Universidade de Brasília

Instituto de Ciências Biológicas

Departamento de Biologia Celular

Programa de Pós-Graduação em Biologia Molecular

Júlia Alves Luz

Fluxos de trabalho em bioinformática para análise de DNA extracromossômico

2025

Universidade de Brasília

Instituto de Ciências Biológicas

Departamento de Biologia Celular

Programa de Pós-Graduação em Biologia Molecular

Júlia Alves Luz

Fluxos de trabalho em bioinformática para análise de DNA extracromossômico

Dissertação apresentada ao Programa de Pós-Graduação em Biologia Molecular do Departamento de Biologia Celular da Universidade de Brasília, como requisito parcial para obtenção do Título de Mestre (a) em Biologia Molecular.

Orientador: Georgios Joannis Pappas Junior

2025

Júlia Alves Luz

Fluxos de trabalho em bioinformática para análise de DNA extracromossômico

Trabalho desenvolvido no Laboratório de Bioinformática do Departamento de Biologia Celular da Universidade de Brasília, sob a orientação do Prof. Dr. Georgios Joannis Pappas Junior.

Banca Examinadora

Dr. Georgios Joannis Pappas Junior	Presidente da banca
Dr. Robert Neil Gerard Miller	Membro titular
Dr. Robert Edward Pogue	Membro titular
Dr. Ricardo Henrique Kruger	Membro suplente

"*Computers are like Old Testament gods; lots of rules and no mercy.*" - Joseph Campbell, The Power of Myth

Agradecimentos

Existem pessoas demais a quem eu preciso agradecer, e não existem palavras, tempo ou espaço que me permitam expressar toda a minha gratidão a todos.

Gostaria de agradecer aos meus pais por todo o suporte que me proporcionaram até então. Às minhas irmãs, por entenderem. Aos meus amigos de fora do laboratório, pelas risadas - tanto doces quanto amargas. Aos meus amigos de dentro do laboratório, pelo café e a companhia. Aos meus cachorros, pelo amor e devoção incondicionais. Às equipes do HAB e do HUB, à minha fisioterapeuta, aos funcionários da manutenção da UnB, ao ar-condicionado e aos inventores da Pregabalina e do Metilfenidato de liberação lenta.

À minha máquina de desenvolvimento no laboratório, a minha querida Uracila, que suportou muitos "erros de BIOS" sob a minha tutela e aos servidores Helix e DNA, que só me deixaram na mão de vez em quando.

Ao meu orientador, Professor Dr. Georgios Pappas, por ter me acolhido e direcionado durante a minha jornada ao mundo da bioinformática; pelo suporte, pela companhia e pelo *expertise*. Sua orientação foi fundamental para o meu desenvolvimento acadêmico, e espero que possa continuar aprendendo e descobrindo cada vez mais junto ao senhor.

À Bianca Simonassi-Paiva, por não só graciosamente ceder os dados utilizados no projeto, mas por todo o seu apoio, bom humor e companheirismo. Não sabemos tudo, mas fazemos o que dá. Aos Professores Marcelo Brigido e Andrea Maranhão, que me acolheram como parte do laboratório 1, muito obrigada por todo o carinho. Ao professor Roberto Togawa, por toda a ajuda com os servidores quando eles resolveram ser malcriados. Ao meu colega de laboratório, Pedro Barros, meu veterano da bioinformática que me auxiliou muitíssimo, principalmente no início da minha aprendizagem. Eu prometo que um dia te deixo formatar meu computador!

A todos os diferentes professores e alunos de diversos laboratórios que participaram dessa jornada, direta ou indiretamente.
À CAPES pelo apoio financeiro.

Resumo

DNAs circulares extracromossômicos (eccDNAs) são moléculas de DNA circularizadas, nucleares encontradas em todos os organismos eucarióticos investigados até agora. Desde sua descoberta na década de 1960, certas características dos eccDNAs foram esclarecidas; sua replicação independente fora dos cromossomos, variação em tamanho e quantidade e diversidade de conteúdo genético - alguns até abrigam genes codificadores de proteínas e impulsionam a amplificação da expressão. Essas características levaram os eccDNAs a ganharem destaque como potenciais alvos de pesquisa sobre câncer e envelhecimento, devido à sua aparente correlação com a instabilidade genômica. Recentemente, estratégias de sequenciamento de próxima geração foram usadas para caracterizar eccDNAs em diferentes espécies e tipos de células, mas até agora, relativamente poucos programas foram publicados para analisar eccDNAs a partir de dados de sequenciamento. Cada programa é único em termos de requisitos de instalação, parâmetros de execução e formatos de saída, dificultando a usabilidade, a comparação de ferramentas e a interpretação de dados para o usuário final. Nosso objetivo foi desenvolver um pipeline para automatizar a execução de quatro ferramentas diferentes de detecção e processamento de dados de eccDNA sequenciado por leituras-longas (Oxford Nanopore): *FLEC, ecc_finder, cyrcular-calling* e *CReSIL*. Utilizamos a linguagem *Nextflow* para coordenar a execução e o processamento de resultados para cada programa e utilizamos contêineres computacionais para garantir uma instalação simples. Testamos o pipeline usando dados de sequenciamento de nanoporos de amostras de fibroblastos humanos e vesículas extracelulares. Com as configurações padrão, resultados mostraram que o número de eccDNAs previstos variou significativamente entre os programas. Comparações e consolidações adicionais dessas previsões resultaram em um conjunto de eccDNAs consenso para cada amostra, que pode ser usado para anotação e interpretação biológica posterior. Esperamos que o desenvolvimento desta ferramenta facilite o processamento e interpretação de dados relacionados à pesquisa dos eccDNAs por sequenciamento de leituras-longas, ajudando a elucidar mais de suas características e permitir a eventual exploração de seu potencial biotecnológico, terapêutico e diagnóstico, contribuindo assim com a pesquisa sobre eccDNAs.

Palavras-chave: DNA circular extracromossômico, Bioinformática, Nextflow, Sequenciamento *Oxford Nanopore,* leituras-longas*, ecc_finder, cresil, flec, circular-calling, pipelines, workflows*

Abstract

Extrachromosomal circular DNAs (eccDNAs) are nuclear, circularised DNA molecules found in all eukaryotic organisms investigated so far. Since their discovery in the 1960s, certain characteristics of eccDNAs have been illuminated; their independent replication outside of chromosomes, variation in size and quantity, and diversity of genetic content - some even harbouring protein coding genes and driving expression amplification. These characteristics have led to eccDNAs gaining the spotlight as potential cancer and ageing research targets, due to their apparent correlation with genomic instability. Next-generation sequencing strategies have recently been used to characterise eccDNAs in different species and cell types, yet so far, relatively few programs have been published to analyse eccDNAs from sequencing data. Each program is unique in terms of installation requirements, execution parameters and output formats, making usability, tool comparison and data interpretation difficult for the final user. Our objective was to develop a pipeline that automates the execution of four different long-read (Oxford Nanopore) eccDNA data detection and processing tools: *FLEC*, *ecc_finder*, *cyrcular-calling* and *CReSIL*. We utilised the Nextflow domain-specific language to coordinate the execution and processing of results for each program and leveraged computational containers to ensure seamless installation. We tested the pipeline using nanopore sequencing data from human fibroblast and extracellular vesicle samples. Under default settings, preliminary results show the number of predicted eccDNAs varied among the programs. Further comparison and consolidation of these predictions resulted in a consensus set of eccDNAs for each sample, which can be used for annotation and further biological interpretation. We hope the development of this tool will facilitate the installation, validation and reproducibility of long-read eccDNA research data, helping elucidate more of their characteristics and allowing for the eventual exploration of their biotechnological, therapeutic and diagnostic potential, thus contributing to eccDNA research.

Key words: Extrachromosomal circular DNA, Bioinformatics, Nextflow, Oxford Nanopore Sequencing, long-reads, ecc_finder, cresil, flec, circular-calling, pipelines, workflows

Lista De Ilustrações

Lista De Tabelas

Lista De Abreviaturas E Símbolos

- 3D-SIM – *Tridimensional Structured Illumination Microscopy*
- ATAC-seq – *Assay for Transposase-Accessible Chromatin with Sequencing*
- BED – *Browser Extensible Data,* formato de arquivo
- BFB – *Breakage-Fusion-Bridge*
- bp – Pares de bases (*Base pairs*)
- cfDNA – *Cell-Free DNA*
- CPG – (C-p-G) em uma Sequência de DNA
- CPU – *Central Processing Unit*
- CReSIL – *Construction-based Rolling-circle-amplification for eccDNA Sequence Identification and Location*
- CWL – *Common Workflow Language*
- DAG – Grafo Acíclico Direcionado
- DMs – Duplas Minutas
- DNA – Ácido Desoxirribonucleico
- eccDNAs – DNA Circular Extracromossômico
- eccPOP – *eccDNA Pipeline of Pipelines*
- ecDNA – DNA Extracromossômico Circular Longo
- EVs – Vesículas extracelulares
- FAIR – *Findability, Accessibility, Interoperability, Reusability*
- FASTA – ("*FAST-All*") Formato de Arquivo de Sequências
- FLEC – *Full-Length eccDNA caller*
- FISH – *Fluorescence in Situ Hybridization*
- GB – *Giga Byte*
- GPU – *Graphical Processing Unit*
- GRCh – *Genome Reference Consortium (Human)*
- GUI – *Graphical User Interface*
- HD – *Hard Drive*
- HPC – *High Performance Computing*
- kb – Kilobases
- MEV – Microscópio Eletrônico de Varredura

- microDNA – Micro DNA
- NGS – Sequenciamento de Nova Geração (*Next Generation Sequencing*)
- ONT – *Oxford Nanopore Technologies*
- PCR – *Polymerase Chain Reaction*
- RAM – *Random Access Memory*
- RCA – *Rolling Circle Amplification*
- RNA – Ácido Ribonucleico
- scEC – *Single-Cell Extrachromosomal Circular DNA and Transcriptome Sequencing*
- spcDNA – *Small Polydispersed Circular DNA*
- TSV – *Tab Separated Value,* arquivo separado por Tab
- UCSC – *University of California, Santa Cruz*
- WDL – *Workflow Description Language*
- WMS – *Workflow Management Systems*
- YAML – *Yet Another Markup Language*, uma linguagem de serialização de dados

Sumário

1. Introdução

1.1. DNA circular extracromossômico

Em procariotos o genoma está geralmente codificado em uma molécula circular de DNA, com alguns milhões de bases, onde as extremidades 5'-fosfato e 3'-OH da dupla fita encontram-se ligadas covalentemente, formando uma molécula circular sem extremidades livres. Outras moléculas episssômicas de destaque em procariotos são os plasmídeos, os quais são menores, com milhares de bases, mas que mantêm a característica de circularidade. Já em organismos eucarióticos a característica marcante do genoma é a sua distribuição em diversos cromossomos lineares, ou seja, as extremidades 5' e 3' são livres. Essas extremidades possuem uma organização distintiva com repetições em *tandem* de milhares de bases associadas a um complexo proteico (shelterina), as quais são denominadas telômeros, marcadores das "pontas" dos cromossomos que primordialmente os protegem da ação de exonucleases e evitam que o sistema de reparo de DNA os julgue como quebras na dupla fita de DNA (Shay & Wright, 2019). A circularidade do DNA é tida como uma propriedade distintiva de procariotos e, interessantemente, as moléculas circulares de DNA proeminentemente encontradas em eucariotos são aquelas dos genomas plastidiais (mitocôndrias e cloroplastos), os quais provavelmente tem sua origem associada a endossimbiontes procarióticos (Vosseberg et al., 2024).

Entretanto, ainda no século XX, estudos utilizando microscopia eletrônica que visavam elucidar a organização do DNA nos cromossomos eucarióticos revelaram moléculas circulares não-cromossômicas em células espermáticas de javali e embriões de trigo (Hotta & Bassel, 1965). No mesmo ano, as "duplas minutas" de DNA (DMs) foram identificadas em células neoplásicas malignas em pacientes pediátricos, assim nomeadas pois eram circulares e com tendência a aparecer "em pares" (Cox et al., 1965).

Nas décadas seguintes, esses DNAs circulares foram encontrados em diversos eucariotos, como fungos (Hull et al., 2019), leveduras (Møller et al., 2015), moscas (Stanfield & Lengyel, 1979), plantas (Molin et al., 2020; Peng et al., 2022), e aves (Møller et al., 2020), sugerindo que o seu aparecimento é uma característica compartilhada em células eucarióticas.

Essa prevalência levou os pesquisadores a designar uma nova classe de polinucleotídeos funcionais em eucariotos: os DNA circulares extracromossômicos (eccDNAs, do inglês 'extrachromosomal circular DNAs'). Estas podem ser definidas de maneira geral como moléculas nucleares circularizadas de DNA de fita dupla que são originadas a partir do DNA cromossômico, coexistindo de maneira dinâmica e independente dele nas células, e possuindo grande heterogeneidade em termos de tamanho, sequência e número de cópias (Gaubatz, 1990).

1.1.1. Classificação dos eccDNAs

Ao longo dos anos, autores se referiram a diferentes tipos e tamanhos de eccDNAs com nomenclaturas distintas, causando divergências de interpretação entre pesquisadores. O tamanho máximo que eccDNAs podem ter ainda não foi estabelecido, e de maneira geral acredita-se que possam ter entre centenas até milhões de pares de bases (T. Wang et al., 2021). De acordo com seus tamanhos, (Liao et al., 2020) propôs a seguinte classificação em quatro sub-categorias distintas:

a) "small polydispersed DNA" (spcDNA) (100 bp-10 kb), são eccDNAs que contém genes ou elementos repetitivos e que são encontrados em taxas elevadas em células instáveis, como em tecidos tumorais, como células HeLa e de pacientes portadores de anemia de Fanconi (Stanfield & Lengyel, 1979);

b) círculos teloméricos (T-circles) (múltiplos de 738 bp), como o nome sugere, são derivados dos telômeros e provavelmente derivados do seu processo de encurtamento ao longo de divisões celulares. Suas funções fisiológicas permanecem desconhecidas, embora se sugira que tenham relevância em processos de manutenção dos telômeros (Mazzucco et al., 2020);

c) microDNA (100–400 bp), o tipo mais abundante de eccDNA. Apesar do seu tamanho, podem conter fragmentos de genes (exons e regiões não traduzidas), regiões regulatórias e sítios onde a cromatina está aberta. Foi observado que estes podem ser transcritos, notadamente, em tumores, e podem ser utilizados como biomarcadores (P. Kumar et al., 2017);

d) ecDNA (milhões de bp), esta categoria se refere a DNAs circulares muito grandes, na ordem de megabases. Estas moléculas são características de células de

câncer e podem conter diversos genes completos e oriundos de diferentes cromossomos, sendo pontos de estudo para a amplificação focal no número de cópias de oncogenes e potenciais vetores para o desenvolvimento de cânceres e resistência à drogas (Pecorino et al., 2022). As duplas minutas, observadas nos primórdios dos estudos de DNAs circulares, fazem parte do grupo dos ecDNAs.

Em humanos, a maioria dos eccDNAs são menores que 25 kb e carregam genes completos ou fragmentos distribuídos ao longo do genoma (Møller et al., 2018). A mesma observação vale para camundongos e galinhas (Dillon et al., 2015), reforçando a prevalência dos pequenos DNAs circulares (até 10 kb) em células normais de vertebrados. Sendo assim, passaremos a designar todas as faixas de tamanho como eccDNAs, exceto a classe ecDNA (item "d" acima) que, apesar de ser uma subcategoria, é uma característica importante células neoplásicas e demanda distinção.

1.1.2. Funções e biogênese dos eccDNAs

Enquanto sua ampla incidência nos organismos parece ser indicativa de relevância funcional, diversos aspectos da formação e a função dos eccDNAs ainda são elusivos. Sabe-se, por exemplo, que estas moléculas podem ser transcritas, mas não se sabe detalhes desse mecanismo (Paulsen et al., 2019).

Embora os seus mecanismos de geração também não sejam bem compreendidos, diversos modelos para a sua biogênese vêm sido propostos nos últimos anos, e são geralmente relacionados à reparação de DNA ou à instabilidade cromossômica de maneira geral (R. Li et al., 2022; Zuo et al., 2022).

