

List of online bioinformatics tools and software used for capacity building (status January 2018)

This document describes the most commonly used software and algorithms for processing whole genome sequencing. It is divided into categories, which describe the key processes for analysing short read data. Tools of particular interest will be tag with a specific character (historical†, commonly used*, easy to run#, etc). We are aware that the list is not complete, and that we present the status as of January 2018. It should be also taken into account that the area is continuously under development and new tools, not included here, will be released.

Most of the presented tools are command line based. In order to use them, you will need to install them on your infrastructure. We highly recommend that you ensure to have proper settings for your infrastructure (i.e. storage capacity and memory to run tools/software) as some of them require a lot of resources. In case you want to try these software and do not have infrastructure, we can recommend you to run Bio-Linux using a Virtual Box¹

Quality Assessment and Trimming

This is the process by which the quality of fastq files is determined and subsequent optional trimming of the data to trim or remove poor quality reads is carried out.

- Trimmomatic*
 - <http://www.usadellab.org/cms/index.php?page=trimmomatic>
 - Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bolger, A. M., Lohse, M., & Usadel, B. (2014). *Bioinformatics*, btu170.
 - Windows, Mac OS X and Linux
 - A flexible read trimming tool that will remove Illumina adapters, reads below a certain length and low quality ends of the read
 - *Comments:* Trimming occurs in the order, which the steps are specified on the command line. It is recommended in most cases that adapter clipping, if required, is done as early as possible. Options will strongly depend on the data you used i.e. single end, paired end.
- FastQC*#
 - <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
 - Windows, Mac OS X and Linux
 - A quality control tool for assessing the quality of NGS data
 - *Comments:* tool available both online/command line. If running more than few sample using the command line is recommended. Interpretation of the results is linked to the sequencing method used. The online documentation details all the warnings and how to interpret them.
- Seqtk
 - <https://github.com/lh3/seqtk>
 - Windows, Mac OS X and Linux
 - Tool for processing sequences in the FASTA or FASTQ format that can be used for adapter removal and trimming of low-quality bases

¹ Installation of Virtual Box tutorial is included at the end of this Appendix.

- FastX
 - http://hannonlab.cshl.edu/fastx_toolkit/
 - Windows, Mac OS X and Linux
 - Toolkit for FASTQ and FASTA pre-processing that can be used for trimming, clipping, barcode splitting, formatting and quality trimming.

Assembly

This is the process of joining short/long reads into longer contigs (contiguous lengths of DNA) without the need for a reference sequence.

- VelvetK
 - <http://www.vicbioinformatics.com/software/velvetk.shtml>
 - Windows, Mac OS X and Linux
 - Perl script to estimate best k-mer size to use for your Velvet de novo assembly.
- VelvetOptimiser
 - <http://www.vicbioinformatics.com/software/velvetk.shtml>
 - Mac OS X and Linux
 - Perl script to assist with optimising the assembly.
 - *Comments:* optimisation can be made using different metrics (e.g. with best N50, best coverage...)
- KmerGenie
 - <http://kmergenie.bx.psu.edu/>
 - Informed and Automated k-Mer Size Selection for Genome Assembly. Chikhi R., Medvedev P. HiTSeq 2013.
 - Windows, Mac OS X and Linux
 - Best k-mer length estimator for single-k genome assemblers like velvet.
- Khmer
 - <http://khmer.readthedocs.io/en/v2.0/>
 - The khmer software package: enabling efficient nucleotide sequence analysis. Crusoe et al., 2015. F1000 <http://dx.doi.org/10.12688/f1000research.6924.1>
 - Linux and Mac OS X
 - Set of command-line tools for dealing with large and noisy datasets to normalise and scale the data for more efficient genome assembly.
- Minia
 - <http://minia.genouest.org/>
 - Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Chikhi, Rayan and Rizk, Guillaume. Algorithms for Molecular Biology, BioMed Central, 2013, 8 (1), pp.22.
 - Windows, Mac OS X and Linux
 - Short-read assembler based on a de Bruijn graph for low-memory assembly.
- SPAdes^{*,#}
 - <http://cab.spbu.ru/software/spades/>
 - SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing, Anton Bankevich, Sergey Nurk, Dmitry Anipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A.

