Júlia Beck Raíces

VERSÃO CORRIGIDA.

O original encontra-se disponível no Instituto de Biociências.

Modelo de Seleção Haplóide para Evolução de Genes Novos

Haploid Selection Model for New Genes Evolution

São Paulo

22 de agosto de 2017

Júlia Beck Raíces

VERSÃO CORRIGIDA.

O original encontra-se disponível no Instituto de Biociências.

Modelo de Seleção Haplóide para Evolução de Genes Novos

Haploid Selection Model for New Genes Evolution

Dissertação apresentada ao Instituto de Biociências da Universidade de São Paulo, para a obtenção de Título de Mestre em Ciências, na Área de Genética e Biologia Evolutiva.

Universidade de São Paulo – USP

Instituto de Biociências

Programa de Pós-Graduação em Genética e Biologia Evolutiva

Orientador: Maria D. Vibranovski

São Paulo

22 de agosto de 2017

Raíces, Júlia Beck
Modelo de Seleção Haplóide para Evolução de Genes Novos / Júlia Beck Raíces; orientadora Maria Dulcetti Vibranovski. -- São Paulo, 2017.

137 f. + anexo

Dissertação (Mestrado) - Instituto de Biociências da Universidade de São Paulo, Departamento de Genética e Biologia Evolutiva.

1. Genes novos. 2. Seleção haplóide. 3. Evolução. 4. Cromossomo X. I. Vibranovski, Maria Dulcetti, orient. II. Título.

Catalogação da Publicação
Serviço de Biblioteca do Instituto de Biociências

Júlia Beck Raíces

VERSÃO CORRIGIDA.

O original encontra-se disponível no Instituto de Biociências.

Modelo de Seleção Haplóide para Evolução de Genes Novos

Haploid Selection Model for New Genes Evolution

Dissertação apresentada ao Instituto de Biociências da Universidade de São Paulo, para a obtenção de Título de Mestre em Ciências, na Área de Genética e Biologia Evolutiva.

Trabalho aprovado. São Paulo, _____ de _____ de 2017:

_____

Maria D. Vibranovski
Orientadora

_____

Professor(a) Dr(a).

_____

Professor(a) Dr(a).

São Paulo
2017

Aos tormentos passados, presentes, e futuros. Todos melhores por serem ao seu lado.

Agradecimentos

Esta dissertação jamais teria chegado à termo sem o apoio e ajuda de diversas pessoas, tanto dentro como fora da academia. Aqui listo algumas, mas certamente não todas.

Primeiro de tudo, agradeço à Maria, minha orientadora, por todos os anos que trabalhamos juntas, e por ter me ensinado como fazer ciência. A agradeço também por criar para mim e todas as outras um exemplo de que também podemos ser grandes pesquisadoras. Para já tirar da lista, agradeço à FAPESP e ao CNPQ pelo apoio financeiro para que eu pudesse fazer esse trabalho. A FAPESP financiou este projeto sob o número 2014/17149-0. E as opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do(s) autor(es) e não necessariamente refletem a visão da FAPESP.

Agradeço também à todas as pessoas do Vibranovski Lab pelas conversas, ideias e apoios. Camila, Gabriel, Mari, Mara e Carol: vocês foram essenciais para que eu conseguisse terminar essa dissertação.

Aproveito deixa do laboratório para agradecer todas as pessoas que frequentam o Laboratório de Drosofilídeos, onde passei a maior parte dos últimos 5 anos. Ao Professor Vilela pela ajuda e heurística essenciais para que eu aprendesse a mexer com as moscas. À Professora Mori pelas conversas, risadas e comidas maravilhosas. Aos técnicos Francisco e Lopes, sem os quais eu não saberia nem mexer minha mesa no laboratório, muito menos cuidar das linhagens de moscas. Ao Augusto por todos os fins de semana e noites passadas no laboratório para conseguirmos terminar nossos trabalhos, e todos os descarregos para conseguirmos (sobre)viver a academia.

Agradeço também a todas as pessoas envolvidas no *Journal Club* e *Evolução no porão*, suas ideias e *inputs* tanto para o meu trabalho quanto para os de outras pessoas, que foram muito importantes para melhorar essa dissertação. Agradeço imensamente aos professores Diogo Meyer, Gabriel Marroig, e Rodrigo Cogni, da minha banca de qualificação. Suas ideias, sugestões e críticas fizeram desse trabalho muito melhor, e de mim uma cientista melhor.

Pela ajuda para programar e pensar sobre estatística agradeço Chalom. Pelos momentos de pensar sobre a vida acadêmica e sobre nosso papel no mundo, assim como pelo apoio e carinho em todos os momentos que mais precisei agradeço Chico, Oz, Lo, Érica, Anne e Rodrigo. Sem vocês eu nunca chegaria aqui (nem em nenhum outro lugar, pra ser honesta). Por me ouvir reclamar do que dava errado em todos os passos, e pelos

cafés mais divertidos e cheios de aprendizagens (e por cuidar tão bem da Siri) agradeço ao Pedro. Por me mostrar como é ser cientista pelo mundo, e por me proporcionar diversos momentos de (auto-)crítica, crescimento, e descanso, agradeço a todas as pessoas que falam sobre carreira acadêmica e sobre suas pesquisas no Twitter. Agradeço imensamente ao meu psiquiatra e minha psicóloga, sem quem tudo teria acabado muito antes.

Por fim, agradeço minha família pelo apoio, amor, e sugestões. Agradeço o apoio e ajuda desde que eu disse pela primeira vez que queria ser bióloga. E ao Elias, pelas imagens editadas, os textos corrigidos, as risadas, o apoio mútuo, os bolos, todos os gatos bonitos, e por sempre me suportar e apoiar, mesmo no meio das minhas infindáveis crises.

"I am a dire wolf, prey stalking, lethal prowler"

"I am a hunter, horse-mounted, wolf-stabbing"

"I am a horsefly, horse-stinging, hunter-throwing"

"I am a spider, fly-consuming, eight legged"

"I am a snake, spider-devouring, poison-toothed"

"I am an ox, snake-crushing, heavy footed"

"I am an anthrax, butcher bacterium, warm-life destroying;"

"I am a world, space-floating, life nurturing."

"I am a nova, all-exploding... planet-cremating"

"I am the Universe – all things encompassing, all life embracing"

"I am anti-life, the beast of judgment. I am the dark at the end of everything.

The end of universes, gods, worlds... of everything."

"I am hope."

The Sandman*, Neil Gaiman*

Resumo

Genes novos, definidos como aqueles presentes em um grupo ou espécie mas ausentes em seu grupo irmão e grupo externo, são conhecidos por terem mais marcadores de seleção positiva que genes antigos. Sabe-se também que a seleção positiva ocorre de forma mais rápida em sistemas haplóides do que em sistemas diplóides. Aqui unimos esses dois sistemas para propor um modelo de seleção haplóide de genes novos. Para isso utilizamos dados de expressão, idade evolutiva, e assinatura de seleção provenientes de genes de *Drosophila melanogaster*. Mostramos que genes novos adquirem uma vantagem seletiva se expressos nas fases tardias da espermatogênese, que são haplóides. Não só há mais genes novos com alta expressão nas fases haplóides (meiótica e pós-meiótica) da espermatogênese em relação à fase diplóide (mitótica), mas também os genes novos possuem expressão mais acentuada que genes antigos nessas fases haplóides. Mostramos que os genes com alta expressão nas fases haplóides possuem mais marcadores de seleção positiva, e.g. valores de dN/dS, *alpha* e para outros modelos que estimam seleção positiva. Dessa forma, propomos um modelo que explica a maior expressão de genes novos nos testículos (fases haplóides da espermatogênese) e como tais genes se fixam na espécie. Por fim, explicamos a maior incidência de genes extremamente novos ligados ao X e com expressão preferencial em machos. Isso ocorre por que genes ligados ao cromossomo X em machos tem expressão funcionalmente haplóide, visto que o X está em hemizigose em todas as células somáticas de machos. Essa situação torna benéfico para genes muito novos ligados ao X que sua seleção em machos ocorra, pois em todas as células tais genes tem o benefício da seleção haplóide. Em particular, genes extremamente novos, ao contrário de genes antigos, do cromossomo X são capazes de burlar a inativação do cromossomo sexual durante a meiose. Isso os torna um bom sistema para novos alelos recessivos e com antagonismo sexual serem expressos e selecionados.

Palavras-chaves: genes novos, seleção haplóide, evolução, cromossomo X.

Abstract

New genes, those present in one group or species but absent in their sister group and outgroup, are frequently under positive selection, as shown by their higher rates and values of positive selection markers. It is also known that beneficial genes in haploid systems tend to be fixed more quickly than in diploid ones. Here we propose to merge this two systems by proposing a model for new genes selection in haploid systems. To do so we use *Drosophila melanogaster* 's spermatogenesis process. We show that new genes have a selective advantage if they are expressed in the haploid (meiotic and post-meiotic) phases of spermatogenesis. This is shown not only by the greater proportion of new genes with high expression in those phases against the diploid (mitotic) phase, but also by the intensity of the expression of new genes being greater than that of old genes during the haploid phases. We also show that genes with higher expression in the haploid phases present more markers of positive selection. They have both a greater value of dN/dS, *alpha* and a greater proportion of genes that best fit a model with selection than one without it. Therefore, we propose a model explaining the higher expression of new genes in the testis (haploid phases of spermatogenesis) and how those genes become fixed in the population and species. At last we explain the abundance of new male-biased genes on the X. This enrichement is due to the functionally haploid expression of the X-linked genes in males. This assures an advantage to those genes, as they will benefit from haploid selection system in the X on males. It is then beneficial for X-linked genes to be selected uppon in the males, as it will resemble haploid selection. Also, extremely new X-linked genes, as opposed to old ones, can bypass the sex cromosome inactivation, being a good system for new antagonistic recessive alleles to be expressed and sellected upon.

Key-words: new genes, haploid selection, evolution, X-chromosome.

Lista de ilustrações

Figura 1 – Relações filogenéticas na identificação de um gene novo. 20
Figura 2 – Principais mecanismos de surgimento de genes novos. 21
Figura 3 – Distribuição do surgimento de genes novos ao longo da evolução. . . . 23
Figura 4 – Número de gene novos essenciais ao longo dos principais ramos evolutivos em *Drosophila*. 24
Figura 5 – Proporção de genes novos originados por diferentes mecanismos em *Drosophila* 24
Figura 6 – Proporção de genes novos originados por duplicações. 25
Figura 7 – Pseudogenizacao de uma cópia gênica. 26
Figura 8 – Possíveis caminhos evolutivos após o surgimento de um novo gene por duplicação 27
Figura 9 – Representação dos processos e células presentes em cada uma das fases da espermatogênese de Drosophila 30
Figura 10 – Expressão de retrogenes e retropseudogenes nas fases da espermatogênese de camundongo. 31
Figura 11 – Testículo de *Drosophila melanogaster* 36
Figura 12 – Métodos de Vibranovski et al. (2009) 37
Figura 13 – Classes de expressão durante espermatogênese 38
Figura 14 – Expressão fase-a-fase ao longo da espermatogênese 39
Figura 15 – Análise de sintenia 41
Figura 16 – Filogenia de *Drosophila* e o surgimento de genes novos ao longo da mesma 41
Figura 17 – Distribuição β 42
Figura 18 – Simulação de fixação de alelo em populações haplóides e diplóides . . . 50
Figura 19 – Esquema de como se dá a expressão de genes autossômicos e ligados ao X ao longo da espermatogênese. 52
Figura 20 – Comparação da espermatogênese com a oogênese em *Drosophila melanogaster* 53
Figura 21 – Proporção de genes autossômicos super-expressos em cada fase da espermatogênese 54
Figura 22 – Esquema da meiose e ploidia das células nesse processo 55
Figura 23 – Expressão de genes autossômicos novos e antigos ao longo da espermatogênese 56
Figura 24 – Marcadores de seleção para genes autossômicos novos e antigos e dos grupos haplóide e mitótico. 58

Figura 25 – Proporção de genes novos autossômicos e ligados ao X ao longo da espermatogênese 61
Figura 26 – Comportamento esperado da expressão de genes novos autossômicos e ligados ao X durante a espermatogênese 62
Figura 27 – Localização cromossômica de genes preferencialmente expressos em machos 66
Figura 28 – Expressão de genes autossômicos e ligados ao X, novos e antigos ao longo da espermatogênese 67
Figura 29 – Inativação dos genes novos no cromossomo X 68

Lista de tabelas

Tabela 1 – Genes novos e antigos nas classes de Vibranovski et al. (2009) 40
Tabela 2 – Número de genes autossômicos e ligados ao X com alta expressão em cada fase da espermatogênese 44
Tabela 3 – Tamanhos amostrais dos grupos de genes utilizados 45
Tabela 4 – Tabela reproduzida parcialmente da tabela S1 do material suplementar de Zhang et al. (2010a) 46
Tabela 5 – Tabela reproduzida parcialmente da tabela S4 do material suplementar do material suplementar de Zhang et al. (2010a) 46
Tabela 6 – Tabela parcialmente reproduzida da Tabela Suplementar S1 de Vibranovski et al. (2009) 47
Tabela 7 – Tabela parcialmente reproduzida de Stanley e Kulathinal (2016) 47
Tabela 8 – Tabela feita com os dados de Vibranovski et al. (2009) e Zhang et al. (2010a) para análise 48
Tabela 9 – Valores de *alpha* e suas comparações usando o programa DoFE (*Distribution of Fitness Effect*) 60

Sumário

Lista de ilustrações 15

Lista de tabelas 16

1 INTRODUÇÃO 19
1.1 Genes novos 19
1.2 Espermatogênese 29

2 OBJETIVOS 33

3 MATERIAIS & MÉTODOS 35
3.1 Expressão gênica na espermatogênese 35
3.2 Idade dos genes 39
3.3 Marcadores de Seleção 40
3.4 Análises realizadas 43

4 RESULTADOS: MODELO DE SELEÇÃO HAPLÓIDE DE GENES NOVOS 49
4.1 Expressão de Genes Novos 53
4.2 Assinaturas de seleção 57
4.3 Genes Autossômicos e Ligados ao X 60
4.4 Modelo Matemático 62

5 RESULTADOS: LOCALIZAÇÃO CROMOSSÔMICA DE GENES PREFERENCIALMENTE EXPRESSOS EM MACHOS 65

6 CONCLUSÕES 71

REFERÊNCIAS 73

APÊNDICE A – PERL 81

APÊNDICE B – R SCRIPTS 89
B.1 Estatísticas e Gráficos 89
B.2 Simulações 117

ANEXO A – MODELO MATEMÁTICO 125

1 Introdução

A evolução de espécies é um assunto que fascina naturalistas (DARWIN, 1859; LAMARCK, 1809). Como as espécies surgem e como mudam ao longo do tempo é assunto de estudo e deslumbre para grande parte das pessoas que estudam as mudanças da vida na terra, seja nas áreas de botânica, zoologia ou paleontologia(DARWIN, 1859; LAMARCK, 1809; HUXLEY, 1942; MAYR, 1972; MAYR, 1982). Atualmente conhecemos algumas das situações e processos que levam espécies a aparecer. Um dos processos que pode levar a diferenciação de populações e eventualmente ao surgimento de espécies é pelo surgimento e fixação de genes novos (CRUZ; DAVIES, 2000; HOLLAND et al., 1994; SEEHAUSEN et al., 2014), que até hoje gera questões e incita novas pesquisas sobre o assunto (para mais detalhes ver Kaessmann, Vinckenbosch e Long (2009), Kaessmann (2010) e Kaessmann (2009)). Para entendermos um pouco melhor a importância desse processo usamos neste trabalho a expressão gênica durante a espermatogênese de *Drosophila melanogaster*. A importância de genes novos e sua expressão na espermatogênese e testículos só se torna clara quando compreendemos o que são e como surgem os genes novos.

1.1 Genes novos

Genes novos são aqueles que apareceram recentemente no genoma de um grupo ou espécie, e por isso estão presentes em uma espécie ou um clado, mas estão ausentes no grupo irmão e no grupo externo, como exemplificado na Figura 1. Hoje é sabido que o ganho de genes não é rara (LONG et al., 2003). Sabe-se também da importância do surgimento de genes novos na evolução do genoma, e, a longo prazo, das espécies (CHEN; KRINSKY; LONG, 2013; KAESSMANN, 2010; KAESSMANN; VINCKENBOSCH; LONG, 2009; KAESSMANN, 2009). Por exemplo, Chen, Zhang e Long (2010) mostraram que genes novos se tornam rapidamente essenciais. Já na espécie humana, Zhang et al. (2010b) mostraram que genes novos são mais expressos que genes antigos nos estágios iniciais do desenvolvimento do cérebro, tendo provavelmente papel importante na evolução de comportamentos cognitivos complexos e na estrutura desse órgão. No grupo dos drosofilídeos, Zhang et al. (2010a) mostraram que genes novos são bastante comuns ao realizarem um levantamento dos genes de *Drosophila melanogaster*, determinando também suas idades através da comparação de genes ortólogos em outras espécies do gênero.

Um dos primeiros trabalhos a caracterizar genes novos foi o de Long e Langley (1993). Nele, o gene *jingwei* é caracterizado como um "gene processado", ou seja, que foi inserido no genoma a partir de um transcrito de RNA. Hoje nos referimos à tais genes como retrogenes ou genes retrotranspostos. Esses genes são originados por uma retrotransposição,

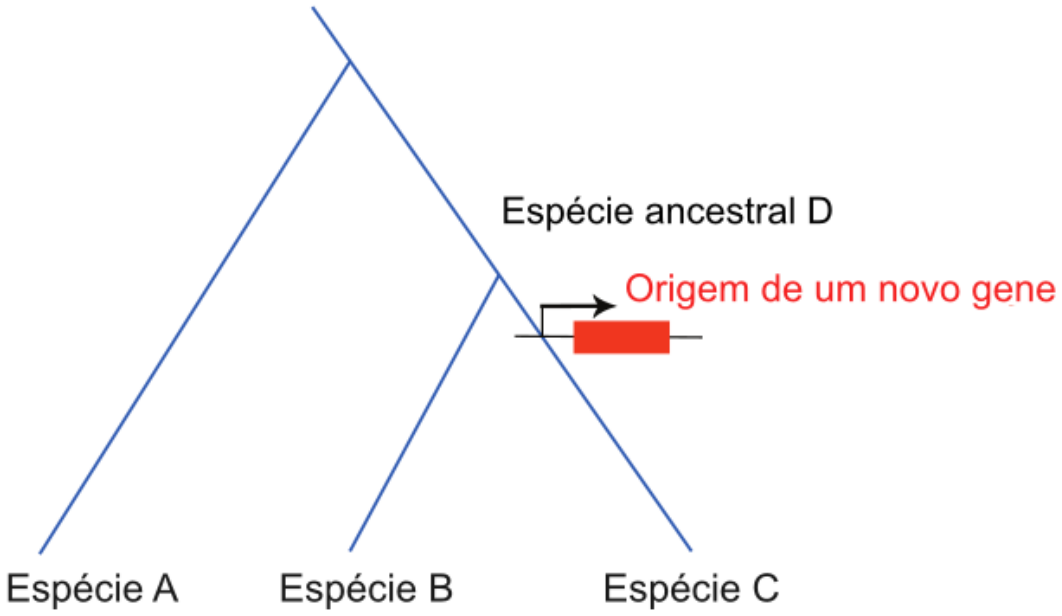


Figura 1 – **Relações filogenéticas na identificação de um gene novo.** Um gene novo está presente em um grupo ou espécie (espécie C no desenho), mas não é identificado no grupo irmão (espécie B) ou no grupo externo (espécie A). Neste caso, o novo gene só está presente na espécie C, estando portanto ausente nas espécies A e B. Figura reproduzida e traduzida de(CHEN; ZHANG; LONG, 2010)

ou seja, da inserção de um RNA mensageiro (RNAm) no genoma e possuem algumas características únicas como: i) ausência ou baixo número de íntrons; ii) presença de uma cauda poli-A (ou poli-T); e iii) ausência de promotores ou outros mecanismos de iniciação da transcrição (KAESSMANN, 2009). Embora tais características sejam comuns em retrogenes jovens, conforme eles envelhecem tais características podem desaparecer, dificultando sua identificação. Além de diversas outras características comuns à retrogenes, o gene *jingwei* possui expressão acentuada nos testículos, o que é comum à genes novos de forma geral (BABUSHOK et al., 2007; LEVINE et al., 2006; BETRÁN; THORNTON; LONG, 2002; MARQUES et al., 2005; SOUMILLON et al., 2013; VIBRANOVSKI et al., 2009).

O surgimento de genes novos a partir de retrotransposição de um RNA no genoma não é raro nem em *Drosophila* (BAI et al., 2007; BAI; CASOLA; BETRÁN, 2008; METTA; SCHLÖTTERER, 2010), nem em mamíferos (SOUMILLON et al., 2013; POTRZEBOWSKI et al., 2008; EMERSON et al., 2004), incluindo humanos (VINCKENBOSCH; DUPANLOUP; KAESSMANN, 2006), e nem em plantas (WANG et al., 2006), sendo tão comuns que levaram Kabza, Ciomborowska e Makałowska (2014) a criar a base de dados para retrogenes de animais, o RetrogeneDB (KABZA; CIOMBOROWSKA; MAKAŁOWSKA, 2014). De forma resumida, os retrogenes são orginados quando um RNAm

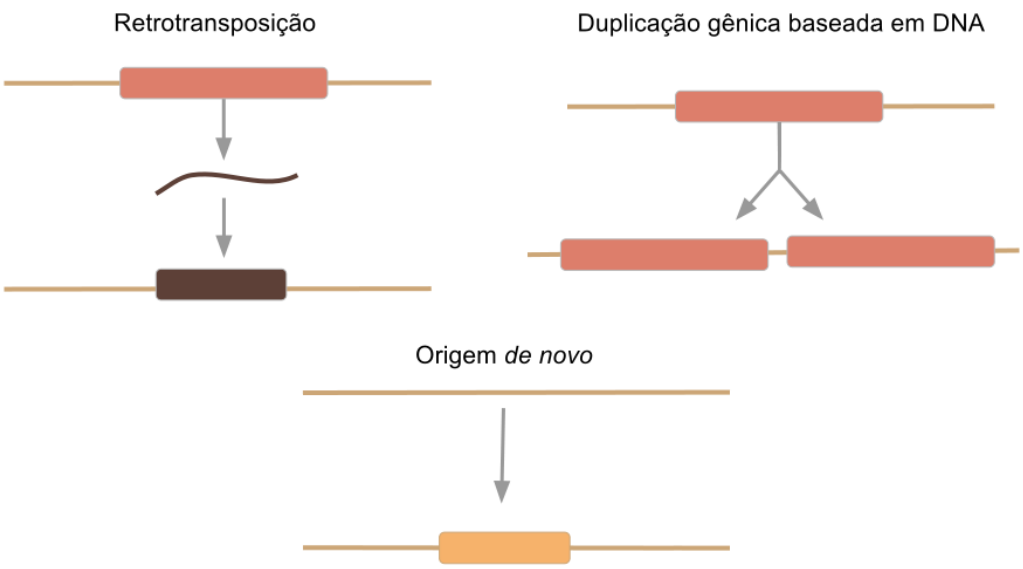


Figura 2 – **Principais mecanismos de surgimento de genes novos.** Retrotransposição: um novo gene surge devido à retrotranscrição de uma molécula de RNAm de um gene parental, em que uma molécula de cDNA que é inserida de volta ao DNA nuclear, formando um gene curto, sem introns e sem a maquinaria genética necessária para a sua expressão. Duplicação Baseada em DNA: um trecho de DNA contendo um ou mais genes é duplicado. Nesse tipo de duplicação os introns e a maquinaria de expressão costumam ser duplicados junto ao gene. Origem *de novo*: uma região genômica que antes não codificava nenhum gene passa a expressar um novo gene, diferente de todos os outros presentes naquele genoma.

é inserido em alguma parte do genoma por meio de uma retrotransposição do mesmo (Figura 2). Esse processo costuma ocorrer com RNA mensageiro (RNAm) já maduro, ou seja, com os introns removidos e com marcadores como a cauda de adeninas. Desta forma, a molécula de RNAm com estas características é retrotranscrita numa molécula de cDNA que pode ser inserida no genoma formando um gene novo com as características descritas acima. Além disso, esses genes não costumam ter toda a maquinaria de expressão quando são retrotranspostos, pois esta não está presente no RNAm, e sim na região do genoma próxima ao gene. Por isso os retrogenes podem, por exemplo, acabar utilizando a maquinaria de expressão de genes próximos (KAESSMANN; VINCKENBOSCH; LONG, 2009).

Outro tipo de duplicação que pode dar origem a novos genes são as duplicações baseadas em DNA (Figura 2) (KAESSMANN; VINCKENBOSCH; LONG, 2009; CHEN; KRINSKY; LONG, 2013). Neste processo todo o gene costuma ser duplicado, muitas vezes incluindo não só exons e introns mas também a maquinaria de expressão. Estas duplicações são mais facilmente identificadas como tal, pois em muito se assemelham ao gene parental, o que nem sempre ocorre em duplicações de RNA (KAESSMANN, 2010). Ambos os tipos

de duplicações são capazes de gerar genes que se tornam essenciais ou importantes para o desenvolvimento de novos grupos, como é o exemplo do gene retrotransposto *GLUD2* relacionado ao metabolismo de glutamato em hominídeos (ROSSO et al., 2008a), ou a duplicação baseada em DNA de uma ribonuclease que auxilia macacos herbívoros a absorver proteínas, a *RNASE1B* (ZHANG; ZHANG; ROSENBERG, 2002).

Por fim, uma das formas menos compreendidas de surgimento de novos genes é a origem *de novo* (Figura 2). Esse tipo de gene aparece quando uma região do genoma que anteriormente não codificava nenhum gene passa a ser expressa. Nessas situações a região do novo gene passa a transcrever um RNA e/ou traduzir uma proteína, esse transcrito ou proteína é então selecionado e fixado levando a origem de um gene funcional. Há duas possibilidades para como se inicia a expressão de tal região. Uma sugere que primeiro há uma ORF (*Open Reading Frame*, o início do quadro de leitura de um gene) e os elementos regulatórios envolvidos na transcrição são adquiridos posteriormente. A outra possibilidade é que o genoma seja em geral transcrito, e tais transcritos adquirem novas mutações e são selecionados, gerando novos genes (SCHLOTTERER, 2015). Embora não sejam muito comuns (CHEN; ZHANG; LONG, 2010; ZHANG, 2003), há alguns casos descritos de genes com origem de novo funcionais, incluindo o *BSC4* em leveduras que está envolvido com o reparo de DNA (CAI et al., 2008).

Apesar de mesmo hoje não sabermos ao certo a função do gene *jingwei* (LONG; LANGLEY, 1993), sabemos a função de diversos outros genes novos (HEIDMANN et al., 2009; DUPRESSOIR et al., 2009; DAI et al., 2008; WANG et al., 2002; CAI et al., 2008; ROSSO et al., 2008b; ROSSO et al., 2008a; ZHANG; ZHANG; ROSENBERG, 2002). Alguns genes novos estão envolvidos no aparecimento de novos grupos e espécies (para mais detalhes ver Kaessmann, Vinckenbosch e Long (2009) e Kaessmann (2010)). Genes novos envolvidos na formação da placenta em *Theria*, por exemplo, foram encontrados em coelhos e camundongos (HEIDMANN et al., 2009; DUPRESSOIR et al., 2009). Nesses casos, retrogenes foram fixados e auxiliam na formação da placenta nessas espécies. Genes novos podem também estar envolvidos em características próprias da espécie, como mostrado em Cai et al. (2008), Kleinjan et al. (2008) e Zhang, Zhang e Rosenberg (2002). O gene *sphinx*, por exemplo, também surgiu de uma retroposição (inserção de um RNAm no genoma), e está envolvido no comportamento de corte de *Drosophila melanogaster* (WANG et al., 2002). Além de mostrar um gene novo restrito a um grupo menor, o exemplo também mostra uma característica não-morfológica, mas sim comportamental sendo selecionada e fixada em uma espécie (WANG et al., 2002). Estes são apenas alguns dos casos conhecidos de genes novos com funções importantes nos grupos em que surgem. Para mais exemplos, discussões, e informações veja Kaessmann (2010), Kaessmann, Vinckenbosch e Long (2009), Kaessmann (2009), Zhou et al. (2008), Zhou e Wang (2008), Bai et al. (2007), Fablet et al. (2009) e Long et al. (2003).

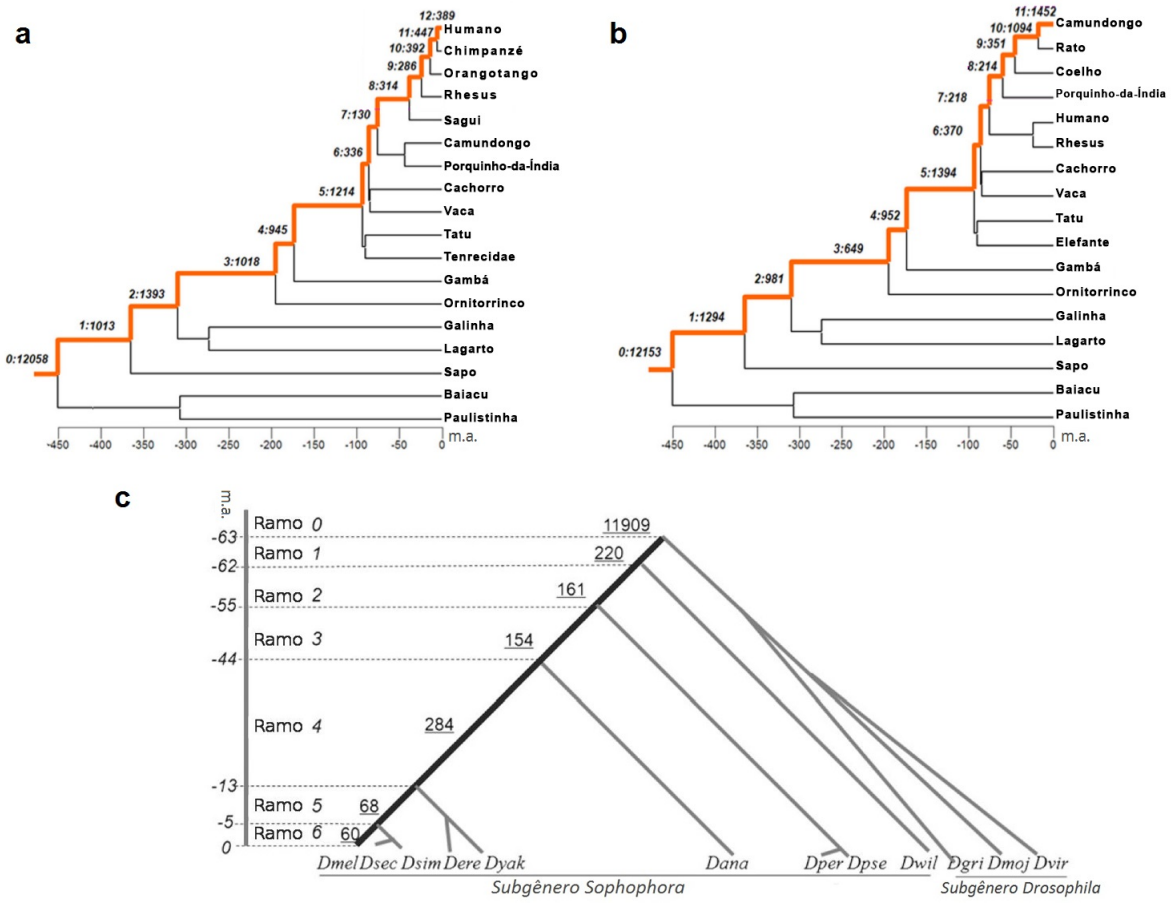


Figura 3 – **Distribuição do surgimento de genes novos ao longo da evolução.** Quantidade de genes novos surgidos em diferentes ramos evolutivos de vertebrados (a e b) e *Drosophila* (c). Em cada ramo vemos a quantidade de genes que devem ter aparecido para os ramos levando a camundongo (a), humanos (b) e *D. melanogaster* (c). As imagens (a) e (b) foram reproduzidas e traduzidas de Zhang et al. (2010b), e a imagem (c) foi reproduzida e traduzida de Zhang et al. (2010a). m.a.: milhões de anos.

Genes novos são assim comuns em todos os principais grupos de organismos, como em *Drosophila*, mamíferos e plantas. Na Figura 3 podemos ver a frequência de genes novos em diferentes ramos ao longo da filogenia de vertebrados nas partes A e B, e de *Drosophila* na parte C. Apesar de sua constância, os genes novos surgidos por processos diferentes apresentam diferentes frequências no genoma. Chen, Zhang e Long (2010) mostraram não apenas que em *Drosophila* duplicações baseadas em DNA são a forma mais comum de origem de genes novos, como podemos ver nas Figuras 4 e 5, mas também que a proporção de genes essenciais no genoma independe da idade de tais genes ou da forma como estes foram originados (CHEN; ZHANG; LONG, 2010; ZHOU et al., 2008). São definidos como genes essenciais aqueles sem os quais o indivíduo ou é inviável (morre sem atingir o estágio adulto) ou é infértil (não consegue gerar prole). Em sua revisão sobre evolução por duplicação gênica Zhang (2003) mostra que essa tendência de grande

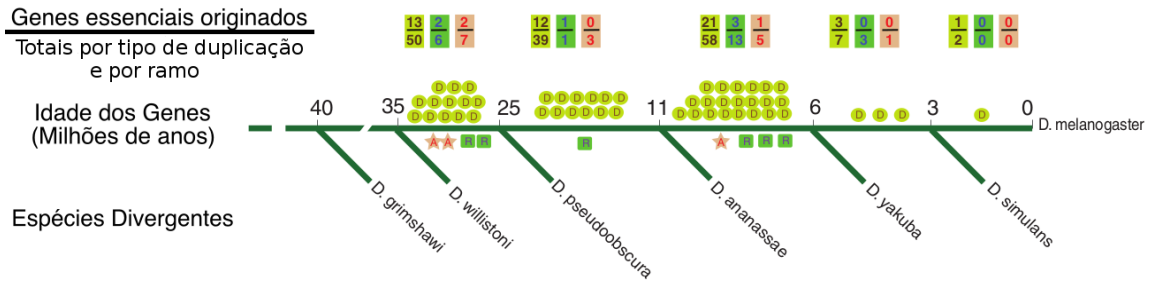


Figura 4 – **Número de gene novos essenciais ao longo dos principais ramos evolutivos em *Drosophila*.** Surgimento de genes por diferentes processos ao longo da filogenia de *Drosophila*. Em amarelo, marcado com D, estão representados os genes que surgiram por duplicações baseadas em DNA; em verde, marcados com R, são as retrotransposições; e em vermelho, marcados com A, as origens de novo. Na parte superior a proporção de genes essenciais que surgiram em cada ramo e por cada processo em relação ao total de genes originados no ramo por cada tipo de processo. Figura reproduzida e traduzida de Chen, Zhang e Long (2010).

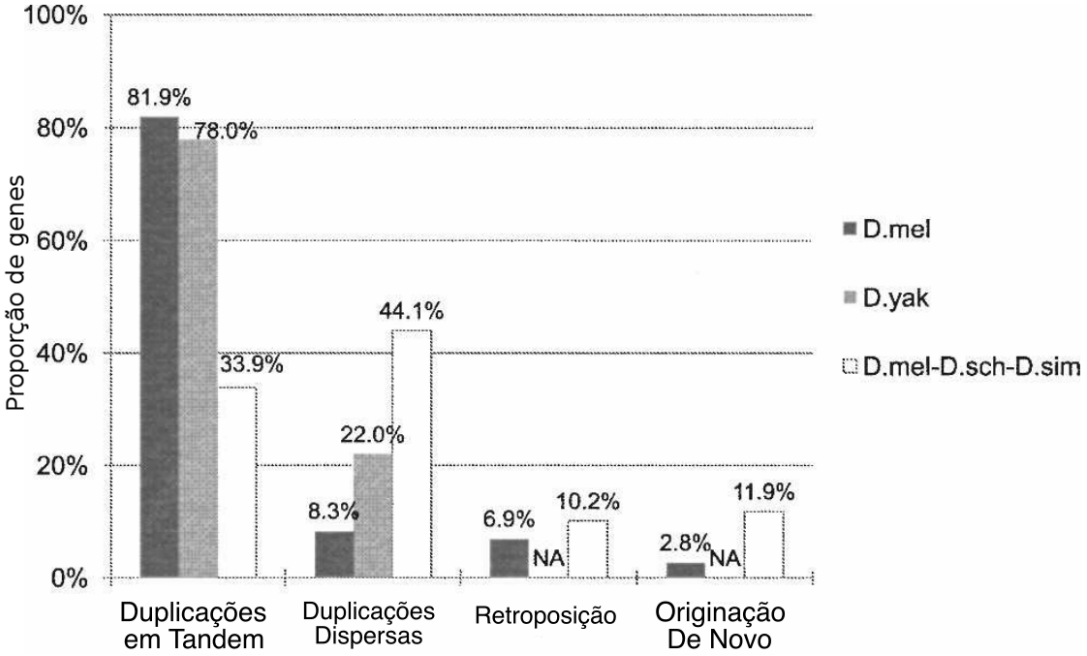


Figura 5 – **Proporção de genes novos originados por diferentes mecanismos em *Drosophila*.** Contribuição relativa de cada mecanismo de origem de genes novos. Para todos os grupos de *Drosophila* representados as duplicações baseadas em DNA (em tandem e dispersas) são a forma mais comum de origem de novos genes. D. mel: *D. melanogaster*; D. yak: *D. yakuba*; D.sch: *D. sechelia*; e D.sim: *D. simulans*. Figura reproduzida e traduzida de Zhou et al. (2008).

proporção de genes novos originados por duplicação não ocorre apenas em *Drosophila*, mas sim em diversas espécies dos três domínios de seres vivos, como podemos ver na Figura 6, incluindo mamíferos (ZHANG et al., 2010b).

