

UNIVERSIDADE FEDERAL DE MINAS GERAIS

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

Ana Mattioli Laborne do Nascimento

CARACTERIZAÇÃO E EVOLUÇÃO DE DNAs SATÉLITES EM *DROSOPHILA ANTONIETAE* E *D. SERIDO* (GRUPO *REPLETA*, *CLUSTER BUZZATII*) E EM GENOMAS DA ZONA HÍBRIDA

Belo Horizonte

2023

Ana Mattioli Laborne do Nascimento

CARACTERIZAÇÃO E EVOLUÇÃO DE DNAs SATÉLITES EM *DROSOPHILA ANTONIETAE* E *D. SERIDO* (GRUPO *REPLETA*, *CLUSTER BUZZATII*) E EM GENOMAS DA ZONA HÍBRIDA

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

Belo Horizonte

2023

043

Nascimento, Ana Mattioli Laborne do.

Caracterização e evolução de DNAs satélites em *Drosophila antonietae* e *D. serido* (grupo *repleta*, *cluster buzzatii*) e em genomas da zona híbrida [manuscrito] / Ana Mattioli Laborne do Nascimento. – 2023.

109 f. : il. ; 29,5 cm.

Orientador: Dr. Gustavo Campos e Silva Kuhn.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genômica. 2. Bioinformática. 3. DNA Satélite. 4. *Drosophila*. 5. Centrômero. I. Silva Kuhn, Gustavo Campos e. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 575



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

ATA DE DEFESA DE DISSERTAÇÃO

ATA DA DEFESA DE DISSERTAÇÃO	346/2023 entrada
ANA MATTIOLI LABORNE DO NASCIMENTO	1º/2021 CPF: 094.474.776-07

Às quatorze horas do dia **29 de agosto de 2023**, reuniu-se a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: "**Caracterização e evolução de DNAs satélites em *Drosophila antonietae* e *D. serido* (grupo repleta, cluster buzatii) e em genomas da zona híbrida**", requisito para obtenção do grau de Mestre em **Genética**. Abrindo a sessão, o Presidente da Comissão, **Gustavo Campos e Silva Kuhn**, 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
Fabício Rodrigues dos Santos	UFMG	567.487.446-87	APROVADA
Fernando de Faria Franco	UFSCar	217.105.428-02.	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, 29 de agosto de 2023.

Gustavo Campos e Silva Kuhn- UFMG

Fabício Rodrigues dos Santos- UFMG

Fernando de Faria Franco- UFSCar



Documento assinado eletronicamente por **Fernando de Faria Franco, Usuário Externo**, em 30/08/2023, às 08:53, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Gustavo Campos e Silva Kuhn, Chefe de departamento**, em 30/08/2023, às 08:56, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Fabricio Rodrigues dos Santos, Membro**, em 30/08/2023, às 10:33, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2577514** e o código CRC **290FCA69**.



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

FOLHA DE APROVAÇÃO

"Caracterização e evolução de DNAs satélites em *Drosophila antonietae* e *D. serido* (grupo repleta, cluster buzatii) e em genomas da zona híbrida"

Ana Mattioli Laborne do Nascimento

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Gustavo Campos e Silva Kuhn
UFMG

Fabício Rodrigues dos Santos
UFMG

Fernando de Faria Franco
UFSCar

Belo Horizonte, 29 de agosto de 2023.



Documento assinado eletronicamente por **Fernando de Faria Franco, Usuário Externo**, em 30/08/2023, às 08:53, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Gustavo Campos e Silva Kuhn, Chefe de departamento**, em 30/08/2023, às 08:56, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Fabricio Rodrigues dos Santos, Membro**, em 30/08/2023, às 10:33, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2577546** e o código CRC **4CC35BBF**.

A todos os trabalhadores brasileiros, que, por meio de seu árduo trabalho e dedicação, financiaram essa pesquisa.

AGRADECIMENTOS

Em especial, agradeço ao meu orientador Prof. Dr. Gustavo Campos e Silva Kuhn por todos os ensinamentos, conselhos, oportunidades e confiança durante os cinco anos em que estive no laboratório. Esse período no Laboratório de Citogenômica Evolutiva foi fundamental para meu desenvolvimento e crescimento acadêmico e pessoal. Agradeço à professora Marta Svartman por sua disponibilidade, atenção, ensinamentos, convivência e contribuição para minha formação acadêmica e profissional.

Às bancas examinadoras da dissertação e da disciplina de seminários, por terem aceitado ler e discutir o meu trabalho e por contribuírem para minha formação.

A toda minha família, em especial, minha mãe, Adriana Laborne, meu pai, Wagner Nascimento, meu irmão, Pedro Mattioli, meu padrasto, Aloizio Belém e minha tia, Fernanda Laborne, por todo esforço investido na minha formação profissional e pessoal. Muito obrigada por todo apoio durante a minha caminhada e por sempre acreditarem em mim.

Ao meu noivo, Lucas Capellini, com quem compartilho os grandes e pequenos momentos da minha vida. Muito obrigada pela compreensão, paciência, cuidado, incentivo e carinho nesses cinco anos.

Aos meus queridos amigos do LCEv, muito obrigada por tudo que me ensinaram sobre ciência e sobre a vida. Bráulio, Erick, Rafa, Pedro, Rada, Alice e Mi vocês foram fundamentais na minha trajetória acadêmica e pessoal, além de tornaram os dias mais leves. Sinto muita saudade e muito orgulho de cada um de vocês.

Em especial, agradeço ao Bráulio e Pedro por todos os ensinamentos, discussões, paciência e contribuições para o meu trabalho. Mais uma vez, à Rada, obrigada por tudo! Pelas discussões que contribuíram diretamente para essa dissertação, pela amizade, companheirismo e por estar ao meu lado, especialmente durante os últimos dois anos.

Aos professores e funcionários da UFMG e do Programa de Pós-Graduação em Genética. Nominalmente, à Raíssa e Izabella pelo auxílio e carinho. Ao Daniel que sempre ajudou com a manutenção dos estoques de *Drosophila*.

Aos meus amigos Anna, Júlia, Carlos, Fábio e Vivian por tornarem mais leve a minha caminhada desde o início da graduação. Sou extremamente grata pela amizade e carinho compartilhados.

Aos amigos e amigas do colégio, do vôlei e da vida pelo carinho, apoio e suporte durante minha caminhada.

À CAPES, FAPEMIG e CNPq, agências financiadoras do meu e de outros projetos conduzidos no LCEv. À CAPES pela bolsa de pesquisa concedida e aceite de prorrogação do prazo de defesa devido à pandemia da COVID.

A todos os professores que contribuíram, de alguma forma, para minha formação pessoal e profissional, desde a pré escola até o mestrado. Obrigada por acreditarem na educação!

A todos aqueles que contribuíram direta e indiretamente para a realização deste trabalho e que fizeram parte desse processo.

RESUMO

Os DNAs repetitivos são os componentes mais abundantes dos genomas dos eucariotos e podem ser classificados de acordo com a sua organização no genoma em dispersos ou em tandem. Os DNAs satélites (satDNAs) são organizados em arranjos de milhares de cópias em tandem, encontrados principalmente na heterocromatina. SatDNAs apresentam um padrão de evolução que é chamado de evolução combinada, em que as cópias de uma mesma família de satDNA apresentam variabilidade intraespecífica menor do que interespecífica. Apesar de não serem sequências codificadoras de proteínas, os satDNAs podem exercer papéis funcionais no genoma como na regulação gênica, modulação da cromatina, manutenção de centrômeros funcionais, fertilidade masculina e organização espacial dos cromossomos. O desenvolvimento de novas técnicas de sequenciamento, como Sequenciamento de Nova Geração (NGS), junto com o desenvolvimento de ferramentas de bioinformática, proporcionaram novas possibilidades para o estudo mais completo e detalhado de DNAs satélites. Neste trabalho nós usamos a pipeline do RepeatExplorer2 para identificação *de novo* dos satDNAs de *Drosophila antonietae* e *D. serido* (*cluster buzzatii*, grupo *repleta*) usando *reads* brutas do sequenciamento das duas espécies e de seis genomas da zona de contato entre elas. Nós identificamos cinco satélites e um minissatélite compartilhados por ambas as espécies. Entre os satDNAs, dois já haviam sido descritos nas duas espécies (pBuM e DBC-150), um já descrito e encontrado em outras espécies do grupo *repleta* (CDSTR138) e dois satélites novos são descritos pela primeira vez neste trabalho (SAT197/230 e SAT109/104). O minissatélite CDSTR198 também foi encontrado nas duas espécies. Nossos resultados de FISH demonstraram que o CDSTR138 (e não o pBuM, como anteriormente sugerido) é o principal componente centromérico de *D. serido*, *D. seriema*, *D. gouveai* e *D. koepferae*. Além disso, os novos satélites foram encontrados exclusivamente nos cromossomos sexuais. O SAT197/230 está presente em três arranjos no cromossomo Y, sendo um deles centromérico, e o SAT109/104, que está presente na região centromérica do cromossomo X. O CDSTR198 foi confirmado como um minissatélite em *D. serido*, visível apenas nos braços eucromáticos de cromossomos politênicos. Todos os satélites centroméricos analisados são enriquecidos de pequenas díades simétricas e têm propensão a formar estruturas secundárias estáveis, características consideradas essenciais para função centromérica. Entre os satélites analisados, quatro apresentaram padrão de evolução combinada e foram considerados bons marcadores taxonômicos para identificação das espécies. As análises de genomas da zona híbrida sugerem que dois genomas podem ser classificados como *D. antonietae* e quatro como *D. serido*. Nosso estudo confirma a eficiência da pipeline RepeatExplorer2 na identificação de satDNAs em *Drosophila* e proporciona novos insights na evolução do genoma e na estrutura do centrômero nessas espécies. Também fornece novos marcadores genéticos para identificação de espécies, que são particularmente úteis em estudos envolvendo a zona híbrida.

Palavras-chave: DNAs satélites, citogenômica, *Drosophila*, RepeatExplorer2, centrômero e zona híbrida

ABSTRACT

Repetitive DNAs are among the most abundant components of eukaryotic genomes and can be classified according to their organization in the genome as dispersed or tandemly arranged. Satellite DNAs (satDNAs) are organized in arrays of thousands of tandemly organized copies, predominantly found in heterochromatin. SatDNAs exhibit a pattern of evolution known as concerted evolution, where copies within the same satDNA family display lower intraspecific variability compared to interspecific one. Despite satellites do not encode proteins, they can play functional roles in the genome, such as gene regulation, chromatin modulation, maintenance of functional centromeres, male fertility and spatial organization of chromosomes. The development of new sequencing techniques, such as Next Generation Sequencing (NGS), together with the advent of new bioinformatics tools, have provided new possibilities for a comprehensive and detailed study of satDNAs. In the present work, we used the RepeatExplorer2 pipeline for de novo identification of satDNAs in *Drosophila antonietae* and *D. serido* (*buzzatii* cluster, *repleta* group). We used raw sequencing reads from both species and from six genomes from the contact zone between them. We identified five satDNAs and one minisatellite shared between the two species. Among the satDNAs, two were previously described in both species (pBuM and DBC-150), one was previously described and found in other species of the *repleta* group (CDSTR138), and two novel satDNAs were identified for the first time in this study (SAT197/230 and SAT109/104). The minisatellite CDSTR198 was also found in both species. Our FISH experiments revealed that CDSTR138 (and not pBuM, as previously suggested) is the main centromeric component in *D. serido*, *D. seriema*, *D. gouveai*, and *D. koepferae*. Additionally, the newly identified satDNAs were exclusively found on the sex chromosomes. The SAT197/230 is present on the Y chromosome and organized in three arrays, one of which is centromeric, and the SAT109/104 is present in the centromeric region of the X chromosome. The CDSTR198 was confirmed as a minisatellite in *D. serido*, visible only on the euchromatic arms of the polytene chromosomes. All the analyzed centromeric satDNAs showed enrichment of small dyads symmetries and were predicted to form stable secondary structures, which are considered essential characteristics for centromeric function. Among the analyzed satDNAs, four exhibited a pattern of concerted evolution and were considered good taxonomic markers. Our analysis of genomes from the hybrid zone suggests that two genomes can be classified as *D. antonietae* and four as *D. serido*. Overall, our study confirms the efficiency of the RepeatExplorer2 pipeline for identifying satDNAs in *Drosophila* and provides new insights into genome evolution and centromere structure in these species. Furthermore, it reveals new genetic markers for species identification, which will be valuable for studies involving the hybrid zone.

Palavras-chave: satellite DNAs, cytogenomics, *Drosophila*, RepeatExplorer2, centromeres e hybrid zone

LISTA DE FIGURAS

Figura 1. Esquema representativo da localização cromossômica dos DNAs satélites

(Adaptado de Kuhn e col., 2021).

Figura 2. Esquema representativo do processo de evolução combinada de satDNAs.

(Adaptado de Plohl e col., 2012)

Figura 3. Filogenia do *cluster buzzatii*. *Drosophila antonietae* e *D. serido* em destaque.

Figura 4. Idiograma do cariótipo básico das espécies do *cluster buzzatii*.

Figura 5. Distribuição geográfica de *D. serido* e *D. antonietae*, incluindo a zona de contato secundário em vermelho.

Figura 6. Representação esquemática da família de satDNA pBuM.

Figura 7. Idiograma representativo da localização cromossômica dos elementos repetitivos identificados em *D. antonietae* e *D. serido* (Kuhn e col. 2007, 2008; Franco e col. 2008; Kuhn e Heslop-Harrison 2011).

LISTA DE ABREVIATURAS

BLAST – “Basic Local Alignment Search Tool” (Ferramenta de busca e alinhamento básico local)

CTR – “Central Tandem Repeats” (Repetições em tandem centrais)

DNA – ácido desoxirribonucleico (ADN)

FISH – Fluorescence in situ hybridization (Hibridização in situ fluorescente)

HZ – “hybrid zone” – zona híbrida

HOR – “Higher-order repeats” (repetições de ordem superior)

Kbs – kilobases

MEGA - Molecular Evolutionary Genetics Analysis (software)

NCBI – National Center for Biotechnology Information (instituição)

NGS – “Next-generation sequencing” (sequenciamento de nova geração)

Pb/bp – pares de base/“base pairs”

PCR – reação em cadeia da polimerase

Reads – fragmentos gerados após sequenciamento de DNA

satDNAs – DNAs satélites

TEs – Elementos transponíveis

R – Linguagem de programação

SUMÁRIO

1. INTRODUÇÃO	9
1.1 Os DNAs satélites	9
1.2 O gênero <i>Drosophila</i>	13
1.2.1 O grupo <i>repleta</i> e as espécies do <i>cluster buzzatii</i>	14
1.2.2. As espécies <i>D. antonietae</i> e <i>D. serido</i>	16
1.3 Estudo de DNAs satélites em <i>D. antonietae</i> e <i>D. serido</i>	18
1.4 Identificação <i>de novo</i> de DNAs satélites usando a pipeline RepeatExplorer2	20
2. OBJETIVOS	22
2.1 Objetivo Geral.....	22
2.2 Objetivos Específicos	22
2.2.1 Usar dados de sequenciamento de NGS para identificar, quantificar e caracterizar todos os DNAs satélites de <i>D. antonietae</i> e <i>D. serido</i>	22
2.2.2 Determinar a localização cromossômica dos novos DNAs satélites identificados.	22
2.2.3. Utilizar os DNAs satélites para classificar seis genomas da zona híbrida. De posse destas informações, poderemos avaliar se os genomas correspondem as espécies <i>D. antonietae</i> ou <i>D. serido</i> , ou ainda se constituem híbridos entre as duas espécies.	23
3. MANUSCRIPT	24
4. CONCLUSÕES.....	90
5. REFERÊNCIAS	92

1. INTRODUÇÃO

1.1 Os DNAs satélites

O genoma da maioria dos eucariotos é composto, em grande parte, por sequências de DNA repetitivo não codificador de proteínas (Walker, 1971; Plohl e col., 2012; Garrido-Ramos, 2017). Essa porção repetitiva dos genomas varia significativamente entre as espécies, tanto em termos de organização, quanto em abundância, podendo representar, em alguns casos, mais da metade do genoma total de um organismo (Plohl e col., 2012). De acordo com a sua organização, os DNAs repetitivos podem ser classificados em dispersos, como os elementos transponíveis (TEs) e famílias gênicas, ou em tandem, como os mini, micro e DNAs satélites (satDNAs) (Charlesworth e col., 1994; Biscotti e col., 2015; López-Flores, 2012; Garrido-Ramos, 2017). Os satDNAs são caracterizados por longos arranjos contendo milhares de sequências de DNA repetidas em tandem. O tamanho das cópias dos satDNAs varia de 2pb até vários Kb, tendo os satélites mais comuns cópias com tamanho entre 100 e 800pb (Garrido-Ramos, 2017; Lower e col., 2018; Thakur e col., 2021).

Em geral, os grandes arranjos de satDNAs são encontrados em regiões cromossômicas ricas em heterocromatina, como o centrômero, telômero e o cromossomo Y, que é predominantemente heterocromático (Charlesworth e col., 1986, 1994; Kuhn e Sene, 2005). No entanto, existem exemplos de pequenos blocos de repetições de satDNAs encontrados em regiões eucromáticas (Kuhn e col., 2012; Feliciello e col., 2015; Ruiz-Ruano e col., 2016; Garrido-Ramos, 2017) (figura 1).

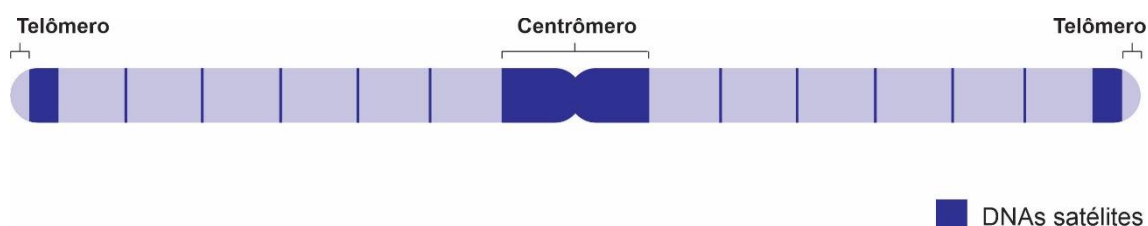


Figura 1. Esquema representativo da localização cromossômica dos DNAs satélites (adaptado de Kuhn e col., 2021)

Dentro de uma mesma espécie, é possível observar a presença de diferentes famílias de DNAs satélites. Geralmente, um mesmo satDNA é compartilhado apenas por espécies próximas filogeneticamente. Em alguns casos, os satélites podem ser espécie-específicos (Garrido-Ramos, 2017). Em termos gerais, o número de famílias de satélites varia consideravelmente entre os organismos. Por exemplo, o gafanhoto *Locusta*

migratória apresenta uma grande diversidade, com a presença de 62 famílias de DNAs satélites identificadas, enquanto que o genoma humano possui uma diversidade menor, com 9 famílias de DNAs satélites (Ruiz-Ruano e col., 2016; Levy e col., 2007; Miga e col., 2015; Garrido-Ramos, 2017).

Por não serem sequências codificadoras de proteínas, seria esperado que os satDNAs evoluíssem de forma neutra ao longo do tempo, levando à diversificação das cópias e à criação de novas variantes (Garrido-Ramos, 2015; Plohl e col., 2012). De fato, os DNAs satélites estão entre os componentes do genoma que apresentam evolução mais rápida. No entanto, dentro da espécie, cópias de um mesmo satDNA são muito homogêneas, resultado de mecanismos moleculares como conversão gênica, recombinação desigual, transposição e amplificação por círculo rolante (Plohl e col., 2012). Consequentemente, repetições de uma mesma família de satDNA apresentam menor variabilidade intraespecífica quando comparadas entre espécies diferentes (Plohl e col., 2012; Ganley, 2013). Esse processo evolutivo que resulta em um padrão de homogeneização entre repetições dentro de uma mesma espécie é denominado evolução combinada (figura 2) (Zimmer e col., 1980; Dover, 1982).

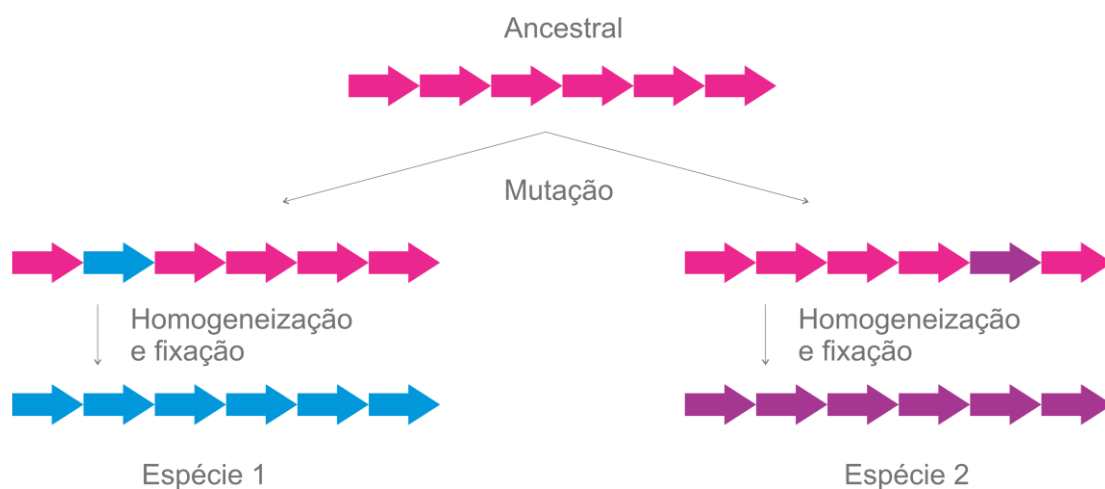


Figura 2. Esquema representativo do processo de evolução combinada em um arranjo de repetições em tandem. (Adaptado de Plohl e col., 2012)

Embora os DNAs satélites não codifiquem proteínas, existem exemplos mostrando que eles podem desempenhar papéis funcionais no genoma, incluindo na regulação gênica (Menon e col., 2014; Joshi e Meller, 2017), na manutenção de centrômeros funcionais (Rošić e col., 2014), na modulação da cromatina (Melters e col., 2013; Dawe e Henikoff, 2006), na fertilidade masculina (Mills e col., 2019) e na organização espacial dos cromossomos (Pathak e col., 2013; Jagannathan e col., 2018, 2019). Além disso, existem evidências de que satDNAs podem contribuir para eventos

de especiação (Mestrovic e col., 1998; Malik e Henikoff, 2001). Por exemplo, satDNAs com uma divergência de aproximadamente 30% foram identificados como responsáveis pela desregulação da segregação cromossômica em híbridos entre *D. melanogaster* e *D. simulans*, resultando em isolamento reprodutivo pós-zigótico entre estas espécies (Ferree e Barbash, 2009).

Os DNAs satélites foram descobertos há mais de 60 anos por meio de experimentos de ultracentrifugação do DNA em gradiente de densidade com cloreto de cério (Kit, 1961; Gall e col., 1971; Plohl e col., 2012). Ao longo do tempo, outras técnicas e tecnologias foram desenvolvidas e incorporadas aos estudos de satélites. O uso de enzimas de restrição para a digestão de satDNAs, análises “dot-blot”, “Southern Blot” e hibridação *in situ* proporcionaram novas perspectivas e conhecimentos sobre essa parte do genoma. Essas técnicas ainda são amplamente utilizadas para a identificação e caracterização de DNAs satélites em diferentes organismos (Castagnone-Sereno e col., 2008; Ruiz-Ruano e col., 2016; Utsunomia e col., 2017; Gatto e col., 2018).

As metodologias experimentais utilizadas na era pré-genômica para a identificação de DNAs repetitivos apresentavam limitações práticas, de tempo e enviesamento, o que dificultava a identificação completa do conjunto de satDNAs de um organismo, conhecido como satelitoma (Garrido-Ramos, 2017). No entanto, o advento do sequenciamento de nova geração (*Next Generation Sequencing*), juntamente com o desenvolvimento de novas ferramentas de bioinformática, revolucionaram os estudos de DNAs repetitivos, preenchendo várias lacunas de conhecimento. Essas tecnologias agora permitem a identificação de potencialmente todos os DNAs repetitivos de um genoma, incluindo o satelitoma (Dias e col., 2014; Fu e col., 2019; Silva e col., 2019).

A era do “big data” teve um impacto significativo nos estudos genéticos com a difusão das técnicas de sequenciamento em todo o mundo. Nesse contexto, merecem destaque o primeiro sequenciamento de um organismo multicelular, *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998), o de *Drosophila melanogaster* (Adams e col. 2000) e o Projeto Genoma Humano (Lander e col., 2001; Venter e col. 2001).

Um crescente número de estudos envolvendo animais e plantas evidenciaram a existência de relações evolutivas entre repetições em tandem, incluindo os satDNAs, e os elementos transponíveis (Meštrović e col. 2015). De modo geral, foi proposto que os TEs

facilitam a dispersão de satDNAs (Meštrović e col. 2015; Zattera e Bruschi 2022; Paço e col. 2019; Kuhn e col. 2021). No grupo *virilis* (*Drosophila*), por exemplo, foram identificadas famílias de satDNAs que tiveram origem a partir da estrutura interna de elementos transponíveis (Dias e col. 2014, 2015; Silva e col. 2023). Por exemplo, o transposon de DNA, Tetris, possui uma repetição em tandem de 220 pb como parte das suas repetições terminais invertidas (TIR – *Terminal Inverted Repeat*) que foi encontrada formando longos arranjos similares aos de satDNAs, sugerindo expansão dessa repetição em tandem (Dias e col. 2014). Outra associação entre satDNAs e TEs, descrita no grupo *virilis*, envolve a família de elementos transponíveis DINE, a qual compreende uma das famílias de TEs mais abundantes de *Drosophila* (Locke e col. 1999; Kapitonov e Jurka 2007; Yang e Barbash 2008; Thomas e col. 2014). Um grupo de DINE, DINE-TR1, contém repetições em tandem centrais (CTRs – *Central Tandem Repeats*) com monômeros de 150 pb. Essas CTRs foram encontradas formando longos arranjos semelhantes aos de DNA satélites nos genomas de *D. virilis* e *D. americana*, sugerindo um evento de expansão destas CTRs (Dias e col. 2015). Além disso, recentemente Silva e col. (2023) relataram mais um evento de satDNA derivado da expansão das CTRs presentes no DINE-TR1 em espécies de *Drosophila* do grupo *montium*.

Em um estudo com mais de 100 espécies, as quais não tiveram seus DNAs centroméricos caracterizados, foi observada uma abundância significativamente maior de uma repetição em tandem específica, em comparação com outras repetições em tandem do genoma (Melters e col. 2013). Dessa forma, esse estudo propôs que o satDNA mais abundante de um genoma deve corresponder ao satélite centromérico.

No contexto do satDNAs centroméricos, é comumente observada a presença de sequências palindrômicas curtas (tipicamente com <10 pb), com díades simétricas (Talbert e Henikoff 2020). Dentro das regiões centroméricas, as sequências palindrômicas passam por mudanças conformacionais, adotando estruturas de DNA não-B como *stem-loops*. Foi sugerido que essas estruturas de DNA desempenham um papel essencial na captação de proteínas específicas, que atuam como identificadores e reguladores centroméricos (Kasinathan e Henikoff 2018; Talbert e Henikoff 2020).

Além disso, Prytkova e col. (2011) demonstraram que uma periodicidade a cada 10 pares de bases de nucleotídeos AA favorece fortemente a formação de nucleossomos, estabilizando assim a histona H3 específica do centrômero (cenH3). Vale ressaltar que uma única volta de DNA corresponde a aproximadamente 10 pares de bases, e sequências

com periodicidade de 10 pares de bases em dinucleotídeos WW, onde W representa A ou T, facilitam o enovelamento de nucleossomos, reduzindo a energia de dobramento envolvida no processo de envolvimento (Struhl e Segal 2013).

Vários satDNAs centroméricos exibem um padrão com uma periodicidade de aproximadamente 10 pares de bases de nucleotídeos AA/AT/TT. Por exemplo, em *D. melanogaster*, ocorre uma expansão de repetições em tandem curtas de 5 a 12 pb contendo dinucleotídeos AA (Talbert e col. 2018). Em *Gallus*, foi observada uma repetição de 21 pb com dinucleotídeos AA e TT separados por 10 pb (Shang e col. 2010). Além disso, uma periodicidade de dinucleotídeos AA também está presente nos satDNAs centroméricos do milho, *Zea mays*, que consiste de 155 pb (Gent e col. 2011).

Alguns monômeros de satDNAs podem formar estruturas complexas de repetições conhecidas como *higher-order repeats* (HOR). As HOR são resultado de amplificações e homogeneização de dois ou mais monômeros adjacentes (Plohl e col. 2012). O modelo de HOR melhor estudado são os arranjos do satélite *alpha*, localizado na região centromérica de todos os cromossomos humanos (Lower e col. 2018). O satDNA *alpha* pode assumir duas organizações principais, sendo encontrado na forma de monômeros individuais (~170 pb) e em HOR, que apresentam de 2 a 34 repetições organizadas em tandem (Sujiwattanasat e col. 2015; Plohl e col. 2014; Willard e Waye 1987; Warburton e Willard 1990; Miga e col. 2014). Nos centrômeros de *D. melanogaster* e *D. simulans*, também foi relatada a presença de satDNAs organizados em HOR (Talbert e col. 2018; Thakur e col. 2021).

1.2 O gênero *Drosophila*

O gênero *Drosophila* abrange mais de 1.600 espécies descritas, que apresentam ampla variação em termos de coloração, tamanho corporal, meios de desenvolvimento larval e tamanho de seus genomas, entre várias outras características (Powell, 1997; Markow e O'Grady, 2005; Boulesteix e col., 2006; Bosco e col., 2007 O'Grady e DeSalle, 2018). As relações filogenéticas da família Drosophilidae ainda são objeto de estudo, e o gênero é considerado parafilético. Portanto, a história filogenética entre as espécies do gênero ainda é tema de debate entre taxonomistas (O'Grady e DeSalle, 2018).

Entre os subgêneros de *Drosophila*, estão *Drosophila* e *Sophophora*, que se divergiram há cerca de 40-60 milhões de anos (Russo e col., 1995; Tamura e col., 2004). A espécie mais conhecida e estudada do gênero, *D. melanogaster*, pertence ao subgênero

Sophophora, que integra outras 344 espécies. Por outro lado, o subgênero *Drosophila* abriga mais de 80% das espécies do gênero, mas ainda é o menos investigado (O’Grady e DeSalle, 2018). *Drosophila melanogaster* foi a primeira espécie do gênero a ter seu genoma sequenciado, no ano 2000, e como resultado de projetos colaborativos, várias outras espécies do gênero tiveram seus genomas sequenciados (Richards e col., 2005; *Drosophila* 12 Genomes Consortium col., 2007; Bronksi e col., 2020; Conner e col., 2021; Kim, Wang e col., 2021; Moreyra e col., 2023).

A disponibilidade de vários genomas sequenciados de *Drosophila* possibilita o estudo de diversos aspectos dos satDNAs, como a investigação de suas origens, estrutura, organização, modos de dispersão, função, evolução e os impactos dos satélites na arquitetura dos genomas.

1.2.1 O grupo *repleta* e as espécies do *cluster buzzatii*

As espécies do grupo *repleta* são consideradas excelentes modelos para estudos ecológicos e evolutivos, uma vez que representam uma das maiores radiações dentro do gênero (Patterson e Stone, 1952; Throckmorton, 1975; Vilela, 1983; Wasserman, 1992; Markow e O’Grady, 2005). A maioria das espécies desse grupo é cactofílica e usa tecidos de cactos em decomposição como sítio de desenvolvimento larval (Ruiz e Heed, 1988; Ruiz e col., 1990). Não é possível determinar o número preciso de espécies do grupo, tendo em vista que alguns táxons foram propostos com sinonímia, além de ser um grupo com alto número de espécies crípticas, cujos limites entre as espécies são pouco conhecidos (Oliveira e col., 2005 e 2008). No entanto, estima-se que o grupo compreenda mais de cem espécies descritas (Oliveira e col. 2012).

O grupo *repleta* é subdividido em seis subgrupos: *mulleri*, *hydei*, *mercatorum*, *repleta*, *fasciola* e *inca*. Estes subgrupos estão separados de *D. melanogaster* por mais de 40 milhões de anos (Powell, 1997) e são ainda subdivididos em complexos, clusters e subclusters com base nos padrões cromossômicos, na genitália masculina (o edeago) e nas associações ecológicas (Patterson, 1943; Wharton, 1944; Wasserman, 1962; Rafael e Arcos, 1989).

O *cluster buzzatii* faz parte do subgrupo *repleta* e é composto por sete espécies: *D. borborema*, *D. serido*, *D. seriema*, *D. gouveai*, *D. buzzatii*, *D. antonietae* e *D. koepferae* (Manfrin e Sene, 2006) (figura 3). A distribuição geográfica desse *cluster* está diretamente associada à presença e distribuição de seus cactos hospedeiros na América

do Sul. Em função dessa associação ecológica, as espécies estão distribuídas em áreas de vegetação xerófila (Manfrin e Sene, 2006).

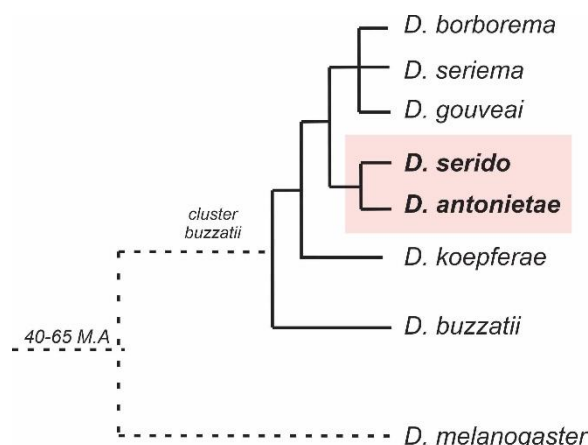


Figura 3. Filogenia do *cluster buzzatii* (adaptado de Kuhn e col. 2021).

O *cluster buzzatii* é um grupo monofilético que tem as relações filogenéticas entre suas espécies bem estudadas a partir de dados de inversões cromossômicas fixas (Ruiz e Wasserman, 1993), do gene nuclear *Xhantine dehidrogenase* (Rodriguez-Trelles e col., 2000) e do gene nuclear *period* (Franco e col., 2010), o que facilita o estudo da evolução dos DNAs satélites dessas espécies (Manfrin e col., 2001; Santos e col., 2009; Franco e col., 2010). Com exceção de *D. buzzatii*, que se tornou subcosmopolita, as demais seis espécies são endêmicas da América do Sul (Fontdevila e col., 1981; Barker e col., 1985; Manfrin e Sene, 2006).

O cariótipo básico das espécies do *cluster buzzatii* consiste em um par de cromossomos sexuais, quatro pares de autossomos telocêntricos com mesmo tamanho, e um par de microcromossomos (Baimai e col., 1983; Kuhn e col., 1996) (figura 4).

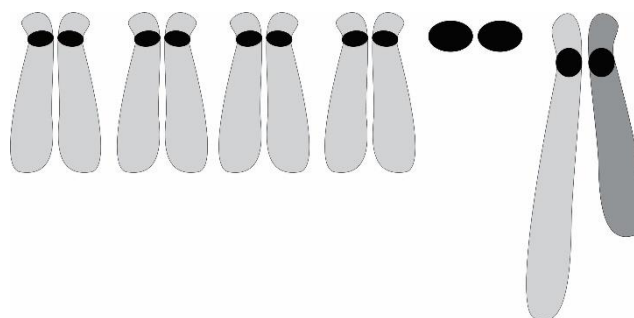


Figura 4. Idiograma do cariótipo básico das espécies do *cluster buzzatii*.

No contexto de satDNAs nas espécies do grupo *repleta*, *D. buzzatii*, *D. mojavensis* e *D. seriema* apresentam contribuição de satélites para o DNA genômico total que varia de 1,5 a 2,9% (Lima e col. 2017). É uma contribuição baixa quando comparamos com o conteúdo de satDNAs em outras espécies do gênero, em que os DNAs satélites podem

corresponder a mais de 20% do DNA genômico (Bosco e col. 2007; Craddock e col. 2016; Garrido-Ramos 2017).

