UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

Bianca Manfré

Investigação de assinaturas de seleção positiva em genes associados à localização do hospedeiro em espécies do grupo *repleta* de *Drosophila*

São José do Rio Preto

2022

Bianca Manfré

Investigação de assinaturas de seleção positiva em genes associados à localização do hospedeiro em espécies do grupo *repleta* de *Drosophila*

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciências Biológicas, junto ao Conselho de Curso de Bacharelado em Ciências Biológicas], do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Financiadora: CNPq 0736156509099361

Orientador: Profª. Drª. Claudia Marcia Aparecida Carareto
Coorientador: MSc Daniel Siqueira de Oliveira

São José do Rio Preto
2022

Manfré, Bianca
M276i Investigação de assinaturas de seleção positiva em genes associados à localização do hospedeiro em espécies do grupo repleta de Drosophila / Bianca Manfré. -- São José do Rio Preto, 2022
90 p. : il., tabs., fotos, mapas

Trabalho de conclusão de curso (-) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientadora: Claudia Marcia Aparecida Carareto
Coorientador: Daniel Siqueira de Oliveira

1. evolução. 2. assinaturas de seleção. 3. espécies cactofílicas. 4. mudança de hospedeiro.
I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências, Letras e Ciências Exatas, São José Rio Preto. Dados fornecidos pelo autor (a).

Essa ficha não pode ser modificada.

Bianca Manfré

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciências Biológicas, junto ao Conselho de Curso de Bacharelado em Ciências Biológicas], do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Financiadora: CNPq 0736156509099361

Comissão Examinadora

Profª. Drª. Claudia Marcia Aparecida Carareto
UNESP – Câmpus de São José do Rio Preto
Orientador

Profª. Drª. Camila Vieira Curti
USP – Câmpus de Pirassununga

Profª. Drª. Priscila Aparecida Casciatori Frassatto
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto
27 de janeiro de 2022

AGRADECIMENTOS

Agradeço imensamente todo o conhecimento adquirido, toda a aprendizagem de vida e de ciência que ganhei da Profa. Dra. Claudia Carareto, que aceitou me orientar, me apoiou e me ensinou muito, e que faz de seu ambiente laboratorial um local de extremo profissionalismo e seriedade e de sua equipe uma família.

Agradeço também a todos meus amigos de graduação: ao William, Cecília e Maryanna pelo incrível companheirismo dentro e fora do laboratório, ao Daniel, que me ensinou muitíssimo, e aos meus fiéis amigos Bárbara, Gabriella, Letícia, Luiz e Kamila, pela amizade, pelo apoio e pelas aventuras.

Agradeço ao Cnpq, sob o processo 0736156509099361 por conceder a bolsa PIBIC, permitindo a realização desse trabalho.

Por fim agradeço à minha família por sempre ter me apoiado e acreditado em mim.

Todos fizeram essa minha experiência na graduação Inesquecível.

RESUMO

As mudanças de hospedeiro observadas no grupo *repleta* de *Drosophila*, no qual o estado de caráter ancestral é o uso de cactos do gênero *Opuntia* como sítios de hospedagem e reprodução, para o uso de cactos colunares quimicamente mais complexos em diversas espécies, envolveu a evolução de proteínas quimiossensoriais que habilitassem a localização e reconhecimento do sítio hospedeiro. Os processos evolutivos e o papel da seleção natural positiva nesse processo adaptativo são ainda pouco explorados. Utilizando genes quimiossensoriais de proteínas de ligação odorante (*Obp*s) e receptores odorantes (*Or*s) para representantes do grupo *repleta*, foi possível identificar sítios com assinaturas de seleção positiva em cerca de 84% dos genes *Obp* e 79% dos genes *Or* analisados. Análises sobre a localização desses sítios na estrutura dessas proteínas permitiu identificar padrões de substituição de aminoácidos que provavelmente afetaram a transdução do sinal nos receptores ORs e o reconhecimento das moléculas odorantes por OBPs. Os resultados deste estudo suportam o papel da seleção positiva na evolução de genes quimiossensoriais cujos produtos proteicos atuam na localização do hospedeiro em espécies cactofílicas.

Palavras-chaves: espécies cactofílicas; mudança de hospedeiro; proteínas de ligação odorante; receptores odorantes, assinaturas de seleção

ABSTRACT

The host changes observed in the *repleta* group of *Drosophila*, in which the ancestral character state is the use of cacti of the genus *Opuntia* as hosting and reproduction sites, for the use of chemically more complex columnar cacti in several species, involved the evolution of chemosensory proteins that enable localization and recognition of the host site. Evolutionary processes and the role of positive natural selection in this adaptive process are still poorly explored. Using chemosensory genes of odorant binding proteins (*Obps*) and odorant receptors (*Ors*) for representatives of the *repleta* group, it was possible to identify sites with positive selection signatures in about 84% of the *Obp* genes and 79% of the *Or* genes analyzed. Analysis of the location of these sites in the structure of these proteins allowed us to identify patterns of amino acid substitution that probably affected signal transduction at OR receptors and the recognition of odorant molecules by OBPs. The results of this study support the role of positive selection in the evolution of chemosensory genes whose protein products act on host localization in cactophilic species.

Keywords: cactophilic species; host change; odor binding proteins; odorant receptors, selection signatures.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CDS	sequência codificadora do gene
GR	receptores gustativos
IR	receptores ionotrópicos
Ka	substituições não sinônimas
Ks	substituições sinônimas
Mya	milhões de anos atrás
pp	probabilidade posterior
p-value	valor de p
OBP	proteínas de ligação odorantes
OR	receptores odorantes
OSN	neurônios sensoriais olfativos

LISTA DE SÍMBOLOS

ω	ômega
π	pi
$\chi 2$	Qui-quadrado

SUMÁRIO

1 INTRODUÇÃO

A especialização a hospedeiros é um processo capaz de levar à especiação ecológica (BOUGHMAN, 2002), e a radiação do grupo *repleta* de *Drosophila* com mudança de hospedeiro, e consequente adaptação, evidencia tal processo (MARKOW, 2019).

Mudanças no habitat, no clima, aumento ou diminuição na predação e/ou competição, e mudanças na disponibilidade de recursos, interferem diretamente na íntima relação inseto-hospedeiro, forçando muitas vezes o deslocamento geográfico do inseto e localização de outra planta hospedeira. Um novo hospedeiro implica em novas adaptações, que vão garantir a sobrevivência do inseto nesse novo contexto, reveladas em sua morfologia (tamanho, cor), fisiologia (sistema sensorial, mecanismos de desintoxicação) e em seu comportamento (preferência pelo hospedeiro, acasalamento) (DATE, 2013; ETGES, 1998; FEDER, 1998; MIYATAKE; SHIMIZU, 1999; DEKKER, 2006). Essas recentes adaptações definem a especialização a um hospedeiro, que são então favorecidas pela seleção natural a fim de garantir uma maior aptidão (DATE, 2013; JAENIKE, 1978). Um subproduto ou consequência direta dessas novas adaptações e especializações é a evolução das barreiras reprodutivas (HANSON, 2019) decorrentes da ação da seleção divergente que poderão resultar em especiação (DATE, 2013; BOUGHMAN, 2002). É necessário, portanto, assimilar os mecanismos pelo qual a seleção divergente age sobre fenótipos diferentes, para que então seja elucidada a especiação ecológica.

As evidências sobre a origem sul americana do grupo *repleta* são recentes. Tinha-se a princípio que a colonização de *Drosophila* ocorreu da Ásia para o Novo Mundo, com radiações importantes durante o Mioceno, e que a região trans-vulcânica mexicana teria sido o grande centro de diversificação de *repleta*, tudo isso devido a um maior conhecimento de espécies do grupo nos EUA e México (OLIVEIRA et al, 2012; PATTERSON; STONE, 1952; THROCKMORTON, 1975). Contudo, após um aumento de coletas na América do Sul, a real diversidade de *repleta* começou a ser desvendada, e mais peças sobre a origem do grupo foram encaixadas (OLIVEIRA et al, 2012). Igualmente, a origem dos cactos hospedeiros foi limitada à América do Norte, particularmente em razão de uma alta diversidade e endemismo de cactos colunares mexicanos, com o sul do México representando um centro mais recente de radiação dessas espécies (OLIVEIRA et al, 2012). No entanto, análises filogenéticas e sistemáticas mais atuais de cactos pertencentes a família *Opuntioideae*, apontam a origem sul-americana dessas

plantas, além de uma divergência tardia na América do Norte (GRIFFITH; PORTER, 2009; NYFFELER; EGGLI, 2010).

O mapeamento da distribuição geográfica tanto das espécies do grupo *repleta* quanto de seus cactos hospedeiros demonstra uma grande representatividade de subgrupos desses drosofilídeos entre o norte e o sul do continente americano, além de sugerir eventos de dispersão intercontinental através do istmo do Panamá (OLIVEIRA et al, 2012).

A revelação da real diversidade de abundância das espécies do grupo *repleta* na América do Sul, combinada com a origem de seus cactos hospedeiros, indicam a origem sul-americana do grupo. Os dados biogeográficos e históricos indicam a origem tanto de *repleta* quanto de seus cactos em meados do Mioceno, na então seca e isolada América do Sul. A diversificação de ambos foi seguindo ao norte, à medida que o continente avançava, colonizando assim as ilhas caribenhas e a América do Norte. Concomitantemente, o sítio reprodutivo de *repleta* de flores e frutos tropicais, de diferentes formas, foi sendo substituído para cactos como *Opuntia*, donos de folhas planas (OLIVEIRA et al, 2012).

Confinado ao Novo Mundo, o grupo *repleta* divergiu de seu grupo irmão *virilis* há cerca de 21 Mya (million years ago - milhõoes de anos atrás), e diversificou dentro do próprio grupo há cerca de 16 Mya. Representando uma das maiores radiações em *Drosophila*, supõe-se que o grupo tenha aparecido depois do surgimento dos principais grupos de cactos, devido à elevação andina há cerca de 17 Mya, quando o interior do continente fez-se mais seco e quente. A radiação entre o complexo *mulleri* norte-americano e *buzzatii* sul-americano é datado de aproximadamente 11,3 Mya (OLIVEIRA et al, 2012). Baseando-se na distribuição atual dos gêneros ancestrais de cactos, concluiu-se que possivelmente a origem dentro das subfamílias *Opuntoideae*, *Cactoideae* e *Pereskioideae*, e seu centro de radiação seriam as terras áridas do Peru e da Bolívia (OLIVEIRA et al., 2012; EDWARDS et al., 2005).

A divergência do grupo *repleta* no Novo Mundo envolveu diversas adaptações, dentre elas, as relacionadas com a realização do ciclo de vida se mostram como sendo uma das mais bem-sucedidas. A grande maioria dos representantes do grupo *repleta* estão intimamente associados a cactos, utilizando-os como sítios de reprodução/hospedagem. Essas espécies ditas saprófagas (RANE et al, 2019) são cactofílicas, por se alimentarem e se reproduzirem em tecidos de cactos em decomposição (RUIZ; HEED, 1988; RUIZ, 1990). Após um mapeamento de 65 espécies deste grupo, associando a filogenia das espécies ao uso do cacto hospedeiros, observou-se a utilização comum do cacto *Opuntia* em toda a filogenia, a alteração independente

de escolha para cactos colunares e a existência de espécies polimórficas que utilizam ambos os tipos de cactos. Concluiu-se a partir da distribuição do uso dos dois tipos de cactos que, a utilização de *Opuntia* é uma condição considerada basal, e a diversificação para cactos colunares, quimicamente mais complexos, um estado derivado (OLIVEIRA et al., 2012).

A transição de hospedeiro para cactos colunares ocorreu inúmeras vezes ao longo da filogenia de clados norte e sul-americanos do grupo *repleta*. A grande maioria das espécies desse grupo são especialistas, já que utilizam os tecidos necróticos dos cactos para alimentação e reprodução. Por outro lado, *Drosophila hydei* e *Drosophila repleta* são exemplos de espécies generalistas, já que utilizam frutas, vegetais podres e fezes de animais, além dos cactos (RANE et al., 2019). Para elucidar tal mudança, tomemos como exemplo, dois subgrupos de espécies dentro do grupo *repleta*: *mercatorum* e *mulleri*, o qual é subdividido em complexos. No subgrupo *mercatorum*, o segundo mais basal na filogenia de *repleta*, todas as três espécies que o compõem, utilizam exclusivamente *Opuntia*. Já no subgrupo *mulleri*, no complexo *mulleri*, encontram-se espécies que utilizam exclusivamente *Opuntia*, como a espécie mais basal do complexo, *D. navojoa*, e duas outras espécies, *D. arizonae* e *Drosophila mojavensis*, que são polimórficas e utilizam tanto *Opuntia* quanto cactos colunares. Contudo, nesse subgrupo, cada uma das quatro subespécies de *D. mojavensis* utilizam *Opuntia*, mas usam preferencialmente cactos colunares, com uma particularidade, exclusivamente um único tipo de cacto colunar: *D. m. sonorensis* utiliza o cacto colunar "órgão de tubo" (*Stenocereus thurberi*); *D. m. mojavensis* utiliza o "cacto-barril" (*Ferocactus cylindraceus*); *D m. wrigleyi* utiliza cacto tipo "pera-espinhosa" (*Opuntia spp.*) e *D. m. baja* utiliza o cacto agria (*S. gummosus*). Por outro lado, do complexo *buzzatti* do subgrupo *mulleri*, nove das dez espécies são polimórficas assim como *D. buzzatti* (FELLOWS; HEED, 1972; JENNINGS; ETGES, 2010; MARKOW; HOCUTT, 1998; RUIZ; HEED, 1988).

A mudança de hospedeiro geralmente se dá entre hospedeiros filogeneticamente mais próximos (TILMON, 2008; DATE, 2013), o que não necessariamente significa uma maior similaridade química entre as plantas (DATE, 2013). Apesar disso, vários subgrupos pertencentes ao grupo *repleta* foram capazes de superar diversas adversidades, como a presença de metabólitos tóxicos, e invadir desertos neotropicais, comumente inóspitos para outros representantes de *Drosophila* (WASSERMAN, 1982; HASSON, 2019).

A íntima e bem-sucedida associação entre planta e inseto se dá pela interação entre características físico-químicas da planta, que determinam a adequação do inseto à ela, e as

características do inseto (RUIZ; HEED, 1988). Esses insetos, ditos fitófagos, acabam se especializando e apresentam adaptações para com o hospedeiro em sua fisiologia, morfologia e comportamento, que são favorecidas pela seleção natural e garantem um maior valor adaptativo (FOGLEMAN; ABRIL; 1990; FUTUYAMA; MORENO, 1988; JAENIKE, 1978). A oviposição é garantida por dois atributos principais relacionados à planta: localização e adequação. Como já comentado, a adequação depende das características físico-químicas da planta, que diferem principalmente na quantidade de nitrogênio e água, componentes estritamente relacionados com as taxas de crescimento larval; além de metabólitos secundários que auxiliam na localização do hospedeiro pelo inseto; ou apresentar elevados níveis de toxicidade. A localização do hospedeiro está associada tanto à aparência e abundância da planta, quanto à sensibilidade e ao comportamento do inseto (RUIZ; HEED, 1988).

A utilização de cactos como hospedeiros por espécies de *Drosophila*, resultou no surgimento de adaptações que refletem a especificidade para com esses organismos. Os cactos apresentam desafios aos insetos como a presença ou não de fitoquímicos tóxicos em seus tecidos necróticos (FOGLEMAN; ABRIL; 1990), além de serem encontrados em ambientes áridos e semiáridos, sendo o estresse hídrico, consequência das altas temperaturas, uma pressão seletiva importante. Exemplificando, foi demonstrado que a presença de glicosídeos triterpênicos em cactos, resultou em uma significativa redução de aptidão de *D. mojavensis* e *Drosophila nigrospiracula* e inibiu o crescimento de algumas leveduras e bactérias (KIRCHER, 1977; STARMER et al., 1980; PHAFF et al., 1985; FOGLEMAN; ARMSTRONG, 1989). Tem-se também que um dos principais componentes do cacto agria, o esterol *macdougallin*, é responsável pela inibição do desenvolvimento larval em insetos (CESPEDES et al. 2005; MATZKIN, 2006).

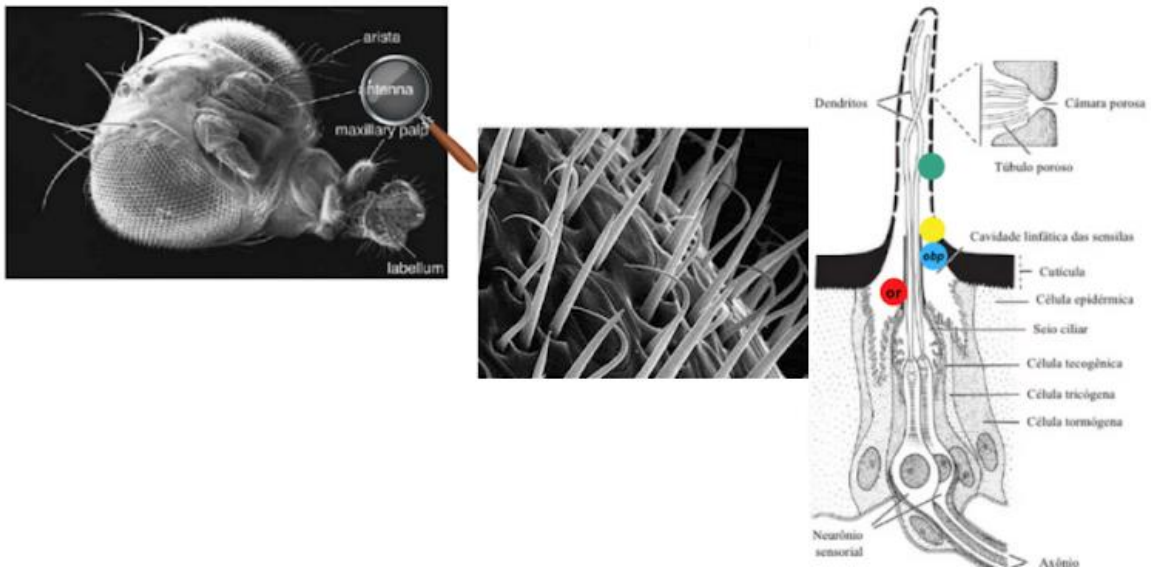
O surgimento de uma alternativa à *Opuntia* e a consequente migração para esse novo hospedeiro implicou em diversas novas adaptações, que possibilitaram aos representantes cactofílicos de *repleta* superar os desafios físico-químicos que os cactos colunares apresentam. Quando comparado aos cactos colunares, *Opuntia* possui um maior teor de água e apresenta níveis significativos de açúcares livres como glicose, frutose e galactose, ausentes nos colunares (RUIZ; HEED, 1988). A presença dos açúcares livres está correlacionada com um menor tempo de desenvolvimento larval, pois eles servem como fonte primária de carbono ou fonte energética que permitem o crescimento microbiano (BRAZNER et. al, 1984). Já os cactos colunares produzem metabólitos secundários como alcalóides, glicosídeos de triterpeno, ácidos graxos e fitosteróis, muitos tóxicos para as larvas de algumas espécies de *Drosophila*,

implicando em mais algumas adaptações que garantam a sobrevivência e o sucesso reprodutivo (RUIZ; HEED, 1988).

A decomposição dos tecidos dos cactos realizada pelos microorganismos, bactérias e leveduras (HASSON et al., 2019), principalmente sobre os açúcares, resulta em produtos voláteis que estão, ao que tudo indica, relacionados com a seleção dessas plantas hospedeiras (FOGLEMAN, 1982). Outros produtos da fermentação incluem ésteres, ácidos graxos e álcoois com peso molecular baixo. As concentrações e composição desses produtos voláteis variam a cargo de diversos fatores como altas concentrações de glicose e ramnose (açúcares importantes na produção de voláteis), assim como a temperatura (FOGLEMAN; ABRIL; 1990), espécie de cacto, idade do tecido, umidade (DATE, 2013), além do conjunto/diversidade de microorganismos que variam sazonalmente e entre cactos (FOGLEMAN; ABRIL; 1990). A presença de xenobióticos, especialmente alcalóides, é uma das razões que definem a especificidade da relação planta-hospedeiro. Diferentes estudos demonstram que os alcalóides de feniletilamina, como mescalina e tricocerina, presentes do cacto colunar *Trichocereus terscheckii*, são mais prejudiciais a *D. buzzatti* promovendo um baixo desempenho da mosca nesse hospedeiro, do que para sua espécie críptica *Drosophila koepferae*. Esses alcalóides foram também encontrados em espécies pertencentes ao gênero *Trichocereus*, como *T. panachanoi*, e no peiote (*Lophophora williamsii*) (HASSON, 2019; CORIO et al. 2013; PADRÓ et al. 2014; SOTO et al. 2014).

O estabelecimento da relação entre o inseto e seu sítio de hospedagem se inicia com a identificação da planta hospedeira, realizada pelo reconhecimento e discriminação dos voláteis produzidos pelos cactos hospedeiros, detectados pelo sistema olfativo (DATE, 2013). Em *Drosophila*, o sistema olfativo está circunscrito ao palpo maxilar e ao terceiro segmento antenal (STOCKER, 1994), onde estão localizadas as sensilas olfativas, estruturas contendo os dendritos dos neurônios olfativos banhados em linfa (Figura 1). É pela ativação dos neurônios sensoriais olfativos (OSNs) que se desenrola a codificação olfativa: os odorantes adentram a sensila através de seus poros na parede cuticular, se dissolvem na linfa que contém proteínas de ligação odorante (OBPs) e assim ativam os OSNs (SMITH, 2001). As antenas acomodam cerca de 1.200 OSNs, enquanto os palpos maxilares abrigam por volta de 120 (STOCKER, 1994). Os OSNs, em sua maioria, expressam um receptor olfativo (OR), que vão receber os odorantes das OBPs, e um co-receptor *Orco*, envolvido no ajuste e localização dos ORs nos dendritos, além de estar envolvido na transdução de sinal (LARSSON et al. 2004; SATO et al. 2008; WICHER et al. 2008; DATE, 2013).

Figura 1. Sistema olfativo de *Drosophila* (Modificado de BENTON; DAHANUKAR, 2011) ; Secção Longitudinal de Sensila multiporosa (Modificado de GRAY,1960)



Modificações na estrutura proteica de OBPs, ORs e até mesmo de receptores ionotrópicos (IRs), podem alterar a capacidade de detecção de alguns odores, devido às diferenças nos mecanismos de transdução de sinal e propriedades de ligação dos receptores (McBRIDE; ARGUELLO, 2007; ROLLEMANN et al, 2010). A especificidade de *D. sechellia* para com seu hospedeiro, o fruto da planta morinda, que é tóxico para outras espécies de *Drosophila*, foi vinculada ao aumento de expressão de *Or*22a, responsável pelas respostas a alguns odorantes do fruto dessa planta (KOPP et al. 2008; DATE, 2013). STORKHULL et al. (2005) mostrou como a expressão incorreta de *Or*43a pode afetar o comportamento de *D. melanogaster* em evitar benzaldeído.

As *Obp*s correspondem a uma família com mais de 50 genes polimórficos expressos nas mais variadas sensilas, que evoluíram por duplicação, perda e seleção dos genes. São pequenas proteínas globulares (entre 135-220 aminoácidos), organizadas em agrupamentos (*cluster*), com seis cisteínas conservadas, sintetizadas pelas células acessórias instaladas à volta dos neurônios, e que formam a maioria das proteínas constituintes da linfa dos insetos. Essas proteínas podem estar envolvidas com algumas funções como inativação de estímulos e no estabelecimento do código olfativo, também como outras funções fisiológicas, já que não estão restritas aos tecidos olfativos e gustativos. Sua principal incumbência é a de ligar as moléculas

odoríferas nos poros das sensilas quimiossensoriais, e transportá-las até os receptores olfativos (ORs) (SANCHEZ-GRACIA; ROZAS, 2008; MARKOW, 2019; JOHNSTUN, 2021).

A detecção de odor pelas OBPs acontece de forma combinatória, onde uma única OBP afeta as respostas de múltiplos odores, e um único odorante é detectado por mais de uma OBP (SWARUP et al. 2011). Acetofenona, benzaldeído, hexanol são alguns dos compostos conhecidos por possuírem afinidade com OBPs (WANG et al. 2007; ARYA et al. 2010; WANG et al. 2010; SWARUP et al. 2011). *Obp*76a, ou LUSH, está associada a respostas aos odorantes de benzaldeído ou etanol, e a falta dessa proteína priva a identificação desses compostos (KIM et. al, 1998; WANG et al. 2001; HEKMAT-SKAFE, 2002).