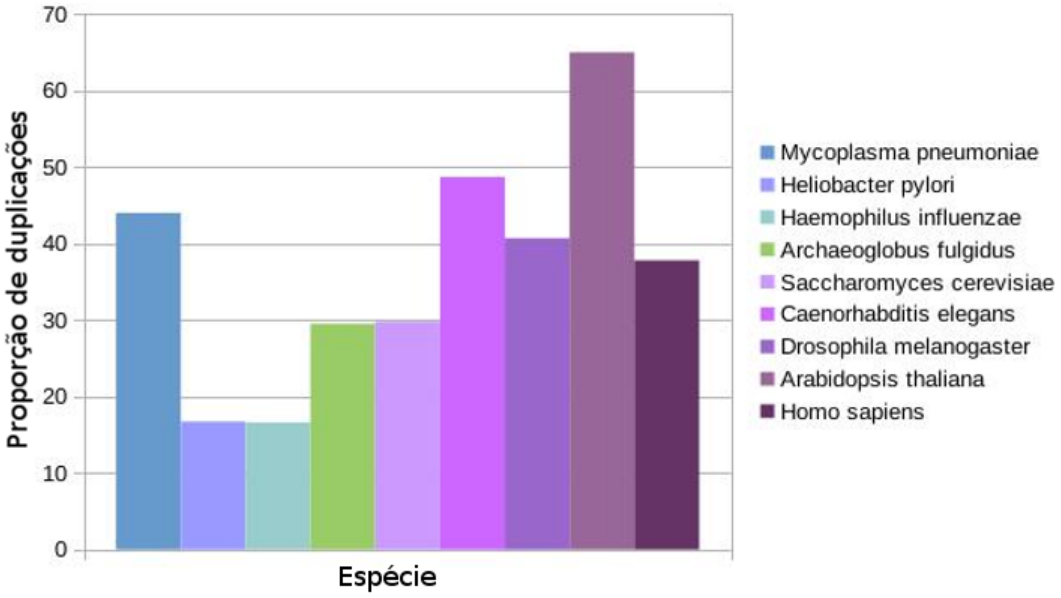


Figura 6 – **Proporção de genes novos originados por duplicações.** Dados obtidos para diferentes espécies abrangendo os três domínios de seres vivos. Apesar de os dados de cada espécie não poderem ser comparados diretamente —pois foram agrupados de diferentes trabalhos na revisão de Zhang (2003)—a figura mostra que esta é uma forma muito comum para a origem de novos genes, de forma que faz sentido que tratemos principalmente dela em nosso modelo. Figura feita a partir de dados obtidos de Zhang (2003).

Sendo a duplicação uma forma muito comum de surgimento de genes novos, Figuras 4, 5 e 6, aparece em consequência a pergunta de como se mantém evolutivamente dois genes iguais, e possivelmente com a mesma função, no genoma. Em 1972 Ohno trabalhava neste problema e mostrou que na maior parte das vezes se espera que uma das cópias acumule mutações deletérias e acabe sendo perdida por não adicionar nenhuma vantagem adaptativa (OHNO, 1972). Quando uma das cópias é perdida, ela se torna um pseudogene (Figura 7). Um pseudogene se forma quando mutações deletérias, como códons de parada prematuros, são acumuladas ao longo do gene, tornando-o cada vez menos ativo, ou seja, seus transcritos e proteínas se tornam menores e com funções mais restritas, até que se tornem inexistente. Isso o transforma em um pseudogene, que embora possa continuar sendo transcrito já não possui função nas células ou nos tecidos.

Alternativamente, quando um gene não se torna um pseudogene pode ter alguns destinos, como exemplificado na Figura 8. As modificações sofridas pelo gene podem levar a neofuncionalização, subfuncionalização ou conservação do gene. Quando ocorre uma neofuncionalização uma das cópias do gene adquire uma nova função. Esta função pode tanto representar uma nova característica que funciona no mesmo ambiente em que o parental é funcional, ou até mesmo a mesma função exercida pelo gene parental, mas em um novo tecido ou nova organela. Rosso et al. (2008a) evidenciaram um evento de

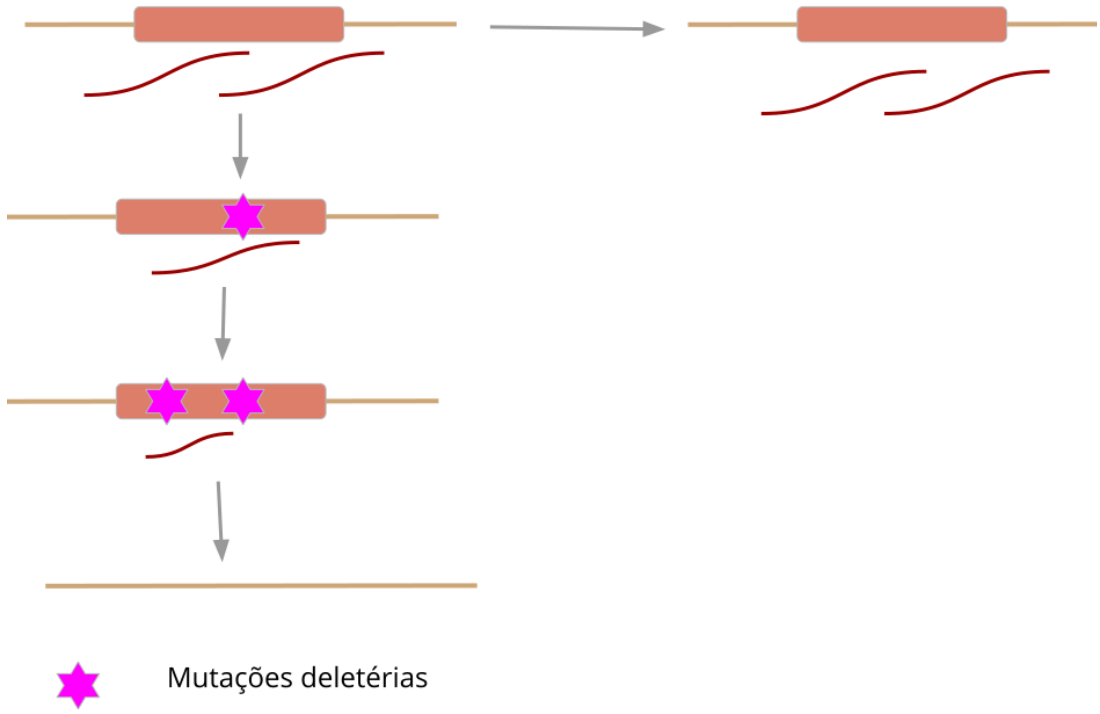


Figura 7 – **Pseudogenizacao de uma cópia gênica.** Quando um novo gene surge ele pode, de forma geral, seguir dois caminhos: pode ser fixado (parte superior); ou pode sofrer pseudogenização (parte inferior). Na pseudogenização um gene antes funcional perde sua função. Isso ocorre devido ao acúmulo de mutações deletérias ao longo do gene, o que leva à diminuição da expressão ou eficiência do produto gênico.

neofuncionalização quando identificaram que o gene *GLUD2* se encontra expresso exclusivamente em mitocôndrias enquanto o gene parental *GLUD1* se encontra em mitocôndrias e citoplasma. A enzima *GLUD* (glutamato desidrogenase) está envolvida no metabolismo celular, e a enzima derivada está potencialmente envolvida em adaptações que envolvam o desenvolvimento do cérebro (ROSSO et al., 2008a). Quando processos como esse ocorrem, uma nova adaptação é formada, e alguns indicadores de seleção podem ser identificados no gene, como foi mostrado para a glutamato desidrogenase e para genes envolvidos com o ciclo celular em humanos (ROSSO et al., 2008a; ROSSO et al., 2008b).

Já quando ocorre uma subfuncionalização dos genes ambas as cópias (o gene parental e o novo) passam a dividir a função que antes era exercida apenas pela proteína parental. Dessa forma, cada gene se especializa em uma parte da função antes exercida apenas pelo parental, de forma que nenhuma adapatação nova é formada. O que ocorre, na verdade, é uma nova divisão da função já existente. Em paulistinha (*Danio rerio*), por exemplo, o fator transcricional *pax6*, envolvido no desenvolvimento do cérebro, olhos, pâncreas, e trato olfatório, se encontra em duas cópias, enquanto nos outros vertebrados é encontrado em cópia única. Nesse peixe, os dois genes *pax6a* e *pax6b* são necessários para o desenvolvimento adequado dos órgão citados (KLEINJAN et al., 2008).

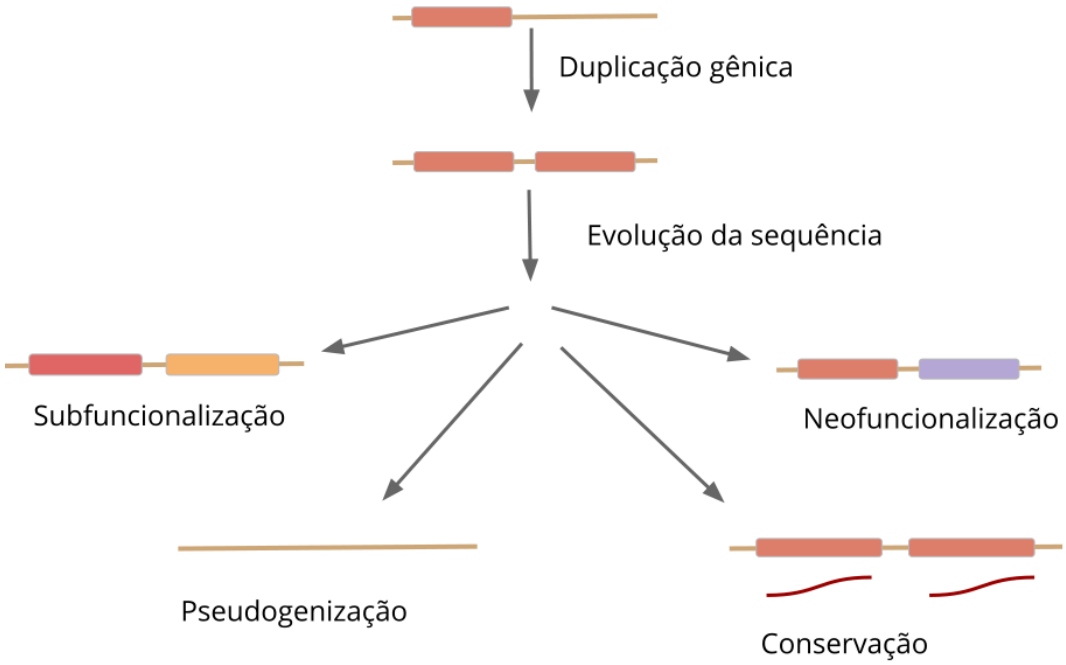


Figura 8 – **Possíveis caminhos evolutivos após o surgimento de um novo gene por duplicação** (ZHANG et al., 2010b; ZHOU et al., 2008; CHEN; ZHANG; LONG, 2010; ZHANG, 2003). Pseudogenização: perda da função por acúmulo de mutações deletérias. Conservação: manutenção da mesma função que seu gene parental pois mais cópias do mesmo gene se torna vantajoso ao organismo. Subfuncionalização: o novo gene passa a dividir a função antes realizada apenas pelo gene parental. Neofuncionalização: novo gene adquire uma nova função, seja sendo expresso em uma nova região, seja por ter um novo produto gênico.

Por fim, quando há a conservação dos genes duplicados ambas as cópias permanecem realizando a mesma função. Essa situação pode ocorrer, por exemplo, quando uma maior taxa de expressão do gene duplicado é benéfica para o organismo, como ocorre em *Ipomoea purpurea*, onde há 3 duplicações em tandem (ou seja, uma logo após a outra no genoma) do gene envolvido no metabolismo de antocianina (Des Marais; RAUSHER, 2008). Des Marais e Rausher (2008) mostraram que as três sequências são conservadas e entre duas delas o local de expressão também é conservado. A pesquisa mostrou também que tal conservação gênica deve ocorrer devido às pressões seletivas que mantém as cópias iguais.

Saber em quais tecidos diferentes genes são expressos permite sabermos onde tais genes possuem função, visto que a expressão gênica se relaciona à função do mesmo naquele ambiente (LIAO; WENG, 2015). Para saber as possíveis funções dos genes novos foram feitos diferentes estudos para identificar processos nos quais estejam envolvidos. Nesses estudos diversos grupos encontraram uma grande quantidade de genes novos sendo altamente expressos nos testículos (BABUSHOK et al., 2007; LEVINE et al., 2006; BETRÁN; THORNTON; LONG, 2002; MARQUES et al., 2005; SOUMILLON et al., 2013; VIBRANOVSKI; ZHANG; LONG, 2009; CUI et al., 2015). Tais estudos identificaram este

padrão tanto para mamíferos (como em Soumillon et al. (2013)), quanto em *Drosophila* (vide Vibranovski, Zhang e Long (2009)), e até mesmo em plantas, onde genes novos de dicotiledôneas e de monocotiledôneas são mais expressos durante o processo de formação do pólen (CUI et al., 2015). A identificação de um processo ou órgão no qual genes novos estão intensamente envolvidos é muito interessante, pois tal processo pode estar intrinsecamente ligado ao aparecimento e fixação de tais genes.

Independente de como surgiu, qual tipo de destino teve, ou onde é mais expresso, um gene pode ser fixado ou perdido na população ou espécie. A fixação ou perda pode ocorrer devido a deriva ou a seleção. Quando o gene é fixado por deriva isso quer dizer que ele foi fixado ao acaso, e não obrigatoriamente é benéfico ao organismo. A deriva ocorre principalmente em populações pequenas, pois nelas é mais fácil que o acaso leve a fixação de um gene qualquer, independente de seu efeito nos indivíduos daquela população (GILLESPIE, 2010). A seleção pode ser direcional e leva genes benéficos ao organismo e/ou grupo a serem fixados na população e/ou espécie. Quando um gene traz vantagens adaptativas aos indivíduos que o possuem, ele é positivamente selecionado, e por isso a seleção ocorre no sentido de fixá-lo na população. Tais genes devem ter assinaturas de seleção positiva, como maiores valores de dN/dS e *alpha*. Porém, não é apenas a seleção de um gene que pode criar tais marcadores. Características da história natural do grupo, e eventos como o efeito carona podem também ser responsáveis por isso. Por exemplo, genes que são fixados pelo efeito carona (ou seja, por estarem próximos a um gene que está sendo selecionado por seu efeito positivo) podem apresentar alguns marcadores de seleção, embora não estejam sendo diretamente selecionados. Características da história evolutiva do grupo também podem levar a existência de marcadores de seleção positiva em genes que não foram positivamente selecionados (SELLA et al., 2009; EYRE-WALKER, 2006).

Essas assinaturas de seleção positiva mostram, em geral, que o gene está sendo modificado em relação ao parental, levando a novas características para aquele gene. Dois exemplos de medidas de seleção positiva são o dN/dS e o *alpha*. O dN/dS mostra a relação entre substituições não-sinônimas (dN), ou seja que modificam a proteína final codificada pelo gene, e substituições sinônimas (dS), ou seja, que não modificam a proteína produzida. Valores muito próximo a 1 de dN/dS indicam que não há seleção agindo sobre o gene, dado que de forma geral o número de substituições sinônimas e não-sinônimas são iguais, indicando a provável ausência de forças seletivas na manutenção ou mudança do gene. Valores menores que 1 indicam seleção negativa, ou seja, que a manutenção da proteína codificada esta sendo selecionada, o que é indicado pelo fato de que existem mais substituições sinônimas que não-sinônimas. Por fim, valores maiores que 1 indicam que genes que codificam proteínas diferentes da parental estão sendo selecionados, ou seja, indica que há seleção positiva.

O *alpha* tem um cálculo um pouco mais elaborado e preciso. As substituições, como

o nome sugere, estão fixadas na espécie, e por isso só mostram as mudanças que aconteceram entre espécies. Os polimorfismos, por outro lado, ainda estão presentes na espécie em diferentes formas (ou alelos). O valor de *alpha* procura relacionar essas características (as substituições e os polimorfismos), pois é esperado que populações com muitos polimorfismos possuam mais substituições, sem que isso indique obrigatoriamente que está ocorrendo seleção. Ao correlacionar os polimorfismos sinônimos e não-sinônimos e as substituições sinônimas e não-sinônimas o *alpha* proporciona um indicador mais confiável de seleção positiva. Independente da forma como a seleção foi identificada, diversos genes novos possuem esse tipo de assinatura de seleção (ROSSO et al., 2008a; LONG; LANGLEY, 1993; CAI et al., 2008; ZHANG; ZHANG; ROSENBERG, 2002; DUPRESSOIR et al., 2009; WANG et al., 2002), incluindo genes como a família *morpheus* de hominídeos (JOHNSON et al., 2001), e o gene *iris* em *D. melanogaster* (MALIK; HENIKOFF, 2005).

1.2 Espermatogênese

A maior expressão de diversos genes nos testículos levanta a questão de qual o processo biológico que ocorre nesse órgão que facilita ou estimula a expressão de genes novos. Sabemos que o processo biológico que está acontecendo nos testículos (e gametófito) é a espermatogênese, que consiste na formação do espermatozóide (ou anterozóide/grão de pólen). Em diversos trabalhos podemos ver uma alta expressão de genes novos na espermatogênese, principalmente nas fases finais da espermatogênese (SOUMILLON et al., 2013; CUI et al., 2015). Os espermatozóides competem diretamente pela fecundação do óvulo, o que cria então um ambiente no qual um gene novo vantajoso pode ser rapidamente selecionado se conferir vantagem ao espermatozóide que o possui.

A gametogênese e em especial a espermatogênese pode ser dividida em três fases principais, como mostrado na Figura 9. A primeira fase é a mitótica, na qual as espermatogônias passam por divisões mitóticas, para aumentar o número de células precursoras de espermatozóides. Em tal fase as células são arredondadas e grandes. Na fase meiótica ocorrem as duas divisões da meiose dos espermatócitos (Meiose I e Meiose II), formando as espermátides. Nessa fase há a diminuição da ploidia das células, que se tornam haplóides e menores. Por fim, a última fase é a pós-meiótica, na qual as células se diferenciam de espermátides para espermatozóides maduros (FULLER, 1993; LINDSEY; TOKUYASU, 1980).

Em *Drosophila* há na fase Meiótica a inativação do cromossomo sexual (Cromossomo X), como documentado em Kemkemer, Hense e Parsch (2011) e Vibranovski et al. (2012a). Esse processo cria assim um ambiente no cromossomo X que não é favorável a genes com alta expressão nos testículos. Durante a inativação meiótica do cromossomo sexual (*Meiotic Sex Chromosome Inactivation em inglês* - ou MSCI) marcadores epigenéticos

silenciam a expressão de boa parte do cromossomo X. Esse processo explica o movimento de duplicação de genes mais expressos no testículo do cromossomo X para os autossomos, como descrito em Zhang et al. (2010a) e Betrán, Thornton e Long (2002).

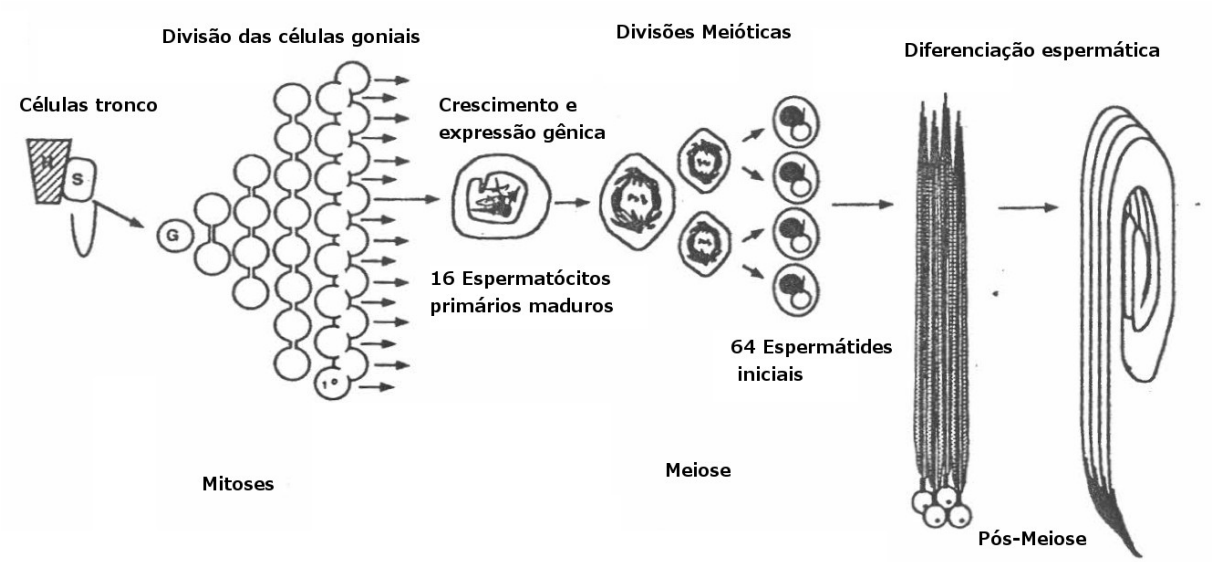


Figura 9 – **Representação dos processos e células presentes em cada uma das fases da espermatogênese de *Drosophila*.** A primeira fase (Mitótica) consiste de divisões celulares mitóticas das espermatogônias que aumentam o número de células gonadais. A fase Meiótica é onde as divisões meióticas ocorrem, tornando os espermatócitos haplóides. Por fim, na fase final (Pós-Meiótica) ocorre o alongamento e mudanças estruturais das espermátides para que se tornem espermatozóides maduros. Figura modificada e traduzida de Fuller (1993).

A expressão no testículo pode ser vantajosa para genes novos por diversas razões. Uma das razões para tal é a competição entre machos devido à competição intrasexual, por exemplo competição espermática, ou o conflito sexual (STOCKLEY, 1997; SNOOK, 2005). Outra possível razão refere-se à fertilidade, pois espermatozóides diferentes podem ter vantagens na competição espermática ou na sua habilidade de fecundar o óvulo. Por fim tais vantagens podem levar até à especiação dado que diferenças nos espermatozóides podem ser mais vantajosas ou comuns em um grupo ou população, levando à separação desses grupos (WEDELL; GAGE; PARKER, 2002; LANDRY et al., 2003). Tais adaptações estariam marcadas por mutações adaptativas e marcadores de seleção positiva.

Em 2013, Soumillon e colaboradores mostraram que durante as fases tardias da espematogênese de camundongo há um aumento na expressão de genes retrotranspostos, sejam eles funcionais ou não (retrogenes e retropseudogenes, respectivamente). A autora propõe então que o aumento na expressão de genes novos durante tais fases da espermatogênese ajude na fixação destes, e por isso a maior expressão dos mesmos nos testículos seria tão comum. A explicação dada pela autora e colaboradores para essa maior expressão

é que há uma maior permissividade da cromatina durante a espermatogênese permitindo o aparecimento e fixação de novos genes, o que ocorreria, portanto, devido ao acaso 2013.

O argumento dado por Soumillon et al. (2013), não explica completamente a fixação de genes novos. Embora a maior permissividade da cromatina de fato possa levar a uma maior expressão de genes nessa fase, a fixação dos mesmos só pode ocorrer por seleção ou deriva. A hipótese mutacional de Soumillon et al. (2013) não explica alta expressão e presença de retrogenes funcionais, o que fortalece a ideia de que a seleção também tem um forte papel na fixação de novos genes durante a espermatogênese, principalmente devido ao fato de os retrogenes funcionais parecem ter expressão mais intensa que os retropseudogenes, como podemos ver na Figura 10.

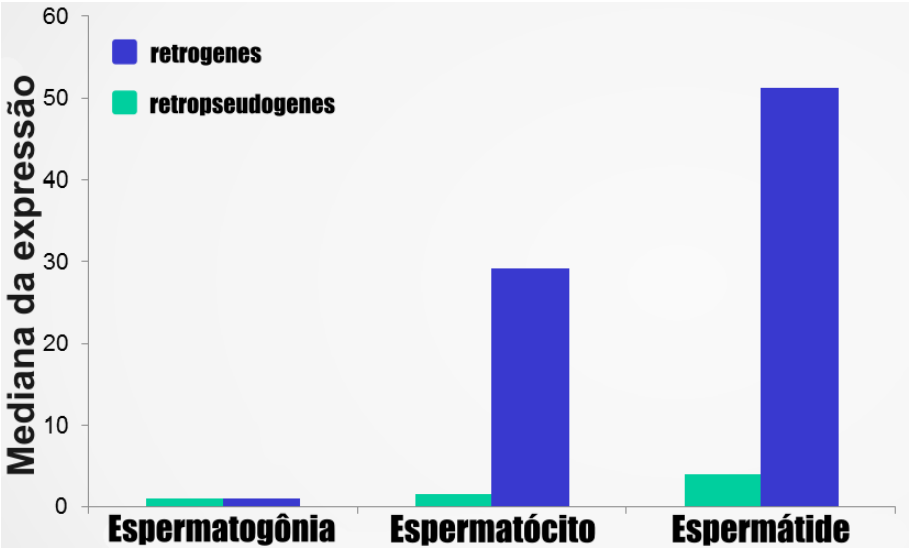


Figura 10 – **Expressão de retrogenes e retropseudogenes nas fases da espermatogênese de camundongo.** Ambos retrogenes funcionais e retropseudogenes tem um aumento de expressão nas fases tardias da espermatogênese. No entanto, esse aumento parece ser muito mais intenso para os retrogenes funcionais que para os retropseudogenes. Tal padrão argumenta a favor da hipótese de que há vantagens seletivas para genes com maior expressão nas fases haplóides, e que este aumento não ocorre somente devido a características da cromatina mais aberta durante as fases tardias da espermatogênese. Figura produzida a partir dos dados de mediana de expressão de Retrogenes e Retropseudogenes de Soumillon et al. (2013).

Sabemos que a expressão de genes novos nas fases tardias da espermatogênese não é exclusiva de camundongos, sendo também encontrada em plantas, como *Arabidopsis* e *Oriza sativa* (CUI et al., 2015). Em *Drosophila* assim como observado inicialmente em plantas e mamíferos, genes novos são altamente expressos em testículos. Se for o caso de também em *Drosophila*, os genes novos serem mais expressos nas fases tardias da espermatogênese, tal evidência aponta contra a hipótese exclusivamente mutacional, pois o tamanho populacional efetivo de *Drosophila* é muito maior que o de camundongos (SELLA

et al., 2009; PETIT; BARBADILLA, 2008; SHAPIRO et al., 2007; BUTLER, 1980) para que a deriva tenha uma ação significativa. Além disso, outros trabalhos mostram que o genoma de tais animais é muito compacto, possuindo poucos pseudogenes, apontando para uma expressão mais ligada à função e menos ligada à permissividade da cromatina durante a espermatogênese (PETROV, 2002). Assim, a hipótese de que genes novos são mais expressos nos testículos devido a questões mutacionais que facilitam a expressão gênica —como a cromatina mais aberta e permissiva, facilitando a transcrição —pode se aliar à hipótese adaptativa: uma vez que se os genes são transcritos e tem produtos benéficos eles podem ser selecionados.

Partindo da premissa de que há processos mutacionais, mas que a seleção natural tem papel importante sobre a origem e evolução dos genes novos e sua expressão gênica durante a espermatogênese esse trabalho foi desenvolvido. A espermatogênese de *Drosophila melanogaster* pode ser dividida em três fases principais: mitose, meiose e pós-meiose. As fases tardias da espermatogênese englobam a meiose e pós-meiose consideradas fases parcialmente ou totalmente haplóides, respectivamente, pela constituição cromossômica de suas células. Neste trabalho, elaboramos um modelo de seleção haplóide de genes novos. Nesse modelo propomos que ao serem mais expressos durante as fases tardias da espermatogênese os genes novos são expostos à seleção em um ambiente haplóide, no qual a fixação de alelos recessivos é mais rápida. Assim genes novos que sejam mais expressos nas fases haplóides dos machos seriam selecionados —e provavelmente fixados —mais rapidamente. A competição espermática também é de grande importância aqui, pois diversos espermatozóides competem pela fertilização do óvulo, e nesse ambiente pode haver a fixação destes novos genes. Para testar este modelo, elaboramos e testamos suas principais predições através dos dados de expressão, idade evolutiva e assinatura de seleção dos genes novos de *Drosophila melanogaster*.

2 Objetivos

Visto a grande diversidade de hipóteses e trabalhos existentes que versam sobre os mecanismos e funções da maior expressão gênica na espermatogênese - principalmente em suas fases finais - é interessante um trabalho que reúna novas análises e perspectivas sobre as questões relacionadas a esse tema afim de construir uma nova e mais abrangente perspectiva sobre o tema.

Assim, o grande objetivo deste trabalho é verificar a importância da seleção de genes novos para a fixação dos mesmos durante a espermatogênese em *Drosophila melanogaster* e verificar a expressão e seleção de tais genes. Para tal, os objetivos específicos são:

- Propor modelo de seleção haplóide para genes novos.
- Elaborar as predições do modelo.
- Testar as predições do modelo utilizando dados de expressão, idade evolutiva e assinatura de seleção dos genes novos de *Drosophila melanogaster*.
- Em colaboração com o Professor Dr. Paulo Otto formalizar matematicamente a viabilidade do modelo.

3 Materiais & Métodos

Neste trabalho os dados utilizados foram obtidos de artigos já publicados, e com dados públicos. Aqui descrevo como os dados que utilizo foram obtidos dos artigos originais e quais modificações e análises fiz para utilizá-los neste projeto. Assim, boa parte do que será aqui apresentado não foi criado por mim, exceto onde explicitado.

3.1 Expressão gênica na espermatogênese

Vibranovski et al. (2009) obtiveram a expressão em cada fase da espermatogênese dissecando testículos de *D. melanogaster* e obtendo partes enriquecidas com células de cada uma dessas fases. Nos testículos de *Drosophila melanogaster* há a separação espacial do processo temporal que é a espermatogênese, Figura 11. A parte apical do testículo é enriquecida com células da fase mitótica, contendo principalmente células arrendondadas e pequenas. A parte medial é rica em células na fase meiótica, representada por células maiores e arredondadas. E a parte distal possui uma maior quantidade de células da fase pós-meiótica, que são grandes e alongadas (VIBRANOVSKI et al., 2009). Para obter amostras enriquecidas com células de cada região do testículo (e fase da espermatogênese análoga) os testículos foram dissecados, conforme a Figura 11, em três partes: uma para cada fase da espermatogênese. Todos os testículos foram obtidos de machos de *Drosophila melanogaster* coletadas no Arizona, EUA, que são livres de *Wolbachia*. Para todas as fases, os testículos foram dissecados em PBS, e a vesícula seminal foi retirada. Para a obtenção da fase mitótica as células apicais do testículo foram isoladas com o auxilio de alfinetes entomológicos para realizar o corte. Para obtenção da fase pós-meiótica, o testículo foi cortado com os alfinetes na parte distal, obtendo-se assim a maior parte do mesmo, com a p̈arte em espiralm̈ostrada na Figura 11. Por fim, para a obtenção da fase meiótica as células da parte proximal foram obtidas da seguinte forma: a parte distal foi retirada para diminuir a pressão do testículo; então a parte apical foi retirada e as células foram espremidas para fora ao se aplicar pressão na direção posterior-anterior. Para cada testículo apenas uma das regiões foi utilizada para evitar contaminação e para que as amostras fossem estatisticamente independentes. Para cada região de 250 a 500 testículos foram dissecados.

Para obter a expressão em cada fase, o RNA foi isolado das amostras com o kit Pico Pure ™(Arcturus). O RNA de três réplicas biológicas foi hibridado ao chip de microarranjos de genes da Affymetrix *Drosophila* Genome 2.0. Na Figura 12 podemos ver um esquema simplificado de como os dados foram obtidos. Uma vez obtidos os dados de expressão, cada gene teve a expressão comparada entre as fases (entre a fase mitótica e

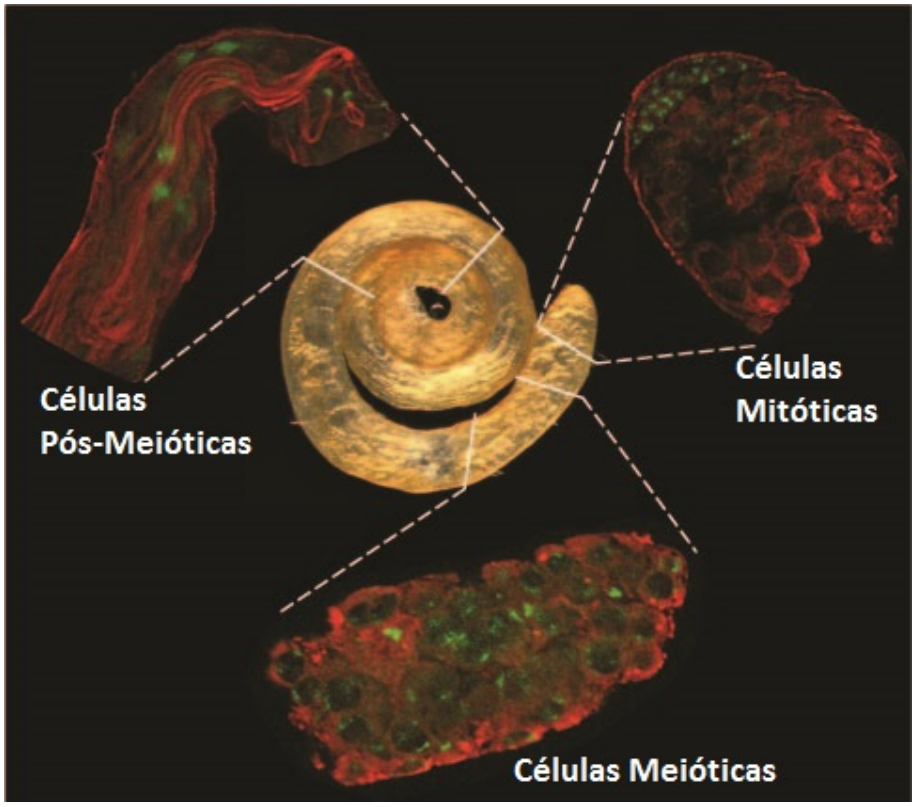


Figura 11 – **Testículo de *Drosophila melanogaster*.** Testículo de *Drosophila melanogaster* no qual podemos ver a separação física das três fases da espermatogênese. Na parte apical temos um enriquecimento de células da fase mitótica; a parte medial é enriquecida com células meióticas; e por fim, na parte distal se encontram mais células pós-meióticas. Em vermelho podemos ver o citoplasma, e em verde o DNA das células de cada região. Imagem modificada e traduzida de Vibranovski et al. (2009).

meiótica; meiótica e pós-meiótica; e entre mitótica e pós-meiótica). Nessas comparações os genes foram categorizados como sub-expressos (se a expressão na primeira fase era menor que na segunda), super-expressos (se a expressão era maior na segunda fase) e igualmente expressos (no caso da expressão ser igual nas duas fases). Usando essas categorias de genes foi possível criar classes de expressão representando todos os genes e como se dá sua expressão ao longo da espermatogênese. Na Figura 13 podemos ver as classes possíveis que foram criadas a partir da comparação da expressão obtida para cada gene nas diferentes fases da espermatogênese.

As classes foram criadas levando em conta a comparação da expressão entre as fases. Como podemos ver na Figura 14, as comparações entre as fases são independentes. Assim é possível que classes tenham comparações impossíveis, por exemplo: um gene super-expresso na meiose em relação à mitose; super-expresso na pós-meiose em relação à meiose; e sub-expresso na pós-meiose em relação à mitose. Essas classes são praticamente impossíveis: no exemplo seria necessário que a expressão na mitose fosse ao mesmo tempo

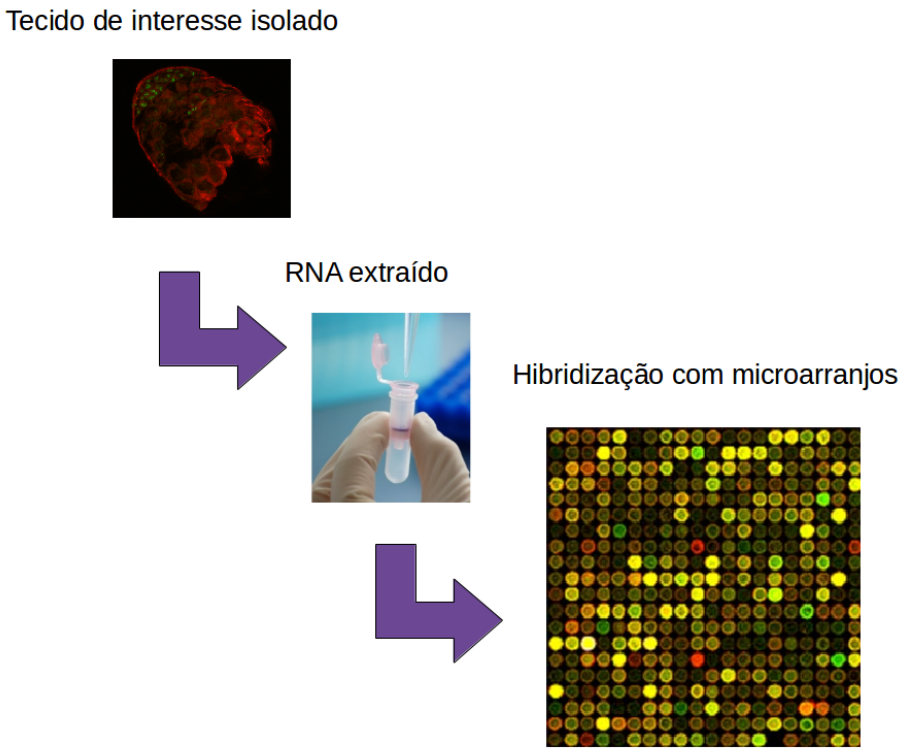


Figura 12 – **Métodos de Vibranovski et al. (2009).** Esquema dos métodos utilizados por Vibranovski et al. (2009). Após o isolamento do tecido de interesse, o RNA do mesmo foi extraído e hibridado à placa de microarranjos para se obter a expressão gênica total em cada fase da espermatogênese.

maior e menor que na pós-meiose. Por essa razão essas classes não foram consideradas para análise. Com as classes possíveis criei os grupos de genes mitóticos - genes que possuem maior expressão na mitose em comparação à meiose e pós-meiose (classes 6, 8 e 10) -, genes meióticos - que possuem maior expressão na meiose comparada às outras duas classes (classes 4, 7 e 12) -, ou genes pós-meiótico - com maior expressão na pós-meiose em relação a mitose e meiose (classes 1, 2 e 3) -, como assinalado na Figura 13. A classe 13 representa genes cuja expressão não se altera significativamente entre as diferentes fases da espermatogênese, independente de se essa expressão é alta ou baixa. Muitos genes dessa classe representam *house keeping genes*, ou seja genes envolvidos nas funções essenciais da célula e ciclo celular. Como podemos ver na Tabela 1, há poucos genes novos na classe 13, o que é esperado para *house keeping genes*, pois tais genes devem ser conservados para executarem suas funções sem adversidades. As classes não mostradas na Figura 13, numeradas de 14 a 19, são aquelas praticamente impossíveis de ocorrer. Essas classes não foram consideradas para análise, visto que provavelmente se devem à erros de medição. As classes 14 e 15 possuem um número considerável de genes, porém os testes realizados apenas com essas fases mostrarem que elas tem o mesmo padrão encontrado nas fases de 1 a 13, fazendo com que a exclusão daquelas não afete as conclusões obtidas neste trabalho.