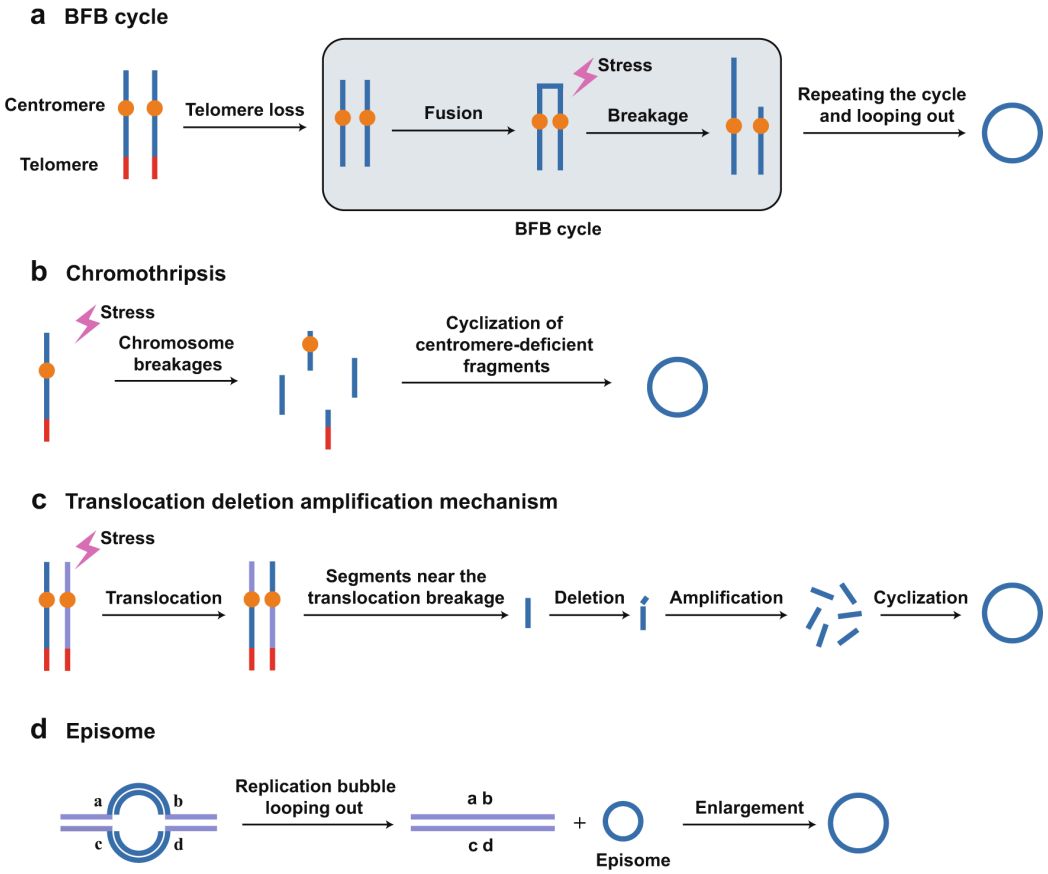

Figura 1 : Diferentes mecanismos de formação de eccDNAs: Yang, L., Jia, R., Ge, T. et al. Extrachromosomal circular DNA: biogenesis, structure, functions and diseases. Sig Transduct Target Ther 7, 342 (2022). Retirado de: https://doi.org/10.1038/s41392-022-01176-8

Dentre os mecanismos propostos, alguns dos mais reconhecidos pela literatura são indicados e detalhados a seguir:

A. Quebra-fusão-ponte (BFB, breakage-fusion-bridge): O ciclo BFB se inicia devido a perda da região telomérica de um cromossomo. Durante a replicação, são geradas duas cromátides irmãs que não possuem telômeros e que se fundem uma à outra, e há a subsequente formação de uma ponte dissentrica na anáfase. Devido a presença de dois centrômeros, a quebra da ponte é dividida, e as células filhas recebem cromátides desiguais. Como as duas cromátides resultantes também não possuem telômeros, o ciclo BFB é continuado. Dependendo da localização e tamanho da quebra, fragmentos resultantes podem ser circularizados (Ling et al., 2021).
B. Cromotripse: Quebra cromossômica seguida pela reorganização aleatória, levando a rearranjos genômicos complexos em determinadas regiões

cromossômicas. Quando o processo de quebra resulta em grupos de quebras de fita dupla, seguidas de reparo de DNA e replicação aberracional pode-se criar condições para a formação de eccDNAs (Shoshani et al., 2021)

C. Translocação-deleção-amplificação: Rearranjos cromossômicos ocorrem próximos ao sítio de translocação. Os segmentos próximos aos breakpoints são amplificados, retidos ou deletados, e podem assim formar eccDNAs (Röijer et al., 2002; T. Wang et al., 2021).
D. Episoma: Ocorre o deslize de um segmento de DNA (*DNA slippage*) durante a replicação ou a formação de um *R-loop* durante a transcrição. eccDNAs são criados a partir das clivagens dessas estruturas e sua subsequente amplificação (Storlazzi et al., 2010), (Carroll et al., 1988).

Diversas características dos eccDNAs já foram elucidadas. Sabe-se, por exemplo, que eccDNAs têm tamanhos e distribuições quantitativas altamente heterogêneas (dos Santos et al., 2023), e que a maioria dos eccDNAs tem menos de 1000 bp (Shibata et al., 2012), e muito poucos possuem mais que 25 kb (Møller et al., 2018). Também se acredita que possam conter sequências capazes de codificar proteínas, elementos transponíveis e regiões regulatórias, que podem ter impacto na expressão gênica (Zuo et al., 2022) .

Devido a heterogeneidade do tamanho, composição, origem e expressão dos eccDNAs, a caracterização dos processos celulares com envolvimento dos mesmos ainda estão por ser descobertos. Algumas de suas funções conhecidas até o momento incluem associações com o câncer, apoptose, ativação do sistema imune e processos de comunicação intercelular (P. Kumar et al., 2017; R. Li et al., 2022; Paulsen et al., 2018; Qiu et al., 2021; Turner et al., 2017). Moléculas de eccDNA também aparentam estar presentes em maiores quantidades em células geneticamente instáveis que em células normais, sendo possível a indução de eccDNAs nas últimas após o uso de um agente carcinogênico (Cohen et al., 1997).

Em plantas, foi reportado que eccDNAs estão relacionados à resistência a herbicidas (Molin et al., 2020). Também demonstrou-se que nematóides fitoparasitas que se alimentam de células de plantas possuem um mecanismo que suprime a internalização de eccDNAs de plantas e impede a troca de material genético via transferência horizontal de genes (Ko et al., 2024).

1.1.3. Potencial terapêutico e diagnóstico dos eccDNAs

O potencial de eccDNAs como biomarcadores para diversas doenças e condições fisiológicas já foi previamente sugerido (Zhu et al., 2017), devido não só ao seu envolvimento nesses processos, mas à sua estabilidade em relação ao DNA linear devido a seu formato estrutural circular e resistência a ação de exonucleases (Sin et al., 2021). No entanto, se estas moléculas também apresentam potencial modulador, como sugere a literatura (Paulsen et al., 2019), suas aplicações em potencial podem se expandir para a abordagem terapêutica.

No estudo do câncer e da senescência, processos biológicos que têm conexão entre algumas de suas características fundamentais (R. Li et al., 2022), há especulação de que moléculas que regulam um destes processos tenham potencial modulador sobre o outro, com a senescência celular levando a uma acumulação de eccDNA nos núcleos das células (Qiu et al., 2021). eccDNAs têm sido associados com a iniciação, progressão da malignidade e evolução heterogênea do câncer. Em células tumorais, também observa-se o aumento da incidência de diversos tipos de eccDNAs contendo oncogenes ou mesmo apenas *enhancers*, o que sugere seu envolvimento na fisiologia anormal destas células, abrindo oportunidades para diagnóstico e terapias (Noer et al., 2022; M. Wu & Rai, 2022).

Sua presença no serum sanguíneo já foi apresentada como uma possibilidade para a biópsia líquida não-invasiva para o diagnóstico de tumores (P. Kumar et al., 2017). Mais recentemente, a análise de eccDNAs no plasma de pacientes com hipertensão pulmonar revelou altas concentrações de eccDNA derivados de uma região no cromossomo 2 associada com a incidência da doença, apresentando grande potencial como diagnóstico não-invasivo (C. Zhang et al., 2024).

Além do potencial diagnóstico, amplificações causadas por ecDNAs também podem servir como marcadores prognósticos. Em pacientes com adenocarcinoma gástrico de cárdia, a detecção de uma amplificação focal do gene ERBB2 advinda de ecDNA foi associada a um prognóstico favorável, enquanto a amplificação do oncogene EGFR parece ter impacto negativo no prognóstico (X.-K. Zhao et al., 2021). Além disso, eccDNAs são mais prevalentes em cânceres agressivos, como

neuroblastoma, com pacientes que possuíam amplificações de ecDNAs tendo prognósticos mais desfavoráveis que aqueles que não as possuíam (Kim et al., 2020).

A indução da apoptose em células animais também é capaz de levar a um aumento na formação de eccDNAs que apresentam potencial de causar uma forte resposta imunológica, com a força da resposta aparentando estar mais ligada à sua estrutura circular que a sequência em si, sugerindo potencial para aplicação biotecnológica como imunoestimulante (Y. Wang, Wang, et al., 2021) .

Assim, o estudo de eccDNAs se mostra extremamente importante em suas possíveis aplicações clínicas e funcionais.

1.1.4. Técnicas para o estudo de eccDNAs

Segundo (T. Wang et al., 2021) diversas técnicas laboratoriais podem ser usadas para se analisar e visualizar a estrutura e localização dos eccDNAs, tais como:

1. Microscopia: Corantes de DNA comuns associados a microscopia óptica podem ser utilizados para observar alguns sinais da sinalização cromossomal durante a fase M do ciclo celular, com alguns deles sendo originários de eccDNAs com peso molecular mais alto. Tecnologias de alta resolução e assistidas por computador, como a microscopia 3D-SIM já foram utilizadas por (S. Wu et al., 2019) para a visualização de análise da arquitetura de ecDNAs. Além disso, tanto a microscopia eletrônica de transmissão quanto a microscopia eletrônica de varredura podem ser utilizados para a visualização de eccDNAs, bem como a combinação da microscopia de luz confocal a sinais de MEV no mesmo campo para análise de imagens.
2. Centrifugação por gradiente de densidade: A centrifugação utilizando gradiente de densidade de cloreto de césio foi utilizada em pesquisas iniciais sobre ácidos nucleicos e identificou eccDNAs grandes em células HeLa (van Loon et al., 1994) mas devido a necessidade de grandes amostras, potencial de destruir estruturas *supercoiled* e a limitada abundância de eccDNAs, esse método é pouco utilizado atualmente.
3. *ATAC-see*: Uma tecnologia de imagem assistida por transposases que utiliza visualização in situ, *cell-sorting* e *deep sequencing* de genomas acessíveis para

identificar elementos visualizados. A combinação de tecnologias de imagem e epigenética dessa técnica com a citometria de fluxo permite uma análise quantitativa automatizada, enquanto a separação celular ocorre em função da acessibilidade da cromatina. A utilização desta técnica (S. Wu et al., 2019) revelou a acessibilidade - possivelmente diferencial - entre a cromatina dos cromossomos e a dos ecDNAs, com esta sendo mais compactada e estando mais acessível na fase G1 da interfase na primeira e menos compactada e mais acessível na metáfase na segunda.

4. Hibridização in situ por fluorescência (FISH): O uso de sondas fluorescentes de eccDNA em amostras celulares pode ser utilizado para a observação da distribuição de eccDNAs nas células (deCarvalho et al., 2018).
5. Eletroforese bidimensional: A eletroforese bidimensional pode ser utilizada para a verificação indireta da estrutura circular dos eccDNAs, bem como para caracterização molecular (Cohen et al., 2008).

1.2. Histórico das Tecnologias de Sequenciamento de DNA

Embora as primeiras investigações iniciais quanto aos eccDNAs tenham sido feitas através de técnicas laboratoriais, estruturais e microscopia eletrônica, o advento das tecnologias de sequenciamento e, principalmente, a popularização de técnicas de sequenciamento de nova geração impulsionou o progresso da pesquisa sobre essas moléculas ao permitir desvendar, com alta resolução, o seu conteúdo genético e distribuição nas células.

O sequenciamento de DNA pela metodologia Sanger ou dideoxi (Sanger et al., 1977), ou técnica de terminação de cadeia, utilizava análogos químicos de desoxirribonucleotídeos (dNTPs), os dideoxinucleotídeos (ddNTPs) marcados radioativamente. Por não possuírem a hidroxila na extremidade 3', e sendo portanto incapazes de se ligar com o 5' fosfato da próxima dNTP, a mistura desses dNTPs marcados resultava em fitas de DNA de todos os tamanhos possíveis sendo produzidas conforme a fita se alonga e os dNTPs são incorporados. Ao realizar quatro reações paralelas, cada uma contendo um tipo de ddNTP, e correr os resultados por um gel de poliacrilamida, era então possível inferir a sequência de nucleotídeos da fita molde ao

se observar as bandas radioativas correspondentes às bases nas canaletas paralelas do gel. A acurácia e robustez do sequenciamento Sanger o tornaram muito popular, e o método sofreu diversas modificações e a eventual automatização.

As máquinas de sequenciamento de primeira geração só eram capazes de obter leituras menores que ~700 bases, com cerca de 67 mil bases sendo sequenciadas por hora. A análise de fragmentos maiores através delas necessitou de adaptações técnicas por parte dos pesquisadores, como o uso do sequenciamento *shotgun*, onde fragmentos de DNA sobrepostos eram clonados e sequenciados separadamente e então "colados" computacionalmente em uma única sequência contígua longa, as *contigs* (Staden, 1979).

A segunda geração de técnicas de sequenciamento, também conhecidas como técnicas de nova geração (NGS, *Next Generation Sequencing*), envolveu o desenvolvimento do pirosequenciamento (Nyrén & Lundin, 1985), que oferecia vantagens em relação ao método Sanger ao permitir o uso de nucleotídeos naturais ao invés de dNTPs e poder ser observado em tempo real. Licenciado pela 454 Life Sciences, os equipamentos produzidos eram capazes de usar a paralelização em massa de reações de sequenciamento, com leituras de até 400 pares de base por fosso e sequenciando cerca de 6.5 milhões de bases em uma hora (Margulies et al., 2005).

Já a tecnologia de sequenciamento Ilumina utiliza o chamado "sequenciamento por síntese", no qual a definição da sequência de bases é dada através do uso de nucleotídeos com bloqueadores fluorescentes, em uma *flowcell*. Após cada etapa da síntese, um computador fotografa o *chip*, e determina qual base foi adicionada a partir do comprimento de onda detectado em cada posição. (Heather & Chain, 2016).

Embora o uso do sequenciamento Ilumina tenha aumentado enormemente a quantidade de bases sequenciadas em paralelo, por utilizar fragmentos curtos de DNA, os dados de sequenciamento de grandes trechos de DNA ainda precisam passar por uma "remontagem" (*reassembly*) de *contigs*. Isso pode ser um desafio e gerar ambiguidade, principalmente em regiões com variações estruturais ou de baixa complexidade (Hu et al., 2021), e se torna um problema ao se trabalhar com eccDNAs, que muitas vezes tem sequências repetitivas e que vem de múltiplas regiões da cromatina. Isso inviabiliza a resolução da junção entre os fragmentos curtos para

reconstruir a sequência contígua de bases do genoma, no contexto do processo denominado como montagem de genoma.

1.2.1. Sequenciamento de leituras longas

Mais recentemente, novas tecnologias de sequenciamento foram criadas, tais como a Pacific Biosciences (PacBio) e Oxford Nanopore, as quais inauguraram a terceira geração de sequenciamento de DNA (Dijk et al., 2018). Estas possuem duas características muito importantes: São capazes de produzir leituras longas, com média de 5.000 bases, mas algumas leituras podem superar 200.000 bases; e são de realizar o sequenciamento, em tempo real, de uma única molécula de DNA, não envolvendo amplificação de sinal por PCR (Polymerase Chain Reaction).

As duas tecnologias, no entanto, ainda sofrem um grande problema: altas taxas de erros (de 5 a 15%) na nomeação de bases (*basecalling)*. Felizmente, a incidência de erros é na maioria das vezes aleatória, o que implica que pode existir um protocolo de autocorreção de bases, caso exista cobertura de sequenciamento suficiente para se obter repetidas leituras independentes da mesma base (Salmela et al., 2017).

O sequenciamento de leituras longas da Oxford Nanopore Technologies (ONT) tem como princípio a passagem de uma fita simples de ácido nucleico por um nanoporo através uma membrana, usando a aplicação de uma voltagem diferencial constante de tal maneira que as moléculas de DNA ou RNA sejam deslocadas do lado negativamente carregado para o lado positivamente carregado. A velocidade desse deslocamento é limitada por uma proteína motora que “catraca” a fita (Y. Wang, Zhao, et al., 2021).

Os nucleotídeos presentes na fita de DNA afetam a resistência elétrica do poro, e de tal forma as medições das mudanças da corrente após o movimento de cada base ao longo do tempo podem ser utilizadas para desvendar a sequência de DNA que passa pelo poro. O sinal elétrico da corrente é o dado bruto (“raw data”) que é coletado pelo sequenciador ONT. A nomeação das bases é realizada pela interpretação dos sinais elétricos brutos para uma sequência de DNA.

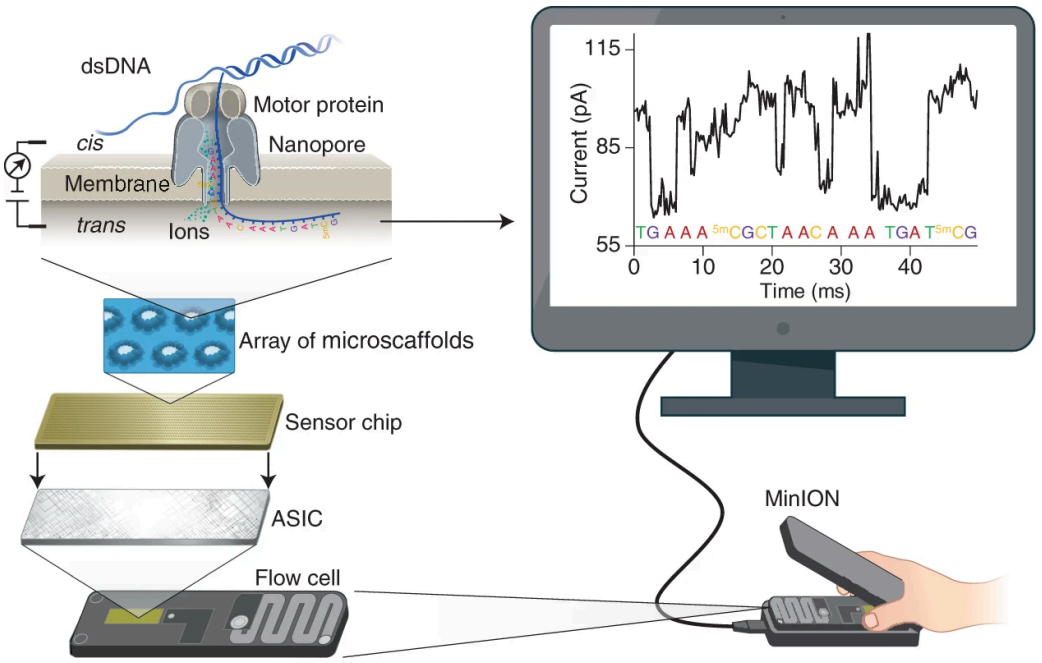

Figura 2: Figura representativa do Sequenciamento Oxford Nanopore. Retirado de: *Wang, Y., Zhao, Y., Bollas, A. et al. Nanopore sequencing technology, bioinformatics and applications.* Nat Biotechnol39,1348–1365(2021) https://doi.org/10.1038/s41587-021-01108-x

A resistência elétrica do poro é determinada pelas bases dos múltiplos nucleotídeos presentes no ponto mais estreito do poro. Há um enorme número de estados possíveis para essas bases: um poro com capacidade para 5 nucleotídeos, como o R9.4, tem 1024 permutações possíveis para um modelo que considera apenas 4 bases de DNA. Quanto mais bases com modificações estão presentes, o número de estados possíveis é ainda maior, de tal forma que o basecalling para equipamentos Oxford Nanopore se torna um problema a ser solucionado por aprendizado de máquina, e é o principal limitador na usabilidade e qualidade do sequenciamento resultante (Wick et al., 2019). Assim, tecnologias de leituras longas, com alto rendimento, são um enorme desafio computacional.

