methods**, Springer Science and Business Media LLC, v. 9, n. 4, p. 357–359, Mar 2012. Disponível em: <<https://doi.org/10.1038/nmeth.1923>>.

LARKIN, A. *et al.* FlyBase: updates to the *drosophila melanogaster* knowledge base. **Nucleic Acids Research**, Oxford University Press (OUP), v. 49, n. D1, p. D899–D907, Nov 2021. Disponível em: <<https://doi.org/10.1093/nar/gkaa1026>>.

LI, H. *et al.* The sequence alignment/map format and SAMtools. **Bioinformatics**, Oxford University Press (OUP), v. 25, n. 16, p. 2078–2079, Jun 2009. Disponível em: <<https://doi.org/10.1093/bioinformatics/btp352>>.

LOHE, A. R.; BRUTLAG, D. L. Multiplicity of satellite DNA sequences in *Drosophila melanogaster*. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 83, n. 3, p. 696–700, Fev 1986. Disponível em: <<https://doi.org/10.1073/pnas.83.3.696>>.

LOHE, A. R.; BRUTLAG, D. L. Adjacent satellite DNA segments in *Drosophila*. **Journal of Molecular Biology**, Elsevier BV, v. 194, n. 2, p. 171–179, Mar 1987. Disponível em: <[https://doi.org/10.1016/0022-2836\(87\)90366-4](https://doi.org/10.1016/0022-2836(87)90366-4)>.

LOHE, A. R.; BRUTLAG, D. L. Identical satellite DNA sequences in sibling species of *Drosophila*. **Journal of Molecular Biology**, Elsevier BV, v. 194, n. 2, p. 161–170, Mar 1987. Disponível em: <[https://doi.org/10.1016/0022-2836\(87\)90365-2](https://doi.org/10.1016/0022-2836(87)90365-2)>.

LOHE, A. R.; HILLIKER, A. J.; ROBERTS, P. A. Mapping simple repeated dna sequences in heterochromatin *Drosophila melanogaster*. **Genetics**, Oxford University Press (OUP), v. 134, n. 4, p. 1149–1174, Ago 1993.

MAHADEVARAJU, S. *et al.* Dynamic sex chromosome expression in drosophila male germ cells. **Nature Communications**, Springer Science and Business Media LLC, v. 12, n. 1, Fev 2021. Disponível em: <<https://doi.org/10.1038/s41467-021-20897-y>>.

MAIMON, I.; GILBOA, L. Dissection and staining of *Drosophila* larval ovaries. **Journal of Visualized Experiments**, MyJove Corporation, n. 51, Mai 2011. Disponível em: <<https://doi.org/10.3791/2537>>.

MARÇAIS, G.; KINGSFORD, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. **Bioinformatics**, v. 7, n. 6, p. 764–770, Jan 2011. ISSN 1367-4803. Disponível em: <<https://doi.org/10.1093/bioinformatics/btr011>>.

MARKOW, T. A.; O'GRADY, P. M. Phylogenetic relationships of Drosophilidae. In: **Drosophila**. Elsevier, 2006. p. 3–64. Disponível em: <<https://doi.org/10.1016/b978-012473052-6/50001-9>>.

MARKOW, T. A.; RICKER, J. P. Developmental stability in hybrids between the sibling species pair, *drosophila melanogaster* and *drosophila simulans*. **Genetica**, Springer Science and Business Media LLC, v. 84, n. 2, p. 115–121, 1991. Disponível em: <<https://doi.org/10.1007/bf00116551>>.

MCILROY, M. D. A Research UNIX Reader: Annotated Excerpts from the Programmer's Manual, 1971-1986." Computing Science. **Bell Laboratories**, 1987.

MCKEE, B. D.; HONG, C. S.; DAS, S. On the roles of heterochromatin and euchromatin in meiosis in *drosophila*: mapping chromosomal pairing sites and testing candidate mutations for effects on X-Y nondisjunction and meiotic drive in male meiosis. **Genetica**, Springer Science and Business Media LLC, v. 109, n. 1/2, p. 77–93, 2000. Disponível em: <<https://doi.org/10.1023/a:1026536200594>>.

MCMAHON, L. E. SED - A Non-Interactive Text Editor. **Bell Laboratories**, Ago 1978. Disponível em: <<https://wolfram.schneider.org/bsd/7thEdManVol2/sed/sed.pdf>>.