Além disso, é válido ressaltar que o conteúdo de heterocromatina estimado em *D. mojavensis*, de 2%, também é considerado baixo, se comparado com outras espécies de *Drosophila*, com média de 15% (Bosco e col. 2007). As espécies do grupo *repleta* também apresentam genomas menores (genomas de *D. mojavensis* e *D. buzzatii* foram estimados em 150 Mb), em relação a várias espécies do gênero que, geralmente, apresentam genomas maiores que 180 Mb (Bosco e col. 2007; Gregory e Johnston 2008). Dessa forma, essa variação encontrada sugere a hipótese de que a baixa contribuição de satDNAs (<5%) para o genoma das espécies do grupo *repleta* seja resultado de uma pressão seletiva (Lima e col. 2017; Kuhn e col. 2021)

1.2.2. As espécies *D. antonietae* e *D. serido*

Drosophila antonietae e *D. serido* são espécies crípticas, que podem ser identificadas através de marcadores genéticos e, morfológicamente, pelo edeago (Manfrin e Sene, 2006) As populações de *D. antonietae* estão associadas principalmente ao cacto *Cereus hildmaniannus*, mas nas regiões costeiras, usam ocasionalmente o cacto *Opuntia monocantha* (Monteiro e Sene, 1995; Manfrin e Sene, 2006). As populações de *D. serido* usam várias espécies de cactos como sítios de desenvolvimento larval, incluindo *C. hildmaniannus* e *O. monocantha*. Apesar das duas espécies apresentarem distribuição majoritariamente alopátrica, foi descrita uma área de contato secundário entre as espécies na costa de Santa Catarina (figura 5) (Manfrin e Sene, 2006). Existem algumas evidências de eventos de hibridação entre essas espécies nesta área de contato, com introgressão de haplótipos mitocondriais (Ruiz e col., 2000; Kokudai e col., 2011; Barrios-Leal e col., 2021).

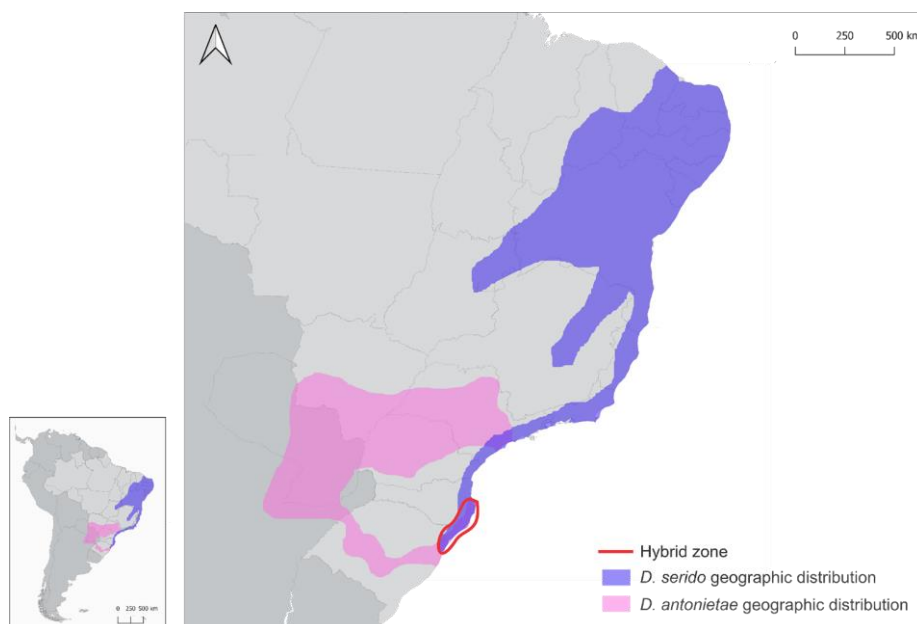


Figura 5. Distribuição geográfica de *D. serido* e *D. antonietae*, incluindo a zona de contato secundário em vermelho.

As áreas de contato secundário, também conhecidas como zonas híbridas (HZ), representam estágios de diferenciação desde o nível de populações até o nível de espécies. Portanto, esses locais proporcionam uma oportunidade única para investigar, em condições naturais, a formação de barreiras reprodutivas (Howard, 1993; Noor, 1999), o papel da seleção natural na manutenção dos limites entre as espécies (Barton e Hewitt, 1981), a arquitetura genética das diferenças entre as espécies (Harrison, 1990; Rieseberg e col., 1999) e até mesmo eventos de especiação (Hewitt, 1998).

A formação e o estabelecimento das zonas híbridas resultam de uma interação complexa entre fatores históricos, genéticos, ecológicos, fisiológicos e comportamentais das espécies em interação (Howard, 1993). Dessa forma, várias abordagens genéticas foram propostas para explicar a dinâmica dessas zonas, uma vez que a sua formação depende de fatores históricos, mas também de fatores intrínsecos ao táxon (Moore, 1977; Rand e Harrison, 1989).

Dentro do gênero *Drosophila*, o estudo de eventos de hibridação é ressaltado principalmente por existirem diversos agrupamentos de espécies com pouca diferenciação morfológica, o que indicaria pouca diferenciação genética e maior probabilidade de hibridização, e pela facilidade de criação das espécies em laboratório (Bock, 1978).

Em grande parte, os estudos de hibridização em *Drosophila* foram realizados com espécies nas quais não se têm evidências de eventos de hibridação na natureza (Markow,

2000). A multiplicidade de barreiras reprodutivas entre espécies do gênero é uma regra entre as espécies de *Drosophila*, como pode ser exemplificado pelos sete tipos de barreiras reprodutivas entre *D. pseudoobscura* e *D. persimilis* (Dobzhansky, 1951). Desse modo, os experimentos de laboratórios não conseguem reproduzir fielmente a diversidade de condições naturais dos cruzamentos entre as espécies. Portanto, é fundamental direcionar a investigação para organismos que tenham sido amplamente estudados em termos evolutivos, a fim de aproveitar os modelos genéticos existentes e o conhecimento prévio sobre essas espécies. Nesse contexto, as espécies do *cluster buzzatii* *D. antonietae* e *D. serido* representam um modelo biológico para o estudo de zonas de contato em populações naturais.

1.3 Estudo de DNAs satélites em *D. antonietae* e *D. serido*

Em trabalhos anteriores usando as metodologias da era pré-genômica, foram identificados em *D. antonietae* e *D. serido* duas famílias de DNAs satélites: o pBuM e o DBC-150 (Kuhn e Sene, 2005; Kuhn e col., 2007). O pBuM foi identificado, pela primeira vez, em *D. buzzatii* e foi o primeiro satDNA descrito nas espécies do grupo *repleta* (Kuhn e col., 1999). Dentro do *cluster buzzatii*, existem duas variantes principais do pBuM: variantes de 190 pb, chamadas de *alpha* ou pBuM-1, e variantes de 370 pb, chamadas de *alpha/beta* ou pBuM-2, formadas por uma sequência *alpha* seguida de uma sequência não-homóloga de 180 pb, denominada de *beta* (figura 6). O pBuM apresentou características de evolução combinada, sendo um potencial marcador taxonômico entre as duas espécies. O DBC-150, com repetições de 150pb, por sua vez, não apresentou padrões de evolução combinada (Kuhn e col., 2007).

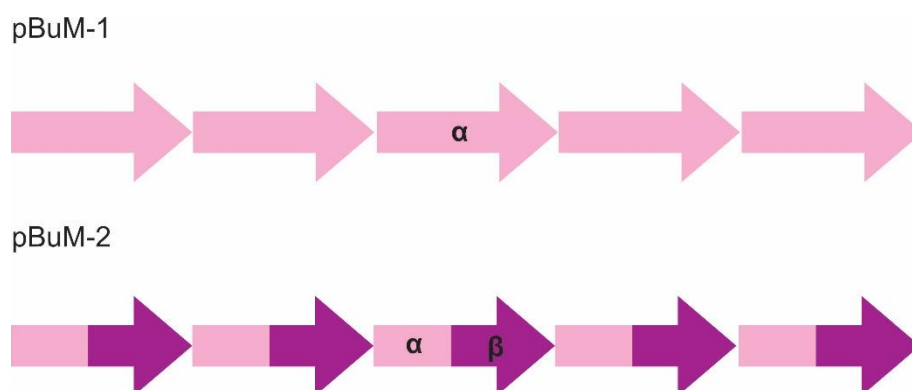


Figura 6. Representação esquemática da família de satDNA pBuM.

Estudos de hibridação fluorescente *in situ* (FISH) em *D. serido* e *D. antonietae* mostraram que o DBC-150 está localizado exclusivamente no par de microcromossomos, enquanto o pBuM está localizado na região centromérica de todos os cromossomos, mas

ausente no cromossomo Y (Kuhn e col., 2007, 2008). Com base nestas observações, foi proposto que o pBuM, provavelmente, é o principal componente do centrômero destas espécies (Kuhn e col. 2008).

Além do pBuM e do DBC-150, destacam-se a identificação do elemento transponível *PERI* e o satDNA SSS139 (Franco e col. 2008; Kuhn e Heslop-Harrison 2011). O *PERI* é um TE que compartilha várias estruturas e características com o DINE-1, um TE abundante e amplamente encontrado no gênero *Drosophila* (Locke e col. 1999), incluindo a presença de repetições em tandem internas (Kuhn e col. 2021).

Em *D. serido*, o *PERI* foi observado nos seis pares cromossômicos, estando presente na região proximal dos quatro pares de cromossomos telocêntricos e do cromossomo X, além de estar distribuído ao longo do cromossomo Y, mas ausente nos telômeros. Sinais mais distais também foram observados no cromossomo X e em um par de autossomos, enquanto sinais mais fracos foram detectados nos microcromossomos (Kuhn e Heslop-Harrison, 2011). Resultados de experimentos de hibridação *in situ* em fibras de DNA estendidas levantaram a hipótese de que as repetições em tandem internas do *PERI* sofreram amplificação em *D. serido* (Kuhn e Heslop-Harrison, 2011), similar ao relatado com elementos DINE-1 de outras espécies de *Drosophila* ((Dias e col. 2015; Silva e col. 2023).

O satDNA SSS139 foi encontrado em todas as espécies do *cluster buzzatii*, com exceção de *D. buzzatii* (Franco e col. 2008). Este satDNA possui sequências ricas em AT (69%) e monômeros de 139 pb, que apresentam repetições internas de 14 a 16 pb, sugerindo que este satDNA teve origem a partir da duplicação de sequências menores. A localização cromossômica do SSS139 ainda não foi determinada (Franco e col. 2008; Kuhn e col. 2021). Além disso, em *D. seriema*, ao contrário do esperado, o SSS139 não foi encontrado em tandem, apesar de representar 0,5% do genoma (Lima e col. 2017).

A figura 7 mostra a distribuição cromossômica conhecida dos elementos repetitivos encontrados em *D. serido* e *D. antonietae*. O sequenciamento recente do genoma de *D. antonietae* e *D. serido* fornece uma perspectiva promissora para uma investigação mais aprofundada dos satDNAs destas espécies, bem como para a identificação e estudo de novos.

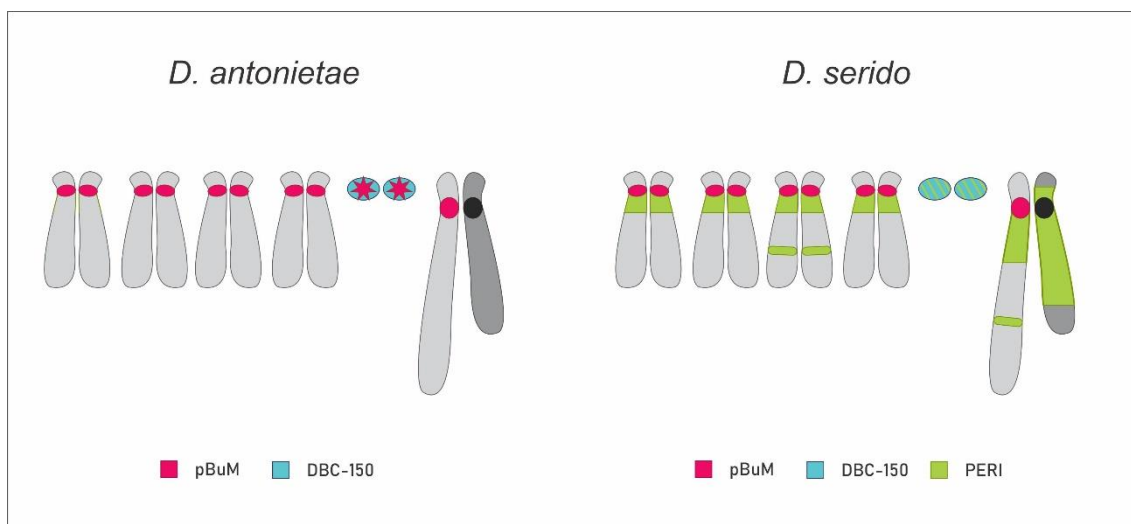


Figura 7. Idiograma representativo da localização cromossômica dos elementos repetitivos identificados em *D. antonietae* e *D. serido* (Kuhn e col. 2007, 2008; Franco e col. 2008; Kuhn e Heslop-Harrison 2011).

1.4 Identificação *de novo* de DNAs satélites usando a pipeline RepeatExplorer2

Apesar do desenvolvimento de novas tecnologias de sequenciamento e de novas ferramentas de bioinformática terem revolucionado os estudos genômicos, o processamento de *reads* curtas depois do sequenciamento ainda apresenta limitações e barreiras na montagem das regiões repetitivas, em função da alta repetitividade e da longa extensão das cadeias de satDNAs (Grady e col., 1992; Sun e col., 2003; Plohl e col., 2014). Uma alternativa para contornar esses problemas foi por meio da identificação de DNAs satélites a partir de *reads* brutas não montadas, geradas pelo sequenciamento de DNA. Novák e colaboradores (2013 e 2017) implementaram essa abordagem nas *pipelines* computacionais RepeatExplorer e TAREAN (*Tandem Repeat Analyzer*).

O RepeatExplorer e o TAREAN foram desenvolvidos para identificar elementos repetitivos em genomas eucariotos. Ambos foram desenvolvidos com base na montagem de grafos de Brujn (Compeau e col., 2011). Em primeiro lugar, as *reads* são agrupadas em *clusters* de acordo com a similaridade de sequências, depois são reorientadas para a formação de *k-mers* e, por fim, são geradas as sequências consenso. Especificamente no TAREAN, os grafos gerados são pontuados em relação à presença de cadeias em tandem e os *clusters* identificados são classificados como satélites *high confidence* ou *low confidence*. Essas estimativas são determinadas de acordo com os índices “*Connected componet index (C)*” e “*Pair completeness index (P)*”. O índice C está relacionado às sequências repetidas em tandem. O índice P mede a razão entre o número de *paired-end reads* completas e “quebradas”, que diz respeito ao tamanho dos arranjos contínuos das

repetições em tandem, característica importante para a classificação correta de satDNAs (Novák e col., 2017).

Enquanto que o RepeatExplorer detecta diferentes tipos de elementos repetitivos, dispersos ou em tandem, o TAREAN por sua vez identifica apenas elementos repetitivos organizados em tandem. A pipeline RepeatExplorer2 é uma nova versão da pipeline computacional RepeatExplorer (Novák e col., 2020). O *workflow* básico foi mantido, com novos módulos que fazem a anotação automática dos clusters com base na similaridade com sequências referência do banco de dados. Além disso, a *pipeline* também executa uma versão integrada da ferramenta TAREAN.

A *pipeline* RepeatExplorer2 requer como *input* um arquivo FASTQ único, formado por *single* ou *paired-end reads* geradas pelo sequenciamento de genoma, pela abordagem *shotgun*. Para dar início as análises, é preciso que as *reads* possuam tamanho uniforme (100-200pb) e que o número total de *reads* seja menor que a cobertura de 1x do genoma (a cobertura recomendada pelos autores é de 0.01 – 0.05x). É preciso pré-filtrar as *reads*, com base na análise de qualidade, sendo aceitável um *score* de qualidade maior ou igual a 10 em 95% das bases e sem a presença de “Ns”. Somente *reads* completas são incluídas na análise.

Nos últimos anos, muitos trabalhos têm usado as *pipelines* supracitadas para o estudo de sequências repetitivas em várias espécies de eucariotos (García e col., 2015; Ruiz-Ruano e col., 2016; Robledillo e col., 2018; Sena e col., 2020; Valeri e col., Dias e col., 2021). A disponibilização recente de dados genômicos de *D. antonietae* e *D. serido* possibilita a investigação de potencialmente todos os DNAs satélites dessas espécies, contribuindo, dessa forma, para o preenchimento de lacunas e elucidação de vários aspectos sobre os satDNAs. Além disso, foram sequenciados seis genomas da zona híbrida, o que proporciona uma oportunidade para usar os satélites como marcadores taxonômicos e contribuir para estudos evolutivos nesta zona híbrida.

2. OBJETIVOS

2.1 Objetivo Geral

O presente trabalho teve como objetivo estudar todo o conjunto de DNAs satélites de duas espécies próximas filogeneticamente do *cluster buzzatii*: *D. antonietae* (populações de Tibari, PR e Sertãozinho, SP0) e *D. serido* (populações de Lajes Pintadas, RN e Milagres, BA). Além disso, teve como objetivo usar os DNAs satélites encontrados como marcadores taxonômicos em seis genomas da zona de contato secundário em Florianópolis, SC.

2.2 Objetivos Específicos

2.2.1 Usar dados de sequenciamento de NGS para identificar, quantificar e caracterizar todos os DNAs satélites de *D. antonietae* e *D. serido*.

De posse dessas informações, pretendemos testar as seguintes hipóteses:

2.2.1.1 *Existem mais DNAs satélites nestas duas espécies do que os dois (pBuM e DBC-150) previamente identificados na era pré-genômica.*

2.2.1.2 *A contribuição dos satDNAs para o conteúdo de DNA genômico deve ser baixa (<5%), assim como encontrado em outras espécies do grupo repleta.*

2.2.1.3 *Alguns dos DNAs satélites encontrados podem estar evolutivamente relacionados com elementos transponíveis.*

2.2.1.4 *O satDNA mais abundante deve corresponder ao satDNA centromérico.*

2.2.1.5 *As repetições internas do elemento transponível PERI devem ter sofrido ampliações e expansões, formando arranjos de DNAs satélites em D. serido.*

2.2.2 Determinar a localização cromossômica dos novos DNAs satélites identificados.

De posse dessas informações, pretendemos testar as seguintes hipóteses:

2.2.2.1 Os satDNAs devem ter localização predominante na heterocromatina.

2.2.2.2 Os DNAs satélites centroméricos, mesmo sendo não-homólogos, devem apresentar características em comum, como o tamanho da repetição, conteúdo AT e propensão a formar estruturas secundárias.

2.2.2.3. A taxa de evolução combinada pode ser diferente entre DNAs centroméricos e não-centroméricos e entre satélites presentes em um cromossomo em relação a vários cromossomos.

2.2.2.4. Novos DNAs satélites possivelmente se localizam na heterocromatina do Y e na heterocromatina não-centromérica do X.

2.2.3. Utilizar os DNAs satélites para classificar seis genomas da zona híbrida. De posse destas informações, poderemos avaliar se os genomas correspondem as espécies *D. antonietae* ou *D. serido*, ou ainda se constituem híbridos entre as duas espécies.

3. MANUSCRIPT

“New insights on the satellitome and centromeres of the two closely related species, *Drosophila antonietae* and *D. serido* (*buzzatii* cluster, *repleta* group)”

Autores: Ana Mattioli Laborne do Nascimento, Dora Yovana Barrios-Leal, John S. Heslop-Harrison, Maura Helena Manfrin & Gustavo C.S. Kuhn

Abstract

Satellite DNA (satDNA) is a highly repetitive DNA composed of tandemly arranged copies typically found forming long arrays in the heterochromatic regions of chromosomes. SatDNAs can make up a significant portion (>25%) of the genomic DNA. SatDNAs are known to evolve rapidly and may contribute to the establishment of genetic incompatibilities between incipient species. *Drosophila antonietae* and *D. serido* are two closely related cactophilic species that belong to the *buzzatii* cluster within the *repleta* group. Their satDNAs were first investigated in pre-genomic era, when studies revealed the presence of only two satDNAs, pBuM and DBC-150. FISH experiments revealed that pBuM was likely the main centromere component of all chromosomes except the Y, while DBC-150 was restricted to microchromosomes. The availability of the sequenced genomes of these two species provided an opportunity to revisit and expand our knowledge of their satDNAs and, consequently, in *Drosophila* in general. In order to identify and quantify satDNAs of each species, we used WGS Illumina reads and RepeatExplorer2. Our results revealed that both species share the same collection of five satDNAs. Three of them (pBuM, DBC-150 and CDSTR138) have been previously described in other species from *repleta* group. The remaining two are novel and were named: SAT197/230 and SAT109/104. We also identified CDSTR198, a euchromatic minisatellite DNA previously found in other species from *buzzatii* cluster. The contribution of satDNA to the total genomic DNA of both species is relatively low (<5%). In comparison, many other *Drosophila* species have higher satDNA content (often >20%). The low satDNA content in both species may indicate possible selective constraints on genome size or heterochromatin content. Our maximum likelihood trees showed four satellites (pBuM, DBC-150, CDSTR138 and SAT197/230) with a species-specific clustering of repeats, indicating a concerted mode of molecular evolution. The calculated interspecific divergence for each satDNA ranges from 0.58%-14.4%. Such low levels of satDNA divergence might have been one of the factors that contributed for

establishment of a hybrid zone between the two species, as revealed in previous studies. Our FISH experiments revealed that CDSTR138 is likely the main centromeric satDNA in *D. serido*, while pBuM is more distally located, towards the heterochromatic short arms of chromosomes. SAT197/230 and SAT109/104 were found to be restricted to the sex chromosomes. The SAT197/230 is present on the Y chromosome and is organized in three arrays, one in the centromere, and SAT109/104 is present in the centromeric region of the X chromosome. In addition, we found that all centromeric satDNAs are enriched for dyad symmetries and expected to form stable secondary structure, feature considered important for centromeric functions. We confirmed CDSTR198 is a minisatellite in *D. serido*, because it is scattered distributed in the euchromatic arms of polytene chromosomes. Overall, our study confirms the efficiency of RepeatExplorer2 pipeline for identifying satDNAs in *Drosophila* and provides new insights into genome evolution and centromere structure in these species. Furthermore, it revealed new genetic markers for species identification, especially in studies involving the hybrid zone. In this context, our analysis of genomes from the hybrid zone suggests that individuals can be classified as either *D. antonietae* or *D. serido*, with no evidence of hybridization in the sample we studied.

Introduction

Eukaryotic genomes harbor a rich diversity of DNA sequences, including both unique and repetitive elements (Britten and Kohne 1968). Repetitive DNA, which constitutes a significant portion of the genome, exhibits remarkable variation in terms of organization and abundance among species and commonly outmatches 50% of the genome (Plohl *et al.* 2012). Classification of repetitive DNA is primarily based on its genomic arrangement, with two prominent categories: i) the dispersed elements, such as transposable elements (TEs) and gene families; and ii) tandemly arranged sequences, including mini, micro and satellite DNAs (satDNAs) (López-Flores and Garrido-Ramos 2012; Garrido-Ramos 2017).

Satellite DNA is a class of repetitive DNA that exhibits several distinguishing features. They display variations in nucleotide sequence, sequence complexity, repeat unit length and abundance. However, despite this diversity, satDNAs share two key properties: the capacity to form extensive arrays and their organization as tandem repeats (Plohl *et al.* 2012).

SatDNAs are the major component of heterochromatic regions of chromosomes, such as centromeres and telomeres and are also enriched on the highly heterochromatic Y chromosome (Charlesworth *et al.* 1986, 1994; Kuhn and Sene 2005). Recent studies have challenged the notion that satDNAs are confined solely to heterochromatin, since evidence has been accumulated indicating the presence of small arrays of satDNAs in the euchromatin (Kuhn *et al.* 2012; Feliciello *et al.* 2015; Ruiz-Ruano *et al.* 2016; Garrido-Ramos 2017).

Several different satellite DNAs may be present in a single species. For instance, species such as the pea, *Pisum sativum*, can possess up to 15 satDNAs families, while the locust, *Locusta migratoria*, boasts 62 families (Macas *et al.* 2007, Ruiz-Ruano *et al.* 2016). On the other hand, the human genome harbors nine satellite families (Levy *et al.* 2007, Miga *et al.* 2015). Despite this extensive variation, a pattern emerges where a single or a few satDNA families tend to be predominant within each species. In the human genome for example, the centromeric *alpha* satDNA stands out as the most abundant, constituting over half of the total satellite DNA content (Lander *et al.* 2001; Levy *et al.* 2007; Miga 2015).

Typically, satDNA families are shared among closely related species, as observed in the genomes of various beetle species belonging to the Tenebrionidae family (Meštrovic *et al.* 1998, Mravinac *et al.* 2002, Mravinac *et al.* 2005). These species exhibit shared satDNA families, although one or two major families tend to be the most abundant ones. It is also noteworthy that satDNAs families can be species-specific, in accordance with the fact that satellites are among the fastest evolving component of the genomes. (Garrido-Ramos 2017).

Satellite DNAs undergo rapid molecular evolution between species but are highly homogenous within species, primarily due to molecular mechanisms such as unequal crossing-over, gene conversion, transposition and rolling circle amplification (Henikoff *et al.* 2001; Plohl *et al.* 2012). These mechanisms contribute to the homogenization of satDNAs within a genome, ultimately leading to the fixation of specific sequences within a population or species (Dover 1982, 1986). In this way, repeats belonging to the same satDNA family present lower intraspecific variability compared to interspecific ones (Plohl *et al.* 2012; Ganley 2013). This process of evolution is called concerted evolution (Zimmer *et al.* 1980; Dover 1982). The mechanisms underlying concerted evolution play a crucial role in shaping the structure and evolution of satDNA families, leading to their distinct patterns of sequence conservation and variability within and between species.

Although satellite DNAs do not encode proteins, they may play functional roles in the genome. Several studies have shown that satDNAs are involved in gene regulation (Menon *et al.*, 2014; Joshi and Meller, 2017), maintenance of functional centromeres (Rošić *et al.*, 2014), modulation of chromatin (Melters *et al.*, 2013; Dawe and Henikoff, 2006), male fertility (Mills *et al.*, 2019), and spatial organization of chromosomes (Pathak *et al.*, 2013; Jagannathan *et al.*, 2018, 2019). Furthermore, satDNAs may contribute to speciation events (Mestrovic *et al.*, 1998; Malik and Henikoff, 2001). For example, satDNAs with an approximate divergence of 30% have been identified as responsible for the disruption of chromosomal segregation in hybrids between *D. melanogaster* and *D. simulans*, resulting in post-zygotic reproductive isolation (Ferree and Barbash, 2009).

With the advent of next-generation sequencing technologies, the genomes of several *Drosophila* species have been sequenced, allowing more in-deep studies of repetitive DNA elements (Richards *et al.* 2005; Drosophila 12 Genomes Consortium 2007; Bronski *et al.* 2020; Conner *et al.* 2021; Kim *et al.* 2021; Moreyra *et al.* 2023). Importantly, the development of novel bioinformatics tools has facilitated the

identification of repetitive DNAs from raw sequencing data (García *et al.* 2015; Ruiz-Ruano *et al.* 2016; Robledillo *et al.* 2018; Silva *et al.* 2019, 2023).

In the *Drosophila* genus, satDNAs have been found to constitute a significant proportion of the genome, accounting for over 20% in several species, including the well-studied *D. melanogaster*, and can even reach up to 70%, as reported in *D. cyrtoloma* (Bosco *et al.* 2007; Craddock *et al.* 2016). In *Drosophila*, satDNA repeat units exhibit a wide size range, spanning from a few base pairs, often less than 10bp, to approximately 400bp (Palomeque and Lorite 2008; Melters *et al.* 2013; Silva *et al.* 2023).

Despite the fact that the *Drosophila* subgenus encompasses over 80% of all known *Drosophila* species, satDNAs studies have been predominantly focused on *D. melanogaster* and other species from the *Sophophora* subgenus (O’Grady and DeSalle 2018). Within the *Drosophila* subgenus, the *virilis-repleta* radiation is recognized as one of the most extensive and diverse, with over 200 species (Throckmorton 1975; O’Grady and DeSalle 2018).

The New World *repleta* group has undergone substantial divergence from the *D. melanogaster* group, with an estimated separation time of over 40 MY (Powell 1997). Most species from this group are cactophilic, relying on fermenting cactus tissues as a resource to complete their life cycles (Ruiz e Heed, 1988; Ruiz *et al.* 1990). However, it is noteworthy that some species within *repleta* group exhibit a broader ecological versatility, using a diverse range of resources and occupying habitats spanning from wet and tropical to temperate forests (Ruiz and Heed 1988; Ruiz *et al.* 1990; Oliveira *et al.* 2012). This ecological plasticity highlights the adaptive flexibility and ecological success of species from *repleta* group across different environments and resource availability.

Studies on satDNA in species from the *repleta* group have been relatively limited. Significant attention has been directed towards investigating satDNAs in seven species belonging to the *buzzatii cluster* (Kuhn *et al.* 2021). The *buzzatii cluster* is a monophyletic group within *repleta* group that comprises seven sibling species: *D. buzzatii*, *D. koepferae*, *D. serido*, *D. antonietae*, *D. seriema*, *D. gouveai* and *D. borborema* (Manfrin and Sene, 2006). These species are characterized by their cactophilic nature and are predominantly endemic to South America, with the exception of *D. buzzatii*, which has achieved a cosmopolitan distribution (Barker *et al.* 1985; Fontdevila, 1989; Manfrin and Sene, 2006). Notably, the short divergence time of approximately 1-2 MY among most

species provides an interesting opportunity to investigate early stages of satDNA evolution (Fransak, 2011; Oliveira et al. 2012; Kuhn et al. 2021).

In species belonging to the *buzzatii* cluster, three non-homologous satDNA families have been identified and studied so far. The first satDNA to be described in the *repleta* group was the pBuM family, initially discovered in *D. buzzatii* (Kuhn et al. 1999). The pBuM satDNA family can be further classified into two subfamilies known as pBuM-1 and pBuM-2. The first is characterized by tandemly arranged *alpha* repeats, which consists of 190bp in length, while the latter is composed of tandemly arranged *alpha/beta* repeats with a longer repeat unit length of 370bp. As typically found for satellite DNAs in general, the pBuM satDNA family displays a slightly AT-rich nature (Kuhn and Sene 2005; Kuhn et al. 2008).

The presence of pBuM-1 copies has been observed in species belonging to both the *mulleri* and *buzzatii* complexes within the *repleta* group. Specifically, pBuM-1 repeats have been found in *D. mojavenensis*, from the *mulleri* complex, and in *D. borborema*, *D. serido*, *D. antonietae*, *D. koepferae* and *D. buzzatii*, from *buzzatii* complex. On the other hand, the pBuM-2 repeats appear to be exclusive to the *buzzatii* complex and have been identified in *D. borborema*, *D. gouveai*, *D. seriema*, *D. serido*, *D. antonietae*, *D. koepferae* and *D. buzzatii* (Kuhn et al. 2007, 2008; Lima et al. 2017).

The DBC-150 satDNA family has been only detected in species belonging to the *buzzatii* complex (Kuhn et al. 2007). The DBC-150 repeats are primarily composed of units with a length of 150bp and exhibit a slightly GC-rich nucleotide composition (Kuhn et al. 2007, 2021).

The CDSTR138 satDNA, characterized by repeats units of 138bp in length and an AT-rich nucleotide composition, has been identified in *D. seriema*, within the *buzzatii* cluster (Lima et al. 2017). However, no information regarding the presence or characterization of CDSTR138 in other species from the *buzzatii* cluster is available to date (Kuhn et al. 2021).

The species *D. antonietae* and *D. serido* are two closely related species within the *buzzatii* cluster, that can be distinguished by various characteristics, including the metaphasic chromosomes, alloenzymatic patterns, mitochondrial haplotype, satDNAs and morphological features, such as the male genitalia, the aedeagus (Baimai et al. 1983; Manfrin et al. 2001; Mateus and Sene 2003; Kuhn and Sene 2005; Franco et al. 2006).

Although *D. antonietae* and *D. serido* primarily exhibit an allopatric distribution, there is evidence of a secondary contact zone occurring in the southern coast of Brazil, specifically in Santa Catarina (Manfrin and Sene 2006). The identification of this sympatric region was achieved through the study of the aedeagus morphology from individuals collected in Santa Catarina and Rio Grande do Sul states. Furthermore, studies using genetic markers such as chromosome inversions and mitochondrial DNA suggest the occurrence of hybrid individuals between *D. antonietae* and *D. serido* in this region (Ruiz *et al.* 2000; Kokudai *et al.* 2011). In addition, Barrios-Leal *et al.* (2021) provided further evidence supporting the occurrence of asymmetric bidirectional introgression between both species. Their investigation employed multiple sources of markers, including mitochondrial and nuclear sequences, microsatellite loci, as well as morphometric of wings and male genitalia.

In the present work, we took advantage of the availability of sequenced genomes from *D. antonietae* and *D. serido* to conduct a comprehensive investigation and characterization of the entire repertoire of satellite DNAs, referred as the “satellitome”, in both species. Moreover, we aimed to use satDNAs as markers to investigate the taxonomic status of six genomes of the hybrid zone between *D. antonietae* and *D. serido*.

Material and Methods

RepeatExplorer2 analyses

We employed the RepeatExplorer2 pipeline, a well-established method for characterizing repetitive DNA elements, to investigate *D. antonietae* and *D. serido* satellite DNAs (Novák *et al.* 2020). RepeatExplorer2 uses graph-based clustering of Next Generation Sequencing (NGS) raw reads and operates on the principle that a distinct cluster represents each repetitive element. Additionally, the pipeline also performs an integrated version of TAREAN (Tandem Repeat Analyzer) tool, which is specifically designed for satDNA identification, based on topology of their clusters graph (Novák *et al.* 2017). All clusters resulted from RepeatExplorer are analyzed and investigated for the presence of circular structures, which is a characteristic of tandem repeats.

For our analyses, we used the raw sequencing data from 10 *Drosophila* genomes (Table 1; Figure 1). These Illumina paired-end reads were generously shared with us by Dr. Maura Helena Manfrin (University of Sao Paulo). All the data processing and analyses were performed on Galaxy Platform (Afgan *et al.* 2018).

To ensure the quality of the sequencing reads, we initially employed “FastQC”. Subsequently, the sequences were converted to a single fastqsanger format using “FASTQ Groomer” tool (Sanger and Illumina 1.8+). Further preprocessing of the paired-end reads was carried out using the “Preprocessing of paired-end reads” tool. This involved specific settings such as read trimming to approximately 100bp in length, with a quality cutoff of 10 and a threshold of 95% of reads above the cutoff. The resulting interlaced output file was used as input in “RepeatExplorer2 clustering” tool. Within this tool, we selected the following parameters: Pair-end reads: yes; Read sampling: no; Select taxon and protein domain database version (REXdb): Metazoa version 3.0; Advanced options: no; Select queue: long; Modify parameters: default. The resulting HTML archives for each individual were downloaded and analyzed separately.

Table 1. Strains and sample number of genomes of *D. antonietae*, *D. serido* and from the hybrid zone with their respective collection sites.

		Nº of analyzed			
	Strains	genomes	Latitude	Longitude	Locality
<i>D. antonietae</i>	R00	1	-24.6200	-50.52	Guarapuava - PR
	R73	1	-21.18	-48.09	Sertãozinho - SP
<i>D. serido</i>	R37	1	-6.28	-38.18	Lajes Pintadas - RN
	J92	1	-11.2000	-39.9000	Milagres - BA
Hybrid Zone	N61	4	-27.5971	-48.4290	Barra da Lagoa - SC
	R75	2	-27.5805	-48.4572	Lagoa da Conceição - SC

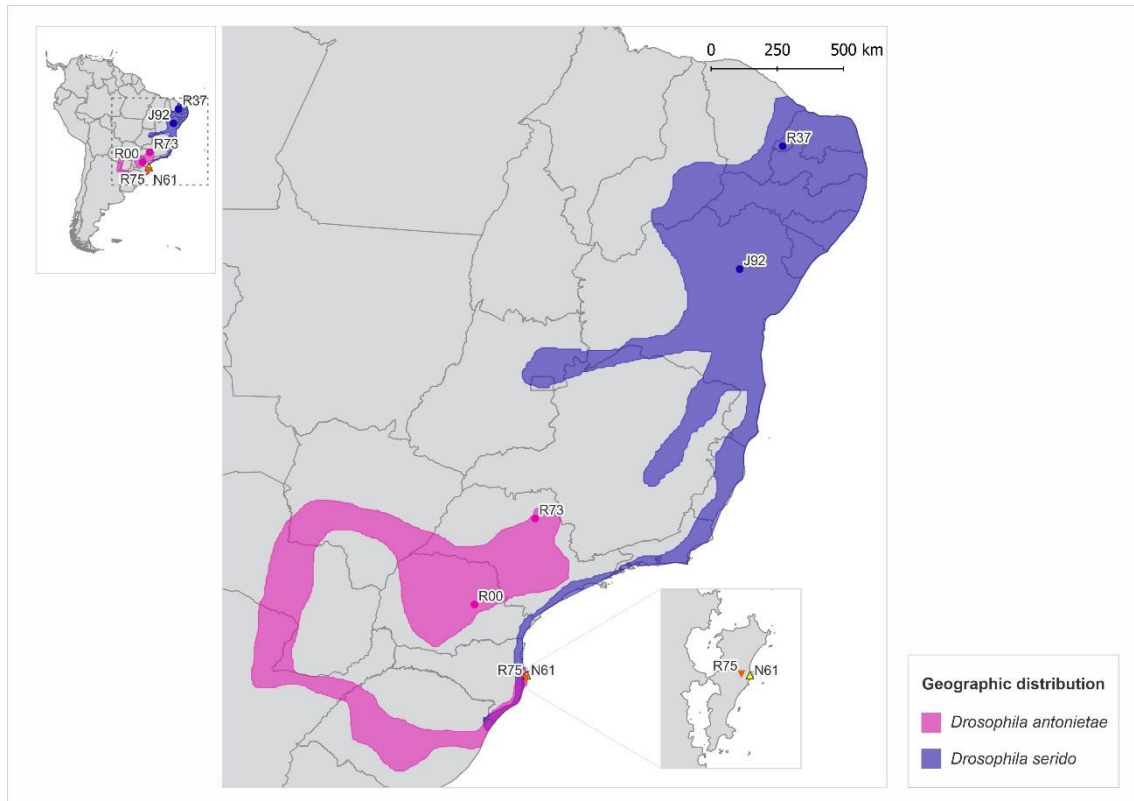


Figure 1. Geographic distribution of *D. antonietae* (pink) and *D. serido* (blue). The locality of each genome analyzed is represented by their respective id (R73, R00, R37, J92, R75 and N61).

