Universidade de Brasília
Instituto de Biologia
Departamento de Biologia Celular

Felipe Marques de Almeida

Desenvolvimento de *pipelines* de genômica bacteriana e sua aplicação em isolados do Hospital Universitário de Brasília

Brasília
Fevereiro, 2020

Felipe Marques de Almeida

Desenvolvimento de *pipelines* de genômica bacteriana e sua aplicação em isolados do Hospital Universitário de Brasília

Dissertação apresentada ao Departamento de Biologia Celular do Instituto de Biologia da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre em Biologia Molecular com ênfase em Biologia Computacional.

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

Brasília

Fevereiro, 2020

Dedico este trabalho a meus familiares, amigos e amada que estiveram sempre presentes e me apoiaram durante toda esta jornada.

Agradecimentos

A trajetória percorrida para a conclusão deste mestrado foi repleta de desafios, incertezas, tristezas, mas também muitos momentos de alegria. A passagem por este caminho só foi possível devido à participação e a contribuição de muitas pessoas que foram indispensáveis para que eu fosse capaz de encontrar a melhor direção em diversos momentos deste caminho. A estas pessoas, dedico este trabalho.

Especialmente ao meu orientador, Professor Doutor Georgios Pappas que, durante toda esta jornada, sempre acreditou em mim e, através de seu rigor científico, orientação exemplar, interesse permanente e exigência saudável contribuiu para o enriquecimento e desenvolvimento deste trabalho em todas as suas etapas e, com grande paciência e dedicação, foi importantíssimo para me moldar como pesquisador científico.

Agradeço igualmente ao meu colega de laboratório Rodrigo Theodoro pelo apoio e pelas discussões realizadas que ajudaram a moldar vários pontos deste trabalho.

Agradeço enormemente à Kissia Batista pelo carinho, apoio e companheirismo que foram fundamentais para que eu me mantivesse focado na realização deste trabalho e não me deixou desanimar quando as coisas não aconteciam como o desejado.

Por último, quero agradecer à minha família e amigos pelo apoio incondicional e que, com alguns esforços, se fazem presentes durante a Defesa deste trabalho. Agradeço, especialmente à minha mãe Waleska Oliveira e pai Edson Almeida pelos incansáveis esforços em disponibilizar os meios e as condições necessárias para que meu irmão e eu possamos nos dedicar aos nossos estudos.

"*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

Avanços nas tecnologias de sequenciamento de DNA têm revolucionado estudos de genômica bacteriana por permitir montar genomas completos, a nível de cromossomo, de maneira rápida e barata, tornando viável pesquisas de genômica populacional. Atualmente, a análise computacional dos dados de sequenciamento é o principal obstáculo que impede a sua utilização em ambientes clínicos. Para solucionar este problema, desenvolvemos três *pipelines* baseados em contêineres computacionais capazes de montar e anotar genomas a partir de dados de sequenciamento de múltiplas plataformas, permitindo a identificação de genes de resistência, fatores de virulência, prófagos e elementos integrativos. No geral, um genoma pode ser montado e anotado em um *laptop*, em menos de um dia. Cada *pipeline* é modular, e leva em conta diferentes cenários analíticos prontamente configurados pelo usuário. A utilização de contêineres permite que estes *pipelines* sejam executados em qualquer sistema operacional e não necessitem da instalação manual de seus componentes. A aplicação destes *pipelines* em isolados bacterianos do Hospital Universitário de Brasília possibilitou a caracterização de uma cepa hipermucoviscosa de *Klebsiella pneumoniae* ST11-K64 multirresistente a antibióticos contendo diversos fatores de virulência. Os resultados deste trabalho reeditam alertas quanto ao surgimento de patógenos de alto risco devido à convergência de genes de resistência e virulência e reforçam a necessidade da criação de programas de vigilância contínua de patógenos. Demonstra-se ainda o grande potencial da genômica como uma valiosa aliada na batalha contra infecções bacterianas. Em conclusão, estes *pipelines* oferecem uma ferramenta computacional simples e direta, capaz de facilitar a inclusão de rotinas de sequenciamento genômico de bactérias em sistemas de saúde.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ANI *Average Nucleotide Identity*

BAM *Binary Alignment Map*

cgMLST *core gene Multilocus sequence typing*

cKp *Klebsiella pneumoniae* clássica

CWL *Common Workflow Language*

DSL *Domain-specific language*

ESBL *Extended Spectrum Beta-Lactamase*

GFF *General File Format*

HTML *Hypertext Markup Language*

HUB Hospital universitário da UnB

hvKp *Klebsiella pneumoniae hipervirulenta*

KNIME *Konstanz Information Miner*

KO Ortologias KEGG

Kp *Klebsiella pneumoniae*

LPS Lipopolissacarídeo

MDR *Multidrug resistant*

MGE *Mobile genetic element*

MIC *Minimum inhibitory concentration*

MLST *Multilocus sequence typing*

N50 Tamanho da menor sequência de contig usada para se obter 50% do genoma

NGS *Next Generation Sequencing*

OMS Organização Mundial da Saúde

ONT *Oxford Nanopore Technologies*

PacBio *Pacific Biosciences*

PCR *Polymerase Chain Reaction*

PFGE *Pulsed field gel electrophoresis*

RFLP *PCR restriction fragment length polymorphism*

RND *Resistance-nodulation-division*

ST *Sequence type*

STEC *Shiga-toxin producer Escherichia coli*

T4SS Sistema de secreção tipo IV

T6SS Sistema de secreção tipo VI

VFDB *Virulence Factor Database*

Sumário

1 Introdução

1.1 Histórico da resistência a agentes antimicrobianos

Infecções bacterianas estiveram sempre presentes ao longo da História da humanidade. Durante milhares de anos inúmeras pandemias assolaram humanos, como a sífilis, hanseníase, cólera e tuberculose. Muitas vezes, estas pandemias foram tão significativas ao ponto de moldar a História, causando o declínio de cidades e nações. Um dos casos mais emblemáticos é o da peste bubônica, que assolou o mundo na idade média e resultou na morte de dezenas de milhões de pessoas em toda a Europa (Mohr, 2016).

A "era dos antibióticos" se iniciou em 1928 com o primeiro antibiótico natural descoberto por Alexander Fleming (Fleming, 1929). Entretanto, em 1945, começaram a surgir relatos de cepas bacterianas resistentes a antibióticos, como nos casos de *Staphylococcus aureus* resistentes a penicilina (Plough, 1945) e estreptomicina (Waksman et al., 1945).

Percebeu-se, já na década de 1990, que a evolução dos mecanismos de defesa empregados pelas bactérias estavam superando o ritmo de descoberta de novas moléculas para combate a estes patógenos (Lewis, 2013). Isto, resultou em um clamor dos cientistas para uma aliança global para atacar este problema (Neu, 1992; Gold e Moellering, 1996). Atualmente, os mesmos alertas vêm sendo reeditados em face a emergência de novas formas de resistência que estão literalmente esgotando as alternativas terapêuticas (Dodds, 2017; Baker et al., 2018). Este panorama levou a Organização Mundial de Saúde (OMS) a considerar a resistência a antibióticos como uma das maiores ameaças à saúde, segurança

alimentar e desenvolvimento [1].

1.2 Os mecanismos de ação de antibióticos

Antibióticos são agentes quimioterapêuticos que tem ação bactericida ou bacteriostática. Tem origem natural, produzidos por fungos (ex: *Penicillium* sp) ou bactérias (ex: *Streptomyces* sp), ou através de síntese em laboratórios. Quimicamente formam um grupo diverso de moléculas de tamanho pequeno, que podem ser divididos em categorias estruturais, como os beta-lactâmicos (penicilinas, carbapenêmicos) ou aminoglicosídeos (estreptomicina), dentre outros (Kapoor et al., 2017).

De acordo com o mecanismo de ação os antibióticos são classificados em três grandes grupos de atuação: (i) inibição da síntese de parede celular; (ii) inibição da síntese de ácidos nucleicos; e (iii) inibição da síntese de proteínas. A seguir, segue uma breve descrição das famílias de antibióticos no contexto destes mecanismos de ação.

1.2.1 Inibição da síntese de parede celular

A principal classe de antibióticos que age através da inibição da síntese de parede celular são os beta-lactâmicos, como as penicilinas, cefalosporinas, carbapenêmicos e monobactâmicos. Beta-lactâmicos se ligam irreversivelmente às proteínas de ligação a penicilina (PBPs, "penicilin binding proteins") e interrompem a última etapa de transpeptidação da camada de peptídeoglicano, perturbando a síntese de parede celular (Kapoor et al., 2017).

Glicopeptídeos, como a vancomicina, também inibem a síntese da parede celular ao se ligar na porção D-ala-D-ala da cadeia peptídica crescente, impedindo a transpeptidação e interrompendo a elongação e ligação cruzada da matriz peptídoglicana (Blair et al., 2015).

1.2.2 Inibidores da síntese de ácidos nucleicos

Quinolonas, sulfonamidas, trimetropina e rifampicina são exemplos de drogas inibidoras da síntese de ácidos nucleicos. Eficientes contra Gram-negativos e Gram-positivos,

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

interferem no crescimento, replicação e sobrevivência bacteriana.

As quinolonas inibem a DNA girase (topoisomerase) bacteriana. Isto, impede que o superenrolamento seja relaxado e, desta forma, impede a replicação ou transcrição do DNA. A rifampicina, por sua vez, age inibindo a transcrição bacteriana ao se ligar à RNA polimerase procariótica (Kapoor et al., 2017).

Já a sulfonamida e a trimetropina, inibem dois pontos diferentes do metabolismo de ácido fólico, necessário para a síntese de ácidos nucléicos. A sulfonamida inibe a enzima dihidropteroase sintetase por apresentar maior afinidade que o substrato natural e, a trimetropina, age inibindo a enzima dihidrofolato redutase (Kapoor et al., 2017).

1.2.3 Inibidores da síntese de proteínas

Enquadram-se como inibidores da síntese de proteínas as tetraciclinas, macrolídeos, cloranfenicóis e aminoglicosídeos. Estes antibióticos possuem um espectro de ação considerável e são efetivos contra várias bactérias Gram-negativas e Gram-positivas.

As tetraciclinas e os cloranfenicóis, são agentes que previnem a ligação do aminoacil-tRNA ao sítio A do ribossomo bacteriano, pela ligação ao rRNA 16S e 23S respectivamente. Por perturbarem a cavidade peptidil transferase da região 23S, macrolídeos interferem diretamente na etapa de translocação da tradução por causar um desacoplamento prematuro da cadeia peptídica nascente (Kapoor et al., 2017).

Os aminoglicosídeos como a gentamicina e tobramicina, por sua vez, têm pouca atividade contra bactérias anaeróbias. Estes antibióticos geralmente interferem na etapa de elongação na subunidade ribossômica 30S, causando erros e a terminação prematura da tradução do mRNA (Kapoor et al., 2017).

1.3 Bases moleculares da resistência antimicrobiana

O uso de antibióticos em concentrações subletais possibilita a seleção de clones capazes de contrapor as condições desfavoráveis a sua sobrevivência (Blair et al., 2015). Seu uso corriqueiro e indiscriminado na saúde humana, veterinária e produção animal tem como consequência impor uma pressões de seleção regular que, o que acelera a taxa de fixação de mutações ou aquisição exógena de genes com o consequente estabelecimento de

novos mecanismos de resistência antimicrobiana em bactérias (Dodds, 2017; Durão et al., 2018).

Ainda que possam impactar a fisiologia da bactéria, mutações capazes de interferir na eficiência de antibióticos surgem estocasticamente pelo processo de seleção natural (Bell e MacLean, 2018). Mutações adaptativas são geradas constantemente, em taxas de até 10^{-5} por célula por geração, contrastando com eventos de transferência lateral de genes, os quais dependem de contextos ecológicos definidos, como, por exemplo, a presença de bactérias doadoras (Hughes e Andersson, 2017; Durão et al., 2018).

Devido a estruturas inerentes e adaptações funcionais, algumas bactérias são intrinsecamente resistentes a alguns antimicrobianos. Por exemplo, pela incapacidade de ultrapassar a membrana externa de bactérias Gram-negativas, a vancomicina é, normalmente, efetiva somente contra bactérias Gram-positivas (Peterson e Kaur, 2018).

Em termos moleculares, existem diversas estratégias para uma bactéria contrapor a ação de antibióticos, envolvendo os seguintes mecanismos gerais (Blair et al., 2015; Santajit e Indrawattana, 2016; Iovleva e Doi, 2017):

- Prevenção de acesso ao alvo pela minimização de concentrações intracelulares de drogas através de mecanismos de efluxo ou permeabilidade reduzida
- Modificação, ou proteção, do alvo da droga no hospedeiro por meio de modificações pós-traducionais ou mutações genéticas
- A inativação direta da droga por modificação química ou hidrólise

Estas características podem surgir através da mutação de genes no cromossomo bacteriano ou pela acquisição através da transferência lateral de genes, mediada por elementos genéticos móveis, como plasmídeos, transposons e integrons (Woodford e Ellington, 2007; Gillings, 2017). Na figura 1.1 encontra-se uma esquematização das principais drogas utilizadas atualmente, seus mecanismos de ação e principais mecanismos de resistência.

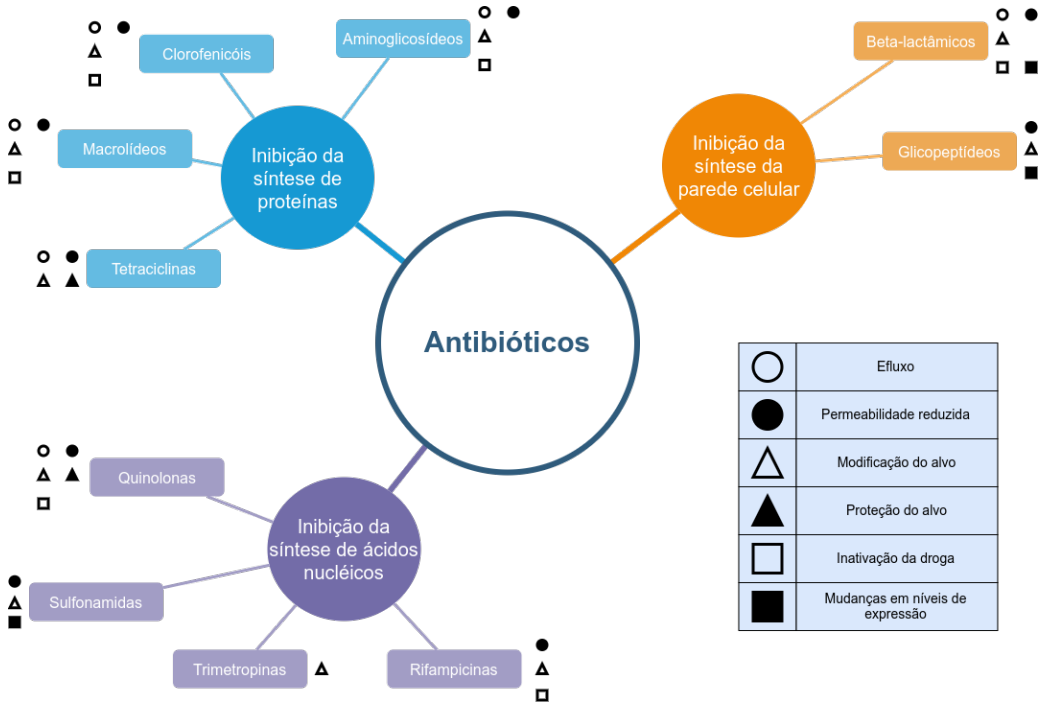


Figura 1.1: Representação esquemática das principais classes de antibióticos da atualidade e seus respectivos mecanismos de ação e resistência antimicrobiana

A seguir, são apresentados os detalhes moleculares, genéticos e fisiológicos de cada uma destas categorias.

1.3.1 Minimização de concentrações intracelulares

A permeabilidade das membranas bacterianas influencia diretamente a concentração intracelular efetiva de antibióticos, através do balanço das taxas de influxo e efluxo dos mesmos.

Na membrana externa de bactérias Gram-negativas existem proteínas, particularmente da família das porinas, que controlam o influxo de pequenas moléculas importantes para a fisiologia celular mas, eventualmente, agentes nocivos como antibióticos. Diversos relatos descrevem que a redução na expressão ou mutações de porinas são correlacionadas com a emergência de fenótipos resistentes a carbapenemas (Hao et al., 2018; Kong et al., 2018).

Bombas de efluxo de antibióticos são complexos proteicos importantíssimos para a resistência intrínseca de bactérias Gram-negativas a algumas drogas. Quando superex-

pressas, podem conferir altos níveis de resistência a antibióticos (Du et al., 2018). Este é um mecanismo de resistência comum em isolados clínicos e, bactérias com este fenótipo vêm sendo isoladas de pacientes desde a década de 1990 (Everett et al., 1996; Pumbwe e Piddock, 2000; Kosmidis et al., 2012).

Altos níveis de expressão dessas bombas são geralmente adquiridos através da mutação de componentes da rede de regulação de sua expressão. Alternativamente, a expressão de bombas de efluxo pode ser acionada por estímulos ambientais, por exemplo, a partir da ligação de uma molécula ao repressor transcricional do gene e consequente diminuição de sua repressão (Baucheron et al., 2014; Du et al., 2018). Ademais, algumas destas são capazes de transportar uma grande variedade de substratos estruturalmente similares e são conhecidos como bombas de efluxo de resistência múltipla (MDR). Destas, a família de bombas RND (do inglês, "resistance-nodulation-division") é a melhor caracterizada.

1.3.2 Modificação do alvo do antibiótico

Outra estratégia evasiva verificada em bactérias consiste em mutações que induzem modificações estruturais na proteína que é o alvo para o antibiótico, interferindo na eficiência da ligação deste, sem alterar drasticamente a função da proteína (Blair et al., 2015). Como grande parte destes alvos são codificados por genes presentes em múltiplas cópias nas bactérias, a mutação em uma dessas cópias pode ser suficiente para a produção de variações fenotípicas de tolerância ao antibiótico (Peterson e Kaur, 2018).

Além de mutações e modificações estruturais, a proteína alvo pode também ser protegida através de metilações pontuais que evitam a ligação de múltiplos antibióticos, mas mantém a função proteica, como é o caso da metilação mediada pela Cfr rRNA metiltransferase que confere resistência a cloranfenicol e lincosamidas (Long et al., 2006).

1.3.3 Inativação dos antibióticos

A resistência antimicrobiana também pode ser adquirida através da capacidade de destruir ou inativar antibióticos. Uma das principais formas de inativação de antibióticos é através da hidrólise. Por exemplo, antibióticos beta-lactâmicos como penicilinas, ce-

falosporinas, monobactâmicos e carbapenêmicos são hidrolisados por diversas classes de enzimas nomeadas beta-lactamases (Santajit e Indrawattana, 2016).

Alternativamente, antibióticos podem ser inativados através da transferência de grupamentos químicos que podem estericamente impedir a ligação do antibiótico ao seu alvo. Por exemplo, devido ao seu tamanho, aminoglicosídeos são bastante susceptíveis a modificações químicas, principalmente por ação de acetil-transferases, fosfo-transferases e nucleotidil-transferases (Romanowska et al., 2013).

1.4 Bactérias Gram-negativas resistentes a carbapenêmicos

Os mecanismos de aquisição de resistência podem ocorrer de maneira autônoma e cumulativa. Consequentemente, ao longo do tempo, diversos genes podem ser encontrados em um mesmo organismo, principalmente em plasmídeos, resultando em fenótipos de resistência a múltiplas drogas (MDR, do inglês "Multidrug Resistance").

O fenótipo MDR é especialmente relevante para um grupo de bactérias denominado "ESKAPE" (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumanii*, *Pseudomonas aeruginosa* e *Enterobacter* spp.). Este grupo é considerado pela OMS como uma das maiores ameaças à saúde mundial em função de estarem intimamente associados a ambiente hospitalares e apresentarem novos paradigmas na aquisição e transmissão de mecanismos de resistência (Pendleton et al., 2013; Theuretzbacher, 2017). Além disto, estes organismos estão relacionados ao maior risco de mortalidade e ao aumento de custos de saúde em países em desenvolvimento, como o Brasil (Founou et al., 2017).

Estas bactérias são extremamente plásticas quanto aos mecanismos de resistência a antibióticos e, por isso, o desenvolvimento de novas drogas para este grupo tem falhado (Brown e Wright, 2016). Uma das últimas alternativas terapêuticas são as carbapenemas, eficazes contra infecções severas de bactérias Gram-negativas. No entanto, mais e mais cepas resistentes a essa classe de antibióticos vêm sendo reportadas (Iovleva e Doi, 2017; Theuretzbacher, 2017; van Duin e Doi, 2017).

1.5 A bactéria *Klebsiella pneumoniae*

Klebsiella pneumoniae são bactérias encontradas em abundância no ambiente, em solos, esgotos, água e plantas. Estas bactérias são capazes de colonizar uma grande diversidade de hospedeiros animais, incluindo humanos, podendo ser encontrados em diferentes partes do corpo como o trato respiratório, urinário e pele (Wyres e Holt, 2016). Em média, o tamanho do genoma de *K. pneumoniae* é de 5,5 Mb com aproximadamente 5.500 genes. Destes genes, somente 2.000 são presentes em todos os indivíduos da espécie ("core genes") e a taxa acumulativa de genes acessórios identificados com a aumento do número de genomas sequenciados indica um pangenoma aberto, com vasto "pool" gênico (Wyres e Holt, 2016).

A bactéria *Klebsiella pneumoniae* (Kp) é uma das principais causas de infecções hospitalares no mundo, sendo o trato urinário o sítio de infecção mais comum. Uma vez adquirida, esta bactéria coloniza principalmente as mucosas, geralmente as do trato gastrointestinal.

Grande parte dos genes de resistência em *Klebsiella pneumoniae* são adquiridos horizontalmente, principalmente carreados em plasmídeos. *Klebsiella pneumoniae* são capazes de carrear múltiplos plasmídeos ao mesmo tempo, cada um com um conjunto diferente de resistência a antimicrobianos. Desta maneira, através da transferência direta de plasmídeos entre cepas de *K. pneumoniae* e outras enterobactérias, *Klebsiella pneumoniae* são capazes de rapidamente desenvolver fenótipos de multirresistência a antibióticos (Wyres e Holt, 2016). Aliado à resistência a antimicrobianos, enfrenta-se hoje, mundialmente, problemas relacionados ao aumento da virulência de *Klebsiella pneumoniae* pelo acúmulo de genes de virulência, principalmente em plasmídeos (Wyres et al., 2019). Este cenário é bastante preocupante pois a convergência de genes de virulência e resistência pode resultar no surgimento de cepas de *Klebsiella pneumoniae* MDR altamente virulentas, implicando no aumento da dificuldade de tratamento destas infecções (Wyres et al., 2019).

1.5.1 Fatores de virulência em Klebsiella pneumoniae

Em geral *K. pneumoniae* é considerada comensal e oportunista e, apesar de não se saber os mecanismos exatos pelo qual uma colonização se desenvolve em infecção, conhece-se diversos fatores de virulência de extrema importância para esta progressão. 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).

