UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
FACULDADE DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

RODRIGO DE BRITTO PONTES RODRIGUES PARÁ

UM ESTUDO COMPARATIVO DE FERRAMENTAS DE BINNING DE METAGENOMAS UTILIZANDO DADOS SIMULADOS MICROBIANOS

BELÉM
2019

RODRIGO DE BRITTO PONTES RODRIGUES PARÁ

UM ESTUDO COMPARATIVO DE FERRAMENTAS DE BINNING DE METAGENOMAS UTILIZANDO DADOS SIMULADOS MICROBIANOS

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Pará como exigência parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Prof.ª Dr.ª Regiane Silva Kawasaki Francês

Coorientador: MSc. Renato Renison Moreira Oliveira

BELÉM
2019

RODRIGO DE BRITTO PONTES RODRIGUES PARÁ

UM ESTUDO COMPARATIVO DE FERRAMENTAS DE BINNING DE METAGENOMAS UTILIZANDO DADOS SIMULADOS MICROBIANOS

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Pará como exigência parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Prof.ª Dr.ª Regiane Silva Kawasaki Francês

Coorientador: MSc. Renato Renison Moreira Oliveira

Aprovado em: ___ /___ /____

Conceito: ____

BANCA EXAMINADORA:

_____

Prof.ª Dr.ª Regiane Silva Kawasaki Francês (Orientadora - UFPA)

_____

MSc. Renato Renison Moreira Oliveira

_____

Prof.ª Dr.ª Danielle Costa Carrara Couto

_____

Prof. Dr. Vinicius Augusto Carvalho de Abreu

AGRADECIMENTOS

Este trabalho não teria sido possível sem o apoio de diversas pessoas ao meu redor. Gostaria de expressar aqui minha profunda gratidão a todo mundo que acompanhou o meu processo.

Primeiramente, agradeço à UFPa, à Faculdade de Computação e a todos os professores que contribuíram na minha formação.

A meus pais, que nunca deixaram faltar nada e que investiram todos os recursos possíveis e inimagináveis para que eu tivesse uma boa formação no decorrer da minha vida.

Ao LaBioCad, em especial à Professora Regiane Kawasaki, ciente da minha situação e das minhas dificuldades, por ter me aceitado e me acolhido no laboratório e ter estado sempre disposta a me ajudar a qualquer momento.

Ao Renato Oliveira, e ao Pedro Henrique, por serem sempre pacientes e atenciosos ao tirarem minhas muitas dúvidas, seja pessoalmente, seja pelo Whatsapp. Sem vocês, esse trabalho definitivamente não teria acontecido.

À Elyene, minha psicóloga, que estabeleceu todas as bases para que eu pudesse ter forças para seguir em frente quando minha mente dizia o contrário.

Ao Lucas Pacheco, meu companheiro de vida, que há quase 4 anos caminha junto comigo e me viu nos meus piores e nos meus melhores momentos, e que sempre acreditou em mim mais do que eu mesmo.

Agradeço aos amigos que a dança, mais precisamente a Companhia Mirai me trouxe ou me reaproximou: Bel, Franco, Rafa, Renan, Renato, e muitos outros. Os ensinamentos que tive enquanto artista moldaram o meu caráter e certamente caminharão comigo até o fim da minha vida.

A todos os meus outros amigos, que estiveram presentes e me ajudaram direta e indiretamente nessa caminhada:

À Duana Aquino, não só por toda a força dada durante a realização dessa pesquisa, mas também pelos muitos anos de amizade e conexão, sem os quais eu provavelmente não seria a pessoa que eu sou hoje;

À Alana Krás, por fazer dissipar minhas angústias através dos melhores conselhos, e também pelo auxílio prestado no trabalho;

Ao Danilo Serrão, que conseguia me entender sem que eu precisasse falar muito;

Ao Alfredo Filho, por ser uma das pessoas mais ocupadas e ao mesmo tempo mais companheiras que já conheci.

A todas as pessoas citadas acima, e também tantas outras que fizeram parte da minha caminhada, os meus sinceros agradecimentos.

RESUMO

A Metagenômica é uma derivação do ramo tradicional da Genômica que estuda o genoma de comunidades de microrganismos retirados diretamente do seu habitat, eliminando a necessidade do cultivo laboratorial. A reconstrução da cadeia de DNA de um genoma é feita através do processo de sequenciamento onde são criados fragmentos menores de DNA, que serão posteriormente montados em sequências maiores a fim de se chegar no genoma original. Para dados metagenômicos, esse processo de montagem é mais trabalhoso, dado que vários genomas de organismos diferentes estão presentes na amostra. Separar as sequências obtidas no processo de montagem em categorias distintas de acordo com o nível taxonômico dos organismos é um processo denominado *binning*, e é importante para facilitar a posterior análise da composição da amostra, para a descoberta de relações filogenéticas e de novos genes. Este trabalho busca comparar os resultados de ferramentas de *binning* através da definição de um *pipeline* de análise de dados metagenômicos, utilizando dois conjuntos de dados simulados de 10 e 100 espécies de bactérias. Utilizou-se o *software Metaquast* para a comparação de três softwares de montagem (*IDBA_UD*, *Megahit* e *MetasSPAdes*) e dois *softwares* de *binning* (*MetaBAT-2.12.1* e *MaxBin-2.2.4)*. A qualidade dos *bins* gerados por cada ferramenta foi comparada através da construção de uma Matriz de Confusão e da comparação das métricas de Acurácia, Sensibilidade, Especificidade e Precisão. Como resultado, o *MetaBAT* se sobressai ao *MaxBin* na qualidade dos *bins* gerados em ambos os conjuntos de dados, provando ser uma boa opção na realização do *binning* de metagenomas.

Palavras-chave: Bioinformática; Metagenômica; Montagem de metagenomas; Binning de metagenomas; Dados simulados.

ABSTRACT

Metagenomics is a derivation of the traditional branch of Genomics that studies the genome of communities of microorganisms extracted directly from their habitat, eliminating the need for laboratory cultivation. The reconstruction of the DNA strand of a genome is done through the sequencing process where smaller fragments of DNA are created, which will later be assembled into larger sequences in order to reach the original genome. For metagenomic data, this assembly process is more laborious, since several genomes of different organisms are present in the sample. Separating the sequences obtained in the assembly process into distinct categories according to the taxonomic level of organisms is a process called binning and is important to facilitate the subsequent analysis of the composition of the sample, for the discovery of phylogenetic relationships and new genes. This work seeks to compare the results of binning tools by defining a pipeline of metagenomic data analysis using two simulated datasets of 10 and 100 species of bacteria. Metaquast software was used to compare three assembly softwares (IDBA_UD, Megahit and MetasSPAdes) and two binning softwares (MetaBAT-2.12.1 and MaxBin-2.2.4). The quality of the bins generated by each tool was compared by constructing a Confusion Matrix and comparing the Accuracy, Sensitivity, Specificity and Accuracy metrics. As a result, MetaBAT excels at MaxBin in the quality of the bins generated in both sets of data, proving to be a good choice in performing binning of metagenomes.

Keywords: Bioinformatics; Metagenomics; Metagenomic Assembly; Metagenomic Binning; Simulated data.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

OTU - *Operational Taxonomy Unit*

NCBI - *National Center for Biotechnology Information*

NGS - *Next Generation Sequencing*

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTO

A Bioinformática pode ser definida como a junção das duas grandes áreas da Biologia e Informática voltada para desenvolvimento de ferramentas, algoritmos, métodos e análises computacionais que sejam capazes de tratar dados biológicos, como sequências de DNA e proteínas (LESK, 2008).

Para que seja possível analisar o DNA de um organismo e determinar a sequência de nucleotídeos de seus cromossomos, é necessário inicialmente fragmentar o DNA em porções menores para depois reconstruí-lo. Esse processo é realizado através da fragmentação química, física ou enzimática da molécula de DNA em porções menores (VERLI, 2014). A fragmentação acontece de maneira aleatória e é denominada de sequenciamento *shotgun*. Reconstruir os fragmentos menores de DNA em porções maiores, de forma a se ter o genoma original do organismo, é um processo denominado Montagem.

O desenvolvimento de tecnologias de sequenciamento fez com que crescesse de forma exponencial a quantidade de dados biológicos gerados, tornando necessário o desenvolvimento de tecnologias e técnicas computacionais que sejam capazes de tratar estes dados.

Dentro da Bioinformática, a Metagenômica ainda é um ramo recente que precisa de uma atenção especial, quando comparado à Genômica tradicional. A Metagenômica estuda a informação genética contida em uma comunidade de microrganismos, como os presentes em amostras ambientais (solos, água, trato intestinal) (WOOLEY et al., 2010).

O processo de montagem de um metagenoma é relativamente mais complexo do que de um genoma tradicional. Em uma amostra metagenômica, há a presença de muitos organismos. Sequenciar e montar o genoma destes organismos simultaneamente é uma tarefa difícil e exige o desenvolvimento de novas técnicas e ferramentas, haja vista que as ferramentas estabelecidas para montagem de genomas tradicionais não são aptas a tratar dados metagenômicos (PENG et al., 2011)

O *binning*, um passo posterior à montagem, consiste em agrupar as sequências de DNA montadas de acordo com uma determinada unidade taxonômica (e.g. espécie) (WOOLEY et al., 2010).

Este trabalho tem como objetivo efetuar a comparação de duas ferramentas de *binning*, de forma a validar um *pipeline* na análise de dados metagenômicos, utilizando dados simulados, e permitir que este *pipeline* possa ser utilizado com dados metagenômicos reais, a fim de descobrir novos microrganismos e genes.

1.2 JUSTIFICATIVA

Os softwares de montagem originais, como o SPades (BANKEVICH et al., 2012), e o Velvet (ZERBINO; BIRNEY, 2008), voltados para um único genoma, já foram inicialmente testados no *pipeline* de análise metagenômica com parâmetros bem especificados, porém com resultados poucos satisfatórios, produzindo vários erros de montagem devido à presença de regiões comuns no genoma das espécies (NAMIKI et al., 2012; PENG et al., 2011).

Novas ferramentas específicas foram desenvolvidas voltadas para a utilização no *pipeline* de dados metagenômicos, porém é necessário que se estabeleçam paradigmas na análise metagenômica através da verificação do desempenho das ferramentas de montagem e *binning* de metagenomas. Para tal, o uso de dados simulados se faz necessário devido ao conhecimento sobre a maioria das espécies microbianas ainda ser muito pequeno. É necessário recorrer a dados gerados sinteticamente, compostos por espécies já conhecidas, para que se afira o desempenho destas ferramentas através da comparação do *output* gerado por elas com os genomas de referência já confirmados. Essa análise poderá permitir posteriormente o reconhecimento do *pipeline* enquanto um modelo válido na análise de metagenomas, permitindo sua aplicação em dados reais.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Analisar e comparar a eficácia das ferramentas de *binning* de metagenomas em dados de diferente complexidade utilizando espécies microbianas simuladas, determinando as ferramentas e parâmetros necessários para a obtenção de bons resultados de *binning*, permitindo sua aplicação em contextos de dados reais.

1.3.2 Objetivos Específicos

- Realizar a montagem para o conjunto de dados simulados de 10 e 100 espécies;
- Realizar o *binning* para o conjunto de dados simulados de 10 e 100 espécies;
- Verificar a qualidade das montagens geradas através de ferramentas de validação e análise estatística;
- Verificar a qualidade dos *bins* gerados através de ferramentas de validação e análise estatística;
- Estabelecer critérios que aprimorem o desempenho das ferramentas através da identificação de parâmetros e características de sequenciamento e de montagem;
- Designar um *pipeline* de análise de dados metagenômicos.

1.4 ESTRUTURA DO TRABALHO

Além do presente capítulo referente à Introdução, este trabalho é composto de mais 4 capítulos:

- **Capítulo 2 – REFERENCIAL TEÓRICO:**

Este capítulo aborda os conceitos básicos referentes a metagenomas, sequenciamento, montagem e *binning*, bem como discorre sobre estudos previamente realizados utilizando dados simulados. Também são abordadas as ferramentas de montagem e *binning* presentes na literatura.

- **Capítulo 3 – MATERIAIS E MÉTODOS:**

Este capítulo aborda a metodologia utilizada na realização da montagem, do *binning* e na posterior análise dos resultados de dados metagenômicos, bem como quais conjuntos de dados foram empregados para verificar a performance destas ferramentas.

- **Capítulo 4 – RESULTADOS E DISCUSSÃO:**

Este capítulo apresenta os resultados do desempenho de duas ferramentas de *binning* (*MetaBAT* e *MaxBin*). Também estabelece um comparativo entre a performance das duas ferramentas em dois ambientes diferentes de dados simulados (10 e 100 espécies).

- **Capítulo 5 – CONCLUSÃO:**

Neste capítulo são estabelecidas as considerações finais a cerca deste estudo, bem como sugestões de possíveis novas pesquisas que possam vir a aperfeiçoar os resultados presentes neste trabalho.

2 REFERENCIAL TEÓRICO

Este capítulo possui a finalidade de apresentar o referencial teórico da pesquisa, contextualizando o problema central da pesquisa e conceituando termos que serão utilizados no decorrer do texto.

2.1 HISTÓRICO

Dados de microrganismos são obtidos através de análise laboratorial e em seguida armazenados em máquinas. Moléculas de DNA se traduzem a uma sequência de caracteres na tela do computador. Algoritmos implementados por softwares especializados tratam esses dados e permitem reconstruir genomas, predizer genes, e descobrir relações evolutivas entre organismos. Ao Bioinformata, é exigido o casamento entre a computação, estatística e matemática para o manuseio adequado desses softwares, e o conhecimento da Biologia molecular para a correta interpretação e análise destes dados (PROSDOCIMI, 2002).

A Genômica é um ramo da biologia que estuda o genoma dos organismos através das descobertas de seus genes e de suas funcionalidades. Quando se ouve falar em genoma, a primeira imagem que talvez venha à mente é o projeto em larga escala de sequenciamento do genoma humano. O Projeto Genoma Humano foi um grande esforço realizado por pesquisadores de diversos países que possuía o propósito de mapear e conhecer os genes, e consequentemente o genoma do ser humano (NIH, 2018). O projeto foi concluído em abril de 2013 e revelou o conhecimento de cerca de 20.500 genes presentes no DNA do ser humano.

Para armazenar as sequências de DNA, a NCBI mantém o *GenBank,* uma base de dados existente desde 1982 e de livre acesso. Até a data atual, a quantidade de bases presentes no banco de dados duplicou a cada 18 meses. Até outubro de 2018, o GenBank conta com cerca de 200 milhões de sequências submetidas por pesquisadores (GENBANK, 2018).

Segundo Wooley et al. (2010), acreditava-se que compreender a complexidade do organismo dos seres humanos residia em sequenciar e compreender o genoma humano. Porém, considerando-se que existem mais células bacterianas habitando o corpo de uma pessoa do que suas próprias células (SAVAGE, 1977 apud WOOLEY et al. 2010), entende-se que não apenas o conhecimento do genoma humano é necessário, mas também o conhecimento dos microrganismos presentes em nossos corpos.

Os microrganismos, presentes praticamente em todo lugar, são essenciais para qualquer fonte de vida devido serem fonte primária de nutrientes e os principais decompositores de matéria orgânica morta. A vida humana e sua qualidade está diretamente relacionada, e é profundamente afetada por eles, perpassando por diversos aspectos, como agricultura, medicina, indústria alimentícia, dentre outros. Possuir o conhecimento sobre os genes desses organismos se faz, portanto essencial (WOOLEY et al., 2010).

Existem limitações naturais quanto às técnicas de sequenciamento existentes necessárias para se sequenciar um genoma. As técnicas para sequenciamento de um único genoma não são adequadas para o sequenciamento de muitos microrganismos pelos seguintes fatores:

1. Cerca de 99% dos microrganismos não são cultiváveis em laboratório (CHITSAZ et al., 2011 apud PENG et al., 2012)
2. Uma cultura de organismos cultivados em laboratório não é capaz de representar a real diversidade e interação que existe nas comunidades de microorganismos, haja vista que eles raramente existem separadamente, pelo contrário, espécies interagem entre si e com os seus habitats (WOOLEY et al., 2010).

Dessa forma, é necessário que se adotem novas tecnologias e abordagens que permitam o sequenciamento desses organismos de maneira não tendenciosa.