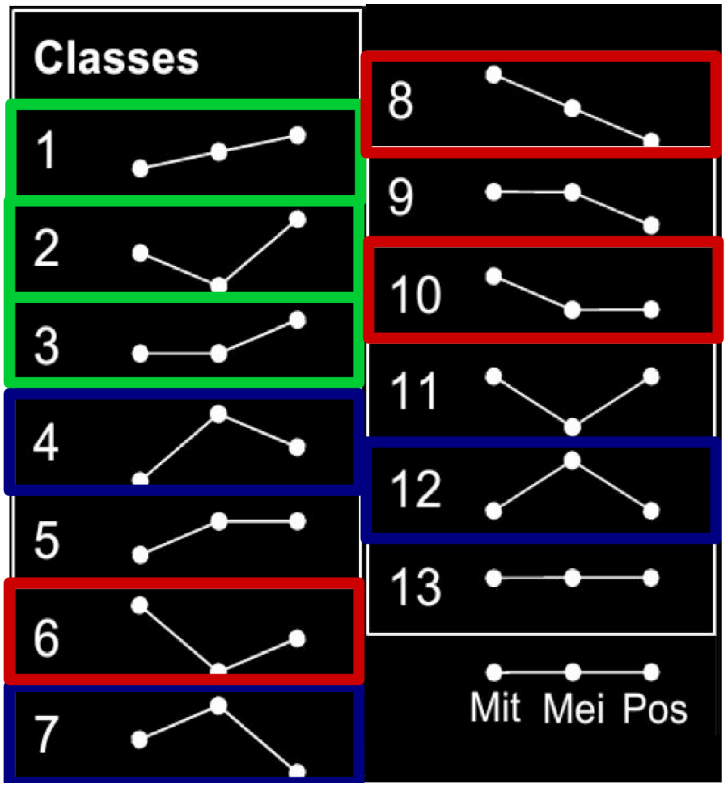


Figura 13 – **Classes de expressão durante a expermatogênese.** Podemos ver na imagem a expressão relativa comparada entre as fases da espermatogênese para 13 classes. As demais classes obtidas por Vibranovski et al. (2009) não estão representadas e não foram utilizadas para as análises. Podemos ver em verde as classes correspondentes ao grupo de genes pós-meióticos. Em azul os genes meióticos. E em vermelho os genes mitóticos.

Como poucos genes estão presentes nas outras classes impossíveis essa exclusão não é relevante para o trabalho.

É importante ressaltar algumas características das três fases da espermatogênese que auxiliarão na interpretação dos resultados obtidos. A fase mitótica (correspondente a primeira fase da espermatogênese) é aquela na qual ocorrem divisões mitóticas das espermatogônias que permanecem diplóides ao longo de toda essa fase, e ao final de tais divisões os espermatócitos são formados. A fase meiótica é a segunda fase da espermatogênese em nossa divisão e é onde ocorrem as divisões meióticas dos espermatócitos, sendo que os espermatócitos primários sofrem a primeira divisão meiótica (meiose I, na qual são separados os cromossomos homólogos) e já então se tornam haplóides, pois não possuem mais cromossomos homólogos, apenas um de cada cromossomo duplicado. Ao final da primeira divisão meiótica são formados os espermatócitos secundários, que então passam pela segunda divisão meiótica (meiose II, na qual as cromátides irmãs são separadas), e formam as espermátides. Por fim, na fase pós-meiótica as células já sofreram todas as divisões meióticas - e são portanto haplóides - e sofrem o processo de diferenciação celular,

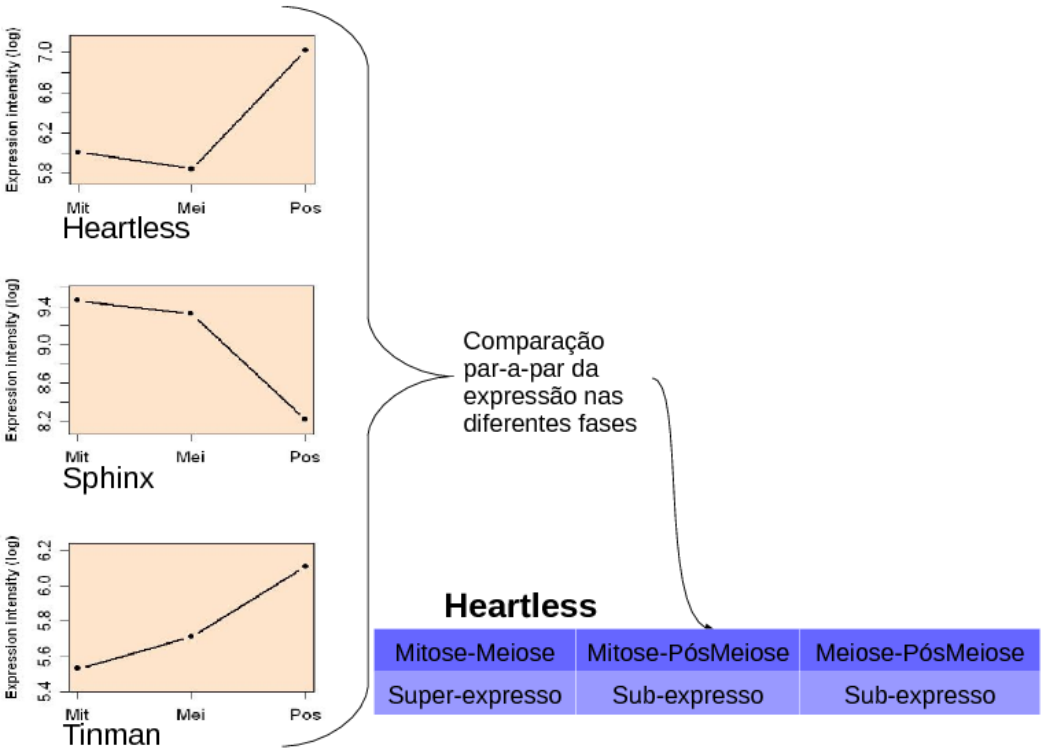


Figura 14 – **Expressão fase-a-fase ao longo da espermatogênese.**. Exemplo de como é determinada a relação da expressão entre duas fases da espermatogênese para um gene. À esquerda na imagem temos três exemplos de genes e suas expressões em cada fase. Na tabela a direita podemos ver as classificações das comparações par-a-par para um gene (*Heartless*), e como isso se reflete no comportamento da expressão do gene. Imagem modificada de Vibranovski et al. (2009).

no qual as espermátides formadas ao fim da meiose se diferenciam em espermatozóides maduros.

3.2 Idade dos genes

Para agrupar os genes como novos ou antigos usamos os dados de Zhang et al. (2010a). No trabalho os autores usam comparações entre os genomas de espécies de *Drosophila* para verificar a idade dos genes. Para verificar a idade de cada gene os autores cruzaram os dados genômicos de *Drosophila melanogaster* com outras espécies de drosofilídeos dos subgêneros *Sophophora* e *Drosophila*, a fim de acessar o ancestral comum mais antigo que provavelmente possuía esse gene (ZHANG et al., 2010a). Assim, os autores conseguiram identificar as idades dos genes. Um exemplo da análise de sintenia feita pelos autores está na Figura 15. Com esse método foi possível identificar a idade dos genes de *Drosophila melanogaster*, como podemos ver na Figura 16. Para as análises realizadas

Classe	Grupo	Autossômicos		Ligados ao X	
		Novos	Antigos	Novos	Antigos
6	Mitótica	2	177	0	48
8		45	2239	20	379
10		18	670	2	166
9	Mitótica-Meiótica	19	527	7	93
4	Meiótica	21	111	5	11
7		64	441	19	65
12		34	314	4	44
5	Meiótica-PósMeiótica	80	525	16	83
1	Pós-Meiótica	57	523	8	76
2		16	428	1	89
3		11	319	0	78
13	Constante	245	2155	21	466
11	O V	4	315	0	94
14	Impossíveis	29	292	6	40
15		10	201	1	49
16		3	38	0	16
17		2	33	0	5
18		4	43	0	4
19		0	55	0	5

Tabela 1 – **Genes novos e antigos nas classes de Vibranovski et al. (2009).** Números de genes novos e antigos, ligados ao X ou autossômicos em cada classe numérica. As classes estão organizadas em Grupos, estes foram usados para as diferentes análises mostradas. Podemos ver que as classes que formam o grupo "Impossíveis"possuem poucos genes, e por isso não são essenciais para a análise do comportamento de genes na espermatogênese.

nesse trabalho consideramos como genes novos todos os genes com menos de 63 milhões de anos, e como genes antigos os com mais de 63 milhões de anos, ou seja genes que estão presentes tanto em representantes do subgênero *Sophophora* como do subgênero *Drosophila*. Com essa separação é possível obter um número suficiente de genes autossômicos novos e antigos.

3.3 Marcadores de Seleção

Ao comparar a sequência gênica de cada gene entre diferentes espécies, no caso *D. melanogaster* e *D. simulans*, Zhang et al. (2010a) conseguiram identificar e quantificar as substituições sinônimas (que codificam o mesmo aminoácido) e não-sinônimas (que codificam aminoácidos diferentes) entre tais espécies. Essas medidas são identificadas por dS e dN, respectivamente. Com essas medidas pude calcular o dN/dS de cada gene, ou

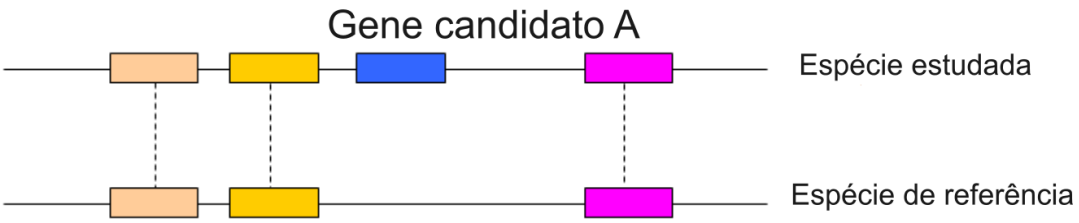


Figura 15 – **Análise de sintenia.** Para verificar a presença de um gene em um grupo ou espécie Zhang et al. (2010a) identificaram a presença ou ausência de cada gene na mesma posição cromossômica em *Drosophila melanogaster* e outras espécies do grupo. A presença do gene indica que ele surgiu antes da separação de *D. melanogaster* e o outro grupo analisado. Já a ausência do mesmo indica que sua origem é posterior a tal separação. Figura reproduzida e traduzida de Zhang et al. (2010a).

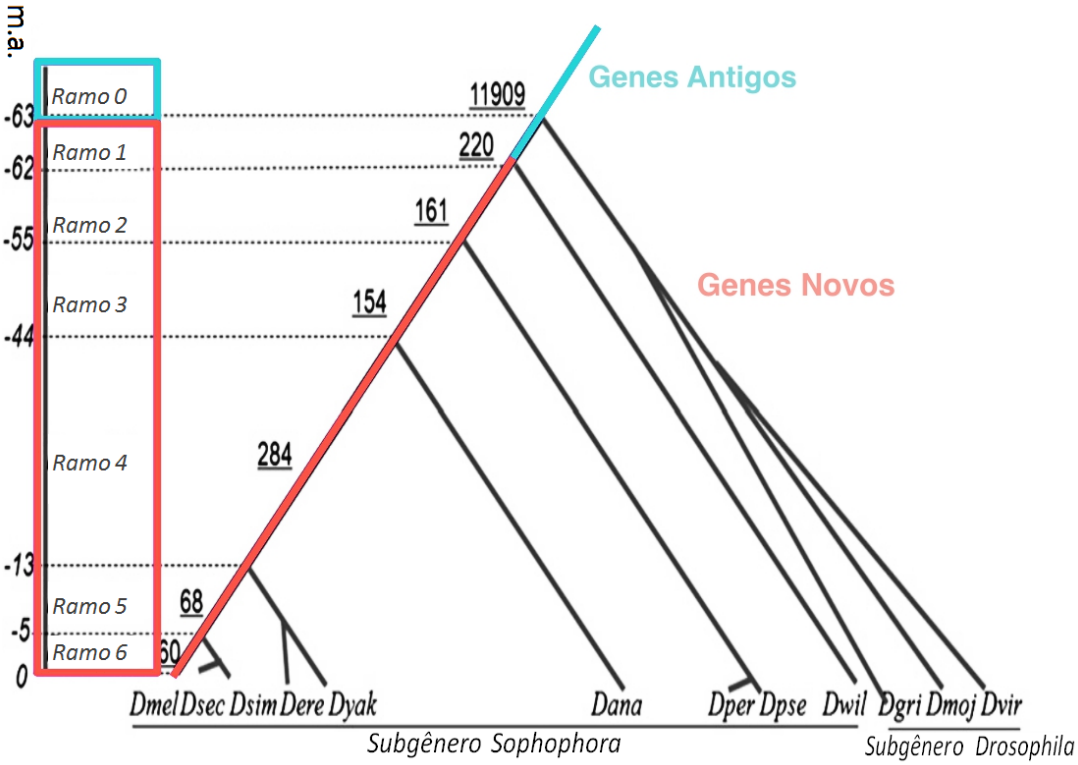


Figura 16 – **Filogenia de *Drosophila* e o surgimento de genes novos ao longo da mesma.** Ramos da filogenia de *Drosophila melanogaster*, suas idades no eixo y em milhões de anos (m.a.), e o número de genes originados em cada ramo (sublinhados próximos ao nó ao qual se referem). Genes novos são comuns ao longo da filogenia de *Drosophila*, dado que mesmo em intervalos de tempo relativamente curtos (como os 5 milhões de anos do Ramo 6) podem surgir diversos genes (no caso, 60 novos genes surgiram no Ramo 6). Figura modificada e traduzida de Zhang et al. (2010a).

seja a relação das substituições sinônimas e não sinônimas.

Porém, o dN/dS não é o método mais adequado para se identificar as assinaturas

de seleção positiva, pois pode ser afetado por características da história evolutiva do grupo, como interações epistáticas, seleção dependente de frequência, vantagem do heterozigoto, variação nas pressões seletivas e efeito carona (para mais sobre o tema verificar Sella et al. (2009), Eyre-Walker (2006), Stanley e Kulathinal (2016) e Kryazhimskiy e Plotkin (2008)). Assim utilizamos também os dados disponíveis em Stanley e Kulathinal (2016). No banco de dados FlyDIVaS (STANLEY; KULATHINAL, 2016)estão disponíveis dados atualizados de ortologia, e seleção de genes em *Drosophila melanogaster*. Nele são apresentadas comparações entre diferentes modelos de seleção, como podemos ver na Tabela 7. Utilizamos tais comparações para identificar genes positivamente selecionados.

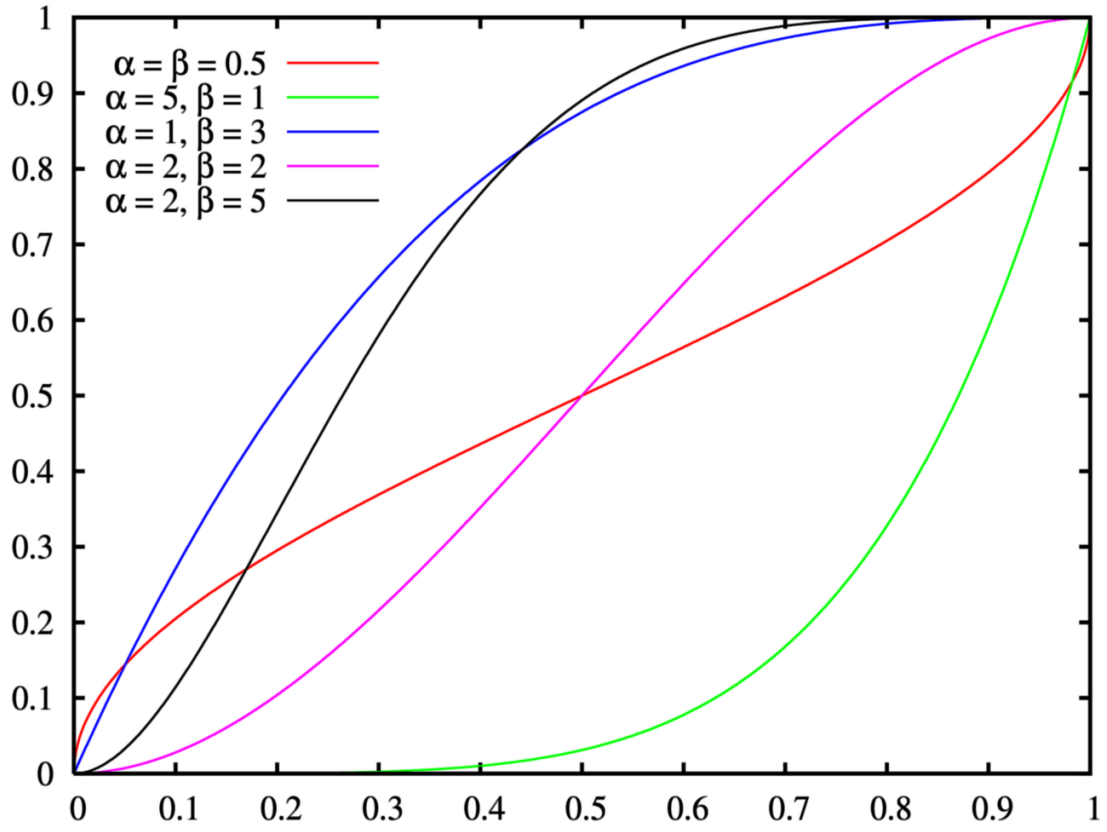


Figura 17 – **Distribuição** β**.** Diferentes curvas possíveis para uma distribuição β com diferentes valores para suas variáveis a e b. Uma função β é dada pela fórmula: $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$.

A primeira comparação contrapõe um modelo de evolução neutra (M1a) e um modelo de seleção positiva (M2a); a segunda comparação é de um modelo em que ω possui uma distribuição β (Figura 17) ao longo dos loci (M7) *versus* um modelo de distribuição beta com ação de seleção positiva possível, $\beta + \omega > 1$ (M8); a última comparação é de um modelo de $\beta + \omega > 1$ (M8) *versus* o modelo de $\beta + \omega = 1$ (M8a). Tais modelos tiveram os valores de dN, dS, e ω ($\omega = dN/dS$) calculados pelo programa PALM (YANG, 1997). Assim, o modelo neutro M1 (proposto inicialmente por Nielsen e Yang (1998)) considera todos os *loci* com duas opções de valores de ω: ou $\omega = 0$, ou $\omega = 1$. O modelo com seleção

M2 (NIELSEN; YANG, 1998) considera também que os *loci* possam ter $\omega > 1$, adicionando assim uma categoria para *loci* com seleção positiva. Para comparar os modelos os autores usam o *Likelihood-Ratio Test*, ou teste da taxa de verossimilhança (LRT), de forma a verificar qual dos modelos se adéqua melhor aos dados. Assim, se é possível rejeitar o modelo M1 em favor de M2 há seleção positiva no *locus*. No modelo nulo M7 se assume que ω tem uma distribuição β entre os *loci*, e o modelo com seleção M8, que assume que ω possui uma distribuição β entre os *loci*, mas adiciona a possibilidade de alguns *loci* terem seleção positiva ao adiciona ω a essa distribuição ($\beta + \omega > 1$) (SWANSON; NIELSEN; YANG, 2003; WONG et al., 2004). Por fim, o modelo neutro M8a estabelece um modelo com distribuição β em que o valor de ω para os *loci* é sempre 1, de forma a não levar em conta possíveis *loci* sob seleção positiva (SWANSON; NIELSEN; YANG, 2003; WONG et al., 2004). A discussão extensa dos modelos assim como a derivação de suas fórmulas podem ser encontradas em Yang (1997), Swanson, Nielsen e Yang (2003) e Wong et al. (2004).

3.4 Análises realizadas

Com os dados obtidos dos artigos Vibranovski et al. (2009) e Zhang et al. (2010a) foi possível criar programas em Perl e R para cruzar as tabelas e fazer as comparações estatísticas apresentadas neste trabalho, programas apresentados como apêndices. As tabelas complementares de Vibranovski et al. (2009) e Zhang et al. (2010a) foram cruzadas a partir da identidade única de cada gene, de forma que foi possível compilar dados sobre idade, localização cromossômica, marcadores de seleção e expressão para cada gene individualmente.

Utilizei as tabelas disponibilizadas por Zhang et al. (2010a) nos materiais suplementares (mostradas parcialmente nas Tabelas 5 e 4) e por Vibranovski et al. (2009) (mostrada parcialmente na Tabela 6) e as cruzei de acordo com o identificador do gene (na coluna 1 da Tabela Suplementar 1 de Zhang et al. (2010a) - reproduzida parcialmente na Tabela 4, coluna 14 da Tabela Suplementar 4b de Zhang et al. (2010a) - reproduzida parcialmente na Tabela 5, e coluna 4 da Tabela Suplementar de Vibranovski et al. (2009) - reproduzida parcialmente na Tabela 6) de forma a construir uma nova tabela com informações sobre a expressão, localização cromossômica, idade, e marcadores de seleção, como mostrado parcialmente na Tabela 8. Isso foi feito com um programa em Perl desenvolvido por mim, que pode ser visto no Apêndice A, e também se encontra disponível em <https://github.com/juliaraices/final_publishable>. Após esse passo, os dados da nova tabela foram analisados e os gráficos aqui apresentados foram criados com o programa em R também escrito por mim, e que pode ser visto no Apêndice B.1, e está disponível em <https://github.com/juliaraices/final_publishable>. Dessa forma foi possível fazer as aná-

	Idade	Mitóticos	Meióticos	Pós-Meióticos	Total
Autossômicos	Novos	65 (24.25%)	119 (44.40%)	84 (31.35%)	268 (100%)
	Antigos	3086 (59.1%)	866 (16.58%)	1270 (24.32%)	5222 (100%)
	Total	3151 (57.40%)	985 (17.94%)	1354 (24.66%)	5490 (100%)
Ligados ao X	Novos	22 (37.29%)	28 (47.46%)	9 (15.25%)	59 (100%)
	Antigos	593 (62.03%)	120 (12.55%)	243 (25.42%)	956 (100%)
	Total	615 (60.59%)	148 (14.58%)	252 (24.83%)	1015 (100%)

Tabela 2 – **Número de genes autossômicos e ligados ao X com alta expressão em cada fase da espermatogênese.** Genes autossômicos e ligados ao X com alta expressão em cada fase da espermatogênese.

lises e imagens disponível nesta dissertação e responder nossa pergunta sobre a veracidade da seleção de genes novos durante as fases haplóides da espermatogênese em *Drosophila melanogaster*. Com a tabela criadas por mim foi possível identificar genes de diferentes grupos, em particular os genes com maior expressão durante a fase mitótica (Mitóticos), na fase meiótica (Meióticos), na fase pós-meiótica (Pós-Meióticos), ou expressão igual em todas as fase (Constantes). Podemos ver o número de genes em alguns desses grupos, e como eles se dividem entre autossomos e cromossomo X, e entre novos e antigos nas tabelas 2 e 3.

Além dos scripts elaborados por mim, também foi utilizado o programa DoFE - Distribution of Fitness Effects, criado e mantido por Adam Eyre-Walker (EYRE-WALKER, 2010), e pode ser encontrado em <http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker/Website/Software.html>. O programa implementa os métodos de análise de fitness de diferentes artigos. Para este trabalho utilizamos o método de Bierne e Eyre-Walker (2004). O método fornecido por Bierne e Eyre-Walker (2004) possui como premissas que todas as substituições e polimorfismos sinônimos estejam sob evolução neutra, que a seleção é igual em todo o genoma, e que mutações não-sinônimas presentes na população ou espécie são neutras ou benéficas. Embora essas premissas não sejam sempre verdadeiras, elas não afetam significativamente os resultados quando isoladas, como discutido em Bierne e Eyre-Walker (2004). O modelo parte de 4 parâmetros (para cada locus) para verificar a proporção de loci com seleção positiva a partir de um modelo de verossimilhança (*likelihood*). Os parâmetros usados no modelo são os já conhecidos dN, dS, pN, pS.

	Novos	Antigos
Grupo Haplóide	283	2661
Grupo Mitótico	65	3086
Grupo Constante	245	2155

Tabela 3 – **Tamanhos amostrais dos grupos de genes utilizados.** Na tabela podemos ver a quantidade de genes novos e antigos em cada classe de grupos utilizados para as análises.

id	chrom	branch	bias	tissue number	testis value	ovary value	adj p-value
CG1903	chrX	0	female	12	1.94	3.92	6.65E-18
CG7223	chr3R	0	unbiased	11	4.39	4.14	0.3502403856
CG7895	chr3R	0	unbiased	0	1.92	1.97	0.7662113311
CG14728	chr3R	0	unbiased	0	2.3	2.83	0.0247927025
CG5354	chr2L	0	female	3	5.2	6.69	1.10E-09
CG8598	chr3L	0	female	10	6.29	9.01	9.47E-15

Tabela 4 – **Tabela parcialmente reproduzida da tabela S1 do material suplementar de Zhang et al. (2010a).** Para 6 dos 12855 genes da tabela suplementar S1 de Zhang et al. (2010a), podemos ver os seus valores para todas as características apresentadas na tabela original do trabalho. Os dados que utilizei para criação da tabela final foram: o identificador único do gene ('id'), o braço cromossômico onde o gene se encontra ('chrom'), o ramo filogenético em que o gene surgiu ('branch'), e se e qual o viés de expressão do gene ('bias').

seq number	seq length	ds	ps	dn	pn	ln	ls	fet_pvalue	g_pvalue	neutrality	alpha	name	id
4	1761	33	11	10	0	1209.5	341.5	0.101263	NaN	0	1	CG1903-RC	CG1903
5	1212	18	22	1	0	602.6	186.4	0.463415	NaN	0	1	CG7223-RA	CG7223
5	813	8.7	19.5	3.3	24.5	427.1	133.9	0.102674	0.0821	3.26667	-2.266667	CG7895-RA	CG7895
5	1491	13.5	5	7.5	3	865.1	283.9	1	0.92885	1.08	-0.08	CG14728-RA	CG14728
6	1005	18	34	7	4	703.2	211.7	0.096126	0.07684	0.302521	0.697479	CG5354-RA	CG5354
7	1002	14	11	3	3	441.3	131.7	1	0.7912	1.27273	-0.272727	CG8598-RB	CG8598

Tabela 5 – **Tabela parcialmente reproduzida da tabela S4 do material suplementar do material suplementar de Zhang et al. (2010a).** Para 6 dos 9956 genes da tabela suplementar S4b de Zhang et al. (2010a), podemos ver os seus valores para todas as características apresentadas na tabela original do trabalho. Os dados que utilizei para criação da tabela final foram: o identificador único do gene ('id'), e os valores de *dn*, *ds*, *pn*, *ps* e *alpha* (nas colunas com esses mesmos nomes).

...	Gene Symbol	CG	Chr	Mitosis	Meiosis	Post meiosis	Mitosis Postmeiosis	Meiosis Postmeiosis	Mitosis Meiosis	CLASS	...
...	sno	CG1903	arm_X	5.16307	4.75682	4.97827	Over	Equal	Over	10	...
...	htl	CG7223	arm_3R	6.01026	5.83766	7.02124	Under	Under	Over	2	...
...	tin	CG7895	arm_3R	5.53099	5.70908	6.11102	Equal	Equal	Equal	13	...
...	sad	CG14728	arm_3R	6.23704	6.54592	6.59023	Equal	Equal	Equal	13	...
...	pie	CG5354	arm_2L	9.15809	9.33837	6.69014	Over	Over	Equal	9	...
...	eco	CG8598	arm_3L	8.76638	8.72333	7951	Over	Over	Equal	9	...

Tabela 6 – **Tabela parcialmente reproduzida da Tabela Suplementar S1 de Vibranovski et al. (2009).** Para 6 dos 18081 genes da tabela suplementar de Vibranovski et al. (2009), podemos ver os valores para parte das características apresentadas na tabela original do trabalho (vemos os dados de 10 das 25 colunas da tabela original). Os dados que utilizei para criação da tabela final foram: o identificador único do gene ('CG'), o braço cromossômico em que se encontra ('Chr'), os valores de expressão do gene na mitose, meiose e pós-meiose (respectivamente as colunas 'Mitosis', 'Meiosis' e 'Post meiosis'), as comparações par-a-par da expressão entre as fases ('Mitosis Postmeiosis', 'Meiosis Postmeiosis' e 'Mitosis Meiosis'), e a classse de expressão do genes ('CLASS').

cg	lnl_1	lnl_2	pos_12	lnl_7	lnl_8	lnl_8a	pos78	pos_88a
CG10949	-2860.849301	-2859.224955	does not	-2863.279256	-2859.318134	-2859.318134	does not	does not
CG3621	-528.703523	-528.703447	does not	-528.709673	-528.709748	-528.709748	does not	does not
CG10307	-2153.123358	-2153.123358	does not	-2152.919955	-2152.919964	-2152.919964	does not	does not
CG16741	-496.522096	-496.522096	does not	-496.55519	-496.510499	-496.507450	does not	does not

Tabela 7 – **Tabela parcialmente reproduzida de Stanley e Kulathinal (2016).** Para 4 dos 9232 genes presentes na tabela obtida do banco de dados FlyDIVaS (STANLEY; KULATHINAL, 2016) para as comparações do subgrupo *melanogaster*. Podemos observar os valores de 10 das 27 colunas originais. Em tais colunas vemos o identificador do gene, e as comparações entre os diferentes modelos de seleção apresentados no trabalho: 1, 2, 7, 8 e 8a.

id	Mitosis	Meiosis	PostMeiosis	...	XorA	Class	Group	...	age	bias	dn	ds	dnds	...	alpha
CG10949	7.66262	7.49633	6.81879	...	A	8	Mitotic	...	old	female	NA	NA	NA	...	NA
CG13761	7.40785	6.67446	6.58836	...	X	10	Mitotic	...	old	female	NA	NA	NA	...	NA
CG9655	10.2577	9.44319	8.8344	...	A	8	Mitotic	...	old	female	1	1	1	...	NaN
CG3621	10.7387	10.5457	10.1274	...	X	8	Mitotic	...	old	female	1	5	0.2	...	1
CG10307	9.24145	10.7523	11.9622	...	A	1	PostMeiotic	...	old	male	5.3	17.7	$\tilde{0}.3$	...	0.558333
CG16741	10.0144	11.4761	12.3173	...	A	1	PostMeiotic	...	new	male	2	1	2	...	1

Tabela 8 – **Tabela feita com os dados de Vibranovski et al. (2009) e Zhang et al. (2010a) para análise.** Tabela obtida após cruzar as tabelas acima exemplificadas com o programa em Perl exibido no Apêndice A. Aqui podemos ver para 6 genes os valores de 13 das 28 colunas obtidas. Em tais colunas há tanto novas características determinadas a partir dos dados publicados em Vibranovski et al. (2009) e Zhang et al. (2010a) (por exemplo a coluna 'age'), como dados presentes em tais tabelas (como 'Class' e 'dnds').

4 Resultados: Modelo de Seleção Haplóide de Genes Novos

Há muito se sabe que alelos novos vantajosos são mais rapidamente fixados em populações haplóides (JOSEPH; KIRKPATRICK, 2004; LENORMAND; DUTHEIL, 2005). Isso se deve, entre outros fatores, ao *fitness* do alelo não ser mascarado por outro alelo, como no caso de alelos recessivos em indivíduos heterozigotos em sistemas diplóides. Dessa forma, em sistemas haplóides a seleção e fixação desse elemento se torna mais rápida, como podemos ver na Figura 18. Também é sabido que genes novos estão sujeitos a maior frequência de seleção positiva (LONG et al., 2003; FAY; WYCKOFF; WU, 2002; SELLA et al., 2009; EYRE-WALKER, 2006), o que os torna bons candidatos para o estudo da interação desses dois modelos.

Como vimos nos capítulos anteriores, a maior parte dos genes novos surge por duplicações baseadas em DNA ou RNA (ZHANG, 2003; ZHOU et al., 2008; CHEN; ZHANG; LONG, 2010). Por isso, tais genes novos possuem funções iguais a de genes antigos já fixados, e, em geral, devem ser perdidos devido ao acúmulo de mutações deletérias em uma das cópias (OHNO, 1972).

Genes novos - originados de duplicações ou não - são intensamente expressos nos testículos, onde há expressão gênica promíscua (BABUSHOK et al., 2007; LEVINE et al., 2006; BETRÁN; THORNTON; LONG, 2002; MARQUES et al., 2005; SOUMILLON et al., 2013; VIBRANOVSKI; ZHANG; LONG, 2009; CUI et al., 2015). Isso pode facilitar a expressão de tais genes que podem então ser fixados na população por seleção ou deriva. Aqui propomos um modelo para genes novos em que a seleção é mais importante que a deriva (ao contrário do proposto por Soumillon et al. (2013)) e que foi desenvolvido e testado ao longo do meu mestrado.

O modelo de seleção de gametas (haplóides) em populações de indivíduos diplóides parte da premissa de que alelos benéficos são mais rapidamente fixados em populações haplóides que diplóides, principalmente se forem recessivos na população diplóide (JOSEPH; KIRKPATRICK, 2004; LENORMAND; DUTHEIL, 2005), conforme podemos ver nas simulações apresentadas na Figura 18. Dessa forma, gametas seriam células em que a seleção de alelos novos vantajosos de genes novos pode ser mais rápida devido a haploidia das mesmas. Para criar esse modelo também pressupomos que a seleção será mais fácil de ser percebida nos genes novos, pois esses ainda podem estar sob regimes de seleção apesar de já terem sido fixados na espécie. Esse modelo pode ser aplicado para alelos novos e recessivos vantajosos de genes novos duplicados de seus genes parentais. Aqui trataremos majoritariamente da seleção de genes novos. A partir do modelo de seleção haplóide de

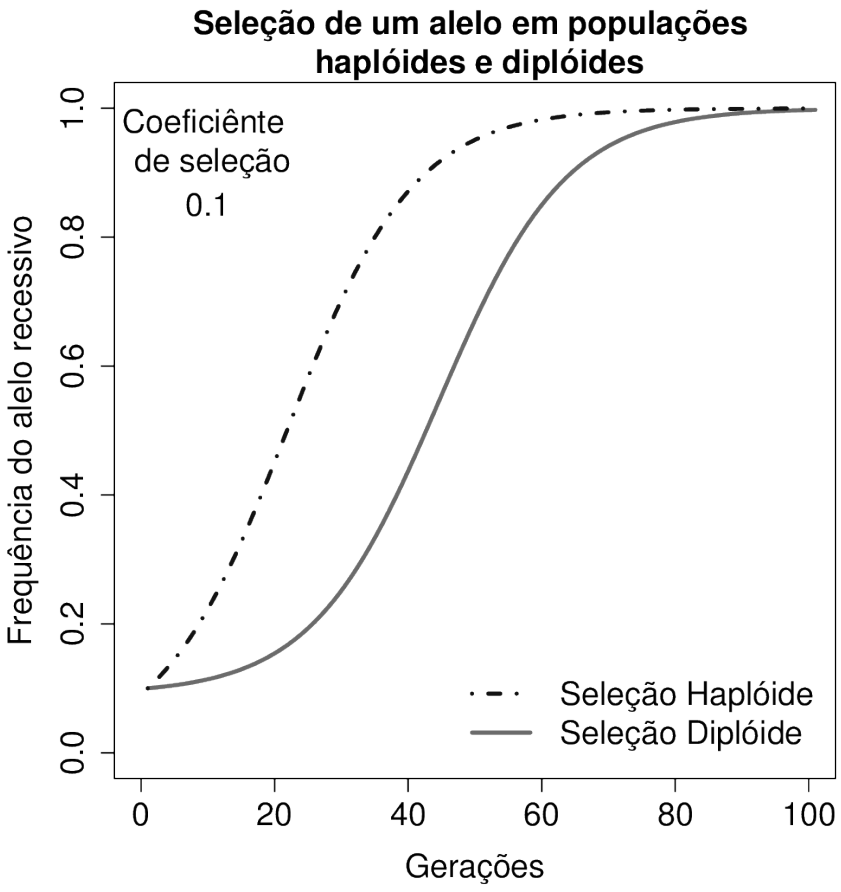


Figura 18 – **Simulação de fixação de alelo em populações haplóides e diplóides.** Simulação de fixação de um alelo recessivo vantajoso na população diplóide, e o mesmo alelo na população haplóide. Embora o coeficiente de seleção seja o mesmo nas duas populações, a fixação na população diplóide é muito mais lenta, devido a possibilidade de este alelo recessivo não estar sendo expresso e/ou selecionado em indivíduos heterozigotos. A figura foi obtida a partir de código na linguagem de programação R, que pode ser encontrado no Apêndice B.2;

genes novos pudemos construir três predições:

- Genes novos devem ser mais expressos nas fases meiótica e pós-meiótica (tardias) da espermatogênese do que na fase mitótica.
- Genes mais expressos nas fases tardias da espermatogênese devem apresentar mais assinaturas de seleção positiva.
- Apenas genes autossômicos devem apresentar tal aumento de expressão nas fases tardias da espermatogênese.

A primeira predição afirma que devemos encontrar um maior aumento de expressão dos genes novos nas fases tardias (meióticas e pós-meióticas) da espermatogênese. Nosso

modelo assume que alelos vantajosos de genes novos são fixados mais rapidamente em sistemas haplóides de seleção. Desta forma, para que um alelo esteja sob seleção haplóide, é necessário que ele seja expresso na fase haplóide do sistema, pois somente desta forma poderá ter impacto no *fitness* e portanto ser alvo da seleção em forma haplóide. Se alelos de genes novos são mais rapidamente fixados por causa da seleção haplóide, esperamos que uma maior proporção de genes novos sejam expressos com maior intensidade nas fases meióticas e pós-meióticas, que são haplóides, do que genes antigos. Como esses últimos, genes antigos, já se encontram fixados na população há algum tempo, e geralmente não acumulam substituições adaptativas, não há mais vantagem para tais genes serem expressos nas fases haplóides da espermatogênese, exceto se possuírem alguma função específica nesse estágio. Embora genes novos também já estejam fixados na população, esse evento ocorreu mais recentemente, por isso há mais chances de ainda possuírem características que os levaram a ser fixados. Desta forma, esperamos encontrar tanto uma maior proporção de genes novos com alta expressão nas fases haplóides em comparação à fase diplóide, como uma maior expressão de tais genes nessas fases em comparação à expressão dos genes antigos nessas mesmas fases.

Já a segunda predição indica que os genes mais expressos nas fases tardias da espermatogênese terão mais assinaturas de seleção positiva. O nosso modelo assume que somente alelos novos recessivos adaptativos se fixarão mais rápido em sistemas haplóides do que diplóides. Desta forma, devemos encontrar mais assinatura de seleção positivas em genes com maior expressão haplóide do que diplóide que corresponde as fases meiótica/pós-meióticas e mitótica da espermatogênese, respectivamente. Durante a fixação de genes novos é comum que estes possuam assinaturas de seleção positiva (LONG et al., 2003; SELLA et al., 2009; EYRE-WALKER, 2006). Assim, os genes mais expressos nas fases haplóides, apesar de já fixados, devem ter mais marcadores de seleção que os genes de outros grupos. Dessa forma, se espera em geral que genes novos possuam mais indicadores de seleção positiva, em particular para genes do grupo haplóide em relação aos genes mais expressos na fase mitótica.

