Universidade de Brasília
Instituto de Biologia
Departamento de Biologia Celular

Felipe Marques de Almeida

Desenvolvimento de protocolos computacionais escalonáveis para a análise completa de genomas bacterianos e sua aplicação em um estudo comparativo e retrospectivo de cepas multirresistentes de Brasília

Brasília
23 de Agosto, 2024

Felipe Marques de Almeida

Desenvolvimento de protocolos computacionais escalonáveis para a análise completa de genomas bacterianos e sua aplicação em um estudo comparativo e retrospectivo de cepas multirresistentes de Brasília

Tese de Doutorado apresentada ao Departamento de Biologia Celular do Instituto de Biologia da Universidade de Brasília como requisito para a obtenção do título de Doutor em Biologia Molecular com ênfase em Biologia Computacional.

Área de Concentração: Bioinformática
Orientador(a): Prof. Dr. Georgios Joannis Pappas Júnior

Brasília

23 de Agosto, 2024

Dedico este trabalho a meus familiares, amigos e amada que estiveram sempre presentes e me apoiaram ao longo da batalha.

Agradecimentos

A trajetória percorrida para a conclusão deste Doutorado foi repleta de desafios e incertezas. Sua conclusão só foi possível devido à participação e a contribuição de muitas pessoas e, a estas, dedico este trabalho.

Ao meu orientador, Professor Doutor Georgios Pappas que, durante o processo, sempre acreditou em mim e me incentivou. Através de seu rigor científico, orientação e exigência saudável, contribuiu para o enriquecimento e desenvolvimento deste trabalho em todas as suas etapas. Com grande paciência e dedicação, ele foi importantíssimo para meu crescimento como pesquisador científico.

Agradeço enormemente à minha esposa, Kissia Batista pelo carinho, apoio e companheirismo que foram fundamentais para que eu me mantivesse focado e não me deixou desanimar e sucumbir às adversidades enfrentadas durante o processo.

Por último, quero agradecer à minha família e amigos pelo total apoio recebido. Agradeço, especialmente à minha mãe, Waleska Oliveira, e pai, Edson Almeida, pelos imensuráveis esforços em disponibilizar os meios e as condições necessárias para que seus filhos se dedicassem aos estudos.

"*Forte não é aquele que nunca vai cair, é aquele que sempre vai conseguir se levantar*"

Lenilson Xavier

"*A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo*"

Albert Einstein

"*A ciência nunca resolve um problema sem criar pelo menos outros dez*"

George Bernard Shaw

"*A ciência consiste em substituir o saber que parecia seguro por uma teoria, ou seja, por algo problemático*"

José Ortega y Gasset

Resumo

Os avanços nas tecnologias de sequenciamento de DNA remodelaram a genômica bacteriana, permitindo montagens em nível de cromossomo por uma fração de custo e tempo. No entanto, a plena realização deste potencial depende de recursos computacionais para analisar os dados. Por isso, desenvolvemos três *pipelines* eficientes e padronizados que compreendem o controle de qualidade de dados brutos de sequenciamento, montagem de genoma e extensa anotação genética com relatórios gráficos. Em particular, fornecemos um módulo de anotação especializado para bactérias envolvidas em infecções relacionadas à assistência à saúde (IRAS), que identifica genes de resistência a antimicrobianos (ARG), fatores de virulência, profagos e elementos integrativos. Somado a isso, desenvolveu-se uma aplicação web para integralização de todos os resultados do *pipeline* de anotação. Para demonstrar seu uso, sequenciamos e analisamos três cepas de *Klebsiella pneumoniae* ST11 provenientes do Hospital Universitário de Brasília. Detectamos a presença dos genes *bla*NDM-1, *bla*CTX-M-15, *bla*TEM-1 e *bla*OXA beta-lactamases nos três isolados, bem como vários outros ARGs distribuídos entre cromossomos e plasmídeos. Além disso, também foram detectados genes de virulência frequentemente relacionados a linhagens hiper virulentas, como Salmochelina. Por fim, realizamos uma análise genômica comparativa retrospectiva utilizando outras linhagens de CRKP isoladas de Brasília e de outras localidades brasileiras. De modo geral, os resultados indicam que o fenótipo de resistência está mudando, com isolados recentes exibindo a coexistência de múltiplas carbapenemases, e que elementos genéticos móveis podem estar desempenhando papel chave nessas observações. Além disso, os resultados reiteram alertas quanto a emergência de clones de alto risco apresentando a convergência de genes de virulência e resistência, ressaltando a urgência em adotar a vigilância genômica de rotina para combater a disseminação dessas características de alto risco.

Abstract

Advances in DNA sequencing technologies have reshaped bacterial genomics, enabling chromosome-level assemblies at a fraction of the cost and time. However, the full realization of this potential depends on computational resources to analyze the data. Therefore, we have developed three efficient and standardized *pipelines* that comprise quality control of raw sequencing data, genome assembly, and extensive gene annotation with graphical reporting. In particular, we provide a specialized annotation module for bacteria involved in healthcare-associated infections (HAIs), which identifies antimicrobial resistance genes (ARGs), virulence factors, prophages, and integrative elements. In addition, we have developed a web application to integrate all the results of the annotation *pipeline*. To demonstrate its use, we sequenced and analyzed three *Klebsiella pneumoniae* ST11 strains from the University Hospital of Brasília. We detected the presence of the *bla*NDM-1, *bla*CTX-M-15, *bla*TEM-1 and *bla*OXA beta-lactamase genes in the three isolates, as well as several other ARGs distributed between chromosomes and plasmids. In addition, virulence genes frequently associated with hypervirulent strains, such as Salmochelin, were also detected. Finally, we performed a retrospective comparative genomic analysis using other CRKP strains isolated from Brasília and other Brazilian locations. Overall, the results indicate that the resistance phenotype is changing, with recent isolates exhibiting the coexistence of multiple carbapenemases, and that mobile genetic elements may be playing a key role in these observations. Furthermore, the results reiterate warnings regarding the emergence of high-risk clones presenting the convergence of virulence and resistance genes, highlighting the urgency of adopting routine genomic surveillance to combat the spread of these high-risk traits.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AAC *Aminoglycoside acetyltransferase*

ANI *Average Nucleotide Identity*

AQI *Assembly Quality Index*

ARG *Antimicrobial Resistance Gene*

ASA3P *Automatic Bacterial Isolate Assembly, Annotation and Analyses Pipeline*

BGI *Beijing Genomics Institute*

BIGSdn *Bacterial Isolate Genome Sequence Database*

BLAST *Basic Local Alignment Search Tool*

BUSCO *Benchmarking Universal Single-Copy Orthologs*

CARD *Comprehensive Antibiotic Resistance Database*

CDC *Centers for Disease Control and Prevention*

CG *Clonal group*

CLSI *Clinical & Laboratory Standards Institute*

CpKP *Carbapenem-producing Klebsiella pneumoniae*

CPS *Capsular polysaccharides*

CRAQ *Clipping Reveals Assembly Quality*

CRKP *Carbapenem-resistance Klebsiella pneumoniae*

CSV *Comma-separated values*

CWL *Common Workflow Language*

DDBJ *DNA Data Bank of Japan*

DFAST *DDBJ Fast Annotation and Submission Tool*

DSL *Domain-specific language*

ECDC *European Centre for Disease Prevention and Control*

ESBL *Extended Spectrum Beta-Lactamase*

ESKAPE *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumanii, Pseudomonas aeruginosa* e *Enterobacter spp.*

FAIR *Findable, Accessible, Interoperable, Reusable*

GFF *General File Format*

HAI *Healthcare-associated infection*

HUB Hospital Universitário da UnB

hvKp *Hypervirulent Klebsiella pneumoniae*

ICE *Integrative conjugative element*

IRAS Infecções Relacionadas à Assistência à Saúde

IS *Insertion Sequence*

JSON *JavaScript Object Notation*

KEGG *Kyoto Encyclopedia of Genes and Genomes*

Kp *Klebsiella pneumoniae*

KPC *Klebsiella pneumoniae carbapenemase*

LAMP/UnB Laboratório de Análises Moleculares de Patógenos da UnB

LPS Lipopolissacarídeo

MDR *Multidrug resistant*

MGE *Mobile genetic element*

MIC *Minimum inhibitory concentration*

MLST *Multilocus sequence typing*

NCBI National Center for Biotechnology Information

NDM *New Delhi metallo-β-lactamase*

NGS *Next Generation Sequencing*

OMS Organização mundial da saúde

ONT *Oxford Nanopore Technologies*

OXA *Oxacilinase β-lactamase*

PacBio *Pacific Biosciences*

PCR *Polymerase Chain Reaction*

PDR Pandrug resistant

PGAP NCBI Prokaryotic Genome Annotation Pipeline

POSIX *Portable Operating System Interface*

RND *Resistance-nodulation-division*

SQL *Structured Query Language*

SRA *Sequence Read Archive*

ST *Sequence type*

T6SS *Type VI secretion system*

UnB Universidade de Brasília

VFDB *Virulence Factor Database*

WSL *Windows Subsystem for Linux*

XDR *Extensively drug-resistant*

YAML *Yet Another Markup Language*

Sumário

1 Introdução

1.1 A resistência a antimicrobianos

A História da humanidade tem sido marcada por pandemias bacterianas, como a peste bubônica, que na idade média, causou a morte de milhões de pessoas na Europa (Mohr, 2016). Em 1928, a descoberta da penicilina por Alexander Fleming deu início a uma nova era de controle e tratamento de infecções por meio de compostos químicos que eliminam ou interferem com o crescimento de microrganismos, genericamente denominados de agentes antimicrobianos (Fleming, 1929). Porém, logo em 1945, os primeiros relatos de bactérias resistentes a estes compostos começaram a aparecer (Plough, 1945; Waksman et al., 1945).

Ao longo da metade final do século XX, diversas moléculas com ação antimicrobiana foram descobertas ou desenhadas, estabelecendo um verdadeiro arsenal contra bactérias patogênicas. No entanto, fatores como o uso indiscriminado de antimicrobianos, impulsionam o processo contínuo de evolução bacteriana, e aceleram a seleção de linhagens resistentes, dificultando a contenção de infecções (Waddington et al., 2022; Djordjevic et al., 2023; Castañeda Barba et al., 2023). A rápida disseminação da resistência a antimicrobianos compromete a medicina moderna, colocando em risco o tratamento de infecções comuns e complexas (Waddington et al., 2022; Walsh et al., 2023; Silva et al., 2023).

Em função disso, e pelo falta de desenvolvimento de novos antimicrobianos, nas últimas décadas houve esforços para a criação de uma aliança global para o combate a resistência a antimicrobianos (Neu, 1992; Gold e Moellering, 1996). Este problema atingiu tamanha proporção, que, atualmente, a resistência a antimicrobianos é classificada como uma das maiores ameaças à saúde pública global [1] (Dodds, 2017; Baker et al., 2018;

[1] www.who.int/en/news-room/fact-sheets/detail/antibiotic-resistance

Walsh et al., 2023; Silva et al., 2023). Isso suscita a mobilização de esforços para uma compreensão aprofundada dos mecanismos moleculares de resistência a antimicrobianos, bem como estratégias inovadoras para abordar esse problema premente.

1.2 Bases moleculares da resistência

Bactérias, naturalmente possuem uma grande plasticidade genética que permite o emprego de uma ampla gama de mecanismos para responder e adaptar rapidamente à perturbações ambientais (Blair et al., 2015; Walsh et al., 2023). Estas ameaças ambientais são diversas, e incluem a exposição a moléculas que podem comprometer sua existência, como os antimicrobianos. A resistência pode acontecer naturalmente por características inerentes à bactéria, como em bactérias Gram-negativas que são naturalmente resistentes à vancomicina, devido à membrana externa que impede a passagem da molécula (Peterson e Kaur, 2018; Ghai, 2023). Alternativamente, podem ser adquiridas através de duas formas principais: (i) mutação; e (ii) aquisição de DNA através de transferência horizontal (Dodds, 2017; Durão et al., 2018; Walsh et al., 2023).

Genes e mutações de resistência podem aparecer tanto em cromossomos quanto em elementos genéticos móveis (MGE, do inglês "mobile genetic element"), como plasmídeos e transposons. Mesmo com potencial impacto negativo na fisiologia bacteriana, mutações capazes de interferir na eficiência de antibióticos surgem estocasticamente e, consequentemente, podem dar origem a populações resistentes através de sua fixação pelo processo de seleção natural (Hughes e Andersson, 2017; Bell e MacLean, 2018; Durão et al., 2018; Gifford et al., 2023).

Estas características podem ser propagadas de forma vertical (entre gerações) através do processo de divisão celular ou de forma horizontal entre uma célula doadora e uma receptora, fora do processo de divisão, geralmente pelo processo de conjugação (Bethke et al., 2022).

A transferência horizontal de genes é um mecanismo essencial na evolução bacteriana, permitindo a rápida aquisição e disseminação de novos traços adaptativos, como por exemplo, a resistência a antimicrobianos, que podem ou não ser fixados na população por seleção natural (Lee et al., 2022). Processo, este, que é potencializado pela cons-

tante pressão seletiva imposta por ações antrópicas, que influenciam na velocidade destes eventos (Dodds, 2017; Durão et al., 2018; Walsh et al., 2023).

Desta forma, tem-se um arcabouço dinâmico de genes que estão em constante processo de aquisição e eliminação, que constituem o que denominamos de pangenoma bacteriano (Lee et al., 2022). Pangenoma, é essencialmente o repertório completo de genes de um clado filogenético e é comumente sub-divido em duas partes: (i) os genes "core" que estão presentes em todas as linhagens do clado e; (ii) os genes acessório, ou genes dispensáveis, que estão presentes em N linhagens mas não em todas (Vernikos et al., 2015). Este conceito permite o entendimento e mensuração da variabilidade genética de espécies, através do cálculo do número necessário de genomas para obtenção de um pangenoma fechado, isto é, que contenha todo o repertório genético da espécie (Vernikos et al., 2015). De maneira geral, diversas espécies de patógenos bacterianos apresentam um pangenoma aberto, que ainda necessitariam um número muito grande de genomas para identificar todo o repertório genético (Vernikos et al., 2015).

Neste cenário, ressalta-se a importância destes elementos genéticos móveis para potencializar e permitir a rápida aquisição de traços adaptativos deste dinâmico pool genético proporcionado pelo pangenoma (Lee et al., 2022). Por exemplo, em 2022, foi detectado na China um caso de transmissão horizontal de genes KPC-3 através de plasmídeos IncX8 entre diferentes espécies de Enterobacterales (Chen et al., 2022).

1.2.1 Mecanismos de resistência a antimicrobianos

Em termos moleculares, bactérias contrapõem a ação de antimicrobianos concentrando mutações em genes envolvidos em mecanismos celulares que incluem: (i) minimização da concentração intracelular da molécula; (ii) modificação, ou proteção, do alvo da droga, ou; (iii) inativação química da molécula antimicrobiana (Iovleva e Doi, 2017; Smith et al., 2023). Uma representação esquemática de classes de drogas e seu enquadramento em termos de mecanismo de ação encontra-se na Figura 1 e uma descrição pormenorizada dos mesmos é apresentada a seguir.

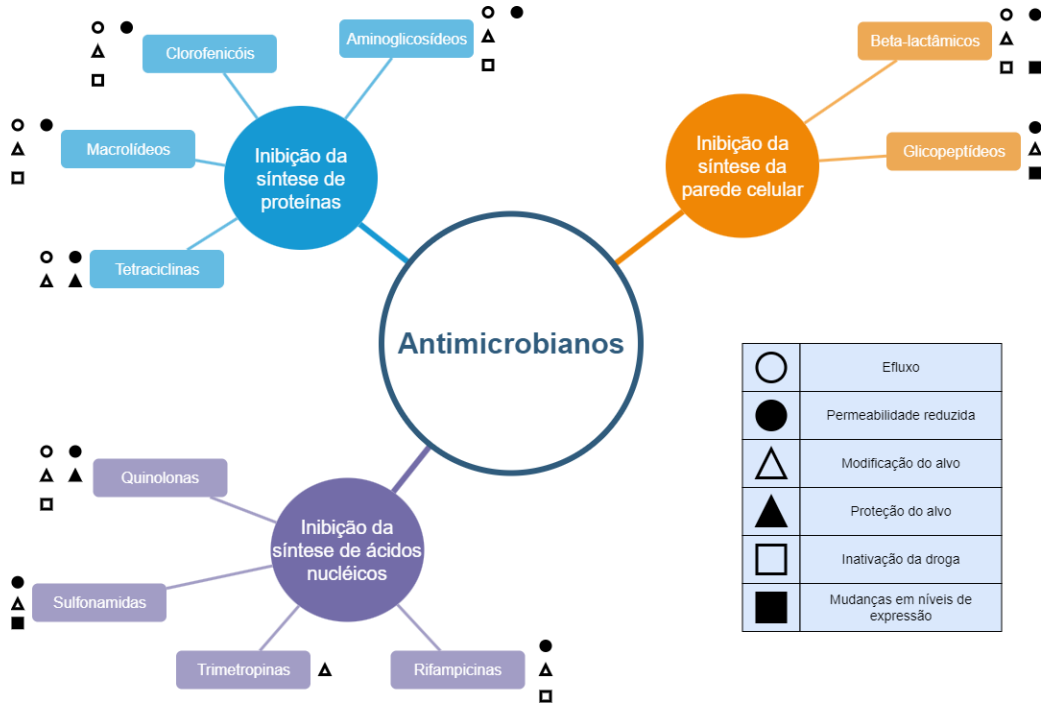

Figura 1: Representação esquemática das principais classes de antimicrobianos da atualidade, separados e coloridos por seus respectivos mecanismos de ação. As formas geométricas representam os principais mecanismos de resistência conhecidos para cada classe. Fonte: de Almeida (2020).

1.2.1.1 Minimização de concentrações intracelulares

Bactérias são capazes de controlar a concentração de certas moléculas na célula através do balanço de suas taxas de influxo e efluxo. Em bactérias Gram-negativas, proteínas localizadas na membrana externa, denominadas porinas, criam canais na membrana que permitem o transporte passivo de pequenas moléculas, como por exemplo, compostos da classe de β-lactâmicos e algumas fluoroquinolonas (Ghai, 2023). A presença de mutações nessas porinas e a redução de sua expressão são geralmente correlacionadas à resistência a carbapenemas (Hao et al., 2018; Kong et al., 2018; Ghai, 2023).

Além das porinas, outros componentes estão envolvidos na minimização de concentrações de antimicrobianos em células bacterianas, como por exemplo, as bombas de efluxo e diversas enzimas de degradação (De Oliveira et al., 2020; Ghai, 2023). As enzimas de degradação serão detalhadas em seção específica sobre inativação de moléculas antimicrobianas (Subsubseção 1.2.1.3).

As bombas de efluxo são responsáveis pela extrusão ativa de antimicrobianos

(De Gaetano et al., 2023). Dentre as famílias de bombas de efluxo melhor caracterizadas podemos destacar a superfamília de bombas de divisão celular de nodulação de resistência (RND, do inglês "resistance-nodulation-division"), altamente relevante para bactérias Gram-negativas, como por exemplo, *Escherichia coli* e *Pseudomonas aeruginosa* (Gaurav et al., 2023). Finalmente, existem bombas de efluxo capazes de transportar uma grande variedade de moléculas, denominadas "bombas de efluxo de resistência a múltiplas drogas (MDR)".

Estas bombas de efluxo, quando super expressadas ou em combinação com outros mecanismos de resistência, podem conferir altos níveis tolerância e até mesmo resistência (Du et al., 2018; Silva et al., 2023; De Gaetano et al., 2023). Sua super expressão é geralmente adquirida através da mutação ou ligação de moléculas a seus reguladores (Baucheron et al., 2014; Du et al., 2018). Dois exemplos de resistência resultantes da super expressão destas bombas são: (i) a resistência a Tigeciclina em *Staphylococcus aureus* ligada à bomba MepA (Kim et al., 2021) e (ii) a resistência a aminoglicosídeos e fluoroquinolonas em *Salmonella enterica* relacionada à bomba MdtK (Nishino et al., 2009).

1.2.1.2 Modificação do alvo

Outra forma pela qual bactérias podem adquirir resistência é modificando os alvos dos antimicrobianos. Essas modificações interferem na eficiência da ligação da droga, mas ainda permitem que estes alvos mantenham sua função fisiológica (Blair et al., 2015; Smith et al., 2023). A modificação destes alvos proteicos pode ser realizada de diversas formas, incluindo modificações genéticas, pós-transcricionais e pós-traducionais.

Estas modificações podem atingir altos níveis de efetividade. Sabe-se, por exemplo, que a presença de mutações em uma única cópia de genes destas moléculas alvo já pode ser suficiente para promover tolerância à droga (Peterson e Kaur, 2018). Além de mutações, modificações químicas, e até mesmo modificações epigenéticas, são também bastante relevantes para o desenvolvimento de fenótipos de resistência (Smith et al., 2023; Wang et al., 2023).

A resistência a fluoroquinolonas pode ocorrer comumente através de mutações es-

pontâneas nos genes *gyrA* e *parC*, que alteram o sítio de ligação da quinolona e reduzem sua afinidade, com consequente redução de suscetibilidade (Tang e Zhao, 2023). No entanto, modificações químicas em componentes não-proteicos também tem efeito destacado. Um exemplo de resistência desta categoria é a resultante de metilações pontuais no RNA ribossômico mediadas pela Cfr rRNA metiltransferase, as quais afetam a ligação de antimicrobianos, gerando resistência a diferentes antimicrobianos (Long et al., 2006). Outro exemplo bastante relevante é a alteração das moléculas de lipopolissacarídeos, induzida pela ativação de sistemas de resposta ao estresse do envelope celular, que acarreta na resistência à polimixina, um antimicrobiano de último recurso (Schaenzer e Wright, 2020; Abavisani et al., 2023).

1.2.1.3 Inativação de moléculas antimicrobianas

A degradação ou inativação direta de moléculas antimicrobianas é talvez uma das estratégias de resistência mais comumente empregadas por patógenos bacterianos (De Oliveira et al., 2020). Em bactérias Gram-negativas, os principais mecanismos envolvem a produção de enzimas que destruam estes agentes antimicrobianos ou que neutralizem estas drogas através de modificações covalentes nas mesmas (De Oliveira et al., 2020).

As enzimas denominadas β-lactamases são talvez as melhor descritas desta categoria. Estas enzimas concentram-se no periplasma bacteriano e são capazes de hidrolisar drogas β-lactâmicas, como cefalosporinas e carbapenemas, conferindo resistência (Bush, 2023).

Baseando-se nas estruturas primárias destas enzimas, pode-se dividir as β-lactamases em quatro classes: A, B, C e D. As enzimas das classes A, C e D, são chamadas serina β-lactamases e possuem um sítio ativo de serina, enquanto que enzimas de classe B, denominadas metalo-β-lactamases, são enzimas zinco-dependentes (Lupo et al., 2022). Além desta primeira classificação por homologia, estas enzimas são também organizadas em subgrupos funcionais baseados nos diferentes perfis de substratos (Bush, 2023). Entre as principais enzimas representantes para cada uma dessas classes, temos:

- a "*Klebsiella pneumoniae* carbapenemase" (KPC) e a CTX-M, enzimas Classe A;

- a "New Delhi metallo-β-lactamase" (NDM), uma enzima Classe B;
- a AmpC, uma enzima Classe C e;
- a "Oxacillinase β-lactamase" (OXA), uma enzima Classe D.

Além da destruição, como mencionado anteriormente, a simples transferência enzimática de grupamentos químicos aos antimicrobianos pela ação de diversas enzimas, como por exemplo as transferases, pode impedir sua ligação ao alvo. Este, é geralmente o caso de aminoglicosídeos que devido ao seu tamanho, são bastante suscetíveis a modificações químicas (Romanowska et al., 2013; De Oliveira et al., 2020). Um exemplo são as enzimas denominadas AAC ("aminoglycoside acetyltransferases") que catalisam a acetilação de aminoácidos específicos na molécula antimicrobiana (De Oliveira et al., 2020). Ressalta-se que a resistência à aminoglicosídeos não ocorre somente pela modificação da molécula, mas pode também ocorrer através da modificação do sítio ribossômico alvo (Miller e Arias, 2024).

1.3 Bactérias podem acumular múltiplos fatores de resistência

Em termos gerais, bactérias são capazes de acumular diversos mecanismos de resistência a diferentes antimicrobianos, culminando em fenótipos de resistência múltipla. Estes fenótipos, por sua vez, podem ser classificados de acordo com a amplitude de resistência agregada. Para isso, internacionalmente segue-se as definições propostas por Magiorakos et al. (2012) que permitem classificar isolados em três grandes grupos:

- MDR, do inglês "Multidrug-resistant" para isolados resistentes a pelo menos um agente de três ou mais classes de antimicrobianos;
- XDR, do inglês "Extensively drug-resistant" para isolados resistentes a pelo menos um agente em quase todas as classes, sendo suscetível para duas ou menos e;
- PDR, do inglês "Pandrug-resistant" para isolados resistentes a todos os agentes de todas as classes.

.

No geral, estas classificações de resistência podem ser observadas em uma grande diversidade de linhagens bacterianas, incluindo Gram-negativas e Gram-positivas (Jubeh et al., 2020; Šámal et al., 2022). Entre elas, linhagens MDR são mais comuns, enquanto PDR são mais raras. Por exemplo, um estudo recente com 1.385 isolados bacterianos de culturas de urina observou que 50% das linhagens era MDR, 27% XDR e apenas 1% PDR (Šámal et al., 2022). Além disso, neste conjunto de dados, Gram-positivas representaram menos de 20% do total de isolados.

Apesar disso, é importante ressaltar que atualmente existem novos antimicrobianos não contemplados nessas categorias definidas por Magiorakos et al. (2012). Adicionalmente, países tem realidades diferentes quanto ao uso e disponibilidade de antimicrobianos e, por isso, a definição destas categorias pode ser relativa. Assim, destaca-se que talvez seja necessário revisar estas definições ou trabalhar com definições personalizadas para cada país.

1.4 O patógeno *Klebsiella pneumoniae* e sua relevância mundial

No contexto global de multirressitência, destaca-se um grupo de bactérias denominado "ESKAPE", que compreende as espécies *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumanii*, *Pseudomonas aeruginosa* e *Enterobacter* spp. As bactérias deste grupo são extremamente plásticas e notoriamente conhecidas pela sua alta capacidade de adquirir fenótipo MDR, o que as permite serem classificadas como uma das maiores ameaças à saúde pública mundial (Pendleton et al., 2013; Theuretzbacher, 2017; Ghorbani et al., 2023).

Dentre as espécies deste grupo, ressaltaremos a bactéria *Klebsiella pneumoniae*, espécie destacada como patógeno de prioridade crítica pela Organização Mundial da Saúde (OMS) em 2024[2] e, foco deste trabalho.

Esta espécie é considerada ubíqua em termos geográficos e de nichos, sendo en-

[2] https://www.who.int/publications/i/item/9789240093461

contrada em forma livre no ambiente e comensal e patogênica em humanos, onde podem ser encontradas em diferentes partes do corpo como trato urinário, respiratório, fígado, e outros (Wyres e Holt, 2016). Esta espécie é uma das principais causas de infecções relacionadas à assistência à saúde (IRAS) no mundo, sendo o trato urinário o sítio de infecção mais comum (Wyres e Holt, 2016; Walsh et al., 2023; Miller e Arias, 2024).

Com um genoma de aproximadamente 5,5 Mb e por volta de 5.500 genes, a plasticidade da espécie *Klebsiella pneumoniae* é evidenciada pelo seu pangenoma aberto, que indica um vasto "pool gênico" com apenas aproximadamente 35% de seus genes considerados "core" (Wyres e Holt, 2016). Um estudo populacional global da espécie baseado no genoma "core" ressalta a alta variabilidade da *K. pneumoniae*, incluindo a observação de altos níveis de diversidade em uma mesma região (Heng et al., 2024).

A resistência a antimicrobianos na espécie é movimentada majoritariamente de forma horizontal através de plasmídeos (Lee et al., 2022). Além disso, isolados desta espécie são capazes de carrear múltiplos plasmídeos simultaneamente, permitindo o rápido desenvolvimento de diferentes fenótipos, particularmente de multirresistência (Wyres e Holt, 2016; Heng et al., 2024; Miller e Arias, 2024).

Somado a isso, enfrenta-se também o problema do acúmulo de genes de virulência em *Klebsiella pneumoniae* e o consequente surgimento de linhagens altamente virulentas (Wyres et al., 2019; Feng et al., 2023). Nos próximos parágrafos, abordaremos um pouco mais a fundo este contexto de resistência e virulência da espécie, e o seu cenário no Brasil.

1.4.1 Bases moleculares da virulência

A virulência é definida como a medida relativa do grau de dano que um microrganismo é capaz de causar ao hospedeiro (Pirofski e Casadevall, 2012). Existem diversos componentes que contribuem para este fenótipo, desde pequenas moléculas até complexos proteicos, que são importantes para a progressão da infecção, principalmente em ambientes hospitalares Riwu et al. (2022).

Por isso, as bases moleculares da virulência da espécie *Klebsiella pneumoniae* são um foco mundial, com muitos autores trabalhando concomitantemente para sua caracterização e compreensão. Tipicamente associadas a IRAS, a bactéria *K. pneumoniae* possui

diversos fatores de virulência muito importantes para a progressão da infecção. São eles: a cápsula bacteriana, o lipopolissacarídeo (LPS), os sideróforos, as fímbrias, as bombas de efluxo e o sistema de secreção tipo VI (T6SS) (Martin e Bachman, 2018; Gorrie et al., 2022).

A ilustração dos autores Paczosa e Mecsas (2016) contida na Figura 2 apresenta talvez um dos esquemas mais difundidos e estabelecidos das principais características e fatores de virulência geralmente encontrados em linhagens *Klebsiella pneumoniae* virulentas e hipervirulentas. Uma descrição detalhada destes fatores de virulência será apresentada a seguir.

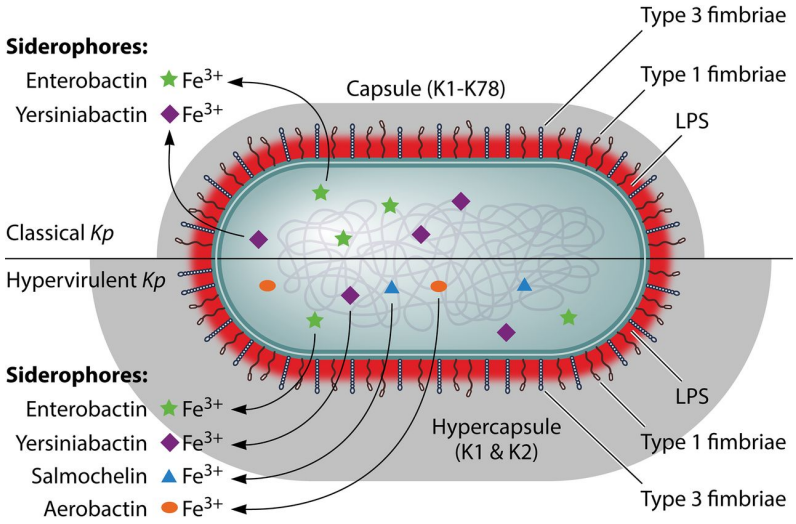

Figura 2: Esquema dos quatro fatores de virulência melhor caracterizados em linhagens *Klebsiella pneumoniae* clássicas e hipervirulentas. São eles: a cápsula, o LPS, as fímbrias e os sideróforos. As principais diferenças propostas são a presença de hipercápsula mais frequentemente encontradas em linhagens hipervirulentas, além da menor variedade de serotipos K encontrados e; a produção de múltiplos sideróforos, e especialmente dos sideróforos salmochelina e aerobactina que são raramente encontrados em linhagens clássicas. Fonte: (Paczosa e Mecsas, 2016).

A cápsula bacteriana (CPS), sintetizada pelo lócus *cps*, protege a bactéria de mecanismos de defesa do organismo hospedeiro, como por exemplo, a fagocitose por macrófagos (Cortés et al., 2002; Miller e Arias, 2024). Tradicionalmente, podem ser diferenciadas por variantes alélicas dos genes *wzi* e *wzc*, genes bastante conservados no lócus *cps* (Martin e Bachman, 2018). Diferenças na sequência dos mesmos permitem classificar a bactéria em 77 sorotipos, nomeados de K1 a K77. Mais recentemente, introduziu-se uma nova técnica de sorotipagem do lócus baseada nas combinações de presença e ausência de genes me-

nos conservados do lócus, permitindo uma classificação de maior resolução em diferentes K-loci (KLs), nomeados a partir do identificador KL101 (Wyres et al., 2016).

O lipopolissacarídeo (LPS) refere-se a um conjunto de moléculas presentes na membrana externa de bactérias Gram-negativas formadas por um lipídeo A conservado, um oligossacarídeo e um antígeno O variável, que é um polissacarídeo. Além de contribuir para a estabilidade estrutural da bactéria, estas moléculas apresentam função importante ao auxiliar na resistência à ação do sistema complemento do sistema imune do hospedeiro (Cortés et al., 2002; Martin e Bachman, 2018; Zhu et al., 2021). O LPS é considerado uma endotoxina e notoriamente reconhecido por mediar o choque séptico causado por bactérias. Sua sorotipagem em bactérias é realizada por variações no antígeno O. Sabe-se hoje que mais de 80% das infecções causadas por *Klebsiella* são provenientes dos sorotipos O1, O2, O3 e O5 (Cortés et al., 2002; Martin e Bachman, 2018; Choi et al., 2020; Zhu et al., 2021; Wantuch et al., 2023). Alguns estudos mostram que certos sorotipos estão relacionados a uma maior virulência e letalidade (Zhu et al., 2021).

O ferro é um elemento crucial para os processos metabólicos da célula bacteriana e, durante a infecção, bactérias tem acesso limitado à esta molécula devido a sua concentração limitada em fluidos extracelulares do hospedeiro (Zhu et al., 2021). Devido à alta afinidade que possuem ao ferro, os sideróforos são essenciais para as bactérias durante infecções, auxiliando-as na captação de ferro, garantindo suas necessidades fisiológicas (Paczosa e Mecsas, 2016; Zhu et al., 2021). A espécie *Klebsiella pneumoniae* pode codificar diferentes sideróforos, com variados níveis de afinidade ao ferro. São eles, a Enterobactina (Ent), Salmochelina (Sal), Aerobactina (Aer) e Yersiniabactina (Ybt) (Paczosa e Mecsas, 2016; Zhu et al., 2021). Dentre elas, a Enterobactina é intrínseca e é codificada por todas as bactérias da espécie *Klebsiella pneumoniae*. Enquanto isso, considerados não universais, os outros sideróforos são geralmente associados à fenótipos de virulência aumentada por propiciarem um "fitness" aumentado durante a infecção (Martin e Bachman, 2018; Paczosa e Mecsas, 2016; Zhu et al., 2021; Russo et al., 2021; Gorrie et al., 2022).

Durante a infecção, bactérias necessitam garantir aderência à superfícies por diversos motivos, como locomoção, formação de microcolônias, biofilme, entre outros. Nesse contexto, as fímbrias, ou *pili*, são moléculas cruciais para a infecção bacteriana (Paczosa e

Mecsas, 2016; Zhu et al., 2021). Destaca-se, em *K. pneumoniae*, dois tipos de fímbria: 1 e 3. Fímbrias do tipo 1 (*fim*), expressas quase que universalmente por isolados de *Klebsiella pneumoniae*, são essenciais para a infecção pois proporcionam esta habilidade de aderir a superfícies bióticas e abióticas (Paczosa e Mecsas, 2016). Por outro lado, particularmente encontradas em enterobactérias, as fímbrias do tipo 3 (*mrk*) são cruciais para a formação de biofilme (Stahlhut et al., 2013; Paczosa e Mecsas, 2016; Martin e Bachman, 2018; Zhu et al., 2021). Vale ressaltar toda a sinergia que ocorre entre estes fatores de virulência na bactéria, uma vez que a atividade da fímbria tipo 3 e a formação de biofilme são influenciadas pela disponibilidade de ferro (Zhu et al., 2021).

Bombas de efluxo, apesar de serem importantes mecanismos de resistência como abordado na Subsubseção 1.2.1.1, são também importantes para a virulência da bactéria. Certas bombas de efluxo, como a bomba AcrAB são consideradas também fatores de virulência pois auxiliam na sobrevivência durante a infecção bacteriana através da extrusão de peptídeos antimicrobianos do hospedeiro (Paczosa e Mecsas, 2016; Martin e Bachman, 2018).

Bactérias normalmente produzem toxinas e moléculas efetoras que estão associadas, por exemplo, à competição com outras bactérias em situações de estresse, ou à sua virulência durante uma infecção. Para que façam uso deste arcabouço, contam com um complexo com múltiplas proteínas, denominado sistema de secreção tipo VI (T6SS), que funciona como uma nano-seringa que serve como canal injetor destas moléculas em outras células (Liu et al., 2017; Martin e Bachman, 2018; Zhu et al., 2021). Ao permitir esta injeção, este sistema garante a capacidade de hidrolisar a parede celular, membrana ou ácidos nucleicos de outras células. É um sistema que reage a situações de estresse, e pode ser, por exemplo, induzido positivamente na presença de antimicrobianos (Liu et al., 2017; Martin e Bachman, 2018; Zhu et al., 2021). Alguns estudos sugerem que o T6SS pode conferir vantagens para a sobrevivência de linhagens hipervirulentas (Zhu et al., 2021).

1.4.2 A hipervirulência em Klebsiella pneumoniae

Descrito pela primeira vez na década de 1980 no sudeste asiático (Catalán-Nájera et al., 2017), linhagens hipervirulentas de *Klebsiella pneumoniae* (hvKp) representam uma

preocupante parcela da população desta espécie. Contrastando com linhagens "clássicas" associadas ao ambiente hospitalar e consideradas unicamente oportunistas, as linhagens hipervirulentas são capazes de causar infecções de alto risco de vida em indivíduos adultos saudáveis adquiridas em comunidade (fora de ambientes hospitalares) (Marr e Russo, 2019; Zhu et al., 2021; Ali et al., 2023; Heng et al., 2024). Dentre elas, destacam-se o abcesso piogênico do fígado, endoftalmite, meningite e bacteremia. Além disso, infecções causadas por hvKp são capazes de se difundir em múltiplos sítios, como cérebro e olhos (Catalán-Nájera et al., 2017; Zhu et al., 2021).

Durante muito tempo, a hipervirulência foi caracterizada pela identificação do fenótipo de hipercápsula, através do "String test" positivo. Consequentemente, esta prática, induziu pesquisadores a considerar o fenótipo de hipermucoviscosidade junto com alguns sorotipos capsulares específicos como um marcador da hipervirulência (Paczosa e Mecsas, 2016; Zhu et al., 2021; Ali et al., 2023). Porém, com os crescentes relatos de "excessões" a regra, entende-se hoje que apesar de potencializar a virulência e ser frequentemente identificado em linhagens hvKp, a hipermucoviscosidade é um fenótipo a parte e, independente da hipervirulência, além de não ser exclusivo dos sorotipos capsulares K1 e K2 como se acreditava Figura 2 (Chuang et al., 2013; Liu et al., 2014; Cubero et al., 2016; Luo et al., 2014; Cubero et al., 2016; Wu et al., 2017; Catalán-Nájera et al., 2017; Zhu et al., 2021; Gorrie et al., 2022; Ali et al., 2023).

Em resumo, este fenótipo hipervirulência é muito mais complexo do que se pensava, e está associado à diversos fatores diretamente relacionados ao genoma acessório da espécie (Catalán-Nájera et al., 2017; Martin e Bachman, 2018; Zhu et al., 2021; Gorrie et al., 2022; Ali et al., 2023). Em razão disto, desde 2019, colaborações internacionais têm sido realizadas com o intuito de estabelecer novos biomarcadores do fenótipo e melhores protocolos para identificação de hipervirulência (Marr e Russo, 2019; Zhu et al., 2021; Heng et al., 2024).

Como resultado destes diversos estudos, atualmente temos um leque mais atualizado de biomarcadores deste fenótipo. No geral, utiliza-se hoje para identificação do fenótipo a presença dos genes *iuc* (aerobactina), *iro* (salmochelina), *rmpA* e/ou *rmpA2* em plasmídeos (Marr e Russo, 2019; Russo e Marr, 2019; Zhu et al., 2021; Ali et al., 2023). Por exemplo, sabe-se que apesar destas linhagens serem capazes de produzir todos

os diferentes sideróforos, a aerobactina parece ser o biomarcador que desponta devido à sua alta contribuição para virulência em condições experimentais, além de ser prevalente em hvKp, sendo encontrada em ≈90% destas linhagens (Zhu et al., 2021).

De todo modo, o trabalho de catalogação de biomarcadores de hipervirulência e desenvolvimento de técnicas para a correta identificação de linhagens de alto risco é contínuo e seu constante aprimoramento é de extrema importância para a saúde pública global (Marr e Russo, 2019; Russo e Marr, 2019; Zhu et al., 2021; Gorrie et al., 2022; Ali et al., 2023; Heng et al., 2024).

1.4.3 A multirresistência em Klebsiella pneumoniae

Linhagens de *Klebsiella pneumoniae* multirresistentes a antimicrobianos (MDR-Kp) são globalmente conhecidas por causarem infecções severas e de alta mortalidade (Gorrie et al., 2017; Bengoechea e Pessoa, 2018; Miller e Arias, 2024). Este problema é ainda mais acentuado pelo fato de que muitos destes isolados são produtores de carbapenemases (CpKp, do inglês "Carbapenem-producing *K. pneumoniae*"), enzimas capazes de hidrolisar um amplo espectro de fármacos β-lactâmicos, como penicilina, aztreonam, cefalosporinas e carbapenemas. Somado a isso, observa-se um incremento anual na prevalência de isolados produtores de carbapenemases (Palmeiro et al., 2019; Calderaro et al., 2021; Dos Santos et al., 2024; Cuicapuza et al., 2024).

Entre as carbapenemases, as predominantemente detectadas nestes isolados são a KPC, NDM e OXA. Por isso, estas três enzimas constituem importantes marcadores do fenótipo de resistência estendida (XDR, do inglês "extensively drug resistant") nestes microrganismos (Theuretzbacher et al., 2021). Dentre elas, os genes da família NDM, observados em *K. pneumoniae* pela primeira vez em 2009 (Yong et al., 2009), são de bastante relevância clínica devido ao seu alto espectro de resistência e a velocidade de disseminação.

A NDM é uma metalo-β-lactamase classe B que, além de apresentar uma rápida disseminação por todo o mundo, confere resistência a todas as drogas β-lactâmicas, exceto o aztreonam (Boyd et al., 2020). A transferência horizontal de genes tem papel fundamental na disseminação desta família gênica, e ela ocorre principalmente por plasmídeos de

nove grupos de incompatibilidade ("Inc type") específicos, como o IncX3, IncFIB, IncI1, entre outros (Dong et al., 2022). Somado a isso, plasmídeos carreadores de NDM geralmente codificam múltiplos genes de resistência, aumentando o risco de co-transmissão (Dong et al., 2022).

Em termos de similaridade genética, linhagens de *K. pneumoniae* pertencentes ao grupo clonal 258 (CG258) são prevalentes mundialmente e intimamente associadas à disseminação de resistência (Wyres et al., 2015). Membros deste grupo tipicamente carreiam plasmídeos contendo β-lactamases de espectro estendido (ESBLs, do inglês "extended spectrum β-lactamases") e carbapenemases classe A, como CTX-M e KPC. Por isso, são geralmente considerados de alta relevância clínica pelo seu potencial de desencadear surtos e por sua ligação com o surgimento de linhagens XDR e até mesmo pan-resistentes (PDR) (Wyres et al., 2015; Lee et al., 2016, 2017; Wang et al., 2018; van Dorp et al., 2019; Li et al., 2022; Nakamura-Silva et al., 2022).

Este grupo CG258 é um grande complexo clonal de bactérias que compreende diversas subdivisões baseadas em tipagem de sequência por múltiplos lócus (MLST). O MLST é talvez uma das técnicas mais difundidas, baseada na identificação de diferentes combinações de alelos de certos genes essenciais (*housekeeping*), que permite a definição de perfis alélicos denominados "Sequence type" (ST) (Lee, 2017). Dessa forma, o CG258 representa um conglomerado de diferentes STs como o ST11, ST14, ST17, ST20, ST29 e outros (De Campos et al., 2018; Nakamura-Silva et al., 2021).

Entre estes diferentes STs, destaca-se as linhagens ST11, um grupo de *K. pneumoniae* MDR de alto risco, produtoras de CTX-M-15 ou CTX-M-14 e de alto sucesso de disseminação, principalmente na Ásia e América do Sul (Qi et al., 2011; D'Andrea et al., 2014; He et al., 2022). A determinação exata do surgimento deste ST é bastante desafiadora pois o esquema de MLST para *Klebsiella pneumoniae* só foi introduzido em 2005 (Diancourt et al., 2005). Todavia, é proposto que os membros do CG258 são linhagens descendentes de um ancestral comum, hoje ST11, que subsequentemente diversificou em diversas linhagens através de recombinações (Wyres et al., 2015).

Atualmente, o ST11 é considerado como um dos três STs de *K. pneumoniae* MDR mais difundidos globalmente e considerado, na China, como o mais perigoso no país (He et al., 2022). No Brasil, um dos primeiros relatos de *K. pneumoniae* ST11 ocorreu em

2006, com isolados de Pernambuco e Minas Gerais (Seki et al., 2011). Apesar de incerto, é possível que este possa ser também o primeiro relato deste ST na América Latina, uma vez que outros membros do grupo CG258 começaram a ser relatados na região por volta do mesmo período (Munoz-Price et al., 2013).

1.4.4 Convergência de genes de resistência e virulência

Em *Klebsiella pneumoniae*, a plasticidade do genoma acessório é de extrema relevância para o surgimento de linhagens de alto risco, contendo ambos os fenótipos de hipervirulência e multirresistência (Holt et al., 2015; He et al., 2022; Liu et al., 2022; Shelenkov et al., 2023). Estas linhagens apresentam taxa de mortalidade muito maior que linhagens "clássicas" e, por isso, impõem um tremendo desafio para a saúde pública global. Não só uma preocupação, esta convergência de fenótipos já é uma realidade com o número de isolados XDR-hvKp aumentando constantemente a cada ano devido principalmente à movimentação de plasmídeos (Holt et al., 2015; Lee et al., 2017; Navon-Venezia et al., 2017; Gu et al., 2018; Lam et al., 2021; Li et al., 2021; Zhu et al., 2021; Liu et al., 2022; He et al., 2022; Liu et al., 2022; Shelenkov et al., 2023).

