Gabriel Nassar Reich Goldstein

Identificação de genes novos de *Drosophila* utilizando *machine learning*

Identifying *Drosophila* new genes using machine learning

Orientadora: Profa. Dra. Maria Dulcetti Vibranovski

São Paulo

2022

Gabriel Nassar Reich Goldstein

Identificação de genes novos de *Drosophila* utilizando *machine learning*

Identifying *Drosophila* new genes using machine learning

Exemplar Corrigido de Tese apresentada ao Instituto de Biociências da Universidade de São Paulo, para a obtenção de Título de Doutor em Genética e Biologia Evolutiva.

Orientador(a): Profa. Dra. Maria Dulcetti Vibranovski

São Paulo

2022

Goldstein, Gabriel Nassar Reich
Identificação de genes novos de Drosophila utilizando machine learning / Gabriel Nassar Reich Goldstein ; orientadora Maria Dulcetti Vibranovski -- São Paulo, 2022.
88 p.

Tese (Doutorado) -- Instituto de Biociências da Universidade de São Paulo. Programa de Pós-Graduação em Genética e Biologia Evolutiva.

1. Genes novos. 2. Machine learning. 3. Transcriptômica. I. Vibranovski, Maria Dulcetti, orient. II. Título.

Comissão Julgadora:

_____	_____
Prof(a). Dr(a).	Prof(a). Dr(a).
_____	_____
Prof(a). Dr(a).	Prof(a). Dr(a). Maria Dulcetti Vibranovski Orientador(a).

Agradecimentos

À minha orientadora profa. Dra. Maria D. Vibranovski, pelo apoio, carinho e colaboração durante todo este projeto de doutorado. Espero que você já saiba o quanto eu te admiro e aprecio seu carinho, sua inteligência e sua dedicação, mas não custa repetir que tive muita sorte de ter você com orientadora. Suas ideias e conhecimentos foram essenciais não só para a execução deste projeto, mas para minha formação como cientista que espero carregar para sempre.

Ao meu orientador na China, o prof. Dr. Yong E. Zhang não só pela imensa contribuição científica neste projeto, mas também por me receber de braços abertos em seu laboratório em Pequim. Suas contribuições geraram grandes mudanças positivas neste trabalho e sua dedicação sempre me servirá como exemplo.

À FAPESP (processos 2016/09378-4 e 2018/12585-7) e o CNPq, instituições que deram o apoio financeiro e institucional que tornaram possível a execução deste projeto e a minha dedicação exclusiva a este doutorado.

À USP e em especial a pós-graduação em Genética e Biologia Evolutiva, pelo apoio institucional e as diversas ajudas que me foram dadas durante estes anos como aluno da USP.

Aos membros do laboratório no Brasil e na China, cujos comentários e ajudas tornaram possível diversas áreas deste projeto. Não só de ciência é feito um laboratório, então agradeço também o convívio e as amizades que me ajudaram demais, principalmente nos momentos mais difíceis. Em especial, gostaria de agradecer aos amigos Eduardo Dupim, Camila Avelino, Carolina de Athayde Mendonça, Frederico Monfardini e Isabela Almeida.

Aos outros colaboradores deste projeto, que cederam moscas ou sequências, em especial o prof. Dr. Antônio Bernardo de Carvalho e o Dr. Nicholas Van Kuren.

Aos grandes amigos que levo na minha vida Gabriel Khattar, Rodrigo Balbi, Felipe Costa, Gabriel Faerstein. Carrego sempre o amor e o carinho por vocês no coração, e sempre penso na sorte que tenho de ter vocês na minha vida.

Aos meus pais Sandra e Nilton, por me apoiaram tanto financeiramente quanto emocionalmente nesta escolha difícil que é a vida acadêmica. Posso dizer com certeza que muitas das minhas melhores características puxei de vocês, e só algumas das ruins! Obrigado de verdade por acreditar em mim e nesse sonho estranho, além de tentar entender minhas explicações obtusas sobre o que estou fazendo.

À Letizia Maria Taboada Roa, minha namorada e companheira. Sempre te digo que minha maior sorte na minha vida foi ter te encontrado, mas repito agora por escrito só para ter certeza que ficará registrado. Só tenho a agradecer estes anos de carinho e amor e sei com certeza que este trabalho não seria possível sem seu apoio nas minhas horas mais difíceis.

Sumário

Resumo

Existe uma classe de genes que surgiram recentemente na história de um táxon: os genes novos. Estes genes são assim classificados pois, apesar da sua presença em um táxon, estão ausentes em um táxon irmão e grupos externos. Para identificar genes novos em um genoma é necessário datar todos os genes de uma espécie focal em relação ao ponto na filogenia do táxon no qual cada gene se originou. O principal método de datação de genes para identificação de genes novos utiliza sintenia e parcimônia ao comparar genomas de espécies relacionadas para datar todos os genes de uma espécie focal. Apesar da precisão do método, ele é extremamente dependente da montagem e anotação do genoma de interesse, o que limita sua aplicação para espécies modelo que tem uma anotação manual e curada. Existem uma série de características biológicas que são sabidamente diferentes entre genes novos e antigos em uma grande diversidade de táxons analisados, como humanos, camundongos e plantas. Um exemplo disso é o perfil de expressão destes grupos, já que genes novos se expressam majoritariamente na gametogênese masculina e genes antigos são expressos de maneira generalista. Com estes fatos em mente, propomos neste trabalho um método de identificação de genes novos que utiliza informações biológicas para separar genes novos de antigos por meio do uso de *machine learning*. Para isso, coletamos informações de bases de dados e geramos informações de dados de expressão, ortologia e *dn/ds* para *D. melanogaster*, a espécie do gênero que teve seus genes novos datados e possibilita o treinamento de um modelo de *machine learning* supervisionado. Além destas informações, utilizamos dados de ortologia para eliminar genes antigos enquanto perdemos poucos genes novos. Isto é possível pois genes antigos possuem, em média, mais espécies com ortólogos do que genes novos já que surgiram antes na história evolutiva do táxon. Primeiramente testamos se as informações de bases de dados seriam capazes de informar um modelo de *machine learning* que separasse genes novos de antigos. Para isto, geramos diversos modelos com diferentes níveis de complexidade e combinações diferentes de variáveis, chegando a um modelo que teve 0.702 de *precision* (fração de instâncias relevantes entre as instâncias recuperadas) e 0.733 *recall* (fração de instâncias relevantes que foram recuperadas). Após esta etapa, precisávamos gerar um modelo que se aproximasse à realidade esperada em espécies sem informações disponíveis em bases de dados, como *D. melanogaster*. Assim, fizemos testes semelhantes com diferentes conjuntos de variáveis, no entanto, utilizamos dados que nós mesmos geramos neste trabalho. Após a realização destes testes, geramos um modelo com 0.508 de *precision* e 0.718 de *recall*, demonstrando que é possível, mesmo com dados gerados em experimentos próprios, identificar e classificar genes novos em *D. melanogaster*. Para verificar se o método que estamos propondo funciona em outras espécies do gênero *Drosophila*, datamos os genes de outra espécie para identificar seus genes novos. Utilizamos o método baseado em sintenia e parcimônia na espécie *D. pseudoobscura* e identificamos 1523 genes novos e 12648 genes antigos.

Abstract

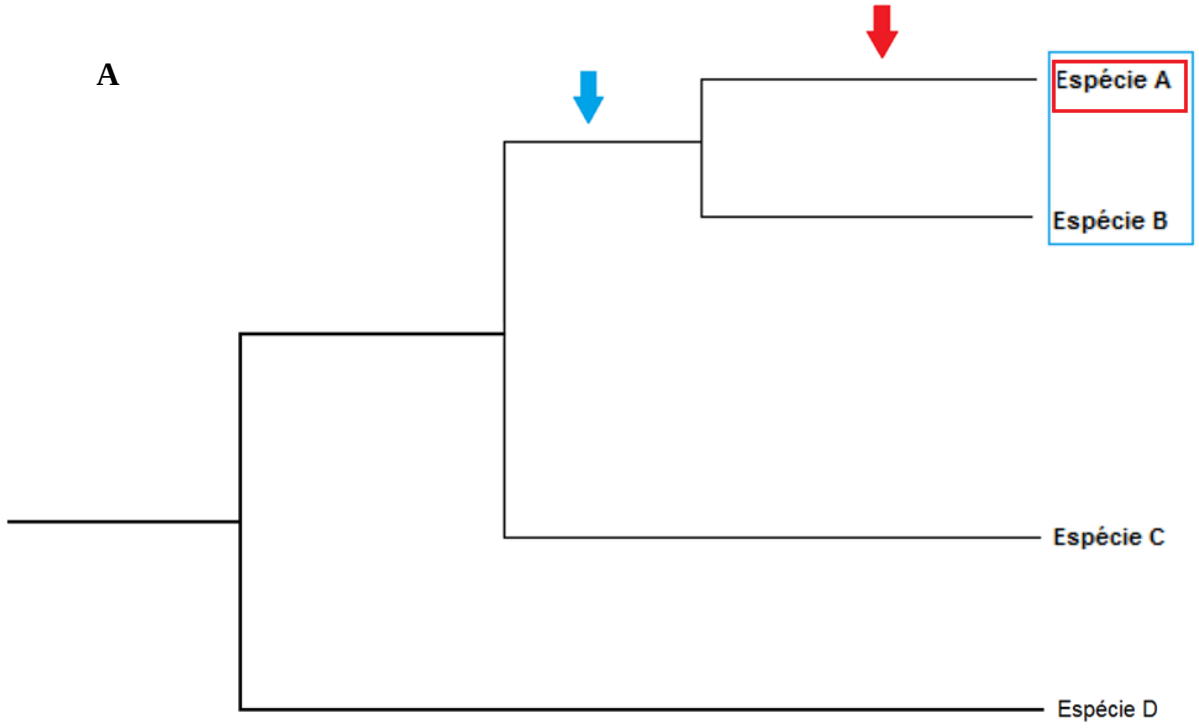
There is a class of genes that emerged recently in the history of a taxon: new genes. These genes are so classified because, despite their presence in a taxon, they are absent in a sister taxon and outgroups. To identify new genes in a genome it is necessary to date all genes of a focal species to the point in the phylogeny of the taxon in which each gene originated. The main gene dating method for identifying new genes uses synteny and parsimony when comparing genomes of related species to date all genes of a focal species. Despite the precision of the method, it is extremely dependent on the assembly and annotation of the genome of interest, which limits its application to model species that have a manual and curated annotation. There are a number of biological characteristics that are known to differ between new and old genes in a wide range of analyzed taxa, such as humans, mice and plants. An example of this is the expression profile of these groups, since new genes are mostly expressed in male gametogenesis and old genes are expressed in a general way. With these facts in mind, we propose in this work a new gene identification method that uses biological information to separate new genes from old ones through the use of machine learning. For this, we collected information from databases and generated expression, orthology and *dn/ds* data information for *D. melanogaster*, the species of the genus that had its new genes dated and makes it possible to train a supervised machine learning model. In addition to this information, we use orthology data to eliminate old genes while losing few new genes. This is possible because old genes have, on average, more species with orthologs than new genes, since they appeared earlier in the evolutionary history of the taxon. First, we tested whether information from databases would be able to inform a machine learning model that would separate new genes from old ones. For this, we generated several models with different levels of complexity and different combinations of variables, reaching a model that had 0.702 precision (fraction of relevant instances among retrieved instances) and 0.733 recall (fraction of relevant instances that were retrieved). After this step, we needed to generate a model that approximated the reality expected in species without information available in databases, such as *D. melanogaster*. So, we did similar tests with different sets of variables, however, we used data that we generated ourselves in this work. After performing these tests, we generated a model with 0.508 precision and 0.718 recall, demonstrating that it is possible, even with data generated in our own experiments, to identify and classify new genes in *D. melanogaster*. To verify whether the method we are proposing works in other species of the Drosophila genus, we date the genes of another species to identify its new genes. We used the method based on synteny and parsimony in the species *D. pseudoobscura* and identified 1523 new genes and 12648 old genes.

1. Introdução

1.1. Genes novos

Considerando a diversidade de seres vivos que surgiram a partir de um só ancestral comum, somos obrigados a nos perguntar como tantas características biológicas surgiram ao longo do tempo. Esta é uma questão importante no estudo da biologia evolutiva: como surgem as novidades evolutivas que definem a diversidade biológica que observamos? Existem mecanismos que contribuem para o surgimento de novidades evolutivas, como elementos móveis (Astrid et. al., 2008) e mudanças nas regiões regulatórias (Koshikawa et. al., 2015). No entanto, esta pergunta pode ser parcialmente respondida pelo surgimento de genes novos que exercem novas funções em um organismo. Apesar de sabermos que genes novos representam apenas uma fração da geração de diversidade biológica, é importante entender o surgimento, evolução e papel biológico destes genes para formar uma imagem completa deste processo.

Genes novos são definidos filogeneticamente, sendo genes que estão presentes em um táxon e ausentes no seu táxon irmão e no grupo externo (Kaessman, 2010). Uma consequência da definição de genes novos é que ela é filogenética, e não necessariamente relacionada ao tempo cronológico. Isto significa que a idade cronológica de dois genes de espécies diferentes pode ser igual, mas um ser considerado um gene novo e outro não dependendo da filogenia de cada grupo. Da mesma maneira, um gene pode ser mais velho do que outro, mas ambos se encaixam na categoria de genes novos. A figura 1 dá um exemplo fictício de dois genes novos com idades diferentes e um exemplo real dos genes novos *Artemis* e *Apollo*.



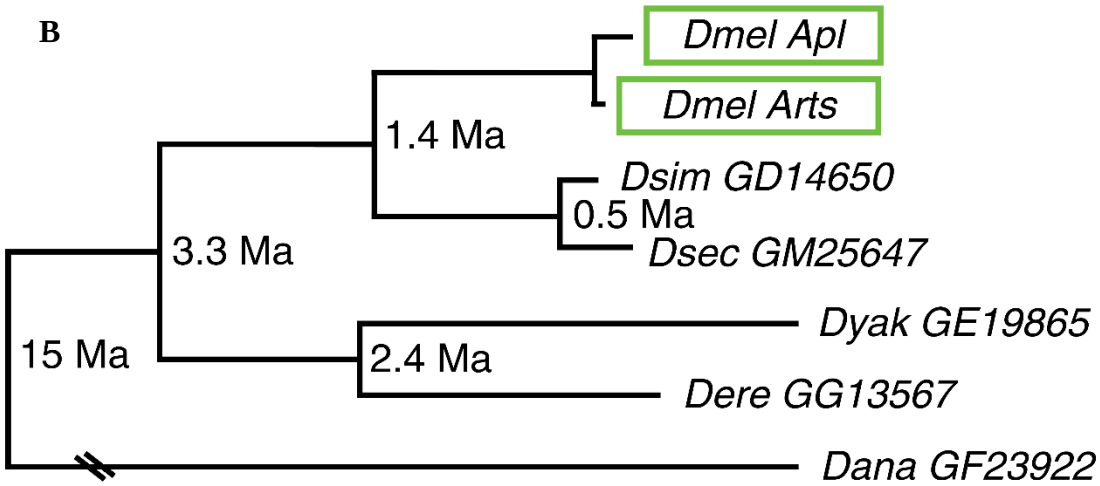


Figura 1: Genes novos em contexto filogenético fictício (A) e real (B). A) O gene vermelho surge após as separações das espécies A e B, portanto, é um gene novo exclusivo da espécie A. O gene azul surge antes da separação do grupo que tem as espécies A e B, mas pode ser considerado um gene novo deste táxon. A presença de grupos externos à espécie focal permite diferenciamos os ganhos das perdas de genes. Por exemplo, se o gene vermelho tivesse sido perdido na espécie B e não fosse um gene novo, estaria presente em um dos grupos externos. **B)** Os genes *Apollo* (Apl) e *Artemis* (Arts) surgem a partir de uma duplicação em tandem de uma região do cromossomo 3L de *D. melanogaster*. Em todas as outras espécies representadas, existe apenas um gene ortólogo, que é então duplicado em *D. melanogaster*. Os números representam o tempo em milhões de anos (Ma) e as abreviações as diferentes espécies de *Drosophila*, como *Dmel*, *D. melanogaster; Dsim*, *D. simulans*; *Dsec*, *D. sechellia; Dyak*, *D. yakuba*; *Dere*, *D. erecta*; *Dana*, *D. ananassae* (adaptado de VanKuren e Long, 2018).

Desde a década de 1990 o conhecimento sobre genes novos aumenta junto da quantidade de estudos realizados na área. Os primeiros trabalhos estavam focados em descrever casos específicos de genes novos, como o trabalho de Long e Langley em 1993 sobre o gene *Jingwei* (Long e Langley, 1993). Neste artigo, os autores descrevem o gene identificado a partir de uma sequência que se acreditava ser pseudogênica e sua formação a partir de uma sequência quimérica. Este gene, localizado no cromossomo 3 das espécies *D. teisseri* e *D. yakuba*, surgiu a partir de um mRNA processado do gene *Adh* com o recrutamento de outros éxons que pertencem ao gene *Yellow emperor* (Long e Langley, 1993, Long *et. al.*, 1999). Estudos como este examinaram genes novos caso a caso, examinando facetas como mecanismos de origem, padrões de expressão e quais forças evolutivas estão agindo sobre esta classe de genes.

Avanços tecnológicos possibilitaram estudos de genômica e transcriptômica, com quantidades massivas de dados por artigo. Com os estudos de genes novos não foi diferente, onde antes se estudavam casos individuais passou a ser possível olhar para todos os genes novos de determinada espécie e examinar padrões biológicos mais amplos. Zhang e colaboradores, por exemplo, utilizaram genomas de 12 espécies do gênero *Drosophila* para datar os genes de *D. melanogaster* e encontrar seus genes novos. Os autores deste trabalho alinharam os genomas disponíveis na época ao genoma mais bem montado do gênero, o da espécie *D. melanogaster* e, assim, verificam a presença de ortólogos de cada gene em cada uma das espécies utilizadas. Assim, foi possível determinar em qual ramo da filogenia que leva a espécie focal cada gene se originou e sua idade estimada em milhões de anos. Os detalhes deste método serão explicados em maior detalhe posteriormente, mas este trabalho serve para destacar a importância do avanço tecnológico no estudo de genes novos (Zhang *et. al.*, 2010).

Apesar da expansão do conhecimento sobre genes novos, ainda existem questões importantes na área que não foram resolvidas, como a função exercida por estes genes, por exemplo. Alguns trabalhos vêm apresentando as possíveis funções biológicas destes genes com o auxílio de tecnologias de edição genética como CRISPR/CAS-9. Um caso como este é apresentado no artigo de VanKuren e Long de 2018 (VanKuren e Long, 2018), no qual os autores demonstraram que o

par de genes duplicados em tandem *Apollo* e *Artemis* podem estar resolvendo um caso de antagonismo sexual na espécie *D. melanogaster*. Ou seja, o gene *Apollo* é essencial para a fertilidade do macho, mas tem efeito negativo na fertilidade de fêmeas enquanto o oposto é verdadeiro para *Artemis*. Assim, após a duplicação e o efeito da seleção natural, cada gene perdeu a capacidade de ser expresso nos tecidos nos quais teriam efeitos negativos.

Outro trabalho utilizando tecnologia CRISPR/CAS-9 para investigar a função de um gene novo publicado por Jiang e colaboradores (Jiang *et. al.*, 2017) demonstrou que o novo retrogene *Rpl10l* é essencial para a fertilidade de machos em camundongos. Este gene surgiu a partir da retrotranscrição do gene *Rpl10*, que está no cromossomo X. Os autores foram capazes de demonstrar que o retrogene *Rpl10l* é essencial para a divisão meiótica em machos ao compensar o silenciamento do gene *Rpl10* durante a meiose.

Apesar destes avanços, apenas uma quantidade pequena de espécies teve todos os seus genes datados, possibilitando a identificação dos genes novos. Durante este trabalho, vamos detalhar as diferenças biológicas entre genes novos e antigos, assim como a importância destas diferenças para que alcancemos o nosso objetivo, que é propor uma nova maneira de identificar genes novos a partir de características biológicas com o uso de *machine learning*.

1.2. Como surgem os genes novos

Existem um grande número de processos genéticos que podem levar ao surgimento de genes novos, como duplicações mediadas por DNA, retrotransposição, transferência lateral e surgimento *de novo* (Long *et. al.*, 2003). É importante entender o funcionamento destes mecanismos no estudo de genes novos, pois o surgimento do gene afeta sua história evolutiva, podendo determinar quais barreiras o gene deve ultrapassar para se fixar em uma população. Por exemplo, um gene resultante de uma duplicação mediada por DNA pode ter a região promotora do seu gene parental, que é o gene antigo a partir do qual surgiu o gene novo, mas isso não é possível para um retrogene.

Os retrogenes surgem a partir da retrotransposição do mRNA processado de um gene pré-existente, e a reinserção deste novo fragmento de DNA em uma região do genoma (Zhou *et. al.*, 2008). Já que o mRNA processado já sofreu *splicing*, estes genes não possuem os íntrons do gene original e, como normalmente são inseridos em outras regiões do genoma, também não compartilham a região regulatória. Por estes fatores, é mais fácil determinar quem é o gene parental e o gene filho nestes casos, já que o gene derivado não possui os íntrons compartilhados pelo gene original e seus ortólogos em outras espécies.

Outro processo que faz com que o gene novo não compartilhe a região regulatória com a sequência original é a origem *de novo*. Estes genes surgem a partir de sequências não gênicas, ou seja, partes do genoma que anteriormente não produziam mRNA ou proteínas e, portanto, não possuíam promotores (Zhou *et. al.*, 2008).

O mecanismo mais comum, no entanto, são as duplicações mediadas por DNA (figura 2), que podem ocorrer por ação de transposons e eventos recombinatórios, entre outros. Estas duplicações podem ser em tandem, ou seja, a nova cópia está do lado da antiga e tem tamanho variável, podendo englobar apenas um gene ou grandes pedaços do genoma.

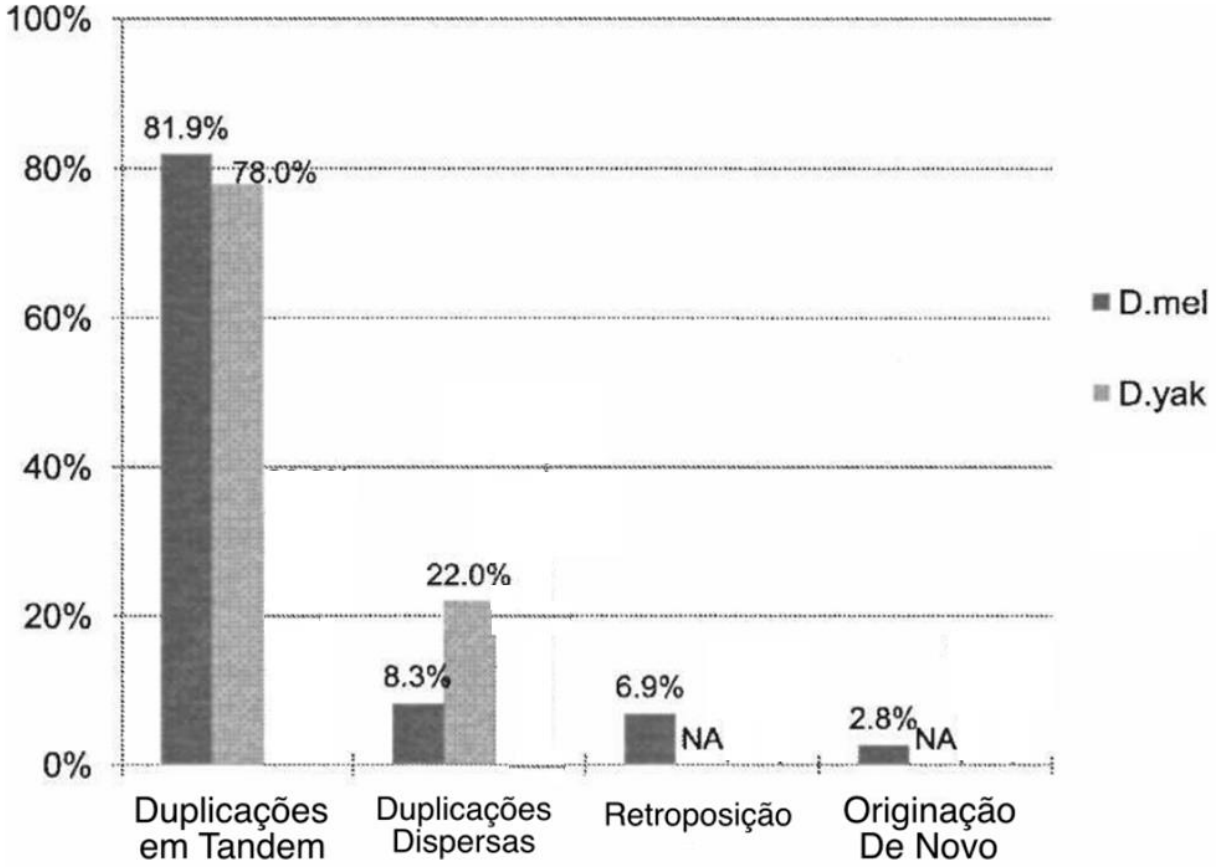


Figura 2: Distribuição de genes novos por seu mecanismo de origem de duas espécies de *Drosophila*. O mecanismo mais comum de originação de genes novos é a duplicação, seja em tandem ou não. *D. mel* representa *D. melanogaster* e *D. yak*, *D. yakuba*, uma espécie próxima (adaptado de Zhou *et. al.*, 2008).

Para a maioria dos genes novos que acabaram de surgir, o caminho evolutivo mais provável é a pseudogenização e o seu subsequente desaparecimento. Se um gene novo não possui uma região regulatória ou é inserido em algum lugar do genoma no qual não consegue ser expresso ele não exercerá nenhuma função, e estará livre para acumular mutações deletérias e desaparecer rapidamente do genoma. Isto é especialmente comum com genes derivados de retrotransposição e mecanismos *de novo*, que não terão como carregar a região regulatória de genes parentais. (Long *et. al.*, 2003).

Os genes novos recém-formados derivados de duplicações que compartilham a região regulatória com seus genes parentais podem conseguir se expressar, mas como são redundantes com seus genes parentais também podem acumular mutações deletérias que os levarão para a pseudogenização (Long *et. al.*, 2003). No entanto, apesar do caminho mais provável destes genes que acabaram de surgir seja sua transformação em pseudogene e subsequente desaparecimento, existem algumas maneiras destes genes sobreviverem a este momento inicial.

Uma delas é o recrutamento de novas regiões regulatórias, ou seja, um gene novo pode se relocar para uma outra localização do genoma que possui promotores e passar a ser expresso em outros tecidos. Outra possibilidade é que mutações aleatórias em um gene duplicado o possibilitem exercer outra função diferente do gene parental, ambos estes processos são chamados de neofuncionalização (Kaessman, 2010). A figura abaixo mostra algumas maneiras de um gene novo se fixar em uma população a partir do seu surgimento em um genoma de um indivíduo.

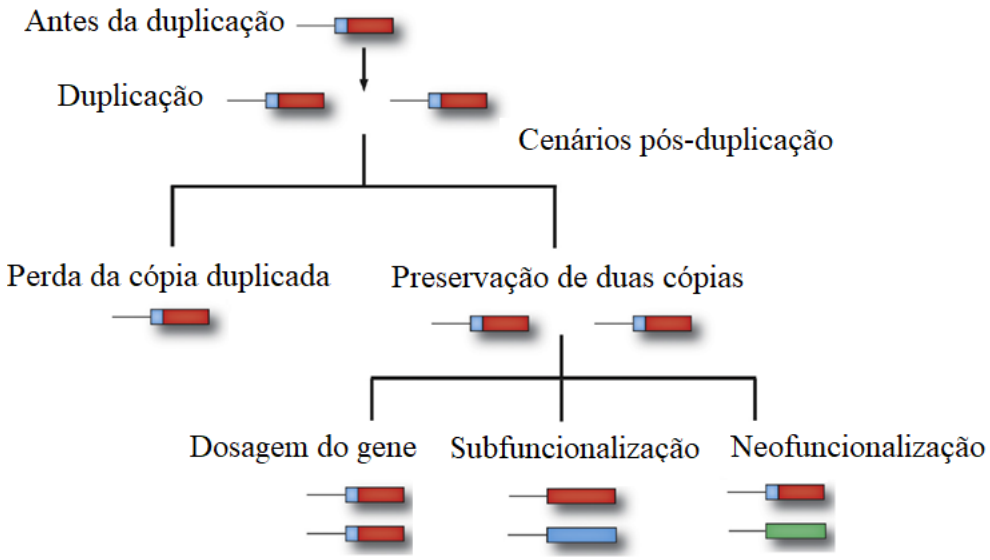


Figura 3: Esquema representando possíveis caminhos de um gene após um evento de duplicação. O caminho mais comum é a perda da cópia duplicada após o acúmulo de mutações aleatórias, mas existem outras possibilidades para a manutenção de ambas as cópias. Cada cor representa uma função exercida por um gene, de maneira que o gene com duas cores realiza duas funções. Após a subfuncionalização, por exemplo, cada cópia mantém apenas uma das duas funções ancestrais, antes exercidas por um só gene (adaptado de Voordeckers e Verstrepen, 2015).

1.3.Ortologia

Para estudar a evolução de genes novos e compreender a evolução de genomas, é necessário esclarecer os conceitos de ortologia e paralogia, assim como as implicações dos métodos de classificação e identificação de genes relacionados.

Para dois genes serem considerados ortólogos, eles devem estar localizados em espécies diferentes e ser derivados de um só gene no ancestral comum destas duas espécies. Ou seja, dois genes ortólogos estão, necessariamente, localizados em duas espécies diferentes e surgiram a partir do mesmo gene, se diferenciando através do processo de especiação.

Em contrapartida, genes parálogos são aqueles que surgiram a partir de uma duplicação de um gene dentro do mesmo genoma. Estes genes podem estar localizados na mesma espécie, ou seja, derivados de uma duplicação mais recente, ou em espécies diferentes no caso de duplicações mais antigas (Fitch, 1970). A figura 4 mostra um esquema de como funciona paralogia e ortologia.

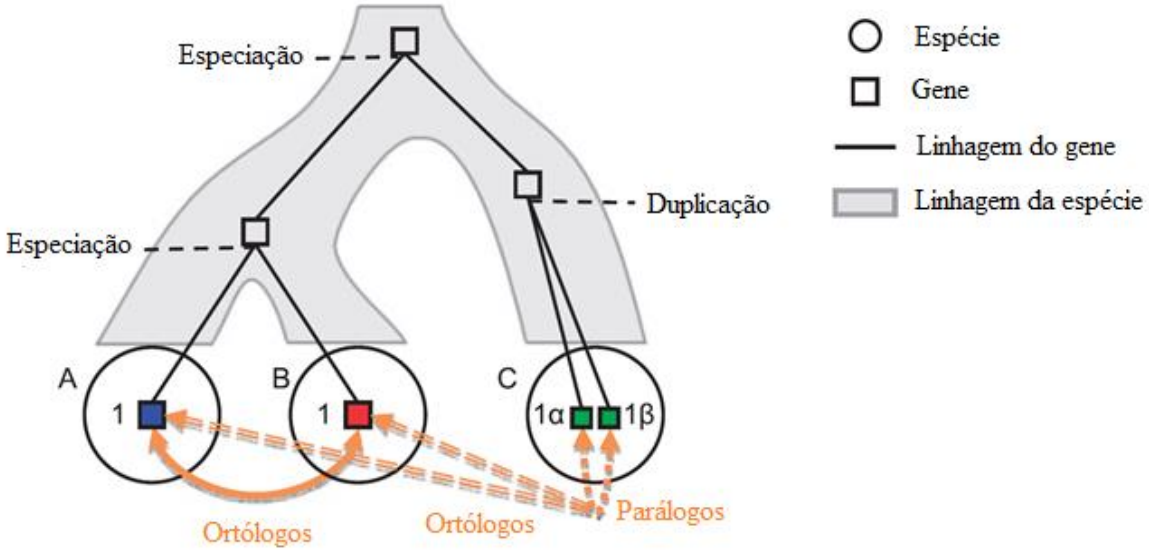


Figura 4: Surgimento de genes ortólogos e parálogos em uma filogenia de três espécies. Os genes azul e vermelho são ortólogos entre si e dos genes verdes, já que se separaram através de eventos de especiação. Os genes verdes são parálogos entre si, já que derivam de um evento de duplicação dentro da mesma espécie (adaptado de Koonin *et. al.*, 2011).

Existem diversas maneiras de identificar ortólogos entre espécies, mas a maioria delas se baseia em uma busca inicial por similaridade. Após encontrar estes grupos de genes que são semelhantes entre si, cada método progride de um jeito diferente, de acordo com seus algoritmos proprietários.

O método *reciprocal best hits* (*rbh*), por exemplo, encontra ortólogos ao fazer uma busca de BLAST recíproca, ou seja, A x B e B x A e selecionar os pares de genes que são as melhores correspondências entre si nas duas buscas. Apesar de ser capaz de encontrar estes pares de genes ortólogos, este método não consegue identificar relacionamentos entre mais de dois genes e encontra problemas quando os genes fazem parte de famílias gênicas complexas ou possuem parálogos semelhantes, já que a reciprocidade pode não ocorrer (Koonin *et. al.*, 2011).

Outro método importante é aquele utilizado pela base de dados OrthoDB, que também utiliza uma busca por similaridade com *reciprocal best hits* mas é mais complexo, passando por outras etapas de processamento para chegar a seu resultado final (Kriventseva *et. al.*, 2019).

De maneira geral, podemos classificar relacionamentos de ortologia em três tipos: um para um, um para muitos e muitos para muitos. O primeiro tipo é aquele no qual se encontra apenas uma cópia do gene em cada espécie; o segundo tipo ocorre quando um gene é ortólogo de mais de um gene em outra espécie e o terceiro é identificado quando existem múltiplos ortólogos em cada espécie (Howe *et. al.*, 2021). Estes tipos de ortologia estão representados na figura 5 abaixo, nos quadros um, dois e três respectivamente.

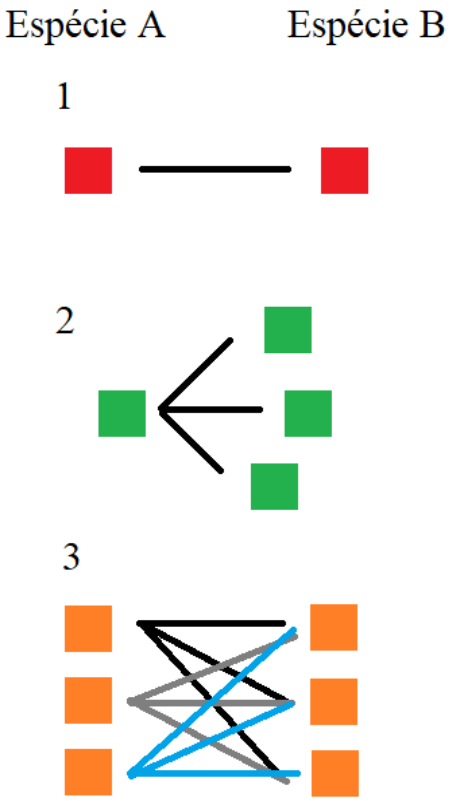


Figura 5: Tipos principais de ortologia entre duas espécies, A e B. Cada cubo representa um gene, enquanto as linhas representam as relações de ortologia. O quadro 1 representa as ortologias um para um, 2 as ortologias um para muitos, enquanto o quadro 3 demonstra as ortologias muito para muitos.

A compreensão destes tipos de ortologia é importante para os estudos de genes novos em geral, e para nosso trabalho especificamente, como será explicado posteriormente. De maneira geral, os tipos de ortologia que não são de um para um podem agrupar genes novos com seus genes parentais, principalmente se tiverem alta similaridade. Este agrupamento pode causar erros na identificação de relações de ortólogos, e pode gerar problemas quando dependemos deste tipo de análise.

1.4.Como se identificam genes novos?

Um dos métodos mais utilizados de identificação de genes novos é feita através do método de datação de genes estabelecido por Zhang e colaboradores em 2010, que por sua vez é uma melhora do método proposto por Zhou e colaboradores em 2008 (Zhang *et. al.*, 2010; Zhou *et. al.*, 2008). Ambos os métodos utilizam a busca de regiões sintênicas e o uso de parcimônia para datar genes de acordo com seu surgimento na filogenia do táxon examinado.