1.3. Enriquecimento e caracterização de eccDNAs por NGS

Por serem moléculas circulares, métodos já estabelecidos para a purificação de plasmídeos procarióticos foram adaptados para enriquecer a fração de eccDNAs a partir de células, tecidos, plasma humano e até mesmo de amostras livres de DNA (cfDNA, *cell free DNA*).

O método *Circle-Seq* de (Møller et al., 2016) merece destaque por ser a base para diversos outros protocolos utilizados para purificar eccDNAs em células eucarióticas. A metodologia envolve as seguintes etapas:

1. Lise celular
2. Purificação do DNA através do uso de cromatografia por coluna, como as utilizadas em *kits* de extração de plasmídeos
3. Remoção de DNA cromossômico linear através do uso de exonuclease, bem como, caso desejado, a remoção de DNA mitocondrial usando enzimas de restrição (PacI) para linearizar esta molécula e torná-la vulnerável à exonuclease
4. Amplificação de DNAs circulares, e
5. Sequenciamento de próxima geração, o qual requer preparo de bibliotecas específicas para a tecnologia selecionada pelo usuário

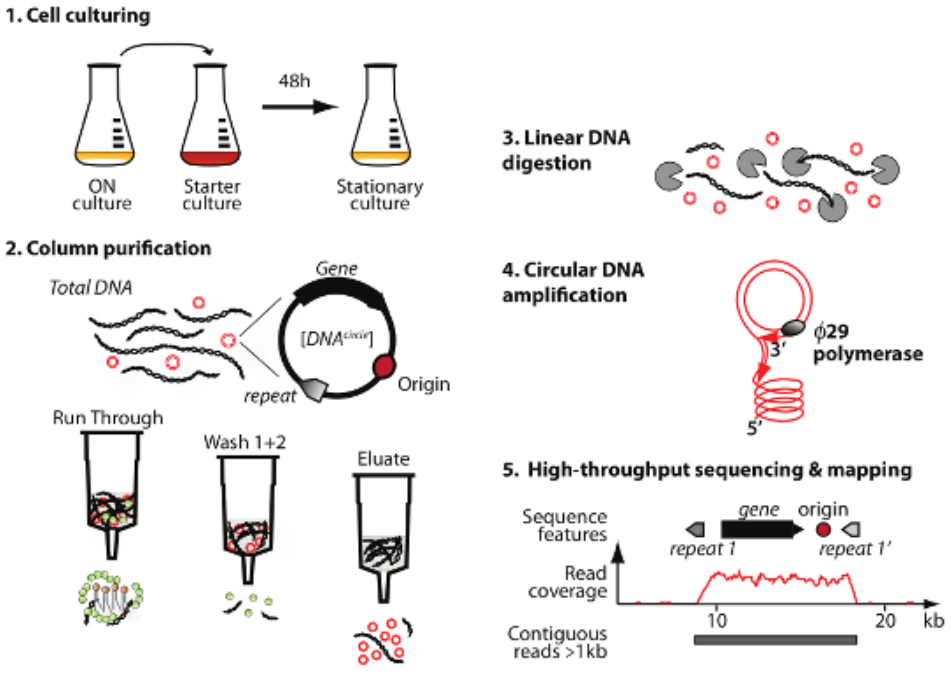

Figura 3: Resumo visual das etapas do protocolo *Circle-seq*: Adaptado de Møller, H. D., Bojsen, R. K., Tachibana, C., Parsons, L., Botstein, D., Regenberg, B. *Genome-wide Purification of Extrachromosomal Circular DNA from Eukaryotic Cells*. J. Vis. Exp. (110), e54239, doi:10.3791/54239 (2016).

Embora outros protocolos já existissem, *Circle-Seq* é um método sensível e que permite a identificação de moléculas maiores. Dois pontos essenciais deste

protocolo são o emprego do tratamento com exonucleases para a remoção de cromossomos lineares e a amplificação de moléculas circulares.

A amplificação dos eccDNAs enriquecidos é feita pela metodologia de "Amplificação por círculo rolante" (RCA, *Rolling Circle Amplification*), um processo enzimático isotérmico no qual um primer curto é amplificado para formar uma fita simples longa a partir de um molde circular e do uso da DNA polimerase do fago φ29 (Ali et al., 2014). Nesse processo, a polimerase continuamente adiciona nucleotídeos ao primer conectado ao círculo molde, criando uma longa fita de DNA contendo uma série de repetições em tandem complementares ao molde (Garafutdinov et al., 2021).

Após a amplificação do eccDNA, a etapa seguinte é o sequenciamento. O protocolo original de *Circle-Seq* sugere a sonificação e fragmentação do DNA para a criação de bibliotecas utilizando a tecnologia Illumina, visto que a mesma pode processar apenas moldes relativamente curtos de DNA.

Atualmente, as tecnologias de terceira geração têm o potencial de permitir a construção das sequências completas de eccDNAs, evitando erros e ambiguidades no mapeamento, podendo ser essenciais para a identificação de ecDNAs (T. Wang et al., 2021) e facilitar seu estudo em geral ao tornar possível a caracterização de toda a sua extensão.

Também vale ressaltar que existe uma profusão de outras técnicas de caracterização de eccDNAs, como o *scEC&T-seq* (*single-cell extrachromosomal circular DNA and transcriptome sequencing*), que analisa o perfil de ecDNAs e sua transcrição em nível de células únicas (Chamorro González et al., 2023) e o *3SEP* (*three-step eccDNA purification*), que incrementa a captura de eccDNAs após a digestão com as exonucleases (Y. Wang et al., 2022).

1.4. Processamento de dados NGS para reconstrução de eccDNAs

Diversos fatores são considerados para determinar a presença de eccDNAs a partir dos resultados de sequenciamento NGS. Mas, fundamentalmente, o processamento depende do tipo de tecnologia de sequenciamento empregada: leituras-curtas (Illumina) ou leituras-longas (Oxford Nanopore).

Esta escolha afeta o transcorrer das análises pois, apesar de gerar mais dados a custos menores, as leituras-curtas não conseguem resolver a estrutura completa dos eccDNAs. Isso já se torna possível com as leituras-longas, a custo de uma maior incidência de erros de nomeação de bases. O que torna o uso das tecnologias de leituras-longas atraente é a possibilidade de se cobrir várias vezes o mesmo círculo de DNA caso seu tamanho seja menor que 10 kb, o que além de tornar a identificação do eccDNA mais robusta, permite corrigir erros de sequenciamento (P. Zhang et al., 2021). Já no caso de leituras-curtas, a presença de regiões de alta profundidade de cobertura indica a amplificação de uma região específica do genoma.

Um fator fundamental para a detecção de eccDNAs é a incidência de leituras divididas, as chamadas *split reads*, que são leituras que são mapeadas em regiões distintas do genoma, mas que devido a circularização da molécula, tornam-se contíguas (Jiang et al., 2023). Uma representação do comportamento dos *split reads* no contexto de moléculas circulares de DNA é representado na Figura 2. A incidência de *split reads* é comumente utilizada como um dos principais indicadores da presença de eccDNAs em métodos bioinformáticos, bem como o número elevado de repetições em tandem - isto é, um grande número de repetições em um padrão constante, o que é considerado indicativo de que a molécula amplificada era, de fato, circularizada.

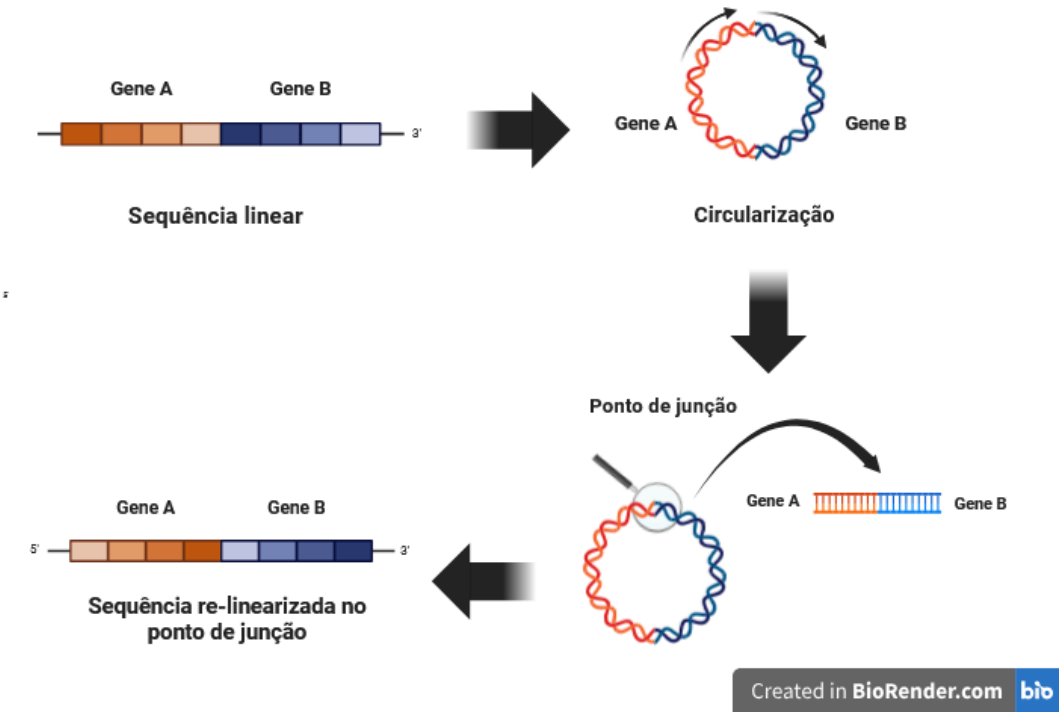

Figura 4: Ilustração representando o mapeamento de leituras ao genoma de referência, demonstrando que a região onde ocorre a circularização do DNA representa uma sequência inédita que não está presente nas sequências lineares dos cromossomos. Quando, durante a leitura há a cobertura do ponto de junção, o efeito no mapeamento é de que a mesma leitura é mapeada em duas regiões distintas. Portanto, esses pontos de junção são indicativos da presença de eccDNAs (Figura criada com *Biorender*[1]).

Ao mesmo tempo que o sequenciamento fornece um levantamento amplo das espécies de eccDNA de uma amostra, ele gera uma copiosa quantidade de dados que requerem um grande esforço computacional para seu devido processamento e contextualização biológica. Assim, forma-se uma lacuna entre a geração dos dados e sua análise, a qual requer mão-de-obra especializada em bioinformática (Y. Wang et al., 2022).

1.5. O uso de *pipelines* em bioinformática

Em função da relevância dos eccDNAs e avanços nas tecnologias de sequenciamento de DNA, verifica-se que ao longo dos anos houve um grande aumento na disponibilidade de dados de sequência que podem fornecer informações relevantes a respeito destas moléculas. Os métodos bioinformáticos baseados nos dados de sequenciamento tornaram-se essenciais para a identificação e análise funcional dos eccDNAs (F. Li et al., 2024).

Com o crescente volume de dados, o grande número de ferramentas computacionais disponíveis e a necessidade de encadeamento de etapas, os fluxos de trabalho (*workflows*) são uma maneira de encadear a saída (*output*) de uma tarefa computacional à entrada (*input*) de outra, em uma sequência de eventos, o que se convenciona chamar de *pipeline* (tubulação) de bioinformática, sendo que *workflow* é uma denominação intercambiável. O uso de *pipelines* permite a padronização de resultados e facilita a execução de processos (Stoudt et al., 2021), sendo essencial para a reprodutibilidade em pesquisa.

Mostra-se pertinente, portanto, o uso de *pipelines* para facilitar o processamento dos dados brutos de sequenciamento de eccDNAs e a interpretação biológica dos dados resultantes. Diversos *pipelines* para a análise de eccDNAs têm surgido nos últimos anos, conforme laboratórios e grupos de pesquisa criam e

[1] https://BioRender.com

relacionam os seus próprios fluxos de trabalho a suas necessidades técnicas específicas.

O tipo de sequenciamento NGS é um dos grandes determinantes para a diferenciação dos *pipelines* para a análise de eccDNAs. Essencialmente, a utilização de leituras-longas (ONT), com milhares de bases por leituras, muitas vezes resolve todo o eccDNA em uma única leitura, enquanto que no caso de leituras-curtas, é necessária uma etapa de montagem (*assembly*) para reconstruir os eccDNAs a partir dos dados de fragmentos de DNA, o que nem sempre resulta na resolução correta da estrutura do eccDNA (Z. Wang et al., 2024).

A seguir, descreve-se diversos *pipelines* disponíveis para o processamento de experimentos de caracterização de eccDNAs, dependendo do tipo de tecnologia usada para gerar os dados de sequenciamento a serem utilizados.

1.6. Pipelines disponíveis para o uso em dados de sequenciamento Ilumina

1.6.1. ECCsplorer

O ECCsplorer (Mann et al., 2022) Utiliza dados de sequenciamento Ilumina de DNA circular amplificado (Circle-seq) para realizar o mapeamento contra o genoma de referência para se detectar dados informativos como a distribuição de leituras, split reads, mapeamento discordante. Depois, sem o uso de um genoma de referência, grupos de leituras de eccDNA amplificado contra dados de controle são usados para detectar o enriquecimento de eccDNA.

1.6.2. CircularDNA_finder

Implementada na linguagem de programação *C*, tornando-a rápida e eficiente, CircularDNA_finder (P. Kumar et al., 2017) utiliza dados de sequenciamento Ilumina enriquecidos para eccDNAs e oferece diversos scripts para a detecção de eccDNAs conforme a necessidade do usuário, dando informações quanto a localização e presença de repetições em tandem características de eccDNAs.

1.6.3. Circle-Map

O Circle-Map (Prada-Luengo et al., 2019) utiliza um arquivo de leituras já alinhadas ao genoma de referência (como um arquivo BAM), e utiliza esses

alinhamentos para detectar onde uma leitura foi separada em dois segmentos (*split-reads*) a fim de detectar os rearranjos cromossômicos característicos de eccDNAs. Caso o alinhador não seja capaz de mapear ambas as *split reads* a leitura - seja pela leitura ser muito curta, ou por alinhar em múltiplos pontos - e as exiba como bases sem mapeamento. O Circle-Map se destaca de outras ferramentas ao ser capaz de re-mapear ambos os segmentos ao utilizar uma abordagem probabilística altamente sensível com o uso de grafos.

1.7. Pipelines disponíveis para o uso em dados de sequenciamento *Nanopore*

1.7.1. *Ecc_finder*

O *Ecc_finder* (P. Zhang et al., 2021) é uma ferramenta robusta e popular implementada em Python e que pode ser utilizada para a montagem e mapeamento tanto de leituras provindas da tecnologia Ilumina quanto ONT, que se propõe a identificar o locus verdadeiro de eccDNAs com base na detecção de *split reads*, leituras discordantes quanto ao mapeamento no genoma e ao número de cópias.

1.7.2. CReSIL

Criada especificamente para o uso em leituras longas de Nanopore, o *CReSIL* (sigla de *Construction-based Rolling-circle-amplification for eccDNA Sequence Identification and Location*) (Wanchai et al., 2022) é uma ferramenta escrita em Python, que detecta regiões de eccDNA em potencial, identifica e verifica eccDNAs a partir desses dados enriquecidos, e realiza a anotação dos eccDNAs identificados.

1.7.3. FLEC (Full-Length eccDNA caller)

O *FLEC* (sigla de *Full-Length eccDNA caller*) (Y. Wang et al., 2022), é uma ferramenta escrita em Python otimizada para dados de Nanoporo, enriquecidas por RCA. As leituras são mapeadas ao genoma de referência e divididas em “subleituras”, e usa-se a ordem e localização do mapeamento para se detectar eccDNAs.

1.7.4. Cyrcular-calling

Com seu principal programa implementado na linguagem *Rust*[2], o *Cyrcular-calling* (Tüns et al., 2022) utiliza-se do WMS *Snakemake* e uma abordagem

[2] https://www.rust-lang.org/

que se utiliza da estatística bayesiana para determinar eccDNAs a partir da construção de grafos. Produz relatórios ricos de resultados que podem ser navegados em página de Web, simplificando a análise para o usuário final.

1.8. Outras ferramentas comuns em bioinformática

1.8.1. *Workflow Management Systems* e *Nextflow*

Conforme as necessidades computacionais e de análise de dados do usuário, os *workflows* podem se tornar cada vez mais complexos enquanto ainda precisam ter performance eficiente, permitir a computação paralela, ser compartilháveis e ter resultados reprodutíveis.

O uso de Sistemas de Gerenciamento de Fluxos de Trabalho (*Workflow Management Systems*, WMS) tem por finalidade facilitar o uso e criação de *pipelines*, bem como iniciativas para a criação de Linguagens Comuns para *Workflows* (como a *Workflow Description Language*, WDL, e a *Common Workflow Language*, CWL), que tem como intuito formalizar a descrição das etapas dos *pipelines*.