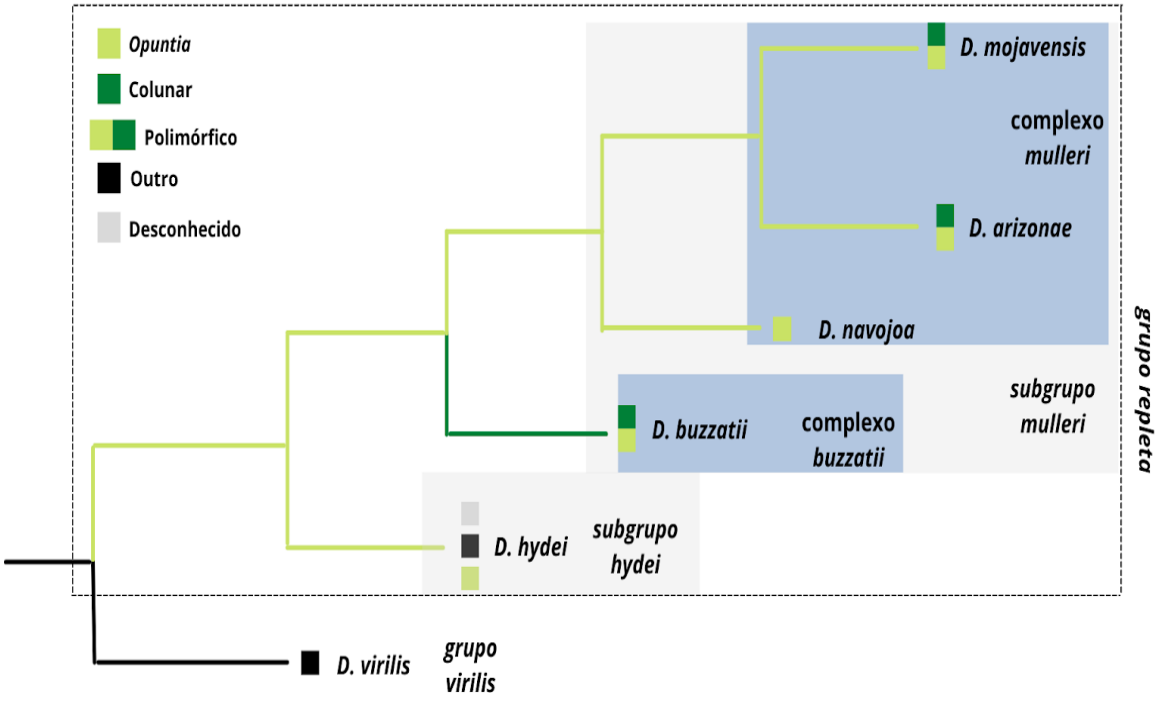
Os *Or*s pertencem a uma família contando com mais de 60 membros (MARKOW, 2019), que podem ser encontrados por todo o genoma (VIEIRA et. al, 2007). São receptores com domínios transmembrana, constituídos por cerca de 400 resíduos de aminoácidos, e que também evoluíram por duplicação, perda e seleção (GOU; KIM, 2007). É interessante notar o tamanho dessa família em *Drosophila* em comparação a outros organismos, já que se estima cerca de 100 *Or*s em peixes (NGAI et al. 1993) e cerca de 1.000 em camundongos e ratos (MOMBAERTS, 1999). Seu principal encargo é receber a molécula odorante das OBPs, transmitir a informação para o sistema nervoso, para que assim o odor possa ser reconhecido e desse modo detectar o ambiente. Os estímulos químicos externos reconhecidos são primeiramente transformados em pulsos elétricos, em seguida são retransmitidos e processados pelo sistema nervoso (GUO; KIM, 2015; MARKOW, 2019).

As populações de *D. mojavensis* do deserto de Mojave (Califórnia, América do Norte) são um excelente exemplo para demonstrar a especificidade e funcionamento desses receptores. A expressão de *Or*71a e *Or*46a aumentam, já que ambos receptores estão associados com 4-propilfenol e 4-metilfenol, aromáticos esses muito presentes em cactos de barril (*Opuntia*), hospedeiro de *D. mojavensis*. Observa-se, por conseguinte, um aumento na expressão de *Or*s sensíveis à aromáticos, e uma diminuição de receptores odorantes sensíveis a ésteres e álcoois, como *Or*42a, *Or*47b e *Or*59b e *Or*85a (DATE, 2013).

Por oferecer grandes vantagens adaptativas, os genes responsáveis pela localização da planta hospedeira e que codificam as proteínas de ligação odorante e receptores olfativos devem ter sido alvos de seleção. Para investigar e demonstrar qual o processo seletivo atuante sobre esses genes, e adicionalmente seus ortólogos em cada espécie, este estudo espera detectar assinaturas de seleção positiva (diversificadora) em genes codificantes de proteínas

associadas à localização do hospedeiro, em espécies do grupo *repleta* que apresentam variações na preferência ao uso de cactos pertencentes a diferentes subgrupos: subgrupo *hydei* (*Drosophila hydei)* e subgrupo *mulleri* (complexo *buzzatii*: *Drosophila buzzatii* e complexo *mulleri*: *Drosophila mojavensis, Drosophila arizonae* e *Drosophila navojoa*) que além de ainda utilizarem *Opuntia*, desenvolveram preferência por tipos específicos de cactos colunares (Figura 2).

Figura 2. Representação esquemática adaptada de OLIVEIRA et al. (2012) do uso do hospedeiro nas espécies do grupo *repleta* e o grupo externo analisadas neste trabalho. Os cactos hospedeiros discutidos são identificados por cores diferentes; a identificação 'outros' refere-se a outros substratos, que não sejam cactos.



Não há estudos anteriores usando uma abordagem ampla para análise do processo seletivo associado aos genes quimioreceptores em espécies cactofílicas como a empregada neste estudo. Três trabalhos analisaram o processo seletivo em algumas espécies cactofílicas, contudo, em abordagens diferentes e utilizando menor número de espécies, ou outras classes de genes. Rane et al. (2019 a), por exemplo, investigaram as mudanças genéticas associadas à adaptação a ambientes áridos utilizando cinco espécies do grupo *repleta* (*D. aldrichi*, *D. mojavensis*, *D. buzzatii*, *D. repleta* e *D. hydei*), no qual o foco do estudo foi a comparação da ocorrência de seleção positiva entre genes ortólogos e parálogos, e também para genes codificadores de proteínas associadas às funções identificadas pela análise de Gene Ontology

(GO) em transcriptomas dessas espécies, e não a genes específicos. Em um segundo estudo (RANE et al, 2019 b), incluindo 14 espécies de *Drosophila* e dentre elas seis cactofílicas, esses autores analisaram genes codificadores de proteínas detoxificadoras (citocromo P450, GSTs, esterases, ABCs e UDPs) tendo como foco também a comparação entre genes duplicados ou não. Já Diaz et al. (2018), utilizaram parte dos métodos empregados neste trabalho, contudo, com um número limitado de genes quimiorreceptores, um receptor gustativo (*Gr*) e três receptores odorantes (*Or*) em apenas duas espécies, *D. mojavensis* e *D. arizonae*.

Para testar a hipótese é que genes que codificam proteínas em espécies que se adaptaram à utilização de cactos com características bioquímicas diferentes, mostram sinais de terem passado por seleção positiva foram realizadas análises do processo seletivo buscando identificar assinaturas de seleção positiva em genes codificadores de proteínas de ligação de odorantes (OBPs) e receptores olfativos (ORs) que são associadas à localização do hospedeiro em cinco espécies do grupo *repleta* adaptadas à utilização de sítios de hospedagem e reprodução com características diferentes.

2 OBJETIVOS

O objetivo deste trabalho foi identificar assinaturas de seleção positiva em genes que codificam proteínas associadas à localização do hospedeiro em cinco espécies do grupo repleta adaptadas à utilização de sítios de hospedagem e reprodução com características diferentes. Para isso, foram formulados os seguintes objetivos específicos:

1. Buscar em bancos de dados públicos sequências de genes codificadores de proteínas de ligação de odorantes (OBPs) e receptores olfativos (ORs), relacionadas à função de localização do hospedeiro em genomas das espécies cactofílicas do grupo *repleta D. mojavensis*, *D. arizonae*, *D. navojoa*, *D. buzzatii* e *D. hydei*, como também em *D. virilis* (grupo *virilis*, grupo externo);

2. Identificar e anotar as CDSs dos genes codificadores das proteínas OBP e OR, e identificar os genes ortólogos entre seis espécies;

3. Identificar genes parálogos em cada genoma, e excluí-los da análise, de modo a utilizar nas análises apenas os genes ortólogos compartilhados pelas 6 espécies;

4. Caracterizar o processo seletivo (seleção negativa, positiva ou evolução neutra) atuantes nos genes ortólogos das duas famílias gênicas.

3 MATERIAL E MÉTODOS

3.1 Anotação dos Genes

Para se obter as sequências do genes que codificam as proteínas OBPs e ORs, identificadas na literatura (SÁNCHEZ-GRACIA; ROZAS, 2008; GOU; KIM, 2007) por serem relacionadas à localização do hospedeiro em espécies de *Drosophila*, utilizou-se genomas sequenciados de espécies que mudaram ou não a utilização do cacto *Opuntia* para cactos colunares: cinco representantes do grupo *repleta*, sendo eles: *D.hydei*, *D. buzzatii*, *D. mojavensis* (ZIMIN et al. 2008), *D. arizonae* (SANCHEZ-FLORES et al., 2016) e *D. navojoa* (SANCHEZ-FLORES et al., 2016); e um representante como grupo externo: *D. virilis* (MILLER et al., 2018). As anotações gênicas para estes genomas estão públicas no repositório NCBI (https://www.ncbi.nlm.nih.gov/) e foram recuperadas com um *script* de linguagem de programação BASH desenvolvido no Laboratório de Evolução Molecular, que é capaz de obter apenas a sequência codificante (CDS) mais longa de cada *Obp* e de cada *Or*, para cada genoma. Desta forma, foram obtidas uma CDS para cada *Obp* e *Or* descrita nestes genomas, exceto para *D. buzzatii*, que não possui genoma público no NCBI. Para utilizar o *script*, primeiramente, foram recuperados manualmente os IDs disponíveis de *Obp*s e *Or*s, as CDSs e as sequências traduzidas de cada espécie disponíveis no NCBI. O *script* filtra os IDs registrados para ambas classes gênicas do arquivo contendo as sequências nucleotídicas das CDSs e as CDSs traduzidas (proteínas codificadas); ao final, tem-se somente as sequências codificantes encontradas das *Obps* e *Or*s presentes no genoma de cada espécie.

Para a anotação genômica de CDSs para *D. buzzatii* foi utilizado como referência o genoma de *D. mojavensis*, tendo em vista a proximidade filogenética de ambas espécies. Antes de tudo, foi necessário filtrar as CDSs resultantes tanto das *Obp*s quanto das *Or*s de *D. mojavensis*, para que se tivesse uma lista somente com os IDs de cada gene, sem as sequências. O segundo passo foi obter o genoma de *D. buzzatii*, sequenciado pela Universidade Autônoma de Barcelona (UAB), disponível em seu site (https://dbuz.uab.cat/download.php). Em seguida, utilizando o genoma total de *D. mojavensis* como referência, alinhou-se o genoma *D. buzzatii* empregando a ferramenta Liftoff. O Liftoff realiza uma anotação genômica através do alinhamento dos éxons a fim de potencializar a paridade da sequência, ao invés de alinhar os genomas inteiros, preservando assim a estrutura gênica e o transcrito. Utilizando anotações em *General Feature Format* (GFF) ou *General Transfer Format* (GTF), foi alinhado genes de um genoma de referência para um genoma alvo (SHUMATE; SALZBERG, 2021). Como o genoma

alinhado pelo Liftoff não conta as sequências anotadas, e sim suas coordenadas, foi empregado o programa Gffread (PERTEA, PERTEA, 2020), para gerar um arquivo com o genoma completo, isto é, com as sequências, de *D. buzzatii*. Esse arquivo final foi então filtrado a partir da lista de IDs de *D. mojavensis*, obtida logo no ínicio, e utilizando a ferramenta seqtk subseq, as sequências CDS de *Obp*s e *Or*s presentes no genoma de *D. buzzatii* foram enfim recuperadas. Seqtk é uma ferramenta capaz de processar sequências no formato FASTA ou FASTQ; já o subseq é um comando pertencente à ferramenta que extrai as sequências do arquivo (SHEN et al, 2016). Samtools é um conjunto de ferramentas que manipulam alinhamentos, podendo classificá-los, mesclá-los, entre outras funções (LI, 2011; 2009); e foi utilizada para organizar as sequências pelos seus tamanhos, para que posteriormente manualmente, as isoformas pudessem ser identificadas com o intuito de selecionar a maior entre elas, em caso de splicing alternativo, a fim de maximizar a cobertura genômica minimizando redundância. As maiores sequências então foram traduzidas pela Emboss Transeq (https://www.ebi.ac.uk/Tools/st/emboss_transeq/).

Por fim, todas proteínas de *D. buzzatii* identificadas como possíveis *Obp*s e *Or*s foram submetidas a análise de domínio proteico com a ferramenta C-DART (GEER et al., 2002), a fim de examinar se estas proteínas teriam o domínio proteico de OBPs e ORs. Com base nessa filtragem, dois genes *Obp*s foram removidos, LOC6579245 (Obp58b) e LOC6579191 (Obp50c). Considerou-se proteínas homólogas às OBPs e às ORs somente as que apresentaram o maior *score* de similaridade com o domínio proteico de OBPs e ORs já conhecidas.

Tabela 1. Lista de espécies *Drosophila* utilizadas neste estudo, classificado em subgênero e grupo, com seus respectivos sítios hospedeiros de reprodução e hospedagem, assim como os números de acesso do NCBI (https://www.ncbi.nlm.nih.gov/genome/) e a qualidade desses genomas. UAB: Projeto genoma de *D. buzzatii* conduzido pela Universidade Autônoma de Barcelona.

Espécies	**Sítios Hospedeiros**	**Acesso do Genoma**	***Scaffolds***	**N50**
Grupo *repleta*				
D. hydei	polífaga	GCA_003285905.2	217	3.367.158
D. arizonae	*Opuntia*, cacto colunares	GCA_001654025.1	3.178	26.536.676
D. m. wrigleyi	*Opuntia*	GCA_000005175.1	6.841	24.764.193
D. navojoa	*Opuntia*	GCA_001654015.2	13.813	389.283
D. buzzatii	Principalmente cacto colunar *F. cylindraceus, Opuntia*	*D. buzzatii* UAB	826	1.380.941
Grupo *virilis*				
D. virilis	ritidomas em decomposição	GCA_007989325.2	45	31.075.311

2.2 Genes ortólogos e parálogos

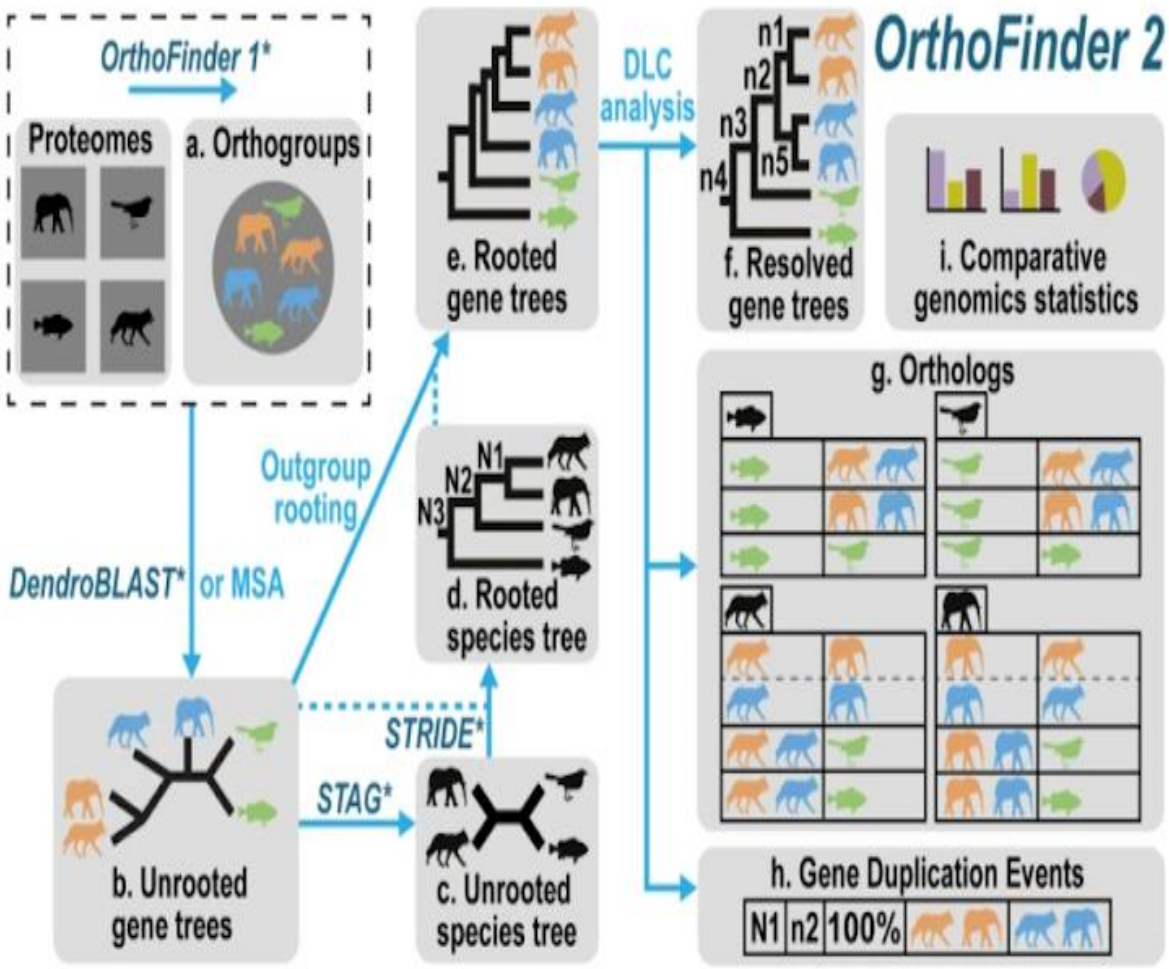
Utilizando todas as sequências proteicas anotadas para os genes das *Obp*s e das *Or*s, os genes ortólogos entre as espécies e parálogos de cada espécie foram preditos por meio do *software OrthoFinder*. O *Orthofinder* fornece uma inferência filogenética de ortólogos, eventos de duplicação de genes (parálogos), árvore dos genes e espécies enraizadas, além de estatísticas comparativas genômicas (EMMS; KELLY, 2019). Define-se como parálogos os genes gerados por evento de duplicação na mesma espécie e ortólogos como os genes originados por evento de especiação a partir de um único gene presente no ancestral comum das espécies em questão (ALTENHOFF; DESSIMOZ, 2012). O *software* também gera ortogrupos, que podem ser definidos por agrupamentos de genes que descendem de um único gene ancestral comum de todas as espécies estudadas (WAPINKS et al., 2007).

Para explicar o funcionamento do *OrthoFinder*, a Figura 3 exemplifica o processo. O *software* é executado por comando, direcionando a entrada de um arquivo FASTA por espécie contendo as sequências proteicas das espécies de interesse. O algoritmo original do programa infere os ortogrupos, e utilizando *DendroBLAST* constrói uma árvore gênica não enraizada para cada ortogrupo. O algoritmo *STAG* então infere a partir do conjunto de árvores dos ortogrupos outra árvore de espécies não enraizada, que com a utilização do algoritmo *STRIDE* acaba sendo enraizada para que então possam ser identificados os eventos de duplicação de genes no conjunto de árvores de ortogrupos. A árvore STAG de espécies

enraizadas é empregada para enraizar as árvores de ortogrupos, assim essas árvores são enraizadas por um algoritmo híbrido combinado modelo de duplicação-perda-coalescente e o método de sobreposição de espécies, para que eventos de duplicação e ortólogos sejam inferidos e estatísticas comparativas gerais sejam calculadas (EMMS; KELLY, 2019).

Os arquivos FASTA contendo as sequências proteicas das OBPs e ORs das sete espécies foram devidamente separados e o *OrthoFinder* foi executado no terminal Linux com um único comando: *orthofinder -f $PATH -S blast*, onde *-f* indica o caminho em que os arquivos fastas estão localizados no ambiente Linux e *-S* o algoritmo utilizado para as comparações proteicas, no qual o usuário pode optar por *Blast* ou *Diamond*, onde *Diamond* é uma otimização do *Blast* e só deve ser selecionado em casos de análises que envolvem alto número de espécies (centenas) e alto número de sequências (milhares).

Figura 3. Visão geral do fluxo de trabalho do OrthoFinder. Retirado de (EMMS; KELLY, 2019). A - Inferência dos ortogrupos. B - Inferência da árvore gênica. C - Inferência da árvore de espécies. D - Enraizamento da árvore de espécies. E - Enraizamento da árvore gênica. F - Inferência dos ortólogos e eventos de duplicação pela utilização do algoritmo híbrido + análise DLC das árvores gênicas. G- Ilustração da tabela de resultados dos genes ortólogos em cada espécie. H - Ilustração da tabela de eventos de duplicação de genes. I - Ilustração da tabela de estatísticas comparativas genômicas gerais.



3.3. Assinaturas de Seleção Positiva

O papel da seleção natural no processo de divergência funcional de sequências ortólogas pode ser evidenciado por meio do valor ω (ômega), uma estimativa do processo evolutivo atuante sobre sequências gênicas decorrente da razão entre as substituições não sinônimas (Ka) e as substituições sinônimas (Ks) (ω =Ka/Ks) que se acumularam ao longo do tempo de divergência entre as sequências nucleotídicas de genes. Os valores de ω positivos (ω>1)

indicam que a seleção natural atua de forma a fixar maior número de substituições de aminoácidos do que sinônimas, ocorrendo em condições favoráveis para a fixação dessas substituições chamada, então, de seleção positiva ou diversificadora. Já os valores de ω negativos (ω<1) evidenciam que as substituições de aminoácidos são desvantajosas, ocorrendo assim maior fixação de substituições nucleotídicas sinônimas do que não sinônimas, indicando a força de seleção natural negativa ou purificadora, que resulta na conservação das sequências de aminoácidos. Já quando a evolução é neutra, espera-se que as taxas de substituições sinônimas e não sinônimas sejam similares e assim o valor ω seja igual a 1 (KIMURA, 1980).

3.3.1 Testes de Seleção Par a Par

O teste de seleção par a par (NEI; GOJOBORY, 1986) é um teste baseado nas sequências de nucleotídeos, onde é calculado o valor Ka/Ks entre pares de espécies analisadas. Os testes foram realizados no software DnaSP6 (ROZAS et al., 2017), a partir das sequências nucleotídicas em FASTA alinhadas no *software* MEGA7 (KUMAR et al., 2016).

O DnaSP6 estima o número de sítios com substituições não sinônimas (Ka), sítios onde ocorreram trocas de nucleotídeos e consequente troca de função dos aminoácidos, e o número de sítios com substituições sinônimas (Ks), ou silenciosas, sítios onde ocorreram ou não trocas de nucleotídeos, mas não ocorreram troca de aminoácidos (função). O *software* também calcula o valor de Pi (π), diversidade nucleotídica entre duas espécies, e ainda a extensão de polimorfismos nas regiões codificadoras e não codificadoras (ROZAS et al., 2017).

3.3.2 Testes de Seleção por Sítio

Para identificar a seleção atuante sobre os genes ortólogos *Obp* e *Or* preditos pelo *Orthofinder* foram então submetidos a testes de seleção por sítio por meio do software CODEML, do pacote PAML X (YANG, 2000; YANG, 2007, YANG, 2013), e dos softwares MEME, FEL e FUBAR do pacote HYPHY (https://www.hyphy.org/), que permitem a identificação dos diferentes valores de ω (Ka/Ks: substituições não sinônimas/substituições sinônimas), pelo método de YANG e NIELSEN (2000), nos diferentes códons das sequências em análise.

3.3.2.1 PAMLX

O PAMLX é um pacote com programas voltados à análise filogenéticas de DNA e sequências proteicas utilizando o método *maximmum likelihood* (ML). Deste pacote foi utilizado o programa codeML (YANG, 2007, YANG, 2013), o qual utiliza como *input* as

sequências de nucleotídeos e a árvore filogenética das espécies construída com as sequências do gene em análise. As sequências de nucleotídeos foram alinhadas e a árvore filogenética foi construída utilizando o *software* MEGA7 (KUMAR et al., 2016). O arquivo FASTA das sequências de nucleotídeos alinhadas de cada gene foi convertido em um arquivo PML compatível com o programa, utilizando o software DAMBE (*Data Analysis in Molecular Biology and Evolution*) (XIA; XIE, 2001**)**. O arquivo com a árvore filogenética MTSX, *output* padrão do MEGA7, foi convertido em arquivo NEWICK no próprio MEGA7.

Os testes de seleção por sítio testa as seguintes hipóteses, com o uso dos diferentes modelos de valores de seleção para sítios: (i) todos os códons apresentam o mesmo valor ω (M0 – modelo nulo); (ii) diferentes códons apresentam dois diferentes valores ω, $0 < \omega < 1$ e $\omega = 1$ (M1- modelos neutro); (iii) semelhante ao M1, suplementarmente permite uma adicional classe de códons com estimativas livres de ω (M2 - modelo de seleção); (iv) estima o valor de ω com uma beta distribuição acima do intervalo (0, 1) (M7 - modelo neutro); (v) adicionando novos parâmetros no M7, onde códons podem apresentar uma categoria extra de sitíos com seleção positiva $\omega > 1$ (M8 - modelo seleção positiva). As hipóteses desenvolvidas para identificação de valor ω, foram testadas pelo teste $\chi 2$, com a comparação dos valores de lnL de cada hipótese. Implementado sobre os modelos M2 e M8, onde há o cálculo de seleção, o BEB (*Bayes empirical Bayes*) (YANG et al., 2005) calcula as probabilidades posteriores (*: >95%, **:> 99%) para classes de sítios, a fim de identificar a significância no teste para sítios detectados sob seleção positiva.

Adicionalmente foi utilizado o *toolkit Environment for Tree Exploration* – ETE (http://etetoolkit.org/, HUERTA-CEPAS et al. 2010) para melhor visualização dos resultados das análises do codeML.

3.3.2.2 HYPHY

A fim de complementar e comparar os resultados com os obtidos com o uso do codeML, foram realizados três testes de seleção disponíveis pelo pacote Hyphy (https://www.hyphy.org/), sendo esses MEME (*Mixed Effects Model of Evolution*, MURREL et al., 2012); FEL (*Fixed Effects Likelihood method*, KOSAKOVSKY POND & FROST, 2005) e FUBAR (*Fast Unconstrained Bayesian Approximation*, MURELL et al., 2013).

O teste MEME foi desenvolvido baseado em métodos de efeitos filogenéticos casuais de uma larga classe de sítios, permitindo a distribuição da variação entre sítios e entre ramos (o efeito causal), sendo um teste capaz de identificar as marcas de seleção positiva tanto

episódica quanto diversificadora. Emprega a abordagem de *maximum likelihood* (ML), com o objetivo de detectar sítios que evoluem sob seleção positiva em uma proporção de ramos (MURELL et al., 2012).

FEL infere diretamente a razão Ka/Ks em cada sítio, utilizando uma abordagem de *maximum likelihood* (ML), sendo a pressão da seleção constante ao longo de toda a filogenia (KOSAKOVSKY POND & FROST, 2005). O modelo do teste deduz independentemente os parâmetros para cada sítio, contudo isso significa que a inferência individual de um sítio não informa expectativas relacionadas a outro, sendo esse efeito um ponto que pode resultar em estimativas sítio-específicas não tão confiáveis (MURELL et al., 2013).

A especificidade de parâmetros estabelecidos por alguns modelos de teste de seleção, caracteriza sítios em um pequeno número de classe de sítios, forçando sítios a pertencerem a uma dessas classes, podendo resultar em dados enganosos. O teste FUBAR utiliza o método hierárquico bayesiano usando Markov chain Monte Carlo (MCMC), que garante resultados robustos contra a especificação de modelos a partir de definir uma média maior de classes de sítios, deixando os parâmetros mais desprendidos, e assumindo que a pressão de seleção é constante ao longo de toda a filogenia (MURELL et al., 2013).

3.3.3 Predições Topológicas e Análises de Alterações nas Propriedades Bioquímicas

O teste PRIME (*PRoperty Informed Model of Evolution*), disponível no pacote HyPhy (https://www.hyphy.org/), foi realizado para analizar as substituições em nível de aminoácidos caracterizando as alterações nas suas propriedades bioquímicas seguindo as cinco propriedades estabelecidas por Atcheley (ATCHELEY et al, 2005): índice de polaridade (P), fator de estrutura secundária (S), volume (V), refratividade e ponto isoelétrico (I).

Os sítios que apresentaram evidências de seleção positiva foram mapeados pelo *software* TMHMM Server v2 (http://www.cbs.dtu.dk/services/TMHMM/), que prediz a topologia de receptores. O diagrama da estrutura 2D desses receptores foram criadas utilizando PROTTER (OMASITS et al., 2014)**.**

Para as proteínas globulares OBPs, sua estrutura putativa 3D foi predita utilizando o Servidor de modelagem automatizada SWISS-MODEL (swissmodel.expasy.org/interactive) (GUEX, 1997). As sequências de aminoácidos de *Drosophila* foram utilizadas como *input* para pesquisar as sequências disponíveis no Protein Data Bank (PDB), e as sequências com a maior

pontuação PSI-BLAST, ou seja, com a maior similaridade, foram utilizadas para a visualização da estrutura 3D, destacando os sítios com relevância nas análises evolutivas.

4 RESULTADOS

4.1 Identificação da Família OBP e OR

A extração total dos genes de *Obp*s nos sete genomas públicos e em *D. buzzatii* por meio do *software* AUGUSTUS permitiu a identificação total de 202 genes *Obp*s e 353 genes *Or*s. Na Tabela Suplementar 1 são apresentados os números dos genes das duas famílias extraídos de cada genoma, que variou de 26 em *D. arizonae* a 42 em *D. virilis* para *Obp*, e de 56 em *D. navojoa*, *D. buzzatii* e *D. hydei* a 65 em *D. virilis* para *Or*.

A diferença no número total de *Obp*s e *Or*s anotados entre os genomas é esperada devido a aspectos biológicos, como ganho e perda gênica, bem como por aspectos técnicos, visto que a qualidade da montagem do genoma, como métricas de cobertura e N50 afetam diretamente a capacidade da predição gênica.

4.2 Análise descritiva

As sequências proteicas anotadas para os genes *Obps* e *Ors* foram utilizadas para predizer ortólogos e parálogos por meio do *software Orthofinder*. Dentre os resultados encontrados com *Orthofinder* está a estatística geral (Tabelas Suplementares 2 e 3). Desta forma, contabilizou-se o número de espécies, o número total de genes, os genes que foram incluídos em ortogrupos e a respectiva porcentagem, os que não foram assinalados e, portanto, não foram agrupados, o número de ortogrupos resultantes e os ortogrupos espécies-específicos (que incluem genes de somente uma espécie). Na presente análise, quase todos os genes anotados e investigados pelo *software* foram atribuídos a um ortogrupo, indicando uma forte relação entre as sequências protéicas analisadas entre as espécies, que são relacionadas de forma relativamente próxima. Os 202 genes *Obps* e os 353 genes *Ors* foram classificados, respectivamente, em 29 e 50 ortogrupos.

4.3 Genes Ortólogos

Definindo genes ortólogos sendo resultantes de eventos de especiação de um único gene em um ancestral comum (ALTENHOFF; DESSIMOZ, 2012), o *Orthofinder* agrupou os genes que descendem de um único ancestral comum em um conjunto de espécies

em ortogrupos. As Tabelas Suplementares 4 e 5 apresentam os números de genes ortólogos compartilhados entre cada par de espécies.

4.4 Genes Parálogos

Definindo parálogos como os genes resultantes de eventos de duplicação na mesma espécie (ALTENHOFF; DESSIMOZ, 2012), o *Orthofinder* encontrou cerca de 5 genes parálogos *Obp,* sendo 3 genes parálogos em *D. hydei* e 2 em *D. virilis*, que podem ser observados na Tabela Suplementar 6. Para *Or,* foram encontrados um total de 12 genes parálogos, sendo 1 em *D. hydei*, 4 em *D. mojavensis*, 1 em *D. navojoa* e 6 em *D. virilis*, informados na Tabela Suplementar 7. Observa-se um maior número de eventos de duplicação em *Or*s do que em *Obp*s, assim como *D. virilis* apresenta um maior número de genes duplicados em ambas classes gênicas do que as espécies do grupo *repleta.*

4.5 Assinaturas de Seleção Positiva

Identificar as regiões de genes codificadores de proteínas que sofreram evolução adaptativa é de grande importância na Biologia Evolutiva. A impressão de sinais (assinaturas) da seleção natural atuante em genes que codificam proteínas é muitas vezes de difícil identificação, porque a seleção é frequentemente transitória ou episódica, ou seja, afeta apenas um subconjunto de linhagens. Os variados métodos, que levam em conta características da evolução das sequências gênicas, propostos para estimar taxas de substituição sinônimas e não sinônimas entre sequências codificadoras de proteínas, se enquadram em duas classes: métodos aproximados e métodos de máxima verossimilhança, que podem ser ainda classificados como baseados em nucleotídeos, ou em códons, de acordo com os modelos de mutação (revisado em KOSAKOVSKY POND; FROST, 2005).

4.5.1 Testes de Seleção Par a Par

Realizou-se uma análise a razão Ka / Ks (razão ω) entre pares de sequências de genes ortólogos usando o teste de Nei-Gojobori (NEI; GOJOBORI, 1986). Apesar da possibilidade de subestimação, este teste foi utilizado pela vantagem oferecida de inferir o processo seletivo entre pares de espécies. Os resultados dos pares (Figuras suplementares 1A e 1B) não mostraram evidências de seleção positiva em todas as comparações entre as sequências de *Obp e Or*, no entanto, existem exemplos (seis *Obp* e 13 *Or*) que indicam seleção purificadora relaxada (1,0 <ω> 0,3), representando divergência Ka em andamento. Nota-se que seis genes *Obp* evoluíram sob um regime seletivo menos restritivo do que os outros 10, que estão sob uma seleção purificadora mais restritiva. Além disso, o par de espécies com maior ω em cada gene

não é o mesmo, mostrando de maneira geral uma relação filogenética mais próxima (Tabela 2). A mesma situação pode ser observada em relação aos genes *Or*: 13 genes apresentam evidências de seleção purificadora relaxada, o par de espécies com maior ω em cada gene não é o mesmo. Nota-se que a maioria das comparações com ω > 0,3 envolvem as espécies *D. mojavensis, D. arizonae* e *D. navojoa* comparadas às demais. O gene *Or* 65b se destaca entre todos, apresentando várias comparações com ω entre 0,5 e 0,75 (Tabela 2).

Tabela 2. Comparação par a par entre as sequências dos genes de *Obp* (à esquerda) e *Or* (à direita) que indicam seleção purificadora relaxada (1.0 < ω > 0.3). Em negrito ω entre 0.5 – 0.75.

Genes *Obp*	Comparação	Comparação	**Genes *Or***
Obp 19b	*D. mojavensis x D. virilis*	*D. mojavensis x D. arizonae*	*Or* 1a
	D. arizonae x D. virilis		
	D. najovoa x D. virilis		
	D. hydei x D. virilis		
Obp 56b	*D. mojavensis x D. hydei*	*D. mojavensis x D. arizonae*	*Or* 9a
	D. arizonae x D. hydei	*D. mojavensis x D. najovoa*	
	D. najovoa x D. hydei	*D. arizonae x D. najovoa*	
	D. buzzatii x D. hydei		
Obp 19d	*D. mojavensis x D. buzzatii*	*D. mojavensis x D. hydei*	*Or* 10a
	D. arizonae x D. buzzatii	*D. arizonae x D. hydei*	
	D. najovoa x D. buzzatii	*D. najovoa x D. hydei*	
	D. hydei x D. buzzatii		
	D. virilis x D. buzzatii		
Obp 69a	*D. mojavensis x D. arizonae*	*D. mojavensis x D. arizonae*	*Or* 42a-like
	D. arizonae x D. najovoa		
Obp 56g	*D. mojavensis x D. arizonae*	*D. mojavensis x D. arizonae*	*Or* 33c
	D. mojavensis x D. najovoa	*D. mojavensis x D. najovoa*	
	D. arizonae x D. najovoa	*D. mojavensis x D. buzzatii*	
		D. mojavensis x D. hydei	
		D. arizonae x D. najovoa	
		D. arizonae x D. buzzatii	
		D. najovoa x D. buzzatii	
		D. arizonae x D. hydei	
		D. arizonae x D. virilis	
Obp 69a	*D. mojavensis x D. arizonae*	*D. mojavensis x D. arizonae*	*Or* 42a-like
	D. arizonae x D. najovoa		
Obp 83a	*D. mojavensis x D. buzzatii*	*D. mojavensis x D. navojoa*	*Or* 42b
	D. arizonae x D. buzzatii		
	D. hydei x D. buzzatii		
		D. mojavensis x D. navojoa	*Or* 42b-like
		D. mojavensis x D. buzzatii	

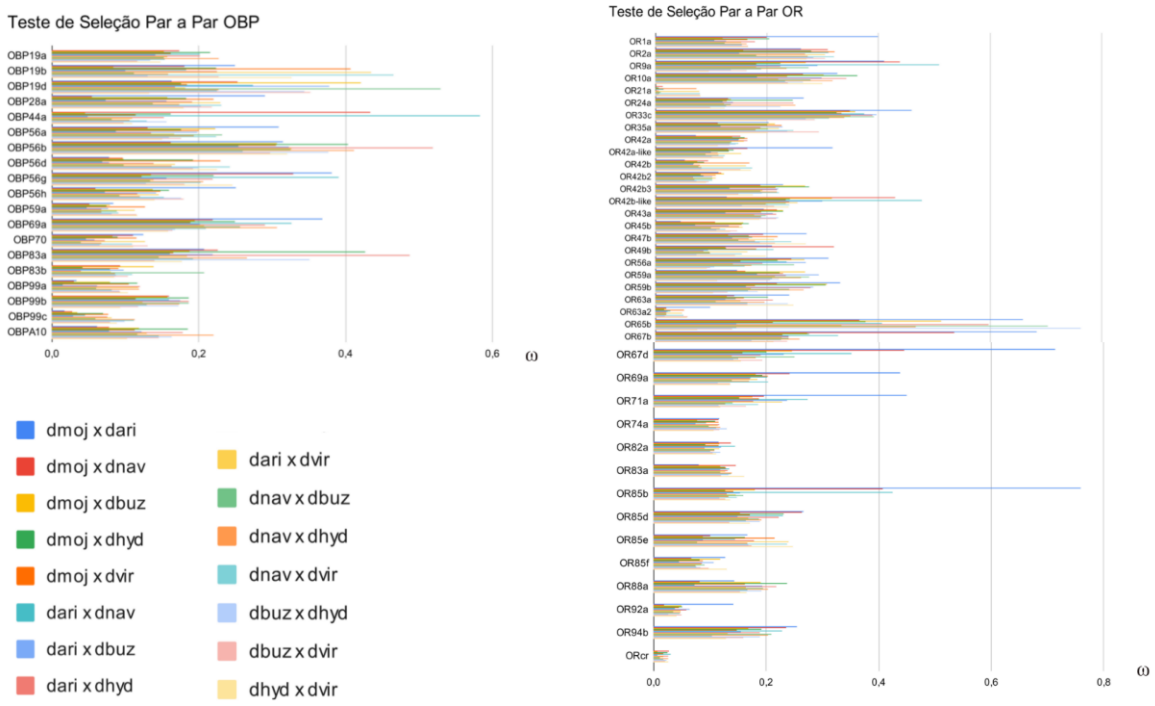
	D. arizonae x D. najovoa	

Tabela 2. Continuação

	D. mojavensis x D. arizonae	*Or* 59b
	D. mojavensis x D. buzzatii	
	D. mojavensis x D. buzzatii	
	D. mojavensis x D. arizonae (>0.5)	*Or* 65b
	D. mojavensis x D. najovoa	
	D. mojavensis x D. buzzatii (>0.5)	
	D. mojavensis x D. hydei	
	D. arizonae x D. najovoa	
	D. arizonae x D. hydei (>0.5)	
	D. arizonae x D. virilis	
	D. najovoa x D. buzzatii (>0.5)	
	D. najovoa x D. hydei	
	D. buzzatii x D. hydei (>0.5)	
	D. mojavensis x D. arizonae	*Or* 67b
	D. mojavensis x D. najovoa	
	D. arizonae x D. najovoa	
	D. mojavensis x D. arizonae	*Or* 67d
	D. mojavensis x D. najovoa	
	D. arizonae x D. najovoa	
	D. mojavensis x D. arizonae	*Or* 69a

Na representação gráfica apresentada na Figura 4, encontra-se ilustradas as variações dos valores de ω dos genes *Obp* e *Or*, com os maiores valores de seleção relaxada envolvendo as espécies do complexo *mulleri* (*D. mojavensis*, *D. arizonae* e *D. navojoa*) quando comparada com as demais.

Figura 4. Representação gráfica dos valores de ω dos genes *Obp* e *Or* estimados em análises de seleção par a par.



4.5.2 Teste de Seleção por Sítio

Uma grande coleção de modelos de substituição de códons usando métodos de ML foi desenvolvida, sendo a maioria derivada do método LRT (NIELSEN; YANG 1998; YANG et al. 2000), uma análise baseada em modelo de seleção natural que muitas vezes categoriza os sítios de em um número relativamente pequeno de classes.

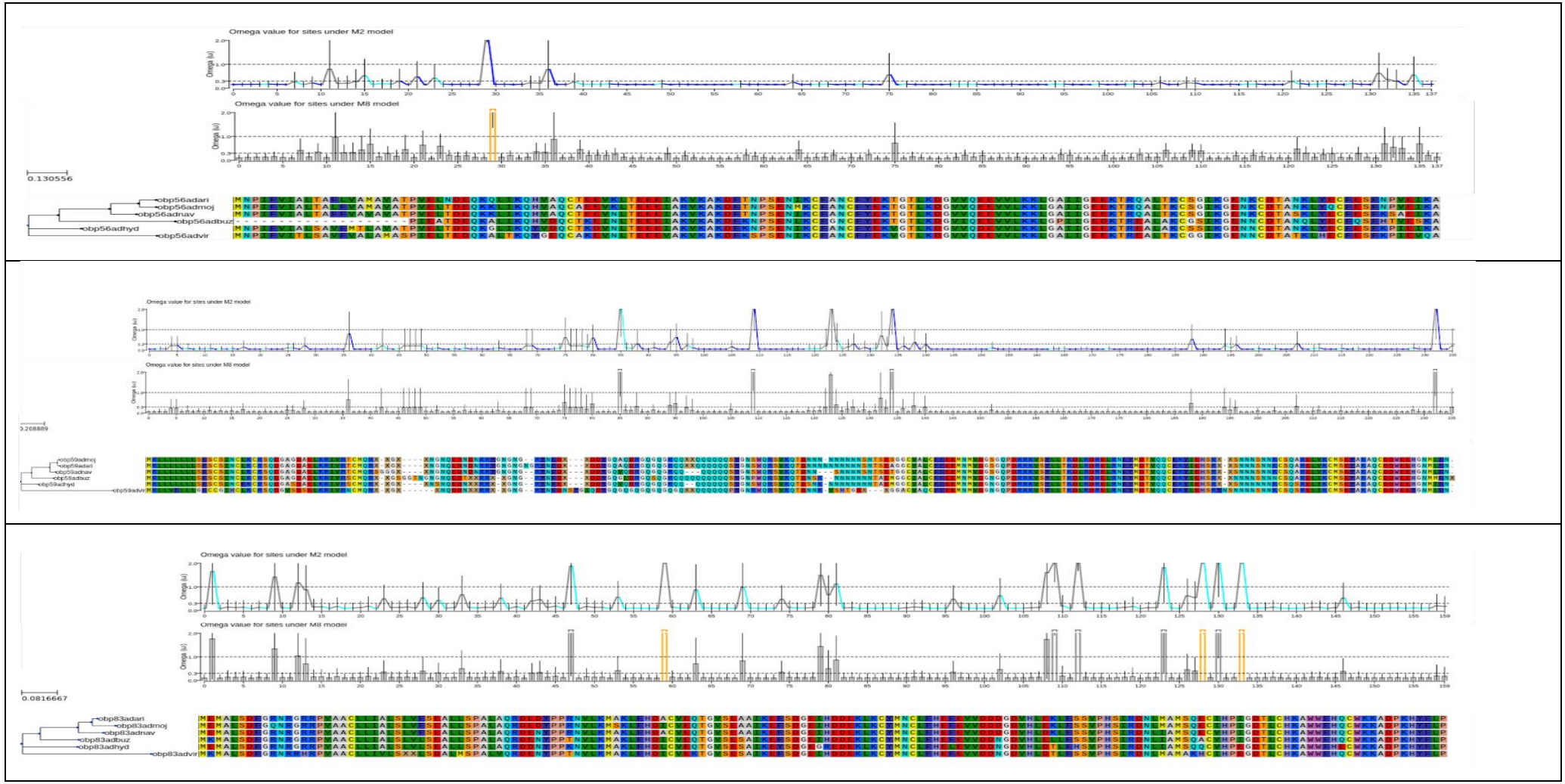
Primeiramente, as assinaturas de seleção positiva foram avaliadas usando o modelo de sítio implementado em codeML (YANG, 2004) comparando as sequências de 19 genes *Obp* e 39 genes *Or* comuns a seis espécies. Embora três *Obp* tenham ω> 1 (15,8%), apenas dois (*Obp*56a: sítio 30 Q e *Obp83a*: sítios 60 A, 129 E, 134 I) têm alterações de aminoácidos classificadas como significativas (*: p> 95%) pelo BEB análise recomendada pelos desenvolvedores de codeML (YANG, 2004) (Tabela 3).

Tabela 3. Assinaturas de seleção positiva de acordo com o teste de seleção por sítio para genes *Obp*. BEB: Bayes empírico Bayes (*p>95%, **: p>99%).

					Pr (w>1)	
	np	**lnL**	**x^2**	**p < 0.05**	**BEB (*p>95%, **: p> 99%)**	
***Obp*56a**						
M1	12	-1,247,656,231				
M2	14	-1,246,142,129	3.1	ns		
M7	12	-1,248,395,748				
M8	14	-1,245,329,068	6.12	*	**30 Q: 0.95 ***	
***Obp*59a**						
M1	12	-1,769,871,883				
M2	14	-1,768,441,329	2.86	ns		
M7	12	-1,772,591,012				
M8	14	-1,767,725,689	9.73	*	86 G: 0.84	110 P: 0.91
					135 M: 0.94	233 M: 0.85
***Obp*83a**						
M1	12	-1,171,072,316				
M2	14	-1,167,904,629	6.36	ns		
M7	12	-1,171,610,182				
M8	14	-1,167,844,584	7.52	*	**60 A: 0.99 ****	110 K: 0.85
					113 S: 0.94	**129 E: 0.96 ***
					131 I: 0.66	**134 I: 0.97 ***

A visualização dos resultados do codeML usando ETE é mostrada na Figura 5, para os três *Obp*s mostrando assinaturas de seleção positiva (*Obp*56a, *Obp*59a e *Obp*83a), incluindo a filogenia, o alinhamento da sequência de aminoácidos e o ω correspondente a cada sítio de aminoácido para modelos M2 (seleção com valor ω estimado livremente) e M8 (seleção positiva). É possível observar que existem vários sítios com ω> 1 nessas proteínas, porém apenas um sítio em *Obp*56a (29) e em *Obp*59a (134), e três sítios em *Obp*83a (59, 128, 133) são significativos de acordo com a análise BEB. Esses sítios com valores significativos são destacados em amarelo. Observe que esses locais são indicados como 30, 135, 60, 129 e 134, respectivamente para os três OBPs na Tabela 3 porque o primeiro aminoácido - metionina - é contado como "1" na saída de codeML, e como "0" o ETE. A visualização do codeML usando o ETE para as demais *Obp*s está em Material Suplementar (Figura Suplementar 2).

Figura 5. Visualização dos resultados M2-M8 da análise codeML para OBP56a, OBP59a e OBP83a fornecida pelo *toolkit* ETE. Mudanças significativas de aminoácidos são destacadas em amarelo.



Proporcionalmente ao maior número de genes *Or* que *Obp*, esta família de genes apresentou proporcionalmente um pouco menos genes com ω> 1 (12,8%), porém esses genes apresentaram maior número de sítios com ω > 1. Porém, em apenas dois destes genes (*Or*42b: sítio 180 I; *Or*42b2: sítio 55 A) as alterações de aminoácidos foram consideradas significativas pela análise BEB (Tabela 4).

Tabela 4. Assinaturas de seleção positiva de acordo com o teste de seleção por sítio para genes *Or*. BEB: Bayes empírico Bayes (*p>95%, **: p>99%). Locais significativos sob seleção positiva em negrito.

					Pr (w>1)
	np	**lnL**	**x²**	**p < 0,05**	**BEB (*p>95%, **: p> 99%)**
***Or*10a**					
M1	12	-4,124,611,687			
M2	14	-4,124,604,477	0.01	ns	
M7	12	-4,129,353,124			
M8	14	-4,124,555,729	9.59	*	5 F: 0.51 18 L: 0.62 66 A: 0.60 95 Y: 0.51 108 Q: 0.51 113 K: 0.61 114 K: 0.52 182 R: 0.74 241 K: 0.74 242 G: 0.64 244 A: 0.65 257 S: 0.82 307 S: 0.60
***Or*42a**					
M1	12	-3,841,548,818			
M2	14	-3,841,548,818	0	ns	
M7	12	-3,841,386,457			
M8	14	-3,838,235,319	6.3	*	125 R 0.72 139 L: 0.86 142 K: 0.87 146 C: 0.65 329 M: 0.65 344 E: 0.92 402 T: 0.77 406 S: 0.65 411 P: 0.62 413 E: 0.73
***Or*42b**					
M1	12	-3,302,882,813			
M2	14	-3,302,882,813	0	ns	
M7	12	-3,296,756,767			
M8	14	-3,294,192,456	5.12	*	4 E:0.77 40 L: 0.628 **180 I: 0.96*** 185 L: 0.61 253 R: 0.82 360 S: 0.59 399 T: 0.75
***Or*42b2**					
M1	12	-3,579,159,660			
M2	14	-3,579,159,660	0	ns	
M7	12	-3,580,899,852			
M8	14	-3,575,215,892	11.36	*	43 R: 0.58 **55 A: 0.96*** 129 E: 0.85 234 S: 0.93 242 L: 0.74 249 S: 0.75 337 V: 0.82 399 D: 0.54
***Or*65b**					
M1	12	-3,942,946,457			
M2	14	-3,941,929,286	0.03	ns	
M7	12	-3,945,102,510			

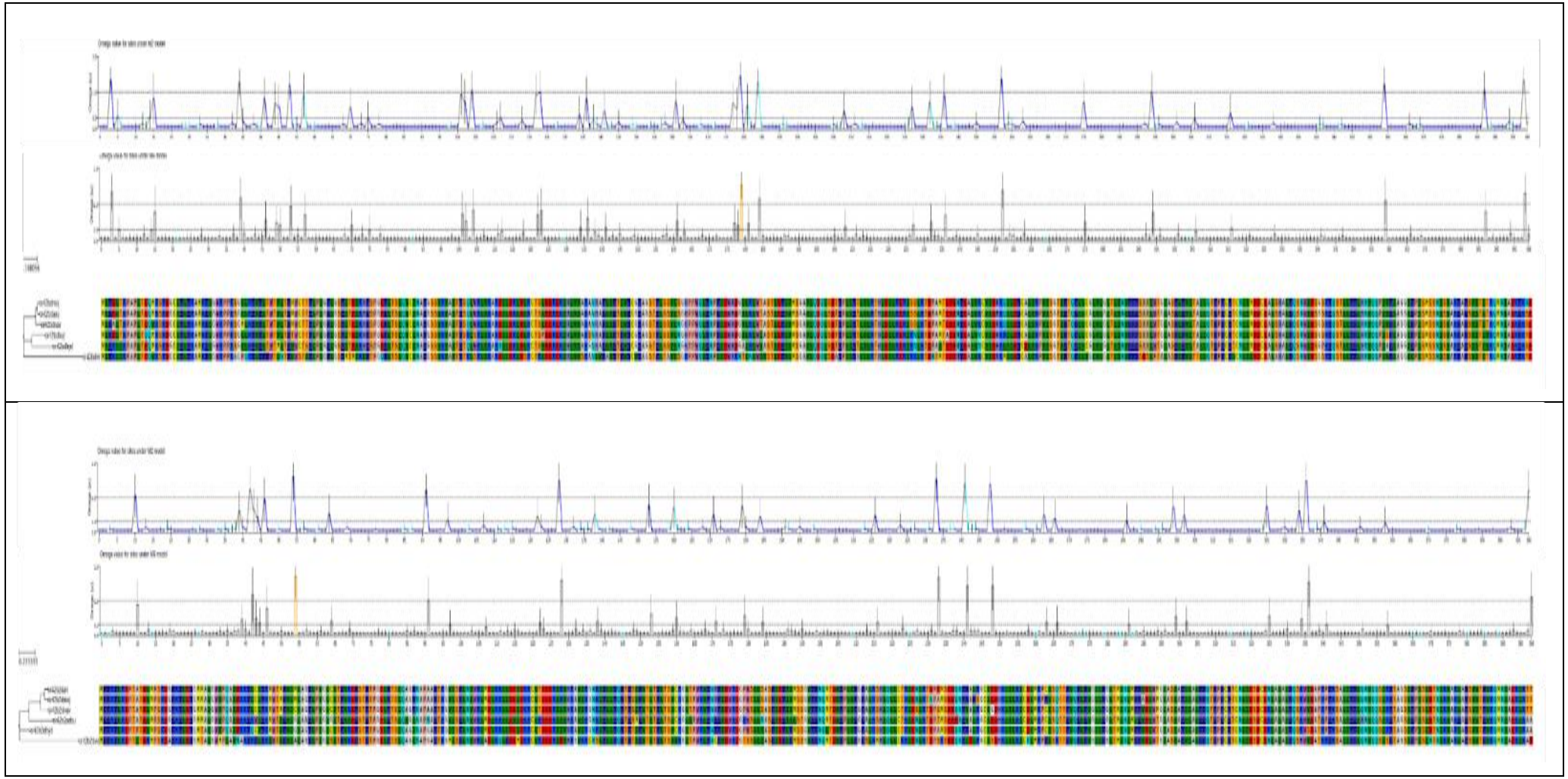
Tabela 4. Continuação.

M8	14	-3,941,994,340	6.21	*	66 I: 0.54	102 G: 0.78	107 S: 0.89
					135 V: 0.58	158 G: 0.59	166 L: 0.76
					173 I: 0.59	188 Y: 0.75	197 T: 0.52
					224 T: 0.71	255 R: 0.59	287 T: 0.59
					321 T: 0.89	324 A: 0.73	

A Figura 6 mostra a visualização dos resultados de codeML para os modelos M2 e M8 para as proteínas *Or*42b, *Or*42b2. Destas proteínas, apenas as substituições de aminoácidos em *Or*42b (sítio 180) e *Or*42b2 (sítio 55) são significativas sob o BEB. A visualização do codeML usando o ETE para as demais *Or*s está em Material Suplementar (Figura Suplementar 3).

.

Figura 6. Visualização dos resultados M2-M8 da análise codeML para OR42b e OR42b2 fornecidos pelo toolkit Environment for Tree Exploration – ETE. Mudanças significativas de aminoácidos são destacadas em amarelo.



Tem sido argumentado que as análises baseadas em modelos de seleção natural muitas vezes categorizam os sítios em um número relativamente pequeno de classes e, embora amplamente usado, argumentou-se que ao considerar cada sítio como pertencente a uma dessas classes, eles colocam restrições irrealistas na distribuição de parâmetros de seleção. Além disso, esses modelos presumem que é constante ao longo do tempo e, portanto, podem ser incapazes de identificar qualquer seleção positiva episódica.

Vários métodos que tentam superar esses problemas potenciais são implementados no pacote HyPhy (https://www.hyphy.org/). Um desses métodos é o FEL (*Fixed Effects Likelihood method*, KOSAKOVSKY POND; FROST, 2005), que usa uma abordagem de probabilidade máxima para inferir taxas de substituição não sinônima (Ka) e sinônima (Ks) por sítio para uma determinada codificação alinhamento e filogenia correspondente. Este método assume que a pressão de seleção para cada sítio é constante ao longo de toda a filogenia. MEME (*Mixed Effects Model of Evolution*, MURRELL et al., 2012) emprega uma abordagem de máxima verossimilhança de efeitos mistos para testar a hipótese de que locais individuais foram sujeitos a uma seleção episódica positiva ou diversificada. FUBAR (*Fast Unconstrained Bayesian Approximation*, MURRELL et al., 2013) usa uma abordagem bayesiana para inferir taxas de substituição não sinônima (Ka) e sinônima (Ks) por sítio para um determinado alinhamento e filogenia correspondente. Este método assume que a pressão de seleção para cada sítio é constante ao longo de toda a filogenia.

As Tabelas 5 e 6 mostram os resultados de todos os métodos utilizados para identificar assinaturas de seleção positiva nos genes *Obp* e *Or*. Ao total é possível observar que a maior concordância entre os resultados dos diferentes testes ocorre entre codeML e FUBAR, possivelmente por utilizarem a abordagem Bayseana para inferir Ka / Ks. Há tal concordância para 10 dos 19 genes *Obp* (62,5%) e 17 dos 31 *Or* (54,8%) com pelo menos um teste indicando seleção positiva. Com base nas comparações das sequências nucleotídicas feitas e nos resultados obtidos, consideramos que o uso dos dois testes em conjunto fornece uma inferência segura da ocorrência de seleção positiva em ramos de uma filogenia gênica.

Tabela 5. Testes de seleção baseado em códons para os genes *Obp.* Valores significativos estão em negrito.

Gene/Codon	**codeML** 95% pp	**FEL** p < 0.05	**MEME** p < 0.05	**FUBAR** pp > 90%	**PRIME *** P < 0.05
***Obp*19b**					
12	0.813	ns	ns	**0.938**	-
29	-	ns	ns	**0.907**	-
136	0.595	ns	ns	**0.922**	-
***Obp*28a**					
52	-	**ns**	**0.011**	ns	-
***Obp*44a**					
52	0.587	**ns**	**0.039**	**0.928**	-
***Obp*56a**					
30	**0.952**	ns	ns	**0.947**	-
***Obp*56b**					
15	-	ns	**0.014**	-	
79	0.669	ns	**ns**	**0.931**	-
***Obp*56d**					
10	-	ns	**0.048**	-	-
***Obp*56g**					
84	-	ns	**ns**	-	**0.020**/ Fator V
95	-	ns	**ns**	-	**0.022**/ Fator I
112	-	ns	**0.037**	-	-
135	-	ns	**ns**		**0.005**/ Fator V
***Obp*56h**					
97	0.879	ns	**ns**	**0.944**	**0.022**/ Fator I

Tabela 5. Continuação.

***Obp*59a**					
135	**0.936**	ns	ns	**0.954**	-
***Obp*69a**					
128	**-**	ns	**ns**	**0.947**	-
141	0.591	**-**	**ns**	**0.925**	-
***Obp*83a**					
60	**0.987**	ns	ns	-	-
113	0.938	ns	ns	**0.931**	-
129	**0.962**	ns	ns	-	-
134	**0.973**	ns	ns	-	0.008/ Fator IV
***Obp*83b**					
11	0.694	ns	ns	**0.927**	-
***Obp*99a**					
141	-	ns	**0.043**	-	-
***Obp*99b**					
20	-	ns	**0.032**	-	-
***Obp*99c**					
4	0.717	ns	**0.038**	**0.902**	-
***Obp*A10**					
15	0.682	ns	**ns**	**0.918**	-
17	-	ns	**0.013**	**-**	-

*: Propriedades dos aminoácidos: Fator I: índice de polaridade; Fator II: estrutura secundária, Fator III: volume molecular, Fator IV: composição do aminoácido, Fator V: carga eletrostática; p: p-value; pp: probabilidade posterior

Atchley et al. (2005) usaram uma análise estatística multivariada em quase 500 atributos de aminoácidos para produzir padrões numéricos interpretáveis de variabilidade de aminoácidos, que foram resumidos por cinco padrões multidimensionais de covariação de atributo que refletem polaridade (Fator I), estrutura secundária (Fator II), molecular volume (Fator III), composição de aminoácidos e diversidade de códons (Fator IV) e carga eletrostática (Fator V). Análises experimentais e analíticas mostraram que existe uma forte base evolutiva para um padrão complexo de covariação envolvendo a extensão da acessibilidade de aminoácidos, polaridade, hidrofobicidade e atributos relacionados (Fator I). Esses padrões observados estavam relacionados à seleção natural, mudança evolutiva e divergência filogenética. O fator IV, que reflete a diversidade de códons e aminoácidos, exibe uma correlação mais fraca, mas ainda altamente significativa, entre os atributos físico-químicos e a mudança evolutiva nos padrões de substituição. De acordo com Atchley et al. (2005), a variação na propensão para formar várias configurações estruturais secundárias (Fator II), tamanho molecular (Fator III) e carga (Fator V) não pode ser atribuída à divergência evolutiva, mas sim a mudanças não evolutivas na estrutura e função. Entre os 10 OBPs com sítios sob seleção positiva, apenas dois (Obp56h e Obp83a) tinham um sítio, cada um com substituições de aminoácidos consideradas evolutivamente relevantes de acordo com esses critérios.

A Tabela 6 mostra os resultados para os 17 genes *Or* com sítios com assinaturas de seleção positiva em concordância entre os testes codeML e FUBAR, bem como os resultados dos outros cinco testes realizados.

Tabela 6. Testes de seleção baseado em códons para os genes *Or*. Valores significativos estão em negrito.

Gene/Codon	**codeML** 95% pp	**FEL** p <0.05	**MEME** p <0.05	**FUBAR** pp > 90%	**PRIME *** p <0.05 / Fator
***Or*1a**					
233	0.631	ns	ns	-	-
375	-		**0.022**	-	-
374	0.715	ns	**0.001**	**0.985**	-
***Or*9a**					
52	0.814	ns	ns	**0.916**	-
212	-		**0.021**	-	-
***Or*10a**					

Tabela 6. Continuação.

5	0.512	ns	**0.083**	**0.926**	-
66	0.600	**0.040**	**0.058**	**0.927**	-
87	-	-	**0.004**	-	-
95	0.507	ns	**0.046**	-	-
108	0.513	-	ns	-	**0.024**/ Fator V
113	0.613	-	-	**0.936**	-
117	-	-	**0.036**	-	-
182	0.739	ns	-	-	-
241	0.736	-	**0.038**	**0.906**	**0.000**/ Fator II **0.046**/ Fator III **0.049**/ Geral
***Or*21a**					
357	-	-	**0.016**	-	-
***Or*33c**				-	
19	-	-	**0.025**	-	-
26	-	-	**0.035**	-	-
131	0.621	ns	-	-	-
177	-	-	**0.011**	-	0.077/ Fator II **0.013**/ Fator IV
189	-	-	**0.011**	-	**0.042**/ Fator III 0.072/ Fator V
226	0.657	ns	-	-	-
358	0.626	**0.023**	**0.034**	**0.918**	**0.004**/ Fator III
***Or*35a**					
57	-	-	**0.014**	-	-
***Or*42a**					
39	-	-	**0.037**	-	-
125	0.717	ns	ns	**0.944**	-

Tabela 6. Continuação.

139	0.863	-	-	**0.925**	-
304	-		**0.034**	-	-
406	0.648	-	-	**0.920**	-
***Or*42a-like**					
52	0.681	ns	ns	**0.957**	-
148	0.648	ns	ns	**0.969**	-
***Or*42b**					
40	0.628	ns	-	**0.915**	-
180	**0.957**	**0.034**	**0.323**	**0.979**	-
388	-	-	**0.032**	-	-
***Or*42b2**					
55	**0.964**	ns	**0.003**	**0.943**	0.058/ Fator I 0.065/ Fator II
***Or*42b3**					
381	-	-	**0.033**	-	-
384	-	-	**0.030**	-	-
397	-	-	**0.034**	-	-
399	-	-	**0.028**	-	-
***Or*42b-like**		-			
2	-	-	**0.022**	-	-
14	-	-	**0.039**	-	-
111	-	ns	**0.020**	-	-
183	0.605	ns	ns	**0.917**	-
323	-	-	**0.045**	-	-
***Or*49b**					
35	-	-	**0.029**	-	-

Tabela 6. Continuação.

***Or*59a**					
5	-	-	**0.013**	-	0.053/ Fator I
63	-	-	**0.038**	-	-
***Or*59b**					
43	-	-	**0.041**	-	-
72	-	-	**0.042**	**0.901**	-
117	-	-	**0.050**	-	0.050/ Fator V
124	0.501	-	**0.011**	-	**0.015**/ Fator V
179	-	--	**0.005**	-	-
313	-		**0.046**	-	0.006/ Fator V
***Or*63a**					
94	-	-	**0.047**	-	-
***Or*63a2**					
485	0.640	-	**0.028**	-	-
***Or*65b**					
102	0.782	ns	**0.003**	**0.963**	**0.014**/ Fator II
107	0.878	ns	ns	**0.953**	0.099/ Fator II **0.029**/ Fator V
158	0.592	**0.047**	**0.200**	-	-
166	0.755	**0.046**	**ns**	**0.968**	**0.001**/ Fator II **0.001**/ Fator IV
173	0.586	ns	**ns**	**0.913**	-
188	0.754	ns	-	**0.961**	-
215	-	-	**0.044**	-	-
287	0.589	ns	**0.026**	**0.948**	-
405	-	-	**0.040**	-	-

Tabela 6. Continuação.

***Or*67b**					
4	-	-	**0.043**	-	-
314	-	-	**0.033**	-	-
***Or*67d**		-			
22	-	ns	ns	**0.937**	-
58	-	-	**0.005**	-	-
232	0.758	ns	**0.032**	**0.921**	-
241	-	-	**0.038**	-	-
265	0.806	ns	**0.031**	**0.962**	-
295	-	ns	**0.070**	-	-
352	-	-	**0.014**	-	
***Or*69a**					
31	-	-	**0.033**	-	0.063/ Fator III
240	-	-	**0.021**	-	-
384	-	-	**0.031**	-	-
***Or*71a**		-			
38	0.536	ns	ns	**0.919**	-
199	0.633	**0.044**	ns	**0.942**	-
386	-	-	**0.004**	-	-
388	-	-	**0.026**	-	-
***Or*74a**				-	
151	-	-	**0.009**	-	-
***Or*83a**				-	
172	-	-	**0.037**	-	-
***Or*85b**				-	
106	-	-	**0.004**	-	**0.023**/Fator III

Tabela 6. Continuação.

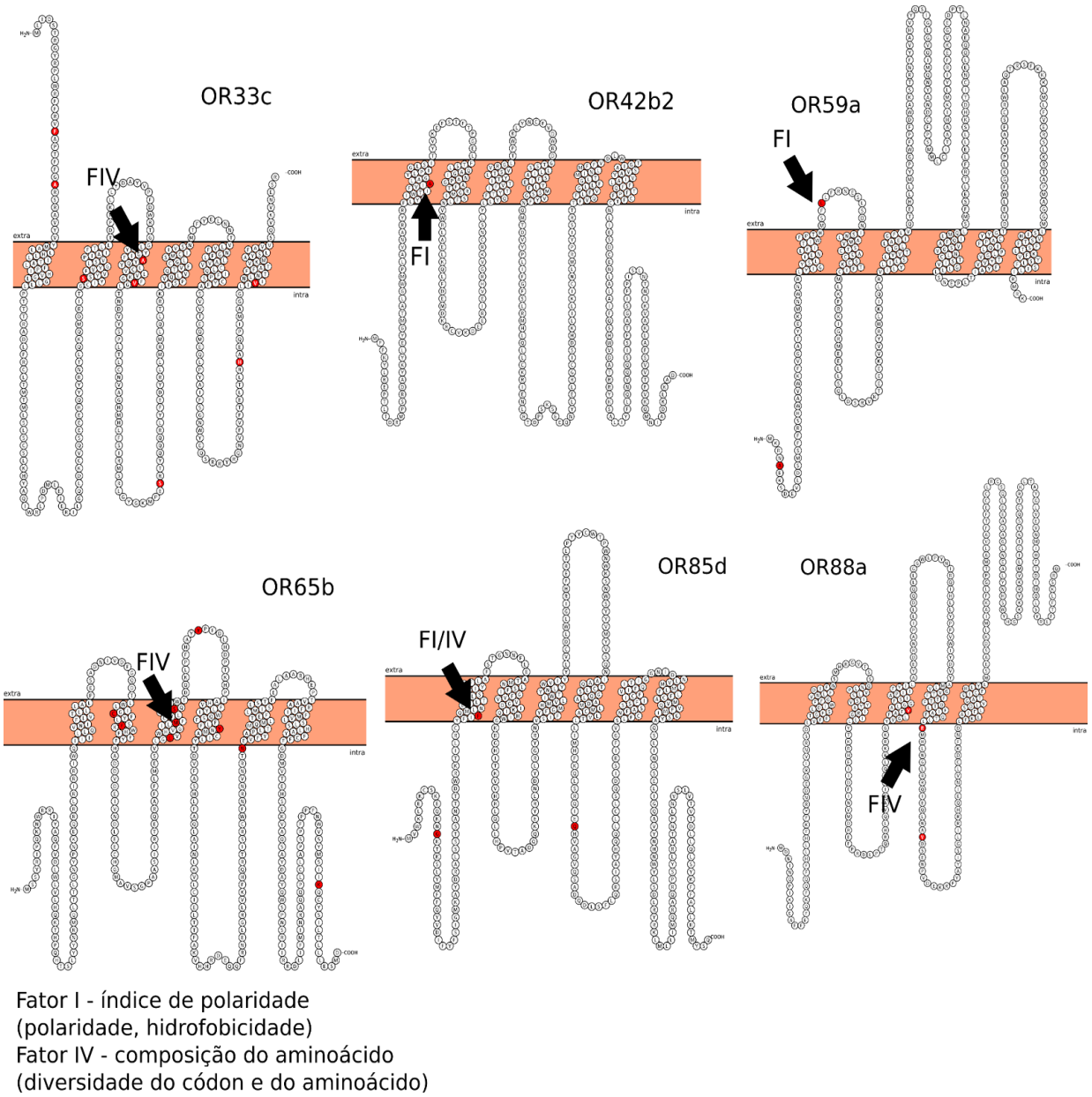
145	-	-	**0.033**	-	0.089 / Geral
181	0.828	**0.041**	**0.015**	**0.982**	-
***Or*85d**					
15	0.506	**0.033**	**0.048**	**0.909**	-
64	0.540	-	**0.001**	**0.932**	0.085/ Fator I 0.069/ Fator IV 0.088/ Geral
253	0.635	**0.035**	ns	**0.946**	-
***Or*85e**					
171	0.687	ns	-	**0.957**	-
277	0.572	-	**0.030**	-	-
***Or*88a**					
144	-	-	**0.042**	-	-
223	-	-	**0.012**	-	0.061/ Fator IV
239	0.505	**0.049**	ns	**0.933**	-
***Or*92a**					
11	-	-	**0.006**	-	-
267	-	-	**0.045**	-	-
***Or*94b**					
34	0.943	ns	-	**0.961**	-
81	-	-	**0.030**	-	**0.009**/Fator III
115	-	-	**0.025**	-	-
Orcr					
58	0.709	ns	**0.001**	-	**0.044**/Geral
111	-	-	**0.046**	-	-

*: Propriedades dos aminoácidos: Fator I: índice de polaridade; Fator II: estrutura secundária, Fator III: volume molecular, Fator IV: composição do aminoácido, Fator V: carga eletrostática; p: p-value; pp: probabilidade posterior

Apenas três genes *Or* com sítios sob seleção positiva (*Or*42b2, *Or*65b e *Or*85d) apresentaram propriedades bioquímicas com relevância evolutiva investigadas pelo PRIME, as quais estão relacionadas à variação nos atributos físico-químicos e diversidade de aminoácidos. Considerando as demais propriedades que também alteram os aminoácidos, alterando a estrutura secundária, o volume ou a carga elétrica da proteína, não consideradas evolutivamente relevantes por Atchley et al. (2005), mas que podem mudar seu funcionamento e alterar negativa ou positivamente sua função, sendo alvos, portanto, da seleção natural, esses números dobram para cinco proteínas incluindo *Or*10a e *Or*33c.

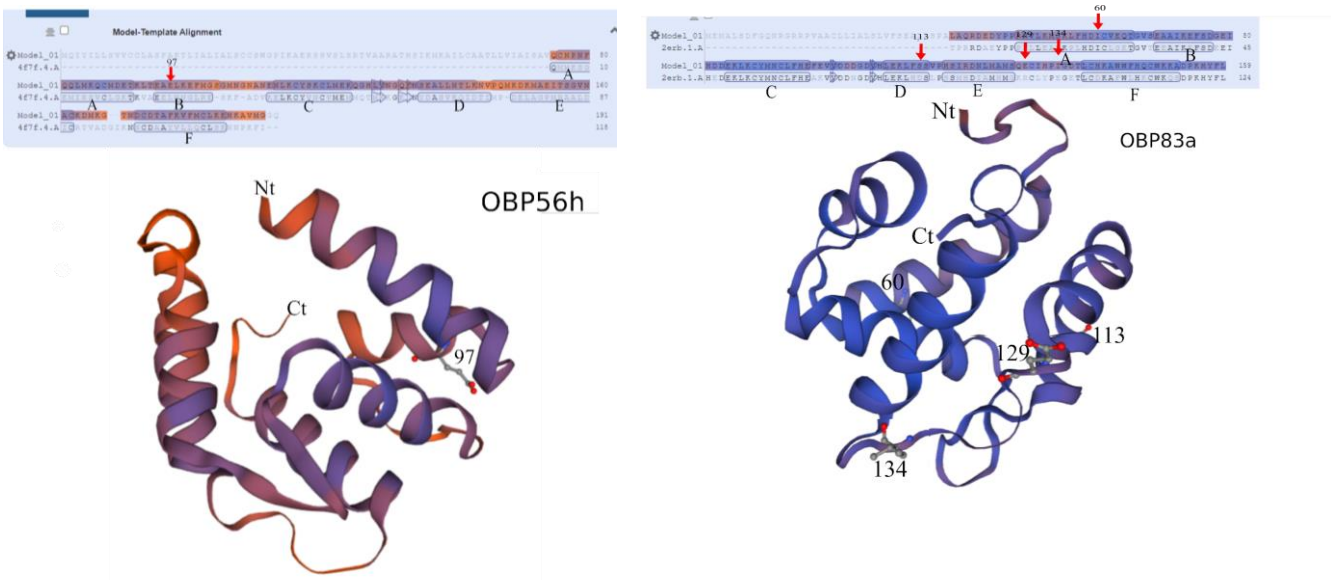
O software TMHMM foi utilizado para predizer as hélices transmembrana das proteínas OR, levando-se em consideração as sequências de aminoácidos de cada gene que apresentou sítios com significância em pelo menos um dos quatro testes de seleção realizados. Os resultados desta etapa foram transferidos para o software PROTTER, onde foi possível visualizar a topologia da proteína. Os locais sob seleção positiva foram identificados manualmente no PROTTER. Um total de 33 sítios sob seleção positiva, de acordo com nosso critério (resultados codeML e FUBAR) foram encontrados entre os 17 genes *Or*. Aproximadamente 45% dos locais sob seleção positiva estão localizados na região citoplasmática do domínio, enquanto apenas 35% e 20% estão localizados na região transmembrana e extracelular do domínio, respectivamente. Apenas três genes *Or* com sítios sob seleção positiva (*Or*42b2, *Or*65b e *Or*85d) apresentaram propriedades bioquímicas com relevância evolutiva investigadas pelo PRIME, que estão relacionadas à variação nos atributos físico-químicos e diversidade de aminoácidos, todos os 4 sítios estão localizados na região transmembrana (Figura 7). A topologia das proteínas OR restantes estão reunidas no Material Suplementar (Figura Suplementar 4)

Figura 7. Topologia prevista de proteínas OR mostrando os domínios transmembrana e uma cauda NH2 citoplasmática. Os sites selecionados positivamente são destacados em vermelho. A topologia foi prevista por TMHMM.



Para predizer a estrutura 3D das proteínas OBPs, e destacar os sítios sob seleção positiva identificados pelas nossas análises e sua localização na estrutura da proteína: α-hélices, *loop* e extremidades (N-terminal e C-terminal) foi utlizado o software SWISS-MODEL (GUEX, 1997). Na Figura 8, estão representadas as OBP56h e OBP83a, ambas com sítios sob seleção que apresentaram mudanças nas suas propriedades bioquímicas que são evolutivamente significativas.

Figura 8. Determinação da estrutura 3D das OBPs. Estão representadas as OBP56h e OBP83a, com os sítios identificados tanto na estrutura proteica quanto no alinhamento, além das caudas amino terminal e carboxil terminal. No alinhamento das sequências, os sítios estão identificados por uma seta e cada letra (A-F) corresponde a uma α-hélice. A estrutura foi prevista pelo SWISS-MODEL.



A determinação da estrutura 3D das proteínas globulares foi necessária para a inferir possível alteração funcional das substituições de aminoácidos relevantes. As sequências com maior similaridade disponíveis no PDB, identificadas no PSI BLAST, correspondem às sequências pertencentes aos genomas de *Phormia regina*, *Anopheles gambiae*, *Aedes aegypti* e *D. melanogaster*. A estrutura resultante é semelhante ao modelo típico (LASKOWSKI et al., 1993), com seis hélices típicas da família OBP em posições equivalentes e dobras semelhantes (Figura 7). A identificação e posterior nomenclatura (A-F) das α-hélices foram baseadas nos resultados de SANCHEZ-GRACIA (2008). A Figura 7 mostra a localização dos sítios sob seleção positiva nas α-hélices. O restante das proteínas OBPs estão reunidas no Material Suplementar (Figura Suplementar 5).

5 DISCUSSÃO

Ao investigar o papel da seleção positiva nos genes de localização do hospedeiro entre espécies generalistas e que mudaram de hospedeiro, encontramos sítios sob seleção em 31 dos 39 genes *Or*s (79%) e 16 dos 19 genes *Obp*s (84%), em pelo menos um dos quatro testes de seleção por sítio realizados. Conforme observado nas espécies aqui investigadas, foram mostradas diferenças significativas de restrição seletiva tanto entre grupos de genes ortólogos quanto entre linhagens, bem como alguns casos de evolução episódica e seleção positiva **(**SANCHEZ-GRACIA et al., 2008).

Essas diferenças podem estar relacionadas a substituições específicas nas proteínas que esses genes codificam em grupos de espécies que permitem um melhor reconhecimento de seus cactos hospedeiros.

Em um estudo de todo o genoma com foco em cinco grupos de réplicas de espécies com tolerâncias térmicas muito diferentes, três espécies cactofílicas altamente tolerantes ao estresse (*D. mojavensis, D. buzzatii* e *D. aldrichi*) e duas espécies dietéticas generalistas menos tolerantes (*D. hydei* e *D. repleta*), investigou-se a seleção positiva por meio de testes de análise seletiva por sítios por ramos (RANE et al., 2019). Em um conjunto de ortólogos 1: 1 variando de 10.745 a 11.669 genes entre as cinco espécies os autores mostraram que o ramo terminal contendo *D. aldrichi* contém um número relativamente alto de genes selecionados positivamente (> 200), e o ramo contendo espécies do complexo de *mulleri* (aqui estudadas) continha relativamente poucos (<200). As assinaturas de seleção positiva pareceram não estar associadas a diferenças relacionadas a tolerâncias térmicas e preferências alimentares, mas a processos regulatórios e de desenvolvimento. Esses resultados são concordantes com os apresentados neste trabalho, no qual as assinaturas de seleção positiva não parecem estar diretamente associadas à preferência pelo hospedeiro.

As OBP são proteínas α-helicoidais globulares, afetadas por mudanças (como o pK, por exemplo), que podem promover divergência funcional, seja pela alteração conformacional, seja pela modificação na especificidade da ligação. Caso seja benéfico para a espécie, a seleção atuará perpetuando essas mudanças (SANCHEZ-GRACIA et al., 2008). Alguns estudos demonstraram (ANDRONOPOULOU, 2006; SANCHEZ-GRACIA, 2003) a alta capacidade de *Obp*s formarem homo ou heterodímeros específicos, produzidos a partir de eventos de duplicação que podem resultar em proteínas com múltiplos domínios de cadeia única que retém características da unidade original, sugerindo um alto potencial complexo combinatório para novas ligações, por exemplo. Portanto, pequenas diferenças nas interações com *Obp* podem resultar em apreciável significado funcional. (VIEIRA et al., 2007). A diversificação funcional de cópias duplicadas é impulsionada pela seleção positiva, e seleção natural pode promover heterodímeros por meio do aumento do potencial combinatório de OBPs (aumentando o espectro de possíveis odorantes alvo ou a especificidade de ligação) (SANCHEZ-GRACIA et al., 2008). Foram identificados em nossas análises cerca de 9 sítios sob seleção positiva localizados na região amino terminal, nos genes *Obp19*b, *Obp*56b, *Obp*56d, *Obp*83b, *Obp*99b, *Obp*99c e *Obp*A10; e cinco sítios localizados na α-hélice A e B, nos genes *Obp*28a, *Obp*56a, *Obp*56h, *Obp*59a e *Obp*83a. Essa posição tem importância na conformação da proteína OBP, pois mudanças em sua sequência podem alterar o tamanho e na forma da cavidade de ligação, pela modificação da posição da primeira ponte dissulfeto (cisteínas normalmente conservadas adequadas para a conformação terciária funcional - SMITH, 2001). Foram encontrados ainda cinco sítios

localizados na região α-hélice F próxima a região C-terminal, nos genes *Obp*19b, *Obp*56g, *Obp*69a e *Obp*83a, e dois sítios localizados na extremidade C-terminal, nos genes *Obp*56g e *Obp*99a. Mudanças conformacionais neste ponto podem desencadear a liberação do ligante perto do receptor odorante, pois essa extremidade se dobra para dentro da proteína formando a parede da cavidade de ligação, onde irá conter os resíduos responsáveis pela interação com a molécula odorante. Substituições nesta parte da proteína, induzidas por ferormônios, já foram relacionadas com a mudança conformacional e consequente disparo de neurônios sensíveis a ferormônios em LUSH em *D. melanogaster* (SMITH, 2001). Outras regiões α-hélice D e E, nas quais foram cerca de quatro sítios sob seleção positiva, nos genes *Obp*56g, *Obp*69a e *Obp*83a,, podem estar envolvidas nas interações proteína-proteína, pela presença de aminoácidos expostos e resíduos hidrofóbicos que cobrem a cavidade de ligação. Sanchez-Gracia (2008) detectou sítios sob seleção positiva em OS-E e OS-F (Obp83a e Obp83b nomeado por HEKMAT-SCAFE et al. 2002, presentes neste trabalho), localizados em posições importantes na conformação da proteína OBP. A região het1, encontrada nas α-hélices D e E com seu *loop* de conexão apresentaram poucos sítios sob seleção positiva, quando comparado com os sítios sob seleção localizados na região variável het2, na porção N-terminal, sugerindo que a região het1 é mais conservada e, portanto, pode ter um importante papel funcional, o que concorda com os resultados de Sánchez-Gracia (2008).

O mapeamento do domínio dos receptores odorantes (*Or*s) indicou que a maioria dos sítios sob seleção positiva estão localizados nas regiões transmembrana e citoplasmática, um padrão inesperado, já que a região citoplasmática está envolvida na interação com mensageiros secundários, início da transdução do sinal e, portanto, esperava-se que estaria sendo conservada. Com exceção de um gene *Or*42b3, o restante não apresentou sítios sob seleção positiva localizados na região terminal da cauda COOH, sugerindo que a região está sendo conservada, concordando com os resultados de GARDINER et al. (2009), já que essa extremidade tem importância no acoplamento com o co-receptor *Or*83b e, portanto, é importante na detecção do sinal. Nossas análises sugerem que a seleção divergente moldou e está atuando tanto na detecção quanto na transdução do sinal. Uma porção considerável dos sítios sob seleção positiva nos genes *Or* mapeia para o domínio transmembrana dessas proteínas, o que está de acordo com um estudo envolvendo 10 *Or* nos 12 genomas de *Drosophila* (GARDINER et al., 2009), no qual apenas 15% dos locais estavam localizados nos terminais NH2 citoplasmáticos. Os nossos achados estão em acordo com DIAZ et al. (2018) que estudando três genes *Or* (*Or*67c, *Or*83c1, *Or*83c2) comuns a *D. mojavensis* e *D. arizonae* encontraram 55% dos sítios sob seleção positiva localizados no domínio citoplasmático, enquanto apenas 24 e 21% nos domínios extracelular e transmembrana, respectivamente. Os autores notaram que este resultado pareceria inesperado, pois este domínio interage com mensageiros secundários envolvidos na transdução de sinal e, portanto, esperava-se que fosse conservado, como já comentado.

Date (2013), por meio de análises de RNAseq e níveis de expressão gênica quimiossensorial entre populações de *D. mojavensis* do deserto de Mojave e da ilha de Santa Catalina (a mesma utilizada neste estudo), demonstrou que no geral 21 genes *Or*, 13 *Obp* e 5 *Ir* (receptor ionotrópico) estavam diferencialmente expressos entre as populações, com diferenças específicas entre os sexos (13 *Or* genes, 5 *Obp* e 2 *Ir*). A população de Mojave utiliza cactos colunares e a de Santa Catalina utiliza *Opuntia*, as diferenças na expressão desses genes quimiossensoriais entre as populações podem evidenciar e contrastar com as diferenças químicas presentes nas plantas hospedeiras as quais essas populações estão adaptadas. De acordo com nossas análises, alguns dos genes envolvidos no reconhecimento do hospedeiro e que estão diferencialmente expressos em Date (2013), também apresentam sítios que estão passando por seleção positiva, sugerindo que a divergência de expressão esteja relacionada com a adaptação nesses hospedeiros.

Apesar de não ter sido possível associar neste trabalho as mudanças de aminoácidos com os hábitos alimentares das espécies, que corroborariam tal hipótese, 10 genes *Or* e 5 *Obp* presentes neste estudo e em DATE (2013), e que apresentaram sítios sob seleção positiva, estão diferencialmente expressos entre populações que utilizam hospedeiros distintos e entre sexos, sugerindo que as mudanças de hospedeiros podem ter sido favorecidas pelas mudanças na expressão de certos genes quimiossensoriais.

Em conclusão, entre as 10 OBPs e as 17 ORs que apresentaram assinaturas de seleção positiva, dois (20%) e cinco (29%), respectivamente, mostraram evidências de divergência funcional em alguns dos aminoácidos substituídos, e dois nas OBPs e três nas ORs, são considerados evolutivamente relevantes. Um estudo de receptores ORs e GRs (gustativos) nos genomas de 12 espécies de *Drosophila*, GARDINER et al. (2009) encontraram 20 genes com evidência de divergência funcional, mas apenas seis foram considerados sob seleção positiva e para o resto dos genes, a divergência rápida foi provavelmente devido ao relaxamento da pressão seletiva. Esses resultados são semelhantes aos observados neste estudo e parecem ser um padrão evolutivo para esses quimiorreceptores.

6 CONCLUSÕES

Aqui fornecemos evidências que apoiam a hipótese de seleção positiva moldando a evolução de sítios de aminoácidos envolvidos em funções moleculares de genes quimiorreceptores em espécies do grupo *repleta* de *Drosophila*. Embora as análises realizadas não tenham permitido correlacionar as mudanças de aminoácidos com os hábitos alimentares foi possível mostrar diferenças tanto entre grupos de genes ortólogos quanto entre linhagens, bem como vários casos de evolução episódica e seleção positiva associados à evolução de famílias de quimiorreceptores OBPs e ORs, responsáveis

pela localização de hospedeiros em espécies cactofílicas, diferenças essas que podem estar associadas a um melhor reconhecimento de seus cactos hospedeiros nas espécies onde foram fixadas.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALTENHOFF, A. M.; DESSIMOZ, C. Inferring orthology and paralogy. **Methods in molecular biology**, 855: 259–279, 2012.

ANDRONOPOULOU, E.; LABROPOULOU, V.; DOURIS, V.; WOODS, D. F.; BIESSMANN, H.; IATROU, K. Specific interactions among odorant-binding proteins of the African malaria vector *Anopheles gambiae*. **Insect Mol Biol**, 15:797-811, 2006.

ARYA, G. H.; WEBER, A. L.; WANG, P.; MAGWIRE, M. M.; NEGRON, Y. L. et al. Natural variation, functional pleiotropy and transcriptional contexts of odorant binding protein genes in *Drosophila melanogaster*. **Genetics,** 186: 1475-1485, 2010.

ATCHLEY, W. R.; ZHAO, J.; FERNANDES, A. D.; DRUKE, T. Solving the protein sequence metric problem. **Proc Natl Acad Sci**, 102:6395–400, 2005.

BENTON, R.; DAHANUKAR, A. Electrophysiological recording from *Drosophila* taste sensilla. **Cold Spring Harbor protocols**, *7*:839-50, 2011.

BOUGHMAN, J. W. How sensory drive can promote speciation. Trends in Ecol & Evol, 17: 571-577, 2002.

BRAZNER, J. et al. Host-plant shifts and adult survival in the cactus breeding *Drosophila mojavensis*, Ecol. Entomol, 9:375-381, 1984.

CESPEDES, C. L.; SALAZAR, J. R.; MARTINEZ, M.; ARANDA, E. Insect growth regulatory effects of some extracts and sterols from *Myrtillocactus geometrizans* (*Cactaceae*) against *Spodoptera frugiperda* and Tenebrio molitor. **Phytochemistry**, 66, 2481–2493, 2005

CORIO, C.; SOTO, I. M.; CARREIRA, V. P.; PADRÓ, J.; BETTI, M. I.; HASSON, E. An alkaloid fraction extracted from the cactus *Trichocereus terscheckii* affects fitness in the cactophilic fly *Drosophila buzzatii* (*Diptera: Drosophilidae*). **Biol J Linn Soc,** 109:342–353, 2013.

DATE, P. **Evolution of host specialization in the cactophilic fly, *Drosophila mojavensis***. Doctorship of Phylosophy in Biological Science. University of Cincinnati, p. 185, 2013.

DEKKER, T.; IBBA, I.; SIJU, K. P.; STENSMYR, M. C.; HANSSON, B. S. Olfactory shifts parallel superspecialism for toxic fruit in *Drosophila melanogaster* sibling, *D. sechellia*. **Curr Biol,** 16: 101-109, 2006.

DIAZ, F.; ALLAN, C. W.; MATZKIN, L. M. Positive selection at sites of chemosensory genes is associated with the recent divergence and local ecological adaptation in cactophilic *Drosophila*. **BMC Evol Biol,** 18, 144, 2019.

EDWARDS, E.J.; NYFFELER, R.; DONOGHUE, M. J. Basal cactus phylogeny: implications of Pereskia (*Cactaceae*) paraphyly for the transition to the cactus life form. **Am. J. Bot.,** 92, 1177–1188, 2005.

EMMS, David M.; KELLY, Steven. OrthoFinder: Phylogenetic orthology inference for comparative genomics. **Genome Biology**, 20(238): 1–14, 2019.

ETGES, J. Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. **Am Nat,** 152 (4): 129-144, 1998

FEDER, J. L. The apple maggot fly, *Rhagoletis pomonella*: Flies in the face of conventional wisdom about speciation. **Species and Speciation Oxford Univ,** 130-144, 1998.

FELLOWS, D.P., HEED, W.B. Factors affecting host plant selection in desertadapted cactophilic *Drosophila*. Ecology, 53: 850–858, 1972.

FOGLEMAN, J. C. The role of volatiles in the ecology of cactophilic *Drosophila*, in: Ecological Genetics and Evolution: The Cactus-Yeast*Drosophila* Model System (J. S. F. Barker, and W. T. Starmer, eds), **Academic Press Australia**, Sydney, pp. 191-206, 1982.

FUTUYMA, D. J., MORENO, G. The evolution of ecological specialization. **Annu Rev Ecol Syst**, 19: 207-234, 1988.

GARDINER, A.; BUTLIN, R. K.; JORDAN, W. C.; RITCHIE, M. G. Sites of evolutionary divergence differ between olfactory and gustatory receptors of *Drosophila*. **Biol Lett**. 5:244–7, 2009.

GEER, L. Y. et al. CDART: Protein homology by domain architecture. **Genome Research**, 12, (10): 1619–1623, 2002. DOI: 10.1101/gr.278202.

GRIFFITH, M. P.; PORTER, J.M. Phylogeny of *Opuntioideae* (*Cactaceae*). **Int. J. Plant Sci.** 170, 107–116.2009

GUEX, N.; PEITSCH, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. **Electrophoresis**, 18(15):2714-2723, 1997.

GUO, S.; KIM, J. Molecular evolution of *Drosophila* odorant receptor genes. **Mol Biol Evol.** 24:1198-1207, 2007.

HASSON, E. et al. Host plant adaptation in cactophilic species of the *Drosophila buzzatii* cluster: fitness and transcriptomics. **J Hered**, 110: 46–57, 2019.

HEKMAT-SCAFE, D. S.; SCAFE, C.R.; McKINNEY, A. J.; TANOUYE, M. A. Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. **Genome Res,** 12: 1357–1369, 2002.

HUERTA-CEPAS, J.; DOPAZO, J.; GABALDÓN, T. ETE: a python Environment for Tree Exploration. **BMC Bioinformatics.** 11:24. 2010.

JAENIKE, J. Host selection by mycophagous Drosophila. Ecology, 59: 1286-1288, 1978.

JENNINGS, J. H.; ETGES, W. J. Species hybrids in the laboratory but not in nature: a reanalysis of premating isolation between *Drosophila arizonae* and *D. mojavensis*. Evolution, 64(2): 587–98, 2010.

JOHNSTUN, J. A.; SHANKAR, V.; MOKASHI, S. S., et al. Functional Diversification, Redundancy, and Epistasis among Paralogs of the *Drosophila melanogaster* Obp50a-d Gene Cluster. **Mol Biol Evol.** 38(5):2030-2044, 2021.

KIM, M. S.; REPP, A.; SMITH, D.P. LUSH odorant-binding protein mediates chemosensory responses to alcohols in *Drosophila melanogaster*. **Genetics**, 150: 711–721, 1998.

KIMURA, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. **Journal of molecular evolution**, v. 16, n. 2, p. 111-120, 1980.

KIRCHER, H. W. Triterpene glycosides and queretaroic acid in organ pipe cactus, Phytochemistry, 16:1078-1080, 1977.

KOSAKOVSKY, P. S. L.; Frost, S. D. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 22:1208-22.

KUMAR, S et al. MEGA 7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. **Molecular Biology and Evolution**, 24: 1870-1874, 2016.

LARSSON, M. C.; DOMINGOS, A. I.; JONES, W. D.; CHIAPPE, M. E.; AMREIN, H. et al. Or83b encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. **Neuron,** 43: 703-714, 2004.

LI, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. **Bioinformatics**, 27(21):2987-93, 2011.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; MARTH, G.; ABECASIS, G.; DURBIN, R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, 25(16):2078-9, 2009.

MARKOW, T. A. Ecological and Evolutionary Genomics: The Cactophilic *Drosophila* Model System. **Journal of Heredity**, 110: 1-3, 2019.

MARKOW, T. A.; HOCUTT, G. D. Reproductive isolation in sonoran desert *Drosophila*: Testing the Limits of the Rules. In: HOWARD, D. J.; BERLOCHER, S. H. (Org.). Endless Forms: Species and Speciation. Oxford: **Oxford Press**, p. 234–244, 1988.

MARKOW, T. A. Host use and host shifts in *Drosophila*. **Sciencedirect**, 31: 139-145, 2019.

MATZKIN, L. M.; WATTS T. D.; BITLER, B. G.; MACHADO, C. A.; MARKOW, T. A. Functional genomics of cactus host shifts in *Drosophila mojavensis*. **Mol Ecol**, 15: 4635-4643, 2006.

McBRIDE, C. S.; ARGUELLO, J. R. Five *Drosophila* genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. Genetics, 177: 1395-1416, 2007.

MILLER, D. E. et al. Highly Contiguous Genome Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing. **G3 (Bethesda)**. 8(10):3131-3141, oct 2018.

MIYATAKE, T.; SHIMIZU, T. Genetic correlations between life-history and behavioral traits can cause reproductive isolation. **Evolution,** 53: 201-208, 1999.

MOMBAERTS, P. Seven-transmembrane proteins as odorant and chemosensory receptors. **Science**, 286(5440):707-711, 1999.

MURRELLl, B.; WERTHEIM, J. O.; MOOLA, S.; WEIGHILL, T.; SCHEFFLER, K.; KOSAKOVSKY-POND, S. L. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genet. 8(7). 2012.

MURRELL, B.; MOOLA, S.; MABONA, A.; WEIGHILL, T.; SHEWARD, D.; KOSAKOVSKY-POND, S. L., et al. FUBAR: A fast, unconstrained bayesian Approximation for inferring selection. **Mol Biol Evol**. 30:1196–205, 2013.

NEI, M.; GOJOBORI, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. **Molecular Biology and Evolution** 3:418-426, 1986.

NIELSEN, R., YANG, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. **Genetics.** 148:929-936.1998.

NGAI, J.; DOWLING, M. M.; BUCK, L.; AXEL, R.; CHESS, A. The family of genes encoding odorant receptors in the channel catfish. **Cell**, 72:657–66, 1993.

NYFFELER, R.; EGGLI, U. A farewell to dated ideas and concepts – molecular phylogenetics and a revised suprageneric classification of the family *Cactaceae*. **Schumannia,** 6, 109–149, 2010.

OLIVEIRA, DCSG; ALMEIDA, FC; O'Grady, PM et al. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila* repleta species group. **Mol Phylogenet Evol**, 64(3):533–544, 2012.

PADRÓ, J.; CARREIRA, V. P.; CORIO, C.; HASSON, E.; SOTO, I. M. Host alkaloids differentially affect developmental stability and wing vein canalization in cactophilic *Drosophila buzzatii*. **J Evol Biol**, 27:2781–2797, 2014.

PATTERSON, J.T.; STONE, W.S. Evolution in the Genus *Drosophila*. 1952

PERTEA, G.; PERTEA, M. GFF Utilities: GffRead and GffCompare. **F1000Research,** 9:304, 2020.

PHAFF, H. J. et al. *Pichia deserticola* and *Candida deserticola*, two new species of yeasts associated with necrotic stems of cacti, Int. J. Syst. Bacteriol, 35:211-216, 1985.

POSADA, D. ModelTest: Média de modelos filogenéticos. **Molecular Biology and Evolution**, 25:1253-1256, 2008.

RANE, R. V. et al. Detoxification Genes Differ Between Cactus-, Fruit-, and Flower-Feeding *Drosophila*. **Journal of Heredity**, 110 (1): 80–91, jan 2019

RANE, R.V. et al. Genomic changes associated with adaptation to arid environments in cactophilic *Drosophila* species. **BMC Genomics**, 20: 52, 2019.

RANE, R. V. et al. Orthonome – a new pipeline for predicting high quality orthologue gene sets applicable to complete and draft genomes. **BMC Genomics**, 18:673, 2017.

RUIZ, A., HEED, W.B. Host plant specificity in the cactophilic *Drosophila mulleri* species complex. J. Anim. **Ecol**, 57: 237–249, 1988.

RUIZ, A., HEED, W.B., WASSERMAN, M. Evolution of the mojavensis cluster of cactophilic *Drosophila* with descriptions of two new species. **J. Hered**, 81: 30– 42, 1990.

SÁNCHEZ-GRACIA, A.; AGUAUDÉ, M.; ROZAS, J. Patterns of nucleotide polymorphism and divergence in the odorant-binding protein genes OS-E and OS-F: analysis in the *melanogaster* species subgroup of *Drosophila*. **Genetics**, 165:1279-1288, 2003.

SÁNCHEZ-GRACIA, A.; ROZAS J. Divergent evolution and molecular adaptation in the *Drosophila* odorant-binding protein family: inferences from sequence variation at the OS-E and OS-F genes. **BMC Evol Biol**, 8: 323, 2008.

SANCHEZ-FLORES, A. et al. Genome Evolution in Three Species of Cactophilic *Drosophila*. **G3 (Bethesda)**, 6(10):3097-3105, oct, 2016.

SHAW, K. H.; JOHNSON, T. K.; ANDERSON, A.; BRUYNE, M.; WARR, C. G. Molecular and Functional Evolution at the Odorant Receptor *Or22* Locus in *Drosophila melanogaster*. **Molecular Biology and Evolution**, 36(5); 919–929, 2019.

SHEN, W.; LE, S.; LI, Y.; HU, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. **PLOS ONE**, 11(10): 163-962, 2016.

SHUMATE, A.; SALZBERG, S. L. Liftoff: accurate mapping of gene annotations. **Bioinformatics**, 2021.

SMITH, D. P. *Drosophila* gustation: a question of taste. **Neuron**, 29: 1–20, 2001

SOTO, I. M.; CARREIRA, V. P.; CORIO, C.; PADRÓ, J.; SOTO, E. M.; HASSON, E. Differences in tolerance to host cactus alkaloids in *Drosophila koepferae* and *D. buzzatii*. **PLoS One**, 9:88-370, 2014.

STANKE, M. et al. AUGUSTUS: A b initio prediction of alternative transcripts. **Nucleic Acids Research**, 34: 435–439, 2006.

STARMER, W. T. et al. Evolution and speciation of host plant specific yeasts. **Evolution**, 34:137-146, 1980.

STOCKER, R. F. The organization of the chemosensory system in *Drosophila melanogaster*: a review. **Cell Tissue Res**, 275: 3–26, 1994.

STORTKUHL, K. F.; KETTLER, R.; FISCHER, S.; HOVEMANN, B. T. An increased receptive field of olfactory receptor Or43a in the antennal lobe of *Drosophila* reduces benzaldehyde-driven avoidance behavior. **Chem Senses**, 30: 81-87, 2005.

SWARUP, S.; WILLIAMS, T. I.; ANHOLT, R. R. Functional dissection of odorant binding protein genes in *Drosophila melanogaster*. **Genes Brain Behav**, 10: 648-657, 2011.

THROCKMORTON, L. The phylogeny, ecology, and geography of *Drosophila*. **Handbook of Genetics**. 421–469, 1975.

TILMON, K. J. Specialization, speciation, and radiation. **University of California Press**, 2008.

VELA, D., RAFAEL, V., Catorce nuevas especies del género *Drosophila* (*Diptera, Drosophilidae*) en el Bosque húmedo montano del volcán Pasochoa, Pichincha. Ecuador. **Rev. Ecuat. Med. Cienc.** Biol. 27, 28–41. 2005.

VIEIRA, F. G, SÁNCHEZ-GRACIA A, ROZAS, J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. **Genome Biol**. 11:235, 2007.

VILELA, C.R., A revision of the *Drosophila* repleta species group (*Diptera, Drosophilidae*). Rev. Bras. **Entomol**. 27, 1–114, 1983.

WANG, P.; LYMAN, R. F.; MACKAY, T. F. C.; ANHOLT, R. R. H. Natural variation in odorant recognition among odorant-binding proteins in *Drosophila melanogaster*. **Genetics,** 184: 759-767, 2010.

WANG, P.; LYMAN, R. F.; SHABALINA, S. A.; MACKAY, T. F. C.; ANHOLT, R. R. H. Association of polymorphisms in odorant-binding protein genes with variation in olfactory response to benzaldehyde in *Drosophila*. **Genetics,** 177: 1655-1665, 2007.

WANG, Y.; WRIGHT, N. J. D.; GUO, H. F.; XIE, Z.; SVOBODA, K.; MALINOW, R.; SMITH, D. P.; ZHONG, Y. Genetic manipulation of the odor-evoked distributed neural activity in the *Drosophila* mushroom body. **Neuron, 29**: 267–276, 2001.

WASSERMAN, M. Evolution and speciation in selected species groups. Evolution of the repleta group. **Academic Press**, 61–139, 1982.

WICHER, D.; SCHAFER, R.; BAURNFEIND, R.; STENSMYR, M.C.; HELLE, R.; HEINEMANN, S.H. et al. *Drosophila* odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. **Nature**. 452: 1007–1011.2008.

XIA, X; XIE, Z. DAMBE: Software Package for Data Analysis in Molecular Biology and Evolution. **Journal of Heredity**, 92(4),371–373, 2001.

YANG, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. **Mol Biol Evol**. 15:568–573,1998.

YANG, Z.; R. NIELSEN. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. **Molecular Biology and Evolution** 17:32-43, 2000.

YANG, Z., W. S. W. WONG; R. NIELSEN. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. **Molecular Biology and Evolution** 22:1107-1118.

YANG, Z. PAML 4: a program package for phylogenetic analysis by maximum likelihood. **Molecular Biology and Evolution**, 24: 1586-1591, 2007.

ZIMIN, A. V. et al. Assembly reconciliation. **Bioinformatics**, 24 (1): 42-45, jan 2008.

8 MATERIAL SUPLEMENTAR

Tabela suplementar 1. Relação espécie e o número de genes que codificam OBPs e ORs encontrados.

Subgênero **Grupo/ Subgrupo**	**Espécie**	**N° de OBPs**	**N° de ORs**
Drosophila			
mulleri/mulleri	*D. mojavensis*	40	63
	D. arizonae	26	57
	D. navojoa	28	56
mulleri/buzzatii	*D. buzzatii*	34	56
hydei	*D. hydei*	32	56
virilis	*D. virilis*	42	65
Total		202	353

Tabela suplementar 2. Análise descritiva do *Orthofinder* para OBP.

Características	
Número de Espécies	6
Número de Genes	202
Número de Genes em Ortogrupos	200
Número de Genes não assinalados	2
Porcentagem de Genes em Ortogrupos	99
Porcentagem de Genes não assinalados	1.0
Número de Ortogrupos	29
Número de Ortogrupos espécie-específico	0
Número de genes em Ortogrupos espécies-específico	0
Porcentagem de genes em Ortogrupos espécies-específico	0

Tabela suplementar 3. Análise descritiva do *Orthofinder* para OR.

Características	
Número de Espécies	6
Número de Genes	353
Número de Genes em Ortogrupos	351
Número de Genes não assinalados	2
Porcentagem de Genes em Ortogrupos	99.4
Porcentagem de Genes não assinalados	0.6
Número de Ortogrupos	50
Número de Ortogrupos espécie-específico	0
Número de genes em Ortogrupos espécies-específico	0
Porcentagem de genes em Ortogrupos espécies-específico	0

Tabela suplementar 4. Relação de OBPs identificadas como ortólogos entre as espécies com relação one-to-one.

	D. arizonae	*D. buzzatii*	*D. hydei*	*D. mojavensis*	*D. navojoa*	*D. virilis*
D. arizonae	-	24	20	26	24	25
D. buzzatii	24	-	20	34	23	31
D. hydei	20	20	-	22	24	24
D. mojavensis	26	34	22	-	26	36
D. navojoa	24	23	24	26	-	27
D. virilis	25	31	24	36	27	-

Tabela suplementar 5. Relação de ORs identificadas como ortólogos entre as espécies com relação one-to-one.

	D. arizonae	*D. buzzatii*	*D. hydei*	*D. mojavensis*	*D. navojoa*	*D. virilis*
D. arizonae	-	48	47	48	50	42
D. buzzatii	48	-	48	50	45	42
D. hydei	47	48	-	45	42	45
D. mojavensis	48	50	45	-	48	40
D. navojoa	50	45	42	48	-	42
D. virilis	42	42	45	40	42	-

Tabela suplementar 6. Genes parálogos OBPs nos genomas das espécies *D. hydei* e *D. virilis* e os ortogrupos aos quais foram associados.

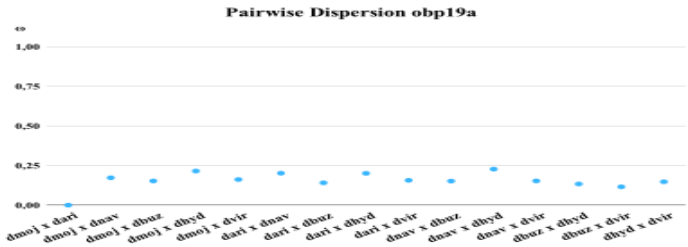
Ortogrupo	**Espécie**	**Gene 1**	**Gene 2**
	D. hydei		
OG0000001		Obp99a LOC115482839	Obp99a LOC111600677
OG0000005		Obp56h LOC111592805	Obp56h LOC115483814
OG0000017		Obp57c LOC111595528	Obp57c LOC111595431
	D. virilis		
OG0000000		Obp46a LOC6625059	Obp46a LOC6624798
OG0000023		Obp50a LOC6627070	Obp50c LOC6627071

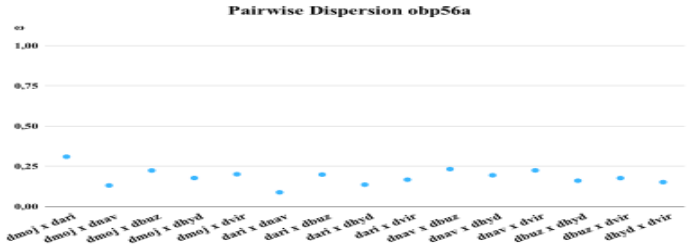
Tabela suplementar 7. Genes parálogos *Or* nos genomas das espécies *D. hydei*, *D. mojavensis, D. navojoa* e *D. virilis,* e os ortogrupos aos quais foram associados.

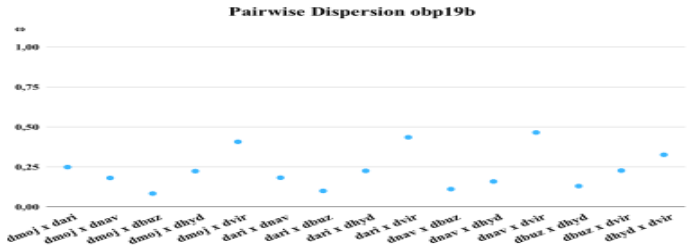
Ortogrupo	**Espécie**	**Gene 1**	**Gene 2**
	D. hydei		
OG0000010		OR85d LOC111598625	OR85d LOC111597551
	D.mojavensis		
OG0000000		OR59b LOC26528727	OR59b-like LOC116803708
OG0000006		OR22c LOC6578143	OR22c LOC6578143
OG0000012		OR47b LOC6578413	OR47b LOC6578414
OG0000041		OR85c LOC6572199	OR85c LOC26528198
	D. navojoa		
OG0000015		OR67d LOC108652442	OR67d LOC108652442
	D. virilis		
OG0000000		OR59b LOC6630620	OR59b LOC26531549
OG0000001		OR42a LOC6636383	OR42a LOC6636382
OG0000005		OR59a LOC6627051	OR59a LOC6626964
OG0000009		OR92a LOC663149	OR92a LOC6631499
OG0000011		OR83c LOC6628121	OR83c LOC116651603
OG0000013		OR45b LOC6636460	OR45b LOC116651034

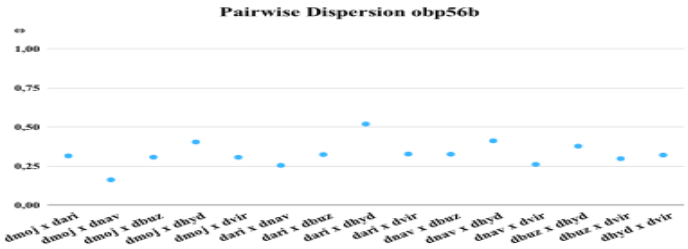
Figura suplementar 1. Testes de Seleção par a par.

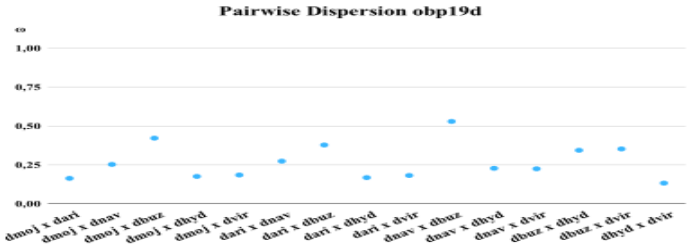
Figura suplementar 1. 1A. Genes *Obp*.

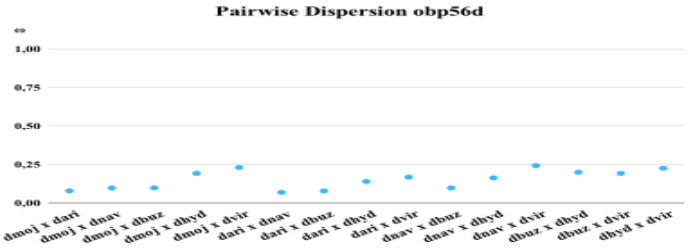


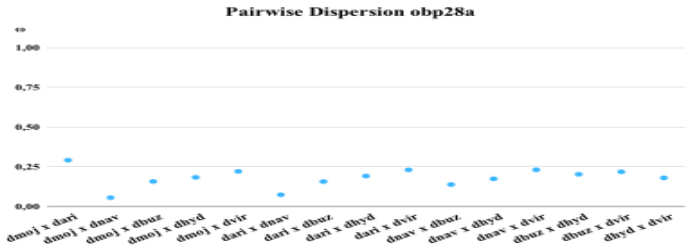


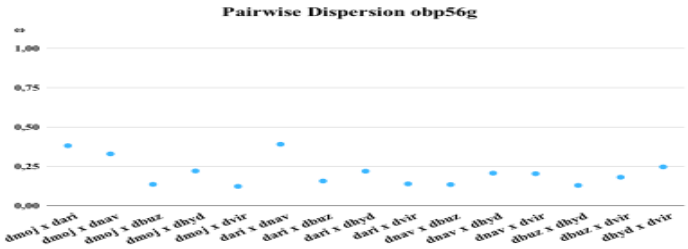


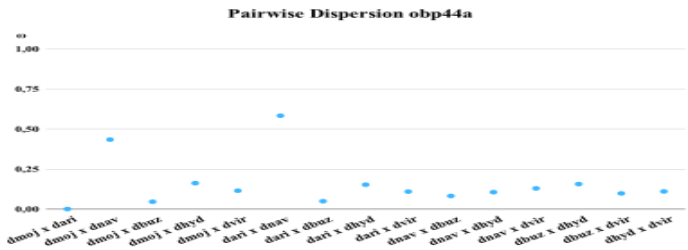


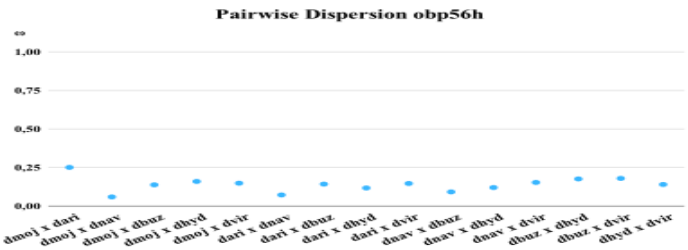




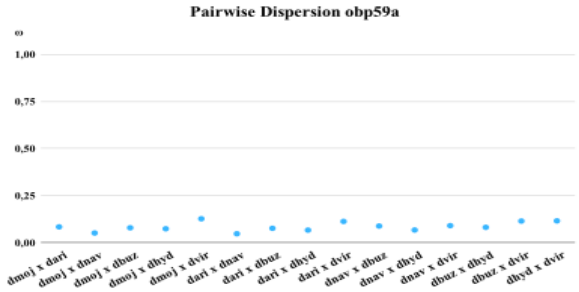


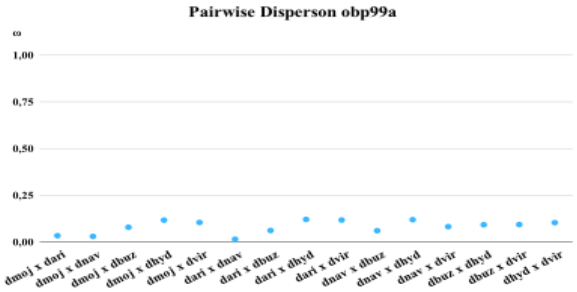


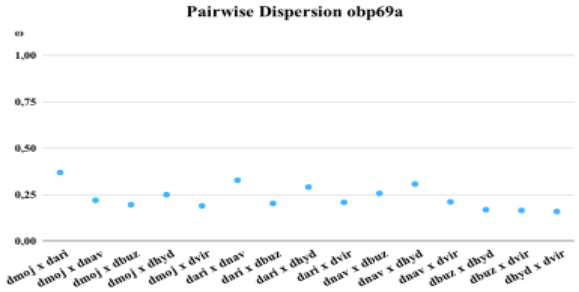


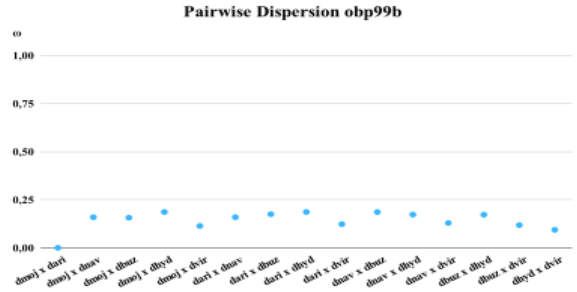


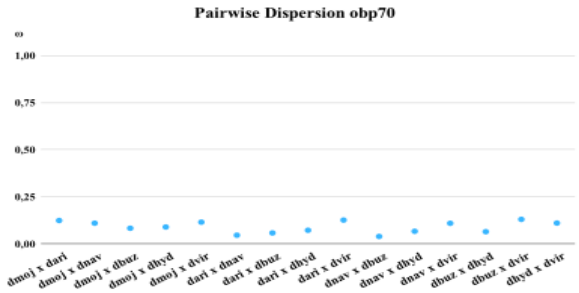
Continuação Figura Suplementar 1. 1A. Genes *Obp*.

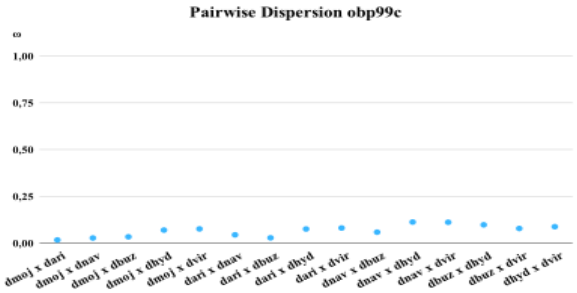


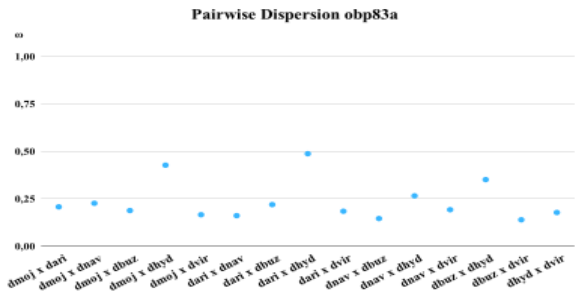


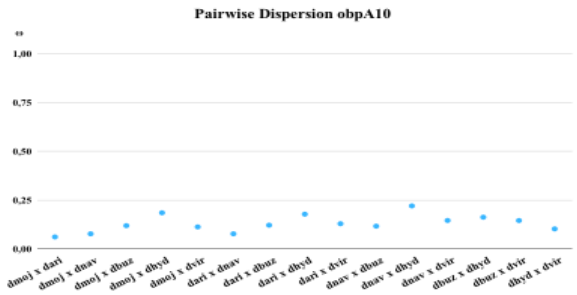












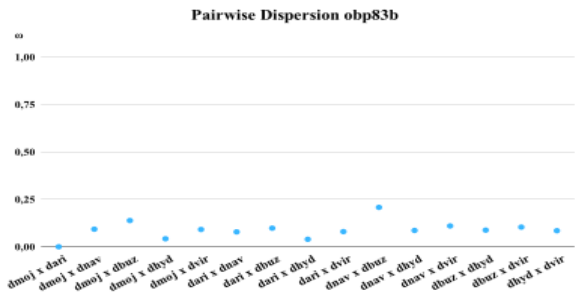
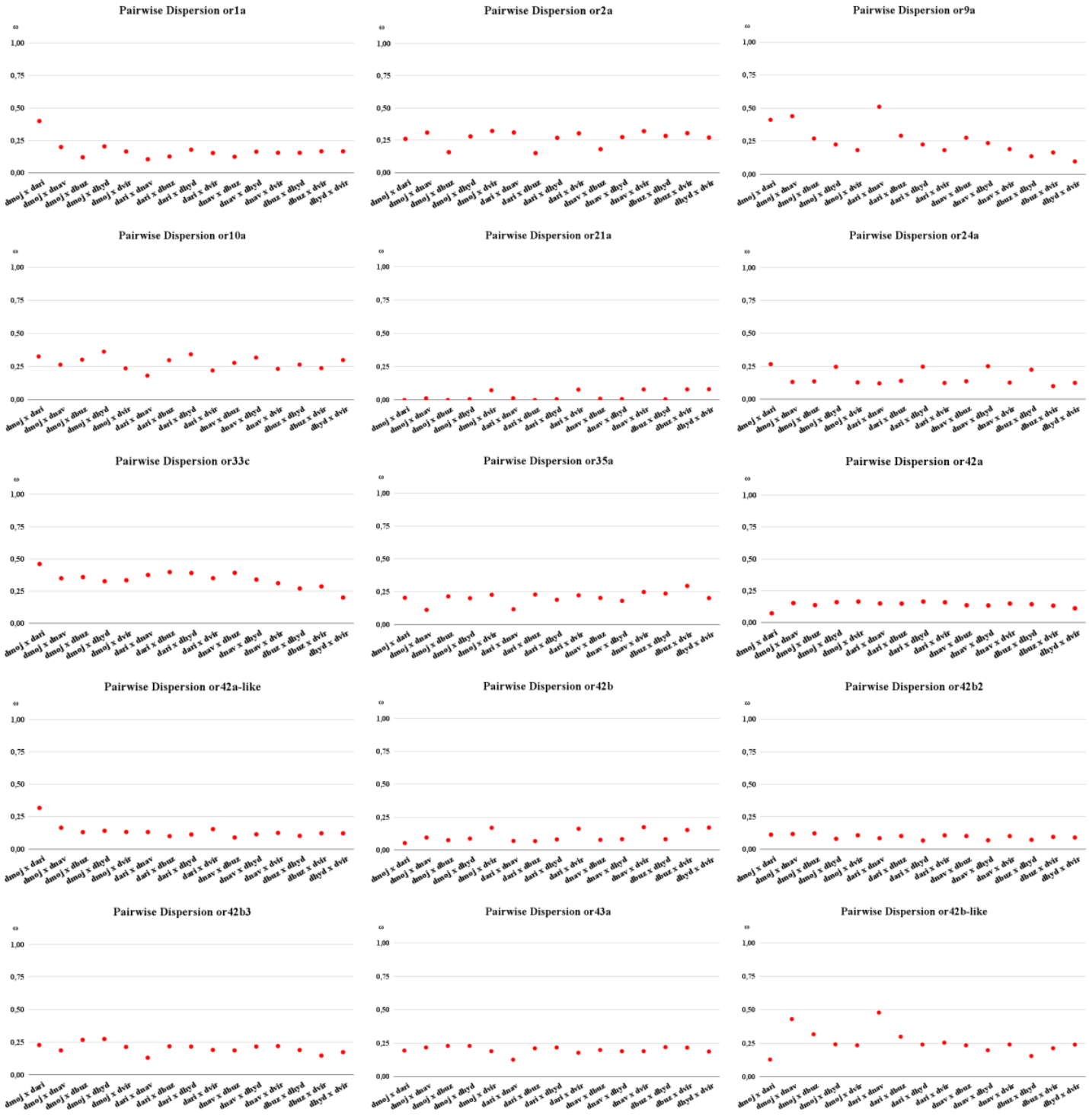
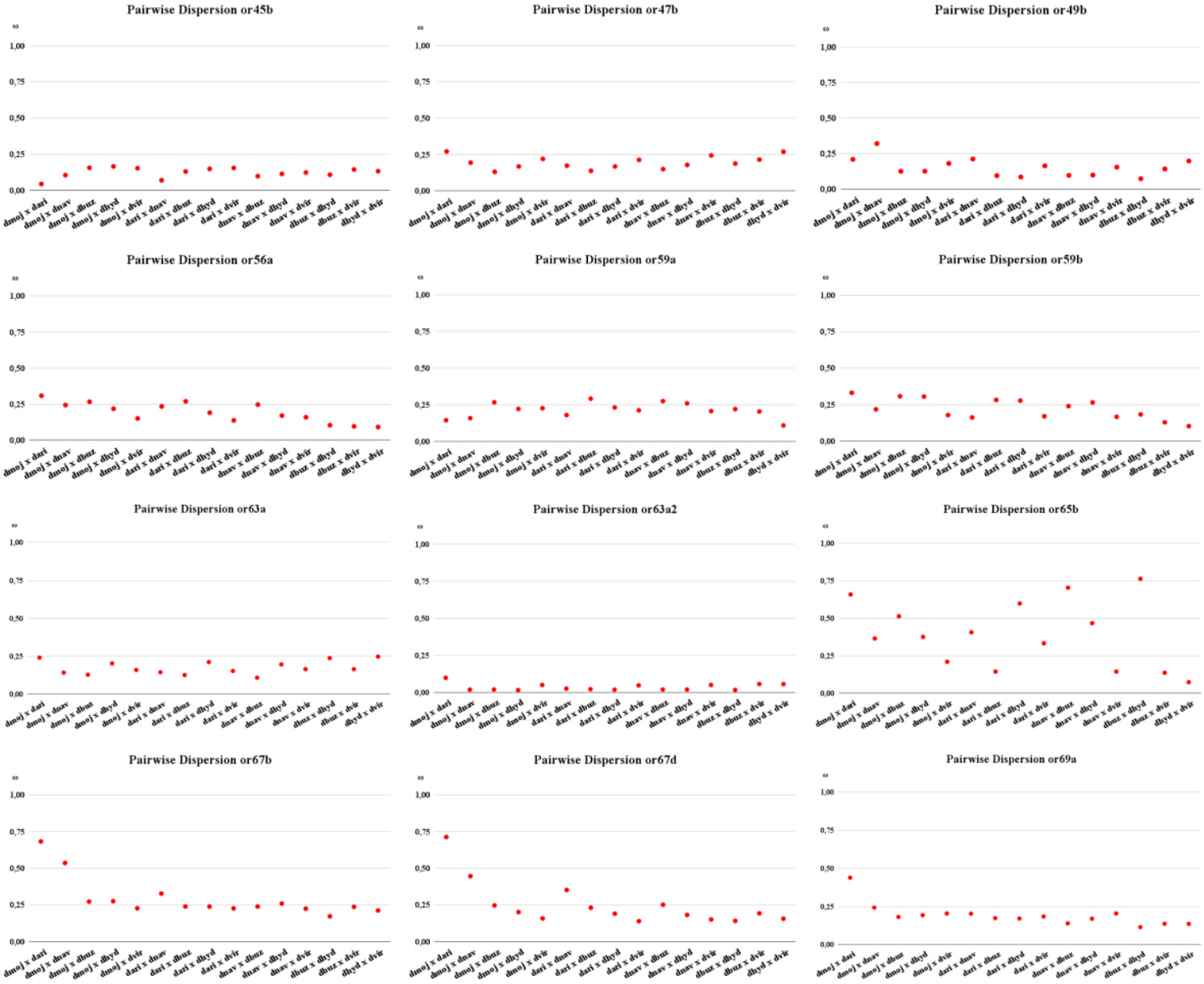


Figura suplementar 1. 1B. Genes *Or*.



Continuação Figura Suplementar 1. 1B. Genes *Or*.



Continuação Figura Suplementar 1. 1B. Genes *Or*.

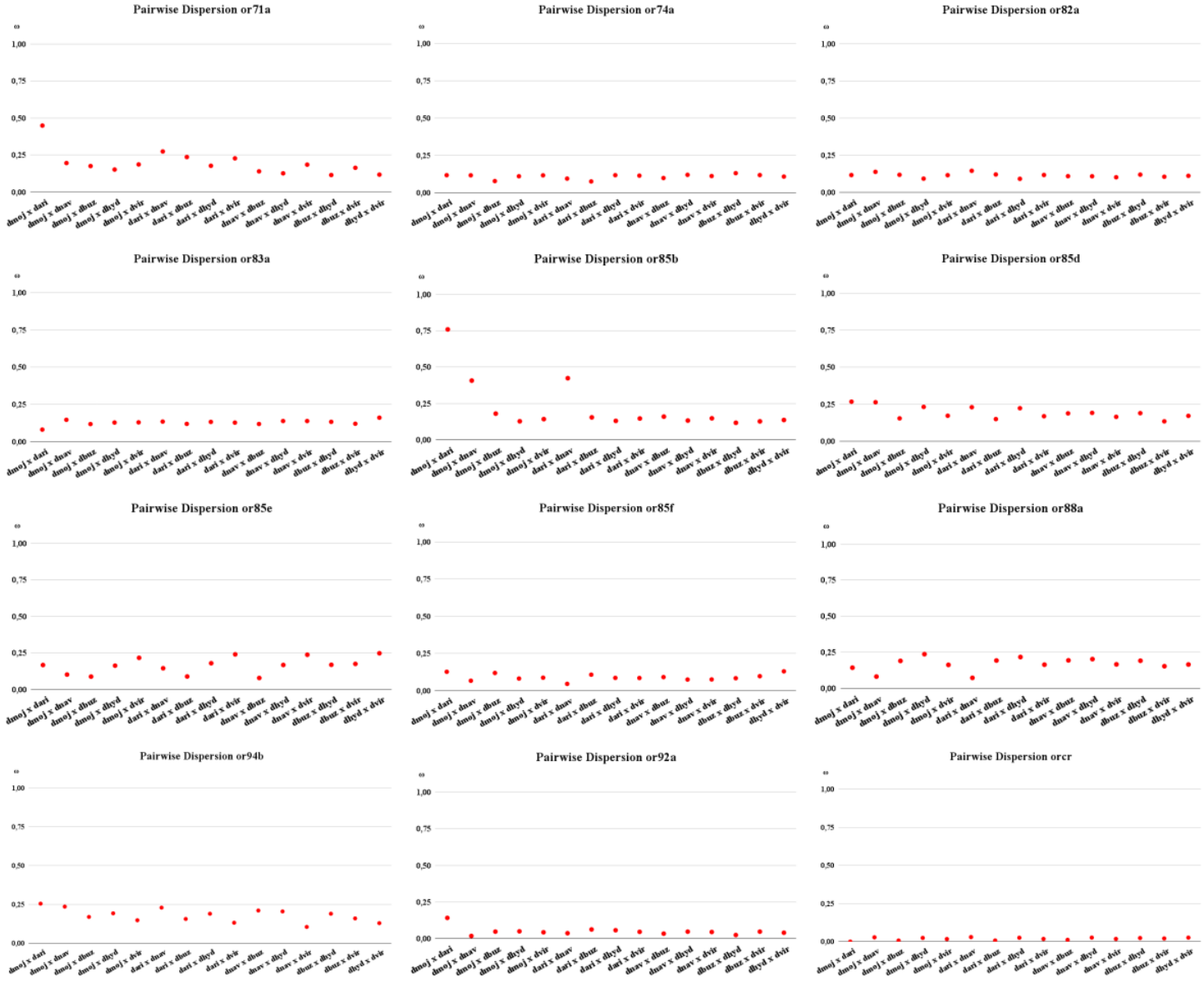
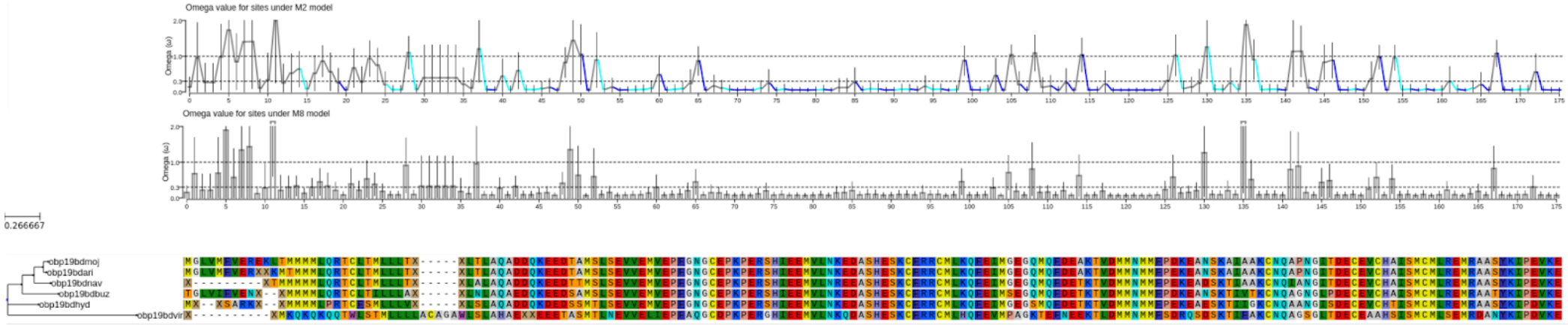
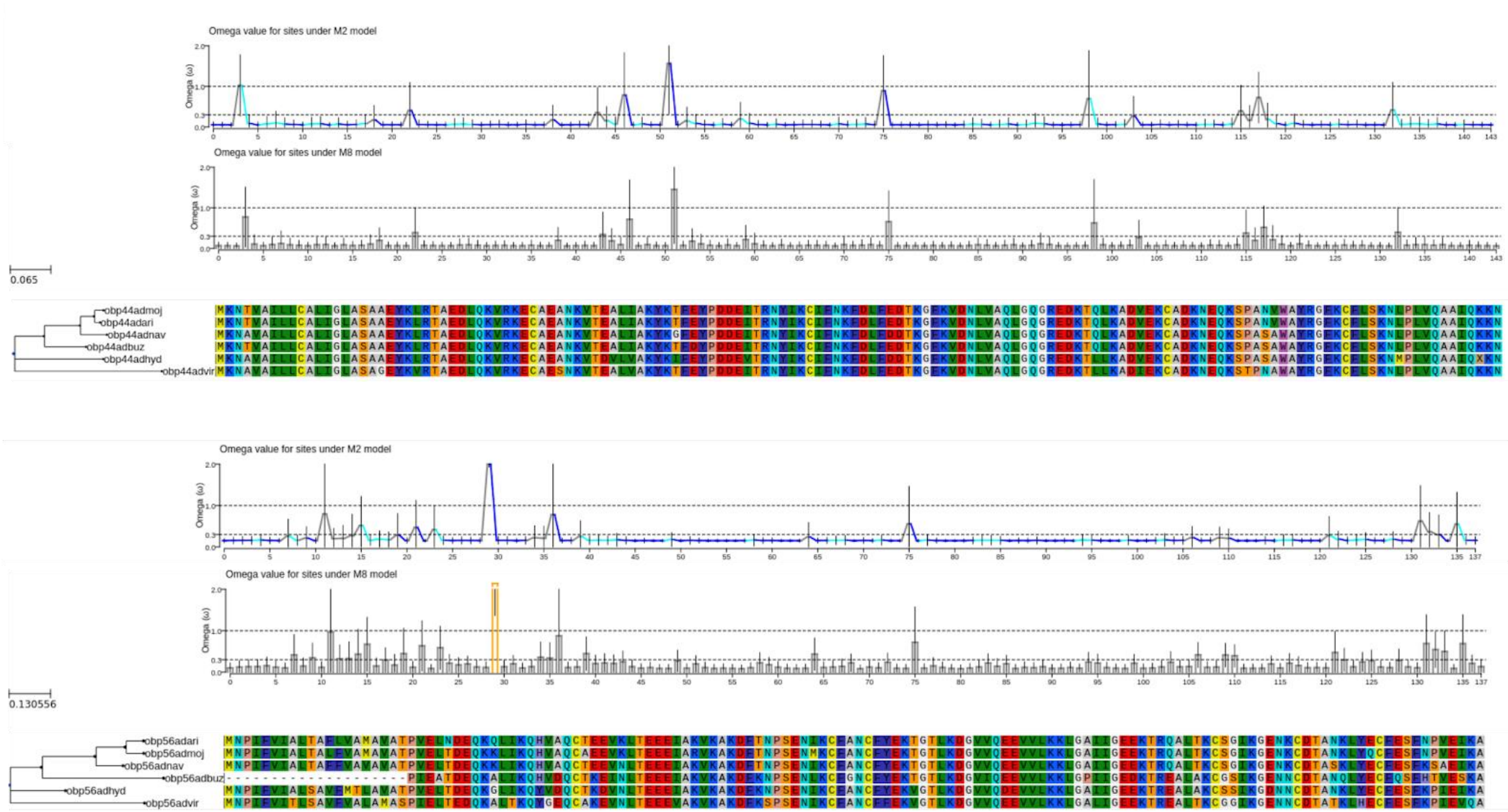


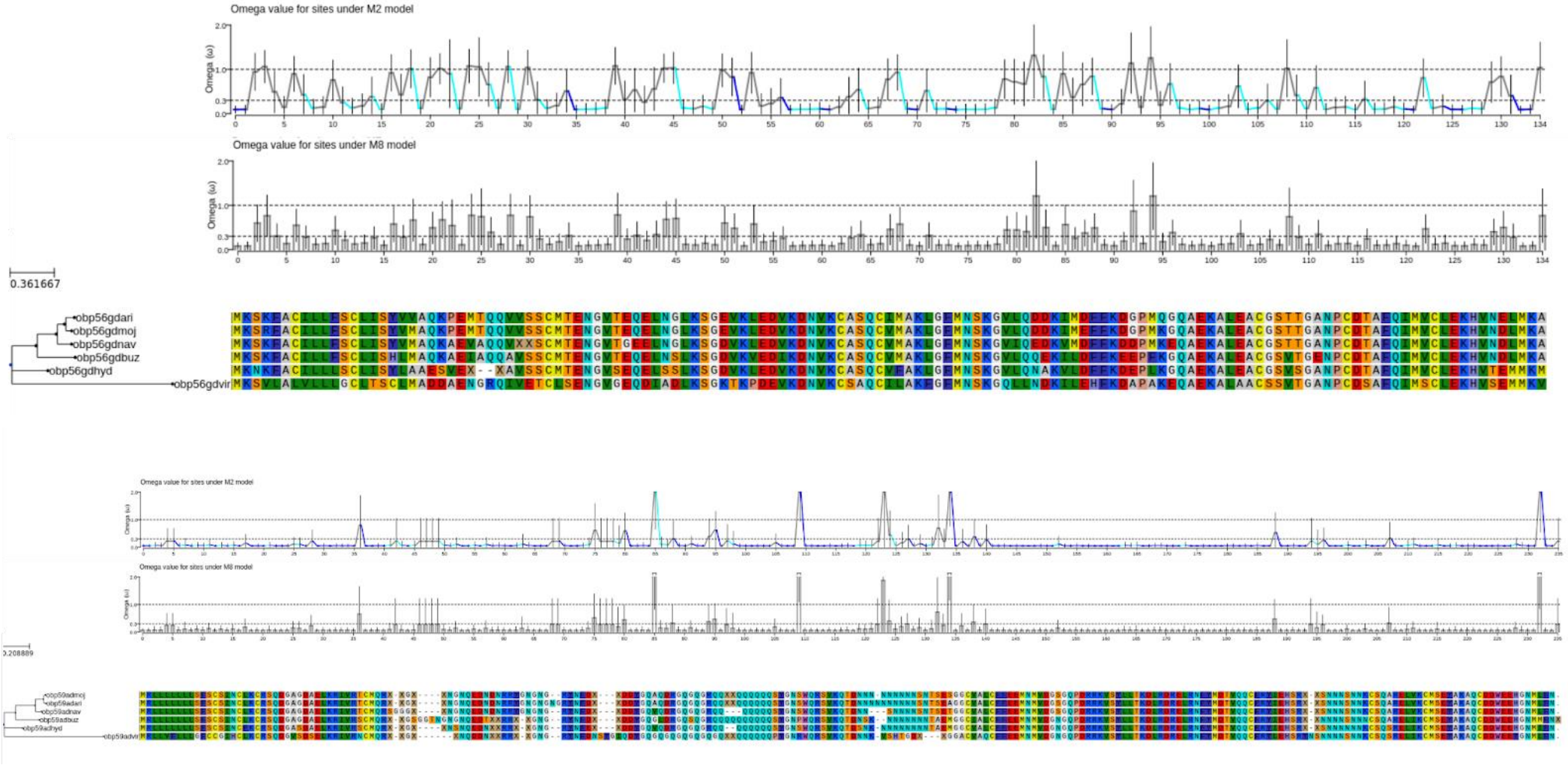
Figura suplementar 2. Visualização dos resultados M2-M8 das análises do codeML para as 16 proteínas OBP que apresentaram sítios selecionados positivamente em pelo menos um dos quatro testes de seleção realizados, disponibilizado pelo *toolkit* ETE. As mudanças de aminoácidos significativas para o teste do Codeml estão em evidência em amarelo.



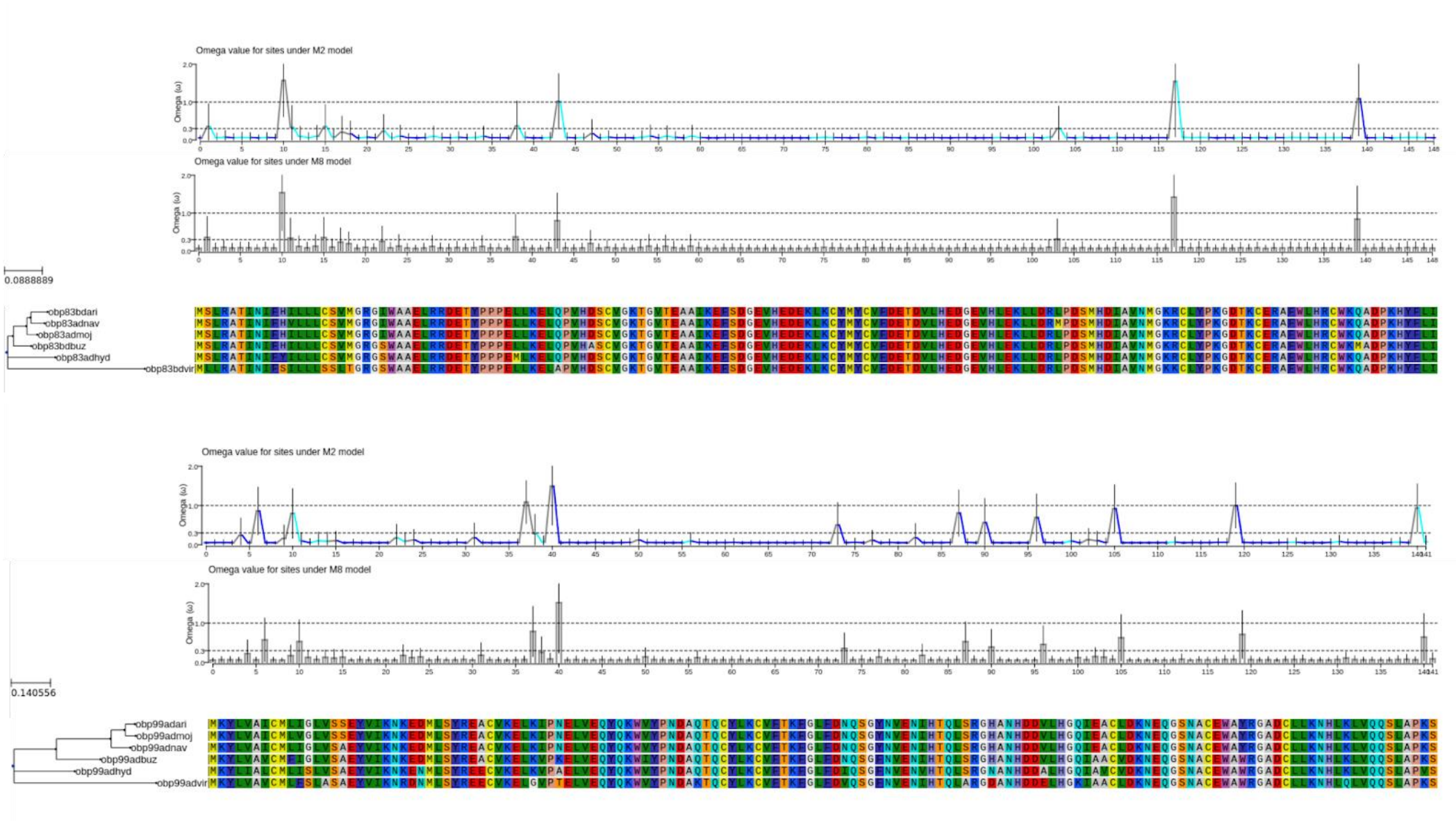
Continuação Figura suplementar 2.



Continuação Figura suplementar 2.



Continuação Figura suplementar 2.



Continuação Figura suplementar 2.

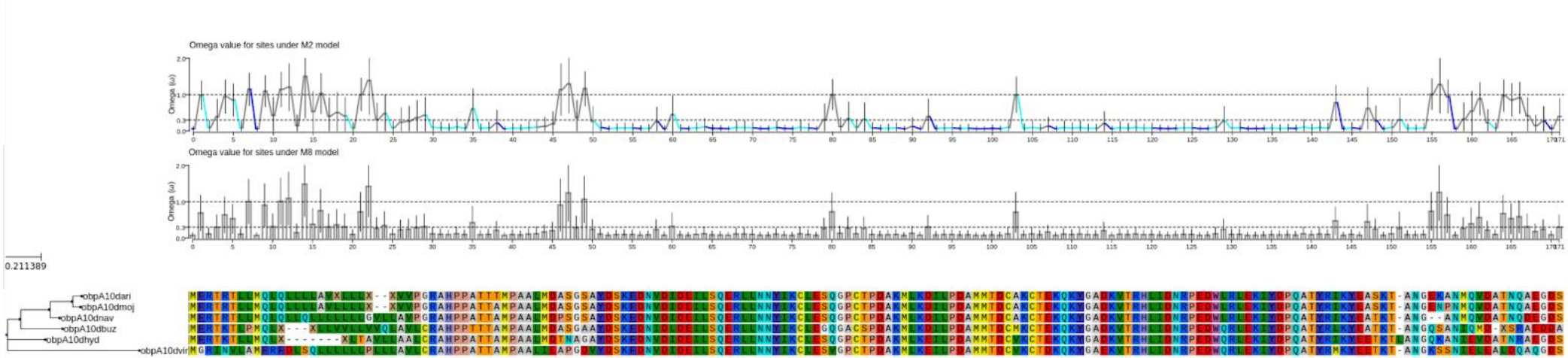
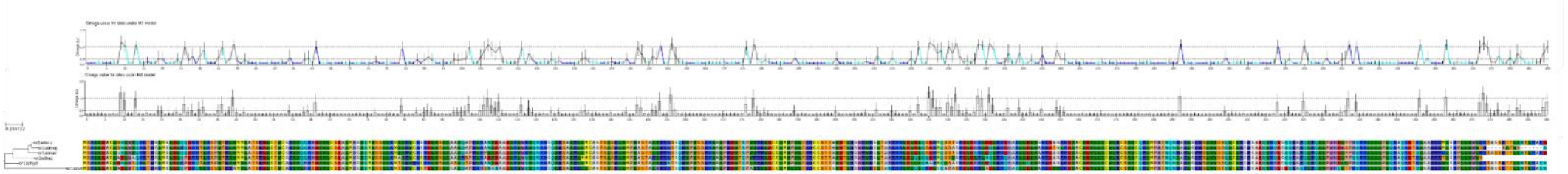
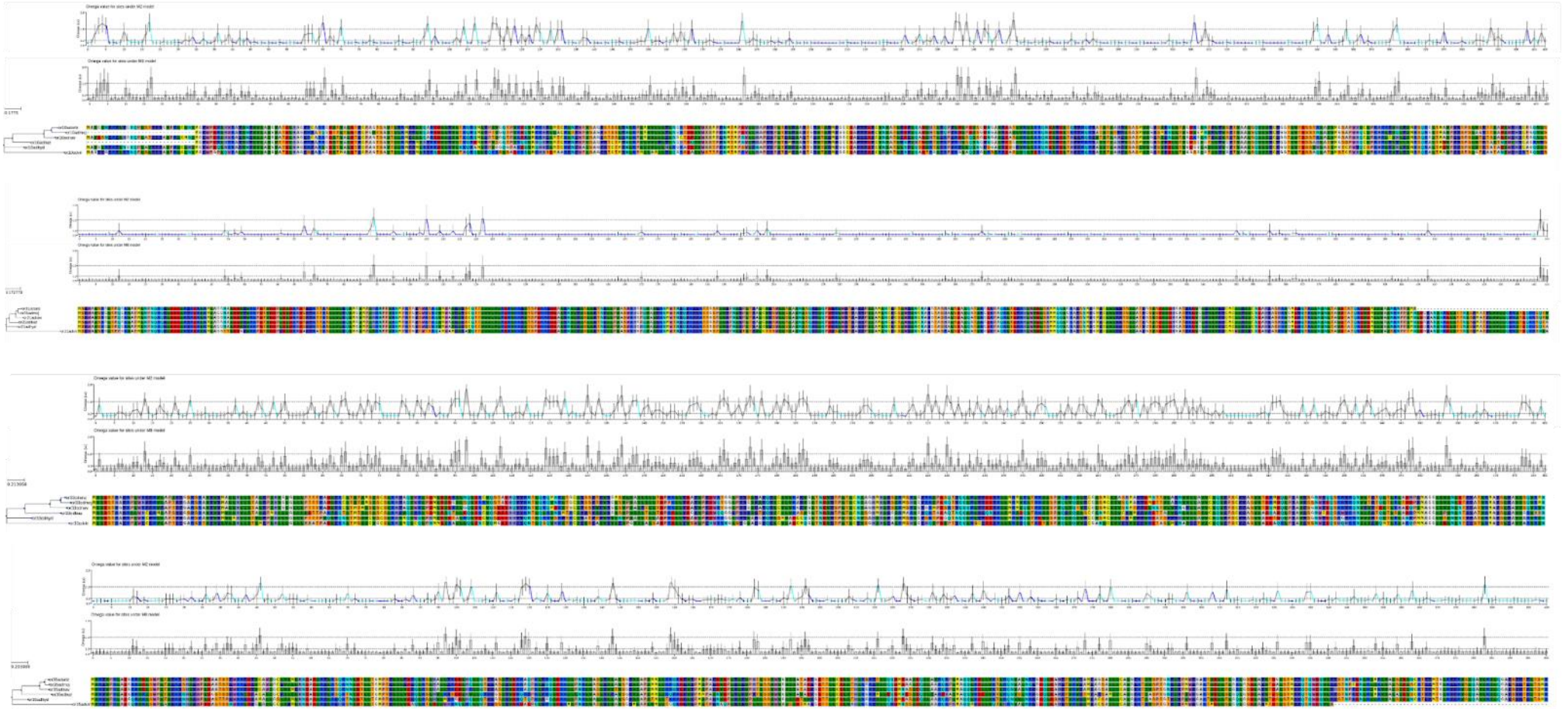


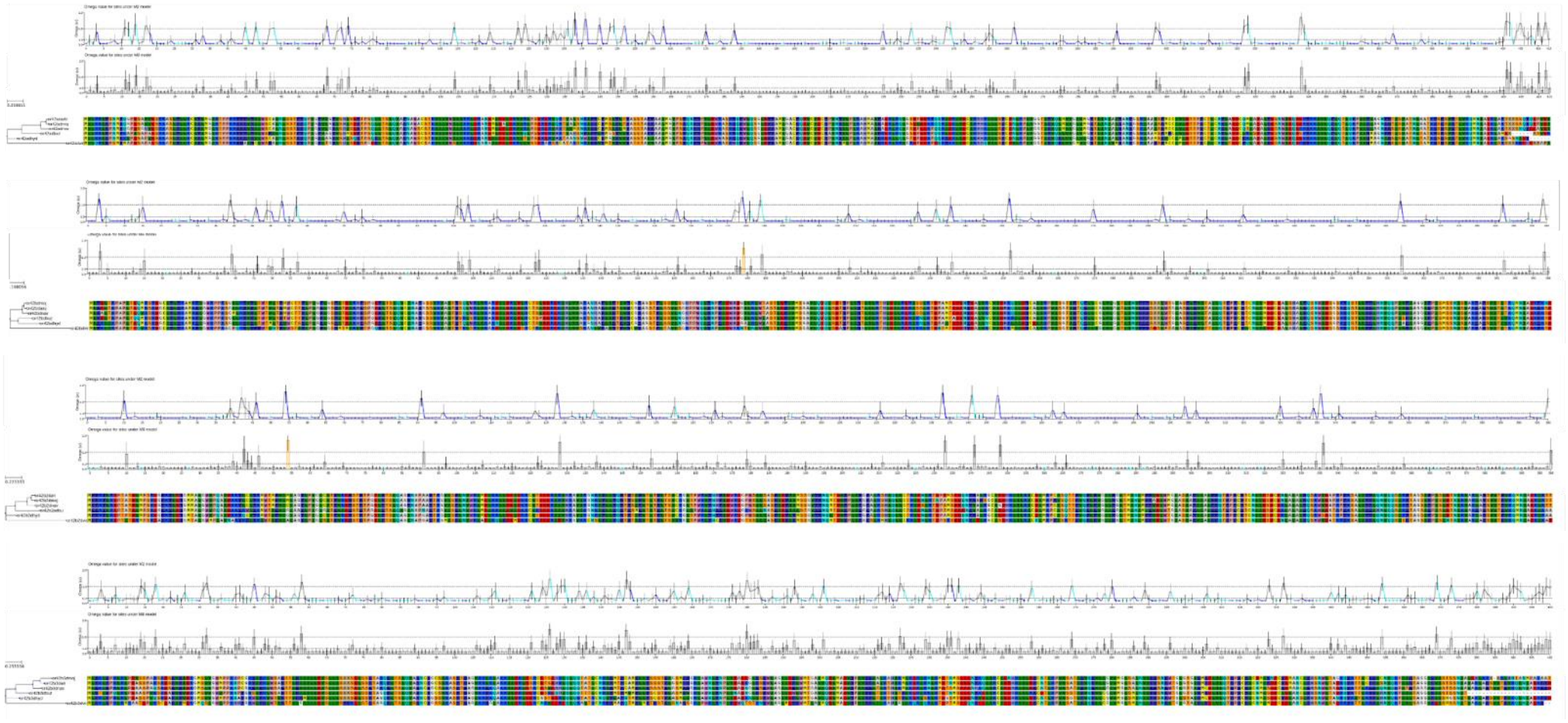
Figura suplementar 3. Visualização dos resultados M2-M8 das análises do codeML para as 39 proteínas OR que apresentaram sítios selecionados positivamente em pelo menos um dos quatro testes de seleção performados, disponibilizado pelo toolkit ETE. As mudanças de aminoácidos significativas para o teste do Codeml estão em evidência em amarelo.



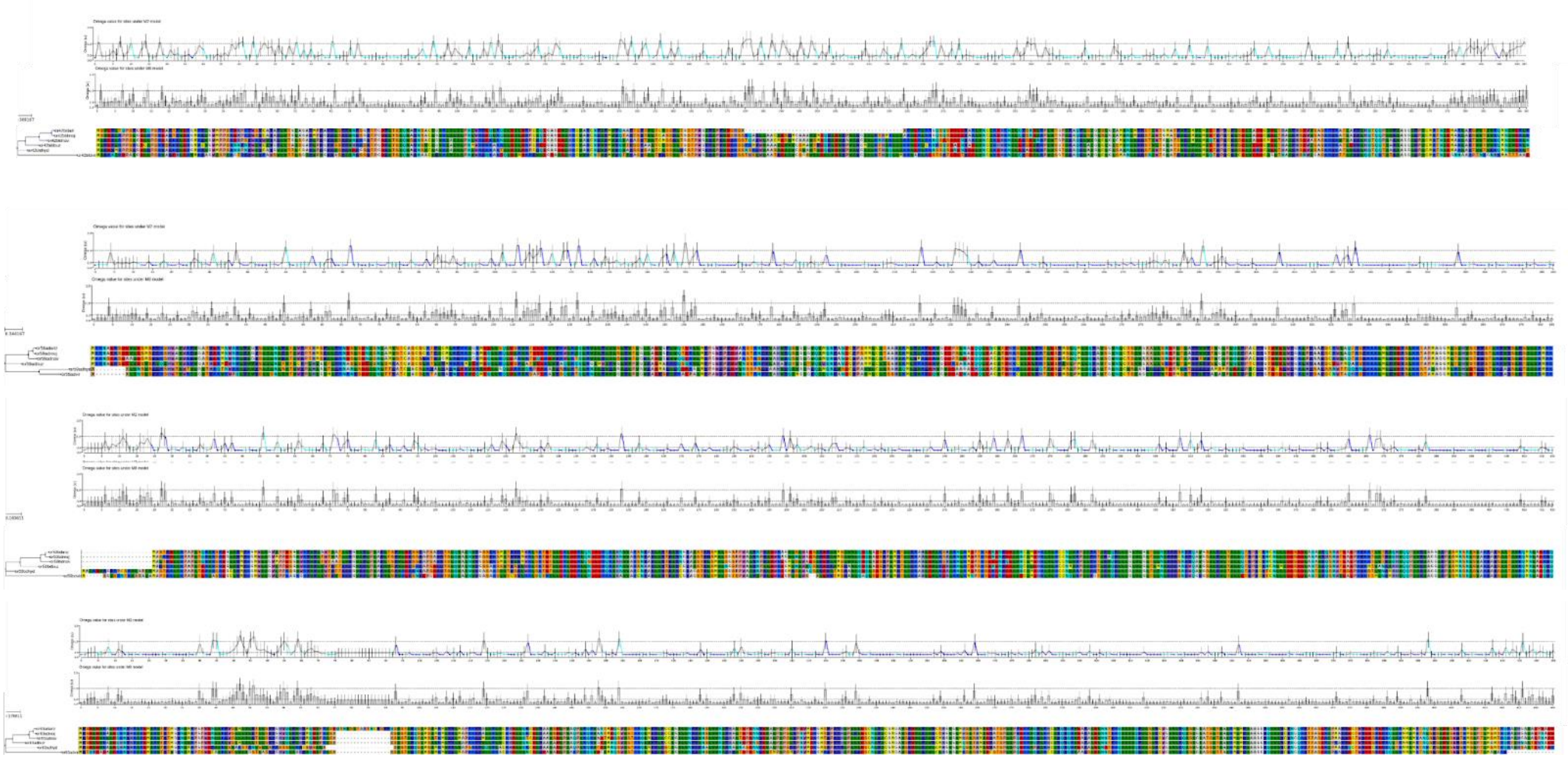
Continuação Figura suplementar 3.



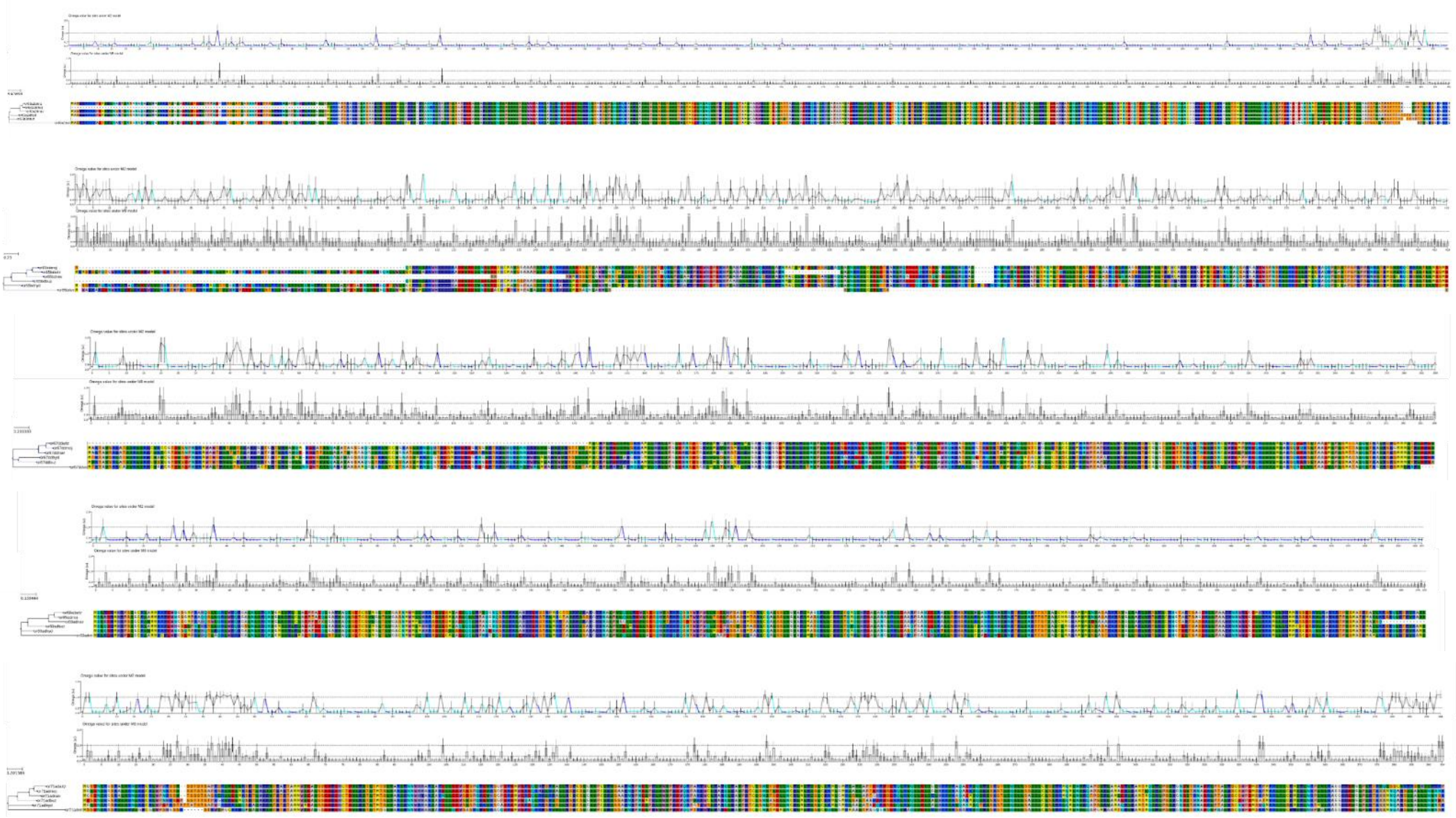
Continuação Figura suplementar 3.



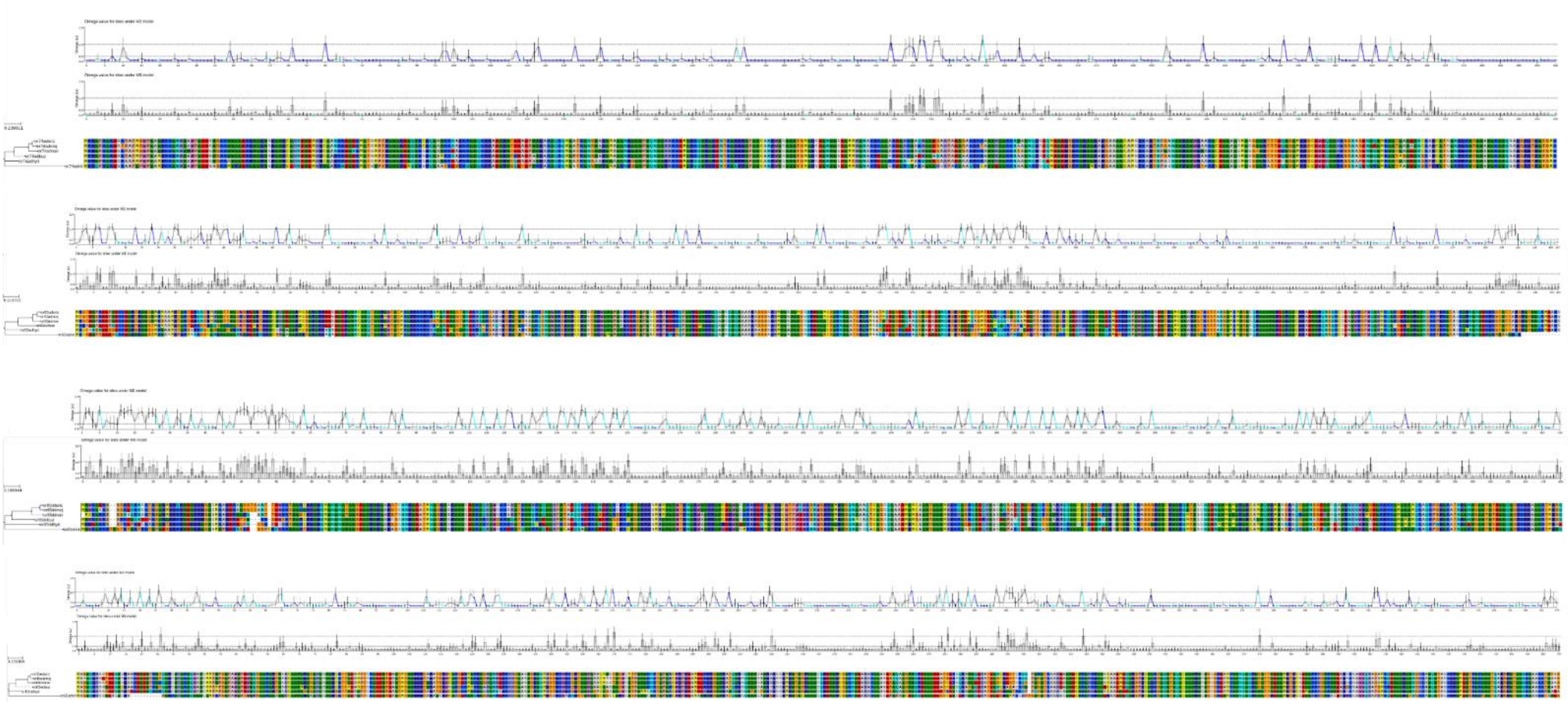
Continuação Figura suplementar 3.



Continuação Figura suplementar 3.



Continuação Figura suplementar 3.



Continuação Figura suplementar 3.

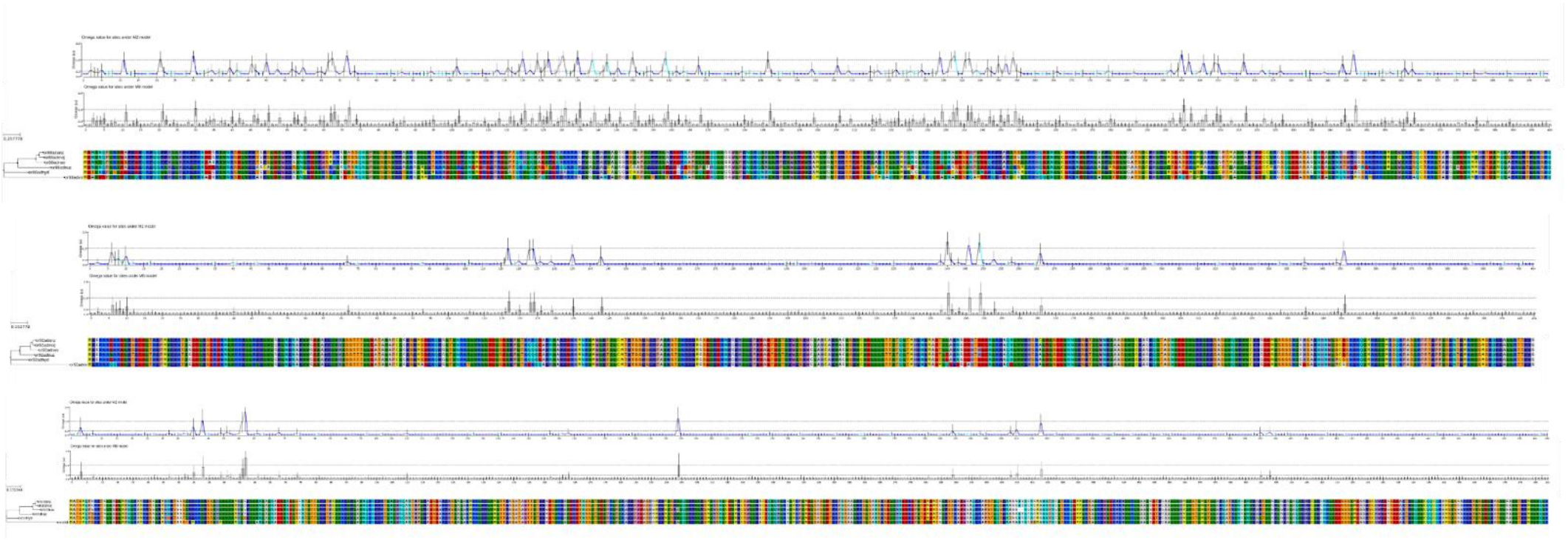
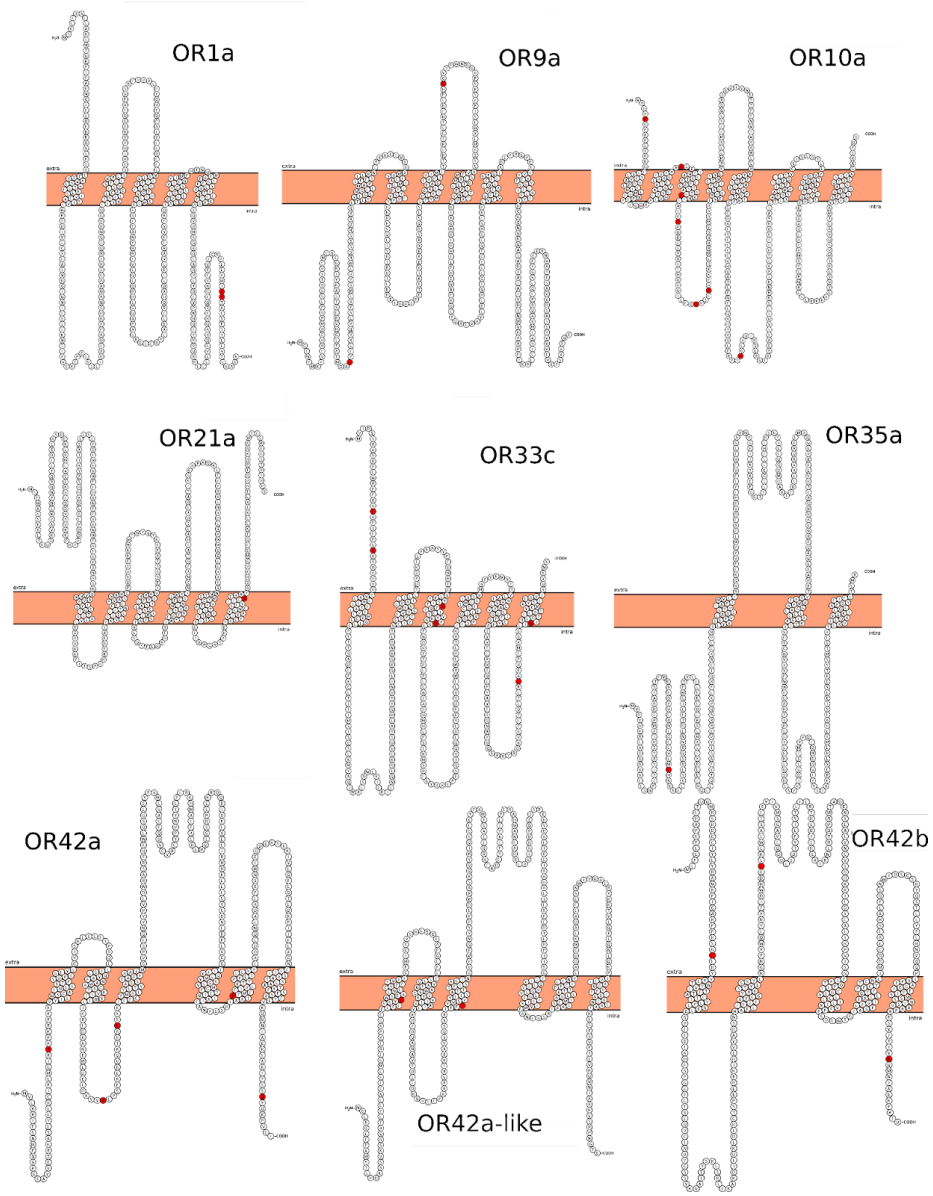


Figura suplementar 4. Predição das hélices transmembranas em 31 proteínas OR, indicando os sítios com mudanças significativas de aminoácidos em pelo menos um dos quatro testes de seleção realizados.



Continuação Figura suplementar 4.

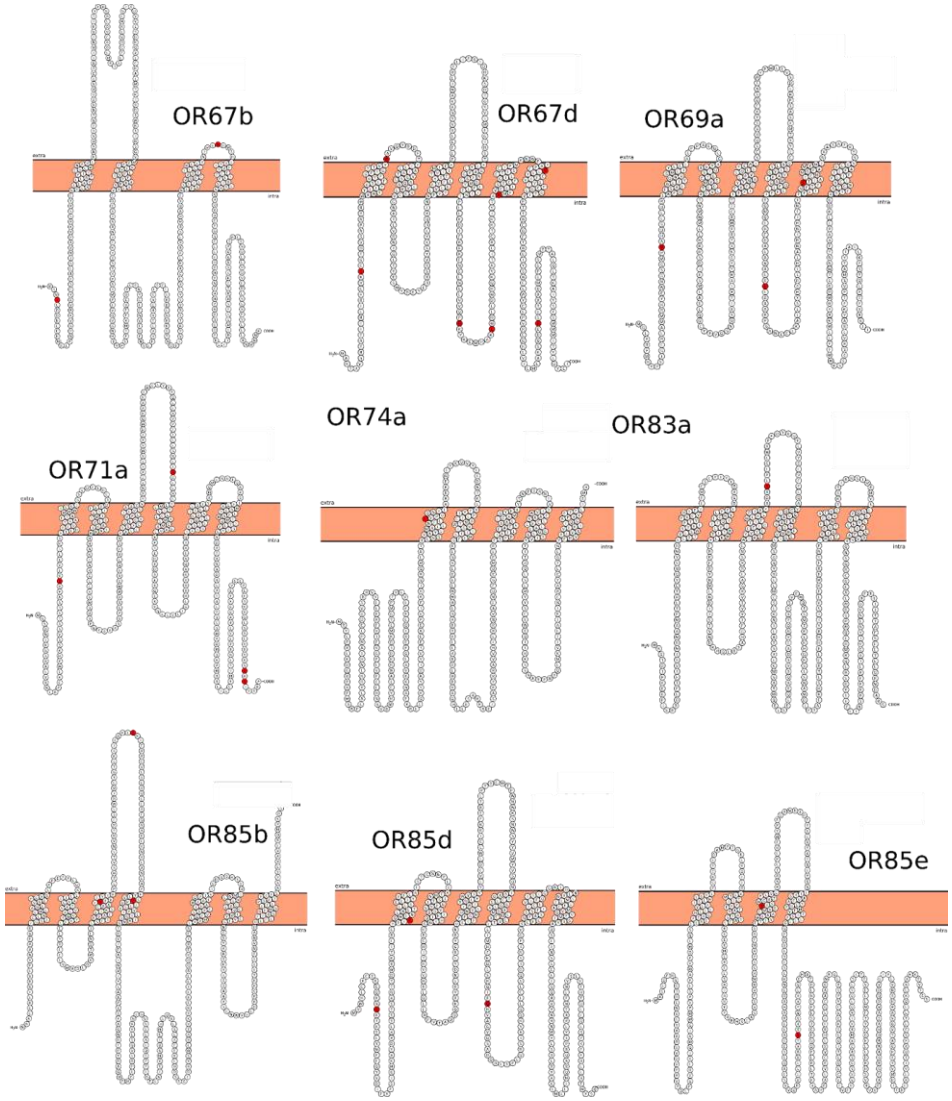
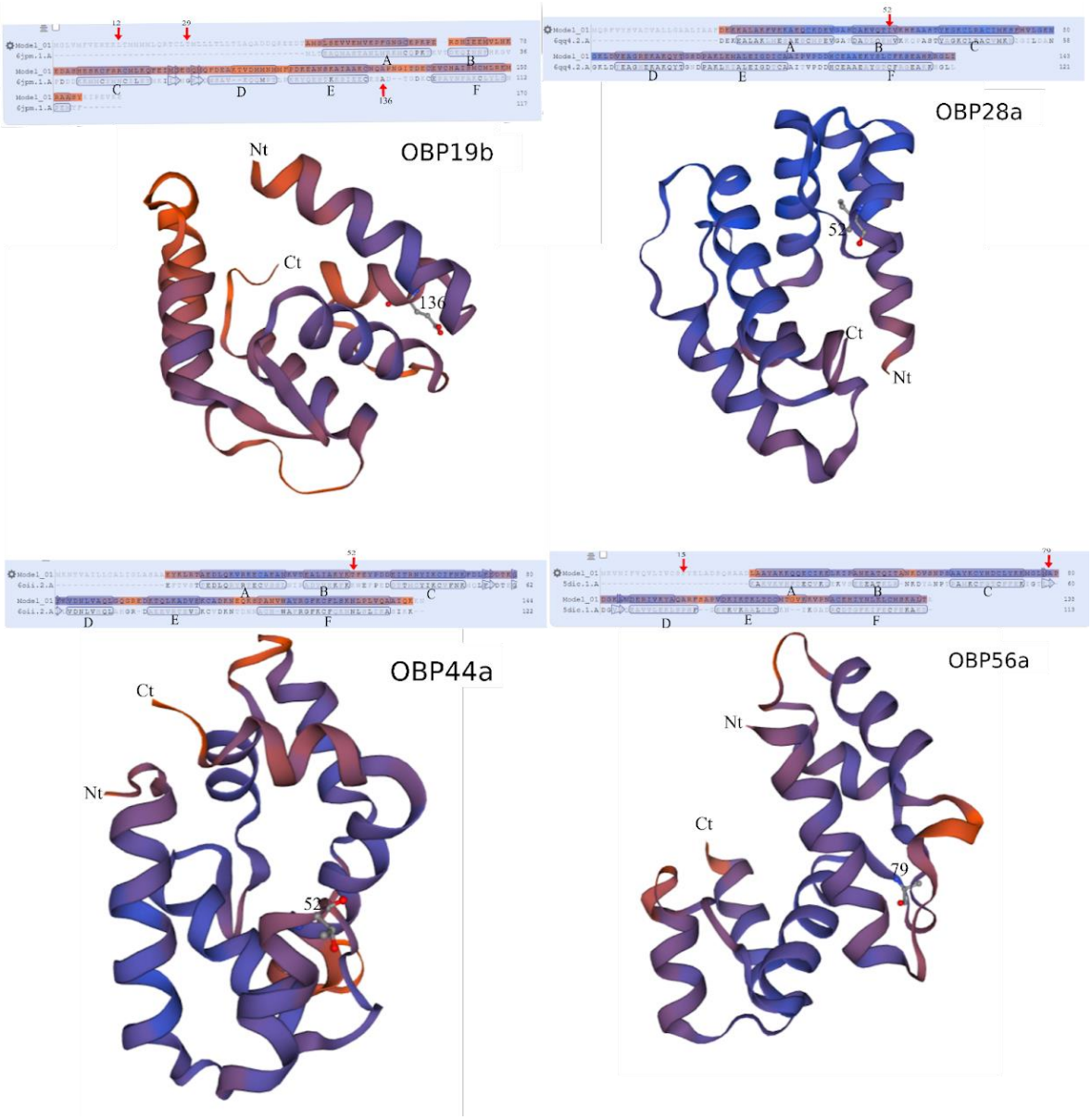
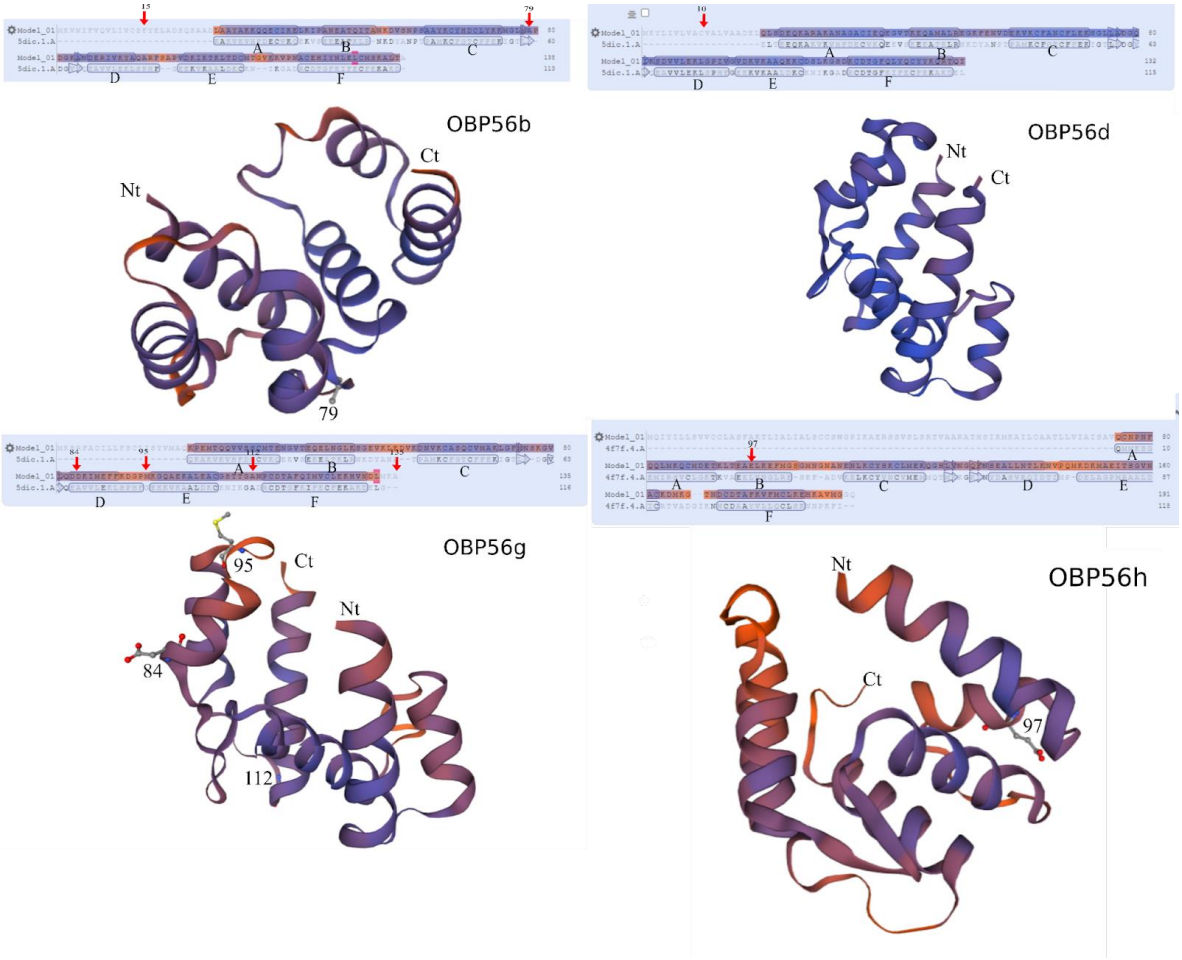


Figura suplementar 5. Predição da estrutura 3D de 16 proteínas OBP, indicando os sítios com mudanças significativas de aminoácidos em pelo menos um dos quatro testes realizados.



Continuação Figura suplementar 5.



Continuação Figura suplementar 5.