2.2 METAGENÔMICA

A análise de microrganismos retirados diretamente do habitat ainda é recente. Na literatura, um dos estudos pioneiros é o de Venter (2004), no qual esforços foram empregados para sequenciar o genoma de microrganismos retirados do Mar dos Sargaços.

Em seu estudo (VENTER, 2004), as amostras foram retiradas diretamente do mar e filtradas, separando as bactérias presentes de possíveis outros organismos eucarióticos e vírus. Como resultado, cerca de 1.2 milhões de genes previamente desconhecidos foram identificados em um conjunto de dados estimado em 1800 espécies.

Tyson (2004) também realizou esforços ao usar o sequenciamento *shotgun* do DNA em uma comunidade de bactérias retiradas de uma região da Califórnia. Também obteve como

resultado a reconstrução quase completa do genoma de bactérias *Leptospirillum* grupo II e arqueobactérias *Ferroplasma* tipo II, e a reconstrução parcial de três outros genomas.

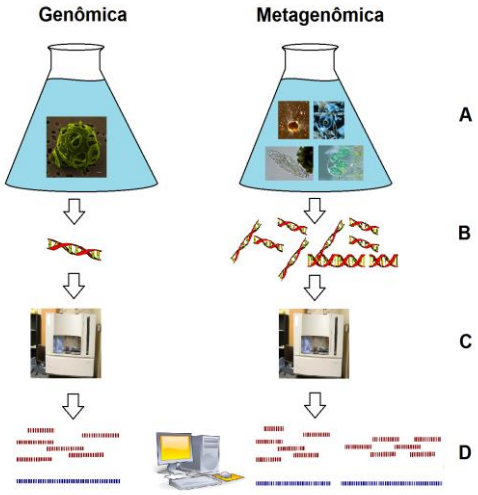
Estes estudos iniciais utilizaram ferramentas convencionais de montagem com apenas algumas modificações realizadas para tratar dados de múltiplos organismos. Para Nurk et al. (2017), estes estudos foram marcos iniciais para o avanço na área da metagenômica, pois a partir dele muito se tem trabalhado no desenvolvimento de novos softwares montadores específicos para a montagem de metagenomas.

A metagenômica pode ser definida como o campo de estudo que analisa o genoma de microrganismos os quais não podem ser isolados nem cultivados laboratorialmente. Segundo Mende et al. (2012, p. 1), "O campo de estudo da metagenômica examina a composição funcional e filogenética de comunidades microbianas em seus habitat naturais, permitindo acesso ao conteúdo genômico da maioria dos organismos que não são facilmente cultiváveis"

Conforme o autor acima explicitou, a Metagenômica, diferentemente da Genômica tradicional, utiliza como base de dados microorganismos retirados do seu próprio habitat, que pode ser o solo, água do mar, ar, intestino de hospedeiro, etc. Outros termos, como *Enviromental Genomics* (Genômica Ambiental) e *Community Genomics* (Genômica de comunidade) também podem ser encontrados na literatura para se referir à Metagenômica.

A Metagenômica enquanto área de estudo só foi possível graças ao desenvolvimento de novas tecnologias de sequenciamento e à redução no custo dessas tecnologias. Essas tecnologias permitem a verificação de espécies até então desconhecidas e não catalogadas, abrindo novos precedentes para a descoberta de relações evolutivas entre espécies e descoberta de novos genes. Ao conjunto de microorganismos extraídos diretamente do seu habitat se dá o nome de *metagenoma*. A Figura 1 estabelece um comparativo entre as diferenças no sequenciamento de um genoma e de um metagenoma:

Figura 1 - Comparação do sequenciamento de um genoma e um metagenoma



Fonte: Metagenômica: o "projeto genoma" dos micróbios (2018).

Enquanto que na genômica, a amostra é retirada de um organismo individualmente, na metagenômica a amostra é retirada de um determinado ambiente, contendo o DNA das diversas espécies presentes naquele ambiente.

2.3 MONTAGEM DE METAGENOMAS

Durante o processo de sequenciamento, o DNA é fragmentado em porções aleatórias menores, denominadas de *reads* (NAMIKI et al., 2012). O tamanho das *reads* geradas pode variar dependendo da tecnologia de sequenciamento utilizada, mas no geral elas variam de 50 a 800 pb (pares de base) (VERLI, 2014). As *reads*, então, podem ser definidas como pequenos fragmentos do DNA contendo *substrings* do genoma analisado.

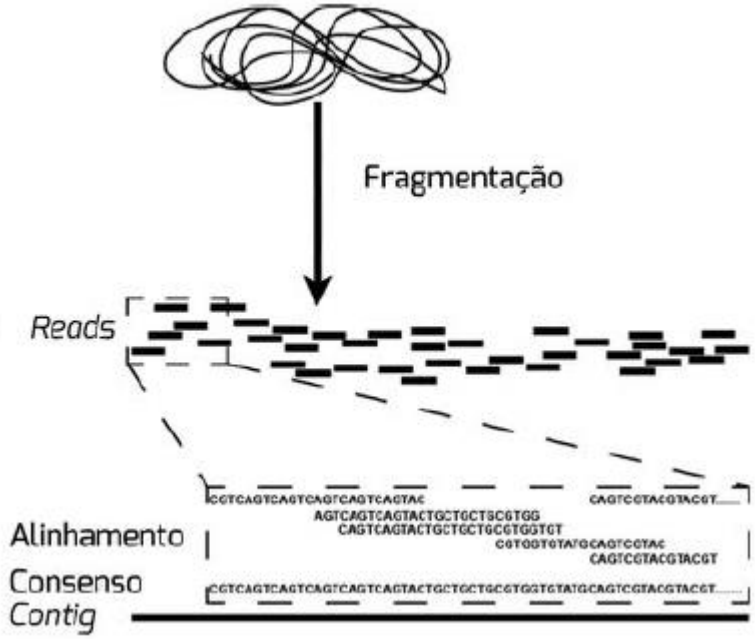
Na metagenômica, o processo realizado para sequenciamento de *reads* é semelhante ao de dados genômicos, porém com uma diferença crucial: ao invés das *reads* serem geradas a partir de uma única espécie, ela é gerada de um conjunto de microorganismos. Assim sendo, essas *reads* pertencem não apenas a uma espécie, mas a todas as espécies presentes naquela amostra. Como consequência, espécies mais abundantes na amostra terão maior material genético, enquanto que espécies menos abundantes possuirão menos material, permitindo dispor somente de uma visão parcial dos genomas presentes na amostra. O sequenciamento, que é denominado *shotgun* e pode ser feito de maneira química, física ou enzimática (VERLI,

2014, p. 63) pode ser chamado de *Enviromental* S*hotgun Sequencing*, ou ESS (WOOLEY et al., 2010), no caso de metagenomas.

Após a obtenção das *reads*, é necessário realizar o caminho inverso para a reconstrução do genoma. Isto é, a partir das *reads*, é preciso uni-las novamente em fragmentos cada vez maiores para tentar reconstruir da maneira mais completa possível o genoma da espécie. A esse processo se dá o nome de *montagem.*

Independente do tamanho ou da tecnologia de sequenciamento utilizada, o procedimento para a montagem é o mesmo. Como explica Verli (2014), para se juntar duas *reads* diferentes é necessário que se tenha uma região de justaposição, ou sobreposição, entre a cadeia de nucleotídeos de cada uma. Quando essas regiões equivalentes se encontram, as *reads* são alinhadas, formando então uma cadeia maior composta dos fragmentos de *reads* anteriores. O objetivo é buscar reconstruir uma grande sequência contínua de nucleotídeos através da sobreposição de várias *reads* diferentes. Essas sequências maiores são denominadas *contigs*. O *contig*, desse modo, é definido como "a união de duas ou mais sequências (*reads)* formadas por sobreposição de elementos comuns a pelo menos duas sequências" (VERLI, 2014, p. 64). Este processo é representado na figura 2 abaixo:

Figura 2 - Representação do processo de montagem de um genoma



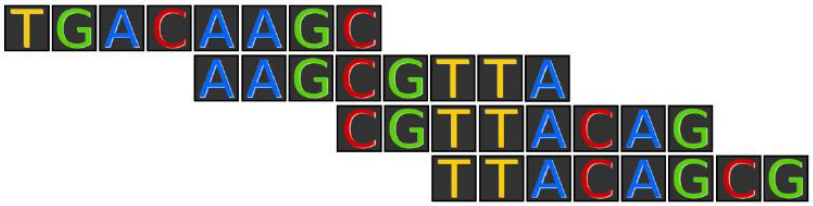
Fonte: VERLI (2014)

Complementando a definição acima, Mende et al. (2012) diz que os primeiros passos na análise de dados metagenômicos é a montagem das *reads* em "*contiguous consensus sequences"* (sequências de consenso contíguas), seguido pela predição de genes.

No que tange aos tipos de montagens existentes, elas podem ser subdivididas em duas categorias: montagem por referência e montagem *de novo*. A montagem por referência ocorre quando se tem disponível o genoma da espécie que está sendo trabalhada. Nesse caso, as *reads* são mapeadas diretamente contra o genoma de referência.

"O termo sequenciamento "*de novo*" vem do latim e significa "desde o princípio"." (MARTINS, 2013). Sendo assim, a montagem *de novo* é utilizada quando não se possui o conhecimento do genoma da espécie trabalhada. Neste caso, o processo a ser realizado, como já citado anteriormente, se dá através da junção de *reads* pelas suas sequências repetidas de nucleotídeos.

Figura 3 - Representação da sobreposição de *reads* através de regiões de repetição



Fonte: Wolf, B. (2018)

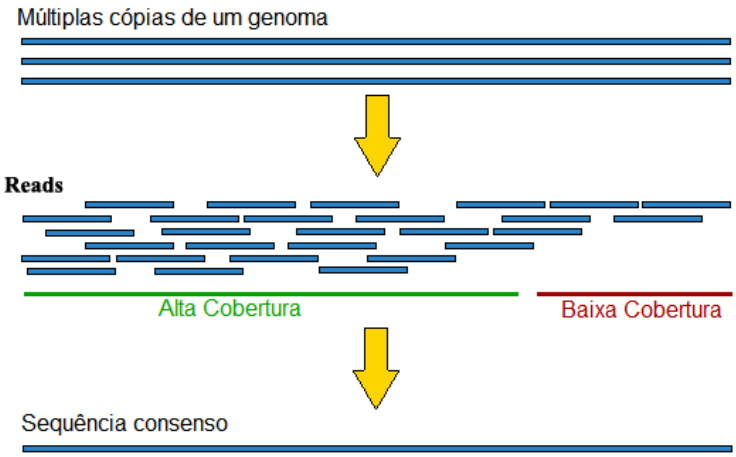
A montagem *de novo* é sempre mais complexa de ser realizada. Muitas vezes o genoma analisado é grande, e também possui muitas regiões repetidas (sequências específicas de nucleotídeos que se repetem várias vezes pelo genoma), ocasionando falhas na montagem (MARTINS, 2013).

Para analisar a qualidade de uma montagem, existem determinados parâmetros que devem ser levados em consideração. Um deles é a cobertura (*coverage)*. Wooley (2010) define a cobertura de um genoma como "o número médio de vezes em que um nucleotídeo é sequenciado". Wooley também ressalta que caso fosse possível sequenciar um genoma com apenas uma única *read*, um valor de cobertura de 1x seria suficiente para o sequenciamento.

A importância da cobertura como métrica de avaliação é dada à sua capacidade de garantir e permitir uma maior precisão na montagem do DNA. Regiões com alta cobertura oferecem garantia de que as *reads* terão regiões repetidas suficientes para se sobrepor, e que essas sobreposições serão distintas o suficiente para garantir a montagem. Isso é especialmente importante no caso de *reads* de comprimento menor, como afirma Martins (2013). *Reads* de

comprimento menor precisam, portanto, de um valor de cobertura maior, para que seja realizada uma montagem de qualidade.

Figura 4 - Cobertura de um genoma



Fonte: Adaptado de "Sequencing-by-Synthesis vs. Single Molecule Sequencing" (2019)

Na Figura 4, é possível observar como em um sequenciamento as regiões de um genoma podem ter coberturas diferentes. Em verde tem-se uma região com uma grande quantidade de *reads* e, portanto, uma cobertura alta, enquanto em vermelho há uma baixa quantidade de *reads*, gerando um valor de cobertura baixo.

Uma outra medida de referência utilizada é o N50. O N50 é uma medida que determina o quanto do genoma é coberto por *contigs* grandes (VERLI, 2014). Um valor de N50 = n significa que 50% ou mais do genoma estão cobertos com contigs de tamanho n ou maior. Por exemplo, ao se analisar 4 *contigs* de comprimentos 1, 2, 4 e 5 kb, totalizando 12 kb, começa-se contabilizando a soma a partir do contig maior para o menor. Quando a soma ultrapassar metade do valor total dos *contigs*, o último *contig* a ser contado é determinado como o menor *contig* necessário para cobrir pelo menos 50% da sequência montada. No exemplo acima, N50 é igual a 4, pois é o menor *contig* necessário para cobrir mais da metade do genoma.

2.3.1 Ferramentas de montagens

Inicialmente, os softwares que foram criados para realizar a montagem de genomas tinham como foco somente realizar a montagem de um único genoma. SPades (BANKEVICH

et al., 2012), Velvet (ZERBINO; BIRNEY, 2008) e IDBA (PENG et al., 2010) são algumas das ferramentas de montagens inicialmente desenvolvidas para esse propósito.

Os montadores atuais trabalham com dois tipos de algoritmos principais: *OLC* (*Overlap Layout Consensus)* e Grafo *de Bruijn (de Bruijn Graph)* (MARTINS, 2013). Em um estudo de Zhang et al. (2011), foi observado que montadores que utilizam como algoritmo o OLC obtém melhores resultados quando utilizados para *datasets* de *reads* curtas ou grandes, de genomas menos complexos. Já montadores de Grafo *de Bruijn* se sobressaíram com *datasets* mais complexos, contendo um número muito mais alto de *reads.*

Namiki et al. (2012) afirma que montadores convencionais para genomas únicos não foram desenvolvidos para tratar genomas múltiplos de sequências de *reads* misturadas com diferentes níveis de cobertura. Logo, faz-se necessário o desenvolvimento de novas ferramentas que sejam capazes de manusear dados com o alto grau de complexidade como os dados metagenômicos.

Alguns dos montadores de genoma único (*Single Genome Assemblers)* receberam módulos específicos para lidar com dados metagenômicos. É o caso do MetaSPades (NURK et al., 2012), IDBA-UD (PENG et al., 2012) e MetaVelvet (NAMIKI et al., 2012). Além disso, o *Megahit* (LI et al., 2015) também é uma ferramenta desenvolvida para tratar dados metagenômicos de alta complexidade. O Megahit, quando comparado com o montador *Minia*, foi o que obteve melhor performance ao montar um *dataset* de 3.3 bilhões de *reads* retirados do solo de uma pradaria (LI et al., 2015). A montagem realizada no estudo de Howe et al. (2014) também foi incluída para efeitos de comparação:

Tabela 1 - Comparativo de Ferramentas de montagem

	MEGAHIT	Howe et al.	Minia
Tempo de Execução (h)	44.1	>488	331.4
Memória de pico (GB)	345	287	29
Tamanho total (Mbp)	4902	1503	1490
Comprimento médio (bp)	633	485	505
N50 (bp)	657	471	488
Maior (bp)	184 210	9397	32 679
# de contigs	7 749 211	3 096 464	2 951 575
# de contigs ≥ 1kbp	841 257	129 513	158 402

Fonte: Adaptado de Li et al. (2015).

A Tabela 1, retirada do estudo de Li, compara o desempenho do *Megahit* com outras ferramentas de montagem. Em tempo de execução, o *Megahit* teve um desempenho aproximadamente 7 vezes mais rápido do que o *Minia*, levando 44.1 horas para realizar a montagem. Também obteve melhores estatísticas com um N50 mais alto e um maior número de *contigs* de tamanho maior que 1000 pares de base, mostrando-se uma ferramenta ideal para tratar um alto volume de dados.