Satellite DNAs identification

The TAREAN pipeline outputs a list of clusters identified as putative satellite DNAs, along with their genomic abundance, lengths and nucleotide sequences of repeats with various other cluster characteristics such as number of reads, satellite probability and similarity with other genetic elements. The putative satDNAs are classified in two categories: High Confidence (HC) or Low Confidence (LC) based on component (C) and pair completeness (P) indexes. The C index assesses the presence of clusters representing tandemly repeated sequences, while the P index evaluates the length of contiguous arrays of monomers. Both C and P indexes are fundamental for correct satellite DNA classification (Novák *et al.* 2017).

We performed RepeatExplorer2 pipeline for all the 10 genomes analyzed and all clusters classified as HC were selected for further studies. We ran BLASTn in public databases (NCBI, RepeatMasker and FlyBase) using consensus sequences of each satDNA identified, to search for similarity with other repetitive elements previously described.

Phylogeny and NMDS

We used the Geneious Prime 2022.0.2 software to extract complete copies of each satDNA from the contigs generated by the RepeatExplorer2 pipeline. We also used Geneious to identify satDNAs shared between *D. antonietae* and *D. serido*.

For the phylogenetic analysis of each satDNA, we aligned all the complete copies obtained from each genome using the MUSCLE method (UPGMA) (Edgar 2004). We constructed maximum likelihood (ML) trees using MEGA11 Software (Tamura 2021) with 1000 bootstraps replicates. Best-fit evolutionary models were predicted for each nucleotide alignment in MEGA11. The phylogenetic trees were edited using TreeGraph 2 and FigTree v1.4.4 (Stöver and Müller 2010; available at <http://tree.bio.ed.ac.uk/software/figtree/>).

The Non-metric multidimensional scaling (NMDS) ordinations were generated by pairwise divergence values. Divergence estimation was carried out using the Poisson correction model in MEGA11, and subsequent analyses were conducted using the R package *vegan*, representing Euclidean distances for three to five dimensions (Tamura 1992; Kumar *et al.* 2018; Dixon 2003). The NMDS and plotting of ordinations were performed in RStudio v4.2.1 (Boston, MA, USA).

Secondary structure and statistics test

We used RNAfold web server (<https://rna.tbi.univie.ac.at/>) to search for optimal secondary structure with a minimum free energy for each identified satDNA. The same parameters outlined by Kasinathan and Henikoff (2018) were employed during the analysis. Predictions were performed on ten randomly sampled sequences generated using the "Shuffle DNA" tool developed by Stothard (2000). To compare the distributions of minimum free energy values between the satDNA sequences and random sequences, we employed the two-sample Kolmogorov-Smirnov test. These statistical tests were conducted using RStudio v4.2.1 (Boston, MA, USA).

DNA extraction and PCR amplification

Total genomic DNA was extracted from a pool of 20-30 adults of *Drosophila serido* (strain J92, Milagres, Bahia – Brazil) with the Wizard® Genomic DNA Purification Kit (Promega). We used the following primers for PCR amplification of satDNAs:

SAT197/230F (GCTAGAAGCAAACAGGTAAAAGATG) and SAT197/230R (GCAACGTCCTGCCTACTCTTA), SAT109/104F (GAACATTAGTTCCTAGTAAATTC) and SAT109/104R (CCATATCATATGTTATGTTCTCTC), pBuMA2F (CGGAGTATTTTTCATTCGAC) and (GGTATGCCATAAAGAAGTCG) (Kuhn et al. 2008), CDSTR138F (CCGATTTTCCATGAACATAACC) and CDSTR138R (CGCAAATATACTCTACTACACC) (Lima et al. 2017), CDSTR198F (GGTAGATTGAACACTAATGGGC) and CDSTR198R (CCTTTCTGCCTTTTTTCTAGCCG) (Lima et al. 2017).

Satellite DNAs were amplified by PCR consisting of initial denaturation step at 95°C for 5 minutes, followed by 30 cycles at 95°C for 1 minute, 47-55°C for 1 minute (based on the melting temperature of each primer) and at 72°C for 1 minute and then final extension for 10 minutes at 72°C. The PCR products were purified using Wizard SV Gel and PCR Clean-Up System (Promega) and sequenced.

Fluorescent in situ Hybridization

Mitotic and polytene chromosomes were obtained from neuroblasts and salivary glands, respectively, of third instar larvae of *D. serido* (strain J92), *D. seriema* (strain D73), *D. gouveai* (strain H6) and *D. koepferae* (strain B26D2), according to methodology described in Baimai (1997) and Ashburner (1989), with some modifications. For metaphase chromosomes preparation, we incubated the neuroblasts in Colcemid for 12 minutes and in the 0.7% trisodium-citrate solution for 15-17 minutes.

For the double-FISH experiments, probes were labeled with antidigoxigenin-Rhodamine and avidin-FITC (Roche Applied Science) and the chromosome hybridizations were conducted as described by Dias *et al.* (2014) with minor modifications, including a 1 minute and 15 seconds treatment with 0.07M NaOH. Chromosomes were counterstained with DAPI (4,6-diamidino-2phenylindole) in antifade reagent (SlowFade, Invitrogen) and analyzed under an Axio Imager A2 epifluorescence microscope equipped with the AxioCamMRm camera (Zeiss). Images were captured with Axiovision (Zeiss) and edited in Adobe Photoshop.

Results and Discussion

Repetitive landscape of *Drosophila antonietae* and *D. serido*

We ran RepeatExplorer2 (Nóvak *et al.* 2020) pipeline on raw Illumina short reads to assess the satellite DNA content in *D. antonietae* and *D. serido*, and also to compare the general repetitive DNA landscape between the two species. We used the sequenced genomes of two strains from *D. antonietae* (R00 from Guarapuava, PR and R73 from Sertãozinho, SP) and two strains from *D. serido* (J92 from Milagres, BA and R37 from Lajes Pintada, RN).

The DNA sequencing yielded a different amount of sequence data for both species. In *D. antonietae*, the sequenced data varied, ranging from ~960 MB in the R00 strain to ~2.6 GB in the R73 strain. On the other hand, sequenced data of *D. serido* was similar in size with ~3.0 GB (R37) and 3.4 GB (J92). All sequenced data exhibited comparable quality after FastQC analysis and were subsequently trimmed to the same length (~100bp), ensuring a consistent pattern for the input files. The RepeatExplorer2 pipeline analyzed 3,908,300 from the 68,666,900 input reads from the R73 strain (*D. antonietae*) and 4,491,120 from the 23,522,604 input reads from the R00 strain (*D. antonietae*). In *D. serido*, the pipeline analyzed 3,569,673 from the 79,158,690 input reads from the R37 strain, and 2,684,156 from the 71,201,362 input reads from J92 strain.

Our analysis revealed that the repetitive DNA fraction in the genome was approximately the same, 21.5%, for both strains of *D. antonietae*. In contrast, for *D. serido*, strain R37 exhibits a repetitive fraction around 17.7%, and 31.5% for strain J92 (Figure 2). All the repetitive DNAs identified in both strains of *D. antonietae* and *D. serido* are presented in Table S1-S4. Our findings are consistent with previous results found for *D. buzzatii*, *D. seriema* and *D. mojavensis* with 13.9%, 20.9% and 13.5% of their genome composed by repetitive DNA, respectively (Lima *et al.* 2017).

The satellite DNA genomic contribution exhibited differences both among and within species. In *D. antonietae*, satDNAs accounted for ~3.9% and ~2.6% in the R73 and R00 strains, respectively. On the other hand, *D. serido* showed a lower satDNA genomic contribution, with ~2.9% and ~1.1% in the R37 and J92 strains, respectively. These findings are also consistent with previous studies that also indicated a low satDNA genomic contribution in other species from the *repleta* group, such as *D. buzzatii* (~1.9%), *D. seriema* (~2.9%) and *D. mojavensis* (2.5%) (Lima *et al.* 2017).

The genomic contribution of repetitive elements classified as TEs by RepeatExplorer2 was found to be higher than the satDNA content in *D. antonietae* and *D. serido*, similar to the results reported by Lima *et al.* (2017) on *D. buzzatii* and *D. seriema*. In *D. antonietae*, TEs accounted for ~8.2% (R73) and ~9.8% (R00) of the genomic DNA, and in *D. serido*, ~6.2% (R37) and ~2.2% (J92). Both species showed lower TE contributions to genomic DNA compared to other species from the *buzzatii* cluster: *D. buzzatii* with ~12% and *D. seriema* with ~18% (Lima *et al.* 2017).

Both *D. antonietae* and *D. serido* exhibited a similar pattern of genomic contribution from different TE orders, with long terminal repeat (LTR) retrotransposons being the most abundant in both species. In *D. antonietae*, LTRs accounted for ~6.5% and ~5.4% in the R00 and R73 strains, respectively, while in *D. serido*, LTRs accounted for ~3.4% and ~1.2% in the R37 and J92 strains, respectively. Interestingly, the genomic contribution of LTR retrotransposons in *D. antonietae* is approximately twice its contribution in *D. serido*. The second most abundant TE order in both species was LINEs, followed by Helitrons.

Previous studies revealed that the most abundant TE in *D. buzzatii* are TIR-transposons, in *D. seriema* are LTRs and in *D. mojavensis* are Helitrons (Lima *et al.* 2017). Conversely, Rius *et al.* (2016) revealed that Helitrons are the most abundant TE of both *D. buzzatii* and *D. mojavensis*. Therefore, we observed similar profiles on the abundance of TE orders between *D. antonietae*, *D. serido* and *D. seriema*, which is the closest related species from the three analyzed by Lima *et al.* (2017). In all three species, the major TE orders found were LTRs, Helitrons and non-LTR/LINEs (Lima *et al.* 2017).

Together with previous studies on repetitive DNAs in *Drosophila* (Silva *et al.* 2019), our study highlights the presence of variation in the repetitive content among different species. Furthermore, our analysis has also revealed significant differences between the two analyzed genomes of *D. serido*. Notably, the sequencing profiles of both *D. serido* analyzed genomes exhibited similar sizes and quality, and RepeatExplorer2 pipeline analyzed a comparable number of reads from both strains, suggesting that the observed variation in the repetitive DNA content between the two strains is not an artefact. It is also worth mentioning that the two sequencing data from *D. antonietae* displayed considerable variation in size, but with a similar quality. Even though, we did not find significant differences in the repetitive content among these strains. Overall, our

findings emphasize the true variation on repetitive DNA content that exists both within and between the two species.

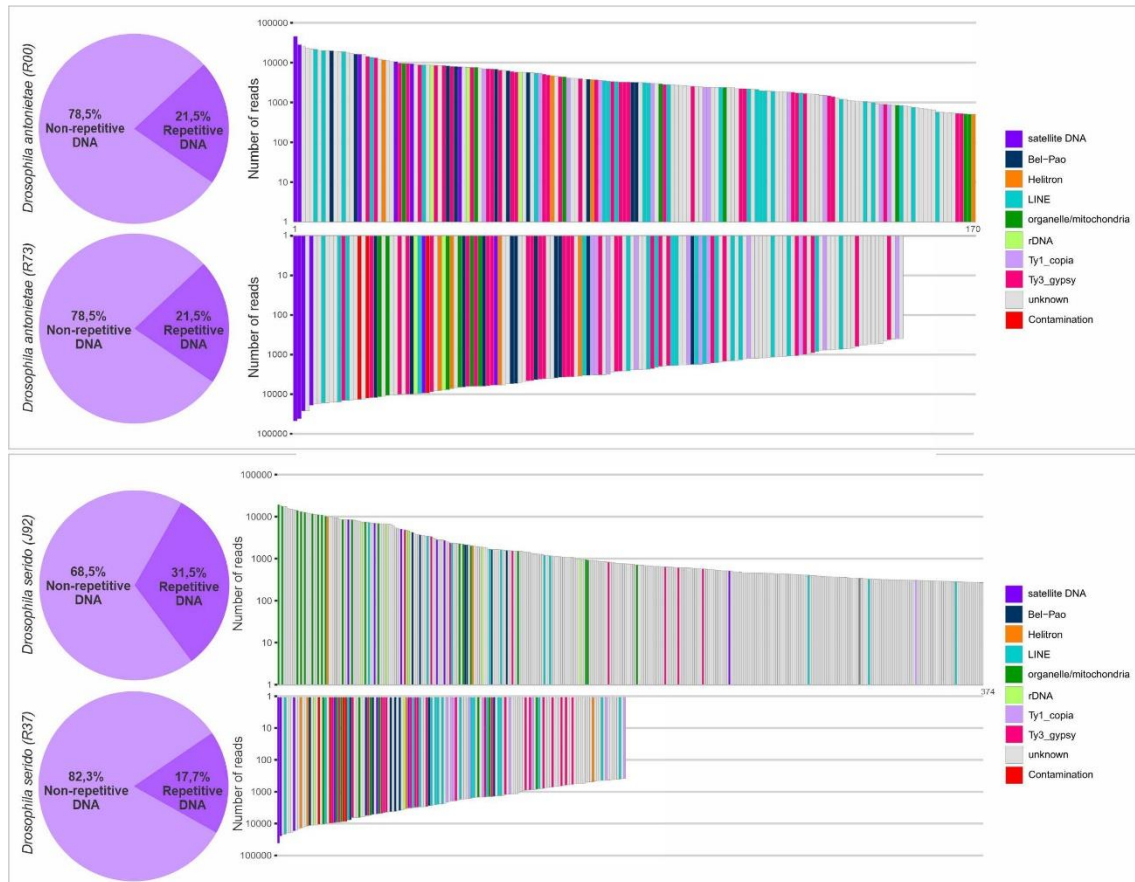


Figure 2. The repetitive DNA landscape in *D. antonietae* and *D. serido*. The strains R73 and R00, from *D. antonietae* are shown in (a), and strains R37 and J92 from *D. serido* are shown in (b).

The satellitome of *D. antonietae* and *D. serido*

The RepeatExplorer2 pipeline includes an integrated version of TAREAN, which enables the identification of potential high confidence satellite DNAs. The resulting consensus sequences of these putative satDNAs were subjected to manual inspection to verify the accurate identification of the satellite monomer sequences. We selected for further analysis satDNAs with more than 1,000 copies in at least one of the analyzed *Drosophila* strain (Table 2). We identified a total of five satDNAs in both *D. antonietae* and *D. serido*. The interspecific divergence calculated for each satellite DNA ranged from 0.6% to 14.4%.

Our analysis revealed that *D. antonietae* and *D. serido* share the same set of five satellite DNAs (Figure 3). Among these satDNAs, three have been previously described in species from the *buzzatii* cluster. The pBuM and DBC-150 have already been identified

in the two species, but CDSTR138, first identified in *D. seriema*, is reported here in the two species for the first time (Kuhn and Sene 2005; Kuhn *et al.* 2007; Lima *et al.* 2017). Additionally, we identified two novel satellites, named SAT197/230 and SAT109/104 based on their consensus length in each species. Our analysis also revealed for the first time in *D. antonietae* and *D. serido* the presence of the minisatellite CDSTR198 (Lima *et al.* 2017; Kuhn *et al.* 2021). Based solely on the abundance of these satDNAs in each species, it is not possible to derive a species-specific profile (Figure 3).

Table 2. General features of the satellite DNAs identified in the strains of *D. antonietae* and *D. serido*.
* = minisatellite DNA

<i>D. antonietae</i> (R73)	Cluster	satellite DNA	Genome Proportion (%)	Consensus Length (pb)	Estimated Copy Number
	1	pBuM	1.234%	370	6,003
	2	SAT197/230	1.100%	197	10,051
	3	DBC-150	0.692%	112	11,121
	5	SAT109/104	0.496%	109	8,191
	33	CDSTR138	0.244%	138	3,183
	51	CDSTR198*	0.154%	198	1,400
<i>D. antonietae</i> (R00)	Cluster	satellite DNA	Genome Proportion (%)	Consensus Length (pb)	Estimated Copy Number
	1	pBuM	1.024%	356	5178
	2	SAT197/230	0.631%	197	5765
	17	DBC-150	0.363%	112	5834
	30	SAT109/104	0.209%	103	3652
	26	CDSTR138	0.236%	138	3078
	42	CDSTR198*	0.177%	198	1609
<i>D. serido</i> (R37)	Cluster	satellite DNA	Genome Proportion (%)	Consensus Length (pb)	Estimated Copy Number
	1	pBuM	1.192%	190	11293
	2	SAT197/230	0.708%	230	5541
	26	DBC-150	0.270%	112	4339
	45	SAT109/104	0.143%	104	2475
	9	CDSTR138	0.489%	138	6378
	59	CDSTR198*	0.093%	199	841
<i>D. serido</i> (J92)	Cluster	satellite DNA	Genome Proportion (%)	Consensus Length (pb)	Estimated Copy Number
	52	pBuM	0.274%	190	2596
	66	SAT197/230	0.201%	230	1573
	38	DBC-150	0.337%	112	5416
	89	SAT109/104	0.108%	104	1869
	85	CDSTR138	0.112%	138	1461
	240	CDSTR198*	0.020%	188	191

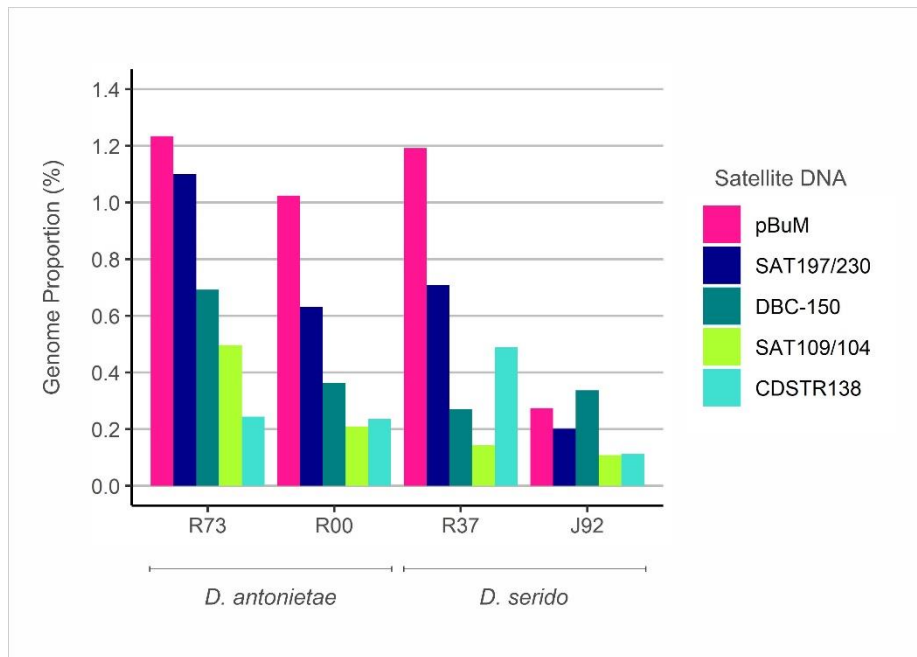


Figure 3. Genome proportion of the five satellite DNAs identified in each strain from *D. antonietae* (R73 and R00) and *D. serido* (R37 and J92).

The pBuM is the most abundant satDNA family in both species, with genomic proportion of 1.0-1.2% in *D. antonietae* and 0.3-1.2% in *D. serido*. The TAREAN pipeline generated a pBuM consensus sequence of 370 bp for *D. antonietae* (both strains), corresponding to *alpha/beta* repeats, and a consensus of 190 bp for *D. serido* (both strains), corresponding to the *alpha* repeats. Careful inspections of the pBuM clusters generated by RepeatExplorer2 revealed that in *D. antonietae* complete copies of *alpha* and *alpha/beta* repeats are present at equivalent frequencies, while in *D. serido*, *alpha* repeats are much more abundant than *alpha/beta* repeats. It was not possible to extract complete *alpha/beta* repeats from *D. serido*, therefore we did not use these copies in subsequent analysis. These findings are in accordance with the observations made by Kuhn *et al.* (2008) based on Southern blotting and FISH experiments, who reported that in *D. antonietae* both *alpha* and *alpha/beta* repeats are present in equivalent abundances, while in *D. serido*, *alpha* repeats are much more abundant than *alpha/beta* repeats.

The DBC-150 satDNA with repeats of 150 bp was previously identified in both species (Kuhn *et al.* 2007). In our analysis, TAREAN generated a consensus length of 112 bp, which differs from the previously reported repeat length. The fact that the pipeline generated a consensus sequence with different length possibly indicates that this DBC-150 variant is the most prevalent in *D. antonietae* and *D. serido*. Through manual analysis of contigs from the corresponding DBC-150 cluster in TAREAN, we further discovered the presence of three main DBC-150 variants, which we named DBC-42 (42 bp), DBC-

112 (112 bp), and DBC-150 (154 bp). It is important to highlight that the DBC-150 variants identified in our study differ from those reported by Kuhn *et al.* (2007). The DBC-42 had an AT content of ~50%, the DBC-112 had ~44.6% and the DBC-150 variant had ~46% AT content. Despite nucleotide differences, the DBC-150 variants described here showed similar features with the DBC-150 previous described, which exhibited ~44% AT content (Kuhn *et al.* 2008, 2021).

The CDSTR138 satDNA was initially described in *D. seriema*, but its presence has not been investigated in other species within the *buzzatii* cluster, except for *D. buzzatii*, where it was found to be absent (Lima *et al.* 2017). Our analysis revealed, for the first time, the presence of CDSTR138 in both *D. antonietae* and *D. serido*. It exhibited similar features to what was previously observed in *D. seriema*, with repeat units 138 bp long and a high AT content of 68.8% (Lima *et al.* 2017). This finding suggests that CDSTR138 underwent amplification after the split of *D. buzzatii* from the remaining species of the *buzzatii* cluster.

The novel satDNA, SAT197/230, with a 70.5% of AT content, exhibited distinct consensus length between *D. antonietae* (197 bp) and *D. serido* (230 bp). However, manual inspection of the correspondent clusters generated by TAREAN confirmed the presence of both variants in both species. Additionally, we identified the SAT109/104 satDNA, which also exhibited distinct consensus repeat lgths between the two species. In *D. antonietae* the main variant was 109 bp long, while in *D. serido* it was 104 bp long. This novel satellite was characterized by a high AT nucleotide content of approximately 78%.

We also detected the presence of the minissatellite CDSTR198, which was previously identified in *D. buzzatii* (Guillén *et al.* 2014) and here reported for the first time in both *D. antonietae* and *D. serido* genomes. However, the contribution of CDSTR198 to the total genomic DNA was low, accounting for 0.15-0.17% in *D. antonietae* and 0.02-0.09% in *D. serido*. The TAREAN analysis generated consensus sequences which varied in length from 188-199bp and with a 64.6% AT content, consistent with previous studies (Guillén *et al.* 2014; Lima *et al.* 2017). The genomic contribution of CDSTR198 in *D. antonietae* was comparable to that found in *D. buzzatii* (0.2%), while in *D. serido*, it was similar to the contribution in *D. seriema* (0.02%) (Lima *et al.* 2017). Consequently, the CDSTR198 exhibited low copy number estimations in both strains of *D. serido*: 841 copies in R37 and 191 in J92 (Table 2). In *D. antonietae*,

on the other hand, the CDSTR198 was found to be more abundant, with estimates of 1,400 copies in the R73 strain and ~1,600, in R00. This finding may suggest a recent burst of this minisatellite in *D. antonieate*.

The SSS139 satDNA, consisting of 139 bp repeats, was previously described in species from the *buzzatii* cluster, except in *D. buzzatii* (Franco *et al.* 2008). To identify clusters with homology to SSS139, we conducted blastn searches in custom databases containing all repetitive DNA clusters identified by the RepeatExplorer2 pipeline. While it was possible to find clusters that showed homology with SSS139, further manual inspection of the contigs in each cluster revealed that SSS139 is not composed of tandem repeats in both *D. antonietae* and *D. serido*. These findings are consistent with previous studies, which also failed to confirm the satDNA status of SSS139 in *D. seriema* (Lima *et al.* 2017). Further investigations are needed to better understand the organization of SSS139 in these *Drosophila* species.

The transposable element *PERI* was previously identified in *D. serido* by Kuhn and Helop-Harrison (2011). *PERI* is a TE that shares features with an abundant and widespread TE in *Drosophila*, the DINE-1 (Locke *et al.* 1999). Interestingly, the DINE-1s have central tandem repeats (CTRs) that have underwent independent amplification leading to satDNA arrays in some *Drosophila* species (Dias *et al.* 2015, Silva *et al.* (2023). The results from genomic DNA digestions and FISH on DNA fibers also suggested that the CTRs of the *PERI* in *D. serido* underwent amplification to satDNA-like arrays (Kuhn and Helop-Harrison, 2011).

Although we identified clusters in the genomes of both *D. antonietae* and *D. serido* that showed sequence homology with *PERI* and its CTRs of 383 bp, we did not observe them organized in long tandem arrays. It is also important to mention that the TAREAN pipeline did not classify these clusters as HC. Therefore, our results revealed no expansion of the internal tandem repeat of *PERI* in *D. serido* as suggested by Kuhn and Heslop-Harrison (2011).

The satellite DNAs identified in *D. antonietae* and *D. serido* revealed a preferential monomer size ranging from 100-370 bp that are enriched with AT-rich nucleotides, except for DBC-150, which displayed an average AT content of 44%. These features seem to be common to satDNAs in species from the *repleta* group (reviewed by Kuhn *et al.* 2021). Moreover, the total contribution of satDNAs to the genomic DNA is

low in *D. antonietae* (2.6-3.9%) and *D. serido* (1.0- 2.9%). These findings are consistent with previous studies in species from the *repleta* group and supports the hypothesis that low satellite DNA contribution may be a consequence of selective constraints (Lima *et al.* 2017; Kuhn *et al.* 2021). In contrast, species like *D. melanogaster* and *D. virilis*, for example, exhibit much higher satDNA genomic contribution, exceeding 20% and ~50%, respectively (Wei *et al.* 2014).

Chromosome localization of satellite DNAs

Species belonging to the *buzzatii* cluster exhibit karyotypes consisting of one pair of sex chromosomes, four pairs of telocentric autosomes of equal size, and one pair of microchromosomes. The karyotype of some species can be distinguished based on the amount of heterochromatin content found in the microchromosomes and in the sex chromosomes (Baimai *et al.* 1983; Kuhn *et al.* 1996; Kuhn *et al.* 2007).

Previous studies showed that in *D. buzzatii*, the pBuM satDNA was found to be predominantly located in the centromere of all chromosomes, except the X chromosome. In *D. seriema*, the pBuM was tentatively reported as the main centromeric component of all autosomes, while CDSTR138 was also detected in the centromeric region of chromosomes 2, 3, 4 and 5 (Lima *et al.* 2017). However, no double-FISH experiments combining probes for both satellites were conducted in this species. In *D. antonietae*, *D. serido* and *D. gouveai* the pBuM was found to be distributed across multiple chromosomes, with a particular concentration in the centromeric regions (Kuhn *et al.* 2008). Additionally, the DBC-150 was found to be restricted to the microchromosomes of *D. gouveai*, *D. seriema*, *D. serido*, *D. koepferae* and *D. antonietae* (Kuhn *et al.* 2007).

Surprisingly, our double-FISH experiments using probes for pBuM and CDSTR138 in mitotic chromosomes of *D. serido* revealed that CDSTR138 is actually the major component of the centromeres of all chromosomes except the Y and the microchromosome (Figure 4), while the pBuM is more distally located towards the heterochromatic short arms of chromosomes. Based on this new finding, we also conducted the same double-FISH experiments on the chromosomes of *D. seriema*, *D. koepferae* and *D. gouveai* (Figure 5). In *D. seriema*, the CDSTR138 was also identified as the main centromeric component of the four autosomes, while the pBuM was more distally located to the short heterochromatic arms and it was also found in the microchromosomes. Similarly, in *D. gouveai*, CDSTR138 was observed in the

centromeric region of the four pair of autosomes and X chromosome, and in the microchromosomes. The pBuM was found in the microchromosomes and exhibited a polymorphic pattern in the remaining autosomes, being present towards the short arm of two pairs of autosomes in certain metaphases (Figure 5), while in others, it was present in only one pair. The CDSTR138 was found in the centromeric region of all autosomes in *D. koepferae* (Figure 5).

Altogether, these results demonstrate that CDSTR138 is actually the major satDNA component of the centromeres in *D. antonietae*, *D. serido*, *D. seriema*, *D. gouveai* and *D. koepferae*. Our findings suggest that, after the split of *D. buzzatii* from the other species from the *buzzatii* cluster, the pBuM shifted its location to the heterochromatic short arms of chromosomes, while CDSTR138 became the centromeric main satDNA.

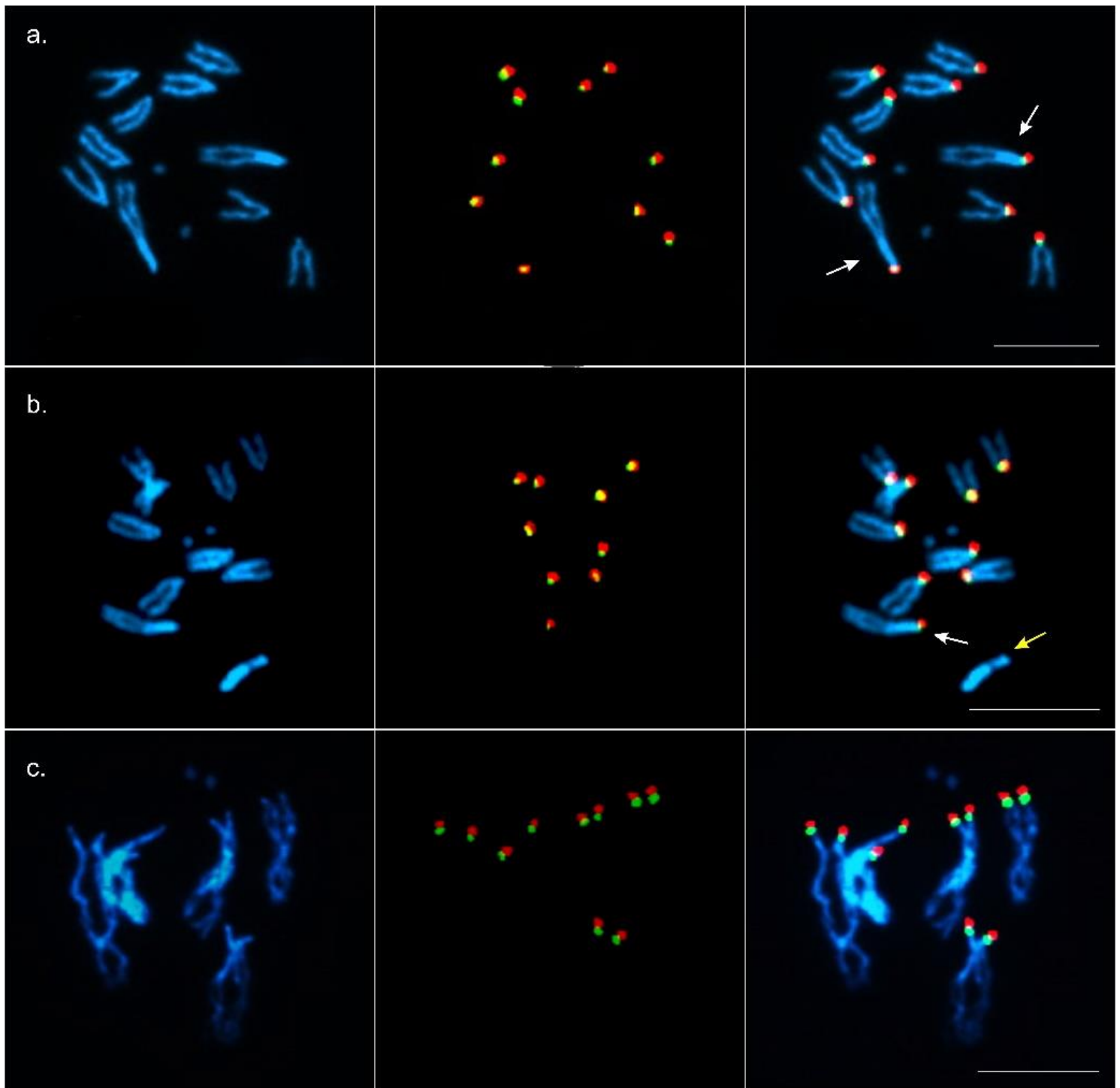


Figure 4. Fluorescent *in situ* hybridization of pBuM (red) and CDSTR138 (green) in chromosomes of *D. serido*. **a.** Female mitotic chromosomes. **b.** Male mitotic chromosomes, no signal was observed in Y chromosome. **c.** Female prometaphasic chromosomes showing almost complete separation between CDSTR138 and pBuM. White arrows point to X chromosome and yellow arrow points to Y. Scale bar represents 5 μ M

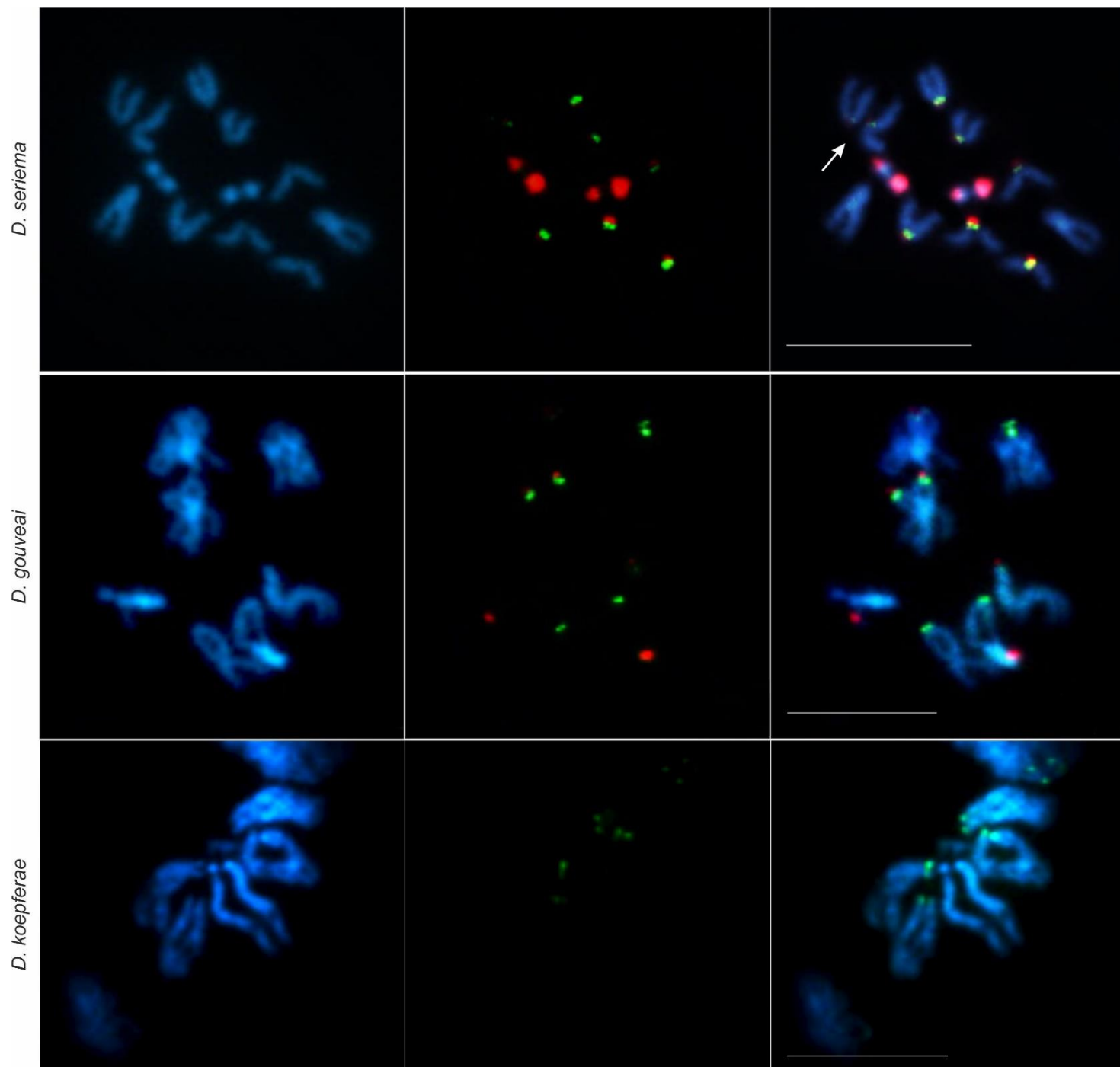


Figure 5. Fluorescent *in situ* hybridization of pBuM (red) and CDSTR138 (green) in chromosomes of *D. seriema*, *D. gouveai* and *D. koepferae*. Arrow points to the pair that showed a weaker signal of both probes in *D. seriema*. Scale bar represents 5 μ M.