Sintenia é o fenômeno da conservação da organização de uma região do genoma através das espécies, ou seja, fragmentos ortólogos estão presentes na mesma região e na mesma ordem em espécies relacionadas. Esta conservação é comum em pequenos blocos, também chamado de micro-sintenia, já que rearranjos cromossômicos são raros e costumam ocorrer em larga escala, preservando a ordem de genes em pequena escala (Heger e Ponting, 2007). A figura 6 mostra a sintenia de uma região genômica em plantas.

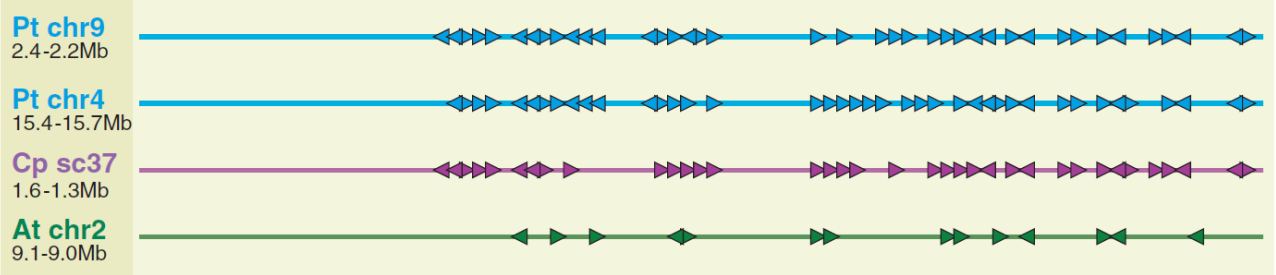


Figura 6: Sintenia entre espécies de três gêneros de plantas. Os genes estão representados por triângulos, com sua orientação demonstrando sua direção transcricional, sendo incluídos apenas os genes com sintenia. Os gêneros representados são *Arabidopsis* (verde), *Carica* (magenta) e *Populus* (azul) (adaptado de Tang *et. al.*, 2008).

Para a datação dos genes em um genoma focal, é utilizado o alinhamento deste genoma com os de espécies relacionadas, feito com BLASTZ e processado posteriormente por outros programas. No caso de espécies modelo como *D. melanogaster*, os arquivos necessários já existem e estão disponíveis na base de dados UCSC (Kent *et. al.*, 2002).

Com estes alinhamentos, é verificada a distribuição de ortólogos de cada gene nas diferentes espécies examinadas, ou seja, para cada gene procura-se a manutenção de sua micro-sintenia da espécie focal com espécies relacionadas. Assim, é possível identificar genes que não possuem sintenia em outras espécies, ou seja, que estão causando uma quebra de sintenia em uma região do genoma, marcando-o como candidato a gene novo (Zhang *et. al.*, 2010). A figura abaixo é um esquema da comparação de sintenia entre *D. melanogaster* e duas espécies relacionadas.

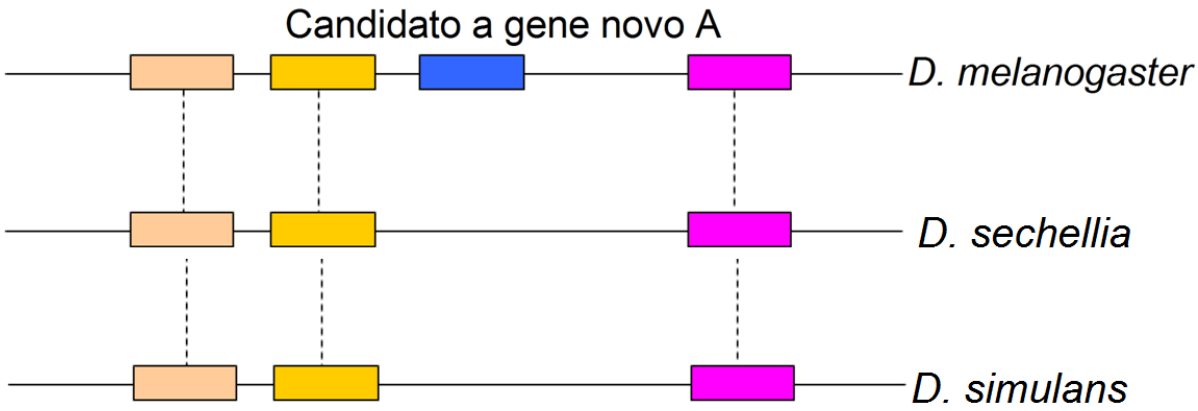


Figura 7: Esquema de comparação de regiões sintênicas do genoma de *D. melanogaster* com espécies relacionadas. Os genes são representados por blocos coloridos e a conservação de sintenia por linhas pontilhadas. O gene "A" em azul é um candidato a gene novo pois está ausente nas outras espécies examinadas (adaptado de Zhang *et. al.*, 2010).

Após a comparação destes alinhamentos, é feita uma contextualização filogenética dos resultados encontrados, assim, o gene é atribuído um grupo de idade dependendo de quais espécies ele possui ortólogo e das distâncias entre estas espécies e a espécie focal. Por exemplo, se um gene possui ortólogo nos grupos externos ele é considerado um gene antigo, e é atribuído a este grupo, enquanto um gene que não possui ortólogo em nenhuma outra espécie é um gene exclusivo e faz parte do grupo de genes mais novos (Zhang *et. al.*, 2010).

Em *D. melanogaster* Zhang e colaboradores utilizaram 12 espécies de *Drosophila* para separar os genes em grupos de 0 a 6. Fazem parte do grupo 0 os genes antigos, presentes nos grupos externos analisados, enquanto os genes novos estão nos grupos 1 a 6. A figura 8 mostra o resultado final da datação de genes de *D. melanogaster* com a distribuição em grupos de idade.

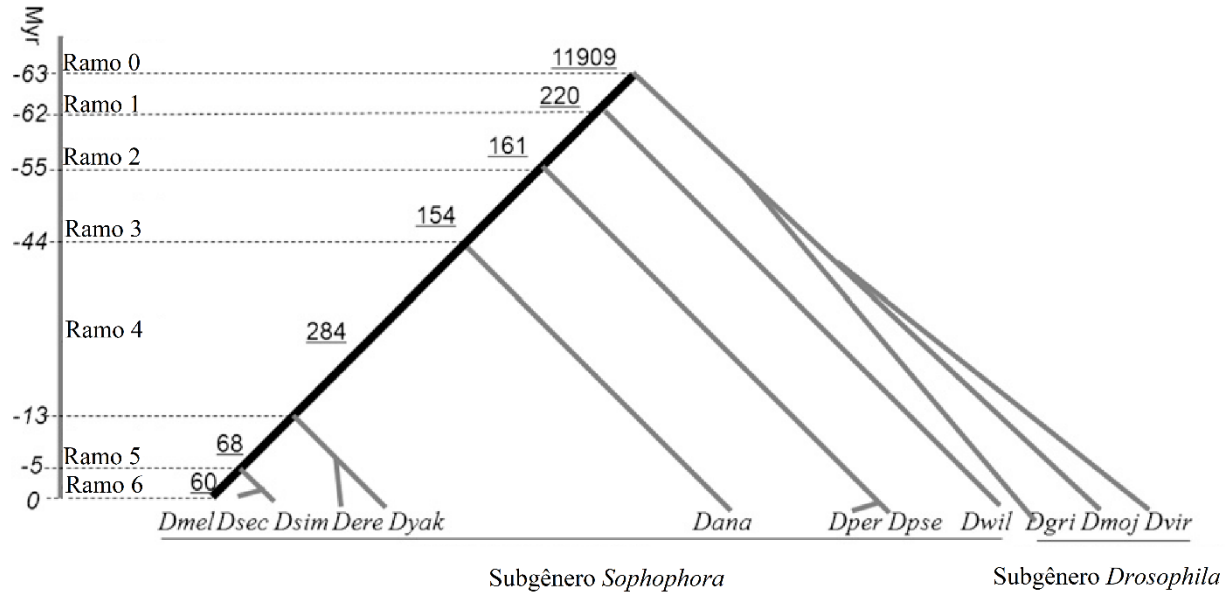


Figura 8: Filogenia do gênero *Drosophila* com as 12 espécies examinadas. Os genes encontrados em cada ramo estão representados pelos números ligados à filogenia, enquanto a escala mostra o tempo evolutivo em milhões de anos (do inglês Myr). Os genes estão separados em grupos de 0 a 6, sendo que os genes antigos fazem parte do grupo 0 e os genes novos do grupo 1 a 6. Os genes são classificados de acordo com a presença de seus ortólogos em outras espécies. Por exemplo, genes do "Ramo 6" não possuem ortólogos em nenhuma outra espécie examinada (adaptado de Zhang *et. al.*, 2010).

Como este método de datação de genes utiliza o alinhamento de genomas e a busca de ortólogos por gene, é necessário que o genoma da espécie focal e de espécies relacionadas estejam sequenciados e anotados. Além disso, tanto a montagem quanto a anotação do genoma da espécie focal precisam ser extremamente bem-feitas já que problemas nestas etapas levam a erros na datação de genes. Por causa destes fatores, a datação de genes foi feita apenas em um pequeno número de espécies modelo (Zhang *et. al.*, 2010 e Shao *et. al.*, 2019).

Após a identificação de genes novos, podemos calcular a taxa de surgimento destes genes em determinado grupo filogenético para associá-los a características particulares do táxon. No entanto, como genes novos são identificados em espécies modelo, taxas como esta tem que ser calculadas utilizando apenas estas espécies.

Com isto, estas taxas são calculadas apenas para o ramo da filogenia que leva a esta espécie, fazendo com que exista um viés na análise e com que seja difícil generalizá-la para todo o grupo filogenético. Assim, não é possível comparar a taxa de aquisição de genes novos de espécies próximas e fica mais difícil associar estas taxas com padrões evolutivos e biológicos.

É importante lembrar que este método de identificação de genes novos depende da qualidade da montagem e anotação do genoma da espécie focal. Portanto, mesmo se identificarmos os genes novos de espécies próximas a espécies modelo só será possível comparar estas listas de genes novos e estas taxas calculadas se as qualidades dos genomas forem semelhantes.

1.5. Expressão de genes novos e antigos

Com a evolução dos estudos sobre genes novos, foi possível observar padrões nas características biológicas de genes novos e genes antigos. Um dos padrões encontrados foi a diferença entre os

perfis de expressão destas duas classes de genes, que se repete em um grande número de táxons (Kaessman, 2010; Witt *et. al.*, 2021).

Primeiramente, genes novos são expressos em poucos tecidos, enquanto genes antigos possuem expressão ampla, sendo ativos em um grande número de tecidos (Zhang *et. al.*, 2010). É importante frisar que existem genes antigos que são tecido-específicos, mas estes são a minoria e não o padrão geral para genes antigos.

A explicação mais provável para esta diferença é o fato de que muitos genes antigos realizam funções essenciais em todos os tecidos, sendo chamados de *housekeeping*. Genes novos, no entanto, surgiram recentemente na história do táxon examinado, e raramente realizam funções generalistas (Kaessman, 2010). A figura 9 divide os genes humanos em 13 grupos de idade e demonstra o número de tecidos nos quais os genes possuem expressão.

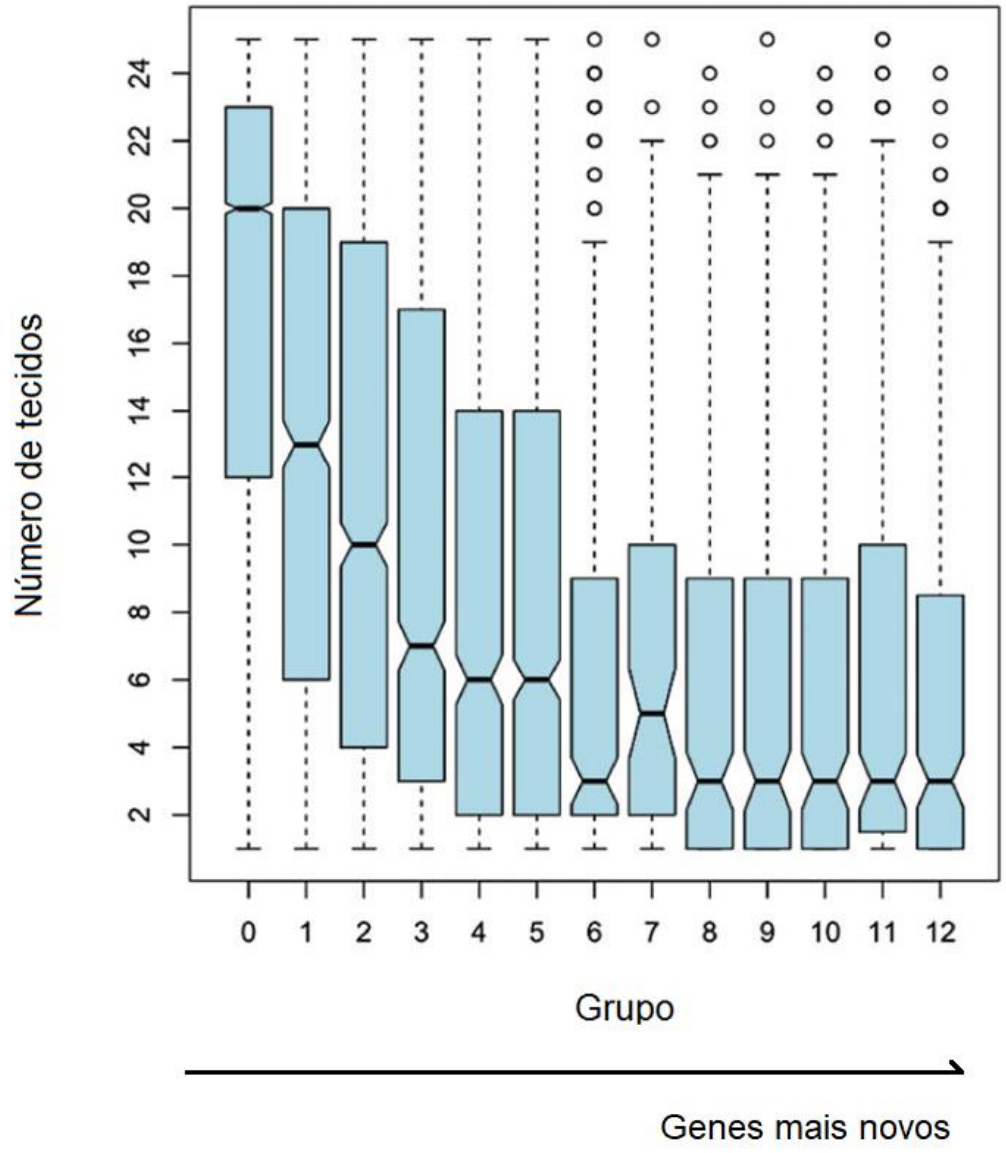


Figura 9: Perfil de expressão de genes de humanos separados por grupos de idade. Os genes estão divididos em 13 grupos, sendo genes pertencentes ao grupo 0 os genes mais antigos, e genes do grupo 12 os mais novos. É possível observar que genes mais novos são expressos em menos tecidos do que os genes mais antigos (adaptado de Zhang *et. al.*, 2010).

Existe outra característica relacionada ao perfil de expressão de genes novos que se repete de maneira praticamente independente do táxon examinado: estes genes são preferencialmente expressos na gametogênese masculina (por exemplo: Vibranovski *et. al.*, 2009; Cui *et. al.*, 2015).

Como no caso anterior, é importante lembrar que genes novos expressos em outros tecidos existem, mas são a exceção da regra mais generalizada.

O motivo pelo qual genes novos são expressos especificamente nestes tecidos ainda é alvo de debate na área, mas existem evidências que nos ajudam a compreender este padrão. A primeira informação importante para esta compreensão é o fato de que o controle da expressão nos testículos de muitos táxons é mais frouxo, permitindo a expressão generalizada de genes. Ou seja, genes novos podem se expressar nestes tecidos simplesmente por uma questão mecânica e não funcional e seriam fixados em uma população através de deriva (Soumillon *et. al.,* 2013). No entanto, este fato não é suficiente para explicar o padrão de expressão em espécies como *D. melanogaster* que possui grande tamanho efetivo populacional, diminuindo a força da deriva na população.

Com isto em mente, foi proposto um modelo que explica a expressão de genes novos em testículos através da seleção haploide, pois os genes novos possuem expressão nas fases finais da espermatogênese, na qual há expressão haploide, em diversos táxons como mamíferos, plantas e em *Drosophila*.

A expressão haploide do genoma que ocorre durante as fases finais da espermatogênese faz com que mutações recessivas benéficas sejam capazes de se manifestar. Assim, genes novos expressos durante este processo teriam chance de manifestar seus fenótipos, e serem selecionados mais rapidamente (Raices *et. al.,* 2019).

Apesar deste debate sobre os motivos pelos quais genes novos possuem este perfil de expressão e o porquê de isto ocorrer, há um consenso sobre a importância da expressão de genes novos na gametogênese masculina e como isso é uma característica peculiar deste grupo de genes (Raices *et. al.,* 2019).

1.6.Evolução de genes novos

Além do perfil de expressão, outra característica comumente observada em genes novos é a maneira a qual estes genes evoluem e se fixam em uma população. Como apresentado anteriormente, existe uma importante relação entre a expressão em testículo e as forças evolutivas mais relevantes na evolução de genes novos. No entanto, existem outros dois fatores importantes nesta discussão: a velocidade de evolução e a importância de seleção positiva para a fixação de genes novos (Kaessman, 2010).

De maneira geral, genes novos evoluem rapidamente com mudanças em sua sequência, estrutura e expressão. Este fenômeno é especialmente visível ao comparar genes novos derivados de duplicações com seus genes antigos parentais e foi observado em diversos organismos, como humanos e *Drosophila melanogaster* (Long *et. al.,* 2003).

O excesso de mudanças não sinônimas em genes novos comparados aos genes antigos ocorre por causa da seleção positiva destas mutações, que podem levar estes genes à neofuncionalização e impedindo que estes genes novos se tornem pseudogenes. Estas mudanças possibilitam, então, o surgimento de novas funções e a expressão do gene em outros tecidos (Long *et. al.,* 2003). A figura 10 traz dois exemplos de diferenças de ritmo de evolução de genes novos e antigos.

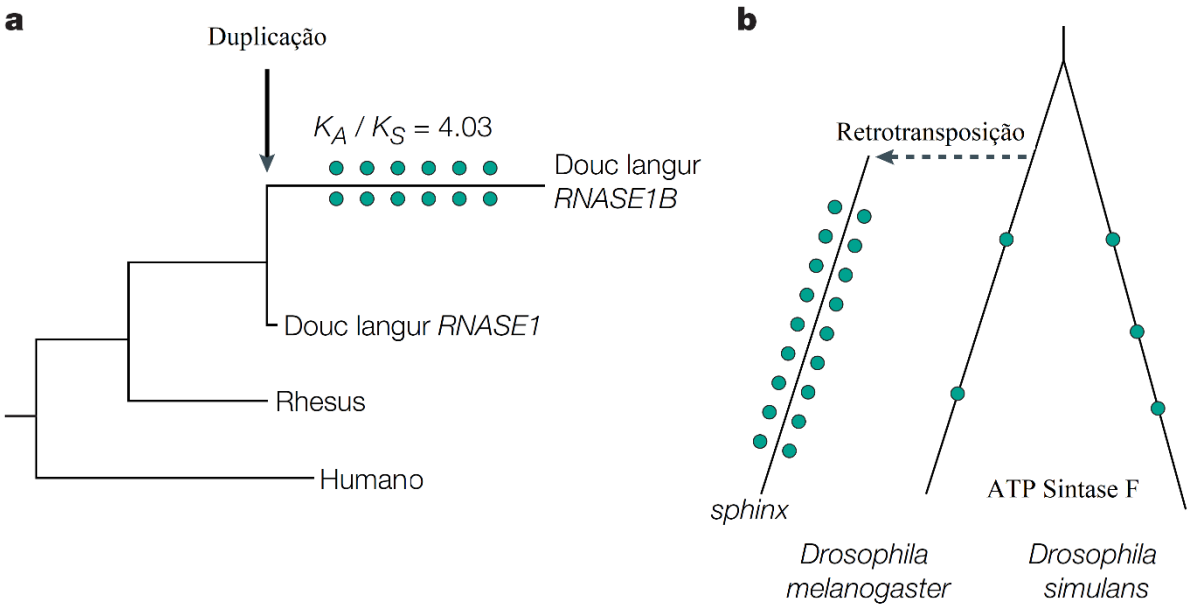


Figura 10: Exemplos de diferença de ritmo de evolução de genes novos e antigos em primatas (a) e *Drosophila melanogaster* (b). Os pontos verdes representam mudanças de nucleotídeos nas sequências dos genes. Podemos observar, em ambos os casos, que os genes novos mudam mais rapidamente do que seus genes parentais (retirado de Long *et. al.*, 2003).

Com a continuação dos estudos sobre evolução de genes novos, acumularam-se evidências da importância da seleção positiva para estes genes, desde exemplos individuais como o do gene *jingwei* em *Drosophila* (Long e Langley, 1999) a estudos de maior escala (Long *et. al.*, 2003). Foram encontrados, também, indícios de duplicações extremamente recentes em *D. melanogaster* sendo fixadas em uma população através de seleção positiva, realçando a importância deste fenômeno para genes novos (Cardoso-Moreira *et. al.*, 2016).

1.7. Machine learning

Machine learning (aprendizado de máquina) é a área que se dedica ao estudo e elaboração de algoritmos matemáticos que são capazes de mudar conforme sua experiência que surgiu a partir do campo de inteligência artificial. Estes algoritmos são capazes de lidar com grandes conjuntos de dados heterogêneos e fazer predições e classificações baseados nas informações dadas pelo usuário (Libbrecht e Noble, 2015).

Para a criação de um modelo de *machine learning*, o usuário deve fornecer ao algoritmo as características relevantes para a classificação desejada, além dos parâmetros relevantes para cada tipo de algoritmo. Podemos dividir os algoritmos existentes em três categorias abrangentes: métodos supervisionados, não supervisionados e semi-supervisionados (Libbrecht e Noble, 2015).

Os métodos supervisionados são treinados com exemplos que tem a classificação conhecida *a priori* e depois aplicados em um conjunto de dados com classificação desconhecida. Portanto, para que seja possível a aplicação destes métodos, é necessário que o usuário já possua exemplos do seu grupo de interesse classificados com outro tipo de método. Por exemplo, para criar um modelo supervisionado capaz de encontrar sítios de início de transcrição em um genoma, é necessário possuir regiões conhecidas com e sem sítios de início de transcrição previamente (Libbrecht e Noble, 2015).

Métodos não supervisionados encontram estrutura e padrões nos dados sem o uso de exemplos com classificação conhecida. Estes métodos devem ser utilizados quando não há informação prévia sobre um conjunto de dados de interesse ou para uma análise exploratória, quando não há interesse

em encaixar os dados em classes específicas e sim descobrir quais classes melhor descrevem os dados disponíveis (Libbrecht e Noble, 2015). Hoffman e colaboradores, por exemplo, apresentaram uma técnica capaz de identificar padrões na estrutura da cromatina em humanos, como sítios de início de transcrição e finais de genes, a partir de dados de sequenciamento de ChIP-seq, DNase-I-seq e FAIRE-seq (Hoffman *et. al.*, 2012).

Os métodos semi-supervisionados são um meio termo entre os métodos exibidos acima, nos quais o algoritmo recebe uma coleção de dados, com uma parcela destes dados tendo classificação conhecida (Libbrecht e Noble, 2015). Este tipo de método foi utilizado por Chen e Yang para identificar potenciais interações entre microRNAs e doenças humanas como câncer de pulmão (Chen e Yang, 2014).

Uma maneira comum de testar a eficiência de um modelo treinado de *machine learning* supervisionado é separar os dados totais em duas partes, treino e teste. Esta separação é feita aleatoriamente, com a maior parte dos dados utilizada para treino e a outra parcela, normalmente em torno de 20%, para teste (Kohavi, 1995). A parcela de treino é utilizada para efetivamente treinar o algoritmo, enquanto a parcela de teste pode ser usada para avaliar o modelo treinado em dados que ele nunca havia encontrado. Um esquema ilustrando a aplicação de um modelo supervisionado está presente na figura 11.

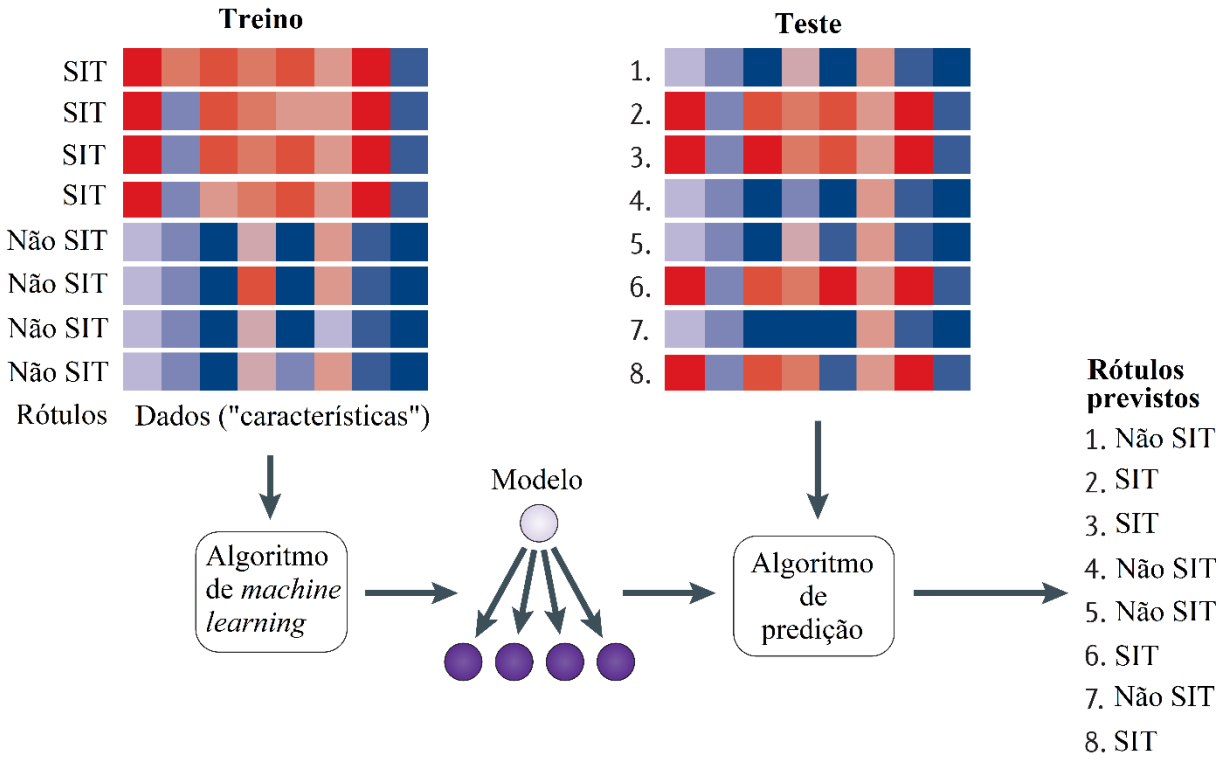


Figura 11: Esquema do funcionamento de um algoritmo de *machine learning* supervisionado para detectar sítios de início de transcrição. É possível observar a separação dos dados em teste e treino, o treinamento do algoritmo e a predição posterior, que é comparada com o que se sabe dos dados para estimar as taxas de erro do modelo. A abreviação SIT significa Sítio de Início de Transcrição. Os quadrados coloridos representam as diferentes características de cada amostra, de maneira que as cores destes quadrados representam as variações destas características nas amostras. Os círculos roxos seriam decisões tomadas pelo modelo treinado (retirado de Libbrecht e Noble, 2015).

Uma situação comumente encontrada na aplicação de *machine learning* é o desequilíbrio de classes, quando uma classe é muito mais abundante do que as outras, o que pode gerar problemas para o treinamento do modelo, já que este é inundado com exemplos da classe majoritária, dificultando seu aprendizado para separar as classes. Em uma situação com apenas duas classes, a classe majoritária possuiria muito mais membros do que a classe minoritária. Um exemplo clássico

deste problema é o exame de um conjunto de compras de cartão de crédito no qual as compras que não são fraudes são muito mais abundantes do que as compras fraudulentas (Guo *et. al.*, 2008).

Se não for possível coletar mais pontos de dados da classe minoritária, existem três maneiras mais comuns de resolver esta situação que são chamadas de *undersampling*, *oversampling* e a geração de dados sintéticos. A primeira abordagem é a amostragem aleatória de membros da classe majoritária de maneira que as duas classes fiquem com o mesmo número de membros, o segundo método é o sorteio com repetição de membros da classe minoritária até que os números se igualem e, por fim, a geração de dados sintéticos utiliza a distribuição de dados da classe minoritária para gerar membros sintéticos desta classe (Guo *et. al.*, 2008).

Cada uma destas abordagens pode gerar problemas para a análise final, e são apropriadas para diferentes tipos de situações. *Undersampling* pode fazer com que informações relevantes da classe majoritária sejam perdidas por causa da seleção aleatória; *oversampling* pode gerar ou reforçar um viés da classe minoritária ao ter membros selecionados várias vezes e a geração de dados sintéticos utiliza informações que não existem no mundo real (Guo *et. al.*, 2008). A figura 12 ilustra como funciona *undersampling* e *oversampling*.

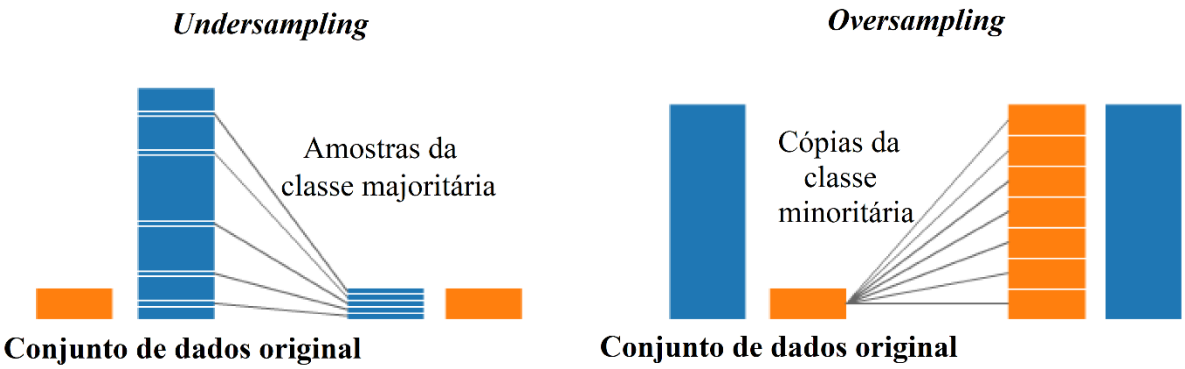


Figura 12: Estratégias de correção de desequilíbrio de classes em uma situação com duas classes. Ambas as abordagens exigem uma seleção aleatória de membros, uma para reduzir a classe majoritária e a outra para aumentar a classe minoritária (retirado de https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392).

Existem diversas maneiras de avaliar a performance de um modelo de *machine learning*, mas a maioria deles começa por examinar uma matriz de confusão. Esta é uma maneira conveniente de comparar as classificações obtidas por um modelo com os resultados esperados de acordo com conhecimento prévio. Em um algoritmo supervisionado tradicional como o exemplo apresentado acima, esta tabela é montada a partir das classificações obtidas com o conjunto de dados de teste. A tabela abaixo é uma representação de uma matriz de confusão tradicional.

	Positivos obtidos	Negativos obtidos
Positivos esperados	Positivos verdadeiros	Falsos negativos
Negativos esperados	Falsos positivos	Negativos verdadeiros

Tabela 1: Matriz de confusão utilizada para análise de resultados de *machine learning*. Com as comparações feitas na matriz é possível obter informações como a taxa de falsos positivos, que são utilizadas em análises posteriores de resultados.

A partir dos dados obtidos com a matriz de confusão, podemos calcular duas medidas importantes na avaliação de um algoritmo de *machine learning*: *precision* e *recall*. A primeira

destas medidas é calculada dividindo o número de positivos verdadeiros pela soma de positivos verdadeiros com falsos positivos (PV / (PV + FP)) e a segunda é a divisão do número de positivos verdadeiros pela soma de positivos verdadeiros e falsos negativos (PV/ (PV + FN)) (Libbrecht e Noble, 2015).

Assim, *precision* nos informa qual a proporção dos elementos selecionados pelo modelo realmente foi classificada corretamente, enquanto *recall* informa qual a proporção dos elementos que deveriam ter sido selecionados foi escolhida pelo algoritmo. A figura abaixo ilustra as medidas *precision e recall* em um esquema gráfico.

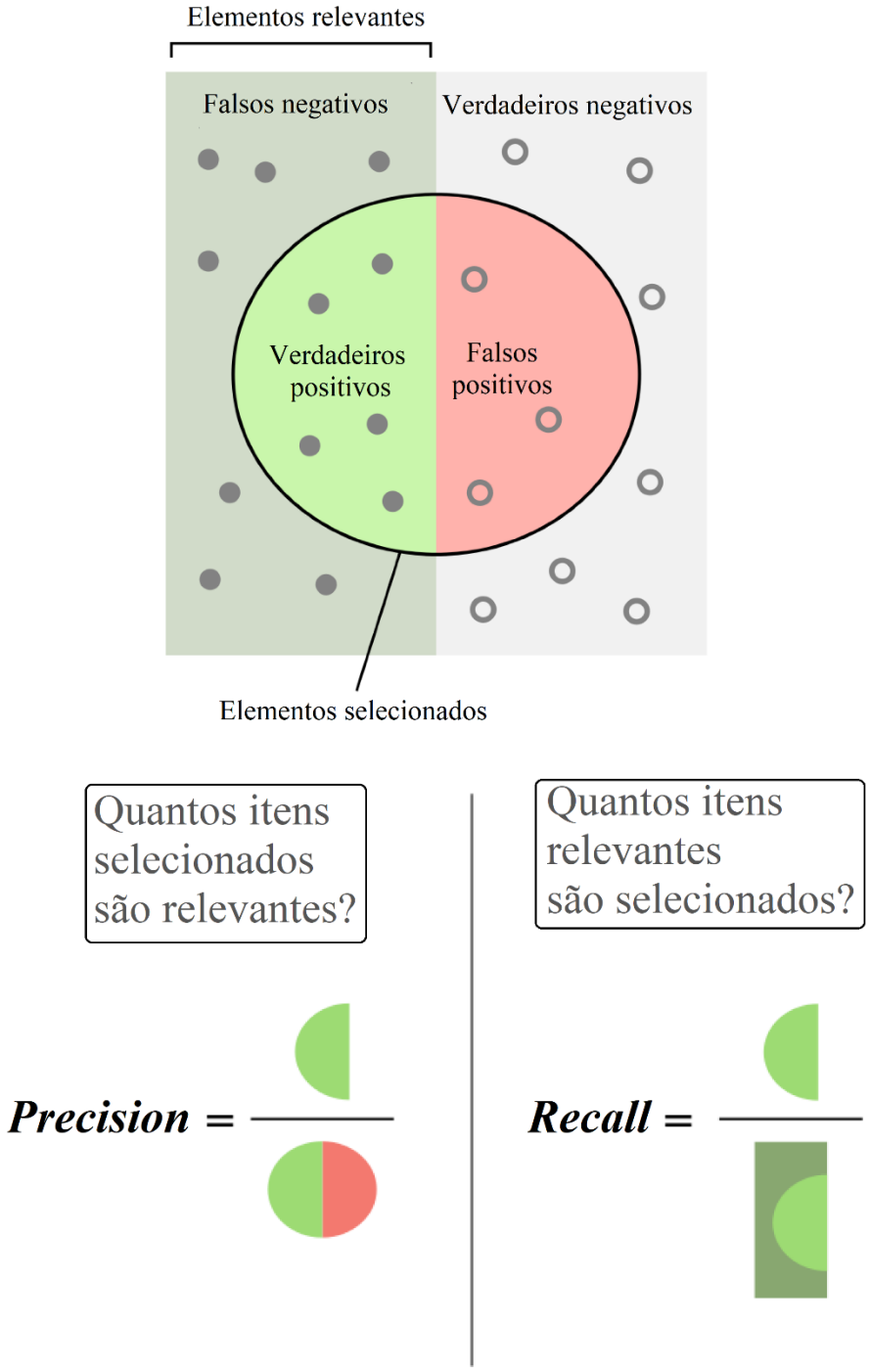


Figura 13: Representação de *precision* e *recall* de uma amostra classificada por *machine learning*. Observa-se como funciona o cálculo destas medidas em um algoritmo de classificação binária, no qual o algoritmo seleciona ou não os elementos de interesse (retirado de https://en.wikipedia.org/wiki/Precision_and_recall).