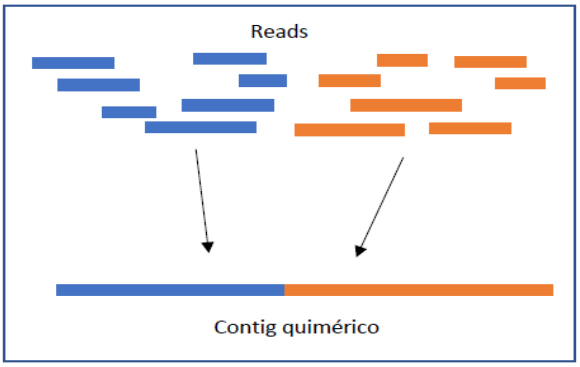
2.4 *BINNING* DE METAGENOMAS

Um dos desafios da Metagenômica é justamente a dificuldade inata ao próprio conceito de metagenoma: por ser uma amostra contendo múltiplas espécies, uma montagem completa de um genoma de uma determinada espécie é extremamente improvável.

Como Wooley (2010) relata, raramente se obtém sequências suficientes para se ter uma boa cobertura dos genomas presentes em uma amostra ambiental. A cobertura portanto permanece incompleta.

Além disso, quando se trabalha com o genoma de uma única espécie, há a garantia de que os *contigs* serão formados por *reads* pertencentes exclusivamente ao DNA cromossômico daquela espécie, com a ausência de contaminantes e de DNA extra cromossômico (WOOLEY, 2010). Na montagem de metagenomas, isso não ocorre. É muito comum que ocorra o alinhamento de *reads* de espécies diferentes gerando *contigs* interespécies denominados de quimera *(chimera*). Isso ocorre devido à presença de regiões repetidas nos genomas dos organismos, causando montagens incorretas. A Figura 5 retrata de forma simplificada a montagem de uma quimera através de *reads* de duas espécies diferentes:

Figura 5 - Montagem de um *contig* quimérico



Fonte: O autor (2018).

Percebe-se que somente a montagem de metagenomas não é suficiente para obtenção dos dados de espécies retirados de uma amostra ambiental. Essas dificuldades são inatas ao próprio processo de se trabalhar com dados de microorganismos diferentes ao mesmo tempo, algo que não se faz presente na análise do genoma de uma espécie única. Faz-se então necessário a realização de um passo além que permita ser feita, após a montagem, a separação dos genomas montados baseado em unidades taxonômicas, amenizando os possíveis erros de montagem ocorridos.

O termo *bin* é definido como "um contêiner ou espaço fechado para armazenamento" (THE FREE DICTIONARY, 2018). *Binning* é, portanto, o ato de armazenar, ou categorizar em *bins*. Na Metagenômica, o termo é utilizado para representar o processo de separação de sequências metagenômicas em unidades taxonômicas. As sequências são normalmente classificadas em unidades taxonômicas (e.g. Gênero, família, etc.) ou agrupadas em conjuntos de sequências que representam grupos taxonômicos baseados em alguma característica compartilhada (SHARPTON, 2014).

Ainda segundo Sharpton (2014), o *binning* tem um papel fundamental na análise de dados metagenômicos. Através do *binning* é possível:

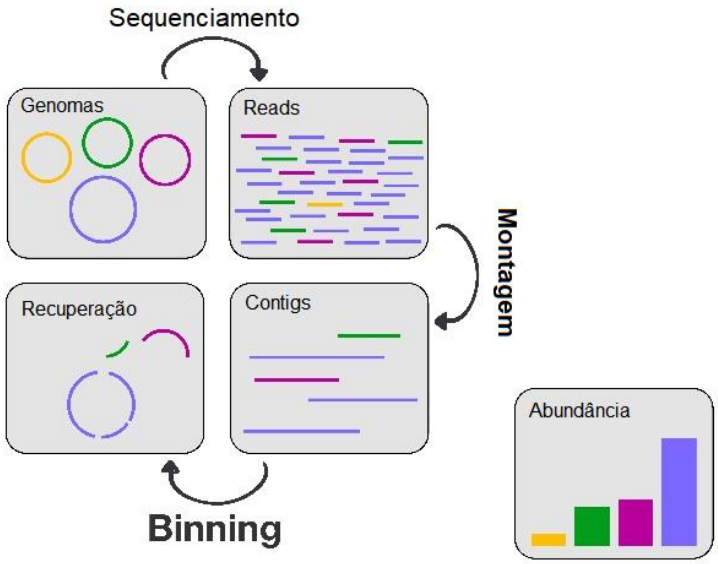
1. Obter informações de novos genomas ainda desconhecidos que são difíceis de identificar através de outros métodos;
2. Obter informações quanto aos tipos e quantidades de unidades taxonômicas presentes;
3. Diminuir a complexidade dos dados trabalhados em etapas pós-realização do *binning*. Caso a montagem seja realizada posteriormente ao *binning*, por exemplo, ela pode ser

realizada de maneira independente em cada um dos *bins* de *reads*, ao invés de em todo o conjunto de dados obtidos.

Como foi descrito acima, pode-se entender que o *binning* é uma etapa que pode ser conduzida antes ou após a realização da montagem. Entretanto, Sharpton (2014) e Sangwan, Xia e Gilbert (2016) reconhecem que a precisão dos algoritmos de *binning* aumenta na medida em que o comprimento das sequências também aumenta. Enquanto que a maioria dos *pipelines* inicialmente desenvolvidos para *binning* trabalham diretamente com *reads*, atualmente realiza-se o processo de montagem primeiro para formar *contigs*, que naturalmente terão maior comprimento que as *reads*.

O *pipeline* utilizado para a análise de metagenomas está sintetizado na Figura 6.

Figura 6 - *Pipeline* representando o processo de análise de metagenomas



Fonte: adaptado de SLIDESHARE (2018)

Na Figura 6, nos dois quadrados superiores tem-se a representação do processo de sequenciamento, isto é, de extrair a amostra de microorganismos e sequenciar seus genomas, gerando *reads* de curto comprimento. O tamanho dos círculos no primeiro quadrado representa a abundância de cada espécie na amostra: quanto maior o círculo maior a quantidade de microorganismos daquela espécie na amostra. A abundância também pode ser observada no quadrado do canto inferior direito. Os quadrados 2 e 3 indicam o processo de montagem, no qual as *reads* são unidas, formando *contigs*. Por fim, os últimos quadrados inferiores

representam o processo de *binning*, no qual os *contigs* são agrupados em *clusters* de acordo com as OTU's (*Operational Taxonomy Units*) presentes naquela amostra.

Como Mikheenko, Saveliev e Gurevich (2016) afirmam, a maioria dos estudos de metagenoma trabalha com montagens *de novo*, sem conhecimento prévio dos organismos estudados, tornando o processo particularmente desafiador. Ao se analisar uma amostra metagenômica, as sequências ali presentes pertencem a todos os organismos da amostra. Como, então, pode-se determinar que duas sequências pertencem ao mesmo organismo quando não se tem nenhuma informação prévia sobre aquelas sequências?

Existem duas técnicas principais utilizadas no *binning* para o agrupamento dos *clusters*: o *binning* por composição (*composition)* (DICK et al., 2009), e o *binning* por cobertura (*coverage*) (WU; YE, 2011).

2.4.1 *Binning* por composição

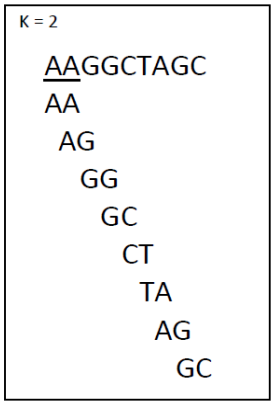
O *binning* por composição toma como premissa a análise das características da sequência baseando-se na composição de seus nucleotídeos. De acordo com Sangwan, Xia e Gilbert (2016), existem dois critérios utilizados nessa análise: a frequência de oligonucleotídeos e o conteúdo GC (*GC Content*) presente na sequência. Um oligonucleotídeo pode ser definido como uma subcadeia de nucleotídeos de uma sequência maior. Já o conteúdo GC é a porcentagem de bases Guanina e Citosina presentes na sequência.

O método consiste em comparar duas sequências de DNA, utilizando como critério a frequência dos oligonucleotídeos presentes na sequência. Em um estudo de Noble, Citek e Ogunseitan (1998), observou-se que existe uma correlação entre a frequência dos tetranucleotídeos (oligonucleotídeos de tamanho 4) e o genoma de determinados organismos. Determinadas regiões da sequência do DNA de um organismo permanecem conservadas no genoma de uma espécie. Essas regiões são denominadas *Sequencies Signatures*. O grau de relação entre dois *contigs* pode consequentemente ser determinado pela frequência com que esses tetranucleotídeos aparecem ao longo da cadeia de caracteres. *Contigs* cujas frequências dos tetranucleotídeos são semelhantes possuem alto grau de probabilidade de pertencerem à mesma unidade taxonômica.

O termo *k-mer* também é utilizado de maneira recorrente na literatura para se referir aos oligonucleotídeos. Um k-mer é definido como uma substring de comprimento k (ALNEBERG

et al., 2014). Na Figura 7, tem-se todos os k-mers de tamanho k = 2 para uma simples sequência de DNA de tamanho oito:

Figura 7 - Representação de k-mers de tamanho 2 de uma sequência de DNA



Fonte: O autor (2018).

O quadro 1, por sua vez, representa a quantidade de vezes em que cada dinucleotídeo aparece na sequência acima. É possível perceber que os dinucleotídeos ‘AG’ e ‘GC’ ocorrem duas vezes, enquanto que os demais (‘AA’, ‘TA’, ‘CT’ e ‘GG’) ocorrem somente uma vez:

Quadro 1 - Frequência dos dinucleotídeos

AG	2
GC	2
GG	1
CT	1
TA	1
AA	1

Fonte: O autor (2018)

Para dinucleotídeos existem um total de $2^4 = 16$ combinações diferentes. No caso de tetranucleotídeo, há 256 (4^4) possibilidades distintas de combinações dos quatro nucleotídeos.

2.4.2 *Binning* por cobertura

A maioria das ferramentas de *binning* utilizam o método por composição de DNA, porém, como afirmam Wu e Ye (2011), é necessário que se tenha sequências de grande comprimento para se obter resultados (no mínimo 800 pares de base).

Em seu estudo, Wu e Ye (2011) afirmam que amostras de metagenoma podem conter organismos com diferentes níveis de abundância. Ao se utilizar o método por análise da composição eventualmente podem ser gerados resultados imprecisos. Nesses casos, a análise por composição de nucleotídeos citada acima pode tornar-se inadequada.

> Procedimentos de sequenciamento shotgun randomizados resultam em amostragens desiguais de diferentes genomas, especialmente quando os níveis de abundância das espécies diferem. Deseja-se descobrir os valores de abundância, bem como os tamanhos do genoma automaticamente e, em seguida, realizar o *bin* nas *reads* por consequência (WU; YE, 2011, p.4)

No estudo de Wu e Ye, foi proposta uma nova ferramenta de *binning* denominada *AbundanceBin* (2011), voltada para realização do *binning* em sequências de curto comprimento e com diferentes níveis de abundância. Esta ferramenta agrupa os *bins* de acordo com a abundância de cada sequência na amostra: sequências agrupadas no mesmo *bin* são provenientes de espécies com abundância igual ou similar. Alneberg et al. (2014) também afirmam que as coberturas das sequências retratam a abundância dos organismos da amostra. Naturalmente, ao se sequenciar uma amostra metagenômica, espera-se que espécies mais abundantes produzam mais *reads*. Dessa forma, a cobertura pode ser utilizada para se estimar qual sequência pertence a qual genoma, e o tamanho dos genomas dos organismos da amostra.

Uma dificuldade relatada no estudo de Wu e Ye é que a ferramenta *AbundanceBin* classifica *reads* de diferentes espécies com abundâncias similares em um mesmo *bin*; ela não é capaz de separar *reads* de espécies com abundâncias similares. O autor afirma que a combinação de métodos de *binning* por cobertura junto com métodos de composição de DNA é capaz de gerar melhores resultados.

2.4.3 Ferramentas de *Binning*

Na atualidade, muitas ferramentas de *binning* se utilizam de uma combinação de técnicas para melhor performance. Quanto ao aspecto de composição de nucleotídeos, CONCOCT (ALNEBERG et al., 2014) GroopM (IMELFORT et al., 2014), *MaxBin* (WU, SIMMONS; SINGER, 2015) e *MetaBAT* (KANG et al., 2015) são ferramentas que se baseiam na frequência de tetranucleotídeos. Estas ferramentas também se utilizam da abordagem de cobertura dos *contigs* para otimização dos resultados.

Quanto ao modelo de cobertura de sequência, segundo Sangwan, Xia e Gilbert (2016), cada ferramenta adota uma abordagem matemática diferente. Enquanto o *MaxBin* se baseia na distribuição de Poisson para calcular a cobertura das sequências, *METABAT* calcula a distância probabilística entre dois *contigs* utilizando uma função de distribuição Normal. De acordo com o autor ainda é incerto qual abordagem traz melhores resultados.

2.5 USO DE DADOS SIMULADOS NA ANÁLISE DE METAGENOMAS

A complexidade de dados metagenômicos e o fato de ainda ser uma área recente traz alguns desafios, sendo um deles a validação da performance das ferramentas e dos métodos utilizados na análise dos dados. Nurk et al. (2012) afirma que testar e validar as ferramentas de metagenoma quanto ao seu desempenho é uma tarefa de alta complexidade, devido a ausência do conhecimento da composição de comunidades microbianas.

Uma forma de validar ferramentas de metagenomas, tanto de montagem quanto de *binning*, é através da utilização de dados simulados, ou sintéticos. Essa afirmação corrobora com Mende et al. (2012), que afirmam que a utilização de dados simulados é o único jeito possível de se validar ferramentas devido a ausência de genomas completamente anotados.

De acordo com Nurk et al. (2012), *datasets* simulados podem ser obtidos através do sequenciamento de bactérias misturadas com genomas conhecidos; obtidos através de dados sequenciados isoladamente, ou simulados a partir de sequências de referência.

Em seu trabalho, Mavromatis et al. (2007) desenvolveu três diferentes conjuntos de dados metagenômicos simulados com diferentes complexidades para validar ferramentas de montagem, predição de gene e *binning*. Esses dados foram construídos através da combinação de *reads* selecionadas aleatoriamente de um conjunto de 113 genomas.

Mende et al. (2012), observou em outro estudo a qualidade das montagens de metagenomas com *reads* sequenciadas por NGS (Illumina e Pirosequenciamento) em comparação com *reads* obtidas através do sequenciamento Sanger. Para isso, foram criados dois simuladores capazes de recriar dados de *reads* de metagenomas: O *iMess* (*interactive MEtagenomic Simulation Software*) para sequenciamento Sanger e pirosequenciamento e o *iMessi*, para sequenciamento Illumina.

Estes simuladores são capazes de reproduzir um ambiente semelhante ao de dados metagenômicos reais, reproduzindo também os erros e os valores de qualidade baseados em dados reais, e estão disponíveis para uso gratuitamente na internet.

Para cada tipo de sequenciamento, Mende et al. (2012) criaram três conjuntos de dados de diferentes complexidades de 10, 100, e 400 genomas para avaliar a qualidade das montagens de metagenomas realizadas nestes dados simulados.

Tabela 2 - Conjuntos de dados simulados gerados pelo *iMess*

Metagenomas Simulados	**MG1**	**MG2**	**MG3**	**MG4**	**MG5**	**MG6**	**MG7**	**MG8**	**MG9**
Tecnologia de Sequenciamento	**Illumina**			**Sanger**			**Pirosequenciamento**		
Número de Genomas	10	100	400	10	100	400	10	100	400
Número de Reads (Milhão)	53.33	53.33	53.33	0.25	0.25	0.25	1.00	1.00	1.00
Tamanho da Sequência (Mb)	4000	4000	4000	200	200	200	255	255	255
Comprimento médio das reads (bp)	75	75	75	800	800	800	255	255	255

Fonte: Adaptado de Mende et al. (2012).

A Tabela 2 compara os três conjuntos de dados (10, 100 e 400 genomas) das três tecnologias diferentes de sequenciamento. Valores como o tamanho médio das *reads* geradas refletem as características de cada tecnologia. *Reads* Illumina possuem um comprimento muito menor em comparação a *reads* Sanger. Em contrapartida, a quantidade de *reads* geradas pelo sequenciamento Illumina é cerca de 200 vezes maior que o Sanger, e 53 vezes maior que as *reads* de pirosequenciamento. Todos os dados acima foram gerados pelos simuladores *iMess* e *iMessi*.

Como se pode perceber, a ausência de informações sobre comunidades microbianas faz com o que o uso de dados simulados se torne imprescindível para a validação dos métodos e softwares utilizados na Metagenômica.