A escolha de WMS é multifatorial. Ao comparar múltiplas WMS entre si, (Larsonneur et al., 2018), relatou que o WMS escolhido condiciona a facilidade do desenvolvimento, compatibilidade e compartilhamento da *pipeline*, com alguns dos mais populares sendo Cromwell-WDL (Voss et al., 2017), Pegasus-mpi-cluster (Deelman et al., 2019), Toil-CWL (Vivian et al., 2017), Snakemake (Mölder et al., 2021) e Nextflow (Di Tommaso et al., 2017). Os dois últimos são os mais popularmente utilizados dentro do campo da bioinformática, possuindo ampla documentação, comunidade ativa e sendo capazes de integrar o uso da conteinerização (definido abaixo), tornando os workflows facilmente compartilháveis entre comunidades e diferentes ambientes de computação.

Uma das principais diferenças entre WMSs existe quanto aos algoritmos usados para se inferir as dependências entre as etapas. Enquanto o Snakemake se baseia na abordagem “make-like” que precisa pré computar o DAG (Grafo Acíclico Direcionado, um conceito matemático de grafo sem ligação entre variáveis, usado para se definir dependências entre programas) por exemplo, o Nextflow usa uma abordagem “top-down”, que é independente do cálculo de um DAG para fazer a descoberta das suas dependências - uma etapa acontece após a outra devido à

existência os seus inputs.

Além disso, a linguagem de implementação do WMS também é capaz de afetar os recursos computacionais necessários: aquelas que utilizam Java tendem a utilizar mais memória RAM, enquanto aquelas baseadas em Python tendem a utilizar mais recursos da CPU, apresentando um maior número de mudanças de contexto por segundo. Assim, também é necessário considerar o ambiente de trabalho na decisão.

Para esse trabalho, a WMS de escolha foi o Nextflow. O Nextflow é uma linguagem específica de domínio (DSL, *Domain Specific Languag*e) que facilita a composição de scripts e permite o uso de *pipelines* já existentes escritas em qualquer linguagem de programação. Usado principalmente nas ciências biológicas, o Nextflow é uma linguagem baseada em Groovy, que por sua vez é uma linguagem baseada em Java, e que usa o paradigma computacional de "fluxo de dados" (dataflow), de forma que as tarefas são iniciadas automaticamente quando os dados necessários são recebidos pelos canais de input. A definição do workflow e a paralelização da execução são implícitas, e cada tarefa é isolada em um contexto de execução próprio. Além disso, o Nextflow possui integração com plataformas de repositórios de software, como o GitHub[3], suporte nativo para computação em nuvem (Di Tommaso et al., 2017), bem como um repositório de ferramentas já existentes, o NF-core (Ewels et al., 2020)[4] .

1.8.2. Gerenciadores de Pacotes e Conda

Programas modernos muitas vezes são criados e construídos a partir de partes já existentes, e são dependentes em outros serviços e aplicações para o seu funcionamento. Cada um desses componentes possui o seu próprio conjunto de dependências que podem entrar em conflito com as dependências de outros componentes, tornando a instalação manual de todo o software necessário onerosa.

O processo de empacotamento de programas surgiu como uma maneira de facilitar o processo da instalação de softwares, através do uso de ferramentas conhecidas como gerenciadores de pacotes. Gerenciadores de pacotes podem ser nativos ao sistema operacional, como o *APT*[5] do Linux Debian, enquanto outros

[3] https://github.com/
[4] https://github.com/nf-core/
[5] https://wiki.debian.org/AptCLI

precisam ser baixados e instalados pelo usuário. Alguns são agnósticos quanto a linguagem (compatíveis com diversas linguagens de programação, como o Conda[6]), enquanto outros oferecem suporte a uma linguagem de programação em específico (como o Pip[7], que oferece suporte a pacotes em Python). Pacotes contém, de maneira geral, o software a ser instalado, uma lista das dependências necessárias e metadados com descrições do software contido neles.

O uso de um gerenciador de pacotes costuma ser simples para o usuário final: o usuário direciona o gerenciador quanto ao software a ser instalado, e o gerenciador realiza todas as etapas necessárias para fazer a instalação do software e das suas dependências. O gerenciador busca o pacote apropriado em um repositório, e utiliza os metadados das instruções do pacote para então completar a instalação. Para a maioria dos pacotes, isso inclui verificar se todas as dependências estão instaladas e, caso contrário, instalar aquelas pendentes (Alser et al., 2024).

Apesar do uso de gerenciadores de pacotes facilitar enormemente a instalação de programas, ainda existem dois problemas comuns a serem contornados: a ocorrência de dependências circulares, que ocorre quando pacotes dependem mutuamente um no outro; e os conflitos de versão, que ocorrem quando dois ou mais pacotes precisam de versões diferentes da mesma dependência.

A estratégia para evitar tais erros utilizada pelo gerenciador de pacotes Conda envolve permitir que o usuário construa ambientes separados manualmente, fazendo com que todas as dependências necessárias para um determinado fim sejam mantidas separadas das dependências de pacotes em outros ambientes.

O gerenciador Conda é comumente utilizado nas ciências biológicas, possuindo um repositório central para as ciências “ômicas”, o Bioconda (Grüning et al., 2018). O uso do Bioconda não só torna essas ferramentas facilmente acessíveis para os usuários, mas também permite que administradores mantenham seus pacotes universalmente consistentes ao estabelecer “guidelines” a serem seguidas para o hosting no repositório. Diversas ferramentas utilizadas pelas ciências ômicas e que podem ser utilizadas para o estudo de eccDNAs estão disponíveis no Bioconda.

[6] https://github.com/conda/conda
[7] https://github.com/pypa/pip

1.8.3. Conteinerização de Software e Docker

Um contêiner é uma unidade de software que "empacota" o código e todas as suas dependências isoladamente, para que programas possam ser executados independentemente do ambiente de computação utilizado. O usuário ou programa que utiliza o contêiner tem acesso a um sistema de arquivos completamente separado daquele do sistema operacional do hospedeiro, podendo este ser criado já com todas as dependências necessárias e passado para novas máquinas sem a necessidade de alterações no código ou novas instalações (Choi et al., 2023). Ou seja, é possível "empacotar" uma série de softwares desenvolvidos para Linux e executá-los em uma máquina Windows de maneira transparente.

Diferente de máquinas virtuais, que oferecem uma versão "abstrata" do hardware de uma máquina física (como CPU, RAM, e armazenamento), contêineres funcionam como instâncias de um software junto com suas dependências que podem correr dentro de uma máquina, sendo essa real ou virtual. Tudo aquilo que existe dentro de um contêiner é descrito utilizando uma imagem de contêiner, um arquivo que inclui as bibliotecas e as dependências do contêiner. De maneira geral, essas imagens são iniciadas por um gerenciador de contêineres, um aplicativo que coordena os componentes necessários para a corrida das imagens do contêiner ao criar um ambiente isolado para cada imagem sobre o sistema operacional do host. Assim, imagens de contêineres tem alta portabilidade, e o isolamento dos sistemas de arquivos resolve tanto os problemas de dependências incompatíveis quanto de dependências circulares.

Embora plataformas de conteinerização já existissem desde o início dos anos 2000, o surgimento da plataforma Docker[8], em 2013, permitiu que o usuário acesse ferramentas para a criação e uso de contêineres de maneira fácil, contribuindo para a sua enorme popularização desde então. Atualmente, o Docker ainda é a plataforma de conteinerização mais popular e que oferece o maior número de imagens de contêineres disponíveis para *download*, apesar da crescente popularidade de plataformas como o Singularity, especialmente para aqueles que usam servidores com configurações de segurança restritivas.

[8] https://www.docker.com/

1.9. Vesículas Extracelulares

As vesículas extracelulares (EVs, do inglês *extracellular vesicles*) são vesículas membranosas derivadas de células e que carregam diversos tipos de moléculas bioativas, agindo como veículos de sinalização tanto em processos homeostáticos normais quanto como consequência de processos patogênicos (van Niel et al., 2018). EVs são secretadas e liberadas nos tecidos adjacentes por diversos tipos de células, e em humanos, uma proporção notável delas é oriunda de células tronco, e estão enriquecidas em fluidos corporais. As EVs são um grupo altamente heterogêneo, e EVs clássicas podem ser divididas em microvesículas (40-1000 nm), que tem sua origem na evaginação da membrana plasmática, exossomas (40-150 nm), que são secretados via exocitose de vesículas intraluminais através da fusão da membrana plasmática com a dos endossomos multivesiculares que as contém, e os corpúsculos apoptóticos, antes vistos como meros resíduos de células que sofreram apoptose mas hoje considerados como tendo um papel importante na comunicação intercelular (Yu et al., 2023).

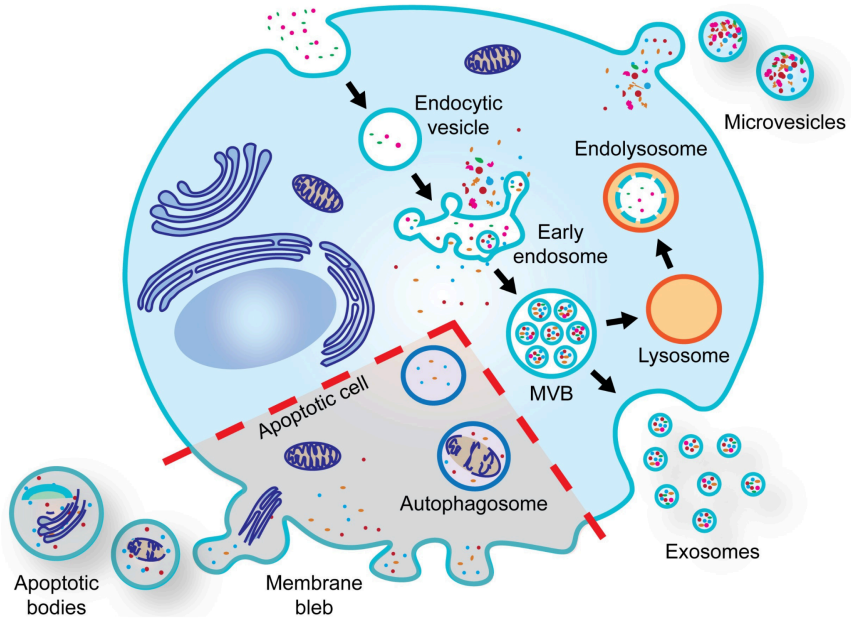

Figura 5: Ilustração da biogênese de vesículas extracelulares clássicas. Retirada de Shahi, S., Kang, T., & Fonseka, P. (2024). *Extracellular Vesicles in Pathophysiology: A Prudent Target That Requires Careful Consideration.* Cells, 13(9), 754. https://doi.org/10.3390/cells13090754

Os papeis das EVs em transporte e comunicação, bem como sua capacidade de encapsular e preservar as características moleculares das suas células-mães, têm

tornado-as candidatas promissoras para o uso de agentes terapêuticos, *drug delivery systems* e como biomarcadores para doenças como o câncer, com a caracterização de seu conteúdo em fluidos biológicos como a urina e sangue tendo potencial papel diagnóstico e prognóstico não invasivo (M. A. Kumar et al., 2024), embora a sua eficácia para tais aplicações tenha permanecido elusiva, com perturbações nos processos de biogênese ou secreção de EVs sendo capazes de afetar a integridade celular e a manutenção da homeostase (Shahi et al., 2024).

Além disso, alguns dos estresses primários considerados como os primeiros sinais do processo de envelhecimento (como instabilidade genômica, erosão telomérica, alterações epigenéticas e proteostase defeituosa) estão associados com uma maior liberação de EVs capazes de espalhar sinais de senescência às células que as recebem, de forma que a alteração na secreção de EVs em si pode ser considerada um marcador para o envelhecimento (Manni et al., 2023). Adicionalmente, a habilidade de EVs de células progenitoras funcionais de promover a regeneração de tecidos sugere que EVs de células tronco (B. Zhang et al., 2016) poderiam ser capazes de reverter ou ao menos prolongar alguns dos processos celulares associados ao envelhecimento (Robbins, 2017).

2. Objetivos Gerais

Com a criação de um protocolo computacional facilmente instalável, capaz de utilizar múltiplos *pipelines*, e com *output* unificado, busca-se diminuir a lacuna entre a geração de dados de eccDNAs e seu processamento, facilitar a execução das análises específicas de caracterização das sequências de eccDNAs *in silico*, esperando-se assim melhor entender a participação dos eccDNAs nos processos biológicos futuramente.

2.1. Objetivos Específicos

- Criação de um *pipeline* unificado contendo quatro *pipelines* de análise de eccDNAs já existentes através do uso do Nextflow, permitindo o uso de um único arquivo de configuração;
- Criação de imagens Docker para os *pipelines* a fim de se permitir a instalação rápida e facilitar a execução de cada um dos *pipelines*;
- Uso do Intervene para a detecção daqueles círculos de eccDNA que são encontrados por todas as *pipelines*;
- Utilização dos *pipelines* desenvolvidos para caracterizar o perfil de eccDNAs encontrados em fibroblastos humanos e vesículas extracelulares.

3. Materiais e métodos

3.1. Recursos computacionais

Para o desenvolvimento local de scripts e testes da *pipeline*, foi utilizada uma máquina que possui CPU Intel i7-3770 (8) @ 3.900GHz, GPU Intel IvyBridge GT2 [HD Graphics 4000] e 16 GB de memória RAM.

Para análises que necessitam de maior capacidade computacional, foi utilizado um servidor Lenovo SR650 com 2 processadores Intel® Xeon® Platinum e 64 GB de memória RAM.

3.2. Softwares utilizados

Foram utilizadas quatro *pipelines* de detecção de eccDNAs a partir de leituras de Oxford Nanopore Technology: *Ecc_finder*[9](v1.0.0) (função *map-ont*)**,** *CReSIL*[10](v1.0.0), *FLEC*[11] (v1.0), *Cyrcular-calling*[12](v2.1.0). Além disso, foi utilizado o software Intervene (v0.6.5). O intervene (Khan & Mathelier, 2017) é uma ferramenta baseada no bedtools, utilizada para realizar a interseção e visualização de múltiplos conjuntos de arquivos de regiões em formato BED. A WMS de escolha para para todo o desenvolvimento foi o Nextflow (v24.03.0). Além disso, para a instalação dos *pipelines* foram criados ambientes utilizando a ferramenta Conda (v23.3.5), bem como contêineres utilizando o Docker Engine (v1.13+). Para a análise estatística posterior dos eccDNAs consenso, foi utilizado o software *PAST*[13] (v4.08) (Hammer et al., 2001).

3.3. Genoma de referência

O genoma de *Homo sapiens* de referência utilizado foi a versão GRCh38, distribuição 110, em formato FASTA, *soft-masked* (elementos repetitivos aparecem em letra minúscula), obtido a partir da plataforma *Ensembl* (Birney et al., 2004). O genoma indexado utilizado pelos *pipelines* foi gerado a partir dele, através de duas ferramentas diferentes, sendo essas Samtools (H. Li et al., 2009) (v1.15) e Minimap2 (H. Li, 2018) (v2.22).

[9] https://github.com/njaupan/ecc_finder
[10] https://github.com/visanuwan/cresil
[11] https://github.com/YiZhang-lab/eccDNA_RCA_nanopore
[12]https://github.com/snakemake-workflows/cyrcular-calling
[13] https://palaeo-electronica.org/2001_1/past/issue1_01.htm

Os arquivos de anotação de genes e de ilhas (CPG) necessários para o uso do *CReSIL* (v1.0.0) foram obtidos do UCSC Table Browser[14], com o output em formato BED (*Browser Extensible Data*), utilizando-se os seguintes filtros:

"*Clade: Mammal, Genome:Human, Assembly:Dec. 2013 (GRCh38/hg38), Group: All Tracks, Track: CpG Islands, Table: cpgIslantExt, Region: Genome*". Para o arquivo BED para a anotação de CPG.

"*Clade: Mammal, Genome:Human, Assembly:Dec. 2013 (GRCh38/hg38), Group: Genes and Gene Predictions, Track: Old UCSC Genes, Table: knownGeneOldV45*" para o arquivo BED para a anotação de genes. Este é arquivo para GRCh38 resultante cujo formato que mais se assemelha ao daquele presente no repositório do *CReSIL*(v1.0.0).

O arquivo BED de anotação de sequências de baixa complexidade e repetições intercaladas é gerado pelo programa RepeatMasker e foi obtido a partir do UCSC. Os filtros utilizados foram: "*Clade: Mammal, Genome:Human, Assembly:Feb.2009 (GRCh37/hg19), Group: Repeats, Track:RepeatMasker, Table:rmsk"*. Até o presente momento, não existe uma versão GRCh38 disponível publicamente. O arquivo de anotação de elementos transponíveis também pode ser obtido diretamente do website do Repeat Masker em formato compactado, o que é feito de forma automática pelo *Cyrcular-calling*(v2.1.0).

3.4. Dados de sequenciamento

Foram utilizados dados de sequenciamento de *Oxford Nanopore MinIon* de eccDNAs isolados a partir de três amostras de fibroblastos primários da derme de mulheres adultas (designadas como grupo Donors, ou Controle), bem como de vesículas extracelulares isoladas dos meios de cultura dessas células (designadas como grupo Vesículas, ou EVs), obtidos em parceria com o Dr. Robert Pogue da Universidade Católica de Brasília e Bianca Simonassi-Paiva (Technological University of the Shanon).

O protocolo utilizado para o enriquecimento de eccDNAs das amostras foi adaptado de (Møller et al., 2018). Em termos gerais: Isolamento do DNA, utilizando

[14] https://genome.ucsc.edu/cgi-bin/hgTables

Kit de Isolamento de Plasmídeos de DNA (*DNA Plasmid Isolation Kit - T1010, New England Biolabs,* EUA); tratamento com enzima de restrição (endonuclease *MssI (pmel) - Thermo Fisher Scientific);* seguido pela degradação de DNA linear através do uso de exonuclease V *(M0345 - New England Biolabs)*. As amostras resultantes foram amplificadas por RCA, preparadas para o sequenciamento e criação de bibliotecas (kit LSK-109 e protocolo NBD104 - ONT, GB*)* e então sequenciadas utilizando o equipamento *MinIon Mk1B,* ONT. O *basecalling* foi realizado utilizando o software *MinKNOW*(v23.11.5) (Simonassi-Paiva B., Manuscrito em Preparação).