Alekseyev, and Pavel A. Pevzner. Journal of Computational Biology 19(5) (2012), 455-477. doi:10.1089/cmb.2012.0021

- Mac OS X and Linux
- Short and hybrid-long read assembler based on a de Bruijn graph that also performs error correction and is a multi-k genome assembler.
- *Comments:* Illumina Paired reads (2*150 and 2*250) need to be assemble with the specific option --careful (see application note for full details) to get the best assembly possible
- Velvet†*
 - <https://www.ebi.ac.uk/~zerbino/velvet/>
 - Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Daniel R. Zerbino and Ewan Birney. Genome Res. May 2008 18: 821-829; Published in Advance March 18, 2008, doi:10.1101/gr.074492.107
 - Linux
 - De novo short read genome assembler with error correction to produce high quality unique contigs.
 - *Comments:* parameters can be difficult to select, some scripts have been developed and are working well to help choose the best parameters. Optimisation of the option should be used: VelvetOptimiser or VelvetK
- Canu
 - <http://canu.readthedocs.io/en/stable/index.html>
 - Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Adam M. Phillippy doi: <http://dx.doi.org/10.1101/071282>
 - Windows, Mac OS X and Linux
 - Long-read assembler designed for high-noise data such as that generated by PacBio or Oxford Nanopore MinION. Canu also performs error correction.
 - *Comments:* specifically designed to work with long read
- Unicycler
 - <https://github.com/rrwick/Unicycler>
 - Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, Kathryn E. Holt , Published in PLoS Comput Biol (2017) <https://doi.org/10.1371/journal.pcbi.1005595>
 - Mac OS X and Linux
 - Unicycler is an assembly pipeline for bacterial genomes. It can assemble Illumina-only read sets where it functions as a SPAdes-optimiser. It can also assemble long-read-only sets (PacBio or Nanopore) where it runs a miniasm+Racon pipeline. For the best possible assemblies, give it both Illumina reads and long reads, and it will conduct a hybrid assembly.
 - *Comments:* use mainly as hybrid assembly for long read associated with Illumina read. Well documented with a Wiki-tutorial <https://github.com/rrwick/Unicycler/wiki/Tips-for-finishing-genomes>

- Bandage[#]
 - <http://rrwick.github.io/Bandage/>
 - Bandage: interactive visualization of de novo genome assemblies. Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bioinformatics (2015) 31 (20): 3350-3352 first published online June 22, 2015 doi:10.1093/bioinformatics/btv383
 - Linux and Mac
 - Program for visualising de novo assembly graphs by displaying connection which are not present in the contigs file for assembly assessment.
 - *Comments:* possibility to use blast inside the software to annotate regions of interest. Can help determine relations between contigs.

Annotation

The process which takes the raw sequence of contigs resulting from assembly and marks it with features such as gene names and putative functions.

- Prokka^{*#}
 - <http://www.vicbioinformatics.com/software.prokka.shtml>
 - Prokka: rapid prokaryotic genome annotation. Seemann T. Bioinformatics. 2014 Jul 15;30(14):2068-9. PMID:24642063
 - Windows, Mac OS X and Linux
 - Software tool for the rapid annotation of prokaryotic genomes.
- RAST
 - <http://rast.nmpdr.org/>
 - The RAST Server: Rapid Annotations using Subsystems Technology. Aziz RK et al.. BMC Genomics, 2008
 - Online tool
 - Fully-automated service for annotating complete or nearly complete bacterial and archaeal genomes.
- Genix
 - http://labbioinfo.ufpel.edu.br/cgi-bin/genix_index.py
 - Online tool
 - Fully automated pipeline for bacterial genome annotation.
- Prodigal
 - <https://github.com/hyattprod/Prodigal/wiki/Introduction>
 - Hyatt, Doug et al. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." BMC Bioinformatics 11 (2010): 119. PMC. Web. 25 Apr. 2018.
 - Windows, Mac OS X, GenericUnix (Linux)
 - Prodigal is a software is a protein-coding gene prediction software tool for bacterial and archaeal genomes
- NCBI Prokaryotic Genome Annotation Pipeline (PGAP)
 - https://www.ncbi.nlm.nih.gov/genome/annotation_prok/
 - Online tool – available for GenBank submitters only
 - PGAP is a pipeline for prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements

Alignment or sequence searching

Tools to align a sequence to other sequences locally or against publically available nucleotide or protein archives.