Muitos estudos se utilizam de dados simulados para testar a eficiências de novas ferramentas desenvolvidas. Namiki et al. (2012) utilizaram dois conjuntos de *datasets*, um simulado e outro real retirado do intestino humano para comparar o desempenho do montador MetaVelvet com outros três montadores: Velvet, SOAPdenovo e Meta-IDBA.

O *MyCC* (LIN; LIAO, 2016) é uma ferramenta de *binning* que também se utilizou de de dados simulados para comparar o seu desempenho junto a outros três softwares de *binning*: CONCOCT, *MaxBin* e *MetaBAT*. No estudo de Lin e Liao (2016), foram utilizados os dados simulados de 10 e 100 genomas retirados do trabalho de Mende et al. (2012).

Garcia-Etxebarria, Garcia-Garcerà e Calafell (2014) também utilizaram o simulador *iMess* para gerar dados sintéticos e avaliar a eficácia de diferentes métodos de classificação de dados metagenômicos.

Para que seja possível analisar a eficácia das ferramentas utilizadas no *pipeline* de metagenomas, é preciso utilizar um software que possa fazer a comparação entre os resultados dos softwares de montagens e *binning*, produza análises estatísticas, gere gráficos, dentre outras funcionalidades. Mikheenko, Saveliev e Gurevich (2016) afirmam que enquanto existem muitas ferramentas de análise de genomas bem estabelecidas e conhecidas, não há ferramentas voltadas para o ambiente específico de metagenomas. O *MetaQUAST* (MIKHEENKO; SAVELIEV; GUREVICH, 2016) foi então desenvolvido para sanar essa deficiência.

2.6 METAQUAST: UMA FERRAMENTA PARA ANÁLISE E VISUALIZAÇÃO DE DADOS METAGENÔMICOS

O MetaQUAST (*Quality Assessment Tool for Metagenome Assemblies*) é uma ferramenta voltada para a avaliação e comparação de dados metagenômicos utilizando genomas de referência (MIKHEENKO; SAVELIEV; GUREVICH, 2016). É uma extensão do Quast, inicialmente desenvolvido para detecção de erros através do alinhamento a um genoma conhecido de referência. Para tratar dados metagenômicos, foram implementadas mudanças como a possibilidade se utilizar uma quantidade ilimitada de genomas de referência e a detecção de *contigs* quiméricos (Ibid.).

Como o *MetaQuast* trabalha primordialmente com genomas de referência para analisar *contigs* que foram gerados, e sabendo-se que ainda há pouca informação quanto às espécies microbianas presentes em metagenomas, o uso de dados simulados em conjunto com o *MetaQuast* é altamente recomendável.

O software é capaz de produzir relatórios e dados estatísticos referentes aos *contigs* que foram submetidos como entrada.

Quadro 2 - Exemplo de tela de visualização do MetaQUAST

Estatísticas sem referência	IDBA_UD	Ray	SOAPdenovo2	SPAdes
# contigs	31 224	10 327	36 468	40 546
Maior contig	305 144	99 107	40 707	189 063
Comprimento total	80 325 286	30 411 921	46 741 224	92 397 329
Comprimento total (>= 1000 bp)	69 223 529	27 080 646	30 720 336	77 823 828
Comprimento total (>= 10000 bp)	34 930 908	13 755 677	2 800 864	33 477 263
Comprimento total (>= 50000 bp)	16 008 349	2 346 322	0	11 409 912
Erros de montagem				
#Erros de montagem	1132	407	831	1240
#Comp. de contigs com erros de montagem	10 448 260	4 115 772	911 826	10 780 557
Mismatches				
# mismatches por 100 kbp	904.95	1054.68	888.21	1401.84
# indels por 100 kbp	31.88	27.7	17.09	51.64
# N's por 100 kbp	238.48	2087.27	3730.51	1425.14
Estatísticas de genoma				
Fração de genoma (%)	12.796	4.386	8.055	11.585
Akkermansia_muciniphila_ATCC	0.003	-	-	0.011
Alistipes_putredinis	1.366	0.595	0.61	1.117
Anaerotruncus_colihominis	2.466	2.067	1.768	2.320
Bacteroides_caccae	5.343	2.643	3.928	5.138
Bacteroides_capillosus	1.173	0.27	0.449	1.05
Bacteroides_cellulosilyticus	1.278	0.952	1.524	0.96
Bacteroides_coprocola	30.532	-	-	-

Fonte: adaptado de Mikheenko, Saveliev e Gurevich (2016).

O Quadro 2 retrata um dos relatórios emitidos pelo metaQUAST, com as diferentes informações referente a quatro montadores. É possível observar a quantidade de *contigs* gerados, tamanho do maior *contig*, tamanho total, quantidade do genoma coberto para (*Genome Fraction*) para cada organismo, dentre outros.

3 MATERIAIS E MÉTODOS

Para este estudo, inicialmente foi realizada uma revisão bibliográfica dos trabalhos que abordavam os conceitos básicos sobre metagenomas, montagens e *binnings*, juntamente com as ferramentas mais comumente utilizadas no *pipeline* de metagenomas.

Para a realização dos testes comparativos de ferramentas de *binning*, foram utilizados os conjuntos de dados de 10 e 100 espécies de bactérias, retirados do artigo *Assessment of metagenomic assembly using simulated next generation sequencing data* (MENDE et al., 2012) *(*Avaliação de montagens metagenômicas utilizando dados simulados de sequenciamento de próxima geração, tradução livre). Estes *datasets* foram gerados através do simulador de dados metagenômicos *iMess (interactive MEtagenomic Simulation Software).* Esses conjuntos de dados são arquivos de texto no formato fastq (formato comumente utilizado para representar sequências de nucleotídeos) contendo as *reads* não tratadas de bactérias. No artigo citado acima, foram geradas *reads* sequenciadas pelo método de Sanger, pelo método de pirosequenciamento e pelo sequenciamento Illumina. Os *dataset* para os três tipos de sequenciamento podem ser encontrados no link *www.bork.embl.de/~mende/simulated_data/.*

Neste estudo foram utilizados os *datasets* de sequenciamento Illumina, contendo *reads* de comprimento médio de 75 pares de base. O tratamento de qualidade foi realizado pela ferramenta PRINSEQ, e em seguida, as *reads* foram submetidas ao processo de montagem. Para esta etapa, três softwares montadores foram utilizados: *IDBA_UD*, *Megahit*, e *MetasSPAdes*, com valores *default* de k-mer (mínimo 20 e máximo 100 para o *IDBA_UD*; mínimo 21 e máximo 99 para o *Megahit* e 21, 33 e 55 para o *MetaSPades*). Em seguida, foi utilizada a ferramenta *Metaquast* para analisar a qualidade da montagem comparando os *contigs* montados com o genoma de referência das espécies.

Através de análises estatísticas no *Metaquast,* como maior cobertura de genoma pelos *contigs* montados e menor número de erros de montagens (*missassemblies*), é definido qual montador gerou o melhor resultado de montagem. Este resultado foi em seguida submetido a duas ferramentas de *binning* diferentes: *MetaBat* e *MaxBin.* Cada software de *binning* produz como saída uma quantidade de arquivos *bins*. O desejável é que esse número seja equivalente ao número de espécies presentes na amostra. Se a amostra possui 10 espécies, espera-se alcançar um resultado cuja saída seja 10 arquivos de *bins* diferentes, cada um representando uma espécie.

Na etapa de *binning*, inicialmente as ferramentas foram rodadas com parâmetros *default*, e posteriormente com algumas mudanças em parâmetros como tamanho do *contig* para tentar alcançar melhores resultados.

Para o conjunto de dado de 10 espécies, o *MaxBin* foi rodado uma única vez e o *MetaBAT* duas vezes: uma sem um arquivo de *depth (*arquivo que contém informação da abundância das espécies na amostra; este arquivo é gerado separadamente) e posteriormente com o arquivo de *depth*.

Para o conjunto de dados de 100 espécies, o *MaxBin* foi rodado duas vezes, uma com os parâmetros *default* e outra com os seguintes parâmetros:

- Tamanho mínimo de *contig* igual a 200;
- *Marker Set* igual a 40.

O *MetaBAT* foi executado três vezes: a primeira execução foi realizada com a configuração padrão da ferramenta. A segunda com os arquivos de *depth* e tamanho mínimo de *bin* de 130000, e a terceira com o arquivo de *depth* e tamanho mínimo de *bin* de 12000. O motivo pelos quais esses parâmetros foram escolhidos serão explicados posteriormente na análise dos resultados.

Cada um dos resultados gerados pelas ferramentas de *binning* foi novamente enviado ao *Metaquast* para avaliação. Para se comparar a eficácia das ferramentas, é necessário ter os genomas de referências das espécies presentes na amostra. Depois de submeter ao *Metaquast* foi possível observar a quantidade de *bins* gerados, qual o grau de completude de cada espécie e a quantidade total de alinhamento realizado.

Entretanto, somente analisar a quantidade de *bins*, ou o grau de cobertura de cada *bin* alinhado ao genoma das espécies não é o suficiente para determinar a eficiência das ferramentas. Considerando-se que o resultado ideal seria um *bin* separado para cada espécie, é importante observar se as seguintes situações ocorreram:

1. Um determinado *bin* alinhou com mais de uma espécie diferente;
2. Dois ou mais *bins* fazem referência a uma mesma espécie;

Levando em consideração que quanto maior a complexidade da amostra, mais improvável que se ocorra a separação correta das espécies, foi criada uma Matriz de Confusão para analisar o desempenho de cada ferramenta ao categorizar uma espécie por *bin*.

A Matriz de Confusão é uma técnica do Aprendizado de Máquinas que permite analisar estaticamente o desempenho de um determinado modelo. Neste caso, deseja-se saber se cada *bin* de fato está sendo mapeado a uma única espécie.

Inicialmente, assumiu-se que cada *bin* corresponde a uma única espécie. Caso um *bin* se alinhe com várias espécies diferentes, calcula-se a espécie com o maior número de *contigs* presentes no *bin*. Essa espécie corresponderá, portanto, àquele *bin*. O Quadro 3 ilustra este processo:

Quadro 3 - *Bins* alinhados ao organismo *Neisseria meningitidis MC58*

Alinhado a"NC_003112.2_Neisseria_meningitidis_MC58_chromosome_complete_genome" |

pior mediano melhor Show heatmap

Estatísticas de Genoma	idbaMetaBAT3.3	idbaMetaBAT3.7
Fração de Genoma (%)	0.187	77.107
Proporção de duplicações	1	1.003
Maior alinhamento	4246	49 978
Tamanho total alinhado	4246	1 756 485
NGA50	-	9572
LGA50	-	68
Erros de montagem (misassemblies)		
# Erros de montagem	0	1
Comprimento dos contigs com erros de montagem	0	15 576
Mismatches		
# mismatches por 100 kbp	0	3.65
# indels por 100 kbp	0	0.57
# N's por 100 kbp	0	0
Estatísticas sem referência		
# contigs	1	170
Maior contig	4246	49 978
Comprimento total	4246	1 757 005
Comprimento total (>= 1000 bp)	4246	1 757 005
Comprimento total (>= 10000 bp)	0	1 106 222
Comprimento total (>= 50000 bp)	0	0

Fonte: Adaptado de MetaQUAST (2018).

Na parte inferior do quadro 3 na linha abaixo de *Statistics without reference* (Estatísticas sem referência), tem-se o número de *contigs* presentes em cada *bin* para a espécie *Neisseria meningiditis MC58*. O *bin* idbaMetaBAT3.7 possui 170 *contigs*, enquanto que idbaMETABAT3.3 possui somente 1. Portanto, o *bin* idbaMetaBAT3.7 é mapeado para *Neisseria meningiditis MC58* por ter a maior correspondência de *contigs* com o organismo.

Após realizado o mapeamento individual de cada organismo para cada *bin*, foi contabilizada a quantidade de *contigs* que caem nas seguintes quatro situações da matriz:

Quadro 4 - Matriz de confusão

	POSITIVO	**NEGATIVO**
VERDADEIRO	VERDADEIRO POSITIVO (VP): *Contig* está no *bin* da espécie *x* e pertence de fato à espécie *x.*	VERDADEIRO NEGATIVO (VN): *Contig* não está no *bin* da espécie *x* e não pertence de fato à espécie *x*
FALSO	FALSO POSITIVO (FP): *Contig* está no *bin* da espécie *x,* mas não pertence à espécie *x*.	FALSO NEGATIVO (FN): *Contig* não está no *bin* da espécie *x,* mas pertence à espécie *x.*

Fonte: O autor (2018).

A Matriz de Confusão foi calculada individualmente para cada *bin*. Em seguida, cada *bin* foi avaliado segundo as métricas de Precisão, Acurácia, Sensibilidade e Especificidade. Para obter a performance geral de cada ferramenta, calculou-se a média das métricas individuais dos *bins*. Por fim, pode-se comparar o desempenho geral das diferentes execuções das ferramentas de *binning*.

As métricas podem ser entendidas da seguinte forma:

Precisão: indica se há presença de contaminantes (*contigs* de outras espécies) em um *bin*. Quantos *contigs* presentes em um *bin* são da espécie desejada?

Fórmula: $PRECISÃO = VP/ (VP + FP)$

Sensibilidade: Frequência que a ferramenta é capaz de colocar no mesmo *bin* os *contigs* de uma espécie. Considerando todos os *contigs* de uma espécie, quantos deles foram classificados no *bin* correto? Fórmula: $SENSIBILIDADE = VP / (VP + FN)$

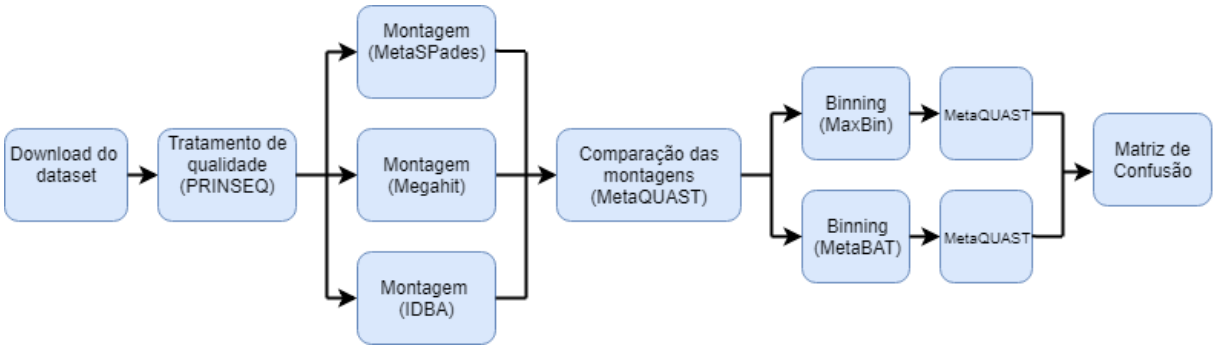
Especificidade: calcula a frequência que a ferramenta é capaz de colocar em outros *bins* os *contigs* que não pertencem à espécie analisada. Dos *contigs* que não pertencem a espécie analisada, quantos deles realmente foram categorizados em outros *bins*?

Fórmula: $ESPECIFICIDADE = VN / (VN + FP)$

Acurácia: indica no geral o quão correta foi feita a classificação pela ferramenta. Ela é calculada através da soma dos Verdadeiros Positivos e Verdadeiros Negativos, dividido pelo total. Fórmula: *ACURÁCIA = (VP+ VN) / (VP + FP + VN + FN)*

Para melhor compreensão, o *pipeline* utilizado na análise dos *binning* de metagenomas pode ser descrito de maneira geral através do fluxograma da Figura 8:

Figura 8 - Representação do *pipelin*e para comparação de ferramentas de *binning*



Fonte: o Autor (2018).

1. Download do conjunto de *reads*;
2. Execução do tratamento de qualidade dos dados baixados;
3. Montagem do *dataset* pelos softwares montadores (IDBA, Megahit e Metaspades);
4. Comparação das montagens utilizando o software *Metaquast*;
5. Melhor resultado escolhido na etapa anterior é submetido aos *softwares* de *binning* (*MetaBAT* e *MaxBin*);
6. Para cada uma das saídas do software de *binning*, é gerado um novo relatório no *Metaquast*;
7. Os melhores resultados de cada ferramenta de *binning* são submetidos à matriz de confusão para efeitos comparativos;

Todos os programas foram rodados em uma máquina de sistema operacional Linux Ubuntu, 64 bits, com 8GB de Memória RAM.