MILLER, A. The internal anatomy and histology of the imago of *Drosophila melanogaster*. In: DEMEREC, M. (Ed.). **Biology of Drosophila**. Cold Spring Harbor Laboratory Press, 2008. p. 420–534. ISBN 978-087969828-7. Disponível em: <flybase.org/reports/FBim0000074.html>.

- MILLER, J. R.; KOREN, S.; SUTTON, G. Assembly algorithms for next-generation sequencing data. **Genomics**, Elsevier BV, v. 95, n. 6, p. 315–327, Jun 2010. Disponível em: <<https://doi.org/10.1016/j.ygeno.2010.03.001>>.
- MOFFITT, J.; ZHUANG, X. RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH). In: **Visualizing RNA Dynamics in the Cell**. Elsevier, 2016. p. 1–49. Disponível em: <<https://doi.org/10.1016/bs.mie.2016.03.020>>.
- MONTGOMERY, L. M. **Anne of Green Gables**. New York: Puffin Books, 2014.
- MULLER, D. K.; MARTIN, C. T.; COLEMAN, J. E. T7 RNA polymerase interacts with its promoter from one side of the DNA helix. **Biochemistry**, American Chemical Society (ACS), v. 28, n. 8, p. 3306–3313, Abr 1989. Disponível em: <<https://doi.org/10.1021/bi00434a028>>.
- MURAKAMI, Y. Heterochromatin and euchromatin. In: **Encyclopedia of Systems Biology**. Springer New York, 2013. p. 881–884. Disponível em: <https://doi.org/10.1007/978-1-4419-9863-7_1413>.
- MURGHA, Y. E.; ROUILLARD, J.-M.; GULARI, E. Methods for the Preparation of Large Quantities of Complex Single-Stranded Oligonucleotide Libraries. **PLoS ONE**, Public Library of Science (PLOS), v. 9, n. 4, p. e94752, Abr 2014. Disponível em: <<https://doi.org/10.1371/journal.pone.0094752>>.
- NGUYEN, S. C.; JOYCE, E. F. Programmable chromosome painting with oligopaints. In: **Imaging Gene Expression**. Springer New York, 2019. p. 167–180. Disponível em: <https://doi.org/10.1007/978-1-4939-9674-2_11>.
- O'GRADY, P. M.; MARKOW, T. A. Phylogenetic taxonomy in drosophila: Problems and prospects. **Fly**, Informa UK Limited, v. 3, n. 1, p. 10–14, Jan 2009. Disponível em: <<https://doi.org/10.4161/fly.3.1.7748>>.
- PADMANABHAN, R.; SARCAR, S. N.; MILLER, D. L. Promoter length affects the initiation of t7 RNA polymerase in vitro: New insights into promoter/polymerase co-evolution. **Journal of Molecular Evolution**, Springer Science and Business Media LLC, v. 88, n. 2, p. 179–193, Dez 2019. Disponível em: <<https://doi.org/10.1007/s00239-019-09922-3>>.
- PARDUE, M. L.; GALL, J. G. Molecular Hybridization of Radioactive DNA to the DNA of Cytological Preparations. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 64, n. 2, p. 600–604, Out 1969. Disponível em: <<https://doi.org/10.1073/pnas.64.2.600>>.
- PARK, K. S.; GODT, D.; KALDERON, D. Dissection and staining of *Drosophila* pupal ovaries. **Journal of Visualized Experiments**, MyJove Corporation, n. 133, Mar 2018. Disponível em: <<https://doi.org/10.3791/56779>>.
- PASSARO, M. *et al.* OligoMinerApp: a web-server application for the design of genome-scale oligonucleotide in situ hybridization probes through the flexible OligoMiner environment. Oxford University Press (OUP), v. 48, n. W1, p. W332–W339, Abr 2020. Disponível em: <<https://doi.org/10.1093/nar/gkaa251>>.

PEDREGOSA, F. *et al.* Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RANGAVITTAL, S. *et al.* DiscoverY: a classifier for identifying y chromosome sequences in male assemblies. **BMC Genomics**, Springer Science and Business Media LLC, v. 20, n. 1, Ago 2019. Disponível em: <<https://doi.org/10.1186/s12864-019-5996-3>>.