A convergência destes fenótipos acontece de duas maneiras. Linhagens hvKp adquirem elementos genéticos móveis (plasmídeos, transposons, etc.) contendo genes de resistência. Ou, quando linhagens multirresistentes adquirem plasmídeos de virulência (Russo e Marr, 2019; Zhu et al., 2021; He et al., 2022). Acredita-se que a taxa de aquisição de plasmídeos de virulência por clones MDR seja maior que a de aquisição de plasmídeos de resistência por clones hipervirulentos (Wyres et al., 2019). Barreiras intrínsecas como a incompatibilidade entre plasmídeos e a super expressão da cápsula bacteriana podem ter papel importante nesta observação (Russo e Marr, 2019; Zhu et al., 2021). Por isso, mesmo que ambos os fluxos de movimentação de genes possam ocorrer, neste contexto, as linhagens MDR representam um alerta maior devido a maior probabilidade de convergência dos fenótipos nestas bactérias.

1.4.5 Monitoramento de linhagens de alto risco no Brasil

Devido ao seu impacto econômico e desafios impostos para a saúde pública global, o monitoramento epidemiológico de linhagens multirresistentes e de virulência aumentada é essencial. No Brasil, linhagens CG258 representam uma grande parte dos relatos associados a IRAS (Azevedo et al., 2019; Aires et al., 2019; Nakamura-Silva et al., 2022). Como discorrido na Subseção 1.4.3, linhagens deste complexo clonal são frequentemente associadas à movimentação de diversos genes de resistência na América do Sul, incluindo genes de alta relevância como o *bla*NDM.

No geral, entre todos os genes de resistência difundidos e detectados em genomas de *Klebsiella pneumoniae* isoladas no Brasil, podemos destacar dentre os mais frequentes, os genes de resistência a β-lactâmicos ($bla_{KPC\text{-}2}$, $bla_{SHV\text{-}11}$, $bla_{OXA\text{-}1/2}$, $bla_{TEM\text{-}1}$, $bla_{CTX\text{-}M\text{-}15}$), fluoroquinolonas (*oqxAB*, *aac(6')lb-cr*), fosfomicinas (*fosA5/6*), sulfonamidas (*sul1*), aminoglicosídeos (*aac(3)-Ila*), trimetropina (*dfrA*) e tetraciclinas (*tetA*) (Azevedo et al., 2019; Aires et al., 2019; Palmeiro et al., 2019; Andrey et al., 2019; Longo et al., 2019; Nakamura-Silva et al., 2022). Apesar de menos frequente, provavelmente devido a sua introdução recente, a detecção de genes *bla*NDM é bastante relevante devido ao vasto espectro de resistência que confere à bactéria como discutido anteriormente (Subseção 1.4.3).

Na América do Sul, o primeiro relato do gene *bla*NDM ocorreu em 2013, na Colômbia, onde autores descreveram um surto de *Klebsiella pneumoniae* produtoras de NDM-1 em uma unidade neonatal (Escobar Pérez et al., 2013). Já no Brasil, o primeiro relato do gene também ocorreu em 2013, porém em um isolado de *Providencia rettgeri* (Carvalho-Assef et al., 2013), mas, logo depois, já encontrado também em *Klebsiella pneumoniae* (Nava et al., 2019). Já entre 2017 e 2018, foi relatado o primeiro surto hospitalar causado por *Klebsiella pneumoniae* produtoras de NDM (Monteiro et al., 2019). Atualmente, como demonstrado por Camargo et al., Enterobacterales produtoras de NDM são detectadas em todas as regiões administrativas do Brasil (Camargo et al., 2022).

Em seu estudo, Camargo et al. ressaltam que o gene *bla*NDM é detectado, majoritariamente, em plasmídeos transferíveis dos grupos de incompatibilidade IncF e IncX3 (Camargo et al., 2022). Estudos recentes têm demonstrado que plasmídeos IncX3 são extremamente relevantes no contexto do carreamento e disseminação de genes de resistência

devido à sua alta estabilidade, baixo impacto no "fitness" bacteriano e eficiente habilidade de conjugação (Guo et al., 2022).

Um pouco mais além, um estudo com isolados de Enterobacterales de 2015 a 2022, incluindo *K. pneumoniae*, ressaltou que as espécies dessa ordem podem estar passando por um processo de mudança na frequência detectada de seus genes de resistência (Kiffer et al., 2023). Baseado nos resultados das frequências de observação dos genes, os autores discorrem que as linhagens parecem estar migrando do gene *bla*KPC para o *bla*NDM, o que ressalta a notória propagação deste gene nos últimos anos (Kiffer et al., 2023). E ainda, corroborando com estudos anteriores (Camargo et al., 2022), os autores sugerem que plasmídeos possuem papel chave no aumento da disseminação do gene *bla*NDM.

Além disso, linhagens pan-resistentes do grupo CG258, incluindo resistência à drogas de último recurso, como a colistina, já foram reportados em diferentes localidades no Brasil (Longo et al., 2019; Conceição-Neto et al., 2022; Fonseca et al., 2023; Gomes et al., 2023). Em paralelo a todo este cenário preocupante de multirresistência, a disseminação de genes de virulência também tem aumentado bastante e, consequentemente, também os relatos de linhagens MDR com virulência aumentada. Incluindo linhagens com fenótipo de hipermucoviscosidade mediada por *rmpA* (biomarcador de hipervirulência), associadas à altas taxas de mortalidade (Azevedo et al., 2019; Ferreira et al., 2019; Andrey et al., 2019; Aires et al., 2019; Nakamura-Silva et al., 2022).

Todo esta conjuntura reitera alertas quanto ao surgimento de linhagens multirresistentes capazes de causar infecções na comunidade, de alta morbidade, mortalidade, e talvez intratáveis. No geral, demonstra-se a necessidade de estudos epidemiológicos de monitoramento e caracterização continuado, como já realizado por alguns grupos (Aires et al., 2019; Azevedo et al., 2019; Gomes et al., 2023).

1.5 A genômica de bactérias patogênicas

O monitoramento continuado de patógenos bacterianos é imprescindível dada à constante evolução destes. Estratégias experimentais para sua detecção e categorização são constantemente aprimoradas e compreendem diversas abordagens. Por exemplo, a cultura de amostras clínicas, a caracterização fenotípica por testes bioquímicos e mor-

fológicos, variações da técnica de PCR (*Polymerase Chain Reaction*), MLST (*multilocus sequence typing*), entre outras (Fournier et al., 2013).

O esquema MLST para *Klebsiella pneumoniae* foi estabelecido em 2005 e é disponibilizado através do banco de dados BIGSdb Pasteur[3] (Diancourt et al., 2005). Além de *Klebsiella pneumoniae*, este banco de dados também possui esquemas MLST para outras bactérias, como *Escherichia coli*, *Bordetella*, entre outros.

Apesar de sua utilidade na caracterização de populações de patógenos, em nível regional ou global, o MLST sozinho não tem resolução suficiente para o estudo e compreensão de sua estrutura genética (Heng et al., 2024). Desta forma, há um demanda por novas abordagens e, neste contexto, a genômica desponta como uma forte aliada.

Em contraste com técnicas experimentais restritas a poucos lócus, como o PCR e MLST, a genômica permite a determinação completa do catálogo de genes, facilitando a investigação de diversos aspectos biológicos, como a identidade do patógeno, dinâmica evolutiva, rotas de introdução e dispersão, bem como seu potencial metabólico (Fournier et al., 2013; Lee, 2017; Djaffardjy et al., 2023; Baker et al., 2023; Djordjevic et al., 2023; Heng et al., 2024).

A possibilidade de obter genomas bacterianos completos proporciona uma visão mais abrangente dos componentes e mecanismos que contribuem para a resistência a fármacos, permitindo uma identificação mais precisa dos genes envolvidos (Djordjevic et al., 2023; Baker et al., 2023; Castañeda Barba et al., 2023). Ao explorar os detalhes do genoma bacteriano, abre-se um leque de estratégias mais eficazes para caracterizar e combater a resistência aos antimicrobianos (Baker et al., 2023; Wheeler et al., 2023).

De fato, a genômica já tem sido efetivamente utilizada para a investigação e controle de surtos e diagnósticos clínicos. Talvez uma das provas atuais mais marcantes de seu uso tenha sido durante a pandemia do vírus SARS-CoV-2 a partir de 2020 (Tosta et al., 2023; Fokam et al., 2023). Destaca-se que, tudo isto só se tornou possível devido aos diversos avanços nas técnicas de sequenciamento de DNA nos últimos anos.

Estes avanços permitiram o estabelecimento da genômica de populações, fornecendo um inventário de ferramentas inédito para a detecção e o estudo populacional de

[3] `https://bigsdb.pasteur.fr/klebsiella/klebsiella.html`

variações genéticas com resolução em nível de bases individuais (Baker et al., 2023; Wheeler et al., 2023; Heng et al., 2024). Neste contexto, diversos estudos têm demonstrado a relevância de experimentos de genômica em larga escala para evidenciar fenômenos importantes.

1.5.1 Genômica de populações

Em 2017, Manson et al. descreveram um estudo de genômica comparativa utilizando 5.310 genomas de *Mycobacterium tuberculosis* dos cinco continentes para investigar a dinâmica evolutiva de linhagens multirresistentes (Manson et al., 2017). Foi detectado que linhagens MDR apresentam majoritariamente resistência a droga isoniazida antes de qualquer outra e, por isso, desenvolveram um diagnóstico para identificar linhagens monoresistentes a isoniazidas como uma estratégia de prevenção ao surgimento de bactérias MDR (Manson et al., 2017).

Ainda no mesmo ano, Goldstone e Smith estudaram 4.022 genomas de *E. coli* e puderam detectar resistência intrínseca a sete classes de antimicrobianos. Além disso, detectaram 118 combinações de genes de resistência a fármacos que nunca ocorriam, à época, simultaneamente em nenhum dos genomas, representando possibilidades terapêuticas para o combate de fenótipos MDR (Goldstone e Smith, 2017).

No ano seguinte, em 2018, Lam et al. investigaram a prevalência, evolução e mobilidade do gene *ybt* em *Klebsiella pneumoniae* (Lam et al., 2018). Identificou-se a presença de um elemento genético móvel, ICEkp, em um terço da população da espécie, sugerindo a plasticidade fenotípica através de eventos de transferência horizontal de genes. Também estabeleceu-se que o gene *ybt* é comumente transmitido através de plasmídeos IncFIB$_k$, conhecidos por co-associar cassetes de resistência e virulência em um único elemento genético de alta estabilidade.

Já em 2019, Wyres et al. compararam >2.200 genomas de *Klebsiella pneumoniae* e caracterizaram os 28 grupos clonais mais comuns da espécie (Wyres et al., 2019). Determinaram que clones MDR apresentam maior risco à saúde pública devido à sua maior taxa de recombinação, ressaltando a necessidade do estabelecimento de sistemas de vigilância genômica para o rastreamento e combate destes clones.

Paralelamente, ainda em 2019, Roe et al. estudaram 107 isolados de *Acinetobacter baumanii* e constataram que diversos genomas analisados haviam sido erroneamente classificados como *A. baumanii*. Demonstraram também que em patógenos de extrema plasticidade, cujos genes de resistência não são conservados entre as linhagens, o uso de dados transcritômicos são essenciais para viabilizar o estudo dos mecanismos de resistência na espécie (Roe et al., 2019).

Alguns anos mais tarde, já em 2022, pesquisadores realizaram uma análise genômica epidemiológica retrospectiva de 420 linhagens de *Klebsiella pneumoniae* produtoras de carbapenemases dos anos de 2009 a 2017, provenientes de 70 diferentes hospitais na China (Li et al., 2022). No estudo, detectaram que o ST prevalente era o ST11, e que a resistência a carbapenêmicos era conferida principalmente pelo gene *bla*KPC, codificado majoritariamente em plasmídeos de 5 grupos de incompatibilidade diferentes. Identificaram também certas combinações entre serotipo e plasmídeos que pareciam garantir uma maior vantagem fenotípica, facilitando sua disseminação no país. Por fim, baseado nos dados genômicos obtidos foram capazes de sugerir tratamentos com combinações de drogas específicas para isolados CG258 e não-CG258.

Ainda em 2022, pesquisadores analisaram 21 isolados de *Klebsiella pneumoniae* do grupo clonal de risco CG258, selecionados entre 2014 a 2016 de quatro hospitais da cidade de Manaus-AM (Nakamura-Silva et al., 2022). A maioria das linhagens foram classificadas como multirresistentes além de oito isolados classificados como hipermucoviscosos. Os autores ressaltam um cenário preocupante observado na cidade que pode ter sido potencializado pelo colapso do sistema público causado na cidade devido a pandemia do COVID-19 em 2020, onde diversos pacientes de diferentes regiões foram admitidos na cidade e submetidos a intensas terapias com antimicrobianos.

Em 2023, Roy Chowdhury et al. realizaram uma análise filogenômica de uma coleção global contendo 925 genomas de *Escherichia coli* ST38, provenientes de 38 países. No estudo, observaram que o co-carreamento dos genes *fyuA* e *irp2* servia como um bom indicador da presença da ilha de hipervirulência de *Yersinia*, uma vez que fora observado em mais de 60% dos genomas. Além disso, detectaram indicativos de eventos de transmissão entre hospedeiros, e entre hospedeiro/ambiente (Roy Chowdhury et al., 2023).

Recentemente, já em 2024, no Reino Unido, Wan et al. realizaram uma análise integrada de genômica e dados de movimentação de pacientes (locais e horários das enfermarias) para investigar uma possível cadeia de transmissão. Os autores foram capazes de demonstrar um surto mediado por plasmídeos, envolvendo múltiplas espécies de Enterobacterales resistentes a carbapenemas e colistina. Não só isso, identificaram enfermarias de alto risco e possíveis rotas de potencial transmissão cruzada, permitindo uma avaliação de risco para evitar tais eventos (Wan et al., 2024).

Hoje, com todos os exemplos disponíveis, é indubitável que a epidemiologia genômica assume um papel de destaque nos esforços para investigar a evolução da resistência e virulência bacteriana, e caracterizar e compreender os mecanismos de resistência, proporcionando dados para o desenvolvimento de terapias mais eficientes (Hendriksen et al., 2019; Kan et al., 2018; Carter et al., 2022; Stockdale et al., 2022; Baker et al., 2023; Djordjevic et al., 2023). Somado a isso, o estabelecimento de redes de vigilância genômica e diagnóstico rápido podem se tornar pilares para esforços de contenção de surtos de doenças infecciosas em estágios precoces de modo a diminuir seus impactos na saúde pública global (Chiu e Miller, 2019). Desta forma, é notório que estas abordagens genômicas torna-se-ão ainda mais frequentes e, hoje, pode-se dizer que o maior gargalo não se encontra mais na geração de dados, mas sim no processamento, armazenamento e integração destes dados (Muir et al., 2016; Stockdale et al., 2022; Baker et al., 2023).

1.6 Processamento e interpretação de dados genômicos

Presenciou-se nos últimos anos uma revolução nas tecnologias de sequenciamento de DNA de nova geração (NGS), reduzindo seu custo exponencialmente e aumentando a qualidade e amplitude das técnicas. Isto transformou a genômica, e tem gradualmente tornado o sequenciamento em um serviço mais acessível para pesquisadores, democratizando o acesso à técnica. Entre as plataformas de sequenciamento atuais, destacam-se como talvez as mais importantes devido sua ampla adoção, as plataformas Illumina, Pacbio e Oxford Nanopore (Koren e Phillippy, 2015).

Em qualquer estudo genômico, estes dados de sequenciamento só podem ser interpretados e biologicamente contextualizados através da bioinformática (Djaffardjy et al.,

2023; Baker et al., 2023). Para isso, diversas ferramentas computacionais são empregadas, sequencialmente, de modo a resolver uma cadeia de etapas analíticas. Atualmente, estas análises bioinformáticas consomem consideravelmente mais tempo e recursos humanos do que o próprio processo de sequenciamento. Além disso, é comum que essas análises sejam subestimadas nos cálculos totais de custos, apesar de representarem uma parcela significativa dos mesmos (Muir et al., 2016; Baker et al., 2023).

Por isso, ao mesmo tempo que as novas técnicas de sequenciamento democratizam a genômica e permitem a sua aplicação regular em laboratórios de microbiologia, também introduzem uma lacuna quanto à análise efetiva dos dados (Muir et al., 2016; Stockdale et al., 2022; Djaffardjy et al., 2023; Baker et al., 2023).

1.6.1 Protocolos computacionais automatizados

Em uma análise de bioinformática, o conceito de protocolo computacional (*pipelines*) é definido pelo fluxo de dados em uma sequência de etapas necessárias para a conversão de dados brutos em resultados analíticos, através do encadeamento de programas (Grüning et al., 2017).

No início, era comum utilizar-se de *scripts* em linguagens de programação para a implementação de *pipelines*. Porém, estes geralmente careciam de pontos importantes para a reprodutibilidade e escalonamento de análises, que são: a abstração de dependências e a habilidade de recomeçar a análise a partir de pontos de interrupção sem a necessidade de reexecutar etapas bem-sucedidas (Leipzig, 2016).

Nos últimos anos, diversos sistemas especializados e automatizados de gerenciamento de *pipelines* que apresentam soluções para estes problemas têm sido desenvolvidos (Tabela 1) e, cabe ao pesquisador, definir qual sistema melhor se enquadra sob suas demandas (Strozzi et al., 2019; Djaffardjy et al., 2023). Estes sistemas fornecem uma sintaxe padronizada e inteligível das etapas a serem executadas de acordo com as dependências de dados entre estes (Perkel, 2019). Ao mesmo tempo, permitem a tolerância de erros (algo bastante frequente em protocolos computacionais) e reinício no ponto de interrupção, proporcionando *pipelines* robustos e escalonáveis (Leipzig, 2016; Perkel, 2019; Djaffardjy et al., 2023). Todas essas características são fatores essenciais para garantir a reprodu-

tibilidade e portabilidade de análises bioinformáticas, o que viabiliza e potencializa os estudos genômicos.

Tabela 1 - Exemplos de alguns dos principais sistemas de gerenciamento de *pipelines* da atualidade, comparando-os quanto a sua linguagem de programação base e se possui ou não interface gráfica agregada.

Nome	Linguagem	Interface gráfica	Referência
CWL	CWL	Não	10.6084/m9.figshare.3115156.v2
Galaxy	Python	Sim	10.1093/nar/gky379
Nextflow	Groovy	Não	10.1038/nbt.3820
Snakemake	Python	Não	10.1093/bioinformatics/bts480

1.6.2 Tecnologia de contêineres

Apesar de toda a melhoria quanto à implementação de tarefas, a instalação dos programas analíticos requeridos ainda pode ser uma grande barreira para a execução de *pipelines*. Uma das características dos programas de bioinformática é que a grande maioria requer o sistema operacional Linux para sua instalação e execução. A necessidade do conhecimento Linux para instalação e gerenciamento de dependências destas ferramentas traz um problema para sua utilização por usuários não especialistas.

Diante deste cenário, um conceito cada vez mais utilizado é o de contêineres computacionais, que consistem em ambientes isolados que encapsulam programas e suas dependências em um ambiente virtual dentro do sistema operacional do usuário (ex: Windows), mas que são executados transparentemente como se estivessem em um sistema operacional de origem (ex: Linux). Dentre as tecnologias de contêineres mais utilizadas atualmente temos o Docker® (Merkel, 2014) e Singularity (Kurtzer et al., 2017).

Este arcabouço permite que desenvolvedores criem *pipelines* com blocos independentes e autossuficientes dedicados a suas aplicações, que podem ser executados em qualquer sistema operacional. Desta forma, tem-se uma grande melhoria na distribuição de programas, pois a instalação de todo o ambiente de execução fica a cargo do desenvolvedor, que disponibiliza o contêiner para a comunidade (Boettiger, 2015; Grüning et al., 2018).

A distribuição de *pipelines* acoplados a contêineres garante máxima portabilidade (execução em diversas plataformas), além de abstrair o complexo processo de instalação das ferramentas pelo usuário final (Boettiger, 2015). Um exemplo de sistema de gerenciamento automatizado de *pipelines* que integra transparentemente os contêineres é o Nextflow (Di Tommaso et al., 2017), que é descrito a seguir.

1.6.3 Nextflow

Nextflow é uma linguagem de domínio específico (DSL, do inglês "Domain Specific Language") desenvolvida especificamente para a descrição e gerenciamento de *pipelines* (Di Tommaso et al., 2017).

Análises de bioinformática geralmente envolvem a utilização de uma grande quantidade de programas. Cada um destes, possuem suas próprias peculiaridades, arquivos de entrada e arquivos de saída diferentes. O Nextflow, executa cada tarefa em um ambiente isolado, o que permite o fácil gerenciamento dos arquivos requeridos por cada programa e a rastreabilidade de erros.

Este sistema, conta com canais dedicados e independentes para a transmissão de arquivos entre processos (tarefas), o que facilita a execução de programas em paralelo. O Nextflow, aplica o paradigma de fluxo de dados e somente inicia uma tarefa quando seus arquivos de entrada são recebidos e, por isso, um processo nunca será executado enquanto seu arquivo de entrada não for gerado. Assim, é possível criar, de maneira simples, um fluxo de trabalho complexo com tarefas interdependentes em paralelo.

Em função de seu sistema de gerenciamento de trabalhos, Nextflow é capaz de tolerar erros de modo que a execução seja pausada assim que algum problema aconteça. Desta maneira, permite-se que o usuário seja capaz de identificar o local exato do problema, corrigi-lo, e reiniciar o *pipeline* do ponto em que parou, sem que seja necessário refazer as análises bem sucedidas.

Além disso, essa ferramenta é completamente integrada às tecnologias Docker e Singularity, permitindo incorporar rapidamente contêineres computacionais de maneira automática. Nextflow se encarrega de montar, executar e desacoplar estes contêineres no momento de sua execução, tudo feito sem a intervenção do usuário, simplificando e

automatizando a execução do *pipeline* e instalação de suas dependências (Di Tommaso et al., 2017; Djaffardjy et al., 2023).

Como um todo, Nextflow disponibiliza um sistema de gerenciamento de trabalhos que facilita a implementação, comunicação entre processos, paralelização, rastreabilidade de erros e o gerenciamento de *pipelines*.

O Nextflow, e outras ferramentas de desenvolvimento de *pipelines* como Snakemake, oferecem uma forma inovadora para a consolidação de análises bioinformáticas, representando o atual direcionamento do desenvolvimento de ferramentas computacionais na área. Isto porque as tratam como verdadeiros protocolos computacionais que aderem às melhores práticas de gerenciamento de dados científicos, os chamados princípios FAIR (do inglês, "Findable, Accessible, Interoperable, Reusable") (de Visser et al., 2023).

1.7 *Pipelines* de genômica bacteriana

Ao longo dos anos, inúmeros *pipelines* especializados em genômica bacteriana foram implementados de diversas maneiras e com diferentes objetivos, podendo ser divididos quanto às formas de distribuição, sistema operacional requerido, tipo de dado aceito, etc. Além disso, não necessariamente são capazes de realizar todas as etapas de estudos genômicos. Alguns, por exemplo, podem ser bastante especializados em algumas tarefas, como PGAP (Li et al., 2020) e DFAST (Tanizawa et al., 2018), específicos para a anotação de genomas. O primeiro, oferecido pelo NCBI, é bastante utilizado para a anotação automática de genomas submetidos ao banco de dados GenBank. Já o segundo, foi desenvolvido para facilitar a submissão de genomas ao banco de dados DDBJ.

Avaliando-os em linha do tempo, é possível perceber que estes *pipelines* foram adotando as características e tendências de cada época, de modo que se percebe um progresso na robustez das ferramentas, como mostrado na Tabela 2. Atualmente, existem *pipelines* como ASA3P (Schwengers et al., 2019), Bactopia (Petit e Read, 2020) e bacass (Peltzer et al., 2019) que agregam estas novas tendências do desenvolvimento e compartilhamento de *pipelines* através de contêineres e sistemas de gerenciamento de execução. Estes são capazes de montar e anotar genomas utilizando leituras curtas e longas. Além disso, o Bactopia possui uma particularidade que permite o *pipeline* ser executado por completo,

ou somente uma única ferramenta por vez.

No entanto, a grande maioria dos *pipelines* são anteriores a estes conceitos, com poucas opções de escolha dos programas incluídos, ou seja, as análises são pré-determinadas e sem flexibilidade. Além disso, cada *pipeline* foi implementado com diferentes objetivos em mente e, consequentemente, produzindo diversos resultados diferentes, ainda que compartilhem a implementação de algumas mesmas análises. Todas estas diferenças de implementações e restrições transformam a escolha do *pipeline* ideal em um processo laborioso e pouco trivial.

Deste modo, ainda existe espaço para *pipelines* genéricos e de simples execução, que apresentem ao usuário múltiplas escolhas de programas, capazes de produzir diversos resultados significativos para a grande maioria dos cenários analíticos, seja no contexto clínico ou de ecologia microbiana, dentre outros. Finalmente, além de executar uma série de programas, os *pipelines* devem proporcionar ao usuário final um ambiente dinâmico de manipulação e investigação dos resultados de modo a facilitar a interpretação dos dados. Neste quesito, a maioria das soluções de software mostrados na Tabela 2 é deficiente e abre uma lacuna para ser explorada no contexto da genômica bacteriana.

Tabela 2 - Exemplos de alguns dos principais *pipelines* de genômica bacteriana disponíveis atualmente. Estes são categorizados na tabela de acordo com o tipo de tarefa realizada (montagem, anotação ou híbrido) e a sua forma de execução (web ou local). Tarefa "híbrida" indica a habilidade de executar diferentes etapas de montagem e anotação.

Nome do programa	Tipo (tarefa)	Execução	Referência (DOI)
ASA3P	Híbrido	Local	10.1101/654319
bacass	Híbrido	Local	10.5281/zenodo.3574476
BacPipe	Híbrido	Local	10.1016/j.isci.2019.100769
BacSeq	Híbrido	Local	10.3390/microorganisms11071769
Bactopia	Híbrido	Local	10.1128/msystems.00190-20
DFAST	Anotação	Local e web	10.1093/bioinformatics/btx713
MEGAnnotator	Híbrido	Local	10.1093/femsle/fnw049
MICRA	Anotação	Web	10.1186/s13059-017-1367-z
MyPro	Híbrido	Local	10.1016/j.mimet.2015.04.006
PATRIC (RAST)	Híbrido	Web	10.1007/978-1-4939-7463-4_4
PGAP	Anotação	Local e web	10.1093/nar/gkw569
TORMES	Híbrido	Local	10.1093/bioinformatics/btz220

2

Objetivos

2.1 Objetivo geral

Desenvolver protocolos computacionais robustos para padronizar e automatizar o pré-processamento, montagem e anotação de genomas procarióticos e utilizá-los para a análise completa de isolados bacterianos multirresistentes a antimicrobianos.

2.2 Objetivos específicos

- Reformular e aprimorar *pipelines* computacionais genéricos de análise de genomas procarióticos prototipados anteriormente (de Almeida, 2020) para torná-los mais robustos e acessíveis para a comunidade
- Conceitualizar e desenvolver aplicação web para integração completa dos resultados de anotação genômica, para prover uma plataforma visual para exploração e consolidação dos resultados
- Aplicar os *pipelines* para a análise e caracterização de dados de sequenciamento de genomas de isolados multirresistentes de *Klebsiella pneumoniae* do Hospital Universitário de Brasília, visando adquirir maior detalhamento genético e melhor compreender essas bactérias causadoras de infecções no hospital, com foco em genes e características de relevância clínica
- Realizar um estudo de genômica comparativa retrospectiva incluindo outras linhagens brasileiras de *K. pneumoniae* para verificar a dinâmica de aquisição e persistência de genes de resistência e virulência

3

Material e Métodos

3.1 Desenvolvimento dos *pipelines*

Os protocolos computacionais desenvolvidos e apresentados neste trabalho são embasados nos protótipos criados durante meu projeto de Mestrado (de Almeida, 2020). Neste trabalho, reformulou-se completamente todos os *pipelines* de modo a aderir a padrões de qualidade de desenvolvimento de software (Ewels et al., 2020), bem como se adicionou uma série de novas funcionalidades para fornecer uma exploração abrangente de diversos aspectos da genômica bacteriana, partindo de dados brutos de sequenciamento até anotação funcional do genoma (Almeida et al., 2023).

Toda a lógica de implementação, módulos analíticos disponibilizados, operação, dependências, versionamento e documentação está completamente delineada em artigo publicado em 2023 (Almeida et al., 2023), incluído neste trabalho como Apêndice.

3.1.1 Implementação

Os protocolos computacionais foram implementados utilizando a linguagem de domínio específico Nextflow (Di Tommaso et al., 2017), uma ferramenta especializada para a definição e gerenciamento de *pipelines* complexos. Foram utilizados diversos conceitos discutidos e estabelecidos pela comunidade nf-core (Ewels et al., 2020). Dentre estes, ressalta-se: (i) suporte para execução através de contêineres; (ii) utilização de versionamentos de código e contêineres; (iii) adoção da estrutura de código estabelecida pela comunidade; (iv) disponibilização de testes mínimos para avaliação da ferramenta; (v) documentação detalhada.

Para o desenvolvimento dos *pipelines*, diversos scripts foram desenvolvidos para intermediar o processamento e compilação de resultados, de modo a serem utilizados

para a geração de arquivos gráficos e de relatórios. Com isso, diversas linguagens de programação foram utilizadas, como Python [1], R [2] e Groovy [3]. Todos os programas, scripts e dependências foram empacotados em imagens Docker® para garantir a distribuição uniforme dos *pipelines* independente do sistema operacional disponível (Merkel, 2014; Boettiger, 2015).

Para garantir flexibilidade e seu uso em diversos cenários, além da análise de genomas procarióticos, o fluxo computacional foi dividido em três *pipelines* independentes. Desta forma, desenvolveu-se os seguintes *pipelines*:

1. ngs-preprocess: para pré-processamento e controle de qualidade de dados brutos de sequenciamento de DNA;
2. MpGAP: para a montagem de genomas;
3. Bacannot: para a anotação de genomas procarióticos

Os dois primeiros podem ser utilizados sequencialmente (ou não), para qualquer projeto de sequenciamento de DNA, independente do organismo. Já o *pipeline* de anotação é específico para genomas procarióticos.

A arquitetura geral dos *pipelines* encontra-se delineada na Figura 3 e a implementação de cada um deles será detalhada de forma sucinta nas próximas seções.

[1] `https://www.python.org/`
[2] `https://www.r-project.org/`
[3] `https://groovy-lang.org/`

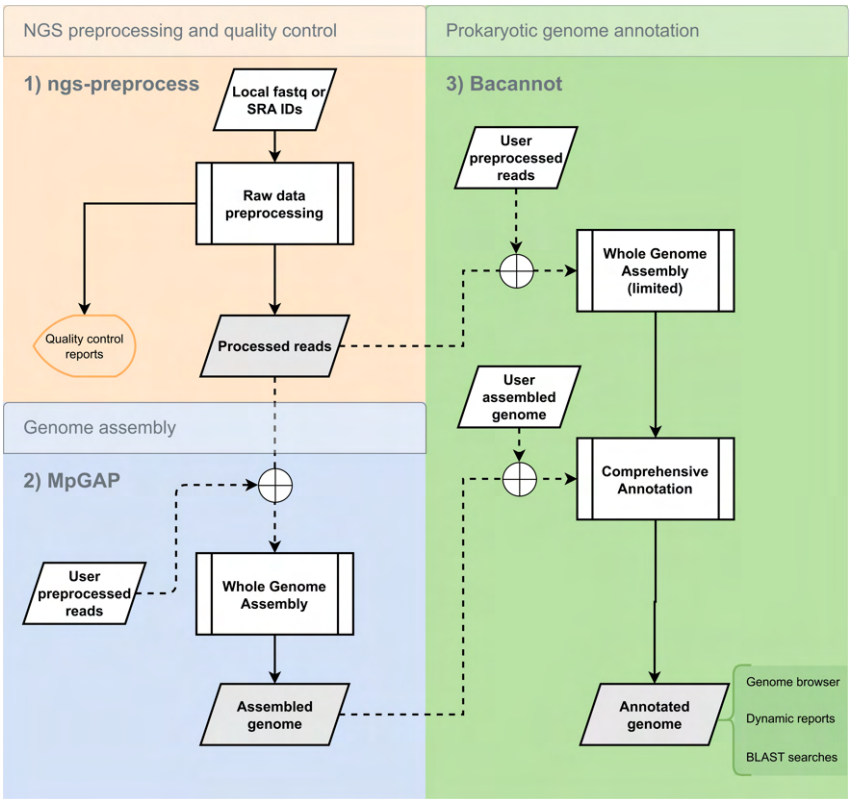

Figura 3: Visão geral do fluxo de trabalho dos *pipelines*. Estes são dividido em três módulos que são executados separadamente, na seguinte ordem: (i) pré-processamento de dados brutos de sequenciamento com o *pipeline* ngs-preprocess; (ii) montagem de genomas com o *pipeline* MpGAP; (iii) anotação de genomas com o *pipeline* Bacannot. Os diferentes arquivos de entrada aceitos pelos *pipelines* são representados por losangos de fundo branco, demarcando os possíveis pontos de início de execução, enquanto losangos de fundo cinza, demarcam o final de cada *pipeline*. Fonte: Almeida et al. (2023).

3.1.2 *Pipeline de pré-processamento: "ngs-preprocess"*

O *pipeline* "ngs-preprocess" é capaz de realizar várias etapas de controle de qualidade necessárias para a avaliação de dados de sequenciamento de DNA de nova geração (NGS). O *pipeline* aceita dados de leituras de sequenciamento de diversas plataformas NGS, como leituras curtas (plataforma Illumina™) e longas (plataformas PacBio e Oxford Nanopore). Os arquivos de sequenciamento podem estar disponíveis em armazenamento local, armazenamentos remotos (ex: o serviço de nuvem S3 da Amazon), ou ainda, diretamente do repositório público "Sequence Read Archive" (SRA[4]). No último caso, caso uma lista de identificadores do SRA seja fornecida, os dados brutos de sequenciamento (no formato FASTQ[5]) são automaticamente baixados localmente.

Todas as etapas analíticas do *pipeline* são determinadas automaticamente dependendo do tipo de leitura utilizado ou por configurações definidas pelo usuário. Essas etapas incluem verificação de contaminação por outros organismos, remoção de adaptadores e regiões de baixa qualidade de bases, demultiplexação, conversão de arquivos e geração de relatórios gráficos (Figura 4). As versões das ferramentas e suas aplicações são detalhadas na Tabela 1 do artigo em Apêndice (Almeida et al., 2023).

O produto final deste *pipeline* é um diretório contendo:

- as leituras de sequenciamento pré-processadas, organizadas por tecnologia de sequenciamento;
- os resultados intermediários de checagem de qualidade e metadados produzidos pelas ferramentas executadas, organizados por amostra;
- um arquivo de entrada ("samplesheet") pré-configurado para encadeamento com o *pipeline* "MpGAP"

Em termos gerais, estes resultados podem ser utilizados em qualquer projeto de sequenciamento NGS, inclusive para análises de transcritomas (RNA-seq) e metagenomas.

[4] https://www.ncbi.nlm.nih.gov/sra
[5] https://en.wikipedia.org/wiki/FASTQ_format

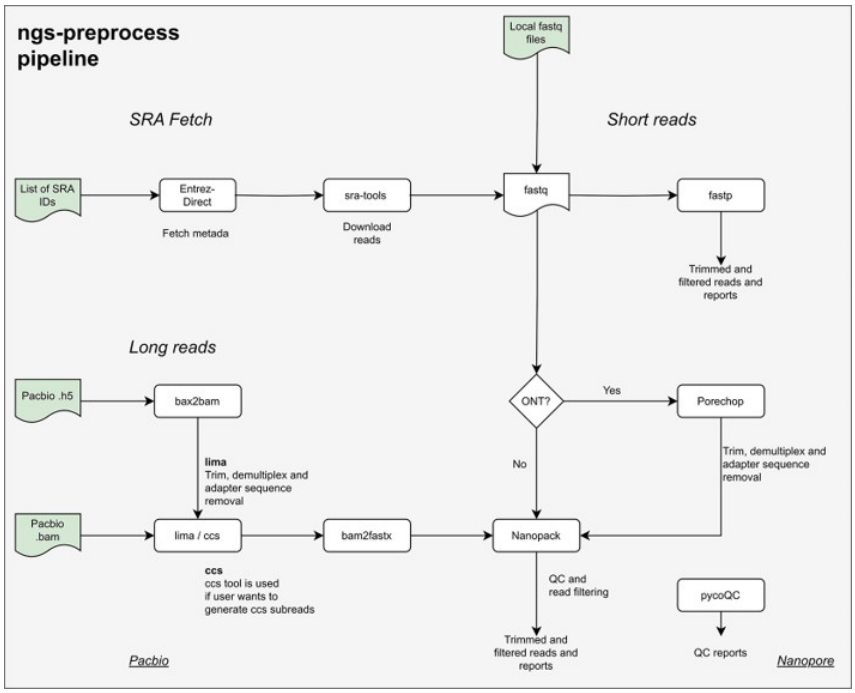

Figura 4: Visão geral do fluxo de trabalho desempenhado pelo *pipeline* "ngs-preprocess" e todas as etapas analíticas disponibilizadas. Detalha-se na figura as diferentes ferramentas e etapas adotadas pelo *pipeline* dependendo do tipo de arquivo bruto recebido e do tipo de tecnologia de sequenciamento. Todos os diferentes arquivos de entrada aceitos pelo *pipeline* e respectivos pontos de início de processamento são representados pelas formas geométricas especiais de fundo verde. Fonte: Almeida et al. (2023).

3.1.3 *Pipeline de montagem de genomas: "MpGAP"*

O *pipeline* "MpGAP" para montagem *de novo* de genomas foi desenvolvido para ser independente da plataforma de sequenciamento e do organismo. Este realiza montagens usando apenas leituras curtas, apenas leituras longas, bem como as denominadas montagens híbridas, que utilizam uma combinação de dados de tecnologias de sequenciamento (Figura 5). Na Figura 5 detalha-se todas as ferramentas de montagem ("assemblers") e de correção de erros de sequenciamento (polimento) que podem ser utilizadas em cada estratégia de montagem.

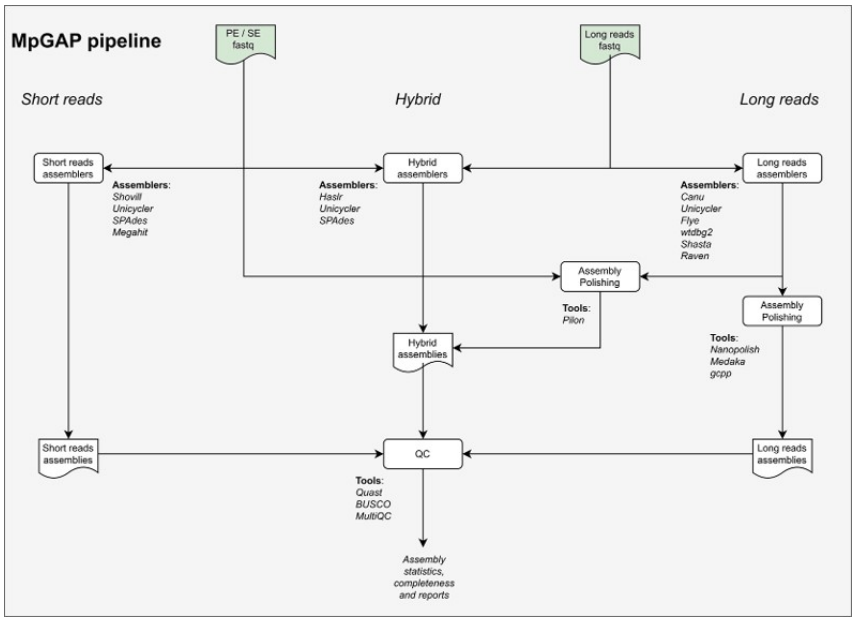

Figura 5: Visão geral do fluxo de trabalho desempenhado pelo *pipeline* "MpGAP" e todas as etapas analíticas disponibilizadas. Todos os diferentes arquivos de entrada aceitos pelo *pipeline* e respectivos pontos de início de processamento são representados pelas formas geométricas especiais de fundo verde. Da esquerda para a direita, exemplifica-se os fluxos de trabalho unicamente de leituras curtas, híbrido e unicamente de leituras longas. Fonte: Almeida et al. (2023).

Dependendo da combinação de dados de entrada fornecida pelo usuário, o *pipeline* seleciona automaticamente todos os modos de montagem possíveis, as quais podem ser específicas para um perfil de leituras (curtas ou longas) ou sua combinação com duas estratégias para a montagem híbrida de genomas:

- estratégia direta onde todos os dados são entregues a montadores especializados que nativamente realizam a montagem híbrida

- estratégia indireta, onde as leituras longas são primeiramente montadas isoladamente por montadores especializados e, em seguida, erros de sequenciamento são corrigidos utilizando leituras curtas de alta qualidade

Por fim, o *pipeline* realiza uma etapa de controle de qualidade onde coleta estatísticas das montagens e as condensa em relatório unificado. As versões das ferramentas e suas aplicações são detalhadas na Tabela 2 do artigo em Apêndice (Almeida et al., 2023).

O resultado final do "MpGAP" é um diretório contendo:

- todos os genomas produzidos, incluindo genomas antes e depois do polimento, organizados por amostra;
- todos os arquivos de checagem de qualidade de cada uma das montagens, incluindo compilação em relatório MultiQC (Ewels et al., 2016) e arquivo de texto para comparação, organizados por amostra;
- um arquivo de entrada ("samplesheet") pré-configurado para encadeamento com o *pipeline* "bacannot"

3.1.4 Pipeline de anotação de genomas procarióticos: "Bacannot"

O *pipeline* "bacannot" é especializado na anotação de genomas procarióticos, abrangendo etapas como a predição de genes, anotação de famílias gênicas, elementos genéticos móveis e identificação de características clinicamente relevantes. A principal característica deste *pipeline* foi seu desenvolvimento focado em auxiliar usuários menos familiarizados com as ferramentas a investigar e interpretar seus resultados. O fluxo de trabalho completo realizado pelo *pipeline*, incluindo etapas obrigatórias e opcionais é resumido na Figura 6.

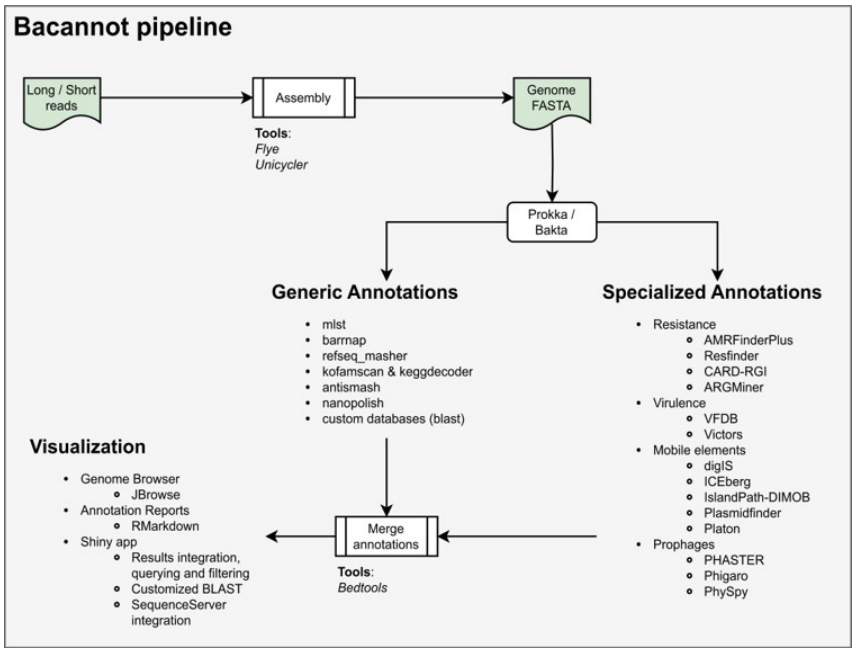

Figura 6: Visão geral do fluxo de trabalho desempenhado pelo *pipeline* "Bacannot" e todas as etapas analíticas disponibilizadas. Detalha-se na figura as diferentes tarefas de anotação: (i) genéricas, isto é, comuns para quaisquer linhagens; (ii) especializadas, que utilizam bancos de dados bastante específicos e geralmente são mais relevantes para linhagens clínicas. Além disso, detalha-se também todos as formas de visualização dos resultados geradas pelo *pipeline*. Todos os diferentes arquivos de entrada aceitos pelo *pipeline* e respectivos pontos de início de processamento são representados pelas formas geométricas especiais de fundo verde. Fonte: Almeida et al. (2023).

Ao final de sua execução, o *pipeline* gera relatórios dinâmicos e interativos de toda a anotação, juntamente com um navegador genômico para garantir a análise dos resultados em seus contextos genômicos (Almeida et al., 2023). Adicionalmente, o *pipeline* processa todos os elementos genéticos anotados e incorpora a anotação completa em um banco de dados relacional SQL[6] (do inglês "Structured Query Language") para permitir a manipulação estruturada dos resultados.

Desta forma, para complementar o *pipeline* visando facilitar a exploração destes resultados, desenvolveu-se uma aplicação web personalizada utilizando o "framework R Shiny" (Chang et al., 2024) que utiliza como base o banco de dados SQL gerado pelo *pipeline*, fornecendo uma plataforma completa para a integração e análise dos resultados.

Esta plataforma oferece aos usuários diversos recursos adicionais como filtragem dinâmica de anotações, bem como uma funcionalidade de busca de similaridade de sequência (BLAST) (Camacho et al., 2009), a qual é integrada com uma interface gráfica para execução e visualização dos resultados produzidos utilizando a ferramenta SequenceServer (Priyam et al., 2019). Além do uso interativo através desta plataforma, usuários podem opcionalmente fornecer bancos de dados personalizados (com genes de interesse) em arquivos FASTA como entrada para o *pipeline*, que utilizará estas sequências para anotação direcionada com BLAST+ (Camacho et al., 2009) e integrará estes resultados adicionais em relatórios dinâmicos, assim como os demais.

Da mesma maneira que as demais dependências do *pipeline*, esta plataforma também encontra-se encapsulada e distribuída através de imagens Docker. As versões das ferramentas e suas aplicações, bem como todos os bancos de dados incluídos são detalhados na Tabela 3 do artigo em Apêndice (Almeida et al., 2023).

3.2 Estudo de caso de isolados bacterianos multirresistentes

Para demonstrar a utilidade dos *pipelines* desenvolvidos realizou-se um estudo de caracterização genômica de isolados bacterianos com relevância para o problema de resistência a antimicrobianos. Os métodos experimentais e computacionais deste estudo são descritos a seguir.