4 RESULTADOS E DISCUSSÕES

Primeiramente, para melhor entendimento, foram estabelecidas algumas convenções quanto às nomenclaturas utilizadas para representar as diferentes execuções das ferramentas.

Os arquivos de *bins* gerados recebem o nome do software montador, seguido do nome da ferramenta de *binning* utilizada (e.g. idbaMetaBAT). Caso uma mesma ferramenta tenha sido executada n vezes, com diferentes parâmetros, representa-se as diferentes execuções com números (e.g. idbaMetaBAT1, idbaMetaBAT2 etc).

4.1 COMPARAÇÃO DAS MONTAGENS

Previamente à realização do *binning*, foi necessário averiguar qual montador obteve um melhor desempenho. O quadro 5, retirado do *MetaQUAST* compara o desempenho de três montadores para a montagem do conjunto de dados de 10 espécies:

Quadro 5 - Comparação dos montadores para conjunto de dados de 10 espécies

Estatísticas de Genoma	IDBA_contig	megahit_contigs	spades_contigs
Fração de Genoma (%)	99.314	99.108	99.051
Proporção de duplicações	1.002	1.002	1.001
Maior alinhamento	1 190 435	1 109 460	633 416
Comprimento total de alinhamento	32 939 637	32 929 197	32 888 301
NGA50	...	...	...
LGA50	...	...	...
Erros de Montagem (misassemblies)			
# Erros de montagem	6	25	13
Comprimento dos contigs com erros de montagem	1 020 007	1 711 948	622 409
Mismatches			
# mismatches por 100 kbp	3.17	4.47	8.36
# indels por 100 kbp	0.23	0.39	0.95
# N's por 100 kbp	0	0	0
Estatísticas sem referência			
# contigs	837	845	817
Maior contig	1 190 639	1 109 460	633 416
Comprimento total	33 713 530	33 704 499	33 661 055
Comprimento total (>= 1000 bp)	33 637 402	33 637 049	33 608 598
Comprimento total (>= 10000 bp)	32 247 795	32 157 122	32 270 292
Comprimento total (>= 50000 bp)	26 672 292	26 599 725	26 195 485

Fonte: Adaptado de MetaQUAST (2018).

Quando alinhado com os genomas de referência das espécies presentes nos dados simulados, o IDBA foi o que obteve melhor cobertura do genoma (*Genome Fraction*) cobrindo um total de 99.3% dos genomas (Quadro 3).

Já o Quadro 6 retrata os mesmos três montadores para o conjunto de dados de 100 espécies:

Quadro 6 - Comparação dos montadores para conjunto de dados de 100 espécies

Estatísticas de Genoma	IDBA_contig	megahit_final.contigs	spades_contigs
Fração de Genoma (%)	50.912	62.717	62.873
Proporção de duplicações	1.007	1.008	1.01
Maior alinhamento	133 685	195 515	144 343
Comprimento total de alinhamento	146 483 092	187 761 413	188 148 203
NGA50	...	...	...
LGA50	...	...	...
Erros de Montagem (misassemblies)			
# Erros de montagem	4507	2766	8925
Comprimento dos contigs com erros de montagem	6 342 356	4 105 226	13 302 933
Mismatches			
# mismatches por 100 kbp	252.45	201.33	269.8
# indels por 100 kbp	9.36	6.77	9.9
# N's por 100 kbp	0	0	0
Estatísticas sem referência			
# contigs	148 330	185 742	183 858
Maior contig	133 685	195 515	144 590
Comprimento total	156 407 370	200 919 170	201 861 317
Comprimento total (>= 1000 bp)	84 109 283	114 907 905	117 190 382
Comprimento total (>= 10000 bp)	11 173 869	12 241 189	12 965 379
Comprimento total (>= 50000 bp)	1 493 172	2 184 831	2 441 224

Fonte: Adaptado de MetaQUAST (2018).

Para o conjunto de 100 espécies, os montadores Megahit e metaSPAdes cobriram uma fração do genoma semelhante. Entretanto, o Megahit obteve uma quantidade menor de erros de montagem (*misassemblies*), se sobressaindo, portanto, ao metaSPAdes.

Dessa forma, para os dados de 10 e 100 espécies foram escolhidas as montagens realizadas pelo IDBA e Megahit, respectivamente.

4.2 *BINNING* – 10 ESPÉCIES

A quantidade de *bins* gerados por cada execução das ferramentas e os parâmetros utilizados para o *dataset* de 10 espécies podem ser visualizados de maneira sintetizada no quadro 7:

Quadro 7 - Quantidade de bins gerados para *dataset* de 10 Espécies

	IdbaMaxBin	**IdbaMetaBAT1**	**IdbaMetaBAT2**
BINS	10	11	10
PARÂMETROS	*default*	*default;* *sem arquivo de depth*	*default;* *com arquivo de depth*

Fonte: O autor (2018).

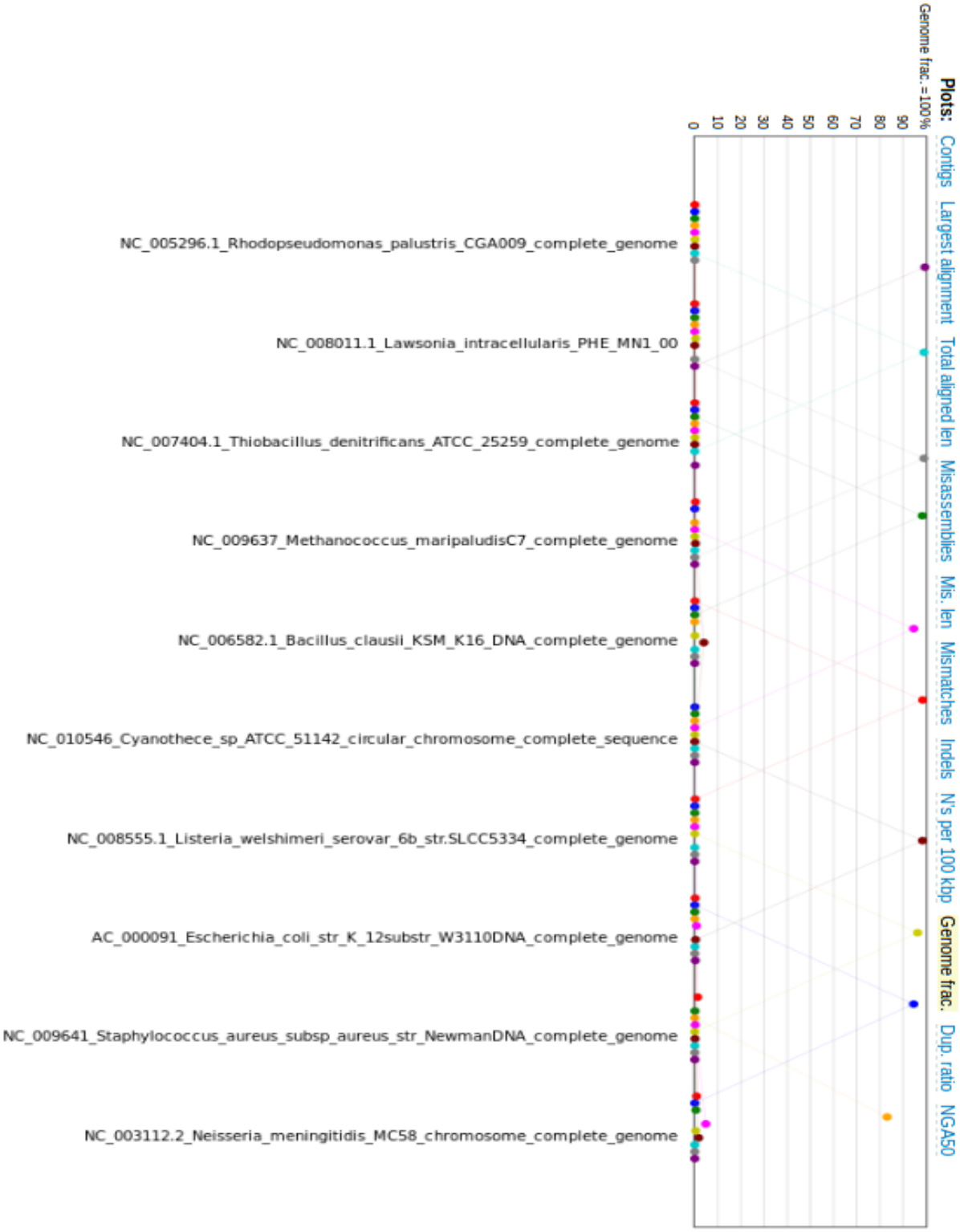
4.2.1 10 Espécies - *MaxBin* (idbaMaxBin)

No conjunto de dados de 10 espécies, o *MaxBin* foi capaz de gerar 10 arquivos *fasta* separados. Para rodar, o *MaxBin* necessita obrigatoriamente de um arquivo contendo a abundância dos organismos da amostra, ou de um arquivo contendo as *reads* sequenciadas.

Para esta ferramenta, foi utilizado o arquivo de 10 espécies contendo as *reads* sequenciadas. Através do arquivo de *reads*, o *MaxBin* é capaz de gerar o arquivo de abundância das espécies utilizando o *BowTie2* (LANGMEAD; SALZBERG, 2012), uma ferramenta que alinha as *reads* com sequências de referência e calcula a cobertura de cada *contig*.

Gráfico 1 - Gráfico de Fração de genoma para idbaMaxBin

.



Fonte: MetaQUAST (2018).

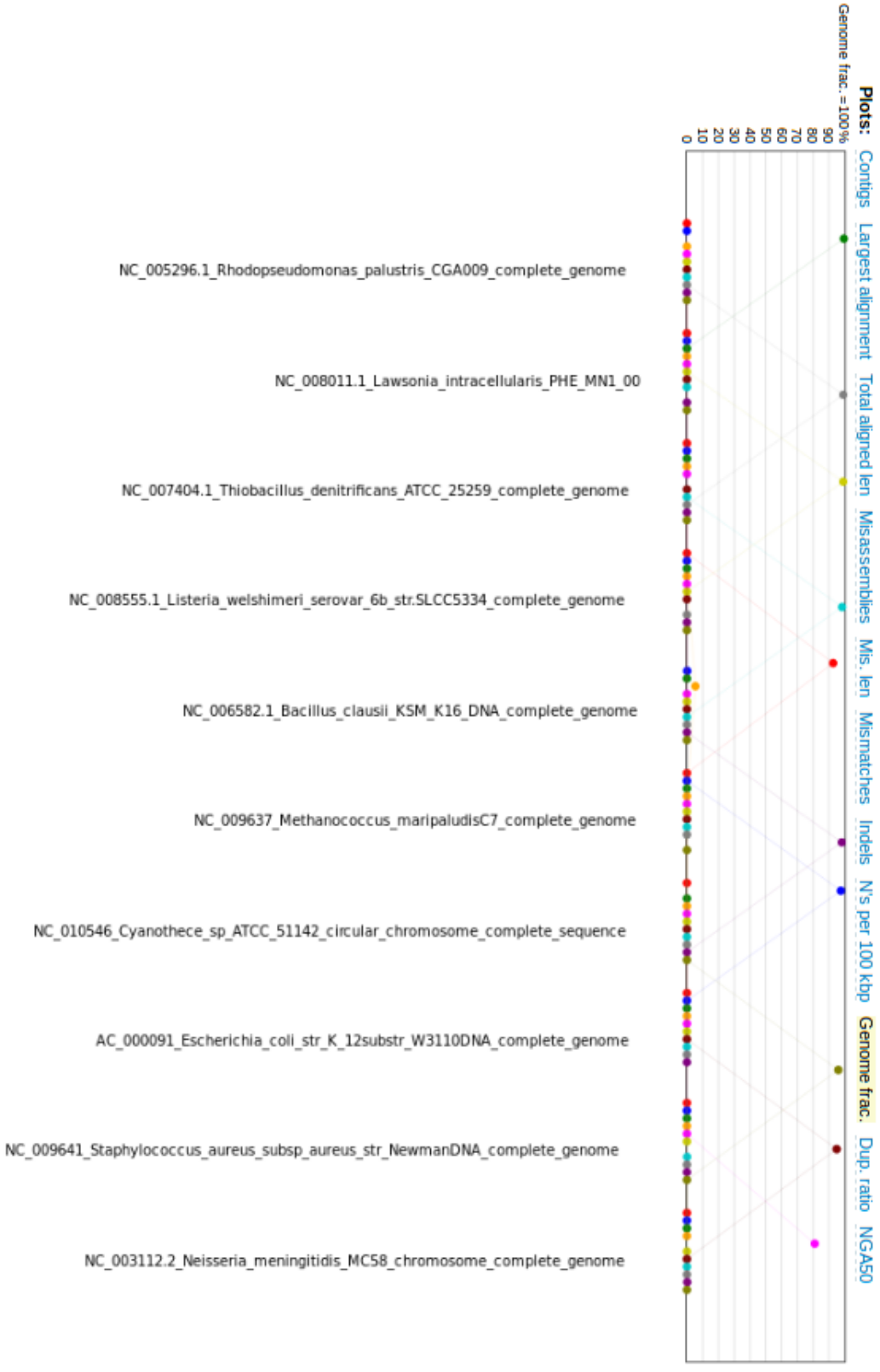
O Gráfico 1 representa o grau de equivalência de cada *bin* com cada genoma de referência que foi submetido ao MetaQUAST, no qual cada círculo de cor diferente representa um *bin* distinto. O *MaxBin* obteve bons resultados, obtendo acima de 94% de equivalência com 9 das 10 espécies presentes. Com a espécie *Neisseria meningitidis MC58*, o grau de equivalência foi de 83%.

4.2.2 10 Espécies - *MetaBAT*

4.2.2.1 IdbaMetaBAT1

Inicialmente, por dificuldades técnicas quanto a geração do arquivo de abundância das espécies (chamado de *depth*), o *MetaBAT* foi executado somente com *input* dos *contigs* do IDBA, algo que o *MaxBin* não permite. Nestas configurações, foram gerados 11 *bins*. Verificou-se pelo MetaQUAST que 10 desses 11 *bins* obtiveram uma correspondência semelhante à execução do *MaxBin*, como mostra o Gráfico 2:

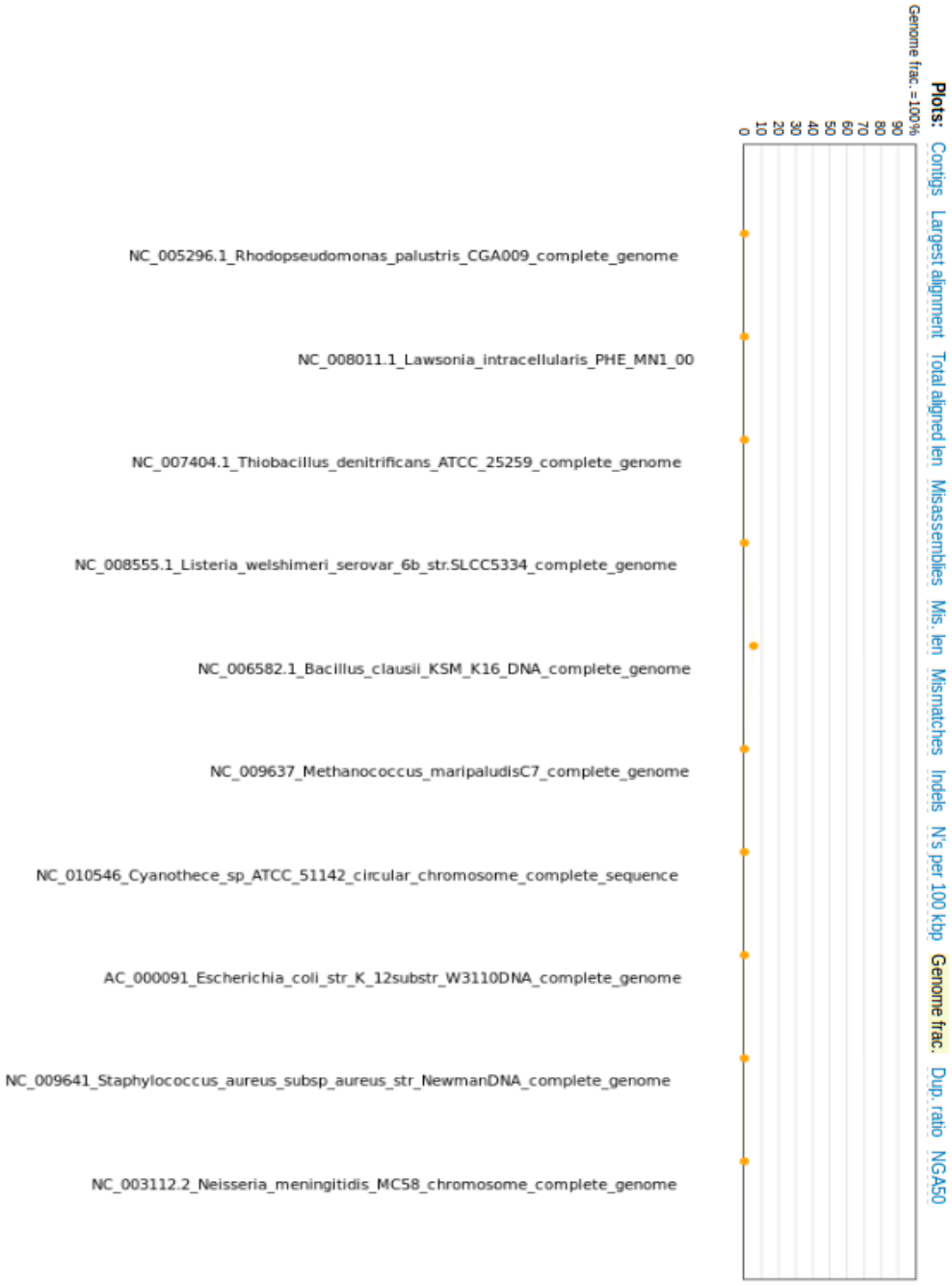
Gráfico 2 - Gráfico de Fração de Genoma para idbaMetaBAT1



Fonte: MetaQUAST (2018).