- BLAST^{†#*}
 - <http://blast.ncbi.nlm.nih.gov/Blast.cgi>
 - Basic local alignment search tool. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman. Journal of Molecular Biology, Volume 215, Issue 3, 5 October 1990, Pages 403-410
 - Windows, Mac OS X and Linux
 - Search tool to find regions of similarity between biological sequences through alignment and calculating statistical significance.
 - Comments: classic methods to search for specific sequence. Different version can be used such as blastn or megablast depending on the similarity between biological sequences. Possibility to create local specific database with makeblastdb.
- MUMmer
 - <http://mummer.sourceforge.net/>
 - Versatile and open software for comparing large genomes. A.L. Delcher, A. Phillippy, J. Carlton, and S.L. Salzberg, Nucleic Acids Research (2002), Vol. 30, No. 11 2478-2483.
 - Windows, Mac OS X and Linux
 - A system for rapidly aligning entire genomes and finding matches in DNA sequences.
- Clustal suite – ClustalO and ClustalW
 - <http://www.clustal.org>
 - Thompson JD, Higgins DG, Gibson TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res., 22, 4673-4680.
 - Sievers et al. (2011) Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega. Molecular Systems Biology, 10.1038/msb.2011.75
 - Windows, Mac OS X and Linux and online (webservers)
 - Software that preforms sequences alignments. Mostly based on sequence weighting, position-specific gap penalties and weight matrix choice.
 - Comments: ClustalO is usually present as performing better (faster and more accurate) than the original version of ClustalW.
- MUSCLE^{*#†}
 - <https://www.drive5.com/muscle/>
 - Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, (5) 113
 - Windows, Mac OS X and Linux and online (webservers)
 - Software for multiple alignment of protein sequences.

Mapping

Alignment of short reads against a reference sequence so that amount of coverage or variations compared to the reference can be assessed.

- BWA^{*#}
 - <http://bio-bwa.sourceforge.net/>
 - Fast and accurate short read alignment with Burrows-Wheeler Transform. Li H. and Durbin R. (2009) Bioinformatics, 25:1754-60. [PMID: 19451168]
 - Windows, Mac OS X and Linux

- Software package for mapping low-divergent sequences against a large reference genome using the Burrows-Wheeler transform algorithm.
- Bowtie 2*#
 - <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>
 - Fast gapped-read alignment with Bowtie 2. Langmead B, Salzberg S. Nature Methods. 2012, 9:357-359.
 - Windows, Mac OS X and Linux
 - Tool for aligning sequencing reads to long reference genomes also based on the Burrows-Wheeler transform algorithm.
- Tablet
 - <https://ics.hutton.ac.uk/tablet/>
 - Using Tablet for visual exploration of second-generation sequencing data. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD and Marshall D. 2013. Briefings in Bioinformatics 14(2), 193-202.
 - Windows, Mac OS X and Linux
 - Comments: Lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments that can be used to view mapping.

Assembly refinement

Process of curating assembly by re-using reads and re-mapping steps.

- Pilon
 - <https://github.com/broadinstitute/pilon/wiki>
 - Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, Ashlee M. Earl (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE 9(11): e112963. doi:10.1371/journal.pone.0112963
 - Windows, Mac OS X, Linux
 - Java based software that automatically improve draft assemblies. Find variation among strains, including large event detection.
 - Comments: assembly need to be performs prior to use the software.
- FGAP
 - <https://github.com/pirovc/fgap>
 - Piro, Vitor C et al. "FGAP: An Automated Gap Closing Tool." BMC Research Notes 7 (2014): 371. PMC
 - Online servers or Linux and Mac OS X
 - FGAP is a tool for closing gaps of draft genome. It uses BLAST to align multiple contigs against a draft genome assembly aiming to find sequences that overlap gaps. The algorithm selects the best sequence to fill and eliminate the gaps.