[6] `https://en.wikipedia.org/wiki/SQL`

3.2.1 Isolamento e teste de suscetibilidade

Foram obtidos três isolados de *Klebsiella pneumoniae* pertencentes à coleção do Laboratório de Análises Moleculares de Patógenos da UnB (LAMP/UnB), coordenado pela professora Tatiana Amabile Campos. As três linhagens utilizadas (KpBSB56, KpBSB60 e ECR) encontram-se descritas na Tabela 3 e foram isoladas a partir de culturas microbianas obtidas de um Hospital terciário (Hospital Universitário de Brasília - HUB/UnB) em Julho de 2021. O isolado KpBSB56 foi obtido de drenagem de secreção de um paciente de 45 anos; KpBSB60 isolado de urina de uma paciente de 80 anos; e ECR isolado de lavagem peritoneal de uma paciente de 41 anos. O sistema MicroScan® (Beckman Coulter) foi utilizado para a identificação microbiana e teste de suscetibilidade a antimicrobianos. Adicionalmente, a resistência a Polimixina B foi testada através do teste de concentração inibitória mínima (MIC), como recomendado pelo CLSI (do inglês, "Clinical & Laboratory Standards Institute").

Tabela 3 - Metadados gerais dos isolados bacterianos analisados neste estudo. Na tabela, apresenta-se a coleção da qual estes isolados fazem parte, a fonte de isolamento e a idade e sexo dos pacientes dos quais estas bactérias foram isoladas.

Isolados	Fonte de isolamento	Idade do paciente	Sexo do paciente	Coleção
KpBSB56	Drenagem de secreção	45	Masculino	LAMP
KpBSB60	Urina	80	Feminino	LAMP
ECR	Lavagem peritoneal	41	Feminino	LAMP

3.2.2 Extração de DNA e sequenciamento

A extração de DNA total foi realizada conforme descrito por Ausubel et al. (1992) e a sua quantificação foi feita utilizando o equipamento Nanodrop™. O sequenciamento de DNA foi realizado utilizando leituras longas e curtas. O sequenciamento de leituras longas por nanoporo foi realizado em Janeiro de 2022 utilizando a plataforma *Oxford Nanopore MinION*, coordenado por Rodrigo de Paula Baptista da Universidade da Geórgia (EUA). As cepas foram multiplexadas com o kit "*Rapid Barcode Prep. Kit*" e sequenciadas em duas "*flowcells*" R9.4. O sequenciamento de leituras curtas foi realizado pela empresa

Beijing Genomics Institute (BGI) (Shenzhen, China) em Agosto de 2022, utilizando a estratégia de leituras pareadas (*paired-end*) com 150 pares de base em uma plataforma DNBseq ®.

3.2.3 Genomas públicos adicionais

Outros isolados brasileiros foram selecionados para permitir análises comparativas com nossas linhagens (Tabela A.1, Material Suplementar). Dentre eles selecionou-se, manualmente:

- todos os genomas de linhagens de *Klebsiella pneumoniae* de um estudo que investigou a prevalência do gene *bla*NDM (New Delhi metallo-β-lactamase) no Brasil de 2015 a 2021 (Camargo et al., 2022)
- um conjunto de dados brutos de sequenciamento de *Klebsiella pneumoniae* isolados em Brasília entre 2010 e 2014 (Lee et al., 2021) para comparação temporal
- duas linhagens sequenciadas por nosso grupo de pesquisa para comparação local, KpBSB31 isolada no mesmo Hospital terciário (De Campos et al., 2018) e KpV3 isolada no lago Paranoá (Janssen et al., 2021)

3.2.4 Análise computacional

3.2.4.1 Pré-processamento

As leituras curtas foram entregues pré-processadas pela empresa BGI, incluindo remoção de sequências adaptadoras, contaminação e sequências de baixa qualidade ($\leq$Q20). As leituras longas brutas foram pré-processadas através do *pipeline* ngs-preprocess v2.4 (Almeida et al., 2023). Através dele, porechop v0.2.4 (Wick et al., 2017) foi utilizado para remoção de sequências adaptadoras, nanofilt v2.8.0 para a filtragem de leituras de baixa qualidade e nanostat v1.6.0 para checagem da qualidade das leituras, ambos parte do pacote NanoPack (De Coster e Rademakers, 2023). As leituras longas foram filtradas baseadas em qualidade ($\geq$10) e tamanho ($\geq$750), através da configuração do *pipeline*

ngs-preprocess com os parâmetros "--lreads_min_length" e "--lreads_min_quality", respectivamente.

3.2.4.2 Montagem de genoma

Leituras curtas e longas foram montadas através do *pipeline* MpGAP v3.1.4 (Almeida et al., 2023). As leituras foram montadas de forma híbrida, onde as leituras longas foram primeiramente montadas com Flye v2.9 (Kolmogorov et al., 2019) seguido da correção de erros (polimento) usando leituras curtas com Pilon v1.24 (Walker et al., 2014). Este modo de montagem é selecionado no *pipeline* ao definir o parâmetro "--hybrid_strategy" como "2".

O isolado KpBSB60, devido à baixa qualidade observada de suas leituras longas, foi adicionalmente montado utilizando somente as leituras curtas, também através do *pipeline* MpGAP com parâmetros padrão. Nesta estratégia, o programa Shovill [7] v1.1.0 foi utilizado tendo o programa Skesa v2.4.0 (Souvorov et al., 2018) como seu montador base ("core assembler"). Em seguida, já sem auxílio do *pipeline*, foi realizado um *scaffolding* destas sequências utilizando as leituras longas disponíveis através do programa LongStich v1.0.4 (Coombe et al., 2021), com parâmetros padrão. Finalmente, outras ferramentas especializadas, como MOB-suite v3.1.7 (Robertson e Nash, 2018) e PLASMe v1.1 (Tang et al., 2023), foram utilizadas com parâmetros padrão para tentar reconstruir os plasmídeos deste isolado.

Os dados públicos de isolados de Brasília (Lee et al., 2021) utilizados para genômica comparativa (Subsubseção 3.2.4.4) encontram-se disponibilizados somente como dados brutos de sequenciamento e, por isso, tiveram seus genomas montados com Shovill e Skesa, como descrito para KpBSB60.

A qualidade das montagens foi avaliada pela quantidade de genes essenciais esperados que são encontrados na montagem. Para isto, utilizou-se o programa BUSCO v5.5.0 (Manni et al., 2021) com parâmetros padrão e seu banco de dados de genes essenciais da Ordem Enterobacterales (versão odb10). Finalmente, a circularização dos genomas foi avaliada com a ferramenta circlator v1.5.5 (Hunt et al., 2015) e erros de montagens foram

[7] https://github.com/tseemann/shovill

verificados com CRAQ v1.0.9 (Li et al., 2023), ambos utilizando seus parâmetros padrão.

3.2.4.3 Anotação de genoma

Todos os genomas utilizados foram anotados usando o *pipeline* bacannot v3.1.5 (Almeida et al., 2023). Os bancos de dados requeridos pelo *pipeline* foram obtidos em Maio de 2022. O *pipeline* executou Prokka v1.14 (Seemann, 2014) como base de sua anotação e o banco de dados da ferramenta foi incrementado com a biblioteca pública do NCBI, PGAP (Li et al., 2020), ao usar o parâmetro "--prokka_use_pgap" no *pipeline*.

A detecção de plasmídeos foi realizada com Platon v1.6 (Schwengers et al., 2020) e PlasmidFinder v2.1.6 (Carattoli et al., 2014). Resfinder v4.1 (Bortolaia et al., 2020) e AMRFinderPlus v3.10 (Feldgarden et al., 2019) foram usados para anotação de resistência. O painel de resistência do programa Resfinder foi selecionado através do parâmetro do *pipeline* "--resfinder_species" que foi definido como "Klebsiella". VFDB (do inglês, "Virulence Factor Database") (Chen et al., 2005) foi utilizado através do *pipeline* para a anotação de virulência.

Adicionalmente, devido aos isolados serem da espécie *Klebsiella pneumoniae*, Kleborate v2.2.0 (Lam et al., 2021) foi utilizado manualmente para complementar os resultados e investigar características específicas de *Klebsiella*.

3.2.4.4 Genômica comparativa

A distância genética entre os genomas analisados e outros genomas do banco de dados NCBI foi calculada com a ferramenta refseq_masher [8] v0.1.2, ainda através do *pipeline* bacannot.

A identidade média de nucleotídeos (ANI, do inglês "Average Nucleotide Identity") entre os genomas deste estudo foi calculado usando FastANI v1.33 (Jain et al., 2018) e a análise de colinearidade e sintenia entre os genomas da ECR e KpBSB56 foi realizada através do *pipeline* MCscan (Tang et al., 2008), parte da biblioteca jcvi v1.3.8, ambos utilizando parâmetros padrão.

Análises filogenéticas foram realizadas por ferramentas que utilizam genomas com-

[8] `https://github.com/phac-nml/refseq_masher`

pletos, sendo elas: SANS serif v2.3_9A (Rempel e Wittler, 2021) que realiza uma reconstrução filogenética livre de alinhamento e Parsnp v1.7.3 (Treangen et al., 2014) que possui abordagem dependente de alinhamento, baseada em polimorfismos de base única. Além dos parâmetros padrão, SANS serif foi configurado para calcular árvore com 1.000 repetições "bootstrap".

Devido ao programa Parsnp não realizar "bootstrap", o seu arquivo de alinhamento de genomas foi utilizado para reconstrução da filogenia com a ferramenta IQTree v2.2.5 (Minh et al., 2020), com seleção de modelo automática ("-m MFP"), "bootstrap" ultra rápido ("-B") com 1000 repetições e parâmetro para evitar super estimação deste método de "bootstrap" ("-bnni"). As figuras de filogenia e mapas de presença e ausência de genes foram geradas com *scripts* R usando as bibliotecas ggtree v3.10.0 (Yu et al., 2016), ggplot2 v3.4.4 (Wickham, 2016) e phytools v2.1-1 (Revell, 2011). As árvores foram desenhadas utilizando a função "midpoint.root()" do pacote phytools.

As análises de co-ocorrência de genes foram realizadas utilizando os pacotes R CooccurrenceAffinity v1.0 (Mainali e Slud, 2022) e Cooccur v1.3 (Griffith et al., 2016). Os gráficos desta análise foram gerados com o pacote R corrplot [9]. O mapa circular do plasmídeo foi desenhado através da ferramenta "CGView Comparison Tool" (CCT) v2.0.3 (Grant et al., 2012). Os gráficos de "cluster" gênicos foram produzidos com o programa Gcluster tool v2.06 (Li et al., 2020).

[9] https://github.com/taiyun/corrplot

4 Resultados e Discussão

4.1 Desenvolvimento de *pipelines* para genômica bacteriana

O desenvolvimento do conjunto de softwares descritos neste trabalho culminou em publicação (Almeida et al., 2023) (em Apêndice), a qual fornece um detalhamento de sua arquitetura e desenvolvimento. Nas próximas seções serão detalhadas a usabilidade e operabilidade dos *pipelines*, bem como publicações que os utilizaram como base.

4.1.1 Formalização de pipelines com o Nextflow

De maneira análoga à protocolos experimentais na Biologia, protocolos realizados no computador, ou "in silico", devem observar diversos critérios de forma a garantir sua consistência e replicação. Na área de bioinformática os protocolos computacionais são denominados *pipelines*, ou *workflows*, e consistem na utilização de diversos programas em etapas bem definidas. Para garantir a sua reprodução faz-se necessário registrar as versões dos programas e todo seu ambiente de execução, como o sistema operacional utilizado.

Ademais, em um *pipeline* todas as etapas devem ser delineadas e conectadas de forma a estabelecer um fluxo coerente de ativação das mesmas. Dependências entre as etapas ocorrem quando a uma etapa B requer a utilização de um arquivo (ou variável lógica) gerado por uma etapa A, estabelecendo a seguinte sequência de execução de tarefas: $A \mapsto B$. Em uma situação comum de um protocolo de bioinformática, existem diversas tarefas que devem ser encadeadas e esta atribuição fica a cargo do desenvolvedor, que deve delinear todas as etapas e respectivas dependências para criar o *pipeline*.

No entanto, existem tarefas que não apresentam dependências entre si, ou seja, seus arquivos de entrada já estão disponíveis. Nestes casos, para efeito de otimização de execução, seria interessante usar o potencial dos múltiplos núcleos ("cores") dos proces-

sadores modernos e executar as tarefas em paralelo.

Por exemplo, uma tarefa C pode requerer arquivos gerados por duas etapas distintas (A e B), mas que não possuem dependências entre si. Para maximizar o potencial do computador, as tarefas A e B podem ser executadas simultaneamente usando núcleos diferentes e, quando ambas forem finalizadas, progredir para a tarefa dependente C, como na representação a seguir:

$$\begin{array}{ccc} A & \| & B \\ \textit{núcleo}_1 & & \textit{núcleo}_2 \\ \searrow & & \swarrow \\ & C & \end{array}$$

A modelagem da situação de paralelismo acima é ainda mais complexa em termos de programação e é geralmente negligenciada em favor de considerar todas as tarefas de maneira linear ($A \mapsto B \mapsto C$), o que não é eficiente.

De modo geral, *pipelines* podem ser escritos em qualquer linguagem de programação, como Python e Bash. No entanto, estas implementações geralmente carecem de uma organização lógica e não conseguem modelar, de maneira simples, as relações de dependência entre tarefas, a sua execução em paralelo, a tolerância a falhas e extensibilidade (Leipzig, 2016).

Em se tratando de *pipelines*, o conceito de "tolerância a falhas" refere-se à capacidade de um protocolo computacional de continuar funcionando mesmo quando ocorrem erros em etapas intermediárias. Estes podem ser tão simples quanto uma interrupção na conexão de rede, até mais complexos como a insuficiência de memória durante a execução de algumas tarefas em paralelo. Por outro lado, o termo "extensibilidade", refere-se a capacidade de um *pipeline* de ser facilmente incrementado com novas funcionalidades, de forma modular, sem a necessidade de grandes modificações na infraestrutura existente. Um exemplo bem difundido na comunidade quanto a este conceito é a plataforma Galaxy (Abueg et al., 2024).

Motivados por estes conceitos e necessidades, optamos por adotar o Nextflow, uma linguagem específica para o desenvolvimento e orquestração de *pipelines* (Di Tommaso

et al., 2017). O Nextflow consiste em uma linguagem específica para definição de *pipelines*, permitindo a criação de tarefas de forma modular e de fácil definição de dependências de entrada e saída entre elas. Adicionalmente, o Nextflow oferece um ambiente robusto de execução de *pipelines*, garantindo uma execução coordenada e eficiente das etapas através do gerenciamento de recursos computacionais (número de CPUs e memória), tolerância a falhas e uso de contêineres.

Para controle de erros, o Nextflow permite que tarefas sejam desenhadas com a possibilidade de reexecução automática em caso falhas temporárias, através do parâmetro `errorStrategy 'retry'`, que reexecuta uma tarefa quantas vezes for definido no *pipeline*. Além disso, caso a execução apresente um erro fatal e que necessite reajuste de parâmetros, o Nextflow permite a sua reexecução com o parâmetro `-resume`, que permite que *pipelines* reutilizem resultados já processados, economizando tempo e recursos. De forma geral, estes recursos avançados fazem com que esta linguagem proporcione uma solução mais robusta e flexível para a execução de processos complexos, assegurando adaptabilidade em ambientes variados (Di Tommaso et al., 2017).

O Nextflow suporta o uso de contêineres para a execução dos *pipelines*, garantindo um ambiente replicável e isolado para execução das tarefas. Por isso, a implementação dos *pipelines* acoplados a contêineres Docker é um ponto chave a se destacar. Os contêineres são microambientes de computação que permitem a instalação de programas e suas dependências (linguagens e bibliotecas de programação, arquivos de configuração) em uma mesma instância. Geralmente este ambiente (contêiner) é baseado no sistema operacional Linux, mas soluções de virtualização, como o Docker e Singularity, permitem sua execução nativa em qualquer outro sistema operacional que o usuário dispõe (ex: Windows). Sendo assim, o desenvolvedor do *pipeline* cria e distribui o seu próprio contêiner, e o Nextflow utiliza de forma transparente os programas instalados no contêiner para consecução das tarefas. Com isso, toda a complexidade de instalação de programas e manutenção das versões sai da alçada do utilizador do *pipeline*, garantindo um ambiente replicável e isolado para a execução das tarefas.

4.1.2 *Implementação dos pipelines*

Em se tratando de genômica bacteriana, análises normalmente envolvem o encadeamento de dezenas de programas dentro de um fluxo de trabalho. É comum ter, para uma mesma tarefa, uma grande diversidade de alternativas de programas com diferentes implementações, vantagens e desvantagens, permitindo que pesquisadores possam escolher ferramentas que melhor se adaptem às suas necessidades específicas. Dessa forma, *pipelines* completos de genômica envolvem diversas opções de tarefas que dependem da natureza dos dados. Por exemplo, dados provenientes de diferentes tecnologias de sequenciamento podem necessitar de etapas de pré-processamento distintas. Mesmo assim, apesar de toda essa variabilidade, existem ainda etapas que são consistentes e universais, como por exemplo, a checagem de qualidade de montagens de genomas.

Neste contexto, visando criar um ecossistema completo de análise automatizada de genomas bacterianos a partir de dados brutos de diversas tecnologias recentes de sequenciamento de DNA, foram desenvolvidos três *pipelines* modularizados e independentes (Tabela 4). Cada um deles, inclui uma diversa seleção de ferramentas computacionais que proporcionam flexibilidade e generalidade, garantindo sua aplicação em diferentes cenários analíticos. De maneira geral, houve um intenso trabalho para estruturar estes *pipelines* de modo a empregar conceitos amplamente adotados pela comunidade nf-core (Ewels et al., 2020), que cataloga diversos *pipelines* padronizados de bioinformática e que adotam os seguintes preceitos: (i) execução por meio de contêineres; (ii) organização do código fonte em uma estrutura normatizada; e (iii) documentação rica em detalhes e exemplos.

Tabela 4 - Descrição dos três *pipelines* desenvolvidos neste trabalho (Almeida et al., 2023). Nos respectivos repositórios encontram-se o código-fonte e documentação, distribuídos na forma de código aberto ("open source").

Nome	Objetivo	Repositório
ngs-preprocess	Pré-processamento de dados brutos	`https://github.com/fmalmeida/ngs-preprocess`
MpGAP	Montagem de genomas	`https://github.com/fmalmeida/MpGAP`
bacannot	Anotação de genomas	`https://github.com/fmalmeida/bacannot`

4.1.3 Descrição dos *pipelines*

O *pipeline* "ngs-preprocess" é capaz de aplicar de forma independente e autônoma, diversas etapas de controle de qualidade e pré-processamento requeridos para a utilização de dados de sequenciamento de próxima geração de diferentes tecnologias de leituras curtas (ex: BGI Genomics e Illumina) e longas (ex: Pacific Biosciences e Oxford Nanopore). Incluindo o controle de qualidade, remoção de sequências adaptadoras, aplicação de filtros, demultiplexação de amostras, conversão de dados e relatórios gráficos para avaliação da qualidade dos dados.

O *pipeline* MpGAP foi desenhado de modo a permitir a montagem de genomas a partir de dados de leituras curtas ou longas, ou combinadas em uma estratégia de montagem híbrida. Dado um arquivo de configuração amplamente documentado[1], o *pipeline* é capaz de selecionar automaticamente os programas mais adequados para a montagem, dada a natureza dos dados disponíveis e a estratégia escolhida. Para isso, o *pipeline* inclui mais de 10 diferentes programas de montagem e alguns programas de correção de erros. Isso possibilita o usuário realizar diversas estratégias de montagem, as quais podem ser posteriormente comparadas através do relatório final gerado pelo *pipeline*, que por sua vez baliza a escolha da montagem final para a análises posteriores.

Por último, o *pipeline* bacannot é especializado na anotação genômica de procariotos. O bacannot permite a execução automática de módulos genéricos de predição de genes, detecção de sequências de rRNA, anotação de metabólitos secundários, cálculo de distância genética com genomas públicos, entre outros. Também inclui módulos específicos para a anotação de genes de virulência e resistência, predição de plasmídeos, anotação de elementos genéticos móveis como ilhas genômicas e elementos integrativos, anotação de prófagos e anotação funcional utilizando sequências ortólogas do banco KEGG (Kanehisa e Goto, 2000; Kanehisa et al., 2022). Por fim, existe a possibilidade do usuário fornecer sequências adicionais (ex: membros de uma família gênica) que são utilizadas para buscas por similaridade, o que permite estender o escopo de anotação. Toda a capacidade analítica do *pipeline* é detalhada em sua documentação web[2], incluindo explicação

[1] `https://mpgap.readthedocs.io/en/latest/`

[2] `https://bacannot.readthedocs.io/en/latest/`

e exemplificação dos resultados interativos produzidos pelo *pipeline*[3].

4.1.4 *Independência e personalização dos pipelines*

A estratégia de desenvolvimento adotada garante a independência dos três *pipelines*. Isso significa que podem ser executados separadamente ou em combinação, em múltiplos contextos. Portanto, esta arquitetura permite a criação de pontos de checagem entre os *pipelines* e a flexibilidade de executar somente as etapas de interesse. Por exemplo, usuários podem utilizar somente o *pipeline* de anotação a partir de um genoma previamente montado. Quando utilizados sequencialmente, os *pipelines* proporcionam uma solução completa para a análise de genomas bacterianos, desde o recebimento de dados brutos de sequenciamento até a geração de relatórios de anotação de diversas características genéticas codificadas no genoma. Além disso, este desacoplamento permite que os dois primeiros *pipelines* mais genéricos, possam ser utilizados para dados de organismos não-procarióticos.

Com enfoque no usuário final, os *pipelines* oferecem uma estrutura mais robusta e abrangente, com uma ampla automação das análises mas, ao mesmo tempo, permitindo a total flexibilidade para personalizar as mesmas conforme desejado. Por exemplo, apesar de automaticamente selecionar todos os métodos de montagem possível dado as diferentes combinações de arquivos de sequenciamento, o *pipeline* MpGAP permite que usuários ativem ou desativem estratégias ou programas de montagem de modo a executar somente os que lhe interessem. Essa versatilidade, garante que os usuários tenham poder de escolha e uma diversidade de resultados para consolidar suas predições. Similarmente, o *pipeline* de anotação também permite selecionar certas análises, como a anotação de metabólitos secundários, de resistência, virulência, etc., além de possibilitar a anotação de genes de interesse não contemplados diretamente pelo *pipeline*, como discutido anteriormente.

4.1.5 *Interface gráfica para exploração de resultados*

Um aspecto distintivo no desenvolvimento dos *pipelines* aqui descritos foi a melhoria na apresentação dos resultados para usuário. A maior parte dos *pipelines* coordena

[3] `https://bacannot.readthedocs.io/en/latest/outputs/`

a execução de diversos programas e organiza suas saídas em pastas. No entanto, estas saídas são geralmente dezenas de arquivos de texto que devem ser posteriormente dissecados para realizar inferências biológicas. O volume de trabalho manual para percorrer os inúmeros arquivos de texto e sumarizá-los é uma atribuição complexa e não trivial para o usuário final.

Diante desta lacuna, os esforços foram direcionados para o desenvolvimento de uma interface gráfica, operada via navegador de internet, para a visualização e interpretação dos resultados do *pipeline* de anotação genômica. Todo este sistema visual consiste em um programa adicional, desenvolvido com a linguagem de programação R e a biblioteca Shiny (Chang et al., 2024), e que realiza o pós-processamento automático dos diversos arquivos de texto gerados possibilitando a integração de todos os resultados. Este, oferece ao usuário diversos recursos adicionais como acesso unificado aos relatórios (como os genes de resistência e virulência preditos), filtragem dinâmica das anotações gênicas e buscas indexadas por palavras-chave ou por similaridade de sequência. Uma visão geral da aplicação encontra-se na Figura 7.

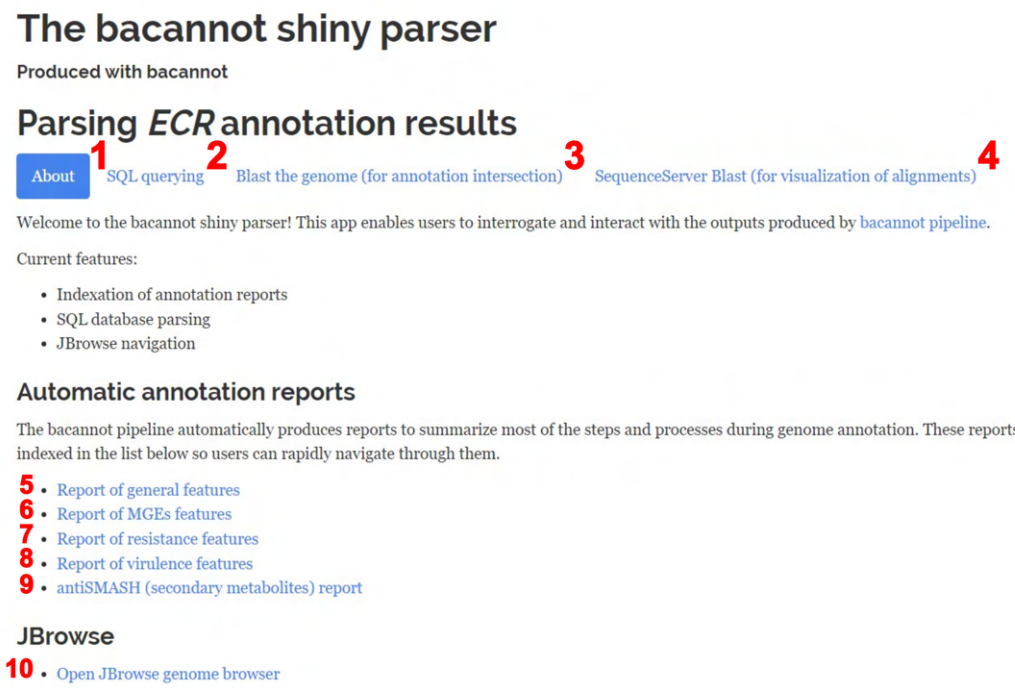

Figura 7: Visão geral das abas e ferramentas disponibilizadas na plataforma web para investigação dos resultados do *pipeline* Bacannot. Na Figura pode-se ver as diversas abas disponibilizadas no *pipeline* (1 a 4) e os links para rápido acesso aos relatórios e aplicações disponíveis (5 a 10). Os números anotados na figura destacam as páginas para acesso: 1. Da página de entrada da aplicação; 2. À ferramenta de filtragem dinâmica dos resultados baseada em texto; 3. À ferramenta para alinhamento e filtragem dinâmica dos resultados baseado em sequência; 4. À aplicação "SequenceServer" (Priyam et al., 2019) para executar e visualizar alinhamentos BLAST (Camacho et al., 2009); 5. Ao relatório automático da anotação genérica; 6. Ao relatório automático da anotação de elementos genéticos móveis; 7. Ao relatório automático da anotação de genes de resistência; 8. Ao relatório automático da anotação de genes de virulência; 9. Ao relatório automático de anotação da metabólicos secundários; 10. Ao navegador genômico.

Todos os elementos genéticos anotados pelo *pipeline* são processados pelo programa que disseca as suas propriedades (posição no genoma, nome do gene, presença no banco de dados CARD, etc.) e as incorpora em um banco de dados relacional que é anexado aos resultados de anotação. Com isso, é possível realizar com eficiência consultas específicas no universo de genes anotados para o genoma em questão. Por exemplo, pode-se inquirir graficamente os resultados por elementos conjugativos e integrativos (ICEs) que pertençam ao cromossomo na posição entre as bases 300.000 a 400.000. O resultado é uma tabela dinâmica no navegador que permite o exame dos componentes filtrados em seu contexto genômico. Os resultados filtrados podem ser baixados pelos usuários em diferentes formatos padrão como GFF[4] para anotação e CSV[5] que pode ser visualizado em um programa de planilha eletrônica. Uma imagem geral desta funcionalidade encontra-se na Figura 8.

Além de condensar todos os relatórios interativos e mecanismos de busca gerados pelo *pipeline* que resultam em tabelas, esta aplicação provê também um navegador genômico com todos os resultados da anotação. Esta visualização gráfica ao longo do genoma facilita um exame de presença de grupos de genes ("gene clusters") e vizinhança com elementos genômicos móveis que podem sugerir mecanismos de mobilização. Esta funcionalidade encontra-se representada na na Figura 9.

Adicionalmente, esta aplicação provê a possibilidade de realizar buscas por similaridade de sequência utilizando a ferramenta BLAST (Camacho et al., 2009) embutida (Figura 10A). Isso permite que o usuário forneça um conjunto de sequências alvo que são interrogadas contra o genoma montado (DNA) ou as proteínas anotadas, o que permite a identificação de homologia, variantes de sequência ou refinamento de anotação (Figura 10B). Por último, disponibiliza-se também o programa SequenceServer (Priyam et al., 2019) permitindo que o usuário visualize o alinhamento de suas sequências alvo (Figura 10C).

[4] `https://www.ensembl.org/info/website/upload/gff.html`

[5] `https://en.wikipedia.org/wiki/Comma-separated_values`

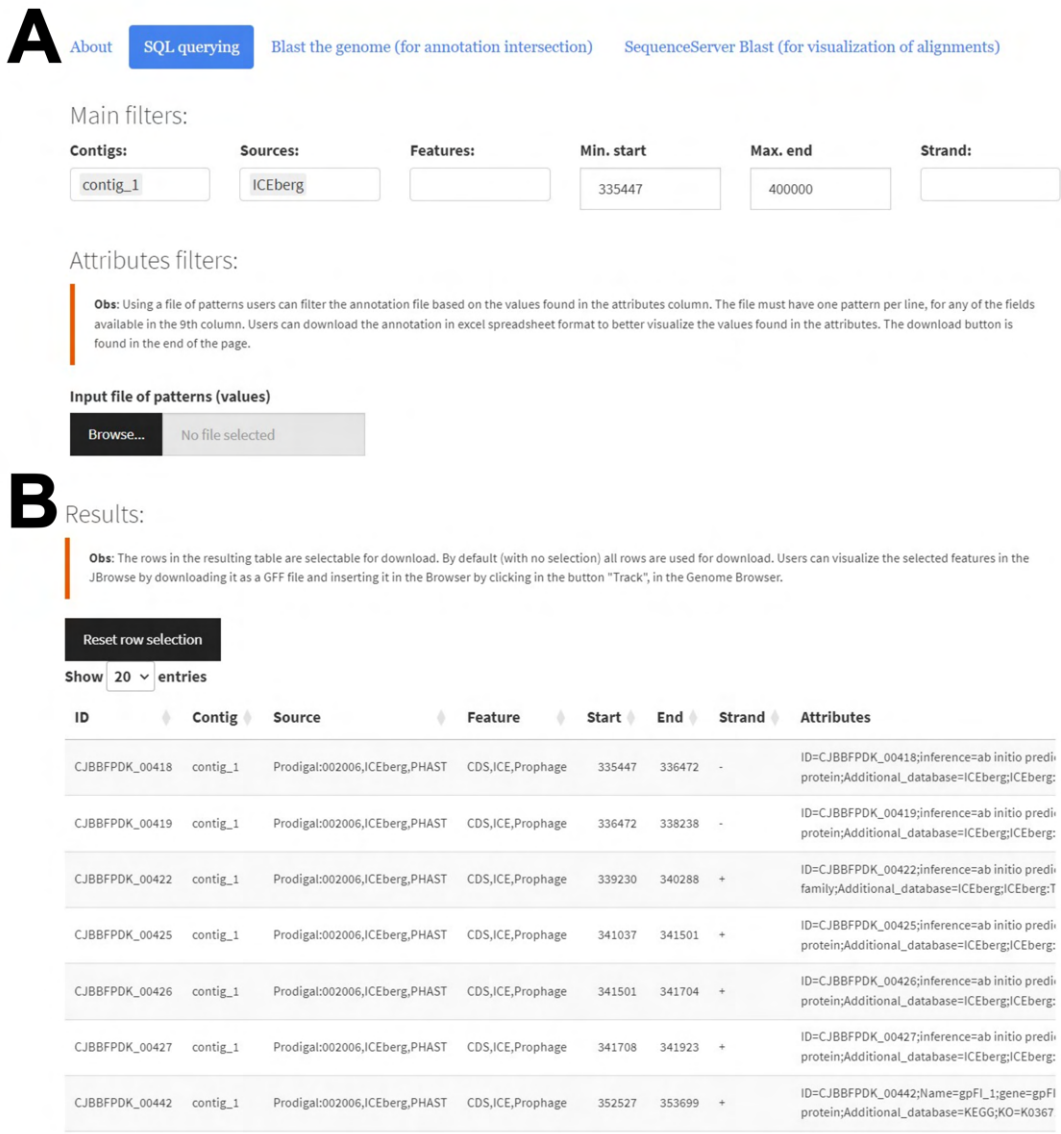

Figura 8: Visão geral da ferramenta de filtragem dinâmica baseada em texto disponibilizada na plataforma web do *pipeline* Bacannot. A figura A ilustra esta ferramenta e seus campos de filtros disponíveis que são acessados ao selecionar a página denominada "SQL querying" na aplicação. Nela, usuários podem utilizar os filtros pré-configurados ("Main filters") para filtrar os resultados por contig, "sources", "features", posição (início e fim) e por orientação ("strand"). Estes campos de anotação são baseados nas colunas do formato GFF[6]. O campo "sources" indica todas as fontes de anotação, por exemplo, "Resfinder", "VFDB", etc. O campo "features" indica todos os tipos de genes anotados, por exemplo, resistência, virulência, etc. Além disso, através da opção "Attribute filters", é possível também que usuários utilizem um arquivo texto com diversos filtros a serem aplicados de forma aditiva, baseado em qualquer par chave-valor presente na coluna nove da anotação GFF final, como por exemplo, identificadores de genes (ID), produto gênico ("product"), entre outros. Na figura B, ilustra-se a tabela contendo todos os resultados que correspondem aos filtros aplicados na figura A.

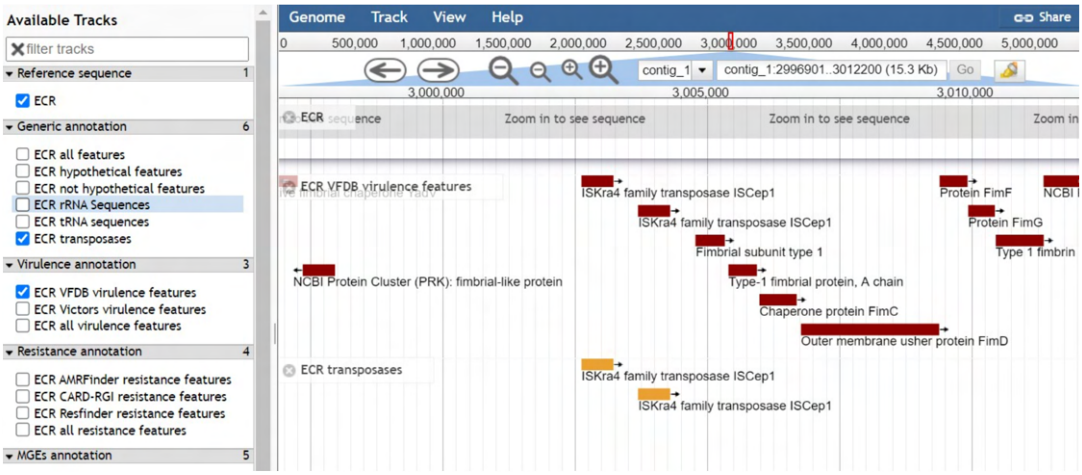

Figura 9: Visão geral da funcionalidade de navegador genômico disponibilizada na plataforma web do *pipeline* Bacannot. Ilustra-se na figura a investigação da anotação do genoma da ECR em seu contexto genômico, destacando a detecção de um grupo de genes de virulência que possuem em sua vizinhança um par de transposases. No menu lateral à esquerda, disponibiliza-se toda a anotação do genoma de forma seccionada, para que o usuário possa selecionar quais anotações visualizar no navegador. Cores são atribuídas a diferentes seções de anotação, como por exemplo, virulência, resistência, anotação genérica, etc.

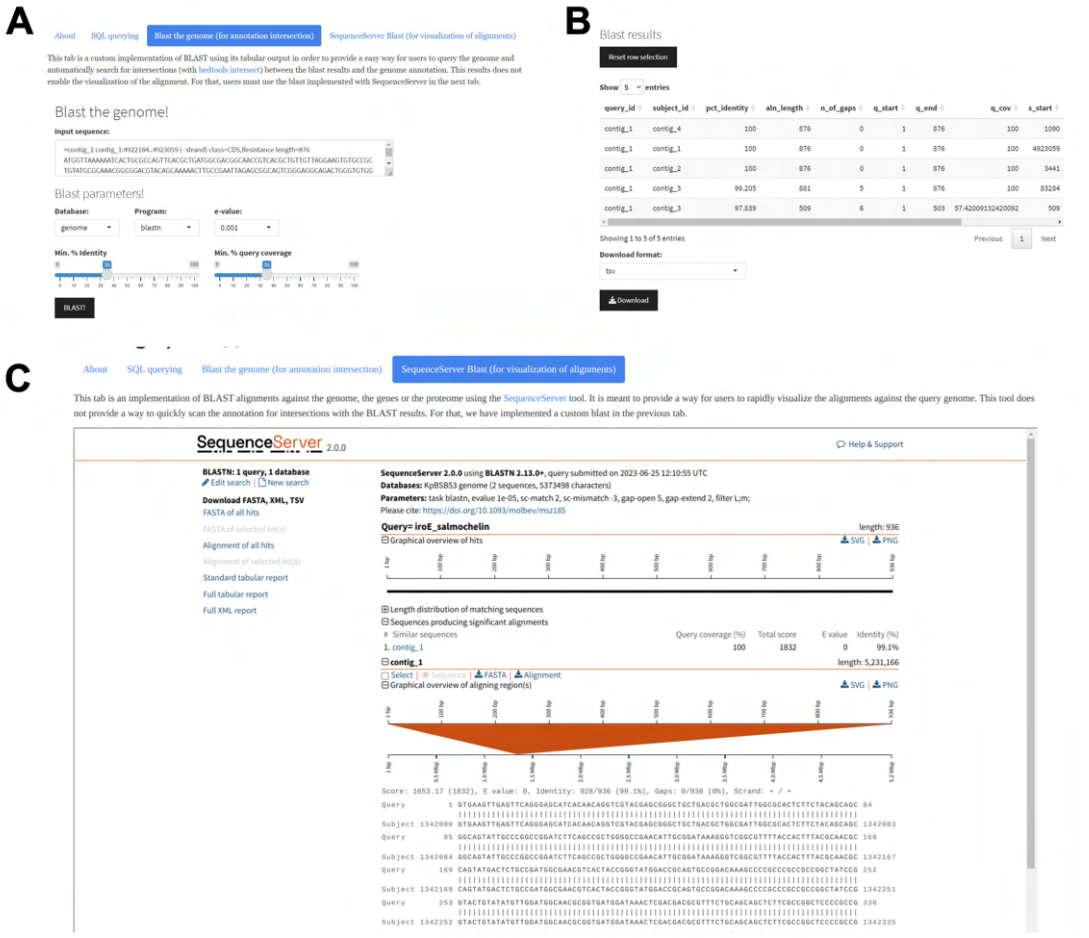

Figura 10: Visão geral da funcionalidade de buscas por similaridade de sequência disponibilizada na plataforma web do *pipeline* Bacannot. As figuras A e B ilustram a funcionalidade de busca por similaridade de sequência utilizando a ferramenta BLAST (Camacho et al., 2009) que permite alinhamento contra o genoma ou as proteínas anotadas (A) e além de apresentar os resultados em tabela, identifica também as anotações geradas pelos *pipelines* que intersectam com as coordenadas genômicas dos resultados BLAST. Na Figura C, ilustra-se o programa SequenceServer (Priyam et al., 2019) disponibilizado na plataforma para que usuários não só executem mas também visualizem o alinhamento de suas sequências alvo.

4.1.6 Visão geral da operação e funcionamento

Os *pipelines* desenvolvidos auxiliam usuários experientes ou com pouca prática em desenvolvimento e gerenciamento de *pipelines* que desejem realizar uma análise genômica bacteriana de maneira rápida e simples, quando comparado ao mesmo esforço sendo realizado de maneira passo a passo e sem coordenação de tarefas. Além disso, a adoção do Nextflow e o uso de contêineres simplifica a execução dos *pipelines*, tornando-a transparente e portável em diversas arquiteturas de computador, incluindo computadores pessoais, servidores com múltiplos processadores e computação em nuvem.

Devido à adoção da linguagem Nextflow, os *pipelines* requerem sistemas POSIX, como Linux e Mac OS X. Mas, também podem ser utilizados no Windows através do WSL2, "Windows Subsystem for Linux". A invocação dos comandos para execução dos *pipelines* deve ocorrer através de um terminal executando um "shell" POSIX (ex: bash, zsh), que são distribuídos de forma padrão em todos os sistemas POSIX acima. Isso significa que não existe uma interface gráfica para o encaminhamento de execução e configuração dos *pipelines*. Em função disso foram elaborados documentos detalhados descrevendo os comandos necessários para execução de cada *pipeline*, os quais encontram-se disponibilizados em seus repositórios GitHub (Tabela 4).

Para executar um *pipeline* no terminal basta indicar o perfil de execução para gerenciamento dos contêineres (Docker ou Singularity) contendo os programas pré-instalados e um arquivo de configuração contendo os parâmetros da corrida (ex: arquivos de sequência). Uma linha de comando genérica que executaria os *pipelines* seria:

```
nextflow run [pipe] -profile docker -c conf.txt
```

Onde as opções são:

```
[pipe]       = nome do pipeline (ver a seguir)
alternativa = -profile singularity
[conf.txt]   = arquivo de configuração
```

Para utilização, os *pipelines* requerem a pré-instalação da ferramenta Nextflow e uma das tecnologias de contêineres, Docker ou Singularity, além da opção de ambientes

virtuais Conda[7] para os *pipelines* de pré-processamento e montagem. Neste documento, serão apresentados os passos para sua utilização com Docker. Para instalação dos *pipelines*, os seguintes passos são necessários:

1. Instalar Java $\geq$ 17 conforme documentação oficial[8]
2. Instalar Docker conforme documentação oficial[9].
   - Para Linux, o modo de execução não-*root* também se faz necessário.
     Os passos estão descritos na documentação oficial[10].
3. Instalar Nextflow

```
curl -s https://get.nextflow.io | bash
```

4. Instalar os *pipelines*

```
nextflow pull fmalmeida/ngs-preprocess
nextflow pull fmalmeida/mpgap
nextflow pull fmalmeida/bacannot
```

Os três últimos comandos geram, na máquina, cópias dos *pipelines* prontos para serem utilizados. Durante a execução, Nextflow automaticamente fará o gerenciamento dos contêineres Docker.

Para uma melhor distribuição e explanação de como utilizar os *pipelines* em conjunto, criamos um depósito em estilo "Quickstart" no repositório Zenodo, uma plataforma pública de propósito geral para compartilhamento de arquivos (de Almeida e Pappas, 2024). Neste repositório, usuários encontram um conjunto de arquivos de parâmetros e linhas de comando pré-configurados, prontos para baixar e executar para realizar uma análise modelo. Este passo a passo será descrito abaixo para exemplificar a operação destes *pipelines*.

[7] https://conda.io/projects/conda/en/latest/user-guide/getting-started.html
[8] https://www.java.com/en/download/help/download_options.html
[9] https://docs.docker.com/engine/install/
[10] https://docs.docker.com/engine/install/linux-postinstall/

O primeiro passo para a execução deste exemplo, é a obtenção dos arquivos do Zenodo que podem ser obtidos através do botão na página web, ou através da linha de comando "`wget https://zenodo.org/records/12205665/files/supporting_files.zip`". Uma vez baixados, este arquivo deve ser descompactado, e gerará um diretório chamado "`zenodo_sup_material`" contendo os seguintes arquivos (Tabela 5):

Tabela 5 - Detalhamento de todos os arquivos disponibilizados no repositório Zenodo, como material suplementar ao artigo Almeida et al. (2023). Estes arquivos são utilizados para a automação de uma análise de demonstração que utiliza os três *pipelines* em conjunto, para replicar a análise descrita no artigo.

Pipeline	Arquivo	Descrição
ngs-preprocess	`input/sra_ids.txt`	Arquivo de entrada listando identificadores SRA para análise
ngs-preprocess	`preprocess-params.yml`	Arquivo YAML com parâmetros pré-configurados
ngs-preprocess	`run_preprocess.sh`	Linha de comando pré-configurada para execução do *pipeline*
MpGAP	`assembly-params.yml`	Arquivo YAML com parâmetros pré-configurados
MpGAP	`assembly.config`	Exemplo de arquivo de configuração Nextflow para personalizar alocação de recursos computacionais
MpGAP	`assembly_samplesheet.yml`	Arquivo "samplesheet" pré-configurado, descrevendo amostra e arquivos de entrada
MpGAP	`run_assembly.sh`	Linha de comando pré-configurada para execução do *pipeline*
Bacannot	`annotation-params.yml`	Arquivo YAML com parâmetros pré-configurados
Bacannot	`annotation_samplesheet.yml`	Arquivo "samplesheet" pré-configurado, descrevendo amostra e arquivos de entrada
Bacannot	`run_annotation.sh`	Linha de comando pré-configurada para execução do *pipeline*

Com estes arquivos disponíveis, os *pipelines* devem ser executados em sequência, para que os arquivos pré-configurados possam usar o resultado de um, como entrada do próximo. Uma análise de demonstração pode ser executada com os seguintes passos, utilizando as seguintes linhas de comando em um terminal[11] usando um shell POSIX (bash, zsh):

1. Execução do *pipeline* de pré-processamento

```
nextflow run fmalmeida/ngs-preprocess -profile docker -r v2.7.1
   -params-file preprocess-params.yml -latest
```

2. Execução do *pipeline* de montagem

```
nextflow run fmalmeida/mpgap -profile docker -r v3.1.4 -c
   assembly.config -params-file assembly-params.yml -latest
```

3. Obtenção do banco de dados pré-formatado para o *pipeline* de anotação usando a URL do Zenodo[12]

```
wget https://shorturl.at/dUItv -O bacannot_db.tgz
tar zxvf bacannot_db.tgz
```

4. Execução do *pipeline* de anotação

```
nextflow run fmalmeida/bacannot -profile docker -r v3.2
   -params-file annotation-params.yml -latest
```

Nas linhas de comando listadas acima, destaca-se alguns parâmetros importantes:

`-r`

para selecionar a versão do *pipeline* a ser executada

[11] Qualquer programa de terminal do Linux, Mac OSX ou Windows WSL2

[12] Caminho completo: `https://zenodo.org/record/7615812/files/bacannot_db_2023_02_07.tar.gz`

`-profile docker`

para selecionar a execução via Docker

`-c`

para passar um arquivo de configuração Nextflow opcional para personalização da execução

`-params-file`

para passar um arquivo em formato JSON[13] ou YAML[14] para modificação de parâmetros dos *pipelines*

`-latest`

para atualizar a versão selecionada do *pipeline* antes da execução

De maneira geral, o tempo de execução dos *pipelines* é intrinsecamente dependente da capacidade do computador (processador, memória) que abriga a paralelização das tarefas. Além disso, os *pipelines* de pré-processamento e montagem são influenciados pela quantidade de dados a serem analisados e pelo número de programas de montagem selecionados pelo usuário. Uma análise de demonstração utilizando 11 Gb de dados de sequenciamento de DNA de uma cepa bacteriana foi executada em um computador pessoal Linux com 4 CPUs (8 núcleos) e 18 Gb de memória RAM (Almeida et al., 2023). Com estes recursos e esta quantidade de dados, esta análise utilizou aproximadamente:

- uma hora para o pré-processamento dos dados brutos;
- 11 horas para a montagem dos genomas, sendo que a tarefa de polimento de montagem consumiu mais de 70% do tempo total de execução; e
- 40 minutos para a anotação do genoma

Ao final deste passo a passo, todos os resultados gerados estarão organizados em três diretórios, nomeados: `01_PREPROCESSED`, `02_ASSEMBLY` e `03_ANNOTATION`. Desta

[13] `https://www.json.org/json-en.html`
[14] `https://yaml.org/`

forma, o usuário irá reproduzir os resultados obtidos e demonstrados no artigo publicado (Almeida et al., 2023). Para uma referência mais detalhada dos resultados e sua interpretação, recomenda-se a utilização do artigo publicado e dos manuais "online" disponibilizados[15][16][17].