Entretanto, um *bin* adicional gerado não obteve correspondência com nenhuma das 10 espécies presentes na mostra. Este *bin* (idbaBin2) está representado no Gráfico 3:

Gráfico 3 - Gráfico de *bin* isolado gerado pelo MetaBAT



Fonte: MetaQUAST (2018)

Cada um dos pontos acima representa a correspondência deste *bin* (idbaBin2) com as 10 espécies. Através do gráfico é possível observar que só existe correspondência com uma única espécie (*Bacillus Clausii KSM K16*), e este valor é abaixo de 10%, um valor extremamente baixo.

No quadro 8, estão exibidos os *bins* alinhados com o organismo *Bacillus Clausii KSM K16*. É possível observar que o idbaBin2 cobre somente 5.3% do genoma do organismo.

Quadro 8 - Alinhamento dos *bins* com *Bacillus Clausi KSM K16*

Alinhado a "NC_006582.1_Bacillus_clausii_KSM_K16_DNA_complete_genome"

pior mediano melhor Mostrar mapa de calor

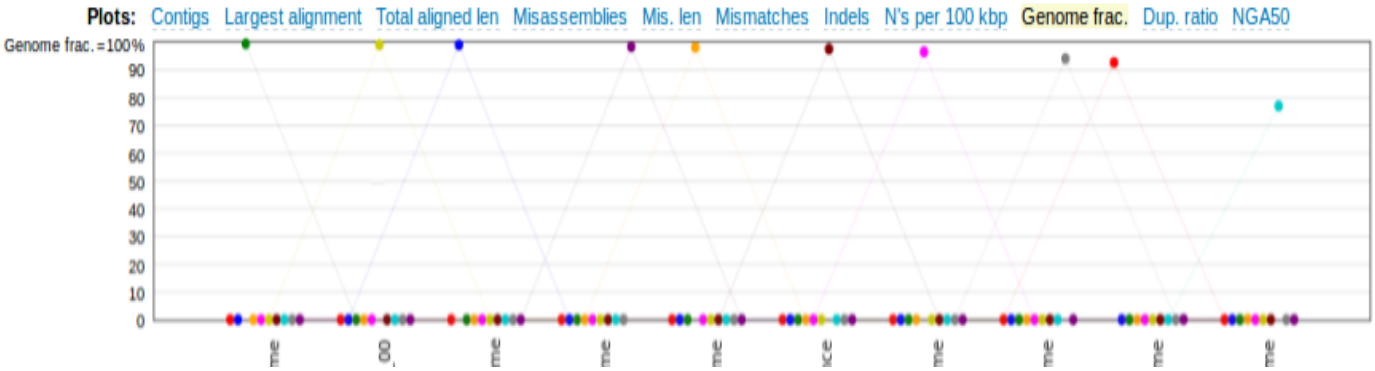
Estatísticas de Genoma	idbaBin.10	idbaBin.11	idbaBin.2
Fração de Genoma (%)	92.759	0.16	5.396
Proporção de duplicações	1	1	1.001
Maior alinhamento	1 109 621	3604	45 456
Tamanho total alinhado	3 992 379	6897	232 242
NGA50	326 749	-	-
LGA50	4	-	-
Erros de montagem (misassemblies)			
# Erros de montagem	0	0	1
Comprimento dos contigs com erros de montagem	0	0	25 168
Mismatches			
# mismatches por 100 kbp	0.08	0	4.31
# indels por 100 kbp	0.03	0	0
# N's por 100 kbp	0	0	0
Estatísticas sem referência			
# contigs	23	2	8
Maior contig	1 109 621	3604	45 456
Comprimento total	3 992 379	6897	232 466
Comprimento total (>= 1000 bp)	3 992 379	6897	232 466
Comprimento total (>= 10000 bp)	3 979 349	0	225 376
Comprimento total (>= 50000 bp)	3 876 629	0	0

Fonte: Adaptado de MetaQUAST (2018).

4.2.2.2 IdbaMetaBAT2

Posteriormente, o arquivo de *depth* foi gerado e o *MetaBAT* foi novamente executado, desta vez com o arquivo de abundância. Com o arquivo de *depth*, o *metaBAT* foi capaz de separar os *contigs* em 10 *bins*. O *MetaBAT* também foi capaz de obter cobertura acima de 90% de 9 das 10 espécies presentes na amostra, como mostra o gráfico 4:

Gráfico 4 - Gráfico de fração do genoma para idbaMetaBAT2



Fonte: MetaQUAST (2018).

Ainda que alguns *bins* tenham alinhado com mais de um organismo, o idbaMetaBAT2 produziu um número de *bins* correspondente à quantidade de espécies presentes na amostra. Pode-se afirmar que cada *bin* foi mapeado a uma única espécie. Esse fato comprova a importância de se utilizar um arquivo de cobertura nas ferramentas de *binning*.

4.3 *BINNING* - 100 ESPÉCIES

De forma análoga aos resultados de 10 espécies, o quadro 9 representa os diferentes resultados e parâmetros utilizados para cada uma das execuções do *MaxBin* e *MetaBAT*:

Quadro 9 - Quantidade de *bins* gerados para *dataset* de Cem Espécies

	MegahitMaxBin1	**MegahitMaxBin2**	**MegahitMetaBAT1**	**MegahitMetaBAT2**	**MegahitMetaBAT3**
BINS	33	76	15	24	96
PARÂMETROS	*default*	*min-_contig_length = 200; markerset = 40*	*default; com arquivo de depth*	*s = 130000; com arquivo de depth*	*s = 12000; com arquivo de depth*

Fonte: O autor (2018).

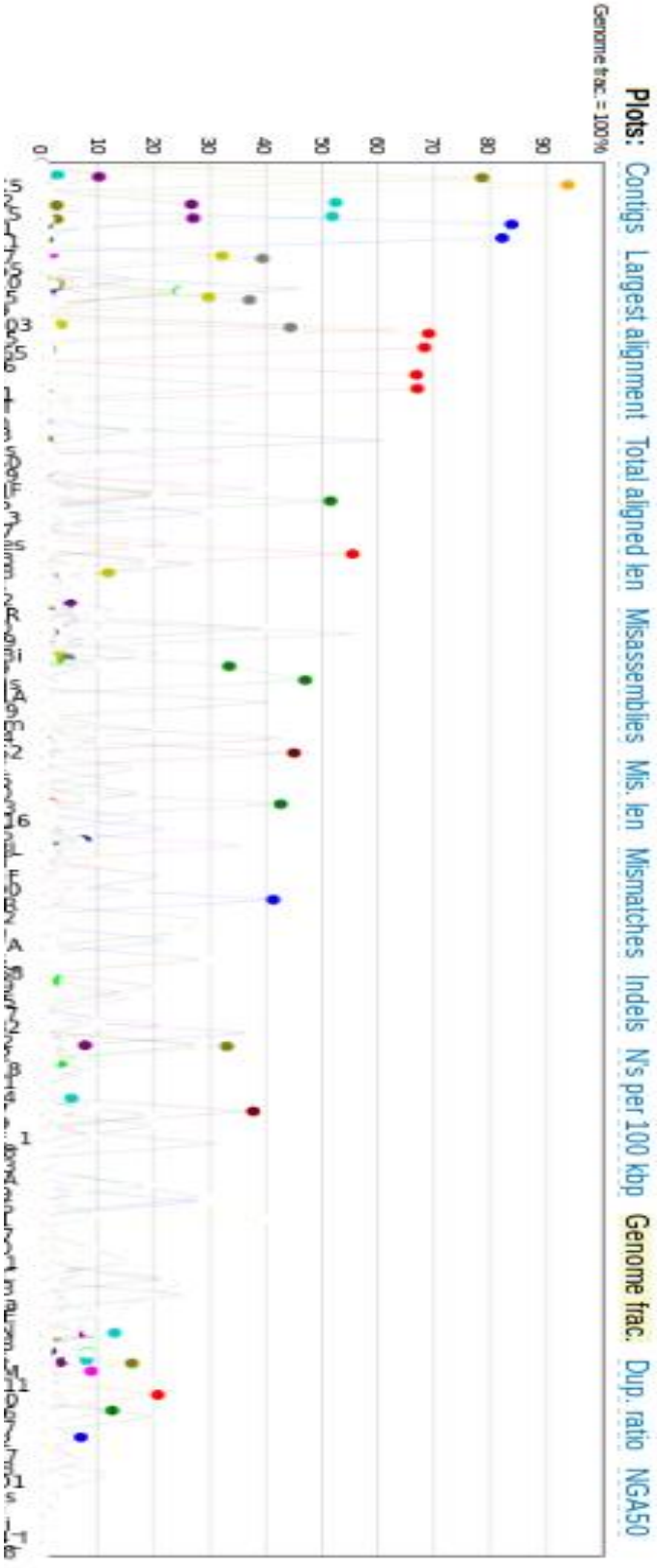
4.3.1 100 Espécies - *MaxBin*

Para o conjunto de dados de 100 espécies, com os parâmetros *default* o *MaxBin* foi capaz de gerar somente 33 *bins* (MegaHitMaxBin1), uma quantidade extremamente baixa considerando a quantidade de organismos presentes na amostra. Devido a esse fato, a ferramenta foi executada novamente com os parâmetros *min_contig_length* e *markerset* setados para 200 e 40, respectivamente. Para chegar a esses valores, verificou-se através da ferramenta PRINSEQ que o menor *contig* gerado pelo Megahit possuía tamanho de 200 pares de base. Como as características da montagem refletem na qualidade do *bin*, definiu-se como 200 o tamanho mínimo de *contig* presente nos *contigs* do Megahit.

Segundo o arquivo *README* do *MaxBin*, o parâmetro *markerset* é definido em 107 como *default*, e representa os marcadores genéticos presentes na maioria das bactérias. É descrito que uma outra opção é utilizar o *markerset* com valor 40, o que tende a dividir a saída em um número maior de *bins*.

Com os parâmetros *min_contig_length* = 200 e *markerset* = 40, o *MaxBin* produziu 76 bins (MegaHitMaxBin2). Esse resultado comprova a importância de se conhecer as características da montagem para utilizar configurações da ferramenta de *binning* que melhor se adequem a elas. O gráfico 5 demonstra os resultados de fração de genoma para os 76 *bins* gerados:

Gráfico 5 - Gráfico de Fração de genoma para megahitMaxbin2 de 100 espécie



Fonte: MetaQUAST (2018).

4.3.2 100 Espécies - *MetaBAT*

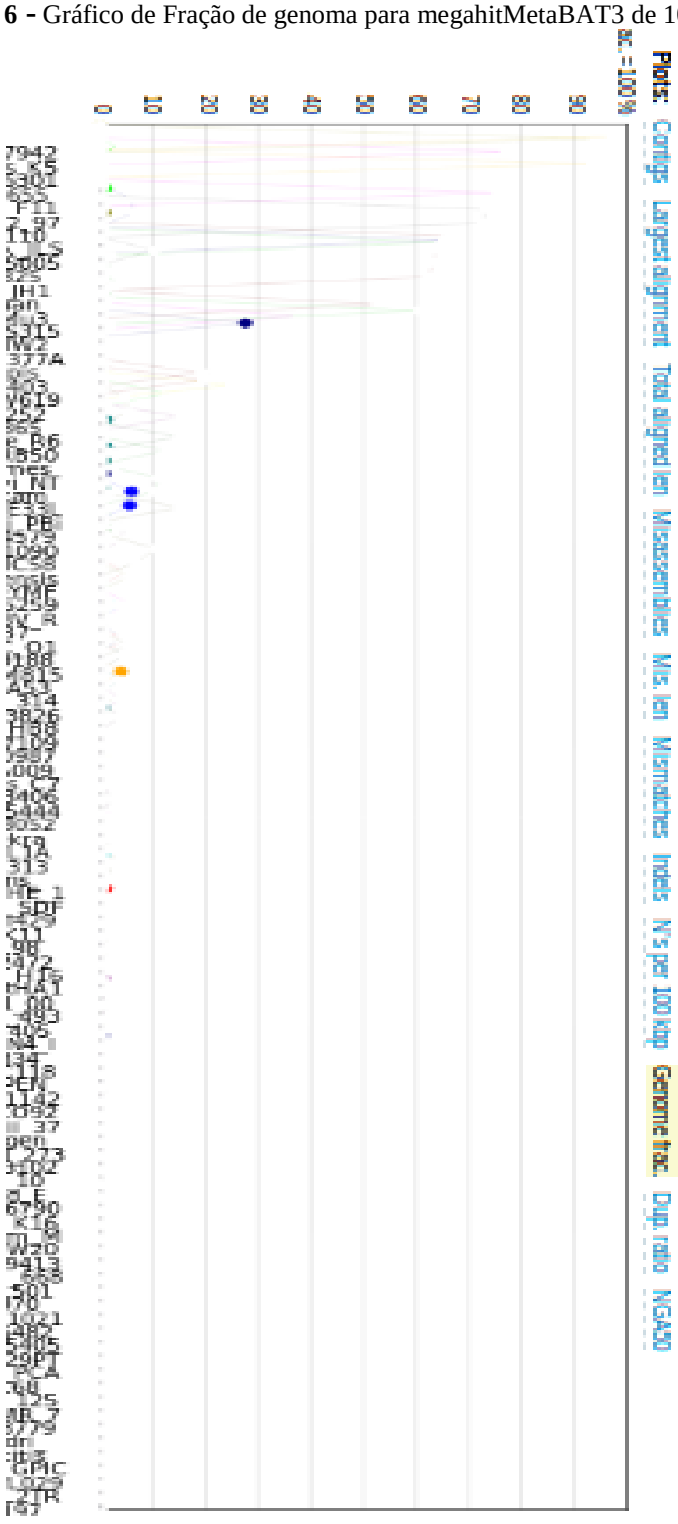
Diferentemente do *MaxBin*, o *MetaBAT* não permite que o tamanho mínimo do *contig* seja configurado para 200. O tamanho mínimo para definir o parâmetro de menor *contig* para realizar o *bin* é de 1500. Isso ocorreu devido a decisões próprias dos desenvolvedores da ferramenta. Kang et al. (2015) afirma que incluir *contigs* menores melhora o grau de completude de cada genoma, com a contrapartida de que *contigs* menores possuem maior variação de abundância, o que os torna mais difícil de separar em *bins*.

No primeiro teste realizado com parâmetros default (com tamanho mínimo de *bin* igual a 200000), o *metaBAT* retornou como saída somente 15 arquivos de *bin* (MegaHitMetaBAT1), um valor muito inferior ao esperado.

Para o segundo teste, decidiu-se modificar o parâmetro *-s*, que é definido no site da ferramenta como o menor tamanho de *bin* para output (METABAT, 2018). Para decidir um tamanho razoável de *bin*, verificou-se que entre os *bins* gerados pelo *MaxBin*, o menor *bin* equivalia a 130.000 pares de base. Esse valor foi utilizado como parâmetro no *metaBAT*, gerando somente 24 *bins* (MegaHitMetaBAT2), um número ainda abaixo.