4. Resultados e Discussão

Considerando-se a relevância biológica dos eccDNAs e o crescente interesse na caracterização do seu conteúdo genético, principalmente com a incorporação do uso de amostras advindas de tecnologias de terceira geração, múltiplas ferramentas para a sua análise têm surgido e estão à disposição da comunidade científica. De maneira paradoxal, porém, a instalação e execução dessas ferramentas é frequentemente dificultosa, e atualmente não existe uma ferramenta única que seja amplamente utilizada e tida como padrão para o estudo de eccDNAs (Deng & Fan, 2024).

Assim, embora a grande variedade de programas e *pipelines* pareça, à primeira vista, vantajosa, ela trás consigo uma nova indagação: Qual ferramenta se deve utilizar? Uma abordagem a ser considerada a fim de se evitar a interpretação biológica feita com base em um único algoritmo é o uso de múltiplas ferramentas, seguida pela criação de um consenso, tornando assim possível que resultados sejam comparados entre programas para as mesmas amostras antes de análises subsequentes.

Embora a reprodutibilidade dentro da programação pareça simples, ela é frequentemente complexa quanto a sua aplicação, e o campo da bioinformática não é uma exceção: Em uma análise sistemática de códigos escritos em *R*[15] utilizados na pesquisa científica disponibilizados no *Harvard Dataverse*, apenas 26% desses foram passíveis de execução sem erros (Trisovic et al., 2022). Esse não é um problema recente; em um estudo realizado em 2009 envolvendo análises bioinformáticas de dados de microarranjo, apenas 2 dos 18 artigos analisados puderam ter seus resultados reproduzidos, com os principais razões sendo a ausência da disponibilidade dos dados, discrepâncias devido a anotação incompleta ou falta de especificação das etapas de processamento de dados utilizadas (Ioannidis et al., 2009).

O presente trabalho buscou, assim, não apenas solucionar problemas de reprodutibilidade através da conteinerização de *pipelines* para o estudo de eccDNAs já estabelecidos para o uso em ONT, melhorando assim a sua usabilidade (Bolchini et al., 2009), mas também a criação de uma pipeline integrativa, capaz de unificar a entrada e saída de dados de maneira conveniente para a análise final, apelidada de eccPOP (*eccDNA Pipeline of Pipelines*).

[15] https://www.r-project.org/

4.1. O uso de dados e usabilidade

Apesar de princípios para melhorar a reprodutibilidade e acessibilidade na pesquisa computacional tenham sido cada vez mais aplicados, como os princípios FAIR (Wilkinson et al., 2016), esse gargalo ainda persiste. A sigla FAIR se refere a quatro aspectos que descrevem características essenciais dos assets digitais usados na pesquisa: *Findability* (Facilidade de encontrar dados e metadados, tanto para humanos quanto para máquinas), *Accessibility* (Capacidade de acessar os dados necessários, incluindo autenticação e autorização), *Interoperability* (Capacidade da integração de dados com outros dados, e a utilização com aplicações ou workflows de análise, armazenamento e processamento desses dados) e *Reusability* (Otimização da reutilização dos dados).

Enquanto algumas das ferramentas utilizadas e seus respectivos dados para teste tivessem documentação apropriada, algumas das ferramentas e artigos correspondentes ainda apresentam problemas de acessibilidade. O principal exemplo foi o software *CReSIL* (v1.0.0). Embora a sua instalação através do Conda tenha sido simples, os dados de amostra e de referência que são oferecidos pelo grupo de pesquisa em um repositório público tem descrição excessivamente genérica e não foram encontrados em outros sites. Enquanto no repositório do GitHub se afirma que os dados foram obtidos do UCSCTableBrowser, existem diversas opções para a formatação e obtenção dos dados nessa plataforma, e nenhuma das opções delas corresponde exatamente aos arquivos disponibilizados. Os arquivos BED disponibilizados aparentam ter sofrido alguma modificação que não foi documentada, tornando-se um problema não só para aqueles que gostariam de reproduzir o estudo do artigo original a partir dos dados brutos, mas também para aqueles que desejam utilizar o próprio genoma de referência e são incapazes de obter a mesma formatação da fonte informada. Além disso, a documentação e artigo originais do *CReSIL*(v1.0.0) afirmam que foi utilizada a versão GRCh37 (a atual é a GRCh38) do genoma humano, mas pesquisadores que desejam utilizar versões mais atualizadas do genoma não conseguem reproduzir a formatação em suas próprias versões dos arquivos. Apesar disso, *CReSIL*(v1.0.0) é considerado um dos melhores algoritmos para a análise de leituras longas de eccDNAs (Fang et al., 2024; Gao et al., 2024).

Em dois *workshops* organizados pelo instituto nacional de saúde dos EUA em

2019, dez equipes tentaram reproduzir cinco estudos de bioinformática, mas nenhum foi capaz de reproduzir os resultados publicados, com os principais problemas sendo a ausência de dados, *softwares* e ferramentas, descrições inadequadas, fluxos de trabalho inadequadamente descritos ou complexos de se seguir. Diversas questões interessantes puderam ser levantadas: Por exemplo, julgou-se que parte da dificuldade advinha das diferentes interpretações do que significa a reprodutibilidade: Dados brutos ou processados? Reutilização de Scripts ou a re-construção destes? Recriar completamente o ambiente de computação original ou um que esteja próximo? Além disso, diversas equipes entraram em contato com os autores dos artigos e obtiveram respostas rápidas e esclarecedoras, sugerindo que esses problemas raramente advém de má fé (Zaringhalam & Federer, 2020).

Considerando evitar os mesmos problemas para aqueles que gostariam de utilizar os mesmos arquivos de referência, o arquivo de configuração do Nextflow disponibilizado contém as URLs das fontes de obtenção de cada conjunto de dados de referência utilizados como comentários. Além disso, o *eccPOP* utiliza um único arquivo de configuração do Nextflow, onde devem ser adicionados os caminhos para todos os arquivos necessários para a execução de todas as pipelines.

Outra questão a ser resolvida quanto ao uso de dados é a facilidade da execução da pipeline para múltiplos arquivos de entrada para o processamento em larga escala. Embora a execução cíclica de tarefas possa ser trivial para aqueles familiarizados com programação, ela pode ser um fator limitante para os que não são. A grande maioria dos *pipelines* utilizados tem a sua execução em linha de comando com apenas um arquivo, com a exceção notável sendo o *Cyrcular-calling*(v2.1.0)l.

O *eccPOP* não só integra a execução de todos os *pipelines*, mas também se aproveita da paralelização proporcionada pelo Nextflow ao utilizar um arquivo TSV (separado por Tab) para permitir multiplos arquivos de entrada, contendo em seu primeiro campo o nome do grupo de amostras, que será utilizado para nomear os diretórios onde os resultados serão publicados, nome das respectivas amostras em seu segundo campo, que será utilizado para nomear os resultados, e os caminhos para a localização dos arquivos das amostras de entrada em seu terceiro campo. As amostras de entrada devem estar em formato FASTQ, que é o formato dos arquivos resultantes

do processo de *basecalling* feito por ONT, e que reúne não só as sequências, mas informações sobre qualidade.

Todos os parâmetros de configuração podem ser modificados permanentemente ao se editar o arquivo de configuração, ou diretamente pela linha de comando para alterações pontuais.

4.2. Instalação das *pipelines*

Embora o uso de ferramentas como Conda e Docker tenha simplificado a instalação de muitos programas utilizados em bioinformática, a instalação muitas vezes continua sendo uma barreira considerável para aqueles pesquisadores que não tem familiaridade com informática (Kulkarni et al., 2018).

Cada uma dos *pipelines* utilizados foi inicialmente instalada conforme as instruções de suas respectivas documentações do GitHub. Para o *Ecc_finder*(v1.0.0), foi necessário alterar o arquivo de configuração de dependências (formato YAML, *Yet Another Markup Language*, uma linguagem de serialização de dados) original, pois o ambiente continha mais dependências especificadas que o estritamente necessário, o que ocasionava em conflitos de versão e a eventual falha na instalação.

Esse é um problema relativamente comum, e que ocorre ao se utilizar a função [conda env export] para a geração automática do arquivo YAML de um ambiente Conda sem se especificar a opção [--from-history]. Essa opção garante que apenas os pacotes que foram explicitamente instalados sejam exportados, evitando que pacotes adicionais que precisarão ser resolvidos quanto a suas dependências e que podem ser incompatíveis com plataformas diferentes sejam adicionados ao arquivo.

Além disso, o *Ecc_finder*(v1.0.0) tem como dependência o software de detecção de repetições em tandem *TideHunter*[16], que por sua vez depende do software *abPOA*[17] para o alinhamento de sequências múltiplas. O arquivo binário usado pelo Conda para a instalação do abPOA é incompatível com CPUs mais antigas, fazendo com que a execução do *Ecc_finder*(v1.0.0) falhe[18]. A criação da imagem Docker para o *Ecc_finder*(v1.0.0) resolveu esse problema ao utilizar um arquivo binário

[16] https://github.com/Xinglab/TideHunter
[17] https://github.com/yangao07/abPOA
[18] https://github.com/Xinglab/TideHunter/issues/19

pré-compilado do *TideHunter* como parte do ambiente e permitiu o uso do *Ecc_finder*(v1.0.0) em CPUs de arquitetura mais antiga.

Para a instalação e uso do *Cyrcular-calling*(v2.1.0), embora as instruções da documentação oficial indicassem inicialmente que é possível efetuar a instalação através da ferramenta Snakedeploy[19], os diretórios com as regras e documentos de configuração necessários para o uso da *pipeline* não eram corretamente instaladas. Assim, a instalação do workflow funcional necessita da instalação "manual", através do download do repositório do GitHub, antes de seguir as outras instruções da documentação.

Para o *FLEC*(v1.0) e *CReSIL*(v1.0.0), as instruções de instalação funcionavam corretamente e não foram necessárias modificações.

4.3. Criação de imagens utilizando Docker e Conda em conjunto

Como todos os programas a serem testados utilizam ou podem utilizar o Conda para a sua instalação e ativação, com dependências em dezenas de ferramentas com versões que precisam ser compatíveis a serem instaladas, julgou-se mais eficiente realizar a instalação dentro das imagens Docker através do uso do Conda, permitindo não só a resolução mais rápida, mas a ativação dos binários e adição dos programas ao caminho padrão de executáveis do Linux (variável PATH) é automática.

Embora a criação das imagens utilizando o Conda tenha sido facilitada, as imagens tornaram-se excessivamente grandes. Isso ocorre pois não só o Conda adiciona pacotes baixados ao cache (uma subdivisão da memória que armazena dados frequentemente usados para acesso rápido) automaticamente, mas o próprio ambiente base onde as ferramentas do Conda são instaladas ocupa uma grande quantidade de espaço.

Enquanto o ambiente base do Conda é necessário durante a instalação dos pacotes, quando o código já está em execução, ele não é mais estritamente necessário. Assim, o Conda pode então ser removido da imagem após a instalação. Para tal propósito, seguindo recomendações descritas por (Turner-Trauring, 2020), foi utilizado o software *conda-pack(v0.8.0)*[20].

[19] https://github.com/snakemake/snakedeploy
[20] https://github.com/conda/conda-pack

O *conda-pack* permite que o ambiente Conda seja empacotado em um ambiente "standalone", sem as ferramentas do Conda presentes na base. Assim, após empacotar o ambiente de tal maneira, é possível copiá-lo para uma nova imagem Docker que contém apenas esse ambiente, essencialmente tornando-o um diretório contendo os programas necessários, reduzindo drasticamente o tamanho da imagem resultante e acelerando o tempo de construção das imagens. Isso é possível com o uso de "multi-stage builds", a construção em múltiplos estágios, do Docker. O primeiro estágio, o estágio de construção (*Build Stage*) utiliza uma imagem mais completa como base e com todos os elementos necessários para o desenvolvimento e empacotamento do programa de interesse. O segundo estágio, o estágio de execução (*Runtime Stage*) utiliza uma imagem mínima e que resulta em uma imagem menor, contendo apenas aquilo que é necessário para a execução do programa já compilado.

Para todas as imagens criadas dessa maneira, foi utilizada a imagem continuumio/miniconda3[21] como *Build*, e após o uso do conda-pack, foi utilizada a imagem debian:bookworm[22] como *Runtime*.

A ativação de ambientes Conda é essencial para que o software dentro de tais ambientes funcione corretamente. A ativação de ambientes Conda adiciona programas ao caminho padrão de localização de programas executáveis no Linux (usando a variável de ambiente PATH) e inicializa scripts de ativação contidos no ambiente Conda. Esses scripts de ativação podem criar variáveis de ambiente arbitrárias, que podem ser necessárias para a sua operação correta. Isso se torna complexo quando utilizamos ambientes Conda dentro de contêineres, pois o ambiente Conda pode acabar não sendo ativado quando o contêiner é iniciado, a depender de como o contêiner é inicializado.

Para que os ambientes Conda sejam devidamente ativados dentro do contêiner quando este é inicializado através do Nextflow, foram seguidas as recomendações descritas por (Allain et al., 2022).

Não foi possível finalizar a criação de uma imagem Docker funcional para o *Cyrcular-calling*(v2.1.0). A solução mais simples foi a utilização do ambiente Conda e opções da linha de comando que especificam os arquivos de entrada, de tal maneira

[21] https://hub.docker.com/r/continuumio/miniconda3
[22] https://hub.docker.com/_/debian

que o caminho para o caminho do arquivo de entrada corresponda corretamente ao do arquivo esperado pelo *Cyrcular-calling*(v2.1.0).

4.4. Subworkflows

Após a criação de *workflows* para a corrida de cada um dos *pipelines* conforme suas documentações, utilizando seus parâmetros padrão para especificação de características dos eccDNAs a serem detectados, cada um dos workflows foi subsequentemente formatado como *subworkflows*. O uso da estrutura de *subworkflows* é específica da segunda versão da sintaxe do Nextflow (DSL2) e seu uso faz parte das recomendações descritas para a publicação no *NF-core* para a combinação de ferramentas usadas para uma determinada etapa de análise da pipeline.

A formatação de cada pipeline como um *subworkflow* permitiria que futuramente usuários avançados possam selecionar a *pipeline* desejada ao utilizar *entry points* ("pontos de entrada", nesse contexto, o ponto a partir do qual a execução é inicializada) diferentes, aumenta a modularidade da *pipeline* de maneira geral facilitando modificações. Embora haja documentação para o uso e combinação de *subworkflows*, devido à recência da oficialização da DSL2, que só ocorreu em 2020, ainda existem detalhes essenciais quanto a sintaxe da sua criação que não explicitados em sua documentação, podendo ocasionar em erros[23].

4.5. Relatórios de predição de eccDNAs

Cada um dos *pipelines* utilizados é capaz de gerar o próprio relatório de corrida, com níveis de detalhes variados. O *Cyrcular-calling* (v2.1.0) apresenta os relatórios mais completos dentre os *pipelines* utilizados (Tabela 1) e foi utilizado como base em análises posteriores (Tabela 2).

Através do uso da função "PublishAs" do Nextflow, os resultados e relatórios obtidos por cada programa são publicados em uma estrutura de diretórios comum, a fim de facilitar a organização dos dados para a análise posterior dos resultados de cada *pipeline*.

[23] https://github.com/nextflow-io/nextflow/discussions/5334

event_id	circle_length	regions	gene_names	regulatory_features	num_split_reads	category
213-8	4537	18:2314160-2318697		TF_binding_site	161	regulatory
213-11	4537	18:2314160-2318697		TF_binding_site	168	regulatory
213-0	4537	18:2314160-2318697		TF_binding_site	102	regulatory
213-2	4749	18:2314160-2318697,3:106312869-106313081		TF_binding_site	112	regulatory
213-6	4749	18:2314160-2318697,3:106312869-106313081		TF_binding_site	112	regulatory
213-12	4749	18:2314160-2318697,3:106312869-106313081		TF_binding_site	112	regulatory
213-14	4749	18:2314160-2318697,3:106312869-106313081		TF_binding_site	112	regulatory
145-0	8470	18:29319598-29328068		enhancer,open_chromatin_region	82	regulatory
140-0	5368	18:36899379-36904747	KIAA1328	CTCF_binding_site,TF_binding_site,enhancer	10	coding
138-0	7317	18:41917200-41924517		enhancer	416	regulatory
133-0	6830	18:54112602-54119432		CTCF_binding_site,open_chromatin_region,enhancer	104	regulatory
130-0	12719	18:56299537-56312256		enhancer,CTCF_binding_site,open_chromatin_region	22	regulatory
129-0	7848	18:57233369-57241217			146	other
117-0	3630	18:77855109-77858739		open_chromatin_region	50	regulatory

Tabela 1: Tabela adaptada a partir do output do Cyrcular-calling (v2.1.0) , mostrando resultados detectados para o cromossomo 18 de uma amostra do grupo controle. Essa é uma das tabelas geradas ao final da execução do programa, e mostra um "overview" de diversas características dos círculos de eccDNA encontrados, como região de mapeamento, tamanho, características regulatórias, probabilidade da detecção ser correta, e nomes dos genes codificadores detectados.