Uma vez terminada a execução dos *pipelines*, uma das principais adições deste trabalho é a plataforma web desenvolvida para proporcionar uma interação completa com os resultados do *pipeline* de anotação (Subseção 4.1.5). Esta aplicação web está disponibilizada através de imagens Docker e, após terminar, o *pipeline* Bacannot guarda no diretório contendo os resultados de anotação de cada amostra um *script* chamado `run_server.sh` que automatiza a inicialização e encerramento da aplicação web, respectivamente com os comandos "`./run_server.sh -s`" e "`./run_server.sh -k`". Após inicializada, a aplicação pode ser acessada utilizando um navegador de internet através da URL `http://localhost:3838/`.

Em resumo, o desenvolvimento dos *pipelines* foi delineado de forma a produzir um arcabouço robusto e completo para realização de análises genômicas pré-definidas e padronizadas. Contudo, ainda mantendo a flexibilidade de personalização. Este arcabouço provê uma forma simplificada de instalação e execução, de modo a incentivar sua utilização por usuários menos e mais experientes. O objetivo é que a utilização de tais *pipelines* permita que laboratórios e grupos de pesquisa padronizem suas análises e gastem menos tempo instalando ferramentas e mais tempo interpretando seus resultados.

4.1.7 Comparação com outros pipelines

Após análise da bibliografia corrente e inspeção no repositório GitHub foi possível identificar seis *pipelines* de genômica bacteriana com manutenção ativa e com características comparáveis aos nossos *pipelines*. São eles: ASA3P (Schwengers et al., 2019), TORMES (Quijada et al., 2019), Nullarbor[18], Bactopia (Petit e Read, 2020), Microbe-Annotator (Ruiz-Perez et al., 2021) e MicroPIPE (Murigneux et al., 2021).

[15] `https://ngs-preprocess.readthedocs.io/en/latest/`
[16] `https://mpgap.readthedocs.io/en/latest/`
[17] `https://bacannot.readthedocs.io/en/latest/`
[18] `https://github.com/tseemann/nullarbor`

Embora possuam alguns elementos comuns e módulos de montagem e anotação comparáveis, cada software possui recursos exclusivos adaptados para fins específicos (Tabela A.2, Material Suplementar). No geral, os objetivos e desenhos de cada *pipeline* variam consideravelmente, proporcionando aos usuários opções de alta qualidade para diversos tipos de análises. Destacaremos algumas das características distintivas desses *pipelines*.

Em conjunto, nossos três pipelines fornecem uma análise genômica bacteriana abrangente, desde de dados brutos de sequenciamento até a anotação genômica, semelhante ao que ASA3P, TORMES, Nullarbor e Bactopia oferecem. Em contrapartida, a análise fornecida pelo MicroPIPE e pelo MicrobeAnnotator é bem mais focalizada e menos abrangente. O MicroPIPE é projetado para montagem de genomas e cobre o processo desde a chamada de bases (do inglês, "basecalling") até o polimento do genoma. No entanto, nosso *pipeline* de montagem MpGAP é mais versátil, pois pode acomodar dados de leituras longas (ex: PacBio) que o MicroPIPE não possui, além de permitir uma maior diversidade de montadores e distintas estratégias de montagem. O MicrobeAnnotator, por outro lado, é um *pipeline* que se especializa na anotação metabólica funcional do genoma, usando os bancos KEGG, UniProt, RefSeq e Trembl, e não possuindo módulos extras e mais especializados para anotação de genes de virulência e resistência, como em outros *pipelines*. Também deve ser ressaltado que um de seus módulos, o de anotação de Ortologia KEGG através do Kofamscan, também está disponível no bacannot.

Assim como nossos *pipelines*, o Bactopia é bastante flexível e customizável. Comparado ao bacannot, o Bactopia tem funções adicionais, por exemplo, para análises de pangenoma e filogenéticas. O bacannot não oferece essas ferramentas, mas a padronização de seus resultados permite que usuários os utilizem rapidamente para tais tarefas. Em contrapartida, o Bactopia não possui módulos para a geração de relatórios visuais e ferramentas interativas para uma melhor inspeção dos resultados assim como o bacannot possui.

De maneira geral, quanto aos *pipelines* de anotação, o bacannot se destaca por sua capacidade adicional de anotar várias classes de características genômicas como parte de seu fluxo de trabalho central, incluindo metabólitos secundários, prófagos, ilhas genômicas, elementos integrativos e conjugativos (ICEs) e metilação de DNA, sem exigir execuções adicionais. Esse alcance analítico permite a anotação de características clinicamente re-

levantes, mas fornece também atributos valiosos para isolados não clínicos.

4.1.8 *Exemplos de aplicação dos pipelines*

Após disponibilizados em plataforma pública, estes *pipelines* já se provaram úteis, tendo sido utilizados em outros estudos publicados em colaboração com outros grupos da UnB e também por grupos fora da UnB. Em 2021, foram utilizados para a análise de 5 isolados multirresistentes de *Klebsiella variicola* que haviam sido primeiramente identificados como *K. pneumoniae* (Campos et al., 2021).

Também em 2021, foi realizada a análise de uma linhagem de *K. pneumoniae* multirresistente isolada no lago Paranoá (Brasília-DF), na qual foi identificado um novo elemento genético *non-Tn4401* mobilizando uma carbapenemase *KPC* (Janssen et al., 2021). Já em 2023, os *pipelines* foram utilizados para a montagem e anotação dos genomas de bactérias ambientais. A primeira, uma bactéria classificada como uma nova espécie, *Novosphingobium terrae*, isolada dos solos do Cerrado (Belmok et al., 2023) e a segunda, uma linhagem de *Pantoea stewartii*, isolada de folhas de uma árvore de castanha do Brasil (Rocha et al., 2023). Finalmente, no final de 2023, um grupo externo, independente de colaboração com nosso grupo, publicou uma análise do potencial adaptativo de *Arabidopsis thaliana* em resposta à *Pseudomonas syringae* onde o *pipeline* de anotação foi utilizado (Bartoli et al., 2023).

4.2 Análise dos isolados clínicos de *K. pneumoniae*

Objetivando demonstrar a aplicabilidade dos *pipelines* e ao mesmo tempo direcionar o desenvolvimento dos mesmos, focamos no problema da resistência a antimicrobianos. Em particular, analisamos três isolados bacterianos (denominados KpBSB56, KpBSB60 e ECR) pertencentes à coleção do Laboratório de Análises Moleculares de Patógenos da UnB, obtidos de diferentes fontes de isolamento de infecções de diferentes pacientes no Hospital Universitário de Brasília em 2021 (Tabela 3).

Através do sistema MicroScan (Beckman Coulter™), os isolados foram microbiologicamente identificados como *K. pneumoniae* e demonstraram resistência para quase

todos os antimicrobianos testados, sendo a ECR suscetível ao antimicrobiano Amicacina e KpBSB60 a Gentamicina (Tabela 6). Além disso, todas testaram positivas para produção de carbapenemase e ESBL (do inglês, "Extended Spectrum Beta-Lactamase"). Por fim, o teste de Concentração Mínima Inibitória (MIC, do inglês "Minimum inhibitory concentration") utilizado para determinar a menor concentração de um antimicrobiano necessária para inibir o crescimento de um microrganismo confirmou resistência a Polimixina B em todos os isolados. Desta forma, todas as linhagens foram caracterizadas como extensivamente resistentes (XDR, do inglês "Extensively drug resistant"), como definido por Magiorakos et al. (2012) e de acordo com as diretrizes do Centro Europeu de Prevenção e Controle de Doenças (ECDC) e do Centro de Controle e Prevenção de Doenças (CDC).

Tabela 6 - Resultados de ensaios experimentais de identificação microbiana e suscetibilidade a antimicrobianos realizados através do sistema MicroScan (Beckman Coulter™) e método de Concentração Inibitória Mínima conforme recomendado pelo CLSI[19] (para Polimixina-B)

Isolados	Espécie	Meropenem	Imipenem	Amicacina	Gentamicina	Ceftazidima / Avibactam	Polimixina-B	Carbapenemase	Metalo-β-lactamase
ECR	*Klebsiella pneumoniae*	R	R	S	R	R	R	Positivo	Não testado
KpBSB56	*Klebsiella pneumoniae*	R	R	R	R	R	R	Positivo	Negativo
KpBSB60	*Klebsiella pneumoniae*	R	R	R	S	R	R	Positivo	Negativo

Diferentemente de estudos que utilizam poucos marcadores e técnicas com menor resolução, empregamos o sequenciamento completo de genomas para investigar, com resolução em nível de bases, determinantes de resistência e outros elementos de importância clínica, como fatores de virulência e plasmídeos. A análise destes dados será detalhada nas próximas seções.

4.2.1 Montagem dos genomas

Visando obter genomas de alta qualidade e ao mesmo tempo tentando maximizar a contiguidade destes, utilizou-se uma combinação de tecnologias de sequenciamento de leituras curtas (fornecendo alta cobertura e baixa taxa de erros) e de leituras longas para garantir esta contiguidade.

Para isso, foram utilizadas tecnologias de sequenciamento BGI/DNBseq para gerar leituras curtas e Oxford Nanopore/MinION para gerar leituras longas, produzindo respectivamente $\geq$800X e $\geq$30X de cobertura de dados brutos por linhagem. As leituras curtas foram recebidas já pré-processadas (sequências adaptadoras removidas) pela em-

presa de sequenciamento BGI. Já as leituras longas foram pré-processadas com o *pipeline* ngs-preprocess para remoção de sequências adaptadoras e filtragem por tamanho (≥750) e qualidade (≥Q10, taxa de erro de 10%), resultando em leituras processadas que foram utilizadas para a estratégia de montagem híbrida de genomas.

Através do *pipeline* MpGAP, as leituras longas e curtas foram utilizadas em conjunto para a realização da montagem híbrida dos genomas dos isolados. Nesta abordagem as leituras longas são utilizadas para construção de um arcabouço do genoma, mas que contem erros de nomeação de bases inerentes à tecnologia. Sobre este arcabouço, as leituras curtas (de baixa taxa de erros) são utilizadas para a correção de erros de sequenciamento, em um procedimento chamado polimento de genoma.

A execução do *pipeline* com esta estratégia de montagem resultou em genomas de alta qualidade para as linhagens ECR e KpBSB56 em termos de acurácia, contiguidade e conteúdo genético esperado, conforme pode ser observado na Tabela 7. Para estes dois genomas, a alta contiguidade idealizada durante desenho experimental foi de fato observada, apresentando resolução do cromossomo bacteriano em nível de sequência única, como evidenciado pelo tamanho do maior contig gerado que possui o tamanho esperado do cromossomo para a espécie *Klebsiella pneumoniae* (Tabela 7).

Tabela 7 - Estatísticas gerais da montagem de genomas. Dentre as estatísticas nas colunas, o N50 é uma medida de continuidade da montagem, indicando o tamanho do contig mais curto cuja soma dos tamanhos dos contigs maiores ou iguais a ele representa pelo menos 50% do tamanho total. A métrica BUSCO (Manni et al., 2021) avalia a completude da montagem com base na presença de genes conservados universalmente, mostrando a porcentagem de genes ortólogos esperados que estão completos.

Linhagem	N. Contigs	Tamanho Total (pb)	Maior contig	N50	BUSCOs completos (%)
ECR	4	5.584.349	5.375.438	5.375.438	98,65
KpBSB56	8	5.718.898	5.156.478	5.156.478	98,65
KpBSB60	64	5.715.670	504.103	363.387	98,65

Porém, a montagem com a estratégia híbrida para a cepa KpBSB60 foi marcadamente diferente e resultou em uma montagem muito fragmentada com ≥200 contigs (dado não mostrado), o que não é esperado. Isto muito provavelmente está relacionado a problemas que foram observados durante a extração de DNA e sequenciamento das leituras longas que resultaram em baixa cobertura (≈20X) e leituras não tão longas (750 a 1200

pares de base). Por isso, os dados da KpBSB60 foram montados novamente, utilizando uma estratégia diferente, onde primeiro as leituras curtas foram montadas em diversos contigs de alta acurácia através do *pipeline* MpGAP e, em seguida, independentemente do *pipeline*, estes contigs foram rearranjados e concatenados em uma etapa chamada "scaffolding" utilizando as leituras longas disponíveis através da ferramenta LongStich (Coombe et al., 2021). Com esta estratégia, gerou-se um genoma de melhor contiguidade que anteriormente, apesar de estar aquém em termos de qualidade do que foi obtido para as outras linhagens (Tabela 7).

Apesar de tudo, demonstra-se que mesmo não sendo possível gerar montagens híbridas de forma direta em alguns casos, ainda assim é possível, através de diferentes metodologias, beneficiar-se da alta acurácia das leituras curtas e do alcance das leituras longas, produzindo genomas acurados de maior contiguidade.

Em posse destes genomas, procuramos melhor caracterizar estas montagens através de uma predição *in silico* de plasmídeos para identificar e separar contigs cromossômicos e de plasmídeos. Para isso, utilizamos a ferramenta Platon (Schwengers et al., 2020), que identificou coletivamente entre os genomas, a presença de 29 contigs plasmidiais (Tabela 8), permitindo a separação de sequências. No geral, foi possível obter seis sequências completas de plasmídeos, provenientes das linhagens ECR e KpBSB56. Apesar das tentativas de reconstruir os plasmídeos da KpBSB60, através de ferramentas especializadas como MOB-Suite (Robertson e Nash, 2018) e PLASMe (Tang et al., 2023), não foi possível resolvê-las e tampouco determinar a sequência completa de seus plasmídeos.

Por fim, obteve-se três genomas completos e de boa qualidade, sendo dois circularizados a nível de cromossomo, para prosseguimento da anotação e utilização em análises comparativas.

Tabela 8 - Detecção *in silico* de contigs das linhagens deste estudo que são provenientes de plasmídeos. Classificação de contigs "cromossômicos" e "plasmidiais" foi realizada através da ferramenta Platon (Schwengers et al., 2020).

Linhagem	Número de contigs plasmidiais	Contigs	Tamanho em pares de base
ECR	3	contig_2	6.261
		contig_3	83.939
		contig_4	118.711
KpBSB56	7	contig_2	222.389
		contig_3	261.646
		contig_4	3.843
		contig_6	43.590
		contig_7	9.270
		contig_8	9.076
		contig_9	12.606
KpBSB60	19	scaffold14	108.453
		scaffold20	53.149
		scaffold21	44.496
		scaffold22	40.867
		scaffold24	33.739
		scaffold25	22.927
		scaffold26	21.973
		scaffold30	10.951
		scaffold31	9.294
		scaffold32	9.007
		scaffold34	6.967
		scaffold36	6.687
		scaffold38	6.346
		scaffold39	6.026
		scaffold42	4.438
		scaffold43	3.549
		scaffold44	3.349
		scaffold45	3.070
		scaffold47	2.809

4.2.2 Conservação genômica

Com o intuito de investigar a conservação de genoma entre as linhagens, selecionou-se as sequências cromossômicas de cada um dos genomas. Em seguida, calculou-se a média da identidade de nucleotídeos (ANI; do inglês, "Average Nucleotide Identity") entre os cromossomos. No geral, esta análise revelou altíssima similaridade de sequência entre os cromossomos das linhagens com valores $\geq$99,7. Somado ao resultado BUSCO, este resultado ANI indica que, apesar da fragmentação, o genoma da KpBSB60 está bem completo em termos de conteúdo.

Realizou-se também um estudo da colinearidade entre as linhagens com cromossomos completos resolvidos (KpBSB56 e ECR) para verificação da colinearidade dos mesmos. O resultado desta análise com o programa MCscan revelou uma alta conservação dos blocos, mas com diversas interrupções curtas na colinearidade destes, indicando alguns pontos de divergência entre estas linhagens (Figura 11). Esta alta equivalência entre as linhagens corrobora os resultados de conservação entre os genomas.

Para confirmar que estas observações não são erros de montagem, utilizamos a ferramenta CRAQ para avaliar a qualidade da montagem dos genomas e a ferramenta circlator para avaliar a circularização da sequência. A análise demonstrou que, para as duas linhagens, os cromossomos estão corretamente circularizados e de boa qualidade conforme o índice de qualidade de montagem (AQI, do inglês "Assembly Quality Index") proposto pelos autores da ferramenta CRAQ (Li et al., 2023). Para os dois cromossomos, observou-se um valor AQI de 100 no meio da sequência e de 90,48 nas bordas, sendo que AQI $>$ 90 representa qualidade excepcional. Desta forma, confirma-se que as observações são legítimas.

Uma análise inicial dos pontos de interrupção de colinearidade dos blocos identificou a presença de genes como transposases, recombinases, integrases e outros envolvidos na resposta ao dano no DNA. Uma análise mais detalhada foi realizada para a região de quebra mais evidente, assinalada com uma seta na Figura 11A. Ainda com a ferramenta MCScan, realizou-se uma análise de sintenia guiada para visualizarmos o conteúdo gênico nesta região, a qual revelou que a grande interrupção da colinearidade entre os cromossomos é causada pela inserção de duas transposases IS3/ISKpn18 no genoma da

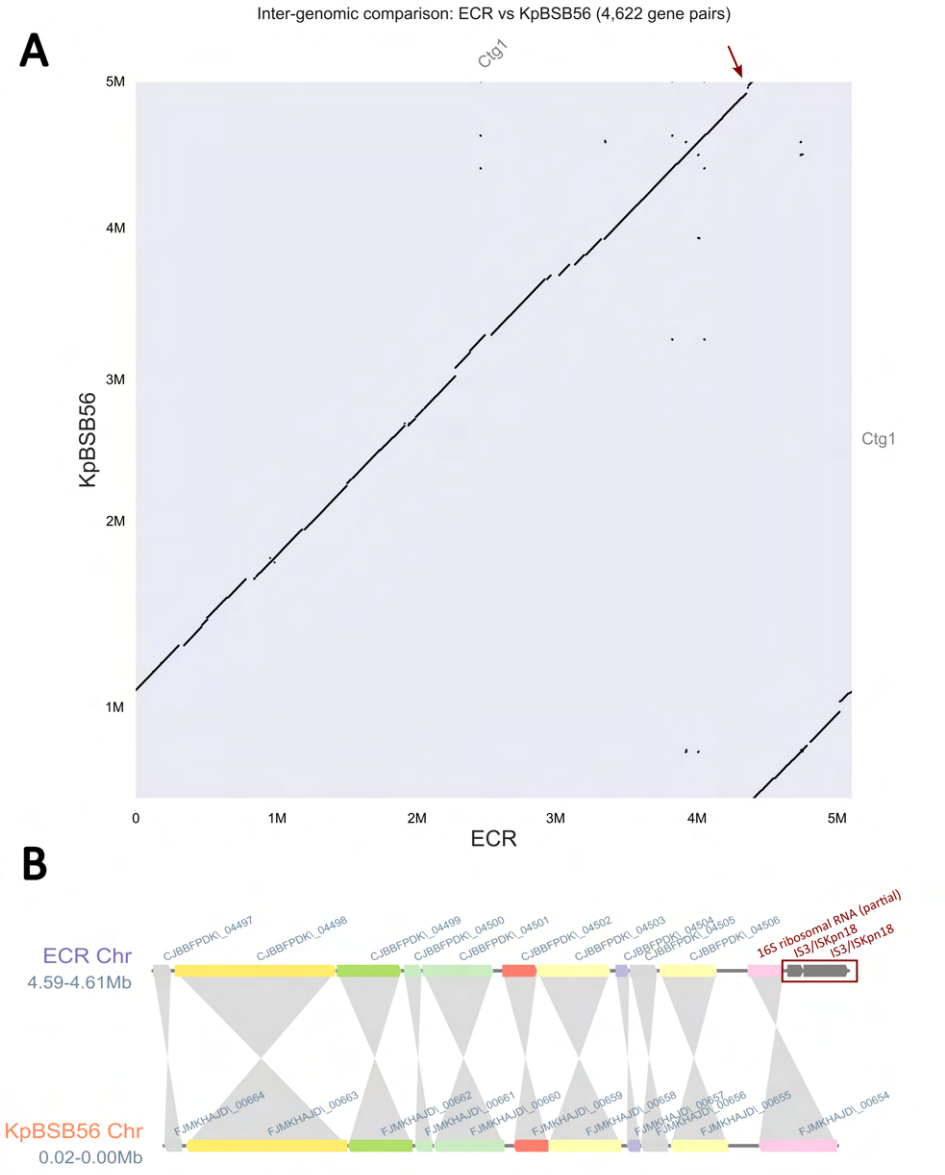

Figura 11: Análise de blocos genômicos colineares entre os genomas da KpBSB56 e ECR. A) *dotplot* do alinhamento das sequências dos cromossomos da KpBSB56 e ECR. Neste destaca-se com uma seta vermelha o ponto de interrupção de colinearidade mais chamativo. B) Detalhamento dos genes adjacentes ao ponto de interrupção destacado. Destaca-se em vermelho os genes encontrados exatamente no ponto de quebra. Os demais genes presentes na figura que fazem parte da região conservada anterior ao ponto de quebra são: hypothetical protein (CJBBFPDK_04497); Cation efflux system protein CusA (CJBBFPDK_04498); Cation efflux system protein CusB (CJBBFPDK_04499); Cation efflux system protein CusF (CJBBFPDK_04500); Cation efflux system protein CusC (CJBBFPDK_04501); Transcriptional regulatory protein CusR (CJBBFPDK_04502); Sensor histidine kinase CusS (CJBBFPDK_04503); hypothetical protein (CJBBFPDK_04504); hypothetical protein (CJBBFPDK_04505); 50S ribosomal protein L16 3-hydroxylase (CJBBFPDK_04506). Resultados foram produzidos através da ferramenta MCscan (Tang et al., 2008).

ECR (Figura 11B). Nesta mesma figura é possível observar a conservação de genes entre as linhagens antes das transposases, incluindo dois operons envolvidos na detoxificação de cobre (cusRS e cusCFBA) (Zulfiqar e Shakoori, 2012). No entanto, o sítio de inserção das transposases no genoma da ECR foi dentro de uma das cópias do rRNA 16S, que foi truncada nesta linhagem.

De forma geral, estas observações sugerem que possam ter ocorrido eventos de movimentação e recombinação do DNA levando ao padrão observado na Figura 11. Interessantemente, nas regiões não-sintênicas, que são aquelas encontradas somente em uma das linhagens, foram detectados elementos integrativos, relaxases, integrases e prófagos. Em resumo, é interessante notar que, apesar de serem provenientes do mesmo Hospital, estas duas linhagens acomodam divergências (Figura 11), sugerindo um processo contínuo de evolução mediado por elementos transponíveis.

4.2.3 Anotação dos genomas

Todos os genomas foram anotados através do *pipeline* bacannot conforme descrição nos métodos (Tabela 9). Os resultados de distância genética obtidos através do *pipeline* corroboram a identificação microbiológica, indicando serem da espécie *Klebsiella pneumoniae*. Em termos de classes de genes, a anotação genômica apresentou números similares, salvo os resultados de rRNA que mostraram uma grande diferença para a KpBSB60 se comparada às demais (Tabela 9). Isto se dá muito provavelmente pelo seu genoma ser embasado nas leituras curtas conforme descrito anteriormente. Por serem menores, as leituras curtas geram montagens fragmentadas que não perpassam a integralidade dos operons de rRNA, o que dificulta sua resolução e impede a distinção de suas múltiplas cópias (de Oliveira Martins et al., 2019).

O *pipeline* bacannot traz suporte automático para a identificação dos grupos clonais de *K. pneumoniae* através da análise computacional de MLST. Entretanto, o *pipeline* não está restrito a esta espécie e permite classificar qualquer espécie presente no banco de dados PubMLST (Jolley et al., 2018).

Tabela 9 - Resumo das características gerais de anotação dos genomas das linhagens deste estudo. As anotações de serotipo K e O foram obtidas através da ferramenta Kleborate (Lam et al., 2021) e as informações de ST, número total de genes e sequências de rRNA e tRNA e a distância genômica para o genoma referência mais próximo, obtidos através do *pipeline* bacannot.

Linhagem	ST	Serotipo K	Serotipo O	CDS	rRNA	tRNA	Genoma mais próximo	Nome da referência	Distância genômica
ECR	11	105	O1/O2v2	5.241	26	86	GCF_000529325.1	*Klebsiella pneumoniae* IS33	0,00257170
KpBSB56	11	3	O1/O2v2	5.447	25	88	GCF_000529325.1	*Klebsiella pneumoniae* IS33	0,00768918
KpBSB60	11	64	O1/O2v1	5.454	8	79	GCF_000445405.1	*Klebsiella pneumoniae* JM45	0,00469623

Os resultados de MLST classificam todas as linhagens analisadas como pertencentes ao grupo clonal ST11 (Tabela 9), um grupo de alto risco de isolados multirresistentes produtores de CTX-M 14 e 15 que se alastrou pelo mundo (Navon-Venezia et al., 2017; He et al., 2022). Este serotipo, parte do grupo clonal 258 (CG258), é considerado uma das principais ameaças à saúde pública global devido à sua relevância para a disseminação de resistência, principalmente através de plasmídeos (Wyres et al., 2015; Nakamura-Silva et al., 2022). A detecção de isolados ST11 é sempre preocupante devido aos constantes relatos de surtos causados por linhagens deste serotipo e sua recente ligação à convergência de fatores de resistência e virulência (Zhan et al., 2017; Gu et al., 2018; Xie et al., 2021; Liu et al., 2022; Mendes et al., 2022; Nicola et al., 2022).

Além disso, estas linhagens foram também avaliadas quanto à seu sorotipo K, referente a cápsula polissacarídea (CPS) e seu sorotipo O, referente ao lipopolissacarídeo (LPS) através da execução manual da ferramenta Kleborate (Lam et al., 2021), a qual não foi integrada ao *pipeline* bacannot, em virtude de poder ser aplicada exclusivamente à espécie *Klebsiella pneumoniae*. Na Tabela 9 pode ser verificado que dentre as nossas linhagens nenhuma foi anotada como K1 ou K2, serotipos comumente associados à hipervirulência (Zhu et al., 2021). Dentre elas, destaca-se a detecção de uma linhagem ST11-KL64 (KpBSB60) que é um grupo de bactérias com distribuição mundial e considerado de extrema importância na expansão de linhagens resistentes a carbapenêmicos (Wang et al., 2023).

Além disso, todas as linhagens apresentaram serotipo O1/O2 que estão entre os se-

rotipos mais comuns associados a infecções globalmente (Choi et al., 2020; Wantuch et al., 2023). Estas classificações, somadas à classificação ST, provêm uma maior resolução na categorização de linhagens para embasar expectativas quanto ao potencial de resistência e virulência (Follador et al., 2016; Choi et al., 2020; Lam et al., 2021; Zhu et al., 2021; Wantuch et al., 2023).

4.2.4 Anotação de fatores de virulência

Fatores de virulência são compostos bacterianos que contribuem para a capacidade de um patógeno de causar infecção. Estes podem estar diretamente relacionados ao aumento do "fitness" bacteriano durante a infecção, mediando eventos como a adesão, invasão celular, colonização, captação de ferro e formação de biofilme (Ho et al., 2014; Coulthurst, 2019; Paczosa e Mecsas, 2016; Zhu et al., 2021; Gorrie et al., 2022). Por isso, visamos também identificar os genes de virulência presentes nos genomas analisados.

Diversos fatores de virulência foram detectados nestes genomas, com todos os genes anotados presentes no cromossomo bacteriano (Tabela 10). Alguns destes fatores como os antígenos K (CPS) e O (LPS) são capazes de provocar respostas imunes no hospedeiro e, por possuírem composição variável, são classificados em sorotipos que representam estas diferentes composições, e foram abordados na Subseção 4.2.3 (Tabela 9). Os demais fatores de virulência não relacionados a serotipagem encontram-se disponíveis na Tabela 10. No geral, a anotação de virulência é bastante conservada e similar entre as linhagens. Dentre os fatores de virulência detectados, a enterobactina, o sistema de secreção tipo VI (T6SS) e as fímbrias do tipo 1 e 3 foram encontradas em todos os genomas. A enterobactina, que permite a captação de ferro, é um sideróforo extremamente prevalente na espécie e considerado intrínseco.

Além disso, outros genes de virulência menos prevalentes também foram detectados nos genomas. Por exemplo, todos os isolados codificam pelo menos um sideróforo adicional além da enterobactina. Esta é uma característica relevante pois, a presença de outros sideróforos é geralmente relacionada à fenótipos de virulência aumentada (Martin e Bachman, 2018; Paczosa e Mecsas, 2016; Zhu et al., 2021; Russo et al., 2021; Gorrie et al., 2022). A KpBSB60 codifica em seu genoma o lócus da Aerobactina, incluindo os

genes *iucABCD* e *iutA* enquanto a ECR e KpBSB56 codificam o lócus completo da Yersiniabactina, detectada em um transposon ICEkp (Tabela 10). A Yersiniabactina é um sideróforo capaz de mediar o escape da resposta inflamatória e potencializar o crescimento e infecção bacteriana (Lam et al., 2018; Zhao et al., 2022). O ICEkp é um elemento autotransmissível frequentemente observado em *Klebsiella pneumoniae* e que pode carregar outros fatores de virulência, além do lócus *ybt*, sendo considerado um dos mecanismos de transferência de virulência mais importantes na espécie (Lam et al., 2018; Farzand et al., 2019; Zhao et al., 2022).

Em termos gerais, esta observação é preocupante pois a presença de sideróforos adicionais, particularmente Aerobactina e Yersiniabactina, é uma característica tipicamente encontrada em isolados de virulência aumentada ou hipervirulentos (Paczosa e Mecsas, 2016; Wyres et al., 2020; Zhao et al., 2022; Gorrie et al., 2022). Apesar da detecção destes variados fatores de virulência, inclusive de sideróforos "não clássicos" sugerir um fenótipo de virulência aumentada, análises experimentais são necessárias para sua mensuração. Além disso, é importante ressaltar que, assim como em algumas de nossas linhagens, isolados de *K. pneumoniae* podem codificar genes *iroE* (Salmochelina) e *iutA* (Aerobactina) em seus cromossomos e, a presença destes genes sozinhos sem o restante do lócus, não é suficiente para observação do fenótipo (Russo e Marr, 2019). Finalmente, os genes reguladores da hipermucoviscosidade *rmpA* e *rmpA2*, que são também usados como indicadores da capacidade de causar infecções comunitárias (De Campos et al., 2018), não foram detectados em nenhum dos genomas.

Tabela 10 - Detecção *in silico* de fatores de virulência codificados nos genomas das linhagens deste estudo. A anotação foi realizada utilizando o banco de dados VFDB (Chen et al., 2005) através do *pipeline* bacannot.

Linhagem	AcrAB	RcsAB	Enterobactina	Aerobactina	Salmochelina	Yersiniabactina	ICEKp	T6SS	Fímbria tipo 1	Fímbria tipo 3
ECR	*acrAB*	*rcsAB*	*entABCEFS*, *fepABCDG*, *fes*		*iroE*	*fyuA*, *irp1*, *irp2*, *ybtAEPQSTUX*	ICEKp4	*tssBCDFGHJKL*, *ompA*	*fimABCDEFGHIK*	*mrkABCDFHIJ*
KpBSB56	*acrAB*	*rcsAB*	*entABCEFS*, *fepABCDG*, *fes*	*iucABCD*, *iutA*	*iroE*	*fyuA*, *irp1*, *irp2*, *ybtAEPQSTUX*	ICEKp3	*tssBCDFGHJKLM*, *ompA*	*fimABCDEFGHIK*	*mrkABCDFHIJ*
KpBSB60	*acrAB*	*rcsAB*	*entABCEFS*, *fepABCDG*, *fes*	*iutA*	*iroE*			*tssBCDFGHJKL*, *ompA*	*fimABCDEFGHIK*	*mrkABCDFHIJ*

4.2.5 Anotação de genes de resistência

Corroborando os resultados experimentais de suscetibilidade a antimicrobianos das linhagens estudadas (Tabela 6), a anotação computacional de genes de resistência detectou, em todas, a presença de genes de resistência para mais de 5 categorias de antimicrobianos, distribuídos entre seus plasmídeos e cromossomos (Tabela 11), ratificando a sua classificação como extensivamente resistentes (XDR). No geral, além dos genes de resistência considerados intrínsecos à espécie (*bla*SHV, *oqx*AB, e *fos*A), a anotação entre os genomas é bem similar, com diversos genes de resistência adquiridos e com quase todas as classes de genes de resistência a antimicrobianos sendo encontradas nas três linhagens.

Interessantemente, as linhagens ECR e KpBSB60 apresentaram suscetibilidade comprovada experimentalmente à Amicacina e Gentamicina, respectivamente, apesar de possuírem genes de resistência a estas classes de drogas. Recentemente, um estudo mostrou que algumas comunidades microbianas podem apresentar genes de resistência chamados "silenciosos", significando que o fenótipo não é observado por diferentes motivos, caracterizados ou não (Deekshit e Srikumar, 2022). Os autores discutem que estes genes podem ter papel importante na adaptabilidade à ambientes clínicos já que diversos fatores podem influenciar a reativação destes genes, como por exemplo, a inserção de sequências de inserção (IS, do inglês "Insertion Sequence"), na região "upstream" destes genes (Deekshit e Srikumar, 2022).

Quanto às β-lactamases, detectou-se em todas as linhagens pelo menos uma ESBL e uma carbapenemase. Todas as linhagens possuem genes das β-lactamases *bla*CTX-M-15, *bla*OXA, *bla*TEM-1 e *bla*NDM-1 em seus genomas. Adicionalmente, detectou-se o gene da *bla*KPC-2 no genoma da KpBSB60, localizada em um plasmídeo, na região *downstream* a uma resolvase de transposon Tn3 e *upstream* a uma transposase ISKpn6 (família da transposase IS1182), sugerindo uma possível mobilidade deste gene através de transposons Tn3.

Tabela 11 - Consenso da anotação *in silico* de genes e mutações de resistência detectadas nos genomas das linhagens deste estudo. Anotação foi realizada utilizando as ferramentas AMRFinderPlus e Resfinder através do *pipeline* bacannot.

Linhagem	Aminoglicosídeo	Betalactâmicos	Diaminopirimidina	Fluoroquinolona	Glicopeptídeo	Macrolídeo	Lincosamida	Ácido fosfônico	Sulfonamida	Tetraciclina	Mutações
ECR	aac(3)-IId, aac(6')-Ib, aac(6')-Ib-cr, aadA1, aph(3")-Ib, aph(3')-VI, aph(6)-Id	blaNDM-1, blaOXA-1, blaOXA-9, blaSHV-11, blaCTX-M-15, blaTEM-1	oqxAB	oqxAB, qnrS1	ble			fosA	sul2	oqxAB, tet(D)	acrR, OmpK36, OmpK37, GyrA, ParC
KpBSB56	aac(3)-II, aac(6')-Ib-cr, aadA1, aadA2, aph(3')-VIa	blaNDM-1, blaOXA-1, blaSHV-11, blaCTX-M-15, blaTEM-1	oqxAB, dfrA12	oqxAB, qnrB1	ble			fosA	sul1, sul2	oqxAB	OmpK35, OmpK36, GyrA, ParC
KpBSB60	aadA1, aadA5, aph(3')-VI	blaKPC-2, blaNDM-1, blaOXA-9, blaSHV-11, blaCTX-M-15, blaTEM-1	oqxAB, dfrA17	oqxAB, qnrS1	ble	mph(A), erm(B)	erm(B)	fosA	sul1, sul2	oqxAB, tet(A)	OmpK35, OmpK36, GyrA, ParC

Recentemente, detectou-se no lago Paranoá uma linhagem ambiental de *K. pneumoniae* (KpV3) que possuía o gene *bla*KPC-2 associado a transposons da mesma família (Janssen et al., 2021). Isso motivou a realização de análises para contrastar as estruturas dos blocos sintênicos que flanqueiam este gene nas nossas linhagens comparando a linhagem ambiental (KpV3) e mais quatro linhagens hospitalares isoladas no estado de São Paulo. Os resultados são mostrados na Figura 12 e mostram que a única característica comum a todas as linhagens é a presença de uma transposase da família IS1182 logo após o gene KPC-2, mas na fita oposta (orientação *tail-to-tail*). A maior conservação da linhagem KpV3 é com a KpBSB60 que compartilha um pequeno bloco incluindo a resolvase de transposon Tn3 antes do *bla*KPC-2 e IS1182.

A Figura 12 revela um comportamento interessante de linhagens com o gene *bla*KPC-2 do DF. Comparada a linhagem analisada neste estudo (KpBSB60), a linhagem KpBSB31, que foi isolada anteriormente no mesmo hospital terciário (De Campos et al., 2018), não compartilha outros genes na região, a não ser os previamente indicados (*bla*KPC-2 e IS1182). No entanto, a linhagem KpBSB31 apresenta uma significativa conservação na mesma região com as linhagens isoladas em hospitais de Tocantins em 2019 e São Paulo em 2021 (Figura 12).

Em resumo, juntas, as linhagens apresentam potencial genético de resistência a virtualmente todas as β-lactamases conhecidas, ressaltando a propensão de linhagens ST11 em acumular genes e mecanismos de resistência (Lee et al., 2016; He et al., 2022). Relatos de isolados carreando múltiplas carbapenemases têm se tornado mais frequente em todo o mundo (Gao et al., 2020; Guo et al., 2023). Estas combinações potencializam os níveis de resistência a carbapenêmicos, dificultando ainda mais o gerenciamento de infecções em ambientes hospitalares, e ainda, ampliam a preocupação quanto à disseminação de genes de resistência e surgimento das chamadas "superbactérias" (Bedenić et al., 2023; Yuan et al., 2024).

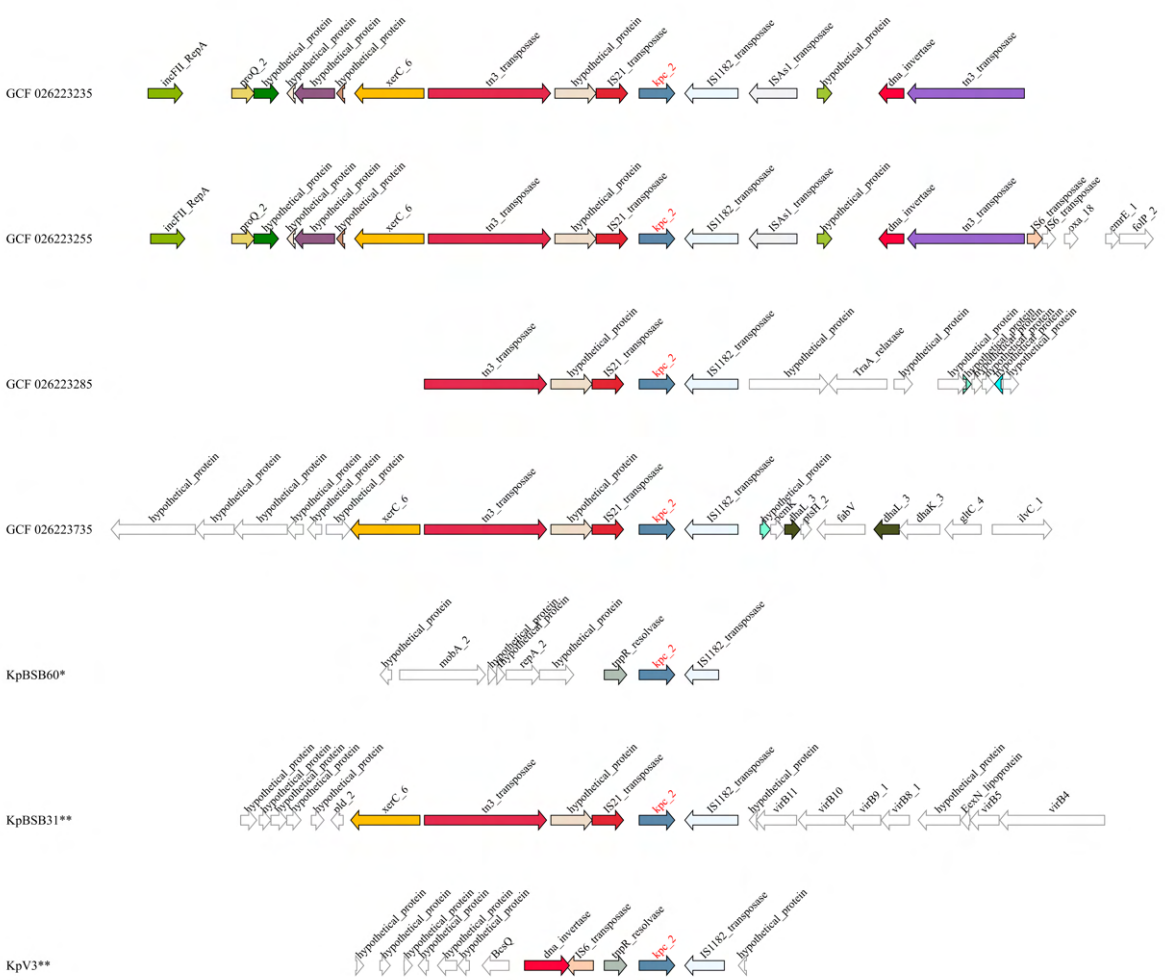

Figura 12: Representação gráfica comparativa do *cluster* gênico da KPC-2 detectado na linhagem KpBSB60 e em outras linhagens brasileiras, gerada através da ferramenta GCluster (Li et al., 2020). Genes considerados conservados entre os genomas são coloridos, baseado em homologia de sequência. Em fonte de texto vermelha, destaca-se o gene que foi utilizado para centralizar a análise e estender o lócus a ser comparado em até 10 genes em cada flanco. Na figura, os genomas marcados com * são provenientes deste estudo, e os marcados com ** são genomas mais antigos também analisados por nosso grupo. As demais linhagens são provenientes dos estados de São Paulo (GCF_026223235; GCF_026223255, GCF_026223735) e Tocantins (GCF_026223285).

4.2.6 Anotação dos plasmídeos

Em posse das informações que nos permitem separar contigs de plasmídeos e de cromossomos (Subseção 4.2.1) visamos aumentar a resolução de nossas anotações quanto aos plasmídeos para identificar os grupos de incompatibilidade ("Inc typing") presentes nas linhagens e categorizar quais dos genes de resistência e virulência discutidos nas seções anteriores estão sendo movimentados por plasmídeos.

Quanto aos genes de interesse clínico, não foi detectado nos plasmídeos, nenhum dos genes de virulência apresentados anteriormente. Em contrapartida, a maioria dos genes de resistência a antimicrobianos não-intrínsecos foram detectados nas sequências de plasmídeos (Tabela 12). Observa-se na tabela a presença de diversas $\beta-$lactamases em plasmídeos, sugerindo a importância dos mesmos no co-carreamento de múltiplos destes genes.

Utilizando o programa PlasmidFinder (Carattoli et al., 2014), que permite a anotação de grupos de incompatibilidade através da detecção de sequências de replicons, identificou-se 19 grupos diferentes em 15 contigs, indicando a presença de plasmídeos com múltiplos replicons (Tabela 12), os quais vêm se mostrando importante disseminadores de resistência em bactérias Gram-negativas (Wang et al., 2021). Nas linhagens ECR e KpBSB56 por exemplo, foram detectados contigs com mais de um replicon carreando diversos genes de resistência a antimicrobianos (Tabela 12). A detecção de plasmídeos com estas características levantam alertas uma vez que estes podem desempenhar papel chave na sobrevivência bacteriana em ambientes de alta pressão de seleção, como hospitais (Wang et al., 2021).

Tabela 12 - Resumo da anotação *in silico* de grupos de incompatibilidade e de genes de resistência a antimicrobianos identificados nos plasmídeos preditos. Os grupos de incompatibilidade foram anotados através da ferramenta PlasmidFinder (Carattoli et al., 2014). Já os genes de resistência são os mesmos detalhados na Tabela 11.

Linhagem	Contig	Grupos de incompatibilidade	Genes de resistência
ECR	contig_2		blaCTX-M-15
	contig_3	IncR	aac(3)-IId, aac(6')-Ib-cr, aph(3")-Ib, aph(6)-Id, blaCTX-M-15, blaOXA-1, blaTEM-1, tet(D), sul2
	contig_4	IncFII(K), IncFIB(pQil)	aadA1, aac(6')-Ib, aph(3')-VI, blaCTX-M-15, blaNDM-1, blaOXA-9, blaTEM-1, ble, qnrS1
KpBSB56	contig_2	IncFIB(K)	aadA1, sul1
	contig_3	IncR, IncFII(pRSB107)	aac(3)-II, aac(6')-Ib-cr, aadA2, blaCTX-M-15, blaOXA-1, blaNDM-1, blaTEM-1, ble, dfrA12, qnrB1, sul1, sul2
	contig_7	ColRNAI	
	contig_9	Col156, IncQ1	aph(3')-VIa
KpBSB60	scaffold14	IncFIB(pKPHS1)	
	scaffold20	IncFII(K)	
	scaffold21	IncFIB(pQil)	blaTEM-1
	scaffold22	IncC	
	scaffold24	IncFIA	
	scaffold25	IncFIB(AP001918), IncFII(pRSB107)	
	scaffold26		tet(A)
	scaffold30	IncQ1	blaKPC-2
	scaffold31	ColRNAI	
	scaffold32		sul2
	scaffold34		blaCTX-M-15, qnrS1
	scaffold36		aph(3')-VI, blaNDM-1, ble
	scaffold38		aadA1, aadA5, blaOXA-9, dfrA17, sul1
	scaffold39		mph(A)
	scaffold42		erm(B)
	scaffold43	Col440I	

4.2.7 Genômica comparativa com isolados brasileiros

Além dos resultados de anotação para as três linhagens estudadas, procuramos também realizar análises de genômica comparativa para entender o contexto destes achados com outras linhagens brasileiras (Tabela A.1, Material Suplementar). Devido ao cenário observado das nossas anotações, utilizamos os seguintes dados genômicos disponíveis em bancos de dados públicos para efeito de comparação:

- linhagens *Klebsiella pneumoniae* produtoras de NDM identificadas no Brasil (Camargo et al., 2022) de 2015 a 2021 para permitir uma contextualização mais aprofundada do gene;
- um conjunto de linhagens MDR isoladas em Brasília para comparação temporal de genomas da mesma região (Lee et al., 2021) e;
- duas linhagens mais antigas, KpBSB31 (De Campos et al., 2018) e KpV3 (Janssen et al., 2021), também analisados por nosso grupo para uma comparação local

4.2.7.1 Análise filogenômica

Para entender melhor como nossas linhagens se relacionam com estas outras linhagens incluídas realizou-se uma análise filogenômica entre as mesmas. Devido à grande quantidade de genomas incluídos, o conjunto foi simplificado, manualmente, para conter somente um ou alguns genomas representativos por ramo. Para comparação e validação, foram utilizadas duas metodologias distintas para a construção de árvores filogenéticas levando em consideração a integralidade dos genomas, ao invés de utilizar poucos genes marcadores: 1. com o programa SANS-Serif (Rempel e Wittler, 2021), que é livre de alinhamento e não necessita de genoma de referência; 2. com o programa Parsnp (Treangen et al., 2014), que é dependente de alinhamento global dos genomas.