Outras medidas importantes para a avaliação de um modelo treinado de *machine learning* são curvas que mostram a relação entre medidas importantes, como uma curva comparando taxas de falsos positivos e falsos negativos, ou *precision* e *recall*. Estas curvas são geradas variando o limite de probabilidade usado pelo algoritmo para classificar um elemento e calculando as medidas durante a variação (Libbrecht e Noble, 2015). A figura 14 apresenta um exemplo de uma curva relacionando falsos positivos e falsos negativos.

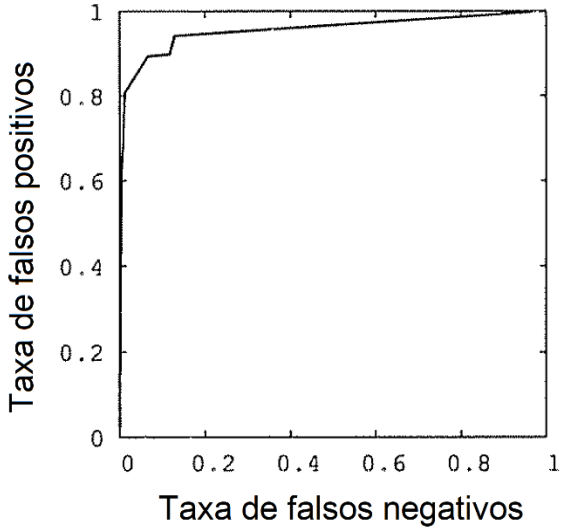


Figura 14: Exemplo de gráfico relacionando falsos positivos e falsos negativos calculados a partir da classificação de um algoritmo de *machine learning*. Esta figura ilustra o fenômeno comum em *machine learning* no qual estratégias para melhorar a taxa de falsos positivos podem piorar a taxa de falsos negativos. Para curvas como esta, quanto maior a área sob a curva melhor o modelo treinado (retirado de Davis e Goadrich, 2006).

1.8. *Random forest*

Random forest é um algoritmo de *machine learning* supervisionado proposto em 1999 por Leo Breiman que o definiu como: "uma combinação de árvores preditoras na qual cada árvore depende dos valores de um vetor aleatório sorteado independentemente e com a mesma distribuição para todas as árvores na floresta" (Breiman, 1999).

Ou seja, o algoritmo constrói um número *n* de árvores de decisão a partir de *x* variáveis escolhidas aleatoriamente de um conjunto de dados, sendo *n* e *x* escolhidos pelo usuário. Cada uma destas árvores faz uma classificação independente dos dados e é feita uma votação para determinar a classe majoritária para cada elemento do conjunto de dados no final deste processo. Desta maneira, cada elemento é classificado de acordo com a maioria das árvores classificatórias. Uma representação do funcionamento do *random forest* está presente na figura 15.

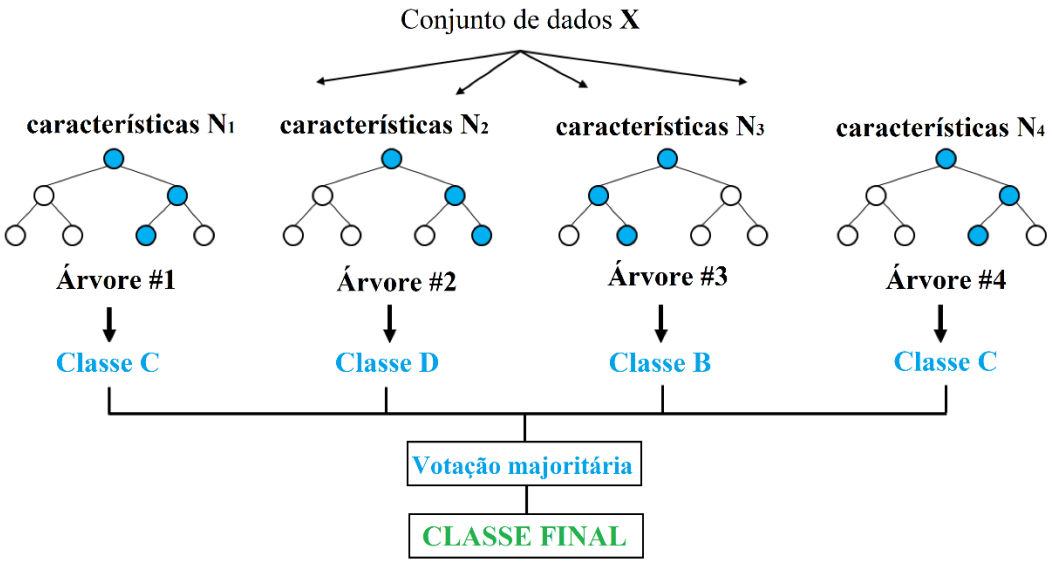


Figura 15: Funcionamento de um algoritmo de *random forest.* Cada árvore é formada por um conjunto de variáveis escolhido aleatoriamente a partir do conjunto total de variáveis e a votação gera a classificação final do modelo. As características são escolhidas aleatoriamente para a geração de cada árvore de decisão, que gera um voto classificatório. Para a decisão da classe majoritária, é feito uma votação das classificações realizadas por cada uma das árvores geradas pelo modelo. Os círculos preenchidos representam as decisões tomadas pelo algoritmo em cada uma das árvores para chegar até a classificação (retirado de http://www.globalsoftwaresupport.com/wp-content/uploads/2018/02/ggff5544hh.png).

Este algoritmo é atrativo por ser eficiente computacionalmente e capaz de examinar grandes conjuntos de dados de maneira rápida, além de produzir resultados melhores do que alguns preditores únicos já que utiliza um grande número de preditores menores, como explicado anteriormente (Baranauskas *et. al*., 2012).

Random forest é utilizado em diversas áreas do conhecimento e em genética já foi usado em diversos tipos de problemas, como prever fenótipos (Aitken *et. al.*, 2012 e Nowicki *et. al.*, 2019), classificar dados de microarranjo (Díaz-Uriarte e Alvares de Andrés, 2006) e mapear características complexas (Bureau *et. al.*, 2003).

1.9. As espécies *D. melanogaster* e *D. pseudoobscura*

O uso de espécies do gênero *Drosophila*, especialmente *D. melanogaster*, em estudos de genética existe há mais de 100 anos e foi popularizado por Thomas Hunt Morgan (1866). Morgan e outros pesquisadores da época começaram a utilizar espécies de *Drosophila* pois são facilmente mantidas em laboratório com custo baixo e possuem um curto ciclo de vida (Roberts, 2006).

Durante este período, drosofilídeos foram utilizados em trabalhos importantes de diversas áreas, como o funcionamento de mecanismos de transmissão genética (Morgan, 1917), construção do conceito de *crossing-over* (Morgan, 1911), estudos em genética de populações (Dobzhansky, 1946), a identificação dos primeiros genes *hox* (Kaufman *et. al.*, 1980), o sequenciamento de genomas (*Drosophila 12 genome consortium*, 2007) entre outros trabalhos de extrema relevância.

Por causa de todos os trabalhos realizados utilizando *Drosophila*, acumulou-se uma grande bibliografia sobre a biologia, genética e evolução de espécies do gênero, além de sequências de genomas, transcriptomas e outros recursos. Muitos destes recursos e informações estão depositados em bases de dados curadas extensivamente, como FlyBase (Larkin *et. al.*, 2021), facilitando ainda mais projetos envolvendo estas espécies, principalmente envolvendo a espécie modelo *D. melanogaster*.

O uso de espécies do gênero *Drosophila* neste trabalho se deve, primeiramente, aos mesmos fatos que levaram os primeiros pesquisadores a utilizarem este organismo: são facilmente mantidos em laboratório, cruzados e dissecados para a extração de mRNA. Em segundo lugar, a diversidade de informações disponíveis em bases de dados, como mencionado anteriormente, possibilita a exploração e investigação destes dados. Por fim, a espécie *D. melanogaster* já teve seus genes datados por Zhang e colaboradores em 2010 e teve seus genes novos identificados.

Portanto, para qualquer método proposto de identificação de genes novos, é necessário utilizar alguma espécie como *D. melanogaster* para verificar a taxa de erro deste novo método em uma espécie cujos genes novos já foram identificados. No caso desta tese, como propomos o uso de *machine learning* para encontrar estes genes, a espécie é utilizada para treinar nosso modelo, como será explicado posteriormente.

Se utilizarmos apenas *D. melanogaster* como ponto de controle do nosso método, não podemos afirmar que ele funcionaria em outras espécies ou que é um método geral para *Drosophila*, já que o modelo de *machine learning* seria treinado e testado em apenas um conjunto de dados. Com isto em mente, escolhemos a espécie *D. pseudoobscura* para servir como um segundo ponto de controle.

Escolhemos esta espécie pois já foram feitas várias rodadas de montagem e anotação do seu genoma (*Drosophila* 12 *genome consortium*, 2007), garantindo a confiabilidade dos dados disponíveis da espécie. Para que fosse possível obter este segundo ponto de controle, datamos os genes da espécie com o método publicado em 2010 por Zhang e colaboradores com o auxílio do Prof. Dr. Yong E. Zhang e membros do seu laboratório em Pequim, na China.

1.10. O contexto da área de genes novos e nossa proposta

Conforme apresentado durante a Introdução, estudos com genes novos demonstram a importância biológica destes genes, assim como seu papel no surgimento de novidades evolutivas e evolução dos genomas. No entanto, apenas um reduzido número de espécies teve seus genes datados e seus genes novos identificados.

Como o método de datação e identificação de genes novos depende de montagens e anotações de genoma de alta qualidade, sua aplicação fica limitada a espécies como *D. melanogaster*, humanos e camundongos, já que não é possível realizar laboriosas etapas de anotação manual para um grande número de espécies.

Com estes fatores em mente, esta tese propõe apresentar um novo método de identificação de genes novos, utilizando as diferenças entre características biológicas de genes novos e antigos como perfil de expressão e ortologia. Por causa da quantidade e diversidade de informações geradas, escolhemos uma abordagem utilizando o treinamento de um modelo de *machine learning*, que fará a classificação de um gene como novo ou antigo.

Para cumprir nossa proposta, dissecamos ovários e testículos de espécies de *Drosophila*, para depois sequenciarmos seu mRNA, analisamos diversas características biológicas de genes novos e antigos e treinamos um modelo de *machine learning* para identificar genes novos.

Nosso modelo foi treinado e testado em *D. melanogaster* já que esta é a única espécie do gênero cujos genes novos são conhecidos. Todavia, era necessário ter outra espécie de *Drosophila* para que pudéssemos verificar se o modelo treinado funcionaria bem nas outras espécies do gênero.

Assim, aplicamos a metodologia já existente de datação de genes em *D. pseudoobscura* e identificamos seus genes novos. Para aprender sobre a aplicação deste método e aprofundar os conhecimentos sobre o processo de datação de genes o aluno foi para Pequim (China), no laboratório do Prof. Dr. Yong E. Zhang, que é o autor do método de identificação de genes novos (Zhang et. al., 2010).

2. Objetivos

2.1. Objetivos Gerais

- Construir um modelo de *machine learning* capaz de identificar genes novos de *Drosophila* utilizando características biológicas.

2.2. Objetivos Específicos

- Observar as diferenças biológicas entre genes novos e antigos.
- Verificar o efeito variáveis relacionadas a características biológicas de genes novos de *Drosophila* em modelos de *machine learning*.
- Montar um modelo de *machine learning* para separar genes novos e antigos utilizando informações de bases de dados de *D. melanogaster*.
- Montar um modelo de *machine learning* para separar genes novos e antigos utilizando informações geradas durante a tese para *D. melanogaster*.
- Datar os genes de *D. pseudoobscura*.
- Identificar os genes novos de *D. pseudoobscura*.

3. Material e métodos

3.1.Criação de moscas

Como apresentado na Introdução e nos Objetivos, precisamos obter dados relacionados às características biológicas que são diferentes entre genes novos e antigos. Esta obtenção passa pela dissecção dos órgãos sexuais e o sequenciamento do seu mRNA que, por sua vez, dependem da manutenção das linhagens em laboratório.

Neste trabalho, foram extraídosos órgãos sexuais e o mRNA de *D. ananassae*, *D. pseudoobscura*, foram dissecadas as espécies *Scaptodrosophila lebanonensis* e *Chymomyza amoena* e foi sequenciado o mRNA de *D. erecta* e *D. virilis* foram dissecadas por Nicholas VanKuren (Universidade de Chicago, EUA) e Mariana Teixeira Kanbe (USP), respectivamente. As outras espécies foram sequenciadas por grupos parceiros para seus próprios trabalhos, e os dados nos foram cedidos para realização deste projeto.

As linhagens das espécies do gênero *Drosophila* utilizadas foram aquelas usadas no projeto de 12 genomas (Clark *et. al.,* 2007), enquanto as espécies de *Scaptodrosophila* e *Chymomyza* foram coletadas por membros do laboratório do Prof. Dr. Antonio Bernardo de Carvalho (UFRJ) e identificadas pelo Prof. Dr. Carlos Ribeiro Vilela (USP) e a Dra. Suzana Casaccia Vaz.

As moscas utilizadas nos experimentos foram criadas em laboratório para que fosse possível a separação de machos e fêmeas virgens. Para cada espécie, mantivemos pelo menos 8 réplicas separadas em garrafas diferentes, de maneira a obter uma grande quantidade de indivíduos, o que é necessário para que tenhamos muitos órgãos extraídos e uma boa quantidade de mRNA no final da extração.

As linhagens são mantidas em garrafas com meio de cultura de fubá, na temperatura de 22 graus e com fotoperíodo de 12 horas, com os indivíduos adultos sendo trocados de garrafa a cada semana. Quando é necessário aumentar a quantidade de indivíduos em uma linhagem, essa troca (chamada de repique) é feita com maior frequência (Ashburner, 1989).

A coleta de indivíduos virgens é importante para garantir que os machos ainda possuam todas as fases da espermatogênese presentes nos testículos já que o macho perde espermatozoides maduros durante a cópula, e poderíamos acabar dissecando testículos sem estes espermatozoides. As fêmeas também devem ser virgens pois são capazes de armazenar espermatozoides de suas cópulas, ou seja, ao dissecar fêmeas não virgens poderíamos acabar contaminando as amostras com espermatozoides (Ashburner, 1989).

Para a coleta de virgens, retiramos os adultos de cada garrafa onde as moscas são criadas, de maneira que todos os indivíduos que emergirem nas próximas horas serão virgens até seu tempo de amadurecimento sexual. Este período de tempo varia de acordo com o ciclo de vida de cada espécie, espécies com o ciclo menor possuem tempo menor, e espécies com ciclo maior tem tempo de desenvolvimento mais lento (Ashburner, 1989).

Após a coleta, os indivíduos adultos são separados em tubos sem meio, e dormem por causa da baixa temperatura ao gelar estes tubos em isopores com gelo. Com o adormecimento dos indivíduos, eles são separados de acordo com o sexo em lupas, de acordo com as características sexuais de drosófilas. No gênero *Drosophila* e em gêneros relacionados, os machos e fêmeas podem ser identificados pela porção terminal do abdômen, onde ficam os órgãos reprodutores, que apresentam claras diferenças entre os sexos (figura 16).

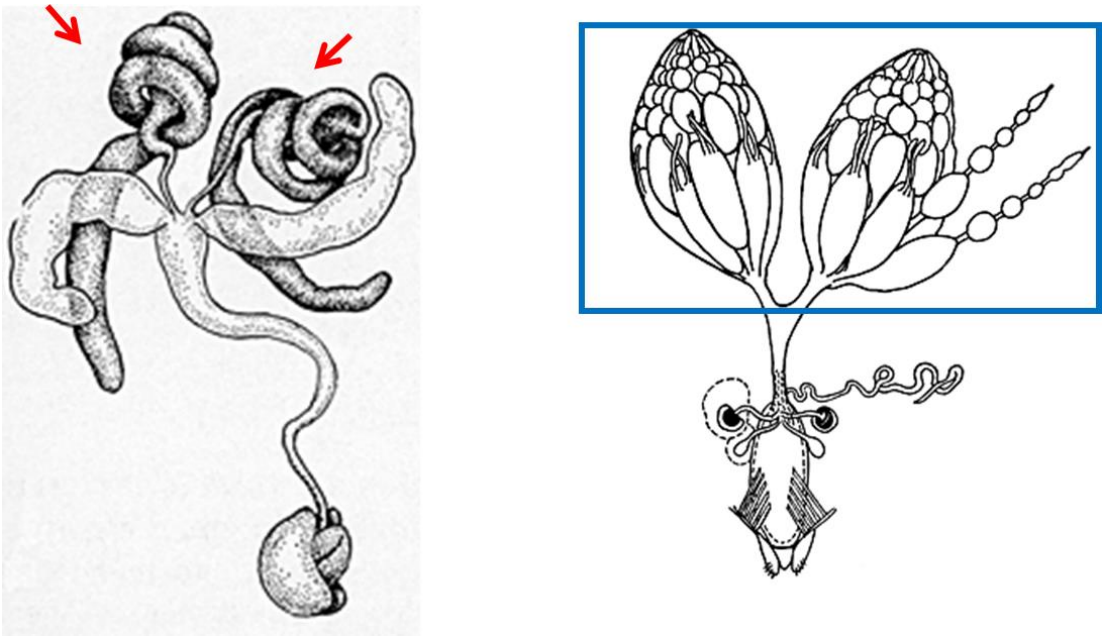


Figura 16: Desenhos representando a terminália do macho (esquerda) e fêmea (direita) de *D. melanogaster*. As setas em vermelho apontam para os testículos, enquanto o retângulo azul marca os ovários. Durante a extração, separamos estes órgãos de interesse dos outros após sua retirada do abdômen (Patterson, 1943, Miller, 1950).

Os indivíduos, agora separados por sexo em tubos diferentes, são envelhecidos de 6 a 10 dias até o dia da dissecção, de maneira a garantir que todas as fases da gametogênese estão presentes nos órgãos sexuais. Este tempo é necessário para a presença de todas as fases da espermatogênese pois este processo começa na fase larval e demora aproximadamente 300 horas (Ashburner, 1989).

Estes tubos são mantidos durante um tempo para garantir que nenhuma pupa ou larva nasçam, o que significaria que os indivíduos não eram virgens ou algum indivíduo foi identificado incorretamente. Se isto ocorrer, todos os órgãos extraídos de indivíduos pertencentes àquele tubo são descartados.

3.2. Dissecção de órgãos sexuais

Nas dissecções, os órgãos reprodutivos são retirados e armazenados em RNA Later (Qiagen) a -20 graus Celsius, o que retarda a degradação de mRNA para que possamos extraí-lo posteriormente. Como o mRNA é instável, os experimentos duram no máximo duas horas e trinta minutos, para não deixar os órgãos extraídos fora do congelador por muito tempo. Em cada experimento, foram extraídos em média de cem a cento e vinte órgãos, para garantir que haverá mRNA suficiente para o sequenciamento.

Para cada disseção, pegamos os indivíduos já separados por sexo e envelhecidos e colocamos as moscas em pequenos tubos de plástico, que são enterrados no gelo para que as moscas entrem em coma por causa do frio. As moscas são colocadas uma a uma em gotas de PBS em lâminas, para que sejam observadas em lupas. O procedimento da dissecção é um pouco diferente para machos e fêmeas, mas se baseiam no fato de que os órgãos sexuais são facilmente reconhecidos, possibilitando sua retirada com pouca chance de erro.

Mais especificamente, a dissecção em é feita com um pequeno corte no abdômen, que é aumentado cuidadosamente com uma pinça de precisão. Em machos, puxamos a porção terminal do abdômen com uma pinça de precisão, fazendo com que os testículos, glândulas acessórias e vesículas seminais saiam da cavidade abdominal. Como o testículo tem morfologia bastante distinta,

ele é facilmente reconhecido mesmo junto dos outros órgãos. Os testículos são então separados do resto do material cuidadosamente separando-os dos outros órgãos nas regiões de junção entre os órgãos, de maneira que os testículos não se rompam, já que isso levaria a perda de material. Os órgãos extraídos são armazenados em RNA later e colocados a -20 graus até o momento da extração do mRNA (figura 17) (https://mnlab.uchicago.edu/sppress/index.php?methods=1).

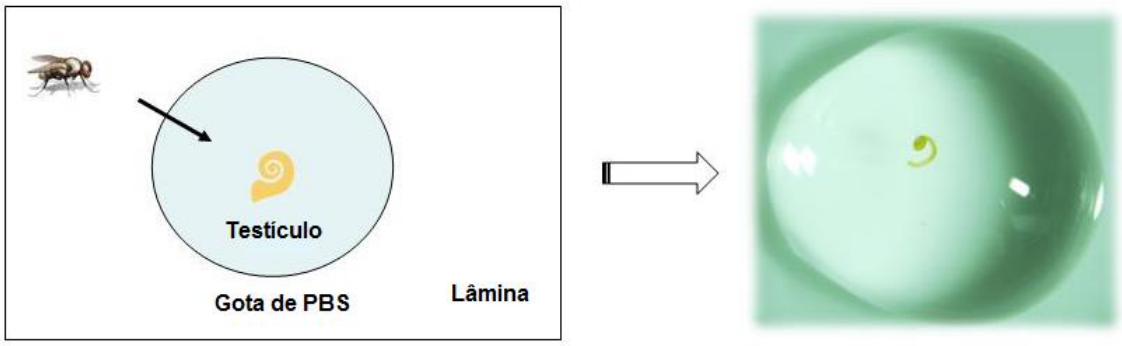


Figura 17: Esquema de uma dissecção para retirada de testículo de *D. melanogaster* (esquerda) e foto de um testículo da espécie. As extrações dos órgãos sexuais são feitas por indivíduo, de maneira que rapidamente sejam retirados estes órgãos de cada mosca em coma (https://mnlab.uchicago.edu/sppress/index.php?methods=1).

A dissecção em fêmeas é feita de maneira semelhante, em uma adaptação do método descrito acima. Ao invés de puxar a região terminal do abdômen, como é feito em machos, quando a abertura feita com auxílio da pinça é grande o suficiente para a saída dos ovários, é feita uma leve pressão no abdômen em direção ao corte, para que os ovários saiam inteiros e nenhum material seja perdido. Os ovários são separados dos outros órgãos que possam sair junto, e armazenados em RNA Later, até o momento da extração do mRNA.

3.3. Análises estatísticas e computacionais

As análises computacionais foram realizadas em um de dois servidores Unix locais, sendo um localizado em São Paulo, no laboratório da Profa. Dra. Maria Dulcetti Vibranovski, com 314Gb de memória RAM, 40 CPUs e 17Tb de espaço de HD e o outro localizado em Pequim, China, no laboratório do Prof. Dr. Zhang E. Yong, com 125G de RAM, 48 CPUs e 55Tb de espaço de memória.

Estes servidores foram acessados remotamente através do programa Putty (www.putty.org) com linhas de comando, que também são utilizadas para escrever os *scripts* menores utilizados para manipular nossos dados. Estes pequenos programas são escritos utilizando linguagens de programação como awk (Aho *et. al.*, 1979), sed (Dougherty e Robbins, 1997), bash (Ramey, 1994), perl (Wall, 1994) e python (Lutz, 2001).

As análises estatísticas e gráficos são feitos no programa R (R Core Team, 2020) e Rstudio (Rstudio Team, 2020), utilizando as bibliotecas apropriadas para cada etapa, como lattice (Sarkar, 2008), psych (Revelle, 2020), ROSE (Lunardon *et. al.*, 2014) e randomForest (Liaw e Wiener, 2002). O programa R foi utilizado tanto nos servidores, quanto em um laptop HP com 16Gb de RAM, 5 CPUs e 1TB de espaço em memória.

Quando publicarmos os artigos relacionados a este trabalho, iremos disponibilizar as sequências, as tabelas formadas durante a execução da tese e os comandos necessários para a reprodução das

análises. Os dados de sequenciamento serão disponibilizados na base de dados SRA (https://www.ncbi.nlm.nih.gov/sra), como é exigido pela maioria das revistas científicas da área.

As outras informações, como *scripts* que escrevemos, tabelas com quantidades extensivas de informação e linhas de comando com os parâmetros dos programas utilizados serão disponibilizadas em um repositório como GitHub (https://github.com/), de maneira a serem acessíveis a todos os interessados.

3.4.Extração de mRNA e sequenciamento

Os órgãos retirados tiveram seu mRNA extraído utilizando o kit Arcturus Pico Pure, sendo utilizados aproximadamente 100 órgãos em cada experimento, para garantir o mínimo de 1 µg de RNA total necessário para o sequenciamento. Após a extração, 1µl do material extraído é usado para a quantificação e controle de qualidade pelo aparelho Nanodrop Lite, para garantir que o material está próprio para sequenciamento.

Os órgãos de quatro espécies, *D. ananassae, D. pseudoobscura, D. virilis* e *D. erecta* foram transportados em RNA Later e tiveram seu mRNA extraído no laboratório do Dr. Manyuan Long, na universidade de Chicago (Estados Unidos). Os transcriptomas destas espécies foram sequenciados na Chicago Functional Genomics Facility, utilizando a tecnologia Illumina com *reads paired-end* de 100pb, por isso foram dissecadas no Brasil, e tiveram seu mRNA extraído nos Estados Unidos.

Após a extração, o material passa por um controle de qualidade com o método Nanodrop, de maneira a garantir que há mRNA em quantidade suficiente para o sequenciamento, e por outro controle de qualidade na instituição que faz o sequenciamento com o BioAnalizer. Nós recebemos os resultados da instituição e os utilizamos para verificar que todas as nossas amostras possuem o mesmo padrão de qualidade e que podem ser sequenciadas. Além da quantificação do mRNA, recebemos da *CFGF* os dados de "RNA *integrity number*" (RIN), e gráficos de florescência, todos usados para o controle de qualidade das amostras.

Após o sequenciamento, as sequências são disponibilizadas para *download*, e são colocadas no servidor local do laboratório, no qual serão feitas as análises subsequentes.

3.5.Bases de dados de *D. melanogaster*

A espécie *D. melanogaster* é um dos principais modelos para insetos, e o modelo para o gênero *Drosophila*. Portanto, não havia necessidade de sequenciar os transcriptomas desta espécie, já que trabalhos anteriores já tinham gerado os dados que desejamos utilizar no projeto. No entanto, foi necessário examinar alguns experimentos já depositados em bases de dados para escolher aqueles que cumpram nossas exigências.

Para isto, pesquisamos no *Sequence Read Archive* (Leinonen *et. al.*, 2010) os experimentos de sequenciamento de transcriptomas de testículo e ovário com tecnologia Illumina e tamanho mínimo de 75 pares de base. Além destas exigências, necessitamos de transcriptomas que tenham sido gerados como parte de um mesmo experimento, e que sejam da linhagem da espécie que foi utilizada para o genoma de referência.

Nós exigimos estas características para garantir que os dados sejam parecidos com aqueles produzidos no projeto, que haja pouca variação entre os experimentos realizados para machos e fêmeas e que possamos utilizar o genoma de referência da espécie durante o projeto. Com estas exigências, chegamos a dois experimentos diferentes que poderiam ser utilizados. Para decidir qual

seria o melhor conjunto de dados, fizemos a montagem de transcriptomas e controle de qualidade como explicado no próximo tópico.

Além das *reads* de *D. melanogaster*, utilizamos diversos dados já disponíveis em bases de dados, como localização genômica, expressão em diversos tecidos, ortologia, *dn/ds*, entre outros. Estes dados foram retirados de *ensembl* (Howe *et. al.*, 2021), FlyBase (Larkin *et. al.*, 2021) e FlyDivas (Stanley e Kulathinal, 2016), que são repositórios curados com uma enorme quantidade de informação genética e evolutiva. FlyDivas contém informação de evolução para os genes de *D. melanogaster*, enquanto o FlyBase inclui uma grande quantidade de dados para diversos dípteros e *ensembl* é uma base de dados generalista com uma grande diversidade de informações.

Para este projeto é essencial o conhecimento dos genes novos de *D. melanogaster*, já que esta é a única espécie do gênero *Drosophila* com seus genes datados. A importância destas informações já foi citada na Introdução, mas será expandida nas seções de Resultados e Discussão. Estas informações de datação estão localizadas na base de dados específica GenTree, que disponibiliza uma tabela com a datação de todos os genes de *D. melanogaster* além de outras informações (Zhang *et. al.*, 2010; Shao *et. al.*, 2019).

3.6.Montagem de transcriptomas

Para que seja possível realizar as análises que o projeto necessita, é necessária a montagem dos transcriptomas *de novo*, ou seja, sem a utilização de um genoma de referência, já que as montagens do genoma de cada espécie são extremamente diferentes. Esta montagem foi feita com o programa Trinity (Grabherr *et. al.*, 2013) e seus parâmetros padrões (tamanho de *k-mer* = 25; número mínimo de vezes que um *k-mer* deve aparecer para ser utilizado = 1) e o programa Trimmomatic (Bolger *et. al.*, 2014), que já vem com o pacote, para lidar com *reads* de baixa qualidade.

Como já conhecemos a lista de genes novos de *D. melanogaster*, utilizamos a lista disponível no GenTree para realizar alguns testes e verificar se era possível recuperar um número maior de genes novos na montagem mudando os parâmetros do montador.

Como muitos montadores de *reads* de NGS, o primeiro passo do Trinity é fragmentar os *reads* em sequências menores de tamanho determinado, chamados de *k-mer*. Estes *k-mers* serão utilizados durante o processo de montagem, ao invés da sequência inteira dos *reads* como era feito em montadores de tecnologias anteriores. Portanto, manipular parâmetros relacionados aos *k-mers* podem afetar a montagem final.

Com estes fatores em mente, tentamos aumentar o tamanho do *k-mer* e o mínimo de vezes que um *k-mer* deve aparecer para ser considerado. Ambos parâmetros aumentam a restrição do montador conforme aumentam, até certo ponto, conforme instruído pelos autores do programa. Da mesma maneira, o montador é menos restritivo ao diminuir-se estes parâmetros (Haas *et. al.*, 2013). No entanto, não foram encontradas diferenças significativas entre as montagens em relação ao número de genes novos encontrados, logo, utilizamos os parâmetros padrões do programa para todas as espécies.

Após as montagens, fizemos alguns controles de qualidade nos transcriptomas, de maneira a garantir que as análises feitas posteriormente sejam confiáveis. O próprio montador Trinity nos dá alguns parâmetros de qualidade após a montagem como N50, que é o tamanho da menor sequência montada que somada as sequências maiores tem 50% do tamanho do transcriptoma, mas estes parâmetros não nos dão uma figura completa da qualidade do transcriptoma.

O primeiro controle de qualidade é verificar qual a porcentagem dos *reads* iniciais que se encontram na montagem final, alinhando o transcriptoma com os *reads* iniciais com alinhadores como o *bowtie2* (Langmed e Salzberg, 2012). Para que seja considerada uma boa montagem, é recomendado que pelo menos 80% das *reads* alinhem na montagem final, de maneira que os *reads*

que não alinham representam transcritos com baixa expressão, *reads* de baixa qualidade ou oriundos de erros de sequenciamento.

Outro método utilizado de controle de qualidade é verificar a reconstrução de genes codificadores de proteína, através da comparação com uma base de dados curada. Esta comparação foi feita através do programa BUSCO (Manni *et. al.*, 2021), que usa sua base de dados de genes curados para verificar quantos deles estão presentes em uma montagem. No caso desta tese, utilizamos a base de dados de genes de *Diptera* disponibilizada pelo BUSCO para esta análise.

Nós utilizamos estes métodos para comparar as montagens de transcriptoma geradas a partir dos diversos *reads* disponíveis para *D. melanogaster* e as montagens das outras espécies. No primeiro caso, estas comparações nos ajudaram a escolher os melhores *reads* possíveis, enquanto no segundo caso nos garantiu que todas as montagens para todas as espécies tinham qualidades semelhantes, já que será necessário compará-las durante este trabalho.

Estes controles de qualidade e as exigências explicadas no tópico anterior nos levaram a escolher os *reads* com o código de acesso SRR384928 e SRR384962 para ovário de *D. melanogaster* e SRR384929, SRR384930 para testículo.

3.7. Quantificação e expressão diferencial

Após a montagem dos transcriptomas, fizemos a quantificação da expressão dos transcritos, e o cálculo da expressão diferencial de cada gene entre testículo e ovário. Estas informações serão utilizadas posteriormente no modelo de *machine learning*, já que são diferenças importantes entre genes novos e antigos. Para a quantificação da expressão, foi utilizado o programa RSEM (Li e Dewey, 2011), enquanto o programa DEseq2 (Anders *et. al.*, 2014) é utilizado para o cálculo da expressão diferencial em conjunto dos alinhadores STAR e kallisto (Dobin *et. al.*, 2013; Bray *et. al.*, 2016). O STAR necessita de um genoma de referência para funcionar e, portanto, foi utilizado apenas em *D. melanogaster*.

Para verificar a expressão dos genes de *D. melanogaster* nas outras espécies, utilizamos os pares de ortólogos identificados por *reciprocal best hits*. A maneira com a qual fizemos a identificação destes ortólogos e como implementamos os programas de encontro de ortólogos será explicada na próxima seção. Com esta lista de ortólogos, montamos uma tabela que contém a informação de viés de expressão de cada gene de *D. melanogaster* para cada uma das outras espécies.

Para montar essa tabela, que contém as informações utilizadas no *machine learning*, examinamos a expressão do gene ortólogo de *D. melanogaster* em cada espécie com o programa DEseq2 que calcula a expressão diferencial entre tecidos. Se um gene de *D. melanogaster* só tem ortólogo identificado no transcriptoma de testículo, ele é considerado como enviesado para testículo e o mesmo é feito para ovário.

Para facilitar nossa análise, ao final do exame de expressão de cada espécie, o gene ortólogo pode ser considerado enviesado para testículo (3), enviesado para ovário (2), sem viés para nenhum tecido (1) e ausente (0), quando não há ortólogo identificado. A tabela 2 é um pequeno exemplo da tabela final gerada dessa maneira.

Gene	Viés em *D. erecta*	Viés em *D. ananassae*	...	Viés em *Scaptodrosophila*
FBgn0000008	3	1	...	2
FBgn0000278	0	2	...	0
FBgn0011260	3	3	...	3

Tabela 2: Exemplo de tabela gerada contendo a informação de viés de expressão dos genes de *D. melanogaster* em outras espécies. Esta tabela é utilizada na etapa de *machine learning*, junto com as outras informações biológicas.

3.8.Busca de ortólogos

Uma diferença importante entre genes novos e antigos é a quantidade de ortólogos, já que por definição espera-se que genes antigos tenham mais ortólogos em um número maior de espécies. Ou seja, ao fazer uma busca de cada gene da espécie focal nas outras espécies, genes antigos têm uma probabilidade maior de ter ortólogos encontrados em múltiplas espécies. Assim, esperamos que esta abordagem nos auxilie a separar genes novos de antigos, melhorando nosso modelo.

Para isso, testamos três métodos de busca de ortólogos em *D. melanogaster*, *Reciprocal Best Hits* (*rbh*), *Reciprocal Smallest Distance* (*rsd*) e *Orthofinder* (Koonin *et. al.,* 1997; Wall *et. al.,* 2003; Emms e Kelly, 2019). Cada um destes métodos encontra ortólogos de uma maneira diferente, portanto testamos os três para verificar qual seria capaz de identificar mais ortólogos corretamente nos nossos dados.

O primeiro método utiliza duas buscas recíprocas de BLAST (Altschul *et. al.,* 1997) utilizando duas bases de dados de sequências fornecidas pelo usuário e separa as sequências que tiveram os melhores *hits* recíprocos. Ou seja, supondo que duas bases de dados, A e B, sejam fornecidas, o programa faz alinhamento por BLAST de A vs B e de B vs A, teremos uma relação de ortologia caso os melhores pares de alinhamentos contenham as mesmas sequências de A e B nas duas etapas de busca. O segundo método tem um procedimento parecido, mas ao invés de utilizar as melhores correspondências recíprocas de alinhamentos de sequências de nucleotídeos, ele procura os pares de sequências com a menor distância entre sequências de proteína.

