UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE

CURSO DE MESTRADO EM CIÊNCIAS DA SAÚDE

Dissertação de Mestrado

Comparação de pipelines computacionais para montagem de genomas de procariotos com dados de *Nanopore*

Adriano Terra da Roza

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciências da Saúde da Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtenção do grau de Mestre em Ciências da Saúde

Orientador: Prof. Dr. Karina dos Santos Machado

Rio Grande, 2026

Ficha Catalográfica

R893c Roza, Adriano Terra da

Comparação de pipelines computacionais para montagem de genomas de procariotos com dados de *Nanopore* / Adriano Terra da Roza – Rio Grande - RS, 2026.

112f.

Dissertação (Mestrado em Ciências da Saúde) - Universidade Federal do Rio Grande, Rio Grande, 2026.

Orientador(a): Drª Karina dos Santos Machado

1. Genômica 2. organismos bacterianos 3. Sequenciamento 4. Nanopore. I.Machado,Karina dos Santos, II. Título.

CDU 616.9:575.113

Catalogação na Fonte: Bibliotecária responsável: Rita Moraes CRB10/2506

UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE
MESTRADO EM CIÊNCIAS DA SAÚDE

ATA DA SESSÃO DE DEFESA PÚBLICA DE DISSERTAÇÃO DE MESTRADO EM CIÊNCIAS DA SAÚDE

ATA

A banca examinadora, designada pela Portaria nº 428/2026 de vinte e três de fevereiro de dois mil e vinte e seis, em sessão presidida e registrada pela orientadora Profa. Dra. Karina dos Santos Machado, reuniu-se no dia vinte e seis de fevereiro de dois mil e vinte e seis, às quatorze horas, na sala de reuniões do C3, Carreiros, FURG e em sala virtual, para avaliar a Dissertação de Mestrado do Programa de Pós-Graduação em Ciências da Saúde, intitulada: "COMPARAÇÃO DE PIPELINES COMPUTACIONAIS PARA MONTAGEM DE GENOMA DE PROCARIOTOS COM DADOS DE NANOPORE" do mestrando Adriano Terra da Roza. Para o início dos trabalhos, a Senhora Presidente procedeu à abertura oficial da sessão, com a apresentação dos membros da banca examinadora. A seguir, prestou esclarecimentos sobre a dinâmica de funcionamento da sessão, concedendo o tempo de até 30 (trinta) minutos para a apresentação da dissertação pelo mestrando, que iniciou às 14 horas e 7 minutos e terminou às 14 horas e 40 minutos. Após a apresentação, passou a palavra aos membros da banca examinadora, para que procedessem à arguição e apresentassem suas críticas e sugestões. Ao término dessa etapa de avaliação, de acordo com os membros da banca examinadora, a dissertação de mestrado avaliada foi Aprovada.

Rio Grande, 26 de fevereiro de 2026.

gov.br Documento assinado digitalmente
KARINA DOS SANTOS MACHADO
Data: 26/02/2026 19:20:52-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Karina dos Santos Machado (Orientadora – FURG)

gov.br Documento assinado digitalmente
LAURIVAL ANTONIO VILAS BOAS
Data: 26/02/2026 19:44:19-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Laurival Antonio Vilas-Boas (Externo – UEL)

gov.br Documento assinado digitalmente
GABRIELA PASQUALIM
Data: 27/02/2026 09:12:18-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Gabriela Pasqualim (Titular – FURG)

gov.br Documento assinado digitalmente
RAIZA DOS SANTOS AZEVEDO
Data: 27/02/2026 09:49:18-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Raíza dos Santos Azevedo (Titular - FURG)

gov.br Documento assinado digitalmente
ADRIANO TERRA DA ROZA
Data: 27/02/2026 10:31:41-0300
Verifique em https://validar.iti.gov.br

CIENTE: _____

Mestrando Adriano Terra da Roza

RESUMO

DA ROZA, Adriano Terra. **Comparação de pipelines computacionais para montagem de genomas de procariotos com dados de *Nanopore***. 2026. 114 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Ciências da Saúde. Universidade Federal do Rio Grande - FURG, Rio Grande.

Desde os primeiros métodos de sequenciamento de DNA, os avanços tecnológicos têm impulsionado a genômica e ampliado aplicações na medicina e na biotecnologia. Entre as tecnologias atuais, o sequenciamento *nanopore* destaca-se pela geração de leituras longas e pela flexibilidade operacional, exigindo, contudo, etapas computacionais robustas para obtenção de montagens genômicas de alta qualidade.

Este trabalho teve como objetivo comparar pipelines computacionais para montagem de genomas de procariotos a partir de dados *Nanopore*. Foi realizada uma revisão sistemática da literatura para identificar as ferramentas mais utilizadas nas etapas de *basecalling*, controle de qualidade, filtragem, montagem e polimento. Com base nesses achados, foi estruturado um fluxo padronizado aplicado a seis organismos bacterianos (*Acinetobacter pittii, Haemophilus haemolyticus, Klebsiella pneumoniae, Shigella sonnei, Staphylococcus aureus e Stenotrophomonas maltophilia*), totalizando 72 combinações por organismo.

As configurações foram comparadas por métricas de completude (BUSCO), contiguidade (N50) e precisão (indels), além do desempenho computacional. O modelo sup de *basecalling* apresentou consistentemente os melhores resultados, com valores de completude (BUSCO) variando de 93,5% a 99,6%, em comparação a valores inferiores nos demais modelos. Em termos de contiguidade, os melhores fluxos atingiram valores de N50 de até 5.134.685 bp, com montagens frequentemente próximas ao tamanho esperado dos genomas. A etapa de correção de erros com Medaka, isoladamente ou combinada com Racon, promoveu redução significativa na taxa de *indels*, com valores mínimos de até 3,48 *indels*/100 kb, em contraste com valores superiores a 100 *indels*/100 kb em cenários menos otimizados. As estratégias de filtragem apresentaram impacto limitado sobre as métricas avaliadas. De forma geral, os resultados indicam que a combinação do modelo sup com polimento baseado em Medaka constitui a abordagem mais eficiente e precisa para a montagem de genomas bacterianos a partir de dados de sequenciamento Nanopore.

Palavras-chave: Genômica, organismos bacterianos, sequenciamento, nanopore.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CNNs	Redes Neurais Convolucionais (*Convolutional Neural Networks*)
CQ	Controle de Qualidade
DBG	Grafos de Bruijn (*de-Bruijn-Graph*)
ddNTPs	Didesoxinucleotídeos trifosfatados (Terminação de cadeia)
DNA	Ácido Desoxirribonucleico (*Deoxyribonucleic Acid*)
dNTPs	Desoxirribonucleotídeos trifosfatados
HMMs	Modelos Ocultos de Markov (*Hidden Markov Models*)
mRNA	RNA Mensageiro
NGS	Sequenciamento de Próxima Geração (*Next Generation Sequencing*)
OLC	Sobreposição-Layout-Consenso (*Overlap-Layout-Consensus*)
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Reação em Cadeia da Polimerase (*Polymerase Chain Reaction*)
RNA	Ácido Ribonucleico (*Ribonucleic Acid*)
RNNs	Redes Neurais Recorrentes (*Recurrent Neural Networks*)
SMRT	Sequenciamento de Molécula Única em Tempo Real (*Single Molecule Real-Time*)
SNVs	Variantes de Nucleotídeo Único (*Single Nucleotide Variants*)
ssDNA	DNA de Fita Simples (*Single-stranded DNA*)
SVs	Variantes Estruturais (*Structural Variants*)

SUMÁRIO

1 INTRODUÇÃO

O sequenciamento de ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA) é uma ferramenta importante na pesquisa biológica e médica, sendo essencial na genômica (DNA) e na transcriptômica (RNA) [106]. Esse processo é resultado de uma série de procedimentos bioquímicos que têm como principal finalidade a determinação da sequência de bases nitrogenadas do DNA, que são: adenina (A), guanina (G), citosina (C) e timina (T), podendo também, em determinadas abordagens, permitir a detecção de modificações químicas nas bases, como metilações [139].

Um progresso considerável foi feito desde que os primeiros métodos para identificar nucleotídeos em cadeias de DNA foram propostos na década de 1970, incluindo o método químico de Maxam e Gilbert [97] e o método de terminação de cadeia desenvolvido por Sanger e Coulson [130]. Este período foi marcado por pesquisas que tinham por objetivo decifrar os primeiros genes e depois genomas inteiros, gerando o campo da genômica [1].

O desenvolvimento de novas tecnologias e sensores tem impulsionado a evolução das plataformas de sequenciamento, resultando em sucessivas gerações de sequenciadores com maior qualidade de leitura, maior volume de material genético processado por ensaio, redução do tempo de execução e melhor custo-benefício [52]. Nesse cenário, a genômica consolidou-se como uma área central para a medicina moderna, evidenciada tanto pelo impacto do Projeto Genoma Humano [1] quanto, mais recentemente, pela rápida geração de dados e aplicações clínicas durante a pandemia de COVID-19, quando o sequenciamento e a análise genômica contribuíram para o desenvolvimento e a implementação das vacinas de mRNA contra o SARS-CoV-2 [114].

O sequenciamento genômico constitui, portanto, uma ferramenta essencial, embora ainda em constante evolução, para a identificação, o diagnóstico e o acompanhamento de doenças, além de desempenhar papel estratégico em aplicações biotecnológicas. Essa abordagem permite a caracterização de microrganismos, a investigação de mecanismos de resistência e virulência e o desenvolvimento de novos produtos e processos industriais, ampliando as possibilidades de inovação em saúde e biotecnologia [137].

Os métodos descobertos por Sanger e Alan R. Coulson [130], e Maxam e Gilbert

[97] na década de 70, que possibilitou o sequenciamento do DNA, também desencadeou uma sequência de inovações tecnológicas. Tais inovações possibilitaram o sequenciamento de genomas completos, requerendo pequenas quantidades de DNA e eliminando as etapas de clonagem bacteriana, anteriormente necessárias [95]. A partir disso, vários projetos-chave ajudaram a impulsionar a área nos anos 80, como por exemplo: o sequenciamento dos vírus bacterianos phi X174 [127, 128] e lambda [129], o vírus animal SV40 [47] e da mitocôndria humana [7]. Esses projetos comprovaram a viabilidade de montar pequenos fragmentos de sequência em genomas completos e demonstraram o valor da caracterização abrangente do conteúdo gênico dos organismos, contribuindo para a proposta do "Projeto Genoma Humano". O sequenciamento do genoma humano foi uma conquista histórica que revolucionou a compreensão da biologia humana, permitindo avanços no entendimento da base genética de características e doenças, bem como da história evolutiva [53, 54]. Esses avanços estavam relacionados à capacidade de que visões globais dos genomas poderiam acelerar muito a pesquisa biomédica, permitindo que os pesquisadores abordassem problemas de forma abrangente e não direcionada por hipóteses prévias; e que a criação de tais visões globais exigiria um esforço comunitário na construção de infraestrutura, diferente de qualquer tentativa anterior na pesquisa biomédica [1].

Há atualmente diferentes tecnologias de sequenciamento. Plataformas de sequenciamento de leitura curta produzem leituras de centenas de pares de bases de comprimento. As principais plataformas de leitura curta são os métodos Sanger e *Next Generation Sequencing (NGS)*. O método Sanger, desenvolvido na década de 1970, é baseado na terminação de cadeia por dideoxinucleotídeos (ddNTPs) e possui alta precisão, no entanto, sua baixa escalabilidade limita sua aplicação a fragmentos pequenos, tornando-o limitado para aplicações em larga escala, como o sequenciamento de genomas completos [130]. Com a introdução do NGS, tecnologias como Illumina revolucionaram o campo, permitindo o sequenciamento massivo de milhões de fragmentos curtos simultaneamente, reduzindo custos e aumentando a velocidade, embora a dependência de montagem de leituras curtas (100-300 pb) possa dificultar a resolução de regiões repetitivas e variantes estruturais. [57].

As tecnologias ONT e PacBio permitem a leitura de fragmentos muito maiores (10 kb e até 100 kb no caso da ONT), sendo altamente vantajosas para montagem *de novo* e análise de variantes complexas. Nos últimos anos, tem se popularizado o sequenciamento *Oxford Nanopore Techmologies* (ONT). Esse sequenciador fornece leituras longas, portabilidade e análise em tempo real [131]. O equipamento MinION, da *Nanopore Oxford Technologies*, lançado em 2014, consiste em detectar de forma ativa diretamente os nucleotídeos de uma fita simples de DNA, pois essa molécula atravessa um nanoporo constituído por proteínas, o qual é estabilizado em uma membrana eletricamente resistente [19]. Uma voltagem é repassada a essa membrana e, desta maneira, os sensores

detectam na corrente iônica pelos nucleotídeos que ocupam o poro em tempo real à medida que a molécula de DNA se desloca, já que ela interrompe a voltagem através do canal [57]. Uma exonuclease é utilizada para que o DNA se mantenha em fita simples e para a clivagem de nucleotídeos [31]. Esse método de sequenciamento apresenta múltiplas vantagens em relação aos outros já existentes, uma dessas vantagens é o tamanho muito menor em relação aos outros sequenciadores o que permite sua utilização em diferentes contextos laboratoriais e aplicações descentralizadas [168]. Com isso, é possível identificar que ao longo do tempo e com o desenvolvimento da tecnologia, esse processo foi sendo modificado e aperfeiçoado, buscando melhor qualidade e menor tempo de sequenciamento [137].

Quando comparamos os métodos de leitura curta (Sanger, Illumina) com os métodos de leitura longa (PacBio, *Nanopore*), essas tecnologias apresentam taxas de erro mais elevadas em comparação com *NGS*, necessitando de estratégias de polimento para aumentar a precisão das sequências. Plataformas baseadas em NGS, como Illumina, apresentam taxas de erro tipicamente inferiores a 0,1%, enquanto tecnologias de leitura longa, como ONT e PacBio, historicamente apresentaram taxas entre 5–15%. Com químicas recentes e modelos de basecalling mais novos, muitos estudos relatam valores na faixa de 2–5%, mas isso depende fortemente do modelo, do tipo de poro e do pipeline. [158, 124].

Como resultado desses avanços no sequenciamento de genes e genomas, uma elevada quantidade de dados passou a ser gerada, o que demandou o desenvolvimento de recursos computacionais capazes de processar, armazenar e analisar essas informações em larga escala [35, 23]. Nesse contexto, surgiram bancos de dados públicos, como o GenBank[1] [34], um repositório público que, na sua Release 269.0 (dezembro de 2025), contém cerca de 49,7 trilhões de pares de bases de sequências distribuídos em aproximadamente 12,3 bilhões de registros de sequências, representando um crescimento contínuo dos dados públicos de nucleotídeos. O intercâmbio diário de dados com o *European Nucleotide Archive* e o Banco de Dados de DNA do Japão (DDBJ) garante uma cobertura mundial [132]. Desde a sua criação há 40 anos, o GenBank foi pioneiro nos princípios da ciência aberta e do compartilhamento de dados, e os apoia como descrito pelos princípios FAIR (permitido, acessível, interoperável e reutilizável) [160]. Como banco de dados público, o GenBank preserva o registro científico, a integridade desse registro e o conhecimento associado a dados de sequência que os pesquisadores geram ao longo do tempo através da reutilização desses dados.

A montagem de um genoma é um processo bioinformático que visa reconstruir a sequência original do DNA de um organismo a partir de pequenos fragmentos lidos por um sequenciador. Esse fluxo de trabalho inicia-se com a extração e coleta de material biológico purificado, seguida pelo preparo de bibliotecas e o sequenciamento propriamente dito. No caso da tecnologia ONT, o processamento exige uma série de etapas

[1] https://www.ncbi.nlm.nih.gov/genbank/

computacionais intensivas: primeiro, o *basecalling* converte os sinais elétricos brutos (arquivos pod5 ou fast5) em sequências de nucleotídeos (fastq), utilizando modelos de redes neurais como Guppy ou Dorado (em níveis de precisão Fast, HAC ou SUP) [111, 110]. Em seguida, os dados passam por um rigoroso controle de qualidade e filtragem (ex: NanoPlot e Filtlong) para remover *reads* curtas ou de baixa qualidade, antes de serem submetidos a algoritmos de montagem de novo, como o Flye, que é responsável por organizar os fragmentos de forma contínua [157]. Para corrigir erros inerentes à tecnologia de poros, o rascunho do genoma é refinado por etapas de polimento (*polishing*) com ferramentas como Racon [151] (baseado em alinhamento) e Medaka [112] (baseado em modelos neurais específicos para a *flow cell*), garantindo que a sequência final atinja alta acurácia e completude biológica.

Em paralelo aos avanços nos métodos de sequenciamento e análise de dados genômicos, doenças bacterianas continuam a ser uma causa importante de morbidade e mortalidade entre humanos e animais. Bactérias patogênicas incluem uma ampla gama de organismos que empregam mecanismos variados na patogênese [144]. O planejamento de intervenções terapêuticas contra doenças bacterianas requer uma boa compreensão dos mecanismos utilizados por esses patógenos para causar doenças [143]. O advento do sequenciamento do genoma, juntamente com avanços na análise bioinformática, proporcionou conhecimentos importantes sobre patógenos bacterianos, incluindo sua evolução, ecologia e mecanismos de patogenicidade. Os genomas bacterianos são dinâmicos e estão expostos a diversos eventos genéticos, incluindo mutações, duplicações, inversões, transposições, recombinações, inserções e deleções. A aquisição de genes por transferência horizontal desempenha papel relevante na adaptação desses organismos, podendo conferir novas capacidades metabólicas, como resistência a antibióticos e fatores de virulência [72]. Nesse contexto, a correta identificação e caracterização do conteúdo gênico dependem diretamente da qualidade das etapas de sequenciamento, montagem e anotação genômica. Limitações nessas etapas podem comprometer a predição da presença ou ausência de genes, afetando a interpretação da dinâmica evolutiva e funcional desses organismos [145].

Os genomas de procariotos constituem um modelo particularmente adequado para a avaliação de pipelines de sequenciamento e montagem genômica [6]. Em geral, apresentam tamanhos relativamente reduzidos, variando aproximadamente entre 0,5 e 10 Mb (Megabases), alta densidade gênica e organização estrutural mais simples quando comparados aos genomas eucariotos. Esses genomas são majoritariamente compostos por regiões codificadoras, frequentemente correspondendo a mais de 85% do total do genoma, e caracterizam-se pela ausência ou raridade de íntrons, o que resulta em uma arquitetura genômica compacta [11]. Além disso, muitos procariotos possuem um único cromossomo circular, frequentemente acompanhado por plasmídeos, facilitando etapas de montagem e anotação genômica [93]. Essas características tornam os genomas procariotos especi-

almente apropriados para análises comparativas de desempenho, acurácia e robustez de diferentes estratégias computacionais aplicadas ao sequenciamento por ONT [123].

Nesse contexto de expansão das tecnologias de sequenciamento de leitura longa, a cresecente disponibilidade dessas tecnologias reforça a necessidade de avaliar criticamente pipelines computacionais e ferramentas bioinformáticas adequadas ao processamento e à análise de dados gerados por sequenciamento *Nanopore*.

Diante da rápida evolução das ferramentas de bioinformática e da diversidade de algoritmos disponíveis para o processamento de leituras longas, torna-se necessário estabelecer critérios claros de escolha para otimizar a acurácia dos genomas montados. Além disso, estudos [56, 165] destacam a falta de consenso na literatura sobre as estratégias de montagem de genomas procariotos mais adequadas, utilizando dados de sequenciamento de *nanopore*. Nesse contexto, o presente trabalho tem como objetivo comparar o desempenho e a qualidade de diferentes pipelines computacionais utilizados na montagem de genomas de procariotos a partir de dados de sequenciamento gerados por plataformas ONT. Busca-se, especificamente, avaliar o impacto de diferentes modelos de *basecalling* e etapas de polimento sobre a continuidade e a completude biológica das sequências consensos, fornecendo subsídios para a seleção de fluxos de trabalho mais eficientes em termos de custo-benefício computacional.

1.1 Objetivo Geral

Comparar o desempenho e a qualidade de diferentes pipelines computacionais utilizados na montagem de genomas de procariotos a partir de dados de sequenciamento gerados por plataformas da *Oxford Nanopore Technologies* (ONT).

1.2 Objetivos Específicos

- Identificar pipelines computacionais e ferramentas de bioinformática utilizadas nas diferentes etapas do processamento de dados do sequenciamento por ONT, com foco na montagem de genomas de procariotos.
- Avaliar, comparativamente, o desempenho dos *pipelines* selecionados quanto a métricas de qualidade genômica, incluindo contiguidade, cobertura, precisão e completude das montagens.
- Desenvolver um *script* automatizado para executar diferentes pipelines de análises desses dados
- Aplicar diferentes *pipelines* a conjuntos de dados públicos de sequenciamento de genomas bacterianos, gerados por tecnologias ONT, visando analisar sua robustez em distintos contextos biológicos.
- Comparar as montagens obtidas com genomas de referência, quando disponíveis, a

fim de verificar a acurácia estrutural e a fidelidade das sequências montadas.

1.3 Motivação

A análise genômica tem sido revolucionada devido ao avanço das tecnologias de sequenciamento NGS, especialmente plataformas de leitura longa, como *Nanopore*. Tais avanços vêm permitindo uma visão mais completa de genomas de procariotos, incluindo elementos repetitivos e plasmídeos. Contudo, a montagem de genomas de procariotos a partir de dados de leitura longa com *nanopore* ainda apresenta desafios, como alta taxa de erro e a escolha de um *pipleine* bioinformático adequado. Estudos recentes [56, 165] destacam a falta de consenso na literatura sobre as estratégias de montagem de genomas procariotos mais adequadas, utilizando dados de sequenciamento de *nanopore*. Além disso, esta pesquisa está alinhada com as atividades desenvolvidas pelo Grupo de Pesquisa Combi-Lab, que atua em biologia computacional, ampliando o escopo de atuação em bioinformática aplicada à microbiologia [30]. Na agricultura e pecuária, contribui para programas de melhoramento genético de plantas e animais, auxiliando na identificação de características desejáveis, como resistência a doenças ou tolerância a condições ambientais adversas.

Nesse contexto, o presente estudo também se alinha à Agenda 2030 para o Desenvolvimento Sustentável, especialmente aos Objetivos de Desenvolvimento Sustentável (ODS) 3 – Saúde e Bem-Estar e 4 – Educação de Qualidade, conforme as diretrizes do Programa de Pós-Graduação em Ciências da Saúde (PPGCS). A contribuição ao ODS 3 ocorre de forma indireta, uma vez que a sistematização e avaliação de pipelines computacionais para montagem de genomas bacterianos favorecem o avanço de análises genômicas aplicadas ao estudo de patógenos, vigilância genômica e resistência antimicrobiana, áreas relevantes para a saúde pública. Em relação ao ODS 4, o trabalho contribui para a formação de recursos humanos qualificados em bioinformática e genômica, fortalecendo a capacitação técnica, a produção científica e a difusão do conhecimento no âmbito acadêmico, além de apoiar a consolidação de competências em biologia computacional aplicadas à microbiologia.

1.4 Organização do Texto

No Capítulo 2 será apresentado o Referencial teórico, onde os principais conceitos sobre sequenciamento, principalmente *nanopore*, serão abordados. Entre eles, estão a linha do tempo e a evolução dessas tecnologias ao longo das gerações, os tipos de arquivos utilizados no processo de sequenciamento, e os principais passos envolvidos na montagem de genomas de procariotos com dados de *nanopore*. Também é apresentada uma seção que indica as principais aplicações dessa tecnologia.

O Capítulo 3 demonstra a Metodologia utilizada neste projeto para realizar a etapa de revisão sistemática da literatura, indicando um passo a passo seguido para desenvolver

uma revisão, que resultou na leitura completa de 53 artigos que comparam diferentes pipelines computacionais para análise de dados de sequenciamento *nanopore*, especificamente para organismos bacterianos. A partir disso, foram identificados os principais fluxos e ferramentas empregados nos estudos. Em seguida, é descrito como ocorreu a instalação dessas ferramentas em dois ambientes distintos, *desktop* e cluster. Nesse capítulo também é abordado como ocorreu a obtenção de dados públicos de organismos procariotos. A partir desses dados ocorreu a preparação de um *script* para automatização do processo e a execução desses fluxos. Por fim é descrito como foi feita a análise desses resutados.

O Capítulo 4 apresenta os Resultados obtidos após a leitura dos 53 artigos selecionados na revisão. Um levantamento da frequência das ferramentas utilizadas para cada etapa de um pipeline para dados de sequenciamento ONT. Nesse mesmo capítulo é apresentado um fluxograma evidenciando a ordem dessas etapas, e os arquivos de entrada e saída para cada uma delas. Com isso, é feita uma análise dos fluxos e ferramentas empregados nesse processo, e a partir disso, 72 fluxos diferentes são analizados para cada organismo. Por fim, é feita uma comparação dos melhores fluxos entre os organismos comparando a qualidade e o custo computacional nos fluxos de montagem.

O Capítulo 5 é a conclusão do trabalho, onde é discutido os principais resultados oriundos dessas 72 combinações de estratégias para o processo de análise dos dados de 6 organismos diferentes.

2 REFERENCIAL TEÓRICO

Neste capítulo, são apresentados os principais sequenciadores, com enfoque no funcionamento do *Nanopore*. Serão apresentados os principais formatos de arquivo e etapas envolvidas no sequenciamento com o Nanopore. Por fim, são descritas algumas aplicações para o uso de sequenciamento com o Nanopore.

2.1 Sequenciadores

A tecnologia de sequenciamento passou por três estágios de desenvolvimento. Diferentes tecnologias têm sido propostas ao longo do tempo para realizar o sequenciamento de genes e genomas, conforme resume a figura 1.

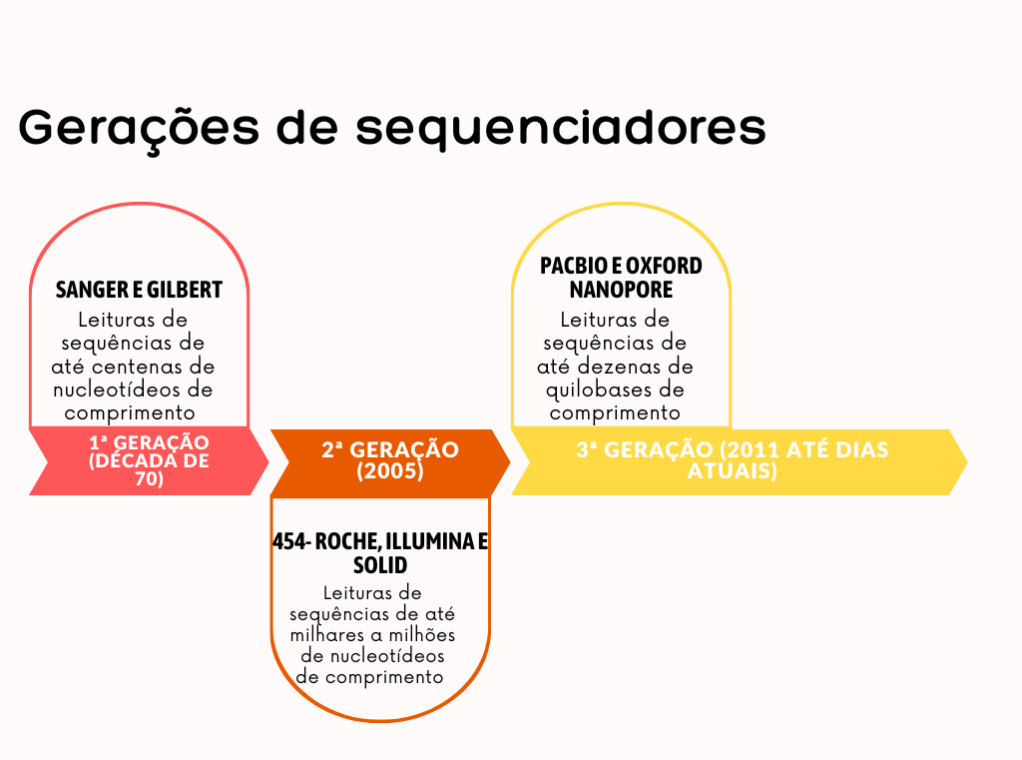

Figura 1: Comparação entre as gerações de sequenciadores quanto ao comprimento das leituras de sequências.

2.1.1 Sequenciamento Sanger (1ª Geração)

O método proposto por Sanger, desenvolvido na década de 1970, é amplamente reconhecido como a principal tecnologia de sequenciamento de primeira geração [161]. Paralelamente, o método químico de Maxam e Gilbert também foi proposto no mesmo período, embora tenha tido uso mais limitado devido à sua complexidade e ao uso de reagentes tóxicos. Segundo Heather et al. (2016) [61], essas abordagens iniciais permitiam a obtenção de leituras relativamente curtas, variando de dezenas a algumas centenas de nucleotídeos, dependendo do método e das condições experimentais. A estratégia de sequenciamento *shotgun*, sugerida por Staden em 1979 [141] baseia-se na fragmentação aleatória do DNA em múltiplos segmentos menores, que são posteriormente sequenciados e remontados computacionalmente. Essa abordagem foi utilizada na montagem de genomas de novo, como no caso do bacteriófago lambda já em 1982 [129]. Após o sequenciamento dos fragmentos, algoritmos de montagem identificam regiões de sobreposição entre as leituras, permitindo a reconstrução da sequência original. [141]. Em 1987, surgiram os primeiros sistemas automatizados de sequenciamento pelo método de Sanger, baseados em detecção por fluorescência, descritos em trabalhos clássicos como os de Hood et al. e Connell et al. [140, 33]. Esses sistemas iniciais permitiam a obtenção de leituras de algumas centenas de bases por reação, sendo que leituras próximas a 1000 pb passaram a ser alcançadas posteriormente, com o aprimoramento das plataformas e das condições experimentais.

Nesse método, o processo começa com a amplificação do DNA alvo por PCR, seguida da reação de sequenciamento, que envolve a adição de nucleotídeos normais (dNTPs) e nucleotídeos modificados, os didesoxinucleotídeos (ddNTPs), que causam a terminação da cadeia de DNA quando incorporados. Cada ddNTP é marcado com um fluoróforo, permitindo a detecção das diferentes bases (A, T, C, G). Após a terminação da cadeia, os fragmentos de DNA são separados por eletroforese capilar, e a sequência é determinada com base na ordem dos nucleotídeos [130]. Em resumo, os *reads* no sequenciamento pelo método de Sanger são gerados pela incorporação aleatória e controlada de ddNTPs durante a síntese do DNA, criando uma população de fragmentos de diferentes tamanhos, que são então separados e lidos com base na fluorescência dos ddNTPs terminais. Uma *read* é a unidade básica de informação gerada por um sequenciador. Imagine que o genoma de uma bactéria é um livro inteiro. O sequenciador não consegue ler o livro do começo ao fim de uma vez só, em vez disso, ele ”lê”pequenos trechos aleatórios do texto. Cada um desses trechos individuais de sequência de DNA (as letras A, C, T e G) que a máquina entrega é chamado de *read*. Essa técnica é altamente precisa e é ideal para sequenciamento de genes individuais ou pequenas regiões do genoma. No entanto, apresenta limitações em termos de custo, tempo e capacidade de gerar grandes volumes de dados, especialmente quando comparada com as tecnologias de sequenciamento NGS.

2.1.2 Next Generation Sequencing (NGS) (2ª Geração)

Os métodos de sequenciamento de segunda geração (2005) revolucionaram o sequenciamento de DNA, permitindo o sequenciamento simultâneo de milhares a milhões de fragmentos de DNA.Esses métodos diferem do sequenciamento Sanger tradicional em sua capacidade de realizar sequenciamento em paralelo. Entre essas tecnologias, destacam-se plataformas como Illumina, que atualmente representam o padrão mais amplamente utilizado, enquanto outras, como 454-Roche e SOLiD, tiveram importância histórica no desenvolvimento do sequenciamento de nova geração [117]. O NGS propõe uma nova maneira de sequenciamento constituindo várias abordagens que dependem da fusão da preparação do molde, determinação da ordem das bases, alinhamento de sequências e montagem do genoma [99].

Para preparar a biblioteca no sequenciamento illumina, o DNA genômico é fragmentado em pedaços pequenos. Adicionam-se adaptadores sintéticos às extremidades dos fragmentos, que permitirão sua fixação em uma superfície sólida (*flow cell*) e o reconhecimento pelos *primers*. Em seguida, os fragmentos de DNA ligam-se a *flow cell*, que contém oligonucleotídeos complementares aos adaptadores. O DNA ligado forma uma ponte com outra região da *flow cell*, e através de sucessivas rodadas de amplificação por polimerase, formam-se *clusters* clonais de fragmentos idênticos, aumentando o sinal detectável, e cada *cluster* é sequenciado simultaneamente. Durante cada ciclo, nucleotídeos marcados com fluoróforos e bloqueadores 3' são incorporados pela DNA polimerase. Após a incorporação de cada base, um laser excita o fluoróforo, e uma câmera registra o sinal fluorescente, identificando a base adicionada, e em seguida, o bloqueador é removido quimicamente, permitindo a adição do próximo nucleotídeo no próximo ciclo. Esse processo é repetido por centenas de ciclos, gerando *reads* curtos (tipicamente de 100 a 300 pb). Por fim, o software da plataforma converte os sinais de fluorescência em sequências de bases nitrogenadas, gerando leituras (*reads*) correspondentes aos fragmentos de DNA sequenciados [57].

Uma grande vantagem do NGS sobre os métodos tradicionais de detecção de mutação é a capacidade de sequenciar múltiplos genes e destacar milhões de variantes simultaneamente. Outras vantagens incluem entrada mínima de DNA, tempo de resposta mais rápido; o NGS revolucionou a velocidade da descoberta genética e genômica e avançou nossa compreensão dos mecanismos moleculares da doença e potenciais opções de tratamento. Avanços técnicos nesses métodos de sequenciamento (substituindo a radiomarcação por corantes fluorescentes e eletroforese em gel por eletroforese capilar) introduziram automação nas abordagens de sequenciamento e aumentaram o volume de sequenciamento para vários milhares de pares de bases em uma única execução [89].

2.1.3 Oxford Nanopore Technologies (ONT) (3ª Geração)

As tecnologias de sequenciamento de terceira geração representam os mais recentes avanços no sequenciamento de DNA, oferecendo novas abordagens que superam as limitações das gerações anteriores. Estas tecnologias permitem o sequenciamento de leitura longa de fragmentos de DNA muito maiores em comparação com métodos anteriores. Os exemplos incluem o sequenciamento PacBio, que usa uma abordagem de *single molecule real-time* (SMRT) com nucleotídeos marcados com fluorescência, permitindo o sequenciamento de leitura longa de fragmentos de DNA de até dezenas de quilobases de comprimento [124]. Outra tecnologia é o sequenciamento *Oxford Nanopore*, baseado na tecnologia *nanopore*, onde uma molécula de DNA de fita simples passa através de um *nanopore* e as mudanças na corrente elétrica são medidas para determinar a sequência do DNA. O sequenciamento Oxford Nanopore fornece leituras longas, portabilidade e análise em tempo real [131]. Com isso, é possível identificar que ao longo do tempo e com o desenvolvimento da tecnologia esse processo foi sendo modificado e aperfeiçoado buscando melhor qualidade e menor tempo de sequenciamento [137]. Essa tecnologia será detalhada na seção 2.2.

2.2 Nanopore - Funcionamento

Oxford Nanopore Technologies (ONT) é uma tecnologia de sequenciamento de molécula única baseada em nanoporos. A terceira geração da tecnologia de sequenciamento de genes foi a combinação entre a tecnologia de engenharia genética e a tecnologia auxiliada por computador. A ONT utilizou de forma inovadora a tecnologia de nanoporos para determinar as sequências de bases a partir das variações na corrente elétrica geradas quando o DNA atravessa o nanoporo. [102].