Por fim, a terceira predição postula que genes novos ligados ao X não devem apresentar aumento de expressão nas fases tardias da espermatogênese. Isso decorre de os genes ligados ao X estarem sempre em hemizigose nos machos, possuindo assim uma situação similar de expressão a aquela vivenciada pelos genes autossômicos nas fases haplóides da espermatogênese (Figura 19). Desta forma, não existe diferencial de seleção em termos de haploidia entre genes ligados ao X expressos nas fases meióticas e pós-meióticas em relação as fases mitóticas.

Sabemos que há uma maior expressão de genes novos nos testículos (órgão responsável pela produção dos espermatozóides, os gametas masculinos) do que nos ovários (órgão responsável pela produção dos óvulos, os gametas femininos) (SOUMILLON et

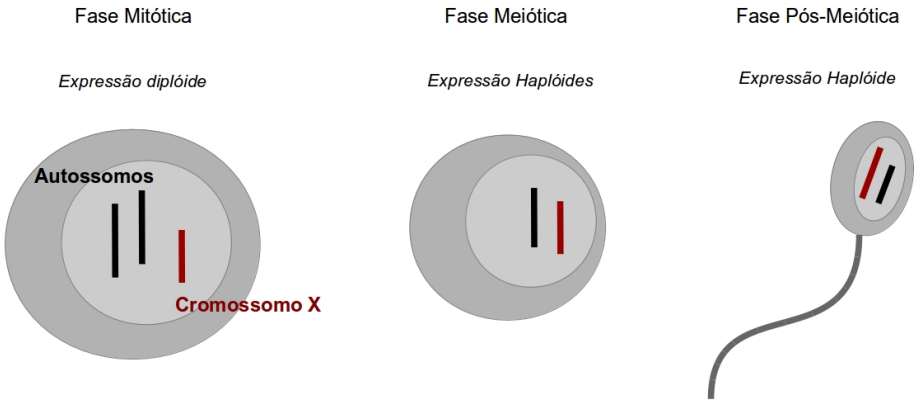


Figura 19 – **Esquema de como se dá a expressão de genes autossômicos e ligados ao X ao longo da espermatogênese.** Como podemos ver, durante a mitose (assim como nas células somáticas) os genes autossômicos possuem expressão diplóide, e os genes ligados ao X têm expressão haplóide, pois estão sempre em hemizigose. Porém, nas fases haplóides (Meiótica e Pós-Meiótica) os genes autossômicos são expressos de forma haplóide, diferente do que ocorre nas células somáticas. Porém, para os genes ligados ao X, sua expressão permanece haplóide, não havendo nenhuma vantagem para a expressão nessas fases tardias da espermatogênese.

al., 2013; VIBRANOVSKI et al., 2009). Em nosso modelo de seleção haplóide para a fixação de genes novos, também esperamos que somente a gametogênese masculina, e não a feminina, possua o sistema propício para a fixação mais rápida de alelos vantajosos. Uma das razões para tal é que durante a formação do óvulo cistos de 16 células são formados (CÁCERES; NILSON, 2005; HE; WANG; MONTELL, 2011; MONTELL, 2003), nos quais apenas uma delas se tornará um óvulo viável, as outras células nutrem o óvulo, dividindo nutrientes e RNAm. Embora nos testículos existam cistos de espermátides ligadas por pontes citoplasmáticas, a transferência de RNAm em tais casos é restrita, enquanto nos ovários as células ainda não se separaram, de forma que todo o RNAm presente no oócito é de mais de uma célula. Como a expressão e RNAm de todas as outras 15 células do cisto (KHAMMARI et al., 2011) contribuem para o *fitness* do óvulo não é possível tratar de seleção haplóide nesse caso. Podemos ver essa distinção com clareza na Figura 20

Verificamos então se as predições do modelo são encontradas em *Drosophila melanogaster*, demonstrando a validade e viabilidade do modelo proposto em um organismo modelo. Para testar tais predições foram usados dados já publicados da expressão dos genes nas diferentes fases da espermatogênese (VIBRANOVSKI et al., 2009), de idade, e de polimorfismos e substituições sinônimas e não sinônimas (pS, pN, dS e dN) dos genes de *Drosophila melanogaster* (ZHANG et al., 2010a; STANLEY; KULATHINAL, 2016).

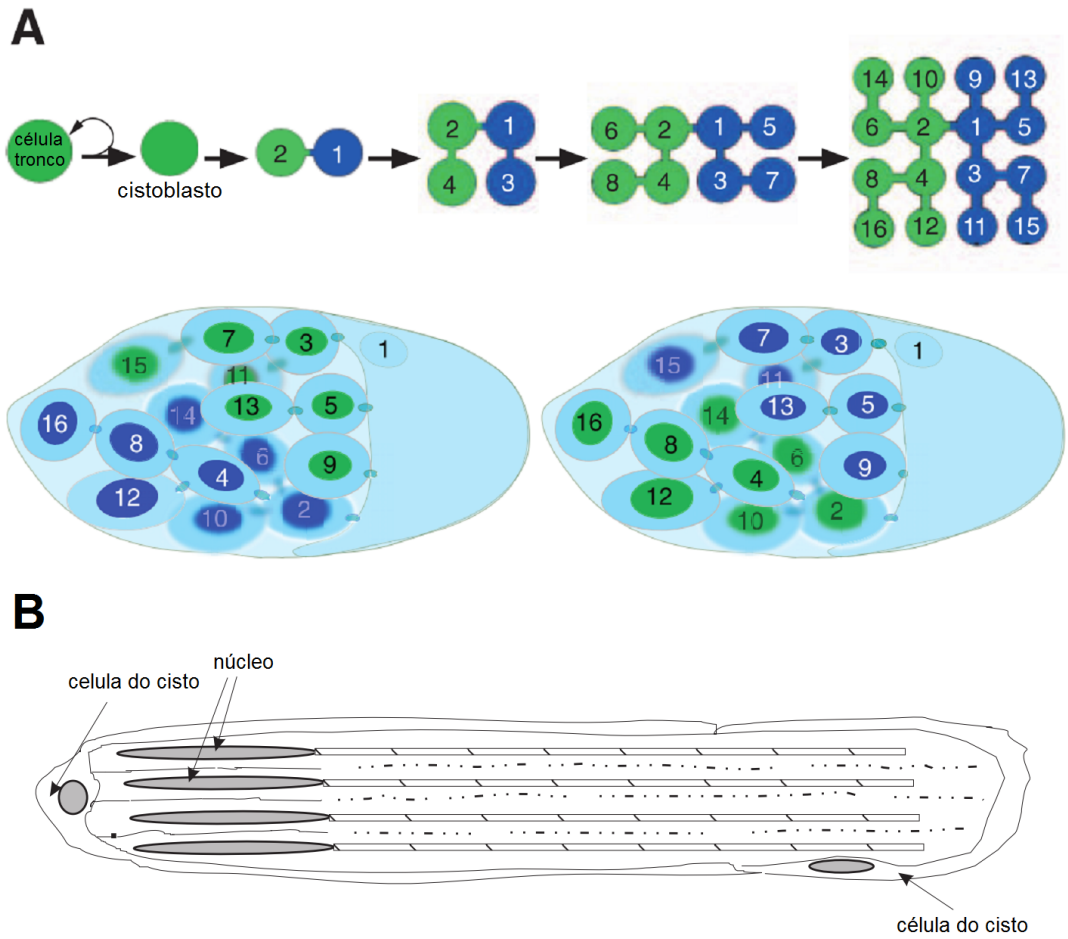


Figura 20 – **Comparação da espermatogênese com a oogênese em *Drosophila melanogaster*.** A figura mostra uma comparação da oogênese e espermatogênese em relação à individualidade de expressão das células gaméticas. A primeira parte da imagem (A) se refere à oogênese, e mostra a formação do óvulo (célula 1) e suas conexões às outras células do cisto. Neste caso todas as células ainda se comunicam amplamente, de forma que as 15 células que não formam o óvulo são responsáveis por nutrí-lo e compartilham nutrientes e RNAm. Esta imagem foi traduzida e adaptada de Cáceres e Nilson (2005). Na segunda parte da imagem (B) está representado o final da espermatogênese, quando da individualização das células espermáticas ao final da fase pós-meiótica. Podemos ver que embora inicialmente as células possuam comunicação extensa, ao final deste processo as pontes citoplasmáticas são pequenas e não permitem uma ampla distribuição de RNAm entre as células. Esta imagem foi traduzida e adaptada de Fabrizio et al. (1998)

4.1 Expressão de Genes Novos

Para testar a primeira predição do modelo - de que uma maior proporção de genes novos deve ser expressa nas fases tardias (haplóides) da espermatogênese - verificamos a proporção de genes novos e antigos com expressão mais acentuada em cada fase da

espermatogênese.

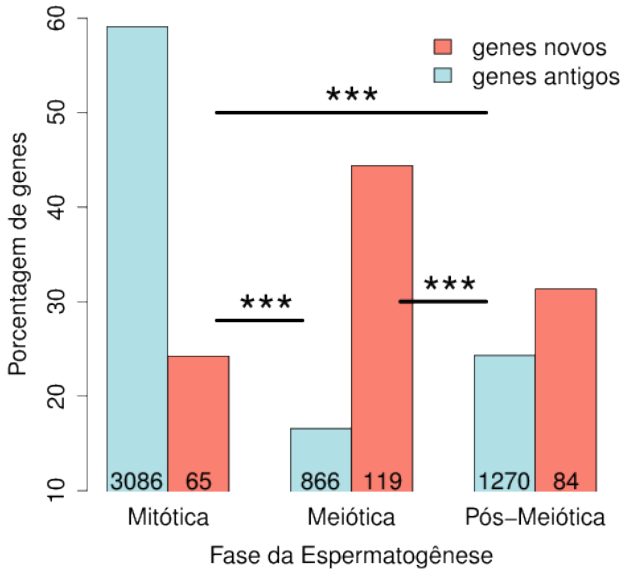


Figura 21 – **Proporção de genes novos e antigos autossômicos super-expressos em cada fase da espermatogênese.** Podemos ver a maior proporção de genes antigos (em azul) na fase mitótica em relação às outras fases, enquanto a maioria dos genes novos (em rosa) tem maior expressão nas fases meiótica e pós-meiótica, com relação à proporção de genes novos na fase mitótica. Foram feitas comparações com o teste exato de Fisher e todas comparações feitas entre os grupos de genes entre as fases tiveram p-valor menor que 0.001 (indicado por ***).

Como podemos ver na Figura 21, há uma maior proporção de genes antigos com alta expressão na fase mitótica, o que se inverte na fase meiótica e pós-meiótica, quando há uma maior proporção de genes novos com alta expressão. Essas proporções corroboram o modelo, pois são exatamente como o predito, tendo uma maior proporção de genes antigos com alta expressão durante a fase diplóide (mitose) e uma maior proporção de genes novos com alta expressão nas fases haplóides (meiose e pós-meiose). A fase haplóide é composta pelas fases Meiótica e Pós-Meiótica. Durante a fase meiótica ocorre a meiose, que torna as células haplóides, como mostrado na Figura 22. Na primeira divisão meiótica (Meiose I) os cromossomos homólogos são separados, o que torna as células efetivamente haplóides em termos de *fitness* do alelo recessivo que não está mais mascarado pelo efeito do alelo do cromossomo homólogo, apesar de possuírem os cromossomo duplicados. Por isso, é lógico considerar toda essa fase como haplóide.

Ademais, não apenas a proporção de genes novos altamente expressos nas fases haplóides é maior que a de genes antigos, a expressão *per se* de tais genes também é mais intensa nessas fases. Como podemos ver na Figura 23, durante a fase mitótica (diplóide)

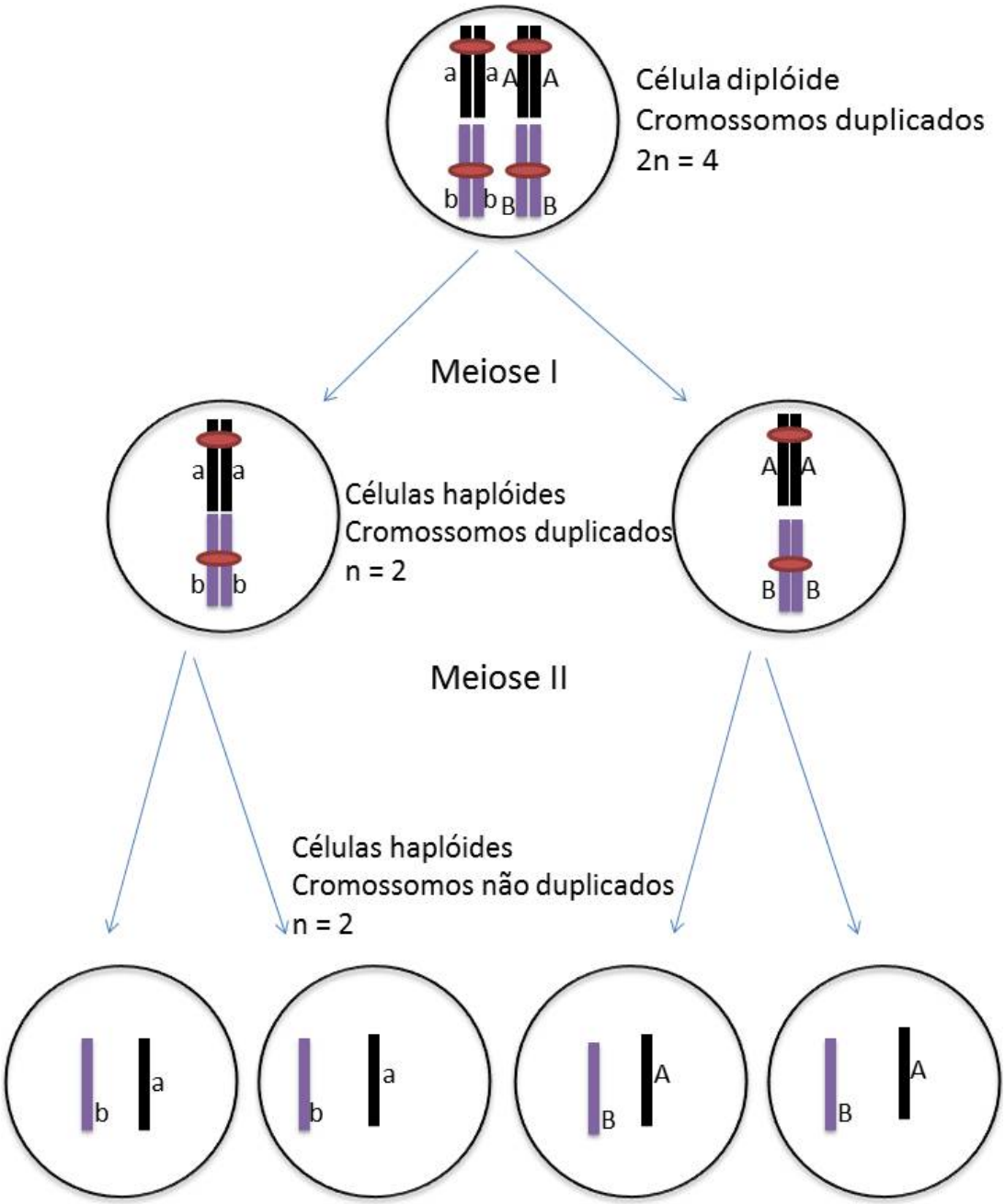


Figura 22 – **Esquema da meiose e ploidia das células nesse processo**. Esquema simplificado da meiose, mostrando a ploidia das células durante esse processo. A célula inicial é diplóide, possuindo seus cromossomos duplicados no início do processo. Ao final da Meiose I (primeira divisão da meiose), as células formadas já são haplóides, embora continuem possuindo os cromossomos duplicados. Porém, como possuem apenas uma cópia de cada cromossomo (e por tanto de cada alelo), podemos considerá-las funcionalmente haplóides.

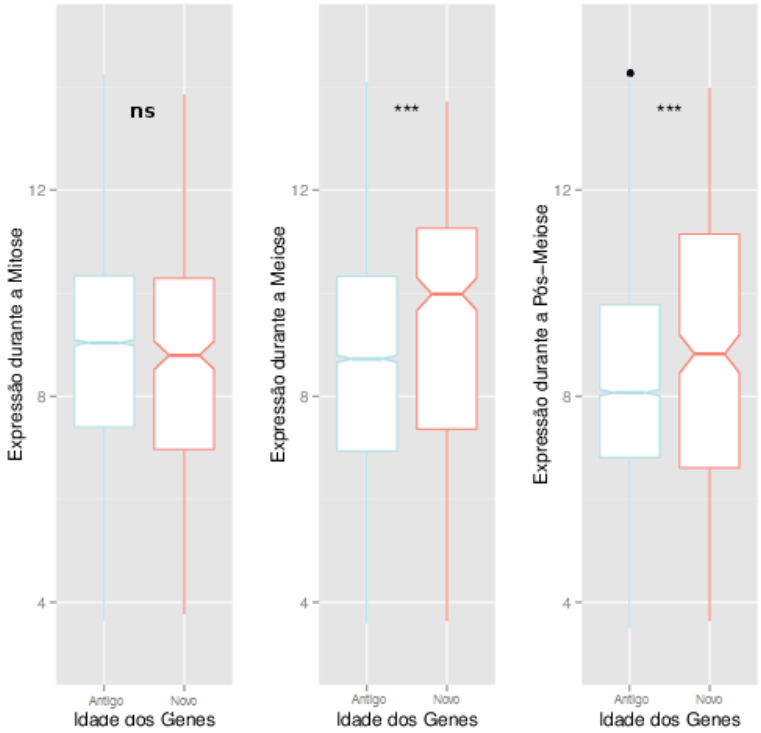


Figura 23 – **Expressão de genes autossômicos novos e antigos ao longo da espermatogênese.** Podemos ver no gráfico a expressão de genes novos (em rosa) e antigos (em azul) nas três fases da espermatogênese analisadas. Durante a fase mitótica não é possível dizer que há maior expressão de genes novos ou antigos, mas nas fases haplóides (meiose e pós-meiose) podemos ver uma maior expressão de genes novos. Para todos os gráficos há 6589 genes antigos e 371 genes novos. Para comparar as classes foi usado o teste de Wilcox-Mann-Whitney, e *** implica p-valor menor que 0.05.

não há diferença na expressão de genes antigos e genes novos, porém nas fases haplóides (meiose e pós-meiose) os genes novos tem maior expressão.

Podemos ver então que há uma maior proporção de genes novos sendo altamente expressos nas fases haplóides, e de genes antigos na fase diplóide. Também vemos que esse aumento não se dá apenas na proporção de genes, mas também em sua expressão, dado que os genes novos também são mais intensamente expressos nas fases haplóides. Com isso em mente podemos ver que de fato os genes novos parecem ter alguma vantagem ao serem expressos nas fases haplóides. Essa vantagem pode, a princípio, ser tanto por conta da seleção que sofrem nessas fases haplóides que lhes confere vantagem evolutiva, e por isso são selecionados; como pode ser devido apenas a transcrição mais promíscua nessas fases e portanto levar a uma maior expressão de todos os genes e alguns pseudogenes. Porém, como mostrarei nas próximas seções, isso é pouco provável, visto que não apenas há uma maior proporção de genes novos sendo mais expressos nessas fases como suas expressões e indicadores de seleção também são maiores para esses genes. Como tais aumentos de

expressão não ocorrem para os genes antigos é pertinente imaginar que não seja devido apenas a maior promiscuidade de expressão nessas fases.

4.2 Assinaturas de seleção

A segunda predição do modelo diz que devemos encontrar maiores assinaturas de seleção para os genes altamente expressos e que exercem função nas fases haplóides do que genes expressos e com função na fase diplóide. Para verificar isso usamos dois indicadores de seleção: o dN/dS e o *alpha*.

Pode-se ver na Figura 24a que os valores de $\frac{dN}{dS}$ para os genes novos é sempre maior que para genes antigos. Isso já é esperado e já foi visto anteriormente (FAY; WYCKOFF; WU, 2002; LONG et al., 2003), visto que genes novos sofrem mais substituições não-sinônimas que os genes antigos, pois possuem menos amarras evolutivas que evitem substituições. Por isso pode ser mais interessante observar indicadores de seleção positiva de forma mais direta, como apartir de modelos.

No entanto, esses primeiros resultados nos mostram um forte indicativo de que há uma vantagem seletiva para genes altamente expressos nas fases tardias da espermatogênese. Também podemos ver as comparações dos valores de dN/dS dos grupos haplóides, que consistem nas fases meióticas e pós-meiótica juntas, e o grupo mitótico, que consiste na fase mitótica da espermatogênese, que é diplóide. As comparações desses grupos nos mostram que caraterísticas são exclusivas dos genes com alta expressão nas fases haplóides e não na fase diplóide, o que ajuda a percebermos as vantagens dos genes que tem alta expressão naquelas fases.

Quando comparamos todos os genes do grupo mitótico aos genes do grupo haplóide (Figura 24b), vemos que os genes do grupo haplóide tem maiores valores de dN/dS, indicando uma correlação entre alta expressão nas fases haplóides e mais marcadores de seleção positiva. Porém, como explicado, valores de dN/dS não são os melhores indicadores de seleção positiva.

Uma outra possibilidade para verificar seleção, é identificar em qual grupo há uma maior proporção de genes sendo positivamente selecionados. Para isso usamos os dados do banco criado por Stanley e Kulathinal (2016). Neste banco de dados estão informados os genes que possuem ou não marcadores de seleção positiva, de acordo com três modelos. Aqui consideramos como genes com seleção positiva aqueles que fossem considerados assim por pelo menos um desses três modelos. Com isso, pudemos observar que para o conjunto de todos genes há uma diferença entre a proporção de genes novos e antigos com assinatura de seleção positiva. Como já visto em trabalhos anteriores (LONG et al., 2003; SELLA et al., 2009; EYRE-WALKER, 2006) há mais genes com assinatura de seleção positiva entre

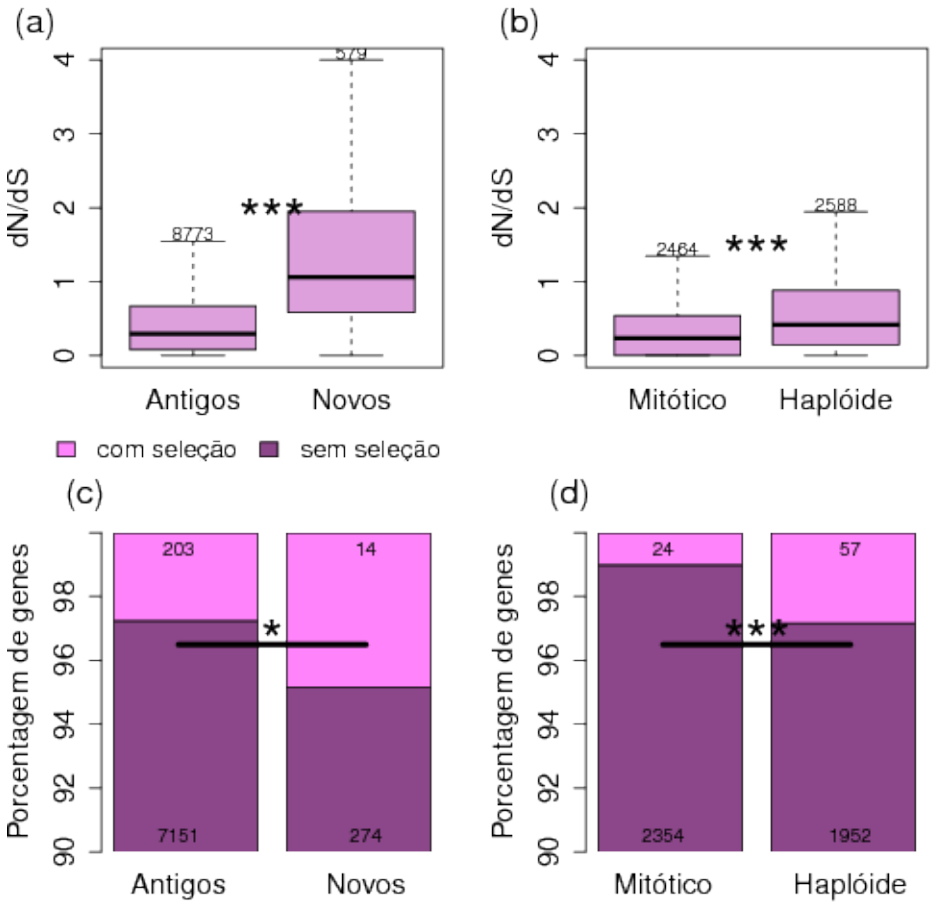


Figura 24 – **Marcadores de seleção para genes autossômicos novos e antigos e dos grupos haplóide e mitótico.** Na figura (a) podemos ver os valores de dN/dS para genes novos e antigos. Podemos ver que, conforme o esperado, genes novos possuem valores maiores de dN/dS. Quando comparamos os genes autossômicos dos grupos haplóide e mitótico entre si (independente das idades dos genes), percebemos maiores valores de dN/dS para o grupo haplóide (b). No gráfico (c) podemos ver a proporção de genes com assinatura de seleção positiva de acordo com Stanley e Kulathinal (2016). Vemos que, novamente, genes novos apresentam mais marcadores de seleção positiva, como esperado. Em (d) vemos a proporção de genes autossômicos com marcadores de seleção positiva nos grupos mitótico e haplóide. Observa-se que no grupo haplóide há mais genes com marcadores de seleção positiva que no grupo mitótico, como esperado por nosso modelo. Seria interessante verificar o comportamento dos genes novos entre os dois grupos, porém os tamanhos amostrais (indicados para cada grupo próximos ao eixo x, em (a) e (b), e próximo ao topo e ao eixo X para os grupos mostrados em (c) e (d)) não permitem análises estatísticas significativas. (***) indica p-valor <= 0.001, (**) indica p-valor <= 0.01, (*) indica p-valor <=0.05.

os genes novos (Figura 24c). Quando comparamos os genes do grupo haplóide e mitótico para enriquecimento de genes com seleção positiva vemos o padrão esperado por nosso

modelo, o grupo haplóide apresenta mais genes com assinaturas de seleção positiva (Figura 24d). Porém, esse resultado pode ser apenas decorrente do fato do grupo haplóide possuir mais genes novos (Figura 21). Devido ao tamanho amostral pequeno (zero genes com seleção positiva para um dos grupos), o teste de comparação entre os grupos haploides e mitóticos usando somente os genes novos não pode ser realizado.

Alternativamente, verificamos que quando comparamos apenas os genes antigos, observamos que há mais genes com assinatura de seleção positiva no grupo haploide do que no grupo mitótico diploide (Teste exato de Fisher: p-valor < 0.01). A presença dessa maior proporção mesmo ao só considerarmos os genes antigos mostra que a maior seleção de genes mais expressos nas fases haplóides não só ocorre nos genes novos, mas também entre os genes antigos, indicando uma grande importância desse tipo de seleção. Além disso, testes utilizando apenas genes com seleção positiva na comparação dos modelos de seleção positiva e neutra (M1 vs M2) mostram o mesmo padrão, em que somente ao compararmos os genes totais ou antigos, há uma maior proporção de genes com assinaturas de seleção positiva no grupo haploide (Teste exato de Fisher: p-valor < 0.01).

O uso de diferentes modelos ao mesmo tempo pode trazer problemas, pois modelos com diferentes premissas podem não ser sobreponíveis. Assim, o uso de programas que analisam marcadores de seleção específicos, como o DoFE (Tabela 9) (EYRE-WALKER, 2010) se faz importante para dar suporte aos resultados encontrados com os modelos anteriores (Figura 24). Este programa foi usado por Zhang et al. (2010a) para verificar assinaturas de seleção positiva, o *alpha*, em genes autossômicos e ligados ao X. Utilizamos o LRT (*Likelihood-Ratio Test*, teste da taxa de verossimilhança) para verificar se as diferenças entre os grupos são esperadas dentro da hipótese nula ou não. Para o teste consideramos a hipótese nula a de que ambos os grupos testados vêm da mesma distribuição, e são por isso iguais. A hipótese alternativa postula então que os grupos possuem distribuições diferentes. Assim, verificamos se a soma dos valores de log da verossimilhança (LL) de dois grupos (valores fornecidos pelo programa DoFE) tem o mesmo comportamento que o grupo que engloba ambos. Ou seja, vemos se o comportamento do grupo que é a junção dos grupos de interesse é igual a simples junção do comportamento dos dois grupos separadamente. Assim, a análise de cada grupo é feita separadamente, e seus valores de log da verossimilhança são armazenados (LL1 e LL2). Então fazemos a análise do grupo total (que engloba os grupos anteriores), e guardamos seus valores de log da verossimilhança (LLt). Desta forma, seguindo os métodos utilizados por Zhang et al. (2010a), utilizamos o teste padrão de taxa de verossimilhança fazendo: $2((LL1 + LL2) - LLt)$, que nos dá o valor do teste dentro da distribuição de χ^2, com um grau de liberdade, dentro da hipótese nula (dado que o *alpha* é igual nos dois grupos).

Utilizando este método, ao analisarmos genes novos e antigos autossômicos, vemos que os genes novos do grupo haplóide possuem valor de alfa significativamente maior

		alpha	Tamanho amostral		
Genes Novos	Grupo Haplóide	0,179	283	***	***
	Grupo Mitótico	0,145	66		
Genes Antigos	Grupo Haplóide	-0,392	2662		***
	Grupo Mitótico	-0,376	3087		

Tabela 9 – **Valores de *alpha* e suas comparações usando o programa DoFE (*Distribution of Fitness Effect*).** Valores de *alpha* para os genes de cada idade e grupo de interesse, apenas para genes autossômicos. Os valores foram calculados com o programa DoFE (Distribution of Fitness Effect). As análises de *likelihood ratio* foram feitas a partir dos resultados do programa DoFE e da análise utilizada por Zhang et al. (2010a). Na tabela vemos os resultados (p-valor) das comparações dos genes novos contra os antigos no grupo haplóide e mitótico separadamente (englobados pelas chaves maiores, mais à direita), e dos genes novos do grupo haplóide contra os do grupo mitótico englobados pela chave menor e mais à esquerda. *** indicam p-valor < 0.01

que os genes antigos do mesmo grupo, e o mesmo ocorre para os genes novos do grupo mitótico (Tabela 9). Além disso, entre os genes novos, aqueles pertencentes ao do grupo haplóide apresentam valores estatisticamente maiores de *alpha* que o grupo mitótico. Esse comportamento apoia a hipótese de que há vantagem seletiva para genes com alta expressão nas fases haplóides da espermatogênese.

Esses resultados mostram a validade e força do modelo de seleção haplóide aqui apresentado. Embora neste trabalho não tenha sido possível obter resultados apenas com genes novos devido ao pequeno tamanho amostral dos grupos depois de todas as condições estabelecidas (sobre sua intensidade de expressão na espermatogênese e idade), os resultados com os grupos totais e genes antigos dão suporte suficiente para o modelo, e comprovam ainda mais sua ampla utilidade para se compreender melhor os processos de seleção e fixação de genes novos.

4.3 Genes Autossômicos e Ligados ao X

Para testar a predição de que somente os genes novos autossômicos devem ser mais expressos nas fases haplóides da espermatogênese, usamos dados disponíveis (VIBRANOVSKI et al., 2009; ZHANG et al., 2010a) sobre a localização cromossômica de cada gene, e os dados já descritos na seção anterior dos grupos de genes de acordo com sua expressão (genes mitóticos, meióticos e pós-meióticos). Podemos ver na Figura 28 que para genes novos existe uma proporção maior de genes autossômicos expressos na

pós-meiose quando comparado aos genes ligados ao X.

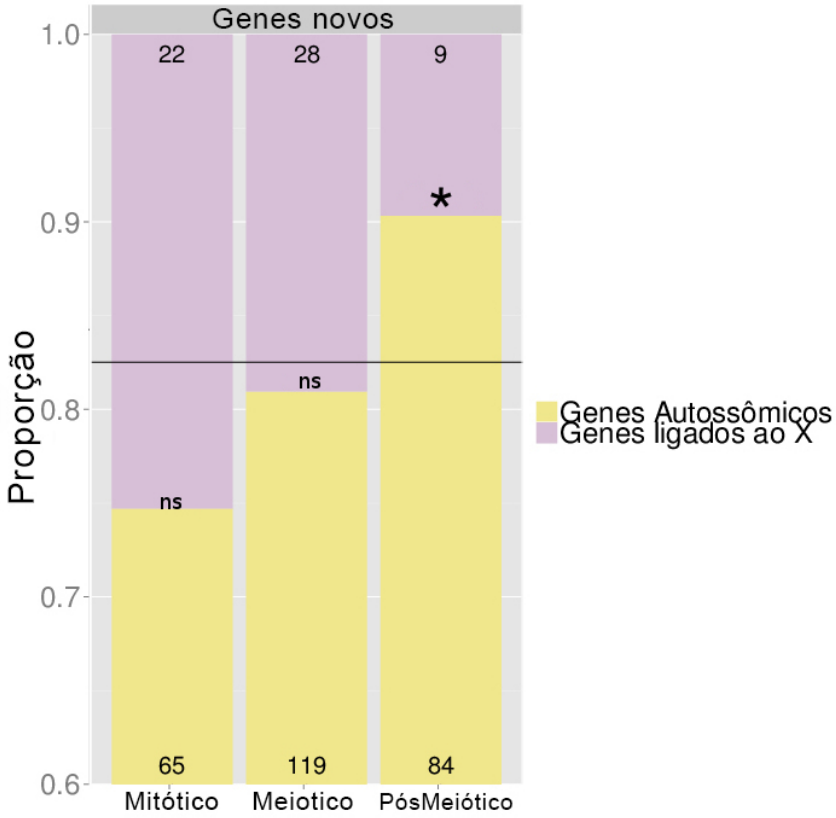


Figura 25 – **Proporção de genes novos autossômicos e ligados ao X ao longo da espermatogênese.** Ao comparar genes novos com maior expressão em cada fase da espermatogênese quanto a sua origem cromossômica, podemos ver uma menor proporção de genes ligados ao X do que autossômicos sendo altamente expressos na pós-meiose. Teste exato de Fisher comparando as proporções de genes ligados ao X e autossômicos mais expressos na pós-meiose em comparação as outras fases mitóticas e meióticas juntas: p-valor (*) < 0.05.

Sabemos que o cromossomo X se encontra em hemizigose nos machos, pois este é o sexo heterogamético, possuindo apenas uma cópia do cromossomo X. Assim não há uma vantagem real para os genes novos ligados ao X serem expressos nas células haplóides da espermatogênese em relação as diplóides, já que não há um homólogo do X nas células diplóides que torne vantajosa para algum gene novo ou alelo recessivo a expressão maior nas células haplóides. Portanto, uma extensão lógica de nossos resultados de seleção durante a meiose e pós-meiose é que não exista uma maior proporção de genes ligados ao X altamente expressos nas fases tardias da espermatogênese. Tais genes novos, se benéficos, serão selecionados com a mesma intensidade nas células haplóides e diplóides (gaméticas e somáticas). Assim, esperamos que os genes novos ligados ao X possuam expressão constante ao longo da espermatogênese, e que a expressão de genes novos autossômicos aumente nas

fases haplóides, como mostrado na Figura 26 .

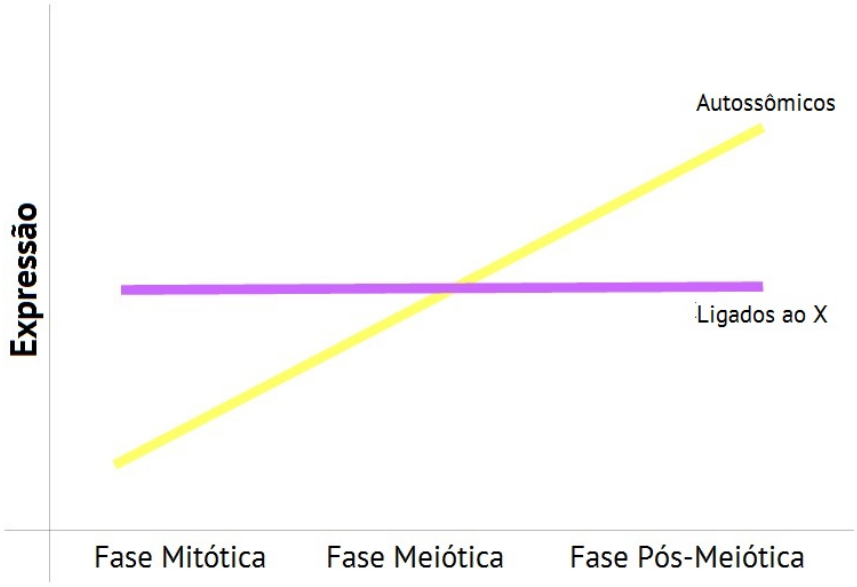


Figura 26 – **Comportamento esperado da expressão de genes novos autossômicos e ligados ao X durante a espermatogênese.** A partir de nossa hipótese, esperamos que genes ligados ao X não possuam vantagem ao serem expressos nas fases haplóides da espermatogênese, pois já se encontram em hemizigose em todas as células dos machos. Assim, esperamos que a expressão dos genes novos ligados ao X não se altere entre as fases mitótica, meiótica e pós-meiótica. Porém, como os genes autossômicos terão vantagem ao serem expressos nas fases meiótica e pós-meiótica, esperamos ver um aumento da expressão dos genes autossômicos nessas fases em relação à fase mitótica.

A demonstração da validade de mais essa predição em *Drosophila* mostra que o modelo proposto é robusto, e traz consigo uma possível razão para a fixação de genes novos que não se deva apenas a deriva, mas também a seleção de tais genes durante a espermatogênese. Desta forma, o modelo proposto inova e traz uma solução plausível para como pode ocorrer a seleção de genes novos nos testículos - onde já se sabe que muitos genes novos são altamente expressos - e, mais especificamente, nas fases tardias da espermatogênese.

4.4 Modelo Matemático

Por fim, para ajudar na elaboração formal do modelo, colaboramos com o professor Dr. Paulo A. Otto que construiu um modelo matemático que levasse em conta as premissas de nosso modelo e mostrasse a vantagem de um gene novo ser selecionado durante a espermatogênese em vez de nas células somáticas (e diplóides) do organismo adulto.

O modelo leva em conta as pontes citoplasmáticas existentes entre espermátides, e a recessividade do novo gene ou alelo. O modelo construído permite modificar esses parâmetros, assim como o coeficiente de seleção nas populações diplóide e haplóide para que possamos verificar a robustez do modelo em diferentes condições.

No modelo elaborado, é possível modificar parâmetros como o coeficiente de seleção para a população diplóide e haplóide, o grau de dominância do alelo e a frequência inicial do alelo recessivo nas populações. O modelo mostra que, dado que a seleção age consideravelmente mais rápido em populações haplóides que diplóides, em qualquer situação em que o coeficiente de seleção for maior na população haplóide ou igual nas duas populações o alelo terá mais vantagem se expresso e selecionado em um ambiente haplóide, como ocorre durante a espermatogênese.

