Análise comparativa de ferramentas de Montagem e *Binning* de metagenomas utilizando dados simulados microbianos

Rodrigo B. P. R. Pará[1], Pedro H. D. M. Rocha[1], Danielle C. C. Couto[2], Renato R. M. Oliveira[1], Regiane Kawasaki[1]

[1]Laboratório de Bioinformática e Computação de Alto Desempenho (LABIOCAD) – Instituto de Ciências Exatas e Naturais - Universidade Federal do Pará (UFPA)

[2]Laboratório Interdisciplinar em Tecnologias, Educação e Computação (LITEC) – Campus Ananindeua - Universidade Federal do Pará (UFPA)

{pedrohenriquedornele, renato.renison, rodrigopara}@gmail.com, {danifc,kawasaki}@ufpa.br

Abstract. *Binning consists of grouping DNA sequences according to taxonomic units, widely used in the Metagenomics, field that studies the genome of communities of microorganisms. New tools are developed for metagenomic pipelines, necessitating the establishment of paradigms in this type of analysis through the verification of the performance of assembly and binning tools. For the comparative tests of this work, data sets of 10 and 100 species of bacteria were used, as well as 3 assembly softwares: IDBA_UD, Megahit and MetasSPAdes, and 2 binning softwares: MetaBAT-2.12.1 and MaxBin-2.2.4. We verified that MetaBAT exceeded MaxBin in the quality of the generated bins.*

Resumo. *Binning consiste em agrupar sequências de DNA de acordo com unidades taxonômicas, muito usado na Metagenômica, campo que estuda o genoma de comunidades de microrganismos. Novas ferramentas são desenvolvidas para pipelines metagenômicos, necessitando que se estabeleçam paradigmas neste tipo de análise através da verificação do desempenho de ferramentas de montagem e binning. Para os testes comparativos deste trabalho foram utilizados conjuntos de dados de 10 e 100 espécies de bactérias, além de 3 softwares montadores: IDBA_UD, Megahit e MetasSPAdes, e 2 softwares de binning: MetaBAT-2.12.1 e MaxBin-2.2.4. Verificou-se que o MetaBAT superou o MaxBin na qualidade dos bins gerados.*

1. Introdução

A análise de dados metagenômicos é um processo mais complexo do que a genômica tradicional, devido a presença de muitos organismos na amostra, diferença de abundância entre estes organismos e sequências de DNA comum entre espécies [NAMIKI et al., 2012]. Sequenciar e montar os genomas destes organismos envolve fragmentar o DNA em pequenas sequências denominadas *reads* e posteriormente agrupá-las em sequências maiores (*contigs*) a fim de reconstruir o genoma original.

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) em arquivos denominados *bins* [WOOLEY et al., 2010]. Através do *binning* é possível obter informações de genomas ainda desconhecidos que são difíceis de identificar por outros métodos, e obter informações quanto aos tipos e quantidades de unidades taxonômicas presentes [SHARPTON, 2014].

A montagem de metagenomas é uma tarefa difícil e exige o desenvolvimento de novas técnicas e ferramentas, haja vista que os *softwares* de montagem de genomas tradicionais não são aptos a tratar dados metagenômicos [PENG et al., 2011].

Softwares de montagem de metagenomas foram desenvolvidas, como *MetaSPades* [NURK et al., 2012], IDBA-UD [PENG et al., 2012] e MetaVelvet [NAMIKI et al., 2012]. Entretanto, é necessário que se estabeleçam novos modelos de análise metagenômica através da verificação da eficácia das ferramentas de montagem em conjunto com o *binning*.

Este trabalho pretende comparar de maneira não enviesada o desempenho de ferramentas de montagem e *binning* de metagenomas. Muitos trabalhos relacionados a comparação de ferramentas são feitos no lançamento de uma ferramenta, com conjuntos de dados distintos, o que não permite uma comparação sólida. É interessante estabelecer um conjunto fixo de dados e a partir daí comparar o comportamento das ferramentas com estes dados. Para tal, o uso de dados simulados contendo genoma de espécies já conhecidas é recomendável pois permite aferir o desempenho destas ferramentas, através da comparação das sequências geradas com os genomas de referências dessas espécies. Essa análise poderá permitir a validação do *pipeline* e das ferramentas, permitindo sua aplicação em amostras com dados reais.

O presente artigo subdivide-se em mais três seções: Materiais e Métodos, abordando metodologia e ferramentas utilizadas no *pipeline*; Resultados e Discussões, contendo a análise comparativa dos resultados das ferramentas, e Conclusão, contendo considerações acerca da pesquisa e sugestões de trabalhos futuros.