We conducted FISH experiments with probes for the novel satDNAs SAT197/230 and SAT109/104, in the chromosomes of *D. serido*. We found that SAT197/230 localization is restricted to the Y chromosome, where it is organized on three major blocks, one situated distally on the short arm, one in the centromere and one in the terminal region of the long arm (Figure 6 and Figure S1). The SAT109/104 is located on the terminal region of the Y chromosome and on the centromeric region of the X chromosome. Based on our findings, it appears that the centromeric region of the X chromosome harbors three distinct satellite DNAs: CDSTR138, pBuM and SAT109/104. These satDNAs are non-homologous. Further investigation with double-FISH experiments using combinations of these three satDNA probes may elucidate the spatial arrangement of these satellites on the X chromosome.

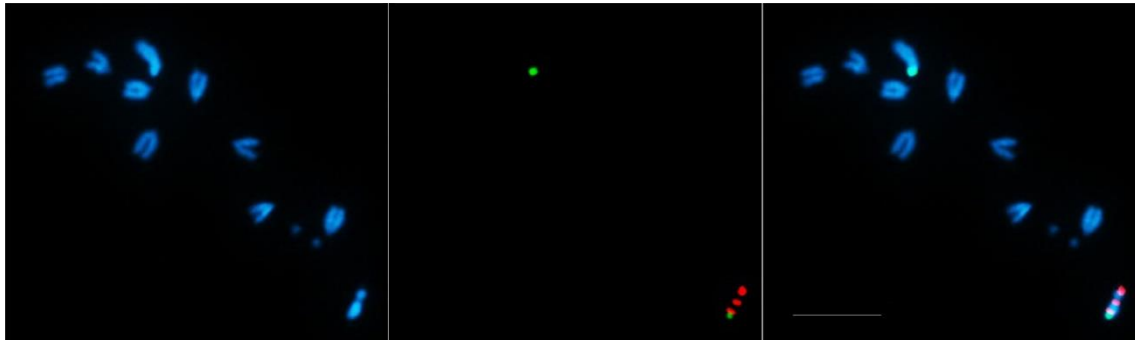


Figure 6. Fluorescent *in situ* hybridization of SAT197/230 (red) and SAT109/104 (green) in mitotic chromosomes of *D. serido*. Scale bar represents 5 μ M.

Our findings showed that CDSTR198 is a minissatellite in *D. serido*. This family has also been described as a minisatellite in other species from the *buzzatii* cluster (Kuhn *et al.* 2021). This conclusion is supported by its low genome proportion and consequently low copy number (191 copies in the J92 strain). Additionally, CDSTR198 was exclusively found scattered distributed throughout the euchromatic arms of polytene chromosomes, what is another feature of minisatellite DNAs (Figure 7).

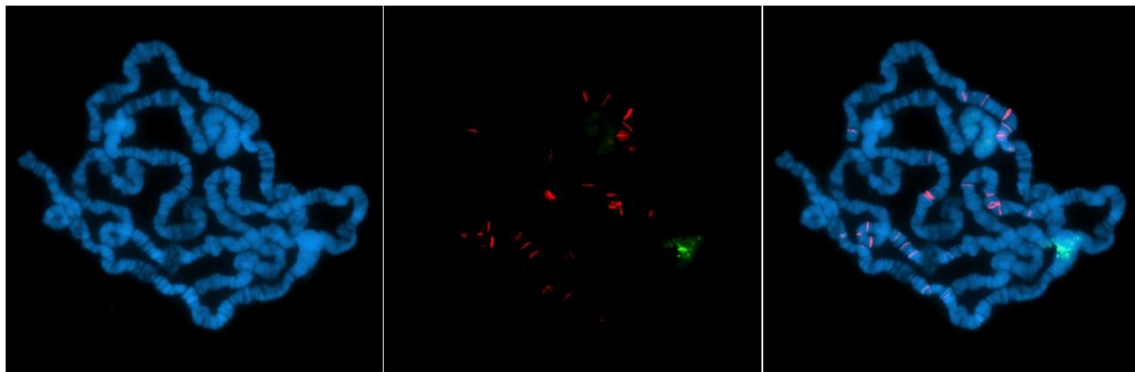


Figure 7. Fluorescent *in situ* hybridization of CDSTR138 (green) and CDSTR198 (red) in polytene chromosomes of *D. serido*. Scale bar represents 10 μ M.

An extensive study of tandem repeats involving more than one hundred animal and plant species suggested that the most abundant satDNA in a genome most likely correspond to the centromeric satDNA (Melters *et al.* 2013). However, our findings showed that the most abundant satDNA in the *buzzatii* cluster, the pBuM, is not the centromeric one (except in *D. buzzatii*), as the less abundant CDSTR138 was found to be the major centromeric DNA component in *D. serido* and *D. seriema*.

Secondary Structure of satDNAs

Satellite DNAs undergo rapid evolutionary changes, but it is noteworthy that the presence of short palindromic sequences (typically less than 10 bp) with dyad symmetries within the satDNA repeats are commonly observed in the centromeric satDNAs of eukaryotes. However, this characteristic has not been observed in the centromeric satDNAs of great apes and mice (Talbert and Henikoff 2020).

Within centromeric regions, palindromic sequences undergo conformational changes, adopting non-B-form DNA structures, such as stem-loops. These DNA structures are believed to play a crucial role in recruiting specific proteins, which serve as identifiers and regulators of the centromere (Kasinathan and Henikoff 2018; Talbert and Henikoff 2020).

Prytkova *et al.* (2011) also showed that a 10-bp periodicity of AA nucleotides strongly favors nucleosome formation, thereby stabilizing the centromere-specific histone H3 (cenH3). It is worth noting that a single turn of DNA corresponds to approximately 10 bp, and sequences with 10 bp periodicity in WW dinucleotides (where W represents A or T) facilitate nucleosomes wrapping by reducing the bending energy involved in the wrapping process (Struhl and Segal 2013).

Several centromeric satellite DNAs exhibit a 10 bp periodicity of AA/AT/TT nucleotides. For instance, in *D. melanogaster*, there is an expansion of short 5-12 bp tandem repeats containing AA dinucleotides (Talbert *et al.* 2018). In *Gallus* a 21-bp repeat with AA and TT dinucleotides separated by 10 bp has been observed (Shang *et al.* 2010). Additionally, an AA dinucleotide periodicity is also present in the centromeric satDNA of maize, *Zea mays*, which consists of 155bp (Gent *et al.* 2011).

Previous studies conducted on the *repleta* group have demonstrated that the centromeric satDNAs, namely DBC-150, CDSTR138 and pBuM, are predicted to form

stable secondary structures (more stable than what would be expected by chance) (Kuhn *et al.* 2021). In addition, the DBC-150 was found to exhibit a higher-order-repeat (HOR) organization, particularly in the form of 3mers. Further investigation into this longer repeated HOR structure revealed that it formed more stable secondary structures compared to the individual repeats (Kuhn *et al.* 2021).

Our analysis of nuclei acid folding prediction revealed that the centromeric satellite DNAs found in *D. serido* and *D. antonietae* were predicted to form stable secondary structures (Figure 8). However, exceptions were observed when the analysis involved individual repeats from DBC-150 and SAT197/230 in *D. serido* and CDSTR138 in *D. antonietae*.

The analysis of the individual repeats of DBC-150 of *D. serido* revealed that it was not predicted to form secondary structures that were significantly more stable than what would be expected by chance. However, based on the previous findings reported by Kuhn *et al.* (2021), we conducted further analysis into the HOR organization of DBC-150. It is also important to note that the DBC-150 variant analyzed here differs from the one previously described by Kuhn *et al.* (2007). Despite these differences, our results are consistent with the previous study, revealing that the DBC-150 showed a more stable secondary structure at the level of HOR organization, as 3mers.

Analysis of PCR products using oligos to amplify SAT197/230 in *D. serido* on agarose gel after electrophoresis revealed a more prominent DNA product at the level of 3mers in contrast to the monomers (data not shown), suggesting a possible HOR organization for SAT197/230 in this species. Accordingly, we investigated whether SAT197/230 would be predicted to form stable secondary structures at the level of HOR, since its individual repeats were not more stable than the expected by chance. We found that the SAT197/230-HOR structure consisting of 3mers displayed a more stable secondary structure than individual repeats. In contrast, in *D. antonietae*, individual repeats of SAT197/230 were found to form stable secondary structures.

The CDSTR138 from *D. antonietae* was predicted to form stable secondary structures as individual repeats, although not significantly more stable than expected by chance. The clusters generated by RepeatExplorer2 pipeline were categorized based on the repeat unit length. Careful inspections revealed clusters consisting of repeats approximately 270-280 bp in length, suggesting that CDSTR138 may be organized at the

level of HOR, as dimers in *D. antonietae*. Based on the manual analysis of the cluster, our results revealed that dimers were predicted to form more stable secondary structures than individual repeats of CDSTR138.

In summary, even though all centromeric satDNAs sequences found in both *D. antonietae* and *D. serido* are different concerning repeat length and nucleotide composition, they are enriched for short dyad symmetries, were predicted to form stable secondary structures and they are predominantly AT-rich. It is worth mentioning that these features appear to be essential for centromeric functions and are conserved between species (Kasinathan and Henikoff 2018; Talbert and Henikoff 2020).

We also looked for WW dinucleotides in the centromeric satDNAs CDSTR138, SAT197/230 and SAT109/104 of *D. serido*. Interestingly, all three satellites displayed a periodicity of AA/AT/TT dinucleotides with an approximately spacing of 10 bp (Figure S2). This feature has been described in several centromeric satDNAs and seem to be important for centromeric function (Prytkova *et al* 2011; Struhl and Segal 2013; Talbert and Henikoff 2020).

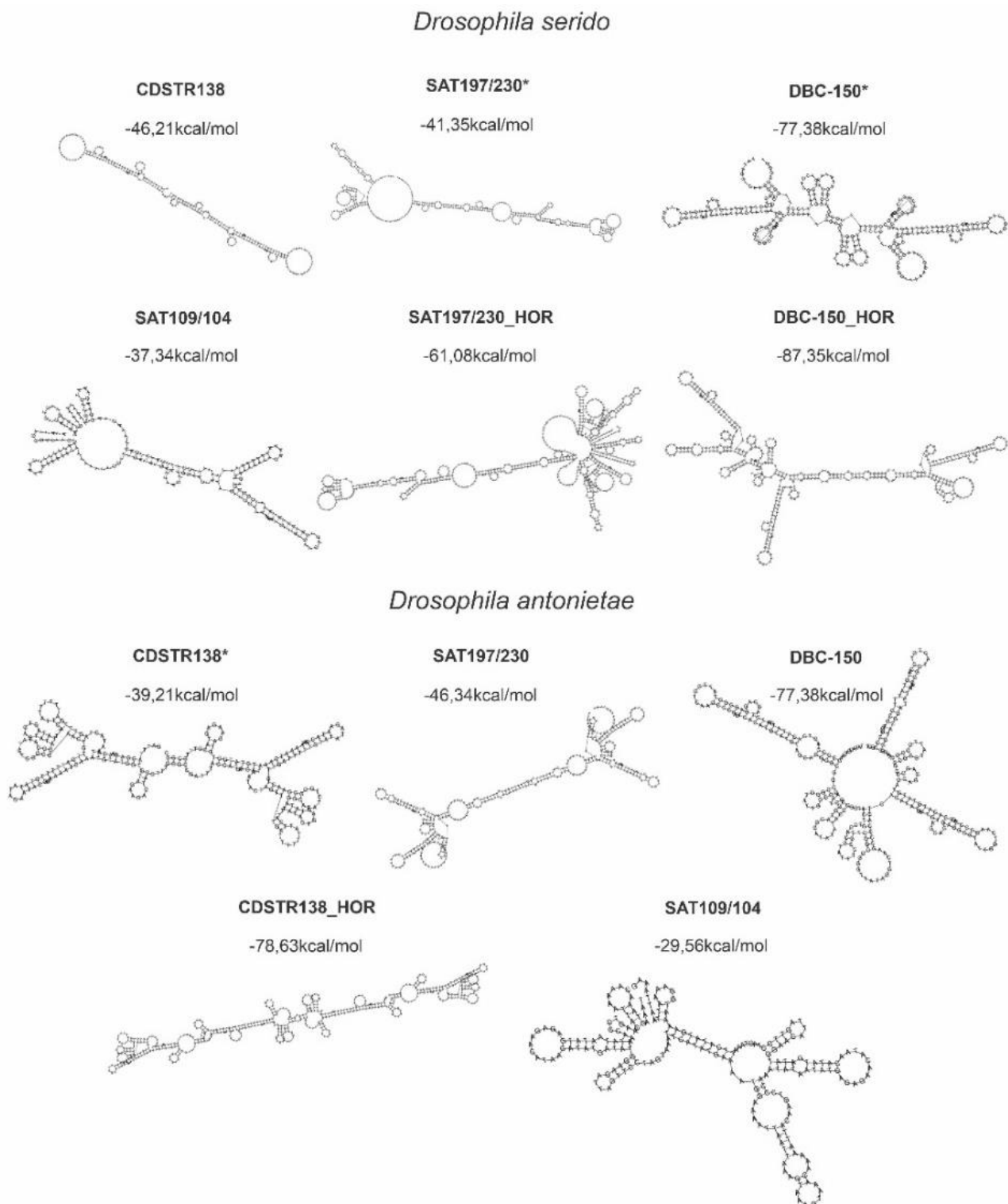
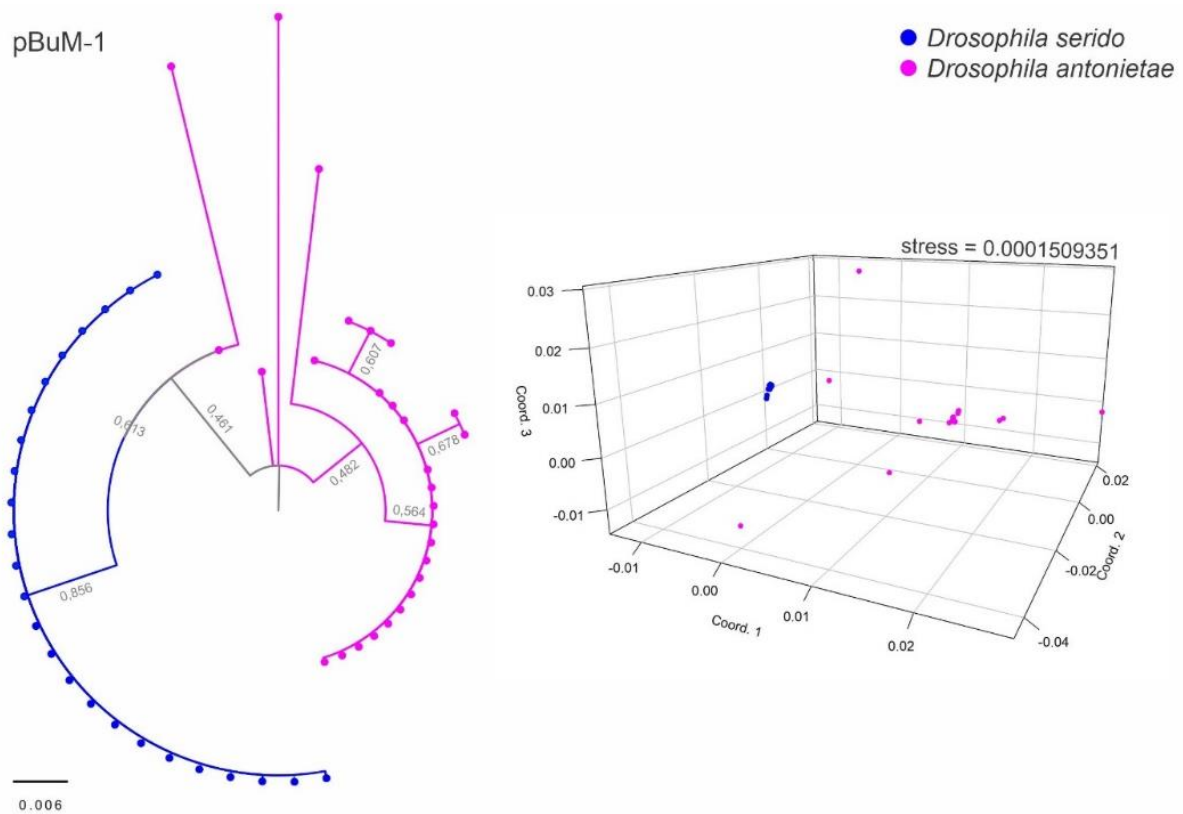


Figure 8. Predicted secondary structures of centromeric satDNAs found in *D. antonietae* and *D. serido*. Minimum free energy (MFE) is shown below each predicted structure. The asterisks indicate that the secondary structure predicted is not more stable than expected by chance.

SatDNA evolution

Satellite DNAs are among the most fast evolving components of the eukaryotic genomes. In contrast, repeats from the same satDNA family often show high intraspecific homogeneity. This pattern of evolution is known as concerted evolution (Zimmer *et al.* 1980; Dover 1982). In order to analyze the mode of evolution of the satDNAs of *D. antonietae* and *D. serido*, we constructed maximum likelihood trees using copies from each satDNA and NMDS.

Previous studies on satellite DNAs from the *repleta* group have shown that the pBuM satDNA family demonstrates a clear pattern of concerted evolution, both for the pBuM-1 and pBuM-2 variants (Kuhn *et al.* 2007; Lima *et al.* 2017). Consistent with these findings, our results from ML trees and NMDS confirmed the concerted evolution pattern in the copies of pBuM from *D. antonietae* and *D. serido* (Figure 9).



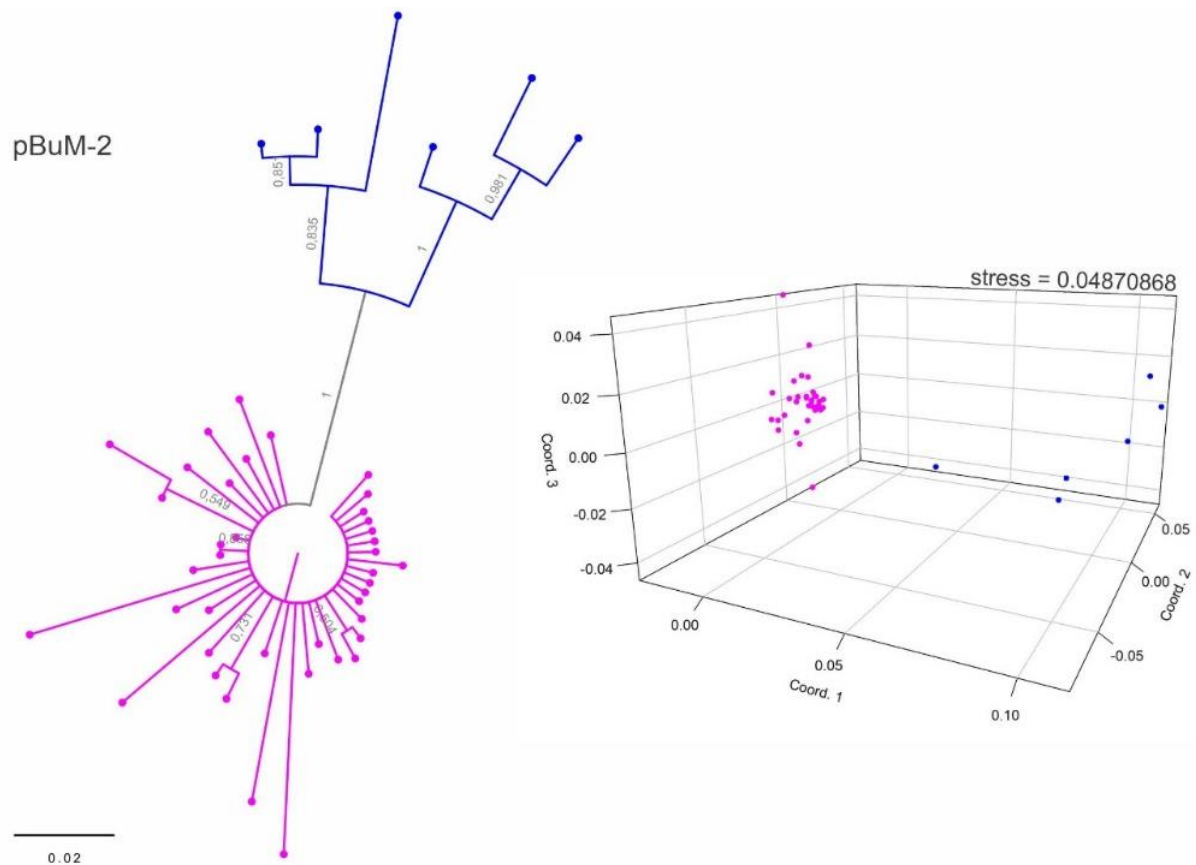


Figure 9. ML trees on the left containing copies of pBuM-1 (up) and pBuM-2 (down) from *D. antonietae* (pink) and *D. serido* (blue). The pBuM-1 tree was inferred by the Jukes-Cantor model with 1000 bootstraps. The pBuM-2 tree was inferred by the Tamura-3-parameter model using a discrete Gamma distribution (+G) with 5 rate categories and 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence among each satDNA is shown on the right. The ordinations represent Euclidian distances for four dimensions.

The evolution of DBC-150 satDNA family was found to be more complex, where copies from each species did not show higher identity when compared between species. Consequently, the repeats did not form species-specific branches on phylogenetic trees (Kuhn *et al.* 2007). Contrary to what was expected, the DBC-150 repeats analyzed here showed a pattern of concerted evolution, with copies from *D. antonietae* and *D. serido* clustering on separate branches and in the NMDS analysis we can see separation between the copies from both species (Figure 10). However, it is important to highlight that the DBC-150 repeats analyzed here belong to different variants from those reported by Kuhn *et al.* (2007).

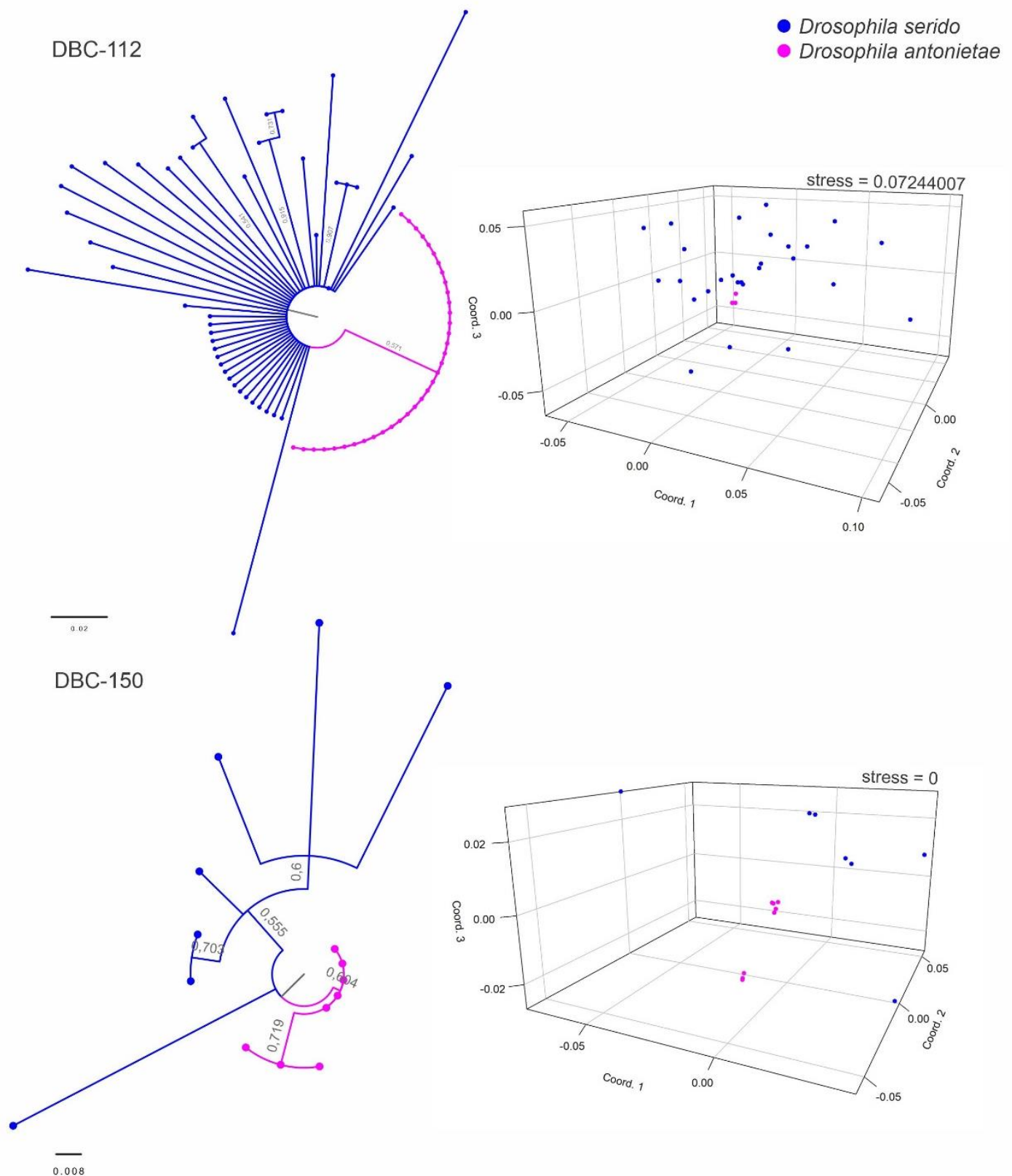


Figure 10. ML trees on the left containing copies of DBC-112 (up) and DBC-150 (down) from *D. antonietae* (pink) and *D. serido* (blue). The DBC-112 tree was inferred by the Kimura-2-parameter model with 1000 bootstraps. The DBC-150 tree was inferred by the Jukes-Cantor model with 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence among each satDNA is shown on the right. The ordinations represent Euclidian distances for tree dimensions for DBC-112 and seven for DBC-150.

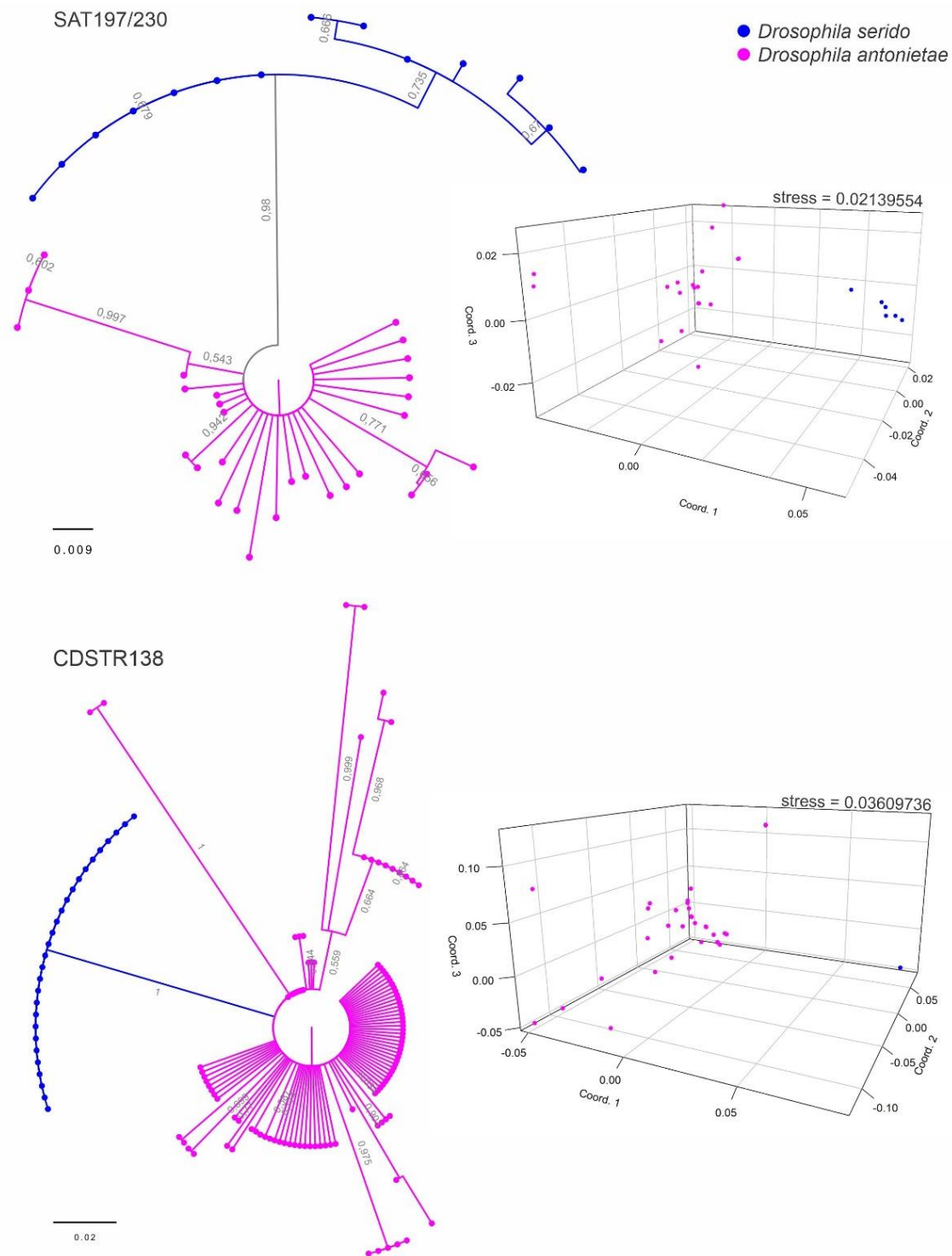


Figure 11. ML trees on the left containing copies of SAT197/230 (up) and CDSTR138 (down) from *D. antonietae* (pink) and *D. serido* (blue). The ML trees were inferred by the Tamura-3-parameter model using a discrete Gamma distribution (+G) with 5 rate categories and 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence among each satDNA is shown on the right. The ordinations represent Euclidian distances for four dimensions.

Finally, both CDSTR138 and SAT197/230 satDNAs showed a concerted mode of evolution, with copies from each satDNA clustering in species-specific branches (Figure

11). These results can also be observed in the NMDS analysis, with copies from *D. antonietae* segregating from copies of *D. serido*.

Based on the chromosome location results in *D. serido*, SAT197/230 is distributed in three major blocks in the Y chromosome, what in theory may allow concerted evolution at the level of each block. Accordingly, both ML tree and NMDS analysis with sequences of SAT197/230 from *D. serido* revealed that the copies were grouped into three major clusters (Figure 12), which may correspond to the three major blocks on chromosome Y where these copies are located. These findings are consistent with previous studies, which demonstrated that homogenization and concerted evolution may occur at the array level (Kuhn *et al.* 2012).

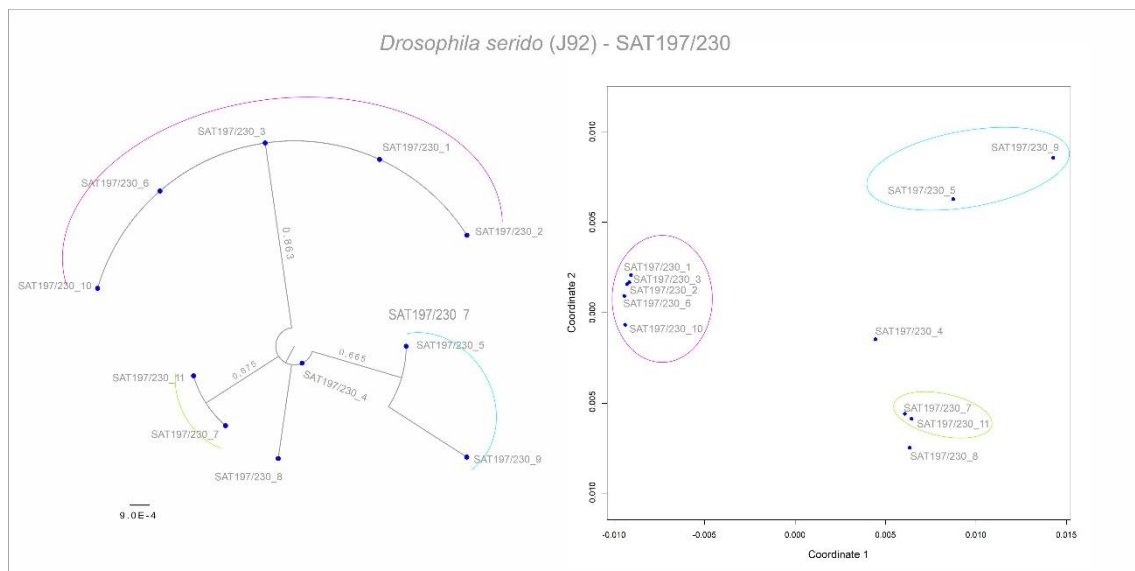


Figure 12. ML trees on the left containing copies of SAT197/230 from *D. serido* (J92). The ML trees were inferred by the Tamura-3-parameter model and 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence among each satDNA is shown on the right. The ordinations represent Euclidian distances for four dimensions.

On the other hand, the SAT109/104 and the minissatellite CDSTR198 did not display the pattern of concerted evolution, since copies of families did not group into species-specific branches on ML trees, nor they were separated by the NMDS analysis (Figure 13). The lack of a concerted evolution pattern in SAT109/104 is unexpected, considering its localization on a single chromosome, which, in theory, could facilitate homogenization and concerted evolution (Strachan *et al.* 1985). On the other hand, the CDSTR198 results aligns with expectations, as it is a minissatellite located in the euchromatin with smaller arrays. The presence of interruptions by nonhomologous sequences may limit recombination between arrays, slowing down the homogenization process (Brutlag 1980; Dover 1982).

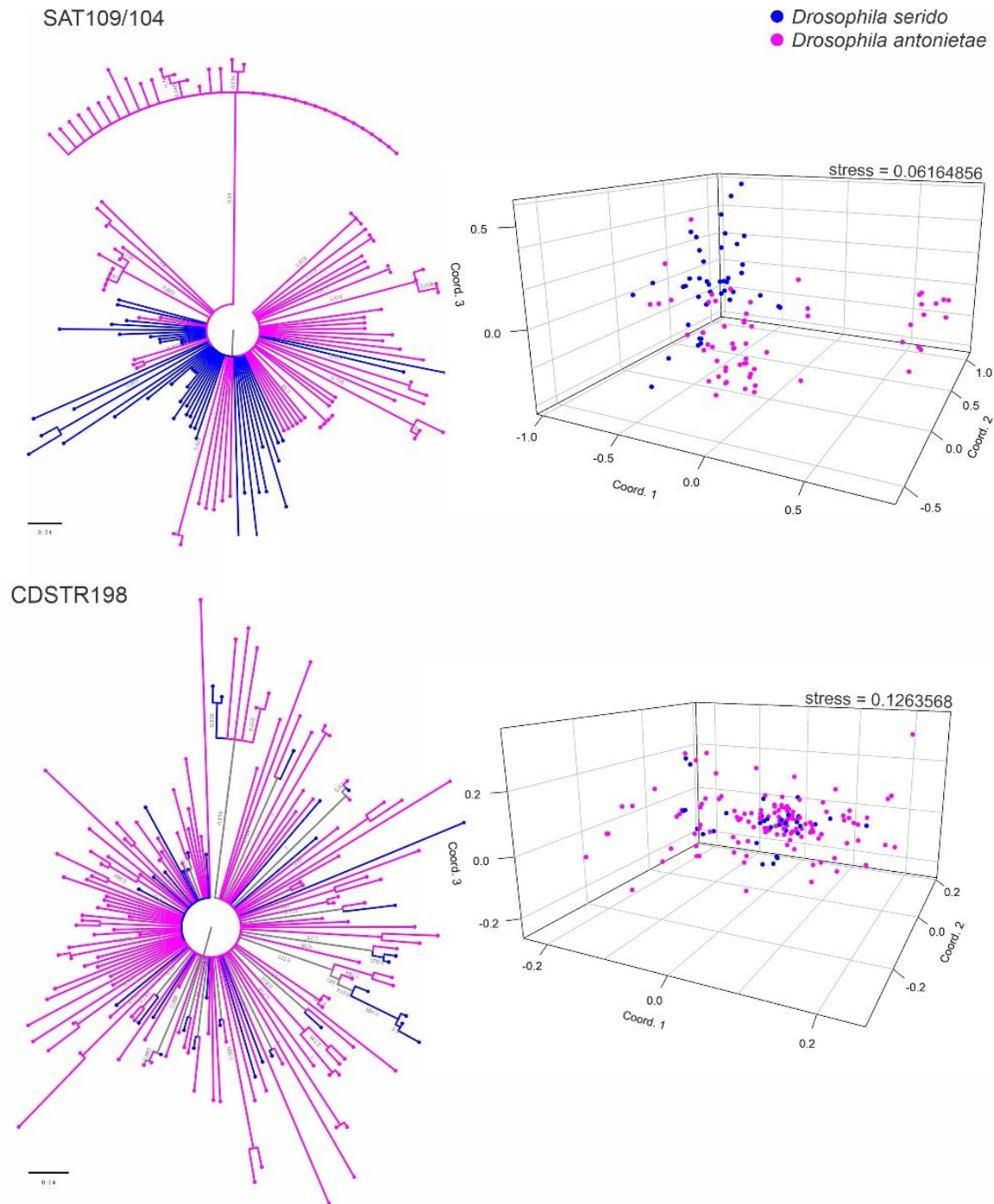


Figure 13. ML trees on the left containing copies of SAT109/104 (up) and CDSTR198 (down) from *D. antonietae* (pink) and *D. serido* (blue). The ML trees were inferred by the Tamura-Nei model using a discrete Gamma distribution (+G) with 5 rate categories and 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence among each satDNA is shown on the right. The ordinations represent Euclidian distances for five dimensions.