O resultado desta análise é apresentado na Figura 13 e, salvo alguns poucos genomas, as árvores geradas são essencialmente iguais. É interessante notar nesta figura que a árvore do programa SANS-Serif (à esquerda) tem maior resolução que a gerada pelo programa Parsnp, o qual viria a sugerir uma expansão clonal de diversas linhagens do DF.

Os motivos para a falta de resolução do Parsnp provavelmente se devem ao processo de seleção dos caracteres para calcular a árvore, que no caso são somente polimorfismos de base única em sub-regiões conservadas em todos os genomas analisados (Treangen et al., 2014). Apesar de ser computacionalmente eficiente, a estratégia do Parsnp não aponta, por exemplo, a diferença nos genomas da KpBSB56 e KpBSB60, enquanto que o programa SANS-Serif identifica caracteres distintivos entre essas. Devido a este fato selecionamos a árvore do programa SANS-Serif para a interpretação das relações filogenéticas entre as diversas linhagens.

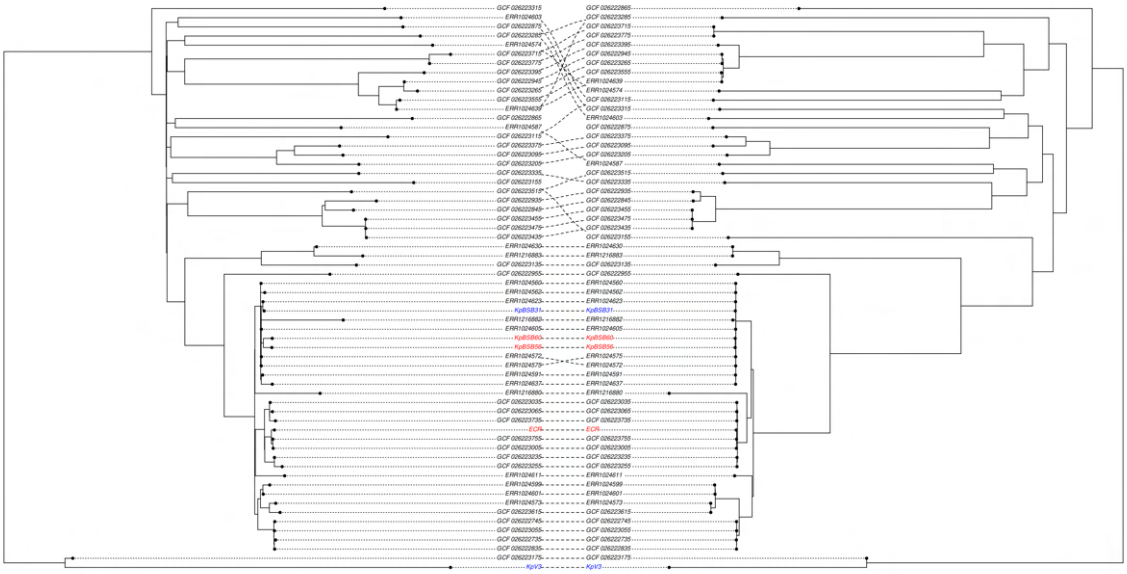

Figura 13: Comparação das diferentes árvores filogenéticas das linhagens de *Klebsiella pneumoniae* incluídas neste estudo. À esquerda, mostra-se a árvore gerada pela ferramenta SANS-Serif que utiliza metodologia livre de alinhamento (Rempel e Wittler, 2021). À direita, mostra-se a árvore gerada pela ferramenta Parsnp+IQTree, que utiliza metodologia dependente em alinhamento. As linhagens isoladas no DF e caracterizadas por nosso grupo neste estudo são apresentadas em vermelho e as em azul são referentes a estudos anteriores.

A árvore final foi anotada de maneira a incluir os "Sequence Types" sendo possível observar um substancial agrupamento de linhagens ST11 (Figura 14), que pertencem ao grupo clonal CG258, o qual representa uma ameaça global em termos de *K. pneumoniae* MDR (Fuga et al., 2020; Nakamura-Silva et al., 2021). Neste clado ST11, observa-se que as linhagens KpBSB56 e KpBSB60 agrupam-se, com a linhagem KpBSB31 e diversas linhagens MDR isoladas em Brasília entre 2010 e 2014 (código iniciando com ERR na Figura 13) (Lee et al., 2021). Os reduzidos comprimentos de ramo no clado indicam uma alta similaridade entre os genomas, mesmo em um período de quase uma década de circulação destas linhagens. Já a ECR, outra linhagem avaliada neste estudo, apesar de pertencer ao ST11 e ter sido isolada no mesmo local que KpBSB56 e KpBSB60, mostra uma similaridade maior com linhagens isoladas no estado de São Paulo entre 2018 e 2021 (Camargo et al., 2022), o que pode indicar uma possível rota de introdução.

Interessantemente, os resultados revelaram que apesar de nossas linhagens serem do mesmo Hospital, somente a KpBSB56 e KpBSB60 agrupam entre si. Sob esta ótica, podemos observar na Tabela 9 que as linhagens ECR e KpBSB56, apesar de serem geneticamente mais próximas da mesma referência, são dispostas distantes entre si na árvore. Isto pode ser uma consequência da diversidade de recombinações (Sakoparnig et al., 2021) (Figura 11), indicando que, apesar de serem da mesma localidade e ST, estas linhagens possuem diferenças genômicas significativas entre si. Também, vale ressaltar que a linhagem KpV3 é geneticamente mais diversa que as demais, provavelmente devido a esta ser a única linhagem ambiental dentre as incluídas na análise.

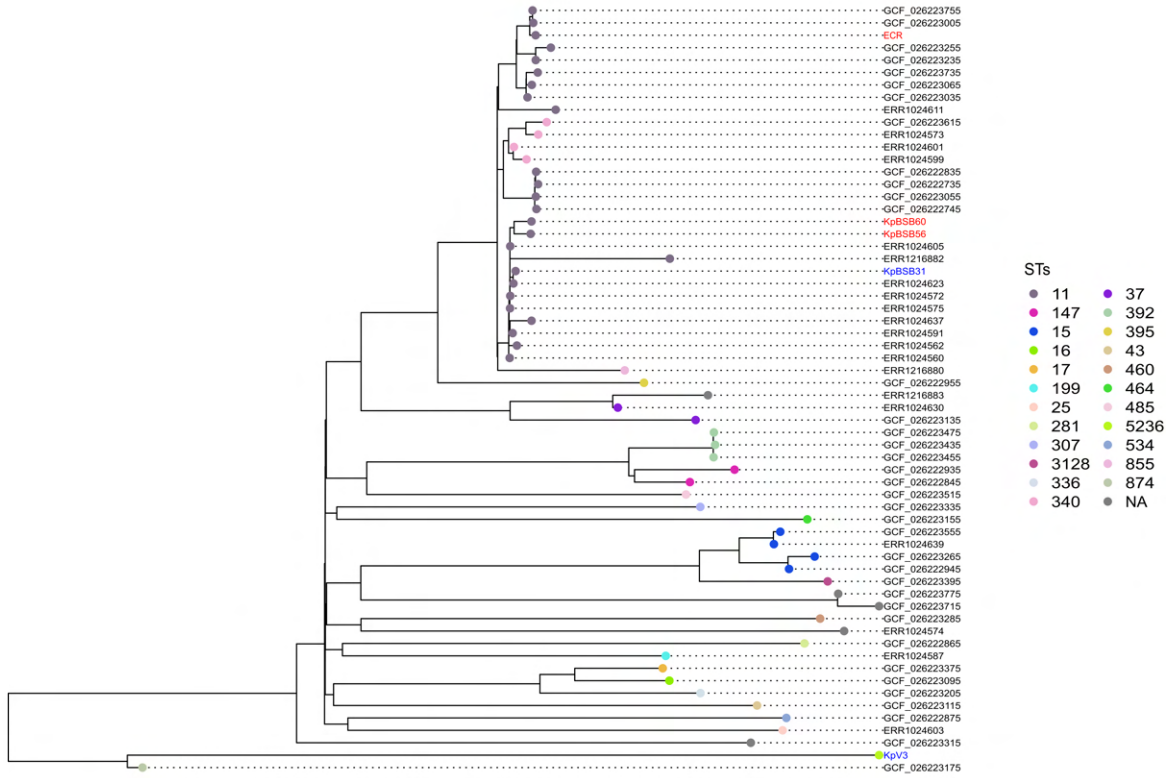

Figura 14: Árvore filogenética dos genomas das linhagens de *Klebsiella pneumoniae* gerada pela ferramenta SANS-Serif (Rempel e Wittler, 2021). Todos os ramos tiveram suporte "bootstrap" maior que 80%. As linhagens isoladas no DF e caracterizadas por nosso grupo neste estudo são apresentadas em vermelho e as em azul são referentes a estudos anteriores. Linhagens com código iniciando com o prefixo ERR foram isoladas em hospitais de Brasília, entre 2010 a 2014 (Lee et al., 2021). As demais linhagens (prefixo GCF) são predominantemente do estado de São Paulo (Camargo et al., 2022). Os círculos terminais são coloridos de acordo com o "Sequence Type"(ST) de cada genoma.

4.2.7.2 Estudo retrospectivo da resistência e virulência

Uma comparação inicial das três novas linhagens (KpBSB56, KpBSB60 e ECR) com as outras duas linhagens (KpBSB31 e KpV3) do mesmo Hospital demonstrou a manutenção de certas características observadas no passado. Primeiro, observamos na linhagem KpBSB60 a presença de uma KPC-2 flanqueada por uma transposase e uma resolvase, assim como identificado na KpV3 (Janssen et al., 2021).

Além disso, observou-se a conservação de diversos fatores de virulência, inclusive a presença de sideróforos adicionais, identificados na linhagem KpBSB31 e outros isolados XDR hipermucoviscosos de 2018 provenientes do mesmo Hospital (De Campos et al., 2018). À época, estes isolados foram classificadas como não-hipervirulentos, porém, todos demonstraram invasividade e toxicidade em células epiteliais Hep-2, e sobrevivência em sangue e soro humano (De Campos et al., 2018). Interessantemente, estes isolados de 2018 e as linhagens ECR e KpBSB60, demonstraram suscetibilidade somente a aminoglicosídeos, sugerindo manutenção de um mesmo perfil de resistência.

Em seguida, passamos para a comparação do resistoma, isto é, o conjunto total de genes de resistência, com todas as outras linhagens brasileiras incluídas. Em um primeiro momento, tentamos avaliar o perfil de resistência em forma de um mapa de presença e ausência ("heatmap") dos genes de resistência detectados (Figura 15). Para uma melhor legibilidade, utilizou-se o mesmo sub-grupo apresentado na árvore filogenética (Figura 14) e genes com muitos alelos foram colapsados em coluna única.

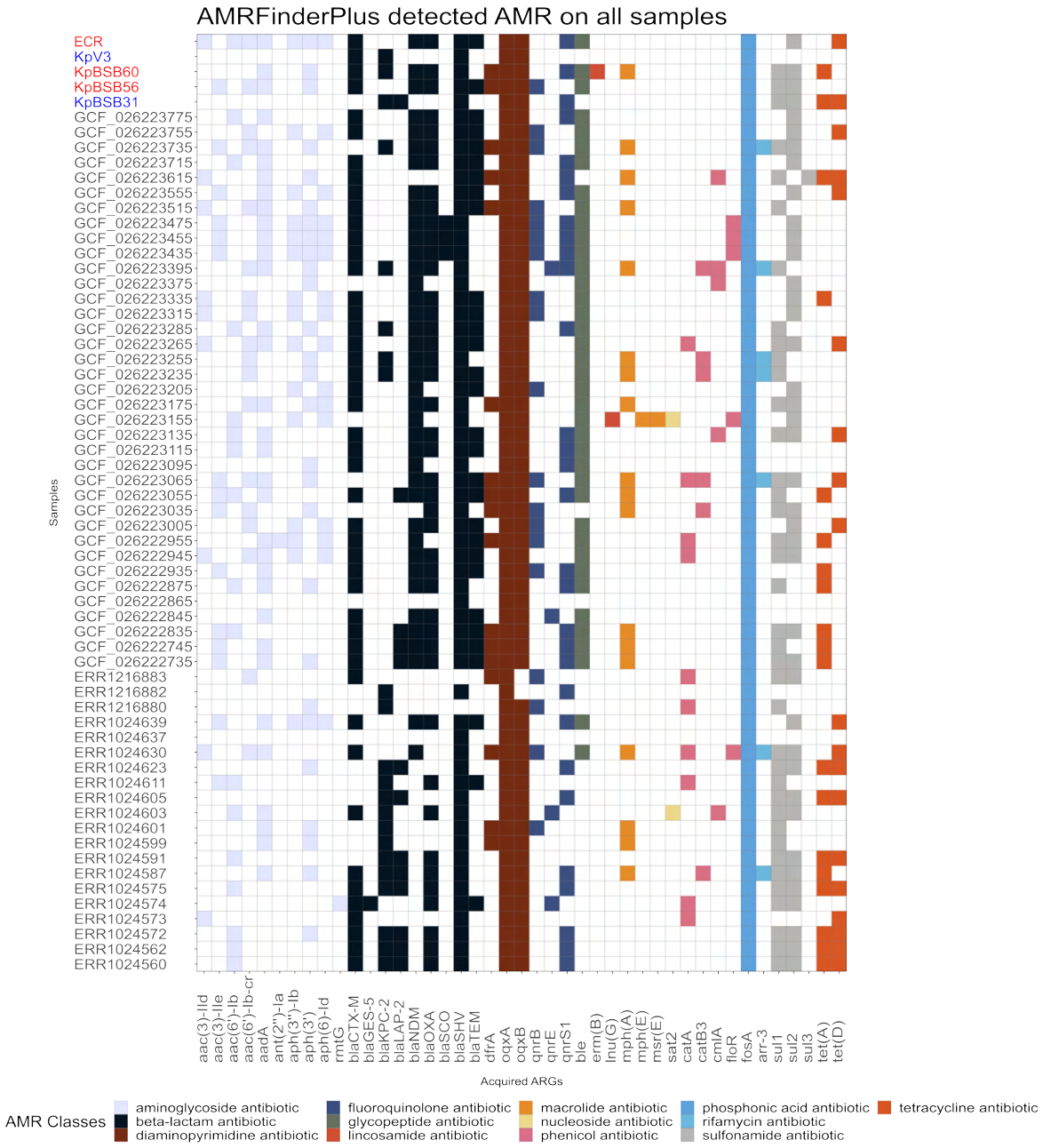

Figura 15: Mapa da presença/ausência de genes de resistência detectados nos genomas analisados. A anotação representa o consenso dos resultados das ferramentas AMRFinderPlus (Feldgarden et al., 2019) e Resfinder (Bortolaia et al., 2020). A presença dos genes é marcada por blocos coloridos conforme as classes de antimicrobianos para as quais estes genes conferem resistência, ou em branco no caso de ausência. Os códigos dos genomas são os mesmos da Figura 14

Pela imagem, não é possível perceber padrões claros de ocorrência dos conjuntos de genes de resistência entre as linhagens, sugerindo uma distribuição aleatória. Porém, é possível destacar três observações:

1. A prevalência de resistência a sulfonamida é elevada e parece estar estabelecida na região. Este cenário se assemelha a uma observação levantada durante um estudo retrospectivo de *K. pneumoniae* multirresistentes em Manaus (Nakamura-Silva et al., 2022);

2. Visualmente, os resultados parecem corroborar com relatos recentes de acúmulo de β-lactamases e maior frequência de co-carreamento de carbapenemases (Gao et al., 2020; Guo et al., 2023);

3. Nota-se uma tendência entre os genes *bla*NDM e *bla*KPC-2, em que os genes *bla*KPC são encontrados mais frequentemente nas linhagens mais antigas (de 2014, códigos ERR), enquanto que o *bla*NDM apresenta um perfil inverso.

Estes resultados respaldam estudos recentes que indicam que o gene *bla*NDM-1 vem se tornando cada vez mais frequente no Brasil desde sua introdução (Camargo et al., 2022; Kiffer et al., 2023). E também apontam para uma possível transição no perfil de resistência a carbapenêmicos, como sugerido em um estudo que avaliou o impacto da pandemia de COVID-19 no padrão de co-ocorrência destes genes no Brasil (Kiffer et al., 2023).

Por isso, decidimos aprofundar a investigação do padrão e dinâmica de ocorrência de pares de genes em nosso conjunto de dados. Procurou-se identificar se existem genes que sempre ocorrem associados, embasados em um arcabouço estatístico. Para isso, realizamos análises de co-ocorrência através dos pacotes na linguagem R: CooccurrenceAffinity (Mainali e Slud, 2022) e Cooccur (Griffith et al., 2016). Para tanto, os genomas foram separados em dois subconjuntos de linhagens, antes e depois de 2019, e as estatísticas calculadas para cada grupo. Os resultados obtidos pelos dois pacotes foram idênticos, e por isso utilizamos para comparação e visualização somente os resultados da ferramenta CooccurrenceAffinity uma vez que esta introduz a métrica α, baseada em estimativa de máxima verossimilhança, que mensura a intensidade da associação entre duas entidades,

enquanto Cooccur somente indica se é positivo, negativo ou aleatório. Este resultado é apresentado na Figura 16.

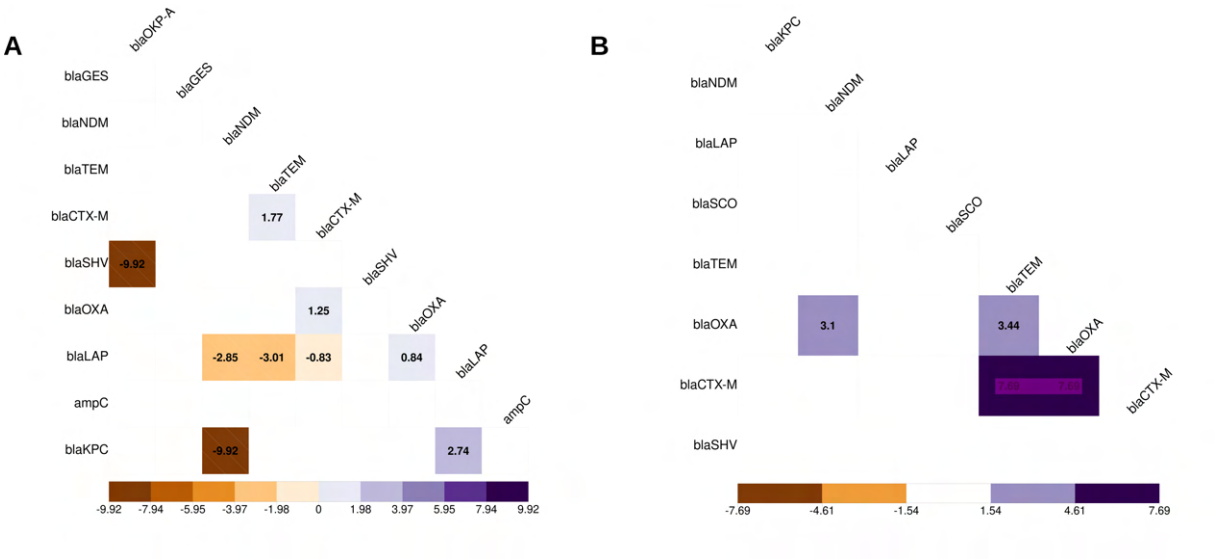

Figura 16: Valores da análise de "afinidade" de co-ocorrência dos genes de β-lactamases detectados nos genomas analisados. Na figura, separam-se os subconjuntos de genomas de antes (A) e depois de 2019 (B). A análise foi realizada através do pacote CooccurrenceAffinity (Mainali e Slud, 2022) que introduz uma nova métrica α capaz de mensurar a intensidade da associação, positiva ou negativa, entre duas entidades. Nesta métrica, o valor 0 representa uma associação aleatória. Somente são mostradas os valores cujo p-valor é menor que 0,05.

Os resultados da análise de co-ocorrência de pares de genes, mostram que os genomas anteriores a 2019 (Figura 16A), apresentam uma certa co-ocorrência positiva, mas não tão intensa, para os genes *bla*LAP, *bla*KPC, *bla*OXA, *bla*CTX-M e *bla*TEM. O valor absoluto dos resultados (métrica α) sugere que, embora estes genes pudessem co-ocorrer, este cenário não era muito frequente, uma vez que 0 significa uma associação aleatória. Interessantemente, o resultado ressalta um α fortemente negativo para a associação *bla*KPC e *bla*NDM, indicando que encontrar um na presença do outro era raro e inesperado.

Em contrapartida, os resultados dos genomas após 2019 (Figura 16B) mostram uma intensa co-ocorrência positiva para os genes *bla*CTX-M, *bla*OXA, *bla*TEM e *bla*NDM. Além disso, a observação de maiores valores absolutos de α sugere que a co-ocorrência destes pares de genes é muito mais provável que anteriormente. Isso indica que estas linhagens podem estar acumulando genes de β-lactamase e carbapenemases, e que esta vem de maneira combinada, assim como também sugerido por outros estudos (Gao et al., 2020; Guo et al., 2023).

A observação dos altos valores α com sinal negativo para a associação *bla*KPC e *bla*NDM aliada a substituição do *bla*KPC pela *bla*NDM na lista de genes positivamente associados do grupo de genomas pós 2019, respalda a sugestão de que no Brasil estas bactérias podem estar mudando seu perfil gênico, favorecendo a NDM (Kiffer et al., 2023). Este cenário pode, por exemplo, estar atrelado à diferenças de custos para o "fitness" bacteriano. Um bom exemplo são os plasmídeos IncX3, que são moléculas de alta estabilidade e custo fisiológico baixo, considerados os principais veículos de transmissão de NDM (Guo et al., 2022). Porém, mais estudos são necessários para entendimento e confirmação dos mecanismos por trás destas observações.

Neste mesmo contexto, percebe-se que a presença dos genes *bla*NDM, *bla*KPC e *bla*OXA ao mesmo tempo não é detectada entre as linhagens mais antigas, enquanto que esta é identificada em seis dos genomas mais recentes (Figura 15). Imagina-se dois cenários possíveis para explicar esta observação: (i) ou estas linhagens encontram-se no processo de reposição, em que acabaram de adquirir o *bla*NDM e vão consequentemente perder o *bla*KPC ou; (ii) ou elas representam um novo evento de acumulação destas carbapenemases.

Algo que poderia ajudar a explicar estas intensas associações observadas entre o gene *bla*NDM e outras β-lactamases, e seu consequente acúmulo nestas linhagens mais recentes, seria a presença de elementos genéticos móveis carreando estes genes em conjunto. Isto permitira aventar a hipótese de que a reposição da KPC pelo NDM seria mediada por trocas de plasmídeos, mudando seu perfil de resistência e balanceando custos de "fitness".

Por isso, buscamos também comparar o perfil dos grupos de plasmídeos detectados no conjunto de dados e investigar a disposição destes genes nos plasmídeos resolvidos de nossas linhagens. Assim como geralmente observado em *Enterobacterales* (Rozwandowicz et al., 2018; Chen et al., 2024) o grupo de plasmídeos IncF foi identificado como o grupo de plasmídeos mais frequentemente detectado entre os genomas *K. pneumoniae* analisados, representando aproximadamente ≈36% de todos os replicons anotados. Plasmídeos IncF são importantes vetores de resistência, frequentemente relacionados à transmissão de genes de resistência a antimicrobianos (Bonnin et al., 2012; Bi et al., 2018; Chen et al., 2024) e considerados como essenciais na expansão recente de carbapenemases (Chen et al., 2024).

Em nosso conjunto de dados, os plasmídeos conjugativos IncFIB(K), IncFII(K) e

IncFII(pCRY) foram os replicons IncF mais comumente detectados. Entre estes, o IncFII(K), um plasmídeo bastante heterogêneo em tamanho que pode ser muito grande e possuir estruturas de mosaico (Bi et al., 2018; Montelongo Hernandez et al., 2022), foi encontrado nos genomas da ECR e KpBSB60. Para ter uma visão geral da anotação e conservação deste plasmídeo entre as linhagens, gerou-se um mapa circular usando a sequência completa do plasmídeo IncFII(K) da ECR como espinha dorsal para alinhamento contra as linhagens KpBSB56, KpBSB60 e outras três linhagens brasileiras produtoras de NDM (Figura 17).

O mapa circular mostra que a sequência é completamente compartilhada entre as linhagens ECR e KpBSB60, sugerindo que ambas possuem este plasmídeo, mas com a ressalva de que não é possível uma confirmação devido à fragmentação da montagem da KpBSB60. Além disso, o fato de várias subseções do plasmídeo serem conservadas entre as diferentes linhagens, mas não a sequência completa, sugere uma integração de diferentes segmentos, destacando o mosaicismo característico deste grupo de plasmídeos (Bi et al., 2018; Montelongo Hernandez et al., 2022). A Figura 17 mostra também que os genes de resistência detectados estão organizados em pequenos grupos cercados por transposases, o que sugere uma capacidade de mobilização. Por fim, o mapa mostra que este plasmídeo nas linhagens ECR e KpBSB60 codifica todas as quatro β-lactamases (NDM-1, CTX-M-15, OXA-9 e TEM-1) que foram encontradas co-ocorrendo positivamente nas linhagens a partir de 2019.

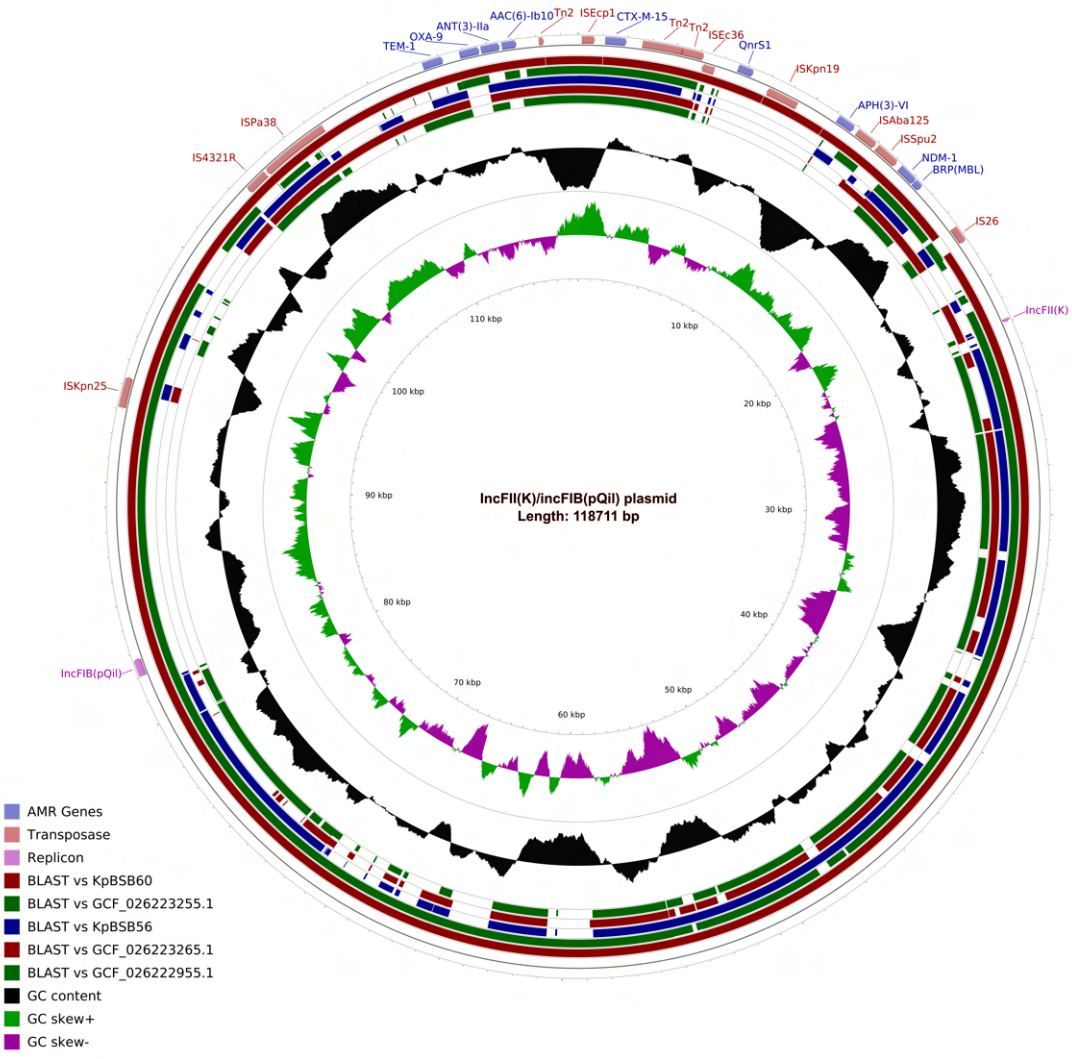

Figura 17: Mapa circular da sequência completa do plasmídeo IncFII(K) identificado na linhagem ECR. Na camada mais externa do mapa, destacam-se os genes de resistência e transposases, e a posição dos replicons detectados neste plasmídeo. As cinco próximas camadas, possuindo blocos retangulares de colorações vermelho escuro, verde escuro e azul escuro, representam os blocos genômicos conservados via alinhamentos BLAST com as cinco linhagens identificadas por seus códigos (após BLAST vs). Nas três camadas mais internas, mostram-se três diferentes representações das variações do conteúdo GC ao longo da sequência. Esta figura foi gerada através da ferramenta "CGView Comparison Tool" (Grant et al., 2012).

5 Conclusões

Apresenta-se neste trabalho o desenvolvimento de três *pipelines* genômicos baseados em contêineres computacionais que fornecem um fluxo de análises completo e automatizado para o estudo de genomas bacterianos, compreendendo as etapas de pré-processamento de dados brutos de sequenciamento, montagem de genomas e anotação genérica e especializada de genomas bacterianos.

Estes foram implementados objetivando mitigar lacunas específicas detectadas entre os diversos *pipelines* existentes para análises de genômica bacteriana. Por isso, desenvolveu-se diversos relatórios gráficos e visualizações interativas para melhorar a interpretação dos resultados. E complementarmente, desenvolveu-se uma interface gráfica acessível pelo navegador de internet que permite aos usuários analisar e refinar todos os resultados de anotação de forma interativa, distinguindo o bacannot de outros *pipelines*.

Além disso, durante o desenvolvimento e concretização destes *pipelines*, estes já foram utilizados em outros estudos que envolveram diferentes espécies de bactérias e contemplam linhagens clínicas e ambientais. Assim, destacam o desenvolvimento de ferramentas genéricas que possam ser úteis em diferentes contextos analíticos (Campos et al., 2021; Janssen et al., 2021; Belmok et al., 2023; Rocha et al., 2023). Salienta-se que os *pipelines* foram adaptados à forma modular de definição de *pipelines* adotada pela comunidade Nextflow, tornando possível estender suas análises de forma simplificada.

De maneira geral, os *pipelines* desenvolvidos proporcionam um sistema que facilita e permite a compilação padronizada de resultados de diversos genomas. Esta padronização abre caminho para a efetivação de estudos de genômica comparativa. Neste contexto, ressalta-se a importância de bancos de dados e plataformas especializadas como o *One Health Brazilian Resistance*[1] (OneBR) que oferece uma rica interface gráfica para

[1] http://onehealthbr.com/

selecionar e explorar metadados de diversos genomas de espécies relevantes para a saúde pública no Brasil. A utilização conjunta dos *pipelines* desenvolvidos nestes trabalho a estas plataformas e iniciativas especializadas permitirá uma melhor integração de dados e a avaliação de aspectos inexplorados da epidemiologia de linhagens multirresistentes, alavancando estudos evolutivos.

Sob esta ótica, sequenciamos e caracterizamos genomicamente três isolados de *Klebsiella pneumoniae* obtidos de um Hospital terciário de Brasília e, incluímos diversos outros isolados brasileiros para realização de uma análise comparativa retrospectiva, objetivando demonstrar a aplicabilidade destes *pipelines*.

Todas as três linhagens deste estudo foram classificadas como linhagens de *Klebsiella pneumoniae* XDR ST11. A identificação de bactérias ST11 é preocupante, pois este é um grupo de bactérias frequentemente associado ao desenvolvimento de surtos bacterianos, com diversos relatos recentes de linhagens de alto risco, apresentando fenótipo convergente de multirresistência e virulência aumentada (Zhan et al., 2017; Gu et al., 2018; Xie et al., 2021; Nicola et al., 2022; Liu et al., 2022; Mendes et al., 2022). Somado a isso, foram detectados três diferentes KL (3, 64 e 105), evidenciando a diversidade das cepas. Particularmente, linhagens ST11-KL64 (KpBSB60) são prevalentes mundialmente e de alta relevância no contexto da expansão de linhagens resistentes a carbapenêmicos (Wang et al., 2023).

Os resultados da anotação destacaram a presença de diversos genes de resistência a antimicrobianos nas três linhagens, com vários destes genes codificados em plasmídeos. Inclusive, detectou-se a presença de plasmídeos com múltiplos replicons (Tabela 12), considerados melhores carreadores de fenótipo de resistência. Em particular, destaca-se a detecção do gene *bla*NDM-1 em ambas as três linhagens sequenciadas. Este é considerado um gene de alto risco que codifica uma carbapenemase que confere resistência a todas as drogas β-lactâmicas, exceto o aztreonam (Boyd et al., 2020).

Em se tratando de β-lactamases, as análises comparativas com todas as linhagens incluídas neste estudo apontaram uma co-ocorrência significativa dos genes *bla*CTX-M, *bla*OXA, *bla*TEM e *bla*NDM no sub-conjunto de dados de linhagens pós-2019. Não só isso, os valores absolutos da análise indicam que estas combinações têm ficado mais frequentes, sugerindo que estas bactérias estejam acumulando estes genes. Além disso, os

resultados corroboram com a observação de outros autores de que, no Brasil, as linhagens parecem estar migrando da KPC para a NDM que está se tornando cada vez mais prevalente (Camargo et al., 2022; Kiffer et al., 2023). Interessantemente, estes genes foram detectados em um mesmo plasmídeo, completamente compartilhado pelas linhagens ECR e KpBSB60.

Realizou-se também uma comparação do viruloma (conjunto total de genes de virulência) das três linhagens deste estudo com outros isolados de 2018 do mesmo Hospital, incluindo a KpBSB31 (De Campos et al., 2018). Esta análise permitiu observar uma certa conservação destes genes, incluindo a detecção de fatores determinantes de virulência aumentada, como os sideróforos Aerobactina e Yersiniabactina que são capazes de aumentar o "fitness" bacteriano durante infecções (Paczosa e Mecsas, 2016; Wyres et al., 2020; Zhao et al., 2022; Gorrie et al., 2022). Além disso, este lócus da Yersiniabactina foi detectado em um transposon ICEKp que é um dos principais fatores de disseminação de virulência na espécie (Farzand et al., 2019; Zhao et al., 2022). Estes resultados reiteram alertas quanto à convergência de genes de resistência e virulência em linhagens da espécie, um cenário que vem se tornando mais prevalente globalmente (Lee et al., 2017; Navon-Venezia et al., 2017; Gu et al., 2018; Li et al., 2021; Liu et al., 2022; He et al., 2022; Liu et al., 2022; Cardoso Almeida et al., 2024). Porém, é necessário realizar análises experimentais para mensurar o real potencial de virulência destas linhagens.

De modo geral, os resultados sugerem que elementos genéticos móveis podem estar desempenhando papel-chave nessas observações destacadas no presente estudo. Além disso, também enfatizam a presença de forte pressão seletiva nos ambientes investigados que além de poder estar favorecendo a manutenção destas moléculas (Wang et al., 2021), pode também estar contribuindo para as observações de substituição da KPC e acúmulo geral de genes destacada por outros autores (Camargo et al., 2022; Kiffer et al., 2023). Porém, a baixa resolução da maioria dos genomas devido a não utilização de leituras longas impede a validação dessa e de outras hipóteses levantadas neste estudo. Assim, constata-se a necessidade de mais estudos que adotem o sequenciamento por leituras longas de modo a obter uma maior resolução do cenário da disseminação de resistência, e virulência, através de moléculas extracromossômicas.

Como um todo, estas análises demonstram a aplicabilidade dos *pipelines* desen-

volvidos para a realização de estudos de genômica comparativa. Além disso, o estudo reitera a necessidade de agir agora para evitar um futuro onde infecções bacterianas sejam intratáveis e, ao mesmo tempo, extremamente capazes de produzir doenças graves. Por fim, enfatiza-se a urgência em adotar a vigilância genômica de rotina em ambientes clínicos utilizando metodologia híbrida, permitindo um estudo mais detalhado destes eventos de expansão de genes e seus mecanismos. Estes sistemas de vigilância podem fornecer informações críticas para ajudar hospitais e governos a desenvolver estratégias para combater a propagação da resistência antimicrobiana e potenciais falhas terapêuticas (Schnall et al., 2019; Hendriksen et al., 2019; on Genomic Surveillance of AMR, 2020; Lam et al., 2021; Deekshit e Srikumar, 2022; Waddington et al., 2022).

Referências Bibliográficas

Abavisani M., Bostanghadiri N., Ghahramanpour H., Kodori M., Akrami F., Fathizadeh H., Hashemi A., Rastegari-Pouyani M., Colistin resistance mechanisms in Gram-negative bacteria: a Focus on Escherichia coli, Letters in Applied Microbiology, 2023, vol. 76

Abueg L. A. L., Afgan E., Allart O., Awan A. H., Bacon W. A., Baker D., Bassetti M., Batut B., Bernt M., Blankenberg D., et al., The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update, Nucleic Acids Research, 2024, vol. 52, p. W83–W94

Aires C. A. M., Pereira P. S., de Souza C. M. R., Silveira M. C., Carvalho-Assef A. P. D., Asensi M. D., Population Structure of KPC-2-Producing Klebsiella pneumoniae Isolated from Surveillance Rectal Swabs in Brazil, Microbial Drug Resistance, 2019

Ali M. R., Yang Y., Dai Y., Lu H., He Z., Li Y., Sun B., Prevalence of multidrug-resistant hypervirulent Klebsiella pneumoniae without defined hypervirulent biomarkers in Anhui, China: a new dimension of hypervirulence, Frontiers in Microbiology, 2023, vol. 14

Almeida F. M. d., Campos T. A. d., Pappas Jr G. J., Scalable and versatile container-based pipelines for de novo genome assembly and bacterial annotation., F1000Research, 2023, vol. 12, p. 1205

Andrey D. O., Dantas P., Martins W. B. S., de Carvalho F. M., Gonzaga L. A., Sands K., Portal E., Sauser J., Cayô R., Nicolas M. F., et al., An Emerging Clone, KPC-2-Producing Klebsiella pneumoniae ST16, Associated with High Mortality Rates in a CC258 Endemic Setting, Clinical Infectious Diseases, 2019

Ausubel F. M., Brent R., Kingston R. E., Moore D. D., Seidman J., Smith J. A., Struhl K., Short protocols in molecular biology, New York, 1992, vol. 275, p. 28764

Azevedo P. A. A., Furlan J. P. R., Gonçalves G. B., Gomes C. N., Goulart R. d. S., Stehling E. G., Pitondo-Silva A., Molecular characterisation of multidrug-resistant Klebsiella pneumoniae belonging to CC258 isolated from outpatients with urinary tract infection in Brazil, Journal of Global Antimicrobial Resistance, 2019, vol. 18, p. 74

Baker K. S., Jauneikaite E., Hopkins K. L., Lo S. W., Sánchez-Busó L., Getino M., Howden B. P., Holt K. E., Musila L. A., Hendriksen R. S., et al., Genomics for public health and international surveillance of antimicrobial resistance, The Lancet Microbe, 2023, vol. 4, p. e1047–e1055

Baker S., Thomson N., Weill F. X., Holt K. E., Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens, Science, 2018, vol. 360, p. 733

Bartoli C., Rigal M., Huard-Chauveau C., Mayjonade B., Roux F., The genetic architecture of the adaptive potential of Arabidopsis thaliana in response to Pseudomonas syringae strains isolated from south-west France, Plant Pathology, 2023, vol. n/a

Baucheron S., Nishino K., Monchaux I., Canepa S., Maurel M. C., Coste F., Roussel A., Cloeckaert A., Giraud E., Bile-mediated activation of the acrAB and tolC multidrug efflux genes occurs mainly through transcriptional derepression of ramA in Salmonella enterica serovar Typhimurium, Journal of Antimicrobial Chemotherapy, 2014, vol. 69, p. 2400

Bedenić B., Luxner J., Car H., Sardelić S., Bogdan M., Varda-Brkić D., Šuto S., Grisold A., Beader N., Zarfel G., Emergence and Spread of Enterobacterales with Multiple Carbapenemases after COVID-19 Pandemic, Pathogens, 2023, vol. 12, p. 677

Bell G., MacLean C., The Search for 'Evolution-Proof' Antibiotics, Trends in Microbiology, 2018, vol. 26, p. 471

Belmok A., de Almeida F. M., Rocha R. T., Vizzotto C. S., Tótola M. R., Ramada M. H. S., Krüger R. H., Kyaw C. M., Pappas G. J., Genomic and physiological characterization of Novosphingobium terrae sp. nov., an alphaproteobacterium isolated from

Cerrado soil containing a mega-sized chromid, Brazilian Journal of Microbiology, 2023, vol. 54, p. 239–258

Bengoechea J. A., Pessoa J. S., Klebsiella pneumoniae infection biology: living to counteract host defences, FEMS Microbiology Reviews, 2018, vol. 43, p. 123

Bethke J. H., Ma H. R., Tsoi R., Cheng L., Xiao M., You L., Vertical and horizontal gene transfer tradeoffs direct plasmid fitness, Molecular Systems Biology, 2022, vol. 19

Bi D., Zheng J., Li J.-J., Sheng Z.-K., Zhu X., Ou H.-Y., Li Q., Wei Q., In Silico Typing and Comparative Genomic Analysis of IncFII(K) Plasmids and Insights into the Evolution of Replicons, Plasmid Backbones, and Resistance Determinant Profiles, Antimicrob Agents Chemother, 2018, vol. 62

Blair J. M. A., Webber M. A., Baylay A. J., Ogbolu D. O., Piddock L. J. V., Molecular mechanisms of antibiotic resistance, Nature Reviews Microbiology, 2015, vol. 13, p. 42

Boettiger C., An introduction to Docker for reproducible research. In Operating Systems Review (ACM) , vol. 49, Association for Computing Machinery, 2015, p. 71

Bonnin R. A., Poirel L., Carattoli A., Nordmann P., Characterization of an IncFII Plasmid Encoding NDM-1 from Escherichia coli ST131, PLoS ONE, 2012, vol. 7, p. e34752

Bortolaia V., Kaas R. S., Ruppe E., Roberts M. C., Schwarz S., Cattoir V., Philippon A., Allesoe R. L., Rebelo A. R., Florensa A. F., et al., ResFinder 4.0 for predictions of phenotypes from genotypes, Journal of Antimicrobial Chemotherapy, 2020, vol. 75, p. 3491

Boyd S. E., Livermore D. M., Hooper D. C., Hope W. W., Metallo-β-Lactamases: Structure, Function, Epidemiology, Treatment Options, and the Development Pipeline, Antimicrobial Agents and Chemotherapy, 2020, vol. 64, p. e00397

Bush K., Classification for β-lactamases: historical perspectives, Expert Review of Anti-infective Therapy, 2023, vol. 21, p. 513–522

Calderaro A., Buttrini M., Martinelli M., Montecchini S., Covan S., Ruggeri A., Rodighiero I., Di Maio A., Galullo M., Larini S., et al., Active surveillance for carbapenemase-producing Klebsiella pneumonia e and correlation with infection in subjects attending an Italian tertiary-care hospital: a 7-year retrospective study, BMJ Open, 2021, vol. 11, p. e042290

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T. L., BLAST+: Architecture and applications, BMC Bioinformatics, 2009, vol. 10, p. 421

Camargo C. H., Yamada A. Y., Souza A. R. d., Reis A. D., Santos M. B. N., Assis D. B. a. d., Carvalho E. d., Takagi E. H., Cunha M. P. V., Tiba-Casas M. R., Genomic Diversity of NDM-Producing Klebsiella Species from Brazil, 2013–2022, Antibiotics, 2022, vol. 11, p. 1395

Campos T. A. d., Almeida F. M. d., Almeida A. P. C. d., Nakamura-Silva R., Oliveira-Silva M., Sousa I. F. A. d., Cerdeira L., Lincopan N., Pappas G. J., Pitondo-Silva A., Multidrug-Resistant (MDR) Klebsiella variicola Strains Isolated in a Brazilian Hospital Belong to New Clones, Frontiers in Microbiology, 2021, vol. 12

Carattoli A., Zankari E., Garciá-Fernández A., Larsen M. V., Lund O., Villa L., Aarestrup F. M., Hasman H., In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing, Antimicrobial Agents and Chemotherapy, 2014, vol. 58, p. 3895

Cardoso Almeida A. P., de Moraes M. A., da Silva A. K. F., Oliveira-Silva M., Nakamura-Silva R., de Almeida F. M., Pappas Junior G. J., Pitondo-Silva A., de Campos T. A., Long-term occurrence of multiple antimicrobial drug resistant Klebsiella pneumoniae isolates harboring virulent potential in a tertiary hospital from Brazil, Brazilian Journal of Microbiology, 2024

Carter L., Yu M. A., Sacks J., Barnadas C., Pereyaslov D., Cognat S., Briand S., Ryan M., Samaan G., Global genomic surveillance strategy for pathogens with pandemic

and epidemic potential 2022–2032, Bulletin of the World Health Organization, 2022, vol. 100, p. 239

Carvalho-Assef A. P. D., Pereira P. S., Albano R. M., Beriao G. C., Chagas T. P. G., Timm L. N., Da Silva R. C. F., Falci D. R., Asensi M. D., Isolation of NDM-producing Providencia rettgeri in Brazil, Journal of Antimicrobial Chemotherapy, 2013, vol. 68, p. 2956

Castañeda Barba S., Top E. M., Stalder T., Plasmids, a molecular cornerstone of antimicrobial resistance in the One Health era, Nature Reviews Microbiology, 2023, vol. 22, p. 18–32

Catalán-Nájera J. C., Garza-Ramos U., Barrios-Camacho H., , 2017 Hypervirulence and hypermucoviscosity: Two different but complementary Klebsiella spp. phenotypes?