O primeiro protótipo, MinION, foi lançado em 2014 [10]. A plataforma atualizada, PromethION, foi lançada em 2015 com rendimento aprimorado. Duas versões do PromethION, denominadas ProtheION 24 e 48, integram 24 e 48 tanques de fluxo, respectivamente. Com o número crescente de tanques de fluxo em comparação ao MinION, o sistema PromethION pode produzir até 7,6 Tb de dados, enquanto o MinION pode gerar apenas 50 Gb em 72 horas de operação. O sistema de reação para sequenciamento de nanoporos é realizado em uma célula de fluxo, na qual dois compartimentos cheios de solução iônica foram separados por membranas contendo 2048 (MinION) ou 12.000 (PromethION) nanoporos [87]. O sequenciamento por ONT baseia-se na passagem de moléculas de DNA em fita simples (ssDNA) através de um nanoporo biológico, geralmente constituído por proteínas formadoras de poro inseridas em uma membrana sintética eletricamente resistente. Associada ao nanoporo, uma proteína motora (*motor protein*) é responsável por controlar a velocidade de translocação da molécula de DNA através do poro, permitindo que o sinal elétrico seja medido de forma estável. À medida que o DNA

atravessa o nanoporo, diferentes combinações de nucleotídeos presentes simultaneamente no interior do poro alteram a corrente iônica de maneira característica, gerando um sinal contínuo que é posteriormente interpretado por algoritmos de basecalling [158]. A imagem 2 mostra um esquema gráfico desse processo.

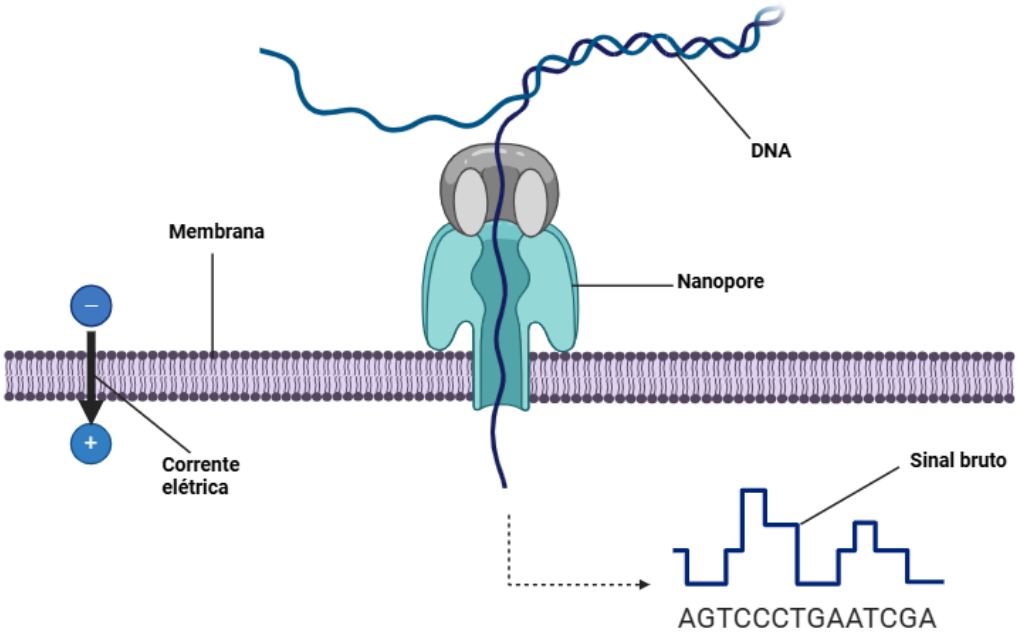

Figura 2: Esquema do processo de sequenciamento ONT

Nos protocolos atualmente utilizados, o sequenciamento por nanoporo é realizado predominantemente a partir de moléculas de DNA em fita simples. Em versões iniciais da tecnologia, estratégias baseadas em leituras bidirecionais (2D reads), nas quais ambas as fitas de uma molécula dupla eram sequenciadas consecutivamente, foram exploradas com o objetivo de aumentar a precisão. No entanto, essas abordagens foram descontinuadas em favor de métodos mais robustos baseados em modelos de aprendizado profundo aplicados a leituras de fita simples [70]

A ONT apresenta vantagens relacionadas à capacidade de sequenciamento de moléculas longas de DNA, alta velocidade de geração de dados e potencial redução de custos. Utilizando ONT, o comprimento das leituras pode atingir valores elevados, com relatos de sequenciamento de fragmentos superiores a 1 milhão de pares de bases.[100].

A seguir serão discutidos os formatos de arquivos utilizados e gerados pelo sequenciamento ONT.

2.2.1 Formatos de arquivos

2.2.1.1 FAST5 e POD5 (ONT)

O arquivo ONT FAST5 armazena leituras de sequenciamento contendo os dados brutos do sinal da série temporal para correntes iônicas e também pode conter a sequência chamada de base. Existem algumas ferramentas disponíveis atualmente especificamente

para visualização do sinal FAST5 e comparações sinal-base, que são úteis para identificar anomalias de sinal em regiões de interesse [90], [115].

O formato ONT POD5 lançado recentemente é um formato de dados mais eficiente, projetado para substituir o FAST5 [98] . Ele difere do FAST5 porque não contém informações de sequência: a sequência chamada de base para arquivos POD5 é armazenada em um arquivo BAM separado.

2.2.1.2 FASTQ

Na área de sequenciamento de DNA, o formato de arquivo FASTQ surgiu como um formato comum para troca de dados entre ferramentas. Ele fornece uma extensão simples para o formato FASTA: a capacidade de armazenar uma pontuação de qualidade numérica associada a cada nucleotídeo em uma sequência [32]. Um arquivo FASTQ consiste em uma ou mais sequências (tipicamente milhões) com cada sequência sendo representada por um identificador, os pares de base de DNA e a qualidade ou confiança de cada par de base individual [13]. Um exemplo do formato FASTQ é mostrado na figura 3.

```
1  @ERR000589.41 EAS139_45:5:1:2:111/1
2  CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
3  +
4  3IIIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
5  @ERR000589.42 EAS139_45:5:1:2:1293/1
6  AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTAAAAGAAAT
7  +
8  IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

Figura 3: Exemplo de um arquivo FASTQ.

Cada leitura é descrita em quatro linhas de texto. A primeira linha, começando com '@', é um identificador de texto arbitrário. O formato é específico da máquina, mas normalmente é um identificador exclusivo construído unindo o nome da célula de execução/fluxo e a localização (coordenadas de faixa, X e Y). A segunda linha contém a sequência de base consistindo de nucleotídeos A, C, G e T com o N ocasional. A terceira linha consiste no caractere '+' opcionalmente seguido por uma cópia do identificador de sequência (geralmente omitido). A quarta e última linha contém as pontuações de qualidade no formato Phred +33 [45], ou seja, cada caractere ASCII representa uma pontuação de qualidade Phred (Q). O valor do caractere é calculado como: ASCII=33+Q, o valor Q é calculado usando a fórmula: $Q = 10 * log_{10}(p)$ onde Q = pontuação de qualidade e p = probabilidade de erro da base chamada. O intervalo possível no Phred+33 é de 33 a 126 [13].

Cálculo exemplo para $Q = 30$:

$ASCII = 33 + 30 = 63$ que corresponde ao caractere '?'

$p = 10^{-Q/10} = 10^{-30/10} = 10^{-3} = 0,001$

Ou seja, há 0,1% de chance de erro na leitura dessa base.

A maioria das ferramentas gera arquivos FASTQ sem quebra de linha da sequência e da *string* de qualidade. Isso significa que cada leitura consiste em exatamente quatro linhas (às vezes linhas muito longas), ideal para um analisador muito simples lidar [32].

2.2.1.3 FASTA

O formato de arquivo FASTA é o formato padrão baseado em texto. Este formato FASTA é um dos mais utilizados para armazenar sequências biológicas, como DNA, RNA e proteínas. Ele foi introduzido no final da década de 1980 com o programa FASTA de busca de similaridade de sequências [116]. Um arquivo FASTA pode conter uma ou múltiplas sequências. Se o arquivo contiver múltiplas sequências, pode ser chamado de arquivo "multi-FASTA" [16]. Os arquivos FASTA devem seguir um formato de duas linhas, o qual pode ser repetido para arquivos contendo várias sequências. Cada entrada em um arquivo FASTA segue uma estrutura simples, como mostra o exemplo da figura 4:

