Foreword

The WG has been established by the European Commission with the aim to promote the use of NGS across the EURLs' networks, build NGS capacity within the EU and ensure liaison with the work of the EURLs and the work of EFSA and ECDC on the NGS mandate sent by the Commission. The WG includes all the EURLs operating in the field of the microbiological contamination of food and feed and this document represents a deliverable of the WG and is meant to be diffused to all the respective networks of NRLs.

Bioinformatics tools for basic analysis of Next Generation Sequencing data

Valeria Michelacci

European Union Reference Laboratory for *E. coli* including Verotoxigenic *E. coli* (VTEC),

Istituto Superiore di Sanità, Rome, Italy

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor HaDEA can be held responsible for them.

1. Introduction

In the framework of the activities of the Inter-EURLs working group on Next Generation Sequencing, an inventory was performed during Summer 2018 on the bioinformatics tools in use for the analysis of next generation sequencing (NGS) data across the National Reference Laboratories (NRLs) networks. The aim of this action was the collection of info on the most commonly used tools, to be provided to the NRLs, and the identification of potential areas of implementation. This inventory was used as the basis for compiling a list of tools routinely used by the majority of NRLs and by EURLs for NGS data analysis. The list is routinely updated in order to keep it inclusive of novel tools developed and to delete eventual tools which are no longer maintained. This document is not to be interpreted as a list of validated tools, but only as an information on those whose use is most spread among the networks of NRLs and EURLs, along with links for an easy and fast access to such tools.

Brief description of analytical steps for basic NGS analysis

Quality check: This step aims to perform a preliminary assessment on the overall quality of the sequences produced. All the tools performing quality check accept raw sequencing files in .fastq format in input.
Trimming: This step is used to remove adaptors sequences and low quality sequences. All the tools performing trimming accept raw sequencing files in .fastq format in input.
Assembly: This step involves the identification of overlapping regions among the sequencing reads included in the sequencing file (.fastq), with the aim of producing longer sequences representative of genomic regions (contigs) compiled in an output file in .fasta format. Some assemblers are specific for long single molecule sequencing reads, such as those produced by PacBio or Oxford Nanopore platforms. Assembly pipelines can use assembly tools to optimize the assembling process. Assembly correction tools correct raw contigs generated by rapid assembly methods by comparing them with consensus sequences generated through reads alignment.
Seven genes Multi Locus Sequence Typing (MLST): This step is used to type bacterial strains according to established schemes of allelic sequences of housekeeping genes.
Virulotyping: The identification of the presence of virulence genes in the sequencing files, through comparison with precompiled databases of sequences of virulence genes.
Serotype identification: Typing of the analysed bacteria by identifying serotype-associated genes in the sequencing files, through comparison with sequences in precompiled databases.
Inference on antimicrobial resistance: This step is used to predict the antimicrobial resistance of bacterial strains from whole genome sequences, through comparison with precompiled databases of antimicrobial resistance genes and with databases of known chromosomal mutations inducing resistance to antimicrobial compounds.

2. Methods

An inventory was prepared by EURL *E. coli* and administered by each participating EURL to the respective network of NRLs in August/September 2018, leaving the participation by the NRLs as voluntary. The EURLs that took part in this survey were EURLs for *Escherichia coli*, *Salmonella*, *Campylobacter* and Antimicrobial

Resistance (AMR). Later on, other tools were added in the list thanks to experience of the participants of the Inter EURLs WG on NGS.

3. Results

A total of 39 NRLs took part in this inventory, including seven NRLs for *E. coli*, 12 NRLs for *Campylobacter*, 10 NRLs for *Salmonella* and 10 NRLs for AMR. The most frequently reported tools in use were used to compile a preliminary list, which was since then widened with the addition of tools suggested by Inter EURLs WG participants.

3.1 Open-source command line tools

Quality check:

- FastQC: Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- QCumber: Available at: https://omictools.com/qcumber-tool

Trimming:

- Trimmomatic:. Available at: http://www.usadellab.org/cms/index.php?page=trimmomatic
- Bbtools: Available at: https://jgi.doe.gov/data-and-tools/bbtools/
- Fastp: Available at: https://github.com/OpenGene/fastp

Assemblers:

- SPAdes: Available at: http://cab.spbu.ru/software/spades/
  Citation: Bankevich, Anton and Nurk, Sergey and Antipov, Dmitry and Gurevich, Alexey A. and Dvorkin, Mikhail and Kulikov, Alexander S. and Lesin, Valery M. and Nikolenko, Sergey I. and Pham, Son and Prjibelski, Andrey D. and 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.
- Skesa: Available at: https://github.com/ncbi/SKESA
  Citation: Alexandre Souvorov, Richa Agarwala, David J Lipman (2018) SKESA: Strategic K-Mer Extension for Scrupulous Assemblies. In Genome Biology 2018; 19: 153.