Além de comprovar matematicamente algumas das características do modelo e sua veracidade, o modelo matemático trouxe ainda mais uma constatação da efetividade do modelo teórico. Conforme mostrado no Anexo A, mesmo que o coeficiente de seleção seja maior na população diplóide que na haplóide, se a frequência inicial do novo gene ou alelo (recessivo na população diplóide) for suficientemente baixa, ainda haverá vantagem para o mesmo ser expresso nos gametas (população haplóide). Os cálculos e simulações matemáticas realizadas pelo professor Dr. Paulo A. Otto mostram que mesmo que a seleção seja mais intensa nas células somáticas (diplóides) que nos gametas (haplóides) em 13% dos casos ainda é vantajoso para esse novo alelo ser expresso nos gametas. Dado que em geral não há razões para o coeficiente de seleção ser diferente entre células haplóides e diplóides, é sensato supor que os novos genes e alelos sejam mais rapidamente fixados se expressos durante a espermatogênese.

Sabemos que tanto genes como alelos novos surgem em frequências muito baixas na população, em geral em apenas um indivíduo, e devido a seleção e deriva tal gene ou alelo vai sendo fixado na população e espécie. Sabendo disso, é possível inferir que genes novos possuam de fato uma vantagem em serem expressos durante a espermatogênese, pois assim podem sofrer seleção mais rapidamente, podendo assim fixar-se na população.

5 Resultados: Localização cromossômica de genes preferencialmente expressos em machos

Sabe-se que genes preferencialmente expressos em machos tendem a se encontrar nos autossomos (RANZ et al., 2003; PARISI et al., 2003; ALLEN; BONDURIANSKY; CHENOWETH, 2013). No entanto, genes preferencialmente expressos nos machos que são muito novos se encontram no X (ZHANG et al., 2010a), como podemos ver na Figura 27. Hipóteses baseadas em antagonismo sexual (GIBSON; CHIPPINDALE; RICE, 2002; RICE, 1992) apresentam diferentes predições sobre a localização cromossômica de genes preferencialmente expressos em machos que dependem da relação de dominância e recessividade do alelo que estiver sob seleção. Como o cromossomo X está em hemizigose nas células masculinas, mas não nas femininas, se espera que genes com um alelo recessivo benéfico para os machos e maléficos para as fêmeas aumentem de frequência na população mais rápido se estiver localizado no cromossomo X do que nos autossomos (BAINES et al., 2008). Isso ocorre porque em machos estes alelos serão sempre selecionados positivamente, e em fêmeas serão selecionados negativamente somente quando estiverem em homozigose. Já os mesmos modelos matemáticos (RICE, 1984) predizem que os alelos dominantes benéficos para os machos e maléficos para as fêmeas se fixam mais rápido se localizados em genes autossômicos. Isso ocorre já que alelos dominantes sofreram mais frequentemente seleção negativa em fêmeas do que seleção positiva em machos. A hipótese de antagonismo sexual dificilmente explicaria sozinha os resultados encontrados por Zhang et al. (2010a) (Figura 27), já que implicaria que alelos de genes novos benéficos para machos e maléficos para fêmeas surgissem com mais frequência em recessividade enquanto que para genes antigos surgissem em dominância.

Outra característica do X que explica a menor ocorrência de genes mais expressos em machos é o MSCI. A inativação do cromossomo sexual (MSCI) ocorre em diversos animais, como camundongos (TURNER et al., 2005; TURNER et al., 2002) e drosofilídeos (VIBRANOVSKI, 2014; VIBRANOVSKI et al., 2012a). O processo consiste na diminuição da expressão do cromossomo X durante a espermatogênese, mais especificamente na fase meiótica da mesma. Essa diminuição da expressão pode ser total ou parcial, sendo esta última presente em *Drosophila* (HENSE; BAINES; PARSCH, 2007; VIBRANOVSKI et al., 2009). Mesmo que tal inativação seja apenas parcial, já é uma desvantagem para um gene mais expresso em machos, pois se estiver no X há uma grande possibilidade de tal gene não ser expresso na meiose.

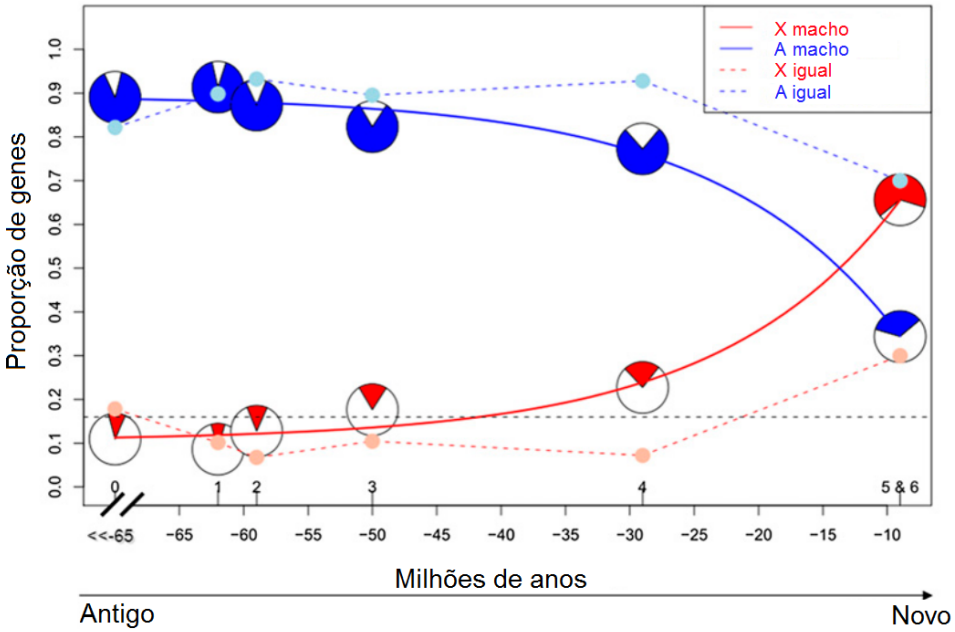


Figura 27 – **Localização cromossômica de genes preferencialmente expressos em machos.** A imagem mostra que embora genes mais expressos em machos sejam mais comuns nos autossomos, quando dividimos esses genes em novos e antigos, podemos ver que esse padrão muda para genes extremamente novos. Para esses genes muito novos, vemos que há um enriquecimento de genes mais expressos em machos no cromossomo X, e não nos autossomos. Na legenda da figura Ẍ machoë Ä machoïepresentam genes mais expressos em machos no cromossomo X e autossomos, respectivamente. As curvas indicadas por Ẍ igualë Ä igualse referem a genes com expressão igual nos machos e fêmeas presentes no cromossomo X e autossomos, respectivamente. Imagem reproduzida e traduzida de Zhang et al. (2010a).

Dado que a localização cromossômica de genes preferencialmente expressos em machos varia com a idade gênica e dado que genes mais expressos em machos em geral são altamente expressos nos testículos em comparação ao ovário, resolvemos analisar a proporção de genes ligados ao cromossomo X e autossomos em diferentes fases da espermatogênese como na Figura 25 do capítulo 4. No entanto, desta vez incluímos para comparação os genes antigos (Figura 28). Como vimos anteriormente, expressos na pós-meiose, existe uma maior proporção de genes novos autossômicos do que ligados ao X. Isto corrobora a nossa hipótese de que a expressão haplóide facilita a fixação de alelos positivamente selecionados. Já para genes ligados ao X, a evolução mais rápida esta relacionada ao estado de hemizigose do cromossomo em todas as células dos machos, fenômeno conhecido em inglês por *faster-X effect* (MEISEL; CONNALLON, 2013; VICOSO; CHARLESWORTH, 2009; KOUSATHANAS; HALLIGAN; KEIGHTLEY, 2014). A divergência mais rápida dos genes ligados ao cromossomo X é conhecida e se dá devido em parte ao menor tamanho populacional efetivo do cromossomo X, que nos machos está em hemizigose (MEISEL; CONNALLON, 2013).

Nos genes antigos (Figura 28), não observamos diferença entre proporção de genes ligados ao X e autossomos na pós-meiose, já que os efeitos da seleção haplóide são esperados para o modelo de genes novos. No entanto, podemos observar que durante a fase meiótica há uma redução na proporção de genes antigos ligados ao X. Esse decaimento é esperado devido a inativação do cromossomo X durante a meiose que ocorre em drosofilídeos (VIBRANOVSKI et al., 2012b; VIBRANOVSKI, 2014).

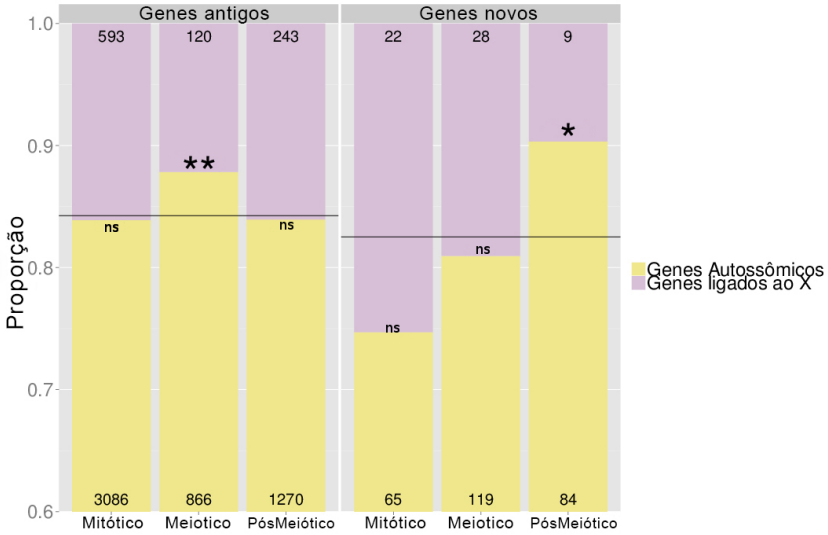


Figura 28 – **Expressão de genes autossômicos e ligados ao X, novos e antigos ao longo da espermatogênese** Ao comparar genes com maior expressão em cada fase da espermatogênese quanto a sua origem cromossômica, podemos ver que para genes antigos há uma quantidade constante de genes ligados ao X com alta expressão em todas as fases da espermatogênese, a diminuição durante a meiose pode ser explicada pela inativação meiótica do cromossomo X (VIBRANOVSKI et al., 2012a; VIBRANOVSKI, 2014). Para genes novos vemos uma diminuição de genes ligados ao X sendo altamente expressos o que é esperado de acordo com nosso modelo de seleção haplóide, pois não há vantagem na expressão em células haplóides para os genes ligados ao X, pois estão sempre em hemizigose. Na figura as comparações significativamente diferentes da média são identificadas por (**) p-valor < 0.01 no teste exato de Fisher, (*) p-valor < 0.05 no teste exato de Fisher.

A proporção de genes no cromossomo X que são altamente expressos na meiose diminui significativamente para os genes antigos, mas não para os novos. Isso mostra a inativação do cromossomo X se faz mais evidente nos genes antigos. Esse maior efeito da inativação do X nos genes antigos provavelmente se dá pois tais genes já devem possuir os marcadores genéticos e epigenéticos necessários para que a inativação dos mesmos seja eficiente. Nos genes novos tais marcadores ainda não devem ter sido totalmente estabelecidos, provavelmente devido ao tempo necessário para que isso aconteça. Assim, os genes antigos já possuem tais marcadores para que sejam inativados durante a meiose, e

os genes novos os obtém com o passar do tempo, e por isso ainda são expressos nessa fase, como mostra a Figura 29.

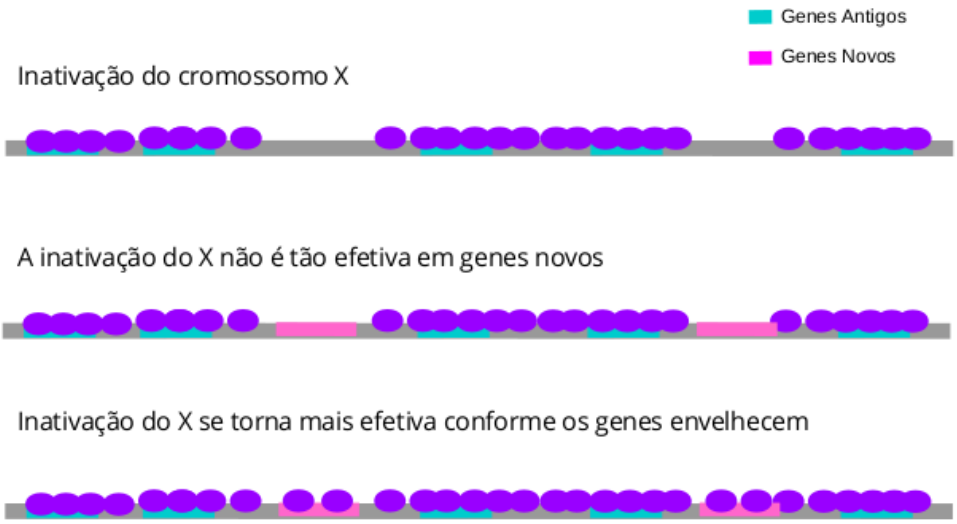


Figura 29 – **Inativação dos genes novos no cromossomo X.** Genes antigos (em azul) já tem marcadores epigenéticos para inativação dos mesmos durante a espermatogênese, porém genes novos (em rosa) ao surgirem no cromossomo X não possuem tais marcadores e por isso são expressos na espermatogênese. Esse fato explica a grande presença de genes novos mais expressos em machos que existe no cromossomo X de *Drosophila* (ZHANG et al., 2010a). Porém, com o passar do tempo os genes novos desenvolvem tais marcadores, sendo então inativados durante o processo, o que explica a migração de tais genes mais expressos em machos para os autossomos como podemos ver em Zhang et al. (2010a).

Nossos dados e análises da localização cromossômica de genes novos e antigos ao longo da espermatogênese conseguem explicar os resultados contraditórios encontrados por Zhang et al. (2010a)), onde foi descrito que genes muito novos e com maior expressão em machos são muito comuns no cromossomo X enquanto que genes mais antigos e com maior expressão nos machos são encontrados nos autossomos. Genes novos preferencialmente expressos em machos (em testículos), se localizados no X, tem a vantagem de não possuírem outra cópia de si no cromossomo homólogo e por isso tem mais chances de serem fixados mais rapidamente, assim como genes autossômicos em fases haplóides. Em outras palavras, dado que o cromossomo X se encontra em hemizigose nos machos sua situação de expressão se assemelha aquela dos gene autossômicos durante as fases haplóides da espermatogênese. Além disso, como podemos ver na Figura 28, os genes novos são capazes de driblar a inativação meiótica do X, o que lhes confere vantagem sobre os genes antigos nessa etapa do processo. Com isso é possível explicar a presença de genes novos com alta expressão em machos no cromossomo X. Tais genes evadem os mecanismos e inativação e podem ser úteis durante a espermatogênese. Assim, conforme tais genes envelhecem e adquirem os marcadores genéticos e epigenéticos de inativação durante a meiose, o cromossomo X

deixa de ser um local vantajoso para a expressão de genes preferencialmente expressos na meiose masculina/testículos/machos (VIBRANOVSKI et al., 2009). Por isso, é possível compreender a presença massiva de genes mais expressos em machos no cromossomo X quando são novos, enquanto os antigos devam ser mais prevalentes nos autossomos.

6 Conclusões

Ao longo deste trabalho pudemos ver a relação entre expressão durante a espermatogênese, idade do gene e marcadores de seleção do mesmo. Pudemos verificar a maior expressão de genes novos nas fases tardias (e haplóides) da espermatogênese. Essa maior expressão pode estar ligada a abertura da cromatina nessas fases (como proposto por Soumillon et al. (2013)), porém a permissividade da cromatina permitiria o aparecimento e posterior fixação não só de genes funcionais mas também de pseudogenes. A verificação do mesmo comportamento em *Drosophila* aponta em outra direção, pois o genoma de *Drosophila melanogaster* é compacto (PETROV, 2002; PETROV; HARTL, 2000) e possui poucos pseudogenes (PETROV; HARTL, 1998; PETROV; HARTL, 2000; HARRISON; GERSTEIN, 2002).

Também vimos que os genes novos do cromossomo X apresentam uma diminuição na proporção de genes com alta expressão em relação aos genes autossômicos nas mesmas fases da espermatogênese. Isso ocorre, pois não há vantagem seletiva para genes que já estão em hemizigose nas células diplóides em serem altamente expresso em células haplóides - para esses genes a pressão seletiva nessas duas situações é a mesma. Isso dá suporte a teoria de que há seleção agindo sobre os genes altamente expressos nas fases haplóides, pois apenas genes autossômicos tendem a ser mais expressos nessas fases, como esperado se há vantagem na seleção haplóide (que ocorre durante as últimas fases da espermatogênese).

Outra conclusão importante desses resultados é a resolução da questão de por quê genes novos com alta expressão em machos são comuns no cromossomo X, se este é inativado durante parte da espermatogênese. Como vimos na figura 28 e 29, para os genes novos a inativação do cromossomo X não é tão eficiente. Isso leva genes muito novos a não terem uma desvantagem efetiva por estarem no X -afinal eles continuam expressos em toda a espermatogênese. Assim, conforme tais genes envelhecem e começam a apresentar marcadores para a sua inativação durante a meiose (Figura 29) há uma pressão seletiva para manter cópias desses genes em autossomos, pois apenas estes serão expressos durante toda a espermatogênese.

Por fim, há algumas evidências de seleção desses genes autossômicos mais expressos nas fases tardias da espermatogênese. Ainda é interessante realizar novos testes para verificar a robustez da presença de seleção nos genes mais expressos nas fases haplóides, pois é extremamente difícil inferir seleção (SOSKINE; TAWFIK, 2010; LONG et al., 2003; SELLA et al., 2009; EYRE-WALKER, 2006; KRYAZHIMSKIY; PLOTKIN, 2008) visto que testes de seleção são muito sensíveis à história natural e características do grupo. Apesar desses percalços, foi possível encontrar marcadores de seleção para os genes

altamente expressos nas fases haplóides, conforme esperado por nosso modelo.

Para dar suporte ao modelo teórico, foi desenvolvido um modelo matemático que demonstra a viabilidade e robustez do mesmo. Um possível empecilho para o modelo é o fato de que há transferência de RNAm entre as espermátides pelas pontes citoplasmáticas que só se fecham ao final da pós-meiose. No modelo matemático isso se reflete na possibilidade de alterar o grau de dominância de um alelo sobre o outro, representando a possibilidade de uma parcela do RNAm presente numa dada célula de um cisto ser proveniente de outra célula do mesmo cisto. Mesmo considerando tais transferências o modelo matemático mostrou que para a maior parte dos casos há vantagem na seleção haplóide. Isso ocorre principalmente quando o gene ainda está em baixa frequência na população, o que sabemos ser o caso para genes que acabaram de surgir.

Essas evidências tornam mais robusta a hipótese de seleção haplóide. Como mostramos, há suficiente suporte teórico e experimental para tal. Assim, podemos concluir que de fato há uma vantagem para genes novos que são altamente expressos nas fases haplóides da espermatogênese. Os genes novos tendem a ser mais expressos que os antigos nessas fases e isso os ajuda a serem fixados na população devido à seleção que sofrem nesses momentos.

Referências

ALLEN, S. L.; BONDURIANSKY, R.; CHENOWETH, S. F. The genomic distribution of sex-biased genes in Drosophila serrata: X chromosome demasculinization, feminization, and hyperexpression in both sexes. *Genome Biology and Evolution*, v. 5, n. 10, p. 1986–1994, 2013. Nenhuma citação no texto.

BABUSHOK, D. V. et al. A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. *Genome Research*, v. 17, n. 8, p. 1129–1138, 2007. Nenhuma citação no texto.

BAI, Y.; CASOLA, C.; BETRÁN, E. Evolutionary origin of regulatory regions of retrogenes in Drosophila. *BMC genomics*, v. 9, p. 241, 2008. Nenhuma citação no texto.

BAI, Y. et al. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. *Genome biology*, v. 8, n. 1, p. R11, 2007. Nenhuma citação no texto.

BAINES, J. F. et al. Effects of x-linkage and sex-biased gene expression on the rate of adaptive protein evolution in drosophila. *Molecular Biology and Evolution*, v. 25, n. 8, p. 1639–1650, 2008. Nenhuma citação no texto.

BETRÁN, E.; THORNTON, K.; LONG, M. Retroposed new genes out of the X in Drosophila. *Genome Research*, v. 12, n. 12, p. 1854–1859, 2002. Nenhuma citação no texto.

BIERNE, N.; EYRE-WALKER, A. The genomic rate of adaptive amino acid substitution in drosophila. *Molecular Biology and Evolution*, v. 21, p. 1350–1360, 2004. Nenhuma citação no texto.

BUTLER, R. G. Population size, social behaviour, and dispersal in house mice: A quantitative investigation. *Animal Behaviour*, v. 28, n. 1, p. 78–85, 1980. Nenhuma citação no texto.

CÁCERES, L.; NILSON, L. A. Production of gurken in the nurse cells is sufficient for axis determination in the Drosophila oocyte. *Development (Cambridge, England)*, v. 132, n. 10, p. 2345–2353, 2005. Nenhuma citação no texto.

CAI, J. et al. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics*, v. 179, n. 1, p. 487–496, 2008. Nenhuma citação no texto.

CHEN, S.; KRINSKY, B. H.; LONG, M. New genes as drivers of phenotypic evolution. *Nature reviews. Genetics*, v. 14, n. 9, p. 645–60, 2013. Nenhuma citação no texto.

CHEN, S.; ZHANG, Y. E.; LONG, M. New genes in Drosophila quickly become essential. *Science*, v. 330, n. 2010, p. 1682–1685, 2010. Nenhuma citação no texto.

CRUZ, F. de la; DAVIES, J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in Microbiology*, v. 8, n. 3, p. 128–133, 2000. Nenhuma citação no texto.

CUI, X. et al. Young genes out of the male: An insight from evolutionary age analysis of the pollen transcriptome. *Molecular Plant*, Elsevier Ltd, v. 8, n. 6, p. 935–945, 2015. Nenhuma citação no texto.

DAI, H. et al. The evolution of courtship behaviours through the origination of a new gene in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, v. 105, n. 21, p. 7478–7483, 2008. Nenhuma citação no texto.

DARWIN, C. *On the origin of species by means of natural selection*. [S.l.: s.n.], 1859. Nenhuma citação no texto.

Des Marais, D. L.; RAUSHER, M. D. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, v. 454, n. 7205, p. 762–765, 2008. Nenhuma citação no texto.

DUPRESSOIR, A. et al. Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proc Natl Acad Sci U S A*, v. 106, n. 29, p. 12127–12132, 2009. Nenhuma citação no texto.

EMERSON, J. J. et al. Extensive Gene Traffic on the Mammalian X Chromosome. *Science*, v. 303, n. 537, 2004. Nenhuma citação no texto.

EYRE-WALKER, A. The genomic rate of adaptive evolution. *Trends in Ecology and Evolution*, v. 21, n. 10, p. 569–575, 2006. Nenhuma citação no texto.

EYRE-WALKER, A. *DoFE 3.0 - Distribution of Fitness Effects Software*. [S.l.]: University of Sussex, 2010. Nenhuma citação no texto.

FABLET, M. et al. Evolutionary origin and functions of retrogene introns. *Molecular Biology and Evolution*, v. 26, n. 9, p. 2147–2156, 2009. Nenhuma citação no texto.

FABRIZIO, J. J. et al. Genetic dissection of sperm individualization in Drosophila melanogaster. *Development*, v. 125, n. 10, p. 1833–43, 1998. Nenhuma citação no texto.

FAY, J. C.; WYCKOFF, G. J.; WU, C.-I. Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature*, v. 415, n. 6875, p. 1024–6, 2002. Nenhuma citação no texto.

FULLER, M. T. Spermatogenesis. In: . [S.l.: s.n.], 1993. Nenhuma citação no texto.

GIBSON, J. R.; CHIPPINDALE, A. K.; RICE, W. R. The X chromosome is a hot spot for sexually antagonistic fitness variation. *Proc R Soc Lond B*, v. 269, p. 499–505, 2002. Nenhuma citação no texto.

GILLESPIE, J. H. *Population Genetics: A Concise Guide*. [S.l.]: JHU Press, 2010. Nenhuma citação no texto.

HARRISON, P. M.; GERSTEIN, M. Studying genomes through the aeons: Protein families, pseudogenes and proteome evolution. *Journal of Molecular Biology*, v. 318, n. 5, p. 1155–1174, 2002. Nenhuma citação no texto.

HE, L.; WANG, X.; MONTELL, D. Shining light on Drosophila oogenesis: live imaging of egg development. *Current opinion in genetics & development*, v. 21, n. 5, p. 612–619, 2011. Nenhuma citação no texto.

HEIDMANN, O. et al. Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin"in a third order of mammals. *Retrovirology*, v. 6, p. 107, 2009. Nenhuma citação no texto.

HENSE, W.; BAINES, J. F.; PARSCH, J. X chromosome inactivation during Drosophila spermatogenesis. *PLoS Biology*, v. 5, n. 10, p. 2288–2295, 2007. Nenhuma citação no texto.

HOLLAND, P. W. H. et al. Gene duplications and the origins of vertebrate development. *Development 1994 Supplement*, p. 125–133, 1994. Nenhuma citação no texto.

HUXLEY, J. *Evolution - The Modern Synthesis*. [S.l.]: George Alien & Unwin Ltd, 1942. Nenhuma citação no texto.

JOHNSON, M. E. et al. Positive selection of a gene family during the emergencce of humans and African apes. *The Holocene*, v. 413, 2001. Nenhuma citação no texto.

JOSEPH, S. B.; KIRKPATRICK, M. Haploid selection in animals. *Trends in Ecology and Evolution*, v. 19, n. 11, p. 592–597, 2004. Nenhuma citação no texto.

KABZA, M.; CIOMBOROWSKA, J.; MAKAŁOWSKA, I. RetrogeneDB - A database of animal retrogenes. *Molecular Biology and Evolution*, v. 31, n. 7, p. 1646–1648, 2014. Nenhuma citação no texto.

KAESSMANN, H. More Than Just a Copy. *Science*, v. 325, n. April 2008, p. 2008–2009, 2009. Nenhuma citação no texto.

KAESSMANN, H. Origins, evolution, and phenotypic impact of new genes. *Genome Research*, v. 20, n. 10, p. 1313–1326, 2010. Nenhuma citação no texto.

KAESSMANN, H.; VINCKENBOSCH, N.; LONG, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature reviews. Genetics*, v. 10, n. 1, p. 19–31, 2009. Nenhuma citação no texto.

KEMKEMER, C.; HENSE, W.; PARSCH, J. Fine-scale analysis of X chromosome inactivation in the male germ line of drosophila melanogaster. *Molecular Biology and Evolution*, v. 28, n. 5, p. 1561–1563, 2011. Nenhuma citação no texto.

KHAMMARI, A. et al. Physiological apoptosis of polar cells during Drosophila oogenesis is mediated by Hid-dependent regulation of Diap1. *Cell Death Differ*, v. 18, n. 5, p. 793–805, 2011. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21113144>. Nenhuma citação no texto.

KLEINJAN, D. A. et al. Subfunctionalization of Duplicated Zebrafish pax6 Genes by cis-Regulatory Divergence. *PLoS Genetics*, v. 4, n. 2, 2008. Nenhuma citação no texto.

KOUSATHANAS, A.; HALLIGAN, D. L.; KEIGHTLEY, P. D. Faster-x adaptive protein evolution in house mice. *Genetics*, v. 196, p. 1131–1143, 2014. Nenhuma citação no texto.

KRYAZHIMSKIY, S.; PLOTKIN, J. B. The population genetics of dN/dS. *PLoS Genetics*, v. 4, n. 12, 2008. Nenhuma citação no texto.

LAMARCK, J.-B. *Philosophie Zoologique*. [S.l.: s.n.], 1809. Nenhuma citação no texto.

LANDRY, C. et al. Recent speciation in the Indo-West Pacific: rapid evolution of gamete recognition and sperm morphology in cryptic species of sea urchin. *Proc R Soc Lond B*, v. 270, n. 1526, p. 1839–1847, 2003. Nenhuma citação no texto.

LENORMAND, T.; DUTHEIL, J. Recombination difference between sexes: A role for haploid selection. *PLoS Biology*, v. 3, n. 3, p. 0396–0403, 2005. Nenhuma citação no texto.

LEVINE, M. T. et al. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America*, v. 103, n. 26, p. 9935–9, 2006. Nenhuma citação no texto.

LIAO, B.-Y.; WENG, M.-P. Unraveling the association between mRNA expressions and mutant phenotypes in a genome-wide assessment of mice. *Proceedings of the National Academy of Sciences*, v. 112, n. 15, p. 201415046, 2015. Nenhuma citação no texto.

LINDSEY, D.; TOKUYASU, K. Spermatogenesis. In: . [S.l.: s.n.], 1980. Nenhuma citação no texto.

LONG, M. et al. The origin of new genes: glimpses from the young and old. *Nature Reviews. Genetics*, v. 4, n. 11, p. 865–875, 2003. Nenhuma citação no texto.

LONG, M.; LANGLEY, C. H. Natural selection and the origin of jingwey, a chimeric processed functional gene in Drosphila. v. 260, n. 5104, p. 91–95, 1993. Nenhuma citação no texto.

MALIK, H. S.; HENIKOFF, S. Positive selection of iris, a retroviral envelope-derived host gene in Drosophila melonogaster. *PLoS Genetics*, v. 1, n. 4, p. 0429–0443, 2005. Nenhuma citação no texto.

MARQUES, A. C. et al. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biology*, v. 3, n. 11, p. 1970–1979, 2005. Nenhuma citação no texto.

MAYR, E. Lamarck revisited. *Journal of the History of Biology*, v. 5, n. 1, p. 55–94, 1972. Nenhuma citação no texto.

MAYR, E. *The growth of biological thought*. [S.l.]: Harvard University Press, 1982. Nenhuma citação no texto.

MEISEL, R. P.; CONNALLON, T. The faster-x effect: integrating theory and data. *Trends in Genetics*, v. 29, n. 9, p. 537–544, 2013. Nenhuma citação no texto.

METTA, M.; SCHLÖTTERER, C. Non-random genomic integration - an intrinsic property of retrogenes in Drosophila? *BMC Evolutionary Biology*, v. 10, p. 114, 2010. Nenhuma citação no texto.

MONTELL, D. J. Border-cell migration: the race is on. *Nature reviews. Molecular cell biology*, v. 4, n. 1, p. 13–24, 2003. Nenhuma citação no texto.

NIELSEN, R.; YANG, Z. Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, v. 148, p. 929–936, 1998. Nenhuma citação no texto.

OHNO, S. *So much "junk"DNA in our genome*. 1972. Nenhuma citação no texto.

PARISI, M. et al. Paucity of genes on the Drosophila X chromosome showing male-biased expression. *Science*, v. 299, n. 5607, p. 697–700, 2003. Nenhuma citação no texto.

PETIT, N.; BARBADILLA, A. Selection efficiency and effective population size in Drosophila species. *Journal of Evolutionary Biology*, v. 22, n. 3, p. 515–526, 2008. Nenhuma citação no texto.

PETROV, D. A. DNA loss and evolution of genome size in Drosophila. *Genetica*, v. 115, n. 1, p. 81–91, 2002. Nenhuma citação no texto.

PETROV, D. A.; HARTL, D. L. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. *Molecular biology and evolution*, v. 15, n. 3, p. 293–302, 1998. Nenhuma citação no texto.

PETROV, D. a.; HARTL, D. L. Pseudogene evolution and natural selection for a compact genome. *The Journal of heredity*, v. 91, n. 3, p. 221–7, 2000. Nenhuma citação no texto.

POTRZEBOWSKI, L. et al. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biology*, v. 6, n. 4, p. 709–716, 2008. Nenhuma citação no texto.

RANZ, M. et al. Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science*, v. 300, n. June, p. 1742–1745, 2003. Nenhuma citação no texto.

RICE, W. R. Sex Chromosomes and the Evolution of Sexual Dimorphism. *Evolution*, v. 38, n. 4, p. 735–742, 1984. Nenhuma citação no texto.

RICE, W. R. Sexually antagonistic genes: experimental evidence. *Science*, v. 256, n. June, p. 10, 1992. Nenhuma citação no texto.

ROSSO, L. et al. Mitochondrial targeting adaptation of the hominoid-specific glutamate dehydrogenase driven by positive darwinian selection. *PLoS Genetics*, v. 4, n. 8, 2008. Nenhuma citação no texto.

ROSSO, L. et al. Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein. *PLoS Biology*, v. 6, n. 6, p. 1281–1291, 2008. Nenhuma citação no texto.

SCHLOTTERER, C. Genes from scratch - the evolutionary fate of de novo genes. *Trends in Genetics*, v. 31, n. 4, p. 215–219, 2015. Nenhuma citação no texto.

SEEHAUSEN, O. et al. Genomics and the origin of species. *Nature Reviews*, v. 15, p. 176–192, 2014. Nenhuma citação no texto.

SELLA, G. et al. Pervasive natural selection in the Drosophila genome? *PLoS Genetics*, v. 5, n. 6, 2009. Nenhuma citação no texto.

SHAPIRO, J. A. et al. Adaptive genic evolution in the Drosophila genomes. *Proceedings of the National Academy of Sciences*, v. 104, n. 7, p. 2271–2276, 2007. Nenhuma citação no texto.

SNOOK, R. R. Sperm in competition: Not playing by the numbers. *Trends in Ecology and Evolution*, v. 20, n. 1, p. 46–53, 2005. Nenhuma citação no texto.

SOSKINE, M.; TAWFIK, D. S. Mutational effects and the evolution of new protein functions. *Nat Rev Genet*, Nature Publishing Group, v. 11, n. 8, p. 572–582, 2010. Nenhuma citação no texto.

SOUMILLON, M. et al. Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports*, v. 3, n. 6, p. 2179–2190, 2013. Nenhuma citação no texto.

STANLEY, C. E.; KULATHINAL, R. J. flydivas: A comparative genomics resource for drosophila divergence and selection. *G3: Genes/Genomes/Genetics*, v. 6, n. 8, p. 2355–2363, 2016. Nenhuma citação no texto.

STOCKLEY, P. Sexual conflict resulting from adaptations to sperm competition. *Trends in Ecology and Evolution*, v. 12, n. 4, p. 154–159, 1997. Nenhuma citação no texto.

SWANSON, W. J.; NIELSEN, R.; YANG, Q. Pervasive adaptive evolution in mammalian fertilization proteins. *Mol. Biol. Evol.*, v. 20, p. 18–20, 2003. Nenhuma citação no texto.

TURNER, J. M. A. et al. Meiotic sex chromosome inactivation in male mice with targeted disruptions of xist. *Journal of Cell Science*, v. 115, p. 4097–4105, 2002. Nenhuma citação no texto.

TURNER, J. M. A. et al. Silencing of unsynapsed meiotic chromosomes in the mouse. *Nature Genetics*, v. 37, p. 41–47, 2005. Nenhuma citação no texto.

VIBRANOVSKI, M. D. Meiotic sex chromosome inactivation in Drosophila. *Journal of genomics*, v. 2, p. 104–17, 2014. Nenhuma citação no texto.

VIBRANOVSKI, M. 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, 2009. Nenhuma citação no texto.

VIBRANOVSKI, M. D.; ZHANG, Y.; LONG, M. General gene movement off the X chromosome in the Drosophila genus. *Genome research*, v. 19, n. 5, p. 897–903, 2009. Nenhuma citação no texto.

VIBRANOVSKI, M. D. et al. Re-analysis of the larval testis data on meiotic sex chromosome inactivation revealed evidence for tissue-specific gene expression related to the drosophila X chromosome. *BMC Biol*, v. 10, n. 1, p. 49; author reply 50, 2012. Nenhuma citação no texto.

VIBRANOVSKI, M. D. et al. Segmental dataset and whole body expression data do not support the hypothesis that non-random movement is an intrinsic property of Drosophila retrogenes. *BMC evolutionary biology*, BMC Evolutionary Biology, v. 12, n. 1, p. 169, 2012. Nenhuma citação no texto.

VICOSO, B.; CHARLESWORTH, B. Effective population size and the faster-x effect: An extended model. *Evolution*, v. 63, n. 9, p. 2413–2426, 2009. Nenhuma citação no texto.

VINCKENBOSCH, N.; DUPANLOUP, I.; KAESSMANN, H. Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, v. 103, n. 9, p. 3220–5, 2006. Nenhuma citação no texto.

WANG, W. et al. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proc Natl Acad Sci U S A*, v. 99, n. 7, p. 4448–4453, 2002. Nenhuma citação no texto.

WANG, W. et al. High rate of chimeric gene origination by retroposition in plant genomes. *The Plant cell*, v. 18, n. 8, p. 1791–802, 2006. Nenhuma citação no texto.

WEDELL, N.; GAGE, M. J. G.; PARKER, G. A. Sperm competition, male prudence, and sperm-limited females. *Trends in Ecology & Evolution*, v. 17, n. 7, 2002. Nenhuma citação no texto.

WONG, W. S. W. et al. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, v. 168, p. 1041–1051, 2004. Nenhuma citação no texto.

YANG, Z. Paml: A program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, v. 13, p. 555– 556, 1997. Nenhuma citação no texto.

ZHANG, J. Evolution by gene duplication: An update. *Trends in Ecology and Evolution*, v. 18, n. 6, p. 292–298, 2003. Nenhuma citação no texto.

ZHANG, J.; ZHANG, Y.-p.; ROSENBERG, H. F. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature genetics*, v. 30, n. 4, p. 411–415, 2002. Nenhuma citação no texto.

ZHANG, Y. E. et al. Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Research*, v. 20, n. 11, p. 1526–1533, 2010. Nenhuma citação no texto.

ZHANG, Y. E. et al. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biology*, v. 8, n. 10, 2010. Nenhuma citação no texto.

ZHOU, Q.; WANG, W. On the origin and evolution of new genes-a genomic and experimental perspective. *Journal of Genetics and Genomics*, Institute of Genetics and Developmental Biology and the Genetics Society of China, v. 35, n. 11, p. 639–648, 2008. Nenhuma citação no texto.

ZHOU, Q. et al. On the origin of new genes in Drosophila. *Genome research*, v. 18, n. 9, p. 1446–1455, 2008. Nenhuma citação no texto.