```
>NC_000962.3:1674202-1675011 Mycobacterium tuberculosis H37Rv, complete genome
ATGACAGGACTGCTGGACGGCAAACGGATTCTGGTTAGCGGAATCATCACCGACTCGTCGATCGCGTTTC
ACATCGCACGGGTAGCCCAGGAGCAGGGCGCCCAGCTGGTGCTCACCGGGTTCGACCGGCTGCGGCTGAT
TCAGCGCATCACCGACCGGCTGCCGGCAAAGGCCCCGCTGCTCGAACTCGACGTGCAAAACGAGGAGCAC
CTGGCCAGCTTGGCCGGCCGGGTGACCGAGGCGATCGGGGCGGGCAACAAGCTCGACGGGGTGGTGCATT
CGATTGGGTTCATGCCGCAGACCGGGATGGGCATCAACCCGTTCTTCGACGCGCCCTACGCGGATGTGTC
CAAGGGCATCCACATCTCGGCGTATTCGTATGCTTCGATGGCCAAGGCGCTGCTGCCGATCATGAACCCC
GGAGGTTCCATCGTCGGCATGGACTTCGACCCGAGCCGGGCGATGCCGGCCTACAACTGGATGACGGTCG
CCAAGAGCGCGTTGGAGTCGGTCAACAGGTTCGTGGCGCGCGAGGCCGGCAAGTACGGTGTGCGTTCGAA
TCTCGTTGCCGCAGGCCCTATCCGGACGCTGGCGATGAGTGCGATCGTCGGCGGTGCGCTCGGCGAGGAG
GCCGGCGCCCAGATCCAGCTGCTCGAGGAGGGCTGGGATCAGCGCGCTCCGATCGGCTGGAACATGAAGG
ATGCGACGCCGGTCGCCAAGACGGTGTGCGCGCTGCTGTCTGACTGGCTGCCGGCGACCACGGGTGACAT
CATCTACGCCGACGGCGGCGCGCACACCCAATTGCTCTAG
```

Figura 4: Exemplo de um arquivo FASTA

- A linha de comentário, que começa com o caractere $>$ e é seguida por informações básicas sobre a sequência. Alguns repositórios possuem exigências de formato (por exemplo, o GenBank), contudo não há um padrão único que se aplique a todos os arquivos FASTA. As informações resumidas geralmente incluem um número de acesso único, o nome do gene, o tipo de sequência e o organismo de origem [16].

- A segunda linha contém as letras que representam a sequência propriamente dita [16].

- Um arquivo FASTA é composto por duas partes: a linha de cabeçalho (que começa com $>$) e a sequência de nucleotídeos ou proteínas. Embora o padrão original não proíba sequências em uma única linha gigantesca, a convenção estabelecida

desde os primórdios da computação biológica (especificamente pelo formato Pearson/FASTA original) sugere um limite. O Limite Padrão: 60 a 80 Caracteres. A maioria dos arquivos FASTA (como os que o Flye ou o Medaka geram) quebram a linha ao atingir entre 60 e 80 caracteres. Isso é devido a legibilidade, pois no passado, os monitores de computador e editores de texto padrão tinham larguras fixas de 80 colunas. Quebrar a linha evitava que o pesquisador tivesse que usar a barra de rolagem horizontal infinitamente. Outro fator é memória de processamento, uma vez que muitos algoritmos de bioinformática antigos liam o arquivo linha por linha. Ter linhas de tamanho controlado evitava estouros de memória RAM ao carregar genomas gigantescos. E por questão de compatibilidade, alguns *softwares* como o NCBI BLAST ou o ClustalW foram desenhados esperando esse formato ”blocagem”[107].

2.2.1.4 Contigs e Scaffolds

Por meio do processo de montagem *de novo*, um genoma é montado computacionalmente a partir de leituras (*reads*) que foram obtidas aleatoriamente e que são sobrepostas. Dependendo de diversos fatores, incluindo a profundidade de cobertura, a metodologia de sequenciamento e a complexidade do genoma, a sequência pode ser montada em um número variável, mas geralmente grande, de *contigs*. Quanto mais fragmentada for a montagem, mais difícil se torna a análise subsequente. Análises da ordem de genes, genômica comparativa ou funcional, bem como investigações sobre padrões de recombinação, todas dependem fortemente de uma montagem com boa continuidade [67].

Devido à complexidade do genoma e às limitações das tecnologias de sequenciamento, os montadores de genoma geralmente produzem primeiro sequências longas fragmentadas (*contigs*). Em seguida, métodos de *scaffolding* são utilizados para inferir as orientações e ordens entre esses *contigs* e para produzir algumas sequências mais longas (*scaffolds*). Após obter um conjunto de *contigs* com a utilização de um montador, a fita de DNA de origem desses *contigs* é desconhecida, assim como suas localizações no genoma. No entanto, muitos tipos de leituras (*reads*) podem conectar diferentes *contigs*, e essas informações podem ser usadas para estimar a distância aproximada entre dois *contigs* e se eles estão na mesma fita. Os métodos de *scaffolding* aproveitam essas conexões para deduzir a ordem e a orientação dos *contigs*. A entrada desses métodos inclui um conjunto de *contigs* e um conjunto de leituras, e sua saída final é um conjunto de *scaffolds*. Um *scaffold* compreende alguns *contigs* ordenados e orientados, e a lacuna entre quaisquer dois *contigs* adjacentes é preenchida com caracteres “N”, que representam bases desconhecidas na sequência. Portanto, o *scaffold* pode tornar a montagem do rascunho mais completa e contígua, sendo uma etapa importante na montagem do genoma [92].

Antes do *scaffolding*, duas questões devem ser consideradas: primeiro a correção

dos *contigs*: geralmente, os *contigs* usados como entrada nos métodos de *scaffolding* podem conter montagens incorretas (*misassemblies* - MA), que podem introduzir conexões incorretas entre *contigs* e complicar o processo de *scaffolding*. Portanto, é importante detectar as MA nos *contigs*, podendo, assim, um *contig* ser dividido em duas partes no local da montagem incorreta [92]. E segundo, a ferramenta de alinhamento: é necessário utilizar uma ferramenta de alinhamento para alinhar as leituras aos *contigs* e produzir as informações de ligação.

Em comparação com as tecnologias de sequenciamento de nova geração (NGS), o sequenciamento de longo alcance, como o sequenciamento baseado em nanoporos da *Oxford Nanopore Technologies* é capaz de produzir *reads* muito mais longas. Na tecnologia de nanoporos, o comprimento da leitura depende do comprimento das moléculas de DNA a serem sequenciadas [149]. O comprimento máximo de leitura obtido por esse método pode ultrapassar 1 Mb [135]. Essas leituras longas podem abranger a maioria das regiões repetitivas do genoma e conectar dois ou mais diferentes, o que melhora significativamente a qualidade do *scaffolding* [5].

Entretanto, a taxa de erro das leituras longas é em torno de 15 % [135] [104], tornando difícil obter alinhamentos precisos entre as leituras longas e os *contigs*. Para essas leituras propensas a erros, uma alta precisão de consenso pode ser alcançada por métodos de correção de erro híbridos ou autocorreção, mas o custo computacional é elevado [164].

2.2.2 Principais Etapas da Análise de dados de um sequenciador Nanopore

Definir claramente as principais etapas, o fluxo adequado (*pipeline*) e os parâmetros das ferramentas utilizadas a cada etapa é o principal objetivo desta proposta. Apesar das vantagens associadas às leituras longas, o sequenciamento por nanoporo apresenta limitações metodológicas importantes. Entre elas destacam-se a presença de erros sistemáticos, a dependência de alta cobertura para obtenção de consenso preciso, a sensibilidade a parâmetros computacionais e a variabilidade de desempenho entre diferentes versões de kits, poros e modelos de *basecalling*. Esses fatores tornam a definição de pipelines padronizados um desafio, reforçando a necessidade de avaliações comparativas e sistemáticas das estratégias computacionais empregadas.

A seguir são apresentadas as principais etapas identificadas até o momento para a análise de dados completa dos dados gerados por um sequenciador *Nanopore*.

2.2.2.1 Basecalling

Basecalling, para dispositivos ONT, é o processo de traduzir o sinal bruto que é gerado durante o sequenciamento em uma sequencia de DNA [123]. Geralmente é o passo inicial para analisar sinais de sequenciamento de *nanopore*. O *basecalling* é uma etapa essencial, responsável principalmente pela saída do sequenciador e pelas bases que ele lê. As ferramentas de *software* que identificam bases a partir de medições de corrente iônica são chamadas de *basecallers* [106].

A ONT fornece pacotes consolidados para *basecalling*, como o Guppy [113]. Atualmente, o *basecalling* de nanopore ainda apresenta uma taxa de erro mais elevada em comparação ao sequenciamento de leitura curta, embora avanços recentes tenham reduzido significativamente esses valores. Estudos reportam taxas de erro na faixa de aproximadamente 2% a 5%, dependendo do modelo de *basecalling*, da química utilizada e da qualidade dos dados. Em contraste, plataformas de leitura curta, como a Illumina HiSeq, apresentam taxas de erro em torno de 0,1% (com a maioria das leituras apresentando Q-score superior a 30). Cada vez mais esforços têm sido direcionados para enfrentar esses desafios e melhorar ainda mais a precisão do *basecalling*. [166].

Uma dificuldade significativa no *basecalling* é que o sinal elétrico em um dado momento é influenciado por múltiplos nucleotídeos presentes simultaneamente no poro. Por exemplo, no poro R9.4 da *Oxford Nanopore Technologies*, aproximadamente cinco nucleotídeos afetam o sinal de corrente a qualquer momento, resultando em 4^5 = 1024 possíveis estados distintos [158].

Diferentemente de outras tecnologias de sequenciamento, o sinal gerado no nanoporo não corresponde à identificação de um único nucleotídeo por vez, mas sim à influência simultânea de um conjunto de nucleotídeos (k-mers) posicionados no interior do poro. Essa característica aumenta substancialmente a complexidade do processo de inferência, pois diferentes combinações de bases podem produzir sinais elétricos semelhantes, contribuindo para erros sistemáticos, especialmente em regiões homopoliméricas [158]

Inicialmente, os algoritmos de *basecalling* utilizavam Modelos Ocultos de Markov (HMMs) para modelar a sequência de eventos e inferir as bases correspondentes. Com o avanço da tecnologia, métodos baseados em aprendizado profundo (*deep learning*), como redes neurais recorrentes (RNNs) e redes neurais convolucionais (CNNs), passaram a ser empregados para melhorar a precisão do *basecalling* . Esses modelos de aprendizado profundo são treinados com dados de sequências conhecidas, permitindo que aprendam a mapear padrões de sinal para sequências de bases com maior precisão.

Todos os *basecallers* modernos usam redes neurais, e essas redes devem ser treinadas usando dados reais. A precisão do *basecaller* pode ser avaliada no nível de leitura (precisão de leitura) ou em termos de precisão da sequência de consenso (precisão de consenso). A precisão de leitura mede a identidade da sequência de leituras de *basecalling* individuais em relação a uma referência confiável. A precisão de consenso mede a identidade de uma sequência de consenso construída a partir de múltiplas leituras sobrepostas originárias da mesma localização genômica. A precisão de consenso geralmente melhora com o aumento da profundidade de leitura, por exemplo, um consenso construído a partir de 10 leituras provavelmente será menos preciso do que um construído a partir de 100 leituras [158].

Além dos algoritmos empregados, a química do nanoporo exerce influência di-

reta na precisão do *basecalling*. Diferentes versões de poros, como R9.4.1 e R10, apresentam geometrias distintas, afetando a resolução do sinal elétrico, especialmente em regiões homopoliméricas. A química R10 foi desenvolvida com o objetivo de melhorar a discriminação de bases nesses contextos, resultando em perfis de erro distintos em relação à R9. Consequentemente, a escolha da química impacta não apenas a taxa de erro observada, mas também a seleção do modelo de *basecalling*, parâmetros de filtragem, estratégias de montagem e etapas de polimento adotadas no pipeline [40].

O *software* Dorado [110] representa a geração mais recente de algoritmos de *basecalling* desenvolvida pela ONT, sucedendo o software Guppy com uma arquitetura otimizada para aceleração via *hardware*, especialmente em GPUs (Unidades de Processamento Gráfico). Diferente de seus antecessores, o Dorado utiliza modelos de redes neurais profundas baseadas em arquiteturas de *Transformers*, o que permite uma tradução mais precisa dos sinais de corrente iônica brutos (arquivos pod5) em sequências de nucleotídeos. Além da velocidade superior, a ferramenta integra funcionalidades avançadas, como a detecção nativa de modificações de bases (metilação) e a capacidade de realizar o *demultiplexing* e o mapeamento de leituras de forma simultânea ao processo de tradução dos sinais [126].

Modelos de Acurácia: *fast*, *hac* e *sup*

A escolha do modelo de *basecalling* no Dorado e no Guppy estabelece um compromisso entre velocidade computacional e precisão das bases (Q-score), sendo categorizado em três níveis principais: *fast*, *hac* (High Accuracy) e *sup* (Super Accuracy). O modelo *Fast* utiliza redes neurais mais simplificadas com menos camadas, priorizando o processamento em tempo real em dispositivos com menor poder computacional, embora apresente uma taxa de erro bruta superior. O modelo *hac*, atualmente o padrão para a maioria das aplicações genômicas, equilibra uma alta taxa de acurácia com um tempo de processamento moderado. Por fim, o modelo *sup* emprega as arquiteturas de redes neurais mais complexas e profundas, exigindo maior capacidade de GPU para processar os mesmos dados; no entanto, ele é capaz de resolver regiões genômicas desafiadoras e minimizar erros em homopolímeros de forma significativamente mais eficaz que os demais modelos [158].

2.2.2.2 Controle de Qualidade das Reads

O controle de qualidade (CQ) é uma etapa crítica no processamento de dados de sequenciamento por *Nanopore*, pois influencia diretamente a confiabilidade das análises *downstream*, como montagem de genoma, mapeamento, detecção de variantes e análises metagenômicas [40]. Devido às características específicas dessa tecnologia, o CQ pode ser dividido em duas abordagens complementares: o controle de qualidade aplicado às leituras brutas e o controle de qualidade das montagens genômicas resultantes.

O controle de qualidade das *reads* visa avaliar e filtrar os dados brutos gerados

durante o sequenciamento. Diferentemente das tecnologias de sequenciamento de curta leitura, os dados produzidos por plataformas *Nanopore* apresentam grande variabilidade no comprimento das leituras, taxas de erro mais elevadas (tipicamente entre 5–15%) e a presença de erros sistemáticos, especialmente em regiões homopoliméricas e em bases modificadas [55].

Nesse contexto, a etapa de CQ das leituras envolve a análise de métricas como distribuição do comprimento das *reads*, qualidade média por leitura (*Phred score*), rendimento total e identificação de padrões anômalos. A filtragem de leituras muito curtas, de baixa qualidade é fundamental para reduzir ruído nos dados e evitar a propagação de erros para as etapas subsequentes [39]. Estudos demonstram que um CQ inadequado nessa fase pode comprometer significativamente a qualidade das montagens e impactar negativamente a interpretação biológica dos resultados [26].

2.2.2.3 Montagem do Genoma

A montagem do genoma é o processo computacional de usar dados de sequenciamento do genoma completo (leituras) para reconstruir a sequência genômica verdadeira de um organismo na maior extensão possível [71]. Ferramentas de software que realizam a montagem (montadores) recebem leituras de sequenciamento como entrada e produzem partes contíguas reconstruídas do genoma (*contigs*) como saída, que por sua vez podem ser agrupadas em estruturas maiores ainda chamadas *scaffolds*. [157]. Esse processo é essencial quando um organismo não possui um genoma de referência ou quando há necessidade de estudar variações genômicas complexas. [78].

A evolução dos algoritmos de montagem acompanhou o desenvolvimento das tecnologias de sequenciamento. Atualmente, existem duas classes de algoritmos amplamente utilizadas: overlap-layout-consensus (OLC) e de-bruijn-graph (DBG) [49], [133]. OLC geralmente funciona em três etapas: primeiro, sobreposições (O) entre todas as leituras são encontradas, então ele realiza um layout (L) de todas as leituras e informações de sobreposições em um gráfico e, finalmente, a sequência de consenso (C) é inferida [86].

O gráfico de Bruijn tem sido usado por abordagens de montagem de leitura curta para representar repetições genômicas como um gráfico de repetição. Estudos anteriores demonstraram o valor dessa abordagem para melhorar a precisão da montagem do genoma [118]. DBG é um algoritmo que trabalha primeiro cortando leituras em $k - mers$ muito mais curtos e então usando todos os $k - mers$ para formar um DBG e finalmente inferindo a sequência do genoma no DBG [68].

As duas principais estratégias, a abordagem de OLC e as abordagens baseadas em gráficos de de Bruijn, têm cada uma suas forças e fraquezas. Enquanto a OLC mantém informações de longo alcance e é robusta contra regiões repetitivas, as abordagens baseadas em gráficos de de Bruijn são eficientes em termos de memória e se destacam com

leituras curtas de alto rendimento.

2.2.2.4 Controle de Qualidade das Montagens

O controle de qualidade das montagens é realizado após a etapa de montagem genômica e tem como objetivo avaliar a consistência estrutural, a completude e a precisão do genoma reconstruído [58]. Diferentemente do CQ das leituras, que foca nas características individuais dos dados brutos, o CQ das montagens analisa propriedades globais do genoma montado.

Entre os principais aspectos avaliados estão métricas como tamanho total da montagem, número de *contigs*, N50, presença de regiões duplicadas ou fragmentadas e a correspondência com genomas de referência ou conjuntos de genes conservados [101]. Essa etapa permite identificar problemas decorrentes tanto da qualidade das leituras quanto de limitações dos algoritmos de montagem, como colapsos de regiões repetitivas ou fragmentação excessiva [138]. Dessa forma, o CQ das montagens é essencial para validar se o genoma obtido é adequado para análises comparativas, anotação gênica e outras aplicações *downstream*.

2.2.2.5 Polimento da Montagem - polishing

Os algoritmos de polimento utilizam, como etapa interna, o alinhamento (mapeamento) das leituras contra a montagem, sendo esse processo fundamental para a identificação e correção de erros [158].

As tecnologias de sequenciamento de terceira geração apresentam altas taxas de erro, e uma grande proporção dos pares de bases nessas leituras longas é identificada incorretamente. Esses erros são propagados para a montagem e afetam a precisão das análises genômicas. Os algoritmos de *polishing* (refinamento/polimento de montagem) minimizam essa propagação de erros, corrigindo ou ajustando os erros na montagem utilizando informações provenientes dos alinhamentos entre as leituras e a montagem (ou seja, informações de alinhamento de leitura para montagem). No entanto, os algoritmos de *polishing* atuais só conseguem refinar uma montagem utilizando leituras de uma determinada tecnologia de sequenciamento ou apenas montagens de pequeno porte. Essa dependência tanto da tecnologia quanto do tamanho da montagem obriga os pesquisadores a (i) executar múltiplos algoritmos de *polishing* e (ii) dividir grandes genomas em pequenos blocos para utilizar todos os conjuntos de leituras disponíveis e conseguir polir montagens de genomas de grande porte. Se o algoritmo de polimento identificar uma dissimilaridade, ele modifica a montagem para torná-la mais semelhante às leituras alinhadas, assumindo que as informações do alinhamento são uma fonte mais confiável. Em outras palavras, a dissimilaridade é atribuída a erros na montagem. Os algoritmos de polimento de montagem partem do princípio de que tais modificações corrigem os erros de uma montagem.[48].

Na etapa de polimento é comum utilizar estratégias combinadas como: (i) duas

rodadas consecutivas de Racon; (ii) uma rodada de Racon seguida por uma rodada de Medaka; e (iii) duas rodadas de Racon seguidas por uma rodada de Medaka. Essas estratégias são comuns na análise de leituras longas, pois as primeiras iterações tendem a corrigir erros mais evidentes, enquanto rodadas adicionais refinam o consenso.

2.2.2.6 Análises adicionais (Identificação de mutação/Detecção de variantes)

Diversos projetos catalogaram de forma abrangente pequenas variantes genéticas [variantes de nucleotídeo único (SNVs)]; entretanto, as variantes estruturais (SVs) continuam amplamente sub-representadas [8]. As SVs são definidas como regiões de DNA maiores que 50 pares de bases que apresentam alterações no número de cópias ou na localização genômica, incluindo variantes no número de cópias (CNVs; deleções e duplicações), inserções, inversões, translocações, elementos móveis, expansões de sequências repetitivas e combinações complexas dessas variações [142]. As SVs contribuem com aproximadamente 3,4 vezes mais nucleotídeos para a variação genética humana do que as SNVs, mesmo sendo numericamente muito menos frequentes [66]. As tecnologias de sequenciamento de terceira geração, como a ONT, produzem sequências que podem ser mapeadas com mais confiança em regiões repetitivas do genoma [69], superando as limitações fundamentais das leituras curtas. As leituras longas podem gerar montagens *de novo* altamente contíguas [100] e estão sendo cada vez mais utilizadas por métodos de análise baseados em referência.

As leituras longas são vantajosas para detecção de SVs devido à maior capacidade de mapeamento em regiões repetitivas e à sua capacidade de abranger SVs inteiramente ou permitir alinhamentos divididos ancorados em lados suficientemente longos. Essas vantagens possibilitam a inferência direta de SVs, frequentemente com precisão nucleotídica nos pontos de quebra [146] [135]. O mapeamento de referência com leitura longa não requer alta cobertura (10–15x) para detecção de SVs, mas grandes inserções não-referenciadas ainda representam um desafio [36] [136] [65]. Regiões anteriormente inacessíveis, como repetições de telômeros e centrômeros, agora podem ser investigadas.

Diversas ferramentas foram desenvolvidas para detectar SVs a partir de leituras longas [36] [136]. Esses métodos apresentam precisão e sensibilidade de aproximadamente 65% em comparação com um conjunto padrão ouro de SVs, sendo que a combinação de métodos pode melhorar a precisão e a sensibilidade [38].

2.2.3 Métricas de avaliação (BUSCO, *indels*, N50)

A avaliação da qualidade das montagens genômicas é uma etapa crítica para determinar a confiabilidade das sequências consenso geradas. Neste trabalho, a qualidade foi mensurada através de três pilares: continuidade, acurácia e completude biológica.

A continuidade da montagem foi avaliada principalmente através da métrica N50. O N50 é definido como o comprimento do menor *contig* tal que a soma de todos os *contigs* de tamanho igual ou superior a ele represente pelo menos 50% do tamanho total

da montagem [58]. Em genomas procariotos, valores elevados de N50, próximos ao tamanho esperado do cromossomo, indicam uma montagem bem resolvida e com baixa fragmentação.

A acurácia da sequência foi monitorada pela frequência de erros de inserções e deleções *indels*. Devido às características inerentes ao sequenciamento por *nanopore*, erros em regiões de homopolímeros são comuns, resultando em *indels* que podem causar mudanças na matriz de leitura (*frameshifts*) em genes codificantes [151]. A redução da taxa de *indels* por 100 kb (quilobases) é um indicador fundamental da eficiência das etapas de polimento com ferramentas como Racon e Medaka.

Por fim, a integridade funcional foi verificada através do software **BUSCO** (*Benchmarking Universal Single-Copy Orthologs*). Esta ferramenta avalia a presença de genes ortólogos de cópia única que são amplamente conservados em linhagens filogenéticas específicas [138]. Uma alta porcentagem de genes completos (BUSCO-C) sugere que a montagem não apenas possui continuidade estrutural, mas também preserva a informação biológica essencial, sendo um indicador robusto da qualidade final para análises de genômica comparativa e anotação funcional.

2.2.4 Aplicações da Tecnologia ONT

2.2.4.1 Montagem de Genomas De Novo

A montagem de genomas *de novo* refere-se à reconstrução de uma sequência genômica sem o uso direto de um genoma de referência. Essa abordagem é especialmente importante para organismos cujo genoma ainda não foi previamente caracterizado, mas também pode ser utilizada quando o genoma de referência disponível é incompleto, distante evolutivamente ou quando se deseja evitar vieses associados ao mapeamento. O uso da tecnologia Nanopore é uma aplicação interessante, pois as leituras longas geradas permitem resolver regiões repetitivas ou sequências complexas que não podem ser montadas utilizando tecnologias de leituras curtas. No entanto, para alcançar uma alta precisão do consenso, é necessário obter muitas leituras de um único organismo, a fim de fornecer uma alta cobertura e, assim, mitigar a taxa de erro da tecnologia [147].

2.2.4.2 Detecção de Variantes Estruturais

Como citado anteriormente, a ONT permite detectar variantes pequenas e grandes, incluindo variantes estruturais, com capacidade superior em regiões repetitivas ou difíceis. Um exemplo dessa aplicação é para o diagnóstico de câncer. No contexto oncológico, a principal contribuição das leituras longas geradas pela ONT está relacionada à caracterização da alta variabilidade genética tumoral, especialmente variantes estruturais complexas e rearranjos genômicos extensos, que são frequentemente difíceis de detectar por tecnologias de leitura curta. Embora ainda não seja amplamente utilizada como ferramenta de rotina clínica, a tecnologia tem demonstrado potencial para estudos exploratórios e de pesquisa translacional voltados à compreensão da heterogeneidade tumoral

[29].

2.2.4.3 Identificação de metilação e modificações epigenéticas

Modificações epigenéticas referem-se a alterações químicas no DNA ou em proteínas associadas, que não modificam a sequência de bases, mas influenciam a expressão gênica. A metilação do DNA é um dos principais mecanismos epigenéticos e desempenha papel importante na regulação da atividade gênica e na resposta celular a diferentes estímulos. Embora existam métodos eficazes para mapear sua ocorrência no genoma, cada um apresenta limitações [139].