A cápsula bacteriana é, talvez, o fator de virulência mais importante para a virulência bacteriana. Sintetizada pelo locus *cps*, ela protege a bactéria da fagocitose por macrófagos (Cortés et al., 2002). Tradicionalmente, sorotipos K são diferenciados por variantes alélicas dos genes *wzi* e *wzc*, genes bastante conservados no locus (Martin e Bachman, 2018). Porém, alguns dos genes do locus *cps* não são conservados e, as combinações de presença e ausência destes genes são dependentes do sorotipo capsular, o que também permite a classificação de Kps em diferentes K-loci (KLs). Nesta classificação, os sorotipos referência K1-77 (diferenciados pelos alelos *wzi* e *wzc*) foram nomeados com seus respectivos números, KL1-77. Enquanto os novos loci baseados somente conteúdo gênico do locus são nomeados a partir do identificador KL101 (Wyres et al., 2016).

O LPS, ou endotoxina, é reconhecido como um potente mediador do choque séptico causado por bactérias. Estas moléculas são formadas por um lipídeo A conservado e um antígeno O variável e são importantes para a resistência de microrganismos ao sistema imune complemento. Variações no antígeno O são reconhecidas como sorotipos O. Das infecções causadas por *Klebsiella*s, 80% são provenientes dos sorotipos O1, O2 e O3 (Martin e Bachman, 2018; Cortés et al., 2002).

Sideróforos são moléculas de alta afinidade ao ferro, por possuírem maior afinidade ao ferro que proteínas de transporte do hospedeiro, ajudam a bactéria a captar ferro durante infecções (Paczosa e Mecsas, 2016). existem diversos sideróforos, de diferentes níveis de afinidade, que podem ser expressos por *Klebsiella pneumoniae*: Enterobactina (Ent), Salmochelina (Sal), Aerobactina (Aer) e Yersiniabactina (Ybt). A Enterobactina é um sideróforo secretado universalmente por todas as bactérias da espécie *Klebsiella pneumoniae*. Por outro lado, a Salmochelina, Aerobactina e Yersiniabactina não são universais e estão intimamente relacionados a maiores níveis de virulência (Martin e Bachman, 2018;

Paczosa e Mecsas, 2016).

Fímbrias, ou *pili*, são moléculas essenciais para a adesão em superfícies e, portanto, cruciais para a infecção bacteriana (Paczosa e Mecsas, 2016). Destacam-se, em *Klebsiella pneumoniae*, dois tipos de fímbria, 1 e 3. Fímbrias do tipo 1 (*fim*), são finas e expressas por ≈90% dos isolados de *Klebsiella pneumoniae*. Estas moléculas são essenciais para o desenvolvimento da infecção pois conferem habilidade de aderir a superfícies bióticas e abióticas (Paczosa e Mecsas, 2016). Fímbrias do tipo 3 (*mrk*) são moléculas adesivas encontradas em enterobactérias. Estas moléculas, em particular, são consideradas cruciais para a formação de biofilme em superfícies bióticas e abióticas (Martin e Bachman, 2018; Paczosa e Mecsas, 2016; Stahlhut et al., 2013).

Bombas de efluxo são complexos proteicos capazes de remover diversos compostos tóxicos do interior da célula. Estes complexos são importantes mecanismos pelo qual bactérias adquirem resistência a antibióticos. Além disso, é demonstrado que a bomba de efluxo AcrAB é (também) um importante fator de virulência, provavelmente por mediar a resistência contra peptídeos antimicrobianos do hospedeiro (Martin e Bachman, 2018; Paczosa e Mecsas, 2016).

Por último, o sistema de secreção tipo VI (T6SS) é um aparato que serve como canal injetor de moléculas efetoras e toxinas em células eucarióticas e procarióticas. Proporcionam vantagem adaptativa por permitir a injeção de células efetoras capazes de hidrolisar a parede celular, membrana ou ácidos nucleicos de outras células. Este sistema é encontrado em diversas bactérias patogênicas e é importante para a competição e virulência bacteriana. Além disso, é um sistema que reage a estímulos de estresse e induzido positivamente na presença de antibióticos (Martin e Bachman, 2018; Liu et al., 2017).

1.5.2 A emergência de K. pneumoniae hipervirulentas

Um novo patovar hipervirulento de *Klebsiella pneumoniae* (hvKp) foi descrito no sudeste asiático na década de 1980 (Catalán-Nájera et al., 2017). As características mais marcantes de hvKps é que são adquiridas da comunidade e a sua capacidade de causar infecções de alto risco de vida em indivíduos adultos saudáveis, principalmente abcesso piogênico do fígado, mas também endoftalmite, meningite e bacteremia. Por outro lado,

as cepas clássicas (cKp) de *Klebsiella pneumoniae* são oportunistas e nosocomiais, afetando pacientes severamente enfermos ou imunodeprimidos (Marr e Russo, 2019). Além disso, hvKps são capazes de desenvolver metástases, difundindo-se para múltiplos sítios de infecção, principalmente cérebro e olhos (Catalán-Nájera et al., 2017). Esta diferença fenotípica é atribuída, principalmente, a variações no conteúdo gênico do genoma acessório da espécie, regulado majoritariamente por plasmídeos e elementos integrativos (Martin e Bachman, 2018).

Por muito tempo, acreditou-se que o fenótipo de hipervirulência poderia ser caracterizado pela identificação do fenótipo de hipercápsula, através do "String test" positivo. Desta forma, a hipermucoviscosidade era considerada um marcador da hipervirulência junto com sorotipos capsulares específicos e a presença de certos sideróforos (Paczosa e Mecsas, 2016). Porém, em 2017, Catalán-Nájera, discute que, embora capaz de potencializar a virulência bacteriana, a hipermucoviscosidade e a hipervirulência são fenótipos diferentes(Catalán-Nájera et al., 2017). Além disso, comentam que estes fenótipos não são exclusivos dos sorotipos capsulares K1 e K2 (Cubero et al., 2016; Chuang et al., 2013; Liu et al., 2014; Luo et al., 2014; Cubero et al., 2016; Wu et al., 2017).

Em 2019, Marr e Russo discutiram sobre novos biomarcadores que têm sido efetivamente utilizados para predizer fenótipos de hipervirulência. Isto, tem instigado pesquisadores a invocarem colaborações internacionais que possibilitem o estabelecimento de um consenso sobre os marcadores da hipervirulência (Marr e Russo, 2019). O sequenciamento de diversas cepas hipervirulentas identificou a presença de dois plasmídeos de virulência conservados: pK2044 e pLVPK. Nestes plasmídeos encontram-se genes, experimentalmente testados, responsáveis pelo fenótipo de hipervirulência: *iuc*, *peg344*, *rmpA*, e *rmpA2* (Marr e Russo, 2019).

Apesar da hipermucoviscosidade sozinha não ser um marcador da hipervirulência, nota-se que a produção de hipercápsula mediada pelos genes *rmpA* e *rmpA2* é hvKp-específica (Russo e Marr, 2019). Somado a isso, cepas hvKp são capazes de produzir quatro diferentes sideróforos (Ent, Ybt, Aer e Sal). Destes, a aerobactina parece ser específica de cepas hipervirulentas. Portanto, atualmente, os marcadores mais eficientes do fenótipo de hipervirulência são os genes *iuc* (aerobactina), *rmpA* e/ou *rmpA2* (Russo e Marr, 2019).

Desta forma, caminha-se para uma catalogação mais atualizada e completa sobre as características epidemiológicas, genéticas e moleculares capazes de eficientemente caracterizar o fenótipo de hipervirulência, dirigindo-se para uma definição mais acurada de biomarcadores da hipervirulência (Marr e Russo, 2019; Russo e Marr, 2019).

1.6 Convergência de genes de resistência e virulência

A maior preocupação acerca do flexível genoma acessório da espécie *K. pneumoniae* é o surgimento de isolados hipervirulentos e multirresistentes. Preocupantemente, isolados que apresentam ambos estes fenótipos já foram detectados (Holt et al., 2015). Mais recentemente, na China, foram identificados isolados hvKp ST11 resistentes a carbapenêmicos que adquiriram porções do plasmídeos de virulência pLVPK (Gu et al., 2018).

Os principais mecanismos pelos quais a convergência destes fenótipos pode acontecer são: (i) através da aquisição de plasmídeos de resistência por cepas hvKp; (ii) através da aquisição e integração de elementos móveis que contenham genes de resistência em cepas hvKp; (iii) através de mutações gênicas; ou através da aquisição de plasmídeos de virulência por cepas cKp multirresistentes (Russo e Marr, 2019).

Porém, nota-se que a taxa de aquisição de plasmídeos de virulência por clones MDR é maior que a de aquisição de plasmídeos de resistência por clones hipervirulentos (Wyres et al., 2019). Além disso, que a taxa de recombinação em cepas hipervirulentas é menor (Wyres et al., 2019). Consequentemente, hvKps não são capazes de adquirir determinantes de resistência antimicrobiana tão rápido quanto cKps e, por isso, especula-se que existam barreiras intrínsecas como a incompatibilidade entre plasmídeos e a superexpressão da cápsula bacteriana (Russo e Marr, 2019). Portanto, a convergência destes fenótipos é mais fácil de acontecer em cepas MDR que, desta forma, representam enorme risco à saúde pública global.

1.7 Incidência e monitoramento de Klebsiellas multirresistentes no Brasil

Devido a limitações terapêuticas, cepas multirresistentes com múltiplos fatores de virulência de *K. pneumoniae* têm sido constantemente associadas a alta morbidade e mor-

talidade. Esta é uma preocupação mundial e, por isso, seu monitoramento epidemiológico é essencial. No Brasil, a maioria das cepas relatadas são de STs (do inglês "Sequence Type") do complexo clonal CG258, que frequentemente é associado a vários genes de resistência e virulência na América Latina (Azevedo et al., 2019; Aires et al., 2019; Palmeiro et al., 2019; Andrey et al., 2019; Longo et al., 2019).

No geral, os genes mais difundidos e frequentemente encontrados nos genomas de *Klebsiella pneumoniae* isoladas no Brasil são genes de resistência a beta-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*). Alarmantemente, crescem os relatos de cepas multirresistentes contendo vários dos genes de virulência (Ferreira et al., 2019; Andrey et al., 2019; Aires et al., 2019).

Apesar de poucos casos no Brasil, existem relatos de cepas hipermucoviscosas MDR, fenótipo associado a altas taxas de mortalidade (Azevedo et al., 2019; Aires et al., 2019). Somado a isso, relatou-se pela primeira vez no Brasil a identificação do gene bla_{NDM} e o primeiro surto hospitalar causado por este (Nava et al., 2019; Monteiro et al., 2019). Por último, foi identificado também a emergência de resistência a colistina (um dos antibióticos de último recurso) em cepa pan-resistente no Rio de Janeiro (Longo et al., 2019).

Todos estes relatos aumentam as preocupações quanto ao surgimento de cepas MDR capazes de causar infecções na comunidade, demonstrando a necessidade de estudos epidemiológicos de monitoramento e caracterização como o realizados por Aires (2019), Azevedo (2019), Palmeiro (2019) e seus colaboradores. De maneira geral, poucos são os estudos que fazem uso da genômica e, mesmo estes, a utilizam apenas em um pequeno subconjunto das amostras.

1.8 A genômica de bactérias patogênicas

Em essência, o universo dos patógenos se apresenta em constante evolução o que implica no seu continuado monitoramento. Estratégias experimentais para detecção dos patógenos vem constantemente sendo aprimoradas e compreendem diversas abordagens

como: a cultura de amostras clínicas, a caracterização fenotípica por testes bioquímicos e morfológicos, espectrometria de massa (MALDI-TOF), variações da técnica de PCR (*Polymerase Chain Reaction*), PFGE (*pulsed-field gel electrophoresis*) e genotipagem por técnicas como RFLP (*PCR restriction fragment length polymorphism*) e MLST (*multilocus sequence typing*) (Fournier et al., 2013).

O MLST é uma técnica de caracterização genética bastante difundida e acessível, que é baseada no sequenciamento de fragmentos (≈500 bp) de genes essenciais (*housekeeping*) e que provê um sistema padronizado e reprodutível de identificação e nomenclatura para cepas de uma espécie (Lee, 2017). As diferentes combinações de alelos destes genes em cada bactéria permitem a definição de perfis alélicos também chamados de "Sequence type" (ST). Em 2005, foi estabelecido um esquema de MLST para *Klebsiella pneumoniae* (Diancourt et al., 2005) que pode ser acessado e analisado através do banco de dados BIGSdb Pasteur[2]. Este banco é constantemente atualizado e contem 4.855[3] diferentes perfis alélicos de MLST para *K. pneumoniae* e somente 986 para *E. coli*, o que evidencia a grande diversidade genética de *K. pneumoniae* (Wyres e Holt, 2016).

O MLST é extremamente útil na caracterização de populações de patógenos em nível regional ou global, mas não tem resolução para distinguir cadeias de transmissão ou mesmo os determinantes genéticos de fenótipos de virulência ou de resistência a antibióticos. O maior nível de discriminação genético possível para detecção de patógenos é, sem dúvida, a sequência completa dos seus genomas. De posse da sequência do genoma pode-se traçar de maneira ampla a dinâmica evolutiva, rotas de introdução e dispersão, identificar bases moleculares para desenvolvimento de resistência a fármacos, além da identificação de plasmídeos e distinção de cepas e até novas espécies de patógenos (Fournier et al., 2013; Lee, 2017).

De fato, a genômica já teve impactos significativos na investigação de surtos e diagnósticos clínicos. Em um dos primeiros casos, sua aplicação tornou possível traçar rotas de dispersão de um surto hospitalar de *K. pneumoniae*, o que não foi possível utilizando técnicas de PFGE ou PCR. A partir das informações genômicas, além da cepa causadora, foi-se capaz de identificar o acúmulo de mutações independentes que

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

[3] Acessado em Janeiro de 2020

culminaram na resistência a droga colistina (Snitkin et al., 2012).

Alguns pesquisadores chegam até mesmo a sugerir que o sequenciamento de genomas irá substituir completamente as técnicas usadas historicamente e revolucionar a prática microbiológica clínica (Boolchandani et al., 2019). Antes disso, a obtenção dos genomas deve ser rápida e custo-efetiva, e este cenário se materializa com recentes avanços tecnológicos no sequenciamento de DNA.

1.8.1 Novas tecnologias de sequenciamento de DNA

O campo da genômica teve um grande impulso quando em 2005 se inaugurou a era das novas tecnologias de sequenciamento de DNA, ou *Next Generation Sequencing* (NGS) (Reuter et al., 2015), que permitiram gerar uma grande quantidade de sequências por uma fração dos custos até então. Na primeira onda do NGS despontou a tecnologia da companhia Illumina que está em constante evolução e é a mais utilizada até hoje (≈90% da geração total de dados de NGS).

A particularidade da tecnologia Illumina é sua capacidade de sequenciar milhões de fragmentos de DNA de forma simultânea e com grande acuidade, mas o tamanho máximo de cada leitura é de 150 bases (leituras curtas). Por sua vez, o tamanho dos fragmentos sequenciados tem influência na capacidade de se reconstruir a sequência original do genoma, e por conseguinte, somente a utilização de dados de Illumina não garante a montagem completa do genoma, mas provê um rascunho de boa qualidade.

Recentemente, novas companhias entraram comercialmente neste ramo, tais como Pacific Biosciences (PacBio) e Oxford Nanopore Technologies (ONT) (Reuter et al., 2015), as quais inauguraram a terceira geração de sequenciamento de DNA. Diferentemente da tecnologia Illumina, estas são capazes de ler grandes fragmentos de DNA (leituras longas) com média de 5.000 bases, mas há relatos de leituras excedendo 200.000 bases. Leituras longas facilitam o esforço computacional para a montagem de genomas, mas tanto a PacBio (Korlach et al., 2010) quanto ONT (Feng et al., 2015) sofrem com significativas taxas de erro na nomeação das bases (≈5%) além de custo por base mais alto do que a tecnologia Illumina. Dentre as tecnologias citadas, o foco do presente projeto está na tecnologia da Oxford Nanopore, por motivos detalhados a seguir.

1.8.2 *Sequenciamento por nanoporos*

Dada uma camada lipídica sobre a qual é aplicada uma diferença de potencial entre os dois lados da membrana, a translocação de uma uma molécula fita simples de DNA por um canal iônico presente nesta membrana gera alteração da corrente elétrica (Kasianowicz et al., 1996). Estas flutuações na corrente são dependentes da natureza das bases que estão passando pelo lúmen canal, o que gera um perfil elétrico que se altera de acordo com o contexto das bases. Consequentemente, os diferentes perfis de bloqueio de corrente podem ser convertidos em informação de sequência. Isto serve como arcabouço teórico para uma elegante técnica de sequenciamento de DNA, o sequenciamento por nanoporos (Wang et al., 2014; Feng et al., 2015). Uma visão esquemática deste processo se encontra na figura 1.2.

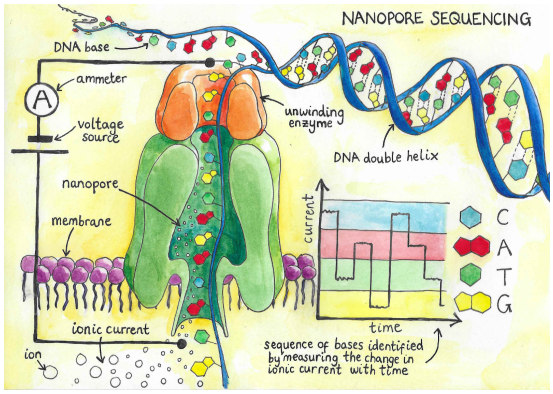


Figura 1.2: Representação esquemática do sequenciamento de DNA por nanoporos, mostrando o bloqueio diferencial de corrente nos poros em função de sequências específicas de DNA que atravessam o poro. Fonte: Retirado do artigo *Pore over this: Advances in DNA sequencing*, do blog *Oxford Scientist*[4].

No começo de 2014, a ONT disponibilizou o primeiro aparato comercial capaz de proceder o sequenciamento por nanoporos: o MinION[5]. Surpreendentemente, este sequenciador é um dispositivo que cabe na palma da mão sendo alimentado por uma porta USB ligada a um computador pessoal. Dentro do MinION existem milhares de nanoporos dispostos em uma malha regular, cada qual endereçado por um registrador eletrônico de voltagem. Desta forma, os milhares de poros podem, independentemente, registrar sequências.

[4] `http://oxsci.org/pore-over-this-advances-in-dna-sequencing/`

[5] `https://nanoporetech.com/`

Sua portabilidade extrema e leitura contínua em tempo real representam uma profunda mudança de paradigma de sequenciamento genômico: ao invés de se enviar amostras para centros de sequenciamento no exterior o dispositivo pode ser diretamente utilizado na clínica. Ressalta-se que o processo leva de 24 a 48 horas da coleta de amostras até a geração de sequências que posteriormente são processadas computacionalmente para obtenção dos genomas.

1.8.3 Avanços do sequenciamento de patógenos por nanoporos

Em uma de suas primeiras aplicações, em 2015, Loman e colaboradores lograram êxito em montar o genoma de *E. coli* em um único "contig" de 4,6 Mb, usando somente dados de sequenciamento de DNA por nanoporos. Este trabalhou demonstrou a possibilidade de reconstruir genomas de alta qualidade (≈95% de identidade) apesar das elevadas taxas de erros de dados de nanoporo (Loman et al., 2015). E com o passar dos anos, estas perspectivas têm se tornado cada vez mais favoráveis.

A identificação de espécies bacterianas tem derivado majoritariamente de técnicas dependentes de cultura. Nos últimos anos, análises independente de cultura baseadas em sequenciamento de DNA têm sido desenvolvidas.

Em Janeiro de 2018, Gong e colaborados descreveram uma metodologia independente de cultura utilizando DNase, centrifugação diferencial e sequenciamento de DNA por nanoporos capaz de identificar a bactéria (*K. pneumoniae*, neste caso) e seus genes de resistência em tempo real, em um caso de abcesso no fígado (Gong et al., 2018). Atualmente, o tratamento desta patologia depende da cultura, isolamento e identificação do patógeno. Porém, este trabalho demonstra que o sequenciamento de DNA por nanoporos é capaz de entregar, em tempo real, uma análise rápida, independente de cultura e precisa que pode ser utilizada para o diagnóstico em casos de abscesso no fígado.

No mesmo ano, Li e colaboradores descreveram um protocolo barato e eficiente para a obtenção de sequências completas de 20 plasmídeos MDR em uma única corrida de sequenciamento por nanoporos. A acuidade final das montagens foi de 97% e a multiplexagem resultou em custo por plasmídeo de menos de US$ 50. O mais importante foi a possibilidade de resolver completamente a estrutura complexa de cassetes de genes de

resistência a antibióticos, muitos com duplicações de genes e elementos transponíveis (Li et al., 2018). Demonstrando a capacidade de se detalhar os mecanismos moleculares responsáveis pela resistência a antibióticos e possibilitar estudos de evolução e transmissão destes mecanismos.

No fim de 2018, Tamma e colaboradores descreveram um estudo comparativo de modo a determinar a eficiência da predição de susceptibilidade antimicrobiana baseada em sequenciamento por nanoporos (Tamma et al., 2019). Neste trabalho, foram selecionados 40 isolados clínicos (sendo 21 *K. pneumoniae* resistentes a carbapenêmicos) que foram analisados de duas formas diferentes: em tempo real e através da montagem dos genomas. As análises em tempo real produziram anotação completa dos genes de resistência em 8 horas, com 77% de precisão enquanto a análise baseada em montagem de genomas durou 14 horas com 92% de precisão. Segundo o grupo, ao avaliar os pacientes dos quais as amostras foram obtidas, este método poderia ter diminuído o tempo de escolha da melhor terapia antimicrobiana em até 26 horas quando comparado com os testes de susceptibilidade tradicionais.

A metodologia padrão para a análise de infecções do trato respiratório inferior é através do isolamento e cultura de patógenos, método com baixa sensibilidade e muito lento para guiar terapias antimicrobianas. Por isso, foi desenvolvida uma outra técnica de diagnose independente de cultura baseada em sequenciamento de DNA por nanoporos (Charalampous et al., 2019), que foi capaz de detectar patógenos e seus genes de resistência em 6 horas, com 96,6% de sensibilidade quando comparados com dados obtidos através das metodologias padrão. Desta forma, demonstraram que a aplicação do sequenciamento de metagenomas utilizando nanoporos pode de maneira rápida e acurada caracterizar patógenos de infecções de trato respiratório inferior e contribuir para a aplicação guiada de tratamentos antimicrobianos.

Cepas de *E. coli* produtoras de toxina Shiga (STEC, do inglês "*Shiga-toxin producer E. coli*") são patógenos alimentares responsáveis por infecções de alta morbidade e mortalidade em todo o mundo (Smith et al., 2019). Por isso, a rápida caracterização de STECs é importantíssima para identificar cepas circulantes que podem ser de grande ameaça à saúde humana. Neste contexto, González-Escalona e colaboradores descreveram a bem sucedida aplicação do sequenciamento de DNA por nanoporos para a rápida

identificação e caracterização de genes de virulência, resistência, fagos Stx (toxina Shiga) e plasmídeos de duas STECs (González-Escalona et al., 2019). Este estudo demonstra a viabilidade do sequenciamento por nanoporos como uma alternativa rápida e econômica para a identificação de marcadores de virulência específicos em genomas de STECs.

Mais recentemente, Taylor e colaboradores descreveram a utilização do sequenciamento de DNA por nanoporos para simultaneamente obter as sequências genômicas e plasmidiais de uma *Salmonella enterica subsp. enterica* e uma *E. coli* O157:H7 (Taylor et al., 2019). Neste estudo, demonstraram o potencial uso da tecnologia em agências de vigilância sanitária para a obtenção rápida e econômica de genomas de alta qualidade de patógenos alimentares para a inferência filogenética, identificação de sorotipo, fatores de virulência e genes de resistência.