ROŠIĆ, S.; KÖHLER, F.; ERHARDT, S. Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. **Journal of Cell Biology**, Rockefeller University Press, v. 207, n. 3, p. 335–349, Nov 2014. Disponível em: <<https://doi.org/10.1083/jcb.201404097>>.

ROSIN, L. F.; NGUYEN, S. C.; JOYCE, E. F. Condensin II drives large-scale folding and spatial partitioning of interphase chromosomes in drosophila nuclei. **PLoS Genetics**, Public Library of Science (PLOS), v. 14, n. 7, p. e1007393, Jul 2018. Disponível em: <<https://doi.org/10.1371/journal.pgen.1007393>>.

ROUILLARD, J.; ZUKER, M.; GULARI, E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. **Nucleic Acids Research**, v. 31, n. 12, p. 3057–3062, Jun 2003. ISSN 0305-1048. Disponível em: <<https://doi.org/10.1093/nar/gkg426>>.

RUALTHANZAUVA, B. “grep was a private command of mine for quite a while before i made it public.” - Ken Thompson. **Wayback Machine**, Fev 2014. Disponível em: <<https://tinyurl.com/2m4xp4rt>>.

SMITH, C. D. *et al.* The Release 5.1 Annotation of *Drosophila melanogaster* Heterochromatin. **Science**, American Association for the Advancement of Science (AAAS), v. 316, n. 5831, p. 1586–1591, Jun 2007. Disponível em: <<https://doi.org/10.1126/science.1139815>>.

STOCKER, H.; GALLANT, P. Getting started. In: **Methods in Molecular Biology**. Humana Press, 2008. p. 27–44. Disponível em: <https://doi.org/10.1007/978-1-59745-583-1_2>.

SULLIVAN, B. A.; BLOWER, M. D.; KARPEN, G. H. Determining centromere identity: cyclical stories and forking paths. **Nature Reviews Genetics**, Springer Science and Business Media LLC, v. 2, n. 8, p. 584–596, Ago 2001. Disponível em: <<https://doi.org/10.1038/35084512>>.

TSAI, J.-H.; YAN, R.; MCKEE, B. D. Homolog pairing and sister chromatid cohesion in heterochromatin in drosophila male meiosis i. **Chromosoma**, Springer Science and Business Media LLC, v. 120, n. 4, p. 335–351, Mar 2011. Disponível em: <<https://doi.org/10.1007/s00412-011-0314-0>>.

VALENTE, G. T. *et al.* B chromosomes: from cytogenetics to systems biology. **Chromosoma**, Springer Science and Business Media LLC, v. 126, n. 1, p. 73–81, Ago 2016. Disponível em: <<https://doi.org/10.1007/s00412-016-0613-6>>.

VANDERLINDE, T. *et al.* An Improved Genome Assembly for *Drosophila navojoa*, the Basal Species in the *mojavensis* Cluster. **Journal of Heredity**, v. 110, n. 1, p. 118–123, Nov 2018. ISSN 0022-1503. Disponível em: <<https://doi.org/10.1093/jhered/esy059>>.

VENTER, J. C. *et al.* The sequence of the human genome. **Science**, American Association for the Advancement of Science (AAAS), v. 291, n. 5507, p. 1304–1351, Fev 2001. Disponível em: <<https://doi.org/10.1126/science.1058040>>.

VIBRANOVSKI, M. D.; KOERICH, L. B.; CARVALHO, A. B. Two new y-linked genes in *drosophila melanogaster*. **Genetics**, Oxford University Press (OUP), v. 179, n. 4, p. 2325–2327, Ago 2008. Disponível em: <<https://doi.org/10.1534/genetics.108.086819>>.

WITTE, A. K. *et al.* Essential role of polymerases for assay performance – impact of polymerase replacement in a well-established assay. **Biomolecular Detection and Quantification**, Elsevier BV, v. 16, p. 12–20, Dez 2018. Disponível em: <<https://doi.org/10.1016/j.bdq.2018.10.002>>.