Chang W., Cheng J., Allaire J., Sievert C., Schloerke B., Xie Y., Allen J., McPherson J., Dipert A., Borges B., , 2024 shiny: Web Application Framework for R

Chen L., Ai W., Zhou Y., Wu C., Guo Y., Wu X., Wang B., Rao L., Xu Y., Zhang J., Chen L., Yu F., Outbreak of IncX8 Plasmid–Mediated KPC-3–Producing Enterobacterales Infection, China, Emerging Infectious Diseases, 2022, vol. 28, p. 1421–1430

Chen L., Yang J., Yu J., Yao Z., Sun L., Shen Y., Jin Q., VFDB: A reference database for bacterial virulence factors, Nucleic Acids Research, 2005, vol. 33, p. D325

Chen R., Li C., Ge H., Qiao J., Fang L., Liu C., Gou J., Guo X., Difference analysis and characteristics of incompatibility group plasmid replicons in gram-negative bacteria with different antimicrobial phenotypes in Henan, China, BMC Microbiology, 2024, vol. 24

Chiu C. Y., Miller S. A., Clinical metagenomics, Nature Reviews Genetics, 2019, vol. 20, p. 341

Choi M., Hegerle N., Nkeze J., Sen S., Jamindar S., Nasrin S., Sen S., Permala-Booth J., Sinclair J., Tapia M. D., et al., The Diversity of Lipopolysaccharide (O) and Capsular

Polysaccharide (K) Antigens of Invasive Klebsiella pneumoniae in a Multi-Country Collection, Front. Microbiol., 2020, vol. 11

Chuang Y. C., Lee M. F., Yu W. L., Mycotic aneurysm caused by hypermucoviscous Klebsiella pneumoniae serotype K54 with sequence type 29: An emerging threat, Infection, 2013, vol. 41, p. 1041

Conceição-Neto O. C., da Costa B. S., Pontes L. d. S., Silveira M. C., Justo-da Silva L. H., de Oliveira Santos I. C., Teixeira C. B. T., Tavares e Oliveira T. R., Hermes F. S., Galvão T. C., et al., Polymyxin Resistance in Clinical Isolates of K. pneumoniae in Brazil: Update on Molecular Mechanisms, Clonal Dissemination and Relationship With KPC-Producing Strains, Frontiers in Cellular and Infection Microbiology, 2022, vol. 12

Coombe L., Li J. X., Lo T., Wong J., Nikolic V., Warren R. L., Birol I., LongStitch: high-quality genome assembly correction and scaffolding using long reads, BMC Bioinformatics Bioinformatics, 2021, vol. 22

Cortés G., Borrell N., de Astorza B., Gómez C., Sauleda J., Albertí S., Molecular Analysis of the Contribution of the Capsular Polysaccharide and the Lipopolysaccharide O Side Chain to the Virulence of Klebsiella pneumoniae in a Murine Model of Pneumonia, Infection and Immunity, 2002, vol. 70, p. 2583

Coulthurst S., The Type VI secretion system: a versatile bacterial weapon, Microbiology, 2019, vol. 165, p. 503

Cubero M., Grau I., Tubau F., Pallarés R., Dominguez M. A., Liñares J., Ardanuy C., Hypervirulent Klebsiella pneumoniae clones causing bacteraemia in adults in a teaching hospital in Barcelona, Spain (2007-2013), Clinical Microbiology and Infection, 2016, vol. 22, p. 154

Cuicapuza D., Loyola S., Velásquez J., Fernández N., Llanos C., Ruiz J., Tsukayama P., Tamariz J., Molecular characterization of carbapenemase-producing Enterobacterales in a tertiary hospital in Lima, Peru, Microbiology Spectrum, 2024, vol. 12

D'Andrea M. M., Amisano F., Giani T., Conte V., Ciacci N., Ambretti S., Santoriello L., Rossolini G. M., Diversity of Capsular Polysaccharide Gene Clusters in Kpc-Producing Klebsiella pneumoniae Clinical Isolates of Sequence Type 258 Involved in the Italian Epidemic, PLoS ONE, 2014, vol. 9, p. e96827

de Almeida F. M., Desenvolvimento de pipelines de genômica bacteriana e sua aplicação em isolados do Hospital Universitário de Brasília, Universidade de Brasília, 2020, Mestrado em biologia molecular, 98

de Almeida F. M., Pappas G. J., , 2024 Scalable and versatile container-based pipelines for de novo genome assembly and bacterial annotation (Sup. Material)

De Campos T. A., Gonçalves L. F., Magalhães K. G., De Paulo Martins V., Pappas Júnior G. J., Peirano G., Pitout J. D. D., Gonçalves G. B., Furlan J. a. P. R., Stehling E. G., et al., A Fatal Bacteremia Caused by Hypermucousviscous KPC-2 Producing Extensively Drug-Resistant K64-ST11 Klebsiella pneumoniae in Brazil, Frontiers in Medicine, 2018, vol. 5, p. 265

De Coster W., Rademakers R., NanoPack2: population-scale evaluation of long-read sequencing data, Bioinformatics, 2023, vol. 39, p. btad311

De Gaetano G. V., Lentini G., Famà A., Coppolino F., Beninati C., Antimicrobial Resistance: Two-Component Regulatory Systems and Multidrug Efflux Pumps, Antibiotics, 2023, vol. 12, p. 965

De Oliveira D. M. P., Forde B. M., Kidd T. J., Harris P. N. A., Schembri M. A., Beatson S. A., Paterson D. L., Walker M. J., Antimicrobial Resistance in ESKAPE Pathogens, Clinical Microbiology Reviews, 2020, vol. 33

de Oliveira Martins L., Page A. J., Mather A. E., Charles I. G., Taxonomic resolution of the ribosomal RNA operon in bacteria: implications for its use with long-read sequencing, NAR Genomics and Bioinformatics, 2019, vol. 2

de Visser C., Johansson L. F., Kulkarni P., Mei H., Neerincx P., Joeri van der Velde K., Horvatovich P., van Gool A. J., Swertz M. A., Hoen P. A. C. t., et al., Ten quick tips for building FAIR workflows, PLOS Computational Biology, 2023, vol. 19, p. e1011369

Deekshit V. K., Srikumar S., 'To be, or not to be'–The dilemma of 'silent' antimicrobial resistance genes in bacteria, Journal of Applied Microbiology, 2022, vol. 133, p. 2902

Di Tommaso P., Chatzou M., Floden E. W., Barja P. P., Palumbo E., Notredame C., Nextflow enables reproducible computational workflows, Nature Biotechnology, 2017, vol. 35, p. 316

Diancourt L., Passet V., Verhoef J., Grimont P. A. D., Brisse S., Multilocus Sequence Typing of Klebsiella pneumoniae Nosocomial Isolates, Journal of Clinical Microbiology, 2005, vol. 43, p. 4178

Djaffardjy M., Marchment G., Sebe C., Blanchet R., Belhajjame K., Gaignard A., Lemoine F., Cohen-Boulakia S., Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems, Computational and Structural Biotechnology Journal, 2023, vol. 21, p. 2075–2085

Djordjevic S. P., Jarocki V. M., Seemann T., Cummins M. L., Watt A. E., Drigo B., Wyrsch E. R., Reid C. J., Donner E., Howden B. P., Genomic surveillance for antimicrobial resistance – a One Health perspective, Nature Reviews Genetics, 2023, vol. 25, p. 142–157

Dodds D. R., Antibiotic resistance: A current epilogue, Biochemical Pharmacology, 2017, vol. 134, p. 139

Dong H., Li Y., Cheng J., Xia Z., Liu W., Yan T., Chen F., Wang Z., Li R., Shi J., et al., Genomic Epidemiology Insights on NDM-Producing Pathogens Revealed the Pivotal Role of Plasmids on blaNDM Transmission, Microbiology Spectrum, 2022, vol. 10, p. e0215621

Dos Santos S., Moussounda M., Togola M., Avoune Nguema E., Matteya C., Bignoumba M., Onanga R., Lekana-Douki J.-B., François P., van der Mee-Marquet N., Carbapenem-producing Enterobacteriaceae in mothers and newborns in southeast Gabon, 2022, Frontiers in Cellular and Infection Microbiology, 2024, vol. 14

Du D., Wang-Kan X., Neuberger A., van Veen H. W., Pos K. M., Piddock L. J., Luisi B. F., Multidrug efflux pumps: structure, function and regulation, Nature Reviews Microbiology, 2018, vol. 16, p. 523

Durão P., Balbontín R., Gordo I., Evolutionary Mechanisms Shaping the Maintenance of Antibiotic Resistance, Trends in Microbiology, 2018, vol. 26, p. 677

Escobar Pérez J. A., Olarte Escobar N. M., Castro-Cardozo B., Valderrama Márquez I. A., Garzón Aguilar M. I., Martinez De La Barrera L., Barrero Barreto E. R., Marquez-Ortiz R. A., Moncada Guayazán M. V., Vanegas Gómez N., Outbreak of NDM-1-Producing Klebsiella pneumoniae in a Neonatal Unit in Colombia, Antimicrobial Agents and Chemotherapy, 2013, vol. 57, p. 1957

Ewels P., Magnusson M., Lundin S., Käller M., MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, 2016, vol. 32, p. 3047–3048

Ewels P. A., Peltzer A., Fillinger S., Patel H., Alneberg J., Wilm A., Garcia M. U., Di Tommaso P., Nahnsen S., The nf-core framework for community-curated bioinformatics pipelines, Nature Biotechnology, 2020, vol. 38, p. 276–278

Farzand R., Rajakumar K., Zamudio R., Oggioni M. R., Barer M. R., O'Hare H. M., ICEKp2: description of an integrative and conjugative element in Klebsiella pneumoniae, co-occurring and interacting with ICEKp1, Scientific Reports, 2019, vol. 9, p. 13892

Feldgarden M., Brover V., Haft D. H., Prasad A. B., Slotta D. J., Tolstoy I., Tyson G. H., Zhao S., Hsu C.-H., McDermott P. F., et al., Using the NCBI AMRFinder Tool to Determine Antimicrobial Resistance Genotype-Phenotype Correlations Within a Collection of NARMS Isolates, bioRxiv, 2019, p. 550707

Feng L., Zhang M., Fan Z., Population genomic analysis of clinical ST15 Klebsiella pneumoniae strains in China, Frontiers in Microbiology, 2023, vol. 14

Ferreira R. L., Da Silva B. C., Rezende G. S., Nakamura-Silva R., Pitondo-Silva A., Campanini E. B., Brito M. C., Da Silva E. M., De Melo Freire C. C., Da Cunha A. F.,

et al., High prevalence of multidrug-resistant klebsiella pneumoniae harboring several virulence and β-lactamase encoding genes in a brazilian intensive care unit, Frontiers in Microbiology, 2019, vol. 10

Fleming A., On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenz*ae*, British journal of experimental pathology, 1929, vol. 10, p. 226

Fokam J., Essomba R. G., Njouom R., Okomo M.-C. A., Eyangoh S., Godwe C., Tegomoh B., Otshudiema J. O., Nwobegahay J., Ndip L., et al., Genomic surveillance of SARS-CoV-2 reveals highest severity and mortality of delta over other variants: evidence from Cameroon, Scientific Reports, 2023, vol. 13

Follador R., Heinz E., Wyres K. L., Ellington M. J., Kowarik M., Holt K. E., Thomson N. R., The diversity of Klebsiella pneumoniae surface polysaccharides, Microbial Genomics, 2016, vol. 2

Fonseca E. L., Morgado S. M., Freitas F. S., Bighi N. S., Cipriano R., Vicente A. C. P., Unveiling the genome of a high-risk pandrug-resistant Klebsiella pneumoniae emerging in the Brazilian Amazon Region, 2022, Memórias do Instituto Oswaldo Cruz, 2023, vol. 118

Fournier P.-E., Drancourt M., Colson P., Rolain J.-M., La Scola B., Raoult D., Modern Clinical Microbiology: New Challenges and Solutions, Nat. Rev. Microbiol., 2013, vol. 11, p. 574

Fuga B., Ferreira M. L., Cerdeira L. T., de Campos P. A., Dias V. L., Rossi I., Machado L. G., Lincopan N., Gontijo-Filho P. P., Ribas R. M., Novel small IncX3 plasmid carrying the blaKPC-2 gene in high-risk Klebsiella pneumoniae ST11/CG258, Diagnostic Microbiology and Infectious Disease, 2020, vol. 96, p. 114900

Gao H., Liu Y., Wang R., Wang Q., Jin L., Wang H., The transferability and evolution of NDM-1 and KPC-2 co-producing Klebsiella pneumoniae from clinical settings, EBioMedicine, 2020, vol. 51, p. 102599

Gaurav A., Bakht P., Saini M., Pandey S., Pathania R., Role of bacterial efflux pumps in antibiotic resistance, virulence, and strategies to discover novel efflux pump inhibitors, Microbiology, 2023, vol. 169

Ghai I., A Barrier to Entry: Examining the Bacterial Outer Membrane and Antibiotic Resistance, Applied Sciences, 2023, vol. 13, p. 4238

Ghorbani M., Emamie A., Zolfaghari P., Zarei A., Prevalence and antibiotic resistance of ESKAPE pathogens isolated from patients with bacteremia in Tehran, Iran, Indian Journal of Medical Specialities, 2023, vol. 14, p. 97

Gifford D. R., Berríos-Caro E., Joerres C., Suñé M., Forsyth J. H., Bhattacharyya A., Galla T., Knight C. G., Mutators can drive the evolution of multi-resistance to antibiotics, PLOS Genetics, 2023, vol. 19, p. e1010791

Gold H., Moellering R., Antimicrobial-drug resistance, The New England journal of medicine, 1996, vol. 335, p. 1445

Goldstone R. J., Smith D. G., A population genomics approach to exploiting the accessory'resistome' of escherichia coli, Microbial Genomics, 2017, vol. 3

Gomes M. Z. R., de Lima E. M., Martins Aires C. A., Pereira P. S., Yim J., Silva F. H., Rodrigues C. A. S., Oliveira T. R. T. e., da Silva P. P., Eller C. M., et al., Outbreak report of polymyxin-carbapenem-resistant Klebsiella pneumoniae causing untreatable infections evidenced by synergy tests and bacterial genomes, Scientific Reports, 2023, vol. 13

Gorrie C. L., Mirčeta M., Wick R. R., Edwards D. J., Thomson N. R., Strugnell R. A., Pratt N. F., Garlick J. S., Watson K. M., Pilcher D. V., et al., Gastrointestinal Carriage Is a Major Reservoir of Klebsiella pneumoniae Infection in Intensive Care Patients, Clinical Infectious Diseases, 2017, vol. 65, p. 208

Gorrie C. L., Mirčeta M., Wick R. R., Judd L. M., Lam M. M. C., Gomi R., Abbott I. J., Thomson N. R., Strugnell R. A., Pratt N. F., Garlick J. S., Watson K. M., Hunter P. C., Pilcher D. V., McGloughlin S. A., Spelman D. W., Wyres K. L., Jenney A. W. J.,

Holt K. E., Genomic dissection of Klebsiella pneumoniae infections in hospital patients reveals insights into an opportunistic pathogen, Nature Communications, 2022, vol. 13

Grant J. R., Arantes A. S., Stothard P., Comparing thousands of circular genomes using the CGView Comparison Tool, BMC Genomics, 2012, vol. 13

Griffith D. M., Veech J. A., Marsh C. J., cooccur: Probabilistic Species Co-Occurrence Analysis inR, Journal of Statistical Software, 2016, vol. 69

Grüning B., Chilton J., Köster J., Dale R., Soranzo N., van den Beek M., Goecks J., Backofen R., Nekrutenko A., Taylor J., Practical Computational Reproducibility in the Life Sciences, Cell Systems, 2018, vol. 6, p. 631

Grüning B. A., Rasche E., Rebolledo-Jaramillo B., Eberhard C., Houwaart T., Chilton J., Coraor N., Backofen R., Taylor J., Nekrutenko A., Jupyter and Galaxy: Easing entry barriers into complex data analyses for biomedical researchers, PLOS Computational Biology, 2017, vol. 13, p. 1

Gu D., Dong N., Zheng Z., Lin D., Huang M., Wang L., Chan E. W. C., Shu L., Yu J., Zhang R., et al., A fatal outbreak of ST11 carbapenem-resistant hypervirulent Klebsiella pneumoniae in a Chinese hospital: a molecular epidemiological study, The Lancet Infectious Diseases, 2018, vol. 18, p. 37

Guo H., Wu Y., Li L., Wang J., Xu J., He F., Global emergence of carbapenem-resistant Klebsiella pneumoniae co-carrying multiple carbapenemases, Computational and Structural Biotechnology Journal, 2023, vol. 21, p. 3557–3563

Guo X., Chen R., Wang Q., Li C., Ge H., Qiao J., Li Y., Global prevalence, characteristics, and future prospects of IncX3 plasmids: A review, Frontiers in Microbiology, 2022, vol. 13, p. 979558

Hao M., Ye M., Shen Z., Hu F., Yang Y., Wu S., Xu X., Zhu S., Qin X., Wang M., Porin deficiency in carbapenem-resistant enterobacter aerogenes strains, Microbial Drug Resistance, 2018, vol. 24, p. 1277

He Z., Xu W., Zhao H., Li W., Dai Y., Lu H., Zhao L., Zhang C., Li Y., Sun B., Epidemiological characteristics an outbreak of ST11 multidrug-resistant and hypervirulent Klebsiella pneumoniae in Anhui, China, Frontiers in Microbiology, 2022, vol. 13, p. 996753

Hendriksen R. S., Bortolaia V., Tate H., Tyson G. H., Aarestrup F. M., McDermott P. F., Using Genomics to Track Global Antimicrobial Resistance, Frontiers in Public Health, 2019, vol. 7

Heng H., Yang X., Ye L., Tang Y., Guo Z., Li J., Chan E. W.-C., Zhang R., Chen S., Global genomic profiling of Klebsiella pneumoniae: A spatio-temporal population structure analysis, International Journal of Antimicrobial Agents, 2024, vol. 63, p. 107055

Ho B. T., Dong T. G., Mekalanos J. J., A view to a kill: the bacterial type VI secretion system, Cell Host & Microbe, 2014, vol. 15, p. 9

Holt K. E., Wertheim H., Zadoks R. N., Baker S., Whitehouse C. A., Dance D., Jenney A., Schultsz C., Kuntaman K., Newton P. N., et al., Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health, Proceedings of the National Academy of Sciences, 2015, vol. 112, p. E3574

Hughes D., Andersson D. I., Evolutionary Trajectories to Antibiotic Resistance, Annual Review of Microbiology, 2017, vol. 71, p. 579

Hunt M., Silva N. D., Otto T. D., Parkhill J., Keane J. A., Harris S. R., Circlator: automated circularization of genome assemblies using long sequencing reads, Genome Biology, 2015, vol. 16, p. 294

Iovleva A., Doi Y., Carbapenem-Resistant Enterobacteriaceae, Clinics in Laboratory Medicine, 2017, vol. 37, p. 303

Jain C., Rodriguez-R L. M., Phillippy A. M., Konstantinidis K. T., Aluru S., High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries, Nature Communications, 2018, vol. 9

Janssen L., de Almeida F. M., Damasceno T. A. S., Baptista R. d. P., Pappas G. J., de Campos T. A., Martins V. d. P., A Novel Multidrug Resistant, Non-Tn4401 Genetic Element-Bearing, Strain of Klebsiella pneumoniae Isolated From an Urban Lake With Drinking and Recreational Water Reuse, Frontiers in Microbiology, 2021, vol. 12

Jolley K. A., Bray J. E., Maiden M. C. J., Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications, Wellcome Open Research, 2018, vol. 3, p. 124

Jubeh B., Breijyeh Z., Karaman R., Resistance of Gram-Positive Bacteria to Current Antibacterial Agents and Overcoming Approaches, Molecules, 2020, vol. 25, p. 2888

Kan B., Zhou H., Du P., Zhang W., Lu X., Qin T., Xu J., Transforming bacterial disease surveillance and investigation using whole-genome sequence to probe the trace, Frontiers of Medicine, 2018, vol. 12, p. 23

Kanehisa M., Furumichi M., Sato Y., Kawashima M., Ishiguro-Watanabe M., KEGG for taxonomy-based analysis of pathways and genomes, Nucleic Acids Research, 2022, vol. 51, p. D587–D592

Kanehisa M., Goto S., KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Research, 2000, vol. 28, p. 27

Kiffer C. R. V., Rezende T. F. T., Costa-Nobre D. T., Marinonio A. S. S., Shiguenaga L. H., Kulek D. N. O., Arend L. N. V. S., Santos I. C. D. O., Sued-Karam B. R., Rocha-de Souza C. M., et al., A 7-Year Brazilian National Perspective on Plasmid-Mediated Carbapenem Resistance in Enterobacterales, *Pseudomonas aeruginosa* , and *Acinetobacter baumannii* Complex and the Impact of the Coronavirus Disease 2019 Pandemic on Their Occurrence, Clinical Infectious Diseases, 2023, vol. 77, p. S29

Kim J., Cater R. J., Choy B. C., Mancia F., Structural Insights into Transporter-Mediated Drug Resistance in Infectious Diseases, Journal of Molecular Biology, 2021, vol. 433, p. 167005

Kolmogorov M., Yuan J., Lin Y., Pevzner P. A., Assembly of long, error-prone reads using repeat graphs, Nature Biotechnology, 2019, vol. 37, p. 540

Kong H. K., Pan Q., Lo W. U., Liu X., Law C. O., fung Chan T., Ho P. L., Lau T. C. K., Fine-tuning carbapenem resistance by reducing porin permeability of bacteria activated in the selection process of conjugation, Scientific Reports, 2018, vol. 8

Koren S., Phillippy A. M., One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly, Current Opinion in Microbiology, 2015, vol. 23, p. 110–120

Kurtzer G. M., Sochat V., Bauer M. W., Singularity: Scientific containers for mobility of compute, PLoS ONE, 2017, vol. 12, p. e0177459

Lam M. M., Wick R. R., Wyres K. L., Gorrie C. L., Judd L. M., Jenney A. W., Brisse S., Holt K. E., Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in Klebsiella pneumoniae populations, Microbial genomics, 2018, vol. 4

Lam M. M. C., Wick R. R., Watts S. C., Cerdeira L. T., Wyres K. L., Holt K. E., A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex, Nat Commun, 2021, vol. 12

Lee A. H. Y., Porto W. F., de Faria Jr C., Dias S. C., Alencar S. A., Pickard D. J., Hancock R. E. W., Franco O. L., Genomic insights into the diversity, virulence and resistance of Klebsiella pneumoniae extensively drug resistant clinical isolates, Microbial Genomics, 2021, vol. 7

Lee C.-R., Lee J. H., Park K. S., Jeon J. H., Kim Y. B., Cha C.-J., Jeong B. C., Lee S. H., Antimicrobial Resistance of Hypervirulent Klebsiella pneumoniae: Epidemiology, Hypervirulence-Associated Determinants, and Resistance Mechanisms, Frontiers in Cellular and Infection Microbiology, 2017, vol. 7, p. 483

Lee C.-R., Lee J. H., Park K. S., Kim Y. B., Jeong B. C., Lee S. H., Global Dissemination of Carbapenemase-Producing Klebsiella pneumoniae: Epidemiology, Genetic Context, Treatment Options, and Detection Methods, Front. Microbiol., 2016, vol. 7

Lee F., Diagnostics and laboratory role in outbreaks, Current Opinion in Infectious Diseases, 2017, vol. 30, p. 419

Lee I. P. A., Eldakar O. T., Gogarten J. P., Andam C. P., Bacterial cooperation through horizontal gene transfer, Trends in Ecology & Evolution, 2022, vol. 37, p. 223–232

Leipzig J., A review of bioinformatic pipeline frameworks, Briefings in Bioinformatics, 2016, vol. 18, p. 530

Li C., Jiang X., Yang T., Ju Y., Yin Z., Yue L., Ma G., Wang X., Jing Y., Luo X., et al., Genomic Epidemiology of Carbapenemase-producing Klebsiella pneumoniae in China, Genomics, Proteomics & Bioinformatics, 2022, vol. 20, p. 1154–1167

Li K., Xu P., Wang J., Yi X., Jiao Y., Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement, Nature Communications, 2023, vol. 14, p. 6556

Li W., O'Neill K. R., Haft D. H., DiCuccio M., Chetvernin V., Badretdin A., Coulouris G., Chitsaz F., Derbyshire M. K., Durkin A. S., et al., RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation, Nucleic Acids Research, 2020, vol. 49, p. D1020

Li X., Chen F., Chen Y., Gcluster: a simple-to-use tool for visualizing and comparing genome contexts for numerous genomes, Bioinformatics, 2020, vol. 36, p. 3871

Li Y., Hu D., Ma X., Li D., Tian D., Gong Y., Jiang X., Convergence of carbapenem resistance and hypervirulence leads to high mortality in patients with postoperative Klebsiella pneumoniae meningitis, Journal of Global Antimicrobial Resistance, 2021, vol. 27, p. 95

Liu C., Du P., Yang P., Yi J., Lu M., Shen N., Emergence of Extensively Drug-Resistant and Hypervirulent KL2-ST65 Klebsiella pneumoniae Harboring blaKPC-3 in Beijing, China, Microbiol Spectr, 2022, vol. 10

Liu C., Yang P., Zheng J., Yi J., Lu M., Shen N., Convergence of two serotypes within the epidemic ST11 KPC-producing Klebsiella pneumoniae creates the "Perfect Storm" in a teaching hospital, BMC Genomics, 2022, vol. 23, p. 693

Liu L., Ye M., Li X., Li J., Deng Z., Yao Y.-F., Ou H.-Y., Identification and Characterization of an Antibacterial Type VI Secretion System in the Carbapenem-Resistant Strain Klebsiella pneumoniae HS11286, Frontiers in Cellular and Infection Microbiology, 2017, vol. 7, p. 442

Liu Y. M., Li B. B., Zhang Y. Y., Zhang W., Shen H., Li H., Cao B., Clinical and molecular characteristics of emerging hypervirulent Klebsiella pneumoniae bloodstream infections in mainland China, Antimicrobial Agents and Chemotherapy, 2014, vol. 58, p. 5379

Long K. S., Poehlsgaard J., Kehrenberg C., Schwarz S., Vester B., The Cfr rRNA methyltransferase confers resistance to phenicols, lincosamides, oxazolidinones, pleuromutilins, and streptogramin A antibiotics, Antimicrobial Agents and Chemotherapy, 2006, vol. 50, p. 2500

Longo L. G., de Sousa V. S., Kraychete G. B., Justo-da Silva L. H., Rocha J. A., Superti S. V., Bonelli R. R., Martins I. S., Moreira B. M., Colistin resistance emerges in pandrug-resistant Klebsiella pneumoniae epidemic clones in Rio de Janeiro, Brazil, International Journal of Antimicrobial Agents, 2019, vol. 54, p. 579–586

Luo Y., Wang Y., Ye L., Yang J., Molecular epidemiology and virulence factors of pyogenic liver abscess causing Klebsiella pneumoniae in China, Clinical Microbiology and Infection, 2014, vol. 20, p. O818

Lupo V., Mercuri P. S., Frère J.-M., Joris B., Galleni M., Baurain D., Kerff F., An Extended Reservoir of Class-D Beta-Lactamases in Non-Clinical Bacterial Strains, Microbiology Spectrum, 2022, vol. 10, p. e00315

Magiorakos A. P., Srinivasan A., Carey R. B., Carmeli Y., Falagas M. E., Giske C. G., Harbarth S., Hindler J. F., Kahlmeter G., Olsson-Liljequist B., et al., Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: An international expert proposal for interim standard definitions for acquired resistance, Clinical Microbiology and Infection, 2012, vol. 18, p. 268

Mainali K. P., Slud E., CooccurrenceAffinity: An R package for computing a novel metric

of affinity in co-occurrence data that corrects for pervasive errors in traditional indices, bioRxiv, 2022

Manni M., Berkeley M. R., Seppey M., Simão F. A., Zdobnov E. M., BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes, Molecular Biology and Evolution, 2021, vol. 38, p. 4647

Manson A. L., Cohen K. A., Abeel T., Desjardins C. A., Armstrong D. T., Barry C. E., Brand J., Ellner J., Pym A. S., Skrahina A., et al., Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance, Nature Genetics, 2017, vol. 49, p. 395

Marr C. M., Russo T. A., Hypervirulent Klebsiella pneumoniae: a new public health threat, Expert Review of Anti-Infective Therapy, 2019, vol. 17, p. 71

Martin R. M., Bachman M. A., Colonization, infection, and the accessory genome of Klebsiella pneumoniae, Frontiers in Cellular and Infection Microbiology, 2018, vol. 8

Mendes G., Ramalho J. a. F., Duarte A., Pedrosa A., Silva A. C., Méndez L., Caneiras C., First Outbreak of NDM-1-Producing Klebsiella pneumoniae ST11 in a Portuguese Hospital Centre during the COVID-19 Pandemic, Microorganisms, 2022, vol. 10, p. 251

Merkel D., Docker: Lightweight Linux Containers for Consistent Development and Deployment, Linux J., 2014, vol. 2014

Miller W. R., Arias C. A., ESKAPE pathogens: antimicrobial resistance, epidemiology, clinical impact and therapeutics, Nature Reviews Microbiology, 2024

Minh B. Q., Schmidt H. A., Chernomor O., Schrempf D., Woodhams M. D., von Haeseler A., Lanfear R., IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era, Molecular Biology and Evolution, 2020, vol. 37, p. 1530–1534

Mohr K. I., History of antibiotics research, Current Topics in Microbiology and Immunology, 2016, vol. 398, p. 237

Monteiro J., Inoue F. M., Lobo A. P. T., Ibanes A. S., Tufik S., Kiffer C. R., A major monoclonal hospital outbreak of NDM-1-producing Klebsiella pneumoniae ST340 and the first report of ST2570 in Brazil, Infection Control and Hospital Epidemiology, 2019, vol. 40, p. 492

Montelongo Hernandez C., Putonti C., Wolfe A. J., Profiling the plasmid conjugation potential of urinary Escherichia coli, Microbial Genomics, 2022, vol. 8

Muir P., Li S., Lou S., Wang D., Spakowicz D. J., Salichos L., Zhang J., Weinstock G. M., Isaacs F., Rozowsky J., et al., The real cost of sequencing: scaling computation to keep pace with data generation, Genome Biol., 2016, vol. 17, p. 53

Munoz-Price L. S., Poirel L., Bonomo R. A., Schwaber M. J., Daikos G. L., Cormican M., Cornaglia G., Garau J., Gniadkowski M., Hayden M. K., et al., Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases, The Lancet. Infectious Diseases, 2013, vol. 13, p. 785

Murigneux V., Roberts L. W., Forde B. M., Phan M.-D., Nhu N. T. K., Irwin A. D., Harris P. N. A., Paterson D. L., Schembri M. A., Whiley D. M., Beatson S. A., Micro-PIPE: validating an end-to-end workflow for high-quality complete bacterial genome construction, BMC Genomics, 2021, vol. 22

Nakamura-Silva R., Cerdeira L., Oliveira-Silva M., da Costa K. R. C., Sano E., Fuga B., Moura Q., Esposito F., Lincopan N., Wyres K., et al., Multidrug-resistant Klebsiella pneumoniae: a retrospective study in Manaus, Brazil, Archives of Microbiology, 2022, vol. 204

Nakamura-Silva R., Oliveira-Silva M., Furlan J. a. P. R., Stehling E. G., Miranda C. E. S., Pitondo-Silva A., Characterization of multidrug-resistant and virulent Klebsiella pneumoniae strains belonging to the high-risk clonal group 258 (CG258) isolated from inpatients in northeastern Brazil, Archives of Microbiology, 2021, vol. 203, p. 4351–4359

Nava R. G., Oliveira-Silva M., Nakamura-Silva R., Pitondo-Silva A., Vespero E. C., New sequence type in multidrug-resistant Klebsiella pneumoniae harboring the bla NDM-1

-encoding gene in Brazil, International Journal of Infectious Diseases, 2019, vol. 79, p. 101

Navon-Venezia S., Kondratyeva K., Carattoli A., Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance, FEMS Microbiology Reviews, 2017, vol. 41, p. 252

Neu H. C., The crisis in antibiotic resistance, Science, 1992, vol. 257, p. 1064

Nicola F., Cejas D., González-Espinosa F., Relloso S., Herrera F., Bonvehí P., Smayevsky J., Figueroa-Espinosa R., Gutkind G., Radice M., Outbreak of Klebsiella pneumoniae ST11 Resistant To Ceftazidime-Avibactam Producing KPC-31 and the Novel Variant KPC-115 during COVID-19 Pandemic in Argentina, Microbiology Spectrum, 2022, vol. 10, p. e03733

Nishino K., Nikaido E., Yamaguchi A., Regulation and physiological function of multidrug efflux pumps in Escherichia coli and Salmonella, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2009, vol. 1794, p. 834–843

on Genomic Surveillance of AMR N. G. H. R. U., Whole-genome sequencing as part of national and international surveillance programmes for antimicrobial resistance: a roadmap, BMJ global health, 2020, vol. 5, p. e002244

Paczosa M. K., Mecsas J., Klebsiella pneumoniae: Going on the Offense with a Strong Defense, Microbiology and Molecular Biology Reviews, 2016, vol. 80, p. 629

Palmeiro J. K., de Souza R. F., Schörner M. A., Passarelli-Araujo H., Grazziotin A. L., Vidal N. M., Venancio T. M., Dalla-Costa L. M., Molecular Epidemiology of Multidrug-Resistant Klebsiella pneumoniae Isolates in a Brazilian Tertiary Hospital, Frontiers in Microbiology, 2019, vol. 10, p. 1669

Peltzer A., Taylor B., Zhou Y., Patel H., nf-core/bacass: nf-core/bacass v1.1.0: "Green Aluminium Shark, Zenodo, 2019

Pendleton J. N., Gorman S. P., Gilmore B. F., Clinical relevance of the ESKAPE pathogens, Expert Review of Anti-Infective Therapy, 2013, vol. 11, p. 297

Perkel J. M., , 2019 Workflow systems turn raw data into scientific knowledge

Peterson E., Kaur P., Antibiotic resistance mechanisms in bacteria: Relationships between resistance determinants of antibiotic producers, environmental bacteria, and clinical pathogens, Frontiers in Microbiology, 2018, vol. 9

Petit R. A., Read T. D., Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes, mSystems, 2020, vol. 5

Pirofski L.-a., Casadevall A., Q&A: What is a pathogen? A question that begs the point, BMC Biology, 2012, vol. 10

Plough H. H., Penicillin resistance of Staphylococcus aureus and its clinical implications, American journal of clinical pathology, 1945, vol. 15, p. 446

Priyam A., Woodcroft B. J., Rai V., Moghul I., Munagala A., Ter F., Chowdhary H., Pieniak I., Maynard L. J., Gibbins M. A., Moon H., Davis-Richardson A., Uludag M., Watson-Haigh N. S., Challis R., Nakamura H., Favreau E., Gómez E. A., Pluskal T., Leonard G., Rumpf W., Wurm Y., Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases, Molecular Biology and Evolution, 2019, vol. 36, p. 2922–2924

Qi Y., Wei Z., Ji S., Du X., Shen P., Yu Y., ST11, the dominant clone of KPC-producing Klebsiella pneumoniae in China, Journal of Antimicrobial Chemotherapy, 2011, vol. 66, p. 307

Quijada N. M., Rodríguez-Lázaro D., Eiros J. M., Hernández M., TORMES: an automated pipeline for whole bacterial genome analysis, Bioinformatics, 2019

Rempel A., Wittler R., SANS serif: alignment-free, whole-genome-based phylogenetic reconstruction, Bioinformatics, 2021, vol. 37, p. 4868

Revell L. J., phytools: an R package for phylogenetic comparative biology (and other things), Methods in Ecology and Evolution, 2011, vol. 3, p. 217–223

Riwu K. H. P., Effendi M. H., Rantam F. A., Khairullah A. R., Widodo A., A review: Virulence factors of Klebsiella pneumonia as emerging infection on the food chain, Veterinary World, 2022, p. 2172–2179

Robertson J., Nash J. H. E., MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies, Microbial Genomics, 2018, vol. 4

Rocha R. T., de Almeida F. M., Pappas M. C. R., Pappas G. J., Martins K., Complete Genome Sequence of Pantoea stewartii RON18713 from Brazil Nut Tree Phyllosphere Reveals Genes Involved in Plant Growth Promotion, Microorganisms, 2023, vol. 11

Roe C., Williamson C. H., Vazquez A. J., Kyger K., Valentine M., Bowers J. R., Phillips P. D., Harrison V., Driebe E., Engelthaler D. M., et al., Bacterial Genome wide association studies (bGWAS) and transcriptomics identifies cryptic antimicrobial resistance mechanisms in Acinetobacter baumannii, bioRxiv, 2019, p. 864462

Romanowska J., Reuter N., Trylska J., Comparing aminoglycoside binding sites in bacterial ribosomal RNA and aminoglycoside modifying enzymes, Proteins: Structure, Function and Bioinformatics, 2013, vol. 81, p. 63

Roy Chowdhury P., Hastak P., DeMaere M., Wyrsch E., Li D., Elankumaran P., Dolejska M., Browning G. F., Marenda M. S., Gottlieb T., et al., Phylogenomic analysis of a global collection of Escherichia coli ST38: evidence of interspecies and environmental transmission?, mSystems, 2023, vol. 8

Rozwandowicz M., Brouwer M. S. M., Fischer J., Wagenaar J. A., Gonzalez-Zorn B., Guerra B., Mevius D. J., Hordijk J., Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae, Journal of Antimicrobial Chemotherapy, 2018, vol. 73, p. 1121–1137

Ruiz-Perez C. A., Conrad R. E., Konstantinidis K. T., MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes, BMC Bioinformatics, 2021, vol. 22

Russo T. A., MacDonald U., Hassan S., Camanzo E., LeBreton F., Corey B., McGann P., An Assessment of Siderophore Production, Mucoviscosity, and Mouse Infection Models

for Defining the Virulence Spectrum of Hypervirulent Klebsiella pneumoniae, mSphere, 2021, vol. 6

Russo T. A., Marr C. M., Hypervirulent Klebsiella pneumoniae, Clinical Microbiology Reviews, 2019, vol. 32

Sakoparnig T., Field C., Van Nimwegen E., Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species, eLife, 2021, vol. 10, p. e65366

Schaenzer A. J., Wright G. D., Antibiotic Resistance by Enzymatic Modification of Antibiotic Targets, Trends in Molecular Medicine, 2020, vol. 26, p. 768–782

Schnall J., Rajkhowa A., Ikuta K., Rao P., Moore C. E., Surveillance and monitoring of antimicrobial resistance: limitations and lessons from the GRAM project, BMC Medicine, 2019, vol. 17, p. 176

Schwengers O., Barth P., Falgenhauer L., Hain T., Chakraborty T., Goesmann A., Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores, Microbial Genomics, 2020, vol. 95

Schwengers O., Hoek A., Fritzenwanker M., Falgenhauer L., Hain T., Chakraborty T., Goesmann A., ASA3P: An automatic and scalable pipeline for the assembly, annotation and higher level analysis of closely related bacterial isolates, bioRxiv, 2019, p. 654319

Seemann T., Prokka: Rapid prokaryotic genome annotation, Bioinformatics, 2014, vol. 30, p. 2068

Seki L. M., Pereira P. S., De Souza M. D. P. A., Conceição M. D. S., Marques E. A., Porto C. O., Colnago E. M. L., Alves C. D. F., Gomes D., Assef A. P. D. C., et al., Molecular epidemiology of KPC-2- producing Klebsiella pneumoniae isolates in Brazil: the predominance of sequence type 437, Diagnostic Microbiology and Infectious Disease, 2011, vol. 70, p. 274

Shelenkov A., Mikhaylova Y., Voskanyan S., Egorova A., Akimkin V., Whole-Genome Sequencing Revealed the Fusion Plasmids Capable of Transmission and Acquisition of Both Antimicrobial Resistance and Hypervirulence Determinants in Multidrug-Resistant Klebsiella pneumoniae Isolates, Microorganisms, 2023, vol. 11, p. 1314

Silva K. P. T., Sundar G., Khare A., Efflux pump gene amplifications bypass necessity of multiple target mutations for resistance against dual-targeting antibiotic, Nature Communications, 2023, vol. 14

Smith W. P. J., Wucher B. R., Nadell C. D., Foster K. R., Bacterial defences: mechanisms, evolution and antimicrobial resistance, Nature Reviews Microbiology, 2023, vol. 21, p. 519

Souvorov A., Agarwala R., Lipman D. J., SKESA: strategic k-mer extension for scrupulous assemblies, Genome Biol, 2018, vol. 19

Stahlhut S. G., Chattopadhyay S., Kisiela D. I., Hvidtfeldt K., Clegg S., Struve C., Sokurenko E. V., Krogfelt K. A., Structural and Population Characterization of MrkD, the Adhesive Subunit of Type 3 Fimbriae, Journal of Bacteriology, 2013, vol. 195, p. 5602

Stockdale J. E., Liu P., Colijn C., The potential of genomics for infectious disease forecasting, Nature Microbiology, 2022, vol. 7, p. 1736–1743

Strozzi F., Janssen R., Wurmus R., Crusoe M. R., Githinji G., Di Tommaso P., Belhachemi D., Möller S., Smant G., de Ligt J., et al.,, 2019 Scalable Workflows and Reproducible Data Analysis for Genomics. Springer New York New York, NY pp 723–745

Tang H., Bowers J. E., Wang X., Ming R., Alam M., Paterson A. H., Synteny and Collinearity in Plant Genomes, Science, 2008, vol. 320, p. 486

Tang K., Zhao H., Quinolone Antibiotics: Resistance and Therapy, Infection and Drug Resistance, 2023, vol. Volume 16, p. 811–820

Tang X., Shang J., Ji Y., Sun Y., PLASMe: a tool to identify PLASMid contigs from short-read assemblies using transformer, Nucleic Acids Research, 2023, vol. 51, p. e83

Tanizawa Y., Fujisawa T., Nakamura Y., DFAST: A flexible prokaryotic genome annotation pipeline for faster genome publication, Bioinformatics, 2018, vol. 34, p. 1037

Theuretzbacher U., Global antimicrobial resistance in Gram-negative pathogens and clinical need, Current Opinion in Microbiology, 2017, vol. 39, p. 106

Theuretzbacher U., Carrara E., Conti M., Tacconelli E., Role of new antibiotics for KPC-producing Klebsiella pneumoniae, Journal of Antimicrobial Chemotherapy, 2021, vol. 76, p. i47

Tosta S., Moreno K., Schuab G., Fonseca V., Segovia F. M. C., Kashima S., Elias M. C., Sampaio S. C., Ciccozzi M., Alcantara L. C. J., et al., Global SARS-CoV-2 genomic surveillance: What we have learned (so far), Infection, Genetics and Evolution, 2023, vol. 108, p. 105405

Treangen T. J., Ondov B. D., Koren S., Phillippy A. M., The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes, Genome Biology, 2014, vol. 15

van Dorp L., Wang Q., Shaw L. P., Acman M., Brynildsrud O. B., Eldholm V., Wang R., Gao H., Yin Y., Chen H., et al., Rapid phenotypic evolution in multidrug-resistant Klebsiella pneumoniae hospital outbreak strains, Microbial Genomics, 2019, vol. 5

Vernikos G., Medini D., Riley D. R., Tettelin H., Ten years of pan-genome analyses, Current Opinion in Microbiology, 2015, vol. 23, p. 148–154

Šámal V., Paldus V., Fáčková D., Mečl J., Šrám J., The prevalence of antibiotic-resistant and multidrug-resistant bacteria in urine cultures from inpatients with spinal cord injuries and disorders: an 8-year, single-center study, BMC Infectious Diseases, 2022, vol. 22

Waddington C., Carey M. E., Boinett C. J., Ellen Higginson Veeraraghavan B., Baker S., Exploiting genomics to mitigate the public health impact of antimicrobial resistance, Genome Med, 2022, vol. 14

Waksman S. A., Reilly H. C., Schatz A., Strain Specificity and Production of Antibiotic Substances: V. Strain Resistance of Bacteria to Antibiotic Substances, Especially to Streptomycin, Proceedings of the National Academy of Sciences, 1945, vol. 31, p. 157

Walker B. J., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., Cuomo C. A., Zeng Q., Wortman J., Young S. K., et al., Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement, PLoS ONE, 2014, vol. 9, p. e112963

Walsh T. R., Gales A. C., Laxminarayan R., Dodd P. C., Antimicrobial Resistance: Addressing a Global Threat to Humanity, PLOS Medicine, 2023, vol. 20, p. e1004264

Wan Y., Myall A. C., Boonyasiri A., Bolt F., Ledda A., Mookerjee S., Weiße A. Y., Getino M., Turton J. F., Abbas H., et al., Integrated Analysis of Patient Networks and Plasmid Genomes to Investigate a Regional, Multispecies Outbreak of Carbapenemase-Producing Enterobacterales Carrying Both blaIMP and mcr-9 Genes, The Journal of Infectious Diseases, 2024

Wang J., Feng Y., Zong Z., The Origins of ST11 KL64 Klebsiella pneumoniae: a Genome-Based Study, Microbiology Spectrum, 2023, vol. 11

Wang Q., Wang X., Wang J., Ouyang P., Jin C., Wang R., Zhang Y., Jin L., Chen H., Wang Z., et al., Phenotypic and Genotypic Characterization of Carbapenem-resistant Enterobacteriaceae: Data From a Longitudinal Large-scale CRE Study in China (2012–2016), Clinical Infectious Diseases, 2018, vol. 67, p. S196

Wang X., Yu D., Chen L., Antimicrobial resistance and mechanisms of epigenetic regulation, Frontiers in Cellular and Infection Microbiology, 2023, vol. 13

Wang X., Zhao J., Ji F., Chang H., Qin J., Zhang C., Hu G., Zhu J., Yang J., Jia Z., et al., Multiple-Replicon Resistance Plasmids of Klebsiella Mediate Extensive Dissemination of Antimicrobial Genes, Front. Microbiol., 2021, vol. 12

Wantuch P. L., Knoot C. J., Robinson L. S., Vinogradov E., Scott N. E., Harding C. M., Rosen D. A., A heptavalent O-antigen bioconjugate vaccine exhibits differential functional antibody responses against diverse Klebsiella pneumoniae isolates, 2023

Wheeler N. E., Price V., Cunningham-Oakes E., Tsang K. K., Nunn J. G., Midega J. T., Anjum M. F., Wade M. J., Feasey N. A., Peacock S. J., et al., Innovations in genomic antimicrobial resistance surveillance, The Lancet Microbe, 2023, vol. 4, p. e1063–e1070

Wick R. R., Judd L. M., Gorrie C. L., Holt K. E., Completing bacterial genome assemblies with multiplex MinION sequencing, Microbial Genomics, 2017, vol. 3

Wickham H., Ggplot2: Elegant graphics for data analysis 2 edn. Use R!, Springer International Publishing Cham, Switzerland, 2016

Wu H., Li D., Zhou H., Sun Y., Guo L., Shen D., Bacteremia and other body site infection caused by hypervirulent and classic Klebsiella pneumoniae, Microbial Pathogenesis, 2017, vol. 104, p. 254

Wyres K. L., Gorrie C., Edwards D. J., Wertheim H. F., Hsu L. Y., Van Kinh N., Zadoks R., Baker S., Holt K. E., Extensive Capsule Locus Variation and Large-Scale Genomic Recombination within the Klebsiella pneumoniae Clonal Group 258, Genome Biology and Evolution, 2015, vol. 7, p. 1267

Wyres K. L., Holt K. E., Klebsiella pneumoniae Population Genomics and Antimicrobial-Resistant Clones, Trends in Microbiology, 2016, vol. 24, p. 944

Wyres K. L., Lam M. M. C., Holt K. E., Population genomics of Klebsiella pneumoniae, Nature Reviews Microbiology, 2020, vol. 18, p. 344

Wyres K. L., Wick R. R., Gorrie C., Jenney A., Follador R., Thomson N. R., Holt K. E., Identification of Klebsiella capsule synthesis loci from whole genome data, Microbial genomics, 2016, vol. 2, p. e000102

Wyres K. L., Wick R. R., Judd L. M., Froumine R., Tokolyi A., Gorrie C. L., Lam M. M., Duchêne S., Jenney A., Holt K. E., Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae, PLoS Genetics, 2019, vol. 15

Xie M., Yang X., Xu Q., Ye L., Chen K., Zheng Z., Dong N., Sun Q., Shu L., Gu D., et al., Clinical evolution of ST11 carbapenem resistant and hypervirulent Klebsiella pneumoniae, Communications Biology, 2021, vol. 4, p. 650

Yong D., Toleman M. A., Giske C. G., Cho H. S., Sundman K., Lee K., Walsh T. R., Characterization of a new metallo-beta-lactamase gene, bla(NDM-1), and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India, Antimicrobial Agents and Chemotherapy, 2009, vol. 53, p. 5046

Yu G., Smith D. K., Zhu H., Guan Y., Lam T. T.-Y., ggtree: an package for visualization and annotation of phylogenetic trees with their covariates and other associated data, Methods Ecol Evol, 2016, vol. 8, p. 28

Yuan P.-B., Dai L.-T., Zhang Q.-K., Zhong Y.-X., Liu W.-T., Yang L., Chen D.-Q., Global emergence of double and multi-carbapenemase producing organisms: epidemiology, clinical significance, and evolutionary benefits on antimicrobial resistance and virulence, Microbiology Spectrum, 2024, vol. 12

Zhan L., Wang S., Guo Y., Jin Y., Duan J., Hao Z., Lv J., Qi X., Hu L., Chen L., et al., Outbreak by Hypermucoviscous Klebsiella pneumoniae ST11 Isolates with Carbapenem Resistance in a Tertiary Hospital in China, Frontiers in Cellular and Infection Microbiology, 2017, vol. 7, p. 182

Zhao L., Xia X., Yuan T., Zhu J., Shen Z., Li M., Molecular Epidemiology of Antimicrobial Resistance, Virulence and Capsular Serotypes of Carbapenemase-Carrying Klebsiella pneumoniae in China, Antibiotics, 2022, vol. 11, p. 1100

Zhu J., Wang T., Chen L., Du H., Virulence Factors in Hypervirulent Klebsiella pneumoniae, Front. Microbiol., 2021, vol. 12

Zulfiqar S., Shakoori A. R., Molecular characterization, metal uptake and copper induced transcriptional activation of efflux determinants in copper resistant isolates of Klebsiella pneumoniae, Gene, 2012, vol. 510, p. 32–38

Apêndice

Apêndice A

Material Suplementar

Tabela A.1 - Genomas públicos incluídos nas análises comparativas deste trabalho.