1.9 Genômica de populações para a caracterização de bactérias patogênicas

Os grandes avanços na obtenção de genomas permitiram o estabelecimento da genômica de populações bacterianas, onde diversos genomas de isolados são sequenciados e comparados, de forma a fornecer um inventário sem precedentes das variações genéticas na população.

Mason e colaboradores descreveram um estudo de genômica comparativa utilizando 5.310 genomas de *Mycobacterium tuberculosis* obtidos dos cinco continentes para investigar a dinâmica evolutiva de cepas multirresistentes (Manson et al., 2017). Descobriu-se padrões de surgimento de cepas multirresistentes de *M. tuberculosis* conservados em todo o mundo. Em todas as cepas MDR, a resistência a isoniazida surgiu antes de qualquer outra e, quando não respeitada esta ordem, as cepas raramente evoluem para fenótipos MDR. Desta forma, à época, o grupo sugeriu o desenvolvimento de diagnósticos capazes de identificar mutações no gene *katG* de modo a identificar cepas monorresistentes a isoniazidas e prevenir o surgimento de bactérias MDR.

Em contraste, 4.022 genomas de *E. coli* foram sequenciados provendo uma visão populacional do resistoma, conjunto de genes de resistência, dessa espécie (Goldstone e Smith, 2017). Com estes dados massivos de genômica foi-se capaz de identificar 7 classes

de antibióticos para os quais a espécie é intrinsecamente resistente e 118 combinações de antibióticos que nunca ocorrem simultaneamente em nenhum dos genomas sequenciados e que, talvez, possam representar novas terapias capazes de mitigar os problemas advindos do surgimento de fenótipos MDR.

Roe e colaboradores divulgaram um estudo integrando dados de genômica e transcritômica de 107 isolados de *Acinetobacter baumanii* para investigar os principais mecanismos de resistência da espécie (Roe et al., 2019). Constatou-se, neste estudo, a dificuldade de sua tipagem visto que muitos dos genomas analisados haviam sido erroneamente classificados como *A. baumanii*, o que demonstra a necessidade de melhora nos diagnósticos clínicos. Demonstrou-se que a vigilância genômica de patógenos extremamente plásticos não será atingida somente com dados genômicos, devido ao fato dos genes de resistência não serem conservados entre as linhagens da espécie e que dados transcritômicos são complementos valiosos para estes estudos visto que, através destes, foram capazes de identificar padrões de expressão de genes de resistência e novos mecanismos de resistência.

Ao compararem 2.498 genomas de *Klebsiella pneumoniae*, Lam e colaboradores investigaram a prevalência, evolução e mobilidade do gene *ybt*, fator de virulência essencial para o desenvolvimento de infecções invasivas (Lam et al., 2018). Identificou-se que o elemento genético móvel (*Mobile genetic element*) *ICEkp* está presente em um terço da população da espécie, sugerindo que este MGE é extremamente dinâmico e mantido na população principalmente através de transferência horizontal. Por último, descobriu-se um novo mecanismo de transmissão do gene *ybt* através de plasmídeos FIB_k, capazes de co-associar cassetes de resistência e virulência em um único plasmídeo de alta estabilidade, o que representa grande risco à saúde pública mundial.

Em Abril de 2019, Wyres e colaboradores descreveram um estudo de genômica evolutiva comparativa utilizando >2.200 genomas de *Klebsiella pneumoniae* para caracterizar os 28 grupos clonais mais comuns da espécie (Wyres et al., 2019). Seus dados mostram que clones MDR apresentam maior taxa de recombinação e são mais diversos que clones hipervirulentos, possivelmente devido a restrições particulares ao fenótipo. Somado a isso, a taxa de aquisição de plasmídeos de virulência por clones MDR é maior que a taxa de aquisição de plasmídeos de resistência por clones hipervirulentos. Desta forma, clones MDR, apresentam maior risco à saúde pública e a urgente necessidade do

estabelecimento de sistemas de vigilância genômica capazes de integrar informações clonal, de resistência e de virulência que são praticamente inviáveis de serem adquiridas se não através da genômica.

Desta forma, fica claro que a epidemiologia genômica é de extrema importância para a investigar a evolução da resistência e virulência bacteriana, para caracterizar e compreender os mecanismos de resistência de bactérias patogênicas, além de se mostrar grande aliada no desenvolvimento de terapias mais eficientes (Hendriksen et al., 2019; Kan et al., 2018). É notório que estas abordagens genômicas torna-se-ão mais frequentes e, hoje, a maior dificuldade se encontra no armazenamento e integração dos dados.

1.10 Análise computacional de dados genômicos

Em qualquer escala de estudo, os dados genômicos gerados pelas tecnologias experimentais só podem ser usados para interpretar fenômenos biológicos com o auxílio da bioinformática. O emprego de diversas ferramentas computacionais faz-se necessário em todas as etapas analíticas, do pré-processamento de dados para remoção de erros de sequenciamento e contaminantes, passando por estratégias para montagem do genoma até chegar em diversos aspectos da anotação gênica. De fato, atualmente as análises bioinformáticas exigem muito mais tempo e mão-de-obra do que o sequenciamento propriamente dito, que no caso de nanoporos pode ser finalizado em até um dia.

Se por um lado o NGS democratiza a genômica e compele a sua aplicação regular em laboratórios de microbiologia, ao mesmo tempo introduz uma enorme lacuna na possibilidade de se proceder uma análise efetiva dos dados (Muir et al., 2016). As análises bioinformáticas requerem especialistas multi-disciplinares para desenhar e implementar verdadeiros protocolos experimentais, mas em outro âmbito, o *in silico*, ao invés de *in vivo* ou *in vitro*.

1.10.1 Protocolos computacionais automatizados

Em uma análise genômica, os dados de NGS são utilizados como arquivos de entrada para diversas ferramentas computacionais, as quais geram resultados que servem de entrada para outros programas, em uma verdadeira linha de montagem que requer

dezenas de etapas. (Grüning et al., 2017). O protocolo computacional que materializa estas é comumente denominado de *pipeline*.

No início, *pipelines* eram produzidos em forma de *scripts* em linguagens de programação como PERL e Python. Apesar de possibilitarem a implementação de variáveis e da lógica de processamento de dados, *scrips* são extremamente básicos e carecem de dois pontos importantíssimos: a dependência, habilidade de definir dependências entre etapas e a reentrância, habilidade de recomeçar a análise do ponto de interrupção sem a necessidade de refazer todo o processo (Leipzig, 2016). Além disso, *scripts* são extremamente particulares ao problema para o qual foram desenvolvidos e, por não possuírem sintaxe pré-definida, tendem a não possuir boa organização e serem pouco reprodutíveis (Perkel, 2019).

Para solucionar estas deficiências, nos últimos anos, novos sistemas automatizados de gerenciamento de *pipelines* vêm sendo desenvolvidos (Strozzi et al., 2019). Estas novas abordagens possuem linguagens e sintaxes pré-estabelecidas (*Domain Specific Languages*, DSL) para a composição de *pipelines* e, por isso, são padronizadas e de melhor compreensão, o que permite a interpretação correta dos passos por qualquer usuário familiarizado com a sintaxe (Perkel, 2019). Somado a isto, estes novos sistemas fazem uso das ideias de reentrada e dependência, que permitem a tolerância de erros (algo bastante frequente em protocolos computacionais) e reinício no ponto de interrupção, proporcionando *pipelines* robustos, escalonáveis e altamente reprodutíveis (Leipzig, 2016; Perkel, 2019).

A bioinformática necessita da padronização de protocolos para facilitar o compartilhamento e desenvolvimento de análises portáteis e reprodutíveis. Atualmente, existem diversos sistemas de gerenciamento de *pipelines*, cada um com suas particularidades, capazes de realizar esta tarefa, como mostrado na tabela 1.1, e, cabe ao pesquisador, definir qual sistema melhor se enquadra sob suas demandas e pretensões.

Tabela 1.1 - Principais sistemas automatizados de gerenciamento de *pipelines* da atualidade

Nome	Linguagem	Interface gráfica	Referência
Bpipe	Groovy	Não	`10.1093/bioinformatics/bts167`
CWL	CWL	Não	`10.6084/m9.figshare.3115156.v2`
Galaxy	Python	Sim	`10.1093/nar/gky379`
KNIME	Java	Sim	`10.1016/j.jbiotec.2017.07.028`
Nextflow	Groovy	Não	`10.1038/nbt.3820`
Snakemake	Python	Não	`10.1093/bioinformatics/bts480`

1.10.2 Contêineres computacionais

Apesar de todos os esforços para facilitar a implementação das tarefas descritas na seção anterior, a própria instalação dos programas requeridos para as análises ainda é uma grande barreira. A maioria dos programas utilizados na bioinformática são desenvolvidos para o sistema operacional Linux e, muitos requerem diversas dependências que exigem conhecimento em Linux para serem devidamente instaladas. Isto, gera um enorme obstáculo para a implementação de *pipelines* padronizados em laboratórios com pouca familiaridade ao sistema.

Existem, por exemplo, iniciativas entre desenvolvedores para o desenvolvimento de programas que tentam facilitar o processo de coleta e instalação de dependências e programas, como o bioconda (Dale et al., 2018). Porém, ainda que automatize a instalação de programas e suas dependências, esta ferramenta ainda demanda conhecimento em Linux, fator que representa obstáculo significativo (Grüning et al., 2018).

Por este motivo, contêineres computacionais estão ganhando cada vez mais espaço na bioinformática. Contêineres são ambientes heterogêneos capazes de abstrair a instalação de programas e dependências. Estes ambientes isolados e portáteis permitem o empacotamento de aplicações e bibliotecas computacionais necessárias para a execução de tarefas específicas (Grüning et al., 2018). São pequenos recipientes de um sistema operacional, restringidos somente às bibliotecas requeridas para sua execução, diferentemente de uma máquina virtual que abriga todo um sistema operacional completo.

Com este arcabouço, um contêiner computacional permite a criação de blocos independentes e autossuficientes dedicados a suas aplicações, que podem ser executados

em qualquer sistema operacional. Assim que um contêiner é criado, o desenvolvedor pode adicionar todos os programas e suas dependências a este, criando um ambiente único e reprodutível que pode ser distribuído para os usuários finais, geralmente chamado de *imagem* (Boettiger, 2015). Além de serem agnósticos em relação ao sistema operacional do usuário, os contêineres podem ser executados em equipamentos variados, desde laptops até super computadores de alta performance com múltiplos nós.

Portanto, é possível obter a portabilidade, reprodutibilidade e escalonamento de *pipelines* através da distribuição destes com seus respectivos programas e dependências requeridas devidamente instalados e encapsulados em contêineres computacionais (Boettiger, 2015). Dentre as tecnologias disponíveis, Docker (Merkel, 2014) e Singularity (Kurtzer et al., 2017) são ótimos exemplos por serem completamente integradas pelo sistema de gerenciamento automatizado de *pipelines* Nextflow.

1.10.3 Combinando pipelines e contêiners: Nextflow

Nextflow é uma linguagem de domínio específico (DSL) que permite o rápido desenvolvimento, adaptação e gerenciamento de *pipelines* (Di Tommaso et al., 2017). Esta DSL disponibiliza um sistema de gerenciamento de trabalhos desenvolvido para executar processos em paralelo, tolerar erros, permitir a rastreabilidade do código e resolver a comunicação entre processos.

A tecnologia Nextflow é completamente integrada à conteinerização Docker, permitindo incorporar imagens de contêineres Docker de maneira automática, prática e rápida. Nextflow se encarrega de montar, executar e desacoplar as imagens Docker no momento de sua execução, de modo que os arquivos da máquina e da imagem continuem acessíveis durante e após a realização de tarefas. Isto é feito sem a intervenção do usuário. Portanto, Nextflow, torna simples a tarefa de gerenciamento de contêineres.

O desenvolvimento de um *pipeline* que se beneficie de ambas estas tecnologias garante de maneira direta sua reprodutibilidade em qualquer computador por manter, na imagem Docker, todas as dependências necessárias para sua execução. Isto simplifica e automatiza a instalação e execução de programas de computador, uma vez que não há necessidade de intensas etapas de configuração e gerenciamento do ambiente computaci-

onal.

1.11 pipelines de genômica bacteriana

Contextualizados com as realidades de cada época, ao longo dos anos, inúmeros *pipelines* especializados em genômica bacteriana foram desenvolvidos. Implementados de diversas maneiras, estas ferramentas podem ser comparadas quanto às formas de distribuição, sistema operacional requerido, tipo de dado aceito, etc. Além disso, não necessariamente um *pipeline* é capaz de realizar todas as etapas de estudos genômicos. Estas ferramentas podem ser divididas em três grupos: (i) de montagem de genomas; (ii) de anotação de genomas; (iii) híbridas. Um resumo sobre diferentes *pipelines* de genômica bacteriana e suas características básicas se encontra na tabela 1.2.

Em 2010, em uma das primeiras tentativas de compartilhamento de *pipelines* portáteis e reprodutíveis de genômica bacteriana, Kislyuk e colaboradores divulgaram um conjunto de *scripts* capaz de montar e funcionalmente anotar genomas bacterianos sequenciados através das plataformas 454, ABI SOLiD, Illumina e Sanger (Kislyuk et al., 2010). No ano seguinte, em 2011, Kumar e colaboradores descreviam uma solução integrada para a anotação e visualização de genomas bacterianos chamada AGeS. Apesar de possuir interface gráfica, este pipeline podia ser instalado e executado somente em sistemas Linux (Kumar et al., 2011).

Pipelines como PGAP (Tatusova et al., 2016) e DFAST (Tanizawa et al., 2018), por exemplo, são ferramentas específicas para a anotação de genomas, distribuídos para a execução local ou através de serviços web. O primeiro, oferecido pelo NCBI é bastante utilizado para a anotação automática de genomas submetidos ao GenBank. Já o segundo, foi desenvolvido para facilitar a submissão de genomas ao banco de dados DDBJ.

De forma análoga, alguns centros de integração de dados como PATRIC (Antonopoulos et al., 2019; Wattam et al., 2018), KBase (Arkin et al., 2018) e MicroScope (Vallenet et al., 2019) disponibilizam, em plataformas *online* de trabalho, diversos *pipelines* e programas que possibilitam que o usuário, além de montar e anotar genomas, seja capaz de integrar seus dados àqueles depositados nestes bancos em análises de genômica comparativa como inferências filogenéticas e exploração de pangenoma.

Tabela 1.2 - Características básicas quanto à tarefa realizada (montagem, anotação ou híbrido) e à forma de execução (web ou local) de diversos *pipelines* de genômica bacteriana

Nome do programa	Tipo (tarefa)	Execução	Referência (DOI)
AGeS	Anotação	Web	10.1371/journal.pone.0017469
ASA3P	Híbrido	Local	10.1101/654319
CloVR-Microbe	Híbrid	Local	10.1038/npre.2011.5887.3
bacass	Híbrido	Local	10.5281/zenodo.3574476
BacPipe	Híbrido	Local	10.1016/j.isci.2019.100769
BugBuilder	Híbrido	Local	10.1101/148783
DeNoGAP	Anotação	Local	10.1186/s12859-016-1142-2
DFAST	Anotação	Local e web	10.1093/bioinformatics/btx713
GAAP	Montagem	Local	10.1186/s12864-016-3267-0
GAMOLA2	Anotação	Local	10.3389/fmicb.2017.00346
Genix	Anotação	Local e web	10.1093/femsle/fnw263
KBase	Híbrido	Web	10.1038/nbt.4163
MEGAnnotator	Híbrido	Local	10.1093/femsle/fnw049
MICRA	Anotação	Web	10.1186/s13059-017-1367-z
MicroScope	Anotação	Web	10.1093/nar/gkz926
MyPro	Híbrido	Local	10.1016/j.mimet.2015.04.006
PATRIC (RAST)	Híbrido	Web	10.1007/978-1-4939-7463-4_4
P-CAPS	Anotação	Local e web	10.1089/cmb.2017.0066
PGAP	Anotação	Local e web	10.1093/nar/gkw569
TORMES	Híbrido	Local	10.1093/bioinformatics/btz220
Tychus	Híbrido	Local	10.1101/283101
Nanopype	Pré-processamento	Local	10.1093/bioinformatics/btz461
Sanger-pathogens pipeline	Montagem	Local	10.1099/mgen.0.000083

Com metodologia bastante diferente dos demais, MICRA, é um *pipeline* web que foca em análises de identificação microbiana e genômica comparativa baseado em mapeamento de leituras e não em montagem de genomas (Caboche et al., 2017). Apesar de identificar genes de resistência e virulência em 10 minutos, MICRA somente aceita dados provenientes das tecnologias Illumina e Ion Torrent e, por ser baseado em mapeamento de leituras, perde-se questões estruturais que podem ser importantes para a biologia da bactéria.

Existem ainda *pipelines* recentes que representam as novas tendências de desenvolvimento e compartilhamento de *pipelines* através de contêineres e sistemas de geren-

ciamento de execução, como Snakemake, e de instalação como bioconda. Nanopype, por exemplo, é um *pipeline* distribuído em contêineres Docker e Singularity, desenvolvido em Snakemake para o pré-processamento de dados de sequenciamento de dna por nanoporos (Giesselmann et al., 2019). TORMES (Quijada et al., 2019) e BacPipe (Xavier et al., 2019) são *pipelines* de montagem e anotação de genomas desenvolvidos especificamente para dados de sequenciamento Illumina, enquanto ASA3P (Schwengers et al., 2019) e bacass (Peltzer et al., 2019) são capazes de montar e anotar genomas utilizando leituras curtas e longas.

Nota-se que vários dos *pipelines* listados na tabela 1.2, são antigos, desatualizados e, consequentemente, com poucas opções de escolha quanto aos programas de montagem de genomas. Todas estas diferenças de implementações e restrições transformam a escolha do *pipeline* ideal em um processo laborioso e pouco trivial.

2

Objetivos

2.1 Objetivo geral

Desenvolver ferramentas computacionais que visam automatizar a montagem, anotação e análise de genomas procarióticos utilizando dados de novas tecnologias de sequenciamento de DNA, demonstrando sua utilização para um isolado bacteriano multirresistente a antibióticos.

2.2 Objetivos específicos

1. Desenvolver os *pipelines* computacionais genéricos, acessíveis e configuráveis para análise de genomas procarióticos
    - Desenvolver um *pipeline* de pré-processamento de dados brutos de sequenciamento de DNA das plataformas Illumina, Pacific Biosciences (Pacbio) e Oxford Nanopore Technologies (ONT)
    - Desenvolver um *pipeline* de montagem de genomas capaz de realizar diversos tipos de montagens utilizando qualquer combinação de dados de sequenciamento das plataformas Illumina, Pacbio ou ONT
    - Desenvolver um *pipeline* de anotação genômica abrangente e também especializado em anotar fatores de virulência, genes de resistência a antibióticos, elementos genéticos móveis e prófagos
    - Integrar o *pipeline* ao modelo de contêineres computacionais Docker para facilitar a instalação e reprodutibilidade

2. Aplicar os *pipelines* desenvolvidos em isolados de *Klebsiella pneumoniae* no Hospital Universitário de Brasília

    - Montar genomas dos isolados utilizando diferentes metodologias
    - Investigar genes de resistência a antibióticos
    - Investigar os fatores de virulência e genes relacionados à hipervirulência
    - Investigar genes relacionados à hipermucoviscosidade

3

Material e Métodos

3.1 Construção de pipelines computacionais

Todos os protocolos computacionais neste trabalho, os *pipelines*, foram desenvolvidos utilizando a linguagem de domínio específico descrita no arcabouço (*framework*) Nextflow (Di Tommaso et al., 2017), o qual organiza e controla a execução de diversas etapas da análise realizadas por programas, quase sempre implementados no sistema operacional Linux.

Estes programas são as dependências do *pipeline* e sua instalação individual é uma tarefa complexa. Para mitigar este problema, utilizamos a tecnologia gratuita de contêineres computacionais denominada Docker (Merkel, 2014; Boettiger, 2015). Esta permite a criação de uma imagem ("máquina virtual") com programas selecionados instalados.

3.2 Pipelines para montagem e anotação de genomas procarióticos

Os processos de montagem e anotação de genomas são complexos e possuem diversas etapas. De modo a garantir flexibilidade ao usuário final, o *pipeline* foi dividido em três módulos independentes, como esquematizado na figura 3.1. Esta arquitetura permite a criação de pontos de checagem da qualidade dos dados e possibilita que usuários executem somente as etapas de interesse.

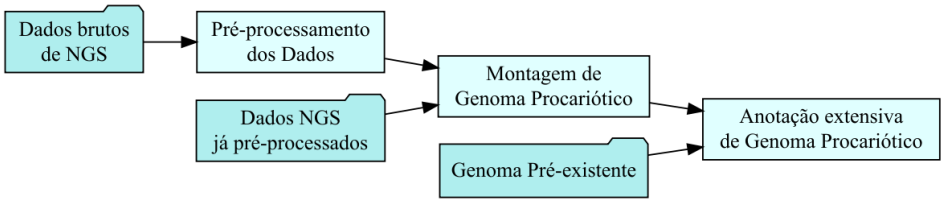


Figura 3.1: Visão geral do fluxo de trabalho do *pipeline*. Este é dividido em três módulos que são executados separadamente, na seguinte ordem: (i) pré-processamento de dados brutos de NGS (ii) montagem de genomas (iii) anotação de genomas

3.2.1 Módulo de pré-processamento de dados brutos

A fim de possibilitar o pré-processamento de dados provenientes de múltiplas plataformas de sequenciamento (Illumina, Pacbio e Oxford Nanopore) foram incorporados ao *pipeline* programas especializados para cada tipo de tecnologia, em função de diferenças quantitativas e qualitativas nas leituras. O fluxo geral de trabalho deste módulo está esquematizado na figura 3.2.

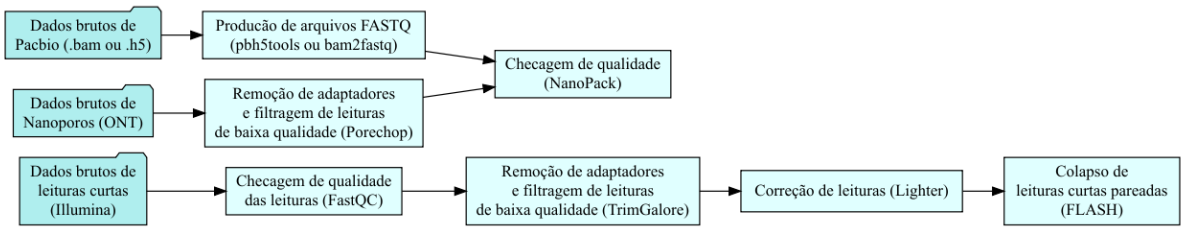


Figura 3.2: Visão geral da etapa de pré-processamento de dados brutos de sequenciamento das plataformas Illumina, Oxford Nanopore e Pacbio.