ZAMORE, P. D.; MA, S. Isolation of *Drosophila melanogaster* testes. **Journal of Visualized Experiments**, MyJove Corporation, n. 51, Mai 2011. Disponível em: <<https://doi.org/10.3791/2641>>.

armU braço cromossômico composto por sequências que apesar de montadas (scaffolds/-contigs) que não foram mapeadas para um cromossomo específico (do inglês *unmapped scaffold*). 21, 27, 28

ASCII código binário (0 e 1) que codifica um conjunto de 128 sinais, sendo 95 sinais gráficos que são imprimíveis (visíveis ao usuário, como letras do alfabeto latino, sinais de pontuação e sinais matemáticos) e 33 sinais de controle que são não-imprimíveis. 33

Bash acrônimo para "**B**ourne-**A**gain **S**hell", o Bash é uma evolução retro-compatível muito mais interativa do Bourne Shell (sh), que permite a execução de sequências de comandos inseridos diretamente na linha de comando ou lidos de arquivos de textos conhecidos como shell *scripts*; assim, o Bash é um interpretador de comandos, um entre os diversos tradutores entre o usuário e o sistema operacional conhecidos como shell. 32, 34, 36

bit computacionalmente, é a menor unidade de informação, tendo dois estados, sendo eles ligado (1) ou desligado (0). 37

bit-array vetor em que a presença ou ausência de determinada informação é codificada em bits, respectivamente em 1 ou 0; também conhecido como bit-vector. 33, 36, 37, 38, 39, 70, 71

cluster consiste em computadores fracamente ou fortemente ligados que trabalham em conjunto, de modo que, em muitos aspectos, podem ser considerados como um único sistema. 38, 69

contig sequência produzida durante o processo de montagem *de novo* a partir da sobreposição de *reads* de sequenciamento formando então uma sequência consenso contígua (do inglês *contiguous*) e contínua em que todas as *reads* pertencem a um e apenas um *contig* e que cada um destes contém pelo menos uma *read*. 19, 21, 27, 28, 29, 32, 37, 38, 58, 59, 62, 63, 64, 65, 69, 71, 72, 73, 74, 75, 79, 80, 81, 82, 93, 94, 96, 99, 103, 104, 108, 114, 134

gap representam espaços inseridos em um alinhamento; biologicamente, pode estar relacionado à inserção ou deleção introduzida em uma ou ambas as sequências alinhadas. 26, 27, 64, 83

k-mer subsequência de monômeros (do inglês *monomers*) de tamanho k contida em uma sequência biológica, ou seja, uma subsequência de nucleotídeos de tamanho k . 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 47, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 69, 70, 71, 77, 78, 83, 84, 85, 86, 88, 89, 90, 91, 93, 94, 97, 98, 99, 100, 101, 114, 131, 133

mismatch alteração pontual entre uma sequência comparada com outra em um alinhamento, biologicamente podendo representar uma mutação pontual. 44, 83

Phred historicamente, o programa Phred foi o primeiro a atribuir de maneira acurada e eficiente pontuações de qualidade específicas para cada base de uma sequência junto à probabilidade de erro; assim, a pontuação de qualidade Phred (*Phred quality score*) é uma medida da qualidade da identificação de cada uma das bases nucleotídicas em sequenciadores automáticos; como resultado, o formato FASTQ codifica pontuações Phred como caracteres ASCII junto com as sequências de *reads*. 30, 33, 34

primer ácido nucleico de fita simples curta utilizado por todos os organismos vivos no início da síntese de DNA, uma vez que as enzimas de DNA polimerase são capazes apenas de adicionar nucleotídeos à extremidade 3' de um ácido nucleico existente, exigindo assim que um *primer* seja ligado ao molde anteriormente; organismos vivos usam *primers* de RNA, enquanto as técnicas de laboratório em bioquímica e biologia molecular que requerem síntese de DNA *in vitro* geralmente usam *primers* de DNA, por serem mais estáveis à variações de temperatura. 49, 50, 51, 113, 117, 118, 119, 123, 124, 125

read sequência inferida de pares de bases (ou probabilidades de pares de bases) correspondendo a todo ou parte de um único fragmento de DNA advindo de um experimento de sequenciamento no qual o genoma é tipicamente fragmentado e cada conjunto destes é sequenciado e produz um conjunto de *reads*, ou leituras, de sequenciamento. 19, 21, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 43, 45, 46, 52, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 71, 73, 74, 97, 98, 99, 100, 101, 102, 103, 104, 107, 109, 127, 129, 133

scaffold sequência produzida a partir da orientação de junção de *contigs*, que originalmente não eram contíguos, em sequências maiores, através da remoção dos *gaps* que existem entre os mesmos. 19, 21, 26, 27, 28, 37, 38, 58, 63, 99