O terceiro método utiliza um algoritmo próprio e mais complexo, construindo e utilizando árvores filogenéticas para encontrar grupos de genes relacionados que os autores chamam de ortogrupos. Com estes ortogrupos, o programa identifica os ortólogos da espécie focal, além de genes relacionados.

Como estamos comparando o genoma de *D. melanogaster* com transcriptomas, temos que tomar cuidado ao exigir alta cobertura nos parâmetros dos programas, já que muitas vezes os transcritos de um gene são apenas parcialmente montados. Da mesma maneira, como estamos comparando espécies com grande distância filogenética, exigir alta semelhança entre sequências pode fazer com que poucos ortólogos sejam encontrados.

Com estes fatos em mente, fizemos alguns testes com os parâmetros de cada programa de maneira a estabelecer como conseguiríamos recuperar o maior número possível de ortólogos nas espécies mais distantes. Após estes testes, vimos que não podemos exigir alta cobertura nem similaridade para recuperar ortólogos de espécies distantes. Ao obter estes resultados, comparamos os melhores resultados de cada um dos programas para escolher aquele que seria melhor para nossos

objetivos, que é eliminar o maior número possível de genes antigos sem perder muitos genes novos, através deste filtro. Este processo será explicado mais detalhadamente na próxima seção.

No final destas comparações, escolhemos o programa *reciprocal best hits*, com os melhores parâmetros de recuperação de ortólogos em espécies distantes. A primeira destas escolhas é utilizar o *tblastx* (Altschul *et. al.,* 1997) como programa de busca, que usa nucleotídeos traduzidos para sua busca. Como utilizamos espécies filogeneticamente distantes, o uso de sequências proteicas auxilia no encontro de ortólogos já que mudam mais lentamente. Os outros parâmetros foram 30% de identidade mínima, o que diminui o número de *hits* errados sem limitar muito o programa e nenhuma exigência de cobertura, já que a montagem de transcriptomas pode estar muito fragmentada.

3.9. Filtro de ortologia

Durante nossos testes com modelos de *machine learning*, pensamos que o grande número de genes antigos presentes no conjunto de dados poderia atrapalhar o desempenho do algoritmo. Para descobrir se isto estava correto, era necessário encontrar uma maneira de eliminar genes antigos antes que as informações fossem dadas ao algoritmo de *machine learning*.

Com este objetivo em mente, buscamos uma diferença fundamental entre genes novos e antigos: genes antigos possuem genes ortólogos em mais espécies. Assim, se fizéssemos uma busca de ortólogos dos genes de *D. melanogaster* nas outras espécies esperávamos que genes antigos teriam mais ortólogos e poderiam ser removidos desta maneira.

Para isto, utilizamos o método de *reciprocal best hits*, detalhado na seção acima para buscar estes ortólogos de *D. melanogaster* no testículo e ovário das outras espécies incluídas neste projeto. Nosso primeiro passo tomado com estas informações foi montar uma tabela contendo todos os genes de *D. melanogaster* e a informação sobre a presença ou ausência de ortólogos em todas as outras espécies.

Com esta informação em mãos, podemos investigar qual é a melhor maneira de eliminar um grande número de genes antigos de acordo com a presença ou ausência de ortólogos nas espécies do incluídas no projeto. Na seção de Resultados mostramos como se comportam os genes de *D. melanogaster* e como aplicamos este filtro, mas é importante entender que genes novos têm menos ortólogos, especialmente em espécies mais filogeneticamente distantes.

Apesar da utilidade deste filtro, e de seu efeito nos nossos resultados (exibido na seção de Resultados), ainda restava uma quantidade considerável de genes antigos na nossa análise. Com isto em mente, pensamos em aplicar uma segunda etapa de filtragem, utilizando uma informação diferente gerada na mesma análise de ortologia: o contexto filogenético.

O primeiro filtro utiliza pouca informação filogenética, já que considerava apenas presença ou ausência. Ou seja, não conseguimos identificar se um gene tinha um ortólogo em duas espécies próximas de *D. melanogaster* ou em duas espécies distantes. Isto é importante pois genes antigos têm maior probabilidade de possuir ortólogos em espécies distantes do que genes novos.

Por estes motivos, construímos uma nova tabela, que agora associa pontos diferentes a cada presença de ortólogos dependendo da distância filogenética de *D. melanogaster*. Além disso, juntamos em alguns grupos espécies com distâncias semelhantes a *D. melanogaster*, de maneira que a presença de ortólogo em qualquer uma destas é considerada apenas uma vez, dando a pontuação referente ao grupo.

Com estas modificações, geramos este segundo filtro, com o qual esperávamos eliminar genes antigos que sobreviveram à primeira etapa de filtragem, mas possuem ortólogos em espécies mais

distantes de *D. melanogaster*. A formação dos grupos em seu contexto filogenético está demonstrada na figura abaixo.

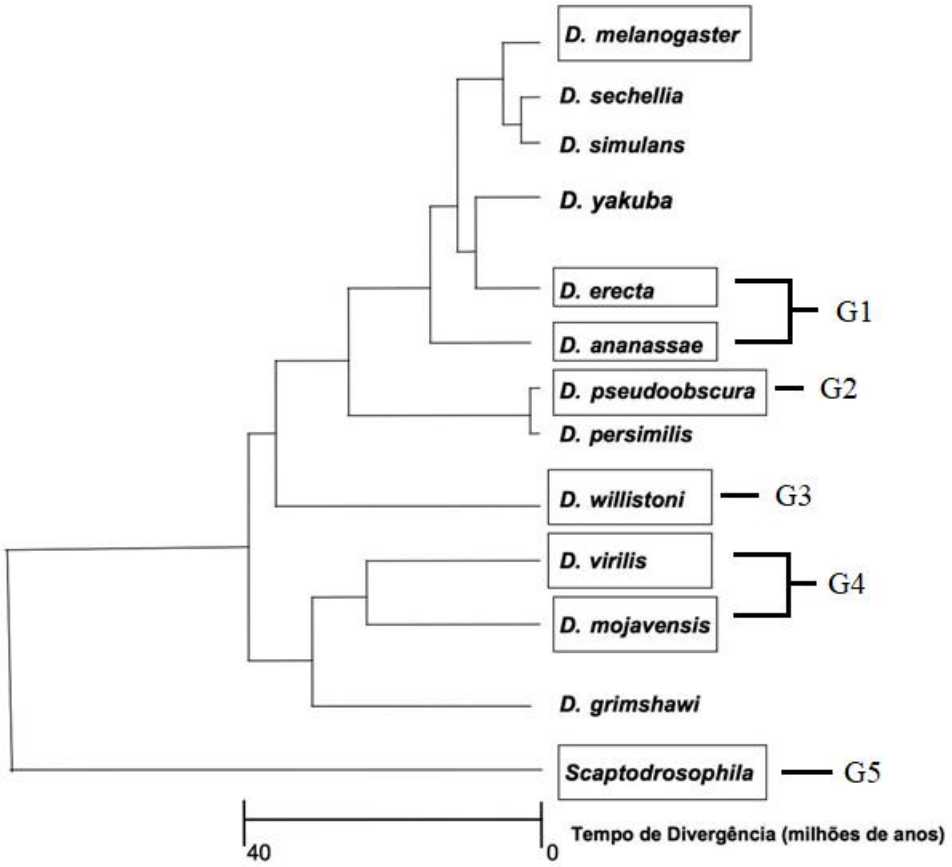


Figura 18: Filogenia simplificada do gênero *Drosophila* e a divisão dos grupos filogenéticos utilizados no segundo filtro. Com esta divisão em grupos de um a cinco (G1 a G5), a presença de um ortólogo em *D. erecta* ou *D. ananassae* (G1) vale menos do que a presença em *D. virilis* ou *D. mojavensis* (G4).

Nesta segunda etapa de filtragem, a presença de um ortólogo no grupo um (G1) vale um ponto; no G2 vale dois pontos; G3 quatro pontos; G4 oito pontos e G5 dezesseis pontos. Assim, a somatória destes valores nos informa não só quais genes possuem mais ortólogos, mas também em quais grupos, já que só há uma maneira de atingir cada valor do somatório. Por exemplo, só há uma maneira de atingir 10 pontos, com ortólogos no grupo 2 e no grupo 4.

Como será visto na seção de resultados, realizamos uma série de investigações para determinar a melhor maneira de utilizar estas duas etapas de filtragem para remover o maior número possível de genes antigos sem perder muitos genes novos, e melhorar os resultados finais obtidos com *machine learning.*

3.10. Cálculo de *dn/ds*

Como genes novos possuem maiores sinais de seleção positiva (Kaessman, 2010), os dados de *dn/ds* podem ajudar a separar estes genes dos genes antigos, e esta variável pode contribuir para melhorar nosso modelo. Com este intuito, utilizamos a base de dados curada FlyDivas (Stanley e Kulathinal, 2016), que tem a informação de *dn/ds* em *D. melanogaster*. Nós utilizamos os dados gerados a partir da comparação dos genes de *D. melanogaster* com as outras espécies do subgrupo *melanogaster* na etapa de *machine learning*.

Após verificar que a adição desta informação melhora o modelo, fomos fazer o cálculo *de novo* destas informações, isto é, com os nossos próprios dados. Esta base de dados fornece somente os dados de seleção positiva e de taxas de substituição não sinônima e sinônima focados em *D. melanogaster*. Como pretendemos aplicar este método em outras espécies do gênero *Drosophila*, precisamos que as variáveis sejam geradas da mesma maneira em todas as espécies.

Para isto, seguimos o que é descrito no artigo do FlyDivas que possui breves instruções de como estes dados foram gerados. De maneira sucinta, utilizamos os ortólogos identificados por *reciprocal best hits*, como descrito anteriormente, alinhamos as sequências utilizando MUSCLE (Edgar, 2004), transformamos o alinhamento para DNA com correção de códon com *translatorX* (Abascal *et. al.*, 2010) e utilizamos o programa PAML (Yang, 1997) para o cálculo das taxas.

Utilizamos os ortólogos de *D. melanogaster* em *D. erecta* para esta análise, já que vimos que se adicionássemos outra espécie e usássemos apenas os genes que tivessem ortólogos nas duas, ficaríamos com um número muito reduzido de genes. Estas sequências são traduzidas para aminoácidos com o programa ORFfinder (Rombel *et. al.,* 2002), com a maior ORF selecionada para cada sequência. O alinhamento é feito com os parâmetros padrão do MUSCLE, e estes alinhamentos são transformados de volta em DNA com o programa *translatorX*. Finalmente, estes alinhamentos vão para o PAML, na função *codeml* com o modelo par a par, que fará o cálculo efetivo do *dn/ds*.

Estes processos devem que ser feitos gene a gene, ou seja, seria praticamente inviável sem algum nível de automatização. Com este intuito, foram escritos alguns *loops* na linguagem de programação *shell* que realizam o mesmo comando para cada um dos genes listados, assim, rodamos apenas um *script* para cada etapa, gerando os resultados em massa. No final, é gerada uma tabela de resultados para cada gene, que são processadas e resumidas com auxílio de um *script* em *python* e unidas em uma grande tabela (tabela 4).

Gene	*dn*	*ds*	*dn/ds*	p
FBgn0000015	0,882604	2,38061	0,370747	0,00337777
FBgn0000459	1,33504	2,48151	0,537994	0,000252156
...				

Tabela 3: Exemplo da tabela gerada ao final da análise de *dn/ds*. Com esta tabela, podemos importar as informações relevantes para uma grande tabela unificada que conterá todas as informações geradas e será utilizada na etapa de *machine learning*. O valor de p é gerado pelo programa PAML através de um teste exato de Fischer com correção para testes múltiplos, e deve ser menor do que 0.05 para ser considerado significativo.

3.11. Implementação de *machine learning*

Para a geração do modelo de *machine learning* utilizamos o programa R (R Core Team, 2020) ou Rstudio (Rstudio Team, 2020) com as bibliotecas ROSE (Lunardon et. al., 2014) e randomForest (Liaw e Wiener, 2002), que vão ser responsáveis por manipular nossos dados, gerar o modelo e testá-lo. Para que possamos gerar um modelo capaz de separar genes novos e genes antigos, utilizamos a espécie-modelo *D. melanogaster*, que tem sua lista de genes novos identificados, possibilitando o treinamento e controle de qualidade do modelo.

Nossos dados são colocados em uma tabela contendo o nome de cada gene da espécie focal, sua idade (novo ou antigo), e todas as informações relevantes para a implementação do modelo (tabela 5). As variáveis utilizadas para o *machine learning* foram: expressão no testículo (TPM), expressão no ovário (TPM), *fold change*, viés de expressão (testículo, ovário ou nenhum viés), presença de parálogo, informações do primeiro filtro de ortologia (presença ou ausência de ortólogos nas oito espécies e soma das presenças), viés de expressão em cada uma das oito espécies (testículo, ovário, nenhum viés ou ausência de ortólogo), tamanho de proteína, *dn*, *ds*, *dn/ds*.

No caso dos nossos dados, temos um problema chamado de desequilíbrio de classes (*class imbalance*), que ocorre quando uma das classes (genes antigos) é muito mais abundante do que a outra (genes novos), distorcendo o resultado final do *machine learning*. Como discutido anteriormente, existem várias maneiras de resolver esta questão, e neste trabalho utilizamos a biblioteca ROSE para implementar estas soluções.

Gene	Idade	Expressão Testículo	Expressão Ovário	...
FBgn0000045	0	3,57	3,57	
FBgn0003886	1	223,72	0,26	
FBgn0013771	1	7,63	11,72	
...				

Tabela 4: Resumo da tabela incluída para o *machine learning*. Cada gene fica em uma linha que contém todas as informações geradas sobre ele, como expressão em TPM dos tecidos. Para o *machine learning*, precisamos transformar a idade em classes, 0 e 1, que representam genes antigos e novos respectivamente.

Durante este trabalho, utilizamos informações retiradas de bases de dados como Flybase (Larkin *et. al.*, 2021), FlyDivas (Stanley e Kulathinal, 2016) e *ensembl* (Howe *et. al.*, 2021), além das informações geradas *de novo* através de nossas análises. De maneira geral, isto é feito pois precisamos ver se a adição da informação melhora nosso modelo, além de ter um ponto para comparar com os dados que nós mesmos estamos gerando. Assim, testamos primeiro os dados derivados de bases de dados curados, que provavelmente são melhores das informações geradas *de novo*, examinando como cada adição afeta os resultados do modelo de *machine learning*.

O primeiro passo para a aplicação do *machine learning* é a divisão dos dados em treino e teste, com proporções de 80% e 20% respectivamente. Os dados usados para treino irão ser utilizados para dar informações para o *machine learning* e torná-lo capaz de, no nosso caso, separar genes novos e antigos, enquanto os dados usados para teste são necessários para verificar a qualidade do modelo final em dados que não tenham sido usados para formá-lo. Esta divisão é feita aleatoriamente, mas é reproduzível, possibilitando que seja testado se ela está afetando indevidamente os resultados.

No caso dos nossos dados, como mencionado anteriormente, a classe de genes antigos é muito mais abundante do que a classe dos genes novos. Existem algumas soluções possíveis para este problema que afeta o treinamento do modelo, como descrito na Introdução. No nosso caso, o método chamado de *undersampling*, que é a retirada aleatória de membros da classe majoritária na hora do treinamento, faz mais sentido biológico já que existem muitos genes antigos, que provavelmente estão trazendo informações redundantes. No entanto, foram feitos testes com outros métodos, de *oversampling*, que é a repetição de membros da classe minoritária, e a geração de dados sintéticos a partir da classe minoritária. Todas estas implementações foram feitas com a biblioteca ROSE do programa R.

Após o teste do modelo nos dados separados anteriormente, o programa nos dá algumas estatísticas que são utilizadas para examinar a qualidade do modelo. Com estas medidas podemos comparar cada modelo treinado, de maneira a verificar quais os conjuntos de dados e parâmetros que geram os melhores resultados.

3.12. Identificação de genes novos de *D.pseudoobscura*

Para que fosse possível aplicar nosso modelo em uma outra espécie com genes novos conhecidos que não *D. melanogaster*, foi necessário identificar os genes novos em *D. pseudoobscura*, já que nenhuma outra *Drosophila* possui esta informação. Com este objetivo, aplicamos o mesmo método utilizado por Zhang e colaboradores em 2010 e 2019 para identificar os genes novos de *D. melanogaster* e humanos (Zhang *et. al.*, 2010; Shao *et. al.*, 2019).

Como *D. pseudoobscura* não é uma espécie modelo, os arquivos necessários para a aplicação do método não estão disponíveis, portanto, temos que gerá-los nós mesmos. Para isto, precisamos detectar as regiões repetitivas com o *RepeatMasker* (Smit *et. al.*, 2013), fazer o alinhamento dos genomas e preparar diversos outros arquivos auxiliares. Para o alinhamento dos genomas utilizamos o programa *lastz* (Harris, 2007), juntando todos os *scaffolds* do genoma em um só arquivo FASTA e registrando os locais de junção de cada *scaffold*, conforme instruído pela metodologia já existente.

No entanto, alguns arquivos gerados ficam grandes demais e precisam ser fragmentados, como o caso de *D. persimilis*. Estas espécies tiveram os arquivos fragmentados em 10 partes e as coordenadas da fragmentação salvas para a reunião posterior. No entanto, mesmo assim o alinhador não estava sendo capaz de completar a operação para esta espécie. Após pesquisar sobre este tipo de problema, fragmentamos o arquivo FASTA em mais de cem partes, e escrevemos um *script* na linguagem *bash* para alinhar e processar os arquivos sem excessivo trabalho manual.

Os parâmetros dos alinhamentos são os mesmos que foram utilizados na comparação de *D. melanogaster* com o resto do gênero *Drosophila*, e foram obtidos a partir de informações disponíveis no UCSC (Kent *et. al.*, 2002). Com os arquivos montados e os parâmetros obtidos, são feitos os alinhamentos da espécie alvo, neste caso *D. pseudoobscura*, com as outras espécies incluídas no projeto, que nesta etapa são as outras espécies do projeto de 12 genomas (*Drosophila 12 genome consortium*, 2007).

Após os alinhamentos, os arquivos resultantes têm formato *lav* e são processados de maneira a recuperarem suas coordenadas originais e gerar arquivos no formato *chain*, usados para o método de datação propriamente dito. Para isto, são utilizados *scripts* em *bash* e *perl* escritos por membros do laboratório do prof. Dr. Yong E. Zhang (Pequim, China), além de programas do UCSC *toolkit* (Kent *et. al.*, 2002). Com isso, passamos por alguns formatos de arquivos intermediários até terminar no formato desejado, *chain*, com as coordenadas verdadeiras. Estes arquivos são formados a partir das comparações de *D. pseudoobscura* com as outras espécies, e são utilizados para gerar os alinhamentos recíprocos.

Este método de datação utiliza a formação de uma base de dados em *mysql* que irá servir para que possamos fazer as comparações entre alinhamentos, encontrar regiões sintênicas, remover repetições e datar os genes. Para isto, utilizamos a base de dados para *mysql* de *D. pseudoobscura* do *ensembl* (Howe *et. al.*, 2021), além dos arquivos de anotação (gtf) e de relação entre gene e transcrito da mesma base de dados.

Com esta base de dados pré-existente como fundação, são utilizados *scripts* nas linguagens *perl* e *mysql* escritos por membros do laboratório do prof. Dr. Yong E. Zhang e programas do *core api*

do *ensembl* (Howe *et. al.*, 2021) para popular a base de dados com nossos dados de alinhamentos entre as espécies, tamanho dos genomas e éxons e a localização dos elementos repetitivos.

Com a ferramenta overlapSelect do UCSC *tools* (Kent *et. al.*, 2002) somos capazes de detectar a sobreposição de éxons nos alinhamentos que geramos anteriormente e verificar se estas sobreposições não se devem a presença de regiões repetitivas. Após esta etapa, um programa em *perl* verifica se o transcrito correspondente a esta sobreposição é o melhor entre as duas espécies.

Por fim, é feita a análise de sintenia, que vai determinar a idade de cada gene ao analisar sua presença ou ausência nas espécies incluídas. Munidos da informação filogenética do grupo, conseguimos determinar a idade de todos os genes de *D. pseudoobscura*, desde genes exclusivos da espécie até aqueles compartilhados por todas as espécies analisadas. A figura abaixo é um esquema simplificado dos passos necessários para a datação de genes em *D. pseudoobscura*.

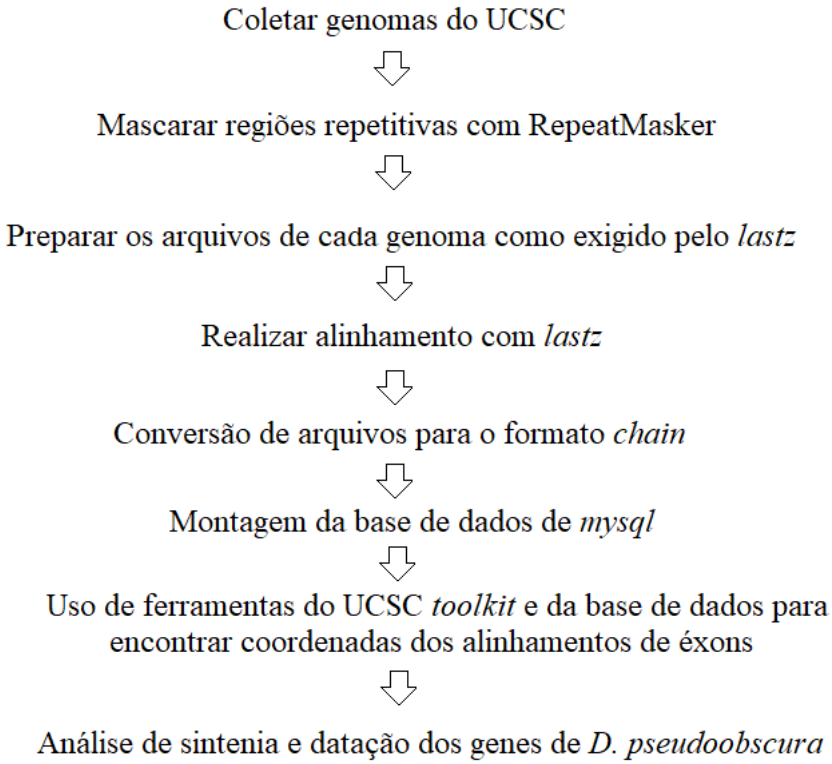


Figura 19: Passos realizados para a datação de genes de *D. pseudoobscura*. Estas etapas foram realizadas em ordem, de maneira a obter as informações relevantes para cada gene da espécie, como descrito nesta seção. Os protocolos para cada passo foram estabelecidos por Zhang e colaboradores em 2010, e os *scripts* e instruções foram cordialmente cedidos por membros do laboratório do Prof. Dr. Yong E. Zhang.

Seguindo o que foi estabelecido por Zhang e colaboradores em 2010, separamos os genes de *D. pseudoobscura* em grupos de idade, de acordo com a presença de ortólogos. Genes que não tem ortólogos em nenhuma espécie são do grupo 4, exclusivo de *D. pseudoobscura*; genes com ortólogo apenas em *D. persimilis*, do grupo 3; se o gene tem ortólogos em espécies do grupo *melanogaster*, do grupo 2; possui ortólogo em *D. willistoni*, grupo1; e, por fim, se o gene possui ortólogo em uma das espécies *D. grimshawi*, *D. virilis* e *D. mojavensis*, grupo 0. Os genes do grupo 0 são considerados genes antigos, e genes dos outros grupos, genes novos. A tabela abaixo é um exemplo da tabela final gerada pelo processo de datação.

Gene	*D. persimilis*	*D. ananassae*	...	*D. virilis*	Idade
FBgn0079870	não	sim		sim	0
FBgn0243605	sim	não		não	3
FBgn0071809	sim	sim		não	2

Tabela 5: Trecho da tabela final gerada no processo de datação. Para cada gene de *D. pseudoobscura*, examinamos a presença de ortólogos em todas as outras espécies para determinar a idade final do gene.

3.13. Verificando os resultados de datação de *D. pseudoobscura*

Após a obtenção da lista de genes novos de *D. pseudoobscura*, realizamos alguns testes de maneira a verificar se os resultados obtidos estão corretos. Para isto, fizemos duas checagens em massa, além de algumas verificações manuais de genes escolhidos aleatoriamente. Em todas as abordagens adotadas, o objetivo era checar se nossa datação agiu da maneira prevista, e se não há erros durante este processo.

Primeiramente, verificamos se os genes que foram chamados de exclusivos de *D. pseudoobscura* possuem ortólogos. Ou seja, esta é uma maneira de verificar se nossa datação não encontrou nenhum ortólogo de um gene, quando este deveria ter ortólogos identificados. Para isto, precisamos utilizar a base de dados *ensembl,* que tem a lista de ortólogos um para um de todos os genes de *D. pseudoobscura* nas outras espécies do gênero. É necessário que utilizemos apenas os ortólogos um para um pois, assim como explicado anteriormente, outros tipos de ortologia confundem as relações entre genes novos e parentais.

Munidos desta tabela, somos capazes de examinar cada gene identificado como exclusivamente de *D. pseudoobscura* e seus ortólogos em cada uma das outras espécies. Para isto, escrevemos um *script* na linguagem *awk* que checa quantos destes genes de *D. pseudoobscura* tem ortólogo em alguma espécie, desde que esta espécie não seja a espécie irmã *D. persimilis*.

Permitimos a presença de ortólogo em *D. persimilis* pois ela é uma espécie extremamente próxima de *D. pseudoobscura*, é a única espécie do projeto que está presente no grupo *obscura* e sua montagem e anotação genômicas tem qualidade mais baixa. Ou seja, se um gene exclusivo do grupo *obscura* não tem seu ortólogo encontrado em *D. persimilis* por algum motivo, ele passa a ser considerado exclusivo de *D. pseudoobscura*.

Nossa outra abordagem em massa foi comparar nossa lista de genes compartilhados entre *D. pseudoobscura* e *D. melanogaster* com a lista obtida por Zhang e colaboradores em 2010. Para isto, utilizamos a lista de ortólogos um para um do *ensembl* de maneira as correspondências entre os genes de *D. pseudoobscura* e *D. melanogaster*. Desta maneira, podemos observar para cada gene compartilhado entre as duas espécies se nós e os trabalhos anteriores os chamam de novos.

Por fim, escolhemos alguns genes aleatoriamente de cada uma das classes e os examinamos mais de perto, verificando sua sintenia no FlyBase com a ferramenta *Gbrowse* (Stein *et. al.*, 2002), no *ensembl* e fazendo BLASTx do *scaffold* onde se encontra com as proteínas de *D. melanogaster*. Estas três estratégias nos ajudam a verificar se nossa datação agiu corretamente caso a caso, ou seja, podemos ver se a quebra de sintenia e a presença de um gene sem ortólogos ocorre como esperado.

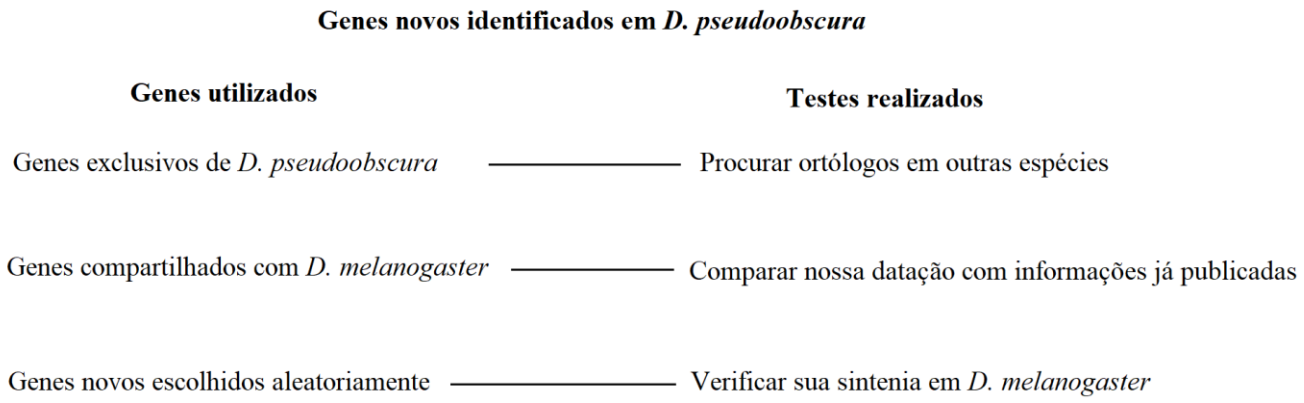


Figura 20: Etapas de verificação realizadas para conferir a datação de genes de *D. pseudoobscura*. Estes processos foram realizados após a datação dos genes de *D. pseudoobscura* e cumprem o mesmo papel de maneiras diferentes. Cada uma destas etapas poderia nos indicar erros cometidos durante o processo de datação, já que usam conjuntos de genes diferentes e examinam questões diferentes.

4. Resultados

4.1. Sequenciamento e controle de qualidade

Nosso objetivo nesta tese foi a identificação de genes novos utilizando suas características biológicas que os diferenciam dos genes antigos. Para isto, precisamos gerar dados relacionados a estas características, que foram apresentadas durante a Introdução (1.3 a 1.6). Portanto, sequenciamos o mRNA de testículo e ovário de diversas espécies do gênero *Drosophila* e de gêneros relacionados, conforme apresentado na seção de Métodos (3.1, 3.2 e 3.4).

Após a extração do mRNA de testículo e ovário, precisamos verificar se não foi cometido nenhum erro durante a dissecção e a extração. Para isto, fazemos um controle de qualidade interno e utilizamos controles gerados pela instituição que faz o sequenciamento, como apresentado na seção de Métodos. Nenhuma das amostras que geramos apresentou sinais de erro durante estas verificações e alguns dos resultados obtidos estão apresentados na tabela e gráfico abaixo (tabela 3 e figura 21).

Espécie	Tecido	mRNA rep 1	mRNA rep 2	RIN rep 1	RIN rep 2
D. erecta	Ovário	307	376	5,3	2
D. erecta	Testículo	99	54	6	4,2
D. ananassae	Ovário	265	479	1,9	1,8
D. ananassae	Testículo	139	111	4.5	6,1
D. pseudoobscura	Ovário	86	195	6,5	4,5
D. pseudoobscura	Testículo	430	92	4,1	7,3
D. virilis	Ovário	401	680	6,3	2,2
D. virilis	Testículo	696	1040	1,5	4,4

Tabela 6: Resultados dos controles de qualidade interno e externo das amostras geradas em 2017. As concentrações de RNA (colunas mRNA rep1 e 2) estão em ug/ul diluidos em 15ul, e RIN é o número de integridade do RNA.

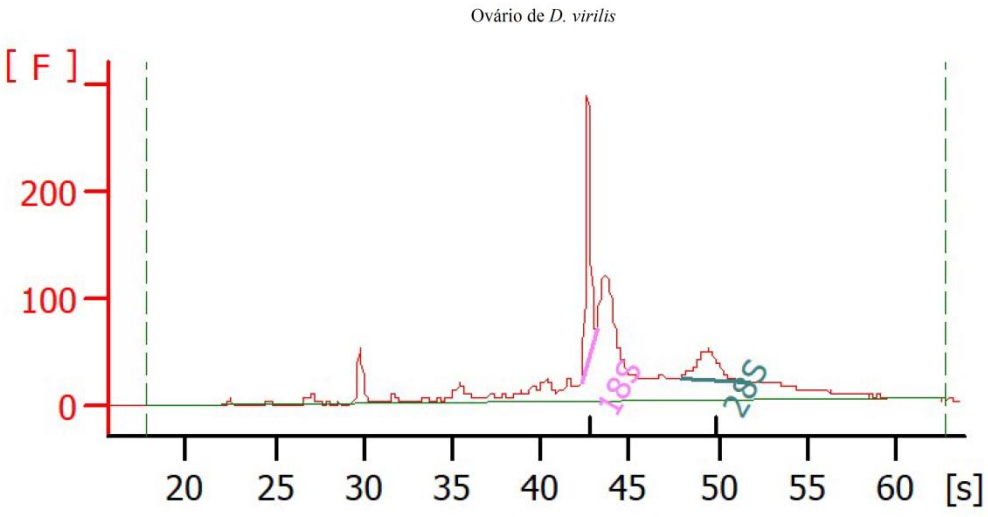


Figura 21: Gráfico de qualidade de uma amostra de ovário de *D. virilis* gerada pela Chicago Functional Genomics Facility com BioAnalyzer 2100. O eixo marcado por [F] representa a fluorescência obtida, enquanto [s] é o tempo decorrido em segundos. As linhas em rosa e verde representam, respectivamente, as subunidades 18S e 28S dos ribossomos.

Após o sequenciamento, as sequências geradas são baixadas do repositório da Chicago Genomics Facility e transferidas para o servidor do laboratório. Após verificar se há algum problema nas sequências utilizando o FASTqc, quantificamos o mRNA e comparamos as réplicas dos tecidos, de maneira a garantir a repetibilidade do experimento.

Para a comparação das réplicas dos tecidos, fizemos testes de correlação entre as réplicas biológicas para ver se os genes tinham valor de expressão semelhante entre as mesmas. Se as réplicas não forem semelhantes entre si, houve algum problema nos experimentos de dissecção, extração ou sequenciamento. Todas as nossas réplicas tiveram resultados semelhantes e satisfatórios (índice de correlação maior do que 0.95) indicando alta consistência dos nossos experimentos, como pode ser visto na figura abaixo das comparações entre as réplicas de ovário de *D. virilis*.

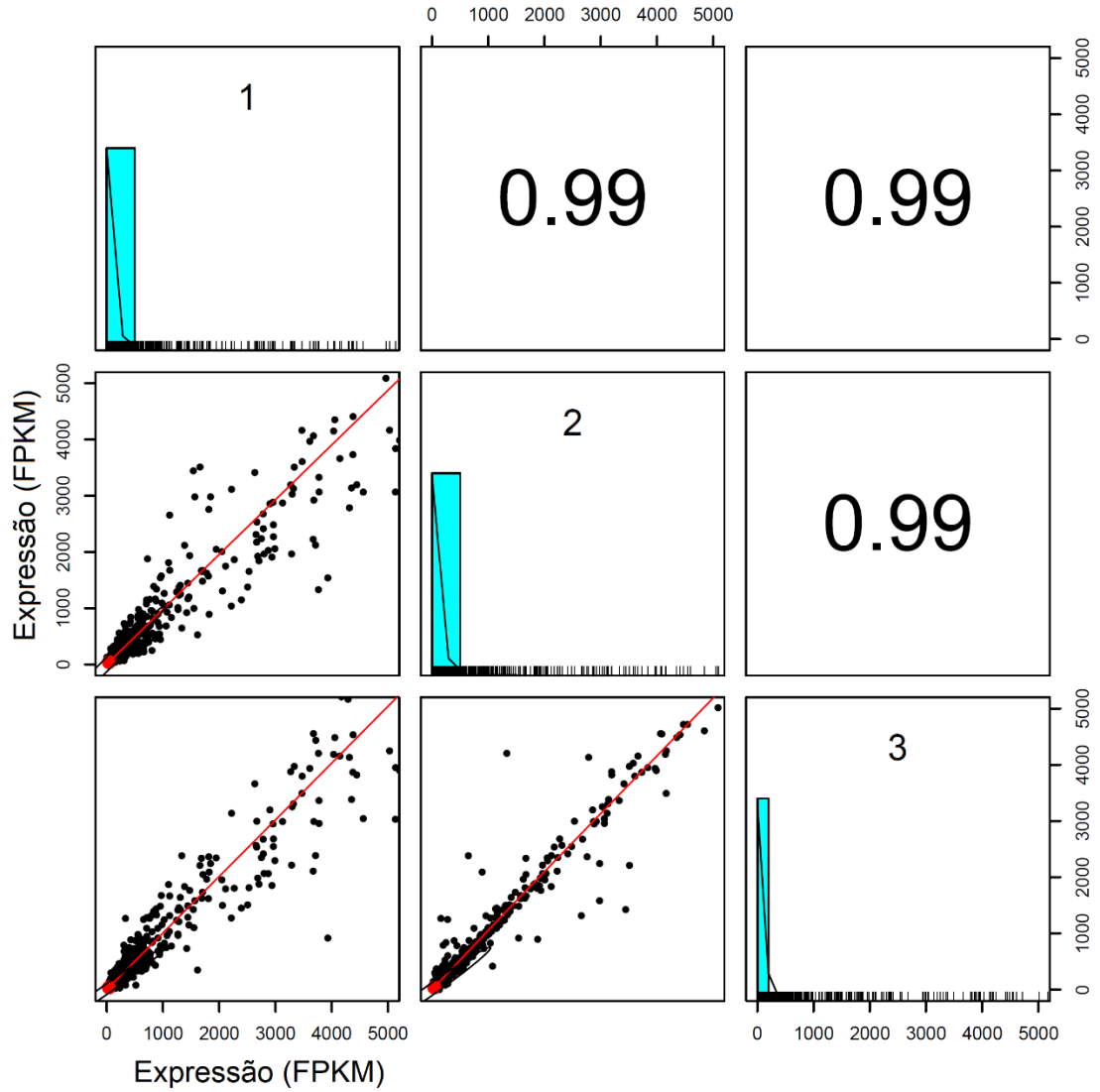


Figura 22: Comparações entre as quantificações de mRNA das réplicas de ovário de *D. virilis*. Os gráficos de distribuição de pontos localizados no canto inferior esquerdo mostram as comparações entre as quantificações de mRNA em FPKM das réplicas em pares, além de uma reta de correlação (vermelho); os histogramas na diagonal mostram as distribuições das quantificações em cada uma das réplicas; os números no canto superior direito são os índices de correlação de Pearson. Assim, podemos ver a alta semelhança entre as réplicas dos experimentos de disseção e sequenciamento para as espécies.