3.2.1.1 Pré-processamento de leituras curtas

Uma análise completa de leituras curtas provenientes da plataforma Illumina compreende as etapas de controle de qualidade, remoção de sequências adaptadoras (de PCR), e opcionalmente correção de erros e colapso de leituras pareadas (leituras *paired-end*). Estas etapas são configuráveis, isto é, são executadas com parâmetros padrão ou modificados

pelo usuário, e envolvem a execução dos seguintes programas:

- FastQC para controle de qualidade das leituras (`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`)
- TrimGalore (`https://github.com/FelixKrueger/TrimGalore`) para a remoção de sequências adaptadoras e de baixa qualidade (Martin, 2011)
- Lighter para a correção de erros (Song et al., 2014)
- FLASH para o colapso de leituras de protocolo *paired end* (Magoč e Salzberg, 2011)

3.2.1.2 Pré-processamento de leituras longas

O pré-processamento de leituras longas, provenientes das plataformas Pacbio e Oxford Nanopore, é composto pelas etapas de conversão dos sinais brutos em arquivos de sequência em formato FASTQ (opcional), remoção de sequências adaptadoras, desmultiplexagem (opcional), avaliação da qualidade e distribuição dos tamanhos de leituras. Para isto, utiliza-se os seguintes programas:

- Porechop (`https://github.com/rrwick/Porechop`) para remoção de sequências adaptadoras e desmultiplexagem de dados ONT
- *bam2fastq* (`https://github.com/PacificBiosciences/bam2fastx`), distribuído pela própria Pacbio, para a produção de arquivos de sequência em formato FASTQ
- pbh5tools (`https://github.com/PacificBiosciences/pbh5tools`) para a conversão de sequências para o formato FASTQ
- Nanopack para produzir resumos das estatísticas gerais (tamanho médio, número de leituras, etc.) e da qualidade média das bases das leituras pré-processadas (De Coster et al., 2018)

Todos os programas são executados com seus parâmetros padrão e podem ser modificados pelo usuário final.

3.2.2 Módulo de montagem de genomas

Desenvolvido de modo a permitir diferentes combinações de tecnologias de sequenciamento, o módulo de montagem de genomas possibilita aproveitar a alta qualidade de

leituras curtas produzidas pelas plataformas Illumina e o longo alcance proporcionado por leituras longas das plataformas Pacbio e Oxford Nanopore. Disponibiliza-se no *pipeline* três maneiras de montar genomas, esquematizadas na figura 3.3. O usuário é capaz de montar o genoma de maneira híbrida mesclando leituras curtas e longas ou utilizando somente leituras curtas ou longas, de acordo com a sua disponibilidade. Todos os programas são executados com os parâmetros padrão e podem ser personalizados pelo usuário através de um arquivo YAML distribuído em conjunto ao *pipeline*.

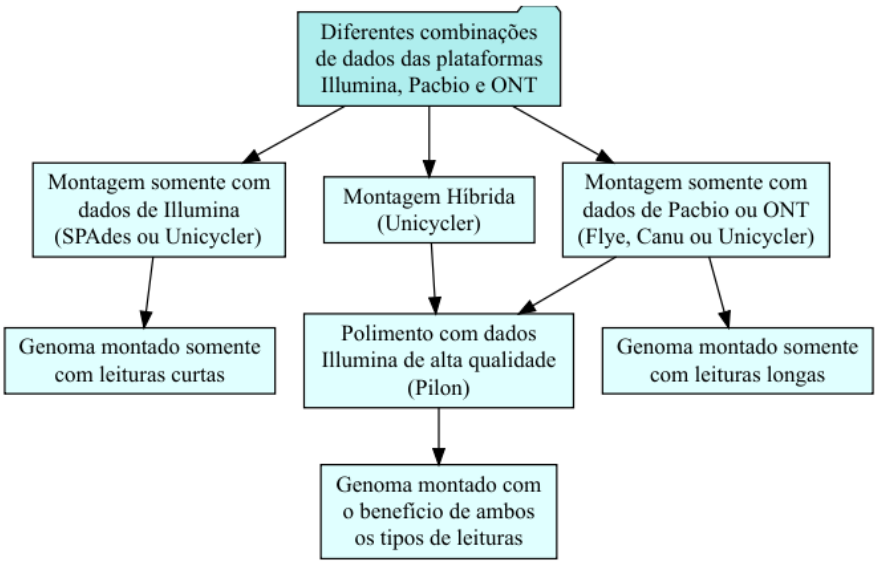


Figura 3.3: Visão geral da etapa de montagem de genomas. É possível montar genomas utilizando diferentes combinações de dados de sequenciamento das plataformas Illumina, Pacbio e Oxford Nanopore.

3.2.2.1 Montagem de leituras curtas

Implementou-se neste *pipeline* a montagem de leituras curtas através dos programas Unicycler (Wick et al., 2017) e SPAdes (Bankevich et al., 2012). Esta etapa pode ser observada no fluxo à esquerda da figura 3.3.

3.2.2.2 Montagem de leituras longas

Para a montagem de leituras longas, escolheu-se os programas Canu (Koren et al., 2017), Flye (Kolmogorov et al., 2019) e Unicycler (Wick et al., 2017). Esta etapa pode ser observada no fluxo à direita da figura 3.3.

Opcionalmente, após a montagem, o usuário pode realizar a correção de bases erroneamente nomeadas no genoma (polimento) utilizando os dados brutos em formato FAST5 (ONT) ou BAM (Pacbio) através dos programa Nanopolish (Loman et al., 2015) e arrow (`https://github.com/pacificbiosciences/genomicconsensus/`), respectivamente.

3.2.2.3 Montagem híbrida

A combinação de leituras curtas e longas pode ser efetuada de duas maneiras distintas, ambas foram implementadas no *pipeline*. Esta etapa pode ser observada no fluxo central da figura 3.3.

Na primeira, os dados Illumina são previamente montados de modo a produzir fragmentos de genoma (contigs) de altíssima qualidade. Em seguida, utilizando as informações de leituras longas, contigs podem ser unidos e ordenados (do inglês "scaffolding"). Em resumo, após a montagem de um genoma fragmentado de alta qualidade, seus fragmentos são consolidados a partir do seu mapeamento com leituras longas. Esta opção de montagem é implementada no *pipeline* através do programa Unicycler.

Já a segunda metodologia consiste na execução de uma montagem primária utilizando somente leituras longas de modo a se obter um genoma bastante contíguo, com mínima fragmentação. Em seguida, através do mapeamento das leituras curtas à montagem, é possível corrigir regiões que contenham erros na sequência de nucleotídeos, produzindo um genoma de alta contiguidade e qualidade. Em resumo, após uma montagem extremamente contígua utilizando somente leituras longas, executa-se uma etapa de correção de sequências (polimento) utilizando leituras curtas. Esta etapa de correção é implementada no *pipeline* através do programa Pilon (Walker et al., 2014).

3.2.3 Módulo de anotação Genômica

O módulo de anotação genômica consiste na incorporação e integração de programas de anotação automática e bancos de dados de extrema relevância para bactérias patogênicas. Deste modo, o *pipeline* desempenha as seguintes etapas de anotação:

1. Anotação genérica com o programa Prokka (Seemann, 2014);

2. Predição de sequências de rRNA com Barrnap (`https://github.com/tseemann/barrnap`);
3. Avaliação de MLST ("Multilocus Sequence Typing") com o programa mlst (`https://github.com/tseemann/mlst`);
4. Anotação através da similaridade de sequências com o programa DIAMOND (Buchfink et al., 2014), utilizando bancos de dados específicos:
    - Os bancos de dados de fatores de virulência Victors (Sayers et al., 2019) e VFDB (Chen et al., 2005);
    - O banco de dados de elementos integrativos e conjugativos ICEberg (Liu et al., 2019);
    - O banco de dados de prófagos PHAST (Zhou et al., 2011)
5. Predição de prófagos com Phigaro (Starikova et al., 2019);
6. Anotação de genes de resistência com AMRFinderPlus (Feldgarden et al., 2019) e RGI (Jia et al., 2017);
7. Predição de ilhas genômicas utilizando o programa IslandPath-DIMOB (Bertelli e Brinkman, 2018)

Fora isso, de maneira opcional, os usuários podem realizar: (i) detecção de bases metiladas com o programa Nanopolish; (ii) anotação de pan-genomas com o programa Roary (Page et al., 2015) utilizando genomas de referência pré-selecionados pelo usuário e; (iii) anotação das ortologias KEGG (KO) utilizando o programa KofamScan (Aramaki et al., 2019). A lista de KOs gerada pode ser posteriormente utilizada para a produção de mapas metabólicos no servidor do KEGG (Kanehisa e Goto, 2000).

Após realizadas as anotações, as informações obtidas são armazenadas em um arquivo GFF (do inglês, "General File Format"). Em seguida, produz-se relatórios em formato padrão para navegadores web (HTML) sobre os genes de virulência, resistência, prófagos e elementos móveis encontrados no genoma. Por último, gera-se um navegador genômico através do programa JBrowse (Buels et al., 2016) que permite a visualização da anotação ao longo do genoma.

3.3 Material biológico e técnicas experimentais

3.3.1 Isolados bacterianos

Os isolados da bactéria *Klebsiella pneumoniae* são provenientes de um estudo anterior (de Campos et al., 2018), obtidos do sangue e swab retal (nomeados Kp31 e Kp34, respectivamente) de um paciente após dezessete dias de hospitalização no Hospital universitário da UnB (HUB). Os isolados foram identificados pelo sistema VITEK-2 e o fenótipo de hipermucoviscosidade foi determinado pelo resultado positivo do *string test*.

3.3.2 Sequenciamento de DNA

Os isolados foram submetidos ao sequenciamento de DNA por nanoporos no centro de Biotecnologia Roy J. Carver (Universidade de Illinois em Urbana-Champaign). O preparo da biblioteca foi realizado utilizando o kit rápido de *barcoding* SQK-RBK004. As bibliotecas foram sequenciadas por 48 horas em uma *flowcell* SpotON R9.4.1 RevC FLO-MIN106 usando um sequenciador GridIONx5.

Adicionalmente, em um estudo anterior, o isolado Kp31 já havia sido submetido ao sequenciamento NGS através da plataforma Illumina NextSeq500, com protocolo *paired-end*, produzindo fragmentos de 150 pares de base, utilizando o kit de preparo Nextera XT DNA (Illumina, San Diego, CA, United States).

3.4 Análise computacional de isolados de K. pneumoniae

3.4.1 Pré-processamento dos dados

O *pipeline* de pré-processamento descrito na seção 3.2.1, foi utilizado para remover sequências adaptadoras e aplicar filtros de qualidade (>20) sobre as leituras Illumina. Para este conjunto de dados, os programas Lighter e FLASH não foram executados.

Em contrapartida, as leituras longas foram recebidas já pré-processadas e desmultiplexadas com os parâmetros padrão do programa Porechop. Portanto, estas tiveram apenas suas qualidades e métricas avaliadas através do programa NanoPack.

3.4.2 Montagem genômica

Leituras curtas e longas foram montadas e polidas utilizando o *pipeline* de montagem descrito na seção 3.2.2. Para ambos os isolados (Kp31 e Kp34, seção 3.3.1), produziu-se montagens "*nanopore-only*" (somente utilizando dados de nanoporo) com cada um dos programas: Unicycler, Canu e Flye. Somente para o isolado Kp31, foram produzidas montagens híbridas e "*Illumina-only*" (somente com dados Illumina) utilizando o programa Unicycler. Montagens "*nanopore-only*" do isolado Kp31 foram polidas utilizando o programa Pilon (Walker et al., 2014), como descrito na seção 3.2.2.3.

Para os programas Canu e Flye, que necessitam da indicação do tamanho aproximado do genoma, seus parâmetros foram configurados como *5.6m* pares de base. O parâmetro `--plasmids` também foi executado para executar o modo de recuperação de plasmídeos do programa Flye.

3.4.3 Avaliação das montagens genômica

A completude das montagens foi avaliada pela quantidade de genes essenciais esperados que são encontrados na montagem. Para isto, utilizou-se o programa BUSCO v4.0.2 (Simão et al., 2015) e o banco de dados de genes essenciais da Ordem Enterobacteriales (odb10).

Adicionalmente, devido a alguns problemas ocorridos durante sua análise, os dados do isolado Kp34 foram montados utilizando o módulo metagenômico do programa Flye, com o parâmetro *–meta*. E, executou-se os programas WhatsHap (Martin et al., 2016), HapCUT2 (Edge et al., 2017), Longshot (Edge e Bansal, 2019) e MetaBAT (Kang et al., 2015) com seus parâmetros padrão.

3.4.4 Anotação genômica

A anotação dos genomas foi conduzida pelo *pipeline* de anotação descrito na seção 3.2.3. Os parâmetros não padrão utilizados estão listados na tabela 3.1. Os parâmetros, da etapa de busca por similaridade usando o programa DIAMOND permitem a definição de valores mínimos permitidos de cobertura e similaridade para anotação de genes. Com os valores utilizados, tem-se uma anotação mais restritiva de genes de virulência e mais

permissiva para prófagos e elementos integrativos que podem ter suas regiões menos semelhantes àquelas presentes nos bancos de dados.

Tabela 3.1 - Parâmetros personalizados utilizados durante a execução do *pipeline* de anotação genômica

Parâmetro	Função	Valor
diamond_virulence_identity	Identidade mínima para genes de virulência	90
diamond_virulence_queryCoverage	Cobertura mínima da referência para genes de virulência	90
diamond_MGEs_identity	Identidade mínima para genes de elementos genéticos móveis	85
diamond_MGEs_queryCoverage	Cobertura mínima da referência para genes de elementos genéticos móveis	75

3.4.5 Análises adicionais

O serviço web do programa PlasmidFinder versão 2.0 (Carattoli et al., 2014) foi utilizado para a predição *in silico* de plasmídeos e seus grupos de incompatibilidade. Os sorotipos de polissacarídeos de superfície em *K. pneumoniae* (K e O) foram preditos através do serviço web do programa Kaptive (Wyres et al., 2016; Wick et al., 2018). A caracterização genética e identificação das cepas foi realizada através de avaliações de MLST e cgMLST, realizadas utilizando os serviços BIGSdb[1] and BacWGSTdb (Ruan e Feng, 2015). A identificação de plasmídeos conjugativos foi realizada através do serviço oriTfinder (Li et al., 2018). A similaridade entre genomas foi mensurada em termos de valores ANI ("Average nucleotide identity") obtidos através do programa fastANI v1.2 (Jain et al., 2018).

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

4

Resultados e Discussão

4.1 Criação de pipelines genéricos e modulares

Em termos gerais, as análises de dados de sequenciamento de DNA por NGS envolvem uma série de etapas computacionais que podem ser materializados em um *pipeline*. Existem diversas maneiras pelas quais os programas podem ter sua execução coordenada. Geralmente os desenvolvedores implementam a lógica do processamento que, mais do que a execução de etapas individuais, envolve outros aspectos como verificar a ocorrência de erros, nomear arquivos intermediários e compor a ligação entre os passos do *pipeline*. Os *pipelines* podem ser implementados em qualquer linguagem de programação, mas em bioinformática, as mais empregadas são Python, Bash e PERL. No entanto, estes *pipelines* deixam a desejar em termos de organização lógica, robustez, tolerância a falhas e extensibilidade (Leipzig, 2016).

Isso nos motivou a adotar um arcabouço computacional específico para o desenvolvimento e orquestração de *pipelines*, o Nextflow (Di Tommaso et al., 2017). Através de uma linguagem específica para compor *pipelines*, o Nextflow permite que criemos uma série de tarefas (*tasks*) e que implicitamente estabeleçamos as dependências entre estas, ou seja, a saída de um passo (arquivo ou valor) serve como entrada para o passo seguinte. Com esta rede de dependências o próprio Nextflow coordena a execução de todo o *pipeline* e fornece a lógica de distribuição de tarefas, controle de erros e possibilidade de retornar a execução em caso de erro.

A meta do presente trabalho consiste na criação de um sistema completo de análise de genomas procarióticos, desde o recebimento de dados brutos de sequenciamento NGS

até a geração de relatórios de anotação de diversas características genéticas codificadas no genoma a ser montado. Nestes processos é possível identificar uma série de etapas que são independentes entre si:

(I) Pré-processamento de dados de sequenciamento
(II) Montagem de genomas
(III) Anotação de genomas procarióticos

Diante desta particularidade, optou-se por implementar três *pipelines* distintos, mas associados, como esquematizado na figura 3.1. Esta estratégia de modularização garante que usuários em diferentes cenários (já tenham o genoma montado, ou já possuam as leituras pré-processadas ou partindo do ponto zero) sejam capazes de utilizar os *pipelines* da maneira que melhor atenda suas necessidades. Por exemplo, usuários de maior expertise podem utilizar somente o *pipeline* de anotação a partir de um genoma previamente montado.

Outro aspecto importante no desenho dos *pipelines* foi a utilização de contêineres Docker, que elimina completamente os problemas de instalação e manutenção das diferentes versões de *software*, além de permitir a execução do *pipeline* em qualquer sistema operacional. Todos os programas requeridos para a execução destes *pipelines* foram previamente instaladas em imagens Docker, as quais constituem um arquivo único e que podem ser distribuídos uma única vez. Para tanto, foram criados repositórios no portal github[1] para cada *pipeline* criado neste trabalho:

- fmalmeida/ngs-preprocess – `https://github.com/fmalmeida/ngs-preprocess`
- fmalmeida/MpGAP – `https://github.com/fmalmeida/MpGAP`
- fmalmeida/bacannot – `https://github.com/fmalmeida/bacannot`

Esta individualização torna sua execução transparente e portátil através da estrutura de gerenciamento de *pipelines* da comunidade Nextflow, tornando a execução tão simples quanto:

```
nextflow run [URL do repositório git] [parâmetros]
```

[1] `https://github.com/`

Com este comando, o Nextflow automaticamente descarrega a imagem Docker correspondente e executa o *pipeline*. Somado a isto, os *pipelines* são distribuídos com arquivos de configuração que permitem ao usuário escolher as configurações de montagem e anotação que melhor se ajustem aos seus dados. Por fim, o *pipeline* de anotação genômica gera relatórios completos e "amigáveis" que podem ser eventualmente customizados pelo usuário.

O foco de nossa implementação são usuários com pouca ou nenhuma expertise em bioinformática que desejem realizar uma análise genômica bacteriana de maneira rápida e simples, evitando laboriosas etapas de instalação de programas e suas dependências. Pelo indicado acima, a execução do *pipeline* é dada por linha de comando que, mesmo para usuários sem conhecimento de ambientes computacionais, é facilmente executada em um terminal, sem uso de interface gráfica. Os mesmos comandos podem ser utilizados em sistemas operacionais Linux, Windows e Mac OS, bastando antes instalar o Nextflow no computador de execução. Por fim, em termos de *hardware*, a execução dos *pipelines* pode ser feita em ambientes desde um *laptop* Windows (com memória RAM $>$ 16 Gb) até servidores Linux com inúmeros processadores.

4.1.1 Comparações com outros pipelines

De maneira geral, *pipelines* distribuídos como ferramenta local são mais reprodutíveis e escalonáveis que *pipelines* somente disponibilizados por serviços web. A quantidade de tarefas em paralelo é limitada pela estrutura computacional local, e não por restrições de acesso aos servidores e velocidade de conexão de Internet.

Por serem implementados em Nextflow, nossos *pipelines* garantem um ambiente de execução robusto, extremamente reprodutível e escalonável. Dentre os *pipelines* específicos para genômica bacteriana presentes na tabela 1.2, destacam-se os seguintes: bacass, TORMES, BacPipe e ASA3P. À exceção do programa bacass, todos os outros, apesar de extremamente robustos quanto a anotação e análises de genômica comparativa, são meramente *scripts* em bash, Python ou Groovy. Portanto, estes programas são pouco customizáveis e apresentam pouca tolerância a erros. Por outro lado, o bacass é extremamente simples e sua anotação é menos robusta que a implementada em nosso *pipeline* com

diversos módulos de anotação especializada. Por último, somente nossa implementação e o programa ASA3P são genéricos capazes de realizar o pré-processamento, montagem e anotação de genomas utilizando dados das plataformas Illumina, ONT e PacBio. Desta forma, comparado a estas ferramentas, a utilização conjunta de nossos *pipelines* representa uma alternativa mais robusta, portátil e reprodutível.

4.1.2 Adversidades do uso de pipelines

Contudo, é importante ressaltar que a utilização exclusiva da genômica e anotação automática para a análise de bactérias possui vieses. Anotações automáticas são geralmente preditivas e baseadas em padrões encontrados no genoma de bactérias. Desta maneira, a qualidade dos dados de sequenciamento e da montagem do genoma é extremamente importante pois podem ser pontos de introdução de erros. Características como inserções e deleções, elementos genéticos móveis e regiões homopoliméricas podem representar grandes problemas para algumas tecnologias de sequenciamento. Estas particularidades podem introduzir erros de montagem que consequentemente podem resultar, por exemplo, na anotação errônea de genes truncados ou com mutações "frameshift" (Baptista e Kissinger, 2019).

Além disso, variações de frequência de nucleotídeos entre genomas podem ser características extremamente impactantes para programas de anotação automática que se baseiam em regras. Isto torna a anotação semi-automática, utilizando dados curados, ponto chave para a diminuição de erros em estudos genômicos. Porém, estas análises são extremamente comparativas e dependem da completude e abrangência de bancos de dados, desta forma, é muito comum ignorarmos novos produtos gênicos pois estes são frequentemente anotados como "putativos" ou de função desconhecida (Danchin et al., 2018).

Portanto, embora bastante promissor, a utilização deste tipo de dado requer bastante cuidado e seu desenho experimental deve ser muito bem planejado, levando em conta a procedência da amostra, a biologia de seu genoma e a tecnologia de sequenciamento.

4.2 Estudo de caso: sequenciamento de isolados clínicos de *K. pneumoniae*

Um paciente de idade entre 60 a 70 anos foi submetido a um procedimento de hemicolectomia no Hospital Universitário de Brasília. Quatro dias após alta, o paciente passou a apresentar insuficiência renal aguda, hipertensão arterial sistêmica, flutter atrial e suspeita de choque séptico. O paciente foi transferido de volta à UTI, onde foi submetido à diálise, monitoração e terapias antimicrobianas. Dezoito dias após a hospitalização, o paciente demonstrou deterioração do seu estado de saúde global e eventualmente veio a óbito em função de uma parada cardíaca (de Campos et al., 2018). Culturas bacterianas foram coletadas a partir do sangue (Kp31) e swab retal (Kp34) e têm sido mantidas no laboratório de análises moleculares de Patógenos do departamento de Biologia Celular, UnB.

Provenientes de um estudo anterior, dispõe-se, para o isolado Kp31, de dados brutos de sequenciamento de DNA originados em plataforma Illumina. Este dados foram depositados em 22 de Janeiro de 2018 no banco de dados ENA ("European Nucleotide Archive") sob o número de acesso ERS2166165 (`https://www.ebi.ac.uk/ena/data/view/ERS2166165`) (de Campos et al., 2018). Adicionalmente, foi realizado pelo nosso grupo um sequenciamento NGS de terceira geração com a tecnologia de nanoporos (seção 3.3.2) para as duas cepas. Os dados brutos foram recebidos no dia 27 de Fevereiro de 2019. Na tabela 4.1 encontram-se as estatísticas gerais dos dados brutos utilizados neste estudo.