Nome	Ano	Localidade	Número de acesso
ERR*	2015	Brasília	PRJEB9325
ECR	2021	Brasília	SAMN41648388
KpBSB56	2021	Brasília	SAMN41648389
KpBSB60	2021	Brasília	SAMN41648403
KpBSB31	2015	Brasília	GCA_022204885
KpV3	2019	Brasília	GCF_019038575
GCF_026223615	2015	Recife	GCF_026223615
GCF_026223555	2016	Jau	GCF_026223555
GCF_026223205	2017	Itatiba	GCF_026223205
GCF_026223155	2017	Rio de Janeiro	GCF_026223155
GCF_026222875	2017	São Caetano do Sul	GCF_026222875
GCF_026222745	2018	Jau	GCF_026222745
GCF_026223755	2018	São Paulo	GCF_026223755
GCF_026223095	2018	Santo André	GCF_026223095
GCF_026223265	2019	Araras	GCF_026223265
GCF_026222945	2019	Araras	GCF_026222945
GCF_026223175	2019	Belo Horizonte	GCF_026223175
GCF_026223135	2019	Botucatu	GCF_026223135
GCF_026223055	2019	Botucatu	GCF_026223055
GCF_026223115	2019	Botucatu	GCF_026223115
GCF_026222935	2019	Botucatu	GCF_026222935
GCF_026222955	2019	Limeira	GCF_026222955
GCF_026223475	2019	Tocantins	GCF_026223475
GCF_026223455	2019	Tocantins	GCF_026223455
GCF_026223435	2019	Tocantins	GCF_026223435
GCF_026223335	2019	Tocantins	GCF_026223335
GCF_026223315	2019	Tocantins	GCF_026223315
GCF_026223285	2019	Tocantins	GCF_026223285
GCF_026222835	2019	Sorocaba	GCF_026222835
GCF_026223065	2020	Araras	GCF_026223065
GCF_026223035	2020	Araras	GCF_026223035
GCF_026223775	2020	Sorocaba	GCF_026223775
GCF_026223715	2020	Sorocaba	GCF_026223715
GCF_026223005	2020	São Paulo	GCF_026223005
GCF_026222865	2020	São Paulo	GCF_026222865
GCF_026222845	2021	Belém	GCF_026222845
GCF_026223395	2021	Campinas	GCF_026223395
GCF_026223375	2021	Campinas	GCF_026223375
GCF_026223735	2021	São Paulo	GCF_026223735
GCF_026223515	2021	São Paulo	GCF_026223515
GCF_026223255	2021	São Paulo	GCF_026223255
GCF_026223235	2021	São Paulo	GCF_026223235
GCF_026222735	2021	São Paulo	GCF_026222735

Tabela A.2 - Comparação de recursos disponíveis em alguns *pipelines* de genômica bacteriana, obtida por tradução livre do artigo

Recurso	Bacannot	Bactopia	ASA3P	MicrobeAnnotator	Nullarbor	TORMES
Tipo de sequenciamento	Leituras curtas e longas	Leituras curtas e longas	Leituras curtas e longas	Leituras curtas	Leituras curtas	Leituras curtas
Tipo de montagem	Leituras curtas, longas e híbrido	Leituras curtas e híbrido	Leituras curtas, longas e híbrido	Apenas leituras curtas pareadas	Apenas leituras curtas pareadas	Apenas leituras curtas pareadas
Leituras "*single-end*"	Sim	Sim	Sim	Não	Não	Não
Fluxo de trabalho	Nextflow	Nextflow	Groovy	Python	Perl + Make	Bash
Retomar se parado	Sim	Sim	Não	Sim	Sim	Não
Compatibilidade com computação de alta performance e nuvem	Sim	Sim	Sim	Não	Não	Não
Processamento em massa a partir de arquivo de configuração	Sim	Sim	Sim	Sim	Sim	Sim
Relatórios de resultados	HTML[a], Navegador de Genoma e Aplicação Shiny	Texto	HTML	Texto e Imagem	HTML	R Markdown
Container disponível	Sim	Sim	Sim	Não	Não	Não
Documentação	Website e Readme	Website	Manual em PDF e Readme	Readme	Readme	Readme
Versão comparada (ano/versão)	v3.2/2022	v2.2.0/2022	v1.3.0/2020	v2.0.5/2021	v2.0.20191013/2019	v1.3.0/2021

[a] `https://en.wikipedia.org/wiki/HTML`

SOFTWARE TOOL ARTICLE

Scalable and versatile container-based pipelines for de novo genome assembly and bacterial annotation. [version 1; peer review: 2 approved, 1 approved with reservations]

Felipe Marques de Almeida [1,2], Tatiana Amabile de Campos[1,3], Georgios Joannis Pappas Jr [1,2]

[1]Departamento de Biologia Celular, Universidade de Brasília, Brasília, DF, 70910-900, Brazil
[2]Programa de Pós-graduação em Biologia Molecular, Universidade de Brasilia, Brasília, FD, 70910-900, Brazil
[3]Programa de Pós-graduação em Biologia Microbiana, Universidade de Brasília, Brasília, DF, 70910-900, Brazil

v1 **First published:** 25 Sep 2023, **12**:1205
https://doi.org/10.12688/f1000research.139488.1
Latest published: 25 Sep 2023, **12**:1205
https://doi.org/10.12688/f1000research.139488.1

Abstract

Background: Advancements in DNA sequencing technology have transformed the field of bacterial genomics, allowing for faster and more cost effective chromosome level assemblies compared to a decade ago. However, transforming raw reads into a complete genome model is a significant computational challenge due to the varying quality and quantity of data obtained from different sequencing instruments, as well as intrinsic characteristics of the genome and desired analyses. To address this issue, we have developed a set of container-based pipelines using Nextflow, offering both common workflows for inexperienced users and high levels of customization for experienced ones. Their processing strategies are adaptable based on the sequencing data type, and their modularity enables the incorporation of new components to address the community's evolving needs.

Methods: These pipelines consist of three parts: quality control, de novo genome
assembly, and bacterial genome annotation. In particular, the genome annotation pipeline provides a comprehensive overview of the genome, including standard gene prediction and functional inference, as well as predictions relevant to clinical applications such as virulence and resistance gene annotation, secondary metabolite detection, prophage and plasmid prediction, and more.

Results: The annotation results are presented in reports, genome browsers, and a web-based application that enables users to explore and interact with the genome annotation results.

Conclusions: Overall, our user-friendly pipelines offer a seamless integration of computational tools to facilitate routine bacterial genomics research. The effectiveness of these is illustrated by examining the sequencing data of a clinical sample of Klebsiella

Open Peer Review

Approval Status ? ✓ ✓

	1	2	3
version 1 25 Sep 2023	? view	✓ view	✓ view

1. **Abdolrahman Khezri**, Innland Norway university of applied sciences, Hamar, Norway

2. **Abhinav Sharma**, Stellenbosch University, Stellenbosch, South Africa
Emilyn Costa Conceição, Stellenbosch University, Stellenbosch, South Africa

3. **Austin G Davis-Richardson**, One Codex, San Francisco, USA

Any reports and responses or comments on the article can be found at the end of the article.

pneumoniae.

Keywords
bacterial genomics, pipelines, nextflow, antibiotic resistance, public health, virulence

This article is included in the Bioinformatics gateway.

Corresponding author: Georgios Joannis Pappas Jr (gpappas@unb.br)

Author roles: Almeida FMd: Conceptualization, Software, Visualization, Writing – Original Draft Preparation; **Campos TAd**: Resources, Writing – Review & Editing; **Pappas Jr GJ**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded in part by a scholarship by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) to FMA and by the grant number 806/2019 from Fundação de Amparo à Pesquisa do Distrito Federal (FAP-DF) to GPJ.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Almeida FMd, Campos TAd and Pappas Jr GJ. **Scalable and versatile container-based pipelines for de novo genome assembly and bacterial annotation. [version 1; peer review: 2 approved, 1 approved with reservations]** F1000Research 2023, **12**:1205 https://doi.org/10.12688/f1000research.139488.1

First published: 25 Sep 2023, **12**:1205 https://doi.org/10.12688/f1000research.139488.1

Introduction

As whole genome sequencing has been established as a routine procedure in research projects worldwide, the computational analysis of sequencing data takes center stage, and it is often the main operational barrier to which biologists stumble (Xuan et al., 2013; Berger and Yu, 2022). Over the years, many open-source software tools were created to tackle different processing steps along intricate computational protocols geared toward different data processing scenarios. Notwithstanding, materializing the data analysis workflow is a non-trivial stage that many biologists face, given challenges ranging from the selection and installation among a vast assortment of computational tools to the logic of enactment of processing steps. Consequently, analyses can be performed in many ways by different groups raising issues of reproducibility and provenance (Grüning et al., 2018; Djaffardjy et al., 2023). Also, individual teams face problems implementing the processing workflow, inventorying the requirements, and optimizing performance and scalability (Wratten et al., 2021; Mölder et al., 2021).

Bacterial whole-genome sequencing has become mainstream in many microbiological settings, such as taxonomy, ecology, and clinical diagnostics. In concert with these applications, several tailored computational workflows, also known as pipelines, were created for bacterial genomics, each with a different underlying design and implementation approach (Petit and Read, 2020). Despite their similarities, each pipeline is unique and may provide different outcomes, considering that they were developed with different components and parameters. Most of the pipelines are designed to work with limited sequencing data types while focusing on specific annotation tasks, such as functional annotation of open reading frames, antibiotic resistance genes, and variant calling using reference genomes (Olawoye et al., 2020; Quijada et al., 2019; Ruiz-Perez et al., 2021; Sserwadda and Mboowa, 2021). Thus, there is still a need for more generic pipelines that give the user an extensive overview of their data while creating visually rich outputs whilst guaranteeing reproducibility.

Here we describe three pipelines built using the workflow composition system Nextflow (Tommaso et al., 2017). These were specifically designed in a modular way to standardize and facilitate bacterial genomic analysis from the standpoint of non-bioinformaticians relieving the issues of installation, configuration, and execution. Altogether, the pipelines are usable in different analytical scenarios, capable of handling data from different sequencing platforms, ranging from small single-genome projects executed on a personal computer to larger multi-genome projects to be executed in cloud computing platforms. We also leverage the use of operating system virtualization, packing the pipelines to use software containers that provide all required supporting programs without the need to install the required operating system and pipeline components.

Together, they offer a seamless exposition of computational tools to provide an easy framework for analyzing and interrogating data in routine bacterial genomics. To illustrate the system's functionality, we provide a full analytical illustration of the processing of a multi-drug-resistant *Klebsiella pneumoniae* strain.

Methods

Implementation

The pipelines have been implemented with Nextflow, a workflow composition and orchestration tool that allows the execution of tasks across multiple heterogeneous computing environments in a portable manner (Tommaso et al., 2017). We have used many concepts discussed and established by the nf-core community and adhered to their development framework (Ewels et al., 2020).

In the pipeline design process, several auxiliary scripts were developed to aid in processing and summarizing intermediate steps and graphical annotation report generation. Several programming languages were used, like Python, R, RMarkdown, and Bash shell scripts. However, the utilization of scripting languages in the pipelines comes with the drawback of increased complexity, as it necessitates the installation of language interpreters and support libraries. To make this as transparent as possible to the end user, all the scripts and core dependencies have been packed into Docker® container images to ensure a consistent distribution and uniform execution regardless of the endpoint hardware and operating system.

We have adopted a modular development approach, resulting in three independent pipelines specializing in the critical steps of a general bacterial genomics pipeline (Figure 1). The first pipeline focuses on data pre-processing and quality control, the second on genome assembly, and the third on genome annotation. Pipelines 1 and 2 can be used independently for any general-purpose sequencing project, while pipeline 3 is optimized for prokaryotic genomes. The pipelines are autonomous, meaning they can be used separately or in combination with other methods, giving users complete control and flexibility (Figure 1). The pipelines provide a comprehensive end-to-end solution for bacterial genomics analyses when used sequentially. When invoked sequentially, the pipelines provide an end-to-end solution for bacterial genomics analyses. The architecture and implementation of each pipeline are detailed in the following sections.

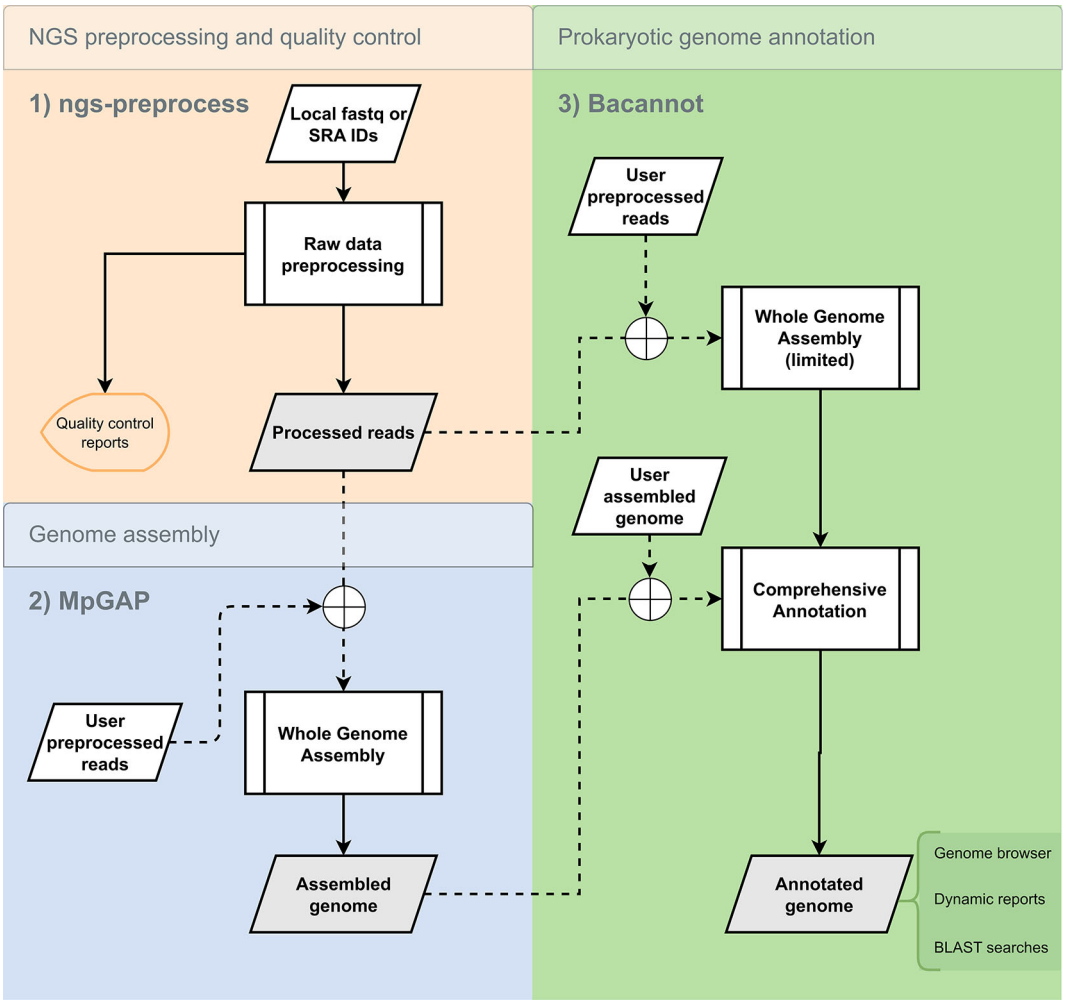

Figure 1. Flowchart of optional sequential execution of the developed pipelines for complete bacterial genomics analysis. The available whole genome assembly module in bacannot is considered limited compared to MpGAP, because it only contains two assembler options. Dashed arrows connected to the crossed circles represent the optional flow of data, highlighting that chaining the pipelines is not required. Gray boxes highlight the pipeline's outputs.

The preprocessing pipeline

The ngs-preprocess pipeline can perform several quality-control steps required for Next-Generation Sequencing (NGS) data assessment. Short or long sequencing reads can be used as input data, and the subsequent steps are determined automatically by the read type or via user configuration settings. These include contamination checking, quality trimming, adapter removal, demultiplexing, file conversion, and graphical report generation (Figure 2). The pipeline accepts data from local storage or deposited in the public repository Sequence Read Archive (SRA). When a list of SRA IDs is given, raw sequencing data (in fastq format) will be automatically downloaded using the entrez-direct and sra-tools tools.

The following steps in the pipeline will be handled automatically based on the sequencing technology indicated in the configuration file. For sample demultiplexing, Porechop (Wick et al., 2017a) is used for Oxford Nanopore Technologies (ONT) reads and lima for PacBio reads. Long-read qualities and statistics are evaluated, and a quality report is generated using NanoPack (Coster et al., 2018). PycoQC (Leger and Leonardi, 2019) can also be used to perform quality control checks on ONT reads. The program fastp (Chen et al., 2018) is used for preprocessing short sequencing reads, including quality assessment, adapter sequence removal, trimming, and reporting. Tools included in the pipeline are summarized in Table 1.

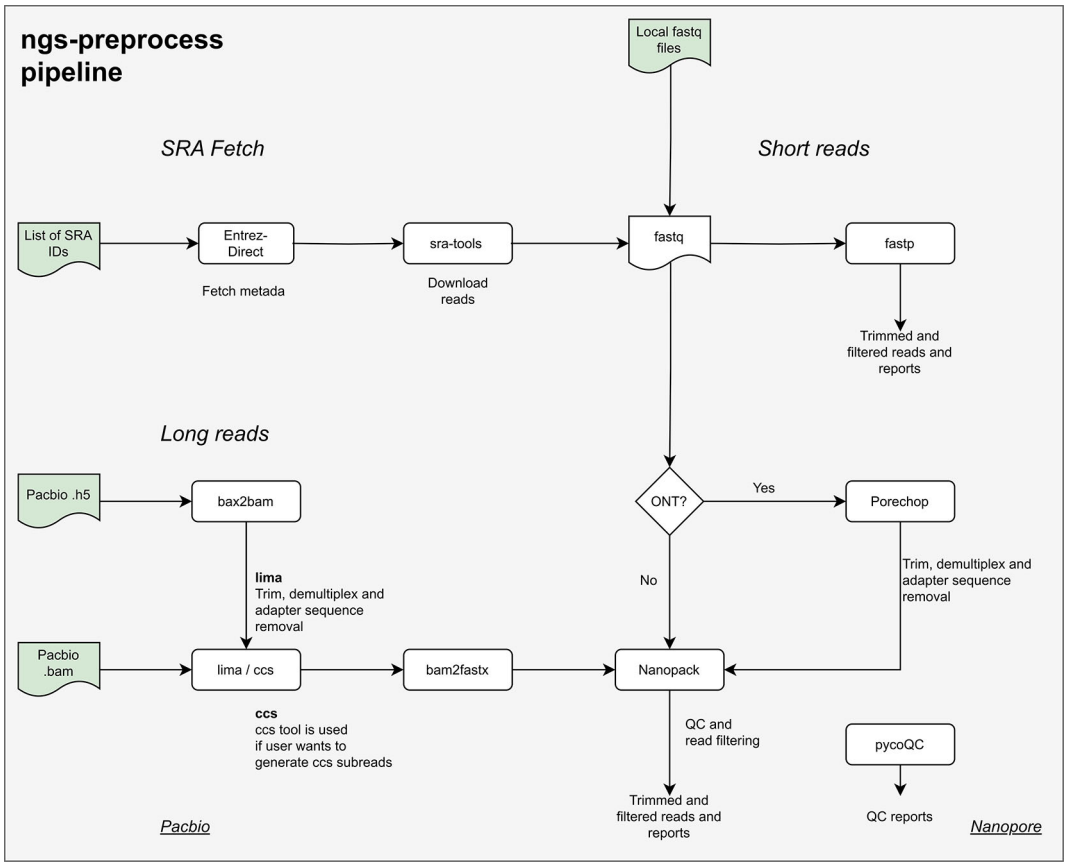

Figure 2. Flowchart of all steps that can be performed by the different workflows available in the ngs-preprocess pipeline.

Table 1. Tools included as part of the ngs-preprocess pipeline (v2.6).

Software	Function	Source code	Version
sra-tools	Data download	github.com/ncbi/sra-tools	3.0.3
entrez-direct	Data download	ncbi.nlm.nih.gov/books/NBK179288/	16.2
fastp	Short-reads processing and reports	github.com/OpenGene/fastp	0.23.2
Porechop	Long-reads processing	github.com/rrwick/Porechop	0.2.4
PycoQC	Nanopore reads reports	github.com/a-slide/pycoQC	2.5.0.3
NanoPack	Long-reads quality control	github.com/wdecoster/nanopack	1.41.0
bam2fastx	Convert PacBio BAM to FASTq	github.com/PacificBiosciences/pbtk	3.1.0
bax2bam	Convert Legacy PacBio to BAM	anaconda.org/bioconda/bax2bam	0.0.9
lima	Demultiplex	github.com/PacificBiosciences/barcoding	2.7.1
ccs	Generate PacBio HiFi	ccs.how	6.4.0

The assembly pipeline

The MpGAP pipeline for *de novo* genome assembly has been designed in an organism and platform-independent manner to perform short or long-read only assemblies, as well as hybrid assemblies using a combination of sequencing technologies (Figure 3). Given the user input, the pipeline automatically selects the assembly mode. When using only short-reads, it performs the genome assembly using any of the programs: SPAdes (Bankevich et al., 2012), Unicycler (Wick et al., 2017b), Shovill and Megahit (Li et al., 2015). On the other hand, when only using long-reads, it uses one or more of the following assemblers: Unicycler, Canu (Koren et al., 2017), Flye (Kolmogorov et al., 2019), Raven (Vaser and Šikić, 2021), Shasta (Shafin et al., 2020), and Wtdbg2 (Ruan and Li, 2019). When both short and long-reads are

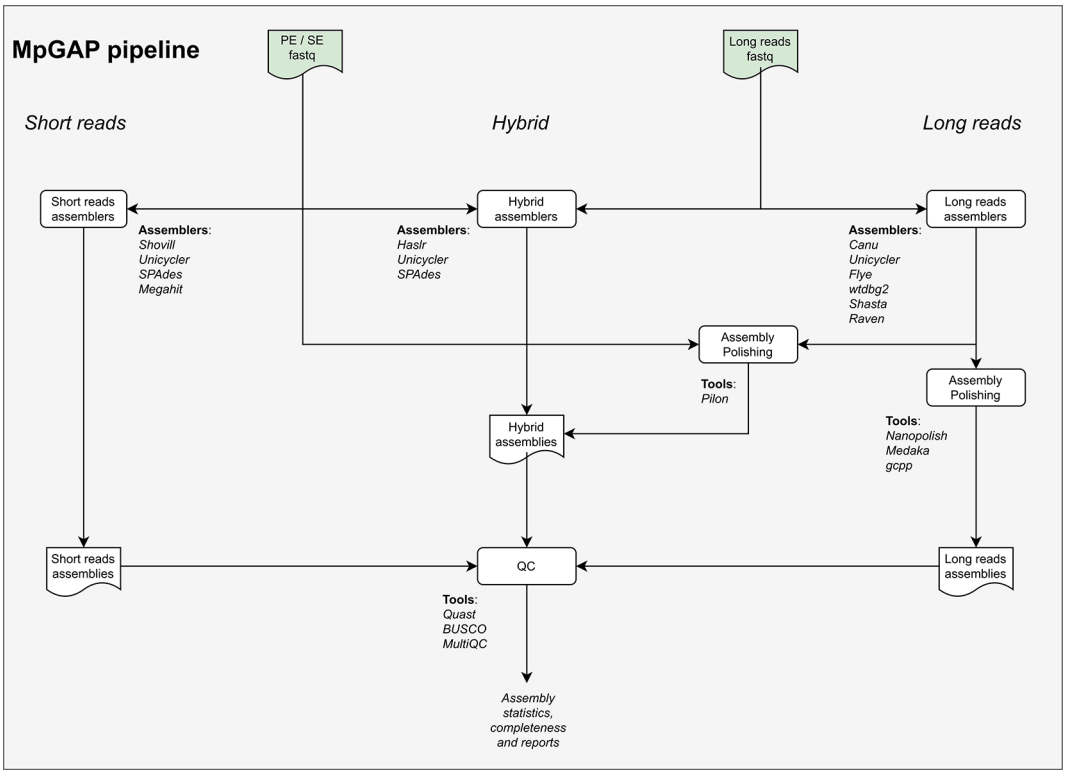

Figure 3. Flowchart of all steps performed by the different workflows available in the MpGAP pipeline.

available, the pipeline is capable of performing two types of hybrid assemblies: (1) A direct hybrid assembly using both short and long-reads sets using HASLR (Haghshenas et al., 2020), SPAdes and Unicycler hybrid assembly modes and (2) A hybrid assembly methodology where long-reads only assembly is produced with one of the assemblers above followed by an error-correction procedure (polishing) using the available short-reads.

MpGAP is capable of polishing long-reads only assemblies using the appropriate tool based on user input: (1) Pilon (Walker et al., 2014) for polishing with short-reads data; (2) Medaka and Nanopolish (Loman et al., 2015) for polishing with nanopore data and (3) GCpp for polishing with PacBio data. Ultimately, assembly statistics are assessed using QUAST (Gurevich et al., 2013) and summarized by MultiQC (Ewels et al., 2016). All the tools that are part of the pipeline are outlined in Table 2.

Table 2. MpGAP pipeline core software components.

Software	Function	Source code	Version
SPAdes	Assembler	github.com/ablab/spades	3.15.3
Unicycler	Assembler	github.com/rrwick/Unicycler	0.4.8
Shovill	Assembler	github.com/tseemann/shovill	1.1.0
Megahit	Assembler	github.com/voutcn/megahit	1.2.9
Haslr	Assembler	github.com/vpc-ccg/haslr	0.8a1
Canu	Assembler	github.com/marbl/canu	2.2
Flye	Assembler	github.com/fenderglass/Flye	2.9
Raven	Assembler	github.com/lbcb-sci/raven	1.6.1
Shasta	Assembler	github.com/chanzuckerberg/shasta	0.8.0
Wtdbg2	Assembler	github.com/ruanjue/wtdbg2	2.5
Pilon	Error correction	github.com/broadinstitute/pilon	1.24

Table 2. *Continued*

Software	Function	Source code	Version
Nanopolish	Error correction	github.com/jts/nanopolish	0.13.2
Medaka	Error correction	github.com/nanoporetech/medaka	1.4.0
Gcpp	Error correction	github.com/PacificBiosciences/gcpp	2.0.2
Quast	Quality control	github.com/ablab/quast	5.0.2
MultiQC	Summary report	github.com/ewels/MultiQC	1.11

The annotation pipeline

The bacannot pipeline is dedicated to annotating prokaryotic genomes. It covers gene prediction, annotation of gene families, mobile genetic elements, and identification of medically relevant features. The pipeline generates dynamic annotation reports and facilitates the navigation of annotated features through a genome browser.

The bacannot workflow is summarized in Figure 4. The pipeline needs a sample sheet file to initiate the process. This file should contain information about the input files, specifying whether they have preprocessed sequencing reads or assembled genomes. If the sample comprises sequencing reads, the pipeline will utilize either Unicycler (Wick et al., 2017b) for short-reads or hybrid assembly or Flye (Kolmogorov et al., 2019) for long-reads assembly.

The process starts with a generic genome annotation using Prokka (Seemann, 2014) or Bakta (Schwengers et al., 2021), followed by the prediction of rRNA sequences using barrnap, identification of closest NCBI RefSeq genome with RefSeq Masher and Multilocus Sequence Type (MLST) assignment with mlst package. Subsequently, plasmid replicons are annotated using the Plasmidfinder (Carattoli et al., 2014) and Platon (Schwengers et al., 2020a) software. Antimicrobial resistance genes are predicted with AMRFinderPlus (Feldgarden et al., 2019), CARD-RGI (Jia et al., 2017), ARGMiner database (Arango-Argoty et al., 2020), and Resfinder (Bortolaia et al., 2020). Virulence genes are annotated with the Virulence Factor Database (VFDB) (Liu et al., 2018) and Victors (Sayers et al., 2018) databases. Prophages are predicted

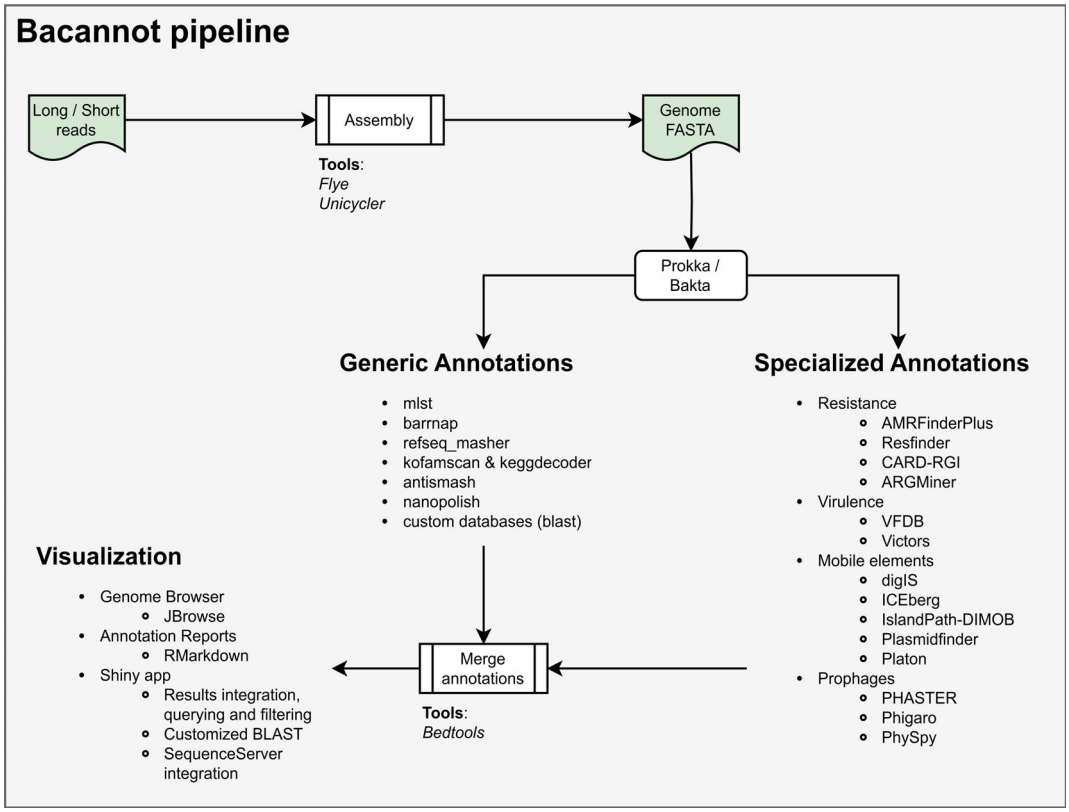

Figure 4. Flowchart of all analytical steps available in the bacannot pipeline.

by Phigaro (Starikova et al., 2019), PhiSpy (Akhter et al., 2012; Edwards et al., 2019), and PHASTER database (Arndt et al., 2016). Genomic islands are predicted with IslandPath-DIMOB (Bertelli and Brinkman, 2018) and plotted with gff-toolbox. Insertion sequences and integrative and conjugative elements (ICEs) are annotated with the ICEberg database (Liu et al., 2018) and the digIS software (Puterová and Martínek, 2021). Orthologs are assigned using KofamScan (Aramaki et al., 2019) and visualized with KEGGDecoder (Graham et al., 2018). Secondary metabolites are annotated with AntiSMASH (Blin et al., 2021). Optionally, users can provide custom databases (with genes of interest) in FASTA files for additional targeted annotations with BLAST+ (Camacho et al., 2009).

The main characteristic of this pipeline is that it is built to aid users less familiar with the tools to investigate and interpret its results with graphically rich reports. The pipeline generates HTML reports summarizing all the results using customizable document templates written in RMarkdown (Xie et al., 2020). A genome browser created with JBrowse (Buels et al., 2016) is also available to explore the annotated features. A custom web application has been developed using the R Shiny framework (Chang et al., 2023) to facilitate further results exploration. It offers users additional features such as dynamic annotation filtering, as well as a built-in sequence similarity search (BLAST) functionality with an interface for executing and visualizing the results produced by SequenceServer (Priyam et al., 2019). The software and databases used in the pipeline are listed in Table 3.

Table 3. Software and databases that have been made part of the bacannot pipeline (v3.2).

Software and Databases	Function	Source code	Version
Unicycler	Assembler	github.com/rrwick/Unicycler	0.4.8
Flye	Assembler	github.com/fenderglass/Flye	2.9
Prokka	Generic annotation	github.com/tseemann/prokka	1.14.6
Bakta	Generic annotation	github.com/oschwengers/bakta	1.6.1
barrnap	rRNA annotation	github.com/tseemann/barrnap	0.9
RefSeq Masher	Find closest reference	github.com/phac-nml/refseq_masher	0.1.2
mlst	Multi-Locus Sequence Typing	github.com/tseemann/mlst	2.22.1
KofamScan	Orthologs annotation	github.com/takaram/kofam_scan	1.3.0
KEEGDecoder	Pathways visualization	github.com/bjtully/BioData	1.3
Nanopolish	Methylation annotation	github.com/jts/nanopolish	0.13.2
bedtools	Data summarization	bedtools.readthedocs.io/en/latest	2.30
gff-toolbox	Data summarization	github.com/fmalmeida/gff-toolbox	0.3
AMRFinderPlus	Resistance annotation	github.com/ncbi/amr/wiki	3.10.30
ARGMiner	Resistance annotation	bench.cs.vt.edu/argminer	-
Resfinder	Resistance annotation	cge.cbs.dtu.dk/services/ResFinder	4.1
CARD-RGI	Resistance annotation	github.com/arpcard/rgi	5.2.1
PHASTER	Prophage annotation	phaster.ca	-
Phigaro	Prophage annotation	github.com/bobeobibo/phigaro	2.3.0
PhySpy	Prophage annotation	github.com/linsalrob/PhiSpy	4.2.21
IslandPath-DIMOB	Genomic Islands prediction	github.com/brinkmanlab/islandpath	1.0.6
Plasmidfinder	Plasmid detection	cge.cbs.dtu.dk/services/PlasmidFinder	2.1.6
Platon	Plasmid detection	github.com/oschwengers/platon	1.6
ICEberg	Integrative and Conjugative elements annotation	db-mml.sjtu.edu.cn/ICEberg	-
digIS	Integrative sequences detection	github.com/janka2012/digIS	1.2
Victors	Virulence annotation	phidias.us/victors	-
VFDB	Virulence annotation	mgc.ac.cn/VFs/main.htm	-

Table 3. *Continued*

Software and Databases	Function	Source code	Version
AntiSMASH	Secondary metabolites annotation	antismash.secondarymetabolites.org/	6.1.1
JBrowse	Results visualization	jbrowse.org/jbrowse1.html	1.16.9
SequenceServer	BLAST visualization	https://sequenceserver.com/	2.0.0

Biological sample, DNA extraction, susceptibility test, and sequencing

The sample KpBSB53 was collected at the University Hospital of Brasilia, Brazil, in April 2016 from the tracheal aspirate of a 41-year-old man. The VITEK 2 system (BioMérieux) was used for microbial identification. Antibiotic susceptibility was tested by the disk diffusion method as described by de Campos et al. (2021). The following antibiotics were tested: Amikacin, Aztreonam, Cefepime, Ceftazidime, Ciprofloxacin, Gentamicin, Imipenem, Levofloxacin, Meropenem, Norfloxacin, Ofloxacin, Piperacillin/tazobactam, Polymyxin, Tobramycin, Ticarcillin/clavulanate (Sensidisc DME - Diagnósticos Microbiológicos Especializados). After 24 h of incubation at 37°C, the sample was classified as susceptible, resistant multiresistant based on what was described by the manufacturer. DNA extraction was performed as described by Ausubel et al. (1992), and extracted DNA was quantified with Nanodrop™. DNA sequencing was performed with both long and short-read technologies. Long-read sequencing was performed using an Oxford Nanopore Technologies MinION Mk1b device, using the rapid barcode kit (SQK-RBK-004) in a R9.4.1 SpotON *flowcell* (FLO-MIN106D). Short-read sequencing was performed by BGI (Shenzhen, China) using paired-end reads with 150 bp on a DNBseq® platform.

Computational analyses

For the processing of the bacterial genome above, the raw long and short-reads have been analyzed sequentially with the developed pipelines. Quality assessment, filtering, and trimming were performed using the ngs-preprocess pipeline v2.6. In our use case, aside from the default parameters, we set the pipeline to correct the short paired-end reads with fastp and filter long-reads based on quality (≥10) and length (≥750) using the parameters *--lreads_min_length* and *--lreads_min_quality*, respectively.

The preprocessed sequencing reads were assembled using the MpGAP pipeline v3.1.4. Genome assembly was performed using the hybrid method where long-reads are first assembled with long-reads assemblers and afterward polished using the short-reads data (*--hybrid_strategy* 2). The hybrid assembly strategy was executed only with the Flye assembler (Kolmogorov et al., 2019) instead of using all the available options to limit the computational burden. The BUSCO (Simão et al., 2015) completion assessment was performed with the bacteria_odb9 dataset, containing 148 expected bacterial genes.

The final polished genomes were annotated using the bacannot pipeline v3.1.5. The pipeline's required databases, such as VFDB (Liu et al., 2018), were downloaded in May 2022. The parameter *--resfinder_species* has been set to "*Klebsiella*". We used Prokka (Seemann, 2014) for generic annotation, and the tool's database was enhanced with the public NCBI Prokaryotic Genome Annotation Pipeline (PGAP) HMM library (Li et al., 2020) by using the *--prokka_use_pgap* parameter.

Operation

The pipeline requires POSIX-compliant systems (e.g., Linux or OS X) or Windows with Windows Subsystem for Linux (WSL2). Pre-installation requirements on the executing computer are Nextflow, and a software management tool (Conda) or a container platform, either Docker or Singularity. Each pipeline requires a configuration file describing the samples and corresponding data to be used. These files also set the specific software to be used in pipeline steps with numerous alternatives (e.g., assembler program) and execution parameters for the whole pipeline or specific software. The configuration files are text files (in YAML format) that should be modified by the user.

With all these requirements satisfied, the execution is conducted non-graphically in a terminal. For example, to trigger the sequential execution of the pipelines, the following command lines should be issued sequentially:

```
nextflow run fmalmeida/ngs-preprocess -profile docker -latest -params-file preprocess-params.yml

nextflow run fmalmeida/mpgap -profile docker -latest -params-file assembly-params.yml

nextflow run fmalmeida/bacannot -profile docker -latest -params-file annotation-params.yml
```

The exemplified command lines would trigger the execution of the latest version of the pipeline's code using docker as the container engine, with all the required parameters configured in the YAML file. Optionally, it is possible to launch specific versions of the pipeline for reproducibility using the "-r" parameter, e.g., "-r v3.2".

In order for users to replicate the analyses of this paper, the configuration files and command lines executed have been made available in a Zenodo project. The configuration file and the sample sheet used for the annotation pipeline have been provided as Supplementary Material (1 and 2) for quick visualization of the expected format. All pipelines have a complete description of workflows and input files in the online manuals.

Results

This paper showcases the utilization of the developed pipelines to conduct a thorough genome examination of an unpublished bacterial sample sequenced with both short and long-reads. The sample, namely KpBSB53 (see Methods), microbiologically classified as *Klebsiella pneumoniae*, was isolated from the tracheal aspirate of a 41-year-old man at the University Hospital of Brasilia. The isolate was susceptible to all the antibiotics tested.

The pipeline assembles and annotates genes for prokaryotes and performs additional annotation steps when examining a clinical sample. This feature makes the tool especially useful for investigating antimicrobial resistance. Using a clinical sample exposes additional tasks included in the pipeline that apply to antimicrobial resistance studies. Subsequently, the KpBSB53 sample will be used to demonstrate the operation of the pipelines to analyze a clinically relevant bacterial genome, as well as a description of the generated results and how to interpret them.

Reads preprocessing and quality control

The process begins with the ngs-preprocess pipeline, which carries out quality control and cleaning actions essential for subsequent genome assembly and annotation tasks. Additionally, it offers the user an important assessment of the sequencing outcomes of the samples.

The output of the ngs-preprocess pipeline is a directory containing the preprocessed sequencing reads organized according to the sequencing technology of the input files, as indicated in the user configuration file. The preprocessed reads can have sequencing adapter or low-quality stretches removed (trimming) or be discarded entirely if a minimum number of good-quality bases is not reached. The pipeline also provides plots and reports for quality control (QC) inspection.

In the particular case of the KpBSB53 sample, the short-reads obtained were of outstanding quality, displaying quality values ≥60 before preprocessing (Figure 5A and B). The nanopore data fluctuated around the expected average quality and length, with median values of ≈10 and ≈4 kb, respectively (Figure 5C and D).

Genome assembly

After running ngs-preprocess pipeline, users can evaluate the quality of the input sequencing reads and ensure that only data with enough quality will be used for genome assembly. Once selected, the user manually supplies the preprocessed sequencing reads (such as short-reads, Nanopore, or PacBio) to the MpGAP pipeline using a sample sheet in a text file (see Methods), allowing users to choose from an assortment of assembly programs and strategies to perform the assembly. The pipeline will automatically select the appropriate assembly program based on the sequencing reads provided.

The KpBSB53 sample illustrates the pipeline's versatility and ability to work with both short and long-reads. For this sample, the Flye assembler (Kolmogorov et al., 2019) was chosen for a hybrid assembly approach (Chen et al., 2020), which initially has errors owing to sequencing technology, later corrected by including short-read data (polishing) using Pilon (Walker et al., 2014).

The pipeline generates a sub-directory for each sample containing the assembly results, including the initial and polished assembly files, and creates a report using MultiQC (Ewels et al., 2016) that includes assembly quality metrics from Quast (Gurevich et al., 2013) and BUSCO (Simão et al., 2015) to facilitate comparison between the assemblies. The user can choose different programs to analyze the same data, and report files are available to aid in selecting the best assembly.

After analyzing the KpBSB53 sample with this pipeline, we found that it consists of two circular contigs - one representing a chromosome (5.2 Mb) and the other a plasmid (142 kb), essentially a full replicon resolution. The BUSCO metrics results highlight the reference quality level of this genome, with 98.65% ortholog completeness in the bacterial chromosome.