2. Materiais e Métodos

Inicialmente foi realizada uma revisão bibliográfica dos trabalhos e ferramentas de Montagem e *Binning* comumente utilizados. Decidiu-se pelo *MetaBAT-2.12.1* e *MaxBin-2.2.4* devido a serem ferramentas atualizadas e com uma boa documentação disponível. Outras ferramentas, como o GroopM [IMELFORT et al., 2014] encontravam-se desatualizadas e por isso foram descartadas. Já o CONCOCT [ALNEBERG et al., 2014] recebeu uma atualização após a realização desta pesquisa em Outubro de 2018, portanto não foi incluído.

Para a realização dos testes comparativos, foram utilizados os conjuntos de dados de *reads* não tratadas de 10 e 100 espécies de bactérias de diferentes complexidades, retirados do artigo *Assessment of metagenomic assembly using simulated next generation sequencing data* [MENDE et al., 2012]. Estes dados foram gerados através do simulador de dados metagenômicos *iMess (interactive MEtagenomic Simulation Software*). Neste trabalho foram utilizadas as *reads* geradas pelo

sequenciamento Illumina, que podem ser encontrados no link: www.bork.embl.de/~mende/simulated_data/.

O tratamento de qualidade foi realizado pela ferramenta PRINSEQ [SCHMIEDER e EDWARDS, 2011]. Em seguida, as *reads* foram submetidas ao processo de montagem, realizado por três *softwares* montadores: *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, através da ferramenta *Metaquast* [MIKHEENKO; SAVELIEV; GUREVICH, 2016] foi feita a análise comparativa da qualidade das montagens, tendo como critérios cobertura do genoma e número de erros de montagem.

O melhor resultado obtido entre os três montadores foi submetido aos *softwares* de *binning*. Cada *software* de *binning* produz como saída um certo número de arquivos *bins*. Inicialmente as ferramentas foram rodadas com parâmetros *default* e posteriormente mudanças nos parâmetros foram feitas com o objetivo de observar o comportamento das ferramentas e a quantidade e qualidade dos *bins* gerados.

Para o *dataset* 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 presentes na amostra; este arquivo é gerado separadamente) e posteriormente com o arquivo de *depth*.

Para o *dataset* de 100 espécies, o *MaxBin* foi rodado duas vezes, uma com os parâmetros default e outra com os parâmetros *min_contig_length* = 200 e *marker_set* = 40. Já o *MetaBAT* foi executado três vezes: a primeira execução com os parâmetros default; 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. Esses parâmetros foram escolhidos para aumentar o número de *bins* gerados na saída da ferramenta.

Cada um dos resultados gerados pelas ferramentas de *binning* foi novamente enviado ao *Metaquast* para avaliação. Através do *Metaquast* é possível observar a quantidade de *bins* gerados, qual o grau de completude de cada espécie e a quantidade total de alinhamento realizado dos *contigs* com os organismos da amostra.

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 eficácia das ferramentas. Considerando-se que o resultado ideal seria ter todos os *contigs* de um *bin* pertencentes a uma única espécie, foi importante observar se ocorreram as seguintes situações: a) Um determinado *bin* alinhou com mais de uma espécie diferente; b) Dois ou mais *bins* fazem referência a uma mesma espécie;

Para analisar a eficácia de cada ferramenta ao categorizar os *contigs* de uma espécie por *bin*, foi criada uma Matriz de Confusão. Deseja-se saber se cada *bin* de fato está sendo mapeado aos *contigs* de uma única espécie.

Para isso, verifica-se todos os *contigs* presentes em um *bin.* A espécie com maior prevalência de *contigs* naquele *bin* será mapeada para àquele *bin*. A Tabela 1 exemplifica esse processo através das estatísticas obtidas do *Metaquast* para espécie *Neisseria meningitidis MC58:*

	Idba.*MetaBAT*3.3	Idba.*MetaBAT*3.7
FRAÇÃO DE GENOMA	0.187	77.107
MAIOR ALINHAMENTO	4246	49978
ERROS DE MONTAGEM	0	1
CONTIGS	1	170

Tabela 1 - *Bins* alinhados ao organismo *Neisseria meningitidis MC58*

Na parte inferior da Tabela 1, observa-se que o *bin* idba*MetaBAT*3.7 possui 170 *contigs* alinhados à *N. meningiditis MC58*., enquanto idba*METABAT*3.3 possui somente 1. Logo, o *bin* idba*MetaBAT*3.7 é mapeado para *N. meningiditis MC58* por ter a maior correspondência de *contigs*. Após realizado o mapeamento individual de cada *bin* para cada organismo, cada *contig* presente no *bin* é classificado de acordo com uma das quatro situações presentes na Tabela 2:

	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*.