O método atualmente considerado ”padrão ouro”para o perfilamento de metilação em DNA é o sequenciamento por bissulfito [103]. O tratamento com bissulfito de sódio converte citosinas não modificadas em uracil, enquanto deixa as citosinas metiladas inalteradas. O tratamento com bissulfito é um processo agressivo, que resulta em uma fragmentação extensa do DNA [103], e quaisquer quebras na molécula inserida impedem a amplificação subsequente, exigindo, assim, altos níveis de DNA de entrada (100 ng). O uso de uma grande quantidade de material pode influenciar a análise, uma vez que o epigenoma é altamente variável e resulta em uma mistura heterogênea de estados epigenéticos. [64].

Trabalhos anteriores demonstraram que a 5-metilcitosina (5-mC) pode ser diferenciada da citosina por meio da análise detalhada dos sinais de corrente elétrica medidos por dispositivos de sequenciamento baseados em nanoporo [81] [134]. No trabalho de Simpson *et al*. (2017) [139], por exemplo, eles descrevem o desenvolvimento de um método para detectar diretamente modificações no DNA, especificamente a 5-mC, utilizando apenas as leituras elétricas do instrumento MinION da ONT, sem necessidade de qualquer tratamento químico.

2.2.4.4 Detecção de patógenos e vigilância epidemiológica

A tecnologia ONT permite a identificação rápida de patógenos em tempo real, diretamente de amostras clínicas ou ambientais, principalmente devido à portabilidade dos dispositivos e à capacidade de geração contínua de dados. No trabalho de Quick *et al*. (2016) [122] os autores criaram um sistema de vigilância genômica que utiliza a tecnologia ONT. Em abril de 2015, esse sistema foi transportado em bagagem de avião padrão para a Guiné e usado para vigilância genômica em tempo real da epidemia de ebola em andamento, mostrando que a vigilância genômica em tempo real é possível em cenários com recursos limitados e pode ser estabelecida rapidamente para monitorar surtos.

Outro exemplo que pode ser citado é o trabalho de Bull *et al.* (2020) [20], que durante a pandemia da Covid-19 mostrou que o sequenciamento ONT permitiu a detecção altamente precisa e reproduzível de SNVs de nível de consenso em isolados de pacientes com SARS-CoV-2.

2.2.4.5 Identificação e caracterização de microbiomas (metagenômica)

Outra aplicação da tecnologia ONT é que ela permite sequenciamento completo de genomas bacterianos e identificação taxonômica precisa em nível de cepa. O sequenciamento de genomas completos de comunidades microbianas (metagenômica) revolucionou a compreensão da evolução e diversidade microbiana, com inúmeras aplicações potenciais na ecologia microbiana, microbiologia clínica e biotecnologia industrial [60]. Aplicações metagenômicas com a tecnologia ONT podem ser analisadas no trabalho de Urban *et al*. (2021) [148] para monitoramento ambiental de ambientes aquáticos. A portabilidade dos dispositivos ONT, como o MinION, permite o monitoramento *in loco* de corpos d'água, facilitando a detecção de contaminantes e patógenos em tempo real. Por exemplo, foi desenvolvido um protocolo de diagnóstico de água doce baseado em sequenciamento por nanoporos, oferecendo uma alternativa rápida e econômica aos testes microbiológicos tradicionais.

Segundo Chen *et al*. (2024) [28], as leituras longas geradas pela ONT facilitam a montagem de genomas metagenômicos, permitindo a reconstrução de genomas bacterianos completos a partir de comunidades complexas, ressaltando sua utilidade na identificação de patógenos e na compreensão de suas características genéticas, especialmente em contextos onde a resolução de genomas individuais a partir de amostras mistas é necessária.

3 METODOLOGIA

A metodologia deste trabalho está resumida na Figura 5 e compreende as etapas de revisão da literatura; identificação dos principais fluxos e ferramentas; instalação das ferramentas em *desktop* e *cluster*; obtenção dos dados públicos dos organismos procariotos; preparação de um *script* para automatização; execução dos fluxos a partir dos dados obtidos dos organismos; e por fim, a análise dos resultados.

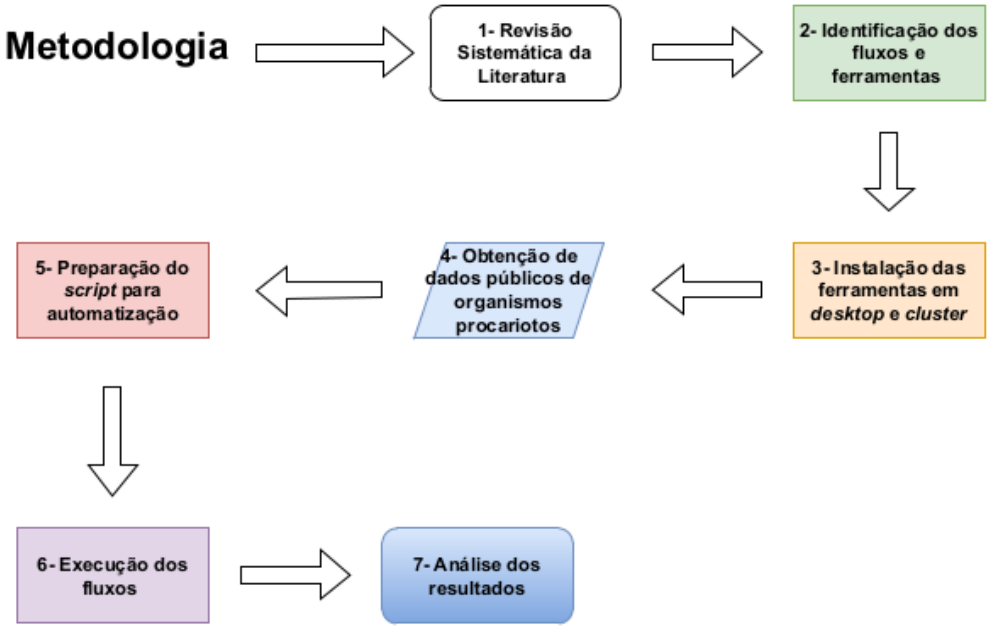

Figura 5: Metodologia do estudo

3.1 Revisão Sistemática da Literatura

A partir do guia proposto por Mariano e colaboradores [96], focado em revisões na área da bioinformática, o presente trabalho utiliza a metodologia proposta, que se

baseia nas etapas: “coleta de referencias”, “avaliação dos dados” e “interpretação das descobertas”.

Para dar início a revisão o primeiro passo é a definição de um protocolo, ou seja, é a etapa onde se define o objetivo do estudo e as perguntas que guiarão o processo da revisão. Após definir esses critérios, o próximo passo é realizar a coleta de referências. Nessa etapa são definidas as plataformas de busca para acessar esses artigos que serão abordados na revisão, também é necessário determinar quais as melhores palavras-chave que são mais adequadas ao objetivo proposto. Feito isso, dados serão coletados nessas plataformas de busca e serão analisados e avaliados, levando em consideração critérios de inclusão e exclusão dos artigos que foram os resultados da primeira busca. Por fim, os artigos selecionados após todos esses processos citados anteriormente serão analisados e discutidos, e então é feita uma avaliação da revisão.

3.1.1 Definição do Protocolo

- Objetivo: identificar e avaliar pipelines computacionais propostos para a montagem de genomas de organismos procariotos a partir de dados de sequenciamento gerados por sequenciadores da ONT.
- Pergunta de pesquisa: Quais pipelines computacionais apresentam o melhor desempenho para as etapas de *basecalling*, filtragem, montagem e polimento de genomas de procariotos oriundas de dados de sequenciamento *nanopore*?

  Quais métricas são empregadas para avaliar a qualidade das montagens?

3.1.2 Coleta de Referências

Para a coleta das referências foram determinadas duas plataformas de busca: Google Acadêmico [1], levando em consideração a abrangência, diversidade de fontes e facilidade de acesso e o PubMed[2] foi levado em conta o seu foco em ciências biomédicas e a qualidade e curadoria da busca. A combinação dessas plataformas garante abrangência e qualidade, permitindo a inclusão de estudos relevantes tanto de um ponto de vista técnico (Google Scholar) quanto de um ponto de vista biomédico (PubMed), cobrindo assim as necessidades da revisão sistemática. As palavras-chave de busca foram definidas utilizando operadores Booleanos, que permitiram delimitar melhor nosso objetivo alvo, de forma que os artigos resultantes relacionados com a questão do trabalho. Deste modo, os operadores negativos “NOT” foram empregados para excluir organismos não procariotos, trabalhos de metagenoma ou que tratassem do sequenciamento de RNA. Abaixo consta a chave de busca utilizada em cada plataforma.

- Google Acadêmico: nanopore AND (pipeline OR workflow) AND (assembly OR genome assembly OR basecalling OR base calling) AND (bacteria OR prokaryote

[1] https://scholar.google.com/
[2] https://pubmed.ncbi.nlm.nih.gov/

OR prokaryotic) NOT metagenome NOT 16s NOT rna NOT eukaryote NOT eukaryotic NOT vírus

- PubMed: nanopore AND (pipeline OR workflow) AND (assembly OR genome assembly OR basecalling OR base calling) AND (bacteria OR prokaryote OR prokaryotic) NOT metagenome NOT 16s NOT rna NOT eukaryote NOT eukaryotic

Foram excluídos artigos que não passaram por revisão de pares e artigos publicados antes 2019, restringindo a revisão a dados mais recentes. Para inclusão, foram considerados artigos de livre acesso e que abordem a montagem de genomas de organismos procariotos e ferramentas relacionadas. Inicialmente, buscas sistemáticas foram conduzidas nas bases de dados PubMed e Google Scholar, abrangendo o período de 2019 a 2024. Para garantir que o conteúdo estivesse atualizado e refletisse os avanços mais recentes na área, uma nova busca foi realizada, abrangendo o período de janeiro a agosto de 2025. A estratégia de busca se manteve combinando termos relacionados a nanoporos, pipeline/fluxo de trabalho, montagem, chamada de bases e procariontes. Operadores booleanos foram usados para restringir a busca a genomas procarióticos e excluir estudos sobre metagenômica, RNA ou organismos eucarióticos.

3.1.3 Seleção dos artigos

A seleção dos estudos foi realizada em múltiplas etapas, com base em critérios de inclusão e exclusão previamente definidos. Inicialmente, foi realizada a triagem dos títulos, seguida da análise dos resumos, leitura diagona e, por fim, da leitura completa dos artigos selecionados.

Foram considerados como critérios de inclusão estudos que abordassem pipelines computacionais para montagem de genomas de organismos procariotos a partir de dados de sequenciamento por Nanopore.

Foram excluídos artigos que não apresentavam descrição metodológica clara, que utilizavam exclusivamente dados de leitura curta, que abordavam dados não relacionados a organismos procariotos ou que focavam em análises de RNA ou metagenômica.

O processo de triagem foi realizado com o auxílio da ferramenta Rayyan [109], que permite a organização e avaliação dos estudos de forma semiautomatizada. Para a etapa de leitura e organização dos artigos selecionados, foi utilizada a ferramenta Zotero [150].

3.2 Identificação dos principais fluxos e ferramentas

As ferramentas selecionadas para compor os *pipelines* foram fundamentadas naquelas identificadas com maior frequência e melhor desempenho na revisão sistemática prévia, além de seguirem as recomendações técnicas da *Oxford Nanopore Technologies* (ONT). Para a etapa crítica de *basecalling*, foram utilizados os softwares Guppy e Dorado,

permitindo a comparação entre diferentes modelos de acurácia. O controle de qualidade e a visualização das métricas das leituras brutas foram realizados com o NanoPlot, seguidos por um refinamento dos dados através do Filtlong e NanoFilt, visando a remoção de reads curtas ou de baixa qualidade que poderiam comprometer a montagem.

Para a etapa de montagem de novo, utilizou-se o Flye, amplamente reconhecido por sua robustez em lidar com genomas bacterianos e regiões repetitivas. O mapeamento das leituras para o polimento foi executado com o Minimap2, enquanto a correção do genoma utilizou o Racon (polimento baseado em alinhamento) e o Medaka (baseado em redes neurais), garantindo uma sequência de consenso de alta acurácia. Por fim, a qualidade das montagens finais foi rigorosamente avaliada através do QUAST, para métricas de continuidade (como N50 e número de contigs), e do BUSCO, para avaliar a completude do genoma com base em ortólogos universais de cópia única para procariotos.

3.3 Instalação das ferramentas em *desktop* e *cluster*

Foram utilizados dois ambientes computacionais distintos, um *cluster* e um *desktop*, de forma complementar ao longo do desenvolvimento e execução do pipeline. Essa estratégia permitiu explorar recursos computacionais especializados para etapas mais custosas computacionalmente, ao mesmo tempo em que viabilizou a execução local das análises subsequentes.

Para garantir a reprodutibilidade dos experimentos e a gestão eficiente das dependências de *software*, toda a infraestrutura computacional foi organizada utilizando o gerenciador de pacotes Conda. Foi criado um ambiente virtual principal contendo as bibliotecas e ferramentas de controle de qualidade (como NanoPlot) e filtragem (como Filtlong). Devido às particularidades de suas bibliotecas de *deep learning* e requisitos específicos de sistema, a ferramenta Medaka foi instalada em um ambiente virtual isolado e dedicado, conforme as recomendações dos desenvolvedores, evitando conflitos de versões e garantindo a estabilidade da execução.

A etapa de *basecalling* integrou o fluxo metodológico do estudo, uma vez que os diferentes modelos (*fast*, *hac* e *sup*) foram considerados sistematicamente na análise. No entanto, por uma decisão logística, essa etapa foi executada separadamente do pipeline automatizado, sendo realizada previamente para todos os organismos no *cluster* do C3 (Centro de Ciências Computacionais) da FURG, com utilização de GPU, por meio das ferramentas Guppy e Dorado. Optou-se por não incluir o *basecalling* diretamente no *script* automatizado para evitar reprocessamento computacional e otimizar o tempo de execução, visto que os dados já haviam sido gerados sob as mesmas condições experimentais para todas as ferramentas e modelos avaliados.

Após a geração dos arquivos *basecalled* no ambiente de cluster, as etapas subsequentes do pipeline, incluindo controle de qualidade, filtragem, montagem e análises comparativas, foram executadas no *desktop* do COMBI-Lab (Laboratório de Biologia Compu-

tacional). O desempenho computacional foi avaliado com base no tempo real de execução (*wall-clock time*), obtido por meio do comando time, possibilitando a comparação direta entre ferramentas e modelos.

Tabela 1: Infraestrutura computacional utilizada no estudo

Ambiente	**Processador (cores)**	**Memória RAM**
Cluster (C3 – FURG)	40	484 GB
Desktop (COMBI-Lab)	24	64 GB

3.4 Obtenção dos dados públicos dos organismos procariotos

A etapa de implementação prática foi iniciada com a seleção e obtenção de dados públicos de sinal bruto (fast5) de procariotos, etapa essencial para viabilizar o *benchmarking* das ferramentas de *basecalling* e montagem. Com base na literatura, foram utilizados os dados disponibilizados por Wick, Judd Holt (2019) [158], que compreendem um conjunto de seis organismos distintos: *Acinetobacter pittii, Haemophilus haemolyticus, Klebsiella pneumoniae, Shigella sonnei, Staphylococcus aureus e Stenotrophomonas maltophilia*. Os autores também disponibilizaram os genomas de referência de cada um dos organismos. Esses dados foram gerados através da tecnologia de sequenciamento MinION, utilizando a *flow cell* R9.4.

Tabela 2: Características gerais dos conjuntos de dados e dos organismos utilizados no estudo

Organismo	Reads	Tamanho (bp)	Ref.	Genoma (Mb)	Gram	Relevância
Acinetobacter pittii	4.467	115.233.809	CP139944.1	3,5–4,0 [a]	-	Patógeno oportunista hospitalar
Haemophilus haemolyticus	9.536	86.772.440	CP131722.1	1,8–2,0 [b]	-	Comensal respiratório
Klebsiella pneumoniae	15.178	243.980.906	CP132713.1	5,0–6,0 [c]	-	Patógeno com resistência antimicrobiana
Shigella sonnei	23.583	491.632.613	CP180015.1	4,5–4,8 [d]	-	Agente de shigelose
Staphylococcus aureus	11.047	491.694.546	OZ059820.1	2,7–3,0 [e]	+	Patógeno clínico (MRSA)
Stenotrophomonas maltophilia	16.010	490.695.014	CP167818.1	4,5–5,0 [f]	-	Patógeno oportunista multirresistente

[a] `https://bridges.monash.edu/articles/dataset/Raw_fast5s/7676174/1?file=14260511`
[b] `https://bridges.monash.edu/articles/dataset/Raw_fast5s/7676174/1?file=14260514`
[c] `https://bridges.monash.edu/articles/dataset/Raw_fast5s/7676174/1?file=14260550`
[d] `https://bridges.monash.edu/articles/dataset/Raw_fast5s/7676174/1?file=14260562`
[e] `https://bridges.monash.edu/articles/dataset/Raw_fast5s/7676174/1?file=14260568`
[f] `https://bridges.monash.edu/articles/dataset/Raw_fast5s/7676174/1?file=14260574`

3.5 Preparação do *script* para automatização

Após a etapa de entendimento manual de cada ferramenta, que consistiu na análise da documentação técnica e ajuste de parâmetros para lidar com o sinal bruto das *flow cells* R9.4, foram desenvolvidos *scripts* em Bash para automatizar o fluxo. Esses *scripts* foram projetados para gerenciar a ativação automática dos ambientes Conda necessários para cada etapa, processando de forma sequencial o *basecalling*, avaliações de qualidade, filtragem, montagem, mapeamento e polimento. A validação dessa automação foi feita

comparando-se os resultados gerados pelos *scripts* nos seis organismos com os resultados obtidos manualmente, assegurando a integridade e padronização da análise de larga escala que resultou nas 72 combinações de pipelines avaliadas para cada organismo.

Ralizou-se a execuçãoo manual do pipeline completo para um desses organismos, servindo como prova de conceito. A escolha das ferramentas para este fluxo fundamentou-se nos resultados da revisão sistemática da literatura, priorizando *softwares* de alta frequência de uso e eficacia comprovada, além de seguir as recomendações técnicas da própria *Oxford Nanopore Technologies*. O processo envolveu a instalação e o estudo detalhado de cada componente, garantindo que o pipeline manual servisse como um padrão ouro (*gold standard*) para a validação dos *scripts* de automação aplicados posteriormente aos demais organismos.

Com o objetivo de promover transparência, reprodutibilidade e reutilização, o *script* desenvolvido para automatização da pipeline foi disponibilizado no repositório Git do laboratório (https://github.com/combilab-furg).

3.6 Execução dos fluxos

O delineamento experimental foi estruturado com o objetivo de avaliar, de forma sistemática, o impacto das etapas de basecalling e pós-processamento na qualidade final das montagens genômicas. Para isso, diferentes combinações de modelos de *basecalling* e estratégias de polimento foram testadas, enquanto as ferramentas e parâmetros de controle de qualidade, filtragem e montagem foram mantidos constantes, de modo a minimizar variáveis de confusão.

Inicialmente, os dados brutos foram submetidos à etapa de *basecalling* utilizando duas ferramentas distintas, Guppy e Dorado. Cada ferramenta foi executada com três modelos de acurácia, *Fast*, *HAC* e *SUP*, resultando em seis conjuntos independentes de (*reads*) para cada organismo analisado. Embora essa etapa não tenha sido incorporada diretamente ao *script* automatizado da *pipeline*, ela faz parte integrante do fluxo metodológico, uma vez que cada modelo de *basecalling* foi considerado como uma condição experimental distinta.

Após o *basecalling*, todos os conjuntos de *reads* foram submetidos às mesmas etapas de controle de qualidade e filtragem. Inicialmente, a qualidade dos dados foi inspecionada utilizando o NanoPlot. Em seguida, as leituras passaram por filtragem de qualidade por meio do NanoFilt, seguida de uma etapa de seleção baseada em qualidade utilizando o Filtlong. Essas etapas foram aplicadas de forma padronizada a todos os conjuntos de dados, garantindo a comparabilidade entre os diferentes modelos de *basecalling*.

As leituras filtradas foram então utilizadas na etapa de montagem de genomas, realizada com a ferramenta Flye, mantendo-se parâmetros fixos para todos os experimentos. A partir das montagens geradas, foram aplicadas diferentes estratégias de polimento, combinando variações no número de rodadas e no uso das ferramentas Racon e Medaka,

totalizando doze abordagens distintas:

1. Montagem bruta (raw, sem polimento);
2. Uma rodada de Racon;
3. Duas rodadas de Racon;
4. Uma rodada de Medaka;
5. Uma rodada de Racon seguida de uma rodada de Medaka;
6. Duas rodadas de Racon seguidas de uma rodada de Medaka.

Dessa forma, a combinação dos seis conjuntos de *reads* gerados na etapa de *basecalling* com as doze estratégias de pós-processamento resultou em um total de 72 montagens distintas para cada organismo. Cada montagem corresponde a um *pipeline* específico, permitindo avaliar de forma isolada e combinada os efeitos do modelo de *basecalling* e do grau de polimento aplicado.

Todas as montagens foram submetidas à avaliação de qualidade estrutural por meio do QUAST e à análise de completude biológica utilizando o BUSCO. O tempo total de processamento (*wall-clock time*) foi registrado para cada pipeline, permitindo a mensuração do custo computacional associado a cada combinação de ferramentas e parâmetros.

De maneira geral, a estratégia aplicada foi estruturada dessa maneira: 2 *basecallers* (Guppy e Dorado) x 3 modelos (*Fast*, *HAC* e *SUP*) x 1 CQ das reads (NanoPlot) x 2 Filtros (NanoFilt e Filtlong) x 1 montador (Flye) x 6 estratégias de polimento (raw, Racon 1x, Racon 2x, Medaka, Racon 1x + Medaka, Racon 2x + Medaka) = 72

3.7 Análise dos resultados

A análise dos resultados foi conduzida com o objetivo de comparar, de forma sistemática, o impacto das diferentes combinações de modelos de *basecalling* e estratégias de pós-processamento na qualidade final das montagens genômicas. Foram avaliados, para cada um dos seis organismos analisados, um total de 72 pipelines, resultantes da combinação entre ferramentas de *basecalling*, modelos de acurácia, controle de qualidade, montagem e estratégias de polimento, mantendo-se o filtro constante. As análises foram organizadas de modo a considerar tanto métricas de desempenho computacional quanto indicadores de qualidade estrutural e completude biológica dos genomas montados.

No total, foram avaliados 432 resultados de montagem genômica, correspondentes a 72 pipelines aplicados a cada um dos seis organismos estudados. Cada pipeline representa uma combinação específica entre ferramenta e modelo de basecalling (Guppy ou

Dorado; Fast, HAC ou SUP) e estratégia de pós-processamento, variando desde montagens sem polimento até abordagens combinadas envolvendo múltiplas rodadas de Racon e Medaka.

O desempenho computacional dos pipelines foi avaliado a partir do tempo total de execução (*wall-clock time*), obtido por meio do comando *time*. Essa métrica foi utilizada para comparar o custo computacional associado aos diferentes modelos de *basecalling* e às estratégias de polimento.

A qualidade das leituras após as etapas de controle de qualidade e filtragem foi avaliada com base nas estatísticas geradas pelo NanoPlot. As etapas de filtragem com NanoFilt e Filtlong foram aplicadas de forma padronizada a todos os conjuntos de dados, garantindo comparabilidade entre os pipelines.

A qualidade estrutural das montagens foi avaliada utilizando o QUAST, considerando métricas como número de *contigs*, tamanho total da montagem, N50 e desalinhamentos em relação ao genoma de referência.

A completude biológica das montagens foi avaliada por meio do BUSCO, utilizando conjuntos de genes conservados específicos para procariotos.

4 RESULTADOS E DISCUSSÃO

Os resultados deste trabalho serão divididos em: Revisão Sistemática da Literatura; Análise do fluxos e ferramentas; Comparação dos melhores fluxos entre os organismos analisados; e Qualidade versus custo computacional nos fluxos de montagem.

4.1 Revisão Sistemática da Literatura

A busca inicial realizada nas bases Google Acadêmico e PubMed resultou na identificação de 420 artigos, sendo 380 provenientes do Google Acadêmico e 40 da PubMed, considerando o período de 2019 a 2024 (Tabela 3).

Busca de dados	**Resultado**
1- Google Acadêmico	380
2- PubMed	40
Total	420

Tabela 3: Resultado das buscas (2019/2024
)

Após a remoção de duplicatas, restaram 365 artigos para a etapa de triagem. A seleção inicial foi realizada com base na leitura dos títulos, resultando na inclusão de 182 artigos para a próxima etapa.

Na etapa de leitura dos resumos, 102 artigos foram selecionados para análise mais aprofundada. Posteriormente, foi realizada a leitura diagonal dos textos (introdução, metodologia e figuras), resultando na seleção de 30 artigos alinhados ao escopo do estudo.

Uma atualização da busca foi realizada no período de janeiro a agosto de 2025, resultando na identificação de 215 novos artigos (Tabela 4). Após a triagem por títulos e resumos, 33 artigos foram selecionados para leitura completa, dos quais 23 atenderam aos critérios de inclusão.

Busca de dados	Resultado
Google Acadêmico	203
PubMed	12
Total	215

Tabela 4: Resultado das buscas (01/2025-08/2025

)

Ao final do processo, foram incluídos 53 artigos na revisão sistemática, abrangendo o período de 2019 a agosto de 2025. O processo completo de seleção dos estudos está apresentado no fluxograma PRISMA (Figura 6).

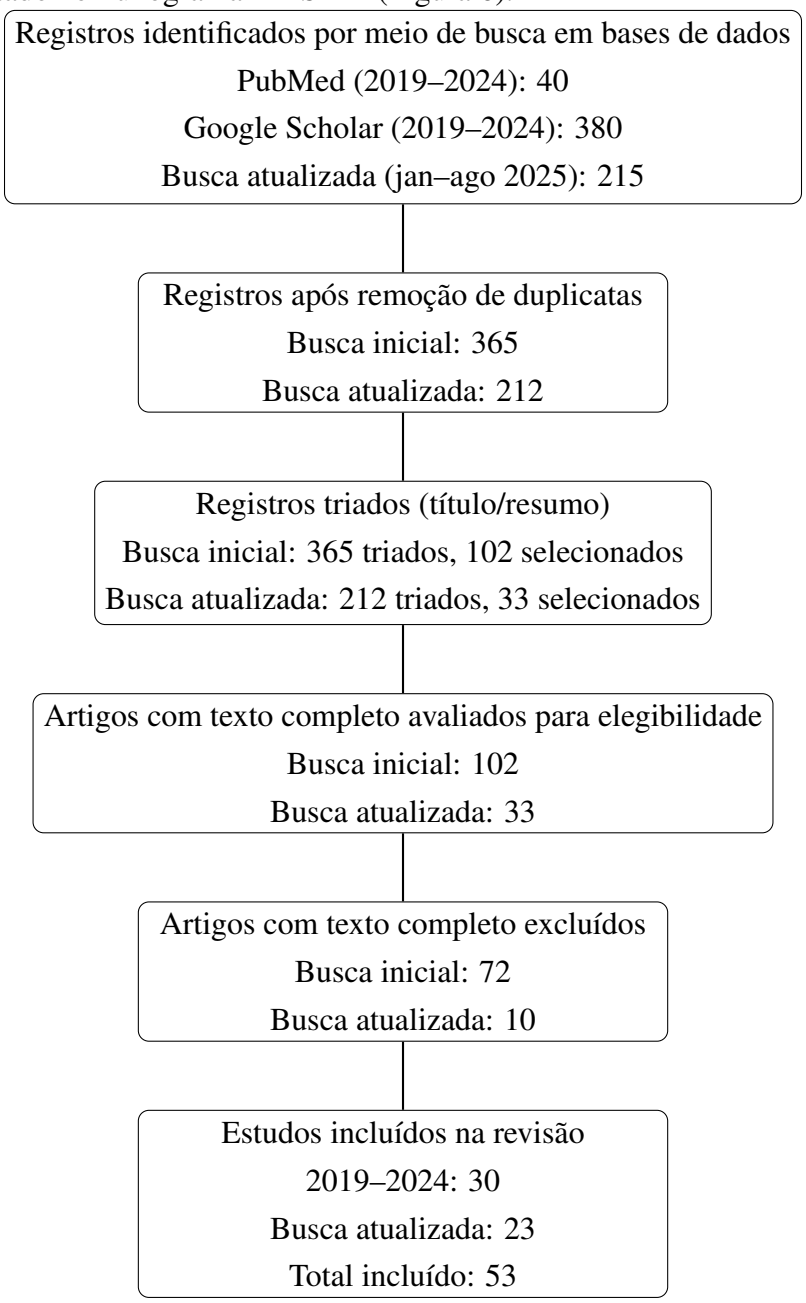

Figura 6: Fluxograma no estilo PRISMA resumindo o processo de seleção dos estudos.

Com base nesses 53 artigos, foram identificados os principais passos e ferramentas envolvidas nos pipelines computacionais descritos para análise de dados gerados por sequenciadores de terceira geração (ONT), com foco em genomas de procariotos.

Um pipeline computacional consiste em uma sequência organizada de etapas de processamento de dados, na qual a saída (output) de uma etapa serve como entrada (input) para a próxima. Em bioinformática, pipelines são amplamente utilizados para automatizar e padronizar análises complexas, como o processamento de dados de sequenciamento [42].

A partir da análise dos estudos selecionados, foi possível identificar as principais etapas recorrentes nos fluxos de trabalho, incluindo basecalling, controle de qualidade, filtragem, montagem do genoma, polimento e avaliação das montagens. Com base nessas informações, foi construído um fluxograma representando essas etapas, conforme apresentado na Figura 7.

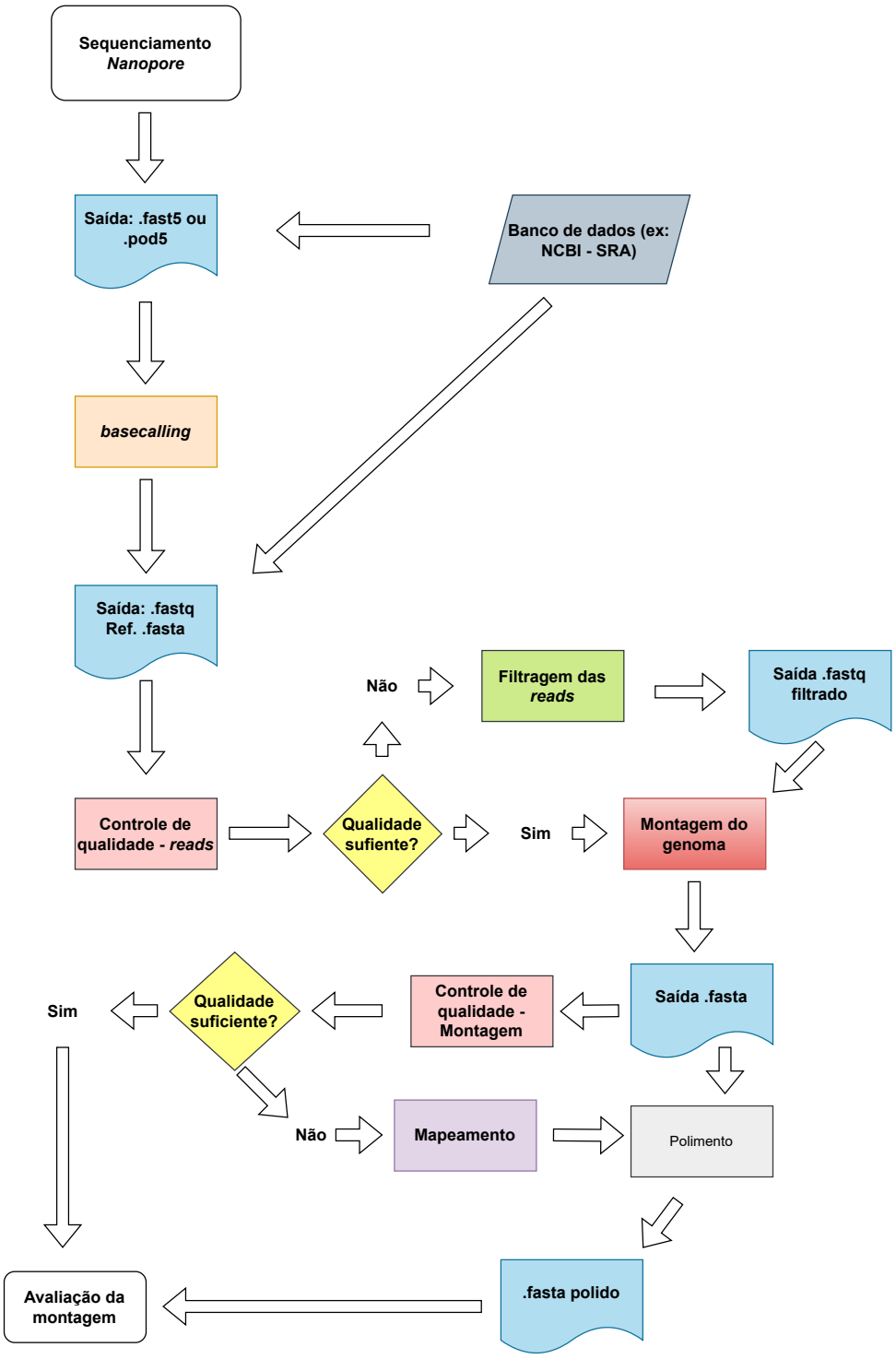

Figura 7: Fluxograma das principais etapas do pipeline

4.1.1 Levantamento da frequência de uso das ferramentas de análise

Além da definição das principais etapas desse processo de análise de dados de sequenciamento gerado com o Nanopore, foi realizado um levantamento das ferramentas utilizadas em cada etapa para medir a frequência de ocorrência das mesmas nos artigos estudados. A partir desse levantamento é possível estabelecer uma visão geral das ferramentas utilizadas em cada etapa envolvida no sequenciamento de *nanopore*. Essa informação possibilita estabelecer possíveis padrões no uso dessas tecnologias. De todos os artigos analisados, 12 artigos utilizam um pipeline computacional específico que foi discutido nesses artigos, realizando todo o processo descrito neste projeto (*Basecalling*, Controle de qualidade, montagem e polimento).

Dos 53 artigos incluídos na revisão sistemática, nem todos reportaram explicitamente as ferramentas utilizadas em cada etapa do pipeline. Dessa forma, a análise de frequência foi realizada apenas com base nos artigos que disponibilizaram essa informação. Para a etapa de basecalling, por exemplo, 43 artigos apresentaram ferramentas claramente identificadas, enquanto 10 não especificaram essa etapa e, portanto, não foram considerados na contagem apresentada na tabela.