Regions (chr:start-end)	gene_names	regulatory_features
1:111530134-111538428	TMIGD3	enhancer
1:215907182-215909429	USH2A	
2:143572582-143580818	ARHGAP15	ıcer,open_chromatin_region,CTCF_binding
2:84657689-84662524	DNAH6	
3:105664063-105667126	CBLB	
3:29475916-29479937	RBMS3	open_chromatin_region,enhancer
3:45823074-45830699	LZTFL1	
3:63561122-63566778	SYNPR	CTCF_binding_site,TF_binding_site
3:77200345-77204001	ROBO2	enhancer
3:77200345-77204001	ROBO2,STAG1	enhancer
4:118230700-118235041	NDST3	
4:143217806-143223934	USP38	open_chromatin_region
5:17095931-17096962	BASP1	enhancer
5:64822685-64825673	CWC27	enhancer
6:166365397-166371465	MPC1	enhancer
6:83152571-83155066	DOP1A,PGM3	enhancer
8:60732210-60734118	CHD7	
9:124690210-124692487	NR6A1	enhancer
10:21595403-21602691	MLLT10	open_chromatin_region,CTCF_binding_site
13:36440766-36443484	CCNA1	open_chromatin_region

Tabela 2: Tabela adaptada do a partir do output do *Cyrcular-calling(v2.1.0)*, mostrando genes encontrados nos eccDNAs de uma amostra do grupo controle. Foram encontrados 21 genes nessa amostra.

Além dos relatórios próprios de cada programa, os *subworkflows* para cada um deles também incluem uma etapa para a criação de arquivos BED. O formato BED é um formato de arquivo de texto utilizado para descrever as coordenadas genômicas e possíveis anotações em associação, e que pode ser acessado tanto por ferramentas específicas quanto por processadores de texto e scripts. Os arquivos BED são gerados a partir da manipulação dos relatórios já gerados por cada programa através de linha de comando, padronizados para o formato mínimo de três colunas: *chr* (Indica o cromossomo), *chromStart* (Indica a coordenada no cromossomo onde a sequência especificada se inicia) e *chromEnd* (Indica a coordenada no cromossomo onde a

sequência especificada termina), permitindo assim que os *outputs* de cada *pipeline* sejam adequadamente comparados entre si.

O software *Intervene* pode então utilizar esses arquivos como *input* para a operação de intersecção entre eles, selecionando assim aquelas coordenadas genômicas que são comuns entre todos os resultados e gerando arquivos BED com esses consensos. O *Intervene* se utiliza da função *intersect* do *software Bedtools*, que considera a existência de sobreposição quando há pelo menos um par de base em comum entre as coordenadas, havendo portanto um certo grau de liberdade para realizar as comparações. Através dos arquivos BED consenso gerados pelo Intervene, os eccDNAs detectados por cada um dos *pipelines* com seus parâmetros padrão naquelas coordenadas genômicas seriam validados entre si, e assim se consideraria que esses círculos de eccDNA seriam "provavelmente verdadeiros" (Tabela 3).

FLEC	***CReSIL***	***Cyrcular***	***ecc_finder***	***Intervene***
18:10402961-10406203				
18:10852629-10860948	18:10852631-10860945	18:10852628-10860944		
18:2314161-2318692	18:2314161-2318697	18:2314160-2318697	18:2314160-2318697	18:2314161-2318697
18:29319599-29328072	18:29319599-29328067	18:29319598-29328068		
18:31869713-31871696				
18:36512301-36513501				
18:36899380-36904747		18:36899379-36904747		
18:41917201-41924527	18:41917201-41924518	18:41917200-41924517	18:41917200-41924518	18:41917201-41924518
18:43957481-43958234				
18:49094415-49101019				
18:52496155-52499195				
18:53909910-53910641				
18:54112603-54119461	18:54112603-54119433	18:54112602-54119432		
18:56299538-56312257	18:56299538-56312257	18:56299537-56312256		
18:57233370-57241235	18:57233370-57241218	18:57233369-57241217		
18:69842849-69845775				
18:73367028-73368823	18:73367028-73368828	18:77855109-77858739		
18:77855110-77858739	18:77855110-77858740			
18:8329323-8332142	18:8329323-8332144	18:8329322-8332143		
18:8419019-8420287				
18:9277386-9282058	18:9277386-9282058	18:9277384-9282057		
Total:21	Total:11	Total:11	Total:2	Total:2

Tabela 3: Tabela comparativa entre eccDNAs encontrados nas mesmas posições do Cromossomo 18 para uma amostra do grupo Controle em cada ferramenta e resultado consenso. (Legenda: Vermelho; Identificado nessa posição por apenas uma *pipeline* (n=9),Amarelo; Identificado

nessa posição por quaisquer dois *pipelines* (n=2), Azul; Identificado nessa posição por quaisquer três *pipelines* (n=8), Verde; Identificado nessa posição por todas as quatro *pipelines* utilizadas (n=2).)

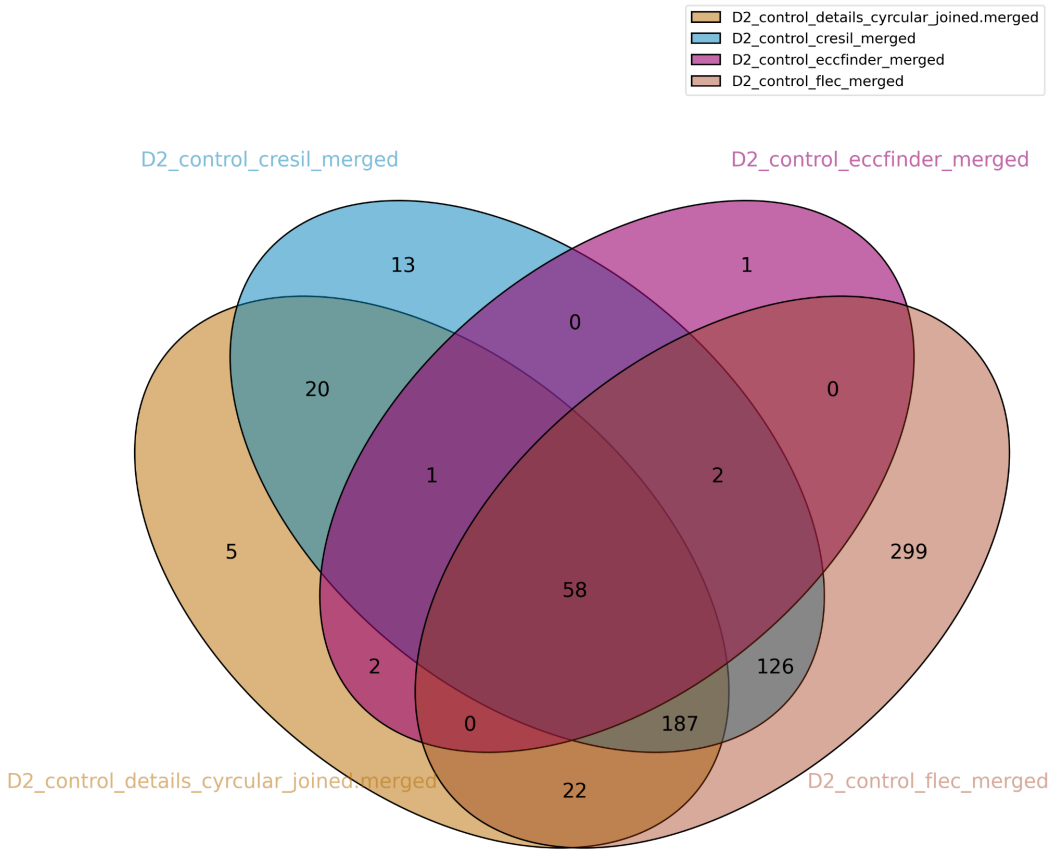

Figura 6: Exemplo de diagrama de Venn gerado pelo Intervene mostrando a sobreposição entre os eccDNAs encontrados por cada um dos *pipelines* com base em sua posição do genoma de referência para a mesma amostra do grupo controle. Nessa amostra, 58 eccDNAs circulares foram encontrados por todos os *pipelines* utilizados.

O Intervene também é capaz de gerar figuras a partir desses dados, como diagramas de Venn, que facilitam a avaliação quanto às diferenças nos resultados quantitativos entre pipelines (Figura 6), tornando o uso do *eccPOP* para comparação entre os resultados das ferramentas mais simples e direto. A utilização dos arquivos de coordenadas genômicas (formato BED) resultantes para a análise comparativa entre os resultados consenso de ambos os grupos também permitiu a análise posterior de algumas características dos eccDNAs encontrados (Figuras 7 e 8).

FLEC(v1.0) detecta grandes números eccDNAs em posições que não são

validadas por nenhuma outra ferramenta. *Ecc_finder*(v1.0.0), *CReSIL*(v1.0.0) e *Cyrcular-calling*(v2.1.0) aparentam ter maior concordância de resultados, e grande parte dos eccDNAs encontrados por essas ferramentas também são detectados por pelo menos uma outra. As menores quantidades de eccDNAs detectados, bem como a grande concordância com múltiplas outras *pipelines*, parece sugerir que *Ecc_finder*(v1.0.0), seguida de *Cyrcular-calling*(v2.1.0), são as ferramentas mais rigorosas quanto a seleção de eccDNAs candidatos, apesar das diferenças entre seus algoritmos; o algoritmo de identificação do *Ecc_finder*(v1.0.0) para dados de leituras longas se aproveita primariamente do número de repetições em tandem geradas pelo RCA para a detecção, enquanto o *Cyrcular-calling*(v2.1.0), utiliza uma abordagem baseada na estatística bayesiana para determinar quais círculos seriam verdadeiros com base na análise das variações estruturais encontradas. *CReSIL*(v1.0.0) e *Cyrcular-calling*(v2.1.0) encontram quantidades similares de eccDNAs, mas *CReSIL*(v1.0.0) tem menos posições validadas, apesar de também utilizar primariamente as informações de *split reads* para determinar eccDNAs verdadeiros.

Embora a abordagem da utilização de múltiplas ferramentas possa trazer maior segurança quanto aos eccDNAs encontrados, a existência de de resultados divergentes também precisa ser observada com cautela. Devido a ausência de um *pipeline* amplamente utilizado como padrão no estudo de eccDNAs, ao compor a sua interpretação biológica com base nos resultados de apenas uma ferramenta, um pesquisador pode por consequência estar sujeito a um possível viés do seu algoritmo de escolha. Ao mesmo tempo, ao aceitar apenas aqueles que tem a interseção de resultados de duas ou mais pipelines, pode-se estar sendo excessivamente exigente e filtrando eccDNAs que seriam de interesse.

Em um estudo de *benchmarking* entre diferentes *pipelines* em termos de acurácia, uso de recursos computacionais, capacidade de identificação, e taxa de duplicação, utilizando dados simulados de diversas técnicas e validação por PCR, (Gao et al., 2024) observou que *CReSIL*(v1.0.0) aparenta ser o *pipeline* mais efetivo para a detecção de eccDNA a partir de dados de leituras-longas quando há profundidade de leitura de, no mínimo, 10X, sendo considerado preferível ao *FLEC* (v1.0) por sua capacidade de detectar repetições em tandem de concatâmeros e de não-concatâmeros que contêm split-reads, enquanto *FLEC*(v1.0) so é capaz de

detectar repetições de concatâmeros. *Ecc_finder*(v1.0.0) teve o menor tempo de execução e utilizou menos memória dentre os *pipelines* testados, mas os eccDNAs identificados apresentaram mais diferenças das sequências de dados simulados que outros *pipelines* (cerca de 66 pb) apesar da identificação ainda ter sido positiva. O estudo não avaliou *Cyrcular-calling*(v2.1.0) ou características relacionadas à usabilidade dos *pipelines*.

Assim, ainda se mostra necessária a averiguação *in situ* e *in silico* para conclusões adicionais quanto a acurácia dos resultados entre as ferramentas, bem como a efetividade da abordagem do uso de consensos para a melhor interpretação biológica.

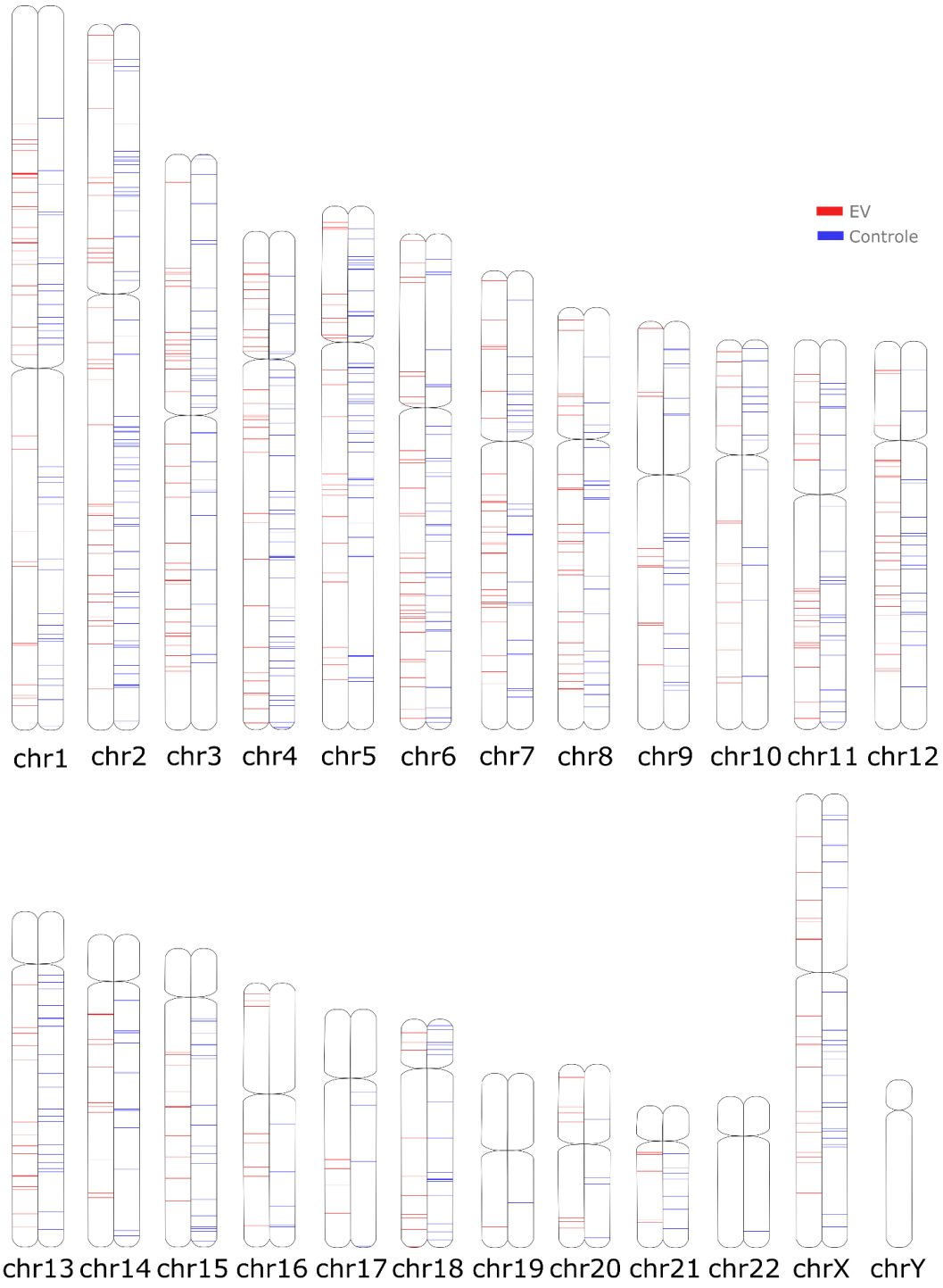

Figura 7: Ideograma comparativo da distribuição das coordenadas genômicas dos eccDNAs consenso para todas as amostras dos grupos controle (bandas em azul) e de vesículas extracelulares (EV, bandas em vermelho).

Em termos da distribuição dos eccDNAs detectados ao longo dos cromossomos (Figura 7), verifica-se que estão amplamente distribuídos. Foram

encontrados 477 eccDNAs distintos no grupo controle e 422 eccDNAs distintos no grupo EV, com todos os cromossomos sendo representados, exceto o cromossomo Y e o cromossomo 22 (com apenas um eccDNA mapeado). Em termos da distribuição dos eccDNAs ao longo dos cromossomos, verifica-se que os eccDNAs estão dispostos uniformemente, sem distinção clara entre posicionamento nos braços dos cromossomos ou em regiões de heterocromatina. Esses resultados são condizentes com relatos da literatura que afirmam que a distribuição dos eccDNAs é aparentemente aleatória (Pecorino et al., 2022).

Além disso, não há grande coincidência entre os eccDNAs encontrados nos dois grupos. Isso pode ocorrer devido a aleatoriedade previamente mencionada, ou ser indicativa de padrões de origens ou mecanismos de produção distintos entre os eccDNAs advindos de ambos os grupos, apesar das EVs serem derivadas dos fibroblastos do grupo Controle. Isso pode sugerir uma relevância biológica quanto ao papel dos eccDNAs presentes nas vesículas extracelulares, porém, o estudo que deu origem aos dados é, até onde temos conhecimento, o primeiro onde eccDNAs foram detectados em vesículas extracelulares, e portanto ainda não existe literatura que possa esclarecer esse notável contraste.