Tabela 4.1 - Estatísticas dos dados brutos de sequenciamento de DNA

Isolado	Tecnologia	N° de leituras	Total de bases (Gb)	Cobertura aproximada
Kp31	Illumina	1.649.060	0,46	$84X$
Kp31	ONT	172.948	1,45	$255X$
Kp34	ONT	903.250	7,30	$1200X$

4.3 Execução dos *pipelines*

De posse dos dados brutos de sequenciamento, procedeu-se a execução dos *pipelines* desenvolvidos neste trabalho. As análises foram realizadas em um laptop Lenovo, Ideapad 310 i7, Linux, com 20 Gb de memória e 4 núcleos. Na tabela 4.2 estão resumidas as estatísticas de execução de cada módulo para o isolado Kp31. Nesta, apresenta-se somente a montagem híbrida pois é a metodologia mais demorada e que mais demanda recursos computacionais. Portanto, constata-se que, em menos de um dia e em um laptop, um genoma bacteriano pode ser completamente analisado com os *pipelines* desenvolvidos.

Tabela 4.2 - Recursos computacionais utilizados durante a execução dos *pipelines*

Etapa	Tempo		Pico de memória(Gb)
Pré-processamento	8	min	0,5
Montagem híbrida	18,5	h	17,0
Anotação	22	min	0,9

Além disso, o Nextflow produz relatórios automatizados da execução que permitem avaliar o desempenho do *pipeline*. Este relatório automatizado e o arquivo de configuração do *pipeline* são exemplificados nas figuras suplementares A.1 e A.2.

A seguir seguem os detalhes da execução e análise das duas cepas estudadas utilizando os *pipelines*. As seções descrevem as etapas de montagem e anotação contrastando as cepas.

4.4 Montagem dos genomas

4.4.1 Isolado Kp31

Como a cepa isolada de amostra de sangue (Kp31) possui dados de diversas tecnologias NGS, Illumina e nanoporos, foi feito um estudo detalhado sobre qual seria a melhor combinação de estratégias e programas de montagem deste genoma. Isso pôde ser feito pois o *pipeline* desenvolvido neste estudo tem flexibilidade de configuração, o que também é um atributo valioso para sua aplicação em outros projetos de pesquisa.

O genoma deste isolado foi então montado com três combinações dos dados de

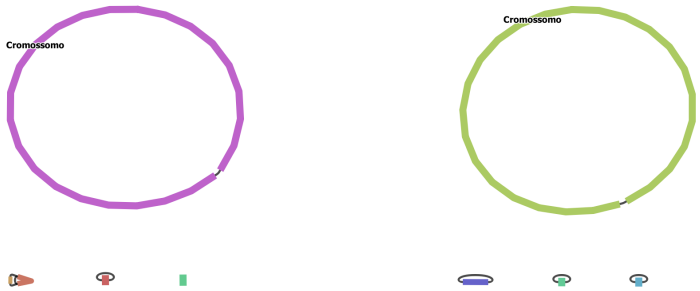
sequenciamento: (i) somente com dados Illumina ("*Illumina-only*"), (ii) somente com dados nanoporo ("*nanopore-only*"), (iii) agregando ambos os dados (Híbrida). No total, foram produzidas 5 montagens diferentes alternando programas e metodologias, cujos resultados encontram-se resumidos na tabela 4.3.

Tabela 4.3 - Estatísticas gerais de todas as montagens de genoma do isolado Kp31

	Programa	Método de Montagem	Número de fragmentos (Contigs)	Tamanho total	N50	GC (%)
(1)	Canu	"*nanopore-only*"	5	5.567.144	5.272.163	57,25
(2)	Unicycler	"*nanopore-only*"	5	5.438.565	5.215.915	56,92
(3)	Flye	"*nanopore-only*"	4	5.490.874	5.274.881	56,90
(4)	Unicycler	Híbrida	73	5.706.735	370.496	57,18
(5)	Unicycler	"*Illumina-only*"	108	5.650.898	203.669	57,20

Para análises genômicas, o valor N50 é uma estatística que representa a qualidade de uma montagem, sendo o valor para o qual, 50% do tamanho total da montagem é representado por contigs maiores ou iguais a este valor. Portanto, quanto maior o N50, mais contíguo é um genoma. Sendo assim, nota-se que, conforme esperado para montagens utilizando somente leituras curtas, obteve-se um genoma bastante fragmentado, com mais de 100 contigs e N50 de ≈200 kb (Tabela 4.3). Isto acontece devido ao tamanho das leituras produzidas pela plataforma Illumina e sua incapacidade intrínseca de resolver completamente estruturas repetitivas.

Por outro lado, a utilização de leituras longas, aumenta significativamente a qualidade das montagens, como observado pela diminuição do número de contigs e aumento do N50 para um valor próximo ao tamanho esperado do genoma de *K. pneumoniae*. De fato, as montagens têm tamanha contiguidade que permitem a identificação, em nível de contigs únicos, da sequência cromossômica e dos possíveis plasmídeos presentes na amostra como observado na Figura 4.1.



(a) (b)

Figura 4.1: Visualização dos grafos de montagem "*nanopore-only*" dos programas Canu (a) e Flye (b). Cada uma das cores representa um único contig de cada montagem. Na imagem, o maior contig circular de cada montagem representa o cromossomo bacteriano e todos os outros pequenos contigs representam possíveis plasmídeos

Na tabela 4.3, percebe-se que, entre as montagens "*nanopore-only*" a produzida pelo programa Flye possui menos contigs que as montagens Canu e Unicycler. Em seu artigo, os desenvolvedores comentam que devido à sua implementação diferenciada dos outros montadores, Flye é capaz de montar genomas com maior acuidade e contiguidade por lidar melhor com regiões repetitivas (Kolmogorov et al., 2019). Para averiguar se realmente esta diferença entre as montagens se dá devido a capacidade de resolução de repetições, as montagens foram primeiramente avaliadas utilizando o programa Quast v5.0.2 (Gurevich et al., 2013) que constatou que todos os contigs das montagens Canu e Unicycler são representados na montagem Flye. Em seguida, foi realizada uma comparação entre as montagens Canu e Flye, usando o programa BLAST (Camacho et al., 2009). O alinhamento é apresentado na figura 4.2, em que são desenhados os cinco contigs da montagem Canu e colore-se os alinhamentos de acordo com os contigs do programa Flye. Assim, observa-se que dois fragmentos da montagem Canu são resolvidos em um único fragmento da montagem Flye (o fragmento identificado pela coloração preta). Isto, de fato sugere que o programa Flye foi capaz de melhor resolver o genoma utilizando somente leituras longas.

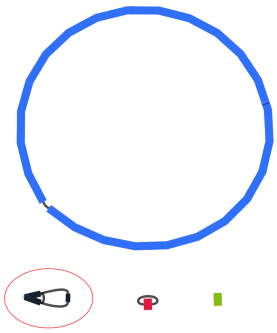


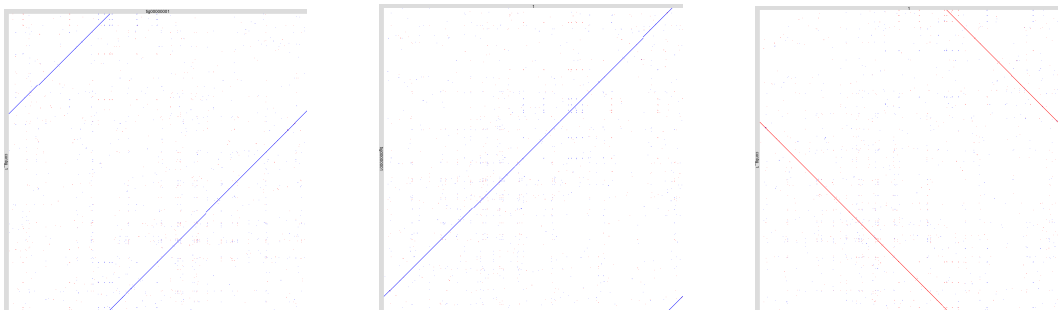
Figura 4.2: Visualização do alinhamento entre as montagens "*nanopore-only*" Flye e Canu. São desenhados os cinco contigs da montagem "*nanopore-only*" Canu. Coloridos, são apresentados os alinhamentos entre os contigs Flye e Canu. Cada uma das cores representa um contig da montagem Flye. Destaca-se no círculo vermelho os dois contigs da montagem Canu que foram resolvidos em um único contig da montagem Flye (identificado pela coloração preta)

Enquanto isso, anomalamente, a montagem híbrida Unicycler (5) mostrou-se extremamente fragmentada (>70 contigs). Devido ao seu algoritmo e resultados divulgados (Wick et al., 2017), esperava-se um cenário de contiguidade semelhante às montagens "*nanopore-only*". Em seu manual, os desenvolvedores comentam que este padrão de resultados pode ser decorrente de dois fatores: (i) quantidade insuficiente de dados Illumina; (ii) leituras de amostras diferentes. Por isso, uma avaliação mais cautelosa destes dados foi realizada (seção 4.4.3).

Adicionalmente, as montagens "*nanopore-only*" foram submetidas a uma etapa de correção de erros utilizando o programa Pilon (Walker et al., 2014). Este programa é capaz de corrigir erros ao nível de bases únicas, utilizando o mapeamento de leituras curtas a montagens "*nanopore-only*". Integrando a alta qualidade, em termos de baixa taxa de erros, das leituras curtas, o Pilon resolve ambiguidades e calcula o nucleotídeo mais provável em cada posição da montagem. Com esta etapa, foram obtidos mais três arquivos de genomas: (i) Kp31 Canu "*nanopore-only*" corrigido, (ii) Kp31 Flye "*nanopore-only*" corrigido, (iii) Kp31 Unicycler "*nanopore-only*" corrigido.

Em bioinformática, "dotplots" são gráficos que permitem a comparação entre duas sequências ao organizar uma em cada eixo. Sempre que se encontra um nucleotídeo idêntico entre as sequências na mesma região do gráfico, desenha-se um ponto. Uma vez que os pontos são desenhados, eles são combinados em linhas. Nota-se que a direção da

sequência no eixo, influência a direção da linha no gráfico. Desta forma, quanto mais idênticas as sequências forem, mais o gráfico se aproxima de uma única linha diagonal. A presença de mutações, inserções e deleções, translocações e repetições afeta o gráfico adicionando linhas adicionais em diferentes configurações. Desta forma, percebe-se na figura 4.3 que existe pouca diferença estrutural global entre os cromossomos das montagens "*nanopore-only*" corrigidas.



(a) (b) (c)

Figura 4.3: Visualização em "dotplot" dos alinhamentos entre os cromossomos das montagens "*nanopore-only*" corrigidas. Representa-se nas figuras: (a) o alinhamento entre as montagens Canu (eixo x) e Flye (eixo y); (b) o alinhamento entre as montagens Unicycler (eixo x) e Canu (eixo y); (c) o alinhamento entre as montagens Unicycler (eixo x) e Flye (eixo y)

A importância da etapa de correção de erros foi investigada através da quantificação das diferenças entre as montagens "*nanopore-only*" e "*nanopore-only*" corrigidas. O efeito global da correção sobre as montagens é descrito na Tabela 4.4, onde se observa que entre $\approx$ 2–4% das bases foram corrigidas. Percebe-se que, dentre as montagens "*nanopore-only*", a realizada pelo programa Canu é a menos afetada pela etapa de correção utilizando Pilon, provavelmente devido à sua etapa intrínseca de correção de leituras (realizada antes da montagem).

Tabela 4.4 - Avaliação global do efeito da etapa de correção de erros no cromossomo usando dados Illumina sobre as montagens "*nanopore-only*". O número total de correções (variantes) foi calculado através do programa snpEff

Montagem	Total de correções
Canu	86,274
Flye	101,688
Unicycler	200,186

Qual seria o efeito da não correção de erros nas etapas posteriores de anotação do genoma? Utilizou-se o programa snpEff (Cingolani et al., 2012) que permite prever o efeito das bases incorretamente nomeadas no contexto de sua consequência funcional dentro de regiões gênicas. Observou-se que para as montagens Canu e Flye, são corrigidos ≈1.400 códons de terminação falsos e ≈35.000 variações "missense", que causam mudança em códons que resultam em alterações de aminoácidos. Para o programa Unicycler, que teve mais bases corrigidas globalmente, o efeito dos erros é mais pronunciado, sendo que ≈6.000 introduzem códons de terminação prematuros e ≈70.000 são variações "missense".

Todos estes resultados sugerem que, considerando montagens somente com leituras longas, os programas Canu e Flye se destacam em comparação ao Unicycler. E, embora as montagens Flye e Canu sejam bastante equiparáveis, adotou-se a montagem Flye como representante das montagens "*nanopore-only*" por ter demonstrado uma aparente melhor capacidade de resolver os contigs. Como um todo, as análises comparativas sobre os efeitos da etapa de correção das montagens "*nanopore-only*" utilizando dados Illumina demonstram a necessidade da utilização conjunta de leituras longas e curtas para permitir a obtenção de um genoma contíguo e com poucos erros.

4.4.2 Isolado Kp34

Por não possuir dados de Illumina, este genoma foi montado utilizando somente dados de nanoporo ("*nanopore-only*") com cada um dos três programas: Canu, Flye e Unicycler. Inesperadamente, os resultados obtidos foram de baixa qualidade, por apresentar dezenas de contigs (Tabela 4.5), ao invés de 4 ou 5 como observado para a cepa Kp31 (Tabela 4.3). Esta grande fragmentação de contigs aliada às montagens com mais de 6 Mb, diante de uma expectativa de genomas na faixa de 4,8 a 5,3 Mb para *K. pneumoniae*, indicam problemas nas amostras enviadas para o sequenciamento, pois as sequências propriamente ditas tiveram boa qualidade de acordo com os diagnósticos obtidos no nosso *pipeline* de pré-processamento (Tabela suplementar A.1).

A alta fragmentação das montagens do isolado Kp34 somado aos seus tamanhos inesperados sugerem que os dados de sequenciamento obtidos não são provenientes de uma única bactéria. Por isso, assumindo a existência de uma amostra não pura, optou-se por

Tabela 4.5 - Estatísticas gerais de todas as montagens de genoma do isolado Kp34

	Programa	Número de fragmentos (Contigs)	Tamanho total	N50	GC (%)
(1)	Canu	47	9.341.367	851.143	57,00
(2)	Flye	34	6.116.267	835.309	56,75
(3)	Unicycler	29	6.286.895	532.490	56,73
(4)	Flye (–meta)	45	6.737.360	375.091	56,30

realizar uma quarta montagem, utilizando a metologia metagenômica do programa Flye (`--meta`), na expectativa de que esta fosse capaz de reconstruir os diferentes genomas presentes na amostra.

Porém, o grafo resultante desta montagem indica que não foi possível separar os genomas. Inclusive, nota-se que os nós são desenhados bem conectados entre si (Figura 4.4). Estes nós, são pequenas sequências nucleotídicas que são produzidas e conectadas pelos programas durante a montagem de um genoma. As linhas entre esses nós, identificam todas as diferentes maneiras possíveis de conectar estas sequências de modo a finalizar o genoma. Desta forma, a identificação de grandes contigs intercalados por bifurcações de pequenos nós, indica a presença de duas bactérias muito similares. Onde os grandes contigs representam as regiões similares entre os dois genomas e os pequenos nós representam as regiões divergentes e bifurcadas devido à dúvida sobre qual o caminho correto a tomar durante a montagem do genoma. Sendo assim, julga-se possível a presença de outra cepa de *Klebsiella pneumoniae* na amostra sequenciada.

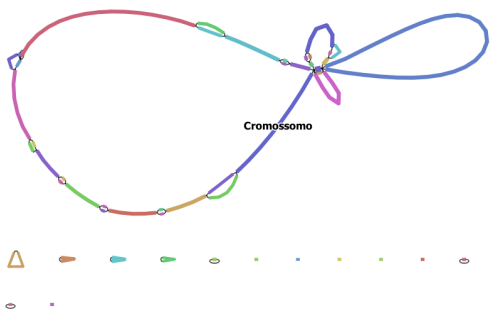


Figura 4.4: Visualização do grafo da montagem Flye (`--meta`) metagenômica do isolado Kp34

Diante desta hipótese, tratou-se esta amostra como um genoma diplóide, em que

cada bactéria representaria um haplótipo. Neste cenário, foram utilizados programas especializados para montagem de haplótipos como WhatsHap (Martin et al., 2016), Longshot (Edge e Bansal, 2019) e HapCUT2 (Edge et al., 2017), visando separar a mistura de genomas. Porém, nenhuma destas tentativas foi bem sucedida. Por último, utilizou-se o programa MetaBAT (Kang et al., 2015), uma ferramenta para a reconstrução de genomas a partir de comunidades microbianas. Todavia, os contigs não foram separados e continuaram agrupados como um único genoma.

Em conclusão, não se obteve uma montagem de boa qualidade que representasse o isolado Kp34, apesar de utilizarmos diversos programas que tiveram sucesso para a cepa Kp31. Diversas hipóteses podem ser aventadas para explicar tal resultado negativo, visto que, diante da utilização de uma tecnologia de sequenciamento de leituras longas e pela cobertura mais que suficiente, era esperada uma montagem de alta qualidade.

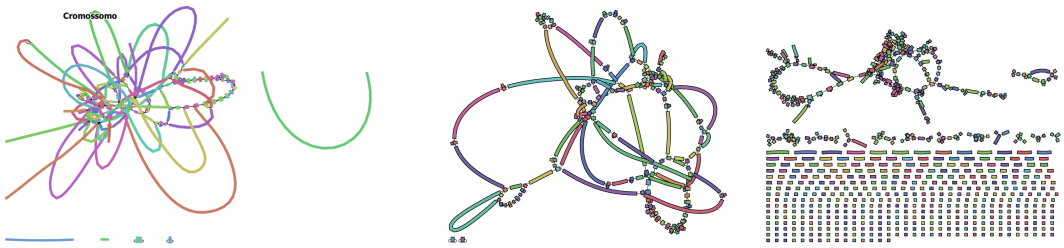
Primeiramente, a cobertura de sequenciamento da cepa Kp34 foi de 1.200 X, enquanto que para Kp31 obteve-se 255 X (tabela 4.1), apesar de teoricamente as duas preparações de DNA terem sido equimolares e utilizar a mesma *flowcell* do dispositivo da Oxford Nanopore. Essa diferença de rendimento pode ser uma das fontes de irregularidade técnica que afetou a qualidade das sequências da Kp34.

Uma outra possibilidade é a contaminação ou identificação incorreta de amostras que foram submetidas à extração de DNA e sequenciamento. Esta contaminação é provavelmente uma cepa de *K. pneumoniae*, pois caso fosse outra espécie provavelmente seria possível separar os genomas, seja com as montagens normais ou com a abordagem metagenômica. Esta visão é reforçada pelo fato de que as montagens resultantes da maioria dos programas possuíam em média 6,3 Mb (Tabela 4.5), superando em ≈1 Mb o tamanho do genoma médio de *K. pneumoniae*. Cepas com genomas muito próximos em termos de sequência não permitem a resolução dos genomas individuais, pois muitas regiões com pequenas diferenças de bases seriam colapsadas pelos algoritmos dos montadores para acomodar a alta taxa de erros de sequenciamento, como observado na Figura 4.4.

As deficiências na montagem desta cepa motivaram uma investigação mais detalhada levando-se em consideração um cenário de que pudesse ter havido erros na identificação de cepas em todos os sequenciamentos realizados, inclusive para a cepa Kp31. Passamos portanto a pormenorizar estes estudos.

4.4.3 *Avaliação da fragmentação da montagem Unicycler híbrida para a cepa Kp31*

Primeiramente, foi avaliada a possível insuficiência de dados Illumina para a montagem híbrida do programa Unicycler. Percebe-se na figura 4.5 que o grafo resultante da montagem Kp31 "*Illumina-only*" não equivale ao considerado ideal pelo programa, mas também não chega a ser desastroso, conforme exemplos presentes no manual do Unicycler. E, mesmo para o cenário considerado desastroso, os desenvolvedores indicam que possivelmente 30 X de leituras longas seriam suficientes para resolver o genoma. Porém, mesmo com ≈255 X de cobertura (tabela 4.1), não foi possível resolver o genoma desta amostra. Portanto, o problema não parece ser decorrente de insuficiência de dados Illumina, mas sim de uma divergência entre as bibliotecas de sequenciamento Illumina e ONT do isolado Kp31.



(a) (b) (c)

Figura 4.5: Comparação entre os grafos da montagem Kp31 "*Illumina-only*" e o recomendado pelo Unicycler. Em (a) é apresentado o grafo resultante da montagem Kp31 "*Illumina-only*", em (b) o grafo considerado ideal pelo Unicycler e em (c) o grafo considerado desastroso

A fim de verificar esta hipótese, as bibliotecas de sequenciamento Illumina e nanoporos da amostra Kp31 foram analisadas conforme apresentado na figura 4.6, utilizando os programas minimap2 v2.16 (r922) (Li, 2018) e bwa v0.7.17 (r1188) (Li, 2013). Desta forma, constatou-se que ≈5% das leituras longas não mapeiam à montagem "*Illumina-only*" e ≈20% das leituras curtas não mapeiam à montagem "*nanopore-only*".

Complementarmente, a comparação destas montagens com o programa Quast indica que cerca de 15% das bases totais não possuem alinhamento entre elas. Esta divergência compreende regiões do cromossomo bacteriano e de sequências de possíveis plasmídeos, como pode ser visto na representação do alinhamento entre estas duas montagens (Figura 4.7) gerado pelo programa mummer (Marçais et al., 2018).

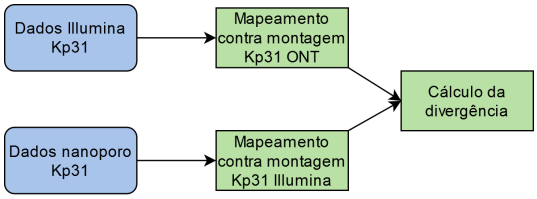


Figura 4.6: Esquematização da análise de divergência entre as bibliotecas de sequenciamento Illumina e nanoporo. As leituras curtas foram mapeadas utilizando o programa bwa. As leituras longas foram mapeadas utilizando o programa minimap2

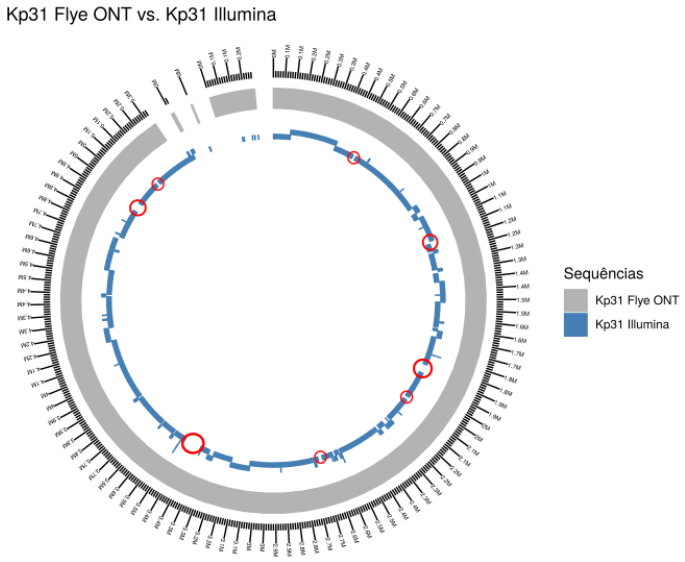


Figura 4.7: Visualização do alinhamento entre as montagens Kp31 Flye "*nanopore-only*" (em cinza) e Unicycler "*Illumina-only*" (em azul). Regiões em círculos vermelho representam partes do cromossomo bacteriano da montagem Flye que não possuem alinhamento. O Alinhamento entre as sequências é representado pelo pareamento entre blocos de cores cinza e azul.