4.1.1.1 Basecallers

A figura 8 apresenta os dados relacionados aos *basecallers* observados nesta revisão. Entre o número total de estudos analisados, cinco artigos focaram especificamente na etapa de *basecalling*. No geral, 27 desses artigos utilizaram o *basecaller* Guppy, tornando-o o mais frequentemente usado de acordo com a revisão. Após o Guppy, o segundo mais frequente foi o Dorado (10 artigos). O Dorado havia aparecido em apenas um artigo até 2024, mas tornou-se muito mais frequente em 2025, aparecendo em nove artigos. Posteriormente, o Albacore apareceu em quatro artigos, enquanto o Chiron e o Flappie foram utilizados em três artigos cada durante a etapa de *basecalling*; o Chiron e o Flappie foram empregados nos mesmos artigos, e dois deles também utilizaram o Albacore. Os *basecallers* DeepNano-Blitz e Bonito apareceram em dois artigos cada. Os demais serviços de chamada de base — Causalcall, Ravvent, Nanonat, Scrappie, Osprey e Heron — apareceram em apenas um artigo cada.

A tabela 5 apresenta um quadro indicando quais artigos utilizaram as ferramentas de *basecalling* descritas na revisão

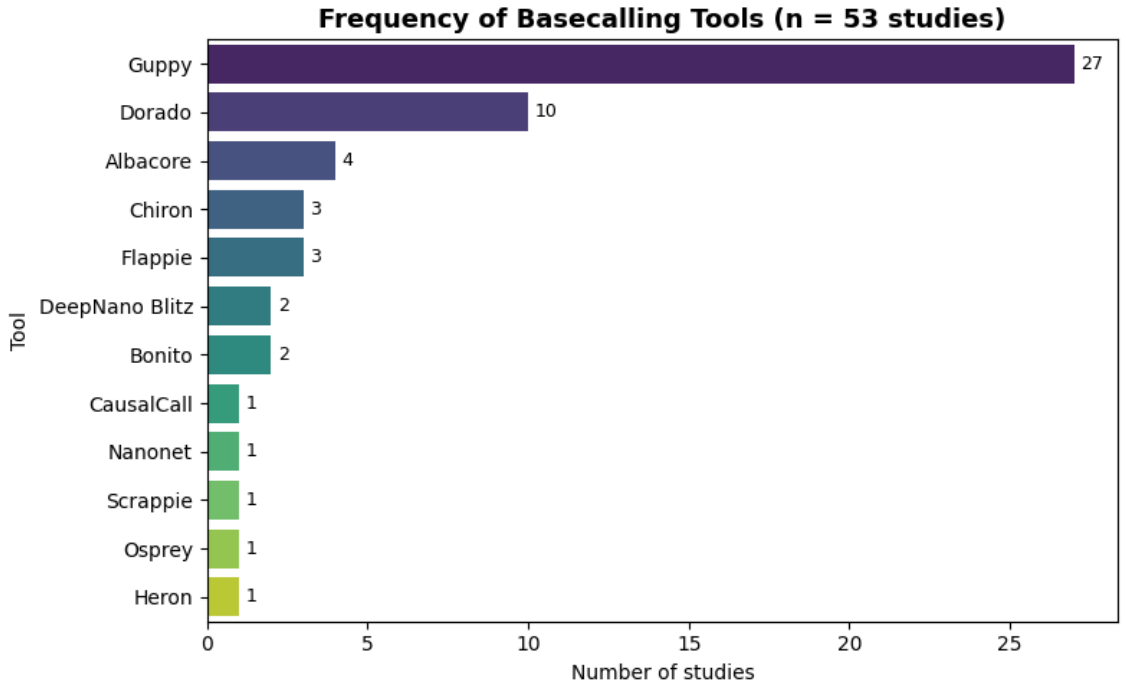

Figura 8: basecallers

Com base na análise de frequência apresentada na Figura 8, observa-se que os *basecallers* Guppy [113] e Dorado [110] foram, de forma consistente, as ferramentas mais utilizadas nos estudos incluídos nesta revisão. O Guppy destacou-se como o *basecaller* mais frequente, sendo amplamente empregado ao longo de diferentes períodos e aplicações, enquanto o Dorado apresentou crescimento expressivo nos estudos mais recentes, especialmente a partir de 2025, refletindo sua adoção como sucessor tecnológico do Guppy. Diante dessa predominância na literatura e de sua relevância atual no ecossistema de sequenciamento da *Oxford Nanopore Technologies*, ambas as ferramentas foram selecionadas para as etapas experimentais subsequentes deste estudo, permitindo uma comparação alinhada às práticas mais consolidadas e atuais da comunidade científica [157].

Além da diversidade de ferramentas identificadas, os estudos analisados também diferenciaram os modelos de *basecalling* empregados, os quais refletem distintos desempenhos entre velocidade de processamento e acurácia. De modo geral, os *basecallers* modernos baseados em redes neurais profundas disponibilizam múltiplos modelos de inferência, comumente classificados como *fast*, *high accuracy* (HAC) e *super accuracy* (SUP). O modelo *fast* prioriza maior velocidade de processamento, à custa de uma menor precisão na identificação das bases. O modelo *HAC* representa um equilíbrio entre desempenho computacional e acurácia, sendo amplamente adotado em análises genômicas rotineiras. Já o modelo *SUP* é otimizado para máxima acurácia, apresentando menores taxas de erro, especialmente em regiões homopoliméricas, embora com maior custo

computacional. Esses modelos têm sido progressivamente incorporados nos fluxos de trabalho de sequenciamento por *Nanopore*, refletindo a maturidade e a evolução contínua dos algoritmos de *basecalling* [158].

Tabela 5: *Basecallers* utilizadaos nos artigos incluídos na revisão sistemática

Artigo	Gup	Dor	Alb	Chi	Fla	DeepN	Bon	Cau	Rav	Nano	Scr	Osp	Her
[154]	X												
[106]	X												
[25]									X				
[162]				X	X			X					
[83]	X												
[15]	X												
[82]	X												
[17]	X					X							
[4]	X												
[84]	X												
[18]	X					X	X					X	X
[79]		X	X	X	X			X		X			
[21]	X												
[152]	X						X						
[125]	X												
[24]	X												
[43]	X												
[105]	X												
[75]	X												
[119]	X												
[158]			X	X	X						X		
[155]	X												
[88]	X												
[62]			X										
[44]	X												
[9]		X											
[167]		X											
[153]		X											
[156]		X											
[120]		X											
[74]	X												
[50]	X												
[108]		X											
[76]	X												
[77]	X												
[91]			X										
[73]		X											
[63]	X												
[46]		X											
[22]	X												
[3]	X												
[2]		X											
[12]		X											

Gup: Guppy; Dor: Dorado; Alb: Albacore; Chi: Chiron; Fla: Flappie; DeepN: DeepNano/DeepNano-Blitz; Bon: Bonito; Cau: Causalcall; Rav: Ravvent; Nano: Nanonet; Scr: Scrappie; Osp: Osprey; Her: Heron.

4.1.1.2 Controle de Qualidade das Reads

A Figura 9 representa a frequência de uso das ferramentas para controle de qualidade de leitura.

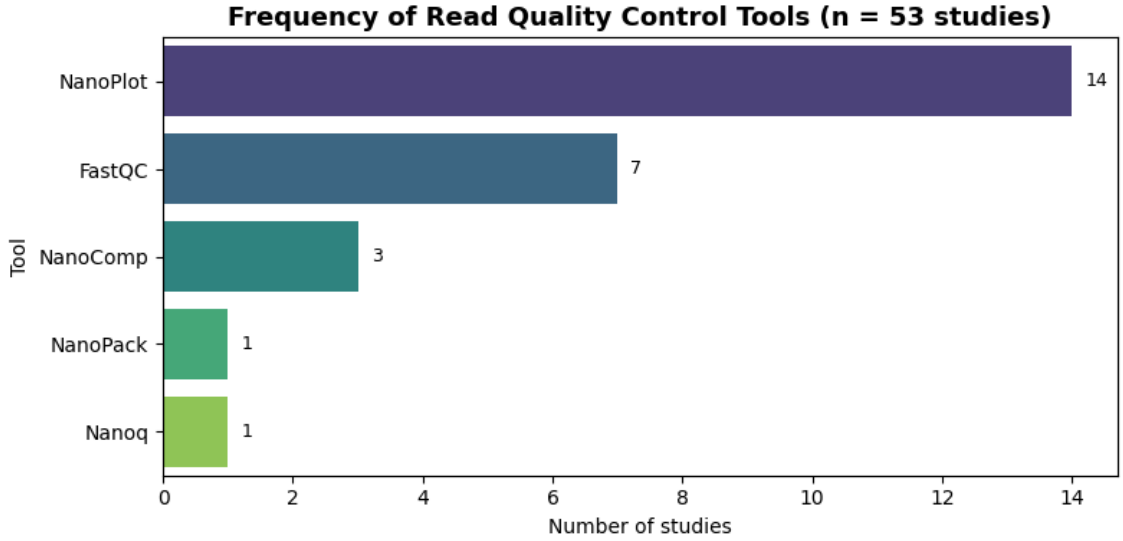

Figura 9: Ferramentas utilizadas para o controle de qualidade das *reads*

Com base na frequência de uso observada nesta revisão sistemática, a ferramenta NanoPlot foi selecionada para as etapas subsequentes de análise. O NanoPlot foi a ferramenta mais empregada para controle de qualidade de leituras, estando presente em 14 dos artigos analisados, o que reflete sua ampla aceitação na comunidade científica para avaliação de dados de sequenciamento por *nanopore*. Essa ferramenta permite a visualização detalhada de métricas fundamentais, como distribuição de comprimento das leituras, qualidade média (Q-score) e rendimento total, sendo amplamente utilizada para caracterizar a qualidade dos dados antes das etapas de filtragem e montagem [39].

Observou-se que uma parcela significativa dos artigos incluídos na revisão não descreve explicitamente uma etapa dedicada ao controle de qualidade das leituras. Essa ausência pode ser atribuída a diferentes fatores. Em muitos estudos, especialmente aqueles focados em pipelines automatizados, o controle de qualidade é incorporado implicitamente a outras etapas do processamento, como a filtragem de leituras ou a própria montagem, não sendo reportado de forma isolada. Além disso, alguns trabalhos assumem que os dados disponibilizados por repositórios públicos já passaram por verificações mínimas de qualidade, optando por não detalhar essa etapa metodológica. Essa prática, embora comum, pode dificultar a reprodutibilidade e a comparação direta entre pipelines, uma vez que métricas iniciais de qualidade influenciam diretamente o desempenho das etapas subsequentes.

A tabela 6 apresenta um quadro indicando quais artigos utilizaram as ferramentas para controle de qualidade descritas na revisão.

Tabela 6: Ferramentas utilizadas para o controle de qualidade das *reads* nos artigos incluídos na revisão sistemática

Artigo	NanoPlot	FastQC	NanoComp	NanoPack	nanoq
[154]	X		X		
[15]	X				
[4]	X				
[84]	X				
[126]	X	X			
[125]		X			
[24]			X		
[41]	X				
[80]	X				
[105]	X				
[37]	X	X			
[94]		X			
[59]		X			
[88]	X				
[62]	X				
[44]		X			
[14]	X				
[50]		X			
[73]			X		
[3]	X				
[121]	X				
[2]					X
[74]				X	

NanoPlot e NanoComp fazem parte do pacote NanoPack; nanoq é uma ferramenta leve para estatísticas básicas de leituras.

Os estudos demonstram que não existe um padrão único estabelecido para o controle de qualidade em dados de sequenciamento por Nanopore, o que contribui para a adoção de abordagens combinadas na literatura. Embora alguns estudos utilizem múltiplas ferramentas para o controle de qualidade das reads, conforme observado na revisão sistemática, neste trabalho foi adotada uma abordagem padronizada com o uso exclusivo do NanoPlot. Essa escolha foi baseada na sua maior frequência de uso entre os artigos analisados, bem como na sua capacidade de fornecer métricas relevantes para dados de sequenciamento por Nanopore, como distribuição de comprimento das leituras e qualidade média. Além disso, a utilização de uma única ferramenta contribui para a padronização do pipeline e evita a introdução de variáveis adicionais na comparação entre os diferentes fluxos analisados.

4.1.1.3 Filtragem das reads

A Figura 10 apresenta a frequência das ferramentas utilizadas para a filtragem das *reads*. Essa etapa é recomendada quando a qualidade das leituras não atinge os critérios ideais para a montagem do genoma, sendo fundamental para a remoção de leituras curtas, de baixa qualidade ou com padrões anômalos que possam comprometer as análises subsequentes.

Entre as ferramentas mais frequentemente utilizadas, destacam-se o Filtlong, presente em 14 artigos, e o NanoFilt, empregado em 12 estudos. Na sequência, o software Porechop apareceu em 9 artigos da revisão, sendo utilizado principalmente para remoção de adaptadores. As ferramentas Chopper e SeqKit foram citadas em 2 artigos cada, enquanto o qcat foi a menos frequente, aparecendo em apenas um dos 53 estudos analisados.

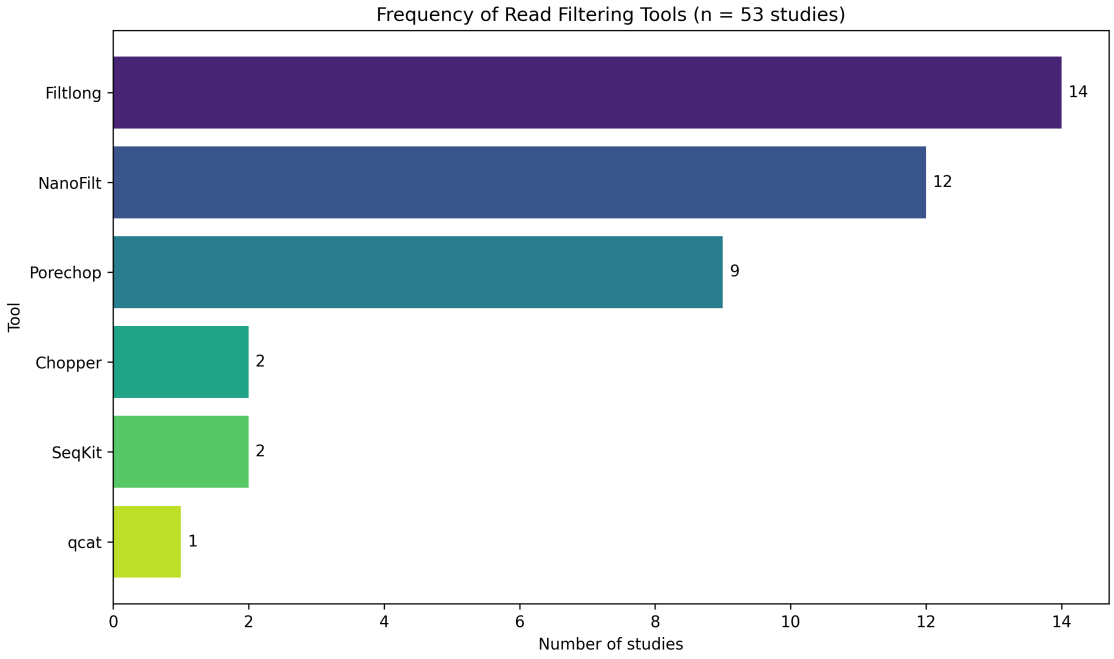

Figura 10: Ferramentas de filtragem das *reads*

Considerando a frequência de uso observada nesta revisão sistemática, bem como a adequação das funcionalidades oferecidas, as ferramentas Filtlong e NanoFilt foram selecionadas para as etapas subsequentes deste estudo. O Filtlong permite a seleção de leituras com base em critérios combinados de comprimento e qualidade, sendo amplamente utilizado em pipelines de montagem com dados de *nanopore* [158]. Já o NanoFilt possibilita a filtragem eficiente das leituras com base em limiares de qualidade e comprimento, integrando-se de forma direta ao ecossistema NanoPack [39].

A tabela 7 mostra quais foram os estudos que utilizaram as ferramentas de filtragem das *reads*.

Tabela 7: Ferramentas utilizadas para a filtragem das *reads* nos artigos incluídos na revisão sistemática

Artigo	Filtlong	NanoFilt	Porechop	Chopper	SeqKit	qcat
[105]	X					
[37]		X				
[155]		X				
[83]			X			
[15]	X	X	X			
[163]		X	X			
[84]	X	X				
[120]				X		
[152]		X		X		
[125]	X					
[24]		X				
[41]	X					
[43]		X				
[80]	X					
[155]	X	X				
[9]	X	X			X	X
[74]	X		X			
[108]	X		X			
[51]	X		X			
[73]		X				
[63]	X					
[46]	X					
[3]		X	X			
[121]			X			
[2]	X		X			
[12]					X	

Filtlong e NanoFilt são ferramentas amplamente utilizadas para filtragem de leituras longas; Porechop é empregado principalmente para remoção de adaptadores; SeqKit é uma ferramenta geral para manipulação de sequências; qcat é utilizado para demultiplexação e filtragem básica de dados de nanopore.

4.1.1.4 Montagem dos Genomas

A Figura 11 apresenta a frequência das ferramentas utilizadas para a montagem de genomas nos estudos incluídos nesta revisão sistemática. O montador Flye foi o mais frequentemente empregado, aparecendo em 32 artigos, seguido por Unicycler (19 artigos) e Canu (14 artigos). Outros montadores também foram relatados, como Raven (11 artigos) e Miniasm (9 artigos), enquanto ferramentas como Shasta, NECAT, wtdbg2 e MaSuRCA apresentaram menor frequência de uso. Diversos montadores adicionais, incluindo Tricycler, NanoPhase, NextDenovo, MetaMDBG, LJA, ybacter, Smartdenovo, SLANG, Redbean e HASLR, foram utilizados em apenas um artigo cada.

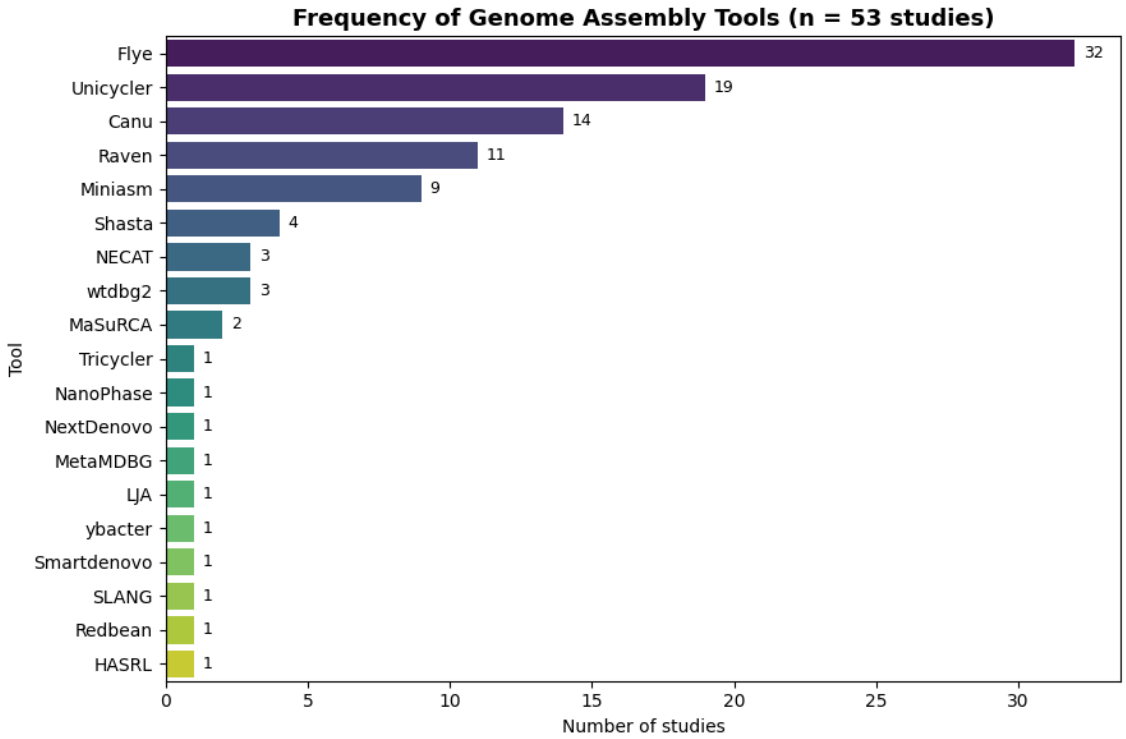

Figura 11: Montadores de genoma utilizados nos estudos incluídos na revisão sistemática

Com base nos resultados da revisão e em testes preliminares realizados neste trabalho, optou-se por utilizar exclusivamente o montador Flye [78] nas etapas subsequentes do pipeline. Essa escolha foi motivada por sua ampla adoção na literatura, robustez no tratamento de leituras longas oriundas de sequenciamento por *Nanopore* e bom equilíbrio entre qualidade da montagem e custo computacional.

O montador Canu, embora amplamente utilizado em estudos anteriores, foi avaliado durante a fase inicial deste trabalho e mostrou-se computacionalmente muito custoso, apresentando elevado consumo de tempo de processamento e recursos computacionais, especialmente em conjuntos de dados maiores. Dessa forma, sua utilização foi descartada no delineamento final do pipeline, visando maior eficiência e reprodutibilidade das análises.

A tabela 8 evidencia quais os montadores de genomas utilizados pelos autores.

Tabela 8: Ferramentas de montagem de genomas utilizadas nos artigos incluídos na revisão sistemática

Artigo	Flye	Unicycler	Canu	Raven	Miniasm	Outros
[154]	X					
[25]		X			X	X
[83]			X			
[15]	X	X	X	X	X	
[82]	X					
[163]		X				
[4]	X	X				
[84]	X	X		X		
[79]			X		X	
[25]	X		X		X	X
[126]	X					
[21]		X				
[152]	X	X	X			
[125]		X				
[24]	X					
[41]	X					
[43]						X
[80]	X		X			X
[105]	X	X	X	X		X
[37]	X	X				
[75]	X		X			X
[157]		X				
[94]	X					
[155]	X		X	X	X	
[59]	X	X	X	X		X
[88]			X			
[62]	X		X		X	
[120]	X					
[167]	X					
[153]	X					
[159]	X	X	X	X		X
[156]	X		X	X	X	X
[9]	X	X				
[14]		X				
[74]		X				
[50]	X					
[108]	X					
[76]	X					
[51]		X				
[77]		X				
[91]					X	
[73]						X
[63]	X	X				
[46]						X
[22]	X					
[3]	X	X		X	X	
[121]		X				
[2]	X					
[12]	X					

Outros: Shasta, NECAT, wtdbg2, MaSuRCA, Tricycler, NanoPhase, NextDenovo, entre outros.

4.1.1.5 Controle da Qualidade das Montagens

A Figura 12 apresenta a frequência das ferramentas utilizadas para o controle de qualidade das montagens de genoma identificadas nesta revisão. As ferramentas mais frequentemente relatadas foram o QUAST, presente em 13 artigos, seguido pelo BUSCO, utilizado em 6 artigos, e pelo CheckM, citado em 5 artigos. Outras ferramentas, como Bandage, NanoPhase (pipeline) e GAEP, apresentaram menor frequência de uso.

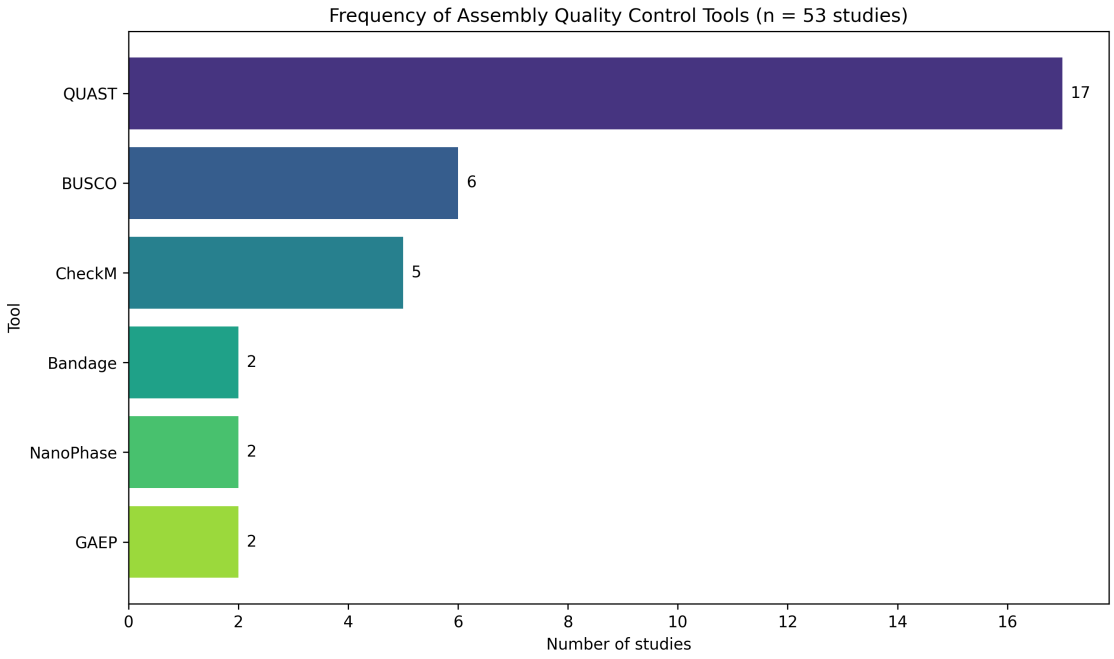

Figura 12: Ferramentas para controle de qualidade da montagem

No presente estudo, as ferramentas selecionadas para a avaliação da qualidade das montagens foram o QUAST [58] e o BUSCO [138], por serem amplamente utilizadas na literatura e fornecerem métricas complementares para a análise da qualidade estrutural e da completude genômica. Essas análises foram realizadas inicialmente após a etapa de montagem e repetidas após o processo de polimento, permitindo avaliar o impacto do polimento na melhoria da qualidade das montagens.

O QUAST fornece métricas estruturais importantes, como o número de *contigs*, o tamanho total da montagem e estatísticas baseadas em N, como o N50, que representa o comprimento do menor *contig* necessário para cobrir 50% do genoma montado. Métricas baseadas em N são amplamente utilizadas para avaliar a contiguidade das montagens, sendo valores mais altos indicativos de montagens mais contínuas.

O BUSCO, por sua vez, avalia a completude genômica com base na presença de genes conservados de cópia única esperados para um determinado grupo taxonômico. Essa análise permite classificar os genes em completos, duplicados, fragmentados ou ausentes, fornecendo uma estimativa da completude biológica da montagem.

Além disso, métricas relacionadas a erros estruturais, como inserções e deleções

(*indels*), são fundamentais para a avaliação da qualidade das montagens, especialmente em dados de sequenciamento por nanoporos, nos quais esse tipo de erro está frequentemente associado às limitações do processo de sequenciamento e *basecalling*. Nesse contexto, o número de *indels* por 100 kb é uma métrica amplamente utilizada para quantificar esses erros. A redução no número de *indels* após o polimento indica uma melhoria na acurácia da sequência montada.

Um provável motivo para a menor frequência das ferramentas de controle de qualidade da montagem nos artigos analisados está relacionado ao foco metodológico de muitos estudos incluídos na revisão. Diversos trabalhos têm como objetivo principal a proposição ou avaliação de pipelines, algoritmos de montagem ou estratégias experimentais, tratando a avaliação da qualidade da montagem como uma etapa secundária ou implícita, frequentemente resumida a poucas métricas ou mesmo incorporada a pipelines automatizados. Outro fator relevante é que estudos voltados a aplicações específicas (por exemplo, vigilância genômica, análise comparativa ou caracterização funcional) podem priorizar análises *downstream*, reduzindo o detalhamento ou a diversidade de ferramentas empregadas na avaliação da montagem.

A tabela 9 indica quais autores utilizaram essas ferramentas em seus trabalhos.

Tabela 9: Ferramentas de controle de qualidade da montagem de genomas utilizadas nos artigos incluídos na revisão sistemática

Artigo	QUAST	BUSCO	CheckM	Bandage	Outros
[27]	X	X			
[83]		X			
[15]	X			X	
[79]	X				
[152]	X				
[24]		X			
[88]		X			
[126]	X				
[41]	X				
[80]	X				
[37]	X				
[94]	X				
[59]	X				
[21]				X	
[125]			X		
[9]	X				
[9]	X				
[14]	X	X			
[74]	X				
[50]			X		
[108]			X		
[91]		X			
[73]					X
[63]	X				
[3]					X
[121]	X				
[12]	X				
[46]			X		
[22]			X		

Outros: NanoPhase (pipeline), GAEP.

4.1.1.6 Mapeamento

A Figura 13 mostra a frequência de ferramentas utilizadas para mapeamento, com a maioria dos artigos utilizando a ferramenta minimap2, em um total de 20 artigos. O MUMmer foi utilizado em 3 artigos, o Bowtie em 2 artigos e as demais ferramentas observadas nesta revisão para esta etapa — BBmap e GraphMap — apareceram em apenas um artigo cada.

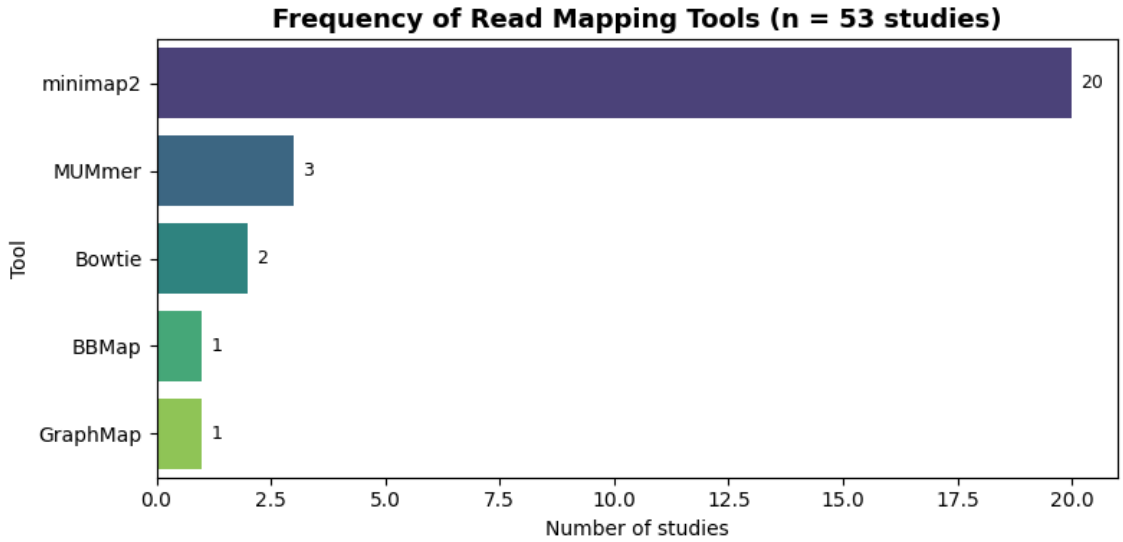

Figura 13: Ferramentas para mapeamento

A etapa de mapeamento foi incluída no delineamento metodológico por ser um componente essencial de estratégias de polimento baseadas em alinhamento, nas quais as *reads* são realinhadas ao genoma montado para correção de erros. Entretanto, neste estudo, o mapeamento não foi executado como uma etapa isolada do fluxo principal, pois ele já ocorre de forma interna nas ferramentas de polimento utilizadas (por exemplo, nas rodadas de Racon e Medaka), que realizam o alinhamento automaticamente como parte de seus próprios procedimentos. Assim, optou-se por não inserir uma etapa explícita de mapeamento no *script* automatizado. Entre as ferramentas identificadas na revisão sistemática, o minimap2 foi a mais frequentemente utilizada e, por isso, foi adotado como referência metodológica para o alinhamento interno nas etapas de polimento [85].

A tabela 10 indica quais ferramentas de mapeamento apareceram na revisão.

Tabela 10: Ferramentas de mapeamento utilizadas nos artigos incluídos na revisão sistemática

Artigo	Minimap2	MUMmer	Bowtie	BBmap	GraphMap
[163]	X	X			
[82]	X				
[17]	X				
[4]	X				
[18]	X				
[79]	X			X	
[25]		X			
[21]	X				
[152]					X
[41]	X				
[80]	X				
[105]	X				
[75]	X				
[158]	X				
[155]	X				
[59]	X				
[88]	X	X			
[62]	X				
[44]	X				
[120]	X				
[167]	X				
[9]	X		X		
[50]			X		

Minimap2 foi a ferramenta de mapeamento mais frequente (20 artigos), seguida por MUMmer (3), Bowtie (2), BBmap (1) e GraphMap (1).

4.1.1.7 Polimento

A etapa de polimento tem como objetivo reduzir erros residuais das montagens geradas a partir de leituras longas, especialmente inserções, deleções e substituições decorrentes das características do sequenciamento por *Nanopore*. Conforme observado na revisão sistemática (Figura 14), os polidores mais frequentemente empregados foram Medaka [112], presente em 25 artigos, e Racon [151], em 17 artigos, sendo portanto selecionados para compor o fluxo experimental deste estudo. O Racon é um polidor baseado em alinhamento de leituras ao genoma montado, realizando correções iterativas de consenso, enquanto o Medaka utiliza modelos de aprendizado profundo treinados para dados *Nanopore* a fim de refinar a sequência consenso e melhorar a acurácia final.

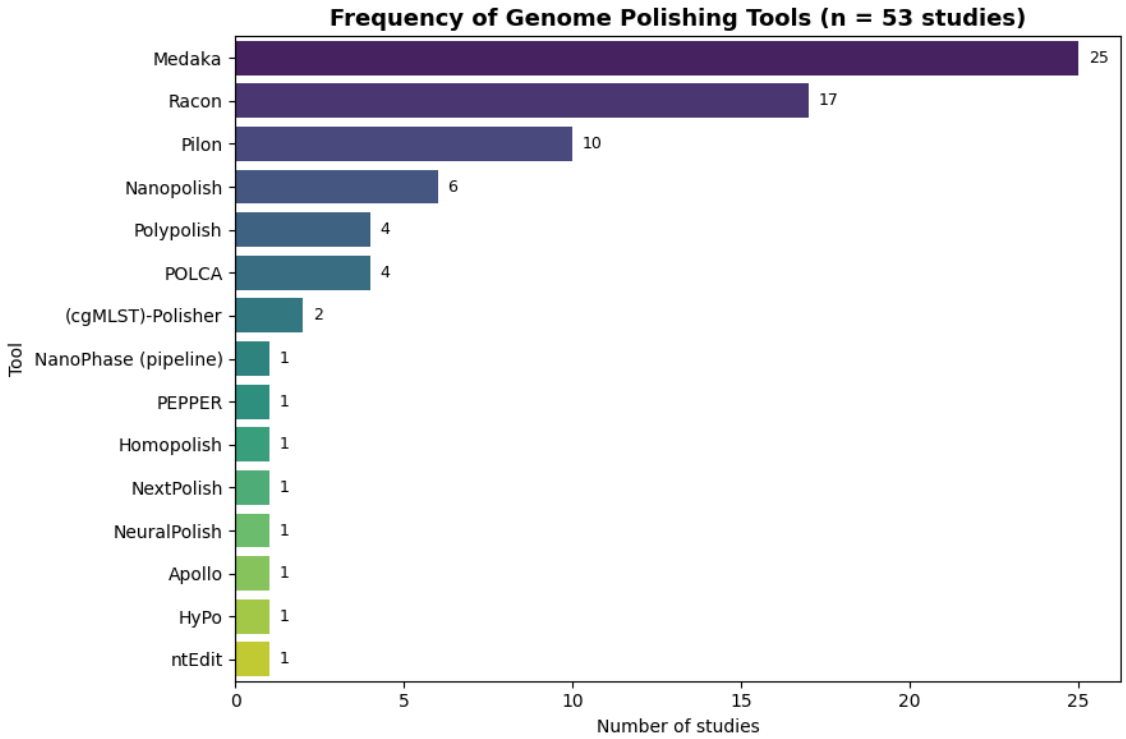

Figura 14: Polidores

No delineamento experimental adotado, foram avaliadas diferentes estratégias de polimento com o objetivo de verificar o impacto do número e da combinação de rodadas de correção sobre a qualidade final das montagens. As seguintes combinações foram testadas: (i) montagem raw (sem polimento) (ii) uma rodada de Racon; (iii) mais uma rodada de Racon na montagem já polida uma vez (iv) uma rodada de Racon seguida por uma rodada de Medaka; e (v) duas rodadas de Racon seguidas por uma rodada de Medaka e (vi) uma rodada de Medaka sozinho. A utilização de múltiplas rodadas de Racon é comum em fluxos de montagem com leituras longas, pois as primeiras iterações tendem a corrigir erros mais evidentes, enquanto rodadas adicionais refinam o consenso. A aplicação do Medaka após o Racon busca realizar um ajuste final com base em modelos específicos para sinais de sequenciamento *Nanopore*, frequentemente resultando em melhorias adicionais na acurácia e redução de *indels*.

Após o polimento, as montagens foram novamente submetidas ao controle de qualidade utilizando as mesmas ferramentas aplicadas na etapa anterior de avaliação (Figura 12), como QUAST e BUSCO. Essa avaliação final permitiu verificar o ganho de qualidade estrutural e biológica obtido com as estratégias de polimento e determinar se as montagens atingiram qualidade adequada para análises posteriores.

A tabela 11 representa a frequência das ferramentas de polimento incluídas nessa revisão.

Tabela 11: Ferramentas de polimento utilizadas nos artigos incluídos na revisão sistemática

Artigo	Medaka	Racon	Pilon	Nanopolish	POLCA	Outros
[154]	X	X				
[27]		X	X			
[162]		X				
[83]	X	X		X		X
[15]	X	X				
[4]	X	X				
[84]	X					
[21]		X	X			
[152]	X	X	X			
[125]			X			
[24]	X		X			X
[80]		X	X			
[105]	X	X				X
[37]	X					
[75]	X					
[158]		X		X		
[94]	X	X	X			X
[155]	X	X				
[59]	X	X	X			
[88]			X	X		
[62]		X		X		
[120]	X					X
[167]	X					X
[153]	X					
[159]	X					
[9]	X					
[14]			X			
[50]			X			
[76]	X	X				
[77]		X	X			
[91]	X				X	
[73]						X
[63]	X				X	X
[46]					X	X
[22]	X					
[3]	X				X	X
[2]	X					
[12]	X					

Outros: NanoPhase (pipeline), PEPPER, Homopolish, NextPolish, NeuralPolish, Apollo, HyPo, ntEdit, cgMLST (pipeline).

4.2 Análise dos Fluxos e Ferramentas

A Figura 15 apresenta o fluxo metodológico adotado neste estudo para avaliação comparativa das diferentes combinações de ferramentas e estratégias aplicadas aos dados de sequenciamento por *Nanopore*. A execução do fluxo foi distribuída em dois ambientes computacionais distintos, de acordo com as demandas específicas de cada etapa.

A etapa de *basecalling*, por exigir alto desempenho computacional e utilização de aceleração por GPU, foi realizada no ambiente de Cluster do C3 (Centro de Ciências

Computacionais) da FURG. Nessa fase, os dados brutos foram processados por duas ferramentas amplamente utilizadas (Guppy e Dorado), cada uma operando com três modelos de acurácia (*Fast*, *HAC* e *SUP*). Embora o *basecalling* não tenha sido incorporado diretamente ao *script* automatizado da *pipeline*, ele integra conceitualmente o fluxo experimental, pois cada modelo gerado constituiu uma condição experimental independente considerada nas análises comparativas.

Após a geração dos arquivos *basecalled* no cluster, todas as etapas subsequentes foram executadas no desktop do COMBI-Lab (Laboratório de Biologia Computacional). Essas etapas incluíram o controle de qualidade das *reads* com NanoPlot, a filtragem com NanoFilt e Filtlong, a montagem genômica utilizando o montador Flye e as diferentes estratégias de polimento. Essa divisão operacional permitiu otimizar o uso dos recursos computacionais, reservando o ambiente com GPU para a etapa mais intensiva e mantendo as demais análises em um ambiente local controlado.

As montagens geradas foram submetidas a seis estratégias de polimento distintas, incluindo montagem bruta (sem polimento), aplicação isolada de Medaka, rodadas sucessivas de Racon e combinações entre essas ferramentas. A combinação entre duas ferramentas de *basecalling*, três modelos de acurácia, dois filtros das *reads* e seis estratégias de polimento resultou em um total de 72 combinações experimentais avaliadas para cada organismo. Esse delineamento possibilitou investigar de forma sistemática o impacto de cada etapa do fluxo na qualidade final das montagens geradas.

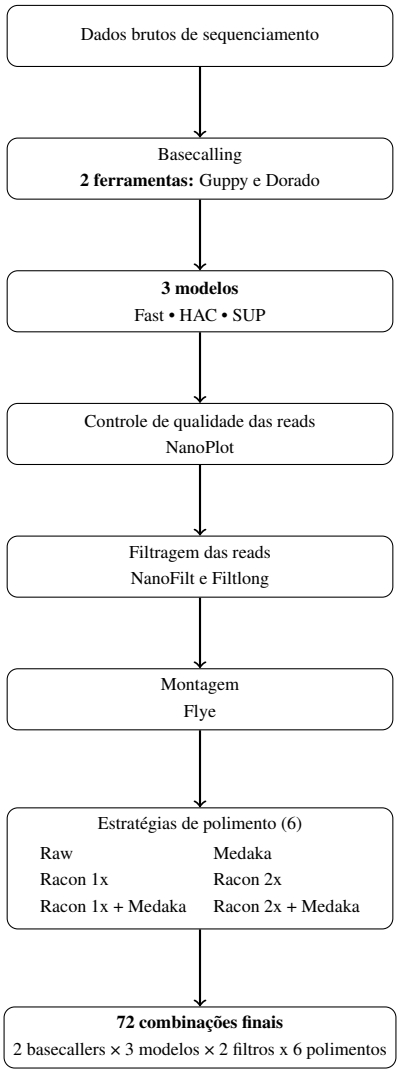

Figura 15: Fluxo metodológico geral e combinação experimental resultando em 72 montagens avaliadas.

Para comparar o desempenho das 72 combinações geradas para cada organismo, foi adotada uma abordagem de ranqueamento multifatorial baseada em métricas estruturais e de precisão amplamente utilizadas na avaliação de montagens genômicas. Foram consideradas quatro variáveis principais: (i) completude genômica estimada pelo BUSCO, (ii) número de *misassemblies*, (iii) taxas de *indels* e (iv) contiguidade da montagem, representada pelo N50.

A completude genômica foi estimada com o BUSCO, ferramenta que avalia a presença de genes conservados universais esperados para o grupo taxonômico analisado, fornecendo uma medida da integridade biológica da montagem. O número de *misassemblies* foi utilizado como indicador de erros estruturais, como inversões, translocações ou junções incorretas de regiões genômicas. As taxas de *indels* foram consideradas como métricas de precisão em nível de base, refletindo a acurácia das sequências montadas. Por fim, a contiguidade foi avaliada pelo N50, que corresponde ao comprimento mínimo dos *contigs* necessários para cobrir 50% do tamanho total da montagem, sendo amplamente

utilizado como indicador de continuidade estrutural.

Cada métrica foi classificada de acordo com sua direção de otimização: valores mais altos foram considerados desejáveis para completude (BUSCO Complete) e N50, enquanto valores mais baixos indicaram melhor desempenho para *indels* e *misassemblies*. Para cada variável, as combinações foram ordenadas e receberam um *ranking* ordinal. O escore final de cada combinação foi calculado pela soma dos *ranks* atribuídos em todas as métricas avaliadas. As combinações com menor escore total foram consideradas aquelas com melhor desempenho global.

Essa estratégia permitiu integrar múltiplas dimensões de qualidade em uma única métrica comparativa, possibilitando a identificação objetiva das combinações de ferramentas e parâmetros que apresentaram melhor equilíbrio entre precisão, completude e contiguidade das montagens geradas.

4.2.1 *Basecalling*

*4.2.1.1 Desempenho computacional do **Guppy** nos modelos fast, hac e sup*