APÊNDICE A – Perl

Durante o desenvolvimento deste projeto foi desenvolvido o seguinte programa em Perl para cruzar as tabelas obtidas de Vibranovski et al. (2009) e Zhang et al. (2010a). Todos os programas apresentados podem ser encontrados em <https://github.com/juliaraices/final_publishable>.

```
# January 2016
# Julia Raices
# Program to do everything I ever did with this tables
#!/usr/bin/perl

use strict; # doesn't let you use variables without declaring
    ↪ them

# declared variables:
my ($exps, $dnds, $age, $lines, $line, $lin, $i, $j, $k, $l, $m,
    ↪  $n, $o, $key, $value_dnds, $value_pnps); # counters and
    ↪ strings
my ($Group, $XorA, $OorN, $controle7, $controle8, $haploid); #
    ↪ strings to be printed
my (@exp, @ages, @dndss); # arrays with raw data
my (%exp_data, %dnds_data, %age_data); # hashs with printable
    ↪ data

# undefine variables, so that they don't - by chance - start not
    ↪  empty
undef $exps;
undef $dnds;
undef $age;
undef $lines;
undef $line;
undef $lin;
undef $i;
undef $j;
undef $k;
undef $l;
undef $m;
```

```
undef $n;
undef $o;
undef $key;
undef $Group;
undef $XorA;
undef $OorN;
undef $value_dnds;
undef $value_pnps;
undef $controle7;
undef $controle8;
undef $haploid;
undef @exp;
undef @ages;
undef @dndss;
undef %exp_data;
undef %dnds_data;
undef %age_data;

# sees what is the argument (after the program was called) and
    ↪ stores it. It should be the names of files to be used by
    ↪ the program
for($i=0; $i<=$#ARGV; $i+=2){
        if($ARGV[$i] eq "-exp"){
                $exps=$ARGV[$i+1];
        }
        if($ARGV[$i] eq "-dnds"){
                $dnds=$ARGV[$i+1];
        }
        if($ARGV[$i] eq "-age"){
                $age=$ARGV[$i+1];
        }
}

# check if files have been designated correctly, if they are not
    ↪ , it tells you how to use the program and exits
if($exps eq "" || $dnds eq "" || $age eq ""){
        print STDERR "Program usage: perl final.pl -exp
            ↪ EXPRESSION_TABLE -dnds DNDS_TABLE -age AGE_TABLE\n
            ↪ \t-exp: Table with genes' expression during each
```

```
            ↪ spermatogenesis␣phase.\n\t-dnds:␣Table␣with␣genes'
            ↪ ␣data␣of␣dN,␣dS,␣pN,␣pS,␣etc.\n\t-age:␣Table␣with␣
            ↪ genes'␣age␣data.\n";
        exit 0;
}

# open log file and adds data from this time program was run
open (LOG, ">>final.log");
print LOG "\n".(localtime)."\nTable␣for␣expression␣data:␣$exps\
    ↪ nTable␣for␣dN/dS␣data:␣$dnds\nTable␣for␣age␣data:␣$age\
    ↪ nOutput␣file:␣final.output\n";

# finds if input files actually exists and are openable, if not,
    ↪  prints error in log and in stderr and exists program
unless(open(EXP, $exps) && open(DNDS, $dnds) && open(AGE, $age))
    ↪ {
        print STDERR "Couldn't␣open␣files,␣please␣check␣if␣files
            ↪ ␣exist␣and␣if␣you␣have␣permission␣to␣read␣them.\n"
            ↪ ;
        print LOG "Error␣opening␣files.\n";
        exit 0;
}

# open output files, and prints header in it
open(OUTPUT, ">final.output");
print OUTPUT
"id\tMitosis\tMeiosis\tPostMeiosis\tMitosisPostmeiosis\
    ↪ tMeiosisPostmeiosis\tMitosisMeiosis\tChromosome\tXorA\
    ↪ tClass\tGroup\tHaploidGroup\tMeiosisControlForHaploids\
    ↪ tPostmeiosisControlForHaploids\tbranch\tage\tbias\tdn\tds\
    ↪ tdnds\tpn\tps\tpnps\tln\tls\tfetValue\tneutrality\talpha\n
    ↪ ";

# get dn and ds data from dNdS table
while(<DNDS>){
        $line=$_;
        chomp $line;
        @dndss=split(/\t/, $line);
        if($line=~/Symbol/){}
```

```
        if($dndss[2] != 0){
                $value_dnds=$dndss[4]/$dndss[2];
        }
        else{
                $value_dnds="NaN";
        }
        if($dndss[3] != 0){
                $value_pnps=$dndss[5]/$dndss[3];
        }
        else{
                $value_pnps="NaN";
        }
        $dnds_data{$dndss[13]}="$dndss[4]\t$dndss[2]\
           ↪ t$value_dnds\t$dndss[5]\t$dndss[3]\t$value_pnps\
           ↪ t$dndss[6]\t$dndss[7]\t$dndss[8]\t$dndss[10]\
           ↪ t$dndss[11]";
        $k++;
        #print "$dndss[13]\n";
}

# get age and bias data from age table
while(<AGE>){
        $lin=$_;
        chomp $lin;
        @ages=split(/\t/, $lin);
        if($lin=~/Symbol/){}
        if($ages[2]==0){
                $OorN="old";
        }
        else{
                $OorN="new";
        }
        $age_data{$ages[0]}="$ages[2]\t$OorN\t$ages[3]";
        $l++;
        #print "$ages[0]\n";
}

# getting id, branch, expression in all spermatogenesis phases,
   ↪ comparisson of expression in each 2 phases, group of each
```

```
    ↪ gene from expression data table
while(<EXP>){
        $lines=$_;
        chomp $lines;
        @exp=split(/\|/, $lines);
        if($lines=~/Symbol/){}
        # make new group for genes according to their classes,
            ↪ and adress if they are Autosomal or X-linked
        #if($exp[3] eq "---"){
        #$exp[3] = $exp[1];
        #}
        if($exp[4] eq "arm_X"){
                $XorA="X";
        }
        else{
                $XorA="A";
        }
        if($exp[11] eq "1" || $exp[11] eq "2" || $exp[11] eq "3"
            ↪ ){
                $Group="PostMeiotic";
        }
        elsif($exp[11] eq "4" || $exp[11] eq "7" || $exp[11] eq
            ↪ "12"){
                $Group="Meiotic";
        }
        elsif($exp[11] eq "5"){
                $Group="MeioticPostmeiotic";
        }
        elsif($exp[11] eq "6" || $exp[11] eq "8" || $exp[11] eq
            ↪ "10"){
                $Group="Mitotic";
        }
        elsif($exp[11] eq "9"){
                $Group="MitoticMeiotic";
        }
        elsif($exp[11] eq "11"){
                $Group="TheV";
        }
        elsif($exp[11] eq "13"){
```

```
                $Group="Equal";
        }
        else{
                $Group="Impossible";
        }
        if($Group eq "Meiotic" || $Group eq "MeioticPostmeiotic"
           ↪  || $Group eq "PostMeiotic"){
                $haploid = "haploid_group";
        }
        else{
                $haploid = "no";
        }
        if($Group eq "Equal" && $exp[6] >= 6.579){
                $controle7 = "control_meiosis";
        }
        else{
                $controle7 = "no";
        }
        if($Group eq "Equal" && $exp[7] >= 7.208){
                $controle8 = "control_postmeiosis";
        }
        else{
                $controle8 = "no";
        }
        $exp_data{$exp[3]}="$exp[3]\t$exp[5]\t$exp[6]\t$exp[7]\
           ↪ t$exp[8]\t$exp[9]\t$exp[10]\t$exp[4]\t$XorA\t$exp
           ↪ [11]\t$Group\t$haploid\t$controle7\t$controle8";
        $j++;
        #print "$exp[3]\n";
}
#$key = $exp[3];
# prints data in output. gives an NA (Not Avaiable) for not
   ↪ avaiable dn/ds data.
foreach $key (keys %exp_data){ # as a foreach was used, there
   ↪ are only uniq copies of each CG in the final output file.
#print "in foreach\n";
        #print "got above if\n";
        if($age_data{$key} ne ""){
                $m++;
```

```
                if($dnds_data{$key} ne ""){
                        print OUTPUT "$exp_data{$key}\t$age_data
                            ↪ {$key}\t$dnds_data{$key}\n";
                        $n++;
                }
                else{
                        print OUTPUT "$exp_data{$key}\t$age_data
                            ↪ {$key}\tNA\tNA\tNA\tNA\tNA\tNA\tNA
                            ↪ \tNA\tNA\tNA\tNA\n";
                        $o++;
                }
        }
}

# print what was done and found in log file
print LOG "\tThere␣were␣$j␣genes␣in␣Expression␣Table.\n\t\t$l␣
    ↪ genes␣in␣Age␣Table.\n\t\t$k␣genes␣in␣dNdS␣Table.\n\t$m␣
    ↪ genes␣were␣found␣in␣both␣the␣Age␣and␣Expression␣Table.\n\t
    ↪ \t$n␣genes␣were␣printed␣with␣data␣from␣all␣three␣tables.\n
    ↪ \t\t$o␣genes␣were␣printed␣with␣data␣only␣for␣the␣
    ↪ Expression␣and␣Age␣tables.\n";

# close files used and exit program.
close LOG;
close OUTPUT;
close EXP;
close AGE;
close DNDS;

exit;
```

APÊNDICE B – R Scripts

Assim como foi criado o programa em Perl colocado no capítulo anterior, também foi desenvolvido um programa em R para as análises estatíticas e construção dos gráficos apresentados, e outro para criar as simulações para fixação de um alelo recessivo em uma população diplóide e numa população haplóide. Todos os programas aqui mostrados estão disponíveis em <https://github.com/juliaraices/final_publishable>.