Sabe-se que foram enviadas para sequenciamento as amostras biológicas Kp31 e Kp34. Porém, os resultados sugerem que as bibliotecas de sequenciamento Kp31 ONT e Kp31 Illumina não são provenientes da mesma amostra. Por isso, levanta-se a hipótese de que houve erro na identificação das mesmas e, na verdade, os dados Kp31 ONT são referentes à amostra biológica Kp34 e vice-versa. Além disso, revisitando os resultados dos dados de sequenciamento da cepa Kp34 (seção 4.4.2) e sua visível contaminação, sugere-se que estes dados sejam provenientes do sequenciamento da "mistura" das amostras biológicas Kp31 e Kp34.

4.4.4 *Avaliação da possível contaminação*

Para checar esta hipótese, as montagens Kp31 "*Illumina-only*" e "*nanopore-only*" foram mapeadas à montagem Kp34 Flye metagenômica, utilizando o programa minimap2. Assim, identificou-se que ≈98,54% das bases totais da montagem Kp31 "*nanopore-only*" e ≈99,83% das bases totais da montagem Kp31 "*Illumina-only*" mapeiam à montagem Kp34. Portanto, estes resultados sugerem que o conjunto de dados de sequenciamento Kp34 realmente seja uma "mistura" entre as amostras. Por último, como representado na figura 4.8, a partir dos resultados do mapeamento, produziu-se dois novos subgrupos do conjunto de dados de sequenciamento Kp34 original: (i) com leituras longas mais relacionadas aos dados de sequenciamento Kp31 Illumina, denominado Kp34/31-illumina (ii) com leituras longas mais relacionadas aos dados de sequenciamento Kp31 ONT, denominado Kp34/31-ont.

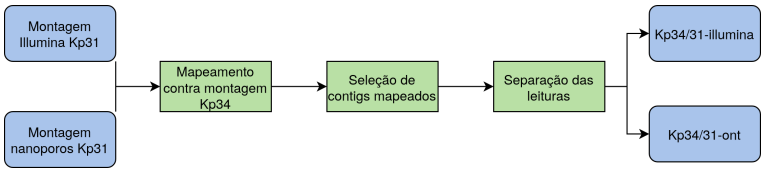


Figura 4.8: Esquematização da estratégia utilizada para dividir os dados de sequenciamento Kp34 em dois subgrupos através do mapeamento dos dados Kp31 Illumina e ONT. Primeiramente, foram identificados na montagem Kp34 metagenômica, quais contigs eram mapeados à cada montagem Kp31. Em seguida, utilizando esses contigs mapeados, foram selecionadas as leituras da biblioteca de sequenciamento Kp34 que produziam estes contigs. E então, os dois subgrupos de leituras foram produzidos.

4.4.5 *Montagem dos dados "descontaminados"*

Realizou-se mais duas montagens utilizando os conjuntos de dados "descontaminados" Kp34/31-illumina e Kp34/31-ont. Os dados Kp34/31-illumina foram montados de forma híbrida pelo programa Unicycler, utilizando os dados de sequenciamento Illumina da Kp31. Já os dados Kp34/31-ont foram montados pelo programa Flye somente com dados disponíveis de nanoporos.

Interessantemente, como pode ser observado na tabela 4.6, somente foi possível montar um genoma de alta qualidade do conjunto de dados Kp34/31-illumina, enquanto a montagem obtida dos dados Kp34/31-ont é bastante fragmentada e com baixo N50. Isto,

talvez seja reflexo da falta de dados extremamente acurados, como os dados Kp31 Illumina, para a separação dos dados de sequenciamento Kp34 (seção 4.4.4) e, consequentemente, os dados Kp34/31-ont acabaram reeditando os problemas da contaminação da amostra. E, embora de posse de uma ótima montagem obtida com os dados de nanoporos da Kp31 utilizando o programa Flye (seção 4.4.1), decidimos não assumir que esta represente de fato a cepa Kp34 pois não possuímos dados de qualidade e confiáveis desta cepa para comparação.

Tabela 4.6 - Estatísticas gerais das montagens "descontaminadas"

Conjunto de dados	Número de fragmentos (Contigs)	Tamanho total	N50	GC (%)
Kp34/31-illumina (híbrido)	7	5.781.326	5.431.784	57,08
Kp34/31-ont	38	5.968.136	1.319.789	57,23

Por fim, as montagens Kp34/31-illumina e Kp31 Unicycler "*Illumina-only*" foram comparadas utilizando o programa Quast. Esta análise mostrou que existe somente uma diferença de 24 kb entre as montagens (0,05% das bases totais), possivelmente devido à resolução de regiões repetitivas através da incorporação de leituras longas. Estes resultados sugerem que estas duas montagens são de uma mesma amostra. Sendo assim, as análises seguintes foram realizadas somente para a montagem híbrida utilizando os dados Kp34/31-illumina, representando a amostra biológica Kp31.

4.4.6 Completude da montagem Kp31 híbrida

Em bioinformática, é possível avaliar a completude de uma montagem genômica utilizando o programa BUSCO (Simão et al., 2015). Este programa, avalia a montagem em termos de conteúdo esperado de genes. Os conjuntos de genes esperados, denominados categorias BUSCO, são determinados baseados na informação evolutiva de genes ortólogos cópia única quase universais. Os resultados desta avaliação de completude são resumidos na tabela 4.7 e sugerem que a montagem híbrida do isolado Kp31 é de alta qualidade, com contiguidade (7 contigs) e completude satisfatória, apresentando 434/440 categorias BUSCO completas.

Tabela 4.7 - Avaliação de completude da montagem Kp31 híbrida (Kp34/31-illumina) utilizando o program BUSCO

	Categorias BUSCO					
	Total	Completos	Completos cópia única	Completos duplicados	Fragmentados	Faltando
Kp31 (híbrido)	440	434	432	2	1	5
GCA_000240185.2 (referência)	440	431	429	2	3	6

4.4.7 Predição in silico de plasmídeos

Depois de avaliada a montagem e definida a sequência cromossômica, voltou-se as atenções para a caracterização dos potenciais plasmídeos, representados pelos pequenos contigs da montagem Kp34/31-illumina. Esta avaliação foi realizada *in silico*, utilizando o programa PlasmidFinder para anotar estas sequências. Esta predição resultou na identificação de 7 replicons: ColRNAI, Col440I, IncA/C2, IncFIB(pKPHS1), IncFII(pCRY), IncN e IncR. Sendo que, dois destes foram identificados no mesmo contig. As posições e identidade dos plasmídeos anotados se encontram na tabela 4.8.

Tabela 4.8 - Resultado da predição *in silico* de plasmídeos do programa PlasmidFinder. A posição genômica das predições são apresentadas no formato *Contig:início..fim*.

Plasmídeo	Identidade (%)	Posição genômica
Contig 2 (116,1 kb)		
IncA/C2	98,56	2:45..461
IncN	99,42	2:29030..29543
Contig 3 (108,4 kb)		
IncFIB(pKPHS1)	95,48	3:1..440
Contig 4 (77,6 kb)		
IncFII(pCRY)	81,32	4:27234..27779
Contig 5 (34,4 kb)		
IncR	100,00	5:7118..7368
Contig 6 (9,2 kb)		
ColRNAI	100,00	6:666..795
Contig 7 (3,6 kb)		
Col440I	94,55	7:2562..2671

Estes plasmídeos foram avaliados através do serviço web oriTfinder (Li et al., 2018) para a rápida identificação de origens de transferência (oriT). Dentre os plasmídeos da Kp31, somente o contig 2 apresenta características de plasmídeos conjugativos, possuindo oriT, relaxase e os "clusters" gênicos do sistema de secreção tipo IV (T4SS) e de proteínas de acoplamento tipo IV (T4CP).

Adicionalmente, foram descarregados todos os plasmídeos de *Klebsiella pneumoniae* depositados no banco de dados NCBI e foi realizada uma busca por similaridade utilizando o programa BLAST para comparar estas sequências. Estes resultados são apresentados na tabela 4.9.

Tabela 4.9 - Análise de similaridade entre plasmídeos da cepa Kp31 e plasmídeos depositados no NCBI.

Contig	*Acession* NCBI	Tamanho total (referência)	Tamanho do alinhamento
2	NZ_KX276209.1	54.609	35.035
3	NZ_CP036373.1	108.291	38.520
4	NZ_CP035205.1	58.460	28.525
5	NZ_CP044378.1	149.953	19.924
6	NZ_CP033630.1	23.185	8.161
7	NZ_CP034764.1	3.674	2.879

Não foi identificado nenhum alinhamento total entre as sequências da Kp31 e aquelas depositadas no NCBI. Desta forma, os resultados sugerem que a cepa Kp31 possui plasmídeos ainda não depositados no banco de dados NCBI. Somado a isso, as sequências da Kp31 não alinham expressivamente à mesma referência, sugerindo não haver "misassemblies" na montagem destes plasmídeos. Estes grandes alinhamentos de >20 kb sugerem que alguns dos plasmídeos identificadas na Kp31 tenham surgido da recombinação com plasmídeos já conhecidos. De fato, são identificadas transposases nas bordas dos alinhamentos dos contigs 2 e 5, corroborando com a hipótese de ter ocorrido um evento de recombinação entre plasmídeos.

Todas estas análises permitem concluir que, em comparação a primeira montagem híbrida realizada para o isolado Kp31, obtém-se, após todas as etapas de correção dos problemas das bibliotecas de sequenciamento, um genoma de qualidade satisfatória (Tabela 4.10). Observa-se a transição de uma montagem fragmentada, com mais de 70 contigs,

para uma montagem bastante contígua de sete fragmentos, sendo um cromossomo e seis plasmídeos.

Tabela 4.10 - Comparação entre as duas montagens híbridas da amostra Kp31. Compara-se os resultados da primeira montagem híbrida da amostra Kp31 e os resultados da segunda montagem híbrida após a reanálise da procedência dos dados.

Montagem		Número de fragmentos (Contigs)	Tamanho total	N50	GC (%)
(1) Kp31	Unicycler Híbrida	73	5.706.735	370.496	57,18
(2) Kp31	Unicycler Híbrida	7	5.781.326	5.431.784	57,08

Contudo, apesar dos esforços, não foram produzidas montagens de qualidade suficiente que pudessem representar o isolado Kp34. Portanto, as análises seguintes de caracterização funcional das sequências através da anotação gênica foram desempenhadas somente para o isolado Kp31, utilizando a segunda montagem híbrida obtida após a reanálise dos dados de sequenciamento.

4.5 Anotação genômica

De posse do genoma completo da Kp31 obtido pela depuração dos dados e protocolos, passou-se para uma análise exploratória das características genéticas e fisiológicas desta cepa.

4.5.1 Contextualização filogenética da Kp31

Sequências genômicas raramente recombinam entre espécies e geralmente apresentam ≈3-4% de divergência nucleotídica entre espécies, desta forma, comparações de indentidade média entre genomas é uma ferramenta bastante útil para a identificação de espécies e cepas (Wyres e Holt, 2016). A população de *Klebsiella pneumoniae* apresenta, em média, uma divergência nucleotídica de 0,5% entre linhagens, indicando que a estrutura populacional como um todo é relativamente clonal (Wyres e Holt, 2016).

Por isso, para realizar a contextualização filogenética desta cepa, foi feita uma análise de genômica comparativa para verificar a similaridade de sequência com 7,688 genomas de *Klebsiella pneumoniae* depositados no NCBI. Esta análise foi feita através do

cálculo de valores ANI ("Average Nucleotide Identity") entre a montagem híbrida Kp31 e os diversos genomas coletados do NCBI, sendo que valores podem variar de 0 a 100%. Os cinco genomas mais próximos à Kp31, identificados pelos maiores valores ANI calculados pelo programa fastANI v1.2 (Jain et al., 2018), são apresentadas na tabela 4.11.

Tabela 4.11 - Genomas identificados como mais próximos filogeneticamente à cepa Kp31 a partir da análise de valores ANI

Código	*Accession* NCBI	ANI	Isolamento		
			Local	Tecido	Data
Kp11_BRA2	GCF_900322605.1	99,90	Brasil (MT)	-	2010
Kp3018	GCF_900322685.1	99,84	Japão	-	2012
Kp11_BRA8	GCF_900322665.1	99,83	Brasil (MT)	-	2006
B35	GCF_002300765.2	99,82	Brasil (MG)	Sangue	2016
B04	GCF_002295145.1	99,81	Brasil (PE)	Sangue	2009

Este nível de identidade é bastante alto e sugere uma distribuição não muito ampla da cepa pois, apesar de um dos genomas ter sido isolado no Japão, a maioria foi isolada no Brasil, o que sugere uma ampla disseminação deste clone no país.

Além do cromossomo, decidiu-se realizar uma análise exploratória quanto a distribuição dos plasmídeos identificados na Kp31. Seriam eles os mesmos identificados nestas outras cepas?

Para isto, foram selecionados os seis pequenos contigs da montagem Kp31 e estes foram alinhados contra estas cinco cepas. Nesta análise, identificou-se que a cepa Kp11_BRA2 parece também possuir todos os plasmídeos identificados na Kp31. As outras cepas, por sua vez, parecem compartilhar somente regiões de alguns plasmídeos, principalmente dos contigs 6 e 7.

Todos estes cinco genomas foram produzidos utilizando somente leituras curtas e, por isso, são bastante fragmentados e não possuem a identificação de quais sequências são cromossômicas ou de plasmídeos. Porém, devido a utilização conjunta de leituras curtas e longas, as sequências do genoma da Kp31 são identificadas e separados em cromossomo e plasmídeos. Isto demonstra a importância da incorporação de leituras longas para a resolução de plasmídeos e consequente identificação de seu conteúdo gênico.

4.5.2 Visão geral da anotação gênica

O *pipeline* de anotação desenvolvido realiza a predição e anotação geral de genes de maneira fácil e rápida utilizando o programa Prokka. Este programa é bastante utilizado para a anotação rápida de genomas procarióticos, e é capaz de predizer e anotar sequências codificadoras e não codificadoras como rRNAs e tRNAs. No geral, foram identificados no genoma da Kp31 5.511 genes, valor próximo ao esperado de ≈5.500 para *K. pneumoniae* (Wyres e Holt, 2016). Dentre estes, são anotadas 33 sequências de rRNA, 167 sequências de tRNA e 1.339 proteínas hipotéticas.

4.5.3 Identificação do "Sequence Typing" e do antígeno capsular

A caracterização do "Multilocus sequence typing" (MLST) para a cepa Kp31, a partir de seu genoma completo, foi feita utilizando o banco de dados BIGSdb Pasteur[2], que a classificou de maneira inequívoca como sendo do tipo ST11. Um esquema mais abrangente de MLST, utilizando de 694 genes essenciais, chamado de "core gene Multilocus sequence typing" (cgMLST), permite a alta resolução de STs e sua agregação em grupos clonais (Wyres e Holt, 2016). A análise de cgMLST realizada através da ferramenta BacWGSTdb (Ruan e Feng, 2015) também classificou a cepa Kp31 como ST11. Adicionalmente, esta análise identificou o isolado Kp11_BRA2 como o mais próximo da Kp31, o mesmo isolado identificado pela análise ANI (4.11), o que demonstra a concordância entre estes métodos.

Complementarmente, *Klebsiella pneumoniae* são geralmente classificadas e categorizadas de acordo com diferentes alelos do antígeno capsular *wzi*, um gene extremamente conservado em *K. pneumoniae* produtoras de cápsula (Brisse et al., 2013). De acordo com a variante alélica deste gene, os isolados são classificados em grupos K. A análise deste locus foi realizada através da ferramenta web do programa Kaptive[3] que identificou na Kp31 o *locus* capsular (KL) 64 e o antígeno capsular *wzi* 64. Portanto, corroborando com a revisão de Catalán-Nájera (Catalán-Nájera et al., 2017), relata-se neste estudo outra *Klebsiella pneumoniae* hipermucoviscosa não K1/K2.

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

[3] `https://kaptive-web.erc.monash.edu/`

Linhagens de *Klebsiella pneumoniae* ST11/CG258 são considerados clones de alto risco, frequentemente associados a epidemias hospitalares em todo o mundo e geralmente apresentando grande diversidade quanto a incorporação e carreamento de genes de virulência e resistência (Kim e Ko, 2019; Jia et al., 2019; van Dorp et al., 2019; Wang et al., 2018). No Brasil, particularmente, este é o grupo clonal mais prevalente e mais disseminado. Clones ST11, geralmente apresentam maiores níveis de resistência que clones não ST11 e frequentemente, surgem relatos da identificação de clones ST11 multirresistentes e até mesmo pan-resistentes a antibióticos (Wang et al., 2018; Fuga et al., 2020; Jia et al., 2019). Este caso chinês é bastante interessante pois assemelha-se ao caso da Kp31, uma *K. pneumoniae* pan-resistente hipermucoviscosa ST11, KL64, isolado de hemocultura de uma paciente hospitalizada de 66 anos, demonstrando a virulência destes clones (Jia et al., 2019).

4.5.4 Avaliação de resistência in silico

O *pipeline* de anotação permite a identificação automatizada de genes de resistência através dos programas AMRFinderPlus (Feldgarden et al., 2019) e RGI (Jia et al., 2017). Estes programas, desenvolvidos pelos bancos de dados NCBI e CARD (Jia et al., 2017), respectivamente, realizam a anotação de genes de resistência baseando-se em informações de homologia e SNPs a genes de seus bancos de dados. Foram identificados genes de resistência a cinco classes diferentes de antibióticos: beta-lactâmicos, fosfomicinas, tetraciclinas, trimetropina e sulfonamidas (Figura 4.9). Este resultado permite a classificação desta cepa como MDR, de acordo com a definição de Magiorakos (2012). Além disso, foram identificados também complexos de efluxo de drogas que podem conferir diferentes níveis de tolerância a múltiplas drogas, dependendo do seu nível de expressão (Du et al., 2018).

Plasmídeos, são moléculas consideradas acessório que podem ser transferidas lateralmente de uma bactéria a outra através do processo de conjugação (Gillings, 2017). A utilização de leituras longas permite a resolução destas sequências e sua diferenciação de sequências cromossômicas (Tabela 4.8). Desta forma, é possível identificar genes de resistência que são carreados em plasmídeos pela Kp31 (Tabela 4.12).

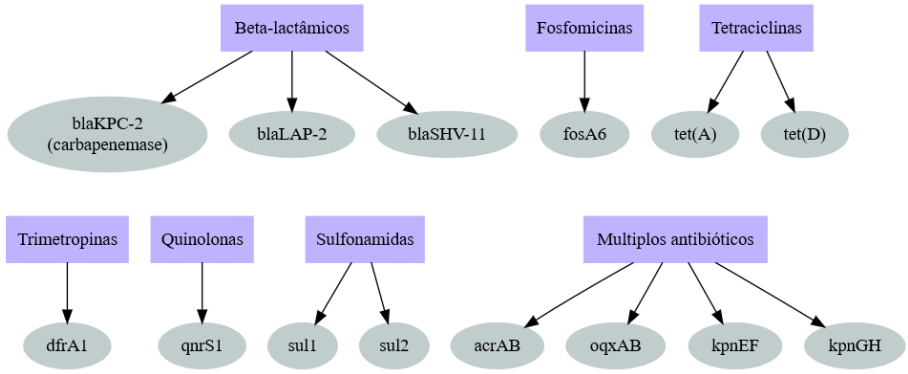


Figura 4.9: Genes de resistência identificados no genoma da Kp31. Na figura são representados os genes de resistência identificados (em oval) e suas classes de antibióticos alvo (em retângulos).

Tabela 4.12 - Genes de resistência do isolado Kp31 identificados em plasmídeos

Plasmídeo	Genes de resistência
IncA/C2 - IncN	*tet(D)*, $bla_{KPC\text{-}2}$, *sul2*
IncFII(pCRY)	*tet(A)*, *qnrS1*, $bla_{LAP\text{-}2}$ *sul1*, *dfrA1*
IncR	*qnrS1* e $bla_{LAP\text{-}2}$

A presença de genes de resistência em plasmídeos permite a rápida variação nos perfis de resistência a antibióticos de bactérias através da perda e ganho de plasmídeos e de eventos de recombinação entre plasmídeos (Gillings, 2017; van Dorp et al., 2019). Estas características podem, por exemplo, ser cruciais para o surgimento de epidemias hospitalares e, desta forma, representam uma grande ameaça à saúde pública (van Dorp et al., 2019). Interessantemente a análise de cgMLST realizada através do banco de dados BacWGSTdb (seção 4.5.3) permite identificar que além dos genes codificados no cromossomo ($bla_{SHV\text{-}11}$ e *oqxAB*), salvo o gene $bla_{LAP\text{-}2}$, todos os genes de resistência identificados em plasmídeos neste estudo, são também identificados no isolado Kp11_BRA2, isolado detectado como mais próximo filogeneticamente da cepa Kp31. O que demonstra a importância de plasmídeos na movimentação e aquisição de características adaptativas.

O gene $bla_{KPC\text{-}2}$ pode ser identificado em diversos contextos genômicos, mas é geralmente carreado em plasmídeos, como identificado na Kp31 (Tabela 4.12). Por exemplo, no Brasil, foi relatado o surgimento, em uma *K. pneumoniae* ST11, de um pequeno plasmídeo IncX3 de 12 kb carreando o gene $bla_{KPC\text{-}2}$ (Fuga et al., 2020). Plasmídeos IncX3 são altamente estáveis e com pouquíssimo custo adaptativo (Liakopoulos et al., 2018). E,

apesar de não conjugativo, o o evento de perda e diminuição no tamanho do plasmídeo relatado neste estudo pode ser relacionado a uma estratégia de disseminação com poucos custos adaptativos, levantando a discussão sobre a importância de pequenos plasmídeos no desenvolvimento de fenótipos de resistência e sua possível mobilização através da recombinação com plasmídeos conjugativos (Fuga et al., 2020). Por outro lado, um estudo recente de Feng e colaboradores (2019) é importante para a demonstração do poder da utilização conjunta de leituras curtas e longas para a resolução plasmídeos com alta qualidade. Neste estudo, os autores relatam uma *K. pneumoniae* ST11 resistente a carbapenêmicos na qual foi identificado três cópias do gene $bla_{KPC\text{-}2}$ em um único plasmídeo, o que seria virtualmente impossível caso somente leituras curtas fossem utilizadas.

4.5.5 Confrontando evidências de resistência in silico com dados experimentais

A cepa Kp31 é proveniente de um outro estudo do grupo da UnB e, por isso, possui-se dados de fenótipos de resistência obtidos experimentalmente (Gonçalves, 2018). Desta forma, torna-se possível comparar os resultados preditivos da genômica aos dados experimentais (Figura 4.10). Nesta figura, as predições computacionais definidas como "Tolerante" são resultantes da anotação das bombas de efluxo que, como mencionado, podem conferir diferentes níveis de tolerância a antibióticos, dependendo de seus níveis de expressão. Esta comparação permite avaliar a consistência da identificação computacional de genes de resistência e sua utilização para a inferência de fenótipos.