Posteriormente, alguns testes foram realizados com valores menores de *bin* para tentar aproximar a quantidade de *bins* à quantidade de organismos. O valor mais aproximado encontrado foi o de 12.000 pares de base como tamanho mínimo. Com esse valor, 96 *bins* foram gerados pelo *MetaBAT* (MegaHitMetaBAT3). O gráfico 6 exibe a fração de genoma do megahitMETABAT3 para 100 espécies:

Gráfico 6 - Gráfico de Fração de genoma para megahitMetaBAT3 de 100 espécies



Fonte: MetaQUAST (2018).

4.4 MATRIZ DE CONFUSÃO

4.4.1 Matriz de Confusão - 10 Espécies

Para o conjunto de dados de 10 espécies, a Matriz de Confusão foi criada para as execuções idbaMaxBin e idbaMetaBAT2, cujas execuções da ferramenta produziram a quantidade exata de *bins* correspondente à cada espécie (Quadro 10):

Quadro 10 - Média das métricas para idbaMetaBAT2 e *MaxBin*

	idbaMetaBAT2	**idbaMaxBin**
ESPECIFICIDADE	99,72%	97,99%
SENSIBILIDADE	99,52%	84,25%
ACURÁCIA	99,71%	96,63%
PRECISÃO	94,34%	75,14%

Fonte: O autor (2018).

É possível observar no quadro 10 que a ferramenta *MetaBAT* se sobressaiu com melhores valores em todas as métricas comparado ao *MaxBin*, obtendo acima de 99% em Especificidade, Sensibilidade e Acurácia, e acima de 94% em Precisão. O *MaxBin* obteve um nível de precisão de 75,14%. Isso significa dizer que a presença de contaminantes nos *bins* gerados pelo *MaxBin* é maior do que os gerados pelo *MetaBAT*.

4.4.2 Matriz de Confusão - 100 Espécies

Para o conjunto de dados de 100 espécies, a Matriz de Confusão foi criada para as execuções megahitMaxBin2 e megahitMetaBAT3, pois essas foram as execuções cujas quantidades de *bins* gerados mais se aproximaram de 100. O quadro 11 compara o resultado das duas execuções:

Quadro 11 - Média das métricas para megahitMetaBAT3 e megahitMaxBin2

	megahitMetaBAT3	**megahitMaxBin2**
ESPECIFICIDADE	99,75%	99,13%
SENSIBILIDADE	39,80%	35,50%
ACURÁCIA	98,32%	98,08%
PRECISÃO	80,73%	37,47%

Fonte: O autor (2018).

O megaHitMetaBAT3 alcançou uma precisão maior que o megahitMaxBin2, com um valor de 80,73%. O valor de 37,47% de precisão para o megaHitMaxBin2 indica que nos *bins* gerados há uma quantidade muito grande de contaminantes. Isso se deve ao fato da quantidade de *bins* gerados ter sido menor (76 *bins*). Ambas as execuções tiveram um nível baixo de sensibilidade (abaixo de 40%). Esse número indica que os *contigs* de uma espécie se encontram dispersos por vários *bins*.

Uma análise importante que pode ser tirada a respeito dos resultados de *binning* diz respeito às características do sequenciamento e da montagem. No trabalho de Mende et al. (2012), os conjuntos de dados simulados de 10, 100 e 400 espécies foram gerados todos com a mesma quantidade de *reads* (53.33 Milhões, como é mostrado na tabela 2). Como o número de *reads* é igual em todos os três conjuntos de dados, à medida em que a quantidade de espécies na amostra aumenta, a cobertura das *reads* geradas diminui. O conjunto de dado de 100 espécies possui uma cobertura relativamente baixa, o que impacta na montagem e consequentemente no *binning*.

No quadro 6 (página 42), é possível visualizar que o montador Megahit cobriu somente 62.7% dos genomas de referências das espécies. A cobertura da amostra metagenômica, portanto, é um fator importante a ser levado em consideração no momento de realização da montagem e do *binning*.

4.5 OUTRAS FERRAMENTAS

Na literatura podem ser encontradas diversas outras ferramentas desenvolvidas voltadas para a realização do *binning*. Dentre elas podem ser citadas o CONCOCT (ALNEBERG et al.,

2014), GroopM (IMELFORT et al., 2014) e MyCC (LIN; LIAO, 2016). O GroopM é uma ferramenta que se encontra com a documentação desatualizada, não sendo possível realizar sua instalação. O CONCOCT também encontrava-se desatualizado, entretanto recebeu uma atualização em Outubro de 2018. Por essa razão, sua análise não pode ser feita em tempo hábil de conclusão deste trabalho. O MyCC, apesar de ter sua última atualização realizada em 2017, possui um processo de instalação relativamente complexo e por isso também foi descartado para este trabalho.

5 CONCLUSÃO

A metagenômica ainda é um campo recente dentro da bioinformática. O seu estudo é essencial para a compreensão de microrganismos que não podem ser cultivados laboratorialmente.

As ferramentas desenvolvidas para aplicação na genômica tradicional não são adequadas para a metagenômica, tornando-se necessário o desenvolvimento de novas ferramentas e métodos que sejam capazes de tratar esses conjuntos de dados mais complexos. É importante que esses métodos sejam validados para que futuramente possam ser aplicados e reconhecidos no tratamento de dados metagenômicos reais. O uso de dados simulados é uma peça essencial para a validação de novos *softwares*, pois através deles é possível analisar o desempenho das ferramentas comparando-os com genomas de referências já previamente conhecidos.

Com isto em mente, através da análise do *MetaBAT* e *MaxBin*, é possível concluir que:

- O *MetaBAT* se sobressai ao *MaxBin* na qualidade dos bins gerados, tanto para o conjunto de dado de 10 espécies quanto para o de 100 espécies;
- O arquivo contendo os dados sobre a abundância dos *contigs* é necessário para que as ferramentas de *binning* retornem um resultado mais preciso;
- A cobertura das *reads* sequenciadas possui um impacto direto na qualidade da montagem e do *binning*

Futuramente, é interessante que novos testes sejam realizados com outras ferramentas, como o CONCOCT, e seu desempenho comparado com as ferramentas presentes neste estudo. Ademais, é importante também verificar e comparar os resultados deste estudo com os dados de 100 espécies utilizando uma cobertura maior de *reads*. É possível que se alcance melhores resultados na montagem e consequentemente melhor resultado no *binning*. Também é de extrema relevância utilizar o conjunto de dados de 400 espécies para observar a eficiência do *MetaBAT*, *MaxBin*, e de possíveis outras ferramentas que virão a ser analisadas.

REFERÊNCIAS

ALNEBERG, J. et al. Binning metagenomic contigs by coverage and composition. **Nature Methods**, v. 11, n. 11, p. 1144–1146, 2014.

BANKEVICH, A et al., SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. **Journal of Computational Biology**, v.19, n. 5, p. 455–477, 2012.

BIN. Dicionário online "The Free Dictionary", 03 mar. 2018. Disponível em: <thefreedictionary.com/bin>. Acesso em: 03 dez. 2018.

DICK, G. J. et al. Community-wide analysis of microbial genome sequence signatures. **Genome Biology**, v. 10, n. 8, 2009.

GARCIA-ETXEBARRIA, K.; GARCIA-GARCERÀ, M.; CALAFELL, F. Consistency of metagenomic assignment programs in simulated and real data. **BMC Bioinformatics**, 2014.

GENBANK, 2018. Disponível em: <http://www.ncbi.nlm.nih.gov/genbank>. Acesso em: 30 Nov. 2018.

HOWE. et al. Tackling soil diversity with the assembly of large, complex metagenomes. **PNAS**, v. 111, n. 16, 2014.

ILLUMINA, 2018. Disponível em: < https://www.illumina.com/>. Acesso pela última vez em: 15/12/2018.

IMELFORT, M. et al. GroopM: an automated tool for the recovery of population genomes from related metagenomes. **PeerJ**, v. 2, p. e603, 2014.

KANG, D. D. et al. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. **PeerJ**, v. 3, p. e1165, 2015.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n.4, p. 357–359, 2012.

LESK, Arthur M. Introdução à Bioinformática, 2. ed. Porto Alegre, RS: Artmed, 2008. 384p. ISBN: 9788536311043.

LI, D. *et al.* Sequence analysis MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. **Bioinformatics**, v. 31, n. January, p. 1674–1676, 2015.

LIN, H. H.; LIAO, Y. C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. **Scientific Reports**, 2016.

MARTINS, A. M. Sequenciamento de DNA, montagem de novo do genoma e desenvolvimento de marcadores microssatélites, indels e SNPs para uso em análise genética de *Brachiaria ruziziensis.* 2013. 198 f. Tese (Doutorado em Biologia Molecular). Instituto de Ciências Biológicas. Universidade de Brasília, Brasília

MAVROMATIS, K. et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. **Nature Methods**, v. 4, n. 6, p. 495–500, 2007.

MENDE, D. R. et al. Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data. **PLos ONE** v. 7, n. 2, 2012.

METAGENÔMICA: o "projeto genoma" dos micróbios. Temas atuais em Biologia. Disponível em: <http://www.temasbio.ufscar.br/?q=artigos/metagen%C3%B4mica-o-%E2%80%9Cprojeto-genoma%E2%80%9D-dos-micr%C3%B3bios>. Acesso pela última vez em: 07/07/2018

METAQUAST. Disponível em <http://bioinf.spbau.ru/metaquast>. Acesso em: 10 Dez. 2018

MIKHEENKO, A.; SAVELIEV, V.; GUREVICH, A. MetaQUAST: Evaluation of metagenome assemblies. **Bioinformatics**, v. 32, n. 7, 2016.

NAMIKI, T. et al. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. **Nucleic Acids Research,** v. 40, n. 20, 2012.

NIH, 2018. Disponível em: <https://www.genome.gov/12011238/>. Acesso pela última vez em: 15/11/2018.

NOBLE, P. A.; CITEK, R. W.; OGUNSEITAN, O. A. Tetranucleotide frequencies in microbial genomes. **Electrophoresis**, v. 19, p. 528–535, 1998.

NURK, S et al. metaSPAdes: a new versatile metagenomics assembler Sergey. **Genome Research**, v. 27, n. 5, p. 824-834, 2017.

PENG, Y. et al. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler, **Research in Computational Molecular Biology**, p. 426-440, 2010.

PENG, Y. et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. **Bioinformatics**, v. 28, n. 11, p. 1420–1428, 2012.

PENG, Y. et al. Meta-IDBA: a de Novo assembler for metagenomic data. **Bioinformatics**, v. 27, p. 94–101, 2011.

PROJECTSTARLIGHT. Sequencing-by-Synthesis vs. Single Molecule Sequencing. Disponível em: <https://projectstarlight.weebly.com/the-competition.html>. Acesso pela última vez em: 12/07/2019

PROSDOCIMI, Francisco et al. Bioinformática: manual do usuário. **Biotecnologia Ciência & Desenvolvimento**, v. 29, p. 12-25, 2002.

ROCHE, 2018. Disponível em: <http://allseq.com/knowledge-bank/sequencing-platforms/454-roche//>. Acesso pela última vez em: 15/12/2018.

SANGWAN, N.; XIA, F.; GILBERT, J. A. Recovering complete and draft population genomes from metagenome datasets. **Microbiome**, v. 4, p. 1–11, 2016.

SHARPTON, T. J. An introduction to the analysis of shotgun metagenomic data. **Frontiers in Plant Science**, v. 5, p. 1–14, 2014.

SLIDESHARE. EREN, A. M. Intro to metagenomic binning. Disponível em: <https://www.slideshare.net/AMuratEren/intro-to-metagenomic-binning>. Acesso pela última vez em: 11/10/2018.

THERMOFISHER, 2018. Disponível em: <https://www.thermofisher.com/>. Acesso pela última vez em: 15/12/2018.

TYSON, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. **Nature**, v. 428, p. 37-43, 2004.

VENTER, J. C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. **Science**, v. 304, n. 5667, p. 66–74, 2004.

VERLI, H. *Bioinformática***:** da biologia à flexibilidade molecular. 1. ed. São Paulo: SBBq, 2014

WATSON, & CRICK, F. H. C. Molecular structure of nucleic acids. **Nature**, v. 171, n. 4356, p.737-738, 1953.

WOLF, B. De novo genome assembly versus mapping to a reference genome Disponível em: < http://beat.wolf.home.hefr.ch/documents/prague.pdf>. Acesso pela última vez em: 08/07/2018

WOOLEY, J. C.; GODZIK, A.; FRIEDBERG, I. A Primer on Metagenomics. **PLoS Computational Biology,** v. 6, n. 2, 2010.

WU, Y. W.; SIMMONS, B. A.; SINGER, S. W. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. **Bioinformatics**, v. 32, n. 4, p. 605–607, 2015.

WU, Y. W.; YE, Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. Lecture Notes in Computer Science. **Journal of Computational Biology**, v. 18, n. 3, p. 523–534, 2011.

ZERBINO, D. e BIRNEY, E., Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. **Genome Research**, v.18, n.5, p. 821-829, 2008.

ZHANG, W. *et al.* A Practical Comparison of DeNovo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. **PLoS ONE**, v. 6, n. 3, p. 97–102, 2011.

ANEXO A – Genomas presentes no conjunto de dados de 10 espécies

Rank	Cromossomo	Cromossomo	Nome	grupo filogenético	Tamanho do Genoma (Mb) [1]	GC %	abundância [2]	Abundância Relativa	Proporção estimada da sequência	Cobertura estimada para Illumina
1	NC_010546	NC_010547	Cyanothece sp. ATCC 51142	Cyanobacteria	5,43	37,9	1000	15,15	23,50	128,53
2	NC_009641	0	Staphylococcus aureus subsp. aureus str. Newm	Firmicutes	2,9	32,9	794,31	12,03	9,97	102,09
3	NC_009637	0	Methanococcus maripaludis C7	Euryarchaeota	1,8	33,3	707,11	10,71	5,51	90,88
4	NC_003112	0	Neisseria meningitidis MC58	Betaproteobacteria	2,3	51,5	656,26	9,94	6,53	84,35
5	NC_006582	0	Bacillus clausii KSM-K16	Firmicutes	4,3	44,8	621,97	9,42	11,57	79,94
6	AC_000091	0	Escherichia coli str. K12 substr. W3110	Gammaproteobacteria	4,65	50,8	596,83	9,04	12,01	76,71
7	NC_008555	0	Listeria welshimeri serovar 6b str. SLCC5334	Firmicutes	2,8	36,4	577,35	8,75	7,00	74,20
8	NC_008011	0	Lawsonia intracellularis PHE/MN1-00	Deltaproteobacteria	1,76	33,1	561,66	8,51	4,28	72,19
9	NC_007404	0	Thiobacillus denitrificans ATCC 25259	Betaproteobacteria	2,91	66,1	548,66	8,31	6,91	70,52
10	NC_005296	0	Rhodopseudomonas palustris CGA009	Alphaproteobacteria	5,47	65	537,65	8,14	12,73	69,10
				soma	34,32		6601,805414	100	100	
1	tamanho do genoma = total de todos os cromossomos e plasmídeos			média	3,432	45,2				
2	calculado usando ni = a (log (i + 1)) - b para 1> i> M e b = 0,5									
	O parâmetro a representa a abundância do genótipo mais abundante, b é um parâmetro relacionado à uniformidade e M é o número de diferentes genótipos na comunidade.									