Assembly statistics and quality assessment

- Quast*#
 - <http://quast.sourceforge.net/>
 - Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies,

Bioinformatics (2013) 29 (8): 1072-1075. doi: 10.1093/bioinformatics/btt086

First published online: February 19, 2013

- Linux, MAC OS X and online servers
- QUAST is a tool design to evaluates assembly. Calculates metrics such as N50, number of contigs, length of assemblies, GC content.
- Comments: this tools accept multiple assemblies and is suitable for comparing assemblies.

Variant Calling

Variant calling is the process by which variants (differences) are identify from sequence data. It usually follows the step of mapping reads against a reference.

- SAMtools*
 - <http://samtools.sourceforge.net/>
 - The Sequence alignment/map (SAM) format and SAMtools. Li H.* , Handsaker B.* , Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) Bioinformatics, 25, 2078-9. [PMID: 19505943]
 - Windows, Mac OS X and Linux
 - Toolkit that provides various utilities for manipulating alignments in the SAM format and also can be used for generating consensus sequences and variant calling
- GATK*
 - <https://software.broadinstitute.org/gatk/>
 - The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernysky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. Genome Res. September 2010 20: 1297-1303; Published in Advance July 19, 2010, doi:10.1101/gr.107524.110
 - Windows, Mac OS X and Linux
 - Toolkit with a primary focus on variant discovery and genotyping.
- Picard
 - <http://broadinstitute.github.io/picard/>
 - Windows, Mac OS X and Linux
 - A set of command line tools (in Java) for manipulating high-throughput sequencing data and formats.
 - Comments: command line only, but helpful to convert/sort and use different output bam, sam...
- Varscan (version 2)
 - <http://dkoboldt.github.io/varscan/>
 - VarScan 2: Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing Genome Research DOI: 10.1101/gr.129684.111
 - Windows, Linux and Mac OS X
 - A set of command line tools running with Java that detects different kind of variants such as Germline variants (SNPs and indels), Multi-sample variants (shared or private) in multi-sample datasets (with mpileup), Somatic mutations, Somatic copy number alterations (CNAs).

Phylogenetic analysis

Assessment of the evolutionary relationship between strains using either distance-based or Bayesian methodologies.

- RaxML*
 - <http://sco.h-its.org/exelixis/web/software/raxml/index.html>
 - RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. A. Stamatakis. Bioinformatics (2014) 30 (9): 1312-1313.
 - Windows, Mac OS X and Linux
 - Randomized Accelerated Maximum Likelihood program for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees.
 - Comments: maximum-likelihood methods give more resolution/accuracy than FastTree but take longer to run. Substitution models can be used as parameters.
- FastTree*#
 - <http://www.microbesonline.org/fasttree/>
 - FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). Molecular Biology and Evolution 26:1641-1650, doi:10.1093/molbev/msp077.
 - Windows, Mac OS X and Linux
 - Comments: Faster tool for speedy inference of approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences. Particularly useful to quickly generate trees.
- CSI Phylogeny*#
 - <https://cge.cbs.dtu.dk/services/CSIPhylogeny/>
 - Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms. Rolf S. Kaas, Pimlapas Leekitcharoenphon, Frank M. Aarestrup, Ole Lund. PLoS ONE 2014; 9(8): e104984.
 - Comments: Online tool, easy to use and configure. Tool to call SNPs, filter the SNPs and to do site validation and inference of phylogeny through a graphical user interface.
- Harvest
 - <https://www.cbcb.umd.edu/software/harvest>
 - The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Treangen TJ, Ondov BD, Koren S, Phillippy AM. Genome Biology, 15 (11), 1-15
 - Windows, Mac OS X and Linux
 - Suite of core-genome alignment and visualization tools for quickly analysing thousands of intraspecific microbial genomes, including variant calls, recombination detection, and phylogenetic trees.
 - Comments: parsnp from this tool can compute trees based on very large number of assembled genomes.
- Gubbins
 - <http://sanger-pathogens.github.io/gubbins/>
 - Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Croucher N. J., Page A. J., Connor T. R., Delaney A. J., Keane J. A., Bentley S. D., Parkhill J., Harris S.R. doi:10.1093/nar/gku1196, Nucleic Acids Research, 2014.
 - Windows, Mac OS X and Linux

- Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) is an algorithm that iteratively identifies loci containing elevated densities of base substitutions while concurrently constructing a phylogeny based on the putative point mutations outside of these regions.
- Comments: detection of recombination and generation of phylogeny. Depending on the number of genomes to analyse, this tool can be really long to run.
- BEAST
 - <http://beast.bio.ed.ac.uk/>
 - Bayesian phylogenetics with BEAUti and the BEAST 1.7. Drummond AJ, Suchard MA, Xie D & Rambaut A (2012) Molecular Biology And Evolution 29: 1969-1973.
 - Windows, Mac OS X and Linux
 - Cross-platform program for Bayesian analysis of molecular sequences using MCMC.
 - Comments: can be use to generate phylogeny based on prior information like time. Useful if you expect some time-relation in your phylogeny but really long to run.
- FigTree*#
 - <http://tree.bio.ed.ac.uk/software/figtree/>
 - Windows, Mac OS X and Linux
 - A graphical viewer of phylogenetic trees and program for producing publication-ready figures of trees.
 - Comments: easy tools to visualise/manipulate trees
- I-TOL*#
 - <https://itol.embl.de/>
 - Letunic I and Bork P (2016) Nucleic Acids Res doi: 10.1093/nar/gkw290 Interactive Tree Of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees
 - Online server
 - I-TOL Interactive Tree Of Life is an online tool for the display, annotation and management of phylogenetic trees.
 - Comments: This is only visualisation. Registration to have a workspace to save/manipulate tree. Really powerful to view large/complex tree. An extensive range of annotation available.
- Mega†
 - <http://www.megasoftware.net/>
 - MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Kumar S, Stecher G, and Tamura K (2016) Molecular Biology and Evolution 33:1870-1874
 - Windows, Mac OS X and Linux
 - Comments: Sophisticated and user-friendly software suite for analysing DNA and protein sequence data from species and populations. Contains building tree algorithms.

Virulence and antimicrobial resistance gene prediction

Inference of potential for a virulent phenotype or resistance to an antimicrobial based on nucleotide sequences.

Virulence prediction

- PathogenFinder
 - <https://cge.cbs.dtu.dk/services/PathogenFinder/>

- PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. Cosentino S, Voldby Larsen M, Møller Aarestrup F, Lund O. (2013) PLoS ONE 8(10): e77302.
- Online tool
- Web-server for the prediction of bacterial pathogenicity by analysing the input proteome, genome, or raw reads provided by the user.

Antimicrobial resistance prediction

- Resfinder
 - <https://cge.cbs.dtu.dk/services/ResFinder/>
 - Identification of acquired antimicrobial resistance genes. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. J Antimicrob Chemother. 2012 Jul 10
 - Online tool
 - Web-server that identifies acquired antimicrobial resistance genes in total or partial sequenced isolates of bacteria.
- ARIBA
 - <https://github.com/sanger-pathogens/riba>
 - ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Martin Hunt, Alison E Mather, Leonor Sánchez-Busó, Andrew J Page, Julian Parkhill, Jacqueline A Keane, Simon R Harris. doi: <https://doi.org/10.1099/mgen.0.000131>
 - ARIBA (Antimicrobial Resistance Identification By Assembly), identifies AMR-associated genes and single nucleotide polymorphisms directly from short reads
 - Comments: can also be used for MLST calling, you need to provide your reference set.
- KmerResistance
 - <https://cge.cbs.dtu.dk/services/KmerResistance-2.2/>
 - Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data
Philip T.L.C. Clausen, Ea Zankari, Frank M. Aarestrup, Ole Lund
Journal of Antimicrobial Chemotherapy. 2016
 - KmerResistance is a tool on a web-server that identifies antimicrobial resistance genes based on read mapping. It examines the co-occurrence of k-mers between the WGS data and a database of resistance genes.
 - Comments: reads mapping based detection of AMR genes is a great alternative to assembly based methods.
- SRST2
 - <https://github.com/katholt/srst2>
 - SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Inouye et al. Genome Medicine. 2014
 - Linux, Mac OS X
 - Short Read Sequence Typing for Bacterial Pathogens (SRST2) is designed to take Illumina sequence data, a MLST database and/or a database of gene sequences (e.g. resistance genes, virulence genes, etc) and report the presence of STs and/or reference genes.
- GeneFinder
 - In-house tool developed by Public Health England (PHE, UK)
 - Genefinder software is a tool to determine presence and absence of genes and retrieve specific sequence variations from NGS paired-end fastq files, using a set of reference sequences in FASTA format