Assemblers for long single molecule sequencing reads, such as those produced by PacBio and Oxford Nanopore Technologies platforms:

- Canu: Available at: https://github.com/marbl/canu
  Citation: Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. In Genome Res. 2017 May; 27(5): 722–736.
- Flye: Available at: https://github.com/fenderglass/Flye
  Citation: Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, Pavel A Pevzner (2019). Assembly of Long, Error-Prone Reads Using Repeat Graphs. In Nature Biotechnology 2019 May;37(5):540-546.

Assembly pipelines:

- INNUCA pipeline: Automatic pipeline for quality check, trimming and assembly of raw sequencing files. Available at: https://github.com/B-UMMI/INNUca
- Unicycler: Available at: https://github.com/rrwick/Unicycler
  Citation: Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, Kathryn E. Holt , Published in PLoS Comput Biol (2017) "Unicycler: resolving bacterial genome assemblies from short and long sequencing reads". https://doi.org/10.1371/journal.pcbi.1005595
- Shovill: Available at: https://github.com/tseemann/shovill

Assembly correction:

- Pilon (on SPAdes assembled contigs): Available at: https://github.com/broadinstitute/pilon/wiki
  Citation: Walker, Bruce J. and Abeel, Thomas and Shea, Terrance and Priest, Margaret and Abouelliel, Amr and Sakthikumar, Sharadha and Cuomo, Christina A. and Zeng, Qiandong and Wortman, Jennifer and Young, Sarah K. and et al. (2014). "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement". In PLoS ONE, 9 (11), pp. e112963. doi:10.1371/journal.pone.0112963
- Racon: Available at: https://github.com/isovic/racon
  Citation: Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research 27(5): 737–746.

Seven genes Multi Locus Sequence Typing (MLST):

- T. Seemann MLST: Available at: https://github.com/tseemann/mlst
- Metric-Oriented Sequence Typer (MOST): Available at: https://github.com/phe-bioinformatics/MOST
- MLST Finder from CGE: Available at: https://bitbucket.org/genomicepidemiology/mlst/src/master/
- BLAST on preinstalled 7 genes MLST databases for the different species: Available at: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download
  Citation: Camacho, Christiam and Coulouris, George and Avagyan, Vahram and Ma, Ning and Papadopoulos, Jason and Bealer, Kevin and Madden, Thomas L (2009). BLAST+: architecture and applications. In BMC Bioinformatics, 10 (1), pp. 421. doi:10.1186/1471-2105-10-421

Virulotyping:

- ABRicate: Available at: https://github.com/tseemann/abricate/
  Accessible through: https://www.iss.it/site/aries
- ARIBA: Available at: https://github.com/sanger-pathogens/ariba
  Citation: Hunt M, Mather AE, Sánchez-Busó L, Page A, Parkhill J, Keane JA, Harris SM. "ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads" Microb Genom. 2017 Oct; 3(10): e000131. doi: 10.1099/mgen.0.000131
- VirulenceFinder: Available at: https://bitbucket.org/genomicepidemiology/virulencefinder/src/master/

- BLAST on preinstalled databases for the different species: Available at: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download
  Citation: Camacho, Christiam and Coulouris, George and Avagyan, Vahram and Ma, Ning and Papadopoulos, Jason and Bealer, Kevin and Madden, Thomas L (2009). BLAST+: architecture and applications. In BMC Bioinformatics, 10 (1), pp. 421. doi:10.1186/1471-2105-10-421
- VFanalyzer: available at http://www.mgc.ac.cn/cgi-bin/VFs/v5/main.cgi?func=VFanalyzer

Serotype identification:

- SISTR: Available at: https://github.com/phac-nml/sistr_cmd
  Citation: Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, et al. (2016) The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. PLoS ONE 11(1): e0147101. Doi: https://doi.org/10.1371/journal.pone.0147101

Inference on antimicrobial resistance:

