Parte 3 do Desafio da Neoprospecta

João Luiz de Meirelles

15 de Julho de 2020

3.1) Montagem e anotação de genoma bacteriano isolado

- **Quality Control (QC)**

Os primeiros passos para realizar a montagem de um genoma com arquivos FASTQ são quality control, filtragem de reads com baixa qualidade e filtragem de adapters. No QC, analisaremos fatores como qualidade PHRED, conteúdo GC e conteúdo de K-mer.

Para definição de reads de baixa qualidade, o threshold PHRED seria de 25, o que significa que a probabilidade de uma base estar incorreta é de aproximadamente 3 em 1000, um bom valor para reads pequenos de Illumina MiSeq. É importante que essas etapas sejam feitas tanto para os reads forward (R1) como para reads reverse (R2) já que nossa amostra teste é paired-end.

Para visualização de parâmetros de qualidade, o software MultiQC [1] seria utilizado. Poderíamos analizar a qualidade de nossos reads em reports HTML e TXT. A etapa de filtragem seria feita utilizando Trim Galore [2], um wrapper de Cutadapt [3], onde nossos reads seriam filtrados por qualidade PHRED.

- **Montagem**

Após as etapas iniciais de quality control e filtragem, teremos que escolher entre montagem de genoma denovo ou montagem por referência para geração dos nossos contigs. Se o isolado bacteriano for de uma espécie já muito estudada e com genomas caracterizados, ou uma espécie que tem proximidade filogenética com outra espécie já caracterizada, podemos utilizar a montagem de genoma por referência. Alguns softwares utilizados para montagem por referência são SMALT [4], Bowtie2 [5] e BWA [6].

Se tivermos poucas informações sobre o isolado bacteriano ou for uma espécie que conhecemos porém sem um genoma montado ou de uma espécie próxima evolutivamente, necessitamos da montagem denovo. Alguns softwares utilizados para montagem denovo são ABySS [7], SOAPdenovo [8], SPAdes [9] e Velvet [10], onde grafos De Bruijn são utilizados como estrutura de dados para representar o overlap entre reads por meio de K-mers. Também existe a possibilidade de utilizar metodos híbridos, aonde contigs gerados pela montagem denovo são alinhados contra um genoma de referência para a geração de supercontigs e scaffolds.

Se um software de montagem denovo for o escolhido, QUAST [11] seria utilizado para testar a qualidade de diferentes valores de K-mer e definir o valor de K-mer final com melhor qualidade baseado em estatísticas de contigs e scaffolds.

- **Verificação de Qualidade**

Utilizando o software QUAST, seria possível verificar parâmetros de qualidade de contigs e scaffold como número, tamanho, N50 , N75, L50 e L75. Para obter a plenitude em termo de conteúdo de genes, podemos utilizar BUSCO [12], que faz uso de bancos de ortólogos para definir plenitude tanto em genomas montados quanto anotados.

Podemos criar um index do genoma montado e mapear os reads de input R1 e R2 com BWA, obtendo a frequência de reads mapeados em nosso genoma montado.

- **Anotação**

Para a anotação, o software utilizado seria Prokka [13]. Com ele, podemos anotar genes codificantes e RNAs não codificantes a partir dos scaffolds montados nos passos anteriores. Arquivos gff/gtf seriam gerados, possibilitando a análise dos genes anotados de cada scaffold.

3.2) Identificação taxonômica do genoma montado

A identificação taxonômica pode ser feita por vários softwares de identificação taxonômica de dados WGS de metagenômica. Um desses é o Kraken [14], que utiliza K-mers para identificação taxonômica de cada contig gerado pela montagem de genoma.

3.3) Automação da montagem do genoma

A automação dos passos 3.1 e 3.2 é possível tendo como input somente os arquivos FASTQ R1 e R2 utilizando uma linguagem de progamação como Python para automatização de programas por linha de comando em sistemas Linux e abertura, escrita e leitura de arquivos.

Para medir a qualidade, o software utilizado seria QUAST, nos possibilitando analisar todas as estatísticas de contigs e conteúdo GC. Se montagem denovo for utilizada, podemos decidir o melhor tamanho de K-mers por meio dos parâmetros de qualidades gerados pelo QUAST para cada K-mer.

Para utilizar montagem de genoma por referência, necessitaríamos do genoma de referência como um input. Sem essa mudanca, não é possivel utilizar uma montagem por referencia.

Podemos enfrentar a contaminação da amostra teste com material genético de outros OTUs. Isso dificulta a automação do pipeline e necessita de etapas de analise pós-montagem pois podem existir scaffolds com contigs de organismos diferentes do isolado esperado.

3.4) Identificação taxonômica de genoma contaminado

Para a identificação taxonômica de isolados contaminados com outras OTUs, podemos utilizar Kraken que tem como output a identificação taxonômica dos contigs resultantes da montagem. Assim, podemos saber quais contigs pertencem a nosso isolado ou a contaminantes.

Podemos analisar o conteúdo GC de contigs. Espécies diferentes podem ter diferentes conteúdo GC, logo, essa estatística pode ser interessante para diferenciar scaffolds advindos de diferentes táxons.

O BLASTn [15] de possíveis pequenos contigs pode ser um bom método, procurando identificar e assinalar taxonomia a contaminantes por meio de similaridade com sequências do NCBI.

3.5) Solucionar amostra contaminada

Como métodos de análise anterior a montagem, podemos filtrar reads baseados em cobertura e conteúdo GC. Também poderíamos usar informação prévia e mapear os reads contra um genoma de referência do nosso isolado esperado antes da montagem. Essa estratégia somente funciona quando o nosso isolado e o contaminantes não são próximos evolutivamente, por conta do alinhamento de reads.

Baseados na anotação dos scaffolds, podemos utilizar a predição de genes para a separação de scaffolds em táxons diferentes. Podemos também utilizar o BLASTn para busca de hits em contigs, dividindo-os em pedaços menores e analisando seus top-hits.

Referências

[1] S. Andrews and A. FastQC, "A quality control tool for high throughput sequence data. 2010," 2015.

[2] F. Krueger, "Trim galore," *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*, vol. 516, p. 517, 2015.

[3] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet. journal*, vol. 17, no. 1, pp. 10–12, 2011.

[4] H. Ponstingl and Z. Ning, "Smalt-a new mapper for dna sequencing reads," *F1000 Posters*, vol. 1, no. L313, 2010.

[5] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, p. 357, 2012.

[6] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[7] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol, "Abyss: a parallel assembler for short read sequence data," *Genome research*, vol. 19, no. 6, pp. 1117–1123, 2009.

[8] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S. M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. W. Lam, and J. Wang, "Erratum to "SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler"[GigaScience, (2012), 1, 18]," *GigaScience*, vol. 4, no. 1, p. 1, 2015.

[9] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, *et al.*, "Spades: a new genome assembly algorithm and its applications to single-cell sequencing," *Journal of computational biology*, vol. 19, no. 5, pp. 455–477, 2012.

[10] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de bruijn graphs," *Genome research*, vol. 18, no. 5, pp. 821–829, 2008.

[11] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "Quast: quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, 2013.

[12] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, "Busco: assessing genome assembly and annotation completeness with single-copy orthologs," *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, 2015.

[13] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, 2014.

[14] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome biology*, vol. 15, no. 3, pp. 1–12, 2014.

[15] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.