Tabela 2 - Matriz de confusão de *contigs* presentes nos *bins*

A Matriz de Confusão foi calculada individualmente para cada *bin*. Em seguida, cada *bin* foi avaliado segundo as seguintes métricas:

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 obter o desempenho geral de cada ferramenta, calculou-se a média das métricas individuais dos *bins*. O *pipeline* utilizado na análise de metagenomas está descrito na Figura 8:

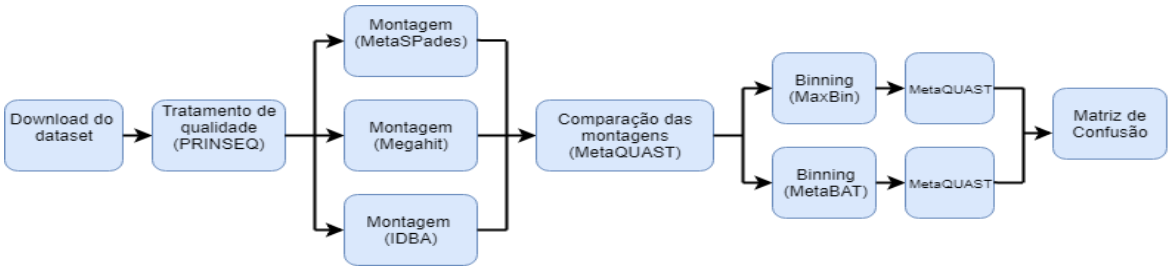

Figura 8 - *Pipeline* para comparação de ferramentas de montagem e *binning*

3. Resultados e Discussões

Os arquivos de *bins* gerados seguem a padronização *nome do montador* seguido pelo *nome da ferramenta de binning.* Diferentes execuções das mesmas ferramentas são indicadas por números (e.g. idba*MetaBAT*1, idba*MetaBAT*2 etc.).

3.1 *Dataset* de 10 Espécies

Para a montagem de 10 espécies, o IDBA foi o que obteve melhor cobertura de genoma (99.3%), como mostra o Quadro 1:

	IDBA_contig	***Megahit*_contig**	**Spades_contig**
FRAÇÃO DE GENOMA	99.314	99.108	99.051
MAIOR ALINHAMENTO	1190435	1100460	633416
ERROS DE MONTAGEM	6	25	13
CONTIGS	837	447	817

Quadro 1 - Comparação dos montadores para *dataset* de 10 espécies

Em seguida, o *binning* foi executado pelo *MetaBAT e MaxBin*. A quantidade de *bins* gerados e os parâmetros utilizados podem ser vistos no Quadro 2:

	Idba*MaxBin*	**Idba*MetaBAT*1**	**Idba*MetaBAT*2**
BINS	**10**	**11**	**10**
PARÂMETROS	***default***	***default*, sem arquivo de depth**	***default*, com arquivo de depth**

Quadro 2 - Quantidade de *bins* gerados para *dataset* de 10 Espécies

No *dataset* 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]. O *MetaBAT* por sua vez obteve um resultado mais preciso quando o arquivo de *depth* foi utilizado, gerando 10 *bins.*

3.2 *Dataset* de 100 Espécies

Para o conjunto de 100 espécies, os montadores *Megahit* e *MetaSPades* cobriram uma fração semelhante do genoma da amostra (aproximadamente 62%). Entretanto, o *Megahit* obteve uma quantidade menor de erros de montagem (2766), sobressaindo-se ao *MetaSPades* (8925), conforme mostra o Quadro 3:

	IDBA_contig	***Megahit*_final_contig**	**Spades_contig**
FRAÇÃO DE GENOMA	50.912	62.717	62.873
MAIOR ALINHAMENTO	133685	195515	144343
ERROS DE MONTAGEM	64507	2766	8925
CONTIGS	814833037	185742	183858

Quadro 3 - Comparação dos montadores para *dataset* de 100 espécies

De forma análoga aos resultados de 10 espécies, o Quadro 4 representa a quantidade de *bins* gerados e os parâmetros utilizados para cada uma das execuções do *MaxBin* e *MetaBAT,* utilizando como entrada os *contigs* gerados pelo *Megahit*:

	***Megahit MaxBin*1**	***Megahit MaxBin*2**	***Megahit MetaBAT*1**	***Megahit MetaBAT*2**	***Megahit MetaBAT*3**
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*