- Resistance Gene Identifier (RGI) on the Comprehensive Antibiotic Resistance Database (CARD) database: Available at: https://github.com/arpcard
  Citations: McArthur A.G., Waglechner N., Nizam F., Yan A., Azad M.A., Baylay A.J., Bhullar K., Canova M.J., de Pascale G., Ejim L., et al. “The comprehensive antibiotic resistance database.” Antimicrob. Agents Chemother. 2013;57:3348–3357;
  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, McArthur AG “CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database”. Nucleic Acids Res. 2017 Jan 4; 45: D566–D573. doi: 10.1093/nar/gkw1004
- ARIBA: Available at: https://github.com/sanger-pathogens/ariba
  Citation: Hunt M, Mather AE, Sánchez-Busó L, Page A, Parkhill J, Keane JA, Harris SM. “ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads” Microb Genom. 2017 Oct; 3(10): e000131. doi: 10.1099/mgen.0.000131
- ABRicate: Available at: https://github.com/tseemann/abricate/
- ARG-ANNOT: Available at: http://backup.mediterranee-infection.com/article.php?laref=282&titre=arg-annot
  Citation: Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain JM. “ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes.” Antimicrob Agents Chemother. 2014;58(1):212-20. doi: 10.1128/AAC.01310-13.
- ResFinder: Available at: https://bitbucket.org/genomicepidemiology/resfinder/src/master/
  Citation: Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. “Identification of acquired antimicrobial resistance genes.” J Antimicrob Chemother. 2012 Jul 10; Bortolaia V, Kaas RF, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AR, Fagelhauer L, Chakraborty T, Neumann B, Werner G, Bender JK, Stingl K, Nguyen M, Coppens J, Xavier BB, Malhotra-Kumar S, Westh H, Pinholt M, Anjum MF, Duggett NA, Kempf I, NykÃ¤senoja S, Olkkola S, Wieczorek K, Amaro A, Clemente L, Mossong J, Losch S,

Ragimbeau C, Lund O, Aarestrup FM. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy, 75(12),3491-3500

- PointFinder: Available at: https://bitbucket.org/genomicepidemiology/pointfinder/src/master/
  Zankari E, Allesøe R, Joensen KG, Cavaco LM, Lund O, Aarestrup FM. (2020) PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. Journal of Antimicrobial Chemotherapy 72(10) 2764-2768.

3.2 Commercial softwares

This list includes suites of commercial softwares integrating solutions for several steps among those illustrated above.

Ridom SeqSphere:
It includes, among others, modules for quality check, trimming and assembly of raw sequencing files.
Information available at: https://www.ridom.de/seqsphere/
Analytical steps mainly mentioned to be operated through this solution in the inventory across NRLs:
Quality check, Assembly, Seven genes Multi Locus Sequence Typing (MLST)

CLC Genomics Workbench:
It includes, among others, modules for quality check, trimming and assembly of raw sequencing files.
Information available at: https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/
Analytical steps mainly mentioned to be operated through this solution in the inventory across NRLs:
Quality check, Assembly

BioNumerics (Applied Maths):
It includes, among others, modules for quality check, trimming and assembly of raw sequencing files.
Information available at: http://www.applied-maths.com/bionumerics
Analytical steps mainly mentioned to be operated through this solution in the inventory across NRLs:
Trimming, Seven genes Multi Locus Sequence Typing (MLST)
Please note that Bionumerics 8.1 is the final version of the software, which will no longer be supported after December 31, 2024 (https://www.applied-maths.com/news/bionumerics-phasing-out).

3.3 Webservers

This list includes webservers offering easy-to-use solutions for several steps of NGS analysis.

ARIES, "Advanced Research Infrastructure for Experimentation in genomics", based on Galaxy system:
Accessible through: https://www.iss.it/site/aries.
Developed by EURL VTEC
Analytical steps mainly mentioned to be operated through this solution in the inventory across NRLs:

- *E. coli* virulotyper:
  Accessible through: https://www.iss.it/site/aries

- *E. coli* serotyper:
  Accessible through: https://www.iss.it/site/aries

Center for genomics epidemiology (CGE):
Accessible through: http://www.genomicepidemiology.org/
Developed by Technical University of Denmark (DTU)
Tools most mentioned to be operated through this solution in the inventory across NRLs:

