

Genomics

DUBii - Module 5

Valentin Loux - Olivier Rué

2020/03/09

```
fastqc -i TOTO/* fastqc -i TOTO/* fastqc -i TOTO/* fastqc -i TOTO/* fastqc -i TOTO/* fastqc -i TOTO/* f
```

toto is bbbfff Andrews (2010)

???

if not asked to lend money.

Toto

fff (Ewels, Magnusson, Lundin, and Käller, 2016)

—

- 1

—

- 2

—

Data

- Shotgun metagenome data of a defined mock community Sevim, Lee, Egan, Clum, Hundley, Lee, Everroad, Detweiler, Bebout, Pett-Ridge, and others (2019)
- Sequencage Illumina HiSeq
- SRX4901583

—

- The Road to Metagenomics
-

Quality Control

- FASTQC (Andrews, 2010)
 - MUTLIQC (Ewels, Magnusson, Lundin, et al., 2016)
-

Preprocess

- FASTP (Zhou, Chen, Chen, and Gu, 2018)
-

Taxonomic affiliation

- KAIJU (Menzel, Ng, and Krogh, 2016) on nr (Wheeler, Barrett, Benson, Bryant, Canese, Chetvernin, Church, DiCuccio, Edgar, Federhen, and others, 2006)
-

Assembly

- metaSPADES (Bankevich, Nurk, Antipov, Gurevich, Dvorkin, Kulikov, Lesin, Nikolenko, Pham, Prjibelski, and others, 2012)
 - MEGAHIT (Li, Liu, Luo, Sadakane, and Lam, 2015)
-

Evaluation

- MetaQUAST
-

Binning / Clustering

- MetaBAT (Kang, Froula, Egan, and Wang, 2015)
 - Validation with CheckM (Parks, Imelfort, Skennerton, Hugenholtz, and Tyson, 2015)
-

Annotation

Genes prediction

- prodigal (Hyatt, Chen, LoCascio, Land, Larimer, and Hauser, 2010)

Taxonomic

- KAIJU

Functional

- KEGG
 - ghostKOALA (Kanehisa, Sato, and Morishima, 2016)
-

Mapping and count

Post-process - Statistical analyses

Visualization tools

References

- Andrews, S. (2010). *FastQC A Quality Control tool for High Throughput Sequence Data*. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Bankevich, A, S. Nurk, D. Antipov, et al. (2012). “SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing”. In: *Journal of computational biology* 19.5, pp. 455-477.
- Ewels, P, M. Magnusson, S. Lundin, et al. (2016). “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19, pp. 3047-3048.
- Hyatt, D, G. Chen, P. F. LoCascio, et al. (2010). “Prodigal: prokaryotic gene recognition and translation initiation site identification”. In: *BMC bioinformatics* 11.1, p. 119.
- Kanehisa, M, Y. Sato, and K. Morishima (2016). “BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences”. In: *Journal of molecular biology* 428.4, pp. 726-731.
- Kang, D. D, J. Froula, R. Egan, et al. (2015). “MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities”. In: *PeerJ* 3, p. e1165.
- Li, D, C. Liu, R. Luo, et al. (2015). “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph”. In: *Bioinformatics* 31.10, pp. 1674-1676.
- Menzel, P, K. L. Ng, and A. Krogh (2016). “Fast and sensitive taxonomic classification for metagenomics with Kaiju”. In: *Nature communications* 7, p. 11257.

- Parks, D. H, M. Imelfort, C. T. Skennerton, et al. (2015). “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes”. In: *Genome research* 25.7, pp. 1043-1055.
- Sevim, V, J. Lee, R. Egan, et al. (2019). “Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies”. In: *Scientific data* 6.1, pp. 1-9.
- Wheeler, D. L, T. Barrett, D. A. Benson, et al. (2006). “Database resources of the national center for biotechnology information”. In: *Nucleic acids research* 35.suppl_1, pp. D5-D12.
- Zhou, Y, Y. Chen, S. Chen, et al. (2018). “fastp: an ultra-fast all-in-one FASTQ preprocessor”. In: *Bioinformatics* 34.17, pp. i884-i890. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty560. eprint: <http://academic.oup.com/bioinformatics/article-pdf/34/17/i884/25702346/bty560.pdf>. URL: <https://dx.doi.org/10.1093/bioinformatics/bty560>.