Para as 13 classes de antibióticos testados experimentalmente, percebe-se na figura 4.10 que a predição genômica foi capaz de identificar corretamente 12 fenótipos, sendo 9 destes baseados na avaliação direta de presença/ausência de genes, e 3 baseados na presença de bombas de efluxo. Para a única predição divergente, no caso dos Aminoglicosídeos, esta é proveniente da predição baseada na identificação de complexos de efluxo de drogas. Complexos de efluxo de drogas não são fatores capazes de determinar fenótipos de resistência a antibióticos, mas sim diferentes níveis de tolerância (Du et al., 2018). Portanto, esta divergência não pode ser considerado um erro da genômica preditiva pois é algo fora de seu escopo, uma vez que ela é baseada na presença e ausência de genes e é incapaz de predizer concentrações mínimas inibitórias (MIC, do inglês "Minimum inhibitory

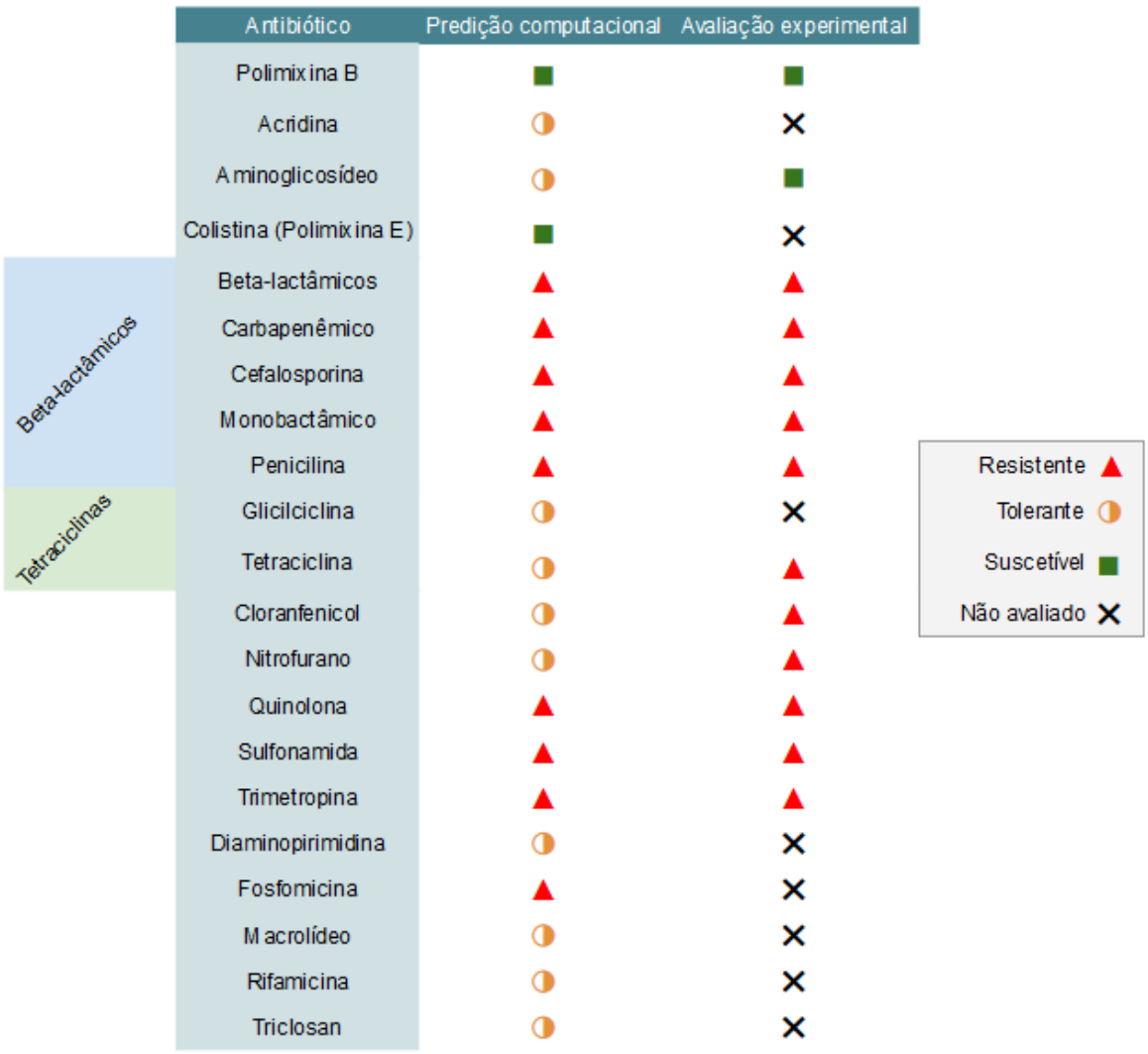


Figura 4.10: Avaliação da predição de resistência do isolado Kp31. As predições computacionais de resistência foram comparadas a resultados obtidos experimentalmente, para 13 classes de antibióticos.

concentration") e, desta forma, não é capaz de predizer níveis de tolerância. Portanto, estes resultados demonstram que a predição e anotação de genes de resistência a partir de sequências genômicas é uma metodologia viável e concisa para genes determinantes de resistência como genes de inativação de antibióticos (seção 1.3).

4.5.6 Avaliação de Virulência

O *pipeline* de anotação, executa automaticamente, a anotação por similaridade de genes de virulência utilizando os bancos de dados VFDB (Chen et al., 2005) e Victors (Sayers et al., 2019). Nesta análise, todos os genes de virulência anotados foram

identificados no cromossomo bacteriano (Tabela 4.13).

Tabela 4.13 - Fatores de virulência identificados no genoma do isolado Kp31

Fator de Virulência	Função
acrAB	Bomba de efluxo de peptídeos antimicrobianos
Cápsula (KL 64)	Evasão imune
rcsAB	Regulação da síntese de polissacarídeos capsulares
LPS (O2v1)	Proteção
Enterobactina (*ent*)	Aquisição de ferro
Yersiniabactina (*ybt*)	Aquisição de ferro
Fímbria tipo I (*fimA*)	Adesão e formação de biofilme
Fímbria tipo 3 (*mrkA*)	Formação de biofilme
T6SS	Sistema de secreção

Um estudo recente na China, investigou 1052 cepas de *Klebsiella pneumoniae* resistentes a carbapenêmicos e identificou que 80% destas bactérias eram cepas ST11, o que demonstra o sucesso adaptativo destes clones e a sua prevalência mundial (Zhang et al., 2019). Além disso, foi constatado que a taxa de carreamento de genes de virulência são maiores em clones K64. Desta forma, este estudo levanta a preocupação quanto ao rápido surgimento de cepas ST11-K64 resistentes e virulentas. Da mesma forma, no presente estudo, identifica-se uma *K. pneumoniae* (Kp31) ST11-K64 MDR, resistente a carbapenêmicos e hipermucoviscosa carreando diversos genes de virulência.

Embora hipermucoviscosa, não foram identificados os genes *rmpA/A2*, geralmente associados a fenótipos de hipermucoviscosidade. Todavia, encontra-se no genoma da Kp31 os ativadores transcricionais da biossíntese de polissacarídeo capsular RcsAB que talvez possam ser fatores chave para o desenvolvimento do fenótipo de hipermucoviscosidade neste isolado (Stout et al., 1991; Su et al., 2018). Bactérias hipermucoviscosas sem a presença dos genes *rmpA/A2* já foram descritas anteriormente (Fang et al., 2004; Cubero et al., 2016; Zhang et al., 2019). Portanto, é óbvio que pouco ainda se sabe acerca deste fenótipo e não exclui a possibilidade de que os genes *rcsAB*, de alguma forma, também sejam cruciais para o desenvolvimento de hipercápsula.

Este isolado não possui nenhum dos biomarcadores atuais da hipervirulência (*iuc*, *rmpA* e *rmpA2*) e, em concordância com revisões recentes, identificamos outra *K. pneumoniae* hipermucoviscosa não hipervirulenta (Catalán-Nájera et al., 2017; Russo e Marr, 2019). Somado a isto, um estudo recente de Zhang e colaboradores 2019, também divulga resultados que fortalecem essa ideia. No estudo, foi identificado que apenas 8% das cepas hipervirulentas apresentavam fenótipo de hipermucoviscosidade, indicando que estes sejam fenótipos diferentes e, desta forma, não podem ser utilizados como sinônimos. Sendo assim, como já abordado por Harada e Doia (2018), uma aliança global ainda necessita ser feita para proporcionar uma definição consensual mais acurada sobre o fenótipo de hipervirulência, em função de facilitar e possibilitar a produção de testes diagnósticos mais eficazes.

Genes de biossíntese de yersiniabactina, particularmente associados a infecções invasivas, são geralmente codificados em um elemento genético móvel chamado ICEKp, o elemento de virulência mais comum em *K. pneumoniae*, e fornece uma via de disseminação de fatores de virulência (Lam et al., 2018). Interessantemente, apesar de não ter sido identificado o elemento ICEkp no genoma da Kp31, os "clusters" gênicos de biossíntese de yersiniabactina e T6SS foram identificados em possíveis ilhas genômicas anotadas pelo programa IslandPath-DIMOB, o que sugere uma maior mobilidade e maior capacidade de transferência lateral destes genes, o que pode culminar na disseminação destes fatores de virulência e consequente surgimento de cepas de altamente virulentas.

Por fim, apesar de não ser hipervirulenta, este isolado reedita o alerta sobre a convergência de fatores de virulência e resistência em uma *K. pneumoniae* hipermucoviscosa, fenótipo considerado de alto risco. Este resultado reforça a necessidade do estabelecimento de programas de vigilância de bactérias patogênicas para o delineamento de medidas de contenção efetivas.

4.6 Relatórios automatizados de anotação

Uma das características mais interessantes do *pipeline* de anotação desenvolvido neste trabalho é a produção de material visual e interativo que permitem a fácil interpretação e exploração dos resultados de anotação. Nesta primeira versão, produz-se

automaticamente pelo *pipeline*: (i) relatórios amigáveis sobre a anotação de genes de virulência, resistência e elementos genéticos móveis, (ii) interface interativa de navegação genômica produzida através do programa JBrowse (Buels et al., 2016).

Os relatórios de anotação são desenvolvidos em formato padrão para navegadores web (HTML) e podem ser visualizados em qualquer navegador web. Estes documentos possuem tabelas e figuras que de maneira clara e coesa apresentam os resultados destas etapas de anotação específica. Todos os genes anotados nos relatórios são diretamente ligados aos seus bancos de dados através de *links*. Além disso, é brevemente informado no texto a metodologia utilizada em cada uma destas etapas. Estas características podem ser observadas na figura 4.11 onde é apresentado pequenas passagens do relatório de anotação de fatores de virulência.

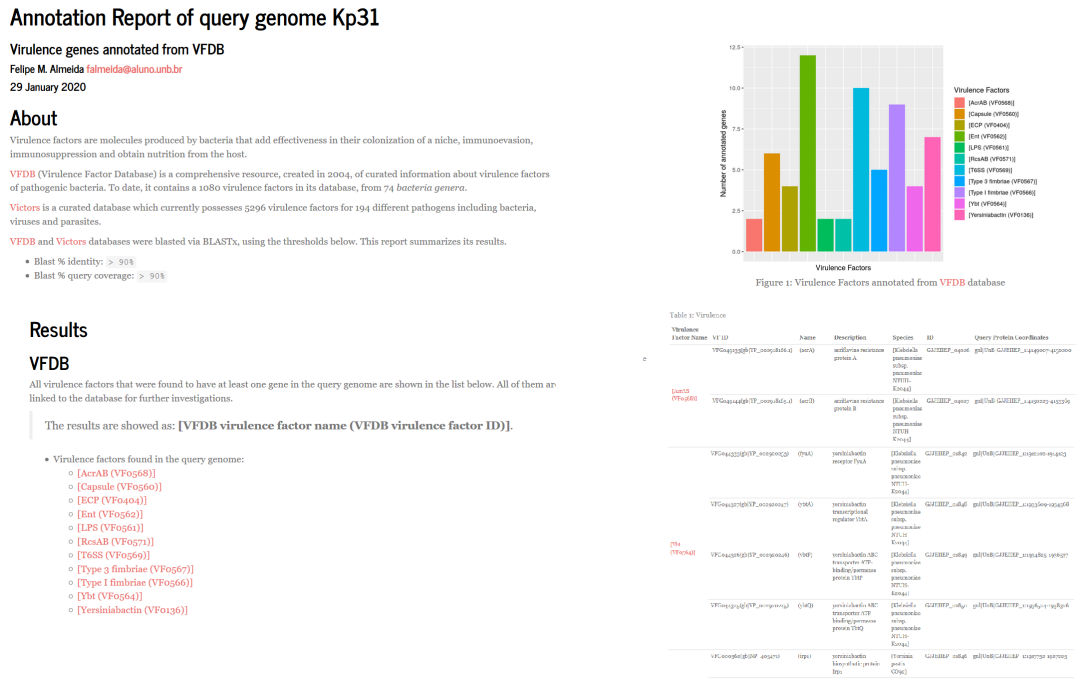
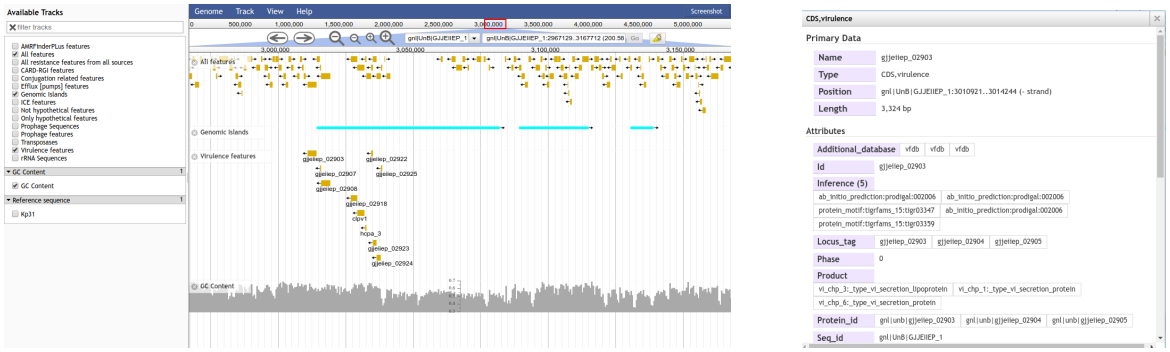


Figura 4.11: Colagem de trechos do relatório automatizado de virulência. Na imagem são mostrados as principais características e blocos informativos presentes nos relatórios automáticos. Todas as informações são explicadas e referenciadas aos bancos de dados.

O navegador genômico é uma ferramenta poderosa e extremamente informativa. Através dela, é possível explorar a anotação de genes ao longo do genoma de maneira visual e interativa. Neste *pipeline*, a produção deste material é desempenhado através da ferramenta JBrowse (Buels et al., 2016). Observa-se na figura 4.12 (a), que este navegador

genômico possui no seu menu lateral diversas faixas que permitem a visualização seletiva de genes divididas por categorias funcionais como rRNAs, tRNAs, genes de virulência, resistência, prófagos, ilhas genômicas, etc. Nota-se na figura 4.12 (b), que esta ferramenta permite, ao clicar em um dos genes, a visualização de todas as informações obtidas pelo *pipeline* sobre aquele gene em particular.



(a) *(b)*

Figura 4.12: Captura de tela do navegador genômico desenvolvido com a ferramenta JBrowse. Na imagem (a) é apresentada a interface principal do navegador. Em seu menu lateral, este navegador possui faixas que permitem a visualização seletiva de diferentes informações e categorias gênicas. Na imagem (b) é apresentada as informações gênicas contidas na ferramenta que podem ser visualizadas ao clicar em qualquer gene.

Estes exemplos demonstram o grande potencial do *pipeline* para rápida anotação e análise de um genoma bacteriano, provendo usuários com relatórios e ferramentas informativas que facilitam a interpretação e investigação dos dados.

5 Conclusões

Demonstrou-se neste trabalho o desenvolvimento de protocolos computacionais aplicados à análise de genomas bacterianos a partir de dados de sequenciamento de próxima geração de DNA.

Optou-se por uma modularização dos *pipelines* para tornar independentes as etapas de pré-processamento, montagem e anotação de genomas, mas mantendo a sua interoperabilidade. Estes *pipelines*, diferentemente de programas já existentes, é genérico e permite que usuários configurem diversas combinações de dados e programas para a análise de genomas procarióticos. O *pipeline* de montagem garante extrema flexibilidade para a produção de genomas somente com leituras longas, curtas ou de forma híbrida, utilizando quatro programas de montagem diferentes. O *pipeline* de anotação, é abrangente e permite, além da anotação genérica de proteínas, a identificação de genes de resistência, virulência e elementos genéticos móveis. Tudo isso, entregue em relatórios ricos em informações quanto ao processo de anotação e resultados obtidos, além da produção de um navegador genômico, extremamente útil para a investigação dos genes preditos.

Reporta-se neste trabalho uma versão totalmente testada e funcional destes *pipelines*, mas salienta-se que existe espaço para melhorias. A tecnologia Nextflow, recentemente divulgou uma atualização que permitirá que usuários incrementem facilmente *pipelines* existentes sem a necessidade de edição do código base. A adaptação dos *pipelines* a esta nova versão os tornará extremamente personalizáveis e extensíveis de modo a melhor atender as necessidades particulares de cada usuário. Além disso, pretende-se desenvolver uma interface gráfica que permita a entrada de dados e execução de forma intuitiva por parte de usuários pouco familiares com a bioinformática.

Estes *pipelines* podem ser utilizados em ambientes computacionais diversos, desde *laptop* Windows (>16 Gb RAM) até servidores Linux com inúmeros processadores. Por isso, são capazes de aproximar a genômica à ambientes clínicos de maneira simples, permitindo a realização, em *laptops*, de análises completas de montagem e anotação de genomas, incluindo a anotação de genes de resistência e a predição *in silico* deste fenótipo. Futuramente, com a incorporação da genômica a prática clínica podem ser desenvolvidos bancos de dados que incrementalmente absorvam os resultados destes *pipelines* de modo a permitir a manutenção de um histórico da prevalência de genes de resistência e virulência de um local ou região. Estes bancos de dados podem ser fundamentais também para estudos de evolução de patógenos e disseminação de características adaptativas.

Um estudo de caso com um isolado de *Klebsiella pneumoniae* obtido no Hospital Universitário da UnB, o Kp31, utilizando os *pipelines* desenvolvidos, demonstrou o grande sucesso na montagem de um genoma de alta qualidade utilizando dados de tecnologias de sequenciamento de leituras curtas (Illumina) e longas (Oxford Nanopore). Os dados de nanoporos permitiram a resolução individualizada de cromossomos e plasmídeos. Porém, demonstrou-se nas análises de correção de sequências que sem os dados de alta qualidade da plataforma Illumina as sequências genômicas montadas possuíam inúmeros erros de sequência que introduzem vieses na anotação gênica, como a introdução de códons de parada prematuros, mutações "frameshift", inserções, deleções, mutações "missense", entre outros. Portando, denota-se a necessidade da aplicação conjunta de leituras curtas e longas e demonstra-se seu grande potencial na resolução de genomas, capazes de produzir montagens extremamente contíguas e de alta qualidade.

Em essência, estes resultados indicam a possibilidade de consolidar protocolos de sequenciamento que permitam fechar genomas bacterianos de interesse clínico com alta qualidade. A quantidade de dados gerados neste trabalho permitem desenhar hoje, um cenário de sequenciamento de 4 bactérias por *flowcell* de ONT e 1 lane de Illumina. Em termos de custo, este cenário chegaria a um total de 750 USD por genoma bacteriano, considerando valores de ≈1000 USD do sequenciamento ONT e 2000 USD do sequenciamento Illumina. Neste cenário, seriam obtidos ≈250 X de dados ONT por genoma bacteriano. Porém, os custos podem ser ainda menores uma vez que os programas de montagem de genomas sugerem a utilização de somente 20-50 X de dados de sequenciamento ONT

para a produção de genomas de alta qualidade. Por isso, a padronização e estipulação de protocolos de sequenciamento multiplex de bactérias utilizando dados Illumina e ONT permitirão a produção de genomas de altíssima qualidade com ótimo custo benefício.

Em termos de anotação, foi possível identificar que a Kp31 faz parte de um patotipo, o ST11, considerado de alto risco à saúde pública mundial por ser frequentemente associado à incorporação conjunta de diversos genes de virulência e de resistência a antibióticos (Wang et al., 2018; Jia et al., 2019; Fuga et al., 2020; van Dorp et al., 2019). A caracterização genômica da Kp31 revela que esta é multirresistente a antibióticos, virulenta e hipermucoviscosa. Estes achados reeditam alertas quanto ao surgimento de cepas bacterianas de altíssimo risco caracterizados pela convergência de genes de resistência e virulência. Complementarmente, a análise deste isolado evidencia alguns debates recentes quanto à utilização sinônima dos termos hipervirulência e hipermucoviscosidade. Os resultados corroboram com a ideia de dois fenótipos diferentes mas complementares introduzido por Catalán-Nájera em 2017. Portanto, denota-se que a hipervirulência é um fenótipo bastante complexo e ainda pouco compreendido que demanda esforços mundiais para a definição de novos biomarcadores e o estabelecimento de um consenso (Harada e Doia, 2018). Além disso, apesar do isolado Kp31 apresentar fenótipo de hipermucoviscosidade, esta cepa não apresenta os genes RmpA/A2, geralmente identificados como responsáveis pelo fenótipo. O que corrobora com outros estudos que também identificam este padrão (Fang et al., 2004; Cubero et al., 2016; Zhang et al., 2019). Este resultado, demonstra a necessidade de mais estudos acerca do fenótipo de hipermucoviscosidade e seus determinantes. Porém, fortalece uma proposição recente sobre a utilização destes genes como marcadores da hipervirulência (Russo e Marr, 2019; Harada e Doia, 2018).

Como um todo, estes resultados reforçam alertas quanto a necessidade do estabelecimento de programas de vigilância de patógenos. A comparação das análises *in silico* do isolado Kp31 com resultados experimentais demonstrou a confiabilidade da genômica para a identificação de genes de resistência e a predição de fenótipos. Portanto, a velocidade e confiabilidade das análises obtidas através dos *pipelines* desenvolvidos neste estudo demonstram seu grande potencial de aplicação em estudos epidemiológicos e disponibilizam arsenal técnico que abre caminho para o desenvolvimento de programas de vigilância através de genomas, a vigilância genômica.