- Assembler 1.2 Accessible through: https://cge.cbs.dtu.dk/services/Assembler/.
  For automatic trimming and assembly. The output consists in assembled contigs in fasta format.
  Citation: Multilocus Sequence Typing of Total Genome Sequenced Bacteria. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM and Lund O. J. Clin. Micobiol. 2012. 50(4): 1355-1361.
- MLST 2.0 (Multi-Locus Sequence Typing): Accessible through: https://cge.cbs.dtu.dk/services/MLST/
  For MLST analysis based on classical schemes of housekeeping *loci*.
  Citation: Multilocus Sequence Typing of Total Genome Sequenced Bacteria. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM and Lund O. J. Clin. Micobiol. 2012. 50(4): 1355-1361.
- SalmonellaTypeFinder:
  Salmonella serotyping
  Accessible through: https://cge.cbs.dtu.dk/services/SalmonellaTypeFinder/
- Virulence Finder 2.0:
  Accessible through: https://cge.cbs.dtu.dk/services/VirulenceFinder/
  Citation: Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. J. Clin. Micobiol. 2014. 52(5): 1501-1510.
- Serotype Finder 2.0:
  Accessible through: https://cge.cbs.dtu.dk/services/SerotypeFinder/
  Citation: Joensen, K. G., A. M. Tetzschner, A. Iguchi, F. M. Aarestrup, and F. Scheutz. 2015. Rapid and easy in silico serotyping of Escherichia coli using whole genome sequencing (WGS) data. J.Clin.Microbiol. 53(8):2410-2426. doi:JCM.00008-15 [pii];10.1128/JCM.00008-15
- SeqSero 1.2:
  Accessible through: https://cge.cbs.dtu.dk/services/SeqSero/
  Citation: Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. "Salmonella Serotype Determination Utilizing High-throughput Genome Sequencing Data." J. Clin. Micobiol. 2015.
- ResFinder:
  Inference on antimicrobial resistance
  Accessible through: https://cge.cbs.dtu.dk/services/ResFinder/
  Citation: Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. "Identification of acquired antimicrobial resistance genes." J Antimicrob Chemother. 2012 Jul 10; Bortolaia V, Kaas RF, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AR, Fagelhauer L, Chakraborty T, Neumann B, Werner G, Bender JK, Stingl K,

Nguyen M, Coppens J, Xavier BB, Malhotra-Kumar S, Westh H, Pinholt M, Anjum MF, Duggett NA, Kempf I, NykÃ¤senoja S, Olkkola S, Wieczorek K, Amaro A, Clemente L, Mossong J, Losch S, Ragimbeau C, Lund O, Aarestrup FM. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy, 75(12),3491-3500; Zankari E, Allesøe R, Joensen KG, Cavaco LM, Lund O, Aarestrup FM. (2020) PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. Journal of Antimicrobial Chemotherapy 72(10) 2764-2768.

3.4 EFSA One Health WGS System and analytical pipeline

On the specific mandate by the European Commission to "Request for the implementation of a 'One Health' system for the collection and analysis of whole-genome sequencing (WGS) data from human and food/animal isolates", European Food Safety Authority (EFSA) and European Center for Disease Prevention and Control (ECDC) have set up two interoperable systems for the collection and sharing of WGS data from non-human and human isolates of *Salmonella, L. monocytogenes*, and *E. coli* officially provided by Member States, allowing the joint analysis of WGS data for for the purpose of multi-country outbreak detection and assessment.

In this context, EFSA developed a pipeline for the automatic analysis of WGS data of *Salmonella, L. monocytogenes*, and *E. coli*, which is not only accessible for the officially appointed users through the EFSA One Health WGS System Portal, but also available for download and setup through the following repository: https://dev.azure.com/efsa-devops/EFSA/_git/efsa.wgs.onehealth**.** Details on the EFSA analytical pipeline are available in the "Guidelines for reporting Whole Genome Sequencing-based typing data through the EFSA One Health WGS System" available as EFSA supporting publication https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2022.EN-7413

3.5 Other webservers dedicated to selected tools

The webservers listed here offer specific tools, as described below.

- PubMLST
  Cited among the most frequently used solutions for Seven genes Multi Locus Sequence Typing (MLST) in the inventory across NRLs. Also offering cgMLST for *Campylobacter*, *Salmonella* and *E. coli*, among others out of the scope of the working group.
  Accessible through: https://pubmlst.org/
  Citation: Jolley KA, Bray JE and Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; peer review: 2 approved]. Wellcome Open Res 2018, 3:124.
- BIGSdb-Lm
  Offering MLST and cgMLST for *Listeria monocytogenes*.
  Accessible through: https://bigsdb.pasteur.fr/listeria/.
- Enterobase

Cited among the most frequently used solutions for Seven genes Multi Locus Sequence Typing (MLST) in the inventory across NRLs. Also offering wgMLST and cgMLST for *Salmonella* and *E. coli*, among others out of the scope of the working group.
Accessible through: https://enterobase.warwick.ac.uk/
Citation: Alikhan NF, Zhou Z,Sergeant MJ, Achtman M (2018) "A genomic overview of the population structure of Salmonella." PLoS Genet 14 (4): e1007261