B.1 Estatísticas e Gráficos

```
# January 2016
# Final R program to make:
# - make controls
# - make graphs 1, 2, 3, 4, suplementals et al

divas <- read.table("input/melsubgroup_analysis_results_flydivas
    ↪ _v1.2", header=T, sep = "\t")
# reads the output
total <- read.table("output/final.output", header=T)
total$age <- factor(total$age, levels=c("old","new"))
total$dnds <- as.numeric(total$dnds)
# cross my output with flydivas output using cg column from
    ↪ flydivas and id from my output
mydivas <- merge(total, divas, by.x="id", by.y = "cg")
a.mydivas <- subset(mydivas, mydivas$XorA=="A") # only autosomal
    ↪  genes
# creates a subgroup of autosomal genes
total.a <- subset(total, total$XorA=="A")
# creates a subset for genes with dN data (positive selection
    ↪ signature)
al <- subset(total, total$dn != "NA")

#################### Figure 1 ####################
# graph with the proportion of genes in each class, and
    ↪ subsequent stats represented by *
# dados a serem usados: todos os genes dos grupos mitotico,
    ↪ meiotico e posmeiotoicos apenas
```

```
mmp <- subset(total.a, total.a$Group=='Mitotic' | total.a$Group
    ↪ =='Meiotic' | total.a$Group=='PostMeiotic')
mmp$Group = factor(mmp$Group) ### "joga fora" os factors vazios
matx = table(mmp$age, mmp$Group)
matx <- matx[,c("Mitotic", "Meiotic", "PostMeiotic")]
matp <- prop.table(matx, 1)*100
# estatisticas
mtmmtx <- matx[,c(1,2)]
mtpmtx <- matx[,c(1,3)]
mpmtx <- matx[,c(2,3)]
chisq.test(matx)
chisq.test(mtmmtx)
chisq.test(mtpmtx)
chisq.test(mpmtx)
fisher.test(matx)
fisher.test(mtmmtx)
fisher.test(mtpmtx)
fisher.test(mpmtx)

########## BW ##########
pdf("figure1_bw.pdf")#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(matp, beside=T, col=c("dimgray","gray"), xlab="
    ↪ Spermatogenesis␣phase", ylab="Percentage␣of␣genes", cex.
    ↪ lab=1.5, cex.axis=1.5, cex.main=1.5, ylim=c(10,60), xpd=F,
    ↪  xaxt="n")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotic", "Meiotic", "
    ↪ Post-Meiotic"), cex.axis=1.5, lwd=0)
text(x=c(3.6,6.6,5), y=c(30,32,52), labels=c("***", "***", "***"
    ↪ ), cex=3)
segments(2.8,28,4.2,28, cex=2, lwd=4)
segments(5.8,30,7.2,30, cex=2, lwd=4)
segments(2.8,50,7.2,50, cex=2,lwd=4)
text(x=c(1.5,2.5,4.5,5.5,7.5,8.5), y=c(rep(11,6)), labels=c(matx
    ↪ [1,1], matx[2,1], matx[1,2], matx[2,2], matx[1,3], matx
    ↪ [2,3]), cex=1.5)
legend(x=7, y=60, inset=c(-5,-0.5),legend=c("new␣genes", "old␣
    ↪ genes"), fill=c("gray","dimgray"), bty="n", cex=1.5, xpd =
    ↪  T)
```

```
dev.off()
pdf("figure1_bw_pt.pdf")#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(matp, beside=T, col=c("dimgray","gray"), xlab="Fase da
    ↪ Espermatogenese", ylab="Porcentagem de genes", cex.lab
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, ylim=c(10,60), xpd=F,
    ↪ xaxt="n")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotica", "Meiotica", "
    ↪ Pos-Meiotica"), cex.axis=1.5, lwd=0)
text(x=c(3.6,6.6,5), y=c(30,32,52), labels=c("***", "***", "***"
    ↪ ), cex=3)
segments(2.8,28,4.2,28, cex=2, lwd=4)
segments(5.8,30,7.2,30, cex=2, lwd=4)
segments(2.8,50,7.2,50, cex=2,lwd=4)
text(x=c(1.5,2.5,4.5,5.5,7.5,8.5), y=c(rep(11,6)), labels=c(matx
    ↪ [1,1], matx[2,1], matx[1,2], matx[2,2], matx[1,3], matx
    ↪ [2,3]), cex=1.5)
legend(x=6, y=60, inset=c(-5,-0.5),legend=c("genes novos", "
    ↪ genes antigos"), fill=c("gray","dimgray"), bty="n", cex
    ↪ =1.5, xpd = T)
dev.off()
########## Color ##########
pdf("figure1_color.pdf")#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(matp, beside=T, col=c("powderblue","salmon"), xlab="
    ↪ Spermatogenesis phase", ylab="Percentage of genes", cex.
    ↪ lab=1.5, cex.axis=1.5, cex.main=1.5, ylim=c(10,60), xpd=F,
    ↪  xaxt="n")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotic", "Meiotic", "
    ↪ Post-Meiotic"), cex.axis=1.5, lwd=0)
text(x=c(3.6,6.6,5), y=c(30,32,52), labels=c("***", "***", "***"
    ↪ ), cex=3)
segments(2.8,28,4.2,28, cex=2, lwd=4)
segments(5.8,30,7.2,30, cex=2, lwd=4)
segments(2.8,50,7.2,50, cex=2,lwd=4)
text(x=c(1.5,2.5,4.5,5.5,7.5,8.5), y=c(rep(11,6)), labels=c(matx
    ↪ [1,1], matx[2,1], matx[1,2], matx[2,2], matx[1,3], matx
    ↪ [2,3]), cex=1.5)
```

```
legend(x=7, y=60, inset=c(-5,-0.5),legend=c("new genes", "old genes"), fill=c("salmon","powderblue"), bty="n", cex=1.5, xpd = T)
dev.off()
pdf("figure1_color_pt.pdf")#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(matp, beside=T, col=c("powderblue","salmon"), xlab="Fase da Espermatogenese", ylab="Porcentagem de genes", cex.lab=1.5, cex.axis=1.5, cex.main=1.5, ylim=c(10,60), xpd=F, xaxt="n")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotica", "Meiotica", "Pos-Meiotica"), cex.axis=1.5, lwd=0)
text(x=c(3.6,6.6,5), y=c(30,32,52), labels=c("***", "***", "***"), cex=3)
segments(2.8,28,4.2,28, cex=2, lwd=4)
segments(5.8,30,7.2,30, cex=2, lwd=4)
segments(2.8,50,7.2,50, cex=2,lwd=4)
text(x=c(1.5,2.5,4.5,5.5,7.5,8.5), y=c(rep(11,6)), labels=c(matx[1,1], matx[2,1], matx[1,2], matx[2,2], matx[1,3], matx[2,3]), cex=1.5)
legend(x=6, y=60, inset=c(-5,-0.5),legend=c("genes novos", "genes antigos"), fill=c("salmon","powderblue"), bty="n", cex=1.5, xpd = T)
dev.off()

##################### Figure 2 ################################
# boxplot with dn/ds and alpha values
# dados a serem utilizados: genes autossomicos com valores de dn/ds e alpha, dos grupos mitotico e haploide (Meiotic + MeioticPostMeiotic + Postmeiotic)
int <- subset(al)#, al$XorA=="A")# & al$Group!="Impossible" & al$Group!="Equal" & al$Group!="TheV")
int$dnds<-as.numeric(int$dnds)
mit <- subset(int, int$Group=='Mitotic' & int$XorA=="A")
hap <- subset(int, int$XorA=="A" & int$Group=='Meiotic' | int$Group=='MeioticPostmeiotic' | int$Group=='PostMeiotic')
wilcox.test(subset(int$dnds, int$age=='old'), subset(int$dnds, int$age=='new'))
```

```
wilcox.test(subset(mit$dnds, mit$age=='old'), subset(hap$dnds,
    ↪ hap$age=='old'))
wilcox.test(mit$dnds, hap$dnds)
#wilcox.test(subset(mit$alpha, mit$age=='old'), subset(hap$alpha
    ↪ , hap$age=='old'))
#flydivas
interest <- subset(a.mydivas)#, a.mydivas$Group!="TheV" & a.
    ↪ mydivas$Group!="Impossible" & a.mydivas$Group!="Equal")
interest.a.old <- subset(interest, interest$age=='old')
interest.a.new <- subset(interest, interest$age=='new')
new.old.a.interest <- matrix(data=c(14, 274, 203,7151), ncol=2)


#new.old.a.interest <- matrix(data=c(length(subset(interest.a.
    ↪ new$id, interest.a.new$pos78=='does' | interest.a.new$pos_
    ↪ 12=='does' | interest.a.new$pos_88a=='does')), (length(
    ↪ interest.a.new$id) - length(subset(interest.a.new$id,
    ↪ interest.a.new$pos78=='does' | interest.a.new$pos_12=='
    ↪ does' | interest.a.new$pos_88a=='does'))), length(subset(
    ↪ interest.a.old$id, interest.a.old$pos78=='does' | interest
    ↪ .a.old$pos_12=='does' | interest.a.old$pos_88a=='does')),
    ↪ (length(interest.a.old$id) - length(subset(interest.a.old$
    ↪ id, interest.a.old$pos78=='does' | interest.a.old$pos_
    ↪ 12=='does' | interest.a.old$pos_88a=='does')))), ncol=2)

new.old.a.interest.p <- prop.table(new.old.a.interest, 2)*100
fisher.test(new.old.a.interest)
# testing between haploid group and mitotic group genes
hap.a <- subset(a.mydivas, a.mydivas$HaploidGroup=='haploid_
    ↪ group')
mit.a <- subset(a.mydivas, a.mydivas$Group=='Mitotic')
mit.hap.a.test <- matrix(data=c(length(subset(hap.a$id, hap.a$
    ↪ pos_12=='does' | hap.a$pos78=='does' | hap.a$pos_88a=='
    ↪ does')), (length(hap.a$id) - length(subset(hap.a$id, hap.a
    ↪ $pos78=='does' | hap.a$pos_12=='does' | hap.a$pos_88a=='
    ↪ does'))), length(subset(mit.a$id, mit.a$pos78=='does' |
    ↪ mit.a$pos_12=='does' | mit.a$pos_88a=='does')), (length(
    ↪ mit.a$id) - length(subset(mit.a$id, mit.a$pos78=='does' |
    ↪ mit.a$pos_12=='does' | mit.a$pos_88a=='does')))), ncol=2)
```

```
mit.hap.a.test <- matrix(data=c(length(subset(hap.a$id, hap.a$
    ↪ pos_12=='does')), (length(hap.a$id) - length(subset(hap.a$
    ↪ id, hap.a$pos_12=='does'))), length(subset(mit.a$id, mit.a
    ↪ $pos_12=='does')), (length(mit.a$id) - length(subset(mit.a
    ↪ $id, mit.a$pos_12=='does')))), ncol=2)

mit.hap.a.test.p <- prop.table(mit.hap.a.test, 2)*100
fisher.test(mit.hap.a.test) # there are more autosomal genes
    ↪ with signatures of poositive selection in the haploid
    ↪ group then in the mitotic group (p = 2.795e-06)
bab <- apply(new.old.a.interest.p, 1, rev)
bab <- apply(bab, 1, rev)
beb <- apply(mit.hap.a.test.p, 1, rev)
beb <- apply(beb, 1, rev)
pdf("figure2_bw.pdf")#, res=100)#, width=10, height=10)
par(mfrow=c(2,2), mar=c(4,4,3,2)+0.1)
## dN/dS ##
boxplot(subset(int$dnds, int$age=="old"), subset(int$dnds, int$
    ↪ age=="new"),
        col=c("ivory3", "ivory3"), ylim=c(0,4), ylab="dN/dS",
            ↪ xaxt="n", cex.lab=1.5, cex.axis=1.5, cex.sub=1.5,
            ↪ outline=F)
axis(1, at=c(1,2), labels=c("Old", "New"), cex.axis=1.5, lwd=0)
legend(x=0, y=-0.87, inset=0.01, legend=c("w/o selection", "w/
    ↪ selection"), fill=c("dimgrey","grey"), horiz=TRUE, cex
    ↪ =1.3, bty="n", xpd=TRUE)
text(x=c(1.5), y=c(2), labels=c("***"), cex=3)
text(x=c(1,2), y=c((boxplot.stats(subset(int$dnds, int$age=='old
    ↪ '))$stats[5]+0.1), (boxplot.stats(subset(int$dnds, int$age
    ↪ =='new'))$stats[5] + 0.1)), labels=c(length(subset(int$
    ↪ dnds, int$age=="old")), length(subset(int$dnds, int$age=="
    ↪ new"))))
mtext("(a)", 3, line=0.5, at=0, cex=1.5, xpd=TRUE)
boxplot(mit$dnds, hap$dnds, col=c("ivory3"), ylim=c(0,4), ylab="
    ↪ dN/dS", xaxt="n", cex.lab=1.5, cex.axis=1.5, cex.sub=1.5,
    ↪ outline=F)
axis(1, at=c(1,2), labels=c("Mitotic", "Haploid"), cex.axis=1.5,
    ↪  lwd=0)
```

```
text(x=c(1.5), y=c(1.5), labels=c("***"), cex=3)
text(x=c(1,2), y=c((boxplot.stats(mit$dnds)$stats[5]+0.1), (
    ↪ boxplot.stats(hap$dnds)$stats[5]+0.1)), labels = c(length(
    ↪ mit$dnds), length(hap$dnds)))
mtext("(b)", 3, line=0.5, at=0, cex=1.5, xpd=TRUE)
### % ###
barplot(bab, col = c("dimgrey","grey"), xpd=F, xlab = "", ylab="
    ↪ Genes␣Percentage", cex.lab=1.5, cex.axis = 1.5, ylim=c
    ↪ (90,100), xaxt="n")
axis(1, at=c(0.65, 1.95), labels=c("Old", "New"), cex.axis=1.5,
    ↪ lwd=0)
text(x=c(1.3), y=c(97), labels=c("*"), cex=3)
segments(0.65,96.5,1.95,96.5, cex=2, lwd=4)
text(x=c(0.65, 0.65, 1.95, 1.95), y=c(90.5, 99.5, 90.5, 99.5),
    ↪ labels=c(new.old.a.interest[2,2], new.old.a.interest[1,2],
    ↪  new.old.a.interest[2,1], new.old.a.interest[1,1]), cex=1)
mtext("(c)", 3, line=1.1, at=0, cex=1.5, xpd=TRUE)
barplot(beb, col = c("dimgrey","grey"), xpd=F, xlab = "", ylab="
    ↪ Genes␣Percentage", cex.lab=1.5, cex.axis = 1.5, ylim=c
    ↪ (90,100), xaxt="n")
axis(1, at=c(0.65, 1.95), labels=c("Mitotic", "Haploid"), cex.
    ↪ axis=1.5, lwd=0)
text(x=c(1.3), y=c(97), labels=c("***"), cex=3)
segments(0.65,96.5,1.95,96.5, cex=2, lwd=4)
text(x=c(0.65, 0.65, 1.95, 1.95), y=c(90.5, 99.5, 90.5, 99.5),
    ↪ labels=c(mit.hap.a.test[2,2], mit.hap.a.test[1,2], mit.hap
    ↪ .a.test[2,1], mit.hap.a.test[1,1]), cex=1)
mtext("(d)", 3, line=1.1, at=0, cex=1.5, xpd=TRUE)
dev.off()
pdf("figure2_bw_pt.pdf")#, res=100)#, width=10, height=10)
par(mfrow=c(2,2), mar=c(4,4,3,2)+0.1)
## dN/dS ##
boxplot(subset(int$dnds, int$age=="old"), subset(int$dnds, int$
    ↪ age=="new"),
        col=c("ivory3", "ivory3"), ylim=c(0,4), ylab="dN/dS",
            ↪ xaxt="n", cex.lab=1.5, cex.axis=1.5, cex.sub=1.5,
            ↪ outline=F)
axis(1, at=c(1,2), labels=c("Antigos", "Novos"), cex.axis=1.5,
    ↪ lwd=0)
```

```
legend(x=0, y=-0.87, inset=0.01, legend=c("sem selecao", "com selecao"), fill=c("dimgrey","grey"), horiz=TRUE, cex=1.3, bty="n", xpd=TRUE)
text(x=c(1.5), y=c(2), labels=c("***"), cex=3)
text(x=c(1,2), y=c((boxplot.stats(subset(int$dnds, int$age=='old'))$stats[5]+0.1), (boxplot.stats(subset(int$dnds, int$age=='new'))$stats[5] + 0.1)), labels=c(length(subset(int$dnds, int$age=="old")), length(subset(int$dnds, int$age=="new"))))
mtext("(a)", 3, line=0.5, at=0, cex=1.5, xpd=TRUE)
boxplot(mit$dnds, hap$dnds,
        col=c("ivory3"), ylim=c(0,4), ylab="dN/dS", xaxt="n", cex.lab=1.5, cex.axis=1.5, cex.sub=1.5, outline=F)
axis(1, at=c(1,2), labels=c("Mitotico", "Haploide"), cex.axis=1.5, lwd=0)
text(x=c(1.5), y=c(1.5), labels=c("***"), cex=3)
text(x=c(1,2), y=c((boxplot.stats(mit$dnds)$stats[5]+0.1), (boxplot.stats(hap$dnds)$stats[5]+0.1)), labels = c(length(mit$dnds), length(hap$dnds)))
mtext("(b)", 3, line=0.5, at=0, cex=1.5, xpd=TRUE)
## % ##
barplot(bab, col = c("dimgrey","grey"), xpd=F, xlab = "", ylab="Porcentagem de genes", cex.lab=1.5, cex.axis = 1.5, ylim=c(90,100), xaxt="n")
axis(1, at=c(0.65, 1.95), labels=c("Antigos", "Novos"), cex.axis=1.5, lwd=0)
text(x=c(1.3), y=c(97), labels=c("*"), cex=3)
segments(0.65,96.5,1.95,96.5, cex=2, lwd=4)
text(x=c(0.65, 0.65, 1.95, 1.95), y=c(90.5, 99.5, 90.5, 99.5), labels=c(new.old.a.interest[2,2], new.old.a.interest[1,2], new.old.a.interest[2,1], new.old.a.interest[1,1]), cex=1)
legend(x=6, y=60, inset=c(-5,-0.5),legend=c("com selecao", "sem selecao"), fill=c("dimgrey","grey"), bty="n", cex=1.5, xpd = T)
mtext("(c)", 3, line=1.1, at=0, cex=1.5, xpd=TRUE)
barplot(beb, col = c("dimgrey","grey"), xpd=F, xlab = "", ylab="Porcentagem de genes", cex.lab=1.5, cex.axis = 1.5, ylim=c(90,100), xaxt="n")
```

```
axis(1, at=c(0.65, 1.95), labels=c("Mitotico", "Haploide"), cex.
    ↪ axis=1.5, lwd=0)
text(x=c(1.3), y=c(97), labels=c("***"), cex=3)
segments(0.65,96.5,1.95,96.5, cex=2, lwd=4)
text(x=c(0.65, 0.65, 1.95, 1.95), y=c(90.5, 99.5, 90.5, 99.5),
    ↪ labels=c(mit.hap.a.test[2,2], mit.hap.a.test[1,2], mit.hap
    ↪ .a.test[2,1], mit.hap.a.test[1,1]), cex=1)
mtext("(d)", 3, line=1.1, at=0, cex=1.5, xpd=TRUE)
dev.off()
########## Color ##########
pdf("figure2_color.pdf")#, res=300)#, width=10, height=10)
par(mfrow=c(2,2), mar=c(4,4,3,2)+0.1)
## dN/dS ##
boxplot(subset(int$dnds, int$age=="old"), subset(int$dnds, int$
    ↪ age=="new"),
        col=c("plum", "plum"), ylim=c(0,4), ylab="dN/dS", xaxt="
            ↪ n", cex.lab=1.5, cex.axis=1.5, cex.sub=1.5,
            ↪ outline=F)
axis(1, at=c(1.5,3.5), labels=c("Old", "New"), cex.axis=1.5, lwd
    ↪ =0)
legend(x=0, y=-0.87, inset=0.01, legend=c("w/o selection", "w/
    ↪ selection"), fill=c("orchid4","orchid1"), horiz=TRUE, cex
    ↪ =1.3, bty="n", xpd=TRUE)
text(x=c(1.5), y=c(2), labels=c("***"), cex=3)
text(x=c(1,2), y=c((boxplot.stats(subset(int$dnds, int$age=='old
    ↪ '))$stats[5]+0.1), (boxplot.stats(subset(int$dnds, int$age
    ↪ =='new'))$stats[5] + 0.1)), labels=c(length(subset(int$
    ↪ dnds, int$age=="old")), length(subset(int$dnds, int$age=="
    ↪ new"))))
mtext("(a)", 3, line=0.5, at=0, cex=1.5, xpd=TRUE)
boxplot(mit$dnds, hap$dnds,
        col=c("plum"), ylim=c(0,4), ylab="dN/dS", xaxt="n", cex.
            ↪ lab=1.5, cex.axis=1.5, cex.sub=1.5, outline=F)
axis(1, at=c(1,2), labels=c("Mitotic", "Haploid"), cex.axis=1.5,
    ↪ lwd=0)
text(x=c(1.5), y=c(1.5), labels=c("***"), cex=3)
text(x=c(1,2), y=c((boxplot.stats(mit$dnds)$stats[5]+0.1), (
    ↪ boxplot.stats(hap$dnds)$stats[5]+0.1)), labels = c(length(
    ↪ mit$dnds), length(hap$dnds)))
```

```
mtext("(b)", 3, line=0.5, at=0, cex=1.5, xpd=TRUE)
## % ##
barplot(bab, col = c("orchid4","orchid1"), xpd=F, xlab = "",
    ↪ ylab="Genes Percentage", cex.lab=1.5, cex.axis = 1.5, ylim
    ↪ =c(90,100), xaxt="n")
axis(1, at=c(0.65, 1.95), labels=c("Old", "New"), cex.axis=1.5,
    ↪ lwd=0)
text(x=c(1.3), y=c(97), labels=c("*"), cex=3)
segments(0.65,96.5,1.95,96.5, cex=2, lwd=4)
text(x=c(0.65, 0.65, 1.95, 1.95), y=c(90.5, 99.5, 90.5, 99.5),
    ↪ labels=c(new.old.a.interest[2,2], new.old.a.interest[1,2],
    ↪  new.old.a.interest[2,1], new.old.a.interest[1,1]), cex=1)
legend(x=6, y=60, inset=c(-5,-0.5),legend=c("w/ selection", "w/o
    ↪ selection"), fill=c("orchid4","orchid1"), bty="n", cex
    ↪ =1.5, xpd = T)
mtext("(c)", 3, line=1.1, at=0, cex=1.5, xpd=TRUE)
barplot(beb, col = c("orchid4","orchid1"), xpd=F, xlab = "",
    ↪ ylab="Genes Percentage", cex.lab=1.5, cex.axis = 1.5, ylim
    ↪ =c(90,100), xaxt="n")
axis(1, at=c(0.65, 1.95), labels=c("Mitotic", "Haploid"), cex.
    ↪ axis=1.5, lwd=0)
text(x=c(1.3), y=c(97), labels=c("***"), cex=3)
segments(0.65,96.5,1.95,96.5, cex=2, lwd=4)
text(x=c(0.65, 0.65, 1.95, 1.95), y=c(90.5, 99.5, 90.5, 99.5),
    ↪ labels=c(mit.hap.a.test[2,2], mit.hap.a.test[1,2], mit.hap
    ↪ .a.test[2,1], mit.hap.a.test[1,1]), cex=1)
mtext("(d)", 3, line=1.1, at=0, cex=1.5, xpd=TRUE)
dev.off()
pdf("figure2_color_pt.pdf")#, res=300)#, width=10, height=10)
par(mfrow=c(2,2), mar=c(4,4,3,2)+0.1)
## dN/dS ##
boxplot(subset(int$dnds, int$age=="old"), subset(int$dnds, int$
    ↪ age=="new"),
        col=c("plum", "plum"), ylim=c(0,4), ylab="dN/dS", xaxt="
            ↪ n", cex.lab=1.5, cex.axis=1.5, cex.sub=1.5,
            ↪ outline=F)
axis(1, at=c(1,2), labels=c("Antigos", "Novos"), cex.axis=1.5,
    ↪ lwd=0)
```

```
legend(x=0, y=-0.87, inset=0.01, legend=c("com selecao", "sem selecao"), fill=c("orchid1","orchid4"), horiz=TRUE, cex=1.3, bty="n", xpd=TRUE)
text(x=c(1.5), y=c(2), labels=c("***"), cex=3)
text(x=c(1,2), y=c((boxplot.stats(subset(int$dnds, int$age=='old'))$stats[5]+0.1), (boxplot.stats(subset(int$dnds, int$age=='new'))$stats[5] + 0.1)), labels=c(length(subset(int$dnds, int$age=="old")), length(subset(int$dnds, int$age=="new"))))
mtext("(a)", 3, line=0.5, at=0, cex=1.5, xpd=TRUE)
boxplot(mit$dnds, hap$dnds,
        col=c("plum"), ylim=c(0,4), ylab="dN/dS", xaxt="n", cex.lab=1.5, cex.axis=1.5, cex.sub=1.5, outline=F)
axis(1, at=c(1,2), labels=c("Mitotico", "Haploide"), cex.axis=1.5, lwd=0)
text(x=c(1.5), y=c(1.5), labels=c("***"), cex=3)
text(x=c(1,2), y=c((boxplot.stats(mit$dnds)$stats[5]+0.1), (boxplot.stats(hap$dnds)$stats[5]+0.1)), labels = c(length(mit$dnds), length(hap$dnds)))
mtext("(b)", 3, line=0.5, at=0, cex=1.5, xpd=TRUE)
## % ##
barplot(bab, col = c("orchid4","orchid1"), xpd=F, xlab = "", ylab="Porcentagem de genes", cex.lab=1.5, cex.axis = 1.5, ylim=c(90,100), xaxt="n")
axis(1, at=c(0.65, 1.95), labels=c("Antigos", "Novos"), cex.axis=1.5, lwd=0)
text(x=c(1.3), y=c(97), labels=c("*"), cex=3)
segments(0.65,96.5,1.95,96.5, cex=2, lwd=4)
text(x=c(0.65, 0.65, 1.95, 1.95), y=c(90.5, 99.5, 90.5, 99.5), labels=c(new.old.a.interest[2,2], new.old.a.interest[1,2], new.old.a.interest[2,1], new.old.a.interest[1,1]), cex=1)
legend(x=6, y=60, inset=c(-5,-0.5),legend=c("com selecao", "sem selecao"), fill=c("orchid4","orchid1"), bty="n", cex=1.5, xpd = T)
mtext("(c)", 3, line=1.1, at=0, cex=1.5, xpd=TRUE)
barplot(beb, col = c("orchid4","orchid1"), xpd=F, xlab = "", ylab="Porcentagem de genes", cex.lab=1.5, cex.axis = 1.5, ylim=c(90,100), xaxt="n")
```

```
axis(1, at=c(0.65, 1.95), labels=c("Mitotico", "Haploide"), cex.
    ↪ axis=1.5, lwd=0)
text(x=c(1.3), y=c(97), labels=c("***"), cex=3)
segments(0.65,96.5,1.95,96.5, cex=2, lwd=4)
text(x=c(0.65, 0.65, 1.95, 1.95), y=c(90.5, 99.5, 90.5, 99.5),
    ↪ labels=c(mit.hap.a.test[2,2], mit.hap.a.test[1,2], mit.hap
    ↪ .a.test[2,1], mit.hap.a.test[1,1]), cex=1)
mtext("(d)", 3, line=1.1, at=0, cex=1.5, xpd=TRUE)
dev.off()

######################## Figure 3
    ↪ ################################### # new ggplot version
    ↪ doesn't work with this
# bibliotecas:
library(ggplot2)
library(plyr)
# dados a seres utilizados: dados totais X e autossomicos dos
    ↪ genes dos grupos mitotico, meiotico e pos-meiotico.
datab <- aggregate(total$id, by = list(total$age, total$XorA,
    ↪ total$Group), FUN=length)
datab <- subset(datab, datab$Group.3=="Mitotic" | datab$Group
    ↪ .3=="Meiotic" | datab$Group.3=="PostMeiotic")
datab$Proportion <- c(length(subset(total$id, total$Group=='
    ↪ Meiotic' & total$age=='old' & total$XorA=='A'))/length(
    ↪ subset(total$id, total$Group=='Meiotic' & total$age=='old'
    ↪ )), #old A mei
                        length(subset(total$id, total$Group=='
                            ↪ Meiotic' & total$age=='new' & total$
                            ↪ XorA=='A'))/length(subset(total$id,
                            ↪ total$Group=='Meiotic' & total$age==
                            ↪ 'new')), #new A mei
                        length(subset(total$id, total$Group=='
                            ↪ Meiotic' & total$age=='old' & total$
                            ↪ XorA=='X'))/length(subset(total$id,
                            ↪ total$Group=='Meiotic' & total$age==
                            ↪ 'old')), #old X mei
                        length(subset(total$id, total$Group=='
                            ↪ Meiotic' & total$age=='new' & total$
                            ↪ XorA=='X'))/length(subset(total$id,
```

```
↪ total$Group=='Meiotic' & total$age==
↪ 'new')), #new X mei
length(subset(total$id, total$Group=='
↪ Mitotic' & total$age=='old' & total$
↪ XorA=='A'))/length(subset(total$id,
↪ total$Group=='Mitotic' & total$age==
↪ 'old')), #old A mit
length(subset(total$id, total$Group=='
↪ Mitotic' & total$age=='new' & total$
↪ XorA=='A'))/length(subset(total$id,
↪ total$Group=='Mitotic' & total$age==
↪ 'new')),#new A mit
length(subset(total$id, total$Group=='
↪ Mitotic' & total$age=='old' & total$
↪ XorA=='X'))/length(subset(total$id,
↪ total$Group=='Mitotic' & total$age==
↪ 'old')), #old X mit
length(subset(total$id, total$Group=='
↪ Mitotic' & total$age=='new' & total$
↪ XorA=='X'))/length(subset(total$id,
↪ total$Group=='Mitotic' & total$age==
↪ 'new')),#new X mit
length(subset(total$id, total$Group=='
↪ PostMeiotic' & total$age=='old' &
↪ total$XorA=='A'))/length(subset(
↪ total$id, total$Group=='PostMeiotic'
↪ & total$age=='old')), #old A pos
length(subset(total$id, total$Group=='
↪ PostMeiotic' & total$age=='new' &
↪ total$XorA=='A'))/length(subset(
↪ total$id, total$Group=='PostMeiotic'
↪ & total$age=='new')), #new A pos
length(subset(total$id, total$Group=='
↪ PostMeiotic' & total$age=='old' &
↪ total$XorA=='X'))/length(subset(
↪ total$id, total$Group=='PostMeiotic'
↪ & total$age=='old')), #old X pos
length(subset(total$id, total$Group=='
↪ PostMeiotic' & total$age=='new' &
```

```
                                    ↪ total$XorA=='X'))/length(subset(
                                    ↪ total$id, total$Group=='PostMeiotic'
                                    ↪ & total$age=='new'))) #new X pos
datab <- rename(datab, c("Group.1"="Age", "Group.2"="Chromosome"
    ↪ , "Group.3"="Group", "x"="Count"))
datab$Group <- factor(datab$Group, levels=c("Mitotic", "Meiotic"
    ↪ , "PostMeiotic"))
x.datab <- subset(datab, datab$Chromosome=="X")
a.datab <- subset(datab, datab$Chromosome=="A")
aall <- length(subset(total$XorA, total$XorA=="A" & total$Group!
    ↪ ="Equal"))/length(subset(total$XorA, total$Group!="Equal")
    ↪ )
newaall <- length(subset(total$XorA, total$XorA=="A" & total$age
    ↪ =="new" & total$Group!="Equal"))/length(subset(total$XorA,
    ↪ total$age=="new" & total$Group!="Equal"))
x.count <- c(subset(x.datab$Count, x.datab$Age=="old" & x.datab$
    ↪ Group=="Mitotic"), subset(x.datab$Count, x.datab$Age=="old
    ↪ " & x.datab$Group!="Mitotic"), subset(x.datab$Count, x.
    ↪ datab$Age=="new" & x.datab$Group=="Mitotic"), subset(x.
    ↪ datab$Count, x.datab$Age=="new" & x.datab$Group!="Mitotic"
    ↪ ))
a.count <- c(subset(a.datab$Count, a.datab$Age=="old" & a.datab$
    ↪ Group=="Mitotic"), subset(a.datab$Count, a.datab$Age=="old
    ↪ " & a.datab$Group!="Mitotic"), subset(a.datab$Count, x.
    ↪ datab$Age=="new" & a.datab$Group=="Mitotic"), subset(a.
    ↪ datab$Count, a.datab$Age=="new" & a.datab$Group!="Mitotic"
    ↪ ))
# estatisticas:
fisher.test(matrix(data=c(subset(datab$Count, datab$Age=='old' &
    ↪ datab$Group=='Mitotic'), sum(subset(datab$Count, datab$
    ↪ Age=='old' & datab$Group!='Mitotic' & datab$Chromosome=='A
    ↪ '')), sum(subset(datab$Count, datab$Age=='old' & datab$
    ↪ Group!='Mitotic' & datab$Chromosome=='X'))), nrow = 2)) #
    ↪ tot: 0.3341 | mmpp: 0.09214 | al: 0.1246
chisq.test(matrix(data=c(subset(datab$Count, datab$Age=='old' &
    ↪ datab$Group=='Mitotic'), sum(subset(datab$Count, datab$Age
    ↪ =='old' & datab$Group!='Mitotic' & datab$Chromosome=='A'))
    ↪ , sum(subset(datab$Count, datab$Age=='old' & datab$Group!=
    ↪ 'Mitotic' & datab$Chromosome=='X'))), nrow = 2)) # tot:
```

```
↪ 0.3389 | mmpp: 0.09631 | al: 0.132
fisher.test(matrix(data=c(subset(datab$Count, datab$Age=='old' &
↪ datab$Group=='Meiotic'), sum(subset(datab$Count, datab$
↪ Age=='old' & datab$Group!='Meiotic' & datab$Chromosome=='A
↪ '')), sum(subset(datab$Count, datab$Age=='old' & datab$
↪ Group!='Meiotic' & datab$Chromosome=='X'))), nrow = 2)) #
↪ tot: 0.001026 | mmpp: 0.001512 | al: 0.002704
chisq.test(matrix(data=c(subset(datab$Count, datab$Age=='old' &
↪ datab$Group=='Meiotic'), sum(subset(datab$Count, datab$Age
↪ =='old' & datab$Group!='Meiotic' & datab$Chromosome=='A'))
↪ , sum(subset(datab$Count, datab$Age=='old' & datab$Group!=
↪ 'Meiotic' & datab$Chromosome=='X'))), nrow = 2)) # tot:
↪ 0.001365 | mmpp: 0.002063 | al: 0.003811
fisher.test(matrix(data=c(subset(datab$Count, datab$Age=='old' &
↪ datab$Group=='PostMeiotic'), sum(subset(datab$Count,
↪ datab$Age=='old' & datab$Group!='PostMeiotic' & datab$
↪ Chromosome=='A')), sum(subset(datab$Count, datab$Age=='old
↪ ' & datab$Group!='PostMeiotic' & datab$Chromosome=='X'))),
↪ nrow = 2)) # tot: 0.6651 | mmpp: 0.4619 | al: 0.4684
chisq.test(matrix(data=c(subset(datab$Count, datab$Age=='old' &
↪ datab$Group=='PostMeiotic'), sum(subset(datab$Count, datab
↪ $Age=='old' & datab$Group!='PostMeiotic' & datab$
↪ Chromosome=='A')), sum(subset(datab$Count, datab$Age=='old
↪ ' & datab$Group!='PostMeiotic' & datab$Chromosome=='X'))),
↪ nrow = 2)) # tot: 0.6855 | mmpp: 0.4933 | al: 0.501
fisher.test(matrix(data=c(subset(datab$Count, datab$Age=='new' &
↪ datab$Group=='Mitotic'), sum(subset(datab$Count, datab$
↪ Age=='new' & datab$Group!='Mitotic' & datab$Chromosome=='A
↪ '')), sum(subset(datab$Count, datab$Age=='new' & datab$
↪ Group!='Mitotic' & datab$Chromosome=='X'))), nrow = 2)) #
↪ tot: 0.06279 | mmpp: 0.0504 | al: 0.01002
chisq.test(matrix(data=c(subset(datab$Count, datab$Age=='new' &
↪ datab$Group=='Mitotic'), sum(subset(datab$Count, datab$Age
↪ =='new' & datab$Group!='Mitotic' & datab$Chromosome=='A'))
↪ , sum(subset(datab$Count, datab$Age=='new' & datab$Group!=
↪ 'Mitotic' & datab$Chromosome=='X'))), nrow = 2)) # tot:
↪ 0.07479 | mmpp: 0.05897 | al: 0.01308
fisher.test(matrix(data=c(subset(datab$Count, datab$Age=='new' &
↪ datab$Group=='Meiotic'), sum(subset(datab$Count, datab$
```

```
↪ Age=='new' & datab$Group!='Meiotic' & datab$Chromosome=='A
↪ ')), sum(subset(datab$Count, datab$Age=='new' & datab$
↪ Group!='Meiotic' & datab$Chromosome=='X'))), nrow = 2)) #
↪ tot: 0.7945 | mmpp: 0.668 | al: 0.6996
chisq.test(matrix(data=c(subset(datab$Count, datab$Age=='new' &
↪ datab$Group=='Meiotic'), sum(subset(datab$Count, datab$Age
↪ =='new' & datab$Group!='Meiotic' & datab$Chromosome=='A'))
↪ , sum(subset(datab$Count, datab$Age=='new' & datab$Group!=
↪ 'Meiotic' & datab$Chromosome=='X'))), nrow = 2)) # tot:
↪ 0.8164 | mmpp: 0.7776 | al: 0.7194
fisher.test(matrix(data=c(subset(datab$Count, datab$Age=='new' &
↪  datab$Group=='PostMeiotic'), sum(subset(datab$Count,
↪ datab$Age=='new' & datab$Group!='PostMeiotic' & datab$
↪ Chromosome=='A')), sum(subset(datab$Count, datab$Age=='new
↪ ' & datab$Group!='PostMeiotic' & datab$Chromosome=='X'))),
↪  nrow = 2)) # tot: 0.016 | mmpp: 0.01609 | al: 0.06117
chisq.test(matrix(data=c(subset(datab$Count, datab$Age=='new' &
↪ datab$Group=='PostMeiotic'), sum(subset(datab$Count, datab
↪ $Age=='new' & datab$Group!='PostMeiotic' & datab$
↪ Chromosome=='A')), sum(subset(datab$Count, datab$Age=='new
↪ ' & datab$Group!='PostMeiotic' & datab$Chromosome=='X'))),
↪  nrow = 2)) # tot: 0.02671 | mmpp: 0.02031 | al: 0.07154
# dados e funcoes pros graficos:
hline.data <- data.frame(z = c(aall, newaall), Age = c("old","
↪ new"))
labs.data <- data.frame(s=c(2, 1, 3), f=c(0.878, 0.74, 0.905), z
↪  = c("**", "", "*"), Age = c("old","new", "new"),
↪ Chromosome=c("A", "A", "A"))
ps <- data.frame(a=c(1,3,1,2), b=c(0.834, 0.834, 0.74, 0.815), d
↪ =c("ns", "ns", "ns", "ns"), Age=c("old", "old", "new", "
↪ new"), Chromosome=c("A", "A", "A", "A"))
texto <- data.frame(x=c(rep(c(1,2,3),4)), y=c(rep(0.99, 6), rep
↪ (0.61, 6)), lab=c(x.count, a.count), Age=c(rep(c(rep("old"
↪ ,3), rep("new", 3)),2)))
facet_names <- list(
    'old'="Old Genes",
    'new'="New Genes")
facet_names_pt <- list(
    'old'="Genes Antigos",
```

```
    'new'="Genes Novos")
facet_labeller <- function(variable, value){
    return(facet_names[value])
}
facet_labeller_pt <- function(variable, value){
    return(facet_names_pt[value])
}
########## BW ##########
pdf("figure3_bw.pdf", width=15, height=10)#, res=300) # width
  ↪ =22, height=10,
ggplot(datab, aes(x=Group, y=Proportion, fill=Chromosome)) +
    geom_bar(position='stack', stat='identity') +
    facet_grid(.~Age, labeller=facet_labeller) +
    coord_cartesian(ylim=c(0.6, 1)) +
    theme(text = element_text(size=30)) +
    scale_x_discrete(name="") +
    geom_hline(aes(yintercept =z), hline.data) +
    geom_text(aes(x=x,y=y,label=lab), texto, size=7, inherit.aes
      ↪ =F) +
    geom_text(aes(x=s, y=f, label=z), labs.data, size=20) +
    geom_text(aes(x=a, y=b, label=d), ps, size=5) +
    scale_fill_manual(values=c("Grey 70", "Gray 50"), name="",
      ↪ breaks=c("A", "X"), labels=c("Autosomal gene", "X-
      ↪ linked gene"))
dev.off()
pdf("figure3_bw_pt.pdf", width=15, height=10)#, res=300) # width
  ↪ =22, height=10,
ggplot(datab, aes(x=Group, y=Proportion, fill=Chromosome)) +
    geom_bar(position='stack', stat='identity') +
    facet_grid(.~Age, labeller=facet_labeller_pt) +
    coord_cartesian(ylim=c(0.6, 1)) +
    theme(text = element_text(size=30)) +
    scale_x_discrete(name="") +
    geom_hline(aes(yintercept =z), hline.data) +
    geom_text(aes(x=x,y=y,label=lab), texto, size=7, inherit.aes
      ↪ =F) +
    geom_text(aes(x=s, y=f, label=z), labs.data, size=20) +
    geom_text(aes(x=a, y=b, label=d), ps, size=5) +
```

```
    scale_fill_manual(values=c("Grey 70", "Gray 50"), name="",
        breaks=c("A", "X"), labels=c("Genes Autossomicos", "
        Genes ligados ao X"))
dev.off()
########## Color ##########
pdf("figure3_color.pdf", width=15, height=10)#, res=300) # width
    =22, height=10,
ggplot(datab, aes(x=Group, y=Proportion, fill=Chromosome)) +
    facet_grid(.~Age, labeller=facet_labeller) +
    geom_bar(position='stack', stat='identity') +
    coord_cartesian(ylim=c(0.6, 1)) +
    theme(text = element_text(size=30)) +
    scale_x_discrete(name="") +
    geom_hline(aes(yintercept =z), hline.data) +
    geom_text(aes(x=s, y=f, label=z), labs.data, size=20) +
    geom_text(aes(x=x,y=y,label=lab), texto, size=7, inherit.aes
        =F) +
    geom_text(aes(x=a, y=b, label=d), ps, size=5) +
    scale_fill_manual(values=c("Khaki", "Thistle"), name="",
        breaks=c("A", "X"), labels=c("Autosomal gene", "X-
        linked gene"))
dev.off()
pdf("figure3_color_pt.pdf", width=15, height=10)#, res=300) #
    width=22, height=10,
ggplot(datab, aes(x=Group, y=Proportion, fill=Chromosome)) +
    facet_grid(.~Age, labeller=facet_labeller) +
    geom_bar(position='stack', stat='identity') +
    coord_cartesian(ylim=c(0.6, 1)) +
    theme(text = element_text(size=30)) +
    scale_x_discrete(name="") +
    geom_hline(aes(yintercept =z), hline.data) +
    geom_text(aes(x=s, y=f, label=z), labs.data, size=20) +
    geom_text(aes(x=x,y=y,label=lab), texto, size=7, inherit.aes
        =F) +
    geom_text(aes(x=a, y=b, label=d), ps, size=5) +
    scale_fill_manual(values=c("Khaki", "Thistle"), name="",
        breaks=c("A", "X"), labels=c("Genes Autossomicos", "
        Genes ligados ao X"))
dev.off()
```

```
######################## Figure 4
    ↪ ##################################
# dados a serem usados: genes totais autossomicos para todas as
    ↪ classes usadas
total.a$Class <- as.numeric(total.a$Class)
classes <- subset(total.a, total.a$Class<=13)
classes$Class = factor(classes$Class) ### "joga fora" os factors
    ↪  vazios
clx = table(classes$age, classes$Class)
clp <- prop.table(clx, 1)*100
########## BW ##########
pdf("figure4_bw.pdf", width = 9)#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(clp, beside=T, col=c("dimgray","gray"), xlab="Numerical␣
    ↪ Class", ylab="Percentage␣of␣genes", cex.lab=1.5, cex.axis
    ↪ =1.5, cex.main=1.5, ylim=c(0,40), xpd=F, xaxt="n")
axis(1, at=c(seq(2,38, by=3)), labels=c(seq(1,13,by=1)), cex.
    ↪ axis=1.5, lwd=0)
legend(x=3, y=35, inset=c(-5,-0.5),legend=c("new␣genes", "old␣
    ↪ genes"), fill=c("gray","dimgray"), bty="n", cex=1.5, xpd =
    ↪  T)
dev.off()
pdf("figure4_bw_pt.pdf", width = 9)#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(clp, beside=T, col=c("dimgray","gray"), xlab="Classes␣
    ↪ Numericas", ylab="Porcentagem␣de␣genes", cex.lab=1.5, cex.
    ↪ axis=1.5, cex.main=1.5, ylim=c(0,40), xpd=F, xaxt="n")
axis(1, at=c(seq(2,38, by=3)), labels=c(seq(1,13,by=1)), cex.
    ↪ axis=1.5, lwd=0)
legend(x=3, y=35, inset=c(-5,-0.5),legend=c("genes␣novos", "
    ↪ genes␣antigos"), fill=c("gray","dimgray"), bty="n", cex
    ↪ =1.5, xpd = T)
dev.off()
########## Color ##########
pdf("figure4_color.pdf", width = 9)#, res=300)
par(mar=c(6,5,4,2)+0.1)
```

```
barplot(clp, beside=T, col=c("powderblue","salmon"), xlab="
    ↪ Numerical Class", ylab="Percentage of genes", cex.lab=1.5,
    ↪  cex.axis=1.5, cex.main=1.5, ylim=c(0,40), xpd=F, xaxt="n"
    ↪ )
axis(1, at=c(seq(2,38, by=3)), labels=c(seq(1,13,by=1)), cex.
    ↪ axis=1.5, lwd=0)
legend(x=3, y=35, inset=c(-5,-0.5),legend=c("new genes", "old 
    ↪ genes"), fill=c("salmon","powderblue"), bty="n", cex=1.5,
    ↪ xpd = T)
dev.off()
pdf("figure4_color_pt.pdf", width = 9)#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(clp, beside=T, col=c("powderblue","salmon"), xlab="
    ↪ Classes Numericas", ylab="Porcentagem de genes", cex.lab
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,40), xpd=F,
    ↪ xaxt="n")
axis(1, at=c(seq(2,38, by=3)), labels=c(seq(1,13,by=1)), cex.
    ↪ axis=1.5, lwd=0)
legend(x=3, y=35, inset=c(-5,-0.5),legend=c("genes novos", "
    ↪ genes antigos"), fill=c("salmon","powderblue"), bty="n",
    ↪ cex=1.5, xpd = T)
dev.off()



########################## Figure 5
    ↪ ##################################
# dados a serem usados: genes totais autossomicos para todas as
    ↪ classes
total.a$Class <- factor(total.a$Class)
ttx = table(total.a$age, total.a$Class)
ttp <- prop.table(ttx, 1)*100
########## BW ##########
pdf("figure5_bw.pdf", width = 13)#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(ttp, beside=T, col=c("dimgray","gray"), xlab="Numerical 
    ↪ Class", ylab="Percentage of genes", cex.lab=1.5, cex.axis
    ↪ =1.5, cex.main=1.5, ylim=c(0,40), xpd=F, xaxt="n")
```

```
axis(1, at=c(seq(2,56, by=3)), labels=c(seq(1,19,by=1)), cex.
    ↪ axis=1.5, lwd=0)
legend(x=3, y=35, inset=c(-5,-0.5),legend=c("new genes", "old
    ↪ genes"), fill=c("gray","dimgray"), bty="n", cex=1.5, xpd =
    ↪  T)
dev.off()
pdf("figure5_bw_pt.pdf", width = 13)#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(ttp, beside=T, col=c("dimgray","gray"), xlab="Classes
    ↪ Numericas", ylab="Porcentagem de genes", cex.lab=1.5, cex.
    ↪ axis=1.5, cex.main=1.5, ylim=c(0,40), xpd=F, xaxt="n")
axis(1, at=c(seq(2,56, by=3)), labels=c(seq(1,19,by=1)), cex.
    ↪ axis=1.5, lwd=0)
legend(x=3, y=35, inset=c(-5,-0.5),legend=c("genes novos", "
    ↪ genes antigos"), fill=c("gray","dimgray"), bty="n", cex
    ↪ =1.5, xpd = T)
dev.off()
########## Color ##########
pdf("figure5_color.pdf", width = 13)#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(ttp, beside=T, col=c("powderblue","salmon"), xlab="
    ↪ Numerical Class", ylab="Percentage of genes", cex.lab=1.5,
    ↪  cex.axis=1.5, cex.main=1.5, ylim=c(0,40), xpd=F, xaxt="n"
    ↪ )
axis(1, at=c(seq(2,56, by=3)), labels=c(seq(1,19,by=1)), cex.
    ↪ axis=1.5, lwd=0)
legend(x=3, y=35, inset=c(-5,-0.5),legend=c("new genes", "old
    ↪ genes"), fill=c("salmon","powderblue"), bty="n", cex=1.5,
    ↪ xpd = T)
dev.off()
pdf("figure5_color_pt.pdf", width = 13)#, res=300)
par(mar=c(6,5,4,2)+0.1)
barplot(ttp, beside=T, col=c("powderblue","salmon"), xlab="
    ↪ Classes Numericas", ylab="Porcentagem de genes", cex.lab
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,40), xpd=F,
    ↪ xaxt="n")
axis(1, at=c(seq(2,56, by=3)), labels=c(seq(1,19,by=1)), cex.
    ↪ axis=1.5, lwd=0)
```

```
legend(x=3, y=35, inset=c(-5,-0.5),legend=c("genes novos", "genes antigos"), fill=c("salmon","powderblue"), bty="n", cex=1.5, xpd = T)
dev.off()


######################### Figure 6 ##################################
# dados a serem usados: genes ligados ao X e autossomicos para as classes de interesse
xatb <- subset(total, total$Group=='Mitotic' | total$Group=='Meiotic' | total$Group=='PostMeiotic')
xatb$Group = factor(xatb$Group) ### "joga fora" os factors vazios
xax <- table(xatb$XorA, xatb$Group)
xap <- prop.table(xax, 1)*100
xaxn <- table(subset(xatb$XorA, xatb$age=='new'), subset(xatb$Group, xatb$age=='new'))
xapn <- prop.table(xaxn, 1)*100
xaxo <- table(subset(xatb$XorA, xatb$age=='old'), subset(xatb$Group, xatb$age=='old'))
xapo <- prop.table(xaxo, 1)*100
########## BW ##########
pdf("figure6_bw.pdf", width = 13)#, res=300)
par(mfrow=c(1,3), mar=c(6,5,4,2)+0.1)
barplot(xap, beside=T, col=c("gray","dimgray"), xlab="Spermatogenesis Phase", ylab="Percentage of genes", cex.lab=1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F, xaxt="n", main = "All genes")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotic", "Meiotic", "PostMeiotic"), cex.axis=1.5, lwd=0)
barplot(xapn, beside=T, col=c("gray","dimgray"), xlab="Spermatogenesis Phase", ylab="Percentage of genes", cex.lab=1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F, xaxt="n", main="New Genes")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotic", "Meiotic", "PostMeiotic"), cex.axis=1.5, lwd=0)
legend(x=1.5, y=60, inset=c(-5,-5),legend=c("autossomic genes", "X-linked genes"), fill=c("gray","dimgray"), bty="n", cex
```

```
    ↪ =1.5, xpd = T)
barplot(xapo, beside=T, col=c("gray","dimgray"), xlab="
    ↪ Spermatogenesis Phase", ylab="Percentage of genes", cex.
    ↪ lab=1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main="Old Genes")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotic", "Meiotic", "
    ↪ PostMeiotic"), cex.axis=1.5, lwd=0)
dev.off()
pdf("figure6_bw_pt.pdf", width = 13)#, res=300)
par(mfrow=c(1,3), mar=c(6,5,4,2)+0.1)
barplot(xap, beside=T, col=c("gray","dimgray"), xlab="Fase da
    ↪ Espermatogenese", ylab="Porcentagem de genes", cex.lab
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main = "Todos os genes")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotica", "Meiotica", "
    ↪ PosMeiotica"), cex.axis=1.5, lwd=0)
barplot(xapn, beside=T, col=c("gray","dimgray"), xlab="Fase da
    ↪ Espermatogenese", ylab="Porcentagem de genes", cex.lab
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main="Genes Novos")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotica", "Meiotica", "
    ↪ PosMeiotica"), cex.axis=1.5, lwd=0)
legend(x=1.5, y=60, inset=c(-5,-5),legend=c("genes autossomicos"
    ↪ , "genes ligados ao X"), fill=c("gray","dimgray"), bty="n"
    ↪ , cex=1.5, xpd = T)
barplot(xapo, beside=T, col=c("gray","dimgray"), xlab="Fase da
    ↪ Espermatogenese", ylab="Porcentagem de genes", cex.lab
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main="Genes Antigos")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotica", "Meiotica", "
    ↪ PosMeiotica"), cex.axis=1.5, lwd=0)
dev.off()
########## Color ##########
pdf("figure6_color.pdf", width = 13)#, res=300)
par(mfrow=c(1,3), mar=c(6,5,4,2)+0.1)
barplot(xap, beside=T, col=c("Khaki","Thistle"), xlab="
    ↪ Spermatogenesis Phase", ylab="Percentage of genes", cex.
    ↪ lab=1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main = "All genes")
```

```
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotic", "Meiotic", "
    ↪ PostMeiotic"), cex.axis=1.5, lwd=0)
barplot(xapn, beside=T, col=c("Khaki","Thistle"), xlab="
    ↪ Spermatogenesis Phase", ylab="Percentage of genes", cex.
    ↪ lab=1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main="New Genes")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotic", "Meiotic", "
    ↪ PostMeiotic"), cex.axis=1.5, lwd=0)
legend(x=1.5, y=60, inset=c(-5,-5),legend=c("autossomic genes",
    ↪ "X-linked genes"), fill=c("Khaki","Thistle"), bty="n", cex
    ↪ =1.5, xpd = T)
barplot(xapo, beside=T, col=c("Khaki","Thistle"), xlab="
    ↪ Spermatogenesis Phase", ylab="Percentage of genes", cex.
    ↪ lab=1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main="Old Genes")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotic", "Meiotic", "
    ↪ PostMeiotic"), cex.axis=1.5, lwd=0)
dev.off()
pdf("figure6_color_pt.pdf", width = 13)#, res=300)
par(mfrow=c(1,3), mar=c(6,5,4,2)+0.1)
barplot(xap, beside=T, col=c("Khaki","Thistle"), xlab="Fase da
    ↪ Espermatogenese", ylab="Porcentagem de genes", cex.lab
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main = "Todos os genes")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotica", "Meiotica", "
    ↪ PosMeiotica"), cex.axis=1.5, lwd=0)
barplot(xapn, beside=T, col=c("Khaki","Thistle"), xlab="Fase da
    ↪ Espermatogenese", ylab="Porcentagem de genes", cex.lab
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main="Genes novos")
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotica", "Meiotica", "
    ↪ PosMeiotica"), cex.axis=1.5, lwd=0)
legend(x=1.5, y=60, inset=c(-5,-5),legend=c("genes autossomicos"
    ↪ , "genes ligados ao X"), fill=c("Khaki","Thistle"), bty="n
    ↪ ", cex=1.5, xpd = T)
barplot(xapo, beside=T, col=c("Khaki","Thistle"), xlab="Fase da
    ↪ Espermatogenese", ylab="Porcentagem de genes", cex.lab
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, ylim=c(0,65), xpd=F,
    ↪ xaxt="n", main="Genes antigos")
```

```
axis(1, at=c(seq(2,8, by=3)), labels=c("Mitotica", "Meiotica", "
    ↪ PosMeiotica"), cex.axis=1.5, lwd=0)
dev.off()


########## Figure 7 ##########
# grafico da expressao de genes novos ao longo da
    ↪ espermatogenese
# bibliotecas necessarias:
library(ggplot2)
library(gridExtra)
# dados usados: genes autossomicos das classes 1 a 10 e 12
total.a$Class <- factor(total.a$Class)
interest <- subset(total.a, total.a$Class!='13' & total.a$Class!
    ↪ ='19' & total.a$Class!='18' & total.a$Class!='17' & total.
    ↪ a$Class!='16' & total.a$Class!='15' & total.a$Class!='14')
interest$Class <- factor(interest$Class)
# estatisticas
wilcox.test(interest$Mitosis~interest$age)
wilcox.test(interest$Meiosis~interest$age)
wilcox.test(interest$PostMeiosis~interest$age)
# graficos:
########## BW Eng ##########
mit <- ggplot(data.frame(interest), aes(x=interest$age, y=
    ↪ interest$Mitosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("dimgrey", "grey")) +
    labs(x="Genes' age", y="Expression during Mitosis") +
    theme(legend.position = "none") + annotate("text", x=c
        ↪ (1.5,1,2), y=c(13.5,3,3), label=c("ns",length(subset(
        ↪ interest$Mitosis, interest$age=="old")),length(subset(
        ↪ interest$Mitosis, interest$age=="new")))) +ggtitle("")
        ↪ +
    scale_y_continuous(limits = c(3, 15))
mei <- ggplot(data.frame(interest), aes(x=interest$age, y=
    ↪ interest$Meiosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("dimgrey", "grey")) +
    labs(x="Genes' age", y="Expression during Meiosis") +
```

```
    theme(legend.position = "none") + annotate("text", x=c
      ↪ (1.5,1,2), y=c(13.5,3,3), label=c("***", length(subset
      ↪ (interest$Meiosis, interest$age=="old")), length(
      ↪ subset(interest$Meiosis, interest$age=="new")))) +
      ↪ ggtitle("") +
    scale_y_continuous(limits = c(3, 15))
pm <- ggplot(data.frame(interest), aes(x=interest$age, y=
  ↪ interest$PostMeiosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("dimgrey", "grey")) +
    labs(x="Genes' age", y="Expression during PostMeiosis") +
    theme(legend.position = "none") + annotate("text", x=c
      ↪ (1.5,1,2), y=c(13.5,3,3), label=c("***", length(subset
      ↪ (interest$PostMeiosis, interest$age=="old")), length(
      ↪ subset(interest$PostMeiosis, interest$age=="new")))) +
      ↪ ggtitle("") +
    scale_y_continuous(limits = c(3, 15))
pdf("figure7_bw.pdf")
grid.arrange(arrangeGrob(mit, mei, pm,  nrow=1, ncol=3))
dev.off()
########## BW Pt ##########
mit <- ggplot(data.frame(interest), aes(x=interest$age, y=
  ↪ interest$Mitosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("dimgrey", "grey")) +
    labs(x="Idade dos Genes", y="Expressao durante a Mitose") +
    theme(legend.position = "none") + annotate("text", x=c
      ↪ (1.5,1,2), y=c(13.5,3,3), label=c("ns",length(subset(
      ↪ interest$Mitosis, interest$age=="old")),length(subset(
      ↪ interest$Mitosis, interest$age=="new")))) +ggtitle("")
      ↪ +
    scale_y_continuous(limits = c(3, 15))
mei <- ggplot(data.frame(interest), aes(x=interest$age, y=
  ↪ interest$Meiosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("dimgrey", "grey")) +
    labs(x="Idade dos Genes", y="Expressao durante a Meiose") +
    theme(legend.position = "none") + annotate("text", x=c
      ↪ (1.5,1,2), y=c(13.5,3,3), label=c("***", length(subset
```

```
        ↪ (interest$Meiosis, interest$age=="old")), length(
        ↪ subset(interest$Meiosis, interest$age=="new")))) +
        ↪ ggtitle("") +
    scale_y_continuous(limits = c(3, 15))
pm <- ggplot(data.frame(interest), aes(x=interest$age, y=
    ↪ interest$PostMeiosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("dimgrey", "grey")) +
    labs(x="Idade dos Genes", y="Expressao durante a Pos-Meiose"
        ↪ ) +
    theme(legend.position = "none") + annotate("text", x=c
        ↪ (1.5,1,2), y=c(13.5,3,3), label=c("***", length(subset
        ↪ (interest$PostMeiosis, interest$age=="old")), length(
        ↪ subset(interest$PostMeiosis, interest$age=="new")))) +
        ↪ ggtitle("") +
    scale_y_continuous(limits = c(3, 15))
pdf("figure7_bw_pt.pdf")
grid.arrange(arrangeGrob(mit, mei, pm,  nrow=1, ncol=3))
dev.off()
########## Color Eng ##########
mit <- ggplot(data.frame(interest), aes(x=interest$age, y=
    ↪ interest$Mitosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("powderblue", "salmon")) +
    labs(x="Genes' age", y="Expression during Mitosis") +
    theme(legend.position = "none") + annotate("text", x=c
        ↪ (1.5,1,2), y=c(13.5,3,3), label=c("ns",length(subset(
        ↪ interest$Mitosis, interest$age=="old")),length(subset(
        ↪ interest$Mitosis, interest$age=="new")))) +ggtitle("")
        ↪ +
    scale_y_continuous(limits = c(3, 15))
mei <- ggplot(data.frame(interest), aes(x=interest$age, y=
    ↪ interest$Meiosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("powderblue", "salmon")) +
    labs(x="Genes' age", y="Expression during Meiosis") +
    theme(legend.position = "none") + annotate("text", x=c
        ↪ (1.5,1,2), y=c(13.5,3,3), label=c("***", length(subset
        ↪ (interest$Meiosis, interest$age=="old")), length(
```

```
        ↪ subset(interest$Meiosis, interest$age=="new")))) +
        ↪ ggtitle("") +
    scale_y_continuous(limits = c(3, 15))
pm <- ggplot(data.frame(interest), aes(x=interest$age, y=
    ↪ interest$PostMeiosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("powderblue", "salmon")) +
    labs(x="Genes' age", y="Expression during PostMeiosis") +
    theme(legend.position = "none") + annotate("text", x=c
        ↪ (1.5,1,2), y=c(13.5,3,3), label=c("***", length(subset
        ↪ (interest$PostMeiosis, interest$age=="old")), length(
        ↪ subset(interest$PostMeiosis, interest$age=="new")))) +
        ↪ ggtitle("") +
    scale_y_continuous(limits = c(3, 15))
pdf("figure7_color.pdf")
grid.arrange(arrangeGrob(mit, mei, pm,  nrow=1, ncol=3))
dev.off()
########## Color Pt ##########
mit <- ggplot(data.frame(interest), aes(x=interest$age, y=
    ↪ interest$Mitosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("powderblue", "salmon")) +
    labs(x="Idade dos Genes", y="Expressao durante a Mitose") +
    theme(legend.position = "none") + annotate("text", x=c
        ↪ (1.5,1,2), y=c(13.5,3,3), label=c("ns",length(subset(
        ↪ interest$Mitosis, interest$age=="old")),length(subset(
        ↪ interest$Mitosis, interest$age=="new")))) +ggtitle("")
        ↪ +
    scale_y_continuous(limits = c(3, 15))
mei <- ggplot(data.frame(interest), aes(x=interest$age, y=
    ↪ interest$Meiosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("powderblue", "salmon")) +
    labs(x="Idade dos Genes", y="Expressao durante a Meiose") +
    theme(legend.position = "none") + annotate("text", x=c
        ↪ (1.5,1,2), y=c(13.5,3,3), label=c("***", length(subset
        ↪ (interest$Meiosis, interest$age=="old")), length(
        ↪ subset(interest$Meiosis, interest$age=="new")))) +
        ↪ ggtitle("") +
```

```
    scale_y_continuous(limits = c(3, 15))
pm <- ggplot(data.frame(interest), aes(x=interest$age, y=
    ↪ interest$PostMeiosis, color=age)) +
    geom_boxplot(notch = T) +
    scale_color_manual(values=c("powderblue", "salmon")) +
    labs(x="Idade dos Genes", y="Expressao durante a Pos-Meiose"
        ↪ ) +
    theme(legend.position = "none") + annotate("text", x=c
        ↪ (1.5,1,2), y=c(13.5,3,3), label=c("***", length(subset
        ↪ (interest$PostMeiosis, interest$age=="old")), length(
        ↪ subset(interest$PostMeiosis, interest$age=="new")))) +
        ↪ ggtitle("") +
    scale_y_continuous(limits = c(3, 15))
pdf("figure7_color_pt.pdf")
grid.arrange(arrangeGrob(mit, mei, pm,  nrow=1, ncol=3))
dev.off()
```

B.2 Simulações

```
# modified code from http://rosetta.ahmedmoustafa.io/selection/

diploid_haploid_deriva <- function(selection_coeficient, number_
    ↪ of_generations, number_of_repetitions, freq_recessive, pop
    ↪ _size){
    fitness_aa = 1 - selection_coefficient
    fitness_ab = 1 - selection_coefficient
    fitness_bb = 1

    fitness_a = 1 - selection_coefficient
    fitness_b = 1

    pdf("simulations_deriva.pdf",)
    par(mar=c(5,5,3,2)+0.1)
    plot(fitness_aa, type="n", xlim=c(0,number_of_generations),
        ↪ ylim=c((freq_recessive -0.1),1), xlab="Generations",
        ↪ ylab="Recessive allele frequency", main="Selection of
        ↪ an allele in Haploid and Diploid populations", cex.sub
        ↪ =1.5, cex.axis=1.5, cex.main=1.5, cex.lab=1.5)
```

```
legend("bottomright", inset=-0.008, legend=c("Haploid␣
    ↪ Selection", "Diploid␣Selection"), col=c("Grey␣7","Grey
    ↪ ␣42"), lwd=3, lty=c(4, 1), bty="n", cex=1.5, horiz = T
    ↪ )
  text(c("Selection␣\n␣coefficient", selection_coefficient
    ↪ ), x=c(number_of_generations/10, number_of_
    ↪ generations/10), y=c(0.98, 0.88), cex=1.5)

for (i in 1 : number_of_repetitions){
  freq_b = freq_recessive
  freq_a = 1 - freq_b
  freq_a_haploid = freq_a
  freq_a_diploid = freq_a
  freq_b_haploid = freq_b
  freq_b_diploid = freq_b
  freq_aa = freq_a^2
  freq_ab = 2 * freq_a * freq_b
  freq_bb = freq_b^2

  genotypes = data.frame(t = 0, cbind(freq_aa, freq_ab,
    ↪ freq_bb))
  alleles_diploid = data.frame(t = 0, cbind(freq_a_diploid
    ↪ , freq_b_diploid))
  alleles_haploid = data.frame(t = 0, cbind(freq_a_haploid
    ↪ , freq_b_haploid))

  for (t in 1 : number_of_generations) {

    x_freq_a = freq_aa + 0.5*freq_ab
    x_freq_b = freq_bb + 0.5*freq_ab

    x_freq_aa = freq_aa * fitness_aa
    x_freq_ab = freq_ab * fitness_ab
    x_freq_bb = freq_bb * fitness_bb

    sum = x_freq_aa + x_freq_ab + x_freq_bb

    freq_aa = x_freq_aa / sum
    freq_ab = x_freq_ab / sum
```

```
        freq_bb = x_freq_bb / sum

        freq_a_diploid = freq_aa + 0.5 * freq_ab
        freq_b_diploid = freq_bb + 0.5 * freq_ab

        A1=rbinom(3,2*pop_size,c(x_freq_bb, x_freq_ab, x_
           ↪ freq_aa))
        x_freq_bb=A1[1]/(pop_size*2)
        x_freq_ab=A1[2]/(pop_size*2)
        x_freq_aa=A1[3]/(pop_size*2)

        sum = x_freq_aa + x_freq_ab + x_freq_bb

        freq_aa = x_freq_aa / sum
        freq_ab = x_freq_ab / sum
        freq_bb = x_freq_bb / sum

        genotypes = rbind(genotypes, data.frame(t, cbind(
           ↪ freq_aa, freq_ab, freq_bb)))

        freq_a_diploid = freq_aa + 0.5 * freq_ab
        freq_b_diploid = freq_bb + 0.5 * freq_ab

        alleles_diploid = rbind(alleles_diploid, data.frame(
           ↪ t, cbind(freq_a_diploid, freq_b_diploid)))
    }
    lines(alleles_diploid$freq_b, col="Grey␣42", lwd=3, cex
       ↪ =1.5)

    for (t in 1 : number_of_generations) {

        x_freq_a = freq_a_haploid
        x_freq_b = freq_b_haploid

        x_freq_a = freq_a_haploid * fitness_a
        x_freq_b = freq_b_haploid * fitness_b

        sum = x_freq_a + x_freq_b
```

```
            freq_a_haploid = x_freq_a / sum
            freq_b_haploid = x_freq_b / sum

            x_freq_a=freq_a_haploid
            x_freq_b=freq_b_haploid

            A1=rbinom(1,pop_size,x_freq_b)
            x_freq_b=A1/(pop_size);
            x_freq_a=1-x_freq_b

            sum = x_freq_a + x_freq_b

            freq_a_haploid = x_freq_a / sum
            freq_b_haploid = x_freq_b / sum

            alleles_haploid = rbind(alleles_haploid, data.frame(
              ↪ t, cbind(freq_a_haploid, freq_b_haploid)))
        }
        lines(alleles_haploid$freq_b, col="Grey 7", lwd=3, lty
          ↪ =4, cex=1.5)

    }
    dev.off()
}

diploid_haploid_determinista <- function(selection_coeficient,
  ↪ number_of_generations, freq_recessive){
    fitness_aa = 1 - selection_coefficient
    fitness_ab = 1 - selection_coefficient
    fitness_bb = 1

    fitness_a = 1 - selection_coefficient
    fitness_b = 1

    pdf("simulations_determinista.pdf",)
    par(mar=c(5,5,3,2)+0.1)
    plot(fitness_aa, type="n", xlim=c(0,number_of_generations),
      ↪ ylim=c((freq_recessive -0.1),1), xlab="Generations",
      ↪ ylab="Recessive allele frequency", main="Selection of
```

```
    ↪ an allele in Haploid and Diploid populations", cex.sub
    ↪ =1.5, cex.axis=1.5, cex.main=1.5, cex.lab=1.5)
legend("bottomright", inset=-0.008, legend=c("Haploid
    ↪ Selection", "Diploid Selection"), col=c("Grey 7","Grey
    ↪ 42"), lwd=3, lty=c(4, 1), bty="n", cex=1.5, horiz = T
    ↪ )
text(c("Selection \n coefficient", selection_coefficient), x
    ↪ =c(number_of_generations/10, number_of_generations/10)
    ↪ , y=c(0.98, 0.88), cex=1.5)

for (i in 1 : number_of_repetitions){
    freq_b = freq_recessive
    freq_a = 1 - freq_b
    freq_a_haploid = freq_a
    freq_a_diploid = freq_a
    freq_b_haploid = freq_b
    freq_b_diploid = freq_b
    freq_aa = freq_a^2
    freq_ab = 2 * freq_a * freq_b
    freq_bb = freq_b^2

    genotypes = data.frame(t = 0, cbind(freq_aa, freq_ab,
        ↪ freq_bb))
    alleles_diploid = data.frame(t = 0, cbind(freq_a_diploid
        ↪ , freq_b_diploid))
    alleles_haploid = data.frame(t = 0, cbind(freq_a_haploid
        ↪ , freq_b_haploid))

    for (t in 1 : number_of_generations) {

        x_freq_a = freq_aa + 0.5*freq_ab
        x_freq_b = freq_bb + 0.5*freq_ab

        x_freq_aa = freq_aa * fitness_aa
        x_freq_ab = freq_ab * fitness_ab
        x_freq_bb = freq_bb * fitness_bb

        sum = x_freq_aa + x_freq_ab + x_freq_bb
```

```
            freq_aa = x_freq_aa / sum
            freq_ab = x_freq_ab / sum
            freq_bb = x_freq_bb / sum

            genotypes = rbind(genotypes, data.frame(t, cbind(
                ↪ freq_aa, freq_ab, freq_bb)))

            freq_a_diploid = freq_aa + 0.5 * freq_ab
            freq_b_diploid = freq_bb + 0.5 * freq_ab

            alleles_diploid = rbind(alleles_diploid, data.frame(
                ↪ t, cbind(freq_a_diploid, freq_b_diploid)))
        }
        lines(alleles_diploid$freq_b, col="Grey␣42", lwd=3, cex
            ↪ =1.5)

        for (t in 1 : number_of_generations) {

            x_freq_a = freq_a_haploid
            x_freq_b = freq_b_haploid

            x_freq_a = freq_a_haploid * fitness_a
            x_freq_b = freq_b_haploid * fitness_b

            sum = x_freq_a + x_freq_b

            freq_a_haploid = x_freq_a / sum
            freq_b_haploid = x_freq_b / sum

            alleles_haploid = rbind(alleles_haploid, data.frame(
                ↪ t, cbind(freq_a_haploid, freq_b_haploid)))
        }
        lines(alleles_haploid$freq_b, col="Grey␣7", lwd=3, lty
            ↪ =4, cex=1.5)

    }
    dev.off()
}
```

```
selection_coefficient = 0.1
number_of_generations = 5*(10^2)
number_of_repetitions = 10^2
freq_recessive = 10^-6
pop_size=10^6

diploid_haploid_deriva(selection_coeficient, number_of_
    ↪ generations, number_of_repetitions, freq_recessive, pop_
    ↪ size)

diploid_haploid_determinista(selection_coeficient, number_of_
    ↪ generations, freq_recessive)
```

ANEXO A – Modelo Matemático da Seleção de Alelos Recessivos em Populações Diplóide e Haplóide.