SNP variação na sequência de DNA que afeta somente uma base nucleotídica na sequência do genoma entre indivíduos de uma espécie ou entre pares de cromossomos de um indivíduo. Mais precisamente, SNP é uma substituição da linha germinativa de um único nucleotídeo em uma posição específica no genoma. 35

suíte em informática, corresponde a uma série de programas de computador concebidos para funcionar em conjunto e/ou em série. 28, 30, 31, 32, 42, 46, 102

tandem repetições em *tandem* são segmentos de DNA com um forte padrão periódico, sendo caracterizadas por uma unidade de repetição curta de 2 a 8 pares de bases em que há um motivo de sequência. Esse motivo pode ter apenas alguns pares de bases alterados e se repetete múltiplas vezes formando a repetição em tandem. 21

thread representa a tarefa que um determinado programa realiza, sendo uma forma de um processo dividir a si mesmo em duas ou mais tarefas que podem ser executadas concorrentialmente. 34

threshold em uma escala, representa o limite para que determinada ação seja realizada. 31, 62

Apêndice A – Memorial descritivo da Autora

Caro leitor invisível,

Nesses pedaços de papel que restaram quero te contar um pouco sobre quem é o ser humano por trás da maior parte das análises, digitando e corrigindo essa dissertação por meses, lendo e relendo, desenhando a maior partes das figuras com seus talentos artísticos duvidosos: sou eu, a Isabela Almeida. Por mais que alguns programas de Pós-Graduação nomeiem uma seção como esta de *Memorial Descritivo*, acredito que este seja um termo muito mórbido, parecendo sugerir que nem estou mais aqui nesse planeta geoide chamado Terra – mas se os sistemas da USP durarem para além da minha vida, um dia isso será verdade. Por essa razão, resolvi escrever uma carta totalmente pessoal – até porque o Programa de Pós-Graduação Interunidades em Bioinformática da USP atualmente não fala nada sobre isso. Então vamos lá!

Nasci em São Paulo/SP e vivi os primeiros cinco anos da minha vida em uma casa mais ou menos do tamanho de uma garagem, que ficava em uma das favelas da cidade. Com as conquistas dos meus pais a vida financeira da minha família melhorou e nós pudemos alugar uma casa que tinha quarto, sala e cozinha. A vida toda continuou melhorando em um crescimento praticamente exponencial, incluindo a nossa mudança para uma cidade bem pequena do interior da Paraíba – Monteiro – e, de lá, para Campina Grande (leve ou razoavelmente maior). Queria te contar esse pedacinho dos meus endereços porque as pessoas e os céus que participaram da minha trajetória tiveram um papel muito importante na minha formação. Foi a vida e seus contextos que me ensinaram a ler textos e interpretar o mundo, a ter preferências, a conversar e a escutar.

Cresci gostando muito de cores, poesia e tabelas. Eu brincava de matemática com os livros do meu pai e também preparava atividades para meus alunos invisíveis usando papel carbono. Acredito que esses aspectos tenham me levado para a ciência: o sonho e a realização de ser uma cientista, analisando dados e encontrando cores e poesia nos resultados.

Uma parte muito importante da minha trajetória na ciência foi cursar meu Ensino Médio Integrado ao Técnico em Mineração no Instituto Federal de Educação, Ciência e

Tecnologia da Paraíba – *Campus* Campina Grande. As colmeias daquele lugar estão registradas em minha memória com muitas recordações e aprendizados – fizeram com que eu tivesse currículo Lattes desde os 15 anos, idade com a qual me tornei bolsista do CNPq pela primeira vez. Esses e inúmeros outros detalhes, como a ABNT fazer parte da minha vida desde a primeira semana do ensino médio, devem ter me ajudado a ser essa cientista meio lapidada.

Me tornei Bacharel em Ciências Biológicas pela Universidade Estadual da Paraíba (também em Campina Grande) em 2018, tendo participado de projetos de iniciação científica em muitas áreas diferentes. Mesmo depois das infinitas aulas de botânica devo confessar que não entendo muito de plantas. Minha paixão são as minúcias, seja qual for o organismo. Deu muito certo, na verdade, porque atualmente as minúcias costumam vir acompanhadas de muitas tabelas e códigos para examinar tabelas: e foi assim que eu me apaixonei pela Bioinformática – depois de um estágio na graduação que foi para lá de fantástico no QIMR Berghofer em Brisbane/Austrália. Sobre esse estágio, só tenho a dizer: participe sempre de congressos, caro leitor invisível, você pode conhecer alguém que vá te oferecer uma oportunidade como essa.