No caso da espécie *D. melanogaster*, baixamos as *reads* que tinham as qualidades que procurávamos, depois montamos os transcriptomas e fizemos alguns controles de qualidade de maneira a escolher a melhor base de dados. Na tabela abaixo estão apresentados alguns resultados do alinhamento do transcriptoma com os *reads* pelo programa *bowtie2* e do BUSCO para duas bases de dados de *D. melanogaster*. Assim, é possível observar que o conjunto de dados número 2 possui resultados consistentemente melhores do que o conjunto de dados número 1.

	Taxa geral de alinhamento do *bowtie*	BUSCOs completos	BUSCOs completos de cópia única	BUSCOs completos e duplicados	BUSCOs fragmentados	BUSCOS ausentes	Número total de BUSCOs buscados
Testículo 1	83,69%	495	163	332	970	1334	2799
Ovário 1	86,69%	1311	823	488	672	816	2799
Testículo 2	94,25%	1023	750	273	635	1141	2799
Ovário 2	91,09%	1697	1174	523	431	671	2799

Tabela 7: Estatísticas de controle de qualidade de montagem de transcriptomas. Para que pudéssemos escolher os melhores dados de *D. melanogaster*, comparamos uma série de resultados até encontrar os melhores *reads* que se encaixavam nas nossas exigências (base de dados número 2 da tabela). A taxa de alinhamento representa a porcentagem dos *reads* que se encontram na montagem final, enquanto o número de BUSCOs completos são as proteínas completas encontradas no transcritpomas montados de acordo com a base de proteínas do BUSCO. Em ambos os casos, os dados derivados do conjunto 2 têm resultados melhores.

Assim, obtivemos os dados de expressão de testículo e ovário de *D. melanogaster* através de bases de dados, selecionadas de acordo com os resultados acima, e os dados das outras espécies com os nossos experimentos de dissecção e sequenciamento.

Com estas informações em mãos, iniciamos a geração das informações biológicas que serão utilizadas para a aplicação no algoritmo de *machine learning*. As seções a seguir detalham os resultados obtidos em cada etapa que gera estas informações, assim como resultados do *machine learning* e da identificação de genes novos de *D. pseudoobscura*, nossa segunda espécie controle.

4.2.Expressão diferencial de genes novos e antigos

Como foi apresentado na seção 1.5 da Introdução, sabemos que que genes novos e antigos possuem perfis de expressão diferentes. No entanto, é importante para nosso projeto verificar se nossos dados reproduzem este padrão e o quão diferente este perfil é considerando os dois grupos de genes como um todo. Esta importância se deve ao fato de que o *machine learning* precisa de variáveis que sejam diferentes entre as classes que tenta separar, portanto precisamos ter certeza de que os padrões esperados são reproduzidos.

Com estes fatores em mente, verificamos a diferença de expressão entre genes novos e antigos utilizando a expressão diferencial calculada com o DEseq2 em *D. melanogaster*. A diferença esperada na proporção de genes novos e antigos foi observada, ou seja, uma porcentagem maior de genes novos apresenta viés para testículo do que os genes antigos, e poucos genes novos tem viés para ovário (tabela 8). Além destas diferenças, observamos que enquanto existem genes antigos que tem expressão preferencial no testículo (3383 de 12013), praticamente não existem genes novos com expressão preferencial no ovário (22 de 1070).

Para que pudéssemos analisar mais a fundo estes padrões, fizemos uma tabela de risco relativo (Sistrom e Garvan, 2004), onde comparamos genes novos e antigos em relação a esta diferença e testamos para ver se um gene novo tem maior probabilidade de ser enviesado para testículo e um gene antigo ser enviesado para ovário (tabela 8).

	Genes novos	Genes antigos	Risco relativo
Viés para testículo	543 (50,7%)	3383 (28,16%)	1,8020
Viés para ovário	22 (2%)	1794 (14,93%)	0,1377

Tabela 8: Quantidade de genes novos e antigos que tem viés de expressão para ovário e testículo e resultado do risco relativo calculado para cada ocorrência. Apesar da ocorrência de genes antigos que tem viés de expressão para testículo, a diferença entre a proporção de genes novos e antigos leva ao alto risco relativo.

Após estas observações, ficamos interessados em descobrir se a dimensão da diferença da expressão entre os tecidos (*fold-change*) é diferente entre genes novos e antigos. Com este objetivo, repetimos este teste separando os genes em classes que possuíam membros apenas com um *fold-change* maior do que um certo valor, como três por exemplo. Como já tínhamos observado que genes novos têm maior probabilidade de ser enviesado para testículo, e genes antigos para ovário, queríamos saber se esta probabilidade aumenta para genes que possuem maior expressão diferencial entre os tecidos.

Observamos isto quando se tratam de genes com expressão enviesada para testículo, conforme limitamos as classes com valores de *fold-change* maior, vemos que o risco relativo também aumenta. No entanto, o mesmo não se repete quando analisamos os genes enviesados para ovário, provavelmente pois o número de genes novos total expressos de maneira enviesada no ovário em cada classe é muito pequeno. Os resultados destas análises podem ser vistos nos gráficos abaixo.

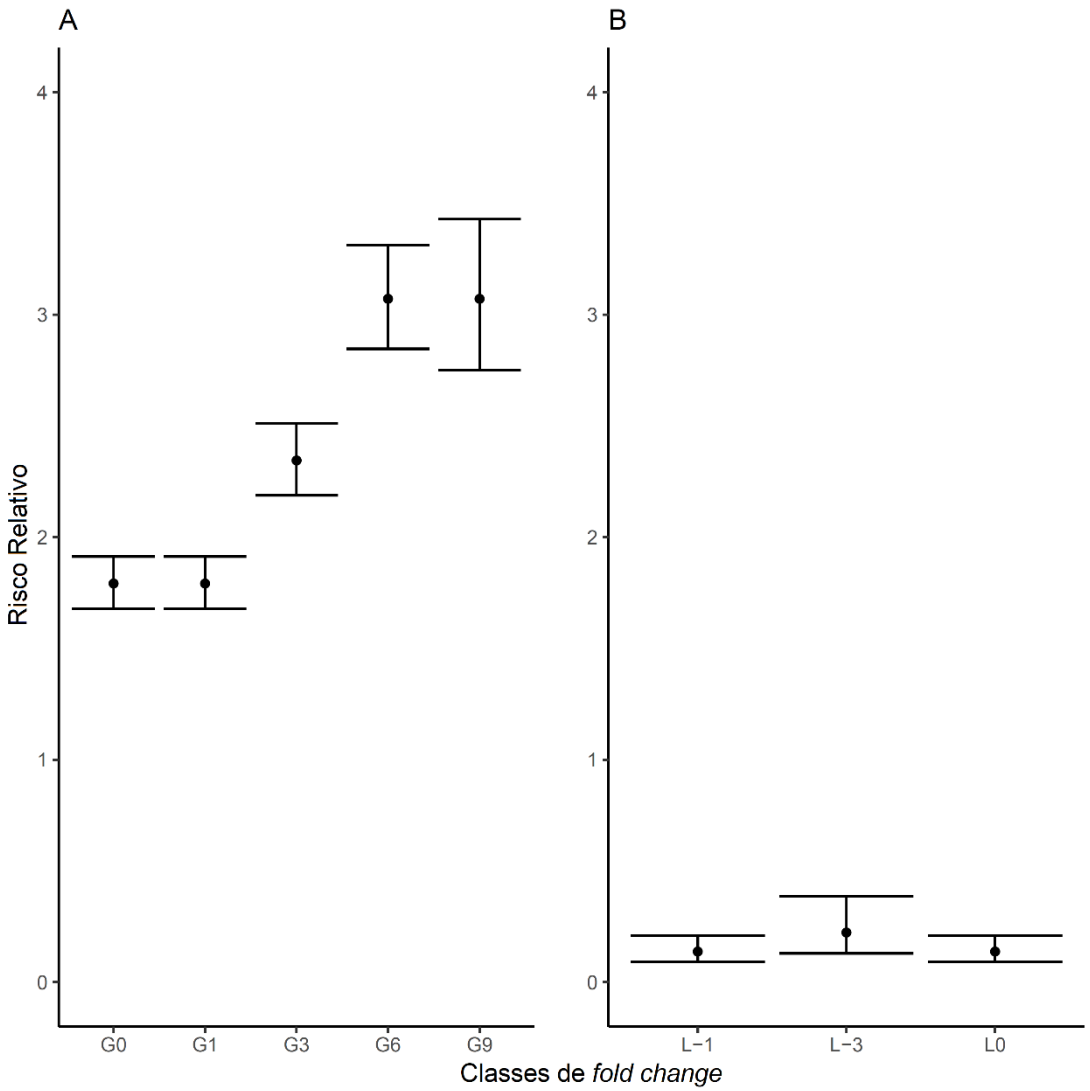


Figura 23: Risco relativo de genes separados por diferentes *folds* de expressão diferencial de testículo (A) e ovário (B). As classes são formadas por genes que possuem expressão diferencial no testículo em relação ao ovário, e possuem log2 de *fold change* maior do que 3, para o grupo G3, por exemplo. G0 representa os genes com *fold change* maior que 0, ou seja, são enviesados para testículo, enquanto L -1 representa os genes com *fold change* menor do que -1, que são enviesados para ovário.

Estes resultados ressaltam a importância das informações relacionadas a padrões de expressão para a separação entre genes novos e antigos. Assim, será importante para o algoritmo de *machine learning* não apenas a informação do viés de expressão, mas também o tamanho da diferença entre os dois tecidos e os níveis de expressão em cada tecido.

4.3. Cálculo de *dn/ds*

Além dos perfis de expressão apresentados acima, outra característica importante que é diferente entre genes novos e antigos é a presença de sinais de seleção positiva em excesso em genes novos quando comparados a antigos. Da mesma maneira como apresentado anteriormente, é importante que sejamos capazes de observar o padrão esperado nos nossos dados para que eles possam ser utilizados no *machine learning*.

Com este fim, calculamos o *dn/ds* dos genes de *D. melanogaster* que tiveram seus ortólogos em *D. erecta* identificados através do método de *reciprocal best hits* como descrito anteriormente.

Após estes cálculos, separamos os genes novos dos genes antigos e fizemos histogramas das duas classes como observado nos gráficos abaixo.

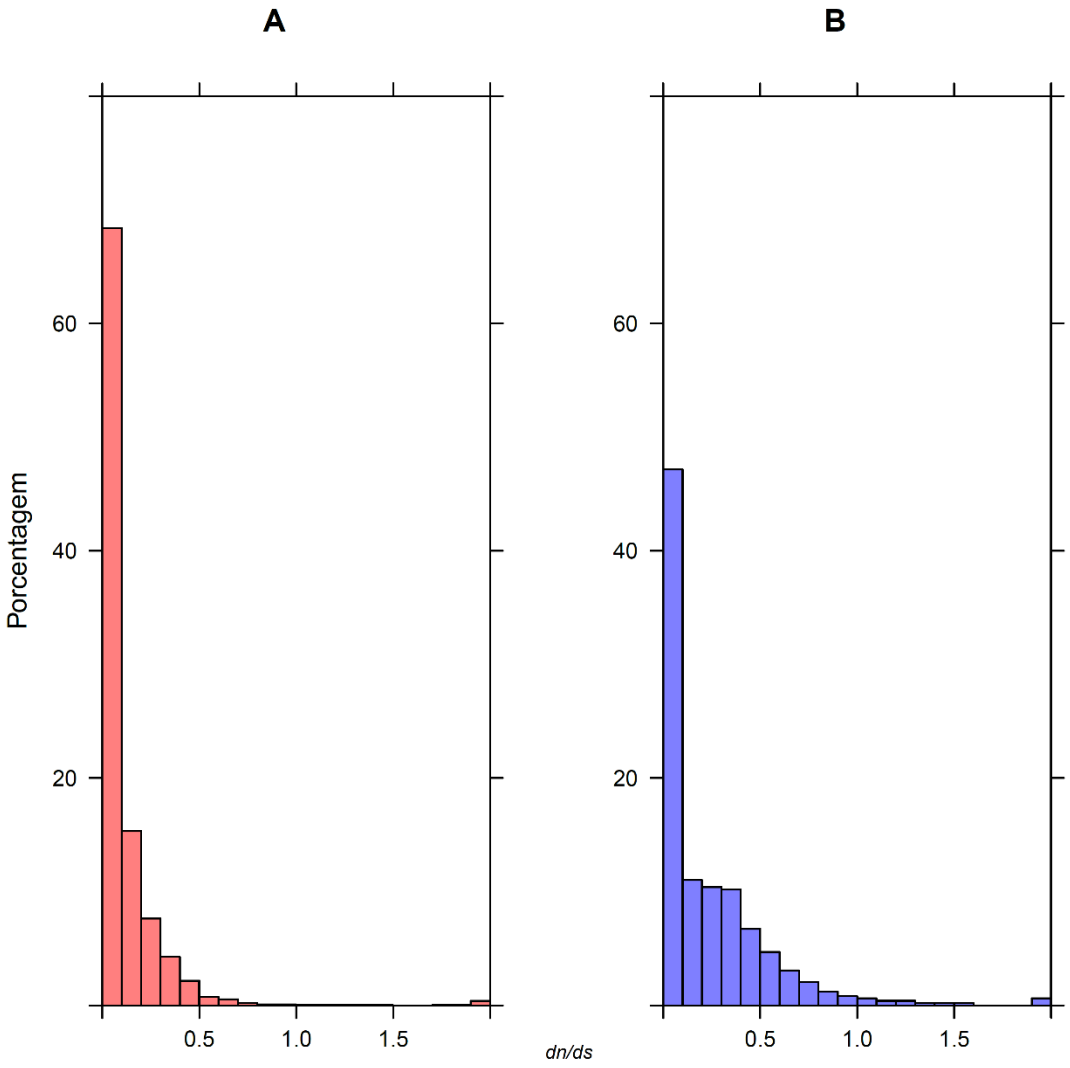


Figura 24: Histogramas da distribuição de valores de *dn/ds* para genes antigos (A) e novos (B) de *D. melanogaster*. Para esta análise foram utilizados 490 genes novos e 7900 genes antigos. Ao examinar os dois gráficos, observa-se que genes antigos possuem, de maneira geral, valores menores de *dn/ds*.

Com estas informações, realizamos um teste de Wilcoxon para verificar se a diferença entre as distribuições de genes novos e antigos era significativa. Este teste deu resultado significativo ($p < 2{,}2^{-16}$), o que pode ser devido ao fato de que genes novos tem, em média, maior sinal de seleção positiva ou à diferença entre as proporções de genes com seleção negativa forte nas duas classes, mais do que 60% para genes antigos e menos do que 50% para genes novos.

portanto, unido a observação dos gráficos foi possível constatar que genes novos têm, como um todo, maior sinal de seleção positiva ou ao menos relaxamento de seleção negativa do que genes antigos.

Após a geração destes dados, e do exame dos padrões encontrados, precisávamos comparar nossos resultados com os dados da base de dados FlyDivas. Existiam diferenças fundamentais entre a maneira que geramos os nossos dados e os dados do FlyDivas, como o fato de que utilizamos transcriptomas, a maneira que encontramos ortólogos e a espécie utilizada na comparação. No entanto, ainda é importante a comparação destes resultados para verificar se, mesmo considerando estas diferenças, os dois resultados teriam certa semelhança. Assim, comparamos todos os valores

de *dn/ds* calculados das duas maneiras nos genes de *D. melanogaster* que aparecem em ambos conjuntos de dados (figura 25).

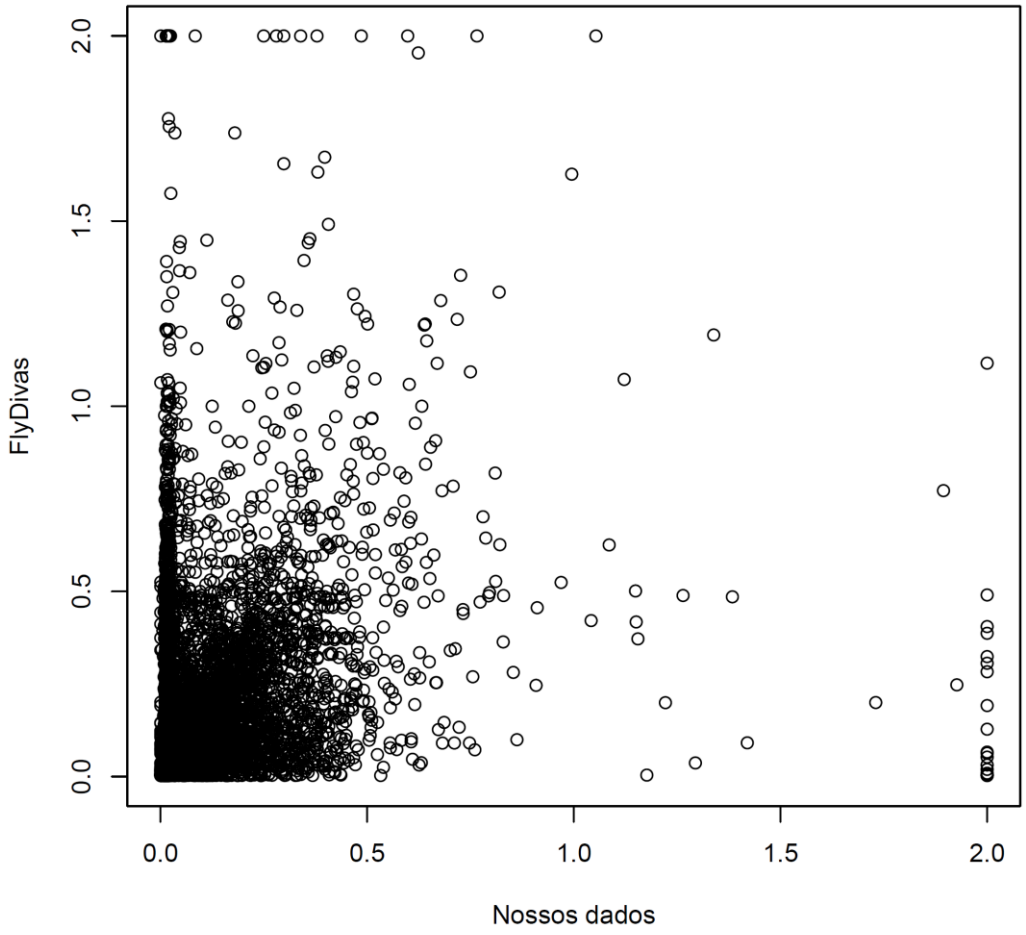


Figura 25: Comparação de valores de *dn/ds* obtidos pela base de dados FlyDivas e gerados neste projeto para 6830 genes de *D. melanogaster*. Como esperado, há uma grande diferença entre os dois dados, já que utilizam informações iniciais diferentes e a base de dados foi curada pelo grupo que a publicou. Além disso, observa-se que os genes têm, em média, valores maiores de *dn/ds* nos dados do FlyDivas.

Apesar desta grande diferença entre os dados de FlyDivas e aqueles produzidos por nós mesmos, é importante frisar que os dois conjuntos de dados apresentam os mesmos padrões para genes novos e antigos, como será visto na Figura 26, e possuem uma fraca correlação positiva (índice de correlação de Pearson = 0,242; $p < 2.2^{-16}$). Ou seja, as distribuições dos valores de *dn/ds* são significativamente diferentes entre genes novos e antigos nos dados de FlyDivas, como pode ser visto nos gráficos abaixo e no teste de Wilcoxon que realizamos ($p < 2{,}2^{-16}$). Assim, mesmo com as diferenças que existem nos dados o padrão biológico para genes novos e antigos se repete.

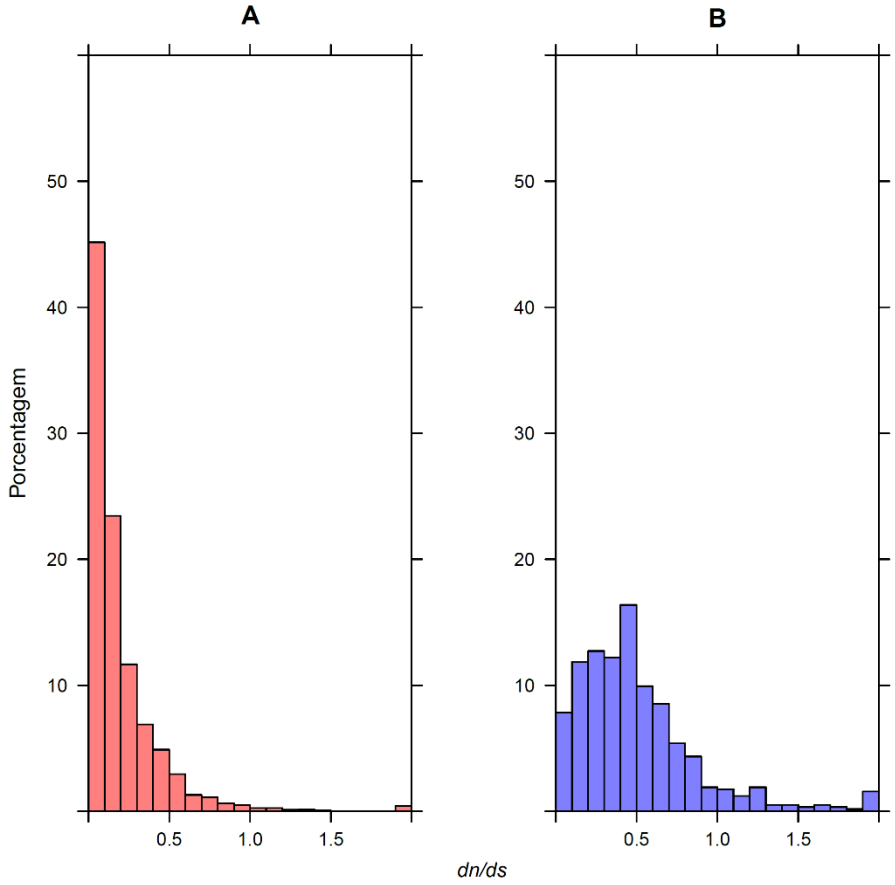


Figura 26: Histogramas de *dn/ds* para genes antigos (A) e novos (B) dos dados retirados da base de dados FlyDivas. Nesta análise, foram utilizados 574 genes novos e 9759 genes antigos. Como esperado, genes novos possuem, em média, valores maiores de *dn/ds* do que genes antigos. Também é possível observar que mais genes possuem valores maiores de *dn/ds* do que nos nossos cálculos *de novo*.

Após as comparações com informações de bases de dados, voltamos aos nossos dados para verificar se as diferenças entre genes novos e antigos se mantinham após o processo de filtragem por ortologia.

Esta verificação é importante pois apenas os genes que não são eliminados no filtro de ortologia serão utilizados no *machine learning* e, portanto, são capazes de afetar nossos resultados finais. Com este intuito, fizemos os mesmos gráficos e testes realizados anteriormente com o conjunto de dados filtrado, que exibiu padrões semelhantes e diferença significativa com teste de Wilcoxon ($p = 1{,}374^{-6}$) (figura 27).

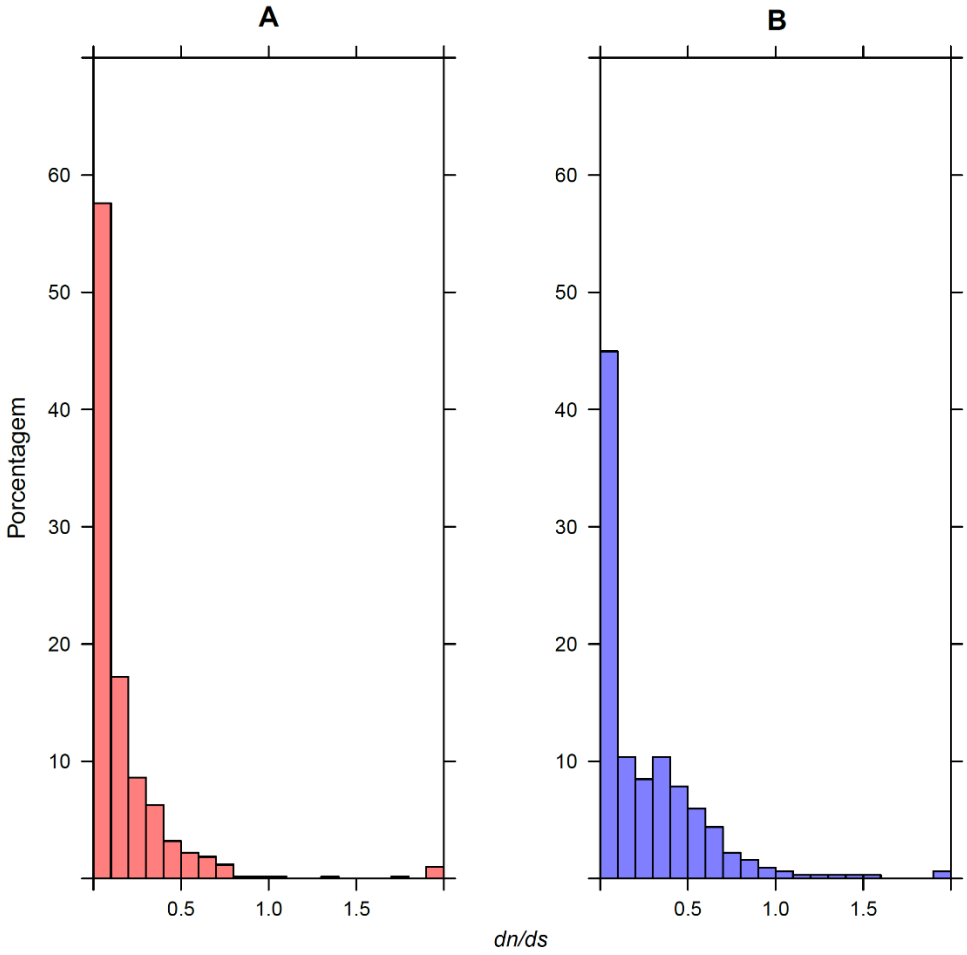


Figura 27: Histogramas da distribuição de valores de *dn/ds* para genes antigos (A) e novos (B) após o filtro de ortologia. Para gerar estes gráficos foram utilizados 321 genes novos e 659 genes antigos. Ao examinar os dois gráficos, observa-se que os padrões observados anteriormente se mantêm e genes novos possuem valores maiores de *dn/ds*, em média.

Ao fim destas análises, vimos que genes novos e antigos mantêm o padrão esperado pela literatura, mesmo quando calculamos estes valores utilizando nossos dados e após a aplicação do filtro de ortologia. Com isso, decidimos utilizar estas informações como variáveis para nosso modelo de *machine learning*, cujos resultados serão apresentados posteriormente.

4.4.Filtro de ortologia

Após alguns testes iniciais com o algoritmo de *machine learning*, decidimos testar se a eliminação de genes antigos antes da etapa de *machine learning* poderia melhorar nossos resultados. Nesta seção, serão apresentados os resultados diretos da aplicação dos filtros de ortologia na eliminação de genes antigos e novos, enquanto a maneira com a qual estes filtros afetam o *machine learning* será apresentada em uma seção posterior (4.6).

Nas seções de Material e Métodos (3.8 e 3.9), explicamos como encontramos os ortólogos de cada gene de *D. melanogaster* e como construímos nossos filtros de ortologia. De maneira geral, precisamos comparar os genes de *D. melanogaster* aos transcriptomas das outras espécies de maneira a encontrar os genes ortólogos em cada espécie.

Após os testes com métodos de encontro de ortólogos (seção 3.8), utilizamos o *reciprocal best hits* para realizar estas comparações. A partir disso, construímos uma tabela que contém a

informação de quais espécies possuem ortólogos para cada gene de *D. melanogaster*, assim como o número total de espécies com ortólogos (tabela 9).

Gene	Idade	*D. erecta*	*D. ananassae*	...	Soma
FBgn0000078	Novo	0	0	...	0
FBgn0001990	Novo	0	0	...	3
FBgn0000357	Antigo	1	1	...	7
FBgn0002842	Antigo	1	1	...	8

Tabela 9: Trecho da tabela montada a partir de informações de ortologia obtidas com *reciprocal best hits.* Todos os genes de *D. melanogaster* são incluídos nesta tabela, que contém a informação de idade, presença (1) ou ausência (0) de ortólogos em cada espécie e número total de espécies com ortólogo para cada gene.

Com esta tabela, podemos ver em quantas espécies cada gene possui ortólogos, e separar os genes antigos e novos. Como utilizamos um (1) para presença, e zero (0) para ausência, podemos somar os valores da tabela para cada gene e ver o número final de ortólogos. Assim, fizemos histogramas para genes novos e genes antigos de maneira a observar a diferença entre as duas classes. Como esperado, os genes antigos possuem, de maneira geral, mais ortólogos encontrados do que os genes novos (figura 28).

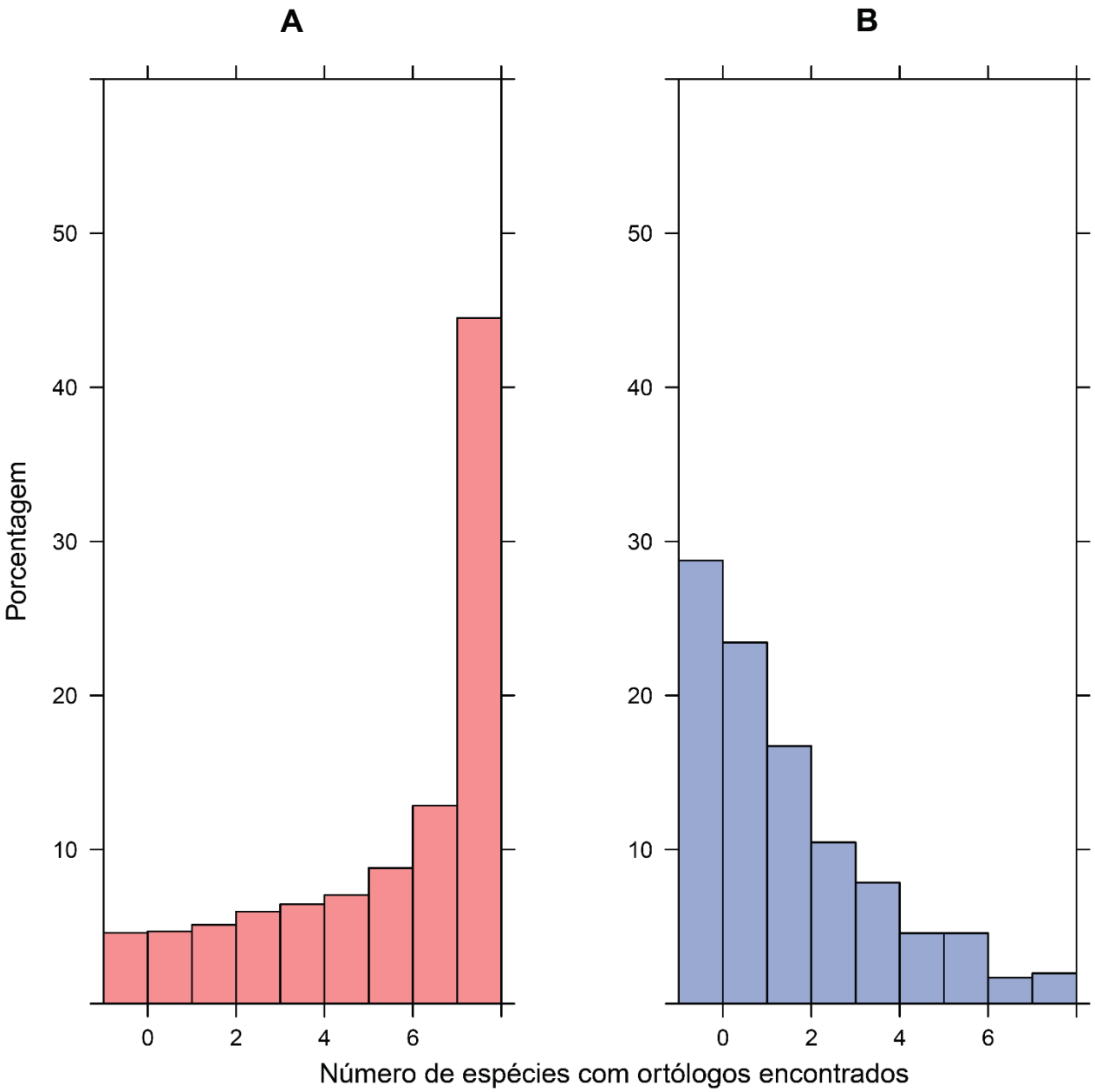


Figura 28: Distribuição do número de espécies com ortólogos encontrados para genes antigos (A) e novos (B). Observa-se que os genes antigos, em sua maioria, têm mais ortólogos encontrados do que genes novos, como esperado. Assim, podemos utilizar esta informação para remover genes antigos da nossa análise (nesta análise foram utilizados 12013 genes antigos e 1070 genes novos).

Com esta informação, investigamos qual seria a melhor maneira de eliminar um grande número de genes antigos, sem perder muitos genes novos. Isto é importante pois facilita a aplicação posterior do *machine learning* e melhora os resultados finais, como será discutido posteriormente. Desta maneira, fizemos diversos cortes eliminando genes que possuíssem ortólogos em um certo número de espécies e verificamos o número de genes novos e antigos eliminados (tabela 10).

Número de espécies com ortólogos	Genes novos eliminados (n = 1070)	Genes antigos eliminados (n = 12013)
3 ou mais	333 (31,12%)	10284 (85,6%)
4 ou mais	221 (20,65%)	9566 (79,63%)
5 ou mais	137 (12,8%)	8792 (73,18%)
6 ou mais	88 (8,22%)	7945 (66,13%)

Tabela 10: Número de genes novos e antigos eliminados baseados no número de espécies com ortólogos, com a exigência de que um dos ortólogos seja uma das espécies externas ao gênero *Drosophila*. Com estes cortes, conseguimos eliminar uma grande quantidade de genes antigos e minimizar a perda de genes novos de maneira a melhorar os resultados do modelo final de *machine learning* (n = total de genes de cada classe).

Após a aplicação deste filtro, nos perguntamos se havia outra maneira de eliminar mais genes antigos em uma segunda etapa, analisando as mesmas informações de maneira diferente. Com este intuito, incluímos contexto filogenético no encontro de ortólogos, ou seja, a presença de um ortólogo em uma espécie próxima a *D. melanogaster* não representa a mesma informação do que a presença de um ortólogo em uma espécie distante. Da mesma maneira, algumas espécies incluídas neste projeto pertencem ao mesmo grupo filogenético, assim, a presença de ortólogo em *D. virilis* e *D. mojavensis* pode ser interpretada como apenas uma informação, já que estas espécies fazem parte de um grupo monofilético.