- CARD: The Comprehensive Antibiotic Resistance Database (not a tool)
 - <https://card.mcmaster.ca/home>
 - CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Jia et al. Nucleic Acids Res. 2017. DOI:10.1093/nar/gkw1004
 - Contain an online pipeline RGI (Resistance Gene Identifier) to identify/query the CARD database for your genomes.
 - Database of resistance genes, their products and associated phenotypes.
 - Comments: useful resource for AMR. RGI need assemblies to run.

Species and serovar identification

Tools and software that uses various algorithms methods to identify a specie by using reads or assembly and predict serovar. These software relies on databases to predict species or serovar.

- Kraken
 - <https://ccb.jhu.edu/software/kraken/>
 - Kraken: ultrafast metagenomic sequence classification using exact alignments. Wood DE, Salzberg SL. Genome Biology 2014, 15:R46.
 - Linux
 - System for assigning taxonomic labels to short DNA sequences, usually obtained through metagenomics studies.
- MetaPhlan2
 - <https://bitbucket.org/biobakery/metaphlan2>
 - MetaPhlan2 for enhanced metagenomic taxonomic profiling. Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower & Nicola Segata. Nature Methods 12, 902-903 (2015)
 - Linux – command line
 - MetaPhlan 2: Metagenomic Phylogenetic Analysis - profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes and Viruses) from metagenomic shotgun sequencing data with species-level. The StrainPhlan module allows to perform accurate strain-level microbial profiling.
- Kmerfinder
 - <https://cge.cbs.dtu.dk/services/KmerFinder/>
 - Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. Hasman H, Saputra D, Sicheritz-Pontén T, Lund O, Svendsen CA, Frimodt-Møller N, Aarestrup FM. J Clin Microbiol. 2014 Jan;52(1):139-46.
 - Online tools and standalone Linux version
 - Tool to identify species from an assembly or reads based on k-mer detection, searching k-mer from a pre-build database (bacterial, fungi viruses...).
 - Comments: the online tool can be used with different database
- SISTR[#]
 - <https://lfz.corefacility.ca/sistr-app/>
 - The Salmonella In Silico Typing Resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. Catherine Yoshida, Peter Kruczkiewicz, Chad R. Laing, Erika J. Lingohr, Victor P.J. Gannon, John H.E. Nash, Eduardo N. Taboada. PLoS ONE 11(1): e0147101. doi: 10.1371/journal.pone.0147101
 - Web based application or standalone version on Linux and Mac OS X

- SISTR is a prediction software that predict serovar predictions from whole-genome sequence assemblies by determination of antigen gene. It also includes MLST, rMLST and cgMLST gene alleles prediction.
- SeqSero
 - <http://www.denglab.info/SeqSero>
 - 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 Microbiol. 2015 May;53(5):1685-92.PMID:25762776
 - Online webserver and command line (Unix based)
 - SeqSero is a pipeline for *Salmonella* serotype determination from raw sequencing reads or genome assemblies
- MOST
 - <https://github.com/phe-bioinformatics/MOST>
 - Tewolde, Rediat et al. "MOST: A Modified MLST Typing Tool Based on Short Read Sequencing." Ed. Nicholas Loman. PeerJ 4 (2016): e2308. PMC. Web. 25 Apr. 2018.
 - Command line (Unix based)
 - MOST is a software derived from SISTR that assign MLST profile and infer Salmonella serotyping from bacterial genomic short read sequence data
 - Comments: require MLST database, detects novel allele if not present in the database, quality of the results assess by different metrics. Can be run in a Galaxy environment.
- Serotypefinder
 - <https://cge.cbs.dtu.dk/services/SerotypeFinder/>
 - 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
 - Online tool
 - SerotypeFinder identifies the serotype in total or partial sequenced isolates of E. coli.

