Prokaryotic Whole Genome Sequencing – Assembly and Functional Annotation of Illumina Reads

Raw Illumina reads from a whole genome sequencing project will be run through an analysis pipeline that includes quality control, read quality and adapter trimming, reference based or *de novo* assembly, gene prediction, and genome functional annotation. We can assemble a genome from pooled Illumina read libraries or single-cell reads.

Quality control will be performed using Trimmomatic [1] and FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to ensure no Illumina adapter sequences, PhiX adapters, or other contaminants are left in the reads and to trim low-quality reads from the assembly. After QC steps have been completed, SPAdes [2] can be used to either assemble using a reference genome (link to FASTA must be provided by client) or to complete a *de novo* assembly. Genome assembly can also be done with other assemblers, but SPAdes has become the *de facto* gold standard for Illumina-based genomics in prokaryotes over recent time. The quality of the genome assembly will be assessed and verified using QUAST [3], and assembly statistics which include N50, contig count, genome size, plasmid count, GC content, and genome size (bp) will be shared in a report. Gene prediction and annotation can be completed via the Prokka pipeline [4] to generate fast and accurate results which will be shared immediately, and then the completed genome will be submitted to NCBI's Prokaryotic Genome Annotation Pipeline [5] due to current NCBI genomic data publication and sharing standards. NCBI has begun requiring authors to submit any novel prokaryote genomes to their PGAP annotation pipeline which can supplement or replace an author's current annotation, and this annotation is included as part of sharing the genomic data to be published into NCBI's databases (SRA, RefSeq, etc). While this is not an issue, the amount of time taken by PGAP may vary drastically and is not up to EICC but is fully at the discretion of NCBI.

Optional analyses can include antibiotic resistance gene searching [6-7], virulence factor searching [8-9], AntiSMASH utilization for discovery of possible novel secondary metabolitic pathways and resistance variation [6], and more.

Sample publication: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6256477/

Requirements:

a. Raw genomic data in form of raw FASTQ reads
b. Reference genome if applicable
c. Type of antibiotic or drug resistance being searched for

Deliverables:

a. Raw read QC report
b. Complete Genome Assembly – FASTA file(s)
    a. Genome Assembly statistics (N50, # contigs, GC content, size in BP, etc.)
c. Complete Gene prediction and annotation – GFF3 format
    a. NCBI compliant PGAP annotation with sample/assembly submission to NCBI SRA

d. Optional further analyses.

Citations

1. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120. *doi.org/10.1039/bioinformatics/btu170*
2. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19(5):455-457.
3. Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, Alexey Gurevich, Versatile genome assembly evaluation with QUAST-LG, *Bioinformatics* (2018) 34 (13): i142-i150. doi: 10.1093/bioinformatics/bty266
4. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. doi: 10.1093/bioinformatics/btu153. Epub 2014 Mar 18.
5. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614-6624. doi: 10.1093/nar/gkw569.
6. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de Los Santos ELC, Kim HU, Nave M, Dickschat JS, Mitchell DA, Shelest E, Breitling R, Takano E, Lee SY, Weber T, Medema MH. 2017. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res 45:W36–W41. doi:10.1093/nar/gkx319GGDc
7. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN, Doshi S, Courtot M, Lo R, Williams LE, Frye JG, Elsayegh T, Sardar D, Westman EL, Pawlowski AC, Johnson TA, Brinkman FS, Wright GD, and McArthur AG. CARD 2017: expansion and model-centric curation of the Comprehensive Antibiotic Resistance Database. (2017). Nucleic Acids Research, 45, D566-573.
8. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, and Jin Q. VFDB: a reference database for bacterial virulence factors. (2005). Nucleic Acids Research Jan 1; D325-D328.
9. Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. BLAST+: architecture and applications. (2008). BMC Bioinformatics 10:421.