In summary, the resulting ML trees and NMDS revealed that four satellite DNAs (pBuM, DBC-150, CDSTR138 and SAT197/230) exhibit the pattern of concerted evolution, with copies forming specie-specific branches on the phylogenetic trees and

segregating species specifically in the NMDS analysis. These satDNAs seem to be promising for their use as taxonomic markers to distinguish *D. antonietae* and *D. serido*, especially in the hybrid zone.

Satellite DNAs in individuals from the hybrid zone

We tested whether pBuM, DBC-150, SAT197-230 and CDSTR138 can be useful as taxonomic markers to classify six individuals collected in the hybrid zone between *D. antonietae* and *D. serido*. Accordingly, we constructed ML trees and NMDS graphics with repeats from *D. serido* and *D. antonietae*, together with repeats extracted from these six individuals (Figure S3-S8).

Analysis of pBuM-1 satDNA revealed that repeats from R75 clustered together with copies from *D. antonietae* (Table 3). On the other hand, repeats from the four genomes from the N61 locality clustered together with copies from *D. serido*. Our analysis of the pBuM-2 in the hybrid zone revealed that all copies from the R75 locality clustered together copies from *D. antonietae*. It was not possible to extract any complete copy of pBuM-2 from the genomes of the N61 locality (Table 3).

Table 3. Numbers of copies from pBuM-1 and pBuM-2 of each genome analyzed (N total). Number of copies, which were clustered together with copies from *D. antonietae* and *D. serido*.

satDNA	Genome	N total	<i>D. antonietae</i>	<i>D. serido</i>
pBuM-1	R73	11	11	0
	R00	12	12	0
	J92	15	0	15
	R37	10	0	10
	N61	10	0	10
	N61A	12	0	12
	N61B	13	0	13
	N61_1	15	0	15
	R75D1_1	12	12	0
	R75D1_2	12	12	0
pBuM-2	R73	7	7	0
	R00	4	4	0
	J92	0	0	0
	R37	0	0	0
	N61	0	0	0
	N61A	0	0	0
	N61B	0	0	0
	N61_1	0	0	0
	R75D1_1	5	5	0
	R75D1_2	5	5	0

The analysis of DBC-150 variants revealed that repeats from genomes in the R75 locality clustered together with copies from *D. antonietae*, with the exception of four

DBC-112 repeats from R75D1_2 that displayed higher similarity with *D. serido*. Repeats from the genomes of the N61 locality grouped with repeats from *D. serido* (Table 4).

Table 4. Numbers of copies from DBC-112 and DBC-150 of each genome analyzed (N total). Number of copies, which were clustered together with copies from *D. antonietae* and *D. serido*.

satDNA	genome	N total	<i>D. antonietae</i>	<i>D. serido</i>
DBC-112	R73	15	15	0
	R00	20	20	0
	J92	28	0	28
	R37	22	0	22
	N61	22	0	22
	N61A	25	0	25
	N61B	23	0	23
	N61_1	32	0	32
	R75D1_1	20	20	0
	R75D1_2	20	16	4
DBC-150	R73	2	2	0
	R00	6	6	0
	J92	5	1	4
	R37	2	0	2
	N61	4	0	4
	N61A	4	0	4
	N61B	0	0	0
	N61_1	6	0	6
	R75D1_1	9	9	0
	R75D1_2	9	9	0

Concerning the SAT97/230, all repeats from the R75 locality were grouped in the same branch as repeats from *D. antonietae*. While, all genomes from the N61 locality clustered together with *D. serido* copies (Table 5).

Table 5. Numbers of copies from SAT197/230 of each genome analyzed (N total). Number of copies, which were clustered together with copies from *D. antonietae* and *D. serido*.

satDNA	genome	N total	<i>D. antonietae</i>	<i>D. serido</i>
SAT197/230	R73	8	8	0
	R00	21	21	0
	J92	11	0	11
	R37	3	0	3
	N61	1	0	1
	N61A	6	0	6
	N61B	4	0	4
	N61_1	10	0	10
	R75D1_1	23	23	0
	R75D1_2	16	16	0

The analysis of the satDNA CDSTR138 revealed that both individuals from the R75 locality were grouped together with *D. antonietae*. On the other hand, in concordance with our findings for the other satellites, repeats from individuals from the N61 locality

exhibited higher identity to *D. serido* (Table 6). Notably, all repeats from *D. serido* formed a distinct branch separate from all copies from the hybrid zone, with a support of 62.7% (Figure S8). This result suggests differentiation of CDSTR138 between the *D. serido* populations.

Table 6. Numbers of copies from CDSTR138 of each genome analyzed (N total). Number of copies, which were clustered together with copies from *D. antonietae* and *D. serido*.

satDNA	genome	N total	<i>D. antonietae</i>	<i>D. serido</i>
CDSTR138	R73	41	41	0
	R00	69	69	0
	J92	18	0	18
	R37	10	0	10
	N61	18	0	18
	N61A	11	0	11
	N61B	11	0	11
	N61_1	27	0	27
	R75D1_1	60	60	0
	R75D1_2	57	57	0

Taken together, our results on satDNAs indicate that the two individuals from the R75 locality belong to *D. antonietae*, while the four individuals from the N61 locality can be classified as *D. serido* (Figure 14). In order to confirm this hypothesis, it will be interesting to combine the results obtained here with satDNAs with studies using other genetic markers in the same sample of individuals.

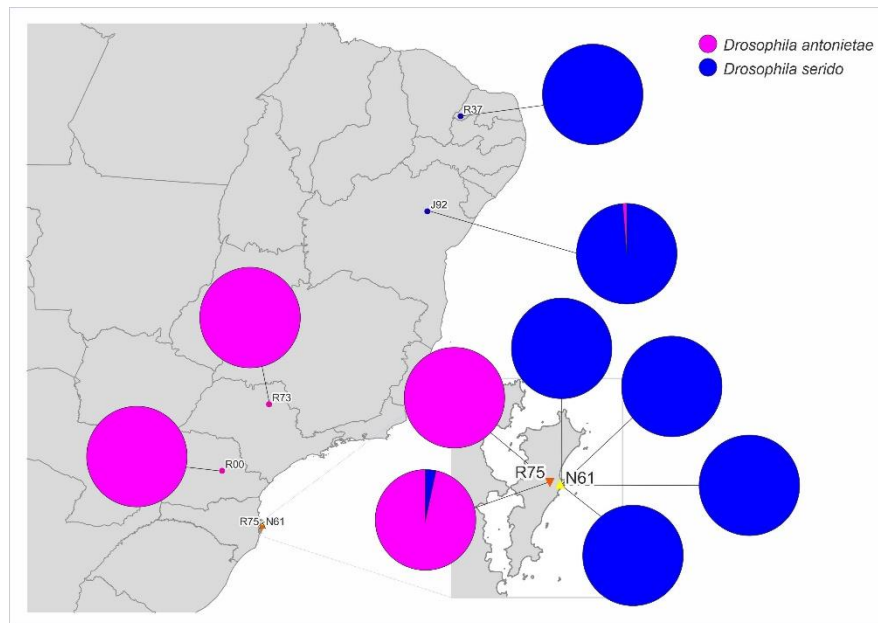


Figure 14. Map showing the geographical collection points of each genome. The pie charts represent the proportion of total copies of satellite DNA that were clustered together with each species.

Conclusions

In the present work, we studied, for the first time, the whole satDNA landscape in *Drosophila antonietae* and *D. serido* based on their sequenced genomes. For this purpose,

we used the RepeatExplorer2 pipeline to identify satDNAs in the genome of both species and in six individuals from the hybrid zone. We identified the same collection of five satellite DNAs in both species. Three of them have been previously described in *repleta* group (Kuhn and Sene 2005; Kuhn *et al.* 2007; Lima *et al.* 2017), while we described here for the first time two satDNAs, named here SAT197/230 and SAT109/104.

The satDNAs families are composed of monomers ranging in size from 100 to 370 bp and both species revealed a low satDNA content. These findings are in accordance with previous results for satellitome investigation in the *repleta* group and suggest that low satDNA content and satDNAs longer than 100 bp might be a feature of the satellite DNAs in species from this group (Kuhn *et al.* 2021). Species from the *repleta* group exhibits small genomes (~150Mb) compared to other *Drosophila* species (usually > 180Mb), and they also have a small heterochromatin content (<3%) compared to other *Drosophila* species (usually >15%) (Bosco *et al.* 2007; Gregory and Johnston 2008). Our findings of a lower contribution of satDNAs to the total genomic DNA, 2.6-3.9% in *D. antonietae* and 1.1-2.9% in *D. serido*, are in accordance with the known positive correlation between the content of satellite DNA, heterochromatin and genome size, that is observed in many *Drosophila* species (Bosco *et al.* 2007; Gregory and Johnston 2008, Kuhn *et al.* 2021).

Our FISH experiments yielded important new insights into the chromosomal distribution of satDNAs. Specifically, the CDSTR138 and not the pBuM (as suggested by previous studies) was found to be the main centromeric component of *D. serido*, *D. seriema*, *D. gouveai* and *D. koepferae*. As pBuM is the main centromeric component of *D. buzzatii*, our results showed a turnover of centromeric satDNAs, from pBuM to CDSTR138, after the separation between *D. buzzatii* and the other species from the *buzzatii* cluster, around 4 Mya (Figure 15) (Kuhn *et al.* 2021).

Contrary to what was hypothesized by Melters *et al.* (2013), we found that the main centromeric component of *D. serido* and *D. seriema*, the CDSTR138, was not the most abundant satDNA

In terms of the two newly discovered satDNAs, we found that both SAT197/230 and SAT109/104 are located on the sex chromosomes of *D. serido*. The SAT197/230 satDNA is localized in three blocks along the Y chromosome, with one located at the centromere. This is the first report of a satDNA localized on the Y chromosome of *D.*

serido. Our analysis did not reveal any evidence of the amplification and expansion of internal repeats from *PERI* in *D. serido*, contrary to previous suggestions (Kuhn and Heslop-Harrison, 2011). Additionally, the SAT109/104 satDNA was found in the centromeric region of the X chromosome and in the terminal region of the Y chromosome. Hence, the centromeric region of the X chromosome from *D. serido* is mainly occupied by CDSTR138 and SAT109/104. Furthermore, our results in *D. serido* showed that CDSTR198 arrays are scattered distributed along the euchromatic arms, suggesting that CDSTR198 in *D. serido* is a minisatellite DNA, as found in other closely related species.

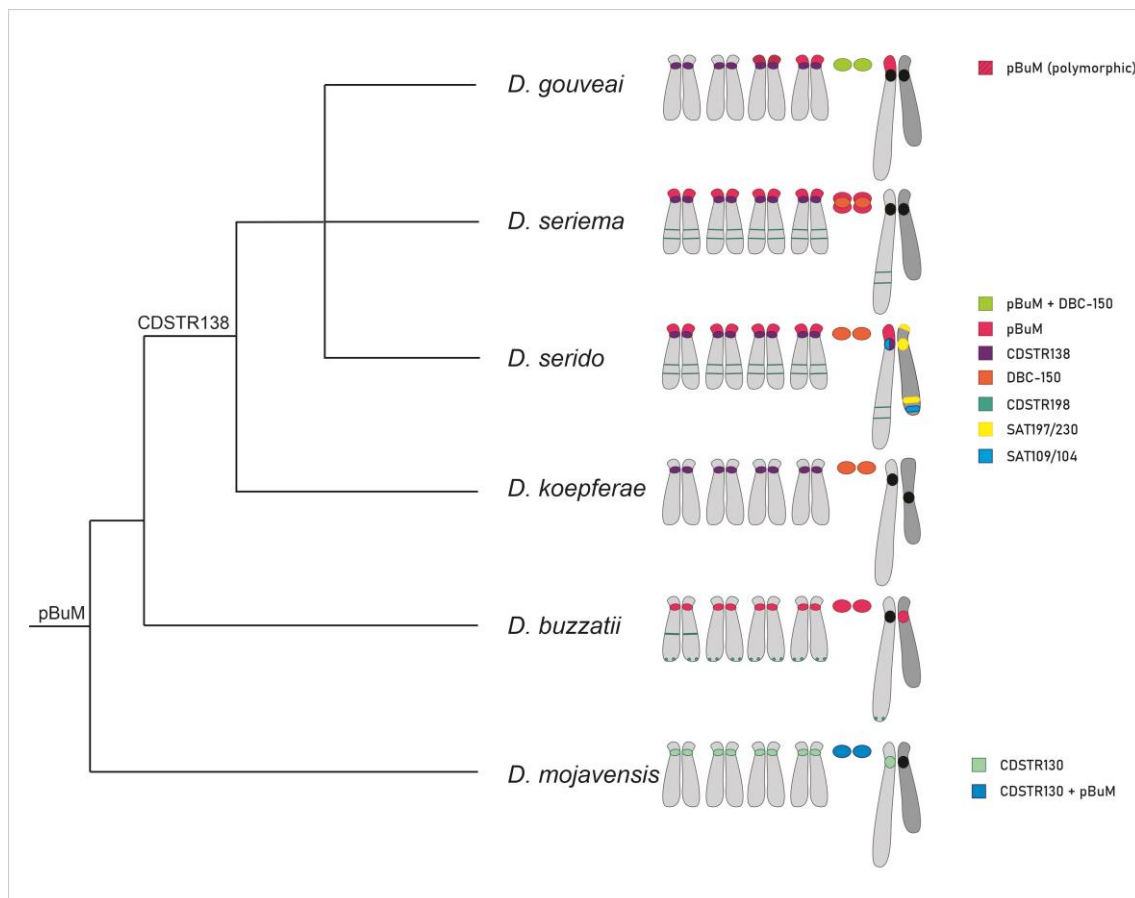


Figure 15. Idiogrammatic representation of the metaphase karyotypes found in *D. gouveai*, *D. seriema*, *D. serido*, *D. koepferae* and *D. buzzatii* and satDNA localization. Data from Kuhn *et al.* (2007, 2008), Lima *et al.* (2017) and the present work. Black-colored chromosome regions represents the centromere.

The centromeric satDNAs analyzed here are enriched for dyad symmetries, their repeats are AT rich and they exhibit a 10-bp periodicity of WW nucleotides. Additionally, these satDNAs were predicted to form stable secondary structures. These features are considered important for centromeric functions and are conserved between species.

Finally, we explored the potential of these satDNAs as taxonomic markers for identifying individuals within the hybrid zone between *D. antonietae* and *D. serido*. Our

study showed that among the five satellite DNAs investigated, pBuM, DBC-150, SAT197/230 and CDSTR138 may be useful as markers to discriminate species. In the context of the hybrid zone, our findings indicate that both genomes from the R75 locality can be classified as *D. antonietae*, while the three genomes from the N61 locality can be classified as *D. serido*. It will be interesting to combine these results with studies using other genetic markers in the same sample of individuals.

The calculated interspecific nucleotide divergence of each satDNA between *D. antonietae* and *D. serido* ranges from 0.6% to 14.4%. These values are relatively low. For example, ~30% nucleotide divergence was reported for the satDNA 1.688 between *D. melanogaster* and *D. simulans*, and such level of divergence seemed to be sufficient to promote hybrid incompatibility between the two species (Ferree and Barbash, 2009). The reduced levels of satDNA divergence between *D. antonietae* and *D. serido* might be one of the reasons that allow the establishment of a hybrid zone between the two species.

Acknowledgments

We thank André Carneiro for several contributions with statistics analysis and Braulio Silva and Pedro Heringer for several discussions during different stages of this work.

References

- AFGAN E et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 2;46(W1):W537-W544. doi: 10.1093/nar/gky379.
- ASHBURNER M et al. (1989) *Drosophila*. A laboratory handbook. Cold Spring Harbor Laboratory Press
- BAIMAI V, SENE FM AND PEREIRA MAQR. (1983) Heterochromatin and karyotypic differentiation of some neotropical cactubreeding species of the *Drosophila* repleta group. *Genetica* 67: 81–92.
- BAIMAI V. (1977) Chromosomal polymorphisms of constitutive heterochromatin and inversions in *Drosophila*. *Genetics*, v. 85, n. 1, p. 85-93
- BARKER, J.S.F., F.M. SENE, P.D. EAST & M.A.Q.R. PEREIRA, (1985). Allozyme and chromosomal polymorphism of *Drosophila buzzatii* in Brazil and Argentina. *Genetica* 67: 161–170

BARRIOS-LEAL D.Y., MENEZES R.S.T., RIBEIRO J.V., BIZZO L., MELO DE SENE F., NEVES-DA-ROCHA J., MANFRIN M.H. (2021) A holocenic and dynamic hybrid zone between two cactophilic *Drosophila* species in a coastal lowland plain of the Brazilian Atlantic Forest. *J Evol Biol.* Nov;34(11):1737-1751. doi: 10.1111/jeb.13934. Epub 2021 Oct 6. PMID: 34538008.

BOSCO, G et al. (2007) Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. *Genetics.* v. 177, n. 3, p. 1277-1290.

BRITTEN, R J.; KOHNE, D E. (1968) Repeated sequences in DNA. *Science*, v. 161, n. 3841, p. 529-540

BRUTLAG D. (1980). Molecular arrangement and evolution of heterochromatic DNA. *Annu Rev Genet.* 14:121–144.

CHARLESWORTH B, LANGLEY CH, STEPHAN W. (1986) The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics.* Apr;112(4):947-62. doi: 10.1093/genetics/112.4.947. PMID: 3957013; PMCID: PMC1202788.

CHARLESWORTH, B., SNIEGOWSKI, P. and STEPHAN, W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371, 215–220.

CRADDOCK EM, GALL JG, JONAS M (2016) Hawaiian *Drosophila* genomes: size variation and evolutionary expansions. *Genetica* 144:107–124

DAWE RK, HENIKOFF S. (2006) Centromeres put epigenetics in the driver's seat. *Trends Biochem Sci.* 2006 Dec;31(12):662-9. doi: 10.1016/j.tibs.2006.10.004. Epub 2006 Oct 30. PMID: 17074489.

DIAS GB, SVARTMAN M, DELPRAT A ET AL (2014) Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*. *Genome Biol Evol* 6:1302–1313

DIXON, P. (2003) VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 2003, 14, 927–930.

DOVER G (1982) Molecular drive: a cohesive mode of species evolution. *Nature* 299:111–117.

DOVER GA (1986) Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. *Trends Genet* 2:159–165. [https://doi.org/10.1016/0168-9525\(86\)90211-8](https://doi.org/10.1016/0168-9525(86)90211-8)

EDGAR, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5, 1–19. <https://doi.org/10.1186/1471-2105-5-113>

FELICIELLO I, AKRAP I, BRAJKOVIĆ J, ZLATAR I, UGARKOVIĆ Đ. (2014) Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*. *Genome Biol Evol.* Dec 19;7(1):228-39. doi: 10.1093/gbe/evu280. PMID: 25527837; PMCID: PMC4316633.

FERREE P.M, BARBASH D.A. (2009) Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. *PLoS Biol.* Oct;7(10):e1000234. doi: 10.1371/journal.pbio.1000234. Epub 2009 Oct 27. PMID: 19859525; PMCID: PMC2760206. Franco *et al.* 2006

FONTDEVILA A. (1989). Founder effects in colonizing populations: the case of *Drosophila buzzatii*. Pp. 74–95 in: Fontdevila A. (ed.), *Evolutionary Biology of Transient Unstable Populations*. – Springer Verlag, Berlin-Heidelberg

GANLEY, A.R.D. (2013) *Concerted Evolution*, Editor(s): Stanley Maloy, Kelly Hughes, *Brenner's Encyclopedia of Genetics (Second Edition)*, Academic Press, Pages 126-130, ISBN 9780080961569

GARRIDO-RAMOS M.A. (2017) Satellite DNA: An Evolving Topic. *Genes* (Basel). Sep 18;8(9):230. doi: 10.3390/genes8090230. PMID: 28926993; PMCID: PMC5615363

GENT, J.I., SCHNEIDER, K.L., TOPP, C.N. *et al.* (2011) Distinct influences of tandem repeats and retrotransposons on CENH3 nucleosome positioning. *Epigenetics & Chromatin* 4, 3 (2011). <https://doi.org/10.1186/1756-8935-4-3>

GREGORY TR, JOHNSTON JS (2008) Genome size diversity in the family *Drosophilidae*. *Heredity* 101:228–238

HENIKOFF S, AHMAD K, MALIK HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293:1098–1102

HJELMEN CE, BLACKMON H, HOLMES VR, BURRUS CG, JOHNSTON JS. (2019) Genome Size Evolution Differs Between *Drosophila* Subgenera with Striking Differences in Male and Female Genome Size in Sophophora. *G3* (Bethesda). Oct 7;9(10):3167-3179. doi: 10.1534/g3.119.400560. PMID: 31358560; PMCID: PMC6778784

JAGANNATHAN M, CUMMINGS R, YAMASHITA YM (2018) A conserved function for pericentromeric satellite DNA. *eLife* 7:e34122

JAGANNATHAN M, CUMMINGS R, YAMASHITA YM (2019) The modular mechanism of chromocenter formation in *Drosophila*. *eLife* 8:e43938

JOSHI SS, MELLER VH (2017) Satellite repeats identify X chromatin for dosage compensation in *Drosophila melanogaster* males. *Curr Biol* 27:1393–1402

KASINATHAN S, HENIKOFF S (2018) Non-B-form DNA is enriched at centromeres. *Mol Biol Evol* 35:949–962

KOKUDAI C.B.S., SENE F.M., MANFRIN M.H. (2011) Sympatry and Asymmetric Introgression between the Cactophilic Species *Drosophila serido* and *Drosophila antonietae* (Diptera: Drosophilidae), *Annals of the Entomological Society of America*, Volume 104, Issue 3, 1 May, Pages 434–442

KUHN G.C, SENE F.M. (2005) Evolutionary turnover of two pBuM satellite DNA subfamilies in the *Drosophila buzzatii* species cluster (repleta group): from alpha to alpha/beta arrays. *Gene*. Apr 11;349:77-85. doi: 10.1016/j.gene.2004.11.032. PMID: 15777676.

KUHN G C S, RUIZ A, ALVES M A R, SENE F M (1996) The metaphase and polytene chromosomes of *Drosophila seriema* (repleta group; mulleri subgroup). *Brazil J Genet* 19:209–216

KUHN GCS, BOLLGONN S, SPERLICH D, BACHMANN L (1999) Characterization of a species-specific satellite DNA of *Drosophila buzzatii*. *J Zool Syst Evol Res* 37:109–112

KUHN G.C., FRANCO F.F., MANFRIN M.H., MOREIRA-FILHO O., SENE F.M. (2007) Low rates of homogenization of the DBC-150 satellite DNA family restricted to a single pair of microchromosomes in species from the *Drosophila buzzatii* cluster.

Chromosome Res. 15(4):457-69. doi: 10.1007/s10577-007-1138-x. Epub 2007 May 15. PMID: 17551842.

KUHN, G C S et al. (2008) Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. *Chromosome Research*, v. 16, n. 2, p. 307-324.

KUHN, G CS et al. (2012) The 1.688 repetitiveDNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. *Molecular biology and evolution*, v. 29, n. 1, p. 7-11.

KUHN G C S, HERINGER P, DIAS G B. (2021) Structure, Organization, and Evolution of Satellite DNAs: Insights from the *Drosophila repleta* and *D. virilis* Species Groups. *Prog Mol Subcell Biol*. 60:27-56. doi: 10.1007/978-3-030- 74889-0_2.

KUMAR S, STECHER G, LI M, KNYAZ C, TAMURA K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol*. 1;35(6):1547-1549. doi: 10.1093/molbev/msy096. PMID: 29722887; PMCID: PMC5967553.

LANDER, E S. (2001) et al. Initial sequencing and analysis of the human genome. 2001.

LEVY, S et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol*. 2007, 5, e254.

LIMA LG, SVARTMAN M, KUHN GCS (2017) Dissecting the satellite DNA landscape in three cactophilic *Drosophila* sequenced genomes. *G3* 7:2831–2843.

LÓPEZ-FLORES I, GARRIDO-RAMOS MA (2012) The repetitive DNA content of eukaryotic genomes. In: Garrido-Ramos MA (ed) *Repetitive DNA*. Karger, Basel, pp 1–28

MACAS, J.; NEUMANN, P.; NAVRATILOVA, A. (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: Comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genom*. 2007, 8, 427.

MALIK H S, HENIKOFF S. (2001) Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*. *Genetics*. Mar;157(3):1293-8. doi: 10.1093/genetics/157.3.1293. PMID: 11238413; PMCID: PMC1461554.

MANFRIN MH, SENE FM (2006). Cactophilic *Drosophila* in South America: a model for evolutionary studies. *Genetica*. Jan;126(1-2):57-75. doi: 10.1007/s10709-005-1432-5.

MANFRIN MH, BRITO ROA AND SENE FM. (2001). Systematics and evolution of the *Drosophila buzzatii* (Diptera; Drosophilidae) cluster using mtDNA. *Ann Entomol Soc Am* 94: 333–346.

MATEUS RP AND SENE FM. (2003). Temporal and Spatial Allozyme Variation in the South American Cactophilic *Drosophila antonietae* (Diptera; Drosophilidae). *Biochem Genet* 41: 219–233.

MELTERS DP, BRADNAM KR, YOUNG HA et al (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* 14:R10

MENON DU, COARFA C, XIAO W, et al (2014) siRNAs from an X-linked satellite repeat promote X-chromosome recognition in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* 111:16460–16465.

MESTROVIĆ N, PLOHL M, MRAVINAC B, UGARKOVIĆ D. (1998) Evolution of satellite DNAs from the genus *Palorus*--experimental evidence for the "library" hypothesis. *Mol Biol Evol*. 1998 Aug;15(8):1062-8. doi: 10.1093/oxfordjournals.molbev.a026005. PMID: 9718733

MRAVINAC, B.; PLOHL, M.; MEŠTROVIĆ, N.; UGARKOVIĆ, D. (2002) Sequence of PRAT satellite DNA “frozen” in some Coleopteran species. *J. Mol. Evol.* 2002, 54, 774–783.

MRAVINAC, B.; PLOHL, M.; UGARKOVIĆ, D. (2005) Preservation and high sequence conservation of satellite DNAs suggest functional constraints. *J. Mol. Evol.* 2005, 61, 542–550.

MIGA, K.H. (2015) Completing the human genome: The progress and challenge of satellite DNA assembly. *Chromosome Res.* 2015, 23, 421–426.

MILLS W K, LEE Y C G, KOCHENDOERFER A M, et al. (2019) RNA from a simple-tandem repeat is required for sperm maturation and male fertility in *Drosophila melanogaster*. *Elife* 8:e48940

MOREYRA NN, ALMEIDA FC, ALLAN C, FRANKEL N, MATZKIN LM, HASSON E. (2023) Phylogenomics provides insights into the evolution of cactophily and host plant shifts in *Drosophila*. *Mol Phylogenet Evol.* 2023 Jan;178:107653. doi: 10.1016/j.ympev.2022.107653. Epub 2022 Oct 28. PMID: 36404461.

NOVÁK P, NEUMANN P, MACAS J. (2020) Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat Protoc.* 2020 Nov;15(11):3745-3776. doi: 10.1038/s41596-020-0400-y. Epub 2020 Oct 23. PMID: 33097925.

NOVÁK, P et al. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, v. 29, n. 6, p. 792-793, 2013

NOVÁK P, ÁVILA ROBLEDILLO L, KOBLÍŽKOVÁ A, VRBOVÁ I, NEUMANN P, MACAS J. (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* 2017 Jul 7;45(12):e111. doi: 10.1093/nar/gkx257. PMID: 28402514; PMCID: PMC5499541.

O'GRADY, PM.; DESALLE, R. (2018) Phylogeny of the genus *Drosophila*. *Genetics*, v. 209, n. 1, p. 1-25.

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

PALOMEQUE T, LORITE P (2008) Satellite DNA in insects: a review. *Heredity* 100:564–573

PATHAK RU, MAMILLAPALLI A, RANGARAJ N et al (2013) AAGAG repeat RNA is an essential component of nuclear matrix in *Drosophila*. *RNA Biol* 10:564–571

PLOHL M, MEŠTROVIĆ N, MRAVINAC B (2012) Satellite DNA evolution. In: Garrido-Ramos MA (ed) *Repetitive DNA*. Karger, Basel, pp 126–152

POWELL, J R. (1997) Progress and prospects in evolutionary biology: the *Drosophila* model. Oxford University Press

PRYTKOVA, T. R., ZHU, X., WIDOM, J., & SCHATZ, G. C. (2011). Modeling DNA-bending in the nucleosome: Role of AA periodicity. *Journal of Physical Chemistry B*, 115(26), 8638-8644. <https://doi.org/10.1021/jp203564q>

ROŠIĆ S, KÖHLER F, ERHARDT S (2014) Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. *J Cell Biol* 207:335–349.

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

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

RUIZ, A., CASSIAN, A.M., KUHN, G.C.S. ALVES M.A.R. e SENE F.M. (2000) The *Drosophila* serido speciation puzzle: Putting new pieces together. *Genetica*, 2000 108: 217-227.

RUIZ-RUANO F.J, LÓPEZ-LEÓN M.D, CABRERO J, CAMACHO J.P.M. (2016) High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci Rep.* 2016;6:28333. Published 2016 Jul 7. doi:10.1038/srep28333

STOTHARD P (2000) The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* 28:1102-1104.

STÖVER BC, MÜLLER KF. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinf.* 11:7

STRACHAN T, WEBB D, DOVER GA (1985) Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. *EMBO J* 4:1701–1708

STRUHL K, SEGAL E. (2013) Determinants of nucleosome positioning, *Nat. Struct. Mol. Biol.* 20. 267–273, <https://doi.org/10.1038/nsmb.2506>

TALBERT PB, HENIKOFF S (2020) What makes a centromere? *Exp Cell Res* 389:111895

TALBERT PB, KASINATHAN S, HENIKOFF S. (2018) Simple and Complex Centromeric Satellites in *Drosophila* Sibling Species. *Genetics*. 2018 Mar;208(3):977-990. doi: 10.1534/genetics.117.300620. Epub 2018 Jan 5. PMID: 29305387; PMCID: PMC5844345.

TAMURA K, STECHER G, AND KUMAR S (2021) MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Molecular Biology and Evolution* 38:3022-3027.

KUMAR S, STECHER G, LI M, KNYAZ C, TAMURA K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 35(6):1547–1549.

THROCKMORTON LH (1975) The phylogeny, ecology, and geography of *Drosophila*. In: King RC (ed) *Handbook of genetics*, vol 3. Plenum Publishing Corporation, New York, pp 421–469

WEI KH, GRENIER JK, BARBASH DA, CLARK AG. (2014) Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. *Proc Natl Acad Sci USA*. 2014; 111(52):18793—8. <https://doi.org/10.1073.pnas.1421951112> PMID: 25512552

ZIMMER EA, MARTIN SL, BEVERLEY SM, KAN YW, WILSON AC. (1980) Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc Natl Acad Sci U S A*. 1980 Apr;77(4):2158- 62. doi: 10.1073/pnas.77.4.2158. PMID: 6929543; PMCID: PMC348671.

Table S1. Description of all clusters retrieved from reads of *D. antonietae* (R73) by RepeatExplorer2. The satDNA families analyzed in this study are highlighted in purple.