Sendo este um pseudo-*Memorial Descritivo*, eu deveria te contar como foi que vim parar na USP: fiz a prova para saber como seria uma seleção de mestrado. No currículo acabamos vendo somente as coisas que deram certo na vida de uma pessoa, mas não é só de currículo que se constrói um profissional. A verdade é que participei de outras seleções para o mestrado, e apenas duas deram certo. A minha favorita era no Max Planck, mas fui desclassificada no meio do processo. Acontece. Para mim foi bem triste porque significou um monte de mudanças que eu não esperava que fossem definitivas, e tudo bem rápido, incluindo me mudar sozinha para São Paulo.

E por mais que eu tenha me matriculado na USP cheia de receios, tive muita sorte: escolhi e fui aceita em um laboratório cheio de colaborações e apoio. Isso se somou à minha formação principalmente porque pude interagir com pessoas com diferentes interesses, sendo muitas vezes auxiliada e algumas das vezes podendo também contribuir com suas pesquisas. Além disso, quando eu era criança brincava muito pelos corredores do IME quando acompanhava meu pai em suas aulas, e gosto de pensar que as árvores desse Instituto me conhecem há muitos anos, apesar de que nossos encontros tenham sido tão esparsados no tempo. Quem diria que eu voltaria para lá uma Bacharel em Ciências Biológicas no intuito de obter um título de Mestre em Bioinformática? Você há de concordar que tem muita beleza nessa história.

Mas nem tudo é só beleza. Um dos motivos para eu querer escrever essa carta era dizer que a Pós-Graduação não é fácil. Tive um mega preparo antes de entrar no mestrado, entrei em um grupo de pesquisa ótimo e mesmo assim me sentia terrivelmente sozinha e

empacada em muitos momentos. Claro que parte disso foi culpa de alguns professores que esquecem que a maior parte dos alunos têm apenas uma graduação e não todas as graduações possíveis dentro de Bioinformática. Por motivos como esse e tantos outros, acredito que seja necessário se iniciar um diálogo mais saudável na Universidade sobre as dificuldades enfrentadas pelos alunos e sobre a liberdade que todos temos de simplesmente jogar tudo para o alto e recomeçar.

Depois de algumas disciplinas encapetadas, diversas horas dissecando *Drosophila melanogaster* e examinando lâminas eu descobri duas coisas: a primeira foi que não gosto de experimentos de bancada e a segunda foi que eu não preciso saber de tudo dentro da Bioinformática para ainda assim gostar de trabalhar na área. A verdade nas entrelinhas é que, mesmo após essas 150 páginas, ainda existem muitas dúvidas sobre minhas reais habilidades.

Idealmente um *Memorial Descritivo* contaria com uma listagem das publicações e títulos honoríficos, mas se você leitor invisível tiver interesse, pode conferir tudo isso no meu currículo Lattes. Cada um desses pontos foi fruto de muita dedicação, de risadas e lágrimas e, algumas vezes, de querer simplesmente desistir ou terminar de vez. E cada um dos nomes que aparecem junto contam, mesmo os extra-oficiais – por isso a seção Agradecimentos é extremamente importante nesse trabalho.

Dediquei com exclusividade praticamente três anos ao mestrado. E cerca de dois terços desse tempo em uma pandemia. Gostaria de deixar aqui registrado os meus sinceros sentimentos por tudo que temos vivido, por todos que foram perdidos. Vivo em quarentena desde março de 2020 (Ainda é verdade é novembro de 2021). E, como cientista e cidadã brasileira, eu não poderia deixar de mencionar minha admiração pelo SUS e, em contrapartida, o desprezo por Aquele-Que-Não-Merece-Nem-Mesmo-Ser-Nomeado e seus seguidores, uma analogia à Voldemort e aos Comensais da Morte de JK Rowling.

Por fim, você precisa se lembrar sempre que por trás de cada pesquisa científica existe uma gama de pessoas, cada qual com suas histórias. Eu sou apenas uma das histórias por trás desta pesquisa e esta pesquisa é apenas um pedaço por trás da minha história. Quem sabe, leitor invisível, um dia as nossas histórias não se cruzem. Se é que já não se cruzaram!

Um grande abraço,

Isabela Almeida