Para que isso fosse possível, modificamos a tabela já existente para que passasse a representar estas informações filogenética. Assim, algumas espécies foram unidas em um mesmo grupo filogenético e a informação deixou de ser apenas zero ou um para um sistema de pontos no qual a presença na espécie vale mais pontos dependendo da sua distância filogenética com *D. melanogaster*. Por exemplo, se um gene está presente em *D. erecta* que é a espécie mais próxima de *D. melanogaster* utilizada neste trabalho ela recebe um ponto, enquanto a presença em *D. ananassae* dá 2 pontos e assim por diante, até a presença em um dos nossos grupos externos *Chymomyza* e *Scaptodrosophila* que dá 32 pontos (tabela 11).

Gene	Idade	*G1*	*G2*	*G3*	*G4*	*G5*	Soma
FBgn0000078	Novo	0	0	0	0	0	0
FBgn0001990	Novo	0	2	4	8	0	14
FBgn0000357	Antigo	1	2	4	8	16	31
FBgn0002842	Antigo	1	2	4	8	16	31

Tabela 11: Exemplo de tabela do novo filtro de ortologia. Os grupos são definidos pelas espécies de acordo com sua distância de *D. melanogaster*. Os grupos são: G1 (*D. erecta, D.*ananassae), G2 (*D.pseudoobscura)*, G3 (*D. willistoni)*, G4 (*D. virilis, D. mojavensis)* e G5 (*Scaptodrosophila e Chymomyza)*.

A escolha deste tipo de pontuação é importante pois um gene só terá mais do que 32 pontos se possuir um ortólogo em uma das espécies de grupos externos e só terá 3 pontos se só possuir ortólogos nas espécies mais próximas de *D. melanogaster*, por exemplo, facilitando nossa análise e a compreensão dos resultados. Assim como feito com o filtro anterior, examinamos as distribuições dos valores de soma (figura 29) para verificar se os genes antigos irão possuir valores maiores de soma das pontuações, como previsto.

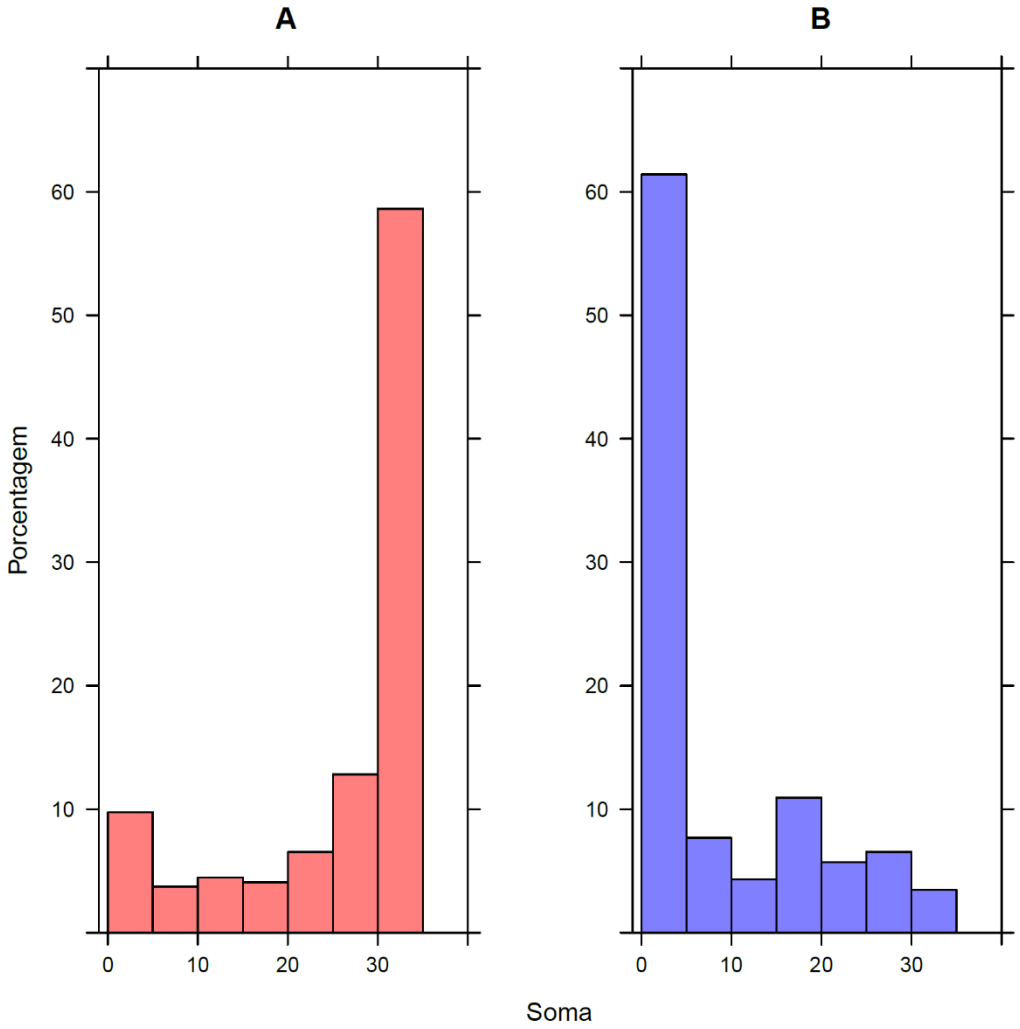


Figura 29: Histogramas das somas dos valores do filtro de ortologia em genes antigos (A) e novos (B). Ao separar genes novos e antigos podemos observar a grande diferença entre estas duas classes em mais este tipo de filtro de ortologia. Estes resultados também seguem os padrões esperados e podem nos ajudar a eliminar mais genes antigos.

Após examinar os nossos dados, fizemos alguns testes utilizando este novo filtro após a aplicação do filtro anterior para que possamos eliminar mais genes antigos com o mínimo de perda de genes novos. Refizemos os histogramas apresentados acima, mas incluindo apenas os genes que não foram eliminados na etapa anterior e fizemos alguns cortes com este segundo filtro assim como no primeiro (figura 30, tabela 12).

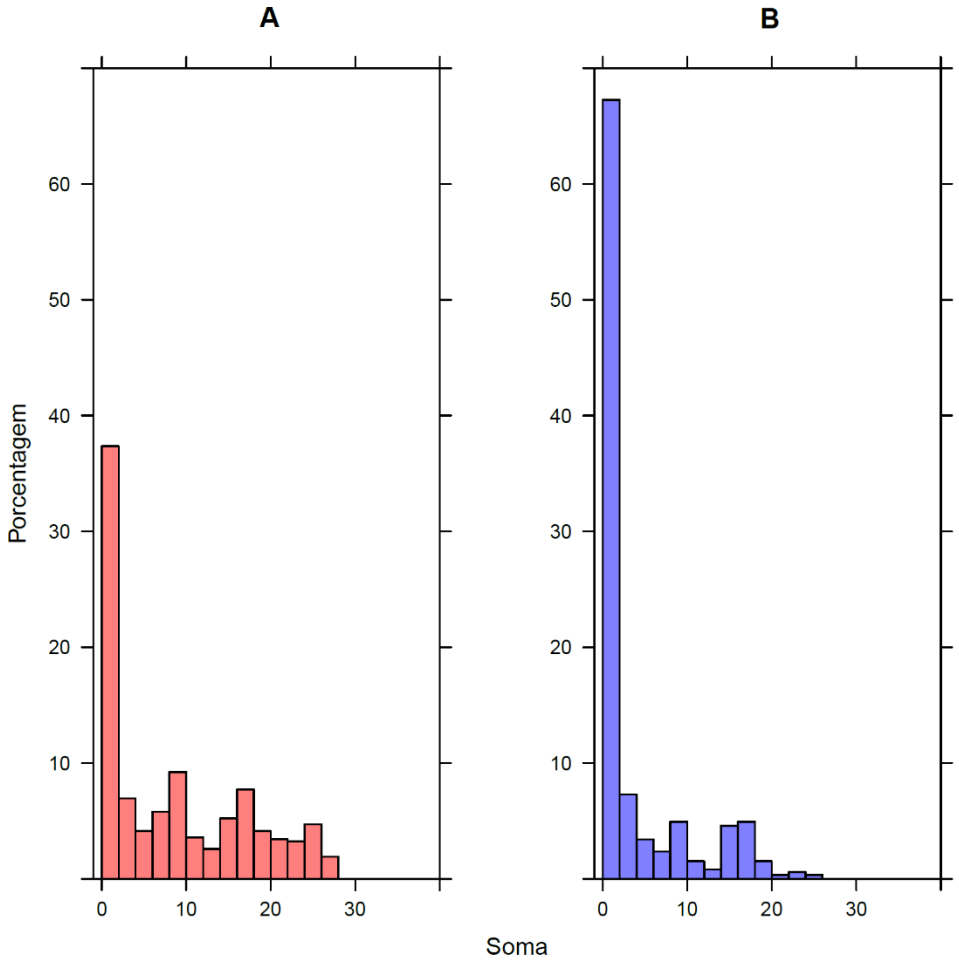


Figura 30: Histogramas das somas dos valores do filtro de ortologia em genes antigos (A) e novos (B) que não foram eliminados no primeiro filtro. Observa-se que mesmo após a passagem do primeiro filtro, mantém-se a diferença entre genes antigos e novos.

Valor limite da soma	Genes novos eliminados (n = 849)	Genes antigos eliminados (n = 2447)
16	105 (12,37%)	743 (30,36%)
17	76 (8,95%)	615 (25,13%)
18	26 (3,06%)	461 (18,83%)
19	24 (2,82%)	426 (17,4%)
20	16 (1,88%)	373 (15,24%)
21	11 (1,29%)	325 (13,28%)

Tabela 12: Número de genes novos e antigos eliminados baseados no segundo filtro, após a passagem do primeiro filtro. Ainda mantemos o mesmo objetivo de eliminar o maior número possível de genes antigos sem perder muitos genes novos (n = número de genes de cada classe que não foram eliminados no primeiro filtro).

O processo total de filtragem por ortologia passa, então a ser feito em duas etapas, primeiro eliminamos genes que estão presentes em quatro ou mais espécies, desde que uma seja *Chymomyza* ou *Scaptodrosophila*; depois eliminamos genes que tenham vinte pontos ou mais no segundo filtro. Assim, com estes filtros ficamos com 833 genes novos e 2074 genes antigos.

4.5.Calculando expressão diferencial em todas as espécies

Existem dois fatores que nos levaram a questionar se genes novos e antigos possuem dinâmicas diferentes de evolução da expressão. Primeiro, genes novos podem estar em um contexto genômico diferente dos seus genes parentais, dependendo de sua origem e história evolutiva. Segundo, foi postulada a hipótese "*Out of testis*" (Kaessman, 2010) que diz que genes novos são expressos nos testículos quando surgem e quando este surgimento é muito recente, e passam a ser expressos em outros tecidos conforme vão envelhecendo.

Utilizando os ortólogos encontrados na busca por *reciprocal best hits*, podemos verificar, para cada gene de *D. melanogaster*, a expressão do seu gene ortólogo encontrado em cada uma das espécies incluídas neste projeto. Após a obtenção destes dados, incluímos a informação da expressão do viés de expressão dos ortólogos no algoritmo de *machine learning*. Assim, para cada gene foi incluída a informação do viés de expressão dos seus ortólogos, ou seja, se eles têm viés para testículo, ovário, nenhum viés ou não foi encontrado na espécie. Estas informações melhoraram os resultados do *machine learning*, como será demonstrado com maior detalhe na próxima seção e, portanto, são importantes para nossa análise.

Depois de ver a melhora dos resultados do *machine learning*, queríamos verificar como evolui o viés da expressão em genes novos e antigos, para ter uma estimativa de quão diferentes são estes padrões nas espécies que estamos examinando e para ver se haveria mais alguma informação que poderia ser utilizada no algoritmo. Para facilitar a visualização e interpretação desses dados, fizemos um contador de mudanças, ou seja, para cada gene de *D. melanogaster* verificamos qual é o seu viés de expressão e para cada padrão diferente encontrado em uma das outras espécies, adicionamos 1 no nosso contador. Assim, nosso contador de mudanças pode ter um valor de zero a oito mudanças, já que temos oito espécies além de *D. melanogaster* neste projeto. Os gráficos abaixo mostram a distribuição dos valores do contador para genes novos e antigos que tem pelo menos uma mudança, antes e depois do nosso filtro de ortologia.

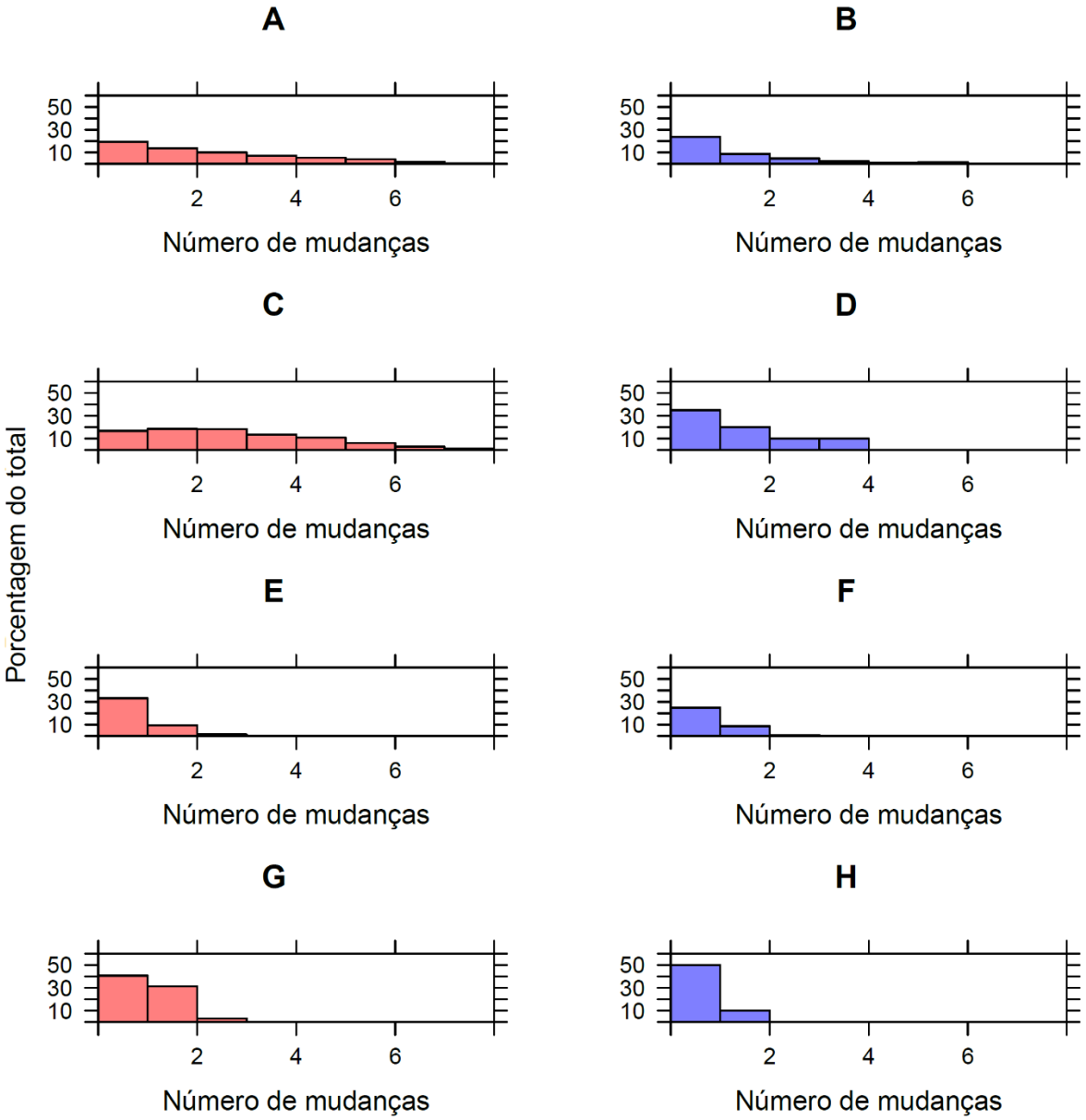


Figura 31: Histogramas da distribuição dos valores do contador de mudanças de expressão em genes antigos (vermelho) e novos (azul). Os gráficos A e B são provenientes dos genes que têm viés em testículo em *D. melanogaster*, enquanto os gráficos C e D são dos genes que têm viés em ovário na espécie. Os gráficos E a H seguem a mesma lógica dos gráficos A a D, porém representam apenas os genes que não foram eliminados no filtro de ortologia.

Após visualizarmos esses resultados, realizamos testes de Wilcoxon para verificar se a diferença entre cada par de distribuições era significativa. Antes da filtragem foram feitos dois testes: para genes que tem viés de testículo em *D. melanogaster* (430 genes novos e 3323 genes antigos; p = $1{,}1^{-12}$) e para genes que tem viés de ovário na espécie (20 genes novos e 1778 genes antigos; p = 0,002917). Após a filtragem por ortologia, fizemos testes semelhantes: para genes com viés em testículo em *D. melanogaster* (273 genes novos e 387 genes antigos; p = 0,5837) e com viés em ovário (10 genes novos e 32 genes antigos; p = 0,2042).

Assim, observa-se que a diferença é significativa (p < 0,05) para os dados antes do filtro de ortologia, mas não significativo para os genes que restam após a filtragem. Esta diferença pode ocorrer por dois motivos: os genes antigos que passam pelo filtro são semelhantes aos genes novos que também passam, ou que o filtro reduz demais o alcance do nosso contador, que passa a ser de zero a três, impossibilitando a observação de diferenças significativas.

Mesmo sabendo que esta diferença não era significativa, decidimos testar a inclusão destes dados no *machine learning* de maneira a verificar se isto afetaria positivamente nossos resultados. Como esperado, esta inclusão não melhora *precision* e nem *recall* dos modelos, confirmado que não seriam úteis para nossa análise.

Apesar destes resultados, ainda queríamos explorar mais possibilidades com estes dados e pensamos que os padrões que encontramos poderiam ser causados pela maneira com a qual estávamos analisando as informações. Assim, fizemos uma série de testes para verificar se havia alguma maneira de mudar este contador de mudanças e outras maneiras de olhar para estes dados de viés de expressão.

No entanto, nenhum dos testes realizados foram capazes de melhorar o modelo de *machine learning*, igual a análise apresentada acima. Ou seja, apenas o dado bruto do viés da expressão dos ortólogos de cada gene nas espécies do projeto foi capaz de auxiliar o algoritmo a separar genes novos e antigos.

A causa mais provável desta diferença entre a utilidade dos dados brutos é que nossas análises estão perdendo informação em relação ao dado inicial, como quando os condensamos em um contador, por exemplo. Portanto, mantemos apenas o uso dos dados brutos no algoritmo de *machine learning*, como explicado no início desta seção.

4.6. *Machine learning*

Munidos destas informações biológicas (seções de 4.1 a 4.5), aplicamos algoritmos de *machine learning* para tentar separar genes novos de genes antigos. Utilizamos *D. melanogaster* como ponto de início, pois esta espécie tem a lista de genes novos identificados que precisamos para treinar e testar nosso modelo. Com isto em mente, nosso primeiro passo foi montar um conjunto de dados ideal que utiliza informações obtidas de bases de dados para que pudéssemos verificar a qualidade de um modelo derivado deste tipo de informação.

Utilizando informações que geramos sobre expressões dos genes em testículos e ovários a partir de dados de mRNA já disponíveis, dados de ortologia das bases de dados FlyBase e *ensembl*, além de dados de *dn/ds* da mesma base de dados, geramos uma série de modelos para verificar a influência da adição destas informações curadas nos resultados finais do modelo.

Para isto, utilizamos os mesmos genes em teste e treino e os mesmos parâmetros, enquanto trocamos os dados fornecidos para o algoritmo. Ou seja, não mudamos as variáveis do algoritmo de *machine learning* e utilizamos o mesmo método de correção de desequilíbrio de classes, que no caso foi o *undersampling*. Desta maneira, podemos ver o efeito que cada tipo de informação tem sobre os resultados finais de separação de genes novos e antigos, sem a complicação adicional que a variação dos parâmetros do algoritmo traria.

Todos os modelos que utilizam dados ideais compartilham os dados de tamanho de proteína e expressão, no entanto, cada conjunto de dados adiciona algumas outras informações. O modelo 1 utiliza informação de identidade e *ds* com os ortólogos provenientes do *ensembl*; o modelo 2 utiliza a presença ou ausência de ortólogos em cada espécie retirados do FlyBase; o modelo 3 junta estas informações e retira genes muito pequenos, com menos do que 50 aminoácidos; o modelo 4 utiliza os mesmos dados do modelo 2, mas elimina os genes que possuem 4 ou mais ortólogos encontrados no nosso filtro de ortologia, desde que um deles seja uma espécie do grupo externo; e o modelo 5

utiliza os mesmos dados do modelo 3, mas apenas os genes que não foram eliminados no nosso filtro em duas etapas. A figura abaixo resume o que significa cada modelo citado.

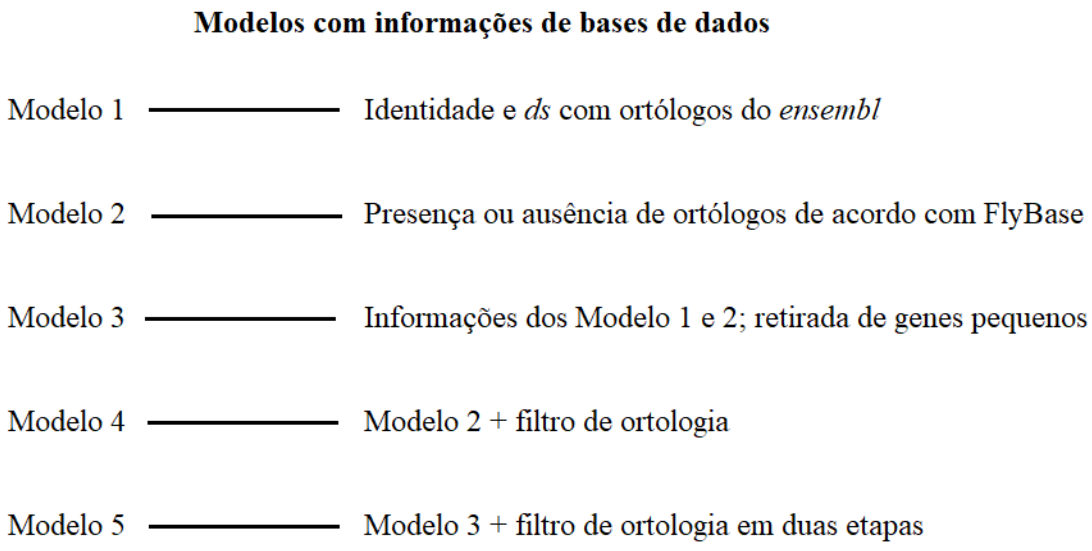


Figura 32: Resumo dos dados utilizados em cada um dos testes com informações de bases de dados. Nós escolhemos variar os dados utilizados por cada modelo enquanto mantemos as variáveis do algoritmo constante para descobrir quais dados são mais importantes para separar genes novos e antigos. Cada conjunto de informações é coletado de bases de dados de acordo com o que já se conhece de genes novos e antigos e colocado em uma tabela para a aplicação do *machine learning*.

É importante lembrar que os modelos 1 e 2 utilizam todos os genes de *D.* melanogaster, o modelo 3 inclui 573 genes novos e 8543 genes antigos, o modelo 4 utiliza 891 genes novos e 2898 genes antigos e, por fim, o modelo 5 tem apenas 414 genes novos e 1008 genes antigos. Cada conjunto de dados foi utilizado para a aplicação do *machine learning* separadamente, utilizando as mesmas variáveis do algoritmo. Os resultados destes diferentes modelos estão apresentados na tabela 13.

Tipo de modelo	*Precision*	*Recall*
Modelo 1	0,339	0,933
Modelo 2	0,334	0,897
Modelo 3	0,447	0,937
Modelo 4	0,702	0,733
Modelo 5	0,615	0,900

Tabela 13: Resultados de *machine learning* para diversos conjuntos de variáveis de bases de dados. Com estas mudanças podemos ver que os dados de ensembl e flybase dão resultados semelhantes, mas a junção destes dados e, principalmente, a inclusão de um filtro de ortologia causam maiores mudanças nos resultados.

Ao observar a tabela acima, nota-se que a adição de variáveis pode mudar drasticamente os resultados obtidos, e que é possível alcançar bons níveis de *recall* e *precision* mesmo sem uma otimização dos parâmetros ou uso de outros modelos quando estamos trabalhando com as bases de dados curadas da espécie. Após estes testes, passamos a tentar formar modelos utilizando os dados gerados por nós mesmos, de maneira a aproximar os resultados do que será encontrado em espécies sem informações disponíveis em bases de dados. Assim, temos uma ideia de como será a performance do método em outros organismos.

Os dados produzidos no projeto e utilizados na criação do modelo são equivalentes aos provenientes de bases de dados na etapa anterior, à exceção dos dados de expressão gênica de cada gene de *D. melanogaster* nas outras espécies incluídas neste projeto que não foram utilizadas nos

testes anteriores. Calculamos a expressão diferencial do gene, seu *dn/ds*, sua ortologia e expressão nas espécies do gênero *Drosophila* como explicado anteriormente para incluir estes dados no treinamento do modelo.

Assim como na etapa anterior, treinamos nosso modelo com diversos conjuntos de dados, além da aplicação dos nossos filtros baseados em ortologia. Ao variar quais informações e quantos genes estamos dando para o *machine learning*, procuramos as combinações que geram os melhores resultados. Com isso, somos capazes de observar quais informações biológicas têm maior importância, assim como o impacto causado por nossos filtros de ortologia.

O modelo 4 é o modelo de bases de dados apresentado anteriormente. Os outros modelos utilizam informações geradas *de novo* e são: 5, que utiliza dados de expressão, tamanho e a primeira etapa do nosso filtro de ortologia; 6, que adiciona informações de *dn/ds* calculados *de novo* ao modelo 5; 7, que adiciona informações de expressão de ortólogos do gene ao modelo 5; 8, que aplica o segundo passo do filtro de ortologia ao modelo 5 e o último modelo, 9, que junta os três modelos anteriores. A figura 33 resume os dados utilizados em cada um dos modelos acima.

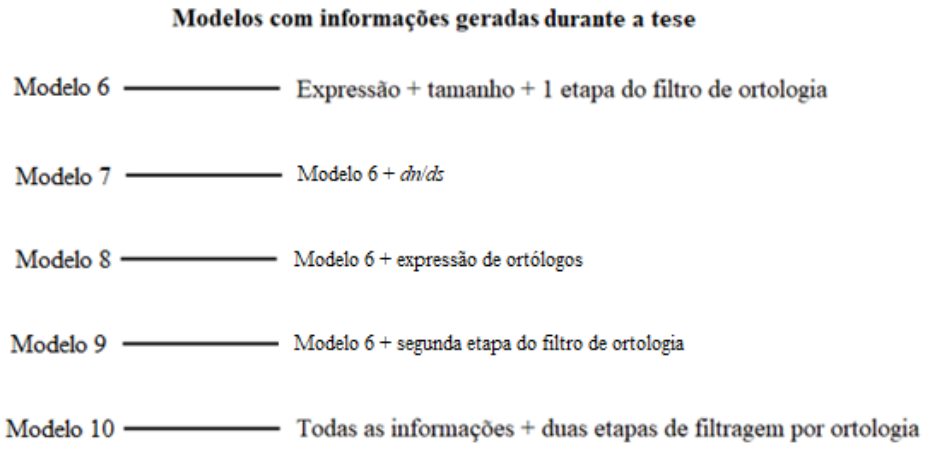


Figura 33: Resumo dos dados utilizados nos modelos com informações geradas durante este projeto. Os dados utilizados foram gerados de acordo com o que já se conhece na bibliografia de genes novos e com o que observamos nos resultados com informações de bases de dados.

Assim como foi explicado acima, cada conjunto de dados foi utilizado para a aplicação do algoritmo de *machine learning* separadamente. Da mesma maneira com que foi feito com informações de bases de dados, não mudamos as variáveis do algoritmo, de maneira a descobrir apenas quais dados são importantes para a separação de genes novos e antigos. Os resultados de *precision* e *recall* do *machine learning* de cada um dos modelos estão na tabela abaixo.

Modelo	*Precision*	*Recall*
5	0,702	0,733
6	0,361	0,667
7	0,398	0,724
8	0,389	0,712
9	0,448	0,718
10	0,508	0,718

Tabela 14: Resultados de *machine learning* com dados gerados *de novo* para a espécie *D. melanogaster*. Assim como quando utilizamos informações de bases de dados, a adição de novas informações e de uma nova etapa de filtragem melhoram nossos resultados finais.

Todos estes resultados apresentados acima foram feitos com os mesmos parâmetros e *undersampling* de maneira a serem comparáveis entre si. No entanto, é importante testar outros

métodos de amostragem, para verificar o efeito de cada um deles no resultado final do nosso modelo. Com este intuito, utilizamos os mesmos dados e genes do modelo 9, mudando apenas os métodos de amostragem entre cada resultado (tabela 15).

Modelo	*Precision*	*Recall*
ROSE	0,360	0,948
Oversampling	0,536	0,592
Undersampling (10)	0,508	0,718

Tabela 15: Resultados dos modelos de *machine learning* utilizando diferentes tipos de estratégia de amostragem. A mudança nos métodos de amostragem tem potencial de mudança radical nos resultados finais, pois modifica drasticamente quais são os dados utilizados pelo algoritmo, como explicado anteriormente. Todos os métodos usam os mesmos dados, variando apenas o método de amostragem. O resultado de *undersampling* é o mesmo do modelo 9 exibido acima.

Como explicado na seção de Materiais e Métodos, ROSE usa dados sintéticos para equilibrar as classes, *oversampling* repete membros da classe minoritária aleatoriamente e *undersampling* retira uma amostra aleatória da classe majoritária. Cada um destes métodos de amostragem gerou um resultado diferente para o mesmo conjunto de dados, reforçando sua importância para o algoritmo de *machine learning* quando se tratam de dados desbalanceados.

Observando a tabela acima, vemos que ao utilizar ROSE conseguimos identificar quase todos os genes novos, no entanto, classificamos muitos genes antigos como novos erroneamente. Ou seja, ao gerar uma lista de genes novos identificados por este método teremos a maioria dos genes novos da espécie, no entanto, teremos uma quantidade muito grande de genes antigos. O segundo método recupera menos genes novos verdadeiros, no entanto, erra menos na classificação de genes antigos como novos. Por fim, o *undersampling* apresenta um resultado mais equilibrado entre os dois, recuperando mais genes novos do que *oversampling* mas errando menos do que ROSE em relação a genes antigos.

Estes resultados mostram que a identificação de genes novos através de *machine learning* é possível, no entanto, ainda é necessário realizar mais testes para que possamos melhorar a *precision* e o *recall*, como será apresentado na seção de Discussão. Mesmo levando isso em consideração, ainda é necessária outra etapa para afirmar com mais categoria que nosso método funciona, que é testá-lo em outra espécie que não seja *D. melanogaster*. É neste ponto que entra o tema da próxima seção, a identificação dos genes novos de *D. pseudoobscura*, que servirá como este segundo ponto de teste do nosso método.

4.7. Identificando novos genes em *D. pseudoobscura*

O método de datação utilizando sintenia e parcimônia criado por Zhang e colaboradores em 2010 e explicado anteriormente data cada gene da espécie focal de acordo com o grupo de idade na qual ele pertence. Construímos uma tabela com estas informações, além da informação da presença do gene em cada uma das espécies utilizadas, de maneira a facilitar a análise de resultados. Assim, cada gene possui a informação de sua idade, em quais espécies este gene está presente e a sua localização no genoma (tabela 16).

Gene	Localização	Ortólogo em *D. persimilis*	...	Ortólogo em *D. virilis*	Idade
FBgn0078177	XL	sim	...	sim	0
FBgn0243605	desconhecido	sim	...	não	3
FBgn0243755	XL	não	...	não	4

Tabela 16: Exemplo da tabela final da etapa de datação em *D. pseudoobscura*. Cada gene anotado na espécie tem seus dados expostos, incluindo localização, presença de ortólogos identificados na datação em cada espécie, e conclusão final da idade na datação.

Assim como nos trabalhos anteriores, consideramos genes novos aqueles que pertencem aos grupos de idade maiores do que 0, resultado em 1523 genes novos e 12648 genes antigos identificados em *D. pseudoobscura*. Mais detalhadamente, foram encontrados 401 genes exclusivos de *D. pseudoobscura* (grupo 4), 382 genes compartilhados entre *D. pseudoobscura* e sua espécie irmã *D. persimilis* (grupo 3), 408 genes presentes na espécie focal e no grupo *melanogaster* (grupo 2) e 331 genes presentes em *D. willistoni* (grupo 1). A figura 34 mostra a quantidade de genes por grupo contextualizada na filogenia do gênero *Drosophila*.

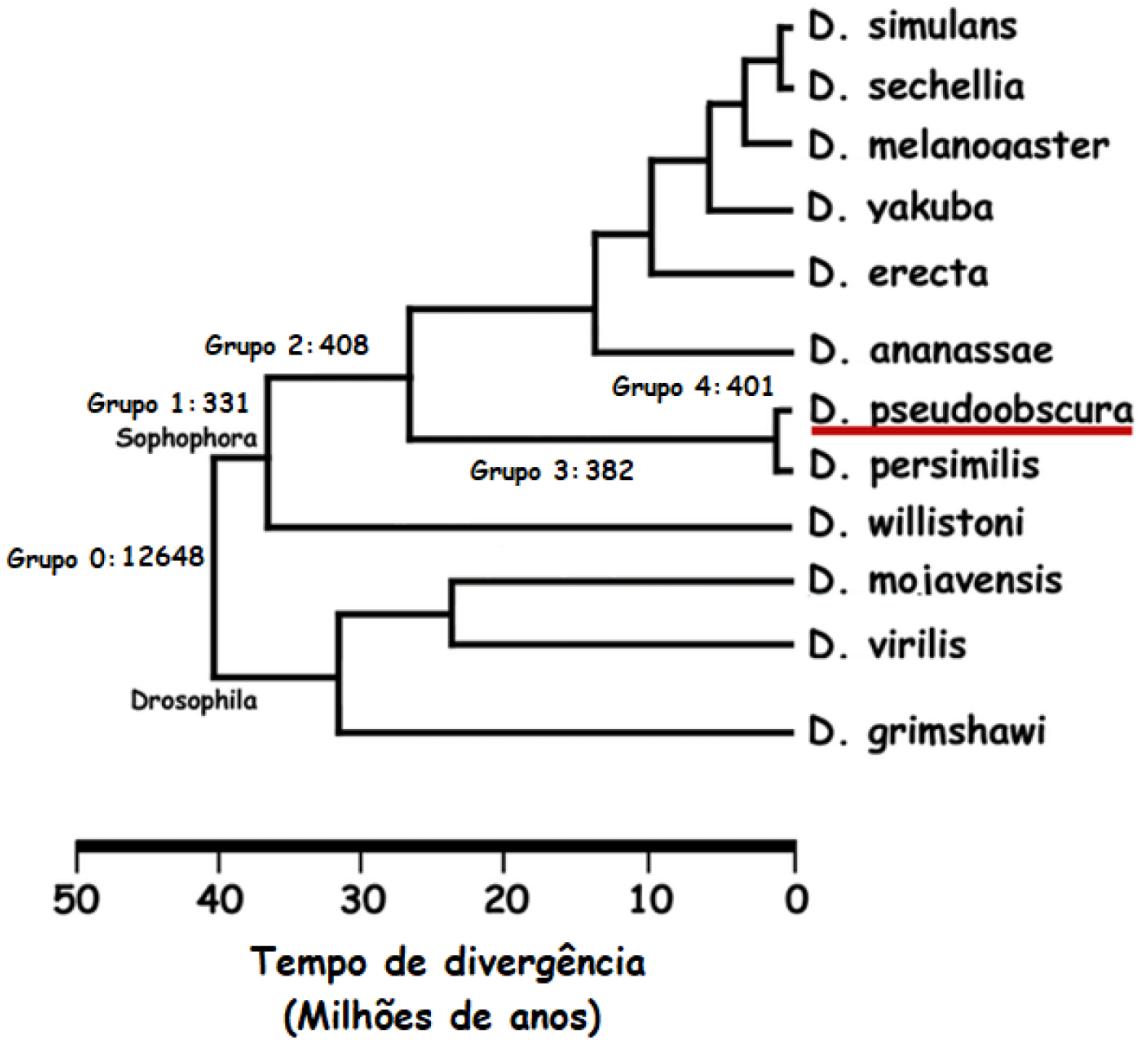


Figura 34: Filogenia do gênero *Drosophila* contendo genes novos identificados em *D. pseudoobscura* em cada ramo. Os números localizados na filogenia representam os genes identificados em cada ramo, enquanto aqueles na escala mostram o tempo em milhões de anos. Nesta imagem, vemos que 401 genes são exclusivos de *D. pseudoobscura*, 382 são compartilhados com *D. persimilis*, 408 são compartilhados com o grupo melanogaster e 331 com *D. willistoni*, enquanto existem 12648 genes antigos (adaptado de flybase.org).