Comparative genomic tools

Comparison of multiple genomes to determine regions of similarity or difference either on a gene-by gene basis or across the whole genome.

- BEDTools
 - <http://bedtools.readthedocs.io/en/latest/index.html>
 - BEDTools: a flexible suite of utilities for comparing genomic features. Aaron R. Quinlan and Ira M. Hall. Bioinformatics (2010) 26 (6): 841-842 first published online January 28, 2010 doi:10.1093/bioinformatics/btq033
 - Mac OS X and Linux
 - Toolkit for the manipulation of genome data for genomic analysis tasks on genomic intervals from multiple files.
- Roary
 - <https://sanger-pathogens.github.io/Roary/>
 - Roary: Rapid large-scale prokaryote pan genome analysis. Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, Julian Parkhill. Bioinformatics, 2015;31(22):3691-3693 doi:10.1093/bioinformatics/btv421.
 - Windows, Mac OS X and Linux

- High speed stand-alone pan genome pipeline, which takes annotated assemblies in GFF3 format and calculates the pan genome.
- Mauve
 - <http://darlinglab.org/mauve/mauve.html>
 - Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Aaron C.E. Darling, Bob Mau, Frederick R. Blattner, and Nicole T. Perna. Genome Res. July 2004 14: 1394-1403; doi:10.1101/gr.2289704
 - Windows, Mac OS X and Linux
 - Interactive genome alignment software that allows for easy browsing of multiple genomes to look for similarities and differences.
- ACT
 - <http://www.sanger.ac.uk/science/tools/artemis-comparison-tool-act>
 - ACT: the Artemis Comparison Tool. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG and Parkhill. Bioinformatics (Oxford, England) 2005;21;16;3422-3. PUBMED: 15976072; DOI: 10.1093/bioinformatics/bti553
 - UNIX, MacOS and Windows
 - Java application for displaying pairwise comparisons between two or more DNA sequences and allowing browsing of detailed annotation
- BRIG
 - <http://brig.sourceforge.net/>
 - BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. NF Alikhan, NK Petty, NL Ben Zakour, SA Beatson (2011). BMC Genomics, 12:402. PMID: 21824423
 - UNIX, MacOS and Windows
 - Image generating software that displays circular blast comparisons between a large number of genomes or DNA sequences
- EasyFig
 - <http://mjsull.github.io/Easyfig/>
 - Easyfig: a genome comparison visualiser. Sullivan MJ, Petty NK, Beatson SA. (2011) Bioinformatics; 27 (7): 1009-1010.PMID: 21278367
 - UNIX, MacOS and Windows
 - Python application for creating linear comparison figures of multiple genomic loci with an easy-to-use graphical user interface (GUI)
- SeqFindR
 - <https://github.com/mscook/SeqFindR>
 - UNIX and MacOS
 - Tool to easily create informative genomic feature plots by detecting the presence or absence of genomic features from a database in a set of genomes

Cloud Services

If infrastructure is not available the cloud based services are worth considering

- **Genomics-Specific**
- MRC CLIMB
 - <http://www.climb.ac.uk/>
 - Microbial bioinformatics cyber-infrastructure.
- Genomics Virtual Laboratory
 - <https://www.gvl.org.au/>
 - A genomics-specific version of Galaxy

- Galaxy
 - <https://usegalaxy.org/>
 - an open source, web-based platform for data intensive biomedical research.
 - **Non-Genomics Specific**
- Amazon Web Services
 - <https://aws.amazon.com>
 - Pay per usage cloud computing managed by amazