UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto de Ciências Biológicas Departamento de Genética, Ecologia e Evolução Programa de Pós-graduação em Genética

Rafaella Ferreira Soares

EVOLUÇÃO DO GENE DA PROTEÍNA CENTROMÉRICA CENP-C EM ESPÉCIES DE DROSOPHILA DO GRUPO MONTIUM

Belo Horizonte 2022 Rafaella Ferreira Soares

EVOLUÇÃO DO GENE DA PROTEÍNA CENTROMÉRICA CENP-C EM ESPÉCIES DE DROSOPHILA DO GRUPO MONTIUM

Dissertação apresentada ao programa de Pós-Graduação em Genética da Universidade Federal de Minas Gerais como pré-requisito obrigatório para obtenção do título de Mestre em Genética, área de concentração Genômica e Bioinformática.

Orientador: Dr. Gustavo Campos e Silva Kuhn Coorientador: Dr. Leonardo Barbosa Koerich

Belo Horizonte 2022

043	Soares, Rafaella Ferreira. Evolução do gene da proteína centromérica Cenp-C em espécie Drosophila do grupo montium [manuscrito] / Rafaella Ferreira Soar 70 f. : il. ; 29,5 cm.	es de res. – 2022.
	Orientador: Dr. Gustavo Campos e Silva Kuhn. Coorientador: D Barbosa Koerich. Dissertação (mestrado) – Universidade Federal de Minas Gerai de Ciências Biológicas. Programa de Pós-Graduação em Genétic	r. Leonardo s, Instituto a.
	1. Genômica. 2. Centrómero. 3. Drosophila. 4. Duplicação Gêni Centromérica A. I. Kuhn, Gustavo Campos e Silva. II. Koerich, Leo Barbosa. III. Universidade Federal de Minas Gerais. Instituto de C Biológicas. IV. Título.	ca. 5. Proteína onardo iências
		CDU: 575

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 – 2680

SEI/UFMG - 1217421 - Ata de defesa de Dissertação/Tese

https://sei.ufmg.br/sei/controlador.php?acao=documento_imprimir_w...



UNIVERSIDADE FEDERAL DE MINAS GERAIS Instituto de Ciências Biológicas Programa de Pós-Graduação em Genética

ATA DE DEFESA DE DISSERTAÇÃO / TESE

ATA DA DEFESA DE DISSERTAÇÃO	320/2022
	entrada
	2°/2019
Rafaella Ferreira Soares	CPF: 113.558.496-67

Às quatorze horas do dia **28 de janeiro de 2022**, reuniu-se remotamente (virtualmente) a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Evolução do gene da proteína centromérica Cenp-C em espécies de Drosophila do grupo montium"**, requisito para obtenção do grau de Mestre em **Genética.** Abrindo a sessão, o Presidente da Comissão, **Gustavo Campos e Silva Kuh**n, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Gustavo Campos e Silva Kuhn	UFMG	260.136.648-62	Aprovada
Leonardo Barbosa Koerich	UFMG	033.549.409-99	Aprovada
Francisco Pereira Lobo	UFMG	012.273.736-94	Aprovada
Maria Dulcetti Vibranovski	USP	076.362.867-00	Aprovada

Pelas indicações, a candidata foi considerada: Aprovada

O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 28 de janeiro de 2022.

Gustavo Campos e Silva Kuhn (UFMG)

Leonardo Barbosa Koerich (UFMG)

Francisco Pereira Lobo (UFMG)

Maria Dulcetti Vibranovski (USP)

Assinatura dos membros da banca examinadora:

SEI/UFMG - 1217421 - Ata de defesa de Dissertação/Tese

https://sei.ufmg.br/sei/controlador.php?acao=documento_imprimir_w...



Referência: Processo nº 23072.204948/2022-16

SEI nº 1217421

31/01/2022 11:06

SEI/UFMG - 1217442 - Folha de Aprovação

https://sei.ufmg.br/sei/controlador.php?acao=documento_imprimir_w...



UNIVERSIDADE FEDERAL DE MINAS GERAIS Instituto de Ciências Biológicas Programa de Pós-Graduação em Genética

FOLHA DE APROVAÇÃO

"Evolução do gene da proteína centromérica Cenp-C em espécies de Drosophila do grupo montium"

Rafaella Ferreira Soares

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Gustavo Campos e Silva Kuhn UFMG

Leonardo Barbosa Koerich UFMG

Francisco Pereira Lobo UFMG

Maria Dulcetti Vibranovski USP

Belo Horizonte, 28 de janeiro de 2022.

	-
coil	-
See	£
assinatura	2
eletrônica	

Documento assinado eletronicamente por **Francisco Pereira Lobo, Professor do Magistério Superior**, em 28/01/2022, às 17:58, conforme horário oficial de Brasília, com fundamento no art. 5º do <u>Decreto nº 10.543, de 13 de novembro de 2020</u>.

Documento assinado eletronicamente por **Maria Dulcetti Vibranovski, Usuário Externo**, em 28/01/2022, às 18:03, conforme horário oficial de Brasília, com fundamento no art. 5º do <u>Decreto</u> nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gustavo Campos e Silva Kuhn, Professor do Magistério Superior**, em 28/01/2022, às 18:12, conforme horário oficial de Brasília, com fundamento no art. 5º do <u>Decreto nº 10.543, de 13 de novembro de 2020</u>.

31/01/2022 11:07

SEI/UFMG - 1217442 - Folha de Aprovação

https://sei.ufmg.br/sei/controlador.php?acao=documento_imprimir_w...

COIL	
No. Contract of the	p ²²
SCI	r,
assinatura	-

Documento assinado eletronicamente por **Leonardo Barbosa Koerich**, **Professor do Magistério Superior**, em 31/01/2022, às 08:53, conforme horário oficial de Brasília, com fundamento no art. 5º do <u>Decreto nº 10.543, de 13 de novembro de 2020</u>.



A autenticidade deste documento pode ser conferida no site <u>https://sei.ufmg.br</u> /<u>sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0,</u> informando o código verificador **1217442** e o código CRC **9CFC37C8**.

Referência: Processo nº 23072.204948/2022-16

SEI nº 1217442

31/01/2022 11:07

Dedico essa dissertação a todos os professores que possibilitam a disseminação e perpetuação do conhecimento científico.

AGRADECIMENTOS

Agradeço primeiramente à minha família que me deu suporte financeiro e apoio emocional para manter meus estudos e que também foram em grande parte responsáveis pelo meu crescimento pessoal e profissional. E não menos importante às minhas cachorrinhas Nina e Pipoca que eu amo tanto e que sempre me fazem companhia, tornando meus dias mais leves e felizes.

Sou muito grata ao meu orientador Gustavo Kuhn que me proporcionou quase cinco anos de formação na área acadêmica e de ensinamentos que levarei para a vida. Nesse período que estive no laboratório de Citogenômica Evolutiva pude me desenvolver não só no aspecto acadêmico, mas também como pessoa. Agradeço à professora Marta Svartman pelas conversas descontraídas e por vários conselhos dados durante esse tempo no laboratório. Ao meu coorientador Leonardo Koerich, agradeço pelas valiosas sugestões para o melhor desenvolvimento da minha dissertação.

À banca examinadora por sua disposição em contribuir para o aprimoramento da minha dissertação e também para minha formação.

Agradeço a todos os funcionários da UFMG e do programa de Pós-Gradução em Genética, em especial à Tulia Marques, à Raíssa, à Vitória e à Dani que nos ajudam nos processos burocráticos.

Ao Daniel que sempre ajudou com a manutenção dos estoques e ao pessoal da limpeza por manterem os espaços da UFMG mais agradáveis.

Aos órgãos de fomento (CNPq, FAPEMIG, CAPES) por financiarem os projetos de pesquisa, em especial ao CNPq e à CAPES que possibilitaram minha permanência por meio das bolsas concedidas.

Aos meus amigos de laboratório Babau, Erick, Aninha, Pedro, Radinha, Licinha, Mi, Zé, Gui, Naiara, Lucas e Gustavinho. Sempre faço questão de dizer que foram essenciais para tornar o nosso ambiente de pesquisa mais agradável. Agradeço também, por terem sido em vários momentos professores aos quais recorri sempre que necessário. Essa galerinha também foi um dos motivos de eu ter certeza que escolhi o laboratório certo para fazer parte.

De novo ao Bráulio, Pedro e Rada agradeço pelas sugestões e discussões que contribuíram diretamente para a produção da minha dissertação.

À minha amiga Lorrayne que desde meu primeiro dia de aula da graduação esteve ao meu lado nos momentos alegres e difíceis.

A minha prima Izabela, que esteve muito presente alegrando meus dias durante esse período de pandemia que coincidiu com o meu mestrado.

A todos os professores que da pré-escola até o mestrado me permitiram, através do compartilhamento de seus conhecimentos, chegar até aqui.

Ao meu psicólogo João que nesses últimos dois anos vem cuidando da minha saúde mental e me ajudando a ter consciência dos fatores causadores de minhas inseguranças e crises de ansiedade para, então, superá-las.

À equipe do "Trem Bão é Ciência" que me faz refletir sobre a importância de devolver o conhecimento de direito à população brasileira, por meio da divulgação científica.

RESUMO

O centrômero é a região necessária para a segregação cromossômica nos eucariotos. Em animais e plantas, centrômeros funcionais se iniciam pela interação entre DNAs satélites e a proteína Cenp-A, que é uma variante da histona H3 (CenH3) exclusiva da cromatina centromérica. Uma outra proteína centromérica, a Cenp-C, se liga à Cenp-A e ao DNA centromérico (cenDNA), o que fornece a base para o recrutamento das demais proteínas do cinetócoro. Dada a importância funcional dos centrômeros, seria esperada uma alta conservação evolutiva de sua estrutura entre espécies. Porém, tanto o cenDNA como Cenp-A/Cenp-C evoluem rapidamente. De acordo com a hipótese do impulso centromérico, isso ocorre devido aos efeitos deletérios associados a expansões do cenDNA, e de seleção positiva em sítios de Cenp-A e Cenp-C que suprimem estes efeitos. Em espécies do subgênero Drosophila foram encontradas duplicações do gene da Cid (CenH3 de Drosophila) (Cid1/Cid6 e Cid5) e da Cenp-C (Cenp-C1 e Cenp-C2). De acordo com padrões diferenciais de expressão destes genes em diferentes tecidos, foi sugerido que Cid1 interage com Cenp-C1 e Cid5 com a Cenp-C2 para função centromérica. Em espécies do grupo montium (subgrupo Sophophora), foram observadas duplicações de Cid (Cid1, Cid3 e Cid4) e estes parálogos também revelaram expressão diferencial dependendo do tecido. O presente trabalho teve como objetivos principais investigar (i) se existem duplicações de Cenp-C nestas espécies e (ii) testar se existe seleção positiva em Cenp-C. Para isto, utilizamos genomas recentemente sequenciados de 23 espécies do grupo montium. Encontramos três novas cópias do gene da Cenp-C, que denominamos de Cenp-C3, Cenp-C4 e Cenp-C5. Porém, todas as espécies possuem apenas uma cópia de Cenp-C no mesmo genoma, com exceção de D. vulcana, que possui duas (Cenp-C1 e Cenp-C5). Esse resultado indica que apenas uma cópia da Cenp-C deve interagir com os três parálogos da Cid presentes nas espécies do grupo montium. Nossas análises mostraram que as três novas cópias se originaram de duplicações independentes de Cenp-C1, com posterior perda de Cenp-C1 na maioria das espécies. Encontramos assinaturas de seleção positiva em quatro aminoácidos da Cenp-C localizados em regiões inter-motivos da Cenp-C. Embora não saibamos quais as regiões da Cenp-C fazem contato direto com o DNA, é possível que a Cenp-C esteja evoluindo sob seleção positiva para suprimir o impulso centromérico. Em conjunto, nossos resultados ajudam a entender as consequências de duplicações gênicas para a estrutura e evolução dos centrômeros.

Palavras chave: Centrômero. Drosophila. Cid. Cenp-A. CenH3. Cenp-C. Duplicação gênica.

ABSTRACT

The centromere is the region necessary for chromosomal segregation in eukaryotes. In animals and plants, functional centromeres are initiated by the interaction between satellite DNAs and the Cenp-A protein, a variant of histone H3 (CenH3) unique to the centromeric chromatin. Another centromeric protein, Cenp-C, binds to Cenp-A and centromeric DNA (cenDNA), providing the basis for the recruitment of other kinetochore proteins. Given the functional importance of centromeres, high evolutionary conservation of their structure among species would be expected. However, both cenDNA and Cenp-A/Cenp-C evolve rapidly. According to the centromere drive hypothesis, this is due to the deleterious effects associated with cenDNA expansions, and positive selection at Cenp-A and Cenp-C sites that suppress these effects. In species of the subgenus Drosophila, duplications of the gene of Cid (CenH3 of Drosophila) (Cid1 or Cid6 and Cid5) and Cenp-C (Cenp-C1 and Cenp-C2) were found. According to differential expression patterns of these genes in different tissues, it was suggested that Cid1 interacts with Cenp-C1 and Cid5 with Cenp-C2 for centromeric function. In the *montium* group species (subgroup Sophophora), duplications of Cid (Cid1, Cid3, and Cid4) were observed and these paralogs also revealed differential expression depending on the tissue. This work had as main objectives to investigate (i) if there are duplications of Cenp-C in these species and (ii) to test if there is positive selection in Cenp-C. For this, we used recently sequenced genomes from 23 montium group species. We found three new copies of the Cenp-C gene, which we called Cenp-C3, Cenp-C4, and Cenp-C5. However, all species have only one copy of Cenp-C in the same genome, except for D. vulcana, which has two (Cenp-C1 and Cenp-C5). This result indicates that only one copy of Cenp-C should interact with the three Cid paralogs present in the montium group species. Our analyzes showed that the three new copies originated from independent duplications of Cenp-C1, with subsequent loss of Cenp-C1 in most species. We found positive selection signatures in four Cenp-C amino acids located in intermediate regions of Cenp-C. Although we do not know which regions of Cenp-C make direct contact with DNA, Cenp-C may be evolving under positive selection to suppress the centromeric drive. Taken together, our results help to understand the consequences of gene duplications for the structure and evolution of centromeres.

Keywords: Centromere. Drosophila. Cid. Cenp-A. CenH3. Cenp-C. Gene duplication.

LISTA DE FIGURAS

Figura 3. Filogenia contendo algumas das espécies do subgênero *Drosophila* e do subgênero *Sophophora* ao qual pertencem os grupos *montium* e *melanogaster* (Adaptado de Teixeira *et al.*

 Figura 6. Esquema representando a questão central do trabalho: quantas cópias da Cenp-C

 interagem com as três cópias da Cid nas espécies do grupo

 montium?

 .29

 Figura 11. Alinhamento dos dois fragmentos do gene da *Cenp-C5* presentes em dois *contigs* (1 e 2) e da sequência definida para o gene da *Cenp-C5* em *D. vulcana* (DvulCenpC5). As barras em cinza mostram sequências de íntrons inferidos pelo algoritmo *Augustus* e por alinhamento com os genes da Cenp-C das outras espécies do grupo *montium*. A barra em azul representa a região deletada de *Cenp-C4* para análises de seleção positiva e construção da

Figura 13. Disposição dos fragmentos do gene da *Cenp-C* de *D. burlai* e *D. punjabiensis* em relação aos 5 locos das cópias dos genes da *Cenp-C*, no cromossomo 3R de *D. triauraria*. As

setas preenchidas e pontilhadas indicam, respectivamente, a presença e ausência dos genes que podem estar inteiros ou fragmentados. A orientação das setas indica a orientação dos genes no cromossomo. O tamanho das seguências de DNA presente entre os locos está representado em preto acima das linhas contínuas. As linhas pontilhadas com setas nas extremidades (azul e verde) mostram localização dos locos indicados pela seta. A distância desses locos em relação ao gene mais próximo (CG14655 ou CG1427) está representada ao lado de cada linha pontilhada com a mesma cor da linha. Abaixo de cada fragmento do gene Cenp-C da é mostrado seu tamanho em pares de

Figura 17. Alinhamento das proteínas Cenp-C das espécies do grupo *montium* mostrando a posição dos motivos proteicos identificados no presente trabalho. As barras em vermelho

representam da esquerda para a direita: Mis12 *binding*, R-rich, DH, NLS, CenH3 *binding* e Cupin. A barra azul-claro representa as sequências removidas para a construção da filogenia dos genes e teste de seleção positiva. A pequena barra em verde-escuro abaixo do motivo CenH3-binding representa o aminoácido R-1101, presente em todas as cópias de Cenp-C do grupo

Figura 22. Árvore de máxima verossimilhança mostrando as relações filogenéticas entre as cópias do gene da *Cenp-C* nas espécies dos subgêneros *Sophophora* e *Drosophila*. O valor

de bootstrap está indicado em cada nó. A escala indica o número de substituições por sítio......52

Figura 24. (A) Representação do elemento TE Harbinger-2_DRh. Setas em preto indicam TIRs e retângulo em cinza indica a sequência interna às TIRs. (B) Evento de transposição por mecanismo cut and paste esperado para duplicação gênica. 1. Duas cópias do TE se insere em ambos os lados da Cenp-C1.2. Uma das TIRs de cada TE sofre mutações impedindo a ligação da transposase. Os elementos são reconhecidos pela mesma transposase e são transpostos para um segundo loco levando a sequência interna a eles. (C) Representação dos locos da Cenp-C1 e da Cenp-C5 com suas sequências flanqueadoras, observados no genoma de D. vulcana. Pontas de seta indicam a posição 3' dos genes. Fragmentos homólogos (65 a 74% de identidade e 54 a 152pb) ao Harbinger-2 DRh indicados pelo Repbase estão em cinza e preto. Os pontilhados delimitam a região de homologia (85 a 95% de identidade e 189 a 190pb) entre as três cópias do elemento Harbinger-2_DRh-like. Linhas pretas e barras em bege indicam, respectivamente, seguências intergênicas e seguências com homologia (211pb e 206pb) entre os dois locos. (D) Representação do que seria esperado para os locos da Cenp-C de D. vulcana, se um elemento Harbinger-2_DRh-like envolvido processo estivesse no de duplicação......56

LISTA DE TABELAS

Tabela 1. Números de acesso das sequências obtidas no Genbank														30	
Tabela	2.	Carac	teríst	ticas	gerais	s dos	gene	es da	a Cel	np-C	das	espéo	cies	do	grupo
montium)														39
Tabela	3.	Sítios	sob	seleç	ão po	ositiva	de a	acordo	com	0	modelo	de	sítios	s-ale	atórios
(CodeMI	_)														62

LISTA DE ABREVIATURAS

- **BEB –** Bayes Empirical Bayes
- **Cal1 –** Chromosome Alignment defect 1
- CCAN Constitutive Centromere-Associated Network
- **cDNA –** DNA complementar
- CDS Sequência codificadora
- CenH3 Histona Centromérica 3
- CenDNA DNA centromérico
- Cenp-A Proteína Centromérica A
- Cenp-C Proteína Centromérica C
- Cid Centromere Identifier
- DH Drosophilid Cenp-C Homolegues
- **DNA –** Ácido desoxirribonucleico
- dN Substituição não-sinônima
- dS Substituição sinônima
- GTR+G+I Generalised Time Reversible + Gamma + Invariable
- HKY+G+I Hasegawa-Kishino-Yano model + Gamma + Invariable
- H3 Histona 3
- HJURP Holliday Junction Recognition Protein
- Kb kilobases
- Mif2p Macrophage migration inhibitory factor-like protein
- KMN Knl1-Mis12-Ndc80
- LTR Repetição Longa Terminal
- Mis12C Complexo Mis12
- **mRNA –** RNA mensageiro
- MV Máxima Verossimilhança
- NCBI Basic Local Alignment Search Tool
- NLS Nuclear Localization Signal
- pb pares de bases
- **R1101 –** arginina 1101
- RNA Ácido ribonucleico
- RNA-Seq RNA Sequencing
- R-rich Ariginine-rich
- satDNA DNA satélite
- **TE –** Elemento Transponível
- TIR Repetição Invertida Terminal

SUMÁRIO

1.	INTRODUÇÃO
1.1.	O DNA centromérico20
1.2.	A função das proteínas Cenp-A e Cenp-C na manutenção do centrômero21
1.3.	A hipótese do impulso centromérico22
1.4.	O gênero <i>Drosophila</i> e as espécies do grupo <i>montium</i>
1.5.	Duplicações da Cenp-A e Cenp-C em espécies do gênero Drosophila
2.	OBJETIVO
2.1.	Objetivos específicos29
3.	METODOLOGIA
3.1.	Identificação das cópias de Cenp-C no genoma das espécies de Drosophila
3.2.	Árvores filogenéticas
3.3.	Análise dos motivos da proteína Cenp-C 32
3.4.	Análises de seleção positiva33
4.	RESULTADOS E DISCUSSÃO
4.1.	Descoberta de três novas cópias do gene da Cenp-C no grupo montium
4.2.	Identificação dos motivos proteicos conservados em cópias da Cenp-C nas espécies
do g	grupo <i>montium</i>
4.3.	As duas cópias da <i>Cenp-C</i> no genoma de <i>D. vulcana</i> 48
4.4.	Relações evolutivas entre as cópias da Cenp-C nas espécies do grupo montium 49
4.5.	Origem das duplicações do gene da Cenp-C nas espécies do grupo montium 53
4.6.	Teste de seleção positiva nas cópias de Cenp-C57
5.	CONCLUSÕES
6.	REFERÊNCIAS BIBLIOGRÁFICAS

1. INTRODUÇÃO

1.1. O DNA centromérico

O centrômero é o loco cromossômico responsável pela segregação dos cromossomos durante a divisão celular em eucariotos. Apesar dessa função importante, os centrômeros são altamente diversos em sequência, organização e tamanho (Talbert & Henikoff 2020). Eles podem ser: (i) pontuais, como em *Saccharomyces cerevisiae*, onde são compostos por sequências curtas e conservadas de aproximadamente 125 pb; (ii) regionais, como em plantas e animais, onde podem atingir centenas de kilobases ou (iii) dispersos ao longo de toda a extensão do cromossomo, como nos nematódeos (Malik & Henikoff 2002). Dentre esses três tipos de centrômeros, o que mais predomina entre eucariotos é o regional. Estes centrômeros são constituídos principalmente por duas classes de DNAs repetitivos: os DNAs satélites (satDNAs) e os elementos de transposição (TEs) (Plohl *et al.* 2014).

Os DNAs satélites são os componentes mais comumente encontrados no centrômero da maioria dos animais e plantas estudados. Trata-se de sequências de DNA repetidas em tandem, que formam tipicamente cadeias de centenas de kilobases de comprimento (Melters *et al.* 2013). Embora a presença de satDNAs nos centrômeros seja uma característica comum a vários táxons de eucariotos, essas sequências são altamente variáveis entre espécies, podendo até ser espécie-específicas. Vários mecanismos foram propostos para explicar essa observação, como: *crossing-over* desigual, conversão gênica e amplificação via círculo rolante com subsequente reinserção. Tais mecanismos são os prováveis responsáveis pela homogeneização nucleotídica entre as cópias e pelas alterações de tamanho das cadeias de cópias em tandem (Hartley & O'Neill 2019).

Outro componente repetitivo que pode ser encontrado no centrômero são os elementos de transposição. Essas sequências são capazes de se moverem ao longo do genoma e por isso estão frequentemente dispersas. Os TEs são classificados conforme seu mecanismo de transposição, podendo ser de Classe I, os retrotransposons, quando se transpõem por intermédio de um RNA, ou de classe II, os transposons de DNA, quando se transpõem por intermédio de um DNA (Wicker *et al.* 2007). Apesar de serem caracterizados pela capacidade de se transpor, a maioria desses elementos não se move ativamente, seja devido a mutações ou silenciamento epigenético (Hartley & O'Neill 2019).

A abundante presença de satélites na maioria dos centrômeros é um indicativo de que essas sequências são importantes para a expansão e estabilização dessas regiões (Plohl *et*

al. 2008). No entanto, a formação de neocentrômeros em regiões desprovidas de DNAs repetitivos indica que essas sequências não são essenciais para os estágios iniciais de formação dos centrômeros (Henikoff *et al.* 2001). O que determina um centrômero funcional é a presença de nucleossomos contendo a histona H3 variante CenH3, o que torna a cromatina do DNA centromérico (cenDNA) distinta de outras partes do genoma (Brown & O'Neill 2014; Dalal *et al.* 2007).

1.2. A função das proteínas Cenp-A e Cenp-C na manutenção do centrômero

A grande diversidade dos centrômeros com relação a sequência de nucleotídeos torna improvável sua identificação por um fator genético universal. No entanto, esse loco é determinado epigeneticamente pela histona CenH3, também conhecida como Cenp-A. Consistente com a importância da Cenp-A na função centromérica é a sua ubiquidade nos centrômeros funcionais, presença em neocentrômeros e sua ausência em centrômeros inativos de cromossomos dicêntricos (Rosin & Mellone 2017). Além disso, a presença de Cenp-A no centrômero é o primeiro passo para a localização dos componentes do cinetócoro (Mckinley & Cheeseman 2016). O cinetócoro é um complexo multiprotéico formado nos centrômeros que fornece a estrutura necessária para que os microtúbulos se conectem aos cromossomos.

Uma segunda proteína associada à centrômeros funcionais é a Cenp-C, um dos componentes do CCAN (*constitutive centromere-associated network*), que são caracterizados por se localizarem no centrômero durante todo o ciclo celular (Maresca 2011). Em vertebrados o CCAN é composto por 16 proteínas (Maresca 2011) enquanto que em *Drosophila* e *Caenorhabditis elegans*, apenas a Cenp-C foi identificada (Liu *et al.* 2016). Essa menor complexidade dos centrômeros de *Drosophila* e *C. elegans*, quando comparado com os dos demais organismos, torna ambos excelentes modelos para compreender a evolução dos componentes centroméricos. De fato, apenas as proteínas Cid (Homóloga da Cenp-A em *Drosophila*), Cenp-C e Cal1, a chaperona que medeia a deposição da Cid na cromatina, são suficientes para a propagação da Cid no centrômero de *Drosophila* (Roure *et al.* 2019). Em humanos, por exemplo, a propagação da Cenp-A requer além da chaperona HJURP e da Cenp-C, o Complexo Mis18 que é composto por três proteínas (Mckinley & Cheeseman 2016).

Foi demonstrado que a Cenp-C se liga diretamente à Cenp-A (Carrol *et al.* 2010; Kato *et al.* 2013; Musacchio & Desai 2017; Roure *et al.* 2019) e ao DNA centromérico (Cohen *et al.* 2008; Politi *et al.* 2002; Trazzi *et al.* 2002). A Cenp-C é também responsável por se ligar ao

complexo Mis12C, um grupo de proteínas que é parte de um complexo maior, o *KMN network*, composto pelos complexos knl1, Mis12 e Ndc80. É no *KMN* que os microtúbulos se conectam aos cromossomos nas divisões celulares (Figura 1) (Maresca 2011; Przewloka & Glover 2009). A Cenp-C é, portanto, uma proteína chave na ligação entre o cinetócoro e a cromatina centromérica.



Figura 1. Esquema do cinetócoro de Drosophila (Adaptado de Przewloka et al. 2007).

Estudos mostraram que a Cenp-A e a Cenp-C são funcionalmente dependentes entre si para manutenção dos centrômeros (Erhardt *et al.* 2008; Falk *et al.* 2015; Moree *et al.* 2011; Orr & Sunkel, 2011; Westhorpe *et al.* 2015). Concordante com isso, um padrão de retenção ou perda de ambas proteínas foi observado em várias linhagens de eucariotos (Hooff *et al.* 2017). Tais achados tornam evidente o papel da Cenp-C e Cenp-A na manutenção de centrômeros funcionais.

1.3. A hipótese do impulso centromérico

O funcionamento adequado do centrômero é essencial para que ocorra a correta segregação dos cromossomos. Concordante com isso, quebras de DNA, rearranjos e aberrações estruturais em regiões centroméricas são comumente observadas em células cancerígenas e em algumas síndromes genéticas (Barra & Fachinetti 2018). Por esse motivo, se esperaria um alto grau de conservação das proteínas centroméricas, Cenp-A e Cenp-C, e do DNA centromérico entre espécies. No entanto, o que se tem observado é que ambos

evoluem rapidamente em animais e plantas (Plohl *et al.* 2008; Talbert *et al.* 2004). Um exemplo dessa rápida evolução é a sequência de nucleotídeos da Cenp-C, que apresenta menos de 5% de conservação entre filos, e que devido à essa grande divergência, pensou-se até estar ausente em *Drosophila* (Talbert *et al.* 2004). Essa maneira paradoxal pela qual o centrômero evolui (i.e. função essencial mas rápida evolução), é explicada pela hipótese do impulso centromérico ou "*centromere drive*" (Figura 2) (Dawe & Henikoff 2006; Henikoff *et al.* 2001).

Na maioria das espécies de animais e plantas, apenas um dos quatro produtos da meiose da fêmea se desenvolve em ovócito. Assim, conforme proposto pela hipótese do impulso centromérico, essa assimetria meiótica das fêmeas possibilita a competição entre cromossomos homólogos para sua inclusão no ovócito. Centrômeros contendo um DNA mais expandido poderão recrutar mais proteínas, como Cenp-A ou Cenp-C, e consequentemente mais microtúbulos. Dessa forma, os cromossomos contendo esses centrômeros expandidos serão preferencialmente transmitidos, em relação aos seus homólogos, para as próximas gerações. Entretanto, existem duas possíveis consequências negativas associadas ao impulso centromérico: (i) a transmissão de mutações deletérias ligadas a esses cromossomos na população (ii) a não disjunção dos cromossomos sexuais em machos e aumento da infertilidade. Mutações nas proteínas centroméricas (Cenp-A e Cenp-C) que reduzam sua afinidade pelo centrômero expandido ou que aumentem a sua afinidade pelo centrômero mais "fraco" poderão suprimir o impulso centromérico. A ocorrência desse processo em ciclos ao longo da evolução pode explicar o padrão atípico de evolução dos centrômeros (Dawe & Henikoff 2006; Henikoff *et al.* 2001; Malik & Bayes 2006).



Figura 2. Modelo do impulso centromérico. A expansão do cenDNA em um dos cromossomos homólogos faz com que mais proteínas centroméricas e microtúbulos sejam recrutados, aumentando a transmissão desse homólogo na meiose feminina. Esse cromossomo aumenta em frequência na

população e a consequência para machos é a não disjunção e infertilidade. Mutações nas proteínas centroméricas que reduzam a afinidade das mesmas pelo cenDNA restaura a paridade entre os cromossomos (Adaptado de Rosin & Mellone 2017).

De acordo com o previsto pela hipótese do impulso centromérico, tanto em *Drosophila* como em *Arabidopisis* foi detectada seleção positiva em regiões da Cenp-A preditas como locais de contato com o DNA do centrômero (Cooper & Henikoff 2004; Malik & Henikoff 2001; Malik *et al.* 2002). Por outro lado, em gramíneas e mamíferos, é na Cenp-C, e não na Cenp-A, onde evidências de seleção positiva foram encontradas. Em mamíferos, regiões sob seleção positiva na Cenp-C incluem aquelas em contato com o DNA centromérico. Já em leveduras, a proteína Mif2p (homóloga à Cenp-C) evolui sob seleção negativa. No entanto, esse é um resultado esperado conforme a hipótese do impulso centromérico, já que leveduras possuem centrômeros com sequência conservada e meiose feminina simétrica (Talbert 2004).

1.4. O gênero Drosophila e as espécies do grupo montium

O gênero *Drosophila* é composto por pequenas moscas, também conhecidas como moscas das frutas, pertences à ordem Diptera e à família Drosophilidae (Hales *et al.* 2015). Esse grupo é composto por mais de 1.600 espécies de *Drosophila* (O'Grady & Desalle 2018). Foi sugerido que o gênero *Drosophila* seja parafilético em relação a vários gêneros, e por isso, as relações filogenéticas entre as espécies do grupo se encontram em debate entre os taxonomistas (O'Grady & Desalle 2018).

O gênero Drosophila apresenta uma extensa literatura em diversas áreas da genética (Markow & O'Grady 2006). De fato, é enorme a lista de trabalhos feitos em Drosophila que serviram e continuam servindo como base para um melhor entendimento da genética como um todo (Ashburner & Bergman 2005; Griffths et al. 2000; Powel 1997). Atualmente, o gênero Drosophila contém genomas sequenciados e montados de mais de 120 espécies, de acordo com 0 National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/drosophila). Esta vasta quantidade de genomas disponíveis possibilita uma enorme quantidade de estudos sobre a evolução de genes e genomas.

Dentre os subgêneros pertencentes ao gênero *Drosophila*, estão os subgêneros *Drosophila* e *Sophophora*, que concentram a maioria das espécies de drosofilídeos estudada no contexto da evolução do centrômero (Kursel & Malik 2017; Malik & Henikiff 2001; Malik *et* al. 2002; Roure et al. 2019; Teixeira et al. 2018). Estima-se que esses subgêneros se divergiram entre 40-62.9 milhões de anos atrás (Russo et al. 1995; Tamura et al. 2004).

O subgênero Sophophora possui 344 espécies (O'Grady & Desalle 2018) divididas em 10 grupos (Yassin 2013). O maior deles é o grupo *montium*, composto por 94 espécies oriundas da Ásia, Australásia e África (Yassin 2018). Inicialmente, esse clado foi considerado como sendo um subgrupo dentro do grupo *melanogaster*. Posteriormente, foi elevado ao nível de grupo dentro do subgênero Sophophora (Figura 3) (Da Lage *et al.* 2007).



Figura 3. Filogenia contendo algumas das espécies do subgênero *Drosophila* e do subgênero *Sophophora* ao qual pertencem os grupos *montium* e *melanogaster* (Adaptado de Teixeira *et al.* 2018).

Com base em dados corológicos e morfológicos (genitália e pigmentação abdominal de machos), as espécies do grupo *montium* foram subdivididas em sete subgrupos (Yassin 2018): *montium*, *parvula*, *orosa*, *kikkawai*, *serrata*, *punjabiensis* e *seguyi*. Duas filogenias recentes propostas para as espécies do grupo *montium* confirmam a monofilia de todos os subgrupos, mas divergem quanto as relações filogenéticas entre os subgrupos (Conner *et al.* 2021; Yassin *et al.* 2016) (Figura 4). Enquanto Yassin *et al.* (2016) consideram o grupo *parvula*

como o mais basal da filogenia, em Conner *et al.* (2021) essa posição é ocupada pelo grupo *montium*. Além disso, na filogenia de Conner *et al.* (2021) o subgrupo *kikkawai* forma um grupo irmão do clado formado pelos subgrupos *serrata, punjabiensis* e *seguyi*. Já em Yassin *et al.* (2016), o grupo *punjabiensis* é o grupo irmão do clado formado pelos subgrupos *serrata, kikkawai* e *seguyi*. Outro ponto importante de ser ressaltado é que o estudo de Conner *et al.* (2021) produziu duas árvores que se diferem quanto ao posicionamento do subgrupo *punjabiensis*. Na primeira, este subgrupo é irmão do subgrupo *seguyi*, já na segunda o grupo *punjabiensis* é o grupo irmão do clado formado pelos subgrupos *serrata* e *seguyi*.

A filogenia de Conner *et al.* (2021) foi construída a partir de 60 genes ortólogos e 42 espécies, contrastando com a de Yassin *et al.* (2016) que utilizou 44 espécies, mas apenas três genes nucleares e um mitocondrial. Embora haja discrepâncias quanto a relação dos subgrupos dentro do grupo *montium*, não há dúvidas quanto a sua monofilia (Da Lage *et al.* 2007; Finet *et al.* 2021; Russo *et al.* 2013; Yang *et al.* 2012). Por isso, o grupo *montium* é considerado um ótimo modelo para estudos de evolução.



Figura 4. Filogenias do grupo *montium*. A. Filogenia proposta por Conner *et al.* (2021) com base em análise Bayesiana e de máximo verossimilhança. B. Filogenia proposta por Yassin *et al.* (2016) com base em análise Bayesiana.

1.5. Duplicações da Cenp-A e Cenp-C em espécies do gênero Drosophila

A homologia entre genes pode se dar por sua origem a partir de um único gene ancestral pela especiação, os chamados genes ortólogos, ou por duplicação gênica, os chamados genes parálogos (Koonin 2005). Genes oriundos de duplicações podem ter quatro destinos ao longo da evolução. O mais comum é a pseudogenização ou eliminação de uma das cópias. Outra possibilidade é a neofuncionalização, um processo mais raro por depender de mutações adaptativas que resultarão em uma nova função para uma das cópias. Uma terceira possibilidade é que o relaxamento da seleção negativa sob os parálogos pode levar à rápida evolução de ambos e à divisão de diferentes funções do gene ancestral entre as cópias, processo denominado de subfuncionalização (Koonin 2005; Rastogi & Liberles 2005). E por último, a retenção de genes duplicados pode ser, às vezes, benéfica por implicar em maiores quantidades de proteína como produto. Nesse caso, esses genes podem ser mantidos por seleção negativa (Zhang 2003).

Duplicações do gene da *Cid* foram identificadas em espécies do gênero *Drosophila* (Kursel & Malik 2017). Essas duplicações estão presentes em *D. eugracilis* do grupo *melanogaster* (*Cid1* e *Cid2*) e nas espécies do grupo *montium* (*Cid1*, *Cid3* e *Cid4*) pertencentes ao subgênero *Sophophora*. Duplicações também foram observadas em espécies do subgênero *Drosophila* (*Cid1* e *Cid5*). Nesse estudo, os parálogos *Cid1*, *Cid3*, *Cid4* e *Cid5* foram expressos em cultura de tecidos e todos eles se localizaram no centrômero, em suas respectivas espécies. Foi demonstrado que a *Cid3* e a *Cid5* apresentam expressão restrita à linhagem germinativa do macho. Por outro lado, a *Cid1* e a *Cid4* são expressas tanto na linhagem germinativa quanto na linhagem somática. Além disso, nas espécies do grupo *montium*, a *Cid3* evolui sob seleção positiva em domínios de possíveis contato com o DNA centromérico. Porém, não foram encontradas evidências de seleção positiva para a *Cid5*. Foi levantada a hipótese de que a *Cid3* e, possivelmente, a *Cid5* podem estar atuando como supressores do impulso centromérico na linhagem germinativa do macho (Kursel & Malik 2017).

Recentemente, nosso grupo descreveu em espécies do subgênero *Drosophila* uma nova duplicação da *Cid* (*Cid6*) e também duplicações da *Cenp-C* (*Cenp-C1* e *Cenp-C2*) (Teixeira *et al.* 2018). Em todas as espécies estudadas duas cópias da *Cid* estão presentes (*Cid1* com Cid5 ou *Cid6* com *Cid5*). Essas duplicações da *Cid* e *Cenp-C* vêm sendo mantidas há pelo menos 50 milhões de anos nas espécies do subgênero *Drosophila*. Nessas espécies, as duplicações da *Cenp-C* apresentam perfis alternados de retenção de motivos proteicos,

possivelmente indicando subfuncionalização. Porém, ambos os parálogos mantém motivos essenciais para localização e função centromérica.

Teixeira *et al.* (2018) relataram que a *Cid6*, que substitui funcionalmente a *Cid1* em algumas espécies, estava expressa em todas as fases do desenvolvimento (embriões, larvas, pupas, adultos machos e adultos fêmeas) analisadas. Já a *Cid5* estava expressa apenas em machos adultos e pupas (que provavelmente darão origem a machos). A *Cid1* foi analisada quanto ao nível de expressão em testículos e se mostrou praticamente silenciada, enquanto que a *Cid5* apresentou altos níveis de expressão. Ao contrário dos parálogos de *Cid*, a *Cenp-C1* e *Cenp-C2* mostraram expressão em quase todas as fases do desenvolvimento, com exceção de larvas. Porém, a *Cenp-C1* foi mais expressa em fêmeas e embriões do que *Cenp-C2*, e esta última mais expressa em adultos machos e pupas do que *Cenp-C1*. Em testículos a expressão de *Cenp-C2* também foi estatisticamente mais alta do que a de *Cenp-C1*. Devido à essas observações, foi sugerido que Cenp-C2 interage com Cid5 e Cenp-C1 interage com Cid1/Cid6 para função centromérica (Figura 5) (Teixeira *et al.* 2018).



Figura 5. Hipótese de interação entre as cópias da Cid e da Cenp-C nas espécies do subgênero *Drosophila* de acordo com Teixeira *et al.* (2018).

A hipótese da interação entre as cópias da Cid e da Cenp-C em espécies do subgênero *Drosophila* levantou a questão de quantas cópias da Cenp-C estariam interagindo com as três cópias da Cid nas espécies do grupo *montium* (Figura 6). Nesse grupo, a única espécie do grupo *montium* estudada, *D. kikkawai*, revelou possuir apenas uma única cópia da *Cenp-C* (Teixeira *et al.* 2018). Por isso, o presente trabalho buscou realizar um estudo mais completo sobre a evolução do gene da *Cenp-C* nas espécies do grupo *montium*, beneficiado pelo recente sequenciamento dos genomas de 23 espécies deste grupo (Bronski *et al.* 2020). A análise da *Cenp-C* em um número maior de espécies poderá ajudar na compreensão da evolução da proteína Cenp-C e consequentemente da evolução do centrômero em *Drosophila* e em outros organismos.



Figura 6. Esquema representando uma questão central do presente trabalho: quantas cópias da Cenp-C interagem com as três cópias da Cid encontradas nas espécies do grupo *montium*?

2. OBJETIVO

Investigar a evolução do gene da proteína centromérica *Cenp-C* em espécies de *Drosophila* do grupo *montium*. Em especial, queremos investigar se existem duplicações de Cenp-C nestas espécies que possam interagir funcionalmente com as duplicações de Cid já relatadas nestas espécies e se existem assinaturas de seleção positiva atuando em *Cenp-C*.

2.1. Objetivos específicos

- 2.1.1. Identificar o gene da proteína centromérica *Cenp-C* nas espécies do grupo *montium*.
- 2.1.2. Investigar se existem duplicações de Cenp-C nas espécies estudadas.
- 2.1.3. Identificar e caracterizar motivos proteicos conservados da proteína *Cenp-C* nas espécies do grupo *montium*.
- 2.1.4. Analisar as relações evolutivas entre as cópias de Cenp-C encontradas.
- 2.1.5. Identificar quais foram os mecanismos envolvidos nas duplicações da Cenp-C.
- 2.1.6. Testar se existe seleção positiva atuando na Cenp-C.

3. METODOLOGIA

3.1. Identificação das cópias de *Cenp-C* no genoma das espécies de *Drosophila*

Foi feita uma busca por tblastn utilizando a proteína Cenp-C de *D. melanogaster* para obtenção, no banco de dados do NCBI, de mRNAs da *Cenp-C* nas espécies do gênero *Drosophila*. As sequências das proteínas codificadas pelos mRNAs retornados eram utilizadas para fazer a busca dos genes de *Cenp-C* no genoma da respectiva espécie. Essas sequências também foram utilizadas para busca em espécies sem mRNA disponível, de acordo com a proximidade filogenética (Tabela 1).

Para determinar a região codificante do gene da *Cenp-C*, foi utilizado o algoritmo preditor de gene *Augustus* (Stanke & Morgenstern 2005). A sequência codificante da *Cenp-C* retornada pelo *Augustus* era validada com mRNA ou RNA-Seq. Essa validação só foi possível para três espécies do grupo *montium* (*D. serrata*, *D. kikkawai* e *D. triauraria*) em que sequências de mRNA ou reads de RNA-Seq estavam disponíveis (Tabela 1). Por isso, regiões com discordância quanto a presença de éxons, entre espécies do grupo *montium*, foram eliminadas de análises de seleção positiva e para construção de árvores filogenéticas (

Figura 7). O genoma de *D. melanogaster*, que contém o maior número de genes anotados em *Drosophila*, foi utilizado como referência para inferências de sintenia. Isso foi feito através da análise da presença ou não dos genes flanqueadores da *Cenp-C* esperados.

Gene	ID do scaffold Genbank ou reads do genoma	ID do mRNA ou RNAseq do Genbank
D. pectinifera Cenp-C3	VNKC01003598.1	Indisponível
D. triauraria Cenp-C3	JABJVT010000005.1	SRR11780982
D. auraria Cenp-C3	VNJW01009994.1	Indisponível
D. tani Cenp-C3	VNJO01011109.1	Indisponível
D. rufa Cenp-C3	VNKH01000388.1	Indisponível
D. lateicornis Cenp-C3	VNKF01009708.1	Indisponível
D. asahinai Cenp-C3	VNJZ01004887.1	Indisponível
D. kanapiae Cenp-C4	VNJM01000860.1/	Indisponível
	VNJM01001987.1	
D. kikkawai Cenp-C1	AFFH02006098.1	XM_017182092.1
D. leontia Cenp-C1	VNKB01009439.1	Indisponível
D. bocki Cenp-C1	VNJY01005756.1	Indisponível
D. serrata Cenp-C1	MTTC01001254.1	XM_020957133.1
D. bunnanda Cenp-C1	VNKE01001283.1	Indisponível

 Tabela 1. Números de acesso das sequências obtidas no Genbank

D. truncata Cenp-C1	VNJQ01004447.1	Indisponível
D. mayri Cenp-C1	VNJN01008118.1	Indisponível
D. birchii Cenp-C1	VNKA01006487.1	Indisponível
D. watanabei Cenp-C1	VNJS01015544.1	Indisponível
D. punjabiensis Cenp-C1	VNJR01008397.1	Indisponível
D. jambulina Cenp-C1	VNJX01008392.1	Indisponível
D. seguyi Cenp-C1	VNJU01009249.1	Indisponível
D. vulcana Cenp-C1	VNJP01006877.1	Indisponível
D vulcana Conn CE	VNJP01005529.1/	Indisponível
D. vuicana Cenp-C5	VNJP01004469.1	
D. bakoue Cenp-C1	VNJL01004438.1	Indisponível
D. nikananu Cenp-C1	VNJV01009586.1	Indisponível
D. melanogaster Cenp-C1	NT_033777.3	NM_169228.3
D. simulnas Cenp-C1	NIFY0100004.1	XM_016174981.1
D. secheilia Cenp-C1	NIFZ0100004.1	XM_032720766.1
D. yakuba Cenp-C1	JAEDAC01000005.1	XM_002096785.2
D. erecta Cenp-C1	QMER02000001.1	XM_026983729.1
D. eugracilis Cenp-C1	AFPQ02005741.1	SRR346729
D. ananassae Cenp-C1	JACRYV010000154.1	XM_001955206.3
D. persimilis Cenp-C1	QMET02000001.1	XM_026986294.1
D. pseudoobscura Cenp-C1	WVEN01000001.1	XM_015182067.2
D. willistoni Cenp-C1	AAQB01009414.1	XM_023180684.1
D. mojavensis Cenp-C1	CH933806.1	SRR6968126
D. mojavensis Cenp-C2	CH933806.1	SRR6968126
D. arizonae Cenp-C1	SRR2070760	SRR2509638
D. arizonae Cenp-C2	SRR2070760	SRR2509638
D. navojoa Cenp-C1	LSRL02000055.1	-
D. navojoa Cenp-C2	LSRL02000228.1	XM_018113591.2
	https://dbuz.uab.cat/blast.php	SRR5145562/
D. buzzatii Cenp-C1	D. buzzatii Freeze 1	SRR5145563
	Scaffolds	
	https://dbuz.uab.cat/blast.php	SRR5145562/
D. buzzatii Cenp-C2	D. buzzatii Freeze 1	SRR5145563
	Scaffolds	
D. seriema Cenp-C1	ERR1976657	-
D. seriema Cenp-C2	ERR1976657	-
D. virilis Cenp-C1	QMEO02000199.1	XM_002056576.3
D. virilis Cenp-C2	QMEO02000199.1	XM_002056451.3
D. americana Cenp-C1	UEJX01001328.1	SRR5279019
D. americana Cenp-C2	UEJX01001683.1	SRR5279019
D. grimshawi Cenp-C1	AAPT01020190.1	XM_001994049.2
D. grimshawi Cenp-C2	AAPT01019320.1	XM_001989754.3
D. busckii Cenp-C1	DULD0100002.1	XM_017991761.1
D. busckii Cenp-C2	DULD0100002.1	XM_017992885.2
Phortica variegata	JXPM01003917.1	SRR1738675



DpecCenpC3 DtriCenpC3 DaurCenpC3 DtanCenpC3 DrufCenpC3 DlacCenpC3 DasaCenpC3 DkanCenpC4 DkikCenpC1 DleoCenpC1 DbocCenpC1 DsetCenpC1 DbunCenpC1 DtruCenpC1 DmayCenpC1 DbirCenpC1 DwatCenpC1 DpunCenpC1 DjamCenpC1 DsegCenpC1 DvulCenpC1 DvulCenpC5

Figura 7. Representação dos genes das espécies do grupo *montium*. Linhas em preto representam os íntrons. Barras coloridas representam os éxons, em que aquelas com cores iguais possuem homologia de sequência. Os éxons em amarelo, rosa escuro, rosa claro e azul turquesa correspondem às sequências retiradas das análises.

3.2. Árvores filogenéticas

Os códons da *Cenp-C* das espécies de *Drosophila* foram alinhados pelo algoritmo *MUSCLE* implementado no software *Geneious* e o alinhamento resultante foi refinado manualmente quando necessário. As árvores filogenéticas foram construídas pelo método de máxima verossimilhança (MV), implementado no software *MEGAX* (*Kumar et al.* 2018). Foi utilizado o modelo de substituição de nucleotídeos GTR+G+I e número de réplicas de *bootstrap* de 1000.

3.3. Análise dos motivos da proteína Cenp-C

Sete motivos na Cenp-C de *D. melanogaster* (Figura 8), caracterizados por Heeger *et al.* (2005), foram utilizados para busca em espécies de *Drosophila* do grupo *montium*. Para isso, as sequências dos aminoácidos da Cenp-C1 de *D. melanogaster* e de mais 9 espécies do grupo *melanogaster* foram alinhadas pelo algoritmo *MUSCLE* implementado no software *Geneious*. As regiões das sequências correspondentes a cada motivo de *D. melanogaster* no alinhamento foram identificadas e extraídas. Para garantir que só motivos conservados seriam utilizados na análise, foram considerados apenas aqueles com identidade mínima de 60%, quando comparado com o mesmo motivo de todas as outras espécies do grupo *melanogaster*. O algoritmo gerador de motivos *MEME* (Bailey *et al.* 2015) foi utilizado para produzir uma matriz que indica a possibilidade de cada aminoácido no motivo. As sequências apresentando deleções ou inserções foram retiradas da análise, pois essas características não são considerados pelo *MEME*. A matriz de possibilidades foi necessária para a busca dos motivos, nas espécies do grupo *montium*, através do algoritmo *MAST* (Bailey & Gribskov 1998). Só foram considerados os motivos que apresentassem *p*-value < 10⁻⁶. As sequências retornadas pelo *MAST* foram utilizadas em conjunto com as sequências da primeira busca para produzir

uma matriz de possibilidades e iniciar novas buscas. Novamente, apenas sequências com p-value < 10⁻⁶, foram consideradas.



Figura 8. Motivos funcionais da sequência da proteína Cenp-C em *Drosophila* identificados por Heeger *et al.* (2005). N e C indicam a região N-terminal e C-terminal da proteína, respectivamente.

3.4. Análises de seleção positiva

A razão entre substituições não-sinônimas e substituições sinônimas (dN/dS) ou omega (ω) permite inferir de que maneira um gene codificador de proteína evolui. Um valor de ω maior que 1 indica que o gene evoluí sob seleção positiva, um valor de ω igual a 1 indica que o gene evolui de maneira neutra e, por último, um valor de ω menor que 1 indica que o gene evoluí sob seleção negativa.

Para determinar se as cópias de *Cenp-C* evoluem sob seleção positiva, foi utilizado o modelo NSsites implementado no programa CodeML do pacote de programas *PAMLX* versão 1.3.1 (Xu & Yang, 2013). Dos modelos utilizados, em três (M1a, M7 e M8a) a razão entre substituições não-sinônimas e substituições sinônimas (dN/dS) não pode exceder 1 enquanto que em dois (M2a e M8) dN/dS é igual ou maior que 1. A análise de seleção positiva é feita comparando os modelos M1a e M2a, M7 e M8 e M8a e M8, sendo o primeiro par de modelos considerado o mais estringente. Sítios sob seleção positiva foram inferidos quando o teste da razão de verossimilhança era significativo e a probabilidade a posteriori BEB (*Bayes Empirical Bayes*) era maior que 80%.

O programa CodeML considera *gaps* como um dado inexistente para os cálculos estatísticos. Por isso foi feito um alinhamento de códons pelo algoritmo *MUSCLE* implementado no software *Geneious* e apenas os sítios em que até duas sequências apresentavam *gaps* foram mantidos. O alinhamento foi refinado manualmente e as regiões fracamente alinhadas foram deletadas. A árvore filogenética foi construída pelo método de máxima verossimilhança (MV), implementado no software *MEGAX* (Kumar *et al.* 2018). Foi utilizado o modelo de substituição de nucleotídeos HKY+G+I e número de réplicas de *bootstrap* de 1000. A árvore e o alinhamento foram utilizados como *input* no programa CodeML para o teste do modelo NSsites.

4. RESULTADOS E DISCUSSÃO

4.1. Descoberta de três novas cópias do gene da Cenp-C no grupo montium

Inicialmente, o gene da *Cenp-C* foi estudado em espécies do subgênero *Sophophora*, onde apenas uma cópia (*Cenp-C1*) foi encontrada e em espécies do subgênero *Drosophila*, onde duas cópias (*Cenp-C1* e *Cenp-C2*) foram encontradas (Teixeira *et al.* 2018). No primeiro subgrupo, a *Cenp-C1* está sempre flanqueada pelo gene *5-HT2B*, enquanto que no segundo subgrupo, a *Cenp-C1* está flanqueada pelos genes *CG1427* e *5-HT2B* e a *Cenp-C2* está flanqueada pelos genes *CG1427* e *5-HT2B* e a *Cenp-C2* está flanqueada pelos genes *CLS* e *RpL27* (Figura 9) (Teixeira *et al.* 2018).



Figura 9. Filogenia das espécies de *Drosophila* do grupo *montium* (Conner *et al.* 2021). Para referência, também foram incluídas *D. melanogaster* do subgênero *Sophophora* e *D. virilis* do subgênero *Drosophila*, (Russo *et al.* 2013). As setas preenchidas e pontilhadas indicam respectivamente a presença e ausência dos genes que podem estar inteiros ou fragmentados. A orientação das setas indica a orientação dos genes. As linhas contínuas indicam presença de sequências intergênicas e as linhas pontilhadas indicam presença de sequências intergênicas e as linhas pontilhadas indicam presença de sequências intergênicas fragmentos não montados.

No presente trabalho, encontramos apenas uma cópia do gene *Cenp-C, Cenp-C1*, nas espécies dos subgrupos *kikkawai, serrata, punjabiensis* e *seguyi*, com exceção de *D. burlai* e *D. vulcana* (Figura 9). Nestes subgrupos, *Cenp-C1* encontra-se na maioria das vezes flanqueado pelos genes *CG1427* (como nas espécies do subgênero *Drosophila*) e *Tim10*. Mas nas espécies do subgrupo *kikkawai*, a *Cenp-C1* é flanqueada pelo gene *OSA* ao invés do *Tim10*. Como esperado, por pertenceram ao subgênero *Sophophora*, em algumas espécies do grupo *montium* o gene *5-HT2B* também foi encontrado próximo ao loco da *Cenp-C1*. Porém, a maioria dos *contigs* analisados não possuía tamanho suficiente para cobrir a localização de *5-HT2B* (Figura 9). Supreendentemente, nas duas espécies do subgrupo *punjabiensis* analisadas, a *Cenp-C1* se encontra em orientação invertida no loco. A causa desta inversão ainda deverá ser melhor investigada.

No subgrupo *montium*, encontramos uma nova cópia da *Cenp-C*, exclusiva deste subgrupo, encontrada em um loco flanqueado pelos genes *CG7785* e *DNAseII*. Denominamos está cópia de *Cenp-C3* (Figura 9). Neste subgrupo apenas a *Cenp-C3* foi encontrada.

Na única espécie analisada do subgrupo *parvula*, a *D. kanapie*, encontramos apenas uma cópia de *Cenp-C*. Trata-se de uma segunda nova duplicação da *Cenp-C*, que denominamos de *Cenp-C4*. Essa cópia se encontra dividida entre dois *contigs*, o *contig1* (19.247 pb) e o *contig2* (4.803 pb), que apresentam uma região de sobreposição de 1.120 pb e com 100% de identidade (Figura 10). Os pontos de quebra desses dois *contigs* correspondem a regiões de íntrons. Uma das extremidades de um terceiro *contig2* (1372 pb), onde se encontra o fim da sequência da *Cenp-C4* (Figura 10). Como o *contig3* apresenta em sua extremidade o fragmento do gene da *Cenp-C4* flanqueado pelo gene *5-HT2A* na posição 3' deduziu-se, com base em sintenia com outras espécies, que o gene da *Cenp-C4* se encontra flanqueado pelos genes *CG9775* e *5-HT2A* (Figura 9).



Figura 10. Alinhamento dos *contigs* 1, 2 e 3 com a sequência definida para o gene da *Cenp-C4* em *D. kanapiae* (DkanCenpC4). O alinhamento das sequências dos *contigs* mostra regiões idênticas (preto), divergentes (cinza) e de gaps (branco). As barras em cinza mostram sequências de íntrons inferidos

pelo algoritmo *Augustus* e por alinhamento com os genes da Cenp-C das outras espécies do grupo *montium*. A barra em azul representa a região deletada de *Cenp-C4* para análises de seleção positiva e construção da árvore filogenética.

Por último, além do gene da *Cenp-C1*, *D. vulcana* (subgrupo *seguyi*) foi a única espécie do grupo *montium* a apresentar uma segunda cópia integra no mesmo genoma. Esse parálogo da *Cenp-C1*, que denominamos de *Cenp-C5*, se encontra em um loco flanqueado pelos genes *atms* e *CG14655* (Figura 9). O gene da *Cenp-C5* foi encontrado dividido entre dois *contigs* (1 e 2) e com extremidades que não se sobrepõe (Figura 11). No entanto, como foi observada sintenia entre o loco dos genes *atms* e *CG14655* nas espécies representadas na Figura 9, as extremidades desses dois *contigs* foram conectadas. Essas extremidades correspondem a uma região de íntron (Figura 11) na sequência da *Cenp-C1* de *D. vulcana*.





É importante ressaltar que os locos das cinco cópias da *Cenp-C* (*Cenp-C1*, *Cenp-C2*, *Cenp-C3*, *Cenp-C4* e *Cenp-C5*) foram investigados em todas as espécies. Isso foi feito através do método de alinhamento gráfico *Dotplot*, utilizando a sequência de cada cópia de *Cenp-C* com seus genes flanqueadores contra seus respectivos locos nas outras espécies que supostamente não possuíam a cópia analisada. Um exemplo dessa análise é mostrado na Figura 12. Não encontramos nenhum vestígio da *Cenp-C1*, *Cenp-C2*, *Cenp-C3*, *Cenp-C4* ou *Cenp-C5* em espécies onde estes genes não foram encontrados. Isso indica que, provavelmente, a *Cenp-C3*, *Cenp-C4* e Cenp-C5 se originaram de duplicações independentes de *Cenp-C1*.



Figura 12. Gráfico *Dotplot* entre *Cenp-C1* de *D. nikananu* e seus genes flanquadores contra o loco da *Cenp-C1* de *D. burlai* e seus genes flanqueadores. A região de similaridade corresponde aos genes *CG1427* e *Tim10* que flanqueiam o gene da *Cenp-C1* de *D. nikananu*. Apenas os genes *CG1427* e *Tim10* estão presentes no loco da *Cenp-C1* no genoma de *D. burlai*.

No caso específico de *D. burlai*, nenhum loco contendo o gene inteiro da *Cenp-C* foi identificado. Porém, foram encontrados vários *contigs* curtos (1.026 pb a 25.030 pb) contendo fragmentos de *Cenp-C*. Quatro desses fragmentos estão flanqueados por diferentes conjuntos de genes, indicando localizações genômicas distintas. Desses quatro fragmentos, os três maiores estão representados na Figura 13.



Figura 13. Disposição dos fragmentos do gene da *Cenp-C* de *D. burlai* e *D. punjabiensis* em relação aos 5 locos das cópias dos genes da *Cenp-C*, no cromossomo 3R de *D. triauraria*. As setas preenchidas e pontilhadas indicam, respectivamente, a presença e ausência dos genes que podem estar inteiros ou fragmentados. A orientação das setas indica a orientação dos genes no cromossomo. O tamanho das sequências de DNA presente entre os locos está representado em preto acima das linhas contínuas. As linhas pontilhadas com setas nas extremidades (azul e verde) mostram localização dos locos indicados pela seta. A distância desses locos em relação ao gene mais próximo (*CG14655* ou *CG1427*) está representada ao lado de cada linha pontilhada com a mesma cor da linha. Abaixo de cada fragmento do gene da *Cenp-C* é mostrado seu tamanho em pares de bases.

O mapeamento de *reads* de sequenciamento de *D. burlai* contra o gene ou a sequência codificante da *Cenp-C1* da espécie mais próxima (*D. nikananu*), também não possibilitou a recuperação da sequência completa do gene da *Cenp-C* nesta espécie. O que se observou foi um resultado ambíguo, com diferentes *reads* mapeando em uma mesma posição. É provável que nenhum gene completo da *Cenp-C* tenha sido montado em um único *contig* de

D. burlai, talvez devido à presença de sequências repetitivas em abundância nas regiões de íntrons, o que pode ter impedido a sua montagem. Por exemplo, no caso do fragmento de *Cenp-C* flanqueado pelo gene *eIF3f1* (Figura 13), o *contig* é justamente interrompido onde se iniciaria um íntron na sequência da *Cenp-C*. Essa mesma explicação pode ser atribuída à separação da *Cenp-C4* e *Cenp-C5* em fragmentos contidos em mais de um *contig*.

De fato, a maioria das espécies do grupo *montium* teve seus genomas montados a partir de *reads* curtas (100 pb) sequenciadas por *Illumina HiSeq 2000* ou *HiSeq 2500 Systems* (Bronski *et al.* 2020). Em contraste, outros sequenciamentos de três espécies do grupo geraram *reads* muito mais longas, como: 3kb e 8kb em *D. kikkawai* por uma combinação de 454 e Tecnologia Ilumina (Chen *et al.* 2014); 8.8 kb em *D. serrata* por *PacBio* (Allen *et al.* 2017); e 10 kb em *D. triauraria* por *Oxford Nanopore* e *Ilumina Hiseq* (Torosin *et al.* 2020), o que facilitou a montagem da *Cenp-C*. Por exemplo, a *Cenp-C1* de *D. serrata* se encontra separada entre dois *contigs* (*contig1*: 31.192 pb e *contig2*: 13.092 pb) no genoma montado por Bronski *et al.* (2020) (Figura 14). Já no genoma montado por Allen *et al.* (2017), a sequência se encontra em um mesmo *scaffold* de 5.769.022 pb. Outro fato que aponta que os genes da *Cenp-C* de *D. burlai* e *D. kanapiae* estão fragmentados por limitação da montagem dos genomas é que a Cenp-C é necessária para manutenção da Cid no centrômero (Erhardt *et al.* 2008; Orr & Sunkel 2011). Por isso, espera-se que cada espécie tenha pelo menos uma cópia do gene.



Figura 14. Alinhamento dos dois fragmentos (*contigs* 1 e 2) do gene da *Cenp-C1* de *D. serrata* montados por Bronski *et al.* (2020) e da sequência completa da *Cenp-C1* presente em apenas um *contig* montado por Allen *et al.* (2017). As barras em cinza mostram sequências dos íntrons definidos com base em RNA-Seq. A barra em azul representa a região deletada da *Cenp-C1* para análises de seleção positiva e construção da árvore filogenética.

Em *D. kanapiae* (subgrupo *parvula*), além do gene da *Cenp-C4*, também foi identificado um fragmento do gene da *Cenp-C* ocupando inteiramente um *contig* (3.140pb), o que impossibilita a identificação do seu loco.

Em *D. punjabiensis* (subgrupo *punjabiensis*), além do gene da *Cenp-C1*, também foram identificados outros seis fragmentos do gene da *Cenp-C*. Um deles ocupa quase inteiramente um *contig* (3.029 pb) e outros três estão presentes em *contigs* nos quais não foram identificados outros genes, sendo, portanto, impossível inferir os locos desses quatro fragmentos. Um quinto fragmento é flanqueado pelo gene *CG1427* que está em um *contig* diferente daquele onde está a *Cenp-C1*. Esse fragmento e o gene da *Cenp-C1* estão

flanqueados na posição 3' por sequências diferentes. Isso, possivelmente, indica que houve uma duplicação parcial de *Cenp-C1* neste loco (Figura 9). E por último, o sexto fragmento foi encontrado flanqueado pelo gene *Rab-23* em uma região diferente do genoma (Figura 13).

Em *D. watanabei* (subgrupo *punjabiensis*), além da *Cenp-C1*, foram identificados outros cinco fragmentos da *Cenp-C* em diferentes *contigs* (2.692pb a 15.737 pb). No entanto, nenhum deles contém outro gene que possibilite a identificação do loco.

Em *D. vulcana* (subgrupo *seguyi*), além da *Cenp-C1* e da *Cenp-C5*, três fragmentos da *Cenp-C* foram encontrados em três diferentes *contigs*. Um desses *contigs* possui mais três outros genes (*Herzog*, *Sh3beta* e *Sec63*) que se localizam no braço esquerdo do cromossomo 3 (3L) de *D. triauraria*.

Em *D. bakoue* (subgrupo *seguyi*), além da *Cenp-C1*, um fragmento da *Cenp-C* foi encontrado em um *contig* contendo o gene (*canoe*) que se localiza no braço direito do cromossomo 3 (3R) de *D. triauraria*.

Por último, em *D. nikananu* (subgrupo *seguyi*), além da *Cenp-C1*, um fragmento da *Cenp-C* foi encontrado. Como nenhum outro gene foi encontrado no *contig* contendo esse fragmento, seu loco não pode ser identificado.

Os genes inteiros de *Cenp-C* encontrados nas espécies do grupo *montium* possuem grande variação de tamanho (5.575 a 10.801 pb) (Tabela 2). Essa variação ocorre, principalmente, devido à diferença de tamanho dos íntrons, já que a sequência codificadora (CDS) varia apenas de 3.513 a 4.650 pb. Esse tamanho variável entre os íntrons pode ser explicado, em parte, pela inserção de TEs nos íntrons, já que estes elementos podem corresponder a 26% (ou 2.819pb) (*D. watanabei Cenp-C1*) ou até 43,3% (ou 4.353pb) (*D. vulcana Cenp-C5*) do tamanho do gene. Em contraste, *Cenp-C* menores possuem menor proporção de TEs, como 1,2% (ou 67 pb) em *D. leontia Cenp-C1* ou 2,4% (ou 135 pb) em *D. pectinifera Cenp-C3*) (Figura 15).

Subgrupos	Gene	N. de éxons	Tamanho do gene (pb)	Tamanho da CDS (pb)
	D. pectinifera Cenp-C3	10	5.610	4.383
montium	D. triauraria Cenp-C3	11	5.907	4.650
	D. auraria Cenp-C3	11	6.667	4.632
	D. tani Cenp-C3	11	8.750	4.485
	D. rufa Cenp-C3	11	6.933	4.596
	D. lateicornis Cenp-C3	11	7.078	4.587
	D. asahinai Cenp-C3	11	7.147	4.602
parvula	D. kanapiae Cenp-C4	11	8.013	4.086
kikkawai	D. kikkawai Cenp-C1	13	6.229	4.149

Tabela 2. Características gerais dos genes da Cenp-C das espécies do grupo montium

	D. leontia Cenp-C1	13	5.575	3.978	
	D. bocki Cenp-C1	11	6.310	4.122	
	D. serrata Cenp-C1	12	6.591	4.149	
	D. bunnanda Cenp-C1	12	7.322	4.245	
serrata	D. truncata Cenp-C1	12	8.324	4.023	
	D. mayri Cenp-C1	11	5.885	3.954	
	D. birchii Cenp-C1	12	5.861	4.101	
punjabiensis	D. watanabei Cenp-C1	12	10.801	4.068	
	D. punjabiensis Cenp-C1	10	7.076	3.654	
	D. jambulina Cenp-C1	11	7.015	3.864	
	D. seguyi Cenp-C1	12	7.039	4.116	
seguvi	D. vulcana Cenp-C1	7	8.552	3.513	
eeguj:	D. vulcana Cenp-C5	10	10.063	3.798	
	D. bakoue Cenp-C1	11	9.264	3.978	
	D. nikananu Cenp-C1	12	8.420	4.023	
DwatCenpC1		00 4,500 5,000 5,500 6,00			10,801
DvulCenpC5					
DpecCenpC3				• •••	
DleoCenpC1					

Figura 15. Representação da sequência do gene da *Cenp-C1* de *D. watanabei, Cenp-C5* de *D. vulcana, Cenp-C3* de *D. pectinifera* e *Cenp-C1* de *D. leontia.* As barras em cinza mostram sequências de íntrons inferidos pelo algoritmo *Augustus* e por alinhamento com os genes da *Cenp-C* de outras espécies do grupo *montium.* As linhas pretas representam os éxons. As barras coloridas representam sequências que possuem similaridade com TEs de acordo com o banco de dados RepBase.

Os cinco locos das cópias da *Cenp-C* descritos acima estão presentes no cromossomo 3R de *D. triauraria* (Figura 13). Genes representando estes 3 locos também estão localizados no cromossomo 3 de *D. melanogaster,* espécie que se divergiu do grupo *montium* há cerca de 28 milhões de anos (Russo *et al.* 2013). Portanto, é provável que os locos onde se encontram as cópias *Cenp-C3, Cenp-C4* ou *Cenp-C5* também estejam presentes no cromossomo 3 em todas as espécies do grupo *montium*.

As evidências obtidas no presente trabalho sugerem que todos as cópias de *Cenp-C* (*Cenp-C3*, *Cenp-C4* e *Cenp-C5*) identificadas nas espécies do grupo *montium* são funcionais. A evidência mais robusta é a essencialidade da *Cenp-C* na função centromérica que requer ao menos uma cópia. Outra evidência é que o algoritmo *Augustus* identificou a sequência codificante esperada para cada duplicação de *Cenp-C* e ausência de stop códons prematuros ao longo das sequências. A terceira evidência é a conservação observada entre as sequências de aminoácidos das cópias da *Cenp-C*, o que inclui motivos proteicos importantes para a função canônica de Cenp-C (ver tópico 4.2).

Como abordado anteriormente, em espécies do subgênero *Drosophila* foi hipotetizado que cada uma das duas cópias de Cid (Cid1 ou Cid6 e Cid5) e de Cenp-C (Cenp-C1 e Cenp-C2) poderiam estar interagindo de maneira específica entre elas como mostrado na Figura 5. Por isso, buscamos investigar com quantas cópias de Cenp-C as três cópias de Cid (Cid1, Cid3 e Cid4) estariam interagindo nas espécies do grupo *montium*.

Apesar de termos encontrado novas cópias de *Cenp-C* nessas espécies, cada espécie possui apenas uma cópia funcional, que equivale a *Cenp-C1*, mas em diferentes locos. Desta forma, podemos considerar *Cenp-C3* e *Cenp-C4* como "ortólogos funcionais" de *Cenp-C1*. Já *D. vulcana* é a única espécie do grupo *montium* que manteve duas cópias de *Cenp-C* no genoma, *Cenp-C1* e *Cenp-C5*.

Os resultados acima indicam que, supreendentemente, apenas uma cópia de Cenp-C é suficiente para interagir funcionalmente com as 3 cópias de Cid (Cid1, Cid3 e Cid4) presentes nas espécies do grupo *montium*. Este resultado contrasta com os obtidos em espécies do subgênero *Drosophila*, onde parálogos divergentes de Cid e Cenp-C co-existem nos mesmos genomas e possivelmente interagem entre si de forma específica. Já em *D. vulcana*, a presença de duas cópias Cenp-C levanta a questão se estas duplicatas interagem com cópias diferentes de Cid.

4.2. Identificação dos motivos proteicos conservados em cópias da *Cenp-C* nas espécies do grupo *montium*

A Cenp-C1 de D. melanogaster possui 7 motivos proteicos funcionais distribuídos ao longo de sua sequência, da região N-terminal para a C-terminal, sendo estas: R-rich (*arginine-rich*), DH (*Drosophilid Cenp-C Homolegues*), AT *hook* 1 (AT1), NLS (*Nuclear Localization Signal*), CenH3 *binding*, AT *hook* 2 (AT2) e Cupin (Figura 16A). Os motivos AT1 e AT2 e o motivo NLS apresentam similaridade com AT *hook* que medeia a ligação ao suco menor do DNA e com o sinal de localização nuclear, respectivamente (Heeger *et al.* 2005).



Figura 16. Motivos funcionais da sequência da proteína Cenp-C em *Drosophila*. (A) Representação dos 7 motivos funcionais identificados por Heeger et al. (2005) e do provável motivo Mis12 binding e suas respectivas posições na sequência da proteína da Cenp-C1 em *D. melanogaster*. O asterisco em preto corresponde a arginina (R1101) da Cenp-C que é idêntica em todas as proteínas. N e C indicam a região C-terminal e N-terminal da proteína, respectivamente. (B) Representação dos motivos presentes na sequência das cópias da Cenp-C, nas espécies do grupo *montium*. (C) Logo gerado pelo *MEME* do provável motivo de ligação do complexo Mis12 (Mis12 binding).

Todos os 7 motivos com exceção do AT1, parecem desempenhar importantes funções, já que a Cenp-C sem tais regiões foi incapaz de prevenir anormalidades fenotípicas em embriões contendo a Cenp-C mutante. A arginina presente na posição 1101 da Cenp-C1 (R1101), que corresponde à região do motivo CenH3 *binding*, se mostrou importante para a localização da Cenp-C no centrómero (Heeger *et al.* 2005). Esse aminoácido nesta posição foi encontrado em Cenp-C de *Drosophila*, plantas, leveduras e vertebrados (Heeger *et al.* 2005; Talbert *et al.* 2004; Teixeira *et al.* 2018). Além disso, o motivo Cupin é necessário para a formação de dímeros de Cenp-C. Apenas após essa dimerização é possível a interação da Cenp-C com a chaperona CAL1 para a deposição de nucleossomos contendo a Cid na cromatina centromérica e também para o recrutamento da própria Cenp-C para esse loco (Roure *et al.* 2019). No entanto, a função específica da maioria dos motivos ainda permanece desconhecida.

Dada a importante função de pelo menos seis dos motivos da Cenp-C determinados por Heeger *et al.* (2005) é esperado que os mesmos estejam conservados em espécies que possuem apenas uma cópia funcional do gene. Como previsto, com exceção dos motivos AT1 e AT2, todos os outros motivos se encontram conservados nas cópias de Cenp-C das

espécies do grupo *montium*, incluindo a arginina presente na posição 1101 (Figura 16B, Figura 17 e Figura 18). A conservação de motivos essenciais, incluindo CenH3 *binding* e o Cupin necessários para localização da Cenp-C/Cid no centrômero é um forte indicativo da manutenção da função centromérica para todas as cópias de *Cenp-C* identificadas nas espécies do grupo *montium*. Embora o motivo AT1 tenha se mostrado não essencial para função da Cenp-C1 em *D. melanogaster*, o mesmo não é verdadeiro para o motivo AT2 (Heeger *et al.* 2005). Desta forma, talvez este motivo não tenha sido detectado nas espécies do grupo *montium* devido ao seu grau de divergência em relação ao de outras espécies.

Enquanto que a região C-terminal da Cenp-C1 em *D. melanogaster* é suficiente para localização centromérica (Heeger *et al.* 2005), a região N-terminal é necessária para o recrutamento de proteínas do cinetócoro (Pzewloka *et al.* 2011). Esse recrutamento ocorre por interação direta entre o Mis12C e o fragmento N-terminal (1-105) da proteína Cenp-C1. Foi demonstrado que dentro desse fragmento, os resíduos de 9-35 possuem determinantes essenciais nessa interação (Liu *et al.* 2016). Outro estudo mostrou que os primeiros 45 aminoácidos da Cenp-C1 de *D. melanogaster* estão envolvidos na interação com Mis12-Nnf1a (subunidades do Mis12C) (Richter *el al.* 2016). Concordante com esses achados, detectamos um motivo na posição N-terminal de todas as Cenp-C1 de *D. melanogaster* (Figura 16C, Figura 17 e Figura 18). A ligação da Cenp-C ao Mis12C é necessária para o recrutamento dos demais componentes do cinetócoro ao centrômero (Liu *et al.* 2016; Richter *et al.* 2016). Por isso, a presença desse sexto motivo (Mis12 binding) conservado na Cenp-C das espécies do grupo *montium* pode representar mais um indício do seu papel na função centromérica.

	1	190	200	300	400	500	600	700	800	90	0	1,000	1,100	1.200	1,300		1,400	1,500	1.600	1,7	700	1,814
Consensus	0000000000	E DI DE		no ne		HERIO			uui –		0000000		∎u nin-		- i III	101	- in 19 10 (19	101800.000 P	- 100 1000	e na de		(IIIII)
Identity	and the first	Land St.	A Issie		MIN THE	UNIT W	de N	Mandal	11 A A		Sec. No		مرا ب ¹⁴ م ا			JP.	مرد ا ز المکر می		N. A.	1. au		(HH
1 DnunCennC1		- 100000000		1000-000				8110-0100			0.0									FEBRUAR DR		
	_					_			_									- T		_		_
2. DwatCenpC1																111				7111111111		
DkikCenpC1		- IIIIIIII														11				10000		
4. DbocCenpC1																			1.000			
5. DleoCenpC1																11				6111111- 1 12		
DvulCenpC1															101	180				1.000		
7. DvulCenpC5											111 111								1.000	COLUMN 12		
 8. DnikCenpC1 																10				STOLET IN		
DbakCenpC1		E IIIIII IIII													11	H				STILLE IN		
DjamCenpC1								81.00								H				a de la compansión de la c		
 DsegCenpC1 				100 00				810000							1	11				1.000		
12. DbunCenpC1		E DE			100000000000000000000000000000000000000											HI-I				A DECEMPENT OF		
13. DsetCenpC1		E DE		1000-000-				BUD DOUD			DEC DE					H-H				e de la compansión de la c		1000
14. DtruCenpC1		E DI DI DI G		1000-000-				810000-000	0.0001						10	11		0.0000000000000000000000000000000000000		e de la companya de la company		
15. DbirCenpC1	1000000000			1000-000							0010-0010					10.0			100.0000	FEBRUAR DE		
16. DmayCenpC1	1000000000	- III III III II		100 100		- FORMERINE					INCOME.					111-1			10.000	FEIDER-10		
17. DkanCenpC4	1000000000	E DE LE D						010 - 010	100 - I		-0.1100-0					HI-I			10.000	ETHER IN THE REPORT		
18 DpecCennC3											000000					DOM: N				and the second		
19 DaurCennC3	10100100															111-1			10.000			
20 DtriCenpC3											-00.41100.0					-141-1				I IIIIIIII		
21 DtanCennC3		1.110.000														111-1				E III III		
27. DrufCennC3																100.00						1000
22. DisconpC3																i se				i i i i i i i i i i i i i i i i i i i		
24 DisconnC3																10.0				COLUMN I		
ZH, DIOUUCHIDUD	the second se				The second se						AND REAL PROPERTY.					- 19 C				ALC: NO DESCRIPTION	the second se	

Figura 17. Alinhamento das proteínas Cenp-C das espécies do grupo *montium* mostrando a posição dos motivos proteicos identificados no presente trabalho. As barras em vermelho representam da esquerda para a direita: Mis12 *binding*, R-rich, DH, NLS, CenH3 *binding* e Cupin. A barra azul-claro representa as sequências removidas para a construção da filogenia dos genes e teste de seleção positiva. A pequena barra em verde-escuro abaixo do motivo CenH3 *binding* representa o aminoácido R-1101, presente em todas as cópias de Cenp-C do grupo *montium*.

Consensus	MSKIPK PRKDÍ TIXER MELELIM ŠQPDADEGQF	AAYLORKLAEKORPKNND - XLECHV	S-EDEDEDVNLIKKRKBA	
1. DpunCenpC1	MSKILNPRKOTTI I EEMEELM SOPDADE GQF	AAYLORKLAEKORPMNND - LLFGNV	S-EDEDEDVNLIKKPMPT	G\$\$\$VVNDD- VASGE AK EPPA P DP AA@\${
2. DwatCenpC1 3. DbikCenpC1 4. ObotCenpC1 5. DieoCenpC1 6. DwulCenpC5 8. DnikCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DsegCenpC1 12. DbuhCenpC1 13. DsetCenpC1 14. DtruCenpC1 15. DbihCenpC1 15. DbihCenpC1 15. DbihCenpC1 19. DbauCenpC1 20. DpurCenpC3 20. DtriCenpC3 21. DtanCenpC3 22. DruTcenpC3 23. DasaCenpC3	SSISTL N.P. RODINI I GEMIERING SP. DAB. GOP SSISTER P.R. COMULTIER MEETING N.C. PARE GOP SSISTER P.R. COMULTIER MEETING N.C. PARE GOP SSISTER P.R. COMULTIER MEETING N.C. PARE GOP SSISTER P.R. RODINI I TEM ENTRY N.C. PARE GOP SSISTER P.R. RODINI V.D.II TELI I.N. S.P. PARE GOP SSISTER P.R. RODINI V.L. I.H. MEETING SSI DABE GOP SSISTER P.R. RODINI V.L. HAM MEETING SSI DABE GOP SSISTER V.R. RODINI V.L. HAM MEETING SSI DABE GOD SSISTER V.R. RODINI V.L. HAM MEETING SSI DABE GOD SSISTER V.R. RODINI V.H. HAM MEETING SSI DABE GOD SSISTER V.C. RODINI V.H. L. DEHNIG SSI	ZANUERZA LAEKKER DUNNEE - LUSSEN VANUERZA LAEKKER DUNNEE - LUSSEN VANUERZA LAEKKER DUNNEE - RUSSEN ZANUERZA LAEKKER DUNNEE - RUSSEN KANUERZA VARKER DUNNEE - RUSSEN KANUERZA VARKER DUNNEE - RUSSEN KANUERZA VARKER DUNNEE - RUSSEN KANUERZA VARKER DUNNEE - RUSSEN VANUERZA VARKER DUNNEE - RUSSEN VARKER DUNNEE - RUSSEN VARKE	S DOFINITION LINERSON P	
Consensus Identity 1. DpunCenpC1	120 10011 Pilizaa - Txisen Letenia ir - 51 11011 Silizaa - Txisen Letenia ir sse	100 R	180 IRQUIRCE E È - UNVIRSISSI A TEND TVSIST IRQUIRCE E È - UNVIRS P EKDAGSIST	210 240 240 250 250 250 250 250 250 250 250 250 25
2. DwatCenpC1 3. DikiKcenpC1 4. ObacCenpC1 5. DieoCenpC1 6. DwulCenpC5 8. DnikCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DegrCenpC1 12. DburCenpC1 13. DikroenpC1 14. DirwicenpC1 15. DbirCenpC1 16. DmayCenpC1 17. DkanCenpC4 18. DpecCenpC3 20. DtriCenpC3 21. DtarCenpC3 22. DruTCenpC3 23. DasaCenpC3 24. DlaCenpC3 24. DlaCenpC3	L 10 TT	R		CGGS 2P130 I = Q VKE I DNNBK MT LSENCE TEK CAGGEP210 I = C VKE I DNNBK MT LSENCE TEK CAGGEP210 I = C VKE P E BN30 I K S E VGETTEK CAGGEP210 I = C VKE P E BN30 I K S E VGETTEK CAGGEP210 I = C VKE E LG N120 MK S E VGETTEK CAGGEP210 I = VF E E I E TT SH I 1 SSE K LBEATK CAGGEP210 I = VF E E I E TT SH I 1 SSE K LBEATK CAGGEP210 I = VF E E I E TT SH I 1 SSE K LBEATK CAGGEP210 I = VF E E I E TT SH I 1 SSE K LBEATK CAGGEP210 I = VF E E I E TT SH I 1 SSE K LBEATK CAGGEP210 I = VF E E I E TT SH I 1 SSE K LBEATK CAGGEP210 I = VF E E I E TT SH I 1 SSE K LBEATK CAGGEP210 I = VF E E I E TS H I 1 SSE K LBEATK CAGGEP210 I = VF E E I E TS H I 1 SSE K LBEATK CAGGEP210 I = VF E I E I SH I I M I N S K K LBEATK CAGGEP210 I = VF E I E I SH I M I N S K K LBEATK CAGGEP210 I = VF E I E I M 120 M I N S K K LBEATK CAGGEP210 I = VF E I E I M 120 M I N S K K LBEATK CAGGEP210 I = VF E I I M 120 M I N S K K LBEATK CAGGEP210 I = VF E I I M 120 M I N S K K LBEATK CAGGEP210 I = VF E I I M 120 M I N S K K LBEATK CAGGEP210 I = VF E I I M 120 M I N S K K LBEATK CAGGEP210 I = VF E I I M 120 M I N S K K LBEATK CAGGEP210 I = VF E I I M 120 M I N S K K LBEATK CAGGEP210 I = VF E I I M 120 K K S K LBEATK CAGGEP210 I = VF E I I M 120 K K M S K LBEATK CAGGEP210 I = VF E I I M 120 K K M S K LBEATK CAGGEP210 I = VF E I I M 120 K K M S K LBEATK CAGGEP210 I = VF E I I M 120 K K M S K LBEATK CAGGEP210 I I = VF E I I M 120 K K M S K LBEATK CAGGEP210 I I S K K K K LBEATK CAGGEP210 I I S K K K K LBEATK CAGGEP210 I I S K K K K LBEATK CAGGEP210 I I I S K K K K LBEATK CAGGEP210 I I I S K K K K LBEATK CAGGEP210 I I I I I I I I I I I I I I I I I I I
Consensus	240 250 250 PAREKSEV2VQTEPKKCLTALDSR	270 280 TVDITAAE AS LMY RERUSSEQNNIL QI	290 310 DYNGXM	370 330 340 35 5PAPGQUULSPSRRBPS
ldentity 1. DpunCenpC1	> A E K P AEKSQV QT EPKK C I TAL D S EI	TIDMVAHASLIMYRQRUSSEQNNI SQI	DCIIGKA	5PARGQUULSPISRR
2. DwatCenpC1 3. DkikCenpC1 4. DbocCenpC1 5. DwatCenpC1 7. DwatCenpC1 9. DbakCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DsegCenpC1 12. DbatCenpC1 13. DsetCenpC1 14. DtruCenpC1 15. DbirCenpC1 15. DbirCenpC1 17. DkanCenpC1 10. DjamCenpC1 20. DtriCenpC3 21. DtraCenpC3 22. DruTcenpC3 24. DlacCenpC3 24. DlacCenpC3			DYING * KB	12.0.2 0.0.4 0.0.5 0.0.7 0.0.7 12.0.1 0.0.5 0.0.7 0.0.7 0.0.7 12.0.1 0.0.5 0.0.7 0.0.7 0.0.7 12.0.1 0.0.7 0.0.7 0.0.7 0.0.7 12.0.1 0.0.7 0.0.7 0.0.7 0.0.7 12.0.1 0.0.7 0.0.7 0.0.7 0.0.7 12.0.1 0.0.7 0.0.7 0.0.7 0.0.7 12.0.1 0.0.7 0.0.7 0.0.7 0.0.7 12.0.1 0.0.7 0.0.7 0.0.7 0.0.7 12.0.1 0.0.7 0.0.7 0.0.7 0.0.7 12.0.1 0.0.1 0.0.7 0.0.7 0.0.7 12.0.1 0.0.1 0.0.7 0.0.7 0.0.7 0.0.7 12.0.1 0.0.1 0.0.7 0.0.7 0.0.7 0.0.7 0.0.7 12.0.1 0.0.1 0.0.7 0.0.7 0.0.7 0.0.7 0.0.7 0.0.7 0.0.7
Consensus Identity 1. DpunCenpC1	O 170 C TEA - CRIMENTATIVEN RH R EN XEMPOTER C TEA - DEREK KONVEN RH K D TVM POTER	390 Alta Hitr Couper Lay VIII Dick food G Alta Hitr Couper Citer Lay 2010, food G Alta Hitr Couper Citer Lay 2010, food G	410 LV/RNLSTRNASTCRUV/VI/INF/RNSDRTAELAHAG LV/RNLSTRNASTCRUV/VI/INF/RSSDRTAELAHAG	ango IEVZEV VARA APSALITIZIUHRANSTIEV SASUQPIA- IEVZEV VARA APSALITIZI HVIBSISTIEGPCARSFQLA-
2. DwatCenpC1 3. DikiCenpC1 4. DbotCenpC1 5. DieoCenpC1 6. DvulCenpC1 7. DvulCenpC5 8. DnikCenpC1 9. DbakCenpC1			LUQNU GRNAAS OP I VVI INTAN SIGENAE MAH VO IUQNU GRNAAS OP I VVI INTAN SIGENAE MAH VO IUQONU GRNAAS OP I VVI INTAN SIGETTE VAH VO IUQONU GRNAAS OP I VVI INTAN SIGETTE VAH VO IUQONU GRNAAS OP I VVI INTAN SIGETTE VAH VO IUQONU GRNAAS OP I VVI INTAK SIGETAE I THAO IUQONU GRNAAS OP I VVI INTAK SIGETAE I THAO IUQONU GRNAAS OP I VVI INTAK SIGETAE I THAO IUQONU GRNAAS OP I VVI INTAK SIGETAE I THAO	DEV P VE ANS AR SALUTELT I VISSETUE GPC (ASS QLIA) DEVISIV LANS AR SALUTELTI VISSETUE GPC (ASS QLIA) DEVISIV LANS AR SALUTELTI VISSETUE STATUS DEVISION CAN BE ALUTELTI VISSETUE SALUS ANS DEVISION CAN BE ALUTELTI VISSETUE SALUS ANS DEVISION CAN BE SALUTELTI VISSETUE SALUS ANS DEVISION CAN BE SALUTELTI VISSETUE VISSA SALUS ANS DEVISION CAN BE SALUTELTI VISSA VISSETUE VISSA SALUS ANS DEVISION CAN BE SALUTELTI VISSA VISSETUE VISSA SALUS ANS DEVISION CAN BE SALUTELTI VISSETUE VISSA SALUS ANS DEVISION CAN BE SALUTELTI VISSETUE VISSA SALUS ANS DEVISION CAN BE SALUTELTI VISSA VI

. Dvuicenpcs	
DnikCenpC1	
DbakCenpC1	JCTMA - QQR SMP IVKYKRHR EEVLIPDTPSRPQTG GSEPEPEYVYPDTOFOLEUGELVQNU SRNAS GPIVV INTAN SKPTPEL THAQPVSVVAS VPSALTPUHRASTEVAS AS LQLA-
0. DjamCenpC1	ICTAP - QKR FLP IVKYKRHR EEVLIPDTPPRTHTR GSEKEPEYVVPDTOFOLE GELVONUSINAS GPIVV INTAN SKPTAELVHAQDE SVVPTVS SALTPLHLASTEVASTSLQLV-
1. DsegCenpC1	ICTAG - RKR SMBLYKYK BHR EEVLIPDTPPQVHTR GSEQEPEYVYPDTOFO, EUGQUVQNU SRNASGPIVVINTAN SKPTADMTHAQPVSVVASEPSPLTPUHRTSTEVASSSUQLA-
2. DbunCenpC1	CTTA - QKRLMPVVKVKRHRDELLVPDTPPRTHTRESAPEPEFEVPDTOFODUGELVQNUSRNASGPLVVINTANSKSTVELPHAQPVSAVASAPSALTELHRASTELHSCSLQPA-
DsetCenpC1	ICTAA - OKR LMPVVK VKRHR EELLVPDTPPRTHTK EAVPEQE FVVPDTQFQDUGELVQNU SRNAS GP I VV INTAN SK SVVD LAHAQPVSMVASAPPALTP UHRASTEVHSAS LQPA-
4. DtruCenpC1	ionad - right mentioned in the second of the second s
5. DbirCenpC1	ICSAD - KKRLMPIVKYKRHREELLVPDTEQRAHTRVSEPEPEYVVPDTOFODUGELVQNUSRNASGPIVVINTANLKPTADLTHAQPVSVAASAPSALTEURCASTQDPSAALQLT-
DmayCenpC1	ir IAD - KKRLMPI VKVRRHREILLUVPDDE PEAQTRASEPEPEYVVPDTOFOD I GETVONUSKNAS GETVV INTANLKETA ELTDAQEVSMSAS APSALTELHCASHOVSSAALQLT-
 DkanCenpC4 	F TAA - QOR U SIZI VIX L RIBHR DHV PVRDDDERRAL ITT GISHPHEREN Y VIR E LOQ L QVU GISHVAN GIRL VIX INITAN PIX 2 TAK
8. DpecCenpC3	
9. DaurCenpC3	
0. DtriCenpC3	\$ PR PAAQKSSMPLIVKLIR PRAHVPDTPPRALTWOSEPEPECVVPDTOPOOLSELVQNUSRNASGPLVVINTANPKPTAERERTOPVSVVVSAPLPSTSUHRAMIVVSSASUPPA-
 DtanCenpC3 	* P AMA P QK P L MP L VK I R R S R A E VP DT P P R A L TWESE P E C V V P DT O P Q D L S EL V Q N L S R NA S GP I VV I NTAS P K P P T E R E HAQP VS P L Q R T STI VV S S A S L M
2. DrufCenpC3	3P AMA - QMP LIMPLIVK IR RPRAE VP ETPPRALTWOSEPEPECIVPDTOPODUSELVONUSRNAS GP I VV INTAS PKPTAEREHAORVSV VAS PUHRASTVVSSAS LIMPA-
3. DasaCenpC3	3P AMA - QMP LIMPLIVK I R R P R VE VP E TPPRA L TWO SEPERE OV VPDTO POOL SELVON L SRNAS GP I VV I NTAS PKP PA ER E HAOPVSV VAS P L HRAST VV S SAS LMPA-
4 Discourses	

Consensus Identity	470 490 470 570 570 570 570 570 570 570 570 570 5
1. DpunCenpC1 2. DwatCenpC1	
3. DkikCenpC1 4. DbocCenpC1 5. DleoCenpC1 6. DvulCenpC1 7. DvulCenpC5	TAANNNS 5227 TV PT 512238 PE 151324 SISTER SILUM PS 2010/2011 ALUDO 1850 EQ251 W2AT OL-20 CR SIBER OWE CULEROVE NG E BUILD- DP DK EN VV KULUTURUE (SILUM VIS 512 FAANNAS 522 I V PASTERAS PE 151324 SISTER SILUM PS 2010/2011 ALUDO 1850 EQ251 W2AT OL-20 CR SIBER OWE CULEROVE NG E BUILD- DP DK EN VV ELUMUTURUS I I 20 WIS 513 VANNINS 5122 A V PASTERAS PE 151324 SISTER SILUM PS 2010/2011 ALUDO 1850 EQ251 W2AT OL-20 CR SIBER OWE CULEROVE NG E BUILD- DP DK EN VV ELUMUTURUS I I 20 WIS 513 VANNINS 5122 A V PASTERAS PE 151324 SILUM PS 2010/2011 ALUDO 1850 EQ251 W2ATA OL-20 CR SIBER OWE CULEROVE NG E BUILD- DP DK EN VV ELUMUTURUS I I 20 WIS 513 VANNINS 5122 A V PASTERAS PE 151324 SILUM PS 2010/2011 ALUDO 1850 EQ251 W2ATA OL-20 CR SIBER OWE CULEROVE NG E BUILD- DP DK EN VV ELUMUTURUS I I 20 WIS 513 ATI NY NS 5122 A VAASTETTS PE 151324 SILUM PS 2010/2011 ALUDO 1850 EQ251 W2ATA OL-20 CR SIBER OWE CULEROVE NG E BUILD - DP DK EN VV ELUMUTURUS I I 20 WIS 513 ATI NY NS 5122 A VAASTETTS PE 151324 SILUM PS 2010 SILU
8. DnikCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DsegCenpC1 12. DbunCenpC1	FAIN YNS SIDAWAARDAW PLUSISIKISSIK SUI IVII NAVORUN HOOISIOSICEDSUIVAN DI -AHHOROD SHOR PCULIK HESLEINE -DPINKEN F VOLUUNUER SHUAANKISU FAIN YNS SIDA
13. DsetCenpC1 14. DtruCenpC1 15. DbirCenpC1 16. DmayCenpC1 17. DkarCenpC4	
18. DpecCenpC3 19. DaurCenpC3 20. DtriCenpC3 21. DtanCenpC3 22. DruiCenpC3	ICP S SPAAN C P S SPAAN
23. DasaCenpC3 24. DlacCenpC3	ASPERALVACSERRANGLANGERENSTESSEUWWWCNVDATHATDOIDSDEOD2SHWAMNU-RPPGGEDTTRERHURRENNR PERUE-DRIEDSSWOUUNUHRENNAARSSE RSPERALVACSERRANLAVEERENSTSSEUWWWGNVDATHATDOIDSDEOD2SHWAMNU-RPPGGEDTTRERHURRENNR PERUF-DRIEDSSWOUUNUHRENNAARSSE RSPERALVACSERRANLAVEERENSTSSEUWWWGNVDATHATDOIDSDEOD2SHWAMNU-RPPGGEDTTRERHURRENNR PERUF-DRIEDSSWOUUNUHRENNUHREN
Consensus	yo IQK TIMPATIVELUNKAR GENETING EQFA GENARMISTIKET DURKRNSLIGSKVYP MISCH KTASIST VRXIII DURKSTIST ALLEUMRE – RNIIK – TINGERUI KG – S – XAADIAU BESEIEMURPR – DEINT XRN K –
ldentity 1. DpunCenpC1	
2. DwatCenpC1 3. DkikCenpC1 4. DbocCenpC1 5. DleoCenpC1 6. DwuCenpC1	
7. DvulCenpC5 8. DnikCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DsegCenpC1	IOK TISPANIZZI NIVABZIALENGEQEACHIVARMINIZYEI HOHBASISIAGMI. HPYNOR KIISIST EORILIDHERSI IORIEHMAR BIXIK - HUNRIKR UTG - SAKA EIRUS EN EMURA SVAINTYRN I K- IOK TISPANIZZI NIVABZIALENGEOZACHIVARMINIZYEI HOHBASISIAGMI. HPYNOROKKISIST VORIGI HENRIK - HUNRIKR UTG - STAADOUR OSE EMURA SVAINTYRN I K- IOK TISPANIZZI NIVABZIALENGEOZACHIVARMINIZYEI HOHBASISIAGMI HPYNOROKKISIST VORIGI HENRIK HUNRIKKI UTG - STAADOUR OSE EMURA SVAINTYRN I K- IOK TISPANIZZI NIVABZIAZI NIVEEZZACHIVARMINIZYEI HOHBASISI GILI HVIKA (KISIST VORIGI HUHBASI I SUINTYRI NIKA IOK TISPANIZZI NIVABZIAZI NIVEEZZACHIVARMINIZYEI HOHBASISI GILI VIVABIA (KISIST VORIGI HUHBASI I SUINTYRI KIKIS NIVEENKI NIVABZIAZI NIVEEZZACHIVARMINIZYEI HOHBASISI GILI VIVABIA (KISIST VORIGI HUHBASI I SUINTYRI KIKISTI HUNRIK DERTINARIVUUNIKARZACHIVAETI NIEGZZACHIVARMINISYEI HOHBASISI GILI VIVABIA (KISIST VORIGI HUHBASISI GILI HUMR- BASIKI I MINIATALI NIVEEZZACHIVARMINIZI HUOHBASISI GILI VIVABIA (KISIST VORIGI HUHBASISI GILI HUMR- BASIKI I MINIATALI NIVEEZZACHIVARMINIZI HUOHBASISI GILI VIVABIA (KISIST KIKISTI HUMR- BASIKI I MINIATALI NIVEEZZACHIVARMINIZI HUDI KARANSI USIKI VIVABIA (KISIST KIKISISI LIHUMR- BASIKI I MINIATALI NIVEEZZACHIVARMINIZI HUDI KARANSI USIKI VIVABIA (KISIST KIKISTI HUMR- BASIKI I MINIATALI NIVEEZZACHIVARMINIZI HUDI KARANSI USIKI KIKI KIKISTI KIKI HUMR- BASIKI I MINIATALI NIVEEZZACHIVARMINIZI HUDI KARANSI USIKI KIKI KIKI KISISTI HUMR- BASIKI I MINIATALI NIVEEZZACHIVARMINIZI HUDI KARANSI USIKI KIKI KIKI KISISTI HUMR- BASIKI I MINIATU NIVABASI ANDI KIKI KIKI KIKI KIKI KIKI KIKI KIKI K
12. DbunCenpC1 13. DsetCenpC1 14. DtruCenpC1 15. DbirCenpC1 16. DmaxCenpC1	IPK NSPANDUDINIVABCIADINGEQEACEU/RMMNYETUDIRXRNSIGNUS LVNGHKIRNSTVORILDUERSIDILEUMRERING KUNTTTAADELPEIREILEPP - GUNTDRIMMQ IOK AN PANDUNXRACHARINGEOACEU/RMMNYETUDIRXRNSIGNUS LVNGHKIRNSTVORILDUERSIDILEUMRERINGEOACEU/RMMNYETUDIRXRNSIGNU IOK SIR SVUDINXRACHARINGEOACEU/RMMNYETUDIRXRNSIGNUS LVNGHKIRNSTVORILDUERSIDILEUMRERISIDILEUMRERISIDILEUMRERIS IOK SIR SVUDINXRACHARINGEOACEU/RMMNYETUDIRXRNSIGNUS LVNGHKIRNSTVORILDUERSIDILEUMRERISIDILEUMRERISIDILEUMRERIS IOK SIR SVUDINXRACHARINGEOACEU/RMMNYETUDIRXRNSIGNUS LVNGHKIRNSTVORILDUERSIDILEUMRERISIDILEUMRERIS IOK SIR SVUDINXRACHARINGEOACEU/RMMNYETUDIRXRNSIGNUS LVNGHKIRNSTVORILDUERSIDILEUMRERISIDILEUMRERIS IOK SIR SVUDINXRACHARINGEOACEU/RAMINYETUDIRXRNSIGNUS LVNGHKIRNSTVORILDUERSIDILEUMRERIS IOK SIR SVUDINXRACHARINGEOACEU/RAMINYETUDIRXRNSIGNUS LVNGHKIRNSTVORILDUERSIDILEUMRERISTIGEITA
17. DkanCenpC4 18. DpecCenpC3 19. DaurCenpC3 20. DtriCenpC3 21. DtanCenpC3	
22. DrufCenpC3 23. DasaCenpC3 24. DlacCenpC3	IR RPKIRSTUVELNIKARISTELINGEGIAE EUVERMISTIVELUOTENKRISTUGUIVEV LIVEHRUR LANGRUDUEGAUGHEUMER-RRUET-WEGURUESWANDES EDDDDERUPPR-DUINNKRKTT- IR RPKIRSTUVELNIKARISTELINGEGIAE EUVERMISTIVELUOTENKRISTUGUIVEV VPLINGERRUR LALGRUDUEGANTORFEUMER-RRUET-WEGURUESWANDES EDDDDERUPPR-DUINNKRKTS- IR RPKIRSTUVELNIKARISTELINGEGIAE EUVERMISTIVELUOTENKRISTUGUIVEV VPLINGERRUR LALGRUDUEGANTORFEUMER-RRUET-WEGURUES IR RPKIRSTUVELNIKARISTELINGEGIAE EUVERMISTIVETUOTENKRISTUGUIVEV VPLINGERRUR LALGRUDUEGANTORFEUMER-RRUET-WEGURUESWANDES EDDDDERUPPR-DUINNKRKTS-
Consensus	710 720 720 720 720 720 720 720 720 720 72
Consensus Identity 1. DpunCenpC1	740 740 740 740 740 740 740 740 740 740
Consensus Identity 1. DpunCenpC1 2. DwatCenpC1 3. DkikCenpC1 5. DiecCenpC1 5. DiecCenpC1	η0 <
Consensus Identity 1. DpunCenpC1 3. DkikCenpC1 4. DkikCenpC1 4. DkikCenpC1 6. DkikCenpC1 6. DkikCenpC1 7. DkikCenpC1 9. DbakCenpC1 10. DjamCenpC1	70 70 <td< td=""></td<>
Consensus Identity 1. DpunCenpC1 2. DwatCenpC1 3. DiekCenpC1 4. DboCcenpC1 5. DiecCenpC1 5. DiecCenpC1 7. DwulCenpC1 9. DbakCenpC1 10. DjamCenpC1 10. DjamCenpC1 11. DsegCenpC1 12. DbunCenpC1 13. DsetCenpC1 14. DtruCenpC1 15. DbinCenpC1 15. DbinCenpC1 15. DbinCenpC1	70 70 <td< td=""></td<>
Consensus Identity 1. DpunCenpC1 2. DwatCenpC1 3. DkikCenpC1 4. DbocCenpC1 5. DiecCenpC1 5. DiecCenpC1 9. DbakCenpC1 9. DbakCenpC1 10. DjamcenpC1 13. DserCenpC1 14. DtruCenpC1 15. DbirCenpC1 16. DmayCenpC1 16. DmayCenpC1 16. DmayCenpC1 17. DkanCenpC4 18. DperCenpC3 19. DaurCenpC3 20. DtriCenpC3 20. DtriCenpC4 20. DtriCenpC4 20. DtriCenpC4 20. DtriCenpC4 20. DtriCenpC4 20. DtriCenpC4 20	10 70<
Consensus Identity 1. DpunCenpC1 3. DkikCenpC1 4. DbocCenpC1 5. DiecCenpC1 5. DiecCenpC1 5. DiecCenpC1 5. DiecCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DsegCenpC1 12. DbunCenpC1 13. DsetCenpC1 14. DtruCenpC1 15. DiecCenpC1 15. DiecCenpC1 16. DiecCenpC1 17. DkarGenpC1 18. DeecCenpC3 20. DtriCenpC3 21. DtarGenpC3 22. DruTCenpC3 23. DasaCenpC3 23. DasaCenpC3 24. DiaCenpC3 24. DiaCenpC3 25. DiaCenpC3 26. DiaCenpC3 27. DiaCenpC4 27. DiaCenpC4 27. DiaCenpC4 27. DiaCenpC4 2	10 10 <td< th=""></td<>
Consensus Identity 1. DpunCenpC1 3. DvikCenpC1 4. DivCompC1 4. DivCompC1 5. DvikCenpC1 5. DvikCenpC1 6. DvikCenpC1 7. DvikCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DsegCenpC1 12. DtrikCenpC1 13. DtrikCenpC1 14. DtrikCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC3 20. DrikCenpC3 21. DaasCenpC3 22. DrikCenpC3 23. DaasCenpC3 24. DlacCenpC3	10 70<
Consensus Identity 1. DpunCenpC1 3. DivikCenpC1 3. DivikCenpC1 4. DiveCenpC1 5. DivicCenpC1 6. DivitCenpC1 7. DvutCenpC5 8. DivikCenpC1 10. DjarnCenpC1 11. DisegCenpC1 12. DivitCenpC1 13. DivitCenpC1 14. DtruCenpC1 15. DivitCenpC1 15. DivitCenpC1 16. DmayCenpC1 17. DkarCenpC3 20. DtarnCenpC3 21. DiarCenpC3 22. DrutCenpC3 23. DasaCenpC3 24. DlacCenpC3 Consensus Identity 1. DpunCenpC1	10 10<
Consensus Identity 1. DpunCenpC1 2. DwatCenpC1 3. DkikKenpC1 4. DboCcenpC1 5. DiecCenpC1 5. DiecCenpC1 7. DwulCenpC1 10. DjancenpC1 11. DsegCenpC1 12. DburCenpC1 13. DsetCenpC1 14. DtruCenpC1 15. DbirCenpC1 16. DmayCenpC1 17. DkanCenpC3 19. DpriCenpC3 22. DruTCenpC3 24. DiaCenpC3 24. DiaCenpC3 24. DiaCenpC1 1. DpunCenpC1 25. DbirCenpC3 24. DiaCenpC3 24. DiaCenpC3 24. DiaCenpC3 24. DiaCenpC1 3. DbirCenpC1 3. Db	10 10 10 10 10 10 10 10 0 - 5 - SNN.SBRK F EES DDD - V-SBR EN FORTKRET Z - SP SSRBOHL NT - OTSGE HE LUIL PVB FINIEHDER RF PGBB (BS NFC Y SIB)
Consensus Identity 1. DpunCenpC1 2. DwatCenpC1 3. DkikCenpC1 4. DboCcenpC1 5. DiecCenpC1 5. DiecCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DsegCenpC1 12. DbutCenpC1 13. DsetCenpC1 14. DtruCenpC1 15. DbitCenpC1 15. DbitCenpC1 17. DsetCenpC1 18. DsetCenpC1 19. DbatCenpC1 19. DbatCenpC3 20. DtriCenpC3 21. DtatCenpC3 22. DtriCenpC3 23. DasaCenpC3 24. DlatCenpC3 24. DlatCenpC1 3. DkikCenpC1 3. DkikCenpC1 3. DkikCenpC1 5. DiecCenpC1 3. DkikCenpC1 5. DiecCenpC1 5. DiecCenpC1	170 1
Consensus Identity 1. DpunCenpC1 2. DwatCenpC1 3. Dkik/CenpC1 4. DbotCenpC1 5. DwatCenpC1 5. DwatCenpC1 9. Dbak/CenpC1 10. DjawCenpC1 11. DsegCenpC1 12. DbutCenpC1 13. DsetCenpC1 14. DsetCenpC1 15. DsetCenpC1 15. DsetCenpC1 16. DmayCenpC1 17. DkatCenpC1 18. DsetCenpC3 20. DtriCenpC3 21. DtriCenpC3 22. DrutCenpC3 23. DastCenpC3 24. DiaCenpC3 24. DiaCenpC3 24. DiaCenpC3 25. DastCenpC3 26. DastCenpC3 27. DtriCenpC3 28. DastCenpC3 29. DastCenpC3 20. DtriCenpC3 20. DtriCenpC3 20. DtriCenpC3 20. DtriCenpC3 21. DtarCenpC3 22. DwatCenpC3 23. DastCenpC3 24. DiaCenpC1 5. DkikCenpC1 5. DkikCenpC1 5. DwatCenpC1 5. DwatCenpC1 10. DjawCenpC1 10. Dja	C - 5 - SKRVSBRK FE BES DD D- MLEPPERPENERF DITK KET ZX - SF SSRRQ HP - X5 - ES SRRQ CTALEBERVER REPCESSR S NRC VSBR - MH-KKRISBRE F. NES DD DP - A - ST BL. LINF TDTK KR TE P - SKRSRRQ HF OT SO S SE SUBL RVBERNBERKER PCBSK S NRC VSBR - I HSKNRSR EF NES DD DP - A - ST BL. LINF TDTK KR TE P - SKRSRRQ HF OT SO S SE SUBL RVBERNBERKER PCBSK S NRC VSBR - I HSKNRSR EF NES DD DP - A - ST BL. LINF TDTK KR TE P - SKRSRRQ HF OT SO S SE SUBL RVBERNBERKER PCBSK S NRC VSBR - I HSKNRSR AR EF NES DD DP - A - ST BL. LINF TDTK KR TE P - SKRSRRQ HLK TS - QK S ED VIT LUBBPVLIR NUELBBANK REP CBSK S SNRC VSBR - SKRSK SKR VSB SKR FOLD A - ULEPPERPERUL FTDTK KR TE P - SKRSRRQ HLK TS - QK S ED VIT LUBBPVLIR NUELBBANK REP CBSK S SNRC VSBR - GS SKR VSB SKR FOLD A - ULEPPERPERUL FTDTK KR TE P - SKRSRRQ HLK TS - QK S ED VIT LUBBPVLIR NUELBBANK REP CBSK S SNRC VSBR - GS SKR VSB SKR FOLD A - ULEPPERPERUL FTDTK KR TE P - SKRSRRQ S SNRC VSB SKR VSB SNRC VS
Consensus Identity 1. DpunCenpC1 2. DwatCenpC1 3. Dkik/CenpC1 4. DioCenpC1 7. DwuTCenpC5 8. Dnik/CenpC1 10. DjamCenpC1 11. DsegCenpC1 12. DbutCenpC1 12. DbutCenpC1 13. DrutCenpC1 14. DrutCenpC1 15. DbitCenpC1 16. DmayCenpC3 20. DrtTCenpC3 21. DasaCenpC3 22. DasaCenpC3 23. DasaCenpC3 24. DlacCenpC1 5. DieCenpC1 5. DieCenpC3 24. DlacCenpC1 5. DieCenpC1 5. DieCenpC3 24. DlacCenpC1 5. DieCenpC1 5. DieCenpC1 5. DieCenpC3 24. DlacCenpC1 5. DieCenpC1 5. DieCenpC1	C - 5 - 5 NRX/BPRX F LES DD/A - MURPHPPIN F KDT:RKT Z A - 5P SSRRR0 1F OTSO 5 SS SD MC TALLEBARVE/NERMER/RK PREME 5 NR CYSTM min

Consensus Identity	
1. DpunCenpC1	
2. DwatCenpC1 3. DkikCenpC1 4. DbucCenpC1 6. DwulCenpC1 7. DwulCenpC5 8. DnikCenpC1 10. DjamCenpC1 10. DjamCenpC1 11. DsegCenpC1 13. DsetCenpC1 14. DtruCenpC1 15. DbirCenpC1 15. DbirCenpC1 16. DmayCenpC1 17. DkanCenpC3 21. DtamCenpC3 22. DrufCenpC3 23. Das3CenpC3 24. DlacCenpC3 24. DlacCenpC3	
Consensus Identity	1,000 1,270 1,000 1,000 1,000 1,000 1,100 1,100 1,100 1,100 1,100 1,100 1,100 1,100 1,100 1,100 1,000
1. Dpuncenpc1	TERESCAPERATE REVAILE REVEL REVEALE OF NAUVEEN LEE FUTURE IN A A REVELATION AND FILSE - MERCUNE - M
2. OwarCenpC1 3. OkiKenpC1 4. OboCenpC1 5. DieoCenpC1 6. DvulCenpC1 7. DvulCenpC1 9. DbakCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DisegCenpC1 13. DserCenpC1 14. DruCenpC1 15. DbirCenpC1 16. DmayCenpC1 16. DmayCenpC1 18. DpecCenpC1 19. DawCenpC1 19. DawCenpC3 19. DawCenpC3 22. DruTCenpC3 23. DasaCenpC3 24. DiaCenpC3	TERSETSTICE FEITWIC QUELE EN
Consensus Identity	1,770 1,180 1,790 1,270 1,270 1,270 1,270 1,270 1,270 1,270 1,270 1,270 1,270 1,270
1. DpunCenpC1	
2. DuarCangC1 3. DikKcenpC1 4. ObacCanpC1 5. DieoCanpC1 6. DvulCenpC5 8. DnikCenpC1 10. DjamCenpC1 10. DjamCenpC1 11. DibmCenpC1 13. DextCenpC1 14. DtruCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC3 20. DtriCenpC3 20.	

Consensus Identity	1,290	1,300	1,310 1, NK	320 1,330 C SE P 11 S D V T P	0 1,340 M∎E⊠	1,350	1,360 - YNEESDET	1,370	1,380 ETTAQSDA-	1,390 1,400 PMINESRLANSPKKSIAMK.
1. DpunCenpC1 2. DwarCenpC1 3. DikKenpC1 4. DbacCenpC1 5. DieoCenpC1 6. DvulCenpC5 8. DnikCenpC1 7. DvulCenpC5 8. DnikCenpC1 10. DjamCenpC1 11. DjamCenpC1 13. DertCenpC1 14. DruCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC1 15. DbirCenpC3 20. DruTcenpC3 21. DtamCenpC3 22. DruTcenpC3 23. DasaCenpC3 24. DlacCenpC3	PTESPYMEO PTESPYMEO SAKKSIHNOS SAKKSIHNOS STKKTSIHNOS STKKSIHNOS STKKSIHNOS STKKSIHNOS STKKSIHNOS	ASGMAPSRESE, AXAKA SQESKI, S AXAKA SQESKI, S AXAK SPS-GPP SAKPSKEPG- SAKPSKEPGPS AXKTSKEPGPSS AXKTSKEPGPS AXKTSKEPGPS	ND OD OD OD OD OD OD OD OD OD O		VIEICE VIEIC		- F (1932 S) D = T - - M (1932 S) N = T - - M (1932 S) D T - - M (1932 S) D = T - - C (1933 K) S A = T - - P (1932 F) D = T -	AQAP AQAPP Q	- ETAQ SNT- APAQAP SHV- APAQAP SHV- - ETAQAP SHV- - ETAQAP SHV- - ETAQSDAH SDA- - ETAQSDAH SDA- ETAQSDAH ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA ETAQSAA 	SWNERKLPISPKKGVA- PWNESRLASPTTSIAK PWNESRLASPTTSIAK PWNESRLASPTTSIAK PWNESRLASPTTSIAK PWNESRLESSKIIGK ISKISKIESSKIIG PWNESRLESSKEIA PWNESRLESSKEIA PWNESRLESSKEIA PWNESRLSSKEIA PWNESRLSSFEIA PWNESRLSSFEIA PWNESRLSSFKIISN PWNESRLSSKEIA PWNESRLSSKEIA PNNESRLSSKEIA SKAPAKKTISNOK LSRSTAKKKTISNOK LSRSTAKKKTISNOK LSRSTSAKKKTISNOK
Consensus Identity 1. DpunCenpC1	1,410 AISKITISIQEEEEL	1,420 QCVSSSQDKSTS QCVSSSQDKP I S	1,430 R L D IVIDNEQP G P VII VIIDSETP	1,440 STSLAARKAREI STSLAARTTRE	1,450 RRT OK DKAKSX- SRT OK NKAQSK-	460 1,470 VDDETTERK LDDETVERK	1,480 PUKPAPRAKSK PUMPAPRVKSQI	1,490 SKEUXK - UX IK SKEUDK LUK	1,500 HSVIIVR PITSL QSVIIVGH I SV	1,510 1, DGS I SSXCVRRSKRG
2. DwatCenpC1 3. DkikCenpC1 5. DieoCenpC1 5. DieoCenpC1 7. DwitCenpC5 8. DnikCenpC1 9. DbatCenpC1 10. DjamCenpC1 11. DsegCenpC1 12. DburCenpC1 13. DsetCenpC1 14. DrutCenpC1 15. DbirCenpC1 16. DmayCenpC1 17. DkarCenpC4 19. DbatCenpC3 20. DrtrCenpC3 21. DtarCenpC3 22. DrutCenpC3 23. DsatCenpC3 24. DiacCenpC3 24. DiacCenpC3		Q (VISSIS) OD KIS 1 Q (VISSIS) OD KIS 1 P (VISSIS) OD KIS 1 P (VISSIS) OD KIS 1 P (VISSIS) OD KIS 1 Q (VISSIS	- GP D UVUD N GO - GP E UVUD N GO - R L D UVUD N GO -	SUS L AAR KADRE SUS L AAR (KARE E SUS L AAR (KARE E SUS L AAR (KARE E SUS L AAR (KARE E SUS RAA (KARE E SUS RAA (KARE E SUS CAAR (KARE E SUS CAAR (KARE E SUS CAAR (KARE E SUS L AAR (KARE E)	SR TOM DK AQSK. RSKK EKTKKDK. SSKEETKKDK SSKEETKKDK DRALMCKAASO DRALMCKAASO DRALMCKAASO DRALMCKAASO DRALMCKAASO DRALMCKAASO DRALMCKAASO DRALMCKAASO ROWCKASS ROWCKASS ROWCKASS ROWCKASS SSTOKENKKTK RTOKENKKTK RSIQKENKKTK RSIQKENKKTK					
Consensus Identity 1. DpunCenpC1	1,520 1 HWP I KNTWCH	,530 1,54 HQDIM EMPEN - TQDIWDMPENR	0 1,550 A − − − K S I MQKAPK A − − − E L PMQRPK	1,560 	1,520 KISIKSV KISITPKISI	1,580 GKKBPTIAATTKKII	1,590 1. - N - P L P - DRGP - N - P L S - DRGP	500 1.6 VCSSTPRNPD	10 1,6 IVEPHOSA	20 1,630 IIG QHSAU IIC QHSAM
2. DwatCenpC1 3. DkikCenpC1 4. DbacCenpC1 5. DleaCenpC1 6. DvulCenpC1 7. DvulCenpC1 9. DbakCenpC1 9. DbakCenpC1 10. DjamCenpC1 11. DsegCenpC1 11. DsegCenpC1 12. DsegCenpC1 14. DtruCenpC1 15. DbirCenpC1 15. DbirCenpC1 16. DmayCenpC1 16. DmayCenpC1 19. DsauCenpC3 20. DtriCenpC3 21. DtamCenpC3 22. DrufCenpC3 23. DsasCenpC3 24. DlacCenpC3			A E L PMQR EX T K S I I Q R 23 S K S I I Q R 23 S K S I I Q R 23 A I CG S TT KIB Q R A CG S TT KIB Q R S K S I M PIRZ A K S I M Q R 25 A S G T V K R 25 S G T V K R 25 				$\begin{array}{c} -N- \frac{1}{2} L S- \frac{1}{2} S C \\ S-K- \frac{1}{2} Q S N S C \\ S-K- \frac{1}{2} Q S N S C \\ S-K- \frac{1}{2} Q S N S C \\ -N- \frac{1}{2} L - \frac{1}{2} S C \\ -N- \frac{1}{2} S C \\ -N$	VCSS11281128 VCSS1		UC
Consensus Identity 1. DpunCenpC1		1,650 E P S A E TD A III S H H E P F A E TD A III S H H	1,660 1 INFERVIS-EGQU	570 1,680 N S N N G N N	0 1,690 QFRINUILLQV RVK⊠ILLQV		1,710 RXAISLİVERMINI KGALSFLERMINI	1,720 IELDGIIDYAFY	1,790 BUKHTAALGY GTKETAAVGY	1,740 1,750 MREKPEQVENKKKVKA
2. DwatCempC1 3. DwatCempC1 4. DbacCempC1 5. DieoCempC1 6. DvulCempC5 8. DnikCempC1 7. DvulCempC5 8. DnikCempC1 10. DjamCempC1 10. DjamCempC1 11. DsegCempC1 13. DselCempC1 14. DruCempC1 15. DbirCempC1 15. DbirCempC1 15. DbirCempC1 16. DmayCempC1 17. DkanCempC3 20. DrufCempC3 20. DrufCempC3 22. DrufCempC3 23. DasaCempC3 24. DjacCempC3 23. DasaCempC3 24. DjacCempC3 23. DasaCempC3 24. DjacCempC3 25. DrufCempC3 25. DrufCempC3 26. DjacCempC3 27. DrufCempC3 27. DjacCempC3 27. DjacCempC4 27. DjacCempC3 27. DjacCempC4 27. DjacCempC4 27. DjacCempC4 27. Dja		E F.AETD AUSH E SAEK H 19170 PSAEK H 19170 PSAEK H 19170 PSAEK H 19170 PSAEK H 19170 PSAETD AUSH FFAETD AUSH PSA TD AUSH PSA TD AUSH PSA TD AUSH PSAETD AUSH PSAETD AUSH PSAETD AUSH PSAETD TUSH PSAETD TUSH PSAET		NR NN E S N NC NN CN S NC NN CN S NC NN CN S TS NP C NN TS NP C NN TS NN G K N TS NN G K N TS NN G K N TS NN G N S NN G N		SINN VOIDS E.R.S.NUIGE SINN VOIDS Y TH.DIGG SINN VO				

	1,690	1,760		1,270		00	17/90	1,600	,	1,610 1,	,614
Consensus Identity	IK VKA×I	PLKFIVQ	GEFAVQ	YGAIDA	- E E E E S (MXEIDKG	SRYSIQN	ILDAVIGV	LLVIRS	Ĥ
1. DpupCeppC1	KVKAS	PI I	GEEAVO	YGPIDG	- BDDBS0	HEIGD	MVEVDKG	SRYSION	DAVIGN	TIVEG	
r sparser per								5 10 10 10 10 10	THE FILL OF		
											۶.,
2 DwatCannC1	KVKACI		GEEAVO	VGPIDG	FEEEST		MVEVDKG	SRVSTON			
3 DividConpc1	KVKESI		GEEAVO	VGALDN	EFFFSC	DECO		EPVSION	TDAWG		
4 DhocCenpC1	KVKESI		GFFAVO	VGAVGS	FFFFSC	REGO	MIFIDIG	EWVSTON			
5 Diecemper	KVKESI	HVKELVÖT	GEEAVO	VGALDN	KEEEESC	REGO	MIFINIG	EWVSTON			
6 DvulCenpC1	KVKGCI	ΡΙΚΕΙΫΟ	GEEAVO	YGSIDD	- F F F F S C	TIREGO	MVEIDKG	SRYSION	DATON	TIVES	п.
7. DvulCenpC5	KIKASI	PLKFIVÔ	GEEAVO	YGALEO	- FFFFSC	REGD	MVFIDKG	NOYSTON	IDATON	LLVR	
8. DnikCenpC1	KIKACI	PLKFVVQ	GEFAVO	YGDIDQ	- EEEESC	REGD	MVEIRKG	SLYSION	DAGG	LMVIRG	
9. DbakCenpC1	KVKDC	PLITFIVOI	ĞEFAVÖ	YGTNDE	- EEEESC	MIREGO	MIEINKG	NRYSION	DAG	LLVIRK	
10. DjamCenpC1	KIKGY	PLKFIVQ	GEFAVQ	YGAIDE	- EEEESC	LHEGD	MVEIDKG	NRYSIQN	ILDTIGN	LLVIRN	
11. DsegCenpC1	KVKAC	ΡΙΚΓΙΧΟ	GEFAVQ	YGAIGE	- EEVESC	LREGD	MVEIDKG	CRYSIQN	ILDAIGN	LLVIRS	
12. DbunCenpC1	KVKACI	P L K F V V Q	GEFAVQ	YGAING	- E E E E S C	REGD	MVEIDKG	SRYSMON	IDAGG	LLVIRS	
13. DsetCenpC1	KVKACI	P L K F V V Q 🛽	GEFAVQ	YGAIDG	- E E E E S C	LILREGD	MIEI <u>DKG</u>	SRYSIQN	IDAGO	LLVIRS	
14. DtruCenpC1	KVKANI	ΡΙΚΕΥΥΟ	GEFAVQ	YGAIDA	- E E E E S C	LREGD	MVEIDKG	RYSIQN	ILNAIGN	ILVIRS	
15. DbirCenpC1	KVKESI	P L K F I V Q 🛛	GEFAVQ	YGVIDA	- E E E E S C	LREGD	MVEIDKG	RYSIQN	VILDA GV	LLVIRS	
16. DmayCenpC1	KVKES/	ALKFIVQ	GEFAVQ	YGAINV	- EEEESC	LILREGD	MVEIDKG	RYSIQN	VILDA GV	ILVIRS	
17. DkanCenpC4	KIKAYI	PLKFUVQU	GEFAVQ	YGPVDG	- DEDESC	IIRVGD	MVETEKG	YRYSIQN	ILDDUGN	LLVRS	
18. DpecCenpC3	RAKVN	PLKFIVYL	GEFAVQ	YEEVDA	- EEEKISS		MIEIDKG	VRYSIQN	ILNDMSM	LLVRC	
19. DaurCenpC3	RAKAH	PLKFIVQE	GEFAVQ	QQGMDA	- EEEESC	RAGD	MIEIDKG	SRYSIQN		LLVIRC	
20. Dtricenpc3	RARAH		GEFAVQ	QQGMDA	EEESC	RAGD	MIEIDKG	SRYSIQN			
21. DtanCenpC3	KAKAY		GEFAVQ	RODLDA	- EEESC	HAGD	MIEIDKG	SRYSION			
22. Drucenpc3	RAKAH		GEFAVQ	RODLDA	EDEESU		MIEIDKG	SRYSIQN.			
23. Dasacenpc3	R A R A H		GEFAVQ	RODLDA	FEESC			SIGN SIQN			
24. Diaccenpcs	RANAH	PLKFIVQE	GEFAVQ	RODLDIA		. I L H A G D	MICIUNU	SRYSIQN		LLVIRC	

Figura 18. Alinhamento amplificado das proteínas Cenp-C das espécies do grupo *montium* mostrando a posição dos motivos proteicos identificados no presente trabalho. As barras em vermelho representam nesta mesma ordem: Mis12 *binding*, R-rich, DH, NLS, CenH3 *binding* e Cupin. A barra azul-claro representa as sequências removidas para a construção da filogenia dos genes e teste de seleção positiva. A pequena barra em verde-escuro abaixo do motivo CenH3 *binding* representa o aminoácido R-1101 idêntico em todas as proteínas.

4.3. As duas cópias da Cenp-C no genoma de D. vulcana

A presença de duas cópias da *Cenp-C* (*Cenp-C1* e *Cenp-C5*) em *D. vulcana* desempenhando a mesma função permitiria teoricamente o relaxamento das pressões seletivas em ao menos uma delas, com subsequente pseudogenização. Porém, várias evidências apontam que ambas as cópias são funcionais. De acordo com o algoritmo *Augustus*, não há presença de stop códons prematuros ao longo da sequência de nenhum dos dois genes. Além disso, há uma maior identidade (81,9%) entre as proteínas codificadas pelo gene da *Cenp-C1* e *Cenp-C5* quando comparadas às proteínas das outras espécies do grupo *montium* (Figura 19). Essa identidade varia de 38,9% (*Cenp-C3* de *D.tani*) a 68,3% (*Cenp-C1* de *D. jambulina*) de identidade em relação à *Cenp-C5*. Por último, seis dos motivos essenciais à função do centrômero estão presentes na sequência da proteína de *Cenp-C1* e *Cenp-C5*, assim como na *Cenp-C* das demais espécies. Em conjunto esses achados apontam para a retenção da função centromérica em ambas as cópias de Cenp-C em *D. vulcana*.



Figura 19. Alinhamento das proteínas codificadas pelo gene da Cenp-C1 e Cenp-C5 de D. vulcana. O alinhamento das sequências dos contigs mostra regiões idênticas (preto), divergentes (cinza) e de gaps

(branco). A barra em azul representa a região deletada da Cenp-C5 para construção da árvore filogenética.

O algoritmo *Augustus* identificou três éxons adicionais no gene da *Cenp-C5* (Figura 20). A sequência de aminoácidos codificada por esses éxons possui homologia com as sequências de aminoácidos encontrados nos éxons de todas as outras espécies do grupo *montium.* Porém, nenhum motivo proteico foi identificado nessa região. Futuros estudos elucidarão se a presença ou ausência desses éxons afetam a função do gene da *Cenp-C5*.



Figura 20. Alinhamento do gene da *Cenp-C1* e *Cenp-C5* de *D. vulcana*. O alinhamento das sequências mostra regiões idênticas (preto), divergentes (cinza) e de gaps (branco). As barras em cinza mostram sequências de íntrons inferidos pelo algoritmo *Augustus* e por alinhamento com os genes da *Cenp-C* das outras espécies do grupo *montium*. A barra em azul representa a região deletada da *Cenp-C5* para construção da árvore filogenética.

Embora vários fatores apontem para a manutenção da função centromérica da *Cenp-C1* e da *Cenp-C5* em *D. vulcana*, serão necessários estudos experimentais futuros para determinar se ambas as cópias codificam proteínas que se localizam no centrômero. A essencialidade de ambas as proteínas em função centromérica poderá ser verificada por nocaute dos genes. Também poderá ser analisado de que maneira essas proteínas interagem com as cópias de Cid encontradas na espécie.

4.4. Relações evolutivas entre as cópias da *Cenp-C* nas espécies do grupo *montium*

Para entendermos as relações evolutivas entre as cópias de Cenp-C encontradas, construímos uma árvore MV contendo todas as sequencias de *Cenp-C* inteiras obtidas no presente trabalho (Figura 21). A topologia da árvore indica que todas as cópias se agrupam conforme os subgrupos propostos por Yassin (2018) (Figura 21). Esta topologia também está de acordo com a robusta filogenia do grupo *montium* proposta por Conner *et al.* (2021) (Figura 9). Portanto, esse resultado indica que, no geral, as relações evolutivas entre as cópias dos genes da *Cenp-C* (Figura 21), refletem a relação evolutiva entre as espécies.



Figura 21. Árvore de máxima verossimilhança mostrando as relações filogenéticas entre as cópias do gene da *Cenp-C* nas espécies do grupo *montium*. O nome dos subgrupos e suas respectivas espécies são mostrados em cores iguais. O valor de *bootstrap* está indicado em cada nó. A escala indica o número de substituições por sítio.

Não existem dúvidas de que o loco de *Cenp-C1* representa a condição primitiva de *Cenp-C* dentro do gênero *Drosophila*. Neste loco, o gene *Cenp-C1* é flanqueado pelos genes *CG1427* e *Tim10*. Nas espécies do subgrupo *montium*, as sequências flanqueadas pelos genes *CG1427* e *Tim10* não possuem homologia com *Cenp-C1* e possuem tamanhos variando de 978 pb (*D. pectinifera*) a 3.051pb (*D. asahinai*). No entanto, o tamanho da *Cenp-C3* varia de 5.610pb (*D. pectinifera*) à 8.750pb (*D. tani*). Esses dados mostram uma possível deleção de *Cenp-C1* das espécies do subgrupo *montium*. Em resumo, a origem da *Cenp-C3* ocorreu depois da separação do subgrupo *montium* dos demais subgrupos, com posterior deleção da *Cenp-C1*.

De todas as espécies analisadas, apenas *D. kanapiae* possui a *Cenp-C4*. Com base nesse dado, provavelmente, essa cópia se originou pelo menos depois da separação do subgrupo *parvula* dos demais subgrupos. Assim como no caso do gene *Cenp-C3*, o gene *Cenp-C4* pode ter se originado a partir de uma duplicação da *Cenp-C1*, seguida de posterior deleção de *Cenp-C1*. Isso porque a sequência contida entre os genes *CG1427* e *Tim10* do loco de *Cenp-C1* em *D. kanapiae* contém apenas 585 pb, enquanto que a *Cenp-C4* possui 8.013 pb. Nesse caso, um indício de duplicação seguido de deleção está mais evidente,

devido à presença de um fragmento do gene *CG1427* no loco onde se encontra a *Cenp-C4* (Figura 9). Esse fragmento está flanqueando a posição 5' da *Cenp-C4* e na orientação esperada, caso tenha se derivado da duplicação do *CG1427* do loco original.

O agrupamento dos genes da *Cenp-C1* e *Cenp-C5* de *D. vulcana*, com alto valor de *bootstrap*, indica que *Cenp-C5* pode ter resultado da duplicação do gene da *Cenp-C1* dentro da linhagem que deu origem a essa espécie (Figura 21). Tendo como base o tempo de divergência estimado por Yassin *et al.* (2016), essa duplicação ocorreu em um passado relativamente recente, há menos de 5 milhões de anos.

Além das três novas cópias inteiras da *Cenp-C* identificadas, duplicações de vários fragmentos do gene também foram observados em sete espécies do grupo *montium*. Em *D. burlai*, pelo menos quatro fragmentos do gene da *Cenp-C* duplicados estão presentes em diferentes localizações genômicas (Figura 13). Porém, como explicado anteriormente, é possível que um deles represente uma cópia inteira do gene da *Cenp-C* que não pôde ser montada, devido às limitações do sequenciamento. Em *D. punjabiensis*, foram encontrados pelo menos dois fragmentos em diferentes localizações genômicas (Figura 13). Em *D. vulcana* e em *D. bakoue* o fragmento encontrado também ocupa uma localização genômica distinta. E por último, em *D. kanapiae*, *D. watanabei* e *D. nikananu* embora nenhum fragmento esteja flanqueado por genes que indiquem seus locos, outros fragmentos duplicados foram encontrados.

Como mencionado anteriormente, a *Cenp-C3* e a *Cenp-C4* substituem funcionalmente a *Cenp-C1* em algumas espécies do grupo *montium*. Por isso, é esperado que em uma filogenia a *Cenp-C1*, *Cenp-C3* e *Cenp-C4* dessas espécies formem um grupo monofilético dentro do subgênero *Sophophora*. Para confirmar essa hipótese foi construída uma árvore ML, contendo a *Cenp-C* das espécies do grupo *montium*, a *Cenp-C1* de outras espécies do subgênero *Sophophora* e a *Cenp-C1* e *Cenp-C2* de espécies do subgênero *Drosophila* (Figura 22). Todos os genes de *Cenp-C* das novas espécies incluídas para construção da árvore, foram anteriormente estudas por nosso grupo (Teixeira *et al.* 2018). Como esperado, conforme a filogenia das espécies do gênero *Drosophila*, a *Cenp-C1* das demais espécies do grupo *montium* formam um grupo monofilético que agrupa com a *Cenp-C1* das demais espécies do subgênero *Sophophora*.



Figura 22. Árvore de máxima verossimilhança mostrando as relações filogenéticas entre as cópias do gene da *Cenp-C* nas espécies dos subgêneros *Sophophora* e *Drosophila*. O valor de *bootstrap* está indicado em cada nó. A escala indica o número de substituições por sítio.

Em resumo, todas as três novas cópias (*Cenp-3*, *Cenp-C4* e *Cenp-C5*) derivam de duplicações da *Cenp-C1* que é o gene ancestral da *Cenp-C* nas espécies do gênero *Drosophila* (Figura 23). A topologia da árvore em acordo com a filogenia das espécies é o resultado esperado para genes duplicados, que agora assumem o papel funcional dos seus respectivos parálogos (*Cenp-C1*) perdidos independentemente em cada subgrupo. Não só duplicações completas, mas também parciais da *Cenp-C*, ocorreram de maneira recorrente nas espécies do grupo *montium*. E por último, a filogenia do gene da *Cenp-C* das espécies do subgênero *Sophophora* e do subgênero *Drosophila* corroboram com a afirmação de que a *Cenp-C3* e *Cenp-C4* sejam ortólogos funcionais da *Cenp-C1* nas espécies do grupo *montium*.



Figura 23. Hipótese de origem e perda dos genes da *Cenp-C* nas espécies do grupo *montium*. Na espécie ancestral que deu origem às espécies do grupo *montium*, a *Cenp-C1* foi duplicada e inserida em um novo loco dando origem à *Cenp-C3*. Em seguida, a *Cenp-C1* foi deletada do loco ancestral. Na espécie ancestral que deu origem às espécies do subgrupo *parvula* ou dentro do subgrupo, a *Cenp-C1* e o gene *CG1427* foram duplicados e inseridos em um novo loco. Em seguida, houve a deleção da *Cenp-C1* no loco ancestral e degeneração do gene *CG1427* no novo loco. E por último em *D. vulcana* houve a duplicação da *Cenp-C1* que foi inserida em um segundo loco dando origem a *Cenp-C5*, ambas as cópias foram mantidas.

4.5. Origem das duplicações do gene da *Cenp-C* nas espécies do grupo *montium*

Duplicações gênicas podem resultar de diferentes mecanismos, como: *crossing-over* desigual, retrotransposição, duplicação cromossômica/genômica que são raras em animais

(Zhang 2003) ou por transposição por intermédio de transposons de DNA (Cerbin & Jiang 2018).

O *crossing-over* desigual geralmente produz cópias em tandem (Zhang 2003). Por isso, se esperaria que um gene duplicado como consequência desse processo estivesse flanqueado pelas mesmas sequências que a cópia que lhe deu origem. Porém, como mostrado anteriormente, cada uma das cópias da *Cenp-C* se encontra flanqueada por genes diferentes (Figura 9), e a Figura 13 mostra que estas cópias se localizam em diferentes locos do cromossomo 3.

A duplicação por retrotransposição pode ocorrer quando um mRNA proveniente de um gene é retrotranscrito em DNA complementar (cDNA) e em seguida inserido no genoma, gerando uma retrocópia do gene que, ainda que raramente, pode ser funcional (Zhang 2003). Algumas assinaturas moleculares são esperadas para o gene duplicado pelos mecanismos citados acima, como: perda de íntrons e sequências regulatórias, presença de cauda poli-A e presença de repetições curtas e diretas flanqueadoras. Nesse caso, esses mecanismos não podem explicar as duplicações do gene da *Cenp-C1*, pois todas as cópias analisadas mantiveram os íntrons em suas sequências.

Retroelementos podem também estar associados a duplicações gênicas por meio do processo de transdução. Esse processo é caracterizado pela captura de um gene hospedeiro como parte do elemento de transposição (Feschotte & Pritham 2007). A transdução, geralmente, ocorre quando o sinal fraco de poliadenilação de um retroelemento, inserido a montante de um gene, é pulado pela maquinaria de transcrição de RNA que, em seguida, reconhece o sinal de poliadenilação no final do gene. O transcrito gerado contendo a sequência do retroelemento e do gene é depois integrado ao genoma. Como esse processo foi responsável por duplicar um gene inteiro três vezes em humanos (Xing *et al.* 2006), seria possível que o mesmo pudesse ser responsável pela duplicação do gene da *Cenp-C1*. No entanto, nesse caso, todas as três duplicações do gene possuíam como característica a perda de íntrons. Outras características observadas nesses casos é a presença de sequências flanqueadoras pertencentes ao loco de origem, presença de cauda poli-A e de repetições curtas e diretas flanqueadoras.

Um segundo processo de transdução gerou uma duplicação atípica de uma sequência de 24.7kb envolvendo um retrotransposons do tipo LTR (Repetição longa terminal) em tomates (Xiao *et al.* 2008). As duas LTRs localizadas nas extremidades de um retrotransposon do tipo LTR são necessárias para o processo de transposição. Foi proposto que o transcrito do retrotransposon se iniciou em sua 5` LTR, mas ao invés de parar na 3`LTR, foi finalizado em uma terceira LTR que flanqueava a região de 24.7kb duplicada. Quatro genes foram duplicados durante o processo, dois deles estavam em orientação antisenso em relação ao retrotransposon e por isso tiveram as sequências de íntron conservadas. No loco duplicado

foram observadas sequências de LTRs flanqueando a sequência duplicada na região 5' e 3' e também repetições curtas e diretas flanqueadoras. Como ocorrido nesse caso, não podemos descartar a possibilidade de que os íntrons das duplicações da *Cenp-C1* tenham sido mantidos em processos de transdução, devido à orientação antisenso do gene em relação ao elemento. Considerando que o retrotransposon e as cópias dos genes duplicados foram mantidos também no loco de origem, é provável que se trate de um evento recente. Por isso, como explicado mais a frente, *D. vulcana* foi a melhor opção para investigar uma possível duplicação envolvendo retrotransposons.

Embora os mecanismos de duplicação via transposons de DNA ainda não sejam claros, evidências sugerem que esses elementos carreguem mais frequentemente genes ou fragmentos de genes do que os retrotransposons (Cerbin & Jiang, 2018). Isso poderia explicar, por exemplo, os diversos fragmentos da *Cenp-C* observados nas espécies *D. burlai*, *D. watanabei* e *D. punjabiensis*. Porém, diferente dos retrotransposons, a identificação da duplicação por elementos de DNA é mais difícil, pois exige o reconhecimento de estruturas intactas (como suas repetições invertidas terminais e as duplicações geradas no sítio alvo) associadas a eles (Cerbin & Jiang, 2018). Além disso, a falta de pressão evolutiva sob as sequências dos transposons causa divergência que pode dificultar sua detecção. Por exemplo, a maioria dos Pack-MULEs (transposons de DNA não-autônomos) ficam irreconhecíveis dentro de 6 milhões de anos (Zhao *et al.* 2018).

Dada a dificuldade de se avaliar eventos antigos de duplicação por intermédio de transposons, procurou-se por evidências de transposons associados à origem de *Cenp-C5* que, entre as cópias encontradas, é a de origem mais recente (~5 milhões de anos). Assim, através da base de dados de elementos repetitivos Repbase, foi feita uma busca por transposons flaqueando as cópias dos genes da *Cenp-C1 e Cenp-C5* de *D. vulcana*. Detectou-se cópias de um elemento de transposição (TE), similar ao Harbinger-2_DRh, flanqueando os dois parálogos (Figura 24A e C). Duas dessas cópias flanqueiam *Cenp-C1*, uma na posição 5' e outra na posição 3', e uma terceira cópia flanqueia *Cenp-C5* na posição 3' (Figura 24C). As três cópias deste elemento Harbinger-2_DRh-*like* conservam a repetição terminal invertida (TIR) de 20 pb em apenas um dos lados da sequência do TE e com identidade de 75% com a TIR (20 pb) do Harbinger-2_DRh.



Figura 24. (A) Representação do elemento TE Harbinger-2_DRh. Setas em preto indicam TIRs e retângulo em cinza indica a sequência interna às TIRs. (B) Evento de transposição por mecanismo *cut and paste* esperado para duplicação gênica. 1. Duas cópias do TE se insere em ambos os lados da *Cenp-C1.2*. Uma das TIRs de cada TE sofre mutações impedindo a ligação da transposase. Os elementos são reconhecidos pela mesma transposase e são transpostos para um segundo loco levando a sequência interna a eles. (C) Representação dos locos da *Cenp-C1* e da *Cenp-C5* com suas sequências flanqueadoras, observados no genoma de *D. vulcana*. Pontas de seta indicam a posição 3' dos genes. Fragmentos homólogos (65 a 74% de identidade e 54 a 152pb) ao Harbinger-2_DRh indicados pelo Repbase estão em cinza e preto. Os pontilhados delimitam a região de homologia (85 a 95% de identidade e 189 a 190pb) entre as três cópias do elemento Harbinger-2_DRh-*like*. Linhas pretas e barras em bege indicam, respectivamente, sequências intergênicas e sequências com homologia (211pb e 206pb) entre os dois locos. (D) Representação do que seria esperado para os locos da *Cenp-C* de *D. vulcana*, se um elemento Harbinger-2_DRh-*like* estivesse envolvido no processo de duplicação.

Os Harbingers compreendem uma superfamília de transposons de DNA que se transpõem pelo mecanismo "*cut and paste*" (Feschotte & Pritham, 2007). Esta transposição se inicia pela associação da transposase com as repetições terminais invertidas (TIR) do elemento (Figura 24A) (Skipper *et al.* 2013). Sendo assim, uma hipótese que explicaria a duplicação da *Cenp-C* em *D. vulcana* seria a de que duas cópias de um elemento do tipo Harbinger teriam se inserido em ambas extremidades 5' e 3' do gene da *Cenp-C1* (Figura 24B.1). Em seguida uma das TIRs de cada uma dessas cópias teria sofrido mutações que impediram a ligação da proteína transposase a essas sequências (Figura 24B.2). Desse modo, os dois elementos Harbinger poderiam ter sido reconhecidos pela mesma transposase. A consequência disso é que a sequência interna a essas duas cópias de TEs, no caso o gene *Cenp-C1*, seria transposto para um segundo loco. Neste caso, seria esperado encontrar duas cópias do Harbinger, uma em cada extremidade da *Cenp-C5* (Figura 24D), mas apenas uma foi encontrada. Em segundo lugar, a sequência entre *Cenp-C1* na posição 3' e a cópia do

Harbinger, e a sequência entre a *Cenp-C5* na posição 3' e a cópia do Harbinger, deveriam ser homólogas, o que não foi observado. Por isso, os dados levantados até o momento não permitem associar o Harbinger-2_DRh-*like* com a origem de *Cenp-C5*.

Em resumo, embora nossas análises conseguiram descartar alguns mecanismos responsáveis pela origem das duplicações de Cenp-C no grupo *montium*, não conseguimos determinar precisamente qual ou quais foram os mecanismos envolvidos.

4.6. Teste de seleção positiva nas cópias de Cenp-C

Para determinar se a *Cenp-C* evolui sob seleção positiva foram utilizados os modelos de sítios-aleatórios (M1a vs M2a, M7 vs M8 e M8a vs M8) que partem do princípio de que *a priori* não sabemos quais sítios da proteína estão conservados ou sob seleção positiva. Como a maioria das espécies possui apenas uma cópia da *Cenp-C*, é esperado que nessas espécies a *Cenp-C* exerça uma mesma função e, portanto, esteja sob efeito das mesmas pressões seletivas.

Em *D. vulcana* dois genes aparentemente funcionais foram encontrados, não sendo possível determinar se algum deles evolui sob o mesmo padrão dos genes das demais espécies. Por esse motivo, a *Cenp-C1* e *Cenp-C5* de *D. vulcana* foram excluídas das análises.

A análise de seleção nos 22 genes da *Cenp-C* de espécies do grupo *montium* revelou quatro aminoácidos, localizados entre motivos proteicos, evoluindo sob seleção positiva (Figura 25 e Figura 26). Esse resultado foi considerado significativo para os modelos M7 vs M8 e M8a vs M8 (Tabela 3). O primeiro aminoácido se encontra na posição 354 entre os motivos R-rich e DH, o segundo e o terceiro se encontram nas posições 624 e 639, respectivamente, entre o motivo DH e NLS e o quarto se encontra na posição 720 entre os motivos NLS e CenH3 *binding*.

Conforme a hipótese do impulso centromérico é esperado que os resíduos da Cenp-C que fazem contato com o cenDNA evoluam sob seleção positiva mitigando possíveis efeitos deletérios associados à expansão desse cenDNA. Nenhum estudo até o momento identificou quais aminoácidos da Cenp-C estão diretamente em contato com o DNA centromérico em *Drosophila*. No entanto, é possível que esses quatro aminoácidos estejam evoluindo sob seleção positiva e suprimindo o impulso centromérico, como observado por Talbert (2004) para a Cenp-C de gramíneas e mamíferos. Como em nossas análises, regiões da *Cenp-C* fracamente alinhadas foram deletadas, é possível que esses quatro aminoácidos não sejam os únicos evoluindo sob seleção positiva nas espécies do grupo *montium*.



Figura 25. Alinhamento das proteínas Cenp-C das espécies do grupo *montium* mostrando os sítios da proteína que apresentaram assinatura de seleção positiva. As barras em vermelho representam da esquerda para a direita os motivos: Mis12 *binding*, R-rich, DH, NLS, CenH3 *binding* e Cupin. A barra azul-claro representa as sequências removidas para a construção da filogenia dos genes e teste de seleção positiva. As pequenas barras em azul-escuro representam os sítios sob seleção positiva.

1. A	1,340 1,350	1.360 1,370	1,380 1,3	390 1,400	1,410 1	.420 1.430	1,440	1,450	1,460 1,	470 1,41	1.4	90 1.50	0 1,510	1,520	1,530	1,540
Identity	- and -		and and				110								a di la c	-
1. DpecCenpC3 Frame 1	TOTO COLLEGE ALLOSSES COLLEGE AL		S C S P A	GCAAATTGTCCTTCA A N C P S	TCGCCA <mark>MEEGCAAA</mark>	TTOTCCOTENTER	AGCGGCAACAG			PAK-	51 		PKS5	RTACACECT	NERA CENSION O	GGC AN M A
2. DaurCenpC3 Frame 1	TORE CRIMES ALEGER CORDURES	T V V S S A	ITRICTURECRECAGE		A A S	TTOTCCTTORCORDO	4		UTCO RE RECED					RATCACEA	NIAT ING C ITO	G N V D
 DtriCenpC3 Frame 1 	TTECHINEALINERACION	TVV55A	S P P A		A A S	TTGCCCTTEDCECE	6		ETCORECTO	PAK			PRSS	R S G		G N V D
 DtanCenpC3 Frame 1 	DECEMBERATER	T V V 5 S A					4,		ETCOLOR	PAK	B(A S	PR 5 5	R 5 G		GGGAACEIGER
 DrufCenpC3 Frame 1 	COLOR COLLEGA A COLLEGA COLLEGA A COLLEGA A COLLEGA COLLEGA A C	T V V S S A	UTRICCERATECONSTRUCTION			TARCE (BICARE)	4	Rigtto		PAK-			RACENCICTED			G N V D
 DasaCenpC3 Frame 1 	PLHRAS	TVVSSA			A	1000 (00000) 	•	A L		PAK			PRST	R S G		G N V D
7. DlacCenpC3 Frame 1	DECENTEATORIGUENDE PLHRAS	T V V 5 5 A	S L M P A		a a a a a a a a a a a a a a a a a a a	TOR COLOR	4			PAK-			PR 5 5	R 5 G		G N V D
 Brame 1 	FORETORICE ADDRESS COND. GALER	T V V S S D	S C P S A		A G D	TTATCATTERTEGTE	4	REGOUNA A	CTGC COLORD	PAT-	5	CGATERATE T	PKSS	R A S	S L V V	
9. DpunCenpC1 Frame 1	P Y R S S	TEGPCA			A A N	V S S P	r			PAK			P K S S	R K S		N N N D
10. DwatCenpC1 Frame 1	P P R S S	TERS PCA	A L A S C ARTTAGRE		REGGCAAA	TTATAATIGEIGITEG	r	RD & GGA		PAK-	50		PKS5	R K 5		AACCENTINGARD
11. DkikCenpC1 Frame 1	FOR TOUR CARDING AND A CARDING	AGC	ANNEGTARCTINN		A A N	TAATAACTERTEGER	••••••••••••••••••••••••••••••••••••••	GACI	P T S	PAK-		E GAR GZ C ER T	PKSS	RKS		A GENATET GER
12. DbocCenpC1 Frame 1	PLRRTS	мс Т	ARCAACCTIC		A A N	TAATAACGENTEGER	• 	A I	CIGCICUTIC V P A 5	PAK		EGAGONGE T	P K S S	R K 5		AGCANTELO
13. DleoCenpC1 Frame 1	FOR TOILING CITCLES ACIANCIES	806С	AND A		TÜGCAAA	TAATAACTINGTOGOO	4	Riggo A A		PAK-	50	CGAA 920 CT	PKSS	RKS		A GORANTING RE
14. DbakCenpC1 Frame 1	PLHRAS	T E V A S A	S L Q L A		A A N	TTATAATTERATEGER	T	GAGI	A A A S	PAK-			PKTS	R T S		N N N D
15. DnikCenpC1 Frame 1		GARETGOCHICIGO E V P S A			CONGCARA	TTATAATIGRIGGIG YN SSS	T	Gili GGC/ A A		PAK	90 		PK 55	R K 5		N N V D
16. DjamCenpC1 Frame 1	FOR TOILING AND TRACOLATORS	T F V A S T			A A N	TTATAATTERTEGER	T	A A		QAK-	.	сстания в	FTAAD ADD TO	RKS		N N V D
17. DsegCenpC1 Frame 1	PLHRTS	T E V A S S	S L Q L A		A A N	TTATAATIGETGGGGG	T	A A		PAK		L E S	PKSS	R K S		N N V D
18. DbunCenpC1 Frame 1	PLH R A 5	TELHSC	S L Q P A		A A N	TTATAATGERIGGE Y N A 5 2	T	REGGA		PAK	90 		PK 5 5	R T 5		N N V D
19. DsetCenpC1 Frame 1	TOES TOESEE ALEGERS CONSERVE P I H R A S	T E V H S A	A 4 Q 1 A 20 D 1 A 20		A A S	TTATAATTITTEGES	T	A V	STOTELES	PAK-	50 		EAGADECADET	R K S		N N L D
20. DtruCenpC1 Frame 1	PVH R A S	TEVPTA	A L O L A		S G N	D N S L P	T	AA		PAK			PKSS	R K S		N N V D
21. DbirCenpC1 Frame 1		T Q D P S A	A L Q L T		A		4,			PAK-	50		PKSS	R N S		GACANTERGER
22. DmayCenpC1 Frame T	PLHCAS	T Q V S S A			A A N	π			S D S	PAK-			P K S S	R N S		N N V D

1. B

	980	990	1,000	1,010	1,020	1,030	1,040	1,050	1,060	1,070	1,080	1,090	1,100	1,110	1,120	1,130	1,140	1,150
Identity																		
1. DpecCenpC3 Frame 1	AZADGRABORK NA	R A	S	ECT GGTACOC	S P L	NAHORAGATHOAA H R A S		CAGNERSIAGO PAAA		A A S	PAK	INCRUTAAAAN SPK	S R	GL	V V G	CANGECARA MAD	GATIZA ISAG M T	D D E
 DtriCenpC3 Frame 1 		R E R T	Q P S		S S L	H R A S		CAGULESULCO	A L V	PAS	P A K	DICODICASALO	CARRENA TT	G L L	UDIGTEGEGAR		CATAATAAQ T	D D E
 DaurCenpC3 Frame 1 	A/(enige Aleree) TA				S P L	H R A S		CACENDERNIC	ABBATTACT	PAS	P A K	S L R	CMBC CCATO	G L L				D D D
4. DtanCenpC3 Frame 1	PTE	R E H A			<mark>PL</mark>	Q R T S			A	A A P	P A K	S P R	CARCERCTO S R	GGGACHECHE	V V G N		CATAA IGAC	D D E
5. DrufCenpC3 Frame 1		R E H A	Q R S	V V A	S P L	HALOROGORALORS		CAGULRSULCO	ABBIGTTORT	A A S	P A K	F P R	CTROBECTO C R	G L L	V M G N		CATAA IGAC	D D E
6. DlacCenpC3 Frame 1			Q P S	V V A	S P L	H R A S		PAA P		A A S	P A K	S P R	CABCECTO S R	G L L	M G N			D D E
7. DasaCenpC3 Frame 1	PACENICE ACTION	R E H A	Q P S	V V A	S P L	H R A S		P A A P		A A S	P A K	S P R		G L L	V V G N			D D E
8. DkanCenpC4 Frame 1			S	A V A		H R A E	T	S A A S	A A V	A A S	PAT	S P K	S R	A ABETTOPTILE				DDE
9. DwatCenpC1 Frame 1		M A H	Q P P	V E A	S P	Y R S S			A G V	A A S	PAK	S P K	S R	S				DDE
Frame 1		M A H	Q P P		S P	Y R S S					PAK	S P K	SR				M T	G D E
Frame 1		V A H	Q P S	V L A	S P L	RRTS		P S A F		PTS	PAK	S P K	SR		V P S N	V D A	MT	DDE
Frame 1		V A H V	Q P S	GLA	S P L	RRTS		P S V P		P A S	P A K	S P K	SR			V D A		DDE
Frame 1		V A H A		PETGGTAPPE	P	R R T S		P S A F		P A S	PAK	S P K	S R R					
Frame 1			Q P S	V V A	S P L	HRAS			A R A	A A S	PAK	S P K	AMENINETA					
Frame 1		THE A			S	DUALISE TREGORALISER								AARCTURCING			T CAREAR	
Frame 1 17. DsegCennC1				V V P		H L A S				A A S	QAK		SR GARINETA				T M V TATAATGAG	DDE
Frame 1 18. DsetCenpC1	AGTUGTTER	M T H A		V V A					A A E	A A C	PAK CARCINA G	S P K	AMENTER				M T	
Frame 1 19. DbunCenpC1	AZZENING TITERAZET					H R A S			TEBIGGARET	S A S	PAK 309200011202	S P K S	AZERTERIA				M T CATAA IGAC	
Frame 1 20. DtruCenpC1	AZGUIGETIGEE					H R A S		PAA F	A E V	A A S	PAK CARCICARC	S P K	A AGENTER TA	AABCTIECTIE				
Frame 1 21. DmayCenpC1		TRACEGE		ATGTORE		H R A S		LASF TAAGUEGOU	A A V	A A S	PAK K	S P K	S R AMERICA				M T CATAATGAC	
Frame 1 22. DbirCenpC1			CBR2 ADDITUDE	M S A	S P L	H C A S		TAABURSUU		- S D S			S R	N S L L				
Frame 1	TAD	L T H A	Q P S	V A A	SPL	RCAS		L T A		- SAS	P A K	SPK	SR	SLL		VDA	MT	DDE

2. A	2.320 2.330	2.340 2.350	2.360	2.370 2	380 2.5	390 2.400	2.410	2.420	2.430	2.440	2.450	2.460	2.470	2.480	2.490	2.500	2.510	2.520
Identity		-	Hink	in the second	in the			. inc	-	للغن	- in	-	ر المراني (المراني ال	-	N. A.	- 10 C	-i-	-rie
1, DpecCenpC3 Frame 1	al Ani tan anaran ani I Mir P P P P	GTING CHICKE AND COM	S L Y D		N H		CGCGATGGCGGC	GCCGCACO	AAGAGATCCA	TEGEEGIG C S	AA <mark>BAXADA<mark>H</mark>GCCA</mark>	TONCITGAGATI	TBCTOBA	EGOCECOMEN	N I A	CRGAG	AG	CG III AA V K
2. DaurCenpC3 Frame 1		GCT TOTOL IS CONTACT SOLO								- A S	GARIANC ATIGODA E N A			- CCCCACA	ZAGA/ABCZZ		TGAAAAGAGA	ATTECE F G
 DtriCenpC3 Frame 1 	NG CARGO CARACTERISTICS	A S L R Y S Q	S L R D	L R Q V	N Q Q	GINN TIGGER (REALIGE) G L D N	<mark>.</mark>			-RCC2dEC	AAREER ACGECA		G	<u>((((//())</u>)	22AE AG2	R I W	TGAAAAGAGA	ATTECS
 DtanCenpC3 Frame 1 		A S L R Y S Q	ATCICTACGCGAT	FRQ M	N R V					C 5	AGENARIGEGAEZ						TGAAAAGGEE	
5. DrufCenpC3 Frame 1					N O V		·c			- 120 22 dillo		TIATTCAGAT		P P P P				ACERICE
6. DasaCenpC3 Frame 1							c			- ଝିଡ଼ିଆ ସିହିତ୍ର ସିହିତ୍ର		ATTGAGAT						
7. DlacCenpC3	Ni Fi CELESCO EC						.c			Constant of the second se		TTATEGAGAT	TGATECAREA	CGCCCCAGA	ZARITORI (Z.			CREECE.
8. DkanCenpC4	311 (FT) - (FT) - (FT) (FT) (FT)		100 - 16 - CODE 1000	व्याहर अवद्य राख्या र	CONTRACTOR		CGTGATAGACGC	GCCACCCC	GAGACTEGEA	TRETENS.	AND THE ADDOLAR	TIATTOAGAT	TOCTORO	COCO TO ALL	CADATEAAT			COMMON (
9. DpunCenpC1								GEAGEAGETER					TEC 1 EEA					ACTE
10. DwatCenpC1							TGAGATA						T BCTBBA	GCCGCCAGA		SCHERCER		ACEE
11. DkikCenpC1								GCCGCTAG				COATTEAGAT	TBCTBBA		ZATIATERCE			
12. DbocCenpC1				III GARGE (ROI)		AT TREAK ANTO		GCCGCTGC				COATTGAGAT	1EC186A		ZABATERGZ			ACTE
13. DleoCenpC1	Nilley (Selected Sel			GARGE (ESSE)			C GAAAAA GGGGC	GCCGCTAG	SAAGATTTCTA	TECCEADA		CONTIGAGAT	TBCTBBAC		ZARATERIZ			ACEE
14. DbakCenpC1				III GARGAACIDI	1 (2011) (2 (1 1 1))		6	TCGCTTT	GTGATTTCTT	TROOM AND G	AA GEEE ALIGOCA	TOATTEGGAT	TECTERA	CECCGCCAGA	2 TEATER (2)	t 17.222 mili		TTEAAE
15. DnikCenpC1													REGER	CGGT	202			ACTIG
16. DjamCenpC1	3111691 (060) DESCRIPTION			GARGE ARTE			G	ccgctgg	AAGATTTGTA	TECCEADA	AAGENE ANGCOA	TOATTGAGAT	TBCTBBA	CECCCCAGA	ZATIATHECZ	TEETE		ACE AD
17. DsegCenpC1				GARGEA (2010	COTEADAE CO		T	CCACTTO	GAGATTTCTA	TECHEN	ACCORDENTION	TIATTGAGAT	A EC TERA	CECCGCCAGA	24EATER/20	то 2ті		
18. DbunCenpC1	NUMBER OF CONSTRUCTION			GARGE ARTE	DE TEADA -	TAR GRANT	.G	CCACTCO	GAGATTTGTA	TECCEADA	AT BANK ANG AC	CONTIGNENT	TECTERA	Gereconan	ZMET MEAST		TGAACCGT	A.
19. DsetCenpC1	31 61 106 DECEM 60 1	AGC ITOLICIE T<mark>SCOLLAGO CO</mark>LLAGO		CAROL AREAL			A	CCGATC-		GAEA	CONSERVENCE	COATTEAGAT	TRCTREA	CREECECACA	2 II THE CO	TECTOCO	TOAACCCTED	ACES
20. DtruCenpC1	NI REI (MARCHART CRAFT		् यतः (स ार्ट्स्स्ट)	GARGE AREAL		व्याप्तः स्वयः व्याप्तः व	.G	ссеттсе	GAGATTTGAA	TREE	A A GEORGE A BIG CITT	TATIFICAGAT	T	nisee Accesso	ZABATER /Z	TERATOR	TAAACCCTER	KARGE .
21. DbirCenpC1	NI 1911 1919 1919 1919 1919			C AR GE ARTE	UND THE AREA IN THE		KII	ACGTTTO	GAGATTTCAA	TECCENT	A REAL ADGET	C/ATTGAGAT	T BTTBBA	CCCCCCAGA	ZALIA CHE GZ	TORTETTE		IT AND CE
22. DmayCenpC1	301 (500) (500) (500 (500 / 500)			CAROLI ARCELO			6	ACGTTCC	GAGATTTCAA	TRECENTA	AADEEE ODGOTO	C (20002 120)	1 B TT BBA		ZADATERICZ		TARACCOTAR	
Frame I			S III R D	E H Q M	N			F F		AN	S A		V		N I G		N R T	A

2. B

	1,810	1,820	1,830	1,840	1,850	1,860	1,870	1,880	1,890	1,900	1,910	1,920	1,930	1,940	1,950	1,960	1,970	1,980
Identity																		
1. DpecCenpC3 Frame 1	P P V	S L R	ACTOCCARTO Y S Q S		GCE AGAE			GAAGAARA E E N	A	TEP	AGAADAHATHGO		AAGARI	E S E	K A	RARGTATICA K V S	Q K P G	AQUETING COMPANY
 DtriCenpC3 Frame 1 	PPA	S L R	ACTICUTOR CHO Y S Q S		GCREERAR		G L D		CGCCATON A		PEN		F G Q		D T S	K V S		
3. DaurCenpC3 Frame 1	P P A	S L R	Atomoto to Al Child		GCERCEARE	TGAAQCAAAT VNQI	AGI TOGACAS S L D	TCGAGAAAA	A	E P	PENN	RAGARTIGAR	F G Q	AGAGAQCAATO	D T S	AAAGTAICA KVS	Q E P T	
 DtanCenpC3 Frame 1 	P P A	S R	A CELEBRA AND		FRQ	MIGMATIQAQI MNRV	D L D			TTGAGCCGC		R S		AGAGAGCAATO	MAAAIIICA E N S	GAAGTTITCA		ECU P
5. DrufCenpC3 Frame 1	P P A	S L R	ACTOCICANTO Y S Q S		F R Q	MIGAATICAAGU MNQV		A ICCOGAAAG	A		PENC	R		AGAGAGCAATO	E T S	K V S	Q D P G	PAR
6. DlacCenpC3 Frame 1	P P A	S L R	ACTORICALATO			ATIGAATICIGAGT M N R V	D F D	ATCACCAAAG	A		PENC	RAGARTING		AGAGAGCAATO		AAAGTTICA KVS		
7. DasaCenpC3 Frame 1	P P A	S L R	A CHECKAMIC		FRQ	MIGAATIGAAGI		ATICAGGAAAG	IGCCATT	TTGAGCCGC		R		CAGAGAGGAATO		AAAGTITCA KV S	CAAGAGGGTGG	E CIIGMECENE PER
 DkanCenpC4 Frame 1 	P P T	S L R	ACTOCICANTO Y S Q S	Q R D	GAAGDICEE				IIGCAATTE		GARTEAUATUA/			AGAGAGIANAID ESN		K S	Q E S E	
9. DwatCenpC1 Frame 1	P P N	S L R	AGECCAALC			M K Q		TIAAGGAAAAA N K <mark>E N</mark>	T P	F E P		A S S		E S N	EGGTGGCT		Q K R E	R Q C
10. DpunCenpC1 Frame 1	P P N	S L R	Y S Q S			M K Q	DLE	NTIAAGGAAAAA	T P	FK P	RATE CONTRACTOR	P S S		E S N	翌G		®A	
11. DkikCenpC1 Frame 1	P P H	S L R	ACTICITERAMIC Y S Q S			N N P S	D S E		IIGCCACCE	E P		PNS		GAGAACAATO	E A S	K A S	CIANCEACERCEASEA	
12. DleoCenpC1 Frame 1	P P H	S L R	AGECTEARED			AUTIMA AGCIAAU	DLE	ATIAGAGAAAAA	D T	TTGAGCCGC		P N S		AGAGAGACACITO	AGCIIII GC	K A S	CAAGAGEGAGA	
13. DbocCenpC1 Frame 1	P P H	S L R	NOTIO I BANATIO Y S Q S		GARGERCER		DLE	R R N		E P		P N S		AGAGAGCAATO	MGCONNICC E A S	K A S		SATING SETTING
14. DbakCenpC1 Frame 1	P P N	S L R	ACTORIANO Y S Q S	L R D			D P E		A	G P				E S N	E A F	K T P	E G P E	R Q S
15. DnikCenpC1 Frame 1	P P N	S R	AGECCAMAEC			M D Q	D S E	300 4 <u>-</u>			G G G G G G G G G G G G G G G G G G G	ACTAATING/		AGAGAGCAAT(MGGEIIIICT E A S			R Q C
16. DjamCenpC1 Frame 1	P P N	S L R	ACTORICANTO Y S Q S		GAR CERAR	M D H		NEAASTANA	A	E P				AGAGACCAATO	E A S	K S	E D P E	R Q C
17. DsegCenpC1 Frame 1	P P D	S L R	ACTORICALING Y S Q S			M D Q	D S N	N K E N	S	E P				E S K	Reconnet K A S	K T S	E G P E	
18. DsetCenpC1 Frame 1	P P S	S L R	ACTICICAMENTO		GAE GEAE				S T	TTGAGCOGO		ACTAATING/		AGAGAGCAAT(MGGGIIIICT E A S			
19. DbunCenpC1 Frame 1	P P N	S L R	NOTOBON OTO Y S Q S		GAR CERAR	M N Q N				E P				P S N	TGTOULUCT V S S	K T S		
20. DtruCenpC1 Frame 1	P P N	S L R	ACTORICALATO		GAS CEALS	N K	D S E T	K E N	A S	E P			G Q	E S N	E A S	K T S		
21. DmayCenpC1 Frame 1	P P N	S L R	Y S Q S		GAR GRARE			K E S	IGCTCCGA A P						MGGGIIIICT E A S	AAAACATCA		
22. DbirCenpC1 Frame 1	P P N	S L R	ATCHLOR CRAVATIO Y S Q S		F R Q	MIGAATRIAZAALI MNQI		N K E N	A P	E P	PENT		RQ	E S D	E A S	KTL	Q D P K	

3. A	3,590 3,600 3,610 3,630 3,640 3,650 3,660 3,670 3,680 3,690 3,700 3,710 3,720 3,720 3,740 3,750 3,760 3,770 3,780
Identity	
1. DpecCenpC3 Frame 1	
2. DaurCenpC3 Frame 1	
3. DtriCenpC3	
4. DtanCenpC3	
5. DrufCenpC3	
6. DasaCenpC3	
7. DlacCenpC3	2014 2010/02 20150 1116-0 2016 2014 2014 2014 2014 2014 2014 2014 2014
8. DkanCenpC4	
9. DpunCenpC1	
10. DwatCenpC1	
11. DkikCenpC1	
12. DbocCenpC1	
13. DleoCenpC1	
14. DbakCenpC1	
15. DnikCenpC1	
16. DjamCenpC1	
17. DsegCenpC1	
18. DbunCenpC1	
19. DsetCenpC1	
20. DtruCenpC1	
21. DbirCenpC1	
22. DmavCenpC1	
3 B	
5.0	3 2,080 2,090 2,100 2,120 2,130 2,140 2,150 2,160 2,170 2,180 2,190 2,200 2,210 2,220 2,230 2,240
Identity	
 DpecCenpC3 	
Frame I	
Prame 1 2. DtriCenpC3 Frame 1	
Frame 1 2. DtriCenpC3 Frame 1 3. DaurCenpC3 Frame 1	
Frame 1 2. DtriCenpC3 Frame 1 3. DaurCenpC3 Frame 1 4. DtanCenpC3 Frame 1	
2. DtriCenpC3 Frame 1 3. DaurCenpC3 Frame 1 4. DtanCenpC3 Frame 1 5. DrufCenpC3 Frame 1 5. DrufCenpC3 Frame 1	
2. DtriCenpC3 Frame 1 3. DaurCenpC3 Frame 1 4. DtanCenpC3 Frame 1 5. DtufCenpC3 Frame 1 6. DlacCenpC3 Frame 1	
2. DtriCenpC3 Frame 1 3. DaurCenpC3 Frame 1 4. DtanCenpC3 Frame 1 5. DtufCenpC3 Frame 1 6. DlacCenpC3 Frame 1 7. DasaCenpC3 Frame 1	
Prame 1 2. DtriCenpC3 Frame 1 3. DaurCenpC3 Frame 1 4. DtanCenpC3 Frame 1 5. DrufCenpC3 Frame 1 5. DrufCenpC3 Frame 1 7. DasaCenpC3 Frame 1 8. DkanCenpC4 Frame 1	
2, DtriCenpC3 Frame 1 3, DaurCenpC3 Frame 1 4, DiamenpC3 5, DrufCenpC3 Frame 1 6, DlacCenpC3 Frame 1 6, DlacCenpC3 Frame 1 8, DkanCenpC4 Frame 1 8, DkanCenpC4 Frame 1 9, DwatCenpC1	
2. DtriCenpC3 Frame 1 3. DaurCenpC3 Frame 1 4. DiangenpC3 Frame 1 6. DiacCenpC3 Frame 1 6. DiacCenpC3 Frame 1 8. DianCenpC4 Frame 1 9. DwarCenpC1 Frame 1 9. DwarCenpC1 Frame 1 9. DwarCenpC1	
2. DtriCenpC3 Frame 1 3. DaurCenpC3 Frame 1 4. DtanCenpC3 Frame 1 5. DtancenpC3 Frame 1 6. DtacCenpC3 Frame 1 8. DtanCenpC4 Frame 1 9. DwatCenpC1 Frame 1 10. DpunCenpC1 Frame 1	
2. DtriCeepc3 frame 1 3. DtriCeepc3 frame 1 4. DtanCenpc3 Frame 1 5. DtriCeepc3 Frame 1 6. Dilacenpc3 Frame 1 8. DisanCenpc3 Frame 1 9. DisanCenpc4 Frame 1 9. DisanCenpc1 Frame 1 10. DipunCenpc1 Frame 1 11. DiskCeenpc1	
2. DtriCenpC3 Frame 1 3. Daur CenpC3 4. Dtard CenpC3 Frame 1 5. DtriCenpC3 Frame 1 6. DlacCenpC3 Frame 1 8. DkanCenpC4 Frame 1 9. DwatCenpC1 Frame 1 9. DwatCenpC1 Frame 1 1. DkikCenpC1 Frame 1 1. DkikCenpC1 Frame 1 1. DkikCenpC1 Frame 1 2. DleoCenpC1 Frame 1 1. DkikCenpC1	
2, DtriCenpC3 Frame 1 3, Daur CenpC3 4, Dian CenpC3 4, Dian CenpC3 Frame 1 5, Druf CenpC3 Frame 1 6, DlacCenpC3 Frame 1 9, DwatCenpC1 Frame 1 10, Dpun CenpC1 Frame 1 11, Dikt CenpC1 Frame 1 12, DieoCenpC1 Frame 1 12, DieoCenpC1 Frame 1 13, DbacCenpC1	
Prame 1 2, DtriCenpC3 Frame 1 3, Daur CenpC3 4, Dian CenpC3 4, Dian CenpC3 Frame 1 5, Druf CenpC3 Frame 1 6, DlacCenpC3 Frame 1 9, DwatCenpC1 Frame 1 10, DpunCenpC1 Frame 1 11, DbunCenpC1 Frame 1 12, DleoCenpC1 Frame 1 13, DbocCenpC1 Frame 1 13, DbocCenpC1 Frame 1 13, DbocCenpC1 Frame 1 14, DbakCenpC1 Frame 1 15, DbakCenpC1 Frame 1 15, DbakCenpC1 Frame 1 15, DbakCenpC1 Frame 1 15, DbakCenpC1	
Prame 1 2, DtriCenpC3 Frame 1 3, DaurCenpC3 Frame 1 4, DianCenpC3 Frame 1 5, DrufCenpC3 Frame 1 6, DlacCenpC3 Frame 1 7, DasaCenpC3 Frame 1 10, DpunCenpC1 Frame 1 11, DkikCenpC1 Frame 1 11, DkikCenpC1 Frame 1 13, DbocCenpC1 Frame 1 14, DbakCenpC1 Frame 1 15, DnikCenpC1 Frame 1 15, DnikCenpC1 Frame 1 15, DnikCenpC1 Frame 1 16, DpunCenpC1 Frame 1 17, DbikCenpC1 Frame 1 17, DbikCenpC1 Frame 1 16, DpunCenpC1 Frame 1 17, DbikCenpC1 Frame	
2. DtriCenpC3 Frame 1 3. DaurCenpC3 Frame 1 4. Janue CenpC3 Frame 1 6. DlacCenpC3 Frame 1 6. DlacCenpC3 Frame 1 8. Dlan CenpC4 Frame 1 10. DpunCenpC1 Frame 1 11. DkikCenpC1 Frame 1 13. DbacCenpC1 Frame 1 14. DbakCenpC1 Frame 1 15. DnikCenpC1 Frame 1 15	
2. DtriCenpC3 Frame 1 2. DtriCenpC3 Frame 1 3. DtricenpC3 Frame 1 5. DtrifCenpC3 Frame 1 5. DtrifCenpC3 Frame 1 7. DasaCenpC3 Frame 1 8. DikanCenpC4 Frame 1 9. DtrifCenpC1 Frame 1 10. DpurCenpC1 Frame 1 11. DkikCenpC1 Frame 1 14. DbakCenpC1 Frame 1 14. DbakCenpC1 Frame 1 15. DbakCenpC1 Frame 1 Frame 1 15. DbakCenpC1 Frame 1 Frame 1 Fr	
2. DtriCenpC3 Frame 1 3. Daniel enpC3 4. DtanCenpC3 Frame 1 5. DtriCenpC3 Frame 1 6. DlacCenpC3 Frame 1 8. DtanCenpC4 Frame 1 8. DtanCenpC4 Frame 1 9. DwatCenpC1 Frame 1 10. DpurCenpC1 Frame 1 11. DtacCenpC1 Frame 1 15. DbacCenpC1 Frame 1 15. DbacCenpC1 Frame 1 15. DbacCenpC1 Frame 1 15. DbacCenpC1 Frame 1 15. DbacCenpC1 Frame 1 15. DbacCenpC1 Frame 1 16. DbacCenpC1 Frame 1 17. DbacCenpC1 Frame 1 18. DbacCenpC1 Frame 1 19. DbacCenpC	
2. DtriCenpC3 Frame 1 3. Daur CenpC3 Frame 1 4. Dtan CenpC3 Frame 1 5. DtriCenpC3 Frame 1 6. DlacCenpC3 Frame 1 8. DkanCenpC4 Frame 1 9. DwatCenpC1 Frame 1 1. DkikCenpC1 Frame 1 12. DlacCenpC1 Frame 1 13. DbikCenpC1 Frame 1 15. DnikCenpC1 Frame 1 15. DnikCenpC1 Frame 1 16. DbjarnCenpC1 Frame 1 16. DbjarnCenpC1 Frame 1 16. DbjarnCenpC1 Frame 1 16. DbjarnCenpC1 Frame 1 19. DbjarnCenpC1	
2. DtriCenpC3 Frame 1 3. Daur CenpC3 Frame 1 5. Druf CenpC3 Frame 1 5. Druf CenpC3 Frame 1 6. DlacCenpC3 Frame 1 8. DkanCenpC4 Frame 1 9. DwatCenpC1 Frame 1 10. Dpb/n CenpC1 Frame 1 12. DleoCenpC1 Frame 1 13. DbacCenpC1 Frame 1 15. DnikCenpC1 Frame 1 10. DjamCenpC1 Frame 1 10. DjamCenpC1 Frame 1 10. Dava CenpC1 Frame 1 10. DiamCenpC1 Frame 1	

Figura 26. Alinhamento de códons da Cenp-C das espécies do grupo montium mostrando os sítios que apresentaram assinatura de seleção positiva. Cada sítio (barras em azul-escuro) é mostrado nas imagens. (1.A) Localização do primeiro sítio antes do refinamento manual do alinhamento. (1.B) Localização do primeiro sítio após refinamento manual do alinhamento. (2.A) Localização do segundo e terceiro sítios antes do refinamento manual do alinhamento. (2.B) Localização do segundo e terceiro sítios após refinamento manual do alinhamento. (3.A) Localização do quarto sítio antes do refinamento manual do alinhamento. (3.B) Localização do quarto sítio após refinamento manual do alinhamento. O refinamento manual do alinhamento foi feito conforme especificado na metodologia.

Genes	M1a vs M2a p-value	M2 % sites with ω > 1 (Avg ω)	M7 vs M8 p-value	M8 % sites with ω > 1 (Avg ω)	M8a vs M8 p-value	M8 BEB (PP%)
Cenp-C1/Cenp- C3/Cenp-C4	1	0.63(0.14)	0.005	0.02 (1.91)	0.016	354 L (87.4) 624 L (80.5) 639 A (93.0) 720 F (89.8)

Tabela 3. Sítios sob seleção positiva de acordo com o modelo de sítios-aleatórios (CodeML)

Modelo de frequência de códons F3X4. Os valores de P < 0.05 estão indicados em negrito.

5. CONCLUSÕES

Em relação aos nossos objetivos, as principais conclusões do presente trabalho foram:

Identificar o gene da proteína centromérica *Cenp-C* nas espécies do grupo *montium* e investigar se existem duplicações de *Cenp-C* nas espécies estudadas.

Foi identificado o gene da *Cenp-C1* e mais três novas cópias (*Cenp-C3*, *Cenp-C4* e *Cenp-C5*) nas espécies do grupo *montium*. Porém *D. vulcana* é a única espécie que possui duas cópias (*Cenp-C1* e *Cenp-C5*) em seu genoma. Várias evidências indicam que todas essas cópias são funcionais. Como em um trabalho anterior foram identificadas três cópias da Cid nessas mesmas espécies, nossos resultados mostram que apenas uma cópia da proteína Cenp-C interage com as três cópias da Cid para função centromérica.

Identificar e caracterizar motivos conservados da proteína *Cenp-C* nas espécies do grupo *montium*.

Dos oito motivos proteicos caracterizados para Cenp-C em *Drosophila*, seis se encontram conservados em todas as cópias de Cenp-C do grupo *montium*. Entre esses seis motivos estão o CenH3 *binding*, Cupin e Mis12 *binding* que apresentam funções essenciais para a função do centrômero. Esse resultado sugere que todas as cópias de Cenp-C encontradas no presente trabalho são funcionais e desempenham função centromérica.

Analisar a relação evolutiva entre as cópias de Cenp-C encontradas.

Todas as três cópias da *Cenp-C* encontradas nas espécies do grupo *montium* (*Cenp-C3, Cenp-C4, Cenp-C5*) derivam de duplicações independentes de *Cenp-C1*, seguida de perda *Cenp-C1* na maioria das linhagens. A árvore filogenética das sequencias de *Cenp-C* de cada subgrupo apresenta topologia que está de acordo com a árvore das espécies. Eventos de duplicações completas ou parciais da *Cenp-C* nas espécies do grupo *montium* ocorreram várias vezes e de maneira independente. E por último, a árvore das espécies do gênero *Drosophila* confirma que a *Cenp-C3* e a *Cenp-C4* são ortólogos funcionais da *Cenp-C1* nas espécies do grupo *montium*.

Identificar quais os mecanismos envolvidos nas duplicações da Cenp-C.

Não foi possível determinar o processo envolvido na duplicação do gene da *Cenp-C*. No entanto, não encontramos evidências de duplicação via *crossing-over* desigual ou por intermédio de TEs.

Testar se existe seleção positiva atuando na Cenp-C.

Encontramos assinaturas de seleção positiva em quatro aminoácidos da Cenp-C das espécies do grupo *montium*. Apesar de não termos nenhuma evidência experimental que determine os sítios em contato com o DNA centromérico em *Drosophila*, não descartamos a possibilidade de que nas espécies do grupo *montium*, a evolução de Cenp-C sofra influência do impulso centromérico.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALLEN, S. L. et al. Single-molecule sequencing of the Drosophila serrata genome. **G3: Genes, Genomes, Genetics**, v. 7, n. 3, p. 781–788, 2017.

ASHBURNER, M.; BERGMAN, C. M. Drosophila melanogaster: A case study of a model genomic sequence and its consequences. **Genome Research**, v. 15, n. 12, p. 1661–1667, 2005.

BAILEY, T. L. et al. The MEME Suite. **Nucleic Acids Research**, v. 43, n. W1, p. W39–W49, 2015.

BAILEY, T. L.; GRIBSKOV, M. Combining evidence using p-values: Application to sequence homology searches. **Bioinformatics**, v. 14, n. 1, p. 48–54, 1998.

BARRA, V.; FACHINETTI, D. The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. **Nature**

Communications, v. 9, n. 1, p. 4340, 18 dez. 2018.

BRONSKI, M. J. et al. Whole genome sequences of 23 species from the drosophila montium species group (Diptera: Drosophilidae): A resource for testing evolutionary hypotheses. **G3: Genes, Genomes, Genetics**, v. 10, n. 5, p. 1443–1455, 2020.

BROWN, J. D.; O'NEILL, R. J. The Evolution of Centromeric DNA Sequences. **eLS**, p. 1–11, 2014.

CARROLL, C. W.; MILKS, K. J.; STRAIGHT, A. F. Dual recognition of CENP-A nucleosomes is required for centromere assembly. **Journal of Cell Biology**, v. 189, n. 7, p. 1143–1155, 2010.

CERBIN, S.; JIANG, N. Duplication of host genes by transposable elements. **Current Opinion in Genetics and Development**, v. 49, n. Figure 1, p. 63–69, 2018.

CHEN, Z. X. et al. Comparative validation of the D. melanogaster modENCODE

transcriptome annotation. Genome Research, v. 24, n. 7, p. 1209–1223, 2014.

COHEN, R. L. et al. Structural and Functional Dissection of Mif2p, a Conserved DNA-binding Kinetochore Protein. **Molecular Biology of the Cell**, v. 19, n. 10, p. 4480–4491, out. 2008. CONNER, W. R. et al. A phylogeny for the Drosophila montium species group: A model clade for comparative analyses. **Molecular Phylogenetics and Evolution**, v. 158, n. April 2020, p. 107061, 2021.

COOPER, J. L.; HENIKOFF, S. Adaptive evolution of the histone fold domain in centromeric histones. **Molecular Biology and Evolution**, v. 21, n. 9, p. 1712–1718, 2004.

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

DA LAGE, J. L. et al. A phylogeny of Drosophilidae using the Amyrel gene: Questioning the Drosophila melanogaster species group boundaries. **Journal of Zoological Systematics** and Evolutionary Research, v. 45, n. 1, p. 47–63, 2007.

DALAL, Y. et al. Structure, dynamics, and evolution of centromeric nucleosomes.

Proceedings of the National Academy of Sciences of the United States of America, v. 104, n. 41, p. 15974–15981, 2007.

DAWE, R. K.; HENIKOFF, S. Centromeres put epigenetics in the driver's seat. **Trends in Biochemical Sciences**, v. 31, n. 12, p. 662–669, 2006.

ERHARDT, S. et al. Genome-wide analysis reveals a cell cycle-dependent mechanism controlling centromere propagation. **Journal of Cell Biology**, v. 183, n. 5, p. 805–818, 2008. FALK, S. J. et al. CENP-C reshapes and stabilizes CENP-A nucleosomes at the centromere. **Science**, v. 348, n. 6235, p. 699–703, 8 maio 2015.

FESCHOTTE, C.; PRITHAM, E. J. DNA transposons and the evolution of eukaryotic

genomes. Annual Review of Genetics, v. 41, p. 331–368, 2007.

FINET, C. et al. DrosoPhyla: Resources for Drosophilid Phylogeny and Systematics.

Genome Biology and Evolution, v. 13, n. 8, p. 1–13, 2021.

HALES, K. G. et al. Genetics on the fly: A primer on the drosophila model system. **Genetics**, v. 201, n. 3, p. 815–842, 2015.

HARTLEY, G.; O'NEILL, R. J. Centromere repeats: Hidden gems of the genome. **Genes**, v. 10, n. 3, 2019.

HEEGER, S. et al. Genetic interactions of separase regulatory subunits reveal the diverged Drosophila Cenp-C homolog. **Genes and Development**, v. 19, n. 17, p. 2041–2053, 2005. HENIKOFF, S. et al. The Centromere Paradox: Stable Inheritance with Rapidly Evolving. **Science**, v. 293, n. 5532, p. 1098–1102, 2001.

HOOFF, J. J. et al. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. **EMBO reports**, v. 18, n. 9, p. 1559–1571, 2017. KATO, H. et al. Reports 11. **Science**, v. 340, n. 6136, p. 1110–1113, 2013.

KOONIN, E. V. Orthologs, paralogs, and evolutionary genomics. **Annual Review of Genetics**, v. 39, p. 309–338, 2005.

KUMAR, S. et al. MEGA X: Molecular evolutionary genetics analysis across computing platforms. **Molecular Biology and Evolution**, v. 35, n. 6, p. 1547–1549, 2018.

KURSEL, L. E.; MALIK, H. S. Recurrent gene duplication leads to diverse repertoires of centromeric histones in Drosophila species. **Molecular Biology and Evolution**, v. 34, n. 6, p. 1445–1462, 2017.

LIU, Y. et al. Insights from the reconstitution of the divergent outer kinetochore of Drosophila melanogaster. **Open Biology**, v. 6, n. 2, 2016.

MALIK, H. S.; BAYES, J. J. Genetic conflicts during meiosis and the evolutionary origins of centromere complexity. **Biochemical Society Transactions**, v. 34, n. 4, p. 569–573, 2006. MALIK, H. S.; HENIKOFF, S. Adaptive evolution of Cid, a centromere-specific histone in Drosophila. **Genetics**, v. 157, n. 3, p. 1293–1298, 2001.

MALIK, H. S.; HENIKOFF, S. Conflict begets complexity : the evolution of centromeres.

Current Opinion in Genetics & Development, v. 12, n. 6, p. 711–718, 2002.

MALIK, H. S.; VERMAAK, D.; HENIKOFF, S. Recurrent evolution of DNA-binding motifs in the Drosophila centromeric histone. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 3, p. 1449–1454, 2002.

MARESCA, T. J. Chromosome segregation: A kinetochore missing link is found. **Current Biology**, v. 21, n. 7, p. R261–R263, 2011.

MARKOW, T. A.; O'GRADY, P. Summary for Policymakers. In: INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (Ed.). . Climate Change 2013 - The Physical Science Basis. Cambridge: Cambridge University Press, 2006. p. 1–30. MCKINLEY, K. L.; CHEESEMAN, I. M. The molecular basis for centromere identity and function. **Nature Reviews Molecular Cell Biology**, v. 17, n. 1, p. 16–29, 2016.

MELTERS, D. P. et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. **Genome Biology**, v. 14, n. 1, p. 1–20, 2013.

MOREE, B. et al. CENP-C recruits M18BP1 to centromeres to promote CENP-A chromatin assembly. **Journal of Cell Biology**, v. 194, n. 6, p. 855–871, 2011.

MUSACCHIO, A.; DESAI, A. A molecular view of kinetochore assembly and function.

Biology, v. 6, n. 1, 2017.

O'GRADY, P. M.; DESALLE, R. Phylogeny of the genus Drosophila. **Genetics**, v. 209, n. 1, p. 1–25, 2018.

ORR, B.; SUNKEL, C. E. Drosophila CENP-C is essential for centromere identity.

Chromosoma, v. 120, n. 1, p. 83–96, 2011.

PLOHL, M. et al. Satellite DNAs between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. **Gene**, v. 409, n. 1–2, p. 72–82, 2008.

PLOHL, M.; MEŠTROVIĆ, N.; MRAVINAC, B. Centromere identity from the DNA point of view. **Chromosoma**, v. 123, n. 4, p. 313–325, 2014.

POLITI, V. et al. CENP-C binds the alpha-satellite DNA in vivo at specific centromere domains. **Journal of Cell Science**, v. 115, n. 11, p. 2317–2327, 2002.

PRZEWLOKA, M. R. et al. Molecular analysis of core kinetochore composition and assembly in Drosophila melanogaster. **PLoS ONE**, v. 2, n. 5, 2007.

PRZEWLOKA, M. R. et al. CENP-C is a structural platform for kinetochore assembly.

Current Biology, v. 21, n. 5, p. 399–405, 2011.

PRZEWLOKA, M. R.; GLOVER, D. M. The kinetochore and the centromere: A working long distance relationship. **Annual Review of Genetics**, v. 43, p. 439–465, 2009.

RASTOGI, S.; LIBERLES, D. A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. **BMC Evolutionary Biology**, v. 5, p. 1–7, 2005.

RICHTER, M. M. et al. Network of protein interactions within the Drosophila inner kinetochore. **Open Biology**, v. 6, n. 2, 2016.

ROSIN, L. F.; MELLONE, B. G. Centromeres Drive a Hard Bargain. **Trends in Genetics**, v. 33, n. 2, p. 101–117, 2017.

ROURE, V. et al. Reconstituting Drosophila Centromere Identity in Human Cells. **Cell Reports**, v. 29, n. 2, p. 464- 479.e5, 2019.

RUSSO, C. A. M. et al. Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). **Zoological Journal of the Linnean Society**, v. 169, n. 4, p. 765–775, 2013.

RUSSO, C. A. M.; TAKEZAKI, N.; NEI, M. Molecular phylogeny and divergence times of drosophilid species. **Molecular Biology and Evolution**, v. 12, n. 3, p. 391–404, 1995. SKIPPER, K. A. et al. DNA transposon-based gene vehicles - Scenes from an evolutionary drive. **Journal of Biomedical Science**, v. 20, n. 1, p. 1–23, 2013.

STANKE, M.; MORGENSTERN, B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. **Nucleic Acids Research**, v. 33, n. SUPPL. 2, p. 465–467, 2005.

TALBERT, P. B.; BRYSON, T. D.; HENIKOFF, S. Adaptive evolution of centromere proteins in plants and animals. **Journal of Biology**, v. 3, n. 4, 2004.

TALBERT, P. B.; HENIKOFF, S. What makes a centromere? **Experimental Cell Research**, v. 389, n. 2, p. 111895, 2020.

TAMURA, K.; SUBRAMANIAN, S.; KUMAR, S. Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks. **Molecular Biology and Evolution**, v. 21, n. 1, p. 36–44, 2004.

TEIXEIRA, J. R. et al. Concurrent Duplication of Drosophila Cid and Cenp-C Genes Resulted in Accelerated Evolution and Male Germline-Biased Expression of the New Copies. **Journal of Molecular Evolution**, v. 86, n. 6, p. 353–364, 2018.

TOROSIN, N. S. et al. 3D genome evolution and reorganization in the Drosophila melanogaster species group. **PLoS Genetics**, v. 16, n. 12, p. 1–29, 2020.

TRAZZI, S. et al. In vivo functional dissection of human inner kinetochore protein CENP-C. **Journal of Structural Biology**, v. 140, n. 1–3, p. 39–48, 2002.

WESTHORPE, F. G.; FULLER, C. J.; STRAIGHT, A. F. A cell-free CENP-A assembly system defines the chromatin requirements for centromere maintenance. **Journal of Cell Biology**, v. 209, n. 6, p. 789–801, 2015.

WICKER, T. et al. Apg Iv. **Nature Reviews Genetics**, v. 8, n. 12, p. 973–982, 2007. XIAO, H. et al. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. **Science**, v. 319, n. 5869, p. 1527–1530, 2008.

XING, J. et al. Emergence of primate genes by retrotransposon-mediated sequence transduction. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 47, p. 17608–17613, 2006.

XU, B.; YANG, Z. PamIX: A graphical user interface for PAML. **Molecular Biology and Evolution**, v. 30, n. 12, p. 2723–2724, 2013.

YANG, Y. et al. Increasing the data size to accurately reconstruct the phylogenetic relationships between nine subgroups of the Drosophila melanogaster species group (Drosophilidae, Diptera). **Molecular Phylogenetics and Evolution**, v. 62, n. 1, p. 214–223, 2012.

YASSIN, A. Phylogenetic classification of the Drosophilidae Rondani (Diptera): The role of

morphology in the postgenomic era. **Systematic Entomology**, v. 38, n. 2, p. 349–364, 2013. YASSIN, A. et al. The pdm3 Locus Is a Hotspot for Recurrent Evolution of Female-Limited Color Dimorphism in Drosophila. **Current Biology**, v. 26, n. 18, p. 2412–2422, 2016. YASSIN, A. Phylogenetic biogeography and classification of the Drosophila montium species group (Diptera: Drosophilidae). **Annales de la Societe Entomologique de France**, v. 54, n. 2, p. 167–175, 2018.

ZHANG, J. Evolution by gene duplication: An update. **Trends in Ecology and Evolution**, v. 18, n. 6, p. 292–298, 2003.

ZHAO, D. et al. The unique epigenetic features of Pack-MULEs and their impact on chromosomal base composition and expression spectrum. n. January, p. 1–18, 2018.