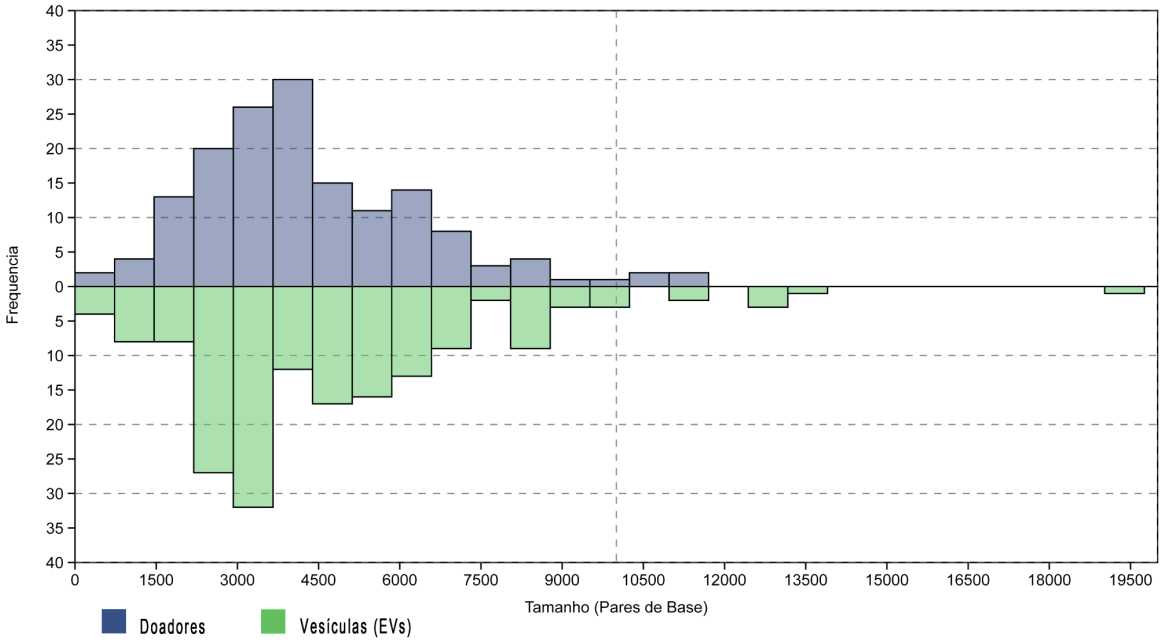

Figura 8: Histograma comparativo entre as distribuições de tamanho dos eccDNAs detectados nas amostras do grupo controle e das vesículas extracelulares (EVs). Produzido com PAST(v4.08)..

Não houve diferença significativa entre as distribuições de tamanho (Figura 8) de ambos os grupos ($p = 0.63$, teste de Kolmogorov-Smirnov). Ambas as distribuições são assimétricas com obliquidade positiva ($v > 0$), de forma que a maioria dos eccDNAs encontrados em ambos os grupos têm tamanho menor que o tamanho médio. Além disso, a cauda dos valores de EVs é levemente mais pesada, havendo mais eccDNAs de tamanhos extremos no grupo de EVs. O menor eccDNA consenso encontrado no grupo Controle possui 585 pb, enquanto o menor eccDNA encontrado em EVs possui 602 pb. O maior eccDNA consenso no grupo Controle possui 11203 pb, em contraste com o maior eccDNA encontrado no grupo de EVs, com 19558 bp. O tamanho médio das sequências encontradas nos Controles foi de 4351 pb, próximo a média de 4724 pb em EVs.

De maneira geral, as distribuições de tamanho dos eccDNAs detectados em ambas as amostras são condizentes com aqueles resultados já encontrados na literatura (Møller et al., 2018), e classificam a sua grande maioria como spcDNA. Além disso, a tendência do método RCA de amplificar preferencialmente DNAs circulares com menos de 10kb (Norman et al., 2014) pode ter limitado o tamanho dos eccDNAs detectados, embora o uso de sequenciamento de leituras-longas combinado ao Circle-seq permita melhor eficiência na detecção de eccDNAs maiores que 10kb (Gao et al., 2024). Nota-se, assim, que a performance de todas as ferramentas de análise de eccDNAs pode ser limitada pelo método experimental utilizado para a geração dos dados..

5. Conclusões

Apesar de sua ampla incidência em eucariotos, os eccDNAs são parte de um campo da biologia molecular ainda pouco explorado, com diversas lacunas sobre os seus mecanismos de biogênese, regulação e sua relação com desenvolvimento normal e patológico das células (Y. Zhao et al., 2022) ainda a serem exploradas. A abordagem bioinformática é essencial para o preenchimento dessas lacunas, mas a sua acessibilidade e reprodutibilidade ainda são fatores limitantes, bem como a ausência de padronização e de protocolos computacionais universalmente aceitos para a realização de estudos na área.

A abordagem envolvendo a combinação entre Nextflow, Conda e Docker facilitou o uso das ferramentas, possibilitou a unificação das configurações de todos os *pipelines* em um único arquivo, permitiu organização da saída em uma estrutura de diretórios simples, facilitou o uso dos programas em diretórios diferentes dos de instalação e até mesmo permitiu o uso e instalação corretos daquelas ferramentas que apresentavam problemas ao permitir que as alterações necessárias no código fossem mantidas. Além disso, erros de versão e manutenção dos *pipelines* foram solucionados, assim, o eccPOP permitiu melhorar a usabilidade dessas ferramentas.

Embora a criação da ferramenta tenha tornado certos aspectos da geração de resultados computacionais para as análises de eccDNAs mais simples, ainda existem aspectos que poderiam ser modificados e melhorados. Um deles é a criação de uma interface gráfica, pois muitos usuários não têm familiaridade com a interface da linha de comando. Permitir o uso exclusivo, ou pelo menos majoritário, da GUI poderia tornar a experiência para o usuário final mais simples, como demonstrado pela popularidade de plataformas como a *Galaxy*[24]. A plataforma da *Sequera Labs*[25] do Nextflow (antigo *Nextflow Tower*) é uma possibilidade que permite o uso em navegador *web* para algumas *pipelines*, porém, programas precisam seguir uma formatação e estrutura de diretórios específica, e o uso público e massivo dessas é limitado e requer pagamento. A criação de outras ferramentas e opções de visualização dos dados também teria sido uma abordagem interessante para facilitar a interpretação biológica dos dados gerados.

A impossibilidade de se conteinerizar o *Cyrcular-calling*(v2.1.0) corretamente para o uso dentro do Nextflow também foi um problema. Enquanto o Snakemake possui suporte

[24] https://usegalaxy.org/
[25] https://seqera.io/platform/

para a integração de outras WMSs (a partir da versão 6.2), permitindo o uso de *workflows* escritos em outros WMSs enquanto se performa pré e pós processamento dentro do Snakemake, através da diretiva [*handover*], o Nextflow não possui tal suporte a integração até o presente momento, tornando esse tipo operação complexa e instável. A maneira como o *Cyrcular-calling*(v2.1.0) é configurado também não permitiu uma boa integração ao *pipeline* geral, o que tornou esse *subworkflow* menos "ajustável" que o de outros componentes.

De maneira geral, o uso de múltiplos *pipelines* permite com que o usuário tenha uma visão mais ampla das possibilidades de resultados oferecidos por cada ferramenta, e o uso dos resultados consenso entre os *pipelines* permite com que o usuário tenha mais confiança nos dados gerados, embora também possa trazer o risco de remover eccDNAs candidatos em excesso, caso sejam usados exclusivamente. A interpretação biológica dos dados gerados ainda depende da validação *in situ*, através da amplificação dos eccDNAs encontrados pela *pipeline,* bem como da validação *in silico* através do uso de sequências falsas de eccDNA, para garantir que a detecção final é acurada e realizar a comparação direta entre os *pipelines* utilizados.

Ainda há espaço para o desenvolvimento de novas ferramentas que possam melhorar o entendimento sobre os eccDNAs. (Zhou et al., 2024) sugere que o uso de aprendizagem de máquina (*deep learning*) seja um possível foco para a caracterização de ecDNAs ou eccDNAs de estrutura complexa de maneira geral, capazes de reduzir a taxa de falsos-positivos, aumentar a acurácia e eficiência do processamento em larga escala e ter aplicações em diversos cenários experimentais para a caracterização funcional dos eccDNA em uma abordagem holística.

Espera-se que a disponibilização dessa *pipeline* permita facilitar a caracterização de eccDNAs por sequenciamento de leituras-longas, a fim de elucidar suas características e permitir a eventual exploração de seu potencial biotecnológico, terapêutico e diagnóstico, contribuindo assim com a pesquisa sobre eccDNAs.

6. Referências

Ali, M. M., Li, F., Zhang, Z., Zhang, K., Kang, D.-K., Ankrum, J. A., Le, X. C., & Zhao, W. (2014). Rolling circle amplification: A versatile tool for chemical biology, materials science and medicine. *Chemical Society Reviews*, *43*(10), 3324–3341. https://doi.org/10.1039/C3CS60439J

Allain, F., Roméjon, J., Rosa, P. L., Jarlier, F., Servant, N., & Hupé, P. (2022). Geniac: Automatic Configuration GENerator and Installer for nextflow pipelines. *Open Research Europe 2022*, 1:76. https://doi.org/10.12688/openreseurope.13861.2

Alser, M., Lawlor, B., Abdill, R. J., Waymost, S., Ayyala, R., Rajkumar, N., LaPierre, N., Brito, J., Ribeiro-dos-Santos, A. M., Almadhoun, N., Sarwal, V., Firtina, C., Osinski, T., Eskin, E., Hu, Q., Strong, D., Kim, B.-D. (B D., Abedalthagafi, M. S., Mutlu, O., & Mangul, S. (2024). Packaging and containerization of computational methods. *Nature Protocols*, *19*(9), 2529–2539. https://doi.org/10.1038/s41596-024-00986-0

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H.-R., Iyer, V., … Clamp, M. (2004). An Overview of Ensembl. *Genome Research*, *14*(5), 925–928. https://doi.org/10.1101/gr.1860604

Bolchini, D., Finkelstein, A., Perrone, V., & Nagl, S. (2009). Better bioinformatics through usability analysis. *Bioinformatics*, *25*(3), 406–412. https://doi.org/10.1093/bioinformatics/btn633

Carroll, S. M., DeRose, M. L., Gaudray, P., Moore, C. M., Needham-Vandevanter, D. R., Von Hoff, D. D., & Wahl, G. M. (1988). Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Molecular and Cellular Biology*, *8*(4), 1525–1533.

Chamorro González, R., Conrad, T., Stöber, M. C., Xu, R., Giurgiu, M., Rodriguez-Fos, E., Kasack, K., Brückner, L., van Leen, E., Helmsauer, K., Dorado Garcia, H., Stefanova, M. E., Hung, K. L., Bei, Y., Schmelz, K., Lodrini, M., Mundlos, S., Chang, H. Y., Deubzer, H. E., … Henssen, A. G. (2023). Parallel sequencing of extrachromosomal circular DNAs and transcriptomes in single cancer cells. *Nature Genetics*, *55*(5), 880–890. https://doi.org/10.1038/s41588-023-01386-y

Choi, Y.-D., Roy, B., Nguyen, J., Ahmad, R., Maghami, I., Nassar, A., Li, Z., Castronova, A. M., Malik, T., Wang, S., & Goodall, J. L. (2023). Comparing containerization-based approaches for reproducible computational modeling of environmental systems. *Environmental Modelling & Software*, *167*, 105760. https://doi.org/10.1016/j.envsoft.2023.105760

Cohen, S., Houben, A., & Segal, D. (2008). Extrachromosomal circular DNA derived from tandemly repeated genomic sequences in plants. *The Plant Journal*, *53*(6), 1027–1034. https://doi.org/10.1111/j.1365-313X.2007.03394.x

Cohen, S., Regev, A., & Lavi, S. (1997). Small polydispersed circular DNA (spcDNA) in human cells: Association with genomic instability. *Oncogene*, *14*(8), Artigo 8. https://doi.org/10.1038/sj.onc.1200917

Cox, D., Yuncken, C., & Spriggs, ArthurI. (1965). MINUTE CHROMATIN BODIES IN MALIGNANT TUMOURS OF CHILDHOOD. *The Lancet*, *286*(7402), 55–58. https://doi.org/10.1016/S0140-6736(65)90131-5

deCarvalho, A. C., Kim, H., Poisson, L. M., Winn, M. E., Mueller, C., Cherba, D., Koeman, J., Seth, S., Protopopov, A., Felicella, M., Zheng, S., Multani, A., Jiang, Y., Zhang, J., Nam, D.-H., Petricoin, E. F., Chin, L., Mikkelsen, T., & Verhaak, R. G. W. (2018). Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nature Genetics*, *50*(5),

708–717. https://doi.org/10.1038/s41588-018-0105-0

Deelman, E., Vahi, K., Rynge, M., Mayani, R., da Silva, R. F., Papadimitriou, G., & Livny, M. (2019). The Evolution of the Pegasus Workflow Management Software. *Computing in Science & Engineering*, *21*(4), 22–36. Computing in Science & Engineering. https://doi.org/10.1109/MCSE.2019.2919690

Deng, E., & Fan, X. (2024). Categorizing Extrachromosomal Circular DNA as Biomarkers in Serum of Cancer. *Biomolecules*, *14*(4), 488. https://doi.org/10.3390/biom14040488

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Dijk, E. L. van, Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, *34*(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Dillon, L. W., Kumar, P., Shibata, Y., Wang, Y.-H., Willcox, S., Griffith, J. D., Pommier, Y., Takeda, S., & Dutta, A. (2015). Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity. *Cell Reports*, *11*(11), 1749–1759. https://doi.org/10.1016/j.celrep.2015.05.020

dos Santos, C. R., Hansen, L. B., Rojas-Triana, M., Johansen, A. Z., Perez-Moreno, M., & Regenberg, B. (2023). Variation of extrachromosomal circular DNA in cancer cell lines. *Computational and Structural Biotechnology Journal*, *21*, 4207–4214. https://doi.org/10.1016/j.csbj.2023.08.027

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278. https://doi.org/10.1038/s41587-020-0439-x

Fang, M., Fang, J., Luo, S., Liu, K., Yu, Q., Yang, J., Zhou, Y., Li, Z., Sun, R., Guo, C., & Qu, K. (2024). eccDNA-pipe: An integrated pipeline for identification, analysis and visualization of extrachromosomal circular DNA from high-throughput sequencing data. *Briefings in Bioinformatics*, *25*(2), bbae034. https://doi.org/10.1093/bib/bbae034

Gao, X., Liu, K., Luo, S., Tang, M., Liu, N., Jiang, C., Fang, J., Li, S., Hou, Y., Guo, C., & Qu, K. (2024). Comparative analysis of methodologies for detecting extrachromosomal circular DNA. *Nature Communications*, *15*(1), 9208. https://doi.org/10.1038/s41467-024-53496-8

Garafutdinov, R. R., Sakhabutdinova, A. R., Gilvanov, A. R., & Chemeris, A. V. (2021). Rolling Circle Amplification as a Universal Method for the Analysis of a Wide Range of Biological Targets. *Russian Journal of Bioorganic Chemistry*, *47*(6), 1172–1189. https://doi.org/10.1134/S1068162021060078

Gaubatz, J. W. (1990). Extrachromosomal circular DNAs and genomic sequence plasticity in eukaryotic cells. *Mutation Research*, *237*(5–6), 271–292. https://doi.org/10.1016/0921-8734(90)90009-g

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

Hammer, O., Harper, D., & Ryan, P. (2001). PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica*, *4*, 1–9.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hotta, Y., & Bassel, A. (1965). Molecular size and circularity of dna in cells of mammals and higher plants*. *Proceedings of the National Academy of Sciences*, *53*(2), 356–362.

https://doi.org/10.1073/pnas.53.2.356

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

Hull, R. M., King, M., Pizza, G., Krueger, F., Vergara, X., & Houseley, J. (2019). Transcription-induced formation of extrachromosomal DNA during yeast ageing. *PLOS Biology*, *17*(12), e3000471. https://doi.org/10.1371/journal.pbio.3000471

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E., & van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*(2), 149–155. https://doi.org/10.1038/ng.295

Jiang, R., Yang, M., Zhang, S., & Huang, M. (2023). Advances in sequencing-based studies of microDNA and ecDNA: Databases, identification methods, and integration with single-cell analysis. *Computational and Structural Biotechnology Journal*, *21*, 3073–3080. https://doi.org/10.1016/j.csbj.2023.05.017

Khan, A., & Mathelier, A. (2017). Intervene: A tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics*, *18*(1), 287. https://doi.org/10.1186/s12859-017-1708-7

Kim, H., Nguyen, N.-P., Turner, K., Wu, S., Gujar, A. D., Luebeck, J., Liu, J., Deshpande, V., Rajkumar, U., Namburi, S., Amin, S. B., Yi, E., Menghi, F., Schulte, J. H., Henssen, A. G., Chang, H. Y., Beck, C. R., Mischel, P. S., Bafna, V., & Verhaak, R. G. W. (2020). Extrachromosomal DNA Is Associated with Oncogene Amplification and Poor Outcome across Multiple Cancers. *Nature Genetics*, *52*(9), 891–897. https://doi.org/10.1038/s41588-020-0678-2

Ko, I., Kranse, O. P., Senatori, B., & Eves-van den Akker, S. (2024). A Critical Appraisal of

DNA Transfer from Plants to Parasitic Cyst Nematodes. *Molecular Biology and Evolution*, *41*(2), msae030. https://doi.org/10.1093/molbev/msae030

Kulkarni, N., Alessandrì, L., Panero, R., Arigoni, M., Olivero, M., Ferrero, G., Cordero, F., Beccuti, M., & Calogero, R. A. (2018). Reproducible bioinformatics project: A community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics*, *19*(10), 349. https://doi.org/10.1186/s12859-018-2296-x

Kumar, M. A., Baba, S. K., Sadida, H. Q., Marzooqi, S. A., Jerobin, J., Altemani, F. H., Algehainy, N., Alanazi, M. A., Abou-Samra, A.-B., Kumar, R., Al-Shabeeb Akil, A. S., Macha, M. A., Mir, R., & Bhat, A. A. (2024). Extracellular vesicles as tools and targets in therapy for diseases. *Signal Transduction and Targeted Therapy*, *9*(1), 1–41. https://doi.org/10.1038/s41392-024-01735-1

Kumar, P., Dillon, L. W., Shibata, Y., Jazaeri, A., Jones, D. R., & Dutta, A. (2017). Normal and Cancerous Tissues Release extrachromosomal circular DNA (eccDNA) into the Circulation. *Molecular cancer research : MCR*, *15*(9), 1197–1205. https://doi.org/10.1158/1541-7786.MCR-17-0095

Li, F., Ming, W., Lu, W., Wang, Y., Dong, X., & Bai, Y. (2024). Bioinformatics advances in eccDNA identification and analysis. *Oncogene*, *43*(41), 3021–3036. https://doi.org/10.1038/s41388-024-03138-6