ANEXO B – Genomas presentes no conjunto de dados de 100 espécies

Rank	Cromossomo	Cromossomo	Nome	grupo filogenético	Tamanho do Genoma (Mb)[1]	GC %	abundância[2]	Abundância Relativa	Proporção estimada da sequência	Cobertura estimada para Illumina
1	NC_007969	0	Psychrobacter cryohalolentis K5	Gammaproteobacteria	3,1	42,2	1000	2,22768389	1,86278127	17,846646
2	NC_009077	0	Mycobacterium sp. JLS	Actinobacteria	6	68,4	794,310867	1,76947352	2,86379498	14,175785
3	NC_007604	0	Synechococcus elongatus PCC 7942	Cyanobacteria	2,75	55,4	707,106781	1,57521038	1,1684708	12,619485
4	NC_004722	0	Bacillus cereus ATCC 14579	Firmicutes	5,42	35,3	656,25952	1,46193876	2,13734718	11,712032
5	NC_002662	0	Lactococcus lactis subsp. lactis Il1403	Firmicutes	2,4	35,3	621,974925	1,38556352	0,89698316	11,100167
6	NC_004344	0	Wigglesworthia glossinidia endosymbiont of Glossina brev	Gammaproteobacteria	0,7	22,5	596,830954	1,3295507	0,25104383	10,651431
7	AC_000091	0	Escherichia coli str. K12 substr. W3110	Gammaproteobacteria	4,65	50,8	577,350269	1,28615389	1,6132159	10,303766
8	NC_010501	0	Pseudomonas putida W619	Gammaproteobacteria	5,8	61,4	561,6626	1,25120672	1,95750856	10,023794
9	NC_003098	0	Streptococcus pneumoniae R6	Firmicutes	2,04	39,7	548,662005	1,22224551	0,67256649	9,7917768
10	NC_002927	0	Bordetella bronchiseptica RB50	Betaproteobacteria	5,3	68,1	537,647493	1,19770866	1,71227558	9,5952047
11	NC_008529	0	Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-3	Firmicutes	1,9	49,7	528,150495	1,17655235	0,60299188	9,4257151
12	NC_008593	0	Clostridium nowi NT	Firmicutes	2,5	28,9	519,844356	1,1580489	0,78093253	9,2774784
13	NC_009632	0	Staphylococcus aureus subsp. aureus JH1	Firmicutes	2,93	32,9	512,493449	1,1416734	0,9023107	9,1462894
14	NC_006156	0	Borrelia garinii PBi	Spirochaetes	0,98	28,1	505,922944	1,12703639	0,29792752	9,0290279
15	NC_010102	0	Salmonella enterica subsp. enterica serovar Paratyphi B	Gammaproteobacteria	4,9	52,1	500	1,11384194	1,4721981	8,9233232
16	NC_010694	0	Erwinia tasmaniensis	Gammaproteobacteria	4,08	53,4	494,621615	1,1018606	1,21264428	8,8273371
17	NC_003923	0	Staphylococcus aureus subsp. aureus MW2	Firmicutes	2,8	32,8	489,706511	1,0909113	0,82393714	8,739618
18	NC_007404	0	Thiobacillus denitrificans ATCC 25259	Betaproteobacteria	2,91	66,1	485,189564	1,08084897	0,84840772	8,6590066
19	NC_008435	0	Rhodopseudomonas palustris BisA53	Alphaproteobacteria	5,51	64,4	481,017893	1,07155581	1,59262306	8,5845562
20	NC_003997	0	Bacillus anthracis str. Ames	Firmicutes	5,23	35,4	477,148036	1,06293499	1,49952945	8,5154923
21	NC_004116	0	Streptococcus agalactiae 2603V/R	Firmicutes	2,2	35,6	473,543899	1,05490611	0,62601263	8,4511705
22	NC_010622	NC_010623	Burkholderia phymatum STM815	Betaproteobacteria	8,7	62,3	470,175211	1,04740174	2,45798456	8,3910507
23	NC_009802	0	Campylobacter concisus 13826	Epsilonproteobacteria	2,15	39,3	467,016372	1,04036485	0,60335197	8,3346761
24	NC_009633	0	Alkaliphilus metalliredigens QYMF	Firmicutes	4,9	36,8	464,045557	1,03374681	1,36633398	8,281657
25	NC_004369	0	Corynebacterium efficiens YS-314	Actinobacteria	3,15	63,1	461,244028	1,02750589	0,87305475	8,2316591
26	NC_007297	0	Streptococcus pyogenes MGAS5005	Firmicutes	1,84	38,5	458,595593	1,02160601	0,50704659	8,1843934
27	NC_009667	NC_009668	Ochrobactrum anthropi ATCC 49188	Alphaproteobacteria	5,22	56,1	456,086174	1,01601582	1,43059789	8,1396087
28	NC_005071	0	Prochlorococcus marinus str. MIT 9313	Cyanobacteria	2,41	50,7	453,703463	1,01070789	0,65703621	8,0970853
29	NC_005296	0	Rhodopseudomonas palustris CGA009	Alphaproteobacteria	5,47	65	451,436648	1,00565815	1,48383055	8,0566302
30	NC_010545	0	Corynebacterium urealyticum DSM 7109	Actinobacteria	2,37	64,2	449,276181	1,00084531	0,63982604	8,0180731

31	NC_006461	0	Thermus thermophilus HB8	Deinococcus-Thermus	2,07	69,5	447,213595	0,99625052	0,55626984	7,9812629
32	NC_010400	0	Acinetobacter baumannii SDF	Gammaproteobacteria	3,45	39,1	445,241354	0,99185699	0,92302775	7,946065
33	NC_002946	0	Neisseria gonorrhoeae FA 1090	Betaproteobacteria	2,15	52,7	443,352718	0,9876497	0,57278021	7,9123592
34	NC_000913	0	Escherichia coli str. K12 substr. MG1655	Gammaproteobacteria	4,6	50,8	441,541642	0,9836152	1,2204772	7,8800376
35	NC_008358	0	Hyphomonas neptunium ATCC 15444	Alphaproteobacteria	3,71	61,9	439,802687	0,97974136	0,9804469	7,849003
36	NC_009641	0	Staphylococcus aureus subsp. aureus str. Newman	Firmicutes	2,9	32,9	438,130939	0,97601723	0,76348778	7,8191679
37	NC_008212	0	Haloquadratum walsbyi DSM 16790	Euryarchaeota	3,18	47,9	436,52195	0,97243291	0,8341293	7,7904529
38	NC_008819	0	Prochlorococcus marinus str. NATL1A	Cyanobacteria	1,9	35	434,971677	0,9689794	0,49660919	7,7627857
39	NC_009617	0	Clostridium beijerinckii NCIMB 8052	Firmicutes	6	29,9	433,476441	0,96564848	1,56284864	7,7361008
40	NC_005085	0	Chromobacterium violaceum ATCC 12472	Betaproteobacteria	4,8	64,8	432,032882	0,96243269	1,24611524	7,7103381
41	NC_010280	0	Chlamydia trachomatis L2b/UCH-1/proctitis	Chlamydiae/Verrucomic	1	41,3	430,637926	0,95932517	0,25876912	7,6854428
42	NC_008255	0	Cytophaga hutchinsonii ATCC 33406	Bacteroidetes/Chlorobi	4,4	38,8	429,288753	0,95631964	1,13501697	7,6613646
43	NC_008340	0	Alkalilimnicola ehrlichei MLHE-1	Gammaproteobacteria	3,3	67,5	427,982769	0,95341032	0,8486702	7,6380571
44	NC_008600	0	Bacillus thuringiensis str. Al Hakam	Firmicutes	5,36	35,4	426,717589	0,9505919	1,37437581	7,6154779
45	NC_010516	0	Clostridium botulinum B1 str. Okra	Firmicutes	4,15	28,2	425,491007	0,94785946	1,06105684	7,5935875
46	NC_009637	0	Methanococcus maripaludis C7	Euryarchaeota	1,8	33,3	424,300986	0,94520847	0,45893028	7,5723497
47	NC_007712	0	Sodalis glossinidius str. 'morsitans'	Gammaproteobacteria	4,29	54,5	423,145639	0,94263472	1,09080553	7,5517306
48	NC_010546	NC_010547	Cyanothece sp. ATCC 51142	Cyanobacteria	5,43	37,9	422,023214	0,94013431	1,37700761	7,5316991
49	NC_008527	0	Lactococcus lactis subsp. cremoris SK11	Firmicutes	2,56	35,8	420,932085	0,93770362	0,64751848	7,5122261
50	NC_008309	0	Haemophilus somnus 129PT	Gammaproteobacteria	2,01	37,2	419,870735	0,93533927	0,50712128	7,4932845
51	NC_009012	0	Clostridium thermocellum ATCC 27405	Firmicutes	3,8	39	418,837753	0,93303811	0,95637802	7,4748493
52	NC_003112	0	Neisseria meningitidis MC58	Betaproteobacteria	2,3	51,5	417,83182	0,93079721	0,57747012	7,4568967
53	NC_009776	0	Ignicoccus hospitalis KIN4/I	Crenarchaeota	1,3	56,5	416,851704	0,92861382	0,32563052	7,439405
54	NC_007512	0	Pelodictyon luteolum DSM 273	Bacteroidetes/Chlorobi	2,36	57,3	415,89625	0,92648537	0,58978969	7,4223533
55	NC_009801	0	Escherichia coli E24377A	Gammaproteobacteria	5,27	50,6	414,964377	0,92440945	1,31407938	7,4057225
56	NC_009616	0	Thermosipho melanesiensis BI429	Thermotogae	1,9	31,4	414,055069	0,92238381	0,4727286	7,3894944
57	NC_008011	0	Lawsonia intracellularis PHE/MN1-00	Deltaproteobacteria	1,76	33,1	413,167375	0,9204063	0,43695716	7,373652
58	NC_002971	0	Coxiella burnetii RSA 493	Gammaproteobacteria	2,03	42,6	412,300396	0,91847495	0,5029328	7,3581794
59	NC_003143	0	Yersinia pestis CO92	Gammaproteobacteria	4,88	47,6	411,453287	0,91658786	1,20653668	7,3430613
60	NC_006274	0	Bacillus cereus E33L	Firmicutes	5,85	35,1	410,625253	0,91474326	1,44344982	7,3282837
61	NC_007292	0	Candidatus Blochmannia pennsylvanicus str. BPEN	Gammaproteobacteria	0,79	29,6	409,815543	0,91293948	0,19454303	7,3138331
62	NC_006512	0	Idiomarina loihiensis L2TR	Gammaproteobacteria	2,84	47	409,023445	0,91117494	0,69801814	7,2996968
63	NC_008346	0	Syntrophomonas wolfei subsp. wolfei str. Goettingen	Other Bacteria	2,94	44,9	408,24829	0,90944814	0,72122683	7,2858629
64	NC_003909	0	Bacillus cereus ATCC 10987	Firmicutes	5,43	35,5	407,489443	0,90775767	1,32958578	7,27232
65	NC_008563	0	Escherichia coli APEC O1	Gammaproteobacteria	5,51	50,3	406,746302	0,90610218	1,34671402	7,2590574
66	NC_009565	0	Mycobacterium tuberculosis F11	Actinobacteria	4,4	65,6	406,018296	0,90448042	1,0734911	7,246065
67	NC_006582	0	Bacillus clausii KSM-K16	Firmicutes	4,3	44,8	405,304885	0,90289116	1,04725023	7,233333
68	NC_007929	0	Lactobacillus salivarius UCC118	Firmicutes	2,1	33	404,605555	0,90133328	0,51056531	7,2208523
69	NC_002945	0	Mycobacterium bovis AF2122/97	Actinobacteria	4,35	65,6	403,919818	0,89980567	1,05580713	7,2086142
70	NC_010628	0	Nostoc punctiforme PCC 73102	Cyanobacteria	9,01	41,4	403,24721	0,89830731	2,18321412	7,1966104

71	NC_008322	0	Shewanella sp. MR-7	Gammaproteobacteria	4,8	47,9	402,587288	0,89683721	1,16118513		7,184833
72	NC_009074	NC_009075	Burkholderia pseudomallei 668	Betaproteobacteria	7	68,3	401,939632	0,89539444	1,16067075		7,1732745
73	NC_000963	0	Rickettsia prowazekii str. Madrid E	Alphaproteobacteria	1,11	29	401,30384	0,8939781	0,26766801		7,1619277
74	NC_000907	0	Haemophilus influenzae Rd KW20	Gammaproteobacteria	1,8	38,1	400,67953	0,89258733	0,43338096		7,1507859
75	NC_009434	0	Pseudomonas stutzeri A1501	Gammaproteobacteria	4,6	63,9	400,066336	0,89122133	1,10583418		7,1398424
76	NC_002967	0	Treponema denticola ATCC 35405	Spirochaetes	2,8	37,9	399,46391	0,88987932	0,67210287		7,1290912
77	NC_005241	0	Ralstonia eutropha H16	Betaproteobacteria	7	66,3	398,871919	0,88856055	1,67776709		7,1185261
78	NC_002944	0	Mycobacterium avium subsp. paratuberculosis K-10	Actinobacteria	4,8	69,3	398,290044	0,88726431	1,14879056		7,1081416
79	NC_008555	0	Listeria welshimeri serovar 6b str. SLCC5334	Firmicutes	2,8	36,4	397,717979	0,88598993	0,66916532		7,0979321
80	NC_002570	0	Bacillus halodurans C-125	Firmicutes	4,2	43,7	397,155434	0,88473676	1,00232824		7,0878926
81	NC_008268	0	Rhodococcus sp. RHA1	Actinobacteria	9,67	67	396,602127	0,88350417	2,30452637		7,0780179
82	NC_002939	0	Geobacter sulfurreducens PCA	Deltaproteobacteria	3,81	60,9	396,057792	0,88229156	0,90674195		7,0683034
83	NC_000922	0	Chlamydophila pneumoniae CWL029	Chlamydiae/Verrucomic	1,2	40,6	395,522172	0,88109837	0,28520179		7,0587443
84	NC_008245	0	Francisella tularensis subsp. tularensis FSC198	Gammaproteobacteria	1,9	32,3	394,995019	0,87992404	0,45096765		7,0493364
85	NC_007795	0	Staphylococcus aureus subsp. aureus NCTC 8325	Firmicutes	2,8	32,9	394,476097	0,87876805	0,66371081		7,0400754
86	NC_010612	0	Mycobacterium marinum M	Actinobacteria	6,62	65,7	393,96518	0,87762988	1,5671696		7,0309573
87	NC_000964	0	Bacillus subtilis subsp. subtilis str. 168	Firmicutes	4,2	43,5	393,462047	0,87650906	0,99300699		7,021978
88	NC_002952	0	Staphylococcus aureus subsp. aureus MRSA252	Firmicutes	2,9	32,8	392,966491	0,87540512	0,68478413		7,013134
89	NC_003047	0	Sinorhizobium meliloti 1021	Alphaproteobacteria	6,74	62,2	392,478308	0,8743176	1,5895556		7,0044216
90	NC_004070	0	Streptococcus pyogenes MGAS315	Firmicutes	1,9	38,6	391,997304	0,87324608	0,44754515		6,9958373
91	NC_002162	0	Ureaplasma parvum serovar 3 str. ATCC 700970	Firmicutes	0,75	25,5	391,523293	0,87219013	0,17644893		6,9873778
92	NC_003361	0	Chlamydophila caviae GPIC	Chlamydiae/Verrucomic	1,18	39,2	391,056094	0,87114936	0,27728172		6,9790398
93	NC_007413	0	Anabaena variabilis ATCC 29413	Cyanobacteria	7,07	41,4	390,595533	0,87012338	1,65938384		6,9708204
94	NC_008709	0	Psychromonas ingrahamii 37	Gammaproteobacteria	4,6	40,1	390,141444	0,86911181	1,07840052		6,9627164
95	NC_006576	0	Synechococcus elongatus PCC 6301	Cyanobacteria	2,7	55,5	389,693666	0,8681143	0,63224773		6,9547251
96	NC_009972	0	Herpetosiphon aurantiacus ATCC 23779	Chloroflexi	6,74	50,9	389,252042	0,8671305	1,57648908		6,9468436
97	NC_004663	0	Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes/Chlorobi	6,33	42,9	388,816423	0,86616008	1,47893293		6,9390692
98	NC_008513	0	Buchnera aphidicola str. Cc (Cinara cedri)	Gammaproteobacteria	0,42	20,1	388,386664	0,86520271	0,09801979		6,9313995
99	NC_009782	0	Staphylococcus aureus subsp. aureus Mu3	Firmicutes	2,9	32,9	387,962624	0,86425809	0,67606438		6,9238318
100	NC_008510	NC_008511	Leptospira borgpetersenii serovar Hardjo-bovis JB197	Spirochaetes	3,9	40,2	387,544169	0,8633259	0,90820938		6,9163638
				Tamanho Total de Geno	371,73		44889,6725	100	100		
				Tamanho total de reads	200						
				Número Total de Reads							
							371,73				
1	tamanho do genoma = total de todos os cromossomos e plasmídeos			média			74,34	40,4			
2	calculado usando ni = a (log (i + 1)) - b para 1> i> M e b = 0,5										
	O parâmetro a representa a abundância do genótipo mais abundante, b é um parâmetro relacionado à uniformidade e M é o número de diferentes genótipos na comunidade.										