N input reads		68666900	N nuclear reads		3823811			
N analyzed reads		3908300	Nn = Na - Nc - No					
N contamination reads		35329						
N organelle reads		49160						
Cluster	Proportion (%)	Proportion adjusted (%)	N reads	Satellite Probability	Consensus length (pb)	C index	P index	Similarity hits
1	1.200	1,234	47175	0.994	370	0.986	0.968	satellite DNA
2	1.100	1,080	41312	0.993	197	0.988	0.959	satellite DNA
3	0.680	0,692	26455	0.895	112	0.959	0.951	satellite DNA
4	0.670	0,690	26376	8.89e-12		0.243	0.179	
5	0.490	0,496	18985	0.752	109	0.982	0.867	satellite DNA
6	0.450	0,460	17586	4.44e-06		0.638	0.0410	
7	0.440	0,449	17163	3.39e-10		0.317	0.115	
8	0.420	0,431	16498	0.01280	1489	0.700	0.512	33.08% Class_I/LINE
9	0.420	0,429	16422	3.54e-06		0.847	0.120	
10	0.400	0,414	15821	4.86e-08		0.442	0.0788	
11	0.400	0,409	15644	1.84e-12		0.235	0.0928	
12	0.400	0,408	15619	1.66e-21		0.000897	0.537	19.77% Class_I/LINE
13	0.370	0,375	14323	0.0250	6321	0.881	0.629	7.81% Class_I/LTR/Ty3_gypsy
14	0.370	0,374	14291	1.16e-05		0.465	0.6050	6.35% Class_I/LINE
15	0.360	0,366	13997	3.00e-09		0.376	0.0578	
16	0.350	0,356	13603	0.00527	505	0.793	0.445	
17	0.340	0,352	13465	6.52e-11		0.285	0.8130	100% Contamination
18	0.340	0,349	13348	2.30e-10		0.320	0.0832	
19	0.320	0,328	12550	2.64e-23		0.000717	0.8020	100% Contamination
20	0.310	0,318	12158	6.01e-20		0.0307	0.6470	9.91% Class_I/LTR/Ty3_gypsy
21	0.310	0,318	12144	1.93e-10		0.261	0.483	8.05% Class_I/LTR/Bel-Pao
22	0.300	0,303	11590	4.71e-23		0.000949	0.774	27.22% organelle/mitochondria
23	0.290	0,296	11313	3.43e-13		0.184	0.238	
24	0.270	0,279	10677	6.84e-23		0.000749	0.754	17.62% organelle/mitochondria
25	0.270	0,276	10555	1.17e-17		0.0691	0.0498	
26	0.260	0,268	10257	1.44e-14		0.158	0.0455	
27	0.260	0,264	10112	4.82e-18		0.0602	0.5360	5.13% Class_I/LTR/Ty3_gypsy
28	0.260	0,262	10024	1.21e-18		0.0445	0.2010	
29	0.250	0,260	9935	0.01970	2700	0.775	0.538	5.09% Class_I/LTR/Ty3_gypsy
30	0.250	0,259	9921	6.09e-06		0.460	0.634	16.99% Class_I/LTR/Bel-Pao
31	0.240	0,250	9569	1.17e-22		0.002620	0.7210	44.80% 45S_rDNA/25S_rDNA
32	0.240	0,250	9549	7.64e-22		0.001050	0.6010	16.18% Class_I/LINE
33	0.240	0,244	9327	0.756	138	0.980	0.881	satellite DNA
34	0.240	0,244	9314	6.84e-23		0.000752	0.7460	100% Contamination
35	0.220	0,226	8627	0.04230	6015	0.912	0.727	3.41% Class_I/LTR/Ty3_gypsy
36	0.210	0,213	8149	9.20e-21		0.003910	0.3530	
37	0.210	0,213	8126	4.49e-14		0.1560	0.5710	40.34% Class_II/Subclass_2/Helitro
38	0.200	0,205	7852	1.39e-22		0.000891	0.7060	37.91% 45S_rDNA/18S_rDNA
39	0.200	0,200	7661	1.65e-22		0.001040	0.6980	18.95% organelle/mitochondria
40	0.180	0,188	7173	1.41e-08		0.334	0.5350	41.80% Class_II/Subclass_2/Helitro
41	0.180	0,180	6873	2.85e-06		0.818	0.0962	
42	0.170	0,174	6644	5.69e-23		0.00166	0.7570	14.78% organelle/mitochondria
43	0.170	0,171	6557	1.18e-06		0.425	0.6050	1.45% Class_I/LTR/Bel-Pao
44	0.160	0,165	6323	0.02880	6011	0.881	0.650	7.32% Class_I/LTR/Ty3_gypsy
45	0.160	0,165	6314	4.29e-22		0.000950	0.6410	4.02% organelle/mitochondria
46	0.160	0,165	6301	1.11e-03		0.591	0.5080	4.75% Class_I/LTR/Ty3_gypsy
47	0.160	0,164	6274	1.65e-22		0.000956	0.7010	13.01% organelle/mitochondria
48	0.160	0,164	6263	2.41e-04		0.565	0.389	2.55% Class_I/LTR/Bel-Pao
49	0.160	0,159	6089	6.21e-09		0.335	0.6550	6.75% Class_I/LTR/Ty3_gypsy
50	0.150	0,154	5891	0.02990	6669	0.854	0.641	6.43% Class_I/LTR/Ty3_gypsy
51	0.150	0,154	5873	1.12e-02		0.676	0.564	minisatellite
52	0.150	0,153	5869	3.56e-17		0.0656	0.4340	5.98% Class_II/Subclass_2/Helitron
53	0.150	0,153	5833	0.01780	227	0.723	0.696	
54	0.140	0,143	5467	1.20e-08		0.416	0.0362	

55	0.140	0,142	5419	1.09e-20		0.007570	0.3970	23.53% Class_I/LTR/Bel-Pao
56	0.140	0,139	5330	0.0169	6052	0.707	0.638	6.30% Class_I/LTR/Bel-Pao
57	0.130	0,131	5028	0.15100	3525	0.937	0.776	
58	0.120	0,125	4798	1.17e-20		0.008130	0.2980	
59	0.120	0,120	4579	0.02510	6126	0.900	0.639	4.74% Class_I/LTR/Ty3_gypsy
60	0.110	0,117	4489	9.61e-11		0.250	0.4190	1.47% Class_I/LTR/Ty3_gypsy
61	0.110	0,113	4310	0.0173	6065	0.775	0.535	6.64% Class_I/LTR/Bel-Pao
62	0.110	0,109	4173	7.64e-22		0.003840	0.5980	6.97% Class_I/LTR/Ty3_gypsy
63	0.100	0,106	4065	5.91e-21		0.009590	0.5440	7.40% Class_I/LTR/Ty3_gypsy
64	0.100	0,103	3927	5.69e-23		0.001020	0.7610	
65	0.100	0,079	3011	6.51e-17		0.12200	0.7910	
66	0.099	0,101	3868	4.56e-20		0.01600	0.6060	14.30% Class_I/LTR/Bel-Pao
67	0.096	0,098	3735	0.0115	4461	0.747	0.477	4.66% Class_I/LTR/Bel-Pao
68	0.094	0,096	3685	9.16e-03		0.6750	0.5450	4.64% Class_I/LTR/Ty3_gypsy
69	0.093	0,095	3636	2.18e-18		0.05280	0.4560	4.04% Class_I/LTR/Ty3_gypsy
70	0.093	0,095	3631	6.64e-18		0.0620	0.4870	2.84% Class_I/LTR/Ty3_gypsy
71	0.090	0,092	3510	1.10e-20		0.005420	0.2590	
72	0.089	0,091	3468	8.27e-20		0.040400	0.6970	19.58% Class_II/Subclass_2/Helitro
73	0.085	0,087	3327	4.69e-03		0.6500	0.6430	4.06% Class_I/LINE
74	0.085	0,086	3307	9.15e-06		0.4630	0.5570	5.29% Class_I/LTR/Bel-Pao
75	0.083	0,085	3250	0.0306	4254	0.957	0.759	16.68% Class_I/LTR/Ty1_copia
76	0.083	0,085	3235	0.01820	3431	0.840	0.570	6.62% Class_I/LTR/Ty1_copia
77	0.082	0,084	3218	1.08e-20		0.0140	0.2540	1.52% Class_I/LTR/Ty3_gypsy
78	0.082	0,084	3218	6.10e-04		0.5720	0.5640	
79	0.080	0,082	3125	1.58e-04		0.5300	0.5260	9.38% Class_I/LTR/Ty1_copia
80	0.071	0,072	2769	5.55e-19		0.0365	0.5060	
81	0.068	0,069	2647	2.64e-21		0.007180	0.6420	7.48% Class_I/LTR/Ty3_gypsy
82	0.067	0,068	2612	2.90e-21		0.004590	0.4900	3.06% Class_I/LTR/Ty3_gypsy
83	0.066	0,067	2581	2.66e-12		0.03450	0.4890	
84	0.066	0,067	2568	1.08e-14		0.1170	0.3690	6.04% Class_I/LINE
85	0.062	0,063	2428	1.29e-21		0.002060	0.5590	
86	0.061	0,062	2373	0.0304	4304	0.933	0.676	18.37% Class_I/LTR/Ty1_copia
87	0.061	0,062	2372	6.04e-21		0.001710	0.4130	
88	0.060	0,062	2356	8.09e-21		0.001290	0.03740	
89	0.060	0,061	2351	2.41e-14		0.1640	0.6070	2.42% Class_I/LINE
90	0.057	0,059	2246	1.47e-21		0.000445	0.5490	7.75% Class_I/LTR/Ty3_gypsy
91	0.053	0,054	2076	5.12e-21		0.001450	0.4290	3.28% Class_I/LINE
92	0.051	0,052	1988	1.14e-03		0.58900	0.5150	7.04% Class_I/LTR/Ty3_gypsy
93	0.050	0,051	1936	8.09e-21		0.00310	0.3710	
94	0.048	0,049	1891	3.72e-16		0.09570	0.5810	5.76% Class_I/LTR/Ty3_gypsy
95	0.048	0,049	1891	0.05640	1407	0.820	0.806	4.12% Class_I/LINE
96	0.048	0,049	1877	2.90e-21		0.0016	0.4890	11.03% Class_I/LINE
97	0.047	0,048	1821	2.94e-05		0.5120	0.3750	
98	0.047	0,048	1818	0.992	206	0.990	0.989	
99	0.046	0,047	1797	0.03300	4185	0.917	0.687	13.91% Class_I/LTR/Ty1_copia
100	0.045	0,046	1777	0.0270	5339	0.899	0.647	10.07% Class_I/LTR/Bel-Pao
101	0.045	0,046	1773	3.91e-21		0.002260	0.4620	4.96% Class_I/LINE
102	0.045	0,046	1771	6.04e-21		0.004260	0.4110	1.64% Class_I/LTR/Ty1_copia
103	0.044	0,045	1717	7.64e-22		0.000582	0.5960	28.48% Class_I/LINE
104	0.044	0,045	1709	1.36e-04		0.5330	0.5870	7.67% Class_I/LTR/Ty1_copia
105	0.041	0,042	1618	2.39e-21		0.007420	0.6510	11.68% Class_I/LTR/Ty3_gypsy
106	0.041	0,042	1597	6.53e-21		0.000626	0.3970	23.36% Class_I/LINE
107	0.039	0,040	1529	6.56e-21		0.009160	0.1310	
108	0.038	0,038	1470	1.25e-10		0.2560	0.5640	3.95% Class_I/LTR/Ty3_gypsy
109	0.037	0,038	1452	1.03e-20		0.002120	0.3260	
110	0.037	0,038	1434	1.47e-21		0.000697	0.5470	2.37% Class_I/LINE
111	0.036	0,037	1408	6.04e-21		0.002130	0.4140	
112	0.035	0,035	1357	2.09e-21		0.000737	0.5180	8.62% Class_I/LINE
113	0.033	0,034	1284	1.09e-20		0.003120	0.3200	
114	0.032	0,033	1243	3.78e-19		0.035400	0.5600	7.32% Class_I/LTR/Ty1_copia
115	0.032	0,032	1241	4.29e-21		0.002420	0.4480	
116	0.032	0,032	1235	2.61e-21		0.000810	0.5040	
117	0.031	0,032	1218	1.14e-20		0.002470	0.3110	
118	0.031	0,032	1212	1.11e-20		0.011600	0.3870	
119	0.030	0,031	1177	0.04630	1929	0.818	0.744	

120	0.029	0,030	1152	3.21e-21		0.000868	0.4810	19.53% Class_I/LINE
121	0.029	0,030	1144	4.69e-21		0.002660	0.4440	
122	0.028	0,029	1109	6.02e-03		0.6580	0.6380	
123	0.028	0,029	1105	0.955	278	0.976	0.935	
124	0.028	0,028	1084	6.84e-23		0.000923	0.7460	9.96% Class_I/LINE
125	0.027	0,028	1066	1.59e-20		0.000938	0.2030	
126	0.027	0,028	1060	1.29e-21		0.000943	0.5610	18.02% Class_I/LTR/Ty3_gypsy
127	0.025	0,026	996	0.0375	4231	0.955	0.717	15.16% Class_I/LTR/Ty1_copia
128	0.025	0,026	984	4.29e-21		0.001020	0.4530	6.50% Class_I/LTR/Ty3_gypsy
129	0.024	0,024	926	0.02780	77	0.758	0.597	
130	0.023	0,024	899	1.39e-22		0.00111	0.7120	9.79% Class_I/LTR/Ty3_gypsy
131	0.021	0,022	835	4.29e-21		0.00120	0.4520	2.40% Class_I/LINE
132	0.020	0,020	777	1.05e-20		0.005150	0.2410	
133	0.019	0,020	762	2.34e-21		0.001310	0.5120	14.57% Class_I/LTR/Ty1_copia
134	0.019	0,020	761	4.69e-21		0.00657	0.0825	
135	0.019	0,020	753	4.02e-21		0.009300	0.0591	
136	0.019	0,019	733	1.27e-19		0.04090	0.2790	
137	0.019	0,019	732	8.21e-23		0.001370	0.7390	15.30% Class_I/LINE
138	0.018	0,019	709	0.69100	440	0.951	0.837	
139	0.018	0,018	703	6.04e-21		0.001420	0.4120	
140	0.017	0,018	678	0.05140	4115	0.917	0.747	
141	0.016	0,016	623	0.0103	4082	0.912	0.554	13.00% Class_I/LTR/Ty3_gypsy
142	0.015	0,015	571	2.61e-21		0.001750	0.4990	
143	0.015	0,015	567	0.02880	698	0.899	0.663	
144	0.014	0,014	542	1.18e-05		0.47800	0.6570	
145	0.014	0,014	535	9.92e-21		0.011200	0.2160	
146	0.014	0,014	533	1.03e-20		0.001880	0.3320	
147	0.013	0,014	523	0.901	456	0.967	0.902	
148	0.012	0,012	453	0.975	367	0.982	0.953	
149	0.011	0,011	422	9.20e-21		0.002370	0.3480	11.37% Class_I/LTR/Ty3_gypsy
150	0.010	0,011	409	1.37e-20		0.002440	0.0623	
151	0.010	0,010	395	3.55e-21		0.002530	0.4740	1.52% Class_I/LTR/Ty1_copia
152	0.010	0,010	395	1.40e-21		0.012700	0.7030	

Table S2. Description of all clusters retrieved from reads of *D. antonietae* (R00) by RepeatExplorer2. The satDNA families analyzed in this study are highlighted in purple.

N input reads	23522604	N nuclear reads		4462360
N analyzed reads	4491120	Nn = Na - Nc - No		
N contamination reads	0			
N organelle reads	28760			

Cluster	Proportion (%)	Proportion adjusted (%)	N reads	Satellite Probability	Consensus length (pb)	C index	P index	Similarity hits
1	1.000	1,024	45695	0.992	356	0.994	0.984	satellite DNA
2	0.630	0,631	28171	0.993	197	0.993	0.962	satellite DNA
3	0.580	0,584	26058	6.64e-07		0.551000	0.207	
4	0.510	0,512	22830	1.16e-06		0.778000	0.0476	
5	0.490	0,496	22113	5.63e-13		0.189000	0.2150	
6	0.480	0,486	21679	2.74e-03		0.624000	0.5310	23.86% Class_I/LINE
7	0.450	0,453	20199	2.97e-09		0.388000	0.0251	
8	0.450	0,450	20084	1.36e-02		0.700000	0.5220	41.80% Class_I/LINE
9	0.440	0,450	20072	8.58e-06		0.64100	0.0676	
10	0.430	0,446	19895	7.11e-03		0.656000	0.5900	6.56% Class_I/LTR/Bel-Pao
11	0.430	0,429	19136	2.89e-07		0.393000	0.4360	
12	0.420	0,424	18918	4.31e-12		0.217000	0.0186	
13	0.420	0,419	18690	0.05430	5039	0.939	0.748	19.10% Class_I/LINE
14	0.410	0,410	18306	1.41e-12		0.222000	0.1190	
15	0.380	0,380	16956	1.31e-06		0.566000	0.0317	
16	0.370	0,368	16414	0.05460	3167	0.826	0.789	19.18% Class_I/LTR/Bel-Pao
17	0.360	0,363	16177	0.994	112	0.992	0.970	satellite DNA
18	0.360	0,359	16038	1.01e-03		0.901000	0.3940	
19	0.320	0,318	14179	1.10e-21		0.010400	0.7210	11.97% Class_I/LTR/Ty3_gypsy
20	0.300	0,306	13635	4.29e-22		0.000953	0.6410	21.86% Class_I/LINE
21	0.290	0,296	13212	0.04520	6055	0.856	0.728	3.54% Class_I/LTR/Ty3_gypsy
22	0.270	0,273	12176	5.50e-19		0.030400	0.3280	
23	0.260	0,261	11664	8.79e-06		0.461000	0.5730	43.89% Class_II/Subclass_2/Helitro
24	0.250	0,253	11278	1.47e-21		0.002660	0.5540	
25	0.240	0,239	10671	1.41e-12		0.214000	0.1760	
26	0.230	0,236	10538	0.851	138	0.985	0.904	satellite DNA
27	0.220	0,218	9730	0.03450	2757	0.882	0.684	5.57% Class_I/LTR/Ty3_gypsy
28	0.210		9557	1.56e-24		0.000314	0.9340	60.10% organelle/mitochondria
29	0.210	0,212	9475	5.44e-05		0.518000	0.6840	11.48% Class_I/LTR/Ty3_gypsy
30	0.210	0,209	9338	0.66100	103	0.988	0.860	satellite DNA
31	0.200	0,198	8856	1.06e-05		0.466000	0.4410	
32	0.200	0,197	8813	0.03220	6047	0.797	0.629	4.19% Class_I/LTR/Ty3_gypsy
33	0.200	0,197	8805	3.69e-22		0.001930	0.6550	3.75% Class_I/LINE
34	0.200	0,197	8789	1.65e-10		0.329000	0.0221	
35	0.190	0,193	8616	4.60e-25		0.000812	0.9820	54.05% 45S_rDNA/18S_rDNA
36	0.190	0,192	8548	0.02700	6214	0.898	0.649	8.82% Class_I/LTR/Ty3_gypsy
37	0.190	0,191	8511	6.74e-05		0.541000	0.3640	
38	0.190	0,190	8478	2.33e-09		0.322000	0.6910	6.24% Class_I/LTR/Ty3_gypsy
39	0.180	0,186	8300	3.91e-21		0.004340	0.4600	28.41% Class_I/LTR/Bel-Pao
40	0.180	0,180	8010	0.02660	6013	0.927	0.661	8.29% Class_I/LTR/Ty3_gypsy
41	0.180	0,179	7986	1.12e-02		0.676000	0.5580	7.96% Class_I/LTR/Bel-Pao
42	0.180	0,177	7897	1.34e-05		0.474000	0.4800	minisatellite
43	0.170	0,174	7760	2.05e-18		0.046700	0.0976	
44	0.170	0,171	7614	1.44e-23		0.003810	0.8280	48.61% 45S_rDNA/25S_rDNA
45	0.170	0,170	7607	2.16e-23		0.003150	0.8070	10.66% Class_I/LTR/Ty3_gypsy
46	0.160	0,165	7383	0.02230	2441	0.738	0.564	
47	0.160		7607	2.16e-23		0.000141	0.9080	20.48% organelle/mitochondria
48	0.160	0,157	7028	0.45600	4254	0.952	0.807	19.41% Class_I/LTR/Ty1_copia
49	0.160	0,157	6991	0.02920	6334	0.923	0.672	8.30% Class_I/LTR/Ty3_gypsy
50	0.150	0,155	6902	2.30e-07		0.401000	0.6780	7.59% Class_I/LTR/Ty3_gypsy
51	0.150	0,154	6860	5.32e-09		0.316000	0.5990	11.68% Class_I/LTR/Bel-Pao
52	0.140	0,144	6440	0.00931	6927	0.720	0.457	5.89% Class_I/LTR/Ty3_gypsy
53	0.140	0,143	6389	0.775	3516	0.960	0.880	
54	0.140	0,140	6239	0.07050	6052	0.942	0.762	9.30% Class_I/LTR/Bel-Pao
55	0.130	0,134	5996	9.81e-23		0.000667	0.7260	5.92% Class_I/LTR/Ty3_gypsy

56	0.130	0,128	5733	0.02730	4270	0.917	0.662	5.30% Class_I/LTR/Ty3_gypsy
57	0.130	0,128	5710	4.91e-24		0.004900	0.8790	64.19% 45S_rDNA/25S_rDNA
58	0.130	0,128	5705	1.09e-20		0.004910	0.3160	
59	0.130	0,127	5681	4.57e-20		0.015700	0.6060	13.71% Class_I/LTR/Bel-Pao
60	0.130	0,126	5627	1.80e-15		0.095300	0.4750	
61	0.120	0,123	5508	8.39e-20		0.022300	0.5460	3.89% Class_I/LINE
62	0.120	0,123	5480	0.02650	3359	0.830	0.613	7.55% Class_I/LTR/Ty1_copia
63	0.110	0,115	5114	0.31800	6804	0.941	0.795	8.10% Class_I/LTR/Ty3_gypsy
64	0.110	0,109	4846	3.69e-22		0.000624	0.6470	6.29% Class_I/LTR/Ty3_gypsy
65	0.100	0,105	4681	1.77e-21		0.024800	0.8370	21.30% Class_II/Subclass_2/Helitro
66	0.100	0,105	4668	1.05e-18		0.051600	0.5720	
67	0.099	0,099	4435	0.756	6527	0.983	0.876	5.41% Class_I/LTR/Ty3_gypsy
68	0.098		4393	3.57e-25		0.000228	0.9860	37.29% organelle/mitochondria
69	0.092	0,092	4116	5.08e-03		0.645000	0.6170	13.68% Class_I/LTR/Ty1_copia
70	0.091	0,091	4077	0.04990	227	0.831	0.756	
71	0.090	0,091	4039	1.54e-20		0.004710	0.2240	
72	0.088	0,088	3938	5.78e-22		0.005840	0.7710	15.54% Class_I/LTR/Ty3_gypsy
73	0.085	0,086	3830	0.00482	1048	0.709	0.415	
74	0.085	0,085	3813	0.03270	4926	0.844	0.651	7.97% Class_I/LTR/Bel-Pao
75	0.083	0,083	3713	2.25e-07		0.382000	0.5140	16.19% Class_II/Subclass_2/Helitro
76	0.082	0,083	3697	0.02020	4276	0.761	0.535	8.76% Class_I/LTR/Ty3_gypsy
77	0.080	0,081	3609	0.04560	4304	0.854	0.733	16.93% Class_I/LTR/Ty1_copia
78	0.078	0,079	3510	2.58e-04		0.568000	0.7450	3.90% Class_I/LINE
79	0.076	0,076	3409	0.07030	1913	0.781	0.941	4.58% Class_I/LINE
80	0.075	0,075	3348	0.13100	5804	0.970	0.780	7.56% Class_I/LTR/Ty3_gypsy
81	0.074	0,074	3307	7.64e-22		0.000907	0.5980	16.99% Class_I/LINE
82	0.072	0,073	3251	3.16e-22		0.001850	0.6590	7.66% Class_I/LTR/Ty3_gypsy
83	0.072	0,073	3242	9.20e-21		0.001540	0.3550	1.54% Class_I/LTR/Ty3_gypsy
84	0.072	0,072	3232	0.02540	4081	0.835	0.614	6.93% Class_I/LTR/Ty3_gypsy
85	0.071	0,072	3210	0.04110	5914	0.836	0.698	9.94% Class_I/LTR/Bel-Pao
86	0.071	0,072	3202	0.04906	5815	0.911	0.749	9.96% Class_I/LTR/Bel-Pao
87	0.071	0,072	3191	1.54e-20		0.004070	0.2150	
88	0.070	0,070	3131	5.76e-22		0.000320	0.6210	20.12% Class_I/LINE
89	0.068	0,069	3059	2.29e-22		0.003270	0.6770	18.67% Class_I/LINE
90	0.067	0,068	3030	8.01e-07		0.422000	0.6510	9.08% Class_I/LTR/Ty1_copia
91	0.067	0,067	2998	1.28e-09		0.295000	0.5170	
92	0.065		2939	9.66e-25		0.000340	0.9540	23.55% organelle/mitochondria
93	0.063	0,063	2823	2.70e-22		0.001420	0.6720	8.08% Class_I/LTR/Ty3_gypsy
94	0.062	0,063	2805	2.16e-17		0.072400	0.5280	2.92% Class_I/LINE
95	0.062	0,062	2788	0.01730	1601	0.848	0.567	
96	0.060	0,061	2710	1.43e-16		0.083500	0.2820	
97	0.060	0,061	2703	1.55e-06		0.452000	0.3160	
98	0.058	0,058	2604	9.00e-17		0.133000	0.8400	
99	0.057	0,057	2557	3.21e-05		0.501000	0.4320	7.51% Class_I/LTR/Ty3_gypsy
100	0.056	0,057	2529	1.14e-20		0.002430	0.3110	
101	0.055	0,058	2576	1.09e-20		0.005210	0.3990	
102	0.055	0,055	2462	2.78e-20		0.065000	0.9450	
103	0.054	0,054	2421	0.03360	4185	0.939	0.699	14.25% Class_I/LTR/Ty1_copia
104	0.054	0,054	2414	1.12e-06		0.423000	0.4340	7.87% Class_I/LTR/Ty1_copia
105	0.053	0,054	2402	3.27e-18		0.049200	0.3140	
106	0.053	0,054	2396	1.09e-20		0.003760	0.3220	
107	0.053	0,054	2393	5.69e-23		0.000418	0.7580	29.59% Class_I/LINE
108	0.053		2386	9.66e-25		0.000419	0.9530	6.83% organelle/mitochondria
109	0.053	0,053	2373	2.02e-19		0.016000	0.4350	
110	0.053	0,053	2366	6.24e-05		0.562000	0.2960	
111	0.050	0,050	2226	2.70e-22		0.001360	0.6740	
112	0.050	0,050	2226	0.03550	4534	0.886	0.690	11.28% Class_I/LTR/Ty3_gypsy
113	0.049	0,050	2209	2.14e-19		0.015600	0.4230	3.26% Class_I/LTR/Ty3_gypsy
114	0.049	0,049	2185	3.91e-21		0.001380	0.4630	7.19% Class_I/LINE
115	0.048	0,049	2176	5.76e-22		0.000460	0.6170	
116	0.047	0,048	2122	2.61e-21		0.002360	0.5040	6.60% Class_I/LINE
117	0.044	0,044	1978	1.66e-21		0.000506	0.5440	13.50% Class_I/LINE
118	0.043	0,044	1953	2.70e-22		0.000512	0.6690	9.06% Class_I/LINE
119	0.043	0,043	1932	6.53e-21		0.000546	0.3970	
120	0.042	0,043	1907	7.64e-22		0.001570	0.6040	14.68% Class_I/LINE

121	0.042	0,042	1869	4.29e-21		0.002720	0.4470	
122	0.041	0,042	1858	5.12e-21		0.001610	0.4340	
123	0.041	0,041	1819	5.69e-23		0.000550	0.7560	
124	0.040	0,041	1812	4.14e-06		0.473000	0.7590	10.26% Class_I/LTR/Ty1_copia
125	0.040	0,040	1800	3.65e-19		0.047800	0.6680	8.72% Class_I/LTR/Ty3_gypsy
126	0.038	0,039	1726	4.28e-05		0.505000	0.5880	10.43% Class_I/LTR/Ty3_gypsy
127	0.038	0,039	1725	5.76e-15		0.173000	0.7930	3.88% Class_I/LINE
128	0.038	0,038	1688	2.00e-03		0.632000	0.7150	7.76% Class_I/LTR/Ty3_gypsy
129	0.037	0,037	1644	5.12e-21		0.002730	0.4310	
130	0.036	0,036	1614	0.985	206	0.998	0.980	
131	0.035	0,036	1589	0.780	1174	0.970	0.872	
132	0.033	0,034	1503	1.81e-12		0.220000	0.6570	
133	0.033	0,034	1502	0.35200	4231	0.983	0.807	18.58% Class_I/LTR/Ty1_copia
134	0.033	0,033	1471	0.04230	6175	0.912	0.733	9.59% Class_I/LTR/Ty3_gypsy
135	0.031	0,031	1400	0.03160	3297	0.868	0.665	9.93% Class_I/LTR/Ty3_gypsy
136	0.029	0,029	1299	9.29e-21		0.007700	0.2230	
137	0.027	0,027	1206	3.93e-24		0.000829	0.8930	10.70% Class_I/LINE
138	0.027	0,027	1200	9.20e-21		0.003330	0.3470	
139	0.025	0,025	1106	1.51e-20		0.000918	0.2340	
140	0.024	0,025	1096	1.57e-20		0.004560	0.2120	
141	0.024	0,024	1087	0.64400	1418	0.935	0.914	
142	0.024	0,024	1077	0.742	4108	0.942	0.934	
143	0.024	0,024	1057	6.64e-22		0.000946	0.6110	18.54% Class_I/LINE
144	0.022	0,023	1008	2.34e-21		0.003970	0.5090	
145	0.022	0,022	992	2.70e-22		0.001010	0.6700	23.49% Class_I/LINE
146	0.021	0,021	923	1.48e-20		0.001090	0.2390	
147	0.020	0,020	905	1.05e-18		0.047500	0.5660	13.70% Class_I/LTR/Ty1_copia
148	0.020	0,020	902	4.29e-22		0.001110	0.6400	18.51% Class_I/LTR/Ty3_gypsy
149	0.020	0,020	884	2.34e-21		0.001130	0.5090	17.99% Class_I/LTR/Ty1_copia
150	0.019	0,019	861	1.50e-20		0.001160	0.0885	
151	0.019		855	1.97e-24		0.001170	0.9170	22.69% organelle/mitochondria
152	0.019	0,019	837	8.21e-23		0.001190	0.7400	22.34% Class_I/LINE
153	0.018	0,018	815	1.09e-20		0.001230	0.3230	
154	0.018	0,018	791	4.29e-22		0.001260	0.6410	
155	0.017	0,017	757	3.76e-10		0.272000	0.4700	14.27% Class_I/LINE
156	0.016	0,016	732	1.15e-20		0.005630	0.3710	
157	0.016	0,016	719	1.30e-20		0.004230	0.2840	
158	0.015	0,015	689	4.29e-22		0.001450	0.6370	
159	0.015	0,015	667	0.925	456	0.970	0.956	
160	0.014	0,015	649	0.771	442	0.958	0.865	
161	0.013	0,013	571	2.16e-23		0.001750	0.8130	21.72% Class_I/LINE
162	0.013	0,013	570	8.64e-21		0.001830	0.3600	
163	0.013	0,013	565	0.950	278	0.975	0.915	
164	0.012	0,012	547	1.08e-17		0.071300	0.6040	
165	0.012	0,012	540	4.69e-21		0.001850	0.4360	
166	0.012	0,012	533	7.56e-21		0.001880	0.3770	2.81% Class_I/LTR/Ty3_gypsy
167	0.012	0,012	525	2.61e-21		0.001900	0.4960	14.48% Class_I/LTR/Ty3_gypsy
168	0.011		514	1.47e-21		0.001950	0.5480	63.23% organelle/plastid
169	0.011		509	7.64e-22		0.001960	0.5960	72.69% organelle/plastid
170	0.011	0,011	508	1.95e-22		0.001970	0.6880	28.54% Class_II/Subclass_2/Helit

Table S3. Description of all clusters retrieved from reads of *D. serido* (R37) by RepeatExplorer2. The satDNA families analyzed in this study are highlighted in purple.

N input reads	79158690	N nuclear reads 3484628	
N analyzed reads	3569673	Nn = Na - Nc - No	
N contamination reads	37565		
N organelle reads	47480		

Cluster	Proportion (%)	Proportion adjusted (%)	N reads	Satellite Probability	Consensus length (pb)	C index	P index	Similarity hits
1	1.200	1,192	41523	0.975	190	0.975	0.954	satellite DNA
2	0.69	0,708	24674	0.993	230	0.987	0.941	satellite DNA
3	0.600	0,615	21430	6.37e-12	3417	0.262000	0.0503	16.84% Class_I/LINE
4	0.600	0,614	21395	0.0197		0.858	0.586	
5	0.580	0,302	10533	1.36e-13		0.210000	0.0230	
6	0.550	0,567	19772	4.48e-04		0.731000	0.2700	
7	0.550	0,564	19639	3.87e-15		0.124000	0.1990	
8	0.530	0,545	18974	11.26e-18	138	0.035600	0.2270	satellite DNA
9	0.480	0,489	17044	0.891		0.984	0.901	
10	0.450	0,463	16135	1.06e-15		0.124000	0.0638	
11	0.410	0,416	14493	1.02e-14		0.160000	0.0205	
12	0.390	0,395	13779	5.92e-11		0.249000	0.5600	
13	0.370	0,383	13341	4.44e-06	6249	0.642000	0.0401	48.15% Class_II/Subclass_2/Helitro
14	0.340	0,352	12256	9.58e-19		0.038700	0.1180	
15	0.330	0,341	11895	2.09e-21		0.003870	0.5150	
16	0.320	0,329	11448	0.0242		0.945	0.649	
17	0.320	0,328	11446	1.28e-06		0.418000	0.4540	
18	0.310	0,323	11241	5.69e-23	6921	0.000801	0.7590	31.13% 45S_rDNA/25S_rDNA
19	0.310	0,319	11115	4.90e-07		0.877000	0.0633	
20	0.300		10564	6.84e-23		0.001140	0.7500	
21	0.290		10519	8.21e-23		0.000475	0.7430	
22	0.280	0,285	9919	2.51e-06		0.426000	0.5380	
23	0.270	0,281	9788	4.19e-16	112	0.111000	0.0428	4.86% Class_I/LINE
24	0.270		9718	8.21e-23		0.000720	0.7450	
25	0.270	0,278	9671	0.0244		0.849	0.613	
26	0.260	0,270	9412	0.775		0.955	0.933	
27	0.260		9141	1.17e-22		0.001090	0.7220	
28	0.250	0,261	9083	8.27e-21	4396	0.012400	0.4840	13.95% organelle/mitochondria
29	0.250		8780	1.17e-22		0.000569	0.7180	
30	0.240		8648	2.70e-22		0.000578	0.6670	
31	0.240		8635	2.29e-22		0.000811	0.6840	
32	0.220	0,223	7774	1.00e-21		0.000772	0.5830	
33	0.210	0,220	7666	0.0144	7342	0.750	0.500	16.63% Class_I/LINE
34	0.180	0,187	6527	2.14e-09		0.297000	0.5490	
35	0.180	0,187	6514	4.30e-12		0.261000	0.0165	
36	0.180	0,186	6464	8.37e-03		0.679000	0.6790	
37	0.180		6415	7.64e-22		0.000779	0.5970	
38	0.170	0,174	6061	1.66e-09	104	0.378000	0.0197	46.16% organelle/mitochondria
39	0.170	0,173	6025	5.71e-13		0.185000	0.3080	
40	0.160	0,161	5625	3.87e-05		0.534000	0.3440	
41	0.160	0,159	5553	0.0334		0.786	0.643	
42	0.150		5444	8.75e-22		0.001100	0.5890	
43	0.140	0,148	5166	3.65e-03	6010	0.641000	0.4830	20.68% organelle/mitochondria
44	0.140	0,144	5006	7.64e-22		0.001200	0.5990	
45	0.140	0,143	4985	0.0368		0.828	0.669	
46	0.140		4852	6.64e-22		0.001240	0.6120	
47	0.130	0,138	4795	0.0258		0.816	0.602	
48	0.130	0,130	4545	0.0151	6243	0.720	0.511	6.69% Class_I/LTR/Ty3_gypsy
49	0.120	0,126	4377	3.41e-19		0.029000	0.4520	
50	0.120	0,125	4344	0.0351		0.954	0.706	
51	0.120	0,122	4251	0.0163		0.885	0.578	
52	0.120	0,122	4234	1.12e-15		0.132000	0.6470	
53	0.120	0,120	4176	0.0266	5983	0.926	0.664	7.87% organelle/mitochondria
54	0.110	0,115	4023	1.44e-20		0.003980	0.2460	
55	0.110	0,112	3888	0.0323		0.931	0.690	
56	0.100	0,107	3743	8.21e-23		0.00801	0.7440	

57	0.099	0,101	3529	1.14e-21		0.000283	0.5660	31.23% 45S_rDNA/25S_rDNA
58	0.095	0,098	3402	3.22e-19		0.053200	0.6760	19.69% Class_I/Subclass_2/Helitro
59	0.091	0,093	3250	0.0370	199	0.918	0.705	minisatellite
60	0.091	0,093	3241	2.54e-05		0.485000	0.4530	2.84% Class_I/LTR/Ty3_gypsy
61	0.089	0,092	3193	5.12e-21		0.001570	0.4310	17.01% Class_I/LINE
62	0.087	0,089	3094	2.61e-21		0.000970	0.5010	17.71% Class_I/LTR/Ty3_gypsy
63	0.086	0,088	3072	0.0149	4839	0.841	0.549	9.11% Class_I/LTR/Ty3_gypsy
64	0.084	0,086	3000	1.29e-21		0.000333	0.5640	5.87% Class_I/LINE
65	0.084	0,086	2996	1.10e-04		0.600000	0.2610	
66	0.083	0,085	2965	4.29e-21		0.000337	0.4520	
67	0.082	0,084	2930	4.11e-21		0.013300	0.5940	7.41% Class_I/LTR/Ty3_gypsy
68	0.078	0,080	2775	4.80e-21		0.011900	0.5730	15.96% Class_I/LTR/Bel-Pao
69	0.077	0,078	2732	1.14e-20		0.004030	0.3080	6.84% Class_I/LINE
70	0.073	0,075	2609	1.05e-04		0.539000	0.7010	
71	0.072	0,074	2575	1.14e-21		0.000388	0.5720	26.52% Class_I/LINE
72	0.070	0,071	2483	2.61e-21		0.001210	0.4970	7.65% Class_I/LINE
73	0.068	0,070	2441	3.92e-14		0.143000	0.4660	
74	0.068	0,070	2429	3.64e-06		0.455000	0.4150	3.05% Class_I/LINE
75	0.066	0,067	2339	8.64e-21		0.000442	0.3590	
76	0.060	0,062	2144	1.14e-20		0.003740	0.3070	4.24% Class_I/LTR/Ty1_copia
77	0.057	0,059	2046	1.14e-20		0.001470	0.3070	
78	0.055	0,056	1968	9.20e-21		0.000509	0.3520	1.88% Class_I/LTR/Ty3_gypsy
79	0.054	0,056	1941	1.06e-20		0.006430	0.4110	3.40% Class_I/LTR/Ty1_copia
80	0.054	0,055	1922	5.76e-22		0.002600	0.6210	7.08% Class_I/LTR/Ty3_gypsy
81	0.050	0,052	1800	8.64e-21		0.002250	0.3620	
82	0.049	0,050	1745	4.29e-21		0.000584	0.4530	18.91% Class_I/LINE
83	0.048	0,050	1727	3.16e-22		0.002320	0.6620	
84	0.047	0,048	1676	1.14e-20		0.000602	0.3080	
85	0.047	0,048	1668	1.14e-20		0.002400	0.3080	
86	0.046	0,047	1641	0.0088	3084	0.845	0.508	7.98% Class_I/LTR/Ty1_copia
87	0.045	0,046	1594	1.86e-21		0.000627	0.5310	16.50% Class_I/LINE
88	0.042	0,044	1517	0.0299	5725	0.905	0.722	14.30% Class_I/LINE
89	0.042	0,043	1501	1.30e-20		0.000667	0.2830	
90	0.042		1491	1.56e-20		0.004020	0.1140	21.33% organelle/mitochondria
91	0.042	0,043	1491	2.43e-06		0.434000	0.4920	
92	0.041	0,042	1476	0.0159	4131	0.831	0.547	11.59% Class_I/LTR/Ty1_copia
93	0.041	0,042	1459	0.0169	3809	0.821	0.546	1.17% Class_I/LTR/Ty3_gypsy
94	0.040	0,041	1433	0.0490	1201	0.849	0.748	6.42% Class_I/LINE
95	0.040		1427	0.5790	112	0.931	0.873	1.68% organelle/plastid
96	0.040	0,041	1424	4.69e-21		0.002130	0.4370	4.63% Class_I/LTR/Ty3_gypsy
97	0.039	0,040	1402	7.30e-03		0.677000	0.4710	7.85% Class_I/LTR/Bel-Pao
98	0.038	0,039	1370	1.62e-20		0.000730	0.1830	
99	0.037	0,038	1338	5.12e-21		0.000747	0.4330	19.28% Class_I/LINE
100	0.037	0,038	1335	1.98e-19		0.033900	0.5340	5.92% Class_I/LINE
101	0.037	0,038	1315	7.73e-21		0.008370	0.1570	
102	0.034	0,035	1229	3.63e-11		0.238000	0.5320	4.23% Class_I/LTR/Ty3_gypsy
103	0.034	0,035	1203	1.60e-20		0.03330	0.1320	
104	0.034	0,034	1200	6.65e-03		0.678000	0.4560	14.00% Class_I/LTR/Ty1_copia
105	0.033	0,034	1188	0.986	255	0.996	0.973	
106	0.033	0,034	1175	7.56e-21		0.004260	0.3760	
107	0.032	0,033	1160	1.48e-20		0.000862	0.2420	
108	0.030	0,030	1061	8.09e-21		0.002890	0.3660	
109	0.027	0,028	974	0.0138	459	0.701	0.648	
110	0.026	0,026	923	1.40e-06		0.434000	0.4030	
111	0.025	0,026	901	6.75e-20		0.025500	0.6380	10.21% Class_I/LTR/Ty3_gypsy
112	0.025	0,026	894	7.04e-21		0.001120	0.3950	
113	0.025	0,025	881	3.39e-19		0.019300	0.2770	3.18% Class_I/LTR/Ty3_gypsy
114	0.024	0,025	863	1.54e-20		0.003490	0.2210	7.53% Class_I/LTR/Ty1_copia
115	0.024	0,024	849	2.70e-22		0.001180	0.6750	
116	0.023		838	2.15e-18		0.050100	0.1130	3.22% organelle/plastid
117	0.023	0,023	809	4.29e-22		0.003710	0.6410	10.88% Class_I/LINE
118	0.022	0,023	794	0.0558	126	0.879	0.784	
119	0.022	0,023	785	1.86e-21		0.001270	0.5300	18.73% Class_I/LTR/Ty3_gypsy
120	0.021	0,022	755	1.25e-20		0.001370	0.2860	
121	0.020	0,021	731	6.13e-04		0.804000	0.3240	
122	0.020	0,021	722	1.15e-20		0.005540	0.3730	

123	0.020	0,021	716	2.61e-21		0.004190	0.4980	5.31% Class_I/LTR/Ty3_gypsy
124	0.019	0,020	682	3.40e-21		0.011700	0.0444	
125	0.019	0,020	681	1.20e-20		0.001470	0.2970	
126	0.019	0,019	661	1.30e-20		0.004540	0.2790	
127	0.018	0,019	660	9.81e-23		0.001520	0.7280	11.36% Class_I/LTR/Ty3-gypsy
128	0.018	0,019	654	7.04e-21		0.004590	0.3860	
129	0.017	0,018	619	7.64e-22		0.001620	0.6040	16.80% Class_I/LTR/Ty3_gypsy
130	0.017	0,018	611	1.14e-21		0.001640	0.5710	
131	0.017	0,017	604	1.35e-20		0.001680	0.2690	
132	0.016	0,017	586	5.57e-21		0.001710	0.4220	7.51% Class_I/LTR/Ty3_gypsy
133	0.016	0,016	560	1.47e-21		0.001860	0.5470	
134	0.015	0,016	550	8.75e-22		0.001830	0.5900	
135	0.015	0,016	550	2.68e-17		0.067300	0.1800	
136	0.015	0,016	548	0.2700	57	0.909	0.839	
137	0.015	0,016	543	8.09e-21		0.001910	0.3680	
138	0.014	0,015	507	1.03e-20		0.005920	0.4240	
139	0.014	0,014	494	5.76e-22		0.002020	0.6200	
140	0.014	0,014	493	1.66e-21		0.002030	0.5410	
141	0.014	0,014	489	9.20e-21		0.002040	0.3470	49.28% Class_II/Subclass_2/Helitro
142	0.013	0,013	459	2.61e-21		0.002180	0.5050	
143	0.013	0,013	451	0.0485	48	0.874	0.748	
144	0.013	0,013	447	0.781	456	0.964	0.886	
145	0.013	0,013	447	2.09e-21		0.002240	0.5150	21.70% Class_I/LINE
146	0.012	0,013	441	9.75e-21		0.002320	0.3360	
147	0.012	0,013	441	2.90e-21		0.002270	0.4950	19.73% Class_I/LTR/Ty1_copia
148	0.012	0,012	431	1.00e-19		0.016200	0.5280	
149	0.012	0,012	423	6.05e-19		0.026000	0.2330	
150	0.012	0,012	419	0.6470	1953	0.964	0.830	
151	0.012	0,012	411	3.21e-21		0.002430	0.4780	
152	0.011	0,012	408	3.95e-17		0.073500	0.3190	
153	0.011	0,012	406	2.34e-21		0.002480	0.5090	2.22% Class_I/LINE
154	0.011	0,011	387	0.1620	2646	0.984	0.792	
155	0.011	0,011	387	4.29e-21		0.002580	0.4550	13.70% Class_I/LTR/Ty1_copia

Table S4. Description of all clusters retrieved from reads of *D. serido* (J92) by RepeatExplorer2. The satDNA families analyzed in this study are highlighted in purple.

N input reads	71201362	N nuclear reads	2521539
N analyzed reads	2684156	Nn = Na - Nc - No	
N contamination reads	0		
N organelle reads	162617		

Cluster	Proportion (%)	Proportion adjusted (%)	N reads	Satellite Probability	Consensus length (pb)	C index	P index	Similarity hits
1	0.720		19268	1.17e-17		1.08e-01	0.864	78.37% organelle/mitochondria
2	0.670	0,715	18026	1.55e-19		6.79e-02	0.897	
3	0.660		17799	6.78e-21		4.71e-02	0.948	6.70% organelle/mitochondria
4	0.650	0,695	17513	2.33e-19		7.94e-02	0.956	
5	0.640	0,680	17139	3.38e-04		7.35e-01	0.263	
6	0.580	0,616	15526	0.01330	1709	0.876	0.560	
7	0.570	0,602	15173	9.15e-21		5.19e-02	0.927	
8	0.570	0,602	15169	5.82e-22		2.87e-02	0.965	
9	0.540	0,576	14517	1.03e-17		6.15e-02	0.283	
10	0.530	0,566	14282	2.79e-21		3.86e-02	0.933	
11	0.520		14060	3.26e-21		3.66e-02	0.925	12.82% organelle/mitochondria
12	0.490	0,522	13161	1.92e-14		1.88e-01	0.882	
13	0.490		13142	5.90e-25		7.61e-05	0.967	20.84% organelle/mitochondria
14	0.470	0,505	12734	3.19e-20		5.61e-02	0.928	
15	0.470		12641	3.65e-20		5.51e-02	0.923	8.98% organelle/mitochondria
16	0.460	0,494	12444	3.18e-13		2.23e-01	0.894	
17	0.450	0,477	12030	2.33e-22		1.81e-02	0.961	
18	0.440	0,465	11727	2.02e-21		4.08e-02	0.954	
19	0.430		11662	9.25e-16		1.64e-01	0.904	56.74% organelle/mitochondria
20	0.420	0,450	11339	5.55e-22		1.67e-02	0.913	
21	0.420	0,446	11243	2.38e-21		4.05e-02	0.944	
22	0.410		11111	8.25e-12		2.54e-01	0.768	45.52% organelle/mitochondria
23	0.410		10909	1.65e-17		1.13e-01	0.829	28.39% organelle/mitochondria
24	0.400	0,433	10924	3.91e-23		1.38e-02	0.936	
25	0.390	0,419	10567	2.78e-20		5.95e-02	0.940	
26	0.380		10183	5.89e-11		2.77e-01	0.833	52.80% organelle/mitochondria
27	0.370	0,398	10029	0.26400	3175	0.927	0.801	18.39% Class_II/Subclass_2/Helitro
28	0.370	0,395	9956	7.31e-19		7.63e-02	0.864	
29	0.370	0,393	9906	0.00361	871	0.884	0.457	
30	0.350	0,374	9436	2.78e-22		1.82e-02	0.945	
31	0.350	0,368	9289	8.38e-20		7.13e-02	0.952	
32	0.350	0,367	9264	2.19e-21		2.77e-02	0.884	
33	0.320	0,345	8705	1.29e-11		2.61e-01	0.830	
34	0.320	0,340	8580	9.15e-21		5.35e-02	0.930	
35	0.320		8571	4.79e-16		1.52e-01	0.862	11.69% organelle/mitochondria
36	0.320	0,339	8555	2.33e-19		7.81e-02	0.958	
37	0.320	0,339	8540	2.71e-23		7.61e-03	0.964	
38	0.320	0,337	8495	5.18e-03		6.64e-01	0.684	satellite DNA
39	0.320	0,336	8474	2.12e-20		5.79e-02	0.957	
40	0.310		8420	5.05e-19		8.11e-02	0.887	18.40% organelle/mitochondria
41	0.310	0,328	8269	4.57e-14		2.01e-01	0.899	
42	0.300	0,321	8106	2.71e-23		8.14e-03	0.957	
43	0.290	0,312	7878	3.94e-22		1.99e-02	0.928	
44	0.290	0,310	7819	5.82e-22		3.39e-02	0.959	
45	0.280	0,296	7463	1.30e-22		1.38e-02	0.867	77.67% 45S_rDNA/18S_rDNA
46	0.280	0,295	7447	5.41e-15		1.77e-01	0.940	
47	0.280		7425	6.70e-17		1.28e-01	0.865	84.86% organelle/mitochondria
48	0.280	0,294	7425	2.71e-23		8.62e-03	0.961	
49	0.270	0,286	7218	6.13e-14		1.80e-01	0.665	4.68% Class_I/LINE
50	0.260	0,281	7079	1.23e-20		4.65e-02	0.910	
51	0.260	0,276	6963	2.12e-20		5.69e-02	0.959	
52	0.260	0,274	6916	0.979	190	0.997	0.988	satellite DNA
53	0.260	0,273	6873	5.63e-03		6.46e-01	0.523	
54	0.260		6851	2.92e-09		3.43e-01	0.837	2.44% organelle/mitochondria
55	0.250	0,269	6787	1.86e-21		2.92e-02	0.892	

56	0.250	0,269	6781	3.31e-22		2.12e-02	0.941	
57	0.250	0,268	6767	3.26e-23		1.39e-02	0.948	
58	0.250	0,264	6653	3.01e-21		2.51e-02	0.856	61.27% 45S_rDNA/25S_rDNA
59	0.250	0,264	6653	3.82e-21		3.86e-02	0.913	
60	0.250	0,262	6610	5.58e-23		1.26e-02	0.918	
61	0.240	0,253	6386	3.28e-18		5.26e-02	0.273	
62	0.220	0,238	5990	7.01e-05		5.36e-01	0.854	
63	0.200	0,215	5411	7.56e-25		3.33e-03	0.961	
64	0.200	0,209	5279	3.26e-23		8.33e-03	0.949	
65	0.190	0,203	5120	1.23e-20		4.79e-02	0.913	
66	0.190	0,201	5058	0.993	230	0.991	0.956	satellite DNA
67	0.180	0,195	4905	1.10e-22		6.53e-03	0.882	
68	0.180	0,190	4802	0.0952	5724	0.900	0.805	5.69% Class_I/LTR/Ty3_gypsy
69	0.170	0,186	4680	3.97e-20		4.96e-02	0.825	58.97% 45S_rDNA/18S_rDNA
70	0.170	0,181	4554	2.83e-22		7.91e-03	0.825	67.08% 45S_rDNA/25S_rDNA
71	0.160	0,170	4296	0.979	1997	0.996	0.989	
72	0.160	0,168	4228	0.6470	5802	0.958	0.834	11.45% Class_I/LTR/Bel-Pao
73	0.150	0,158	3990	2.71e-23		5.51e-03	0.960	
74	0.140	0,151	3803	1.23e-24		2.63e-03	0.943	
75	0.140	0,150	3789	7.56e-25		1.58e-03	0.960	
76	0.140	0,146	3674	1.60e-06		4.23e-01	0.535	3.59% Class_I/LTR/Bel-Pao
77	0.130	0,141	3555	2.71e-23		9.56e-03	0.955	
78	0.130	0,141	3550	3.78e-14		1.99e-01	0.960	
79	0.130	0,137	3457	2.33e-22		1.71e-02	0.958	
80	0.130	0,137	3446	5.69e-23		8.71e-04	0.764	19.15% Class_I/LINE
81	0.130	0,136	3431	2.71e-23		6.41e-03	0.961	
82	0.130	0,133	3360	0.0254	6920	0.744	0.636	5.54% Class_I/LTR/Ty3_gypsy
83	0.120	0,123	3099	9.66e-25		9.68e-04	0.953	
84	0.110	0,117	2956	3.26e-23		1.25e-02	0.951	
85	0.110	0,112	2825	0.767	138	0.978	0.891	satellite DNA
86	0.100	0,111	2791	2.25e-23		7.17e-03	0.968	
87	0.100	0,110	2782	2.25e-23		1.22e-02	0.966	
88	0.100	0,110	2776	2.33e-22		1.66e-02	0.965	
89	0.100	0,108	2713	2.57e-07		4.16e-01	0.862	satellite DNA
90	0.094	0,100	2527	1.10e-22		1.42e-02	0.876	
91	0.093	0,099	2487	0.02740	228	0.707	0.960	
92	0.088	0,094	2364	3.04e-11		2.39e-01	0.550	3.98% Class_I/LTR/Ty3_gypsy
93	0.087	0,093	2345	3.52e-05		5.20e-01	0.771	12.15% Class_I/LINE
94	0.087	0,093	2339	2.71e-23		6.84e-03	0.957	
95	0.087	0,092	2325	3.26e-23		1.20e-02	0.954	
96	0.086	0,091	2307	0.959	3922	1.000	0.999	
97	0.084		2262	3.26e-23		9.28e-03	0.950	14.68% organelle/mitochondria
98	0.083	0,088	2225	3.82e-21		4.18e-02	0.910	
99	0.082		2194	2.71e-23		7.29e-03	0.957	44.76% organelle/plastid
100	0.080	0,085	2145	1.18e-02		6.98e-01	0.717	7.93% Class_I/LTR/Bel-Pao
101	0.079	0,085	2132	3.23e-05		5.13e-01	0.702	5.49% Class_I/LTR/Bel-Pao
102	0.079	0,084	2128	0.55400	4057	0.932	0.878	
103	0.075		2026	2.71e-23		5.43e-03	0.961	29.12% organelle/mitochondria
104	0.075	0,080	2018	9.51e-11		2.57e-01	0.586	8.62% Class_II/Subclass_2/Helitron
105	0.073	0,078	1970	9.66e-25		4.57e-03	0.947	
106	0.072	0,076	1928	9.66e-25		2.07e-03	0.953	
107	0.070	0,075	1887	5.83e-21		4.61e-02	0.957	
108	0.070	0,075	1886	0.986	2743	0.994	0.989	
109	0.070	0,074	1871	5.58e-23		1.23e-02	0.925	56.01% 45S_rDNA/25S_rDNA
110	0.069	0,073	1849	3.21e-23		5.41e-04	0.786	
111	0.068	0,072	1821	2.33e-22		1.59e-02	0.964	
112	0.063	0,067	1692	1.23e-24		4.73e-03	0.945	
113	0.063	0,067	1685	1.40e-16		1.31e-01	0.800	7.42% Class_I/LINE
114	0.063	0,066	1665	8.76e-17		1.04e-01	0.657	9.79% Class_I/LTR/Bel-Pao
115	0.063	0,066	1664	7.56e-25		3.00e-03	0.960	
116	0.063	0,066	1657	2.71e-23		1.27e-02	0.961	
117	0.062	0,066	1656	3.26e-23		1.15e-02	0.948	
118	0.061	0,065	1646	5.90e-25		4.25e-03	0.971	
119	0.061	0,065	1628	0.895	5742	0.965	0.945	17.32% Class_I/LINE
120	0.060	0,064	1605	2.33e-22		1.99e-02	0.957	

121	0.059	0,063	1578	5.90e-25		3.80e-03	0.965	
122	0.058	0,062	1564	0.4110	6101	0.921	0.838	11.06% Class_I/LTR/Bel-Pao
123	0.058	0,062	1558	3.21e-21		6.42e-04	0.478	
124	0.058	0,061	1548	2.71e-23		5.81e-03	0.962	
125	0.057	0,061	1535	0.05580	7342	0.884	0.783	6.06% Class_I/LTR/Ty3_gypsy
126	0.057	0,061	1526	7.56e-25		2.62e-03	0.964	
127	0.057	0,060	1519	9.66e-25		1..97e-03	0.952	
128	0.056		1513	2.16e-23		6.61e-04	0.805	39.66% organelle/mitochondria
129	0.056	0,059	1498	7.56e-25		2.67e-03	0.956	
130	0.056	0,059	1495	2.25e-23		5.35e-03	0.967	
131	0.054	0,057	1445	9.66e-25		3.46e-03	0.950	
132	0.054	0,057	1439	4.80e-20		6.05e-02	0.896	
133	0.052	0,056	1409	3.26e-23		1.14e-02	0.954	
134	0.052	0,056	1409	1.97e-24		4.97e-03	0.925	
135	0.050	0,053	1347	9.20e-21		3.71e-03	0.352	
136	0.049	0,052	1322	1.33e-04		5.46e-01	0.950	
137	0.049	0,052	1307	7.56e-25		3.06e-03	0.962	
138	0.048	0,051	1293	5.90e-25		2.32e-03	0.968	
139	0.048	0,051	1276	1.23e-24		4.70e-03	0.942	
140	0.047	0,050	1272	3.91e-23		1.10e-02	0.942	
141	0.046	0,049	1247	3.26e-23		5.61e-03	0.948	
142	0.045	0,048	1214	7.56e-25		2.47e-03	0.958	3.46% Class_I/LINE
143	0.045	0,048	1202	3.91e-23		9.15e-03	0.939	
144	0.045	0,047	1195	2.25e-23		9.21e-03	0.965	
145	0.043	0,046	1166	4.80e-19		3.52e-02	0.532	19.04% Class_I/LINE
146	0.043	0,046	1160	9.66e-25		2.59e-03	0.953	
147	0.043	0,046	1156	2.71e-23		8.65e-03	0.959	
148	0.043	0,046	1150	9.66e-25		3.48e-03	0.952	
149	0.041	0,045	1137	0.50900	1432	0.945	0.822	
150	0.041	0,044	1113	9.66e-25		8.98e-04	0.949	
151	0.041	0,044	1098	9.66e-25		4.55e-03	0.954	
152	0.041	0,043	1096	2.71e-23		5.47e-03	0.961	
153	0.041	0,043	1096	1.23e-24		3.65e-03	0.943	
154	0.040	0,043	1087	2.71e-23		1.29e-02	0.962	
155	0.040	0,043	1084	7.56e-25		4.61e-03	0.957	
156	0.040	0,043	1076	7.56e-25		2.79e-03	0.964	
157	0.040	0,042	1065	7.56e-25		2.82e-03	0.965	
158	0.038	0,040	1020	9.66e-25		3.92e-03	0.954	
159	0.037	0,039	993	3.26e-23		1.31e-02	0.955	
160	0.037	0,039	990	3.91e-23		5.05e-03	0.945	
161	0.036	0,038	959	3.31e-22		1.56e-02	0.941	57.46% 45S_rDNA/25S_rDNA
162	0.036	0,038	953	1.56e-24		4.20e-03	0.925	
163	0.035	0,038	952	3.26e-23		6.30e-03	0.947	
164	0.035		949	3.91e-23		6.32e-03	0.945	4.21% organelle/mitochondria
165	0.034		922	1.97e-24		1.08e-03	0.917	2.93% organelle/plastid
166	0.034	0,036	904	2.34e-21		3.32e-03	0.509	
167	0.034	0,036	902	5.90e-25		1.11e-03	0.969	
168	0.033	0,036	899	5.90e-25		3.34e-03	0.967	
169	0.033	0,035	882	6.43e-04		5.86e-01	0.951	
170	0.033	0,035	879	7.56e-25		3.41e-03	0.958	
171	0.032	0,034	860	9.66e-25		1.16e-03	0.946	
172	0.032	0,034	853	6.53e-21		3.52e-03	0.396	
173	0.032	0,034	851	3.26e-23		5.88e-03	0.947	
174	0.031	0,033	839	2.25e-23		5.96e-03	0.969	
175	0.031	0,033	834	7.56e-25		3.60e-03	0.958	
176	0.031	0,033	830	0.0236	4864	0.836	0.602	8.19% Class_I/LTR/Ty3_gypsy
177	0.031	0,033	826	2.71e-23		6.05e-03	0.957	
178	0.031	0,033	822	9.66e-25		3.65e-03	0.948	
179	0.030	0,032	818	2.71e-23		9.78e-03	0.962	
180	0.029	0,031	778	3.26e-23		6.43e-03	0.955	
181	0.029	0,031	777	3.91e-23		5.15e-03	0.938	
182	0.029	0,030	765	7.56e-25		5.23e-03	0.947	
183	0.028	0,030	763	7.56e-25		1.31e-03	0.956	
184	0.028	0,030	761	7.56e-25		3.94e-03	0.956	
185	0.028	0,030	752	9.66e-25		1.33e-03	0.948	

186	0.028	0,030	751	3.26e-23		5.33e-03	0.946	
187	0.028	0,030	744	3.26e-23		1.21e-02	0.948	
188	0.027	0,029	738	2.25e-23		5.42e-03	0.968	
189	0.027	0,029	733	3.26e-23		8.19e-03	0.955	
190	0.026	0,028	711	5.90e-25		1.41e-03	0.970	
191	0.026		709	2.71e-23		7.05e-03	0.959	51.48% organelle/mitochondria
192	0.026	0,028	706	3.91e-23		5.67e-03	0.945	
193	0.026	0,028	700	0.959	3577	1.000	1.000	
194	0.026	0,028	698	8.17e-22		3.15e-02	0.944	
195	0.026	0,027	684	3.91e-23		5.76e-03	0.939	
196	0.025	0,027	681	3.58e-10		3.13e-01	0.935	
197	0.025	0,027	679	1.17e-23		1.47e-03	0.840	
198	0.025	0,026	662	7.56e-25		1.51e-03	0.959	
199	0.025	0,026	660	9.66e-25		1.52e-03	0.947	
200	0.024	0,026	657	7.56e-25		1.52e-03	0.955	
201	0.024	0,026	657	2.71e-23		9.13e-03	0.955	
202	0.024	0,026	656	9.66e-25		4.57e-03	0.947	
203	0.024	0,026	655	2.71e-23		1.07e-02	0.961	
204	0.024	0,026	649	9.66e-25		1.54e-03	0.955	
205	0.024	0,025	642	2.71e-23		9.35e-03	0.957	
206	0.024	0,025	638	7.64e-22		1.57e-03	0.595	5.02% Class_I/LTR/Ty3_gypsy
207	0.024	0,025	634	7.88e-23		6.31e-03	0.904	
208	0.023	0,025	629	3.52e-21		2.86e-02	0.850	
209	0.023	0,025	627	3.91e-23		7.97e-03	0.941	
210	0.023	0,025	624	1.17e-23		1.60e-03	0.841	
211	0.023	0,025	619	7.56e-25		1.62e-03	0.959	
212	0.023	0,024	612	3.26e-23		9.80e-03	0.949	
213	0.023	0,024	608	3.69e-22		1.64e-03	0.648	15.46% Class_I/LTR/Ty3_gypsy
214	0.022	0,024	600	3.26e-23		1.33e-02	0.954	
215	0.022	0,024	600	9.66e-25		1.67e-03	0.954	
216	0.022	0,024	600	2.71e-23		8.33e-03	0.961	
217	0.022	0,024	600	7.56e-25		1.67e-03	0.961	
218	0.022	0,024	599	3.26e-23		1.17e-02	0.951	
219	0.022	0,024	596	3.91e-23		1.01e-02	0.935	
220	0.022	0,024	594	1.95e-22		1.68e-02	0.967	
221	0.022	0,023	583	3.26e-23		6.86e-03	0.950	
222	0.022	0,023	583	2.33e-22		1.72e-02	0.956	
223	0.022	0,023	579	5.58e-23		6.91e-03	0.924	
224	0.021	0,023	575	2.71e-23		1.22e-02	0.956	
225	0.021	0,023	574	2.25e-23		8.71e-03	0.966	
226	0.021	0,023	573	4.29e-22		1.75e-03	0.642	14.49% Class_I/LTR/Ty3_gypsy
227	0.021	0,022	567	0.00735	1403	0.727	0.439	
228	0.021	0,022	559	3.91e-23		7.16e-03	0.941	
229	0.021	0,022	558	2.71e-23		1.08e-02	0.965	
230	0.021	0,022	553	0.975	2958	0.976	0.954	
231	0.020	0,022	547	2.71e-23		1.28e-02	0.961	
232	0.020	0,022	543	2.25e-23		1.29e-02	0.967	
233	0.020	0,021	531	2.71e-23		1.13e-02	0.959	
234	0.019	0,021	523	4.60e-25		1.91e-03	0.981	
235	0.019	0,021	523	1.97e-24		1.91e-03	0.923	
236	0.019	0,020	515	3.26e-23		9.71e-03	0.951	
237	0.019	0,020	514	3.26e-23		7.78e-03	0.947	
238	0.019	0,020	513	2.61e-21		1.95e-03	0.500	
239	0.019	0,020	512	3.91e-23		1.17e-02	0.939	
240	0.019	0,020	510	0.02250	188	0.853	0.604	minisatellite
241	0.019	0,020	509	4.68e-23		1.18e-02	0.928	
242	0.018	0,019	490	5.90e-25		2.04e-03	0.968	
243	0.018	0,019	489	2.33e-22		1.64e-02	0.964	
244	0.018	0,019	487	9.66e-25		2.05e-03	0.948	
245	0.018	0,019	480	4.68e-23		6.25e-03	0.928	
246	0.018	0,019	474	2.43e-16		1.50e-01	0.959	
247	0.017	0,018	464	1.56e-24		2.16e-03	0.925	
248	0.017	0,018	464	2.33e-22		1.94e-02	0.958	
249	0.017	0,018	463	4.60e-25		2.16e-03	0.979	
250	0.017	0,018	462	2.71e-23		8.66e-03	0.958	

251	0.017	0,018	461	5.90e-25		2.17e-03	0.970	
252	0.017	0,018	460	9.66e-25		2.17e-03	0.949	
253	0.017	0,018	458	2.25e-23		8.73e-03	0.966	
254	0.017	0,018	452	2.25e-23		6.64e-03	0.974	
255	0.017	0,018	452	6.64e-23		8.85e-03	0.907	
256	0.017	0,018	450	4.91e-24		2.22e-03	0.883	
257	0.017	0,018	449	9.66e-25		2.23e-03	0.952	
258	0.017	0,018	448	9.66e-25		2.23e-03	0.948	
259	0.017	0,018	448	4.68e-23		6.70e-03	0.931	
260	0.017	0,018	446	2.33e-22		1.57e-02	0.965	
261	0.017	0,018	446	3.91e-23		8.97e-03	0.939	
262	0.017	0,018	446	7.56e-25		2.24e-03	0.956	
263	0.016	0,018	442	2.78e-20		5.88e-02	0.939	
264	0.016	0,017	441	2.33e-22		1.59e-02	0.960	
265	0.016	0,017	440	6.53e-21		2.27e-03	0.397	
266	0.016	0,017	438	1.23e-24		2.28e-03	0.938	
267	0.016	0,017	436	9.66e-25		2.29e-03	0.946	
268	0.016	0,017	435	3.91e-23		6.90e-03	0.942	
269	0.016	0,017	434	3.26e-23		6.91e-03	0.946	
270	0.016	0,017	431	2.71e-23		1.16e-02	0.959	
271	0.016	0,017	427	1.95e-22		1.64e-02	0.968	
272	0.016	0,017	425	2.78e-22		1.65e-02	0.950	
273	0.016	0,017	424	3.91e-23		9.43e-03	0.945	
274	0.016	0,017	420	3.21e-23		2.38e-03	0.795	
275	0.016	0,017	419	7.56e-25		2.39e-03	0.958	
276	0.016	0,017	417	9.66e-25		2.40e-03	0.949	
277	0.015	0,016	414	1.23e-24		2.42e-03	0.944	
278	0.015	0,016	412	3.26e-23		7.28e-03	0.953	
279	0.015	0,016	407	4.60e-25		2.46e-03	0.976	
280	0.015	0,016	405	3.45e-22		2.72e-02	0.985	
281	0.015	0,016	404	3.26e-23		7.43e-03	0.952	
282	0.015	0,016	403	9.81e-23		2.48e-03	0.730	14.89% Class_I/LINE
283	0.015	0,016	397	7.56e-25		2.52e-03	0.956	
284	0.015	0,016	396	0.08250	276	0.826	0.932	
285	0.015	0,016	396	2.09e-21		2.53e-03	0.517	
286	0.015	0,016	393	1.23e-24		2.54e-03	0.936	
287	0.014	0,015	385	0.979	255	1.000	0.995	
288	0.014	0,015	384	7.56e-25		2.60e-03	0.959	
289	0.014	0,015	378	2.71e-23		7.94e-03	0.959	
290	0.014	0,015	375	1.86e-23		1.07e-02	0.984	
291	0.014	0,015	373	4.60e-25		2.68e-03	0.984	
292	0.014	0,015	371	2.71e-23		1.35e-02	0.963	
293	0.014	0,015	369	1.23e-24		2.71e-03	0.942	
294	0.014	0,015	369	2.71e-23		8.13e-03	0.963	
295	0.014	0,015	367	3.26e-23		8.17e-03	0.952	
296	0.014	0,015	367	3.26e-23		1.09e-02	0.952	
297	0.014	0,014	365	1.23e-24		2.74e-03	0.941	
298	0.013	0,014	362	0.48700	112	0.903	0.926	
299	0.013	0,014	361	9.66e-25		2.77e-03	0.951	
300	0.013	0,014	360	4.60e-25		2.78e-03	0.978	
301	0.013	0,014	358	7.56e-25		2.79e-03	0.956	
302	0.013	0,014	356	7.56e-25		2.81e-03	0.956	
303	0.013	0,014	350	1.23e-24		2.86e-03	0.944	
304	0.013	0,014	349	1.23e-24		2.87e-03	0.939	
305	0.013	0,014	348	2.71e-23		8.62e-03	0.955	
306	0.013	0,014	348	2.71e-23		8.62e-03	0.955	
307	0.013	0,014	348	2.71e-23		1.15e-02	0.955	
308	0.013	0,014	343	3.91e-23		8.75e-03	0.938	
309	0.013	0,014	342	1.23e-24		2.92e-03	0.943	24.27% Class_I/LTR/Retrovirus
310	0.013	0,013	339	3.91e-23		1.18e-02	0.937	
311	0.012	0,013	335	1.55e-19		7.16e-02	0.903	
312	0.012	0,013	334	3.26e-23		1.20e-02	0.953	
313	0.012	0,013	332	3.13e-24		3.01e-03	0.897	
314	0.012	0,013	325	3.21e-21		3.08e-03	0.484	40.00% Class_I/LINE
315	0.012	0,013	325	4.68e-23		1.23e-02	0.935	

316	0.012	0,013	323	2.71e-23		9.29e-03	0.958	
317	0.012	0,013	322	2.71e-23		9.32e-03	0.963	
318	0.012	0,013	321	2.49e-24		3.12e-03	0.911	
319	0.012	0,013	321	3.47e-21		9.55e-03	0.613	
320	0.012	0,013	319	1.58e-20		3.13e-03	0.115	
321	0.012	0,013	317	7.56e-25		3.15e-03	0.957	
322	0.012	0,013	317	2.71e-23		9.46e-03	0.957	
323	0.012	0,013	317	1.14e-21		2.84e-02	0.921	
324	0.012	0,013	317	3.93e-24		3.15e-03	0.887	
325	0.012	0,013	316	2.33e-22		1.90e-02	0.963	
326	0.012	0,012	315	1.95e-22		1.59e-02	0.969	
327	0.012	0,012	313	4.60e-25		3.19e-03	0.981	
328	0.012	0,012	313	1.95e-22		3.19e-03	0.692	
329	0.012	0,012	312	1.46e-21		3.85e-02	0.975	
330	0.012	0,012	311	1.97e-24		3.22e-03	0.920	
331	0.012	0,012	311	4.60e-25		3.22e-03	0.981	
332	0.012	0,012	309	1.97e-24		3.24e-03	0.919	
333	0.011	0,012	308	2.49e-24		3.25e-03	0.913	
334	0.011	0,012	307	3.94e-22		1.63e-02	0.931	
335	0.011	0,012	307	1.44e-20		3.31e-03	0.248	
336	0.011	0,012	306	2.90e-21		3.29e-03	0.493	
337	0.011	0,012	306	0.959	2560	1.000	1.000	
338	0.011	0,012	305	7.56e-25		3.28e-03	0.955	
339	0.011	0,012	305	6.84e-23		3.28e-03	0.753	20.33% Class_I/LTR/Ty1_copia
340	0.011	0,012	302	1.35e-22		1.66e-02	0.987	
341	0.011	0,012	301	5.90e-25		3.32e-03	0.967	
342	0.011	0,012	300	9.66e-25		3.33e-03	0.948	
343	0.011	0,012	300	7.56e-25		3.33e-03	0.961	
344	0.011	0,012	298	3.26e-23		1.01e-02	0.948	
345	0.011	0,012	297	0.08660	99	0.811	0.967	
346	0.011	0,012	295	1.63e-22		1.69e-02	0.980	
347	0.011	0,012	295	1.97e-24		3.39e-03	0.916	
348	0.011	0,012	295	4.60e-25		3.39e-03	0.980	
349	0.011	0,012	295	9.19e-22		2.37e-02	0.879	
350	0.011	0,012	294	9.66e-25		3.40e-03	0.947	
351	0.011	0,012	293	5.90e-25		3.41e-03	0.966	
352	0.011	0,012	292	2.49e-24		3.42e-03	0.908	
353	0.011	0,012	291	4.60e-25		3.44e-03	0.980	
354	0.011	0,011	288	2.25e-23		1.04e-02	0.973	
355	0.011	0,011	288	5.90e-25		3.47e-03	0.973	
356	0.011	0,011	287	1.23e-24		3.48e-03	0.939	
357	0.011	0,011	286	7.56e-25		3.50e-03	0.959	
358	0.011	0,011	285	5.90e-25		3.51e-03	0.966	
359	0.011	0,011	283	2.49e-24		3.53e-03	0.912	
360	0.011	0,011	282	3.69e-22		3.55e-03	0.649	25.46% Class_I/LINE
361	0.011	0,011	282	7.56e-25		3.55e-03	0.958	
362	0.011	0,011	282	3.31e-22		1.77e-02	0.945	
363	0.010	0,011	280	7.56e-25		3.57e-03	0.958	
364	0.010	0,011	279	2.71e-23		1.08e-02	0.965	
365	0.010	0,011	278	2.90e-21		3.60e-03	0.945	
366	0.010	0,011	277	4.69e-21		3.69e-03	0.443	
367	0.010	0,011	276	8.17e-22		2.90e-02	0.944	
368	0.010	0,011	275	4.60e-25		3.64e-03	0.978	
369	0.010	0,011	273	2.71e-23		1.47e-02	0.964	
370	0.010	0,011	272	7.56e-25		3.68e-03	0.957	
371	0.010	0,011	271	5.58e-23		1.48e-02	0.922	
372	0.010	0,011	270	5.90e-25		3.70e-03	0.971	
373	0.010	0,011	268	1.56e-24		3.73e-03	0.928	
374	0.010	0,011	268	2.25e-23		1.49e-02	0.971	

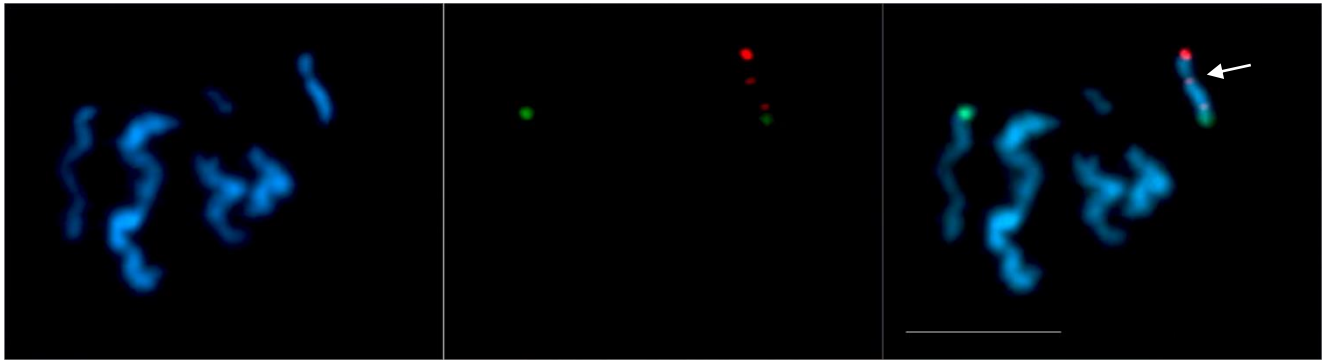


Figure S1. Fluorescent *in situ* hybridization of SAT197/230 (red) and SAT109/104 (green) in prometaphasic chromosomes of *D. serido*. Arrow points to the centromeric array of SAT197/230. Scale bar represents 5 μ M.

```

>CDSTR138_DSDO
CCACTGTTGATTTTTGTATATAGAAATACGAATAAAATCAAATGT
ATGGTGTAGAAGAGTATATTTGCGCCGATTTACCATGAACATAA
CACTAGTTTGATTTTGTATGCATTTATACTGACTATATACATAT
CTGGGC

>SAT197/230_DSDO
ATAAAGAGTAGGCGGAACGTTGCTGGAAGCAACAGGTAAAAGAT
GTGATATATTAAGTCAATTAATGAATATAAATTAGATAACACTG
CTTAGAAAAATGAATAAAAAACAAGACAACCCCTGATTTGTGTAC
ACACATTCAGTAAAATGGTAGTGTAAAAAGGAAAAATTAAAGACAA
GTGAAAACAACCCCTGCTTAGTTACATAAATAGATAAATATAAAA
GAATAGTTAA

>SAT109/104_DSDO
ATATTTGGGACTTGTAAATTTTTTCAAAAAAAAAAATTTTTTTTTC
CTTTCATATGAATTTACTAGGAATTAATGTTCCATATTATATGT
TATGTTCTCTCATAAT
  
```

Figure S2. Example of AA/AT/TT periodicity of the centromeric satellite DNAs from *D. serido*.

Figure S3. ML trees on the left containing copies of pBuM-1 from *D. antonietae*, *D. serido* and the six genomes from the hybrid zone. The ML tree was inferred by the Jukes-Cantor model using 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence is shown on the right. The ordinations represent Euclidian distances for four dimensions.

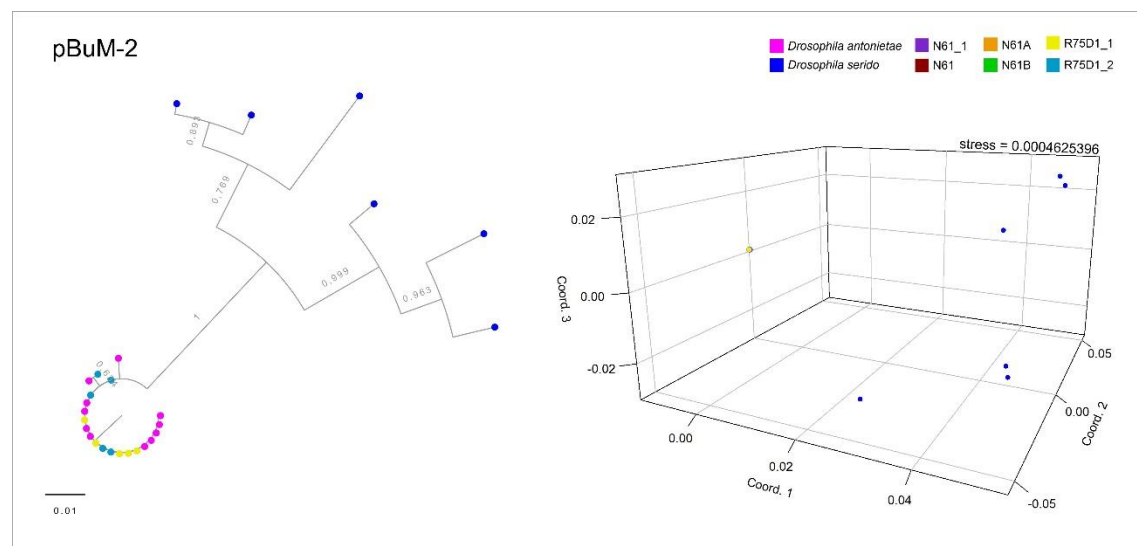


Figure S4. ML trees on the left containing copies of pBuM-2 from *D. antonietae*, *D. serido* and the six genomes from the hybrid zone. The ML tree was inferred by the Tamura-3-parameter model using 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence is shown on the right. The ordinations represent Euclidian distances for four dimensions.

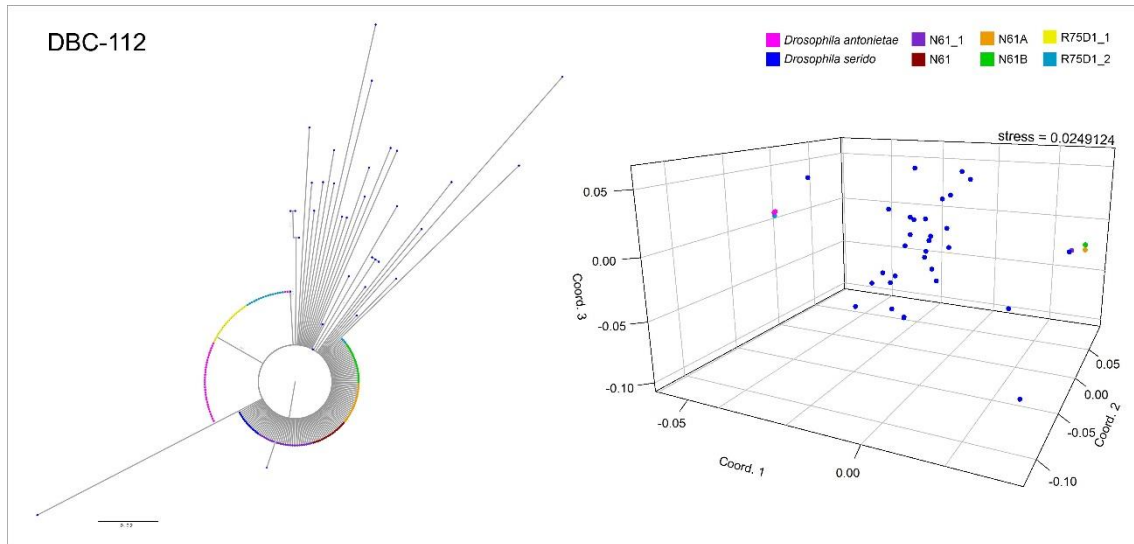


Figure S5. ML trees on the left containing copies of DBC-112 from *D. antonietae*, *D. serido* and the six genomes from the hybrid zone. The ML tree was inferred by the Kimura-2-parameter model using 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence is shown on the right. The ordinations represent Euclidian distances for five dimensions.

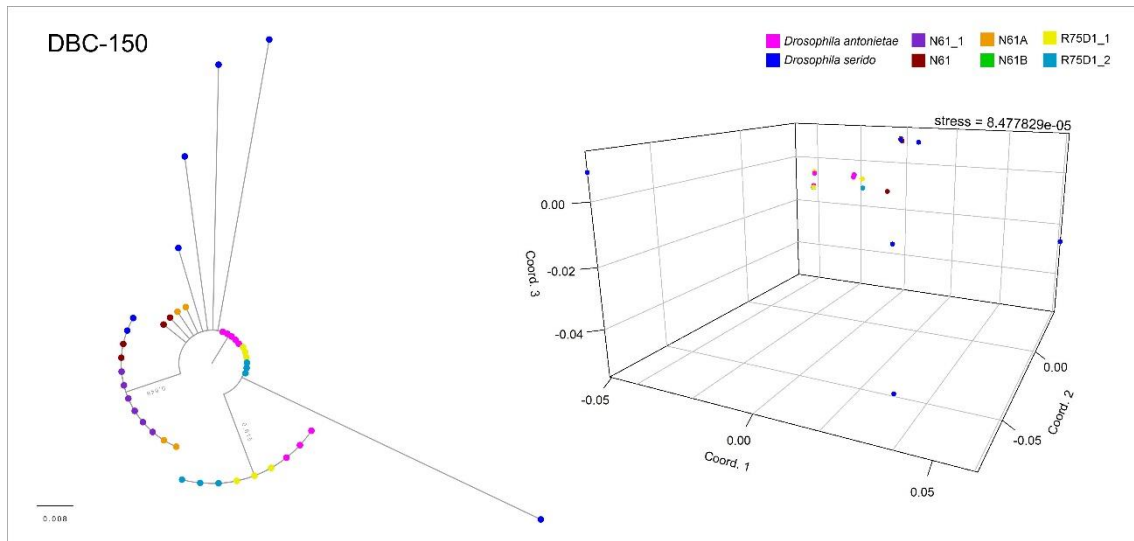


Figure S6. ML trees on the left containing copies of DBC-150 from *D. antonietae*, *D. serido* and the six genomes from the hybrid zone. The ML tree was inferred by the Jukes-Cantor model using 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence is shown on the right. The ordinations represent Euclidian distances for seven dimensions.

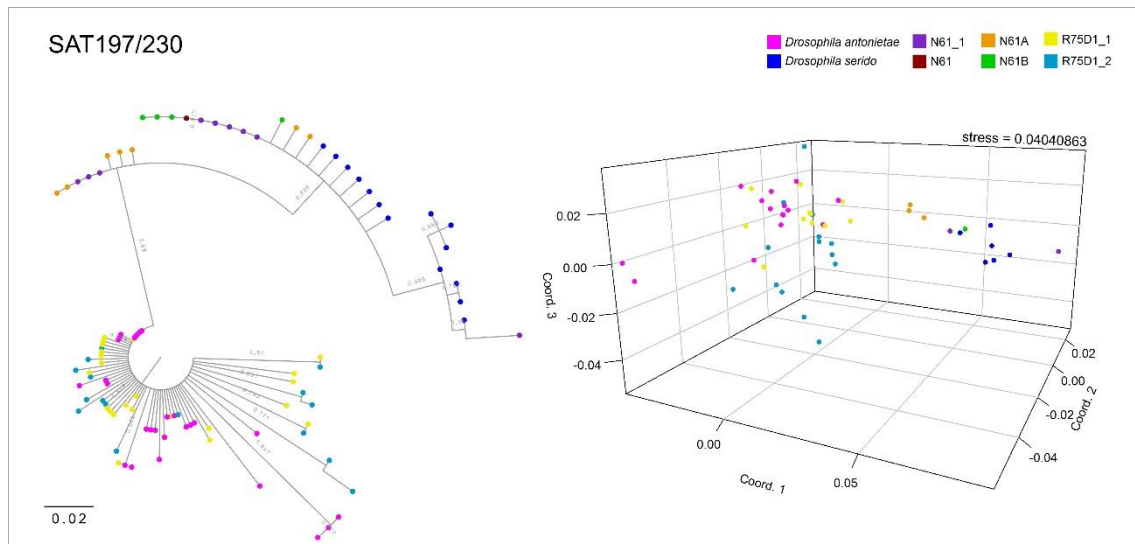


Figure S7. ML trees on the left containing copies of SAT197/230 from *D. antonietae*, *D. serido* and the six genomes from the hybrid zone. The ML tree was inferred by the Tamura-3-parameter model using a discrete Gamma distribution (+G) with 5 rate categories and using 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence is shown on the right. The ordinations represent Euclidian distances for four dimensions.

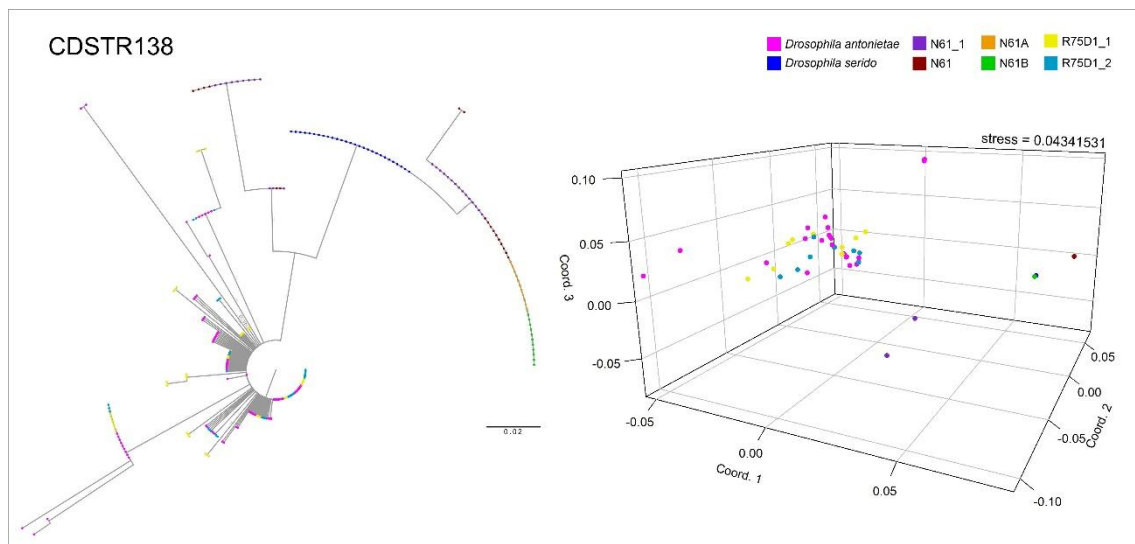


Figure S8. ML trees on the left containing copies of SAT197/230 from *D. antonietae*, *D. serido* and the six genomes from the hybrid zone. The ML tree was inferred by the Tamura-3-parameter model using a discrete Gamma distribution (+G) with 5 rate categories and using 1000 bootstraps. Minimum bootstrap support is 50%. NMDS of evolutionary divergence is shown on the right. The ordinations represent Euclidian distances for four dimensions.

4. CONCLUSÕES

Em relação aos nossos objetivos:

2.2.1 Usar dados de sequenciamento de NGS para identificar, quantificar e caracterizar todos os DNAs satélites de *D. antonietae* e *D. serido*.

A pipeline RepeatExplorer2 foi capaz de identificar cinco DNAs satélites no genoma das duas espécies, incluindo os dois satélites identificados previamente na era pré-genômica, pBuM e DBC-150. Identificamos a presença do satDNA CDSTR138, previamente descrito em outras espécies dos *cluster buzzatii* e descrevemos dois novos satélites. Além disso, a análise revelou que os DNAs satélites contribuem com uma pequena fração do conteúdo genômico em *D. antonietae* e *D. serido*. Esse resultado está em concordância com estudos anteriores em espécies do grupo *repleta*, que apresenta uma baixa contribuição de satDNAs (<5%). Além disso, identificamos a presença do minissatélite CDSTR198, descrito previamente em outras espécies do *cluster buzzatii*. É importante destacar que nenhum satDNA identificado apresentou homologia com elementos transponíveis. Além disso, não foi confirmada a expansão das repetições internas em tandem do TE *PERI* em *D. serido*, como sugerido em trabalhos anteriores.

2.2.2 Determinar a localização cromossômica dos novos DNAs satélites identificados. De posse dessas informações, testar as seguintes hipóteses:

Conforme esperado, os DNAs satélites identificados estão localizados predominantemente na heterocromatina, exceto pelo minissatélite CDSTR198, que está distribuído em pequenos blocos de arranjos eucromáticos. Essa distribuição do CDSTR198 foi confirmada pela sua localização nos braços de cromossomos politênicos de *D. serido*.

O satDNA mais abundante nas duas espécies, o pBuM, ao contrário do que era esperado, não é o principal componente centromérico dos cromossomos. O centrômero é ocupado, principalmente, pelo CDSTR138. Os satDNAs centroméricos apresentam sequências ricas em AT, com tamanhos que variam de 104-230 pb. Além disso, todos os satélites centroméricos apresentam propensão a formar estruturas secundárias estáveis. Os DNAs satélites CDSTR138, SAT197/230, DBC-150 e o pBuM apresentam padrão de evolução combinada e são potencialmente bons marcadores taxonômicos entre as duas espécies. Os novos DNAs satélites identificados, SAT197/230 e SAT109/104, estão

localizados na heterocromatina do Y e na heterocromatina do X, possivelmente na região centromérica.

2.2.3. Utilizar os DNAs satélites para classificar sete genomas da zona híbrida. De posse destas informações, poderemos avaliar se os genomas correspondem as espécies *D. serido* ou *D. antonietae*, ou ainda se constituem híbridos entre as duas espécies.

Os quatros satDNAs que apresentaram padrão de evolução combinada foram usados para classificação de seis genomas da zona híbrida. Os resultados de todos os satélites confirmam que os dois genomas da localidade R75 apresentam padrão de satélite de *D. antonietae*. Enquanto isso, os genomas da localidade N61 foram agrupados com *D. serido*. Os resultados do CDSTR138, satDNA centromérico, revelaram um grau de diferenciação entre as populações de *D. serido*, uma vez que as cópias de *D. serido* foram agrupadas em um ramo separado das cópias dos genomas da zona híbrida.

5. REFERÊNCIAS

ADAMS M D. et al. **The genome sequence of *Drosophila melanogaster*.** Science. 2000. Mar 24;287(5461):2185-95. PMID: 0731132.doi:10.1126/science.287.5461.2185

BAIMAI V, SENE FM AND PEREIRA MAQR. **Heterochromatin and karyotypic differentiation of some neotropical cactubreeding species of the *Drosophila repleta* group.** 1983. Genetica 67: 81–92.

BARKER J S F, SENE F M, EAST P P, PEREIRA M A Q R. **Allozyme and chromosomal polymorphism of *Drosophila buzzatii* in Brazil and Argentina.** Genetica 67: 161-170. (1985)

BARRIOS-LEAL D.Y., MENEZES R.S.T., RIBEIRO J.V., BIZZO L., MELO DE SENE F., NEVES-DA-ROCHA J., MANFRIN M.H. **A holocenic and dynamic hybrid zone between two cactophilic *Drosophila* species in a coastal lowland plain of the Brazilian Atlantic Forest.** J Evol Biol. 2021 Nov;34(11):1737-1751. doi: 10.1111/jeb.13934. Epub 2021 Oct 6. PMID: 34538008.

BARTON N H e HEWITT G M. **Hybrid zones and speciation.** In: Evolution and Specieation, (Atchley, W R & Woodruff, D S., eds.). pp 109-145. Cambridge University Press, Cambridge (1981).

BISCOTTI, M A; OLMO, E; HESLOP-HARRISON, J S P. **Repetitive DNA in eukaryotic genomes.** 2015.

BOSCO, G et al. **Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species.** Genetics. 2007 v. 177, n. 3, p. 1277-1290.

BOULESTEIX, M; WEISS, M; BIÉMONT, C. **Differences in genome size between closely related species: the *Drosophila melanogaster* species subgroup.** Molecular biology and evolution. 2006 v. 23, n. 1, p. 162-167.

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.

C. ELEGANS SEQUENCING CONSORTIUM*. **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** Science, v. 282, n. 5396, p. 2012-2018, 1998.

CASTAGNONE-SERENO, P et al. **Satellite DNA as a versatile genetic marker for *Bursaphelenchus xylophilus*.** In: Pine Wilt Disease: A Worldwide Threat to Forest Ecosystems. Springer, Dordrecht, 2008. p. 187-195.

CHARLESWORTH B, LANGLEY CH, STEPHAN W. **The evolution of restricted recombination and the accumulation of repeated DNA sequences.** Genetics. 1986 Apr;112(4):947-62. doi: 10.1093/genetics/112.4.947. PMID: 3957013; PMCID: PMC1202788.

CHARLESWORTH, B., SNIEGOWSKI, P. and STEPHAN, W. **The evolutionary dynamics of repetitive DNA in eukaryotes.** Nature 371, 215–220 (1994).

COMPEAU P E, PEVZNER P A, TESLER G. **How to apply de Bruijn graphs to genome assembly.** Nat Biotechnol. 2011 8;29(11):987-91. doi: 10.1038/nbt.2023.

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, p. 107061, 2021.

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

DIAS G.B., SVARTMAN M., DELPRAT A., RUIZ A., KUHN G.C. **Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*.** Genome Biol Evol. 2014 May 24;6(6):1302-13. doi: 10.1093/gbe/evu108. PMID: 24858539; PMCID: PMC4079207.

DIAS, C A R et al. **Identification and characterization of repetitive DNA in the genus *Didelphis* Linnaeus, 1758 (*Didelphimorphia*, *Didelphidae*) and the use of satellite DNAs as phylogenetic markers.** Genetics and molecular biology, v. 44, n. 2, 2021.

DOVER, G. **Molecular drive: a cohesive mode of species evolution.** Nature, v. 299, n. 5879, p. 111-117, 1982

DROSOPHILA 12 GENOMES CONSORTIUM et al. **Evolution of genes and genomes on the *Drosophila* phylogeny.** nature, v. 450, n. 7167, p. 203, 2007.

FELICIELLO I, AKRAP I, BRAJKOVIĆ J, ZLATAR I, UGARKOVIĆ Đ. **Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*.** Genome Biol Evol. 2014 Dec 19;7(1):228-39. doi: 10.1093/gbe/evu280. PMID: 25527837; PMCID: PMC4316633.

FERREE P.M, BARBASH D.A. **Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*.** PLoS Biol. 2009 Oct;7(10):e1000234. doi: 10.1371/journal.pbio.1000234. Epub 2009 Oct 27. PMID: 19859525; PMCID: PMC2760206.

FONTDEVILA, A., et al. **“Evolutionary History of *Drosophila Buzzatii*. I. Natural Chromosomal Polymorphism in Colonized Populations of the Old World.”** *Evolution*, vol. 35, no. 1, 1981, pp. 148–57. JSTOR, <https://doi.org/10.2307/2407948>.

FRANCO F F, SILVA-BERNARDI E C C, SENE F M et al. **Intra- and interspecific divergence in the nuclear sequences of the clock gene period in species of the *Drosophila buzzatii* cluster.** J Zool Syst Evol Res 48:322–331(2010)

FU J, ZHANG H, GUO F, MA L, WU J, YUE M, ZHENG X, QIU Z, LI L. **Identification and characterization of abundant repetitive sequences in *Allium cepa*.** Sci Rep. 2019 Nov 14;9(1):16756. doi: 10.1038/s41598-019-52995-9. PMID: 31727905; PMCID: PMC6856378.

GALL J G, COHEN E H, POLAN M L. Repetitive DNA sequences in *Drosophila*. **Repetitive DNA sequences in *Drosophila*.** Chromosoma 33:319–344 (1971)

GANLEY, A.R.D. **Concerted Evolution**, Editor(s): Stanley Maloy, Kelly Hughes, Brenner's Encyclopedia of Genetics (Second Edition), Academic Press, 2013, Pages 126-130, ISBN 9780080961569

GARCÍA, G.; RÍOS, N.; GUTIÉRREZ, V. **Next-generation sequencing detects repetitive elements expansion in giant genomes of annual killifish genus *Austrolebias* (Cyprinodontiformes, Rivulidae).** Genetica, v. 143, n. 3, p. 353-360, 2015.

GARRIDO-RAMOS, M A. **Satellite DNA: an evolving topic.** Genes, v. 8, n. 9, p. 230, 2017.

GATTO, K P. et al. **Sex chromosome differentiation in the frog genus *Pseudis* involves satellite DNA and chromosome rearrangements.** *Frontiers in genetics*, v. 9, p. 301, 2018.

GRADY, D L. et al. **Highly conserved repetitive DNA sequences are present at human centromeres.** *Proceedings of the National Academy of Sciences*, v. 89, n. 5, p. 1695- 1699, 1992.

HARRISON R G. **Hybrid zones: windows on evolutionary process.** *Oxford Surveys in Evolutionary Biology* 7: 69-128 (1990)

HEWITT G.M. **Hybrid zones-natural laboratories for evolutionary studies.** *Trends Ecol Evol.* 1988 Jul;3(7):158-67. doi: 10.1016/0169-5347(88)90033-X. PMID: 21227192

HOWARD, D. J. **Reinforcement: origin, dynamics, and fate of an evolutionary hypothesis.** In: **Hybrid zones and the evolutionary process**, 1993, (Harrison, R. G., ed.). pp. 46-69. Oxford University Press, New York.

JAGANNATHAN M, CUMMINGS R, YAMASHITA Y M (2018) **A conserved function for pericentromeric satellite DNA.** *Elife* 7:e34122.

JAGANNATHAN M, CUMMINGS R, YAMASHITA YM (2019) **The modular mechanism of chromocenter formation in *Drosophila*.** *Elife* 8:e43938.

JOSHI S S, MELLER V H. **Satellite repeats identify X chromatin for dosage compensation in *Drosophila melanogaster* males.** *Curr Biol* 27:1393–1402. (2017)

KIM, B Y. et al. **Highly contiguous assemblies of 101 drosophilid genomes.** *eLife*, v. 10, p. e66405, 2021.

KIT, S. **Equilibrium sedimentation in density gradients of DNA preparations from animal tissues.** *Journal of molecular biology*, v. 3, n. 6, p. 711-IN2, 1961

KOKUDAI C.B.S., SENE F.M., MANFRIN M.H. **Sympatry and Asymmetric Introgression between the Cactophilic Species *Drosophila serido* and *Drosophila antonietae* (Diptera: Drosophilidae),** *Annals of the Entomological Society of America*, Volume 104, Issue 3, 1 May 2011, Pages 434–442,

KUHN G C S, HERINGER P, DIAS G B **Structure, Organization, and Evolution of Satellite DNAs: Insights from the *Drosophila repleta* and *D. virilis* Species Groups.** Prog Mol Subcell Biol. 60:27-56. doi: 10.1007/978-3-030- 74889-0_2. (2021).

KUHN G C S, RUIZ A, ALVES M A R, SENE F M **The metaphase and polytene chromosomes of *Drosophila seriema* (repleta group; mulleri subgroup).** Brazil J Genet 19:209–216 (1996)

KUHN G.C., FRANCO F.F., MANFRIN M.H., MOREIRA-FILHO O., SENE F.M. **Low rates of homogenization of the DBC-150 satellite DNA family restricted to a single pair of microchromosomes in species from the *Drosophila buzzatii* cluster.** Chromosome Res. 2007;15(4):457-69. doi: 10.1007/s10577-007-1138-x. Epub 2007 May 15. PMID: 17551842.

KUHN, G C S et al. **Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster.** Chromosome Research, v. 16, n. 2, p. 307-324, 2008.

KUHN, G C S; SENE, F M. **Evolutionary turnover of two pBuM satellite DNA subfamilies in the *Drosophila buzzatii* species cluster (repleta group): from alpha to alpha/beta arrays.** Gene, v. 349, p. 77-85, 2005.

KUHN, G CS et al. **The 1.688 repetitiveDNA of *Drosophila*: concerted evolution at different genomic scales and association with genes.** Molecular biology and evolution, v. 29, n. 1, p. 7-11, 2012.

KUHN, G. C. S., S. BOLLGÖNN, D. SPERLICH, AND L. BACHMANN. **Characterization of a species-specific satellite DNA of *Drosophila buzzatii*.** J. Zoological Syst. Evol. Res. 1999 37(2): 109– 112

LANDER, E S. et al. **Initial sequencing and analysis of the human genome.** 2001.

LEVY, S et al. **The diploid genome sequence of an individual human.** PLoS Biol. 2007, 5, e254.

LÓPEZ-FLORES, I1; GARRIDO-RAMOS, M. A. **The repetitive DNA content of eukaryotic genomes.** Repetitive DNA, v. 7, p. 1-28, 2012.

LOWER SS, MCGURK MP, CLARK AG, BARBASH DA. **Satellite DNA evolution: old ideas, new approaches.** Curr Opin Genet Dev. 2018;49:70-78. doi:10.1016/j.gde.2018.03.003

MALIK H S, HENIKOFF S. **Adaptive evolution of Cid, a centromere-specific histone in Drosophila.** Genetics. 2001 Mar;157(3):1293-8. doi: 10.1093/genetics/157.3.1293. PMID: 11238413; PMCID: PMC1461554.

MANFRIN M H, BRITO R O A and SENE F M. 2001. **Systematics and evolution of the Drosophila buzzatii (Diptera; Drosophilidae) cluster using mtDNA.** Ann Entomol Soc Am 94: 333–346.

MANFRIN M.H, SENE F.M. **Cactophilic Drosophila in South America: a model for evolutionary studies.** Genetica. 2006 Jan;126(1-2):57-75. doi: 10.1007/s10709-005-1432-5. PMID: 16502085.

MARKOW, T A.; O'GRADY, P. **Drosophila: a guide to species identification and use.** Elsevier, 2005.

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

MENON, D U. et al. **siRNAs from an X-linked satellite repeat promote Xchromosome recognition in Drosophila melanogaster.** Proceedings of the National Academy of Sciences, v. 111, n. 46, p. 16460-16465, 2014.

MESTROVIĆ N, PLOHL M, MRAVINAC B, UGARKOVIĆ D. **Evolution of satellite DNAs from the genus Palorus--experimental evidence for the "library" hypothesis.** Mol Biol Evol. 1998 Aug;15(8):1062-8. doi: 10.1093/oxfordjournals.molbev.a026005. PMID: 9718733

MIGA, K.H. **Completing the human genome: The progress and challenge of satellite DNA assembly.** Chromosome Res. 2015, 23, 421–426.

MIGA, K.H.; NEWTON, Y.; JAIN, M.; ALTEMOSE, N.; WILLARD, H.F.; KENT, W.J. **Centromere reference models for human chromosomes X and Y satellite arrays.** *Genome Res.* 2014, 24, 697–707.

MILLS W K, LEE Y C G, KOCHENDOERFER A M, et al. **RNA from a simple-tandem repeat is required for sperm maturation and male fertility in *Drosophila melanogaster*.** *Elife* 8:e48940 (2019)

MOORE W S. **An evaluation of narrow hybrid zones in vertebrates.** *Quarterly Review of Biology* 52: 263-277 (1977)

MOREYRA NN, ALMEIDA FC, ALLAN C, FRANKEL N, MATZKIN LM, HASSON E. **Phylogenomics provides insights into the evolution of cactophily and host plant shifts in *Drosophila*.** *Mol Phylogenet Evol.* 2023 Jan;178:107653. doi: 10.1016/j.ympev.2022.107653. Epub 2022 Oct 28. PMID: 36404461.

NOOR, M. **Reinforcement and other consequences of sympatry.** *Heredity* 83, 503–508 (1999).

NOVÁK P, ÁVILA ROBLEDILLO L, KOBLÍŽKOVÁ A, VRBOVÁ I, NEUMANN P, MACAS J. **TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads.** *Nucleic Acids Res.* 2017 Jul 7;45(12):e111. doi: 10.1093/nar/gkx257. PMID: 28402514; PMCID: PMC5499541.

NOVÁK P, NEUMANN P, MACAS J. **Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2.** *Nat Protoc.* 2020 Nov;15(11):3745-3776. doi: 10.1038/s41596-020-0400-y. Epub 2020 Oct 23. PMID: 33097925.

NOVÁK, P et al. **RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads.** *Bioinformatics*, v. 29, n. 6, p. 792-793, 2013

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

OLIVEIRA C. 2008 **Causes for sexual isolation in the *Drosophila buzzatii* species cluster.** PhD Thesis. University of Arkansas. Fayetteville

OLIVEIRA D C S G, O'GRADY P M, ETGES W J, HEED W B, DESALLE R. **Molecular systematics and geographical distribution of the *Drosophila longicornis* species complex (Diptera: Drosophilidae).** Zootaxa 1069, 1-32 (2005)

PATHAK RU, MAMILLAPALLI A, RANGARAJ N, et al. **AAGAG repeat RNA is an essential component of nuclear matrix in *Drosophila*.** RNA Biol 10:564–571 (2013)

PATTERSON J T and STONE W S. **Evolution in the Genus *Drosophila*.** MacMillan, New York. 1952

PATTERSON J T. **The *Drosophilidae* of Southwest.** Univ. Tex. Publ. 4313, 7-203. (1943)

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

PLOHL, M.; MEŠTROVIĆ, N.; MRAVINAC, B. **Satellite DNA evolution.** Genome Dyn. 2012, 7, 126–152.

POWELL, J R. **Progress and prospects in evolutionary biology: the *Drosophila* model.** Oxford University Press, 1997.

RAFAEL V, ARCOS G. **Subgrupo inca, un nuevo subgrupo del grupo repleta, com descripción de *Drosophila huancavilcae* n. sp (Diptera, Drosophilidae).** Evol. Biol. 3, 233-243 (1989)

RAND D M and HARRISON R G. **Ecological genetics of a mosaic hybrid zone: mitochondrial, nuclear and reproductive differentiation of crickets by soil type.** Evolution 43: 432-449 (1989)

RICHARDS, S et al. **Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution.** Genome research, v. 15, n. 1, p. 1-18, 2005.

RIESEBERG L H, WHITTON J and GARDNER K. **Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species.** Genetics 152: 713-727 (1999)

ROBLEDILLO A, KOBLÍŽKOVÁ A, NOVÁK P, BÖTTINGER K, VRBOVÁ I, NEUMANN P, SCHUBERT I, MACAS J. **Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing.** Sci Rep. 11;8(1):5838. doi: 10.1038/s41598-018-24196-3. (2018)

RODRÍGUEZ-TRELLES, F et al. **Molecular Evolution and Phylogeny of the *buzzatii* Complex (*Drosophila repleta* Group): A Maximum-Likelihood Approach**, *Molecular Biology and Evolution*, Volume 17, Issue 7, July 2000, Pages 1112–1122, <https://doi.org/10.1093/oxfordjournals.molbev.a026392>

ROŠÍČ S, KÖHLER F, ERHARDT S. **Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division.** J Cell Biol 207:335–349. (2014)

RUIZ A., WASSERMAN M. **Evolutionary cytogenetic of *Drosophila buzzatii* species complex.** Heredity, v. 70, n. 6, p. 582-596, 1993

RUIZ, A., CASSIAN, A.M., KUHN, G.C.S. ALVES M.A.R. e SENE F.M. **The *Drosophila* serido speciation puzzle: Putting new pieces together.** Genetica, 2000 108: 217-227.

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

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

RUIZ-RUANO F.J, LÓPEZ-LEÓN M.D, CABRERO J, CAMACHO J.P.M. **High-throughput analysis of the satellitome illuminates satellite DNA evolution.** Sci Rep. 2016;6:28333. Published 2016 Jul 7. doi:10.1038/srep28333

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.

SANTOS, M H et al. **Esterase-5 gene sequences from the *Drosophila buzzatii* cluster.** DIS, v. 92, p. 44-56, 2009

SENA, R S et al. **Identification and characterization of satellite DNAs in two-toed sloths of the genus Choloepus (Megalonychidae, Xenarthra)**. Scientific reports, v. 10, n. 1, p. 1-11, 2020.

SILVA BSML, HERINGER P, DIAS GB, SVARTMAN M, KUHN GCS. **De novo identification of satellite DNAs in the sequenced genomes of *Drosophila virilis* and *D. americana* using the RepeatExplorer and TAREAN pipelines**. PLoS One. 19;14(12):e0223466. doi: 10.1371/journal.pone.0223466. (2019).

SUJIWATTANARAT, P.; THAPANA, W.; SRIKULNATH, K.; HIRAI, Y.; HIRAI, H.; KOGA, A. **Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids**. Sci. Rep. 2015, 5, 10315.

SUN X et al. **Sequence analysis of a functional *Drosophila* centromeres**. Genome research, v. 13, n.2, p. 182-194 (2003)

TAMURA, K., SUBRAMANIAN, S., KUMAR, S. **Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks**. Mol. Biol. Evol. 21, 36–44. (2004)

THAKUR J, PACKIARAJ J, HENIKOFF S. **Sequence, Chromatin and Evolution of Satellite DNA**. Int J Mol Sci. 2021 Apr 21;22(9):4309. doi: 10.3390/ijms22094309. PMID: 33919233; PMCID: PMC8122249

THROCKMORTON, L H. **The virilis species group**. The Genetics and Biology of *Drosophila*, v. 3, p. 227-296, 1982.

UTSUNOMIA, R et al. **A glimpse into the satellite DNA library in characidae fish (Teleostei, Characiformes)**. Frontiers in genetics, v. 8, p. 103, 2017.

VALERI, M P et al. **Characterization of Satellite DNAs in Squirrel Monkeys genus *Saimiri* (Cebidae, Platyrrhini)**. Scientific reports, v. 10, n. 1, p. 1-11, 2020

VENTER, J. C et al. **The sequence of the human genome**. Science, v. 291, n. 5507, p. 1304-1351, 2001.

VILELA, C. R. **A revision of the *Drosophila repleta* species group (Diptera; Drosophilidae)** Revista Brasileira de Entomologia, v. 27, p. 1-114, 1983

WALKER, P. M. B. **“Repetitive” DNA in higher organisms**, Progress in Biophysics and Molecular Biology, V. 23, 1971, Pages 145-190, ISSN 0079-6107 | doi.org/10.1016/0079-6107 (71)90019-8.

WARBURTON, P.E.; WILLARD, H.F. **Genomic analysis of sequence variation in tandemly repeated DNA. Evidence for localized homogeneous sequence domains within arrays of alpha-satellite DNA**. J. Mol. Biol. 1990, 216, 3–16.

WASSERMAN M and WASSERMAN F. **Inversion polymorphism in island species of Drosophila**. In Hecht, M K. MacIntyre, R and Wallace B. (eds). Evolutionary Biology, 26, 357-381 (1992)

WASSERMAN M. **Cytological studies of replete group of the genus Drosophila. V. The mulleri subgroup**. Univ. Texas Publ, 6205, 85-117.

WHARTON L T. **Interspecific hybridization in the repleta group**. Univ. Tex. Publ. 4445, 175-193. (1944)

WILLARD, H.F.; WAYE, J.S. **Hierarchical order in chromosome-specific human alpha satellite DNA**. Trends Genet. 1987, 3, 192–198.

ZIMMER EA, MARTIN SL, BEVERLEY SM, KAN YW, WILSON AC. **Rapid duplication and loss of genes coding for the alpha chains of hemoglobin**. Proc Natl Acad Sci U S A. 1980 Apr;77(4):2158- 62. doi: 10.1073/pnas.77.4.2158. PMID: 6929543; PMCID: PMC348671