A Tabela 12 apresenta o desempenho computacional da ferramenta Guppy nos três modelos de acurácia (*fast*, *hac* e *sup*) para os seis organismos analisados. O tempo de execução foi mensurado em minutos por meio do tempo real (*wall-clock time*), obtido com o comando *time* em ambiente de cluster com utilização de GPU.

Tabela 12: Tempo de execução (minutos) do Guppy nos modelos *fast*, *hac* e *sup*.

Organismo	**fast**	**hac**	**sup**
Acinetobacter pittii	0,20	2,11	3,20
Haemophilus haemolyticus	0,26	1,52	3,20
Klebsiella pneumoniae	4,53	6,06	8,29
Shigella sonnei	6,45	8,43	11,49
Staphylococcus aureus	2,59	3,24	6,17
Stenotrophomonas maltophilia	1,29	10,34	15,34

*4.2.1.2 Desempenho computacional do **Dorado** nos modelos fast, hac e sup*

A Tabela 13 apresenta o tempo de execução da ferramenta Dorado nos três modelos de acurácia avaliados. Assim como para o Guppy, os tempos foram obtidos em ambiente de cluster com uso de GPU.

Tabela 13: Tempo de execução (minutos) do Dorado nos modelos *fast*, *hac* e *sup*.

Organismo	**fast**	**hac**	**sup**
Acinetobacter pittii	0,14	0,37	2,16
Haemophilus haemolyticus	0,29	1,25	4,50
Klebsiella pneumoniae	0,59	2,42	11,18
Shigella sonnei	1,16	3,47	15,54
Staphylococcus aureus	0,44	2,30	9,00
Stenotrophomonas maltophilia	0,55	2,31	10,12

De forma geral, observou-se aumento progressivo do tempo de execução conforme a complexidade do modelo de *basecalling*, sendo os modelos *sup* os mais custosos computacionalmente em ambas as ferramentas. O Dorado apresentou tempos menores nos modelos *fast* e *hac*, enquanto o modelo *sup* demonstrou aumento expressivo no tempo de processamento para ambas as ferramentas. O Guppy apresentou maior variabilidade entre organismos, sugerindo diferenças na forma como cada ferramenta escala com o volume de dados de entrada.

A comparação apresentada nesta seção considera exclusivamente o custo computacional da etapa de *basecalling*, mensurado em minutos de execução em ambiente de cluster com GPU. Entretanto, a avaliação do desempenho dos modelos não se baseia apenas no tempo de processamento. A qualidade das leituras geradas por cada modelo de *basecalling* foi avaliada indiretamente nas etapas subsequentes do fluxo, por meio do impacto nas montagens genômicas e nas métricas estruturais e de precisão obtidas ao final do pipeline. Dessa forma, a análise do *basecalling* neste trabalho integra duas dimensões complementares: (i) o custo computacional de execução em ambiente de cluster e (ii) a qualidade resultante das montagens geradas ao longo do fluxo automatizado.

Cabe destacar ainda que, como a execução ocorreu em ambiente de cluster com GPU compartilhada, variações na carga de processamento podem influenciar os tempos absolutos observados, embora o padrão geral de aumento de custo entre os modelos se mantenha consistente.

4.2.2 Avaliação individual por organismo

4.2.2.1 Acinetobacter pittii

A avaliação individual de *Acinetobacter pittii* foi conduzida considerando todas as 72 combinações testadas. A análise foi estruturada inicialmente a partir das métricas globais de completude (BUSCO), taxa de *indels* (QUAST) e contiguidade (N50), representadas graficamente para todas as combinações avaliadas.

- Completude genômica (BUSCO)

  A análise da completude (BUSCO *Complete*) evidencia um padrão claro de separação entre os modelos de *basecalling*. Observa-se aumento progressivo da completude ao comparar os modelos *fast*, *hac* e *sup*, independentemente do *basecaller* utilizado. Esses resultados são evidenciados de maeira gráfica na Figura 16

  As combinações envolvendo o modelo *sup* concentraram os maiores valores de BUSCO, atingindo até 98,5%. O modelo *hac* apresentou desempenho intermediário (94–96%), enquanto o modelo *fast* apresentou os menores valores, especialmente nas montagens não polidas.

  O efeito do polimento foi consistente em todos os cenários, com aumento sis-

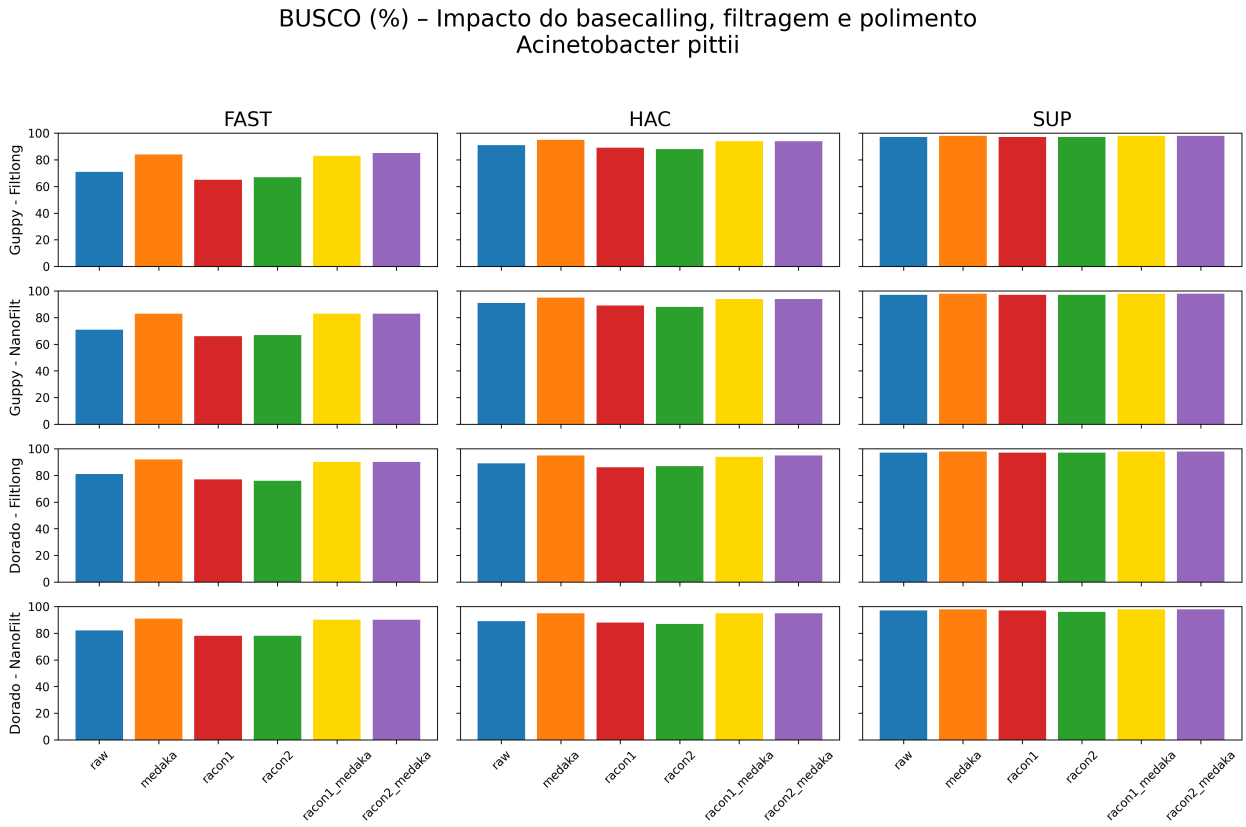

Figura 16: Avaliação da completude genômica das montagens de *Acinetobacter pittii* utilizando BUSCO, considerando diferentes combinações de ferramentas de *basecalling* (Guppy e Dorado), modelos de acurácia (FAST, HAC e SUP), filtros de melhoria das reads (Nanofilt e Filtlong) e estratégias de polimento (sem polimento, Racon, Medaka e combinações).

temático da completude após aplicação de Medaka ou combinações com Racon. Entretanto, o ganho absoluto foi mais expressivo nos modelos de menor acurácia inicial (*fast* e *hac*), enquanto no modelo sup o polimento atuou como refinamento final.

- Taxa de *indels* (QUAST)

  A taxa de *indels* por 100 kb apresentou comportamento inverso ao BUSCO, conforme esperado. O modelo *sup* concentrou os menores valores absolutos de *indels*, seguido por *hac* e, por último, *fast*.

  O polimento promoveu redução substancial dos *indels* em todas as configurações. Nas montagens brutas derivadas do modelo fast, os valores ultrapassavam 70 *indels*/100 kb, sendo reduzidos para aproximadamente 30 após polimento com Medaka. Já no modelo *sup*, os valores iniciais eram significativamente menores (16 *indels*/100 kb) e foram reduzidos para menos de 9 *indels*/100 kb após polimento.

  Observa-se que a aplicação exclusiva de Medaka já foi suficiente para promover a maior parte da correção de erros, enquanto rodadas adicionais de Racon produziram

apenas variações marginais.

Esses resultados são mostrados na Figura 17

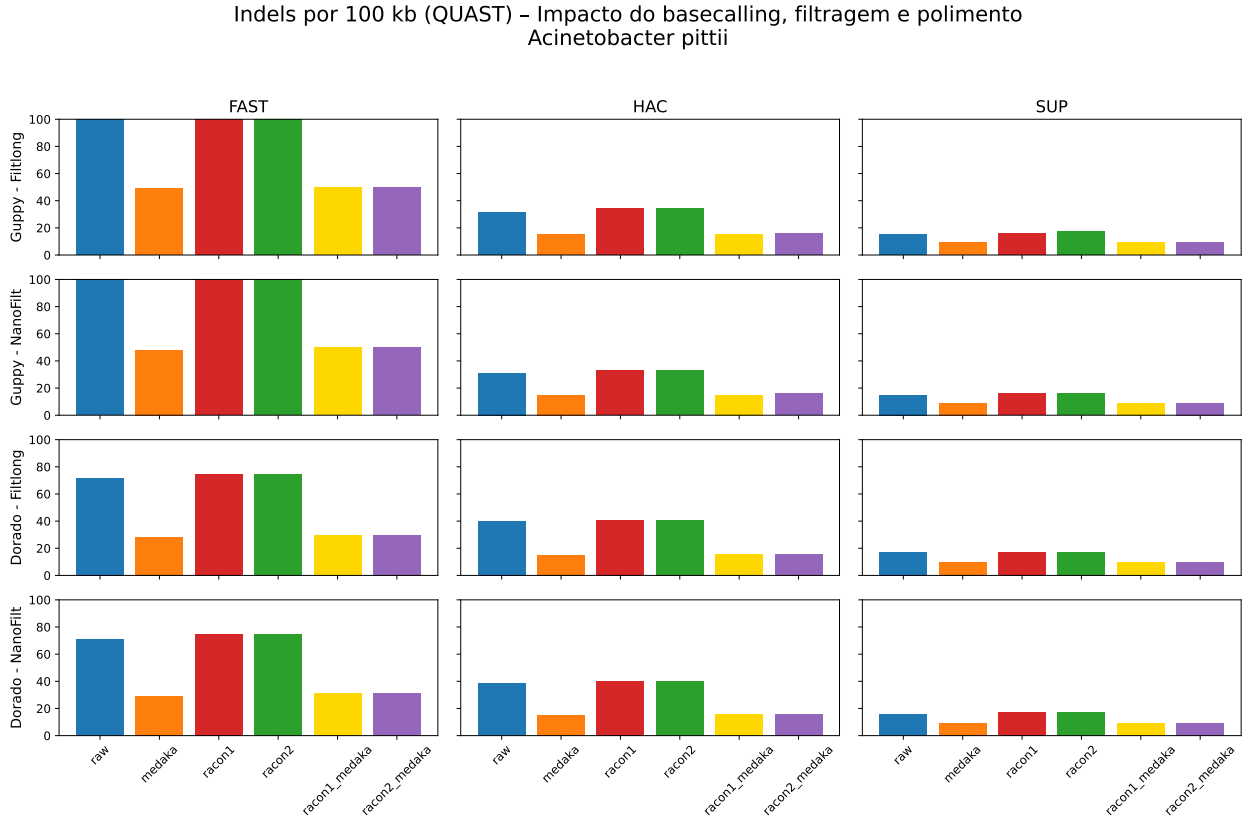

Figura 17: Taxa de *indels* por 100 kb (QUAST) em montagens de *Acinetobacter pittii* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Contiguidade (N50)

  Para facilitar a visualização das diferenças de contiguidade, o N50 foi representado como diferença em relação ao menor valor observado (N50). Essa abordagem evita a compressão visual causada pelos valores absolutos próximos a 3,8 Mb (megabases).

  O padrão observado reforça o comportamento das métricas anteriores:

  As maiores diferenças positivas de N50 concentram-se no modelo *sup*. O modelo *hac* apresenta desempenho intermediário. O modelo *fast* concentra as menores diferenças relativas.

  O polimento promoveu ganhos discretos no N50, sendo seu impacto menos pronunciado do que nas métricas de erro (*indels*) e completude (BUSCO). Isso indica que o polimento atua principalmente na correção de erros de base, com menor influência na estrutura global da montagem.

  Os resultados de N50 podem ser analisados na Figura

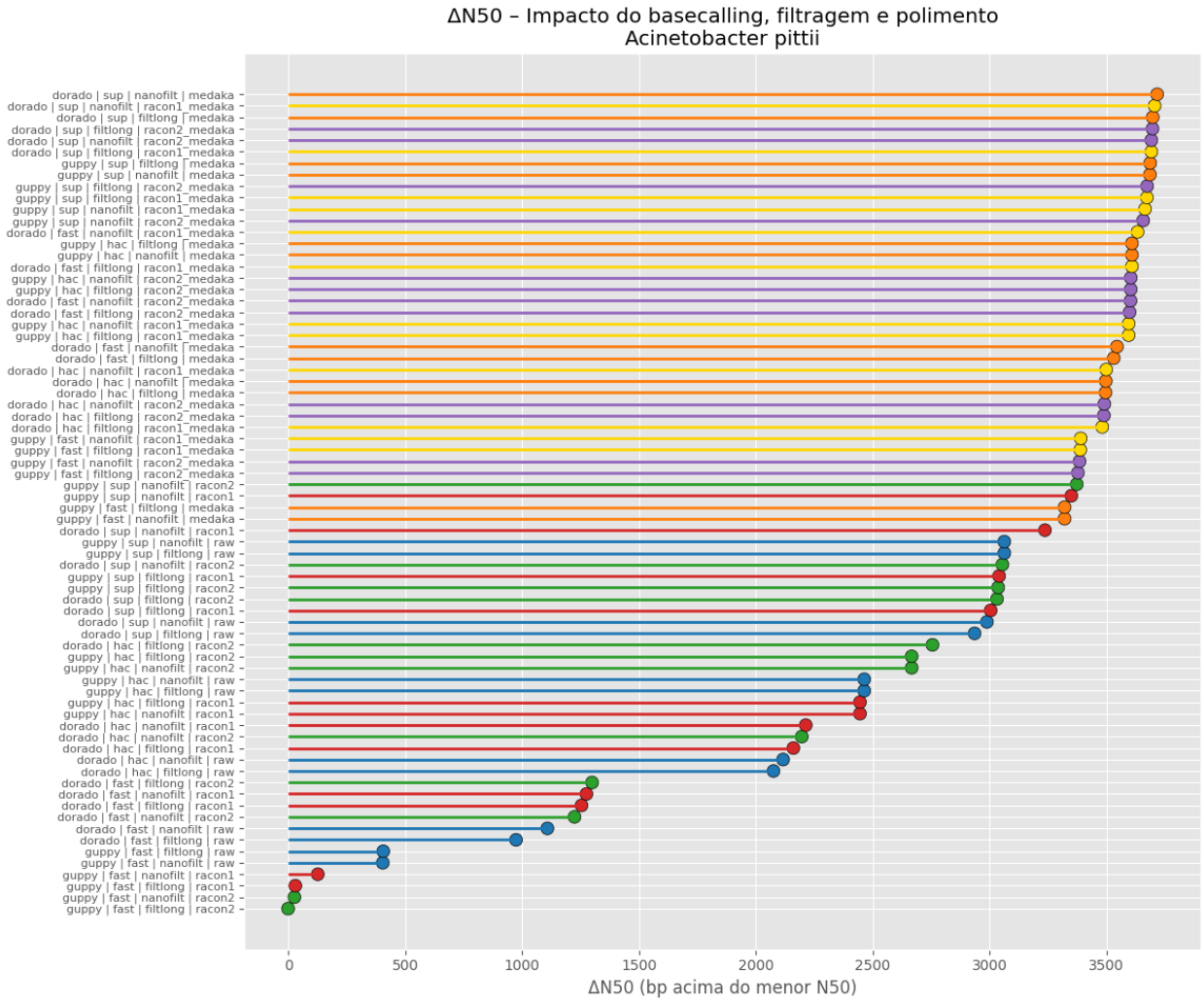

Figura 18: Variação da contiguidade das montagens (N50) para *Acinetobacter pittii* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

N50 = N50 menor N50 Ou seja: quanto maior → melhor que o pior cenário

- Síntese dos melhores fluxos

  A Tabela 14 apresenta as cinco combinações com melhor escore global, considerando o ranqueamento multifatorial.

Tabela 14: Top 5 combinações ranqueadas para *A. pittii* com base no escore global.

Rank	Basecaller	Modelo	Filtro	Polimento	BUSCO (%)	N50 (bp)	Indels/100kb
1	Dorado	sup	NanoFilt	Medaka	98.5	3,815,297	8,94
2	Dorado	sup	NanoFilt	1xRacon+Medaka	98.3	3,815,286	9,44
3	Guppy	sup	NanoFilt	Medaka	98.0	3,815,266	8,86
4	Guppy	sup	Filtlong	Medaka	98.0	3,815,267	8,89
5	Dorado	sup	NanoFilt	2xRacon+Medaka	98.3	3,815,272	9,39

Todas as cinco melhores combinações utilizaram o modelo *sup*, reforçando a influência determinante da acurácia do *basecalling* na qualidade final da montagem.

O filtro NanoFilt esteve presente em quatro das cinco melhores combinações, sugerindo impacto ligeiramente superior em relação ao Filtlong nas métricas integradas.

O melhor fluxo foi: Dorado – *sup* – NanoFilt – Medaka

Essa configuração apresentou:

Maior completude (98,5%)

Elevado N50 (3.815.297 bp)

Baixa taxa de *indels* (8,94/100 kb)

A inclusão de rodadas adicionais de Racon antes do Medaka não resultou em ganhos substanciais. As diferenças entre Medaka isolado e combinações com Racon foram pequenas e não alteraram significativamente o desempenho estrutural ou de precisão.

A análise conjunta das métricas indica que: o modelo de *basecalling* é o principal determinante da qualidade estrutural e da precisão final.

O polimento com Medaka é suficiente para atingir alto padrão de qualidade. Estratégias mais complexas de polimento produzem ganhos marginais. O modelo *sup* amplifica os efeitos positivos do polimento.

4.2.2.2 Haemophylus haemolyticus

As Figuras 19, 20 e 21 sintetizam o impacto combinado das decisões de *basecalling* (Guppy vs. Dorado), do modelo (*fast/hac/sup*), do filtro (NanoFilt/Filtlong) e das estratégias de polimento sobre três métricas-chave: completude (BUSCO), precisão em nível de base (*indels*/100 kb, QUAST) e contiguidade (N50; aqui representado como N50 em relação ao pior cenário).

- Completude genômica (BUSCO)

  Pela Figura 19, observa-se que a completude (BUSCO *Complete*) é fortemente influenciada pelo modelo de *basecalling*: as configurações com modelo *sup* tendem a concentrar os maiores valores, enquanto o *fast* apresenta desempenho mais limitado e maior sensibilidade às variações do pós-processamento. Em todos os cenários, as estratégias envolvendo polimento com Medaka (isolado ou após Racon) aparecem associadas aos melhores patamares de completude, sugerindo ganhos consistentes na recuperação de genes conservados.

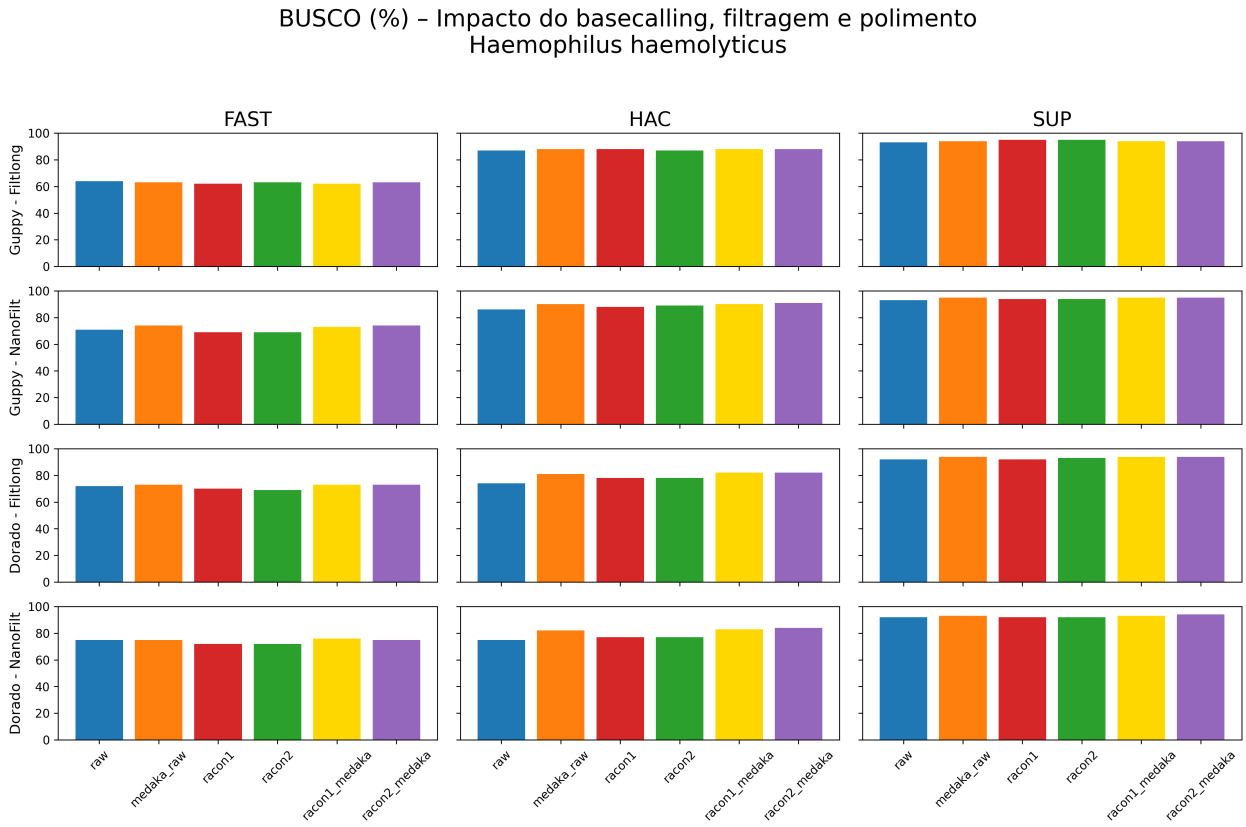

Figura 19: Avaliação da completude genômica das montagens de *Haemophilus haemolyticus* utilizando BUSCO, considerando diferentes combinações de ferramentas de *basecalling* (Guppy e Dorado), modelos de acurácia (FAST, HAC e SUP), filtros de melhoria das reads (Nanofilt e Filtlong) e estratégias de polimento (sem polimento, Racon, Medaka e combinações).

- Taxa de *indels* (QUAST)

  A Figura 20 indica um padrão complementar ao BUSCO: as menores taxas de *indels* tendem a ocorrer com o modelo *sup*, e o polimento reduz os erros em nível de base quando comparado às montagens brutas (*raw*). Em particular, combinações com Racon + Medaka aparecem como estratégias mais robustas quando a meta é reduzir *indels*.

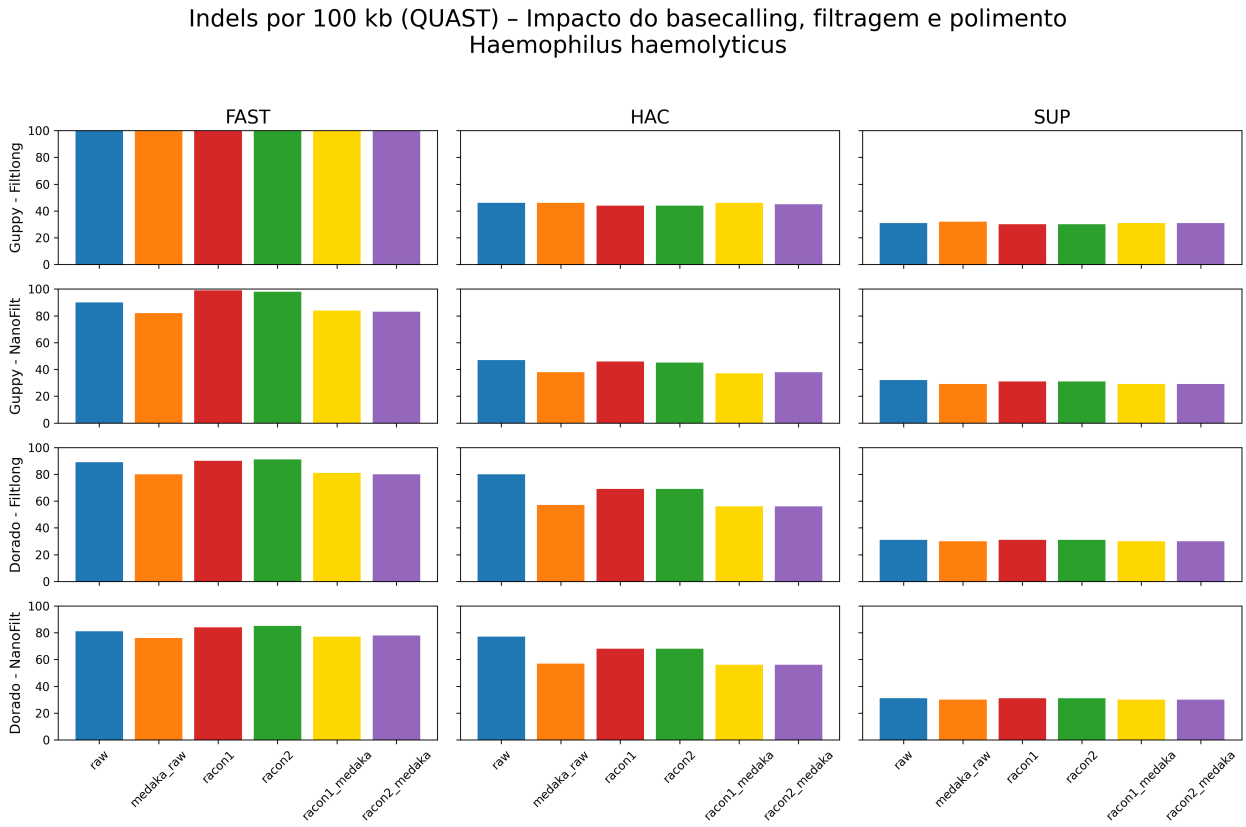

Figura 20: Taxa de *indels* por 100 kb (QUAST) em montagens de *Haemophilus haemolyticus* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Contiguidade (N50)

  Na Figura 21, as combinações com maiores N50 concentram-se nas configurações com modelo *sup*, sugerindo que a maior acurácia do *basecalling* facilita montagens mais contíguas (maior N50), enquanto cenários de menor acurácia (principalmente *fast*) tendem a limitar a contiguidade final.

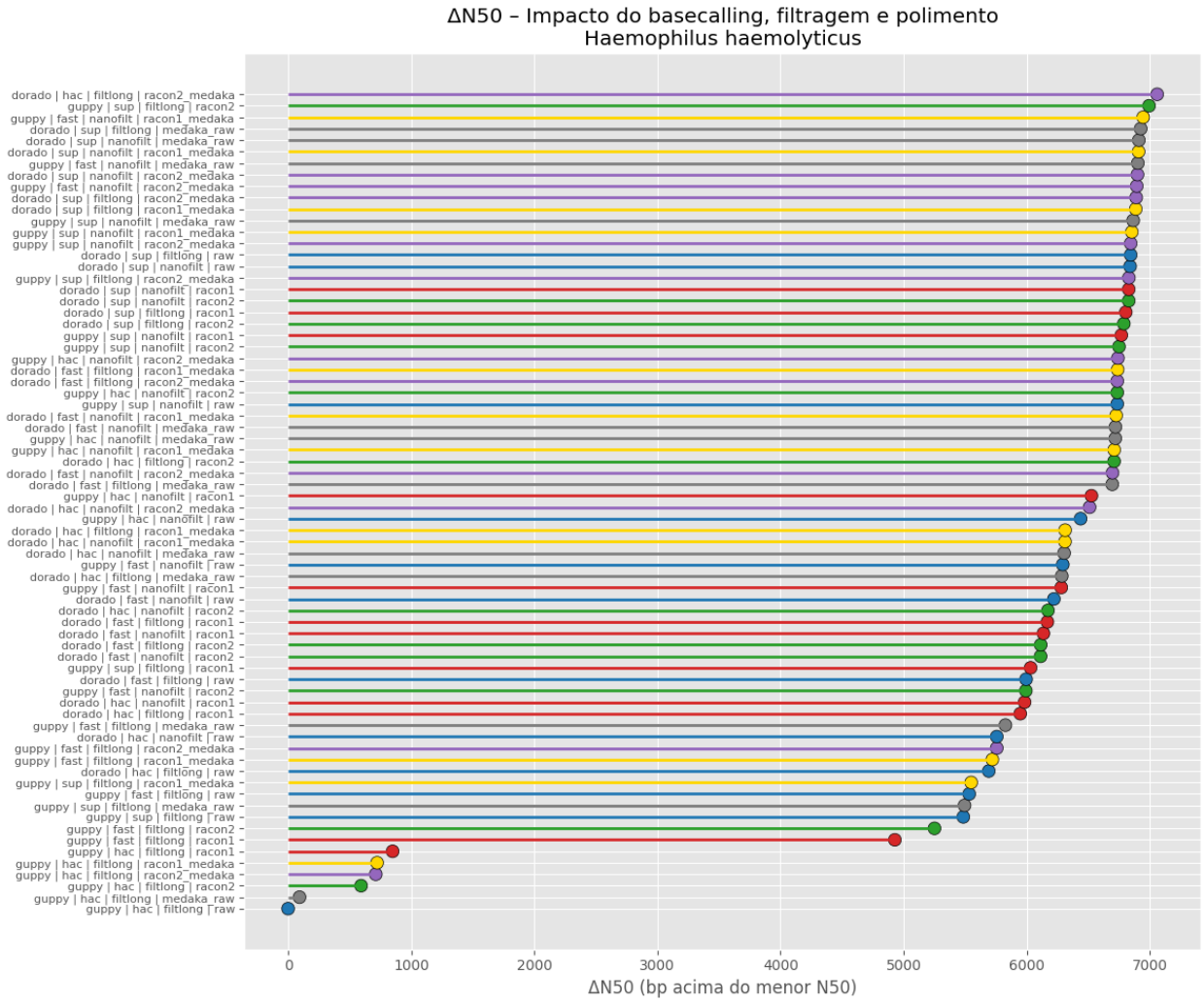

Figura 21: Variação da contiguidade das montagens (N50) para *Haemophilus haemolyticus* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Síntese dos melhores fluxos

  A Tabela 15 apresenta as cinco combinações de melhor desempenho global para *H. haemolyticus*, considerando simultaneamente completude (BUSCO), contiguidade (N50) e precisão (*indels*/100 kb), conforme o escore multifatorial descrito na metodologia.

  Um ponto central é que todas as cinco melhores combinações foram obtidas com Dorado no modelo *sup*, o que reforça que, para este organismo, a acurácia do *basecalling* foi o principal fator determinante do desempenho global. Além disso, as diferenças entre as primeiras posições são pequenas (BUSCO e N50 praticamente invariantes), e o que passa a discriminar melhor as combinações é sobretudo a taxa de *indels*, isto é, precisão em nível de base. Nesse contexto, as estratégias com 1xRacon + Medaka aparecem de forma recorrente no topo, sugerindo que a

combinação sequencial de polidores oferece um balanço favorável no refinamento final.

Tabela 15: Top 5 combinações ranqueadas para *H. haemolyticus* com base no escore global.

Rank	Basecaller	Modelo	Filtro	Polimento	BUSCO (%)	N50 (bp)	Indels/100kb
1	Dorado	sup	Filtlong	1xRacon+Medaka	93.5	2,051,276	30,36
2	Dorado	sup	NanoFilt	1xRacon+Medaka	93.5	2,051,270	30,17
3	Dorado	sup	NanoFilt	2xRacon+Medaka	93.5	2,051,271	30,27
4	Dorado	sup	Filtlong	Medaka	93.5	2,051,273	30,66
5	Dorado	sup	NanoFilt	Medaka	93.3	2,051,267	30,56

4.2.2.3 Klebsiella pneumoniae

- Completude Genômica (BUSCO)

  A Figura 22 de completude genômica (BUSCO) evidencia que a variação entre os modelos de *basecalling* exerce influência direta sobre a qualidade final das montagens. Observa-se um aumento progressivo dos valores de BUSCO *Complete* dos modelos *fast* para *hac* e, sobretudo, para *sup*, independentemente do *basecaller* ou da estratégia de polimento empregada. As combinações baseadas no modelo *sup* concentram os maiores valores de completude, atingindo valores superiores a 97% nas melhores configurações. Nos modelos *fast*, a completude apresenta maior dispersão e valores consistentemente inferiores, enquanto o modelo *hac* ocupa posição intermediária, indicando que a acurácia inicial do *basecalling* é um dos principais determinantes da recuperação de genes conservados.

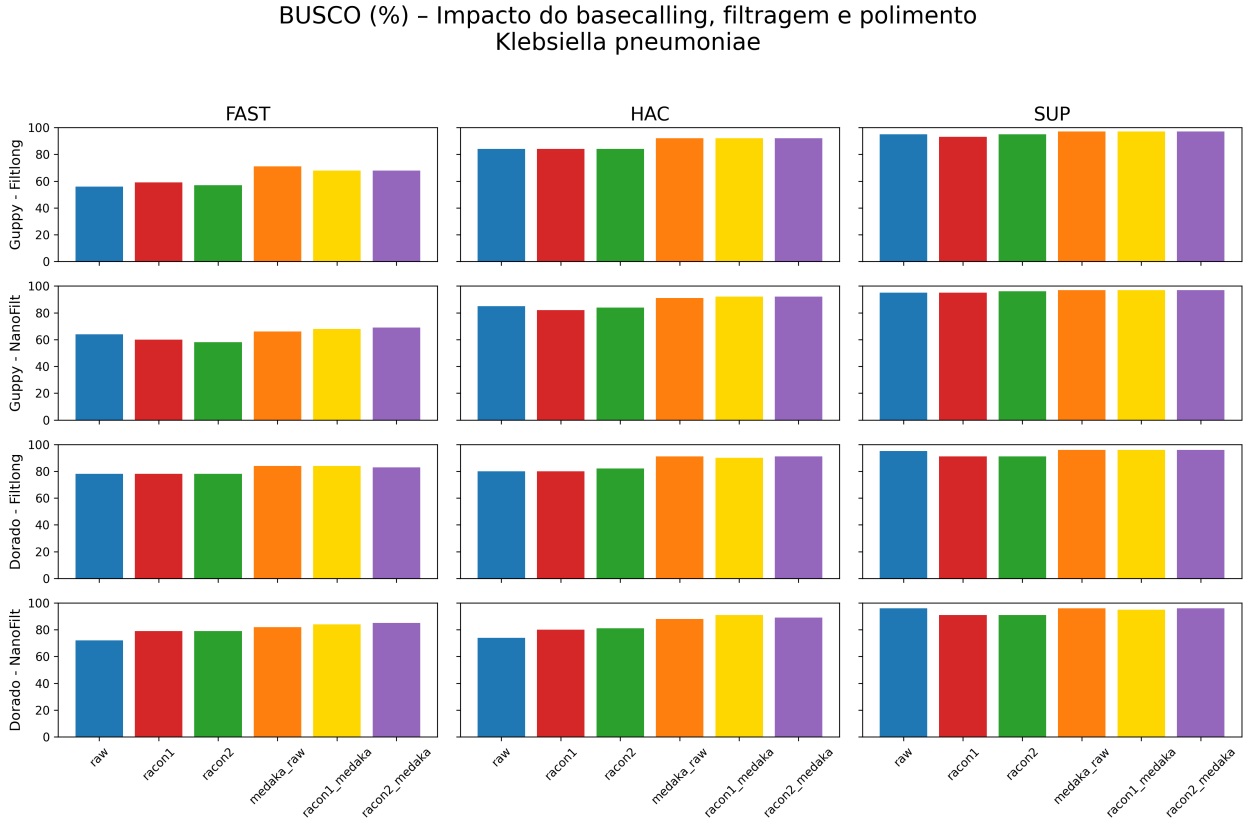

Figura 22: Avaliação da completude genômica das montagens de *Klebsiella pneumoniae* utilizando BUSCO, considerando diferentes combinações de ferramentas de *basecalling* (Guppy e Dorado), modelos de acurácia (FAST, HAC e SUP), filtros de melhoria das reads (Nanofilt e Filtlong) e estratégias de polimento (sem polimento, Racon, Medaka e combinações)

- Taxa de *indels* (QUAST)

  A análise das taxas de *indels* por 100 kb (Figura 23) reforça esse padrão. As montagens derivadas do modelo *sup* apresentaram as menores taxas de erro em nível de base, com redução consistente após o polimento. Nos modelos *fast* e *hac*, as taxas de *indels* são mais elevadas e mostram maior sensibilidade à estratégia de polimento aplicada. Observa-se que o polimento com Medaka, isoladamente ou combinado a rodadas de Racon, promove redução sistemática das taxas de erro, com impacto mais evidente quando aplicado a montagens provenientes do modelo *sup*. Esse comportamento sugere que montagens iniciais mais acuradas permitem maior eficiência das etapas subsequentes de correção.

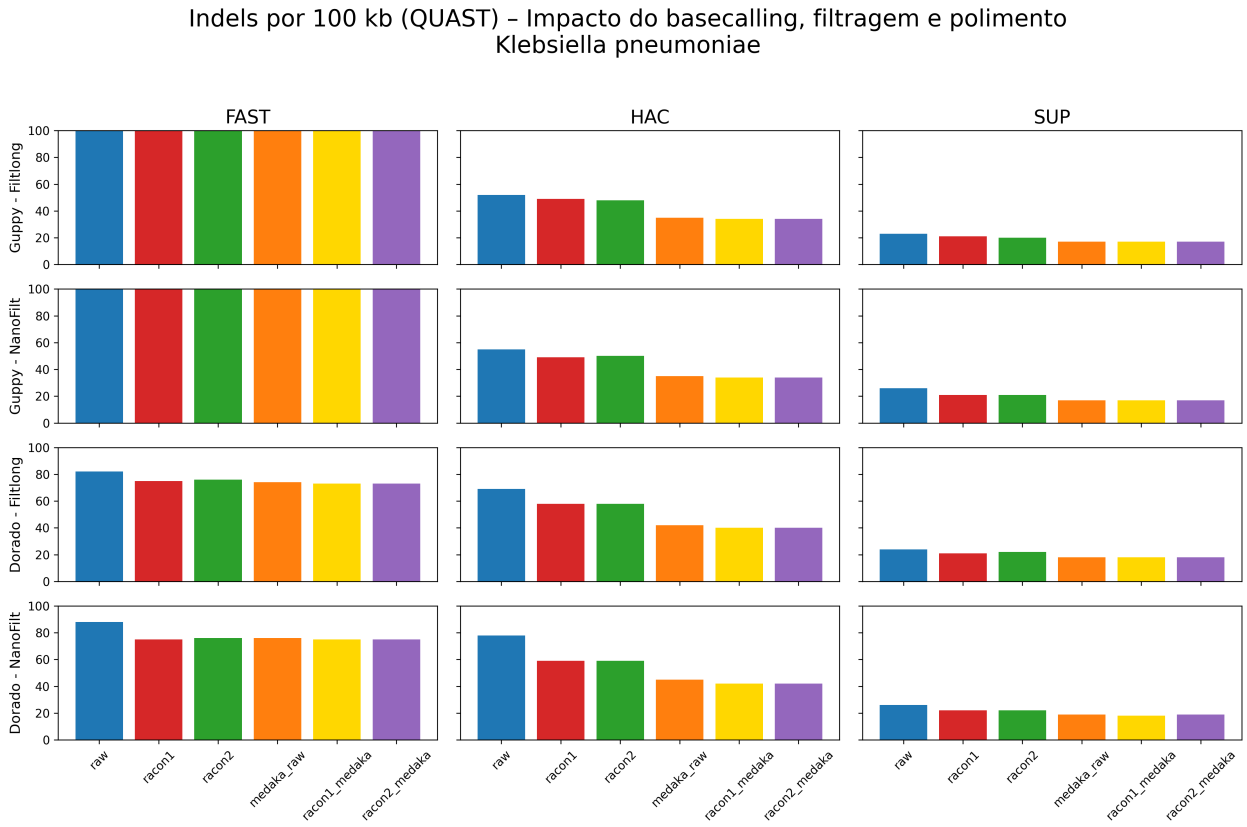

Figura 23: Taxa de *indels* por 100 kb (QUAST) em montagens de *Klebsiella pneumoniae* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Contiguidade (N50)

  A análise de contiguidade por meio do ΔN50 (diferença em relação ao menor N50 observado), obserava na Figura 24 indica que as maiores contiguidades também se concentram nas combinações que utilizam o modelo *sup*. As diferenças entre estratégias de filtragem (NanoFilt e Filtlong) mostraram impacto discreto sobre a contiguidade final, enquanto a escolha do modelo de *basecalling* e a presença de polimento com Medaka exerceram influência mais evidente. As combinações com maior ΔN50 correspondem às montagens mais contínuas e coincidem, em geral, com aquelas que apresentam maiores valores de BUSCO e menores taxas de *indels*, indicando convergência entre métricas estruturais e de precisão.

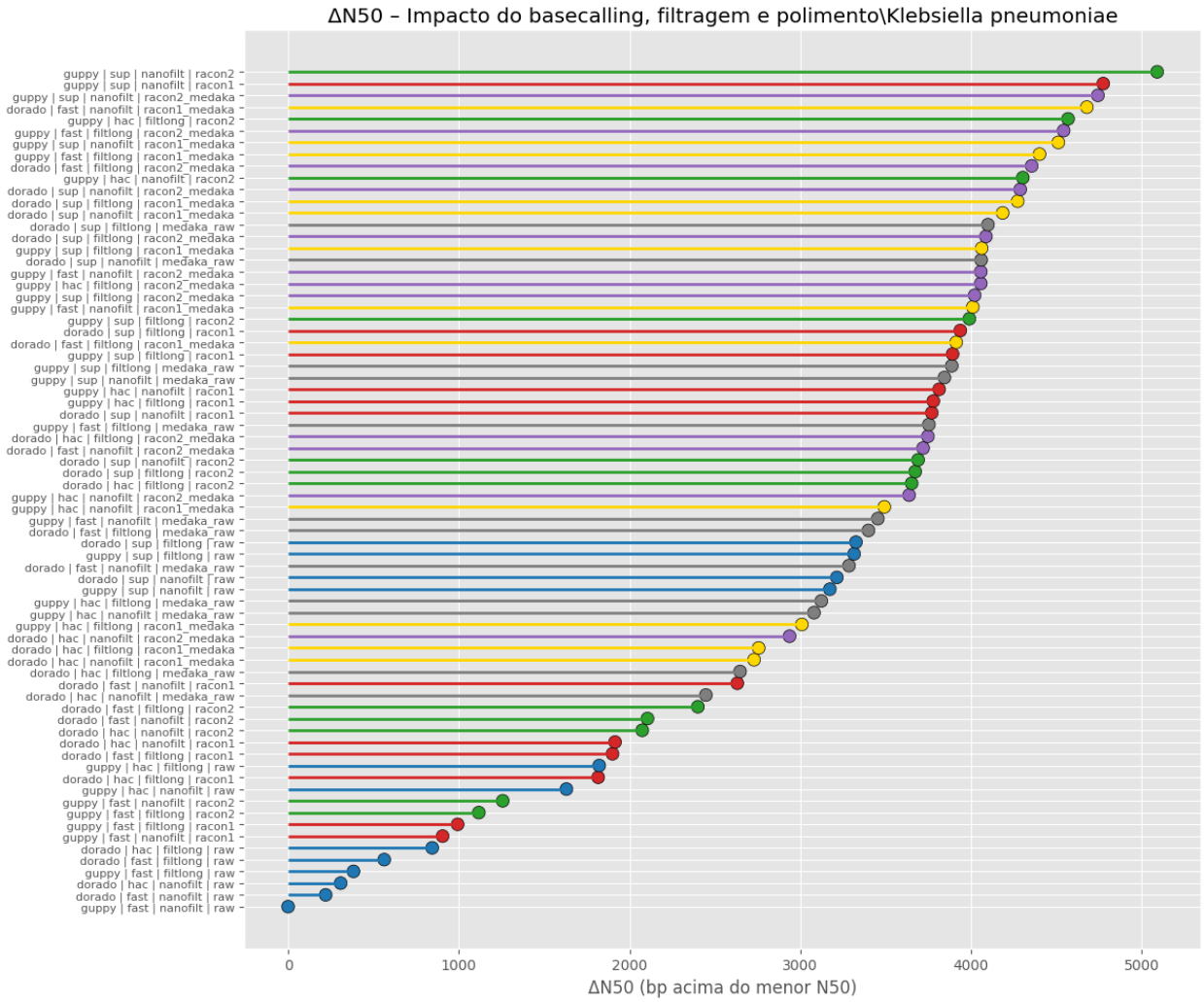

Figura 24: Variação da contiguidade das montagens (N50) para *Klebsiella pneumoniae* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Síntese dos melhores fluxos

  A Tabela 16 sintetiza as cinco combinações com melhor desempenho global. Diferentemente do observado para os organismos anteriores, todas as configurações do top 5 utilizaram o *basecaller* Guppy associado ao modelo *sup*, indicando maior consistência desse conjunto específico para *K. pneumoniae*. As variações entre as combinações concentram-se principalmente nas estratégias de polimento, com destaque para o uso combinado de Racon e Medaka. A melhor configuração global foi obtida com a combinação Guppy–*sup*–NanoFilt–2×Racon+Medaka, que apresentou a maior completude (97,4 %), elevada contiguidade (N50 superior a 5,13 Mb) e uma das menores taxas de *indels* observadas.

Tabela 16: Top 5 combinações ranqueadas para *K. pneumoniae* com base no escore global.

Rank	Basecaller	Modelo	Filtro	Polimento	BUSCO (%)	N50 (bp)	Indels/100kb
1	Guppy	sup	NanoFilt	2xRacon+Medaka	97.4	5,134,685	16,75
2	Guppy	sup	NanoFilt	1xRacon+Medaka	97.4	5,134,453	16,85
3	Guppy	sup	Filtlong	1xRacon+Medaka	96.6	5,134,004	16,79
4	Guppy	sup	NanoFilt	2xRacon	95.7	5,135,032	20,93
5	Guppy	sup	Filtlong	2xRacon+Medaka	96.6	5,133,963	16,71

As diferenças entre as cinco melhores combinações foram relativamente discretas, indicando que, uma vez adotado o modelo *sup*, o impacto adicional das estratégias de filtragem e do número de rodadas de polimento torna-se marginal. Ainda assim, a presença do Medaka esteve associada às menores taxas de erro, enquanto rodadas adicionais de Racon contribuíram principalmente para ajustes finos na precisão local. Em conjunto, os resultados indicam que, para *K. pneumoniae*, a acurácia inicial do *basecalling* e a aplicação de polimento com Medaka constituem os fatores mais determinantes para a qualidade final das montagens, enquanto a escolha do filtro exerce influência secundária.

4.2.2.4 Shigella sonnei

A avaliação das montagens para *S. sonnei* indicou padrões consistentes entre as métricas de completude, continuidade e acurácia, com forte influência do modelo de *basecalling* e do polimento final.

- Completude Genômica (BUSCO)

  Os resultados de completude (Figura 25) mostraram clara superioridade dos modelos *sup* em relação aos modelos *fast* e *hac*, independentemente do *basecaller*. As combinações com Dorado e Guppy em modelo *sup* apresentaram valores próximos ao máximo observado para o conjunto de dados, com diferenças discretas entre estratégias de filtragem. O uso de NanoFilt ou Filtlong não produziu variações expressivas nos percentuais de BUSCO quando o polimento final incluiu Medaka. Em contrapartida, montagens sem polimento final ou com apenas racon apresentaram reduções perceptíveis na completude, especialmente nos modelos *fast* e *hac*. De forma geral, o polimento combinado (Racon seguido de Medaka) foi determinante para atingir os maiores valores de BUSCO.

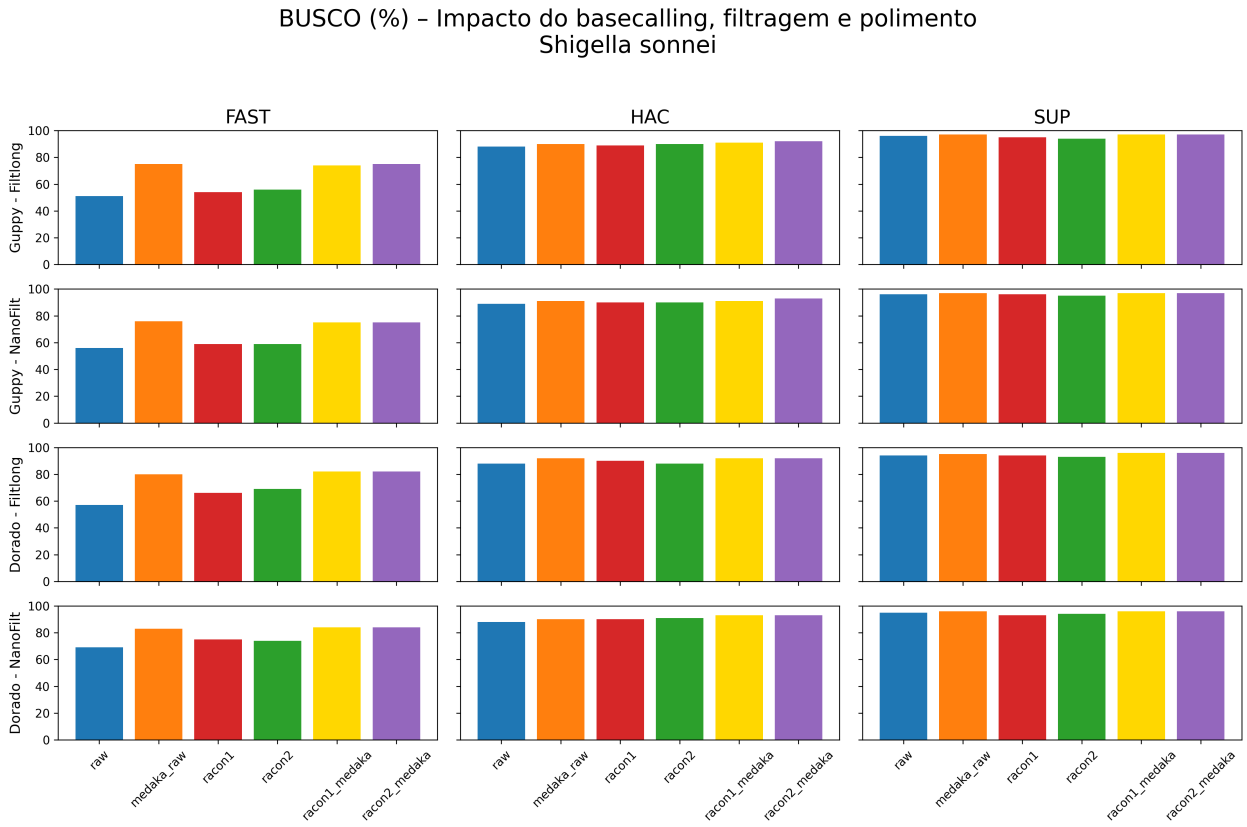

Figura 25: Avaliação da completude genômica das montagens de *Shigella sonnei* utilizando BUSCO, considerando diferentes combinações de ferramentas de *basecalling* (Guppy e Dorado), modelos de acurácia (FAST, HAC e SUP), filtros de melhoria das reads (Nanofilt e Filtlong) e estratégias de polimento (sem polimento, Racon, Medaka e combinações)

- Taxa de *indels* (QUAST)

  A taxa de *indels* (Figura 26) apresentou comportamento inverso ao de completude: os menores valores foram obtidos com modelos *sup* e polimento final com Medaka. O uso isolado de racon reduziu parcialmente os erros, porém a adição de Medaka resultou nas menores taxas observadas. As diferenças entre NanoFilt e Filtlong foram discretas, sugerindo que o impacto da filtragem foi secundário em comparação ao modelo de *basecalling* e ao polimento. Nos modelos *fast*, as taxas de *indels* permaneceram elevadas mesmo após polimento, evidenciando limitação na acurácia inicial das leituras. Nos modelos *hac*, a redução de erros foi intermediária, enquanto o modelo *sup* apresentou os melhores resultados de forma consistente.

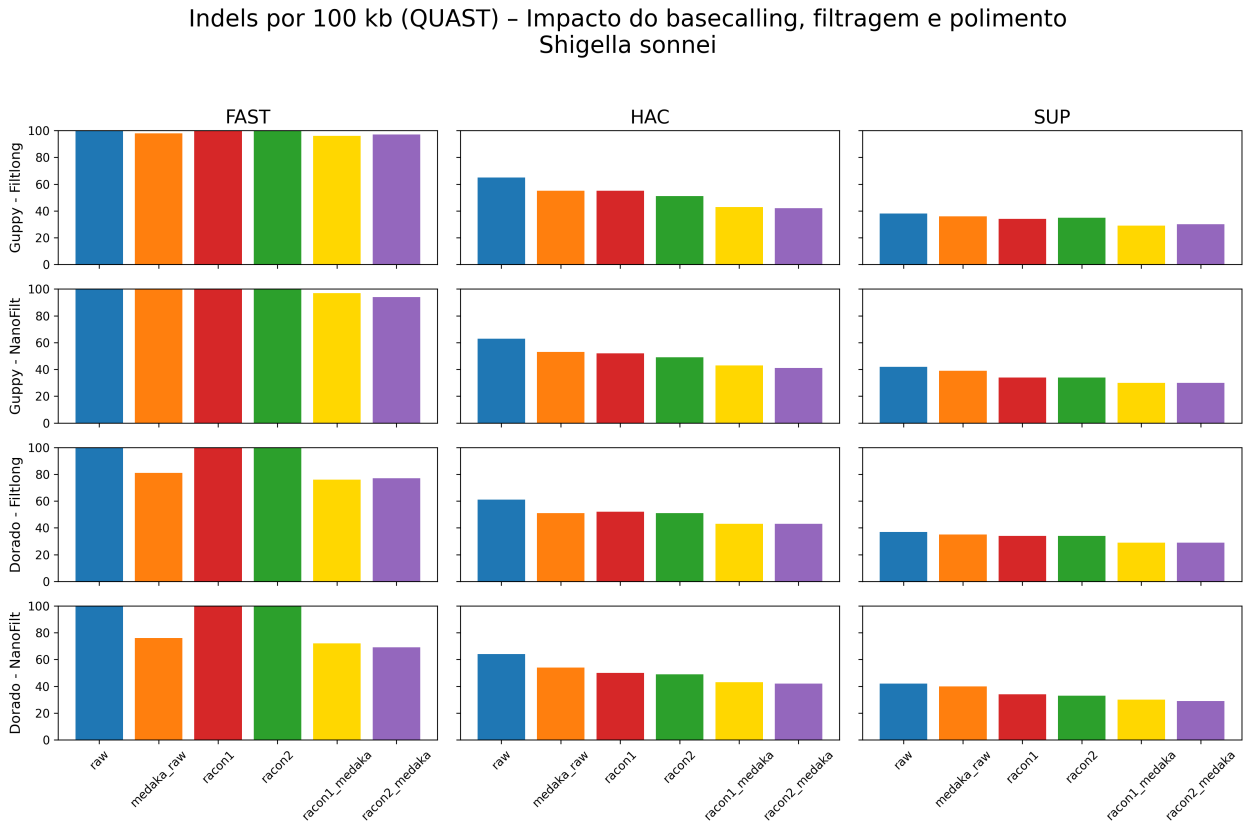

Figura 26: Taxa de *indels* por 100 kb (QUAST) em montagens de *Shigella sonnei* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Contiguidade (N50)

  A análise de contiguidade indicou variações relativamente pequenas entre as combinações, com valores de N50 bastante próximos entre si quando comparadas apenas as melhores montagens, como é observado na Figura 27 . O gráfico de ΔN50 mostrou que as diferenças absolutas entre as combinações de maior desempenho foram modestas, indicando que a contiguidade genômica foi menos sensível às etapas de filtragem e polimento do que as métricas de acurácia. Ainda assim, as maiores diferenças em relação ao pior cenário ocorreram nos modelos *sup*, especialmente quando combinados a polimento com racon seguido de Medaka. Nos modelos *fast*, a variabilidade foi maior e associada principalmente ao tipo de filtragem e à ausência de polimento final.

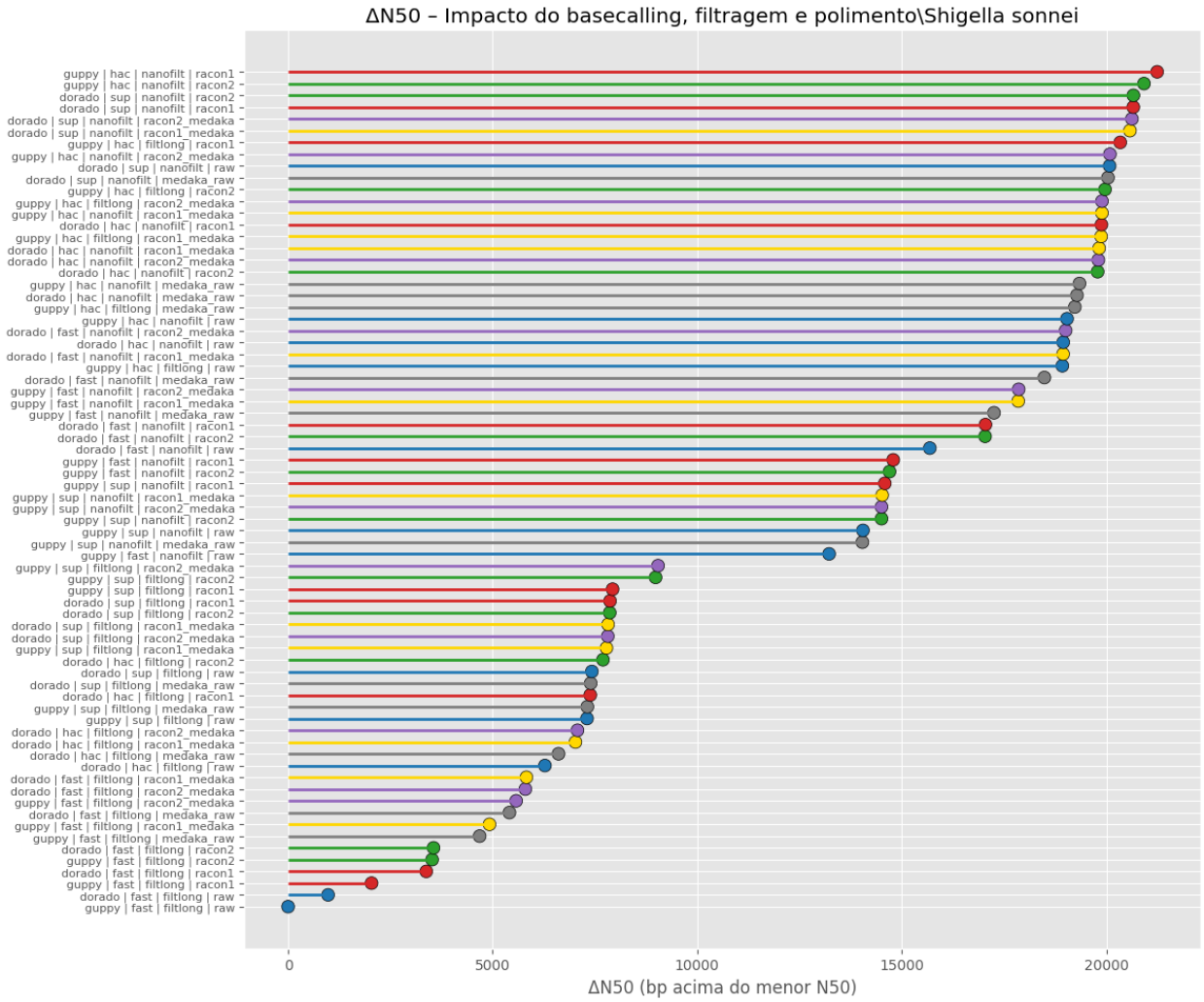

Figura 27: Variação da contiguidade das montagens (N50) para *Shigella sonnei* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Síntese dos melhores fluxos

  De modo geral, os resultados para *S. sonnei* indicam que o principal determinante de qualidade das montagens foi o modelo de *basecalling*, seguido pelo polimento final com Medaka. A filtragem teve impacto secundário e mais evidente apenas nos cenários com menor qualidade inicial de leitura. O top 5 das dos fluxos de *S. sonnei* podem ser observados na tabela 17

Tabela 17: Top 5 combinações ranqueadas para *S. sonnei* com base no escore global.

Rank	Basecaller	Modelo	Filtro	Polimento	BUSCO (%)	N50 (bp)	Indels/100kb
1	Dorado	sup	NanoFilt	2xRacon+Medaka	95.5	4,777,370	29,21
2	Dorado	sup	NanoFilt	1xRacon+Medaka	95.7	4,777,322	30,30
3	Dorado	sup	NanoFilt	2xRacon	93.0	4,777,407	32,63
4	Guppy	sup	NanoFilt	1xRacon+Medaka	97.7	4,771,267	29,43
5	Guppy	sup	NanoFilt	2xRacon+Medaka	97.0	4,771,251	29,21

Entre as combinações ranqueadas, observa-se predominância do modelo *sup* associado ao uso de NanoFilt e polimento final com Medaka. As diferenças de N50 entre as cinco melhores estratégias foram mínimas, reforçando que a seleção do melhor pipeline para *S. sonnei* deve priorizar métricas de acurácia, especialmente taxa de *indels* e completude BUSCO. A presença recorrente de estratégias com 1xRacon+Medaka ou 2xRacon+Medaka no topo do ranking indica que o polimento híbrido foi essencial para otimizar simultaneamente continuidade e acurácia.

4.2.2.5 Staphylococcus aureus

- Completude Genômica (BUSCO)

  Os resultados de completude genômica (Figura 28) apresentaram valores elevados em praticamente todas as combinações testadas, especialmente para os modelos *hac* e *sup*, que atingiram valores próximos de 100% independentemente do filtro ou do polimento aplicado. No modelo *fast*, observou-se maior variação, com completude inferior e maior sensibilidade às etapas de polimento.

  De modo geral, o impacto do *basecaller* foi discreto: Guppy e Dorado exibiram desempenhos muito semelhantes, com leve vantagem para os modelos de maior acurácia. O polimento contribuiu para pequenos ganhos em completude, sobretudo quando Medaka foi aplicado após Racon, mas essas diferenças foram modestas. Isso indica que, para *S. aureus*, a montagem já atinge alta completude em estágios iniciais do fluxo, e o ganho adicional das etapas de refinamento é limitado.

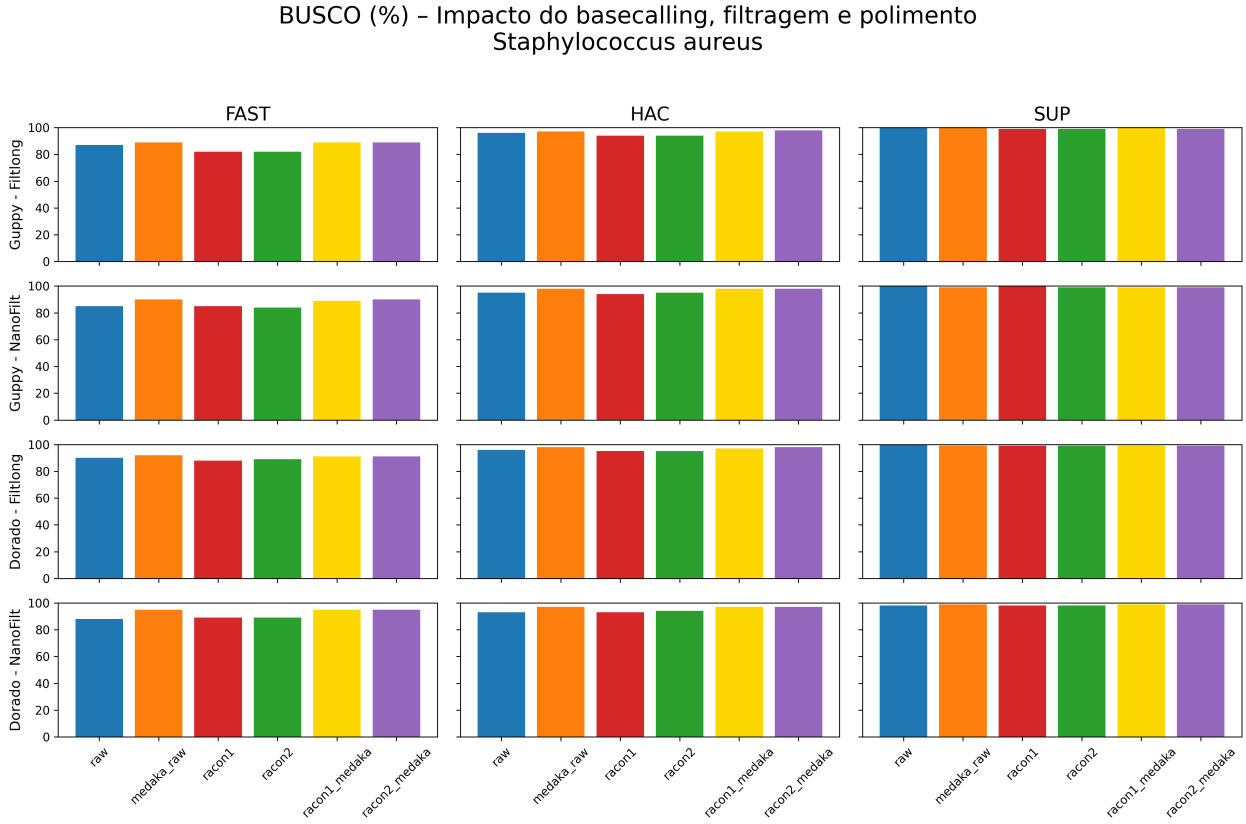

Figura 28: Avaliação da completude genômica das montagens de *Staphylococcus aureus* utilizando BUSCO, considerando diferentes combinações de ferramentas de *basecalling* (Guppy e Dorado), modelos de acurácia (FAST, HAC e SUP), filtros de melhoria das reads (Nanofilt e Filtlong) e estratégias de polimento (sem polimento, Racon, Medaka e combinações)

- Taxa de *indels* (QUAST)

  A taxa de *indels*, representada na Figura 29, foi globalmente baixa em comparação com os demais organismos analisados, com valores concentrados em faixas reduzidas mesmo antes do polimento. A aplicação de Racon e Medaka promoveu redução consistente dos erros, principalmente nas configurações *fast* e *hac*, nas quais o impacto do *basecalling* inicial é mais evidente.

  Nos modelos *sup*, as taxas já eram mínimas desde as montagens iniciais, e o polimento resultou em melhorias marginais. Observou-se que o filtro Filtlong associado ao *basecalling sup* gerou as menores taxas de *indels*, sugerindo que a combinação entre alta acurácia de leitura e filtragem de qualidade reduz substancialmente a necessidade de correções posteriores.

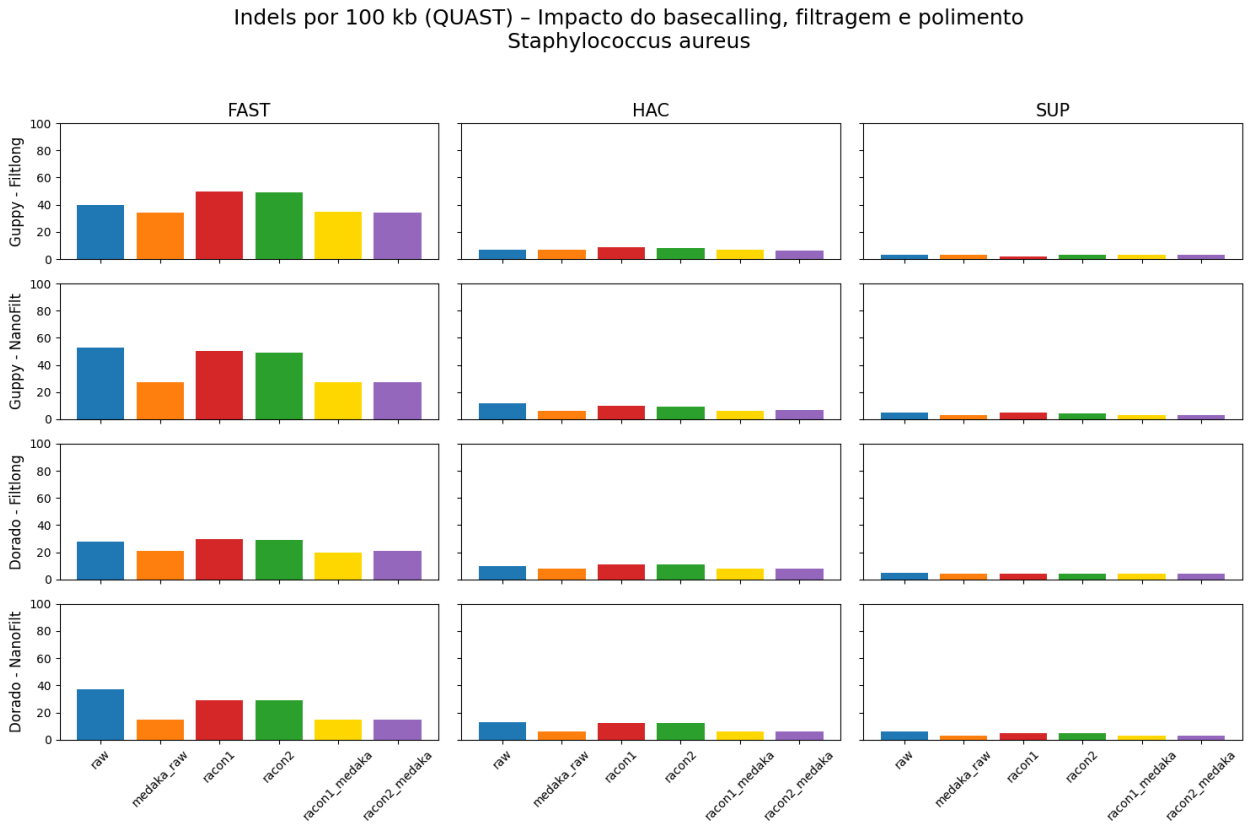

Figura 29: Taxa de *indels* por 100 kb (QUAST) em montagens de *Staphylococcus aureus* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Contiguidade (N50)

  A análise do N50 (Figura 30) evidenciou baixa variação entre as combinações, refletindo a elevada contiguidade obtida em praticamente todos os cenários. As diferenças entre as montagens ficaram restritas a pequenas variações em relação ao menor N50 observado, indicando que a estrutura global das montagens foi estável independentemente das etapas de polimento. Não houve tendência clara de aumento de N50 com o aumento de ciclos de Racon ou com a adição de Medaka, sugerindo que a contiguidade já é maximizada na etapa de montagem inicial. O filtro Filtlong contribuiu para leve melhoria em algumas configurações, mas o efeito foi discreto quando comparado ao observado em organismos com genomas mais fragmentados.

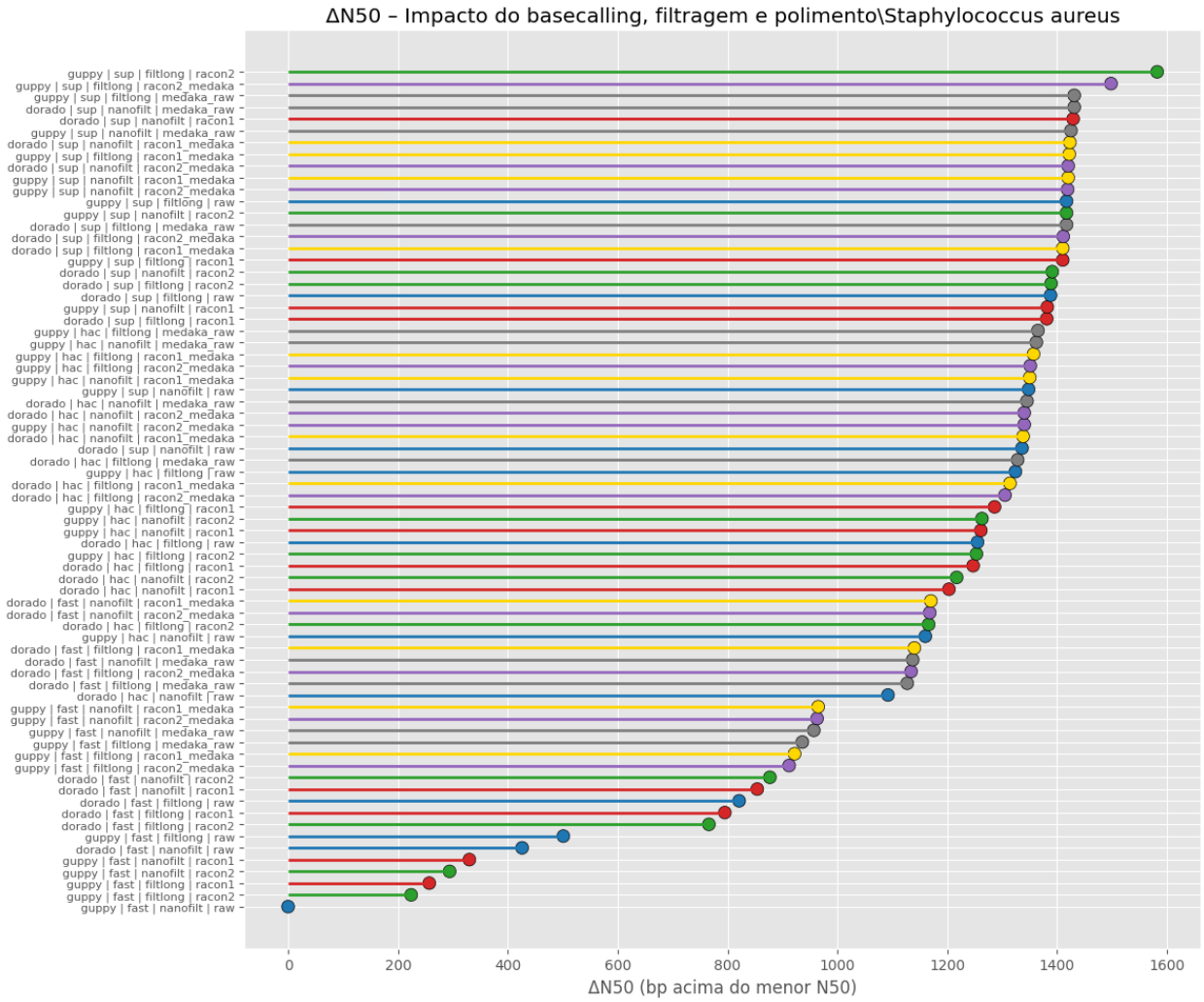

Figura 30: Variação da contiguidade das montagens (N50) para *Staphylococcus aureus* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Síntese dos melhores fluxos

  Para *S. aureus*, os três indicadores convergem para um cenário de alta robustez do pipeline. A completude genômica se mantém elevada em praticamente todas as combinações, a taxa de *indels* é baixa e pouco variável, e o N50 apresenta estabilidade estrutural, conforme indica a tabela 18

  As melhores combinações ranqueadas indicam que o uso de Guppy *sup* com Filtlong já produz montagens de alta qualidade mesmo sem polimento, e que a aplicação de Medaka ou ciclos adicionais de Racon resulta em melhorias pontuais, sem alterações substanciais na qualidade global. Assim, o organismo se mostra pouco sensível a variações no pipeline, permitindo maior flexibilidade na escolha das etapas sem comprometer significativamente os resultados finais.

Tabela 18: Top 5 combinações ranqueadas para *S. aureus* com base no escore global.

Rank	Basecaller	Modelo	Filtro	Polimento	BUSCO (%)	N50 (bp)	Indels/100kb
1	Guppy	sup	Filtlong	raw	99.6	2,902,052	3,48
2	Guppy	sup	Filtlong	Medaka	99.6	2,902,052	3,76
3	Guppy	sup	Filtlong	1xRacon+Medaka	99.3	2,902,052	3,55
4	Guppy	sup	Filtlong	1xRacon	99.1	2,902,052	3,27
5	Guppy	sup	Filtlong	2xRacon	98.9	2,902,052	3,48

4.2.2.6 Stenotrophomona maltophilia