Quadro 4 - Quantidade de *bins* gerados para *dataset* de 100 Espécies

Com parâmetros padrão, O *MaxBin* gerou somente 33 bins, uma quantidade baixa considerando o tamanho da amostra. Na segunda execução, utilizou-se os parâmetros *min_contig_length* e *markerset* setados para 200 e 40, respectivamente. Para chegar nesses valores, verificou-se através da ferramenta PRINSEQ que o menor *contig* gerado pelo *Megahit* possuía tamanho de 200 pares de base. Definiu-se então como 200 o tamanho mínimo de *contig.*

Já o parâmetro *markerset* representa os marcadores genéticos presentes na maioria das bactérias e é definido em 107 como *default*. Uma outra opção é utilizar o

markerset com valor 40, o que tende a dividir a saída em um número maior de bins. Com esses parâmetros, o *MaxBin* gerou 76 bins.

O *MetaBAT* não permite que o tamanho mínimo do *contig* seja configurado para 200. O tamanho mínimo do parâmetro de menor *contig* para realizar o *bin* é de 1500 [KANG et al., 2015]. Devido a isso, outros parâmetros foram utilizados.

No primeiro teste com parâmetros padrão (com tamanho mínimo de *bin* igual a 200000), o *MetaBAT* retornou como saída somente 15 arquivos de *bin* (*MegahitMetaBAT*1). Para o segundo teste, modificou-se o parâmetro *-s* que representa o tamanho mínimo de *bin* gerado. Para decidir um tamanho razoável de *bin*, verificou-se que entre os *bins* gerados pelo *MaxBin*, o menor equivalia a 130000 pares de base. Esse valor foi utilizado como parâmetro no *MetaBAT*, gerando 24 *bins* (*MegahitMetaBAT*2). Testes posteriores foram realizados com valores menores de *bin* para tentar aproximar a quantidade de *bins* à quantidade de organismos, chegando-se ao número de 12000 pares de base como tamanho mínimo. Com esse valor, 96 *bins* foram gerados pelo *MetaBAT* (*MegahitMetaBAT3*).

3.3 Matriz de Confusão (10 espécies)

O Quadro 5 representa a média das métricas da Matriz de Confusão para as melhores execuções das ferramentas de *binning* (aquelas cuja quantidade de *bins* gerados se aproxima da quantidade de organismos na amostra):

	idba*MetaBAT*2	**idba*MaxBin***
ESPECIFICIDADE	99,72%	97,99%
SENSIBILIDADE	99,52%	84,25%
ACURÁCIA	99,71%	96,63%
PRECISÃO	94,34%	75,14%

Quadro 5 - Média das métricas para idba*MetaBAT*2 e *MaxBin*

O *MetaBAT* obteve 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 que há uma quantidade maior de *contigs* contaminantes nos *bins* gerados pelo *MaxBin*.

3.4 Matriz de Confusão (100 espécies)

Para o *dataset* de 100 espécies, a Matriz de Confusão foi criada para as execuções *MegahitMaxBin2* e *MegahitMetaBAT3*. O Quadro 6 compara os resultados:

	***MegahitMetaBAT*3**	***MegahitMaxBin*2**
ESPECIFICIDADE	99,75%	99,13%
SENSIBILIDADE	39,80%	35,50%
ACURÁCIA	98,32%	98,08%
PRECISÃO	80,73%	37,47%

Quadro 6 - Média das métricas para *MegahitMetaBAT*3 e *MegahitMaxBin*2

O *MetaBAT* alcançou melhores resultados que o *MaxBin* em todas as métricas. Entretanto, ambos tiveram baixa sensibilidade, indicando que os *contigs* de uma espécie se encontram dispersos por vários *bins*.

Tanto o *dataset* de 10 quanto o de 100 espécies foram gerados com a mesma quantidade de *reads* [MENDE et al., 2012], o que fez com que o *dataset* de 100 espécies tivesse baixa cobertura dos genomas, impactando na montagem e no *binning*.

4 Conclusão

O *MetaBAT* se sobressaiu ao *MaxBin* na qualidade dos *bins* gerados, tanto para o *dataset* de 10 espécies quanto para o de 100 espécies; e o arquivo contendo os dados sobre a abundância dos *contigs* é necessário para que as ferramentas de *binning* retornem um resultado mais preciso. Além disso, 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 de *binning*, como o CONCOCT, e seu desempenho comparado com as ferramentas presentes neste estudo. Testes posteriores utilizando um valor maior de cobertura para o *dataset* de 100 espécies também se fazem necessários.

Referências

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.

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

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.

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

SCHMIEDER, R.; EDWARDS, R. Quality control and preprocessing of metagenomic datasets. **Bioinformatics**, v. 27, n. 6, p. 863–864, 2011.

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

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