Durante o mestrado consultamos o Professor Paulo Otto para colaboramos na elaboração de um modelo matemático que representasse as situações de um gene ou alelo novo sendo expresso e selecionado em populações haplóide e diplóide. Assim, o modelo matemático aqui apresentado foi elaborado em conjunto com o Professor Paulo Otto, para verificarmos a validade teórica do modelo ao analisarmos os resultados das simulações matemáticas em tais situações.

Aviso: o texto a seguir não foi elaborado pela autora, mas sim pelo Professor Doutor Paulo A. Otto.

During the spermatogenesis in *Drosophila*, transcription takes place even during the post-meiosis period. It is therefore interesting to compare de effect of positive selection acting in the haploid phase with the corresponding one taking place during the diploid phase, in relation to the important process of positive selection and fixation of new alleles. In order to perform this, we shall adopt the following simple selection population genetic model:

(a) Letting $\mathbf{s_1}$ $\mathbf{(0 \leq s_1 \leq 1)}$ and $\mathbf{hs_1}$ ($\mathbf{0 \leq h \leq 1}$) be the coefficients of selection of **AA** and **Aa** cells respectively, and **1** the relative fitness of cells with the genotype **aa** during the diploid phase, $\mathbf{q = q_1}$ and $\mathbf{1-q = 1-q_1}$ the frequencies of alleles **a** and **A** in a given cell generation, $\mathbf{q_1'}$ the frequency value of **a** after a single generation of cell duplication, and $\mathbf{\Delta q_1 = q_1' - q}$, it comes out that

$$\mathbf{q_1' = q[1-(1-q)hs_1]/\{1-(1-q)s_1[1-q(1-2h)]\}}$$
and
$$\mathbf{\Delta q_1 = q(1-q)s_1[1-h-q(1-2h)]/\ \{1-(1-q)s_1[1-q(1-2h)]\}}.$$

In the formulas above, **h** is a dominance measure. When **h = 1**, it means that the fitness values of **AA**, **Aa** and **aa** cells are respectively $\mathbf{W_{AA} = 1-s_1}$, $\mathbf{W_{Aa} = 1-s_1}$ and $\mathbf{W_{aa} = 1}$ and therefore there exists a positive selection mechanism favoring the recessive genotype **aa**. When **h = 0**, it means that the fitness values of **AA**, **Aa** and **aa** cells are respectively $\mathbf{W_{AA} = 1-s_1}$, $\mathbf{W_{Aa} = 1}$ and $\mathbf{W_{aa} = 1}$ and therefore there exists a positive selection mechanism favoring the dominant genotypes **Aa** and **aa**.

When **h = 1** the formulas for $\mathbf{q_1'}$ and $\mathbf{\Delta q_1}$ take form

$$\mathbf{q_1' = q[1-(1-q)s_1]/[1-(1-q^2)s_1]}$$
and
$$\mathbf{\Delta q_1 = q^2(1-q)s_1\ /[1-(1-q^2)s_1]},$$

while in the case **h = 0** the corresponding formulas become

$$\mathbf{q_1' = q/[1-(1-q)^2s_1]}$$
and
$$\mathbf{\Delta q_1 = q(1-q)^2s_1\ /[1-(1-q)^2s_1]}.$$

(b) letting, now, $\mathbf{s_2}$ $\mathbf{(0 \leq s_2 \leq 1)}$ be the coefficient of selection of **A** gametes during the haploid phase, **1** the relative fitness of gametes carrying the **a** allele, $\mathbf{q = q_2}$ and

$1-q = 1-q_2$ the frequencies of gametes with alleles **a** and **A** in a given generation, q_2' the frequency value of **a** after a single generation of competition among gametes, and $\Delta q_2 = q_2' - q$, it comes out that

$$q_2' = q/[1-(1-q)s_2] \ , \ \Delta q_2 = q(1-q)s_2/[1-(1-q)s_2];$$

(c) let $\Delta q_2/\Delta q_1$ be the increment rate, a pertinent variable for comparing the evolutionary gain of frequency (fixation rate) of the allele **a** under the alternative hypotheses of positive selection acting at the haploid and diploid phases of spermatogenesis; its value is

$$\Delta q_2/\Delta q_1 = s_2\{1-(1-q)s_1[1-q(1-2h)]\} / \{s_1[1-(1-q)s_2][1-h-q(1-2h)]\} \ ;$$

when $h = 1$, $\Delta q_2/\Delta q_1$ takes form

$$\Delta q_2/\Delta q_1 = s_2[1-(1-q^2)s_1]/\{s_1q[1-(1-q)s_2]\};$$

and when $h =0$,

$$\Delta q_2/\Delta q_1 = s_2[1-(1-q)^2s_1]/\{s_1(1-q)[1-(1-q)s_2]\};$$

in the interesting case $h = 1/2$, Δq_1 simplifies to

$$\Delta q_1 = q(1-q)s_1/\{2[1-(1-q)s_1]\}$$

and $\Delta q_2/\Delta q_1$ to

$$\Delta q_2/\Delta q_1 = 2s_2[1-(1-q)s_1]/\{s_1[1-(1-q)s_2]\}.$$

If we put $s_2 = s$ and $s_1 = sx$, we obtain the more suitable expression

$$\Delta q_2/\Delta q_1 = \{1-(1-q)sx[1-q(1-2h)]\}/\{x[1-(1-q)s][1-h-q(1-2h)]\},$$

an expression that reduces to

$\Delta q_2/\Delta q_1 = (1-sx+sxq^2)/(xq-sxq+sxq^2)$ when $h = 1$,

to $\Delta q_2/\Delta q_1 = [1-sx(1-q)^2]/[x(1-q)-sx(1-q)^2]$ when $h = 0$,

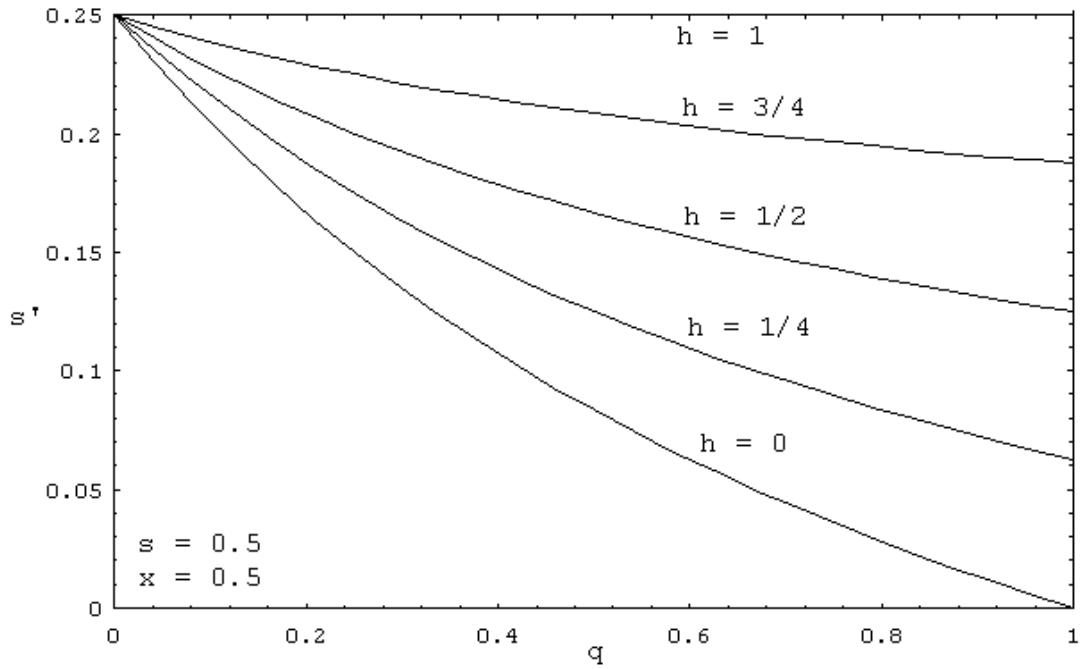
and to $\Delta q_2/\Delta q_1 = 2[1-sx(1-q)]/[x-sx(1-q)]$ when $h = 1/2$.

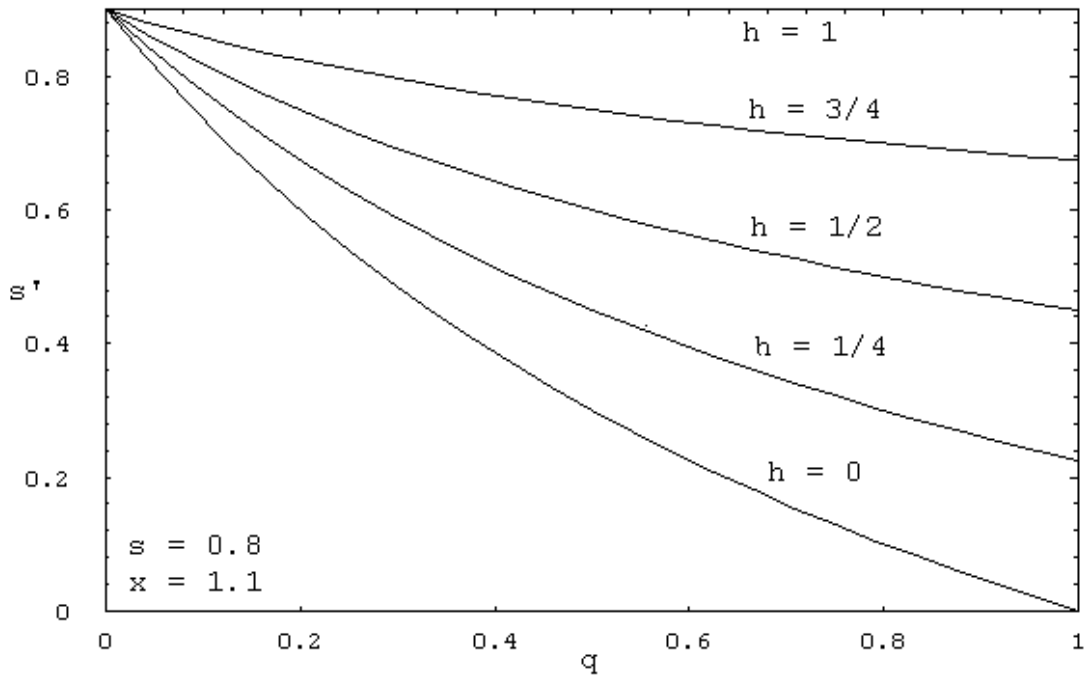
Taking into account the variables **s**, **x**, and **h**, it comes out that the average coefficient of selection of the group of diploid cells with the ***A*** allele (***AA*** and ***Aa***) is

$$s' = [(1-q)^2 sx + 2q(1-q)shx]/[(1-q)^2 + 2q(1-q)],$$

that simplifies to $s' = sx[(1-q)+2qh]/(1+q)$ and that reduces to sx when $h = 1$, to $sx/(1+q)$ when $h = 1/2$ and to $sx(1-q)/(1+q)$ when $h = 0$.

The numerical analysis of the generalized expression $s' = sx[(1-q)+2qh]/(1+q)$ shows unequivocally that for any fixed values of **x**, **s** and **q**, the larger the value of **h**, the larger the value of **s'**, whether $x \leq 1$ or $x > 1$, as the graphs below clearly show.





(d) let now **k** (**0** ≤ **k** ≤ **1**) and **1-k** be respectively the proportions of haploid and diploid cells contributing to the transcription process during spermatogenesis. The frequency **Q'** of the allele ***a*** as a result of the whole process is obtained by averaging (by **1-k** and **k** respectively) the contributions in gene frequency q_1' and q_2' of diploid and haploid cells to the next generation. Since $\Delta Q = Q' - q$, $\Delta q_1 = q_1' - q$ and $\Delta q_2 = q_2' - q$, we obtain immediately

$$Q' = kq_2' + (1-k)q_1' = k(\Delta q_2 - \Delta q_1) + \Delta q_1 + q \ ,$$

$$\Delta Q = Q' - q = k(\Delta q_2 - \Delta q_1) + \Delta q_1$$

and

$$\Delta Q/\Delta q_1 = k(\Delta q_2 - \Delta q_1)/\Delta q_1 + 1 = 1 - k(1 - \Delta q_2/\Delta q_1) \ .$$

Since $\Delta Q/\Delta q_1$ is a linear function of $\Delta q_2/\Delta q_1$, the behavior of $\Delta Q/\Delta q_1$ can be indirectly derived from the behavior of $\Delta q_2/\Delta q_1$, and this is exactly what we first perform in the paragraphs that follow. It is important to keep in mind, in the text below, that the larger the values of the coefficients of selection $s_1 = sx$ and $s_1h = sxh$ (of ***AA*** and ***Aa*** diploid cells) or s_2 (of gametes **A**), the larger will be the

relative fitness values of ***aa*** cells and gametes containing the allele ***a***.

Analysis of special case h = 1 (AA and Aa dominant, aa recessive)

When $x \leq 1$, indicating therefore that $s_2 \geq s_1$ and that the fitness value W_a of haploid gametes ***a*** is equal or larger than the fitness value W_{aa} of homozygous cells ***aa***, the increment rate $\Delta q_2/\Delta q_1 = (1-sx+sxq^2)/(xq-sxq+sxq^2)$ is always larger than unity for any combination of values of q ($0 < q < 1$) and s ($0 < s < 1$), because under these conditions $1-sx > xq-sxq$. *We conclude therefore that when* $x \leq 1$ *(*$s_2 \geq s_1$ *or* $W_a \geq W_{aa}$*) the increment rate* $\Delta q_2/\Delta q_1$ *is always larger than unity, thus indicating that the gain in allele* ***a*** *frequency (rate of fixation of the* ***a*** *allele) is always larger in the case of positive selection during the haploid phase than in the diploid one.*

When $x > 1$, indicating that $s_2 < s_1$ and that the fitness value of ***aa*** cells is larger than that of ***a*** gametes, the inspection of $\Delta q_2/\Delta q_1 = (1-sx+sxq^2)/(xq-sxq+sxq^2)$ shows that if $q > (1-sx)/(x-sx)$, $\Delta q_2/\Delta q_1$ will be smaller than 1; inversely, if $q < (1-sx)/(x-sx)$, $\Delta q_2/\Delta q_1$ will be larger than 1. Unlike the previous case $x \leq 1$, in the present situation sx must be always smaller than 1, so that the domain of s is now $0 < s < 1/x$ instead of $0 < s < 1$. *We conclude therefore that, respected the restriction* $0 < q \leq (1-sx)/(x-sx)$*, even in the apparently non-advantageous situation in which the fitness value* W_{aa} *is larger than* W_a*, the rate of frequency gain (fixation rate) of the* ***a*** *allele is larger under the system of positive selection during the haploid phase than during the diploid one.*

Going back to the generalized expression $\Delta Q/\Delta q_1 = 1 - k(1 - \Delta q_2/\Delta q_1)$, since the domain of k (relative proportion of contributing haploid cells) is $0 < k < 1$, it comes out straightforwardly that

$\Delta Q/\Delta q_1 > 1$ if $\Delta q_2/\Delta q_1 = (1-sx+sxq^2)/(xq-sxq+sxq^2) > 1$,

that is, independently from the relative proportions k and $1-k$ of haploid and diploid cells contributing to transcription during the spermatogenesis process, the gain in allele **a** frequency under positive selection during the haploid phase

is always larger than the corresponding one during the diploid phase. This situation $\Delta q_2/\Delta q_1 > 1$ takes place: a) without any restrictions always that $x \leq 1$, that is when the fitness value of ***a*** gametes is equal or larger than the fitness value of diploid ***aa*** cells ($W_a \geq W_{aa}$); b) when $x > 1$ (fitness value of **aa** diploid cells larger than that of **a** gametes), under the restrictive conditions $0 < q < (1-sx)/(x-sx)$ and $0 < s < 1/x$. However, if $x > 1$ and $(1-sx)/(x-sx) < q < 1$, the gain in allele ***a*** frequency under positive selection in diploid phase is always larger than the corresponding one in haploid phase.

Analysis of special case h = 0 (AA recessive, Aa and aa dominant)

When $x \leq 1$, indicating therefore that $s_2 \geq s_1$ and that the fitness value W_a of haploid gametes ***a*** is equal or larger than the fitness values W_{aa} and W_{Aa} of homozygous and heterozygous cells ***aa*** and ***Aa***, the increment rate $\Delta q_2/\Delta q_1 = [1-sx(1-q)^2]/[x(1-q)-sx(1-q)^2]$ is always larger than unity for any combination of values of q ($0 < q < 1$) and s ($0 < s < 1$), because $1-sx(1-q)^2$ is always larger than $x(1-q)-sx(1-q)^2$, or simply because 1 is always larger than $x(1-q)$.

When, on the contrary, $x > 1$ (situation in which $s_2 < s_1$ and therefore $W_{aa} = W_{Aa} > W_a$), the increment rate $\Delta q_2/\Delta q_1$ will be larger than unity only if $x < 1/(1-q)$, respected the obvious restriction that sx must be always smaller than 1, so that the domain of s is $0 < s < 1/x$ instead of $0 < s < 1$, just like in the corresponding case $x > 1$ when $h = 1$.

Since the domain of k (relative proportion of contributing haploid cells) is $0 < k < 1$, it comes out straightforwardly that

$\Delta Q/\Delta q_1 > 1$ if $\Delta q_2/\Delta q_1 = [1-sx(1-q)^2]/[x(1-q)-sx(1-q)^2] > 1$,

that is, independently from the relative proportions k and $1-k$ of haploid and diploid cells contributing to transcription during the spermatogenesis process, the gain in allele **a** frequency under positive selection during the haploid phase is always larger than the corresponding one during the diploid phase. This situation $\Delta q_2/\Delta q_1 > 1$ takes place: a) without any restrictions always that $x \leq 1$, that is when the fitness value of ***a*** gametes is equal or larger than the fitness value of diploid ***aa*** and ***Aa*** cells ($W_a \geq W_{aa} = W_{Aa}$); b)

when $x > 1$ (fitness value of **aa** and ***Aa*** diploid cells larger than that of **a** gametes), under the restrictive conditions $x < 1/(1-q)$ and $sx < 1$. However, if $x > 1$ and $x > 1/(1-q)$, the gain in allele ***a*** frequency under positive selection in diploid phase is always larger than the corresponding one in haploid phase.

Analysis of special case h = 1/2

When $x \leq 1$, indicating therefore that $s_2 \geq s_1$, the increment rate $\Delta q_2/\Delta q_1 = 2[1-sx(1-q)]/[x-sx(1-q)]$ is always larger than unity for any combination of values of q ($0 < q < 1$) and s ($0 < s < 1$), because the above expression can be rewritten as $\Delta q_2/\Delta q_1 = 2 + 2(1-x)/[x-sx(1-q)]$ and it is easy to determine that the minimum value $\Delta q_2/\Delta q_1$ can achieve is **2** when $x = 1$; for any value of x less than **1**, $\Delta q_2/\Delta q_1$ is always larger than **2**. *We conclude therefore that when* $x \leq 1$ *(*$s_2 \geq s_1$*) the increment rate* $\Delta q_2/\Delta q_1$ *is always larger than unity, thus indicating that the gain in allele* ***a*** *frequency (rate of fixation of the* ***a*** *allele) is always larger in the case of positive selection during the haploid phase than in the diploid one.*

When $x > 1$, indicating that $s_2 < s_1$, the numerical analysis of $\Delta q_2/\Delta q_1 = 2[1-sx(1-q)]/[x-sx(1-q)]$ shows that for any combination of values of ($0 < q < 1$) and ($0 < s < 1/x$) the increment rate $\Delta q_2/\Delta q_1$ will always be smaller than **1.** In fact, replacing x by $1+\delta$, $\delta > 0$, the expression above takes form $\Delta q_2/\Delta q_1 = 2 - 2\delta/\{(1+\delta)[1-s(1-q)]\}$. When s is at its maximum possible value $1/x = 1/(1+\delta)$, the increment rate has value $\Delta q_2/\Delta q_1 = 2 - 2\delta/(\delta+q)$ so that if q is of order of magnitude of δ, $\Delta q_2/\Delta q_1$ will be a bit smaller than unity; if δ is much larger than q, $\Delta q_2/\Delta q_1$ will be a bit larger than zero. When s is near its minimum value **0**, due to the constraint $s = 1/(1+\delta)$, δ must be very large so that $\Delta q_2/\Delta q_1$ takes a value just a bit larger than zero. *We conclude therefore that when* ***h = 1/2*** *the rate of frequency gain (fixation rate) of the* ***a*** *allele is larger under the system of positive selection during the diploid phase than during the haploid one, for any admissible values* ***s, x*** *and* ***q*** *can take.*

Going back to the generalized expression $\Delta Q/\Delta q_1 = 1 - k(1 - \Delta q_2/\Delta q_1)$, since the domain of k (relative proportion of contributing haploid cells) is $0 < k < 1$, it comes out

straightforwardly that in the present case, with no exceptions, $\Delta Q/\Delta q_1 > 1$ if $\Delta q_2/\Delta q_1 > 1$ and $\Delta Q/\Delta q_1 < 1$ if $\Delta q_2/\Delta q_1 < 1$.

Analysis of the behavior of $\Delta q_2/\Delta q_1 = f(q, s, x, h)$ in the general case $W_{AA} = 1-s_1 = 1-sx$, $W_{Aa} = 1-hs_1 = 1-hsx$, $W_{aa} = 1$, $W_A = 1-s_2 = 1-s$, $W_a = 1$

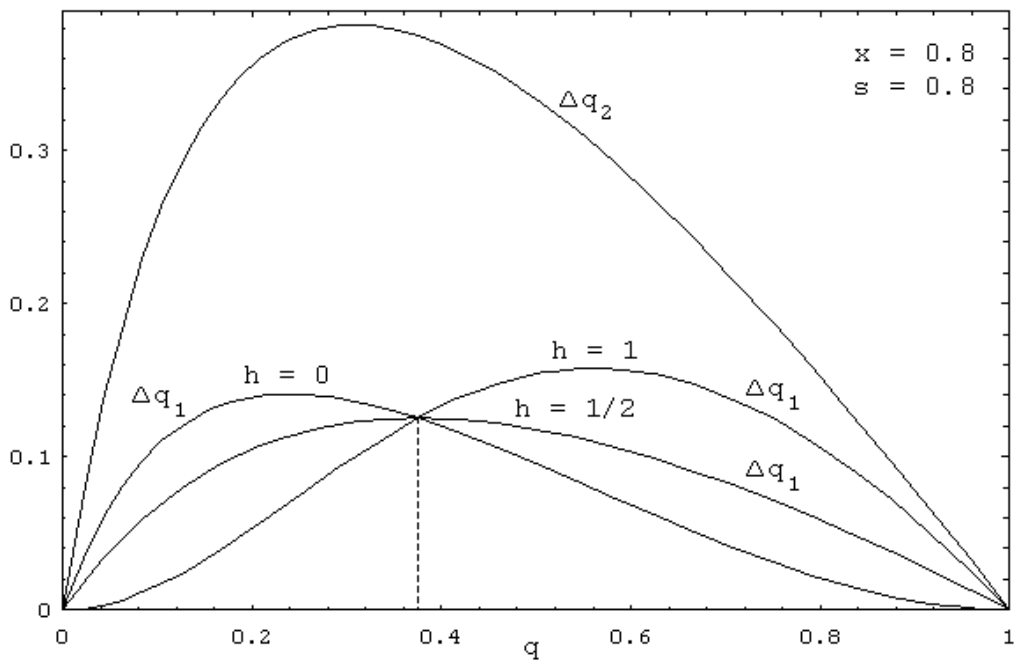
The expression already derived for the increment rate,

$$\Delta q_2/\Delta q_1 = \{1-(1-q)sx[1-q(1-2h)]\}/\{x[1-(1-q)s][1-h-q(1-2h)]\},$$

can be rewritten in the more convenient form

$$\Delta q_2/\Delta q_1 = 1 + \{1-x[(1-q)(1+hs-h)+qh]\}/\{x[1-(1-q)s][1-h-q(1-2h)]\}.$$

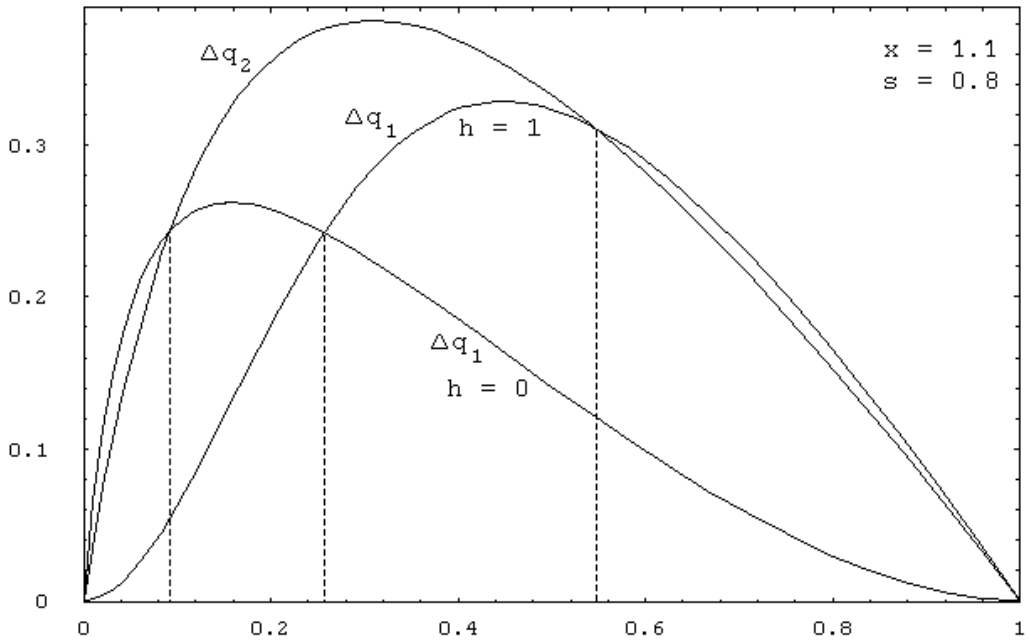
When $0< x \leq 1$, for any combination of values of $0 < s < 1$, $0 \leq h \leq 1$, and $0 < q < 1$, it comes out that both the numerator $1-x[(1-q)(1+hs-h)+qh]$ and the denominator $x[1-(1-q)s][1-h-q(1-2h)]$ of the rightmost part of the last equation are positive quantities. Therefore we conclude that if $0< x \leq 1$, $\Delta q_2/\Delta q_1$ is always larger than one, irrespectively the value h (the dominance factor) can take.



The graph shows unequivocally that for small values of q, when ***h = 1*** *(**aa** completely recessive) the gain in gene frequency* $\Delta q_2/\Delta q_1$ *(fixation rate) of the **a** allele is much larger under the system of positive selection during the haploid phase than during the diploid one, for any admissible value **h** can take.* The graph also shows that for $q = 3/8 = 0.375$, the value of Δq_1 is $1/8 = 0.125$, whatever **h** be **1**, **1/2** or **0**. This results from the fact that when this takes place q is the positive root of equation $q^2sx + 2q(1-sx) - (1-sx) = 0$.

When $x > 1$, $\Delta q_2/\Delta q_1$ will be larger or smaller than unity depending of the value (positive or negative) the expression $1-x[(1-q)(1+hs-h)+qh$ from the equation's rightmost part takes, because $x[1-(1-q)s][1-h-q(1-2h)]$ will be always positive, since $x > 1$, $0 \leq h \leq 1$, $0 < q < 1$ and $0 < s < 1/x$ It is important to stress that, unlike the previous case $x \leq 1$, when $x > 1$ sx must be always smaller than **1**, so that the domain of **s** is now $0 < s < 1/x$ instead of $0 < s < 1$). Extensive numerical analysis of the formulas above shows that when $h > 1/2$, if $q > [1-x(1-h+hs)]/[x(1-2h+hs)]$, $\Delta q_2/\Delta q_1$ will be smaller than unity and if $q < [1-x(1-h+hs)]/[x(1-2h+hs)]$, $\Delta q_2/\Delta q_1$ will be larger than unity. Otherwise, that is when $h < 1/2$, if $q > [1-x(1-h+hs)]/[x(1-2h+hs)]$, $\Delta q_2/\Delta q_1$ will be larger than unity and if $q < [1-x(1-h+hs)]/[x(1-2h+hs)]$, $\Delta q_2/\Delta q_1$ will be smaller than unity. When $h = 1/2$, the formula of the increment rate reduces to $\Delta q_2/\Delta q_1 = 2[1-sx(1-q)]/[x-sx(1-q)]$ and the numerical analysis of this expression shows that when $x > 1$, for any combination of values of ($0 < q < 1$) and ($0 < s < 1/x$) the increment rate $\Delta q_2/\Delta q_1$ will always be smaller than **1.** In fact, replacing **x** by $1+\delta$, $\delta > 0$, the expression above takes form $\Delta q_2/\Delta q_1 = 2 - 2\delta/\{(1+\delta)[1-s(1-q)]\}$. When **s** is at its maximum possible value $1/x = 1/(1+\delta)$, the increment rate has value $\Delta q_2/\Delta q_1 = 2 - 2\delta/(\delta+q)$ so that if **q** is of order of magnitude of δ, $\Delta q_2/\Delta q_1$ will be a bit smaller than unity; if δ is much larger than **q**, $\Delta q_2/\Delta q_1$ will be a bit larger than zero. When **s** is near its minimum value **0**, due to the constraint $s = 1/(1+\delta)$, δ must be very large so that $\Delta q_2/\Delta q_1$ takes a value just a bit larger than zero. *We conclude therefore that when* ***x > 1*** *and* ***h = 1/2*** *the rate of frequency gain (fixation rate) of the **a** allele is larger under the system of positive selection during the diploid phase than*

*during the haploid one, for any admissible values **s** and **q** can take.*



*The graph shows that for small values of q, when **h = 1** (**aa** completely recessive) the gain in gene frequency **$\Delta q_2/\Delta q_1$** (fixation rate) of the **a** allele is much larger than in the case **h = 0** under the system of positive selection during the haploid phase than during the diploid one and t*his is valid when other values of **h < 1** are compared to case **h =1**. But in any case for every combination of s and x for some value of q the gain during the diploid phase will be larger than in the haploid phase when x > 1: this takes place, as seen in the above graph, for the prescribed conditions **x = 1.1** and **s = 0.8**, when **q = (x-1)/x = 1/11 = 0.09091** if **h = 0** and **q = (1-sx)/(x-sx) = 6/11 = 0.54545.** The graph also shows, just like in the case **0< x ≤ 1**, that **Δq_1** takes the same value independently from the value of **h** (**0** or **1**) when **q** is the solution of equation **$q^2sx + 2q(1-sx) - (1-sx) = 0$**. For the prescribed value of **sx = 0.88**, **q = 0.25237**.

If we now take also into account the relative proportion **k** of contributing haploid cells to the transcription process during spermatogenesis, we get finally the (fully) generalized expression

$$\Delta Q/\Delta q_1 = f(q,s,x,h,k) = 1 - k(1 - \Delta q_2/\Delta q_1).$$

Since the domain of **k** (relative proportion of contributing haploid cells) is $\mathbf{0 < k < 1}$, it comes out straightforwardly that $\mathbf{\Delta Q/\Delta q_1 > 1}$ if $\mathbf{\Delta q_2/\Delta q_1 > 1}$, that is, independently from the relative proportions **k** and **1-k** of haploid and diploid cells contributing to the transcription process during spermatogenesis, the gain in allele **a** frequency under positive selection during the haploid phase is always larger than the corresponding one during the diploid phase. This situation $\mathbf{\Delta q_2/\Delta q_1 > 1}$ takes place: (1) without any restrictions always that $\mathbf{x \leq 1}$, that is when the fitness value of ***a*** gametes is equal or larger than the fitness value of diploid cells **aa**, independently from the value **h** takes; and (2) with the following restrictions when $\mathbf{x > 1}$ (fitness value of diploid cells ***aa*** larger than that of ***a*** gametes): $\mathbf{q < [1-x(1-h+hs)]/[x(1-2h+hs)]}$ if $\mathbf{h > 1/2}$, and $\mathbf{q > [1-x(1-h+hs)]/[x(1-2h+hs)]}$ if $\mathbf{h < 1/2}$. If these stringent conditions however do not prevail, the gain in allele ***a*** frequency under positive selection in diploid phase is always larger than the corresponding one in haploid phase.

The following (rounded) figures of percentage values of $\mathbf{\Delta q_2/\Delta q_1}$ larger and smaller than unity were obtained from 1,000,000 computer-assisted random combinations of $\mathbf{2 < x < 20}$, $\mathbf{0 < s < 1/x}$, and $\mathbf{0 < q < 1}$ for each **h** value varying from 0 to 1 with 0.05 intervals.

h	<1	>1
0.00	0.23	0.77
0.05	0.20	0.80
0.10	0.17	0.83
0.15	0.15	0.85
0.20	0.13	0.87
0.25	0.11	0.89
0.30	0.09	0.91
0.35	0.08	0.92
0.40	0.07	0.93
0.45	0.05	0.95
0.50	0.00	1.00
0.55	0.01	0.99
0.60	0.02	0.98
0.65	0.03	0.97
0.70	0.03	0.97
0.80	0.05	0.95
0.85	0.07	0.93
0.90	0.08	0.92

0.95	0.10	0.90
1.00	0.13	0.87
0-1	0.13	0.87

We conclude therefore that, even in the non-advantageous situation when ***x > 1****, for all possible* ***h*** *values can take in the case* ***2 < x < 20****, in 13% of all combinations of values of q, x and s, when* ***x > 1*** *the rate of frequency gain (fixation rate) of the* ***a*** *allele is larger under the system of positive selection during the haploid phase than during the diploid one. Taking into account that this is exactly what always takes place when* ***x ≤ 1****, we have just evidenced the importance of the mechanism of positive selection acting during the haploid phase of spermatogenesis in the process of fixation of new genes.*

Extensive computer-assisted numerical analysis of $\Delta q_2/\Delta q_1$ and $\Delta Q/\Delta q_1$ confirmed unequivocally all the logical intuitive statements of all previous paragraphs.