A avaliação das 72 combinações testadas para *S. maltophilia* revelou um cenário de desempenho elevado e relativamente homogêneo entre os fluxos com modelos *hac* e *sup*, contrastando com a maior variabilidade observada nos modelos *fast*. A interpretação integrada das métricas de completude (BUSCO), precisão (indels) e contiguidade (N50) permite identificar tendências consistentes entre *basecalling*, filtragem e polimento.

- Completude Genômica (BUSCO)

  Nos gráficos de BUSCO (Figura 31), observa-se que as combinações com modelos *hac* e, principalmente, *sup* atingem valores de completude elevados e estáveis, geralmente acima de 95%, independentemente do *basecaller* utilizado. O modelo *fast* apresenta maior dispersão e valores inferiores de completude, sobretudo antes do polimento. A aplicação de etapas de polimento melhora de forma consistente os resultados, com destaque para Medaka isolado ou em combinação com Racon, que eleva a completude em praticamente todos os cenários. Em geral, as combinações com modelo *sup* e qualquer estratégia de polimento atingem os maiores valores de BUSCO, indicando montagem genômica mais completa.

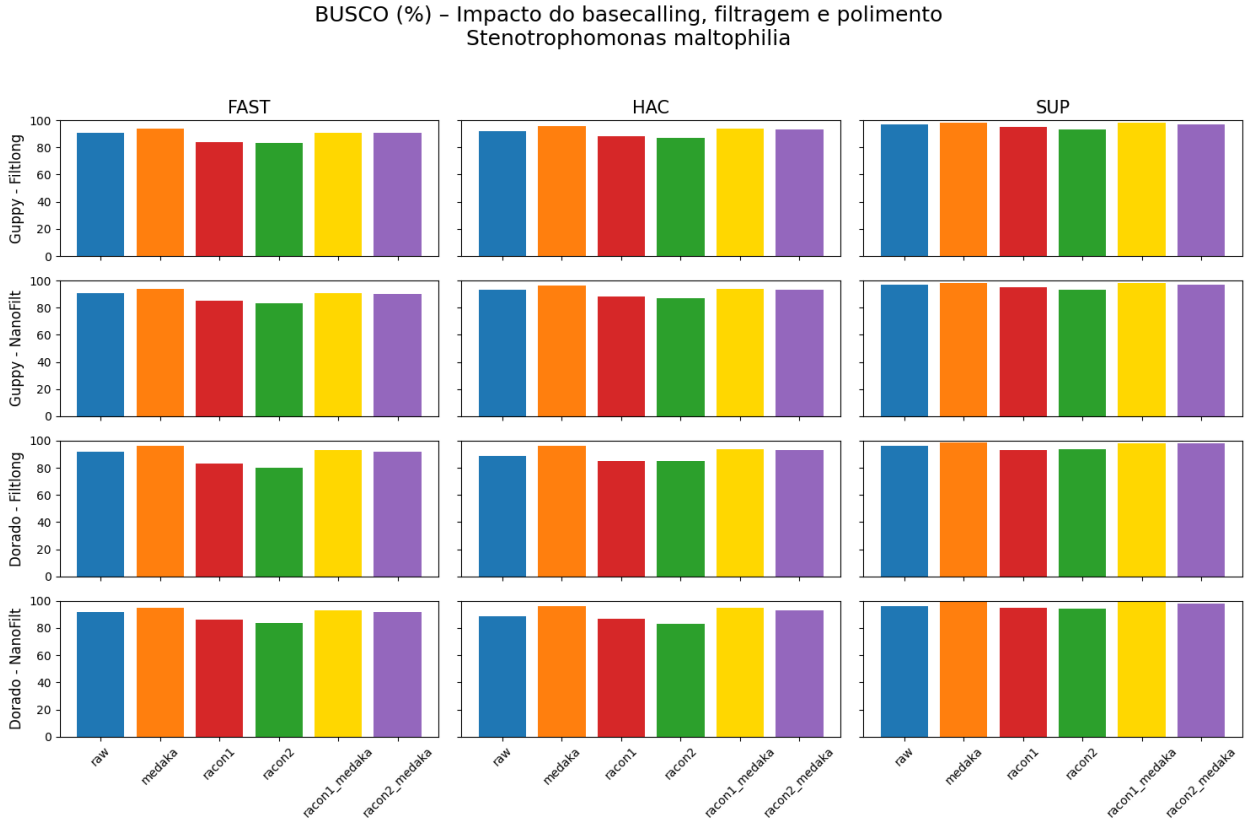

Figura 31: Avaliação da completude genômica das montagens de *Stenotrophomonas maltophilia* utilizando BUSCO, considerando diferentes combinações de ferramentas de *basecalling* (Guppy e Dorado), modelos de acurácia (FAST, HAC e SUP), filtros de melhoria das reads (Nanofilt e Filtlong) e estratégias de polimento (sem polimento, Racon, Medaka e combinações)

- Taxa de *indels* (N50)

  Em relação às taxas de *indels*, os gráficos (Figura 32) indicam uma redução clara após o polimento, principalmente com Medaka e com as combinações Racon + Medaka. As montagens sem polimento (raw) apresentam as maiores taxas de erro, enquanto a aplicação de duas rodadas de Racon seguida de Medaka tende a produzir as menores taxas. Esse padrão é consistente entre *basecallers* e filtros, sugerindo que o polimento tem impacto mais pronunciado na precisão de base do que a etapa de filtragem. O modelo *sup* novamente se destaca por apresentar menores taxas de *indels* após o polimento, refletindo a maior acurácia das leituras *basecalladas* nesse modo.

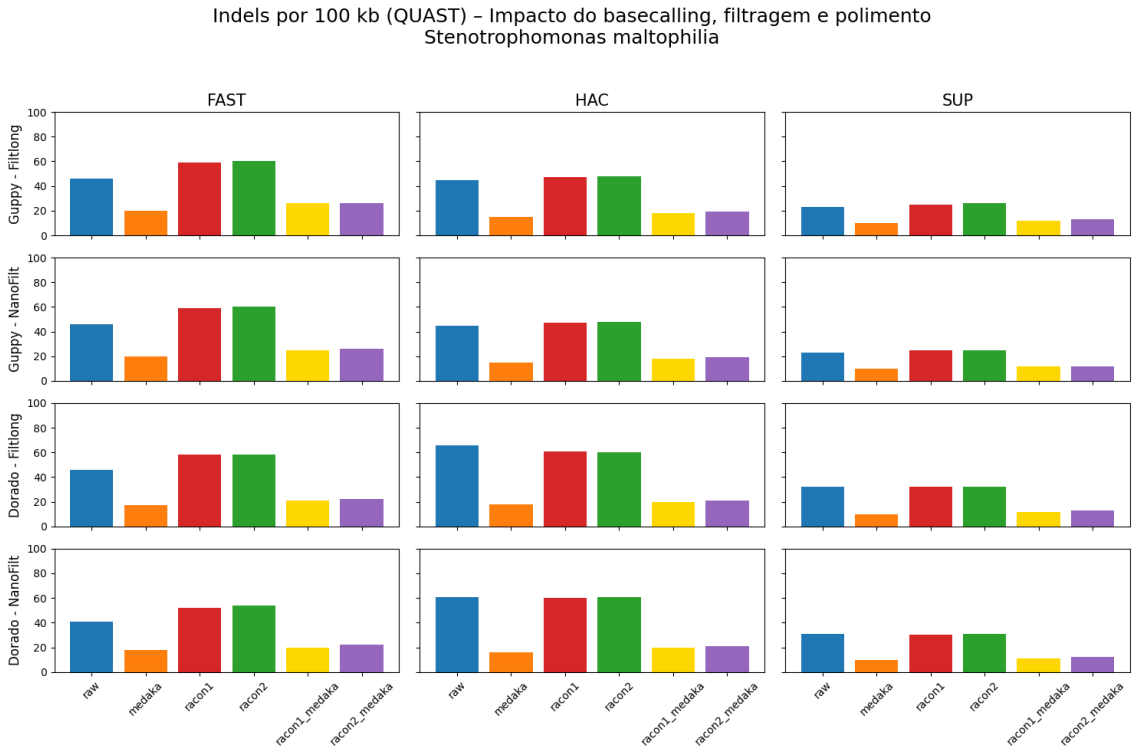

Figura 32: Taxa de *indels* por 100 kb (QUAST) em montagens de *Stenotrophomonas maltophilia* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Contiguidade (N50)

  Em *S. maltophilia*, o maior N50 foi observado em combinações com modelo *fast*, como é evidenciado na Figura 33 . No entanto, essa vantagem de contiguidade não se refletiu em melhor qualidade global da montagem, uma vez que modelos *hac* e *sup* apresentaram maior completude (BUSCO) e menores taxas de erro. Esse resultado sugere que, para esse organismo, o modelo *fast* preservou leituras longas suficientes para favorecer a conexão entre *contigs* durante a montagem, enquanto os modelos mais precisos impactaram principalmente a acurácia da sequência final.

  As combinações baseadas no modelo *fast* apresentam maior variabilidade de N50, incluindo tanto os menores quanto alguns dos maiores valores observados, enquanto os modelos *hac* e *sup* tendem a produzir resultados mais consistentes. Isso sugere que a qualidade do *basecalling* inicial influencia a contiguidade, mas o efeito não é linear e depende da interação com as etapas de filtragem e polimento.

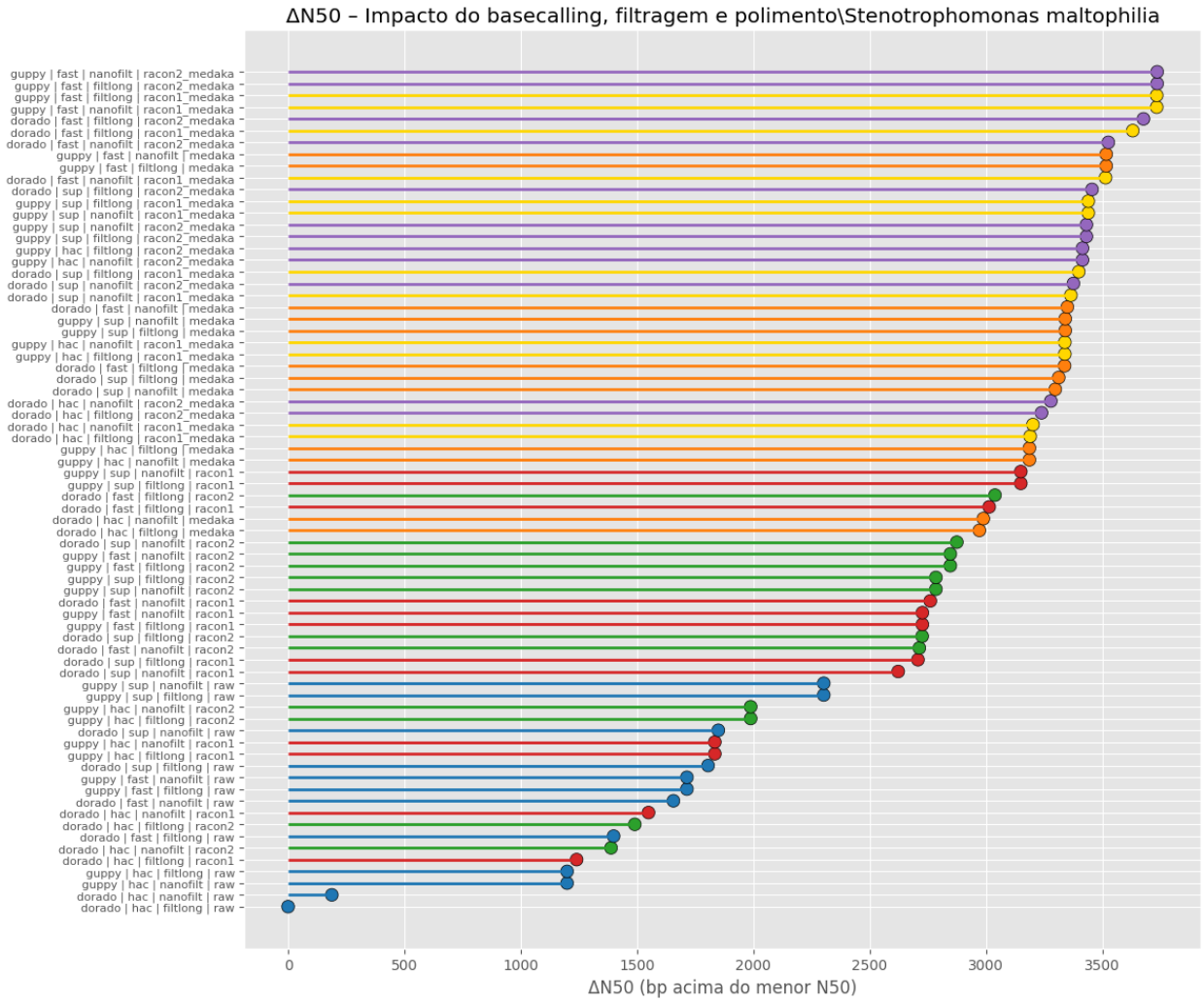

Figura 33: Variação da contiguidade das montagens (N50) para *Stenotrophomonas maltophilia* sob diferentes combinações de *basecalling*, filtragem e estratégias de polimento.

- Síntese dos melhores fluxos

  De modo geral, os três conjuntos de métricas indicam que as melhores combinações para *S. maltophilia* concentram-se em fluxos com modelo *sup*, seguidos de etapas de polimento envolvendo Medaka. A filtragem por NanoFilt ou Filtlong não produz diferenças substanciais entre os melhores resultados, sugerindo impacto secundário em relação ao *basecalling* e ao polimento.

  As cinco combinações mais bem ranqueadas (tabela 19 reforçam as tendências observadas nos gráficos. Todas utilizam o modelo *sup*, e quatro delas empregam Guppy como *basecaller*. O polimento com Medaka, isolado ou combinado com uma rodada de Racon, está presente em todas as melhores combinações, indicando seu papel central na melhoria da precisão e da completude. As diferenças entre NanoFilt e Filtlong são mínimas nas posições superiores, sugerindo que ambas as

estratégias de filtragem são adequadas para esse organismo.

Em conjunto, os resultados indicam que, para *S. maltophilia*, a escolha do modelo de *basecalling* e da estratégia de polimento exerce maior influência sobre o desempenho global do pipeline do que a escolha do filtro de leituras.

Tabela 19: Top 5 combinações ranqueadas para *S. maltophilia* com base no escore global.

Rank	Basecaller	Modelo	Filtro	Polimento	BUSCO (%)	N50 (bp)	Indels/100kb
1	Guppy	sup	NanoFilt	1xRacon+Medaka	98.3	4,803,236	11,60
2	Guppy	sup	Filtlong	1xRacon+Medaka	98.3	4,803,236	11,60
3	Guppy	sup	NanoFilt	Medaka	99.0	4,803,137	9,81
4	Guppy	sup	Filtlong	Medaka	99.0	4,803,137	9,81
5	Dorado	sup	NanoFilt	1xRacon+Medaka	99.0	4,803,161	11,20

4.3 Comparação dos melhores fluxos entre os organismos analisados

A comparação entre os melhores fluxos identificados para os seis organismos avaliados permitiu a identificação de padrões consistentes quanto à escolha das ferramentas, modelos de *basecalling*, estratégias de filtragem e abordagens de polimento, independentemente das diferenças genômicas específicas entre as espécies analisadas.

De modo geral, observou-se convergência clara dos fluxos de melhor desempenho em torno do uso de modelos de *basecalling* de maior acurácia (*sup*), tanto para Guppy quanto para Dorado. Em todos os organismos, as combinações que utilizaram o modelo *sup* figuraram entre as primeiras posições do ranqueamento global, apresentando maiores valores de completude genômica segundo BUSCO, maior contiguidade estrutural (N50 elevado) e menores taxas de erro em nível de base, especialmente *indels* por 100 kb. Esse padrão foi observado de forma consistente em *Acinetobacter pittii, Haemophilus haemolyticus, Klebsiella pneumoniae, Shigella sonnei, Staphylococcus aureus e Stenotrophomonas maltophilia*, indicando que a acurácia do *basecalling* constitui o principal fator determinante da qualidade final da montagem.

Quanto à etapa de filtragem, tanto NanoFilt quanto Filtlong foram capazes de gerar *datasets* adequados para montagem, com diferenças pontuais no impacto sobre métricas iniciais de qualidade. Entretanto, após a aplicação do polimento, as diferenças entre os métodos de filtragem tornaram-se menos evidentes, sugerindo que, em fluxos baseados em modelos *sup* e polimento adequado, a escolha do filtro exerce influência secundária sobre a qualidade final da montagem quando comparada à escolha do modelo de *basecalling* e da estratégia de refinamento.

A etapa de polimento mostrou-se essencial para o aprimoramento da precisão em nível de base em todos os organismos analisados. A aplicação de Medaka, isoladamente ou combinada com uma rodada prévia de Racon, promoveu redução substancial das taxas de *indels* e aumento consistente da completude genômica. Em diversos casos, como

observado para *A. pittii, H. haemolyticus e K. pneumoniae*, a combinação de uma rodada de Racon seguida de Medaka apresentou o melhor equilíbrio entre melhoria das métricas e simplicidade do fluxo, enquanto rodadas adicionais de Racon resultaram em ganhos marginais, sugerindo efeito de saturação no refinamento da montagem.

A análise integrada também evidenciou que, embora modelos *fast* e *hac* apresentem menor custo computacional, os ganhos proporcionados pelo uso do modelo *sup* foram sistemáticos e biologicamente relevantes, especialmente no que se refere à redução de erros estruturais e à recuperação de genes completos. Esse comportamento foi particularmente evidente em organismos com tamanho de dados de sequenciamento maiores, como *K. pneumoniae e S. sonnei*, nos quais diferenças iniciais de acurácia do *basecalling* refletiram-se nas métricas da montagem bruta e nas necessidades de polimento subsequentes.

Em conjunto, os resultados demonstram que, apesar da variabilidade genômica entre os organismos avaliados, os melhores fluxos apresentaram alto grau de similaridade, caracterizando um padrão robusto e reproduzível. Essa convergência reforça a aplicabilidade geral do pipeline automatizado proposto e fornece evidências sólidas para a recomendação de fluxos otimizados baseados em modelos de *basecalling* de alta acurácia, associados a estratégias de polimento eficientes, como abordagem padrão para montagem de genomas bacterianos a partir de dados de sequenciamento *Nanopore*.

A comparação dos melhores fluxos por organismo expressos na Tabela 20 evidencia que, embora todos tenham alcançado elevados níveis de completude genômica e contiguidade estrutural, observou-se variabilidade no desempenho relativo dos pipelines e no impacto das etapas de polimento entre os diferentes organismos analisados. De modo geral, os modelos de *basecalling* do tipo *sup* mostraram-se consistentes entre os organismos, resultando em altos valores de BUSCO e N50 próximos ao tamanho esperado dos genomas. Entretanto, o impacto do polimento foi heterogêneo. Organismos como *Acinetobacter pittii, Klebsiella pneumoniae, Shigella sonnei e Stenotrophomonas maltophilia* apresentaram redução substancial nas taxas de *indels* apenas após a aplicação de ciclos de Racon combinados com Medaka, indicando dependência do polimento para obtenção de maior acurácia nucleotídica. Em contraste, *Staphylococcus aureus* destacou-se por atingir simultaneamente o maior valor de BUSCO (99,6%), a menor taxa de *indels* por 100 kb (3,48) e alta contiguidade estrutural já na montagem bruta, sem necessidade de polimento adicional. Esse desempenho sugere que, para esse organismo, a elevada qualidade intrínseca das leituras geradas pelo modelo *sup*, aliada ao uso de Filtlong, foi suficiente para produzir uma montagem de alta qualidade. Assim, considerando conjuntamente completude genômica, precisão em nível de base, simplicidade do fluxo e menor custo computacional associado à exclusão do polimento, o pipeline aplicado a *Staphylococcus aureus* pode ser definido como o melhor fluxo geral entre os organismos avaliados neste estudo.

Organismo	Basecaller	Modelo	Filtragem	Polimento	BUSCO (%)	Indels / 100 kb	N50 (bp)	Misassemblies
Acinetobacter pittii	Dorado	sup	**NanoFilt**	Medaka	98.5	8,94	3,815,297	0
Haemophilus haemolyticus	Dorado	sup	**Filtlong**	1× Racon + Medaka	93.5	30,36	2,051,276	1
Klebsiella pneumoniae	Guppy	sup	**NanoFilt**	2× Racon + Medaka	97.4	16,75	5,134,685	1
Shigella sonnei	Dorado	sup	**NanoFilt**	2× Racon + Medaka	95.5	29.21	4,777,370	1
Staphylococcus aureus	Guppy	sup	**Filtlong**	Raw (sem polimento)	99.6	3.48	2,902,052	1
Stenotrophomonas maltophilia	Guppy	sup	**NanoFilt**	1× Racon + Medaka	98.3	11.60	4,803,236	0

Tabela 20: Comparação dos melhores fluxos de montagem genômica por organismo.

4.4 Qualidade versus custo computacional nos fluxos de montagem

A avaliação integrada da qualidade das montagens e do custo computacional dos fluxos permitiu analisar de forma crítica o equilíbrio entre acurácia e tempo de processamento no sequenciamento por *Nanopore*. Embora métricas como BUSCO, N50 e taxas de *indels* sejam fundamentais para a escolha do melhor pipeline, o tempo de *basecalling* representa um fator limitante importante, especialmente em cenários com grande volume de dados ou infraestrutura computacional restrita.

4.4.1 Tempo de *basecalling* e impacto no pipeline

O tempo de *basecalling* associado aos diferentes modelos está apresentado na 4.2.1, que sintetizam o custo computacional observado para os modelos *fast*, *hac* e *sup* dos *basecallers* avaliados. Os resultados mostram que os modelos *sup* demandaram consistentemente mais tempo de execução em comparação aos modelos *fast* e *hac*, independentemente do organismo analisado. Esse comportamento reflete a maior complexidade dos modelos *sup*, que utilizam arquiteturas de redes neurais mais profundas e, consequentemente, maior carga computacional.

Apesar dessa diferença, observa-se que o aumento no tempo de *basecalling* está fortemente associado ao modelo escolhido, e não às características específicas dos genomas analisados, uma vez que o padrão de crescimento do tempo foi semelhante entre organismos com diferentes tamanhos genômicos.

4.4.2 Relação entre qualidade final e custo computacional

A comparação entre os melhores fluxos de cada organismo evidencia um *trade-off* claro entre qualidade da montagem e custo computacional. Os fluxos baseados em modelos *sup* apresentaram, de forma consistente, os maiores valores de completude genômica (BUSCO) e as menores taxas de *indels*, resultando em montagens mais precisas em nível de base. Em contrapartida, esses fluxos apresentaram os maiores tempos de *basecalling*.

Por outro lado, os modelos *fast* e *hac* mostraram-se computacionalmente mais eficientes, com tempos de execução significativamente menores, porém produziram montagens com maior taxa de erros, especialmente antes do polimento. Embora etapas adicionais de polimento tenham reduzido essas diferenças, os fluxos baseados em modelos de menor acurácia raramente atingiram o mesmo nível de qualidade observado nos fluxos

sup, mesmo após múltiplas rodadas de refinamento.

Esses resultados indicam que a economia de tempo obtida na etapa de *basecalling* com modelos mais rápidos pode ser parcialmente compensada pelo aumento do esforço computacional nas etapas subsequentes de polimento e validação. Assim, quando considerado o pipeline como um todo, o uso de modelos *sup* tende a concentrar o custo computacional na etapa inicial, enquanto reduz a necessidade de correções extensivas posteriores.

A análise conjunta da qualidade e do custo computacional sugere que não existe um único fluxo universalmente ótimo para todos os cenários. Para aplicações que exigem alta precisão genômica, como estudos comparativos, análise de variantes ou investigação de resistência antimicrobiana, os fluxos baseados em modelos *sup* se mostram mais adequados, mesmo diante do maior tempo de *basecalling*. Em contrapartida, para análises exploratórias ou triagens rápidas, os modelos *hac* ou *fast* podem representar uma alternativa viável, desde que se aceite uma redução na qualidade final da montagem.

Em resumo, a integração das métricas de qualidade com o tempo de *basecalling* fornece uma visão mais realista do desempenho dos *pipelines* avaliados, permitindo que a escolha do fluxo seja orientada não apenas pela acurácia, mas também pelas limitações computacionais e pelos objetivos específicos de cada estudo.

5 CONCLUSÃO

Este estudo teve como objetivo realizar uma comparação de *pipelines* computacionais para a montagem de genomas de procariotos a partir de dados de sequenciamento por tecnologia *Nanopore* com a flow cell R.9.4, avaliando o impacto das escolhas metodológicas ao longo das etapas de *basecalling*, filtragem, montagem e polimento. A partir da análise integrada de seis organismos bacterianos, foi possível identificar padrões consistentes que permitiram tanto a avaliação individual de cada fluxo quanto a proposição de um fluxo geral mais robusto.

De forma predominante entre os organismos analisados, a escolha do modelo de *basecalling* mostrou-se o principal determinante da qualidade final das montagens. Os modelos do tipo *sup*, independentemente do *basecaller* utilizado, apresentaram melhor desempenho em métricas de completude genômica (BUSCO), menor taxa de erros por base, especialmente *indels* por 100 kb, e maior estabilidade estrutural das montagens, quando comparados aos modelos *fast* e *hac*. Esses resultados indicam que ganhos iniciais de acurácia no *basecalling* se refletem diretamente na qualidade da montagem bruta e reduzem a dependência de etapas extensivas de polimento.

As etapas de polimento exerceram papel relevante, sobretudo na redução de erros do tipo *indel*, com efeitos mais pronunciados em montagens derivadas de modelos de *basecalling* menos acurados. Entretanto, para organismos como *Staphylococcus aureus*, o elevado desempenho observado já na montagem bruta com modelo *sup* evidenciou que, em determinados contextos, a aplicação de polimento adicional pode resultar em ganhos marginais, não necessariamente justificando o aumento do custo computacional.

A comparação entre os melhores fluxos de cada organismo revelou que, embora variações específicas estejam associadas às características genômicas individuais, emergiu um padrão geral favorável ao uso de modelos *sup*, combinados com filtragem por NanoFilt ou Filtlong e, quando necessário, uma etapa moderada de polimento com Racon e Medaka. Entre os organismos avaliados, *Staphylococcus aureus* apresentou o fluxo mais eficiente de forma geral, atribuindo elevada completude genômica, baixa taxa de erros e menor complexidade computacional, configurando-se como um caso ótimo de equilíbrio entre qualidade e custo.

De maneira geral, os resultados deste trabalho reforçam que a otimização de pipelines para dados *Nanopore* deve priorizar decisões iniciais, especialmente na etapa de *basecalling*, uma vez que estas exercem influência estrutural e funcional sobre todo o processamento subsequente. As conclusões obtidas fornecem resultados práticos para a escolha de fluxos computacionais em projetos de genômica bacteriana, contribuindo para o uso mais eficiente de recursos computacionais sem comprometimento da qualidade das montagens.

Por fim, este estudo estabelece uma base comparativa sólida que pode ser expandida em trabalhos futuros, incluindo a avaliação de novas químicas de sequenciamento, arquiteturas de *hardware* distintas e a aplicação dos fluxos propostos a genomas mais complexos, ampliando o entendimento sobre o desempenho de *pipelines* computacionais em sequenciamento por *Nanopore*.

6 COMUNICAÇÃO CIENTÍFICA

O sequenciamento genômico tem se tornado uma ferramenta cada vez mais importante para a identificação e monitoramento de bactérias associadas a infecções humanas, especialmente diante do aumento global da resistência aos antibióticos. Entre as tecnologias mais recentes, o sequenciamento por tecnologias *nanopore* se destaca por permitir análises rápidas e portáteis. Entretanto, a qualidade final dos genomas obtidos depende diretamente das ferramentas computacionais utilizadas para processar os dados gerados durante o sequenciamento.

Neste estudo, foram comparados diferentes pipelines bioinformáticos para montagem de genomas bacterianos utilizando dados de sequenciamento utilizando tecnologia *nanopore*. A pesquisa avaliou combinações de ferramentas de *basecalling*, filtragem, controle de qualidade e polimento aplicadas a diferentes espécies bacterianas de relevância clínica, incluindo *Acinetobacter pittii, Klebsiella pneumoniae, Shigella sonnei, Staphylococcus aureus, Stenotrophomonas maltophilia* e *Haemophilus haemolyticus*. O objetivo foi identificar quais estratégias produzem montagens genômicas mais completas, contínuas e com menor quantidade de erros.

Os resultados demonstraram que os modelos de *basecalling* do tipo *sup* apresentaram desempenho consistentemente superior, produzindo genomas com maior completude e menor taxa de erros. Também foi observado que o impacto das etapas de polimento variou entre os organismos analisados. Em algumas bactérias, o uso combinado das ferramentas Racon e Medaka foi essencial para reduzir erros de montagem, enquanto em *Staphylococcus aureus* foram obtidos excelentes resultados mesmo sem etapas adicionais de polimento, reduzindo o custo computacional e simplificando o fluxo analítico.

Os achados deste trabalho contribuem para a otimização de pipelines bioinformáticos aplicados ao sequenciamento com a tecnologia *nanopore*, auxiliando pesquisadores e laboratórios na escolha de estratégias mais eficientes para análise genômica bacteriana. Além disso, os resultados podem favorecer aplicações em microbiologia clínica, vigilância epidemiológica e monitoramento da resistência antimicrobiana, fortalecendo o uso da genômica em saúde pública.

REFERÊNCIAS

[1] 4, U. D. J. G. I. H. T. . B. E. . P. P. . R. P. . W. S. . S. T. . D. N. . C. J.-F. . O. A. . L. S. . E. C. . U. E. . F. M., 9, R. G. S. C. S. Y. . F. A. . H. M. . Y. T. . T. A. . I. T. . K. C. . W. H. . T. Y. . T. T., Genoscope, 10, C. U.-. W. J. . H. R. . S. W. . A. F. . B. P. . B. T. . P. E. . R. C. . W. P., Department of Genome Analysis, I. o. M. B. R. A. . P. M. . N. G. . T. S. . R. A. ., 11, G. S. C. S. D. R. . D.-S. L. . R. M. . W. K. . L. H. M. . D. J., 15, B. G. I. G. C. Y. H. . Y. J. . W. J. . H. G. . G. J., et al. (2001). Initial sequencing and analysis of the human genome. *nature*, 409(6822):860–921.

[2] Abdel-Glil, M. Y., Brandt, C., Pletz, M. W., Neubauer, H., and Sprague, L. D. (2025). High intra-laboratory reproducibility of nanopore sequencing in bacterial species underscores advances in its accuracy. *Microbial Genomics*, 11(3):001372.

[3] Aleksandrzak-Piekarczyk, T., Grzesiak, J., Gawor, J., and Kosiorek, K. (2025). Genome sequences of polar carnobacterium maltaromaticum strains 2857 and 2862 with genes for glycerol and 1, 2-propanediol pathways. *Microbiology Resource Announcements*, 14(6):e00283–25.

[4] Alioto, T. S., Gut, M., Rodiño-Janeiro, B. K., Cruz, F., Gómez-Garrido, J., Vázquez-Ucha, J. C., Mata, C., Antoni, R., Briansó, F., Dabad, M., et al. (2023). Development of a novel streamlined workflow (aacre) and database (incredble) for genomic analysis of carbapenem-resistant enterobacterales. *Microbial Genomics*, 9(11):001132.

[5] Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., and Schatz, M. C. (2019). Ragoo: fast and accurate reference-guided scaffolding of draft genomes. *Genome biology*, 20:1–17.

[6] Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, 21(1):30.

[7] Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–465.

[8] Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., et al. (2019). Characterizing the major structural variant alleles of the human genome. *Cell*, 176(3):663–675.

[9] Bejaoui, S., Nielsen, S. H., Rasmussen, A., Coia, J. E., Andersen, D. T., Pedersen, T. B., Møller, M. V., Kusk Nielsen, M. T., Frees, D., and Persson, S. (2025). Comparison of illumina and oxford nanopore sequencing data quality for clostridioides difficile genome analysis and their application for epidemiological surveillance. *BMC genomics*, 26(1):92.

[10] Bentley, D. L. (2014). Coupling mrna processing with transcription in time and space. *Nature Reviews Genetics*, 15(3):163–175.

[11] Bentley, S. D. and Parkhill, J. (2004). Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.*, 38(1):771–791.

[12] Bogaerts, B., Maex, M., Commans, F., Goeders, N., Van den Bossche, A., De Keersmaecker, S. C., Roosens, N. H., Ceyssens, P.-J., Mattheus, W., and Vanneste, K. (2025). Oxford nanopore technologies r10 sequencing enables accurate cgmlst-based bacterial outbreak investigation of neisseria meningitidis and salmonella enterica when accounting for methylation-related errors. *Journal of Clinical Microbiology*, 63(10):e00410–25.

[13] Bonfield, J. K. and Mahoney, M. V. (2013). Compression of fastq and sam format sequencing data. *PloS one*, 8(3):e59190.

[14] Boondech, A., Ainmani, P., Khieokhajonkhet, A., Boonsrangsom, T., Pongcharoen, P., Rungrat, T., Sujipuli, K., Ratanasut, K., and Aeksiri, N. (2025). Complete genome and comparative analysis of xanthomonas oryzae pv. oryzae isolated from northern thailand. *Access Microbiology*, 7(6):000986–v4.

[15] Boostrom, I., Portal, E. A., Spiller, O. B., Walsh, T. R., and Sands, K. (2022). Comparing long-read assemblers to explore the potential of a sustainable low-cost, low-infrastructure approach to sequence antimicrobial resistant bacteria with oxford nanopore sequencing. *Frontiers in Microbiology*, 13:796465.

[16] Bowman, L., Sheridan, S., Wham, B. E., and Wright, S. (2023). Fasta/fastq data curation primer.

[17] Bož̌a, V., Perešíni, P., Brejová, B., and Vinař, T. (2020). Deepnano-blitz: a fast base caller for minion nanopore sequencers. *Bioinformatics*, 36(14):4191–4192.

[18] Boža, V., Perešíni, P., Brejová, B., and Vinař, T. (2021). Dynamic pooling improves nanopore base calling accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6):3416–3424.

[19] Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., et al. (2008). The potential and challenges of nanopore sequencing. *Nature biotechnology*, 26(10):1146–1153.

[20] Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., Naing, Z., Yeang, M., Verich, A., Gamaarachchi, H., et al. (2020). Analytical validity of nanopore sequencing for rapid sars-cov-2 genome analysis. *Nature communications*, 11(1):6272.

[21] Calvario, R. C. and Shanks, R. M. (2024). Genome sequence generated by hybrid nanopore-illumina assembly of an extensively drug-resistant pseudomonas aeruginosa strain from a keratitis outbreak. *Microbiology Resource Announcements*, 13(2):e01188–23.

[22] Campos-Magaña, M. A., Martins dos Santos, V. A., and Garcia-Morales, L. (2025). Enabling access to novel bacterial biosynthetic potential from ont draft genomic data. *Microbial Biotechnology*, 18(3):e70104.

[23] Cantacessi, C., Campbell, B., Jex, A., Young, N., Hall, R., Ranganathan, S., and Gasser, R. (2012). Bioinformatics meets parasitology. *Parasite immunology*, 34(5):265–275.

[24] Chakrawarti, A., Eckstrom, K., Laaguiby, P., and Barlow, J. W. (2024). Hybrid illumina-nanopore assembly improves identification of multilocus sequence types and antimicrobial resistance genes of staphylococcus aureus isolated from vermont dairy farms: comparison to illumina-only and r9. 4.1 nanopore-only assemblies. *Access Microbiology*, 6(3):000766–v3.