Referências Bibliográficas

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

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., Vasconcelos A. T. R., Medeiros E. A., Walsh T. R., Gales A. C., An Emerging Clone, KPC-2-Producing Klebsiella pneumoniae ST16, Associated with High Mortality Rates in a CC258 Endemic Setting, Clinical Infectious Diseases, 2019

Antonopoulos D. A., Assaf R., Aziz R. K., Brettin T., Bun C., Conrad N., Davis J. J., Dietrich E. M., Disz T., Gerdes S., Overbeek R., Parrello B., Pusch G. D., Santerre J., Shukla M., Stevens R. L., VanOeffelen M., Vonstein V., Warren A. S., Wattam A. R., Xia F., Yoo H., PATRIC as a unique resource for studying antimicrobial resistance, Briefings in Bioinformatics, 2019, vol. 20, p. 1094

Aramaki T., Blanc-Mathieu R., Endo H., Ohkubo K., Kanehisa M., Goto S., Ogata H., KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold, bioRxiv, 2019, p. 602110

Arkin A. P., Cottingham R. W., Henry C. S., Harris N. L., Stevens R. L., Maslov S., Dehal P., Ware D., Perez F., Canon S., Sneddon M. W., Henderson M. L., Riehl W. J., Syed M. H., Thomason J., Tintle N. L., Wang D., Xia F., Yoo H., Yoo S., Yu

D., KBase: The United States department of energy systems biology knowledgebase, Nature Biotechnology, 2018, vol. 36, p. 566

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 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

Bankevich A., Nurk S., Antipov D., Gurevich A. A., Dvorkin M., Kulikov A. S., Lesin V. M., Nikolenko S. I., Pham S., Prjibelski A. D., Pyshkin A. V., Sirotkin A. V., Vyahhi N., Tesler G., Alekseyev M. A., Pevzner P. A., SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing, Journal of Computational Biology, 2012, vol. 19, p. 455

Baptista R. P., Kissinger J. C., Is reliance on an inaccurate genome sequence sabotaging your experiments?, PLoS Pathogens, 2019, vol. 15, p. e1007901

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

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

Bertelli C., Brinkman F. S., Improved genomic island predictions with IslandPath-DIMOB, Bioinformatics, 2018, vol. 34, p. 2161

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

Boolchandani M., D'Souza A. W., Dantas G., Sequencing-based methods and resources to study antimicrobial resistance, Nat. Rev. Genet., 2019, vol. 20, p. 356

Brisse S., Passet V., Haugaard A. B., Babosan A., Kassis-Chikhani N., Struve C., Decré D., wzi Gene Sequencing, a Rapid Method for Determination of Capsular Type for Klebsiella Strains, Journal of Clinical Microbiology, 2013, vol. 51, p. 4073

Brown E. D., Wright G. D., Antibacterial drug discovery in the resistance era, Nature, 2016, vol. 529, p. 336

Buchfink B., Xie C., Huson D. H., Fast and sensitive protein alignment using DIAMOND, Nature Methods, 2014, vol. 12, p. 59

Buels R., Yao E., Diesh C. M., Hayes R. D., Munoz-Torres M., Helt G., Goodstein D. M., Elsik C. G., Lewis S. E., Stein L., Holmes I. H., JBrowse: A dynamic web platform for genome visualization and analysis, Genome Biology, 2016, vol. 17, p. 66

Caboche S., Even G., Loywick A., Audebert C., Hot D., MICRA: An automatic pipeline for fast characterization of microbial genomes from high-throughput sequencing data, Genome Biology, 2017, vol. 18

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

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

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

Charalampous T., Kay G. L., Richardson H., Aydin A., Baldan R., Jeanes C., Rae D., Grundy S., Turner D. J., Wain J., Leggett R. M., Livermore D. M., O'Grady J., Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection, Nature Biotechnology, 2019, vol. 37, p. 783

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

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

Cingolani P., Platts A., Coon M., Nguyen T., Wang L., Land S., Lu X., Ruden D., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3, Fly, 2012, vol. 6, p. 80

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

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

Dale R., Grüning B., Sjödin A., Rowe J., Chapman B. A., Tomkins-Tinch C. H., Valieris R., Batut B., Caprez A., Cokelaer T., Stöcker B. K., Moskalenko O., Bogema D. R., Workentine M. L., Newhouse S. J., Leprevost F. d. V., Arvai K., Köster J., Bioconda: Sustainable and comprehensive software distribution for the life sciences, Nature Methods, 2018, vol. 15, p. 475

Danchin A., Ouzounis C., Tokuyasu T., Zucker J. D., No wisdom in the crowd: genome

annotation in the era of big data – current status and future prospects, Microbial Biotechnology, 2018, vol. 11, p. 588

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. P. R., Stehling E. G., Pitondo-Silva A., 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., D'Hert S., Schultz D. T., Cruts M., Van Broeckhoven C., NanoPack: visualizing and processing long-read sequencing data, Bioinformatics, 2018, vol. 34, p. 2666

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

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

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

Edge P., Bafna V., Bansal V., HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies., Genome research, 2017, vol. 27, p. 801

Edge P., Bansal V., Longshot: accurate variant calling in diploid genomes using single-molecule long read sequencing, bioRxiv, 2019, p. 564443

Everett M. J., Jin Y. F., Ricci V., Piddock L. J., Contributions of individual mechanisms to fluoroquinolone resistance in 36 Escherichia coli strains isolated from humans and animals, Antimicrobial Agents and Chemotherapy, 1996, vol. 40, p. 2380

Fang C. T., Chuang Y. P., Shun C. T., Chang S. C., Wang J. T., A Novel Virulence Gene in Klebsiella pneumoniae Strains Causing Primary Liver Abscess and Septic Metastatic Complications, Journal of Experimental Medicine, 2004, vol. 199, p. 697

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., Tadesse D. A., Morales C., Simmons M., Tillman G., Wasilenko J., Folster J. P., Klimke W., Using the NCBI AMRFinder Tool to Determine Antimicrobial Resistance Genotype-Phenotype Correlations Within a Collection of NARMS Isolates, bioRxiv, 2019, p. 550707

Feng Y., Liu L., McNally A., Zong Z., Coexistence of three bla KPC-2 genes on an IncF/IncR plasmid in ST11 Klebsiella pneumoniae, Journal of Global Antimicrobial Resistance, 2019, vol. 17, p. 90

Feng Y., Zhang Y., Ying C., Wang D., Du C., Nanopore-Based Fourth-Generation DNA Sequencing Technology, Genomics Proteomics Bioinformatics, 2015, vol. 13, p. 4

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., Da Silva Pranchevicius M. C., 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

Founou R. C., Founou L. L., Essack S. Y., Clinical and economic impact of antibiotic resistance in developing countries: A systematic review and meta-analysis, PLoS ONE, 2017, vol. 12, p. e0189621

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

Giesselmann P., Hetzel S., Müller F. J., Meissner A., Kretzmer H., Nanopype: a modular and scalable nanopore data processing pipeline, Bioinformatics (Oxford, England), 2019, vol. 35, p. 4770

Gillings M. R., Lateral gene transfer, bacterial genome evolution, and the Anthropocene, Annals of the New York Academy of Sciences, 2017, vol. 1389, p. 20

Gold H., Moellering R., Antimicrobial-drug resistance, The New England journal of medicine, 1996, vol. 335, p. 1445—1453

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

Gong L., Huang Y. T., Wong C. H., Chao W. C., Wu Z. Y., Wei C. L., Liu P. Y., Culture-independent analysis of liver abscess using nanopore sequencing, PLoS ONE, 2018, vol. 13, p. e0190853

González-Escalona N., Allard M. A., Brown E. W., Sharma S., Hoffmann M., Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing Escherichia coli, PLoS ONE, 2019, vol. 14, p. e0220494

Gonçalves L. F., Caracterização microbiológica e avaliação do papel do PPARγ na colonização e sobrevivência de isolados hipermucoides de Klebsiella pneumoniae em células epiteliais da linhagem HEp-2, Universidade de Brasília, 2018, Dissertação de Mestrado

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., Chen S., 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

Gurevich A., Saveliev V., Vyahhi N., Tesler G., QUAST: quality assessment tool for genome assemblies, Bioinformatics, 2013, vol. 29, p. 1072

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

Harada S., Doia Y., Hypervirulent Klebsiella pneumoniae: A call for consensus definition and international collaboration, Journal of Clinical Microbiology, 2018, vol. 56

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

Holt K. E., Wertheim H., Zadoks R. N., Baker S., Whitehouse C. A., Dance D., Jenney A., Schultsz C., Kuntaman K., Newton P. N., Moore C. E., Strugnell R. A., Thomson N. R., 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

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

Jia B., Raphenya A. R., Alcock B., Waglechner N., Guo P., Tsang K. K., Lago B. A., Dave B. M., Pereira S., Sharma A. N., Doshi S., Courtot M., Lo R., Williams L. E., Frye J. G., Elsayegh T., Sardar D., Westman E. L., Pawlowski A. C., Johnson T. A., Brinkman F. S., Wright G. D., McArthur A. G., CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database, Nucleic Acids Research, 2017, vol. 45, p. D566

Jia H., Chen H., Ruan Z., Unravelling the genome sequence of a pandrug-resistant Klebsiella pneumoniae isolate with sequence type 11 and capsular serotype KL64 from China, Journal of Global Antimicrobial Resistance, 2019, vol. 19, p. 40

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

Kang D. D., Froula J., Egan R., Wang Z., MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities, PeerJ, 2015, vol. 3, p. e1165

Kapoor G., Saigal S., Elongavan A., Action and resistance mechanisms of antibiotics: A guide for clinicians, Journal of Anaesthesiology Clinical Pharmacology, 2017, vol. 33, p. 300

Kasianowicz J. J., Brandin E., Branton D., Deamer D. W., Characterization of Individual Polynucleotide Molecules Using a Membrane Channel, Proc. Natl. Acad. Sci. U.S.A., 1996, vol. 93, p. 13770

Kim S. Y., Ko K. S., Diverse Plasmids Harboring blaCTX-M-15 in Klebsiella pneumoniae ST11 Isolates from Several Asian Countries, Microbial Drug Resistance, 2019, vol. 25, p. 227

Kislyuk A. O., Katz L. S., Agrawal S., Hagen M. S., Conley A. B., Jayaraman P., Nelakuditi V., Humphrey J. C., Sammons S. A., Govil D., Mair R. D., Tatti K. M., Tondella M. L., Harcourt B. H., Mayer L. W., Jordan I. K., A computational genomics pipeline for prokaryotic sequencing projects, Bioinformatics, 2010, vol. 26, p. 1819

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., Walenz B. P., Berlin K., Miller J. R., Bergman N. H., Phillippy A. M., Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation, Genome Research, 2017, vol. 27, p. 722

Korlach J., Bjornson K. P., Chaudhuri B. P., Cicero R. L., Flusberg B. A., Gray J. J., Holden D., Saxena R., Wegener J., Turner S. W., Real-time DNA sequencing from single polymerase molecules., Methods in enzymology, 2010, vol. 472, p. 431

Kosmidis C., Schindler B. D., Jacinto P. L., Patel D., Bains K., Seo S. M., Kaatz G. W., Expression of multidrug resistance efflux pump genes in clinical and environmental isolates of Staphylococcus aureus, International Journal of Antimicrobial Agents, 2012, vol. 40, p. 204

Kumar K., Desai V., Cheng L., Khitrov M., Grover D., Satya R. V., Yu C., Zavaljevski N., Reifman J., AGeS: A software system for microbial genome sequence annotation, PLoS ONE, 2011, vol. 6, p. e17469

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

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

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

Lewis K., Platforms for antibiotic discovery, Nature Reviews Drug Discovery, 2013, vol. 12, p. 371

Li H., , 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

Li H., Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics, 2018, vol. 34, p. 3094

Li R., Xie M., Dong N., Lin D., Yang X., Wong M. H. Y., Chan E. W. C., Chen S., Efficient generation of complete sequences of MDR-encoding plasmids by rapid assembly of MinION barcoding sequencing data, GigaScience, 2018, vol. 7, p. 1

Li X., Xie Y., Liu M., Tai C., Sun J., Deng Z., Ou H.-Y., oriTfinder: a web-based tool for the identification of origin of transfers in DNA sequences of bacterial mobile genetic elements, Nucleic Acids Research, 2018, vol. 46, p. W229

Liakopoulos A., Van Der Goot J., Bossers A., Betts J., Brouwer M. S., Kant A., Smith H., Ceccarelli D., Mevius D., Genomic and functional characterisation of IncX3 plasmids encoding bla SHV-12 in Escherichia coli from human and animal origin, Scientific Reports, 2018, vol. 8, p. 1

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 M., Li X., Xie Y., Bi D., Sun J., Li J., Tai C., Deng Z., Ou H. Y., ICEberg 2.0: An updated database of bacterial integrative and conjugative elements, Nucleic Acids Research, 2019, vol. 47, p. D660

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

Loman N. J., Quick J., Simpson J. T., A complete bacterial genome assembled de novo using only nanopore sequencing data, Nature Methods, 2015, vol. 12, p. 733

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

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

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., Paterson D. L., Rice L. B., Stelling J., Struelens M. J., Vatopoulos A., Weber J. T., Monnet D. L., 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

Magoč T., Salzberg S. L., FLASH: Fast length adjustment of short reads to improve genome assemblies, Bioinformatics, 2011, vol. 27, p. 2957

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., Swaminathan S., Van Der Walt M., Alland D., Bishai W. R., Cohen T., Hoffner S., Birren B. W., Earl A. M., 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 M., Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.journal, 2011, vol. 17, p. 10

Martin M., Patterson M., Garg S., Fischer S. O., Pisanti N., Klau G. W., Schöenhuth A., Marschall T., WhatsHap: fast and accurate read-based phasing, bioRxiv, 2016, p. 085050

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

Marçais G., Delcher A. L., Phillippy A. M., Coston R., Salzberg S. L., Zimin A., MUMmer4: A fast and versatile genome alignment system, PLOS Computational Biology, 2018, vol. 14, p. 1

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

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

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

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

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

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

Page A. J., Cummins C. A., Hunt M., Wong V. K., Reuter S., Holden M. T., Fookes M., Falush D., Keane J. A., Parkhill J., Roary: Rapid large-scale prokaryote pan genome analysis, Bioinformatics, 2015, vol. 31, p. 3691

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

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

Pumbwe L., Piddock L. J., Two efflux systems expressed simultaneously in multidrug-resistant Pseudomonas aeruginosa, Antimicrobial Agents and Chemotherapy, 2000, vol. 44, p. 2861

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

Reuter J. A., Spacek D. V., Snyder M. P., High-Throughput Sequencing Technologies, Mol. Cell, 2015, vol. 58, p. 586

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., Sahl J. W., 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

Ruan Z., Feng Y., BacWGSTdb, a database for genotyping and source tracking bacterial pathogens, Nucleic Acids Research, 2015, vol. 44, p. D682

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

Santajit S., Indrawattana N., Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens, BioMed Research International, 2016, vol. 2016, p. 1

Sayers S., Li L., Ong E., Deng S., Fu G., Lin Y., Yang B., Zhang S., Fa Z., Zhao B., Xiang Z., Li Y., Zhao X. M., Olszewski M. A., Chen L., He Y., Victors: A web-based knowledge base of virulence factors in human and animal pathogens, Nucleic Acids Research, 2019, vol. 47, p. D693

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

Simão F. A., Waterhouse R. M., Ioannidis P., Kriventseva E. V., Zdobnov E. M., BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics, 2015, vol. 31, p. 3210

Smith A. M., Tau N. P., Kalule B. J., Nicol M. P., McCulloch M., Jacobs C. A., McCarthy K. M., Ismail A., Allam M., Kleynhans J., Shiga toxin-producing Escherichia coli O26:H11 associated with a cluster of haemolytic uraemic syndrome cases in South Africa, 2017, Access Microbiology, 2019

Snitkin E. S., Zelazny A. M., Thomas P. J., Stock F., Henderson D. K., Palmore T. N., Segre J. A., Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing, Science Translational Medicine, 2012, vol. 4

Song L., Florea L., Langmead B., Lighter: fast and memory-efficient sequencing error correction without counting, Genome biology, 2014, vol. 15, p. 509

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

Starikova E. V., Tikhonova P. O., Prianichnikov N. A., Rands C. M., Zdobnov E. M., Govorun V. M., Phigaro: high throughput prophage sequence annotation, bioRxiv, 2019, p. 598243

Stout V., Torres-Cabassa A., Maurizi M. R., Gutnick D., Gottesman S., RcsA, an unstable positive regulator of capsular polysaccharide synthesis, Journal of Bacteriology, 1991, vol. 173, p. 1738

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

Su K., Zhou X., Luo M., Xu X., Liu P., Li X., Xue J., Chen S., Xu W., Li Y., Qiu J., Genome-wide identification of genes regulated by RcsA, RcsB, and RcsAB phosphorelay regulators in Klebsiella pneumoniae NTUH-K2044, Microbial Pathogenesis, 2018, vol. 123, p. 36

Tamma P. D., Fan Y., Bergman Y., Pertea G., Kazmi A. Q., Lewis S., Carroll K. C., Schatz M. C., Timp W., Simner P. J., Applying rapid whole-genome sequencing to predict phenotypic antimicrobial susceptibility testing results among carbapenem-resistant klebsiella pneumoniae clinical isolates, Antimicrobial Agents and Chemotherapy, 2019, vol. 63

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

Tatusova T., Dicuccio M., Badretdin A., Chetvernin V., Nawrocki E. P., Zaslavsky L., Lomsadze A., Pruitt K. D., Borodovsky M., Ostell J., NCBI prokaryotic genome annotation pipeline, Nucleic Acids Research, 2016, vol. 44, p. 6614

Taylor T. L., Volkening J. D., DeJesus E., Simmons M., Dimitrov K. M., Tillman G. E., Suarez D. L., Afonso C. L., Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology, Scientific Reports, 2019, vol. 9

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

Vallenet D., Calteau A., Dubois M., Amours P., Bazin A., Beuvin M., Burlot L., Bussell X., Fouteau S., Gautreau G., Lajus A., Langlois J., Planel R., Roche D., Rollin J., Rouy Z., Sabatet V., Médigue C., MicroScope: an integrated platform for the annotation and

exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis., Nucleic acids research, 2019

van Dorp L., Wang Q., Shaw L. P., Acman M., Brynildsrud O. B., Eldholm V., Wang R., Gao H., Yin Y., Chen H., Ding C., Farrer R. A., Didelot X., Balloux F., Wang H., Rapid phenotypic evolution in multidrug-resistant Klebsiella pneumoniae hospital outbreak strains, Microbial Genomics, 2019, vol. 5

van Duin D., Doi Y., The global epidemiology of carbapenemase-producing Enterobacteriaceae, Virulence, 2017, vol. 8, p. 460

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., Earl A. M., Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement, PLoS ONE, 2014, vol. 9, p. e112963

Wang Q., Wang X., Wang J., Ouyang P., Jin C., Wang R., Zhang Y., Jin L., Chen H., Wang Z., Zhang F., Cao B., Xie L., Liao K., Gu B., Yang C., Liu Z., Ma X., Jin L., Zhang X., Man S., Li W., Pei F., Xu X., Jin Y., Ji P., Wang H., 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 Y., Yang Q., Wang Z., The Evolution of Nanopore Sequencing, Front Genet, 2014, vol. 5, p. 449

Wattam A. R., Brettin T., Davis J. J., Gerdes S., Kenyon R., Machi D., Mao C., Olson R., Overbeek R., Pusch G. D., Shukla M. P., Stevens R., Vonstein V., Warren A., Xia F., Yoo H., , 2018 in , Vol. 1704, Methods in Molecular Biology. Humana Press Inc. pp 79–101

Wick R. R., Heinz E., Holt K. E., Wyres K. L., Kaptive web: User-Friendly capsule and lipopolysaccharide serotype prediction for Klebsiella genomes, Journal of Clinical Microbiology, 2018, vol. 56

Wick R. R., Judd L. M., Gorrie C. L., Holt K. E., Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads, PLoS Computational Biology, 2017, vol. 13, p. e1005595

Woodford N., Ellington M. J., The emergence of antibiotic resistance by mutation, Clinical Microbiology and Infection, 2007, vol. 13, p. 5

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

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

Xavier B. B., Mysara M., Bolzan M., Ribeiro-Gonçalves B., T.F Alako B., Harrison P., Lammens C., Kumar-Singh S., Goossens H., A Carriço J., Cochrane G., Malhotra-Kumar S., BacPipe: A Rapid, User-Friendly Whole Genome Sequencing Pipeline for Clinical Diagnostic Bacteriology, SSRN Electronic Journal, 2019, vol. 23, p. 100769

Zhang Y., Jin L., Ouyang P., Wang Q., Wang R., Wang J., Gao H., Wang X., Wang H., Network C. C.-R. E. C., Evolution of hypervirulence in carbapenem-resistant Klebsiella pneumoniae in China: a multicentre, molecular epidemiological analysis, Journal of Antimicrobial Chemotherapy, 2019, vol. 75, p. 327

Zhou Y., Liang Y., Lynch K. H., Dennis J. J., Wishart D. S., PHAST: A Fast Phage Search Tool, Nucleic Acids Research, 2011, vol. 39, p. W347

Apêndice

Apêndice A

Material Suplementar

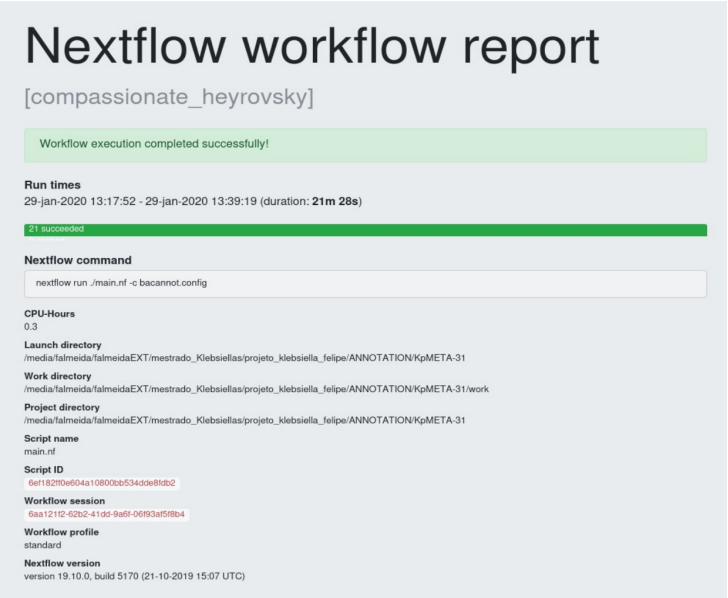


Figura A.1: Representação do relatório de execução de *pipelines* Nextflow. Captura de tela da primeira página do relatório.

Tabela A.1 - Estatísticas gerais da qualidade dos dados de sequenciamento do isolado Kp34

N° de leituras	Tamanho médio	Qualidade média	Quantidade de leituras com qualidade >7
903.250,0	8.096,1	11,2	77%

```
/*
 * Configuration File to run NGS-PreProcess pipeline.
 */

/*
 * General Parameters
 */
    // Output folder name
params.outDir = 'output'
    // Number of threads to be used
params.threads = 2
    // Set true or false to run or not longreads and shortreads pipeline.
params.run_shortreads_pipeline = true
params.run_longreads_pipeline  = false

/*
 * Short Reads Parameters
 */
    // Loading input reads
    // Remember to use wildcards in order to allow the pipeline to get
reads ID.
    // Loading examples:
    // For Paired end reads: params.shortreads = 'SRR6307304_{1,2}.fastq'
& params.reads.size = '2'
    // For Single end: params.shortreads = 'SRR7128258*' &
params.reads.size = '1'
params.shortreads = 'dataset_1/illumina/*_{1,2}.fastq'
    // 1 for single end, 2 for paired ends. Let it blank if not needed.
params.reads_size = 2
    //
    // TrimGalore Parameters
    //
    // These are the parameters used to set the number of bases to clip
from
    // 5' end and 3' end of paired end reads in TrimGalore. 0 < value <
read length.
    // Optional. Quality default is 20 (phred)
    // Clip from 5' end
params.clip_r1 = 0
params.clip_r2 = 0
    // Clip from 3' end
params.three_prime_clip_r1 = 0
params.three_prime_clip_r2 = 0
    // This one might me left blank to use 20 as default or set a integer
params.quality_trim = 20
    //
    // Lighter error correction parameters
    // Set wheter to run or not lighter correction step.
params.lighter_execute = false
    //
    // Which k-mer to use. Check Ligther's manual
(https://github.com/mourisl/Lighter)
params.lighter_kmer = 21
    // Bacterial Genome Size
```

Figura A.2: Representação do arquivo de configuração utilizado pelos *pipelines*. Apresenta-se a primeira página do arquivo de configuração, demonstrando sua sintaxe e informações embebidas no texto.