Ao observar a figura acima, podemos ver que existem três espécies que são consideradas como grupos externos e, portanto, levam um gene a ser considerado como antigo. Desta maneira, queríamos ver quantos genes antigos foram assim classificados por possuírem ortólogos em apenas uma destas espécies, o que faria com que sua datação fosse menos confiável. Dos 12648 genes antigos, apenas 714 se encaixam neste padrão e tem datação como genes antigos duvidosa, mostrando que a maioria dos genes antigos foi datada com alta confiança.

Para genes novos, o que pode mudar em uma análise deste tipo é o grupo no qual o gene é colocado, e não sua classificação em si, já que a presença em uma das espécies externas o levaria a ser classificado como antigo. Apesar da separação em grupos de idade ser importante, não é nossa maior preocupação, já que esta variação não mudaria a classificação de um gene novo para um gene antigo.

Após a finalização da datação, é necessário conferir os genes novos para ver se houve algum erro cometido pelo usuário durante o processo, ou se alguma peculiaridade do genoma ou da anotação atrapalharam o método. Para isto, utilizamos duas abordagens de verificação em massa, checamos se os genes compartilhados em *D. melanogaster* foram chamados de novos por nós e por Zhang *et. al*. em 2010 e conferimos quantos dos genes chamados de exclusivos de *D. pseudoobscura* possuem ortólogos nas bases de dados. Além destas abordagens massivas, selecionamos aleatoriamente alguns genes de cada categoria para examiná-los detalhadamente. Estas etapas foram descritas anteriormente, na seção 3.13 dos Materiais e Métodos e na figura 20.

Primeiramente, separamos os genes que nossos resultados indicam que são novos e compartilhados com *D. melanogaster* para conferir se foram atribuídos como novos em ambas instâncias. Procuramos, então, os ortólogos um para um destes genes de *D. pseudoobscura* e *D. melanogaster* na base de dados *ensembl*. Para cada um dos genes com ortólogos um para um, comparamos nossa datação com a datação encontrada por Zhang em 2010.

Dos 426 genes identificados pela nossa datação como compartilhados entre estas espécies, 194 possuem ortólogos um para um identificados no *ensembl*. Destes, 46 não possuem informação de idade em *D. melanogaster*, provavelmente porque Zhang e colaboradores determinaram que sua datação não era confiável, 109 são chamados de genes novos pelas duas datações e 39 foram identificados como novos neste trabalho, mas antigos no trabalho anterior (figura 35).

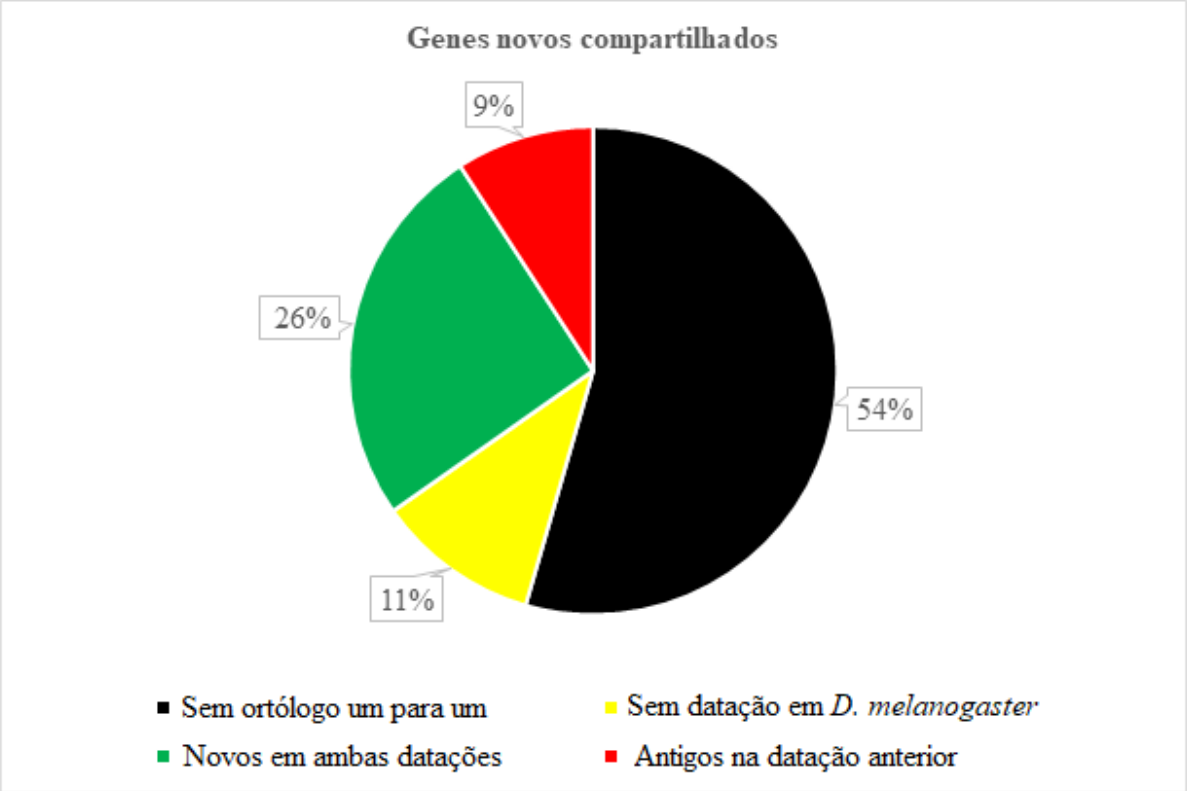


Figura 35: Resultado da comparação entres genes novos compartilhados entre *D. melanogaster* e *D. pseudoobscura* segundo nossa datação. As legendas com as cores representam os resultados obtidos na comparação e a percentagem de cada resultado está localizada no gráfico com a cor correspondente. Mais da metade dos genes analisados não tem ortólogos um para um entre as espécies, o que faz com que não possam ser comparados nesta análise. Dos genes restantes, a maioria foi identificada como genes novos em ambas datações.

Outra etapa da verificação dos resultados da datação é ver se os genes datados como exclusivos de *D. pseudoobscura* possuem ortólogos um para um em outras espécies de acordo com bases de dados curadas, como *ensembl*. Assim, podemos ver quantos destes genes que segundo nossa datação não deveriam ter ortólogos um para um em nenhuma outra espécie possuem ortólogos nesta base de dados e, portanto, provavelmente foram datados erroneamente. Dos 401 genes que foram datados como grupo 4, 23 possuem ortólogos um para um em uma espécie incluída no projeto de 12 genomas (figura 36).

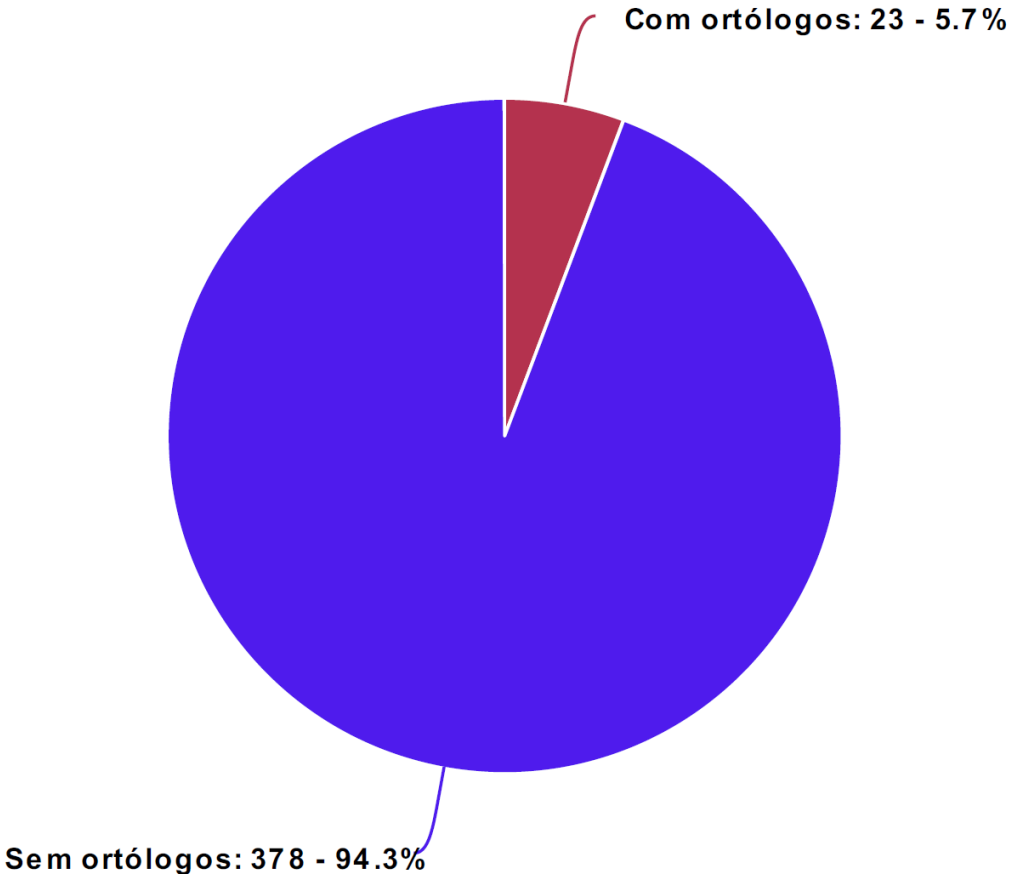


Figura 36: Proporções de genes exclusivos de *D. pseudoobscura* com e sem ortólogos encontrados em outras espécies de *Drosophila* na base de dados *ensembl*. Nesta etapa, testamos se os genes identificados como exclusivos da espécie na verdade teriam sido erroneamente classificados como tal. Para isto, buscamos ortólogos um para um na base de dados *ensembl* nas outras espécies do gênero *Drosophila*. Como a grande maioria destes genes não possui ortólogos, vemos que não há sinais de erro na datação destes genes exclusivos.

Após estas etapas de verificação em massa, separamos aleatoriamente alguns genes de cada categoria: genes que deveriam ter ortólogos em *D. melanogaster* de acordo com nossa datação, mas não nas bases de dados; genes que datamos como novos mas Zhang e colaboradores chamaram de antigos e genes que chamamos de exclusivos de *D. pseudoobscura*.

Para os genes sem ortólogos em *D. melanogaster* queríamos ver se não havia nenhum erro em relação a comparação entre nossas tabelas, de maneira que o gene teria ortólogo mas nós não conseguíssemos encontrá-lo, o que não ocorreu em nenhum gene examinado. Para o segundo grupo de genes, verificamos se não havia nenhum erro na análise da nossa tabela de datação, o que também não foi observado em nenhum caso.

Finalmente, para o terceiro grupo de genes era necessário conferir se nosso método de datação estava agindo corretamente, ou seja, se os genes chamados de exclusivos de *D. pseudoobscura* estão presentes em áreas de quebra de sintenia e são realmente novos. Para isto, escolhemos aleatoriamente dez destes genes e fomos conferir as sintenias em *D. melanogaster*, que é a espécie modelo do gênero com a melhor montagem e anotação, e *D. pseudoobscura* que é a espécie focal.

Em nenhum destes dez casos encontramos qualquer sinal de erro da nossa datação, indicando que o método agiu corretamente. Para maiores detalhes, podemos escolher o gene FBgn0244905 como exemplo. Ao utilizar a sequência de proteína deste gene obtida no *ensembl* para realizar um BLAST utilizando a base de dados *nr* do NCBI, que contém todas as proteínas curadas disponíveis, vemos que o gene de *D. melanogaster* com maior similaridade e identidade com nosso gene é o *arrow*. No entanto, este gene *arrow* possui outro gene marcado como ortólogo nas bases de dados FlyBase e *ensembl*, o gene FBgn0079220.

Com isso, precisamos examinar os dois genes de *D. pseudoobscura* para examinar se conseguimos dizer qual dos dois genes é o ortólogo verdadeiro de *arrow* e se nossa datação funcionou corretamente para o gene FBgn0244905. Para isto, utilizamos região vizinha de cada gene de maneira a verificar a sintenia entre as duas espécies em cada uma das duas localizações.

O primeiro gene que examinamos foi o FBgn0079220, que é marcado como ortólogo verdadeiro do gene *arrow*. Utilizamos a base de dados FlyBase para examinar a região genômica vizinha do gene *arrow* em *D. melanogaster* e o *ensembl* metazoa para visualizar a região do gene de *D. pseudoobscura*. Para determinar se a sintenia está conservada, vemos se os genes vizinhos do *arrow* são os mesmos que estão em volta do FBgn0079220, através das ortologias anotadas para cada gene. Neste caso, é possível observar que os mesmos genes de *D. melanogaster* estão presentes ao redor do gene de *D. pseudoobscura* (figura 37).

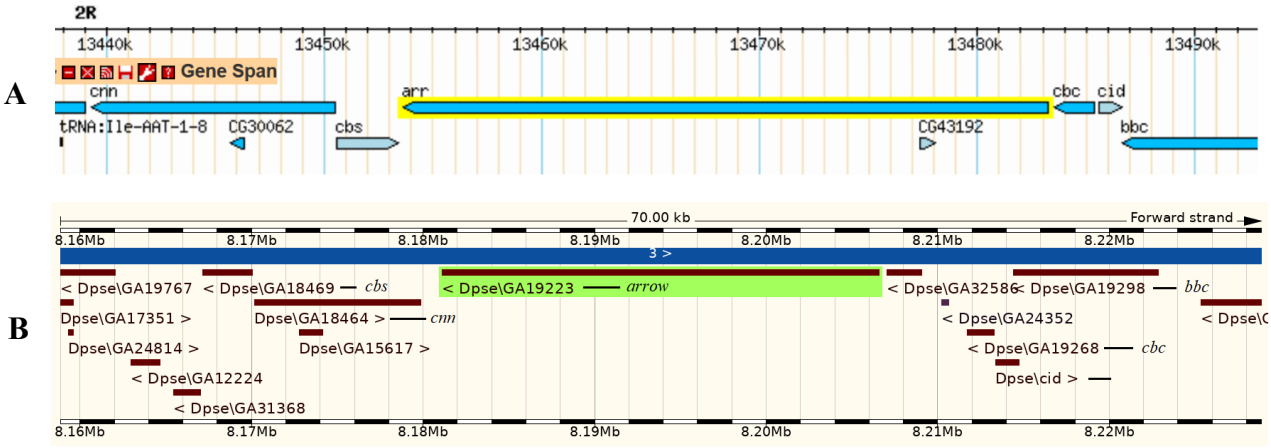


Figura 37: Comparação das regiões genômicas de *D. melanogaster* (A) e *D. pseudoobscura* (B) ao redor do gene *arrow*. A parte A foi retirada do FlyBase, enquanto B foi retirada do *ensembl*. As barras com números representam a localização cromossômica do fragmento, enquanto as barras menores com nomes representam os genes presentes na região de cada espécie. Os nomes dos ortólogos dos genes de *D. pseudoobscura* em *D. melanogaster* estão anotados ao lado do nome de cada gene, quando o gene tem ortólogos documentados. Observamos que os genes flanqueadores de *arrow* em *D. melanogaster* são os mesmos em *D. pseudoobscura*, de acordo com os dados de ortologia disponíveis em bases de dados.

Para o gene FBgn0244905, fomos ao *ensembl biomart* para ver quais eram seus genes vizinhos para que pudéssemos encontrar esta região em *D. melanogaster*, já que ele não tem ortólogo na espécie. Da mesma maneira do que foi feito com o gene anterior, usamos os ortólogos entre as espécies para ancorar nossa análise e examinar as regiões genômicas.

Este gene está localizado entre os genes ortólogos de *schlank* e *Spt-6* de *D. melanogaster*, que estão lado a lado nesta espécie, sem nenhum gene no meio (figura 38). Portanto, nossa datação agiu corretamente ao encontrar um gene que está localizado em uma região de quebra de sintenia e cuja maior similaridade é com um gene que tem um gene ortólogo verdadeiro.

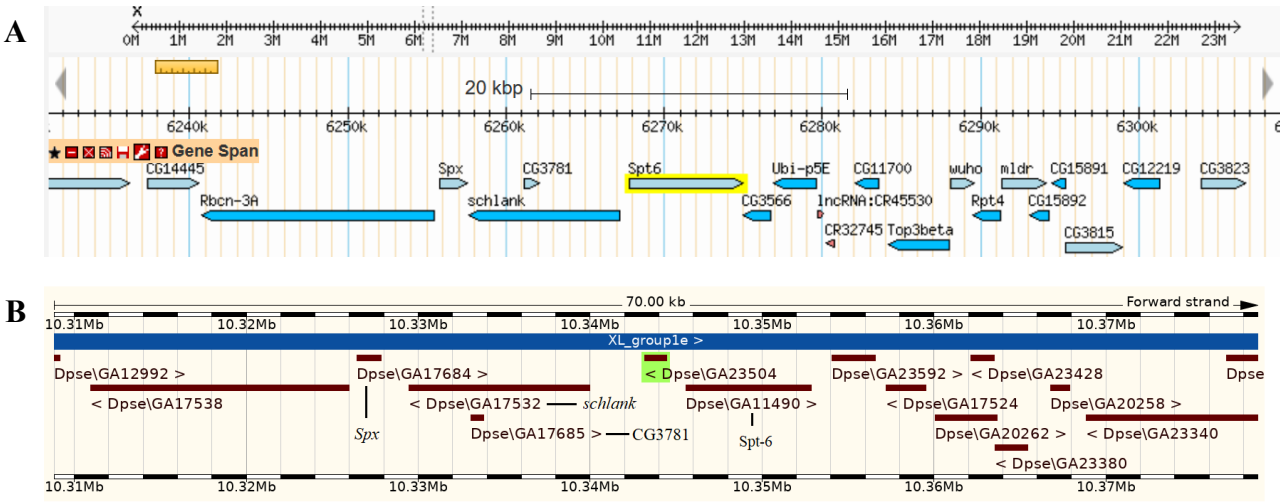


Figura 38: Comparação das regiões genômicas de *D. melaogaster* (A) e *D. pseudoobscura* (B) ao redor do gene FBgn0244905. A parte A foi retirada do FlyBase, enquanto B foi retirada do *ensembl*. As barras com números representam a localização cromossômica, enquanto as barras menores com nomes representam os genes presentes na região. Os nomes dos ortólogos dos genes de *D. pseudoobscura* em *D. melanogaster* estão anotados ao lado do nome de cada gene. É possível observar a quebra de sintenia entre as espécies justamente na região onde se localiza o gene focal, que está em volta de duas regiões conservadas entre os genomas.

Durante esta seção, exibimos os resultados obtidos durante a execução do doutorado, culminando nas etapas mais importantes: a aplicação do algoritmo de *machine learning* e a datação dos genes novos de *D. pseudoobscura*. Na próxima seção (Discussão), todos estes resultados serão analisados mais detalhadamente e colocados em seu contexto teórico e bibliográfico, auxiliando o leitor a compreender como se comparam ao que já se conhece da área assim como sua relevância no estudo de genes novos.

5. Discussão

5.1. Expressão de genes novos e antigos de *D. melanogaster*

Para que seja possível identificar genes novos a partir das suas características biológicas através de *machine learning*, precisamos de uma série de características biológicas quantificáveis que sejam significativamente diferentes entre genes novos e antigos. Essas características serão utilizadas pelo algoritmo de *machine learning* para fazer as classificações dos genes, portanto, são essenciais para o projeto. Portanto, procuramos em artigos científicos da área as características que poderíamos utilizar e que pudessem ser facilmente reproduzíveis ou obtidas.

Uma das características mais importantes que encontramos é a diferença entre os padrões de expressão, como explicado anteriormente, que influenciou diretamente o nosso desenho experimental. Antes de alimentar estas informações no algoritmo de *machine learning*, precisamos realizar análises exploratórias, de maneira a compreender melhor o comportamento das classes de genes e, assim, qual a melhor maneira de transformar estas informações em variáveis. Por exemplo, precisamos saber se as quantificações de expressão são relevantes, ou apenas o tipo de viés de expressão.

Após a busca em bases de dados por *reads* de mRNA de ovário e testículo de *D. melanogaster*, baixamos estes *reads* e fizemos sua subsequente montagem e quantificação, como explicado na seção de métodos. A partir disso, somos capazes de calcular a expressão diferencial entre estes dois tecidos, de maneira a encontrar os genes que são enviesados para testículo ou para ovário. Existem dois resultados particularmente importantes derivados destas análises que foram exibidos na seção de resultados: a diferença entre a quantidade de genes novos e antigos que tem viés para cada tecido e o fato de que o tamanho do viés também afeta a distribuição das classes.

Ou seja, se um gene tiver expressão preferencial no ovário em relação ao testículo ele tem maior probabilidade de ser antigo e, além disto, quanto maior o viés maior a chance deste gene ser antigo. Nós conseguimos observar estes fatos ao construir tabelas de contingência que comparam genes novos e genes antigos em relação a duas condições, como visto em Resultados. A partir destas observações sobre padrões de expressão, escolhemos colocar na nossa tabela de dados que será analisada pelo algoritmo de *machine learning* as seguintes variáveis relacionadas a este tema: expressão em testículo e ovário (TPM), *fold change* e o tipo de viés para cada gene.

Outro ponto importante nestas observações de padrões de expressão é que existe um grupo de genes antigos extremamente semelhante a genes novos. Isto é verdadeiro para todas as características biológicas, e realça a importância do uso de *machine learning* e do filtro de ortologia. O uso de *machine learning* nos permite utilizar todas as informações biológicas, além de ser capaz de analisar e encontrar padrões que ordinariamente passariam despercebidos. Assim, já que nenhuma característica é capaz de separar genes novos de antigos por si só, o algoritmo utiliza estes dados para realizar esta separação.

5.2. Mudanças de expressão dos genes de *D. melanogaster*

A partir do artigo publicado em 2020 por Pollock e Fukushima (Pollok e Fukushima, 2020) que demonstrou que genes com história recente de duplicação apresentam maiores mudanças no padrão de expressão, nos perguntamos como genes novos e antigos seriam afetados. Se por um lado a maioria dos genes novos surge através de duplicações, por outro eles estão presentes em menos

espécies, tendo menos chances de mudar, e normalmente demoram para adquirir expressão fora dos testículos (Zhang *et. al.*, 2010).

Precisávamos, então, examinar estes padrões para decidir se estas informações seriam úteis ao serem inseridas no algoritmo de *machine learning*, portanto, calculamos os valores de expressão e determinamos o viés de todos os genes das 8 espécies deste projeto como descrito em Métodos. Com o auxílio da ortologia derivada da busca feita com *reciprocal best hits*, montamos uma tabela contendo as informações de expressão diferencial de cada gene em cada uma das espécies para ser incluído no conjunto de dados para *machine learning*. Transformamos as informações de viés em valores numéricos, de maneira que a expressão enviesada para testículo passa a valer 3 pontos, para ovário 2 pontos, sem viés 1 ponto e a ausência de ortólogo para o gene 0 pontos, como visto na tabela abaixo.

	D. erecta	*D. ananassae*	*D. pseudoobscura*	*D. willistoni*	*D. virilis*	*D. mojavensis*	*Chymomyza*	*Scaptodrosophila*	Soma
FBgn0000008	3	1	2	2	1	3	1	2	15
FBgn0000014	1	2	0	3	2	2	3	2	15
FBgn0000028	2	0	2	2	2	0	2	2	12

Tabela 17: Dados de viés de expressão de genes de *D. melanogaster* em outras espécies transformadas em valores numéricos para uso no algoritmo de *machine learning*. Cada gene passa a ter um valor numérico associado a cada um dos tipos de viés de expressão em cada espécie, de maneira a dar ao algoritmo informação sobre todas as espécies incluídas no projeto.

Estas informações, ao serem incluídas no treinamento do algoritmo de *machine learning*, melhoraram os resultados do nosso modelo e fazem parte do nosso melhor modelo treinado até o momento. No entanto, queríamos fazer mais investigações sobre este tema, de maneira a tentar responder à pergunta postulada acima, se genes novos ou antigos possuíam mais mudanças no padrão de expressão. Com este intuito, calculamos quantas vezes um gene teve mudança no padrão de expressão que resulta em um número de 0 a 8. Porém, como genes antigos estão presentes em mais espécies do que genes novos, não podemos utilizar os valores absolutos pois introduziriam um viés para cada classe. Assim, dividimos nosso contador de mudanças pelo número de espécies com ortólogos presentes, para equalizar esta medida entre genes novos e antigos.

Para facilitar nossa análise, separamos os genes novos e antigos em três categorias: aqueles que não tem viés entre testículo e ovário em *D. melanogaster*, genes com viés para testículo e genes com viés para ovário. Desta maneira conseguimos comparar a distribuição de genes novos e antigos para cada um dos tipos de viés que podemos ter na nossa análise (figura 39).

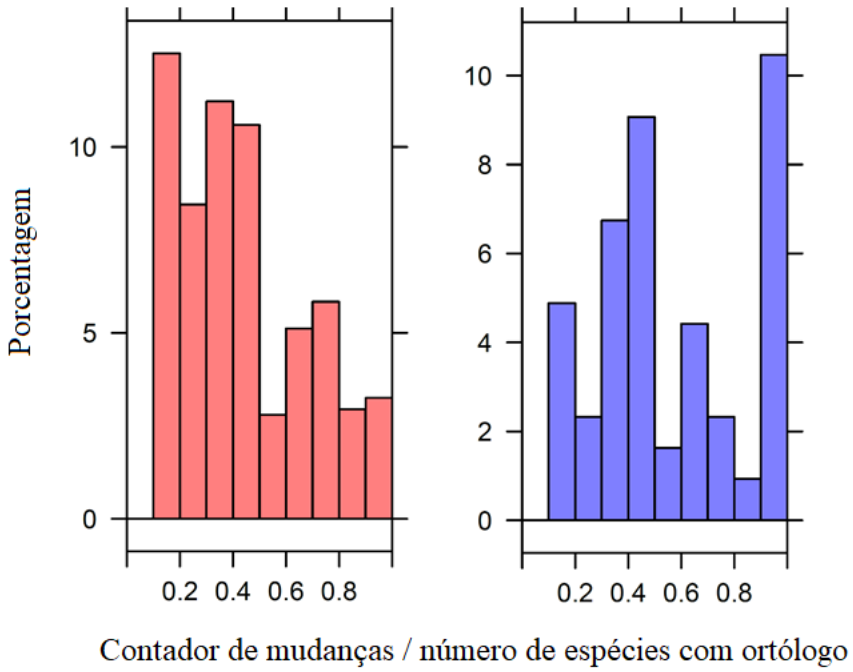


Figura 39: Distribuição do número de mudanças em viés de expressão dividido pelo número de espécies com ortólogo para genes antigos (vermelho) e novos (azul). Os genes apresentados nestes gráficos são aqueles com expressão enviesada para testículo de *D. melanogaster*. Com estas distribuições podemos verificar se há diferença significativa entre genes novos e antigos.

Para verificar se as diferenças entre as distribuições eram significativas, fizemos testes de Wilcoxon com nível de significância de 5% para cada tipo de viés com genes novos e antigos. Destas três categorias: com viés em testículo, ovário e sem viés; genes novos e antigos de *D. melanogaster* são significativamente diferentes em duas. Apenas genes que têm viés em ovário deram resultados não significativos entre genes novos e antigos em *D. melanogaster*, o que provavelmente se deve ao fato de que apenas 20 genes novos têm viés para ovário. Além disso, genes novos possuem média e mediana maior do que genes antigos nas duas classes que possuem diferenças significativas. Assim, como esperávamos, encontramos indícios que genes novos e antigos possuem padrões diferentes de mudança de expressão durante o tempo evolutivo e que genes novos parecem mudar mais do que antigos.

É importante afirmar que estes são apenas alguns indícios que apontam na direção que genes novos possuem maior chance de mudança de padrão de expressão. Para que pudéssemos fazer afirmações mais precisas, seria necessário verificar se isto se repete quando examinamos mais tecidos, pois analisamos apenas testículo e ovário, e outras espécies. Apesar de não podermos afirmar que este é um padrão geral para genes novos, isto ocorre nos nossos dados e pode ser utilizado pelo algoritmo de *machine learning* a separar genes novos de antigos.

5.3. Evolução de genes novos e antigos

Assim como feito para dados de expressão, investigamos os padrões de *dn/ds* de genes novos e antigos, motivados pelas diferenças previamente conhecidas entre estas duas classes. De maneira geral, genes novos possuem maior sinal de seleção positiva, fato que pode ser utilizado para auxiliar o algoritmo de *machine learning* na separação dos genes novos e antigos.

Primeiramente, buscamos informações de *D. melanogaster* já disponíveis, para que pudéssemos ver se seriam úteis no treinamento do nosso modelo. Para isto, utilizamos a base de dados FlyDivas

que já tem disponível dados de *dn/ds* que podemos inserir na nossa tabela. Como visto na seção de Resultados, usamos estes dados na nossa investigação sobre modelos de *machine learning* e vimos que a adição destes foi informativa e melhorou o modelo. Estes resultados nos motivaram a calcular as informações de *dn/ds* utilizando nossos dados de transcriptomas, de maneira a aproximar os resultados daqueles que serão encontrados em espécies não modelo. A partir disto, são necessários dois passos principais: examinar estas informações a fundo, incluindo a comparando com FlyDivas e inserir os dados no treinamento de *machine learning*.

Para o primeiro passo, fizemos três análises principais: vimos se havia diferença significativa entre a distribuição de valores de *dn/ds* de genes novos e antigos, verificamos esta mesma diferença apenas para genes que sobreviveram nosso filtro de ortologia e comparamos nossos resultados com aqueles provenientes do FlyDivas. De maneira geral, podemos observar o padrão esperado de genes novos com valores de *dn/ds* que são maiores do que os genes antigos, em média. No entanto, após a aplicação filtro de ortologia, esta diferença deixa de ser significativa como visto em Resultados. Apesar disto, a inclusão destas informações ainda afeta positivamente os resultados do treinamento do *machine learning*. Portanto, esta informação ainda é importante para nossa análise por mais que as distribuições não sejam significativamente diferentes, já que o algoritmo está fazendo uso desta informação de alguma maneira para separar genes novos de antigos. Isto realça o ponto que vínhamos fazendo: o algoritmo é capaz de analisar e utilizar um volume de informações que não é facilmente visualizado.

Por fim, ao comparar os dados que calculamos com aqueles obtidos na base de dados vemos que tem uma correlação significativa (teste de Pearson, $p < 0.05$) mas com um coeficiente de apenas 0.242 e que têm distribuições significativamente diferentes (teste de Wilcoxon, $p < 0.05$). Nós esperávamos que houvesse diferença entre nossos dados e aqueles de bases de dados, já que os calculamos a partir de transcriptomas e com ortólogos encontrados por *reciprocal best hits*, enquanto FlyDivas tinha genomas e dados de ortologia mais completos.

Observando as médias e medianas das duas distribuições, observa-se que estamos, de maneira geral, subestimando os valores de *dn/ds* quando calculamos do zero (tabela 17). Estas comparações nos ajudam a compreender nossos dados, e a observar as consequências de calcular *dn/ds* utilizando estas informações imperfeitas. Apesar destas diferenças, as informações provenientes de *dn/ds* ainda são importantes para a separação de genes novos e antigos pelo *machine learning*.

	Média	Mediana
FlyDivas	0,198127	0,114924
Nossos dados	0,1086	0,0410

Tabela 18: Sumário das estatísticas descritivas das distribuições de *dn/ds* para os dois conjuntos de dados. Observa-se que os dados do FlyDivas têm média e mediana maior do que os outros, o que provavelmente significa que estamos subestimando estes valores ao calculá-los nós mesmos.

5.4. Obtendo e utilizando informações de ortologia

Na nossa busca por características que nos ajudem a separar genes novos de genes antigos, utilizamos a diferença na quantidade de ortólogos que cada classe de gene possui dentre as oito espécies examinadas. Como visto na seção de Resultados, o uso deste tipo de informação com filtro para eliminar genes antigos melhora substancialmente o modelo de *machine learning*. Parte dessa diferença é explicada simplesmente pelo conceito de genes novos, que são aqueles presentes em um táxon, mas ausente no seu táxon irmão, o que leva a sua presença em menos espécies do que os

genes antigos. Outra parte desta diferença pode ser explicada pela evolução mais rápida de genes novos em relação a genes antigos, o que dificulta o encontro de ortólogos destes genes pelos métodos de busca por similaridade, como aqueles que utilizamos neste projeto.

Independentemente do motivo, a diferença da distribuição do número de ortólogos de genes novos e antigos é significativa e pode ser utilizada para eliminar uma grande quantidade de genes antigos. Esta informação tem, portanto, um grande valor para nossa proposta já que pode ser utilizada como filtro, na eliminação de genes antigos, e como variável para o treinamento do algoritmo.

Para a primeira etapa da nossa filtragem, contamos apenas a presença ou ausência de ortólogos do gene de *D. melanogaster* nas outras espécies para construir uma tabela. A partir disto, escolhemos eliminar os genes que tinham ortólogos em quatro ou mais espécies, desde que uma destas seja uma espécie distante da espécie focal. A exigência da presença em uma espécie distante está diretamente relacionada com o fato de que poucos genes novos possuem ortólogos nestas espécies, enquanto o mesmo não é verdade para genes antigos, o que está relacionado com as características de cada classe. A escolha de quatro ou mais espécies está ligada a dois fatores: a quantidade de espécies nas quais buscamos ortólogos (oito) e o exame dos resultados de genes eliminados (tabela 10 em Resultados).

Para a segunda etapa do filtro, nosso desafio é inserir a informação filogenética de maneira mais completa na comparação de genes novos e antigos. Ainda tínhamos o mesmo objetivo de eliminar o maior número possível de genes antigos sem perder muitos genes novos, mas precisávamos incluir mais da informação filogenética que não estava sendo incluída no primeiro filtro. Ao pensar em um contexto evolutivo, genes novos têm maior probabilidade de ter ortólogos em espécies próximas a espécies focais do que em espécies distantes. Com isto em mente, criamos uma segunda tabela, desta vez dividida em grupos filogenéticos, de maneira que a presença de ortólogo em uma espécie mais próxima confere menos pontos do que em uma espécie mais distante. Como explicado anteriormente, a escolha da pontuação não foi feita ao acaso, e sim de maneira que a soma das pontuações informações de quais grupos o gene possui ortólogos.

A escolha, nesta etapa, de eliminar genes com 20 ou mais pontos foi feita pois a única maneira de um gene atingir esta pontuação é estar presente em um dos grupos mais distantes. Isto ocorre, pois, a presença de um ortólogo em *Chymomyza* ou *Scaptodrosophila* confere 16 pontos ao gene e cada gene que sobreviveu a primeira filtragem só tem ortólgo em no máximo 3 espécies. Além disso, como discutido anteriormente, nosso objetivo com estes filtros sempre é eliminar o maior número possível de genes antigos sem perder muitos genes novos.

Examinando os resultados obtidos, podemos ver que alguns genes novos possuem ortólogos em espécies muito distantes e alguns genes antigos não tem ortólogos mesmo em espécies próximas a *D. melanogaster*. Para compreender o motivo pelo qual isto ocorre, precisamos lembrar que estes filtros foram montados com *reciprocal best hits*, um método de procura de ortólogos baseado em similaridade. Este método pode incorrer em alguns tipos de erros, mas vale lembrar que qualquer método de busca de ortologia tem seus problemas em potencial e este foi o que apresentou nossos melhores resultados.