[25] Chen, Y., Nie, F., Xie, S.-Q., Zheng, Y.-F., Bray, T., Dai, Q., Wang, Y.-X., Xing, J.-f., Huang, Z.-J., Wang, D.-P., et al. (2020a). Fast and accurate assembly of nanopore reads via progressive error correction and adaptive read selection. *BioRxiv*, pages 2020–02.

[26] Chen, Y., Zhang, Y., Wang, A. Y., Gao, M., and Chong, Z. (2021). Accurate long-read de novo assembly evaluation with inspector. *Genome biology*, 22(1):312.

[27] Chen, Z., Erickson, D. L., and Meng, J. (2020b). Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using illumina and oxford nanopore sequencing. *BMC genomics*, 21:1–21.

[28] Chen, Z., Grim, C. J., Ramachandran, P., and Meng, J. (2024). Advancing metagenome-assembled genome-based pathogen identification: unraveling the power of long-read assembly algorithms in oxford nanopore sequencing. *Microbiology Spectrum*, 12(6):e00117–24.

[29] Chi, H., Shi, L., Gan, S., Fan, G., and Dong, Y. (2025). Innovative applications of nanopore technology in tumor screening: An exosome-centric approach. *Biosensors*, 15(4):199.

[30] Choi, K. R., Jang, W. D., Yang, D., Cho, J. S., Park, D., and Lee, S. Y. (2019). Systems metabolic engineering strategies: integrating systems and synthetic biology with metabolic engineering. *Trends in biotechnology*, 37(8):817–837.

[31] Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore dna sequencing. *Nature nanotechnology*, 4(4):265–270.

[32] Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771.

[33] Connell, C., Fung, S., Heiner, C., Bridgham, J., Chakerian, V., Heron, E., Jones, B., Menchen, S., Mordan, W., Raff, M., et al. (1987). Automated dna-sequence analysis. *BioTechniques*, 5(4):342.

[34] Coordinators, N. R. (2015). Database resources of the national center for biotechnology information. *Nucleic acids research*, 44(Database issue):D7.

[35] Costa, E. B., Silva, G. P., and Teixeira, M. G. (2015). Performance evaluation of parallel genome assemblers. In *Proc. of the 7th Int. Conf. on Bioinformatics and Computational Biology (BICOB 2015)*, volume 1, pages 31–38.

[36] Cretu Stancu, M., Van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., De Ligt, J., Pregno, G., Giachino, D., Mandrile, G., Espejo Valle-Inclan, J., et al. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature communications*, 8(1):1326.

[37] Czmil, A., Wronski, M., Czmil, S., Sochacka-Pietal, M., Cmil, M., Gawor, J., Wołkowicz, T., Plewczynski, D., Strzalka, D., and Pietal, M. (2022). Nanoforms: an integrated server for processing, analysis and assembly of raw sequencing data of microbial genomes, from oxford nanopore technology. *PeerJ*, 10:e13056.

[38] De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Sleegers, K., and Van Broeckhoven, C. (2019). Structural variants identified by

oxford nanopore promethion sequencing of the human genome. *Genome research*, 29(7):1178–1187.

[39] De Coster, W., D'hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). Nanopack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15):2666–2669.

[40] Delahaye, C. and Nicolas, J. (2021). Sequencing dna with nanopores: Troubles and biases. *PloS one*, 16(10):e0257521.

[41] Delandre, O., Lamer, O., Loreau, J.-M., Papa Mze, N., Fonta, I., Mosnier, J., Gomez, N., Javelle, E., and Pradines, B. (2024). Long-read sequencing and de novo genome assembly pipeline of two plasmodium falciparum clones (pf 3d7, pf w2) using only the promethion sequencer from oxford nanopore technologies without whole-genome amplification. *Biology*, 13(2):89.

[42] Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319.

[43] Dorfner, M., Ott, T., Ott, P., and Oberprieler, C. (2022). Long-read genotyping with slang (simple long-read loci assembly of nanopore data for genotyping). *Applications in Plant Sciences*, 10(3):e11484.

[44] Elton, L., Aydin, A., Stoker, N., Rofael, S., Wildner, L. M., Abdul, J. B. P. A. A., Tembo, J., Hamid, M. A., Claujens Chastel, M. M., Canseco, J. O., et al. (2025). A pragmatic pipeline for drug resistance and lineage identification in mycobacterium tuberculosis using whole genome sequencing. *PLOS Global Public Health*, 5(2):e0004099.

[45] Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome research*, 8(3):175–185.

[46] Faber, Q., West, J. R., Corriveau, E. J., and Barbato, R. A. (2025). Complete genomic sequences of nine bacillota isolated from alaskan permafrost. *Microbiology Resource Announcements*, 14(9):e00268–25.

[47] Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van De Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G., and Ysebaert, M. (1978). Complete nucleotide sequence of sv40 dna. *Nature*, 273(5658):113–120.

[48] Firtina, C., Kim, J. S., Alser, M., Senol Cali, D., Cicek, A. E., Alkan, C., and Mutlu, O. (2020). Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm. *Bioinformatics*, 36(12):3669–3679.

[49] Flicek, P. and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature methods*, 6(Suppl 11):S6–S12.

[50] Frantsuzova, E., Kocharovskaya, Y., Vetrova, A., Delegan, Y., and Bogun, A. (2025). Complete genome sequence of gordonia rubripertincta 112, a promising degrader of aromatic and aliphatic compounds. *Microbiology Resource Announcements*, pages e00114–25.

[51] Fukuda, A., Murakami, T., Abe, N., Suzuki, Y., Nakajima, C., and Usui, M. (2025). Complete genome sequence of treponema medium isolated from foot of bovine digital dermatitis in japan. *Microbiology Resource Announcements*, 14(10):e00656–25.

[52] Garrido-Cardenas, J. A., Garcia-Maroto, F., Alvarez-Bermejo, J. A., and Manzano-Agugliaro, F. (2017). Dna sequencing sensors: an overview. *Sensors*, 17(3):588.

[53] Gates, A. J., Gysi, D. M., Kellis, M., and Barabási, A.-L. (2021). A wealth of discovery built on the human genome project—by the numbers. *Nature*, 590(7845):212–215.

[54] Gibbs, R. A. (2020). The human genome project changed everything. *Nature Reviews Genetics*, 21(10):575–576.

[55] Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., Tischler, G., Jackson, D. K., Keane, T. M., Li, J., et al. (2017). De novo yeast genome assemblies from minion, pacbio and miseq platforms. *Scientific reports*, 7(1):3935.

[56] Goldstein, S., Beka, L., Graf, J., and Klassen, J. L. (2019). Evaluation of strategies for the assembly of diverse bacterial genomes using minion long-read sequencing. *BMC genomics*, 20:1–17.

[57] Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics*, 17(6):333–351.

[58] Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

[59] Hackl, S. T., Harbig, T. A., and Nieselt, K. (2022). Technical report on best practices for hybrid and long read de novo assembly of bacterial genomes utilizing illumina and oxford nanopore technologies reads. *bioRxiv*, pages 2022–10.

[60] Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*, 68(4):669–685.

[61] Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8.

[62] Heikema, A. P. et al. (2021). Wefacenano: a user-friendly pipeline for complete ont sequence assembly and detection of antibiotic resistance in multi-plasmid bacterial isolates. *Journal Name*.

[63] Hetland, M. A., Winkler, M. A., Kaspersen, H., Håkonsholm, F., Bakksjø, R.-J., Bernhoff, E., Delgado-Blas, J. F., Brisse, S., Correia, A., Fostervold, A., et al. (2025). Complete genomes of 568 diverse klebsiella pneumoniae species complex isolates from humans, animals, and marine sources in norway from 2001 to 2020. *Microbiology resource announcements*, 14(6):e00931–24.

[64] Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. (2012). Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13:1–16.

[65] Hu, L., Liang, F., Cheng, D., Zhang, Z., Yu, G., Zha, J., Wang, Y., Xia, Q., Yuan, D., Tan, Y., et al. (2020). Location of balanced chromosome-translocation breakpoints by long-read sequencing on the oxford nanopore platform. *Frontiers in genetics*, 10:1313.

[66] Huddleston, J., Chaisson, M. J., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27(5):677–685.

[67] Hunt, M., Newbold, C., Berriman, M., and Otto, T. D. (2014). A comprehensive evaluation of assembly scaffolding tools. *Genome biology*, 15:1–15.

[68] Idury, R. M. and Waterman, M. S. (1995). A new algorithm for dna sequence assembly. *Journal of computational biology*, 2(2):291–306.

[69] Jain, C., Rhie, A., Hansen, N., Koren, S., and Phillippy, A. M. (2020). A long read mapping method for highly repetitive reference sequences. *BioRxiv*, pages 2020–11.

[70] Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338–345.

[71] Jr, E. W. M. (2016). A history of dna sequence assembly. *it - Information Technology*, 58(3):126–132.

[72] Juhas, M., Van Der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., and Crook, D. W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS microbiology reviews*, 33(2):376–393.

[73] Kang, M., Harjadi, D., Hoover, E., Dan, H., Adam, B., and Huang, H. (2025). Complete genome sequences of six serratia strains with multiple resistance genes isolated from food products in canada. *Microbiology Resource Announcements*, pages e00301–25.

[74] Kent, A. G., Sula, E., Breaker, E., McAllister, G. A., Gable, P., Chan-Riley, M. Y., Leung, V. H., Peaper, D. R., Jones, S. A., Jones, J. M., et al. (2025). Complete genome sequence of a novel acinetobacter spp. linked to an outbreak of sepsis following apheresis platelet transfusion. *Microbiology Resource Announcements*, pages e01361–24.

[75] Khrenova, M. G., Panova, T. V., Rodin, V. A., Kryakvin, M. A., Lukyanov, D. A., Osterman, I. A., and Zvereva, M. I. (2022). Nanopore sequencing for de novo bacterial genome assembly and search for single-nucleotide polymorphism. *International Journal of Molecular Sciences*, 23(15):8569.

[76] Kirichek, E. A., Afonin, A. M., Kusakin, P. G., and Tsyganov, V. E. (2025). Complete genome sequence of the unique rhizobium johnstonii strain napi. *Microbiology Resource Announcements*, 14(3):e01202–24.

[77] Kolar, O., Appleberry, H., Wolfe, A. J., Kula, A., and Putonti, C. (2025). Complete genome sequence of vaginal swab isolate enterococcus faecalis umb6935b, including two complete plasmids. *Microbiology Resource Announcements*, 14(8):e00368–25.

[78] Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, 37(5):540–546.

[79] Kuśmirek, W. (2023). Estimated nucleotide reconstruction quality symbols of basecalling tools for oxford nanopore sequencing. *Sensors*, 23(15):6787.

[80] Lang, J. (2022). Maeci: A pipeline for generating consensus sequence with nanopore sequencing long-read assembly and error correction. *Plos one*, 17(5):e0267066.

[81] Laszlo, A. H., Derrington, I. M., Brinkerhoff, H., Langford, K. W., Nova, I. C., Samson, J. M., Bartlett, J. J., Pavlenok, M., and Gundlach, J. H. (2013). Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore mspa. *Proceedings of the National Academy of Sciences*, 110(47):18904–18909.

[82] Lee, G., Kim, M.-J., and Shin, J.-H. (2021a). Data on complete genome sequence and annotation of paenibacillus sonchi lmg 24727t. *Data in Brief*, 38:107271.

[83] Lee, J. Y., Kong, M., Oh, J., Lim, J., Chung, S. H., Kim, J.-M., Kim, J.-S., Kim, K.-H., Yoo, J.-C., and Kwak, W. (2021b). Comparative evaluation of nanopore polishing tools for microbial genome assembly and polishing strategies for downstream analysis. *Scientific Reports*, 11(1):20740.

[84] Lerminiaux, N., Fakharuddin, K., Mulvey, M. R., and Mataseje, L. (2024). Do we still need illumina sequencing data? evaluating oxford nanopore technologies r10. 4.1 flow cells and the rapid v14 library prep kit for gram negative bacteria whole genome assemblies. *Canadian Journal of Microbiology*, 70(5):178–189.

[85] Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

[86] Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., and Fan, W. (2011). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, 11(1):25–37.

[87] Lin, B., Hui, J., and Mao, H. (2021). Nanopore technology and its applications in gene sequencing. *Biosensors*, 11(7):214.

[88] Lin, G. and Liu, S. (2022). The canu genome assembly pipeline using nanopore long reads.

[89] Lohmann, K. and Klein, C. (2014). Next generation sequencing and the future of genetic diagnosis. *Neurotherapeutics*, 11:699–707.

[90] Loman, N. J. and Quinlan, A. R. (2014). Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, 30(23):3399–3401.

[91] Lorv, J. S. and McConkey, B. J. (2025). Complete genome sequence resources for pseudomonas syringae strains 508 and tlp2. *PhytoFrontiers™*, pages PHYTOFR–06.

[92] Luo, J., Wei, Y., Lyu, M., Wu, Z., Liu, X., Luo, H., and Yan, C. (2021). A comprehensive review of scaffolding methods in genome assembly. *Briefings in Bioinformatics*, 22(5):bbab033.

[93] Madigan, M. T., Martinko, J. M., Dunlap, P. V., and Clark, D. P. (2008). Brock biology of microorganisms 12th edn. *Int. Microbiol*, 11(1).

[94] Mak, Q. C., Wick, R. R., Holt, J. M., and Wang, J. R. (2023). Polishing de novo nanopore assemblies of bacteria and eukaryotes with fmlrc2. *Molecular Biology and Evolution*, 40(3):msad048.

[95] Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9(1):387–402.

[96] Mariano, D. C., Leite, C., Santos, L. H., Rocha, R. E., and de Melo-Minardi, R. C. (2017). A guide to performing systematic literature reviews in bioinformatics. *arXiv preprint arXiv:1707.05813*.

[97] Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing dna. *Proceedings of the National Academy of Sciences*, 74(2):560–564.

[98] McDermott, J., Guerquin, M., Frazier, Z., Chang, A. N., and Samudrala, R. (2005). Bioverse: enhancements to the framework for structural, functional and contextual modeling of proteins and proteomes. *Nucleic acids research*, 33(suppl_2):W324–W325.

[99] Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31–46.

[100] Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., et al. (2020). Telomere-to-telomere assembly of a complete human x chromosome. *Nature*, 585(7823):79–84.

[101] Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with quast-lg. *Bioinformatics*, 34(13):i142–i150.

[102] Mikheyev, A. S. and Tin, M. M. (2014). A first look at the oxford nanopore minion sequencer. *Molecular ecology resources*, 14(6):1097–1102.

[103] Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic acids research*, 40(17):e136–e136.

[104] Morisse, P., Lecroq, T., and Lefebvre, A. (2020). Long-read error correction: a survey and qualitative comparison. *BioRxiv*, pages 2020–03.

[105] Murigneux, V., Roberts, L. W., Forde, B. M., Phan, M.-D., Nhu, N. T. K., Irwin, A. D., Harris, P. N., Paterson, D. L., Schembri, M. A., Whiley, D. M., et al. (2021). Micropipe: validating an end-to-end workflow for high-quality complete bacterial genome construction. *BMC genomics*, 22(1):474.

[106] Napieralski, A. and Nowak, R. (2022). Basecalling using joint raw and event nanopore data sequence-to-sequence processing. *Sensors*, 22(6):2275.

[107] National Center for Biotechnology Information (2023). Fasta sequence quick start. Disponível em: https://www.ncbi.nlm.nih.gov/genbank/fastaformat/. Acesso em: 19 fev. 2026.

[108] Oshiro, M., Nakamura, K., Sk, R., Tashiro, Y., and Shiwa, Y. (2025). Complete genome sequence of levilactobacillus acidifarinae type strain jcm 15949 (dsm 19394). *Microbiology Resource Announcements*, pages e00374–25.

[109] Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2017). Rayyan—a web and mobile app for systematic reviews. syst rev. 2016; 5: 210.

[110] Oxford Nanopore Technologies (2023a). *Dorado Basecaller Documentation*. Oxford Nanopore Technologies. Available at: https://github.com/nanoporetech/dorado.

[111] Oxford Nanopore Technologies (2023b). *Guppy Basecalling Software Documentation*. Oxford Nanopore Technologies. Available at: https://nanoporetech.com.

[112] Oxford Nanopore Technologies (2023c). Medaka. https://github.com/nanoporetech/medaka.

[113] Oxford Nanopore Technologies (2025). Guppy basecalling software. `https://community.nanoporetech.com`. Accessed: 2025-05-17.

[114] Pattan, V., Kashyap, R., Bansal, V., Candula, N., Koritala, T., and Surani, S. (2021). Genomics in medicine: a new era in medicine. *World Journal of Methodology*, 11(5):231.

[115] Payne, A., Holmes, N., Rakyan, V., and Loose, M. (2019). Bulkvis: a graphical viewer for oxford nanopore bulk fast5 files. *Bioinformatics*, 35(13):2193–2198.

[116] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448.

[117] Pervez, M. T., Hasnain, M. J. U., Abbas, S. H., Moustafa, M. F., Aslam, N., and Shah, S. S. M. (2022). [retracted] a comprehensive review of performance of next-generation sequencing platforms. *BioMed research international*, 2022(1):3457806.

[118] Pevzner, P. A., Tang, H., and Tesler, G. (2004). De novo repeat classification and fragment assembly. In *Proceedings of the eighth annual international conference on Research in computational molecular biology*, pages 213–222.

[119] Player, R., Verratti, K., Staab, A., Forsyth, E., Ernlund, A., Joshi, M. S., Dunning, R., Rozak, D., Grady, S., Goodwin, B., et al. (2022). Optimization of oxford nanopore technology sequencing workflow for detection of amplicons in real time using ont-dart tool. *Genes*, 13(10):1785.

[120] Prior, K., Becker, K., Brandt, C., Cabal Rosel, A., Dabernig-Heinz, J., Kohler, C., Lohde, M., Ruppitsch, W., Schuler, F., Wagner, G., et al. (2025). Accurate and reproducible whole-genome genotyping for bacterial genomic surveillance with nanopore sequencing data. *Journal of Clinical Microbiology*, pages e00369–25.

[121] Puangseree, J., Hein, S. T., Prathan, R., Srisanga, S., and Chuanchuen, R. (2025). Genomic insights into multidrug-resistant salmonella enterica isolates from pet dogs and cats. *Scientific Reports*, 15(1):22104.

[122] Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., et al. (2016). Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232.

[123] Rang, F. J., Kloosterman, W. P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology*, 19(1):90.

[124] Rhoads, A. and Au, K. F. (2015). Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289.

[125] Ruan, Z., Wu, J., Chen, H., Draz, M. S., Xu, J., and He, F. (2020). Hybrid genome assembly and annotation of a pandrug-resistant klebsiella pneumoniae strain using nanopore and illumina sequencing. *Infection and drug resistance*, pages 199–206.

[126] Sanders, B. and Batut, B. (2024). Genome assembly of mrsa from oxford nanopore minion data (and optionally illumina data).

[127] Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison III, C. A., Slocombe, P. M., and Smith, M. (1977a). Nucleotide sequence of bacteriophage φx174 dna. *nature*, 265(5596):687–695.

[128] Sanger, F., Coulson, A., Friedmann, T., Air, G., Barrell, B., Brown, N., Fiddes, J., Hutchison Iii, C., Slocombe, P., and Smith, M. (1978). The nucleotide sequence of bacteriophage φx174. *Journal of molecular biology*, 125(2):225–246.

[129] Sanger, F., Coulson, A. R., Hong, G., Hill, D., and Petersen, G. d. (1982). Nucleotide sequence of bacteriophage λ dna. *Journal of molecular biology*, 162(4):729–773.

[130] Sanger, F., Nicklen, S., and Coulson, A. R. (1977b). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467.

[131] Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., et al. (2023). Next-generation sequencing technology: current trends and advancements. *Biology*, 12(7):997.

[132] Sayers, E. W., Cavanaugh, M., Frisse, L., Pruitt, K. D., Schneider, V. A., Underwood, B. A., Yankie, L., and Karsch-Mizrachi, I. (2025). Genbank 2025 update. *Nucleic Acids Research*, 53(D1):D56–D61.

[133] Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome research*, 20(9):1165–1173.

[134] Schreiber, J., Wescoe, Z. L., Abu-Shumays, R., Vivian, J. T., Baatar, B., Karplus, K., and Akeson, M. (2013). Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual dna strands. *Proceedings of the National Academy of Sciences*, 110(47):18910–18915.

[135] Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018a). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19(6):329–346.

[136] Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., and Schatz, M. C. (2018b). Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods*, 15(6):461–468.

[137] Silva, R. C., Lima, A., and da Silva Souza, L. C. (2022). Principais métodos de sequenciamento de dna. *Scientific Electronic Archives*, 15(10).

[138] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

[139] Simpson, J. T., Workman, R. E., Zuzarte, P., David, M., Dursi, L., and Timp, W. (2017). Detecting dna cytosine methylation using nanopore sequencing. *Nature methods*, 14(4):407–410.

[140] Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B., and Hood, L. E. (1986). Fluorescence detection in automated dna sequence analysis. *Nature*, 321(6071):674–679.

[141] Staden, R. (1979). A strategy of dna sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–2610.

[142] Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.

[143] Tang, C. and Holden, D. (1999). Pathogen virulence genes–implications for vaccines and drug therapy. *British medical bulletin*, 55(2):387–400.

[144] Todar, K. (2004). Todar's online textbook of bacteriology.

[145] Tonkin-Hill, G., Corander, J., and Parkhill, J. (2023). Challenges in prokaryote pangenomics. *Microbial Genomics*, 9(5):001021.

[146] Treangen, T. J. and Salzberg, S. L. (2012). Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46.

[147] Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R., and Corbett, C. R. (2018). Evaluation of oxford nanopore's minion sequencing device for microbial whole genome sequencing applications. *Scientific reports*, 8(1):10931.

[148] Urban, L., Holzer, A., Baronas, J. J., Hall, M. B., Braeuninger-Weimer, P., Scherm, M. J., Kunz, D. J., Perera, S. N., Martin-Herranz, D. E., Tipper, E. T., et al. (2021). Freshwater monitoring by nanopore sequencing. *elife*, 10:e61504.

[149] Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics*, 34(9):666–681.

[150] Vanhecke, T. E. (2008). Zotero. *Journal of the Medical Library Association: JMLA*, 96(3):275.

[151] Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5):737–746.

[152] Vereecke, N., Bokma, J., Haesebrouck, F., Nauwynck, H., Boyen, F., Pardon, B., and Theuns, S. (2020). High quality genome assemblies of mycoplasma bovis using a taxon-specific bonito basecaller for minion and flongle long-read nanopore sequencing. *BMC bioinformatics*, 21:1–16.

[153] Vereecke, N., Yoon, T. B., Luo, T. L., Corey, B. W., Lebreton, F., Mc Gann, P. T., and Dekker, J. P. (2025). An open-source nanopore-only sequencing workflow for analysis of clonal outbreaks delivers short-read level accuracy. *Journal of Clinical Microbiology*, 63(8):e00664–25.

[154] Versmessen, N., Mispelaere, M., Vandekerckhove, M., Hermans, C., Boelens, J., Vranckx, K., Van Nieuwerburgh, F., Vaneechoutte, M., Hulpiau, P., and Cools, P. (2024). Average nucleotide identity and digital dna-dna hybridization analysis following promethion nanopore-based whole genome sequencing allows for accurate prokaryotic typing. *Diagnostics*, 14(16):1800.

[155] Wagner, G. E., Dabernig-Heinz, J., Lipp, M., Cabal, A., Simantzik, J., Kohl, M., Scheiber, M., Lichtenegger, S., Ehricht, R., Leitner, E., et al. (2023). Real-time nanopore q20+ sequencing enables extremely fast and accurate core genome mlst typing and democratizes access to high-resolution bacterial pathogen surveillance. *Journal of Clinical Microbiology*, 61(4):e01631–22.

[156] Wick, R., Howden, B., and Stinear, T. (2025a). Autocycler: long-read consensus assembly for bacterial genomes. bioinformatics 41: btaf474.

[157] Wick, R. R. and Holt, K. E. (2021). Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research*, 8:2138.

[158] Wick, R. R., Judd, L. M., and Holt, K. E. (2019). Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome biology*, 20:1–10.

[159] Wick, R. R., Judd, L. M., Stinear, T. P., and Monk, I. R. (2025b). Are reads required? high-precision variant calling from bacterial genome assemblies. *Access Microbiology*, 7(5):001025–v3.

[160] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., and Baak, A. (2018). & bouwman, j.(2016). the fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):160018.

[161] Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., and Zhang, L. (2020). Review on the application of machine learning algorithms in the sequence data mining of dna. *Frontiers in Bioengineering and Biotechnology*, 8:1032.

[162] Zeng, J., Cai, H., Peng, H., Wang, H., Zhang, Y., and Akutsu, T. (2020). Causalcall: Nanopore basecalling using a temporal convolutional network. *Frontiers in Genetics*, 10:1332.

[163] Zhang, C., Wang, F., Zhu, C., Xiu, L., Li, Y., Li, L., Liu, B., Li, Y., Zeng, Y., Guo, B., et al. (2020a). Determining antimicrobial resistance profiles and identifying novel mutations of neisseria gonorrhoeae genomes obtained by multiplexed minion sequencing. *Science China Life Sciences*, 63:1063–1070.

[164] Zhang, H., Jain, C., and Aluru, S. (2020b). A comprehensive evaluation of long read error correction methods. *BMC genomics*, 21:1–15.

[165] Zhang, P., Jiang, D., Wang, Y., Yao, X., Luo, Y., and Yang, Z. (2021). Comparison of de novo assembly strategies for bacterial genomes. *International Journal of Molecular Sciences*, 22(14):7668.

[166] Zhang, Y.-z., Akdemir, A., Tremmel, G., Imoto, S., Miyano, S., Shibuya, T., and Yamaguchi, R. (2020c). Nanopore basecalling from a perspective of instance segmentation. *BMC bioinformatics*, 21:1–9.

[167] Zidane, N., Rodrigues, C., Bouchez, V., Rethoret-Pasty, M., Passet, V., Brisse, S., and Crestani, C. (2025). Accurate genotyping of three major respiratory bacterial pathogens with ont r10. 4.1 long-read sequencing. *Genome Research*, pages gr–279829.

[168] Zucherato, V. S. (2023). *Aplicação da tecnologia portátil de sequenciamento MinION para análise do metagenoma viral*. PhD thesis, Universidade de São Paulo.