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, R., Wang, Y., Li, J., & Zhou, X. (2022). Extrachromosomal circular DNA (eccDNA): An emerging star in cancer. *Biomarker Research*, *10*(1), 53.

https://doi.org/10.1186/s40364-022-00399-9

Liao, Z., Jiang, W., Ye, L., Li, T., Yu, X., & Liu, L. (2020). Classification of extrachromosomal circular DNA with a focus on the role of extrachromosomal DNA (ecDNA) in tumor heterogeneity and progression. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1874*(1), 188392. https://doi.org/10.1016/j.bbcan.2020.188392

Ling, X., Han, Y., Meng, J., Zhong, B., Chen, J., Zhang, H., Qin, J., Pang, J., & Liu, L. (2021). Small extrachromosomal circular DNA (eccDNA): Major functions in evolution and cancer. *Molecular Cancer*, *20*(1), 113. https://doi.org/10.1186/s12943-021-01413-8

Mann, L., Seibt, K. M., Weber, B., & Heitkam, T. (2022). ECCsplorer: A pipeline to detect extrachromosomal circular DNA (eccDNA) from next-generation sequencing data. *BMC Bioinformatics*, *23*(1), 40. https://doi.org/10.1186/s12859-021-04545-2

Manni, G., Buratta, S., Pallotta, M. T., Chiasserini, D., Di Michele, A., Emiliani, C., Giovagnoli, S., Pascucci, L., Romani, R., Bellezza, I., Urbanelli, L., & Fallarino, F. (2023). Extracellular Vesicles in Aging: An Emerging Hallmark? *Cells*, *12*(4), 527. https://doi.org/10.3390/cells12040527

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Mazzucco, G., Huda, A., Galli, M., Piccini, D., Giannattasio, M., Pessina, F., & Doksani, Y. (2020). Telomere damage induces internal loops that generate telomeric circles. *Nature Communications*, *11*(1), 5297. https://doi.org/10.1038/s41467-020-19139-4

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). *Sustainable data analysis with Snakemake* (10:33). F1000Research. https://doi.org/10.12688/f1000research.29032.2

Molin, W. T., Yaguchi, A., Blenner, M., & Saski, C. A. (2020). The EccDNA Replicon: A Heritable, Extranuclear Vehicle That Enables Gene Amplification and Glyphosate Resistance in Amaranthus palmeri[OPEN]. *The Plant Cell*, *32*(7), 2132–2140. https://doi.org/10.1105/tpc.20.00099

Møller, H. D., Bojsen, R. K., Tachibana, C., Parsons, L., Botstein, D., & Regenberg, B. (2016). Genome-wide Purification of Extrachromosomal Circular DNA from Eukaryotic Cells. *Journal of Visualized Experiments : JoVE*, *110*, 54239. https://doi.org/10.3791/54239

Møller, H. D., Mohiyuddin, M., Prada-Luengo, I., Sailani, M. R., Halling, J. F., Plomgaard, P., Maretty, L., Hansen, A. J., Snyder, M. P., Pilegaard, H., Lam, H. Y. K., & Regenberg, B. (2018). Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. *Nature Communications*, *9*(1), 1069. https://doi.org/10.1038/s41467-018-03369-8

Møller, H. D., Parsons, L., Jørgensen, T. S., Botstein, D., & Regenberg, B. (2015). Extrachromosomal circular DNA is common in yeast. *Proceedings of the National Academy of Sciences*, *112*(24), E3114–E3122. https://doi.org/10.1073/pnas.1508825112

Møller, H. D., Ramos-Madrigal, J., Prada-Luengo, I., Gilbert, M. T. P., & Regenberg, B. (2020). Near-Random Distribution of Chromosome-Derived Circular DNA in the Condensed Genome of Pigeons and the Larger, More Repeat-Rich Human Genome. *Genome Biology and Evolution*, *12*(2), 3762–3777.

https://doi.org/10.1093/gbe/evz281

Noer, J. B., Hørsdal, O. K., Xiang, X., Luo, Y., & Regenberg, B. (2022). Extrachromosomal circular DNA in cancer: History, current knowledge, and methods. *Trends in Genetics*, *38*(7), 766–781. https://doi.org/10.1016/j.tig.2022.02.007

Norman, A., Riber, L., Luo, W., Li, L. L., Hansen, L. H., & Sørensen, S. J. (2014). An Improved Method for Including Upper Size Range Plasmids in Metamobilomes. *PLOS ONE*, *9*(8), e104405. https://doi.org/10.1371/journal.pone.0104405

Nyrén, P., & Lundin, A. (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical Biochemistry*, *151*(2), 504–509. https://doi.org/10.1016/0003-2697(85)90211-8

Paulsen, T., Kumar, P., Koseoglu, M. M., & Dutta, A. (2018). Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells. *Trends in Genetics*, *34*(4), 270–278. https://doi.org/10.1016/j.tig.2017.12.010

Paulsen, T., Shibata, Y., Kumar, P., Dillon, L., & Dutta, A. (2019). Small extrachromosomal circular DNAs, microDNA, produce short regulatory RNAs that suppress gene expression independent of canonical promoters. *Nucleic Acids Research*, *47*(9), 4586–4596. https://doi.org/10.1093/nar/gkz155

Pecorino, L. T., Verhaak, R. G. W., Henssen, A., & Mischel, P. S. (2022). Extrachromosomal DNA (ecDNA): An origin of tumor heterogeneity, genomic remodeling, and drug resistance. *Biochemical Society Transactions*, *50*(6), 1911–1920. https://doi.org/10.1042/BST20221045

Peng, H., Mirouze, M., & Bucher, E. (2022). Extrachromosomal circular DNA: A neglected nucleic acid molecule in plants. *Current Opinion in Plant Biology*, *69*, 102263. https://doi.org/10.1016/j.pbi.2022.102263

Prada-Luengo, I., Krogh, A., Maretty, L., & Regenberg, B. (2019). Sensitive detection of

circular DNAs at single-nucleotide resolution using guided realignment of partially aligned reads. *BMC Bioinformatics*, *20*(1), 663. https://doi.org/10.1186/s12859-019-3160-3

Qiu, G.-H., Zheng, X., Fu, M., Huang, C., & Yang, X. (2021). The decreased exclusion of nuclear eccDNA: From molecular and subcellular levels to human aging and age-related diseases. *Ageing Research Reviews*, *67*, 101306. https://doi.org/10.1016/j.arr.2021.101306

Robbins, P. D. (2017). Extracellular vesicles and aging. *Stem Cell Investigation*, *4*, 98. https://doi.org/10.21037/sci.2017.12.03

Röijer, E., Nordkvist, A., Ström, A.-K., Ryd, W., Behrendt, M., Bullerdiek, J., Mark, J., & Stenman, G. (2002). Translocation, Deletion/Amplification, and Expression of HMGIC and MDM2 in a Carcinoma ex Pleomorphic Adenoma. *The American Journal of Pathology*, *160*(2), 433–440.

Salmela, L., Walve, R., Rivals, E., & Ukkonen, E. (2017). Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*, *33*(6), 799–806. https://doi.org/10.1093/bioinformatics/btw321

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

Shahi, S., Kang, T., & Fonseka, P. (2024). Extracellular Vesicles in Pathophysiology: A Prudent Target That Requires Careful Consideration. *Cells*, *13*(9), Artigo 9. https://doi.org/10.3390/cells13090754

Shay, J. W., & Wright, W. E. (2019). Telomeres and telomerase: Three decades of progress. *Nature Reviews Genetics*, *20*(5), 299–309. https://doi.org/10.1038/s41576-019-0099-1

Shibata, Y., Kumar, P., Layer, R., Willcox, S., Gagan, J. R., Griffith, J. D., & Dutta, A. (2012).

Extrachromosomal MicroDNAs and Chromosomal Microdeletions in Normal Tissues. *Science*, *336*(6077), 82–86. https://doi.org/10.1126/science.1213307

Shoshani, O., Brunner, S. F., Yaeger, R., Ly, P., Nechemia-Arbely, Y., Kim, D. H., Fang, R., Castillon, G. A., Yu, M., Li, J. S. Z., Sun, Y., Ellisman, M. H., Ren, B., Campbell, P. J., & Cleveland, D. W. (2021). Chromothripsis drives the evolution of gene amplification in cancer. *Nature*, *591*(7848), 137–141. https://doi.org/10.1038/s41586-020-03064-z

Sin, S. T. K., Ji, L., Deng, J., Jiang, P., Cheng, S. H., Heung, M. M. S., Lau, C. S. L., Leung, T. Y., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (2021). Characteristics of Fetal Extrachromosomal Circular DNA in Maternal Plasma: Methylation Status and Clearance. *Clinical Chemistry*, *67*(5), 788–796. https://doi.org/10.1093/clinchem/hvaa326

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, *6*(7), 2601–2610. https://doi.org/10.1093/nar/6.7.2601

Stanfield, S. W., & Lengyel, J. A. (1979). Small circular DNA of Drosophila melanogaster: Chromosomal homology and kinetic complexity. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(12), 6142–6146. https://doi.org/10.1073/pnas.76.12.6142

Storlazzi, C. T., Lonoce, A., Guastadisegni, M. C., Trombetta, D., D'Addabbo, P., Daniele, G., L'Abbate, A., Macchia, G., Surace, C., Kok, K., Ullmann, R., Purgato, S., Palumbo, O., Carella, M., Ambros, P. F., & Rocchi, M. (2010). Gene amplification as double minutes or homogeneously staining regions in solid tumors: Origin and structure. *Genome Research*, *20*(9), 1198–1206. https://doi.org/10.1101/gr.106252.110

Stoudt, S., Vásquez, V. N., & Martinez, C. C. (2021). Principles for data analysis workflows. *PLoS Computational Biology*, *17*(3), e1008770.

https://doi.org/10.1371/journal.pcbi.1008770

Trisovic, A., Lau, M. K., Pasquier, T., & Crosas, M. (2022). A large-scale study on research code quality and execution. *Scientific Data*, *9*(1), 60. https://doi.org/10.1038/s41597-022-01143-6

Tüns, A. I., Hartmann, T., Magin, S., González, R. C., Henssen, A. G., Rahmann, S., Schramm, A., & Köster, J. (2022). Detection and Validation of Circular DNA Fragments Using Nanopore Sequencing. *Frontiers in Genetics*, *13*. https://www.frontiersin.org/articles/10.3389/fgene.2022.867018

Turner, K. M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., Li, B., Arden, K., Ren, B., Nathanson, D. A., Kornblum, H. I., Taylor, M. D., Kaushal, S., Cavenee, W. K., Wechsler-Reya, R., Furnari, F. B., Vandenberg, S. R., Rao, P. N., Wahl, G. M., … Mischel, P. S. (2017). Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature*, *543*(7643), Artigo 7643. https://doi.org/10.1038/nature21356

Turner-Trauring, I. (2020, setembro 11). *Shrink your Conda Docker images with conda-pack*. Python⇒Speed. https://pythonspeed.com/articles/conda-docker-image-size/

van Loon, N., Miller, D., & Murnane, J. P. (1994). Formation of extrachromosomal circular DNA in HeLa cells by nonhomologous recombination. *Nucleic Acids Research*, *22*(13), 2447–2452. https://doi.org/10.1093/nar/22.13.2447

van Niel, G., D'Angelo, G., & Raposo, G. (2018). Shedding light on the cell biology of extracellular vesicles. *Nature Reviews Molecular Cell Biology*, *19*(4), 213–228. https://doi.org/10.1038/nrm.2017.125

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A. D., Musselman-Brown, A., Schmidt, H., Amstutz, P., Craft, B., Goldman, M., Rosenbloom, K., Cline, M., O'Connor, B., Hanna, M., Birger, C., …

Paten, B. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, *35*(4), 314–316. https://doi.org/10.1038/nbt.3772

Voss, K., Auwera, G. V. der, & Gentry, J. (2017). Full-stack genomics pipelining with GATK4 + WDL + Cromwell . *F1000Research*, *6*. https://doi.org/10.7490/f1000research.1114634.1

Vosseberg, J., van Hooff, J. J. E., Köstlbacher, S., Panagiotou, K., Tamarit, D., & Ettema, T. J. G. (2024). The emerging view on the origin and early evolution of eukaryotic cells. *Nature*, *633*(8029), 295–305. https://doi.org/10.1038/s41586-024-07677-6

Wanchai, V., Jenjaroenpun, P., Leangapichart, T., Arrey, G., Burnham, C. M., Tümmler, M. C., Delgado-Calle, J., Regenberg, B., & Nookaew, I. (2022). CReSIL: Accurate identification of extrachromosomal circular DNA from long-read sequences. *Briefings in Bioinformatics*, *23*(6), bbac422. https://doi.org/10.1093/bib/bbac422

Wang, T., Zhang, H., Zhou, Y., & Shi, J. (2021). Extrachromosomal circular DNA: A new potential role in cancer progression. *Journal of Translational Medicine*, *19*, 257. https://doi.org/10.1186/s12967-021-02927-x

Wang, Y., Wang, M., Djekidel, M. N., Chen, H., Liu, D., Alt, F. W., & Zhang, Y. (2021). eccDNAs are apoptotic products with high innate immunostimulatory activity. *Nature*, *599*(7884), Artigo 7884. https://doi.org/10.1038/s41586-021-04009-w

Wang, Y., Wang, M., & Zhang, Y. (2022). Purification, full-length sequencing and genomic origin mapping of eccDNA. *Nature Protocols*, 1–17. https://doi.org/10.1038/s41596-022-00783-7

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Wang, Z., Yu, J., Zhu, W., Hong, X., Xu, Z., Mao, S., Huang, L., Han, P., He, C., Song, C., &

Xiang, X. (2024). Unveiling the mysteries of extrachromosomal circular DNA: From generation to clinical relevance in human cancers and health. *Molecular Cancer*, *23*(1), 276. https://doi.org/10.1186/s12943-024-02187-5

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, *20*(1), 129. https://doi.org/10.1186/s13059-019-1727-y

Wu, M., & Rai, K. (2022). Demystifying extrachromosomal DNA circles: Categories, biogenesis, and cancer therapeutics. *Computational and Structural Biotechnology Journal*, *20*, 6011–6022. https://doi.org/10.1016/j.csbj.2022.10.033

Wu, S., Turner, K. M., Nguyen, N., Raviram, R., Erb, M., Santini, J., Luebeck, J., Rajkumar, U., Diao, Y., Li, B., Zhang, W., Jameson, N., Corces, M. R., Granja, J. M., Chen, X., Coruh, C., Abnousi, A., Houston, J., Ye, Z., … Mischel, P. S. (2019). Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature*, *575*(7784), 699–703. https://doi.org/10.1038/s41586-019-1763-5

Yu, L., Zhu, G., Zhang, Z., Yu, Y., Zeng, L., Xu, Z., Weng, J., Xia, J., Li, J., & Pathak, J. L. (2023). Apoptotic bodies: Bioactive treasure left behind by the dying cells with robust diagnostic and therapeutic application potentials. *Journal of Nanobiotechnology*, *21*(1), 218. https://doi.org/10.1186/s12951-023-01969-1

Zaringhalam, M., & Federer, L. (2020). *Data and Code for Reproducible Research: Lessons Learned from the NLM Reproducibility Workshop*. https://doi.org/10.5281/ZENODO.3818329

Zhang, B., Yeo, R. W. Y., Tan, K. H., & Lim, S. K. (2016). Focus on Extracellular Vesicles: Therapeutic Potential of Stem Cell-Derived Extracellular Vesicles. *International Journal of Molecular Sciences*, *17*(2), 174. https://doi.org/10.3390/ijms17020174

Zhang, C., Du, Q., Zhou, X., Qu, T., Liu, Y., Ma, K., Shen, Z., Wang, Q., Zhang, Z., & Zhang,

R. (2024). Differential Expression and Analysis of Extrachromosomal Circular DNAs as Serum Biomarkers in Pulmonary Arterial Hypertension. *Respiratory Research*, *25*(1), 181. https://doi.org/10.1186/s12931-024-02808-z

Zhang, P., Peng, H., Llauro, C., Bucher, E., & Mirouze, M. (2021). ecc_finder: A Robust and Accurate Tool for Detecting Extrachromosomal Circular DNA From Sequencing Data. *Frontiers in Plant Science*, *12*, 743742. https://doi.org/10.3389/fpls.2021.743742

Zhao, X.-K., Xing, P., Song, X., Zhao, M., Zhao, L., Dang, Y., Lei, L.-L., Xu, R.-H., Han, W.-L., Wang, P.-P., Yang, M.-M., Hu, J.-F., Zhong, K., Zhou, F.-Y., Han, X.-N., Meng, C.-L., Ji, J.-J., Chen, X., & Wang, L.-D. (2021). Focal Amplifications Are Associated with Chromothripsis Events and Diverse Prognoses in Gastric Cardia Adenocarcinoma. *Nature Communications*, *12*(1), 6489. https://doi.org/10.1038/s41467-021-26745-3

Zhao, Y., Yu, L., Zhang, S., Su, X., & Zhou, X. (2022). Extrachromosomal circular DNA: Current status and future prospects. *eLife*, *11*, e81412. https://doi.org/10.7554/eLife.81412

Zhou, L., Tang, W., Ye, B., & Zou, L. (2024). Characterization, biogenesis model, and current bioinformatics of human extrachromosomal circular DNA. *Frontiers in Genetics*, *15*. https://doi.org/10.3389/fgene.2024.1385150

Zhu, J., Zhang, F., Du, M., Zhang, P., Fu, S., & Wang, L. (2017). Molecular characterization of cell-free eccDNAs in human plasma. *Scientific Reports*, *7*(1), Artigo 1. https://doi.org/10.1038/s41598-017-11368-w

Zuo, S., Yi, Y., Wang, C., Li, X., Zhou, M., Peng, Q., Zhou, J., Yang, Y., & He, Q. (2022). Extrachromosomal Circular DNA (eccDNA): From Chaos to Function. *Frontiers in Cell and Developmental Biology*, *9*. https://www.frontiersin.org/articles/10.3389/fcell.2021.792555