O tipo de erro que envolve os genes antigos sem ortólogos em diversas espécies é mais fácil de compreender. Estes genes podem ter sido perdidos nas espécies sem ortólogos, mas isto não explicaria todos os casos observados de ausência de ortologia, já que seriam necessários múltiplos eventos para explicar o padrão de alguns genes. Um dos motivos pelos quais este erro pode ocorrer é que o ortólogo do gene não é expresso nem no testículo nem no ovário, que são os tecidos utilizados na busca, fazendo com que o gene não seja detectado. O outro motivo pode ser alguma

falha na busca por similaridade, seja por que o gene antigo evolui muito rapidamente e causam esta falha ou por que fazem parte de uma família gênica com membros semelhantes, causando a ausência de reciprocidade na busca.

No caso dos genes novos com ortólogos em espécies mais distantes, as explicações provavelmente também estão ligadas a problemas com a busca por similaridade. Existem duas maneiras de um gene novo ter um ortólogo encontrado onde não deveria: a busca por similaridade confundiu o gene novo com seu gene parental ou o gene novo faz parte de uma família gênica e fez com que o programa de busca acuse sua presença erroneamente. Em ambos os casos, a semelhança entre o gene novo e seu gene parental pode confundir o *reciprocal best hits.*

Assim, podemos ver que mesmo com os erros encontrados devido ao nosso método de busca de ortólogos, aquilo que esperávamos é alcançado e conseguimos eliminar uma grande quantidade de genes antigos e perder poucos genes novos utilizando o filtro de ortologia. Isto é importante pois melhora significativamente nossos resultados com *machine learning*, como exibido anteriormente na seção de Resultados. Por fim, ainda serão feitos alguns testes com estes filtros, realizando leves mudanças e vendo como cada mudança afeta o modelo treinado.

5.5. *Machine learning* com informações de bases de dados

Para que fosse possível implementar um modelo de *machine learning* capaz de identificar genes novos, precisávamos de duas informações principais: a lista de genes novos de *D. melanogaster* e dados capazes de diferenciar genes novos de antigos. Ao utilizar *D. melanogaster* para testar e treinar o *machine learning* temos não só a lista de genes novos, mas também uma grande quantidade de informações biológicas disponíveis em bases de dados curadas.

Apesar de já conhecermos as principais características biológicas que diferenciam genes novos de antigos, como explicado anteriormente, é necessário verificar como a adição de cada tipo de informação afeta o modelo de *machine learning*. Com este objetivo, utilizamos informações de bases de dados para gerar uma série de modelos com diferentes níveis de complexidade, cujas taxas de erro foram exibidas na seção de Resultados. Estas etapas de investigação são importantes pois nem todas as informações adicionadas contribuem da mesma maneira, já que podem ser mais ou menos diferentes entre genes novos e antigos. Além disto, nos ajuda a descobrir se existem variáveis que se utilizada junto com outras não adicionam informação ao modelo. Assim, esta abordagem nos ajuda a compreender quais variáveis são úteis para a separação de genes novos e antigos e serão aplicadas no modelo real.

Como visto em resultados, a adição destas informações biológicas aumenta a qualidade de predição do modelo, mas o efeito depende de qual variável está sendo adicionada, como previsto. Na seção de resultados não adicionamos os testes que não tiveram sucesso em aumentar a qualidade do modelo, como a informação da segunda etapa de nossa filtragem dentro do algoritmo. Apesar dos resultados dos modelos melhorarem, de maneira geral, com a adição de novas informações e dos filtros de ortologia, não conseguimos alcançar um modelo que separa genes novos e antigos com altas taxas de *precision* e *recall* simultaneamente. Isto ocorre, principalmente, pois existem genes antigos que exibem características extremamente semelhantes a genes novos, como expressão preferencial no testículo, como visto na seção de Resultados. Analisando estes resultados podemos ver a importância da utilização do nosso filtro de ortologia, que gera grandes melhoras nos nossos modelos.

Com estes resultados, podemos ver as consequências dos usos de diversos tipos de informações biológicas no algoritmo de *machine learning*. Ao invés de continuar buscando a otimização dos modelos gerados, decidimos utilizar as informações obtidas sobre o efeito das variáveis na qualidade do modelo e criar um modelo com informações mais próximas do que será encontrado em outras espécies.

5.6. *Machine learning* com dados gerados localmente

Após a realização dos testes com as informações provenientes de bases de dados de *D. melanogaster*, realizamos uma investigação semelhante com informações geradas por nós mesmos, ou seja, dados gerados *de novo* para a mesma espécie. Esta investigação é importante para a construção do método de identificação de genes novos que estamos propondo por dois motivos principais.

Primeiramente, esta abordagem aproxima os dados inseridos no modelo àqueles que são encontrados nas outras espécies que não tem informações disponíveis em bases de dados curadas. Assim, como teremos que gerar as variáveis que serão utilizadas pelos algoritmos de *machine learning* em todas as outras espécies, precisamos fazer testes com a mesma abordagem em *D. melanogaster* para garantir que os resultados do modelo treinado serão semelhantes em outras espécies. Da mesma maneira, para que seja possível treinar o modelo em *D. melanogaster*, espécie na qual já temos a lista de genes novos, e aplicá-lo em outras espécies é necessário que as informações utilizadas para treino e aplicação do modelo tenham sido geradas da mesma maneira. Por causa destes fatores, precisamos investigar o comportamento do modelo de *machine learning* com as variáveis geradas por nós mesmos.

Assim como foi feito na etapa anterior, treinamos uma série de modelos utilizando o algoritmo *random forest* mudando as variáveis incluídas, mas mantendo os parâmetros de maneira a verificar como cada variável afeta o resultado final. Como visto em Resultados o uso de mais informações e dos filtros de ortologia afeta positivamente nossos resultados.

No entanto, alguns dos testes que fizemos não melhoraram os modelos, como a inclusão da informação do segundo filtro de ortologia como variável dentro do algoritmo de *machine learning*, ao invés de ser utilizado apenas como método de filtragem. Em casos como este, estas variáveis não são suficientemente diferentes entre genes novos e antigos ou não adicionam nada que o algoritmo não tinha anteriormente, ou seja, são apenas informações redundantes. Estes testes e seus resultados foram omitidos pois não contribuem para o *machine learning* e não foram incluídos no nosso modelo final.

De maneira geral, os padrões dos modelos de *machine learning* observados para os testes com informações de bases de dados e *de novo* se repetem nos dois tipos de teste, tendo resultados melhores com mais informações e com o uso dos filtros de ortologia. No entanto, as informações de bases de dados são mais informativas para o modelo, gerando resultados consistentemente melhores.

Esta diferença em resultados era esperada, já que as informações de bases de dados são mais completas e mais cuidadosamente curadas. Por exemplo, as informações de ortologia do *ensembl* são criadas a partir de métodos com múltiplos passos utilizando genomas de alta qualidade como referência. Assim, as variáveis correspondentes a estes dados separam melhor genes novos de genes antigos, gerando melhores modelos finais.

Como mencionado anteriormente, os filtros de ortologia tem grande efeito nos nossos resultados, melhorando substancialmente os modelos de *machine learning*. Os testes aqui apresentados foram feitos com quatro espécies no primeiro filtro e 20 pontos no segundo, seguindo a lógica apresentada anteriormente. No entanto, vamos realizar testes com as mesmas variáveis e parâmetros com diferentes pontos de cortes nas duas etapas de filtragem por ortologia.

Com a realização destes testes no futuro, poderemos mostrar como cada tipo de escolha no filtro afeta os resultados, assim como a relação entre a eliminação de genes antigos e as taxas de erro do modelo. Assim, esperamos apresentar para o usuário a opção de eliminar mais genes, e perder mais genes novos, mas ter menos erros no *machine learning* e obter uma lista mais confiável de genes novos de acordo com os parâmetros do filtro de ortologia.

Os nossos dados têm um desequilíbrio de classes, ou seja, nós temos sempre mais genes antigos do que genes novos e por isso precisamos aplicar abordagens para corrigir esta questão antes do treinamento do modelo. Nas seções anteriores, mostramos como cada abordagem funciona, além de suas potenciais consequências e seus resultados quando aplicadas a nossos dados. Por causa da proposta deste projeto, precisamos testar todas as abordagens de maneira a garantir que nosso método funcione da melhor maneira possível e tenha os melhores resultados. Assim, precisamos escolher o que estamos variando em cada tipo de teste, para que os resultados sejam comparáveis. Quando mudamos quais variáveis estão incluídas no modelo, não mudamos os parâmetros ou abordagem de desequilíbrio de classes, por exemplo. Desta maneira, é possível ver como cada tipo de variação afeta nossos resultados.

Portanto, será necessário sempre testar as diferentes abordagens relacionadas ao desequilíbrio de classes, já que podem afetar nossos resultados. A princípio, tanto o *undersampling* quanto *oversampling* podem servir bem nossos propósitos, já que os genes antigos e genes novos possuem padrões claros e que se repetem na maior parte dos genes.

Sempre quando testamos estas seleções aleatórias, é necessário verificar se os genes selecionados ao acaso não afetam desproporcionalmente os resultados. Como estas seleções são aleatórias, existe a possibilidade de afetarem desproporcionalmente o treinamento do modelo. No entanto, não encontramos nenhum caso assim neste trabalho, com nossos testes mostrando alta repetibilidade das taxas de erro do *machine learning*.

Ao analisar nossos resultados, é importante lembrar que nem todas as estatísticas são igualmente importantes para nossa proposta. Idealmente, um modelo treinado de *machine learning* teria *precision* e *recall* altos, acima de 0.9, além de alta área sob curva, ou seja, seria um modelo que identifica quase todos os genes novos e praticamente não classifica genes antigos erroneamente. No entanto, é possível que não alcancemos esses números mesmo após testar a mudança de parâmetros e algoritmos de *machine learning* que ainda precisamos fazer. Isto pode ocorrer pois existe um grupo de genes antigos que possui características extremamente semelhantes aos genes novos, sendo expressos em testículo e evoluindo rapidamente. Assim, se exigirmos que o modelo seja mais restritivo em uma tentativa de não classificar estes genes antigos como novos, provavelmente perderemos genes novos verdadeiros.

Este tipo de consequência é conhecido na área de *machine learning* como *tradeoff* entre *precision* e *recall*, que quer dizer que muitas abordagens utilizadas para melhorar uma das taxas normalmente pioram a outra (Libbrecht *et. al.*, 2015). Como apresentado anteriormente, *precision* é a proporção dos elementos selecionados que são positivos verdadeiros (Positivos verdadeiros/positivos verdadeiros + falsos positivos) e *recall* é a proporção dos elementos que deveriam ser selecionados que foram encontrados pelo modelo (Positivos verdadeiros/positivos verdadeiros + falsos negativos). Assim, talvez seja necessário escolher entre um modelo que tem

mais falsos negativos ou falsos positivos, dependendo do que é melhor para cada abordagem. Se realmente não for possível alcançar baixas taxas de todo tipo de erro, será necessário escolher qual tipo de modelo preferimos para nossa proposta. Neste caso, seria melhor selecionar um modelo que tem menos falsos positivos, para que sejamos capazes de gerar uma lista de genes novos de alta confiança, mesmo que estejamos perdendo genes novos verdadeiros.

Este tipo de análise é importante quando estivermos investigando o efeito de parâmetros e algoritmos de *machine learning* nos nossos dados, o que ainda não foi testado. Nossa prioridade era testar primeiro o efeito das variáveis correspondentes a informações biológicas e, para isso, testamos sempre com os mesmos parâmetros como discutido acima. No entanto, será necessário fazer esta outra rodada de testes para focar nestes fatores e examinar os resultados subsequentes.

Apesar de ainda serem necessárias outras rodadas de testes para resolver as questões apresentadas durante este trabalho, é possível detalhar as etapas necessárias para a identificação de genes novos a partir de *machine learning*. De maneira resumida, é feito o sequenciamento do transcriptoma de ovário e testículo de espécies relacionadas para a obtenção das informações biológicas relevantes, como descrito durante este trabalho. Utiliza-se, então, uma espécie modelo como *D. melanogaster* que tenha os genes novos identificados para treinar o algoritmo que será aplicado nas outras espécies, gerando a lista de genes novos. Um esquema mais detalhado destas etapas para *Drosophila* está apresentado na figura 40.

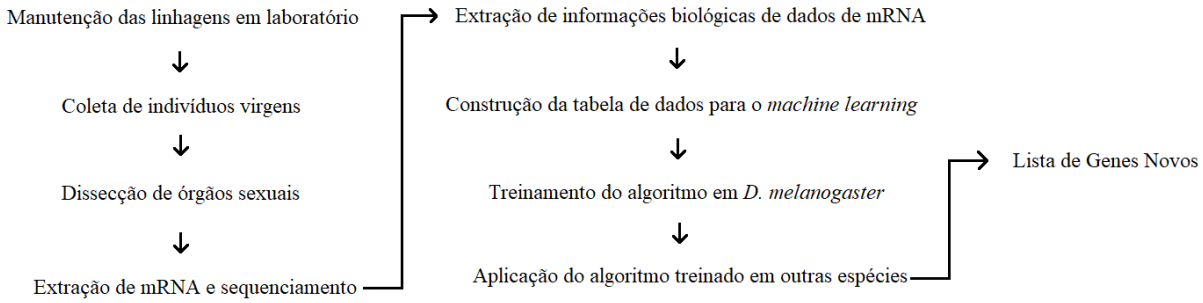


Figura 40: Etapas para a identificação de genes novos através de informações biológicas com uso de *machine learning* em *Drosophila*. Estes passos foram explicados durante este trabalho, assim como o significado dos seus resultados. Caso as espécies de interesse já tenham as informações biológicas ou transcriptomas disponíveis em bases de dados algumas destas etapas não são necessárias.

5.7. Identificando genes novos de *D. pseudoobscura*

Como estamos propondo um novo método de identificação de genes novos baseado em *machine learning*, é necessário que esta predição funcione em outras espécies além da espécie na qual o *machine learning* foi treinado. Para isto, precisamos de uma espécie que tenha seus genes datados e não seja *D. melanogaster*, a qual usamos para treinar o modelo. Como esta é a única espécie do gênero *Drosophila* com os genes novos identificados, precisamos realizar o processo de datação de genes em alguma outra espécie do gênero.

Para isto, escolhemos a espécie *D. pseudoobscura* para ter seus genes datados pelo método já existente e servir como um ponto de controle. Esta espécie servirá para que possamos aplicar o modelo treinado em *D. melanogaster* e treinar um modelo para ser aplicado em *D. melanogaster*.

Assim, será possível verificar a aplicabilidade do método no gênero *Drosophila* de maneira geral, sem medo de um viés forte para *D. melanogaster*.

D. pseudoobscura foi selecionada pois seu genoma passou por algumas rodadas de montagem e anotação, ambos fatores essenciais para a datação de genes utilizando sintenia, como explicado anteriormente. Após a datação dos genes de *D. pseudoobscura*, identificamos 1523 genes novos e 12648 genes antigos de acordo com os parâmetros estabelecidos por Zhang e colaboradores. Para que pudéssemos ter um contexto do que era esperado, olhamos para *D. melanogaster* que tem 1070 genes novos e 12013 genes antigos (Zhang *et. al.*, 2010). Neste contexto, é importante frisar que *D. pseudoobscura* tem mais genes novos do que *D. melanogaster*, no entanto, também tem um número maior de genes identificados no genoma: 1088 genes a mais.

Outra comparação importante que devemos fazer é a quantidade de genes de *D. pseudoobscura* que foram encontrados apenas na espécie e somente na espécie e na sua espécie irmã *D. persimilis* e as informações análogas em *D. melanogaster*. Zhang e colaboradores encontraram apenas 60 genes exclusivos de *D. melanogaster* e 68 compartilhados apenas com suas espécies irmãs (Zhang *et. al.*, 2010), enquanto nós encontramos 401 genes exclusivos e 382 compartilhados com a espécie irmã em *D. pseudoobscura*. A figura abaixo mostra uma comparação entre os resultados das duas espécies.

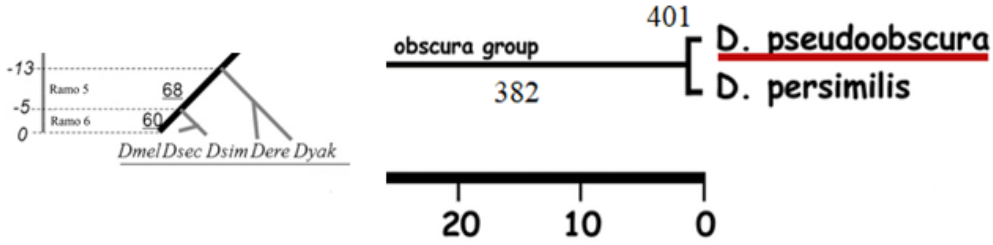


Figura 41: Comparação de resultados de genes exclusivos encontrados para *D. melanogaster* em 2010 e *D. pseudoobscura*. Os números exibidos à esquerda da primeira imagem e abaixo da segunda representam o tempo em milhões de anos, enquanto aqueles junto as filogenias são o número de genes que foram identificados em cada ramo. Assim, é possível ver a grande diferença no número de genes exclusivos identificados nas duas espécies (adaptado de Zhang *et al.*, 2010 e flybase.org).

Existem algumas explicações possíveis para as diferenças encontradas em relação a estes genes exclusivos. A primeira seria que nós cometemos algum erro na aplicação do método, fazendo com que um grande número de genes seja datado erroneamente. No entanto, como discutido na seção sobre os testes realizados com estes genes de *D. pseudoobscura*, não encontramos sinais de erros em massa da nossa datação. A segunda explicação é que houve erros na identificação de genes de *D. pseudoobscura*, na anotação ou montagem. Assim, estes fragmentos chamados equivocadamente de genes sempre serão exclusivos ou presentes apenas na espécie irmã já que, não sendo genes, seriam eliminados rapidamente do genoma durante o processo evolutivo. Isto não é impossível, mas é bastante improvável já que a espécie passou por várias rodadas de montagem e anotação, minimizando este tipo de erro. Alguns testes ainda serão realizados apenas para verificar se não é este tipo de erro que estamos observando, como a comparação com bases de dados de rRNA e a verificação da expressão. A terceira possibilidade seria que a adição de mais espécies do grupo filogenético *obscura* resultaria em menos genes exclusivos de *D. pseudoobscura* e desta espécie com sua espécie irmã, *D. persimilis*. Isto pode acontecer pois a espécie irmã é a única incluída nesta análise que pertence a este grupo, e seu genoma tem problemas de anotação e montagem. Para verificar se é isso que está acontecendo, iremos analisar os genomas de outras espécies próximas, como *D. miranda*, e ver se estes genes classificados como muito novos estão presentes ou não nestas espécies.

Por último, estes resultados são verdadeiros e *D. pseudoobscura* realmente possui não só mais genes novos do que *D. melanogaster*, mas também mais genes exclusivos e genes compartilhados apenas com a espécie irmã. Portanto, as diferenças observadas seriam relacionadas apenas às histórias evolutivas de cada táxon que culminaram na configuração atual do genoma de cada espécie.

Considerando as diferenças encontradas entre *D. pseudoobscura* e *D. melanogaster*, o objetivo das nossas etapas de controle de qualidade era verificar se o método havia sido aplicado corretamente, já que se tratam de muitas etapas complexas que devem ser executadas em uma ordem precisa. Com isto em mente, se isto tivesse acontecido era esperado que encontrássemos um grande número de genes novos e antigos datados erroneamente, já que qualquer erro na aplicação do método geraria datações erradas em todos os grupos de idade.

No entanto, não foi isto que observamos ao realizar nossas etapas de verificação em massa e ao olhar casos individuais. Assim como exibido na seção de resultados, a maioria dos genes exclusivos de *D. pseudoobscura* não tem ortólogos em outras espécies de acordo com *ensembl*, uma grande quantidade dos genes compartilhados em *D. melanogaster* possuem a mesma datação e nossas verificações individuais mostraram apenas padrões esperados para genes novos.

Para a análise dos resultados das datações, é importante lembrar que a única espécie do grupo *obscura* analisada é a espécie irmã, ou seja, para um ramo de aproximadamente 30 milhões de anos, apenas a espécie focal e sua espécie irmã estão representadas. No trabalho de Zhang e colaboradores em 2010 o ramo do grupo *melanogaster* que tem aproximadamente o mesmo tempo evolutivo teve seis espécies analisadas.

Esta diferença se deve à disponibilidade de espécies com genomas montados e anotados no grupo *Drosophila*, que é mais focada em espécies relacionadas à espécie modelo *D. melanogaster*. Nós poderíamos ter adicionado outras espécies, mas introduziríamos dois novos problemas: a maioria das anotações das outras espécies sequenciadas foram feitas com anotação automática e não foram curadas manualmente e perderíamos a comparação direta com o artigo de 2010. Desta maneira, todos os genes que surgiram no ramo do grupo *obscura* são classificados como exclusivos ou como compartilhados apenas com *D. persimilis*, já que não tem outras espécies analisadas neste grupo. Se tivéssemos mais espécies neste ramo, poderíamos separar estes genes em mais grupos, no entanto, eles ainda estariam classificados como genes novos o que é a datação mais importante, principalmente para nosso objetivo de utilizar esta espécie como segundo ponto de controle para nosso método com *machine learning*.

Com estas informações, é possível afirmar que a probabilidade de a datação destes genes ser proveniente de um erro na aplicação do método é baixa e que provavelmente se trata de uma informação biológica verdadeira. Como mencionado, iremos fazer mais alguns testes para verificar se a anotação destes genes teve algum problema e após este processo iremos utilizar a espécie *D. pseudoobscura* para continuar testando nossos modelos de *machine learning*.

6. Conclusões

Neste trabalho propomos um método de identificação de genes novos através de características biológicas analisadas por um algoritmo de *machine learning*. Com este intuito, observamos as diferenças entre genes novos e antigos em relação à expressão, evolução e ortologia. Estes resultados eram esperados de acordo com a literatura, mas precisávamos confirmar seu aparecimento nos nossos dados.

Para treinar o algoritmo utilizamos *D. melanogaster*, que teve seus genes novos identificados e é a espécie modelo do gênero *Drosophila,* contando com uma grande quantidade de informações disponíveis em bases de dados curadas. Com estas informações em mãos, geramos uma série de modelos de *machine learning* com dados curados e gerados por nós mesmos de maneira a separar genes novos e genes antigos.

Como esperado, as informações biológicas significativamente diferentes entre genes novos e antigos contribuem para o funcionamento do algoritmo, com importância especial para os dados de ortologia que são utilizados, também, como filtros. Utilizando a diferença de número de ortólogos e de espécies com o ortólogos entre genes novos e antigos, montamos um filtro em duas etapas para eliminar um grande número de genes antigos enquanto perdemos o menor número possível de genes novos.

Os resultados com informações de bases de dados tiveram a melhor performance, com 0,702 de *precision* e 0,733 de *recall* no melhor modelo, enquanto o modelo com informações geradas nesta tese chegou a 0,508 de *precision* e 0,718 de *recall*. Isto era esperado pois as bases de dados são mais completas e são curadas, fazendo com que sejam geradas variáveis mais informativas. Ainda existem outros testes que podem ser feitos para buscar a melhora destes resultados, como a aplicação de outros tipos de algoritmo de *machine learning* além de *random forest*.

Para que tivéssemos outro ponto de controle do nosso método, datamos os genes de *D. pseudoobscura* utilizando o método publicado por Zhang e colaboradores, identificando 1523 genes novos. Aplicamos estratégias diferentes de controle de qualidade para garantir que não cometemos erros na aplicação do método e não encontramos sinais de problemas. Antes de testarmos o *machine learning* nesta espécie iremos verificar se problemas pré-existentes não podem estar causando o número elevado de genes novos.

Apesar de serem necessárias mais etapas para a finalização deste projeto, já é possível observar as importantes diferenças biológicas entre genes novos e antigos e como podemos utilizar estas mesmas diferenças para identificar genes novos através de *machine learning*.

7. Bibliografia

ABASCAL, Federico; ZARDOYA, Rafael; TELFORD, Maximilian J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. **Nucleic acids research**, v. 38, n. suppl_2, p. W7-W13, 2010.

AHO, Alfred V.; KERNIGHAN, Brian W.; WEINBERGER, Peter J. Awk—a pattern scanning and processing language. **Software: Practice and Experience**, v. 9, n. 4, p. 267-279, 1979.

ALTSCHUL, Stephen F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic acids research**, v. 25, n. 17, p. 3389-3402, 1997.

ASHBURNER, Michael et al. ***Drosophila*. A laboratory handbook**. Cold spring harbor laboratory press, 1989.

BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, 2014.

BÖHNE, Astrid et al. Transposable elements as drivers of genomic and biological diversity in vertebrates. **Chromosome research**, v. 16, n. 1, p. 203-215, 2008.

BRAY, Nicolas L. et al. Near-optimal probabilistic RNA-seq quantification. **Nature biotechnology**, v. 34, n. 5, p. 525-527, 2016.

BREIMAN, Leo. Random forests. **UC Berkeley TR567**, 1999.

BUREAU, Alexandre et al. Mapping complex traits using Random Forests. In: **BMC genetics**. BioMed Central, 2003. p. 1-5.

CARDOSO-MOREIRA, Margarida et al. Evidence for the fixation of gene duplications by positive selection in *Drosophila*. **Genome research**, v. 26, n. 6, p. 787-798, 2016.

CHEN, Xing; YAN, Gui-Ying. Semi-supervised learning for potential human microRNA-disease associations inference. **Scientific reports**, v. 4, n. 1, p. 1-10, 2014.

CLARK, Andrew G. et al. Evolution of genes and genomes on the *Drosophila* phylogeny. **Nature**, v. 450, n. 7167, p. 203-218, 2007.

CUI, Xiao et al. Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. **Molecular Plant**, v. 8, n. 6, p. 935-945, 2015.

DAVIS, Jesse; GOADRICH, Mark. The relationship between Precision-Recall and ROC curves. In: **Proceedings of the 23rd international conference on Machine learning**. 2006. p. 233-240.

DÍAZ-URIARTE, Ramón; DE ANDRES, Sara Alvarez. Gene selection and classification of microarray data using random forest. **BMC bioinformatics**, v. 7, n. 1, p. 1-13, 2006.

DOBIN, Alexander et al. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, v. 29, n. 1, p. 15-21, 2013.

DOBZHANSKY, Th. Genetics of natural populations. XIII. Recombination and variability in populations of *Drosophila* pseudoobscura. **Genetics**, v. 31, n. 3, p. 269, 1946.

DOUGHERTY, Dale; ROBBINS, Arnold. **sed & awk: UNIX Power Tools**. " O'Reilly Media, Inc.", 1997.

EDGAR, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic acids research**, v. 32, n. 5, p. 1792-1797, 2004.

EMMS, David M.; KELLY, Steven. OrthoFinder: phylogenetic orthology inference for comparative genomics. **Genome biology**, v. 20, n. 1, p. 1-14, 2019.

FITCH, Walter M. Distinguishing homologous from analogous proteins. **Systematic zoology**, v. 19, n. 2, p. 99-113, 1970.

FUKUSHIMA, Kenji; POLLOCK, David D. Amalgamated cross-species transcriptomes reveal organ-specific propensity in gene expression evolution. **Nature communications**, v. 11, n. 1, p. 1-14, 2020.

GUO, Xinjian et al. On the class imbalance problem. In: **2008 Fourth international conference on natural computation**. IEEE, 2008. p. 192-201.

GRABHERR, Manfred G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. **Nature biotechnology**, v. 29, n. 7, p. 644, 2011.

HAAS, Brian J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. **Nature protocols**, v. 8, n. 8, p. 1494-1512, 2013.

HARRIS, Robert S. **Improved pairwise alignment of genomic DNA**. The Pennsylvania State University, 2007.

HEGER, Andreas; PONTING, Chris P. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. **Genome research**, v. 17, n. 12, p. 1837-1849, 2007.

HOFFMAN, Michael M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. **Nature methods**, v. 9, n. 5, p. 473-476, 2012.

HOLLIDAY, Jason A.; WANG, Tongli; AITKEN, Sally. Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (Picea sitchensis) using random forest. **G3: Genes| genomes| genetics**, v. 2, n. 9, p. 1085-1093, 2012.

HOWE, Kevin L. et al. Ensembl 2021. **Nucleic acids research**, v. 49, n. D1, p. D884-D891, 2021.

HOWE, Kevin L. et al. Ensembl 2021. **Nucleic acids research**, v. 49, n. D1, p. D884-D891, 2021.

http://www.globalsoftwaresupport.com/wp-content/uploads/2018/02/ggff5544hh.png

https://en.wikipedia.org/wiki/Precision_and_recall

https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392

https://mnlab.uchicago.edu/sppress/index.php?methods=1

JIANG, Long et al. RPL10L is required for male meiotic division by compensating for RPL10 during meiotic sex chromosome inactivation in mice. **Current Biology**, v. 27, n. 10, p. 1498-1505. e6, 2017.

KAESSMANN, Henrik. Origins, evolution, and phenotypic impact of new genes. **Genome research**, v. 20, n. 10, p. 1313-1326, 2010.

KAUFMAN, Thomas C.; LEWIS, Ricki; WAKIMOTO, Barbara. Cytogenetic analysis of chromosome 3 in *Drosophila* melanogaster: the homoeotic gene complex in polytene chromosome interval 84a-B. **Genetics**, v. 94, n. 1, p. 115-133, 1980.

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

KOHAVI, Ron et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **Ijcai**. 1995. p. 1137-1145.

KOSHIKAWA, Shigeyuki et al. Gain of cis-regulatory activities underlies novel domains of wingless gene expression in Drosophila. **Proceedings of the National Academy of Sciences**, v. 112, n. 24, p. 7524-7529, 2015.

KRISTENSEN, David M. et al. Computational methods for Gene Orthology inference. **Briefings in bioinformatics**, v. 12, n. 5, p. 379-391, 2011.

KRIVENTSEVA, Evgenia V. et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. **Nucleic acids research**, v. 47, n. D1, p. D807-D811, 2019.

LANGMEAD, B.; SALZBERG, S. L. Fast Gapped-Read Alignment With Bowtie 2. **Nature Methods**. 2012; 9: 357–9.

LARKIN, Aoife et al. FlyBase: updates to the *Drosophila* melanogaster knowledge base. **Nucleic acids research**, v. 49, n. D1, p. D899-D907, 2021.

LEINONEN, Rasko et al. The sequence read archive. **Nucleic acids research**, v. 39, n. suppl_1, p. D19-D21, 2010.

LI, Bo; DEWEY, Colin N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. **BMC bioinformatics**, v. 12, n. 1, p. 1-16, 2011.

LIAW, Andy et al. Classification and regression by randomForest. **R news**, v. 2, n. 3, p. 18-22, 2002.

LIBBRECHT, Maxwell W.; NOBLE, William Stafford. Machine learning applications in genetics and genomics. **Nature Reviews Genetics**, v. 16, n. 6, p. 321-332, 2015.

LONG, Manyuan et al. The origin of new genes: glimpses from the young and old. **Nature Reviews Genetics**, v. 4, n. 11, p. 865-875, 2003.

LONG, Manyuan; LANGLEY, Charles H. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. **Science**, v. 260, n. 5104, p. 91-95, 1993.

LONG, Manyuan; WANG, Wen; ZHANG, Jianming. Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei, in *Drosophila*. **Gene**, v. 238, n. 1, p. 135-141, 1999.

LOVE, Michael I.; HUBER, Wolfgang; ANDERS, Simon. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome biology**, v. 15, n. 12, p. 1-21, 2014.

LUNARDON, Nicola; MENARDI, Giovanna; TORELLI, Nicola. ROSE: A Package for Binary Imbalanced Learning. **R journal**, v. 6, n. 1, 2014.

LUTZ, Mark. **Programming python**. " O'Reilly Media, Inc.", 2001.

MANNI, Mosè et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. **arXiv preprint arXiv:2106.11799**, 2021.

MILLER, Albert. The internal anatomy and histology of the imago of *Drosophila* melanogaster. **The biology of *Drosophila***, p. 421-534, 1950.

MORGAN, Thomas Hunt. An attempt to analyze the constitution of the chromosomes on the basis of sex-limited inheritance in *Drosophila*. **The Journal of Experimental Zoology**, v. 11, p. 365, 1911.

MORGAN, Thomas Hunt. The theory of the gene. **The American Naturalist**, v. 51, n. 609, p. 513-544, 1917.

OSHIRO, Thais Mayumi; PEREZ, Pedro Santoro; BARANAUSKAS, José Augusto. How many trees in a random forest?. In: **International workshop on machine learning and data mining in pattern recognition**. Springer, Berlin, Heidelberg, 2012. p. 154-168.

PATTERSON, John Thomas. **Studies in the Genetics of *Drosophila*: The Drosophilidoe of the Southwest. III**. The University, 1943.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RAICES, Julia B.; OTTO, Paulo A.; VIBRANOVSKI, Maria D. Haploid selection drives new gene male germline expression. **Genome research**, v. 29, n. 7, p. 1115-1122, 2019.

RAMEY, Chet. Bash, the Bourne− Again Shell. In: **Proceedings of The Romanian Open Systems Conference & Exhibition (ROSE 1994), The Romanian UNIX User's Group (GURU)**. 1994. p. 3-5.

REVELLE, William R. psych: Procedures for personality and psychological research. 2020.

ROBERTS, David B. *Drosophila* melanogaster: the model organism. **Entomologia experimentalis et applicata**, v. 121, n. 2, p. 93-103, 2006.

ROMBEL, Irene T. et al. ORF-FINDER: a vector for high-throughput gene identification. **Gene**, v. 282, n. 1-2, p. 33-41, 2002.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

SARKAR, Deepayan. **Lattice: multivariate data visualization with R**. Springer Science & Business Media, 2008.

SHAO, Yi et al. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. **Genome research**, v. 29, n. 4, p. 682-696, 2019.

SISTROM, Christopher L.; GARVAN, Cynthia W. Proportions, odds, and risk. **Radiology**, v. 230, n. 1, p. 12-19, 2004.

SMIT, A. Smit, AFA, Hubley, R & Green, P. **RepeatMasker Open-4.0**, v. 2015, 2013.

SOUMILLON, Magali et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. **Cell reports**, v. 3, n. 6, p. 2179-2190, 2013.

STANLEY JR, Craig E.; KULATHINAL, Rob J. flyDIVaS: a comparative genomics resource for *Drosophila* divergence and selection. **G3: Genes, Genomes, Genetics**, v. 6, n. 8, p. 2355-2363, 2016.

STEIN, Lincoln D. et al. The generic genome browser: a building block for a model organism system database. **Genome research**, v. 12, n. 10, p. 1599-1610, 2002.

TANG, Haibao et al. Synteny and collinearity in plant genomes. **Science**, v. 320, n. 5875, p. 486-488, 2008.

TATUSOV, Roman L.; KOONIN, Eugene V.; LIPMAN, David J. A genomic perspective on protein families. **Science**, v. 278, n. 5338, p. 631-637, 1997.

VIBRANOVSKI, Maria D. et al. Stage-specific expression profiling of *Drosophila* spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. **PLoS genetics**, v. 5, n. 11, p. e1000731, 2009.

VOORDECKERS, Karin; VERSTREPEN, Kevin J. Experimental evolution of the model eukaryote Saccharomyces cerevisiae yields insight into the molecular mechanisms underlying adaptation. **Current opinion in microbiology**, v. 28, p. 1-9, 2015.

WALL, D. P.; FRASER, H. B.; HIRSH, A. E. Detecting putative orthologs.

Bioinformatics, v. 19, n. 13, p. 1710-1711, 2003.

WALL, Larry et al. **The Perl programming language**. 1994.

WITT, Evan et al. Transcription factors drive opposite relationships between gene age and tissue specificity in male and female *Drosophila* gonads. **Molecular biology and evolution**, v. 38, n. 5, p. 2104-2115, 2021.

www.putty.org

YANG, Shuang et al. Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. **PLoS genetics**, v. 4, n. 1, p. e3, 2008.

YANG, Ziheng et al. PAML: a program package for phylogenetic analysis by maximum likelihood. **Computer applications in the biosciences**, v. 13, n. 5, p. 555-556, 1997.

ZAORSKA, Katarzyna; ZAWIERUCHA, Piotr; NOWICKI, Michał. Prediction of skin color, tanning and freckling from DNA in Polish population: linear regression, random forest and neural network approaches. **Human genetics**, v. 138, n. 6, p. 635-647, 2019.

ZHANG, Yong E. et al. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. **Genome research**, v. 20, n. 11, p. 1526-1533, 2010.

ZHOU, Qi et al. On the origin of new genes in *Drosophila*. **Genome research**, v. 18, n. 9, p. 1446-1455, 2008.