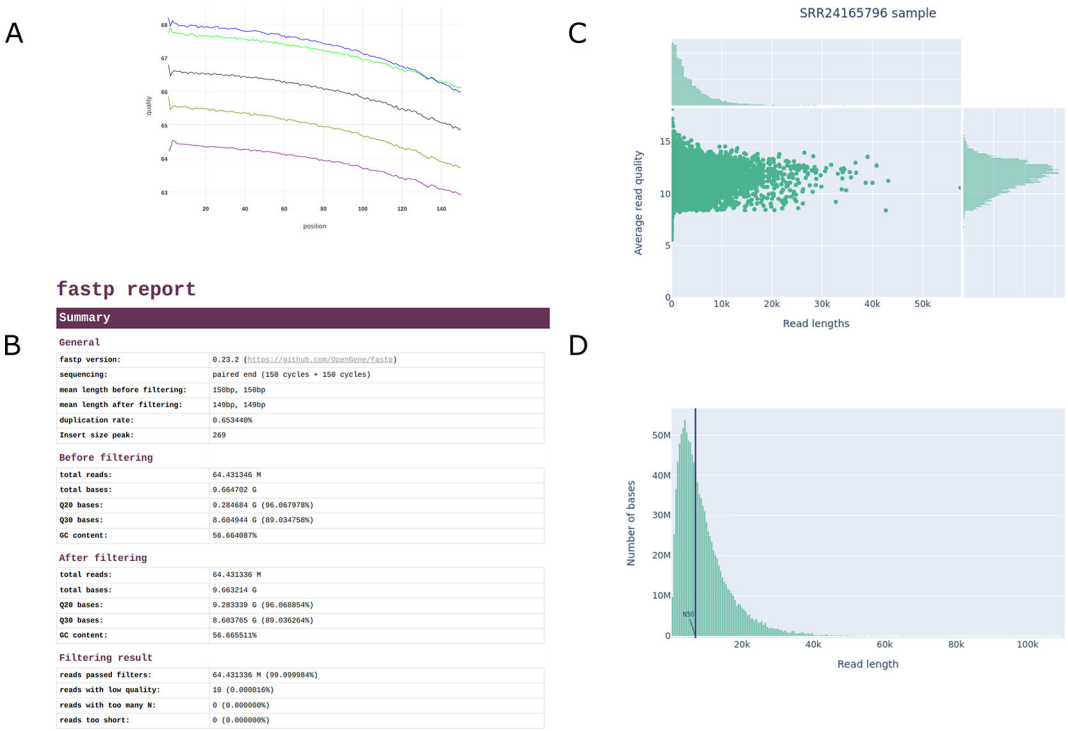

Figure 5. Overview of typical QC outputs generated by the ngs-preprocess pipeline. Figures A and B are generated by the fastp tool and display the base quality of one of the short-read pairs and the summary of reads statistics, respectively. Figures C and D are generated by the NanoPlot tool and display the average read quality per read length and the weighted read length histogram, respectively.

Genome annotation

The assembly reports offer guidelines to assist users in choosing the most suitable assembly alternative for their sample. After making their selection, users can manually indicate the chosen assembly results to the annotation pipeline through the configuration file. Bacannot provides a structured and consistent output, allowing straightforward summarization and examination of its contents. Though multiple annotation stages produce outputs, the primary focus will be on the pipeline's essential results, namely:

1. Complete genome annotation.

2. A web-based application for results visualization and exploration (Figure 6).

3. Automatic HTML reports for resistance, virulence, mobile elements, and annotations from specialized databases (Figure 7).

4. A genome browser for visualization of annotated features (Figure 8).

After all the specialized programs in bacannot have finished their analysis, the results are consolidated into a single General Feature Format file (GFF) and a GenBank format file (GBK) containing the complete genome annotation. These files can be used in other general investigation programs or submitted to NCBI databases. The annotated features are also saved as nucleotide and protein sequences in FASTA format. Moreover, these files are processed internally and presented in the web application as interactive web pages (Figure 6A). This workbench allows users to filter results by text or sequence using the SequenceServer and BLAST applications. The filtered results can be converted into tables of varying formats (Figure 6B and C). The BLAST function included in this workbench allows users to easily annotate other target sequences even after the pipeline is finished. Users can find intersections of alignment results with the genome annotation (Figure 6E and E) or visualize alignments with SequenceServer (Figure 6F).

Figure 6. Overview of the main features available in the web-based application for results exploration made available as part of the bacannot pipeline. Figure A is the homepage of the application for navigation between available features. Figures B and C display the dynamic text-filtering of annotation results. Figures D and E exemplify the interactive results filtering and investigation based on sequence alignment. Figure F shows the SequenceServer tool included in the application for execution and visualization of BLAST alignments.

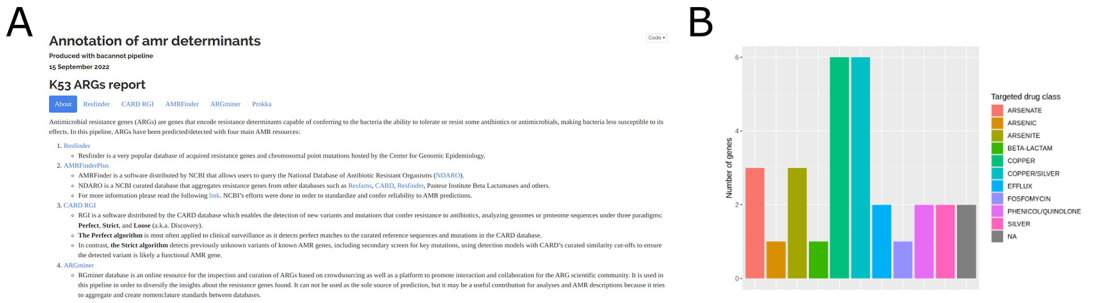

Figure 4: Resistome Predicted using NCBI's AMRFinderplus

Figure 7. Overview of the specialized automatic HTML reports generated by the bacannot pipeline. Figures A and B are screenshots of the antimicrobial resistance (AMR) automatic report, highlighting its homepage containing the annotation description and summary along with a bar plot displaying all the features annotated by the AMRFinderPlus tool. Figure C shows a partial screenshot of the KEGG annotation heatmap autogenerated using KOfamscan and KEGGDecoder tools.

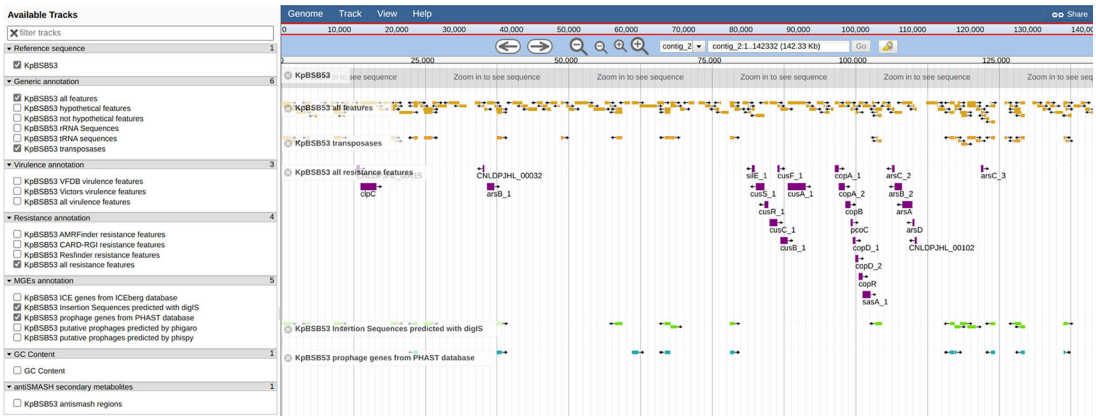

Figure 8. Overview of the automatically rendered genome browser using JBrowse made available as part of the bacannot outputs. For illustration, the annotation tracks showing predicted antibiotic resistance genes and insertion sequences are activated in a region of the KpBSB53 plasmid.

Moreover, detailed reports are generated for feature-specific analyses, such as antibiotic gene prediction, consolidating the results, and providing cross-references to the source databases used for annotation (Figure 7). Besides these clinically-relevant features, the pipeline also has other generic annotation modules that are useful for any bacterial strain, such as prophage and secondary metabolites annotation, KEGG KO annotation, and the possibility of using custom annotation databases or a list of NCBI protein IDs. The KO module, for instance, outputs a text file ready for the KEGG mapper tool to generate pathway figures.

The genome browser allows users to visually explore the annotation results and investigate the genomic context of relevant genes. The browser contains various specialized tracks for the targeted annotations performed by the pipeline, such as for resistance, virulence, prophages, and more. For example, in Figure 8, one can quickly observe a cassette of stress-related resistance genes surrounded by insertion sequences in the plasmid contig.

One of the first results generated by this pipeline is the strain classification based on the alignment-free sequence distance against a database of bacterial genomes (NCBI RefSeq) as calculated by RefSeq Masher. The strain BIDMC 55 (GCF_000692955.1) isolated in the USA was the closest genome to strain KpBSB53 (Mash distance=0.000715202). Additionally, an *in silico* multilocus sequence typing (MLST) is executed using the mlst program against the BIGSdb PubMLST database (Jolley and Maiden, 2010), revealing that our strain belongs to the *Klebsiella pneumoniae* ST 105 group, which has been reported as the driver of a plasmid-mediated outbreak of NDM-1-producing strains in China (Zheng et al., 2016). Using Prokka (Seemann, 2014) for generic annotation, 4937 coding sequences (CDS) have been detected, alongside 25 rRNAs and 87 tRNAs. Another pipeline result is the prediction of plasmid replicons using PlasmidFinder (Carattoli et al., 2014). For the KpBSB53 strain, a single plasmid was predicted and classified as IncFIB(K), a very dynamic replicon that is mostly associated with MDR plasmids in *K. pneumoniae* (Lahlaoui et al., 2015; Tian et al., 2021) and important for virulence as well (Tian et al., 2021).

The pipeline includes several analyses for annotating antimicrobial resistance genes as default. Strain KpBSB53 has only a few acquired antibiotic-resistance genes, namely *blashv*-1, *fosA*, and *oqxAB*, all considered intrinsic to the species and located in the chromosome (Bernardini et al., 2019; Holt et al., 2015). This agrees with our experimental results, which show that this strain is susceptible to all antibiotic classes tested.

Additionally, a set of stress-related genes (Supplementary Material 3) conferring resistance to copper, silver, and other metalloids have been detected in the plasmid sequence (Figure 7). On the other hand, several virulence genes have been detected in the sample's genome using the Virulence Factor Database (VFDB) (Liu et al., 2018). Despite the classical virulence genes normally found in *Klebsiella pneumoniae* strains (Paczosa and Mecsas, 2016), strain KpBSB53 encodes three operons for siderophore biosynthesis: enterobactin (*entABCDEFS*, *fesABCDG*), salmochelin (*iroE*) and aerobactin (*iutA*). Moreover, the type 1 and 3 fimbriae and an *ecp* (*E. coli* common pilus) gene have also been detected. These three fimbriae types are directly related to the adhesion to surfaces, interaction with host cells, and biofilm formation (Alcántar-Curiel et al., 2013). Taken together, the presence of several virulence factors suggests that the KpBSB53 sample may be a hypervirulent strain or at least more virulent than the classical *K. pneumoniae* strains, as discussed by Paczosa and Mecsas (2016).

Table 4. Resource usage metrics for the execution of all pipelines using strain KpBSB53 data, as measured automatically by Nextflow. For bacannot, only the most time and memory-consuming tasks are shown.

Pipeline	Software	Task	Duration (min)	% of pipeline duration	Memory (Gb)
ngs-preprocess	fastp	short-reads preprocessing	7	7.6	2.7
ngs-preprocess	porechop	long-reads preprocessing	70	76.1	4.4
ngs-preprocess	nanopack	filter and quality check	15	16.3	2.1
MpGAP	Flye	assembly	48	5.4	11.3
MpGAP	medaka	long-reads polishing	28	3.18	9.5
MpGAP	pilon	short-reads polishing	625	70.8	10.6
MpGAP	QUAST & Busco	quality check	182	20.6	10.2
MpGAP	MultiQC	reporting	0.2	0.02	0.7
bacannot	Prokka	gene annotation	16	43.2	1
bacannot	Platon	plasmid detection	2	5.4	3.6

Resources usage

Table 4 presents the expected computer resources and timings for the execution of all pipelines on a standard Linux laptop. A Linux Ubuntu 22.04 laptop, with 4 CPUs (8 cores) and 18 Gb RAM, was used in this study. The computational requirements of the ngs-preprocess pipeline are low. However, the time it takes to finish is positively related to the amount of data. A single tool called fastp is used for short-read preprocessing, which takes around 7 minutes to complete and requires approximately 3 Gb of RAM. For nanopore reads, porechop is the most resource-intensive module, taking approximately 1 hour to finish and using approximately 5 Gb of RAM. In terms of the MpGAP assembly pipeline, the execution depth highly depends on the amount of sequencing data provided and is the most resource-intensive step in the workflow. In our hybrid assembly use case, it accounts for half of the overall processing time and requires 11 Gb of RAM (Table 4).

Discussion

In the last decade, advances in DNA sequence generation resulted in a steadfast incorporation of genomics into routine microbiological research practice, ranging from clinical to environmental applications (Didelot and Parkhill, 2022). Despite the significant progress made by the establishment of multiple computational protocols to process genomic data, a considerable gap still needs to be bridged for these to be readily integrated into laboratory practices. The primary reason behind the problem is the lack of necessary bioinformatics foundations, including a shortage of skilled personnel or proper infrastructure.

Our belief is that genomics pipelines should possess certain inherent characteristics to promote their widespread use. These attributes include easy installation and execution, modularity and extensibility, and the generation of user-friendly reports that consolidate the results and foster biological interpretation.

Considering these tenets, we have created a comprehensive set of bacterial genomics pipelines (ngs-preprocess, MpGAP, and bacannot). These specialized pipelines should be invoked in sequence, taking raw data from any sequencing platform and converting them into an annotated genome, emphasizing antimicrobial resistance and virulence genes. The first two modules are not restricted to bacterial data, making it possible to analyze data from other organisms. Conversely, the bacannot pipeline is specific to bacteria but can accept sequencing reads as input, allowing users to assemble and annotate their genome with a single command, which is particularly helpful for those less experienced with bioinformatics.

By following a few simple steps (available online), users can customize the execution of these pipelines and obtain complete results in under a day using a commodity computer. These pipelines utilize technology for modular workflow composition, installation, and execution through virtual containers that package all necessary software requirements. This modularity enables incremental updates in response to community requirements which are transparently pushed to the users unless a specific version is requested for reproducibility purposes.

Table 5. Comparison of selected bacterial genome annotation pipelines.

Feature	Bacannot	Bactopia	ASA3P	MicrobeAnnotator	Nullarbor	TORMES
Sequence technology	Short and Long-reads	Short and Long-reads	Short and Long-reads	Short-reads	Short-reads	
Assembly type	Short-reads, Long-reads and Hybrid	Short-reads and Hybrid	Short-reads, Long-reads and Hybrid	Paired-end short-reads only	Paired-end short-reads only	
Single-end reads	Yes	Yes	Yes	No	No	
Workflow	Nextflow	Nextflow	Groovy	Python	Perl + Make	Bash
Resume if stopped	Yes	Yes	No	Yes	Yes	No
Computing cluster and cloud compatibility	Yes	Yes	Yes	No	No	No
Batch processing from config file	Yes	Yes	Yes	Yes	Yes	Yes
Summary reports	HTML, Genome Browser and Shiny Application	Text	HTML	Text and Image	HTML	R Markdown
Package manager	Nextflow + Github	Nextflow + Conda	Conda	Conda and Brew	Conda	
Container available	Yes	Yes	Yes	No	No	No
Documentation	Website and Readme	Website	PDF Manual and Readme	Readme	Readme	Readme
Source code	https://github.com/fmalmeida/bacannot	https://github.com/bactopia/bactopia	https://github.com/oschwengers/asap	https://github.com/cruizperez/MicrobeAnnotator	https://github.com/tseemann/nullarbor	https://github.com/nmquijada/tormes
Latest release (version/year)	v3.2/2022	v2.2.0/2022	v1.3.0/2020	v2.0.5/2021	v2.0.20191013/2019	v1.3.0/2021

Currently, we are aware of six established bacterial genomics pipelines that are actively maintained and have comparable characteristics to our three pipelines. These include ASA3P (Schwengers et al., 2020b), TORMES (Quijada et al., 2019), Nullarbor, Bactopia (Petit and Read, 2020), MicrobeAnnotator (Ruiz-Perez et al., 2021) and MicroPIPE (Murigneux et al., 2021). Although they have some common elements and comparable assembly and annotation modules, each software has unique features tailored for specific purposes. Their goals and designs vary, providing users with high-quality options for various analytical scenarios. Table 5 outlines the differences in capabilities among the annotation pipelines. We now highlight some of the distinctive features of these pipelines.

Our three pipelines provide a comprehensive bacterial genomics analysis from unprocessed reads to annotation similar to what ASA3P, TORMES, Nullarbor, and Bactopia offer. The analysis provided by MicroPIPE and MicrobeAnnotator is less comprehensive than the other pipelines. MicroPIPE is crafted for genome assembly and covers the process from base-calling to genome polishing. However, our MpGAP genome assembly pipeline is more versatile as it can accommodate data from Illumina, PacBio, and ONT sequencing technologies and has nine assemblers with distinctive assembly strategies. This flexibility enables users to select the most suitable choice for their requirements. MicrobeAnnotator is a pipeline that specializes in functional genome annotation. It offers a KEGG Orthology annotation step through Kofamscan, also available in bacannot. However, MicrobeAnnotator only focuses on the functional annotation of predicted genes using KEGG, UniProt, RefSeq, and Trembl. Unlike other pipelines, it does not include extra modules like virulence and resistance gene annotation.

Analogous to Bactopia, our pipelines are very flexible and customizable. Compared to bacannot, Bactopia has more functionalities but cannot currently produce visual reports for results inspection. Moreover, those with limited bioinformatics experience may find the additional configuration steps more complex. Additionally, Bactopia provides extensions to its core workflow with post-processing tools intended, for example, for pangenome and phylogenetic analyses. Bacannot does not offer these as default, but the standardization of its outputs enables users to adapt them for such tasks.

Bacannot stands out for its added ability to annotate various genomic feature classes as part of its central workflow, including secondary metabolites, prophages, genomic islands, integrative and conjugative elements (ICEs), and DNA methylation, without requiring additional executions. This analytic range permits the annotation of clinically relevant traits and provides valuable attributes for non-clinical samples. As a result, bacannot has been used in various scenarios, including the analysis of clinical samples (de Campos et al., 2021), environmental samples from a lake (Janssen et al., 2021), soil (Belmok et al., 2023), and plant-associated bacteria (Bartoli et al., 2022; Ramírez-Sánchez et al., 2022).

Lastly, bacannot is the only tool that includes a comprehensive set of dynamic reports, presented as a built-in web application, along with a genome browser, offering a unified and interactive platform for interrogation and visualization of the annotation results.

Using the developed pipelines, we analyzed a strain of *Klebsiella pneumoniae* (KpBSB53) isolated from a patient at the University Hospital of Brasilia. We aimed to characterize its antibiotic resistance profile and virulence at the genome sequence level. We used sequencing data from short and long-read platforms to conduct a comprehensive genomic analysis of the strain on a laptop with 18 Gb of memory in less than a day. Our findings revealed that the strain belongs to the ST 105 group, associated with a neonatal unit outbreak in China (Zheng et al., 2016). The annotation of resistance genes only identified components considered intrinsic to the species (Bernardini et al., 2019; Holt et al., 2015). Additionally, mutations in *acrR*, *ompK36*, and *ompK37* genes were found, which may play a role in resistance. These findings are consistent with the experimentally observed susceptibility to the tested antibiotics.

Conclusions

This work showcases three bioinformatics pipelines that, together, provide a complete workflow for a thorough analysis of bacterial genomics using next-generation sequencing data. These computational protocols encompass the entire process, from the initial raw reads to the final genome assembly and gene annotation. It is recommended to execute them in succession, but they can also function independently by incorporating external data provided by the user.

The pipelines were designed to simplify the installation process by incorporating the many specialized software tools required for all stages in the form of virtualized containers. This eliminates the complexity of setup tasks, allowing the pipelines to be deployed and executed with a single command line. We also provide thorough documentation to expand the user base and make them accessible to those without bioinformatics expertise.

We not only focused on processing data but also on creating graphical reports and visualizations to improve result interpretation. To achieve this, we developed a web-based tool that allows users to analyze and refine the results using text or sequence annotation, distinguishing bacannot from other pipelines.

The existence of several comparable pipelines indicates that there is no one-size-fits-all approach to genomics. Generating DNA sequence data is becoming more widespread, but there is a significant challenge in analyzing this data. Our set of tools has been developed to address this issue and aid in the study of bacterial genomics.

Ethical considerations

The studies involving human participants were reviewed and approved by the ethical approval received from the Faculdade de Medicina, Universidade de Brasília, Brasília, DF, Brazil (approval no. CEP/FMUnB 1.131.054; CAEE: 44867915.1.0000.558). The patients provided their written informed consent to participate in this study.

Data availability

Sample sequencing data has been made in NCBI, BioProject PRJNA955456, BioSample: SAMN34178607. Additionally, the code required to reproduce the analysis performed in this paper has been made available in a Zenodo repository containing bash scripts with all required configurations.

Zenodo. Scalable and versatile container-based pipelines for de novo genome assembly and bacterial annotation (Sup. Material) https://doi.org/10.5281/zenodo.7859428.

This project contains the following underlying data:

- input/sra_ids.txt (list of SRA ids for the preprocessing pipeline)
- preprocess-params.yml (pre-set parameters file for the preprocessing pipeline)
- run_preprocess.sh (script to run preprocessing pipeline with pre-set configutations)
- assembly-params.yml (pre-set parameters file for the assembly pipeline)
- assembly.config (pre-set resources configuration file for the assembly pipeline)
- assembly_samplesheet.yml (pre-generated samplesheet for the assembly pipeline)
- run_assembly.sh (script to run assembly pipeline with pre-set configuration)
- annotation-params.yml (pre-set parameters file for the annotation pipeline)
- annotation_samplesheet.yml (pre-generated samplesheet for the annotation pipeline)
- run_annotation.sh (script to run annotation pipeline with pre-set configuration)

Software availability

The pipelines are available as GitHub repositories under the GNU General Public License v3.0. The pipeline and dependencies are easily installed by combining Nextflow and Docker or Singularity, all described in the documentation. The repositories are:

- ngs-preprocess - version controlled https://github.com/fmalmeida/ngs-preprocess
- ngs-preprocess - archived, Zenodo: https://zenodo.org/record/7831610
- MpGAP - version controlled https://github.com/fmalmeida/MpGAP
- Mpgap - archived, Zenodo: https://zenodo.org/record/7046782

- bacannot - version controlled https://github.com/fmalmeida/bacannot
- bacannot - archived, Zenodo: https://zenodo.org/record/7459261
- License: GNU General Public License v3.0 only

Acknowledgments

We thank Rodrigo de Paula Baptista for conducting the nanopore sequencing experiment.

References

Alcántar-Curiel MD, Blackburn D, Saldaña Z, *et al.*: **Multi-functional analysis of Klebsiella pneumoniae fimbrial types in adherence and biofilm formation.** *Virulence.* February 2013; **4**(2): 129–138.
Publisher Full Text

Aramaki T, Blanc-Mathieu R, Endo H, *et al.*: **KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold.** *bioRxiv.* April 2019; 602110.
Publisher Full Text

Arango-Argoty GA, Guron GKP, Garner E, *et al.*: **ARGminer: A web platform for the crowdsourcing-based curation of antibiotic resistance genes.** *Bioinformatics.* February 2020; **36**(9): 2966–2973.
Publisher Full Text

Akhter S, Aziz RK, Edwards RA: **PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies.** *Nucleic Acids Res.* May 2012; **40**(16): e126–e126.
PubMed Abstract | Publisher Full Text | Free Full Text

Arndt D, Grant JR, Marcu A, *et al.*: **PHASTER: A better, faster version of the PHAST phage search tool.** *Nucleic Acids Res.* May 2016; **44**(W1): W16–W21.
PubMed Abstract | Publisher Full Text | Free Full Text

Ausubel FM, Brent R, Kingston RE, *et al.*: *Short protocols in molecular biology.* New York: 1992; vol. **275**: 28764–28773.

Bankevich A, Nurk S, Antipov D, *et al.*: **SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing.** *J. Comput. Biol.* May 2012; **19**(5): 455–477. 10665277.
PubMed Abstract | Publisher Full Text | Free Full Text

Belmok A, de Almeida FM, Rocha RT, *et al.*: **Genomic and physiological characterization of Novosphingobium terrae sp. nov., an alphaproteobacterium isolated from Cerrado soil containing a mega-sized chromid.** *Braz. J. Microbiol.* March 2023; **54**(1): 239–258. 1517-8382, 1678-4405.
PubMed Abstract | Publisher Full Text | Free Full Text

Bartoli C, Rigal M, Mayjonade B, *et al.*: **Unraveling the genetic architecture of the adaptive potential of Arabidopsis thaliana to face the bacterial pathogen Pseudomonas syringae in the context of global change.** *Pathology.* August 2022. Preprint.

Berger B, Yu YW: **Navigating bottlenecks and trade-offs in genomic data analysis.** *Nat. Rev. Genet.* December 2022; **24**(4): 235–250.
PubMed Abstract | Publisher Full Text | Free Full Text

Bertelli C, Brinkman FSL: **Improved genomic island predictions with IslandPath-DIMOB.** In Alfonso Valencia, editor, *Bioinformatics.* July 2018; **34**: pp. 2161–2167.
PubMed Abstract | Publisher Full Text | Free Full Text

Bernardini A, Cuesta T, Tomás A, *et al.*: **The intrinsic resistome of Klebsiella pneumoniae.** *Int. J. Antimicrob. Agents.* January 2019; **53**(1): 29–33.
PubMed Abstract | Publisher Full Text

Blin K, Shaw S, Kloosterman AM, *et al.*: **antiSMASH 6.0: Improving cluster detection and comparison capabilities.** *Nucleic Acids Res.* May 2021; **49**(W1): W29–W35.
PubMed Abstract | Publisher Full Text | Free Full Text

Bortolaia V, Kaas RS, Ruppe E, *et al.*: **ResFinder 4.0 for predictions of phenotypes from genotypes.** *J. Antimicrob. Chemother.* August 2020; **75**(12): 3491–3500.
PubMed Abstract | Publisher Full Text | Free Full Text

Buels R, Yao E, Diesh CM, *et al.*: **JBrowse: A dynamic web platform for genome visualization and analysis.** *Genome Biol.* December 2016; **17**(1): 66. 1474760X.
PubMed Abstract | Publisher Full Text | Free Full Text

Camacho C, Coulouris G, Avagyan V, *et al.*: **BLAST+: Architecture and applications.** *BMC Bioinformatics.* December 2009; **10**(1).
PubMed Abstract | Publisher Full Text | Free Full Text

de Campos TA, de Almeida FM, de Almeida APC, *et al.*: **Multidrug-Resistant (MDR) Klebsiella variicola Strains Isolated in a Brazilian Hospital Belong to New Clones.** *Front. Microbiol.* April 2021; **12**: 604031. 1664-302X.
PubMed Abstract | Publisher Full Text | Free Full Text

Chang W, Joe Cheng JJ, Allaire CS, *et al.*: **Shiny: Web Application Framework for R.** 2023.

Carattoli A, Zankari E, García-Fernández A, *et al.*: **In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing.** *Antimicrob. Agents Chemother.* July 2014; **58**(7): 3895–3903.
PubMed Abstract | Publisher Full Text | Free Full Text

Chen S, Zhou Y, Chen Y, *et al.*: **Fastp: An ultra-fast all-in-one FASTQ preprocessor.** *Bioinformatics.* September 2018; **34**(17): i884–i890.
PubMed Abstract | Publisher Full Text | Free Full Text

Chen Z, Erickson DL, Meng J: **Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing.** *BMC Genomics.* December 2020; **21**(1): 631. 1471-2164.
PubMed Abstract | Publisher Full Text | Free Full Text

De Coster W, D'Hert S, Schultz DT, *et al.*: **NanoPack: Visualizing and processing long-read sequencing data.** *Bioinformatics.* March 2018; **34**(15): 2666–2669.
PubMed Abstract | Publisher Full Text | Free Full Text

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat. Biotechnol.* April 2017; **35**(4): 316–319.
PubMed Abstract | Publisher Full Text

Didelot X, Parkhill J: **A scalable analytical approach from bacterial genomes to epidemiology.** *Philos. Trans. R Soc. Lond. B Biol. Sci.* October 2022; **377**(1861): 20210246. 0962-8436, 1471-2970.
PubMed Abstract | Publisher Full Text | Free Full Text

Djaffardjy M, Marchment G, Sebe C, *et al.*: **Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems.** *Comput. Struct. Biotechnol. J.* 2023; **21**: 2075–2085.
PubMed Abstract | Publisher Full Text | Free Full Text

Edwards R, Katelyn P, Daniel S: **Linsalrob/PhiSpy: Version 3.4 prerelease.** *Zenodo.* October 2019.

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat. Biotechnol.* February 2020; **38**(3): 276–278.
Publisher Full Text

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: Summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* June 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Feldgarden M, Brover V, Haft DH, *et al.*: **Using the NCBI AMRFinder Tool to Determine Antimicrobial Resistance Genotype-Phenotype Correlations Within a Collection of NARMS Isolates.** *bioRxiv.* February 2019; page 550707.
Publisher Full Text

Graham ED, Heidelberg JF, Tully BJ: **Potential for primary productivity in a globally-distributed bacterial phototroph.** *ISME J.* March 2018; **12**(7): 1861–1866.
PubMed Abstract | Publisher Full Text | Free Full Text

Grüning B, Chilton J, Köster J, *et al.*: **Practical computational reproducibility in the life sciences.** *Cell Systems.* June 2018; **6**(6): 631–635.
PubMed Abstract | Publisher Full Text | Free Full Text

Gurevich A, Saveliev V, Vyahhi N, *et al.*: **QUAST: Quality assessment tool for genome assemblies.** *Bioinformatics.* April 2013; **29**(8): 1072–1075. 13674803.
PubMed Abstract | Publisher Full Text | Free Full Text

Haghshenas E, Asghari H, Stoye J, *et al.*: **HASLR: Fast Hybrid Assembly of Long Reads.** *iScience.* August 2020; **23**(8): 101389.
PubMed Abstract | Publisher Full Text | Free Full Text

Holt KE, Wertheim H, Zadoks RN, *et al.*: **Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health.** *Proc. Natl. Acad. Sci.* July 2015; **112**(27): E3574–E3581. 0027-8424, 1091-6490.
PubMed Abstract | Publisher Full Text | Free Full Text

Janssen L, de Almeida FM, Damasceno TAS, *et al.*: **A Novel Multidrug Resistant, Non-Tn4401 Genetic Element-Bearing, Strain of Klebsiella pneumoniae Isolated From an Urban Lake With Drinking and Recreational Water Reuse.** *Front. Microbiol.* November 2021; **12**: 732324. 1664-302X.
Publisher Full Text

Jia B, Raphenya AR, Alcock B, *et al.*: **CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database.** *Nucleic Acids Res.* January 2017; **45**(D1): D566–D573. 13624962.
PubMed Abstract | Publisher Full Text | Free Full Text

Jolley KA, Maiden MC: **BIGSdb: Scalable analysis of bacterial genome variation at the population level.** *BMC Bioinformatics.* 2010 December; **11**(1): 595. 1471-2105.
PubMed Abstract | Publisher Full Text | Free Full Text

Koren S, Walenz BP, Berlin K, *et al.*: **Canu: Scalable and accurate long-read assembly via adaptive *κ*-mer weighting and repeat separation.** *Genome Res.* March 2017; **27**(5): 722–736. 15495469.
PubMed Abstract | Publisher Full Text | Free Full Text

Kolmogorov M, Yuan J, Yu L, Pevzner PA: **Assembly of long, error-prone reads using repeat graphs.** *Nat. Biotechnol.* 2019; **37**(5): 540–546. 15461696.
Publisher Full Text

Lahlaoui H, De Luca F, Maradel S, *et al.*: **Occurrence of conjugative IncF-type plasmids harboring the blaCTX-M-15 gene in Enterobacteriaceae isolates from newborns in Tunisia.** *Pediatr. Res.* January 2015; **77**(1): 107–110. 0031-3998, 1530-0447.
PubMed Abstract | Publisher Full Text

Leger A, Leonardi T: **pycoQC, interactive quality control for Oxford Nanopore Sequencing.** *J. Open Source Softw.* 2019; **4**(34): 1236.
Publisher Full Text

Li D, Liu C-M, Luo R, *et al.*: **MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.** *Bioinformatics.* January 2015; **31**(10): 1674–1676.
PubMed Abstract | Publisher Full Text

Li W, O'Neill KR, Haft DH, *et al.*: **RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation.** *Nucleic Acids Res.* December 2020; **49**(D1): D1020–D1028.
PubMed Abstract | Publisher Full Text | Free Full Text

Liu B, Zheng D, Jin Q, *et al.*: **VFDB 2019: A comparative pathogenomic platform with an interactive web interface.** *Nucleic Acids Res.* November 2018; **47**(D1): D687–D692.
PubMed Abstract | Publisher Full Text | Free Full Text

Loman NJ, Quick J, Simpson JT: **A complete bacterial genome assembled de novo using only nanopore sequencing data.** *Nat. Methods.* June 2015; **12**(8): 733–735.
PubMed Abstract | Publisher Full Text

Mölder F, Jablonski KP, Letcher B, *et al.*: **Sustainable data analysis with snakemake.** *F1000Res.* April 2021; **10**: 33.
Publisher Full Text

Murigneux V, Roberts LW, Forde BM, *et al.*: **MicroPIPE: Validating an end-to-end workflow for high-quality complete bacterial genome construction.** *BMC Genomics.* June 2021; **22**(1): 474.
PubMed Abstract | Publisher Full Text | Free Full Text

Olawoye IB, Frost SDW, Happi CT: **The Bacteria Genome Pipeline (BAGEP): An automated, scalable workflow for bacteria genomes with Snakemake.** *PeerJ.* 2020; **8**: e10121. 2167-8359.
PubMed Abstract | Publisher Full Text | Free Full Text

Paczosa MK, Mecsas J: **Klebsiella pneumoniae: Going on the Offense with a Strong Defense.** *Microbiol. Mol. Biol. Rev.* September 2016; **80**(3): 629–661.
PubMed Abstract | Publisher Full Text | Free Full Text

Petit RA, Read TD: **Bactopia: A Flexible Pipeline for Complete Analysis of Bacterial Genomes.** *mSystems.* 2020; **5**(4).
PubMed Abstract | Publisher Full Text | Free Full Text

Priyam A, Woodcroft BJ, Rai V, *et al.*: **Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases.** *Mol. Biol. Evol.* August 2019; **36**(12): 2922–2924.
PubMed Abstract | Publisher Full Text | Free Full Text

Puterová J, Martínek T. digIS: **Towards detecting distant and putative novel insertion sequence elements in prokaryotic genomes.** *BMC Bioinformatics.* December 2021; **22**(1): 258. 1471-2105.
PubMed Abstract | Publisher Full Text | Free Full Text

Quijada NM, Rodríguez-Lázaro D, Eiros JM, *et al.*: **TORMES: An automated pipeline for whole bacterial genome analysis.** *Bioinformatics.* April 2019; **35**(21): 4207–4212. 1367-4803.
PubMed Abstract | Publisher Full Text

Ramírez-Sánchez D, Gibelin-Viala C, Mayjonade B, *et al.*: **Investigating genetic diversity within the most abundant and prevalent non-pathogenic leaf-associated bacteria interacting with Arabidopsis thaliana in natural habitats.** *Front. Microbiol.* September 2022; **13**: 984832. 1664-302X.
PubMed Abstract | Publisher Full Text | Free Full Text

Ruan J, Li H: **Fast and accurate long-read assembly with wtdbg2.** *Nat. Methods.* December 2019; **17**(2): 155–158.
PubMed Abstract | Publisher Full Text | Free Full Text

Ruiz-Perez CA, Conrad RE, Konstantinidis KT: **MicrobeAnnotator: A user-friendly, comprehensive functional annotation pipeline for microbial genomes.** *BMC Bioinformatics.* January 2021; **22**(1): 11.
PubMed Abstract | Publisher Full Text | Free Full Text

Sayers S, Li L, Ong E, *et al.*: **Victors: A web-based knowledge base of virulence factors in human and animal pathogens.** *Nucleic Acids Res.* October 2018; **47**(D1): D693–D700.
PubMed Abstract | Publisher Full Text | Free Full Text

Schwengers O, Hoek A, Fritzenwanker M, *et al.*: **ASA3P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates.** *PLoS Comput. Biol.* March 2020b; **16**(3): e1007134–e1007115.
PubMed Abstract | Publisher Full Text | Free Full Text

Schwengers O, Jelonek L, Dieckmann MA, *et al.*: **Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification.** *Microb. Genom.* November 2021; **7**(11). 2057-5858.
Publisher Full Text

Schwengers O, Barth P, Falgenhauer L, *et al.*: **Platon: Identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores.** *Microb. Genom.* October 2020a; **6**(10).
PubMed Abstract | Publisher Full Text | Free Full Text

Seemann T: **Prokka: Rapid prokaryotic genome annotation.** *Bioinformatics.* July 2014; **30**(14): 2068–2069. 14602059.
PubMed Abstract | Publisher Full Text

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* June 2015; **31**(19): 3210–3212.
Publisher Full Text

Shafin K, Pesout T, Lorig-Roach R, *et al.*: **Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes.** *Nat. Biotechnol.* May 2020; **38**(9): 1044–1053.
PubMed Abstract | Publisher Full Text | Free Full Text

Sserwadda I, Mboowa G: **rMAP: The Rapid Microbial Analysis Pipeline for ESKAPE bacterial group whole-genome sequence data.** *Microbial Genomics.* June 2021; **7**(6). 2057-5858.
Publisher Full Text

Starikova EV, Tikhonova PO, Prianichnikov NA, *et al.*: **Phigaro: High throughput prophage sequence annotation.** *bioRxiv.* April 2019; page 598243.
Publisher Full Text

Tian D, Wang M, Zhou Y, *et al.*: **Genetic diversity and evolution of the virulence plasmids encoding aerobactin and salmochelin in *Klebsiella pneumoniae.*** *Virulence.* December 2021; **12**(1): 1323–1333. 2150-5594, 2150-5608.
PubMed Abstract | Publisher Full Text | Free Full Text

Vaser R, Šikić M: **Time- and memory-efficient genome assembly with Raven.** *Nat. Comput. Sci.* May 2021; **1**(5): 332–336.
Publisher Full Text

Walker BJ, Abeel T, Shea T, *et al.*: **Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement.** *PLoS One.* November 2014; **9**(11): e112963. 19326203.
PubMed Abstract | Publisher Full Text | Free Full Text

Wick RR, Judd LM, Gorrie CL, *et al.*: **Completing bacterial genome assemblies with multiplex MinION sequencing.** *Microbial. Genomics.* October 2017a; **3**(10).
PubMed Abstract | Publisher Full Text | Free Full Text

Wick RR, Judd LM, Gorrie CL, Holt KE: **Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads.** *PLoS Comput. Biol.* June 2017b; **13**(6): e1005595. 15537358.
PubMed Abstract | Publisher Full Text | Free Full Text

Wratten L, Wilm A, Göke J: **Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers.** *Nat. Methods.* September 2021; **18**(10): 1161–1168.
PubMed Abstract | **Publisher Full Text**

Xie Y, Dervieux C, Riederer E: *R Markdown Cookbook.* Boca Raton, Florida: Chapman and Hall/CRC; 2020.

Xuan J, Ying Y, Qing T, *et al.*: **Next-generation sequencing in the clinic: Promises and challenges.** *Cancer Lett.* November 2013; **340**(2): 284–295.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Zheng R, Zhang Q, Guo Y, *et al.*: **Outbreak of plasmid-mediated NDM-1-producing Klebsiella pneumoniae ST105 among neonatal patients in Yunnan, China.** *Ann. Clin. Microbiol. Antimicrob.* February 2016; **15**(1): 10.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ? ✓ ✓

Version 1

Reviewer Report 14 November 2023

https://doi.org/10.5256/f1000research.152764.r219643

✓ **Austin G Davis-Richardson**
One Codex, San Francisco, California, USA

The authors have developed a suite of open source workflows implemented in Nextflow for microbial WGS preprocessing, de novo assembly, and annotation. The goal of this work is to provide re-usable workflows that can be tailored to the end-user's needs, therefore several interchangeable tools (e.g., de novo assemblers) are wrapped in a common Nextflow framework. Interoperability between bioinformatics tools is a significant challenge therefore these workflows provide a useful resource for tailoring a set of tools to meet one's needs based on sequencing technology.

The workflows are then applied to a real world problem of analyzing genomic sequences from a Klebsiella isolate with recapitulation of antibiotics resistance phenotypes from genomic data as the "ground truth" for validating the workflow. While this is a limited dataset, the purpose of the paper is to not validate all of the bioinformatics tools used in the workflow but to demonstrate how it could be used to generate a useful output from genomic sequencing data.

The paper would benefit from either demonstrating the first two pipelines (excluding annotation) on a non-bacterial dataset as the authors claim that only the annotation pipeline is prokaryote-specific.

I was able to download and start the Nextflow pipelines however I do not have access to sufficient computational resources required to complete the pipelines. Therefore, I am unable to evaluate the outputs of the pipelines at this time. It would be useful for the authors to provide example outputs on Zenodo.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Metagenomics, Microbial Genomics, Microbiome

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 18 Nov 2023

Felipe Almeida

Dear Reviewer,

Thanks for the thoughtful words addressed when reviewing our paper.

We are on full agreement with your suggestion and indeed the paper and the pipelines ngs-preprocess and MpGAP would enormously benefit from us displaying its usage in a non-bacterial dataset. We would like to reply saying that we are going to work on that, together with the comments of Reviewer 2.

First, we are going to generate a new page in the web documentation for this 2 pipelines, showing the analysis of a fungi or plant sequencing dataset. We will make sure that they have the necessary command lines from input to output, so one can reproduce, but also, add an overview of the generated results in the web page.

Finally, we are first going to work as making all these changes and enhancements of providing explanation and example of outputs in the web page itself. Once done, we are going to check how easily can we update the paper to provide an additional Zenodo for the non-bacterial analysis, and also to update the bacterial-analysis Zenodo to have the outputs.

We have already created issues in the pipelines' repositories so we do not forget to address them.

Sincerely,
Felipe.

Competing Interests: None.

Reviewer Report 14 November 2023

https://doi.org/10.5256/f1000research.152764.r218196

Abhinav Sharma
Department of Biomedical Sciences, Stellenbosch University, Stellenbosch, Western Cape, South Africa

Emilyn Costa Conceição
Department of Biomedical Sciences, Stellenbosch University, Stellenbosch, Western Cape, South Africa

Summary:
The authors present 3 modular piplines aimed at (non-)bioinformaticians and applicable in different stages of bacterial analysis such as data download, assembly and annotation and generates reports to facilitate the interpretation of the results. The pipelines also support the analysis of data from multiple sequencing platforms, aiding the reuse of the pipeline for both short and long read sequences, which adds value to the bioinformatical toolkit given the increasing use of long-read sequences and the current prevalence of short-read sequences.

Furthermore, these pipelines have been written in Nextflow and generally follow the best practices of modularization outlined by the nf-core community pipeline template.

Suggestion (ref documentation website):
As far as the design of the pipeline and description of the design choices goes, it is well explained. The interpretation of the generated results has been neatly described for bacannot in the documentation page (https://bacannot.readthedocs.io/en/latest/outputs/), including the directory structure and the relevant links for the tool-specific reference material. However as the pipelines are modular and the user is expected to run them separately, it is recommended that the documentaiton websites for MpGAP pipeline and ngs-preprocess pipelines should also have an "Output" page to facilitate users on the output structure and refer the correct tools-specific links.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Mycobacterium tuberculosis, Reproducible methods, Machine learning, Variant calling, Reference Alignment, Resistance profiling and Scientific workflow managers.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 18 Nov 2023

Felipe Almeida

Dear Reviewer,
Thank you for the words addressed when reviewing the paper.

I would like to reply saying that we will indeed work on the suggestion you made. It is indeed a very important note you have brought to our attention that, even though each pipeline is independent, we forgot to add a proper specification and explanation of all Outputs in the ngs-preprocess and MpGAP pipelines.

We have already created issues on the repositories so we do not forget to address this suggestion.

Sincerely,
Felipe.

Competing Interests: None.

Reviewer Report 18 October 2023

https://doi.org/10.5256/f1000research.152764.r210163

?

Abdolrahman Khezri
Innland Norway university of applied sciences, Hamar, Norway

The paper entitled "Scalable and versatile container-based pipelines for de novo genome assembly and bacterial annotation", takes advantage of several well-known tools for NGS data processing, assembly, and downstream analyses. The GitHub page seems to provide enough information too. However, the paper still needs to improve in some aspects.

The aim of this paper is not clear. The authors developed a pipeline based on existing tools. However, the aim is not easy to find, and the work lacks novelty.

The other drawback of this paper is the fact that the authors tested the pipeline only using a clinical isolate. How this pipeline handles, for instance, gram-positive bacteria or other bacteria with different genome sizes is not clear.

The figures are not accurate. They do not provide enough/precise information regarding the steps, tasks, or tools.

It is necessary to implement the Bandage tool into the pipeline right after assembly to visualize the assembly.

Is there any specific reason the author chose fastp as a quality check tool? Obviously, fastqc provides much more details compared to fastp.

I understand that different tools make different assembly, however many publications already benchmarked them, for instance, it is well known that SPAdes, Flye, and Unicycler provides a better assembly for short, long, and hybrid assembly, respectively. Please see Wick *et al.* (2019[1]) or Khezri *et al.* (2021[2]). However, in my opinion in a pipeline, an author should stick to the best overall tool and avoid providing many tools as pipelines are designed for less experienced users.

In Table 4, some of the tasks such as annotation with Prokka as well as Quast and Busco took a tremendous amount of time which is not how it should be. What is the author's comment on this?

Although the authors presented the differences between their pipeline and other published pipelines, the samples in this study should be tested using other pipelines in order to have a clear demonstration of performance.

Porechop is out of date and no longer supported. This is important when it comes to new chemistry and sequencing kits, where, for instance, 24 barcodes in rapid barcoding kits are not recognized by Porechop. Here it would be better to take advantage of ONT guppy software not only for trimming but also for other tasks.

Very little is written regarding BLAST. How can the user benefit from BLAST in this pipeline? Does it

come with ore-indexed reference databases?

The Prokka output can be used for many different purposes including core genome and pan genome analyses which are important for people working on epidemiology. It would be nice if the author could extend their pipeline to cover this.

References

1. Wick RR, Holt KE: Benchmarking of long-read assemblers for prokaryote whole genome sequencing.*F1000Res*. 2019; **8**: 2138 PubMed Abstract | Publisher Full Text
2. Khezri A, Avershina E, Ahmad R: Hybrid Assembly Provides Improved Resolution of Plasmids, Antimicrobial Resistance Genes, and Virulence Factors in Escherichia coli and Klebsiella pneumoniae Clinical Isolates.*Microorganisms*. 2021; **9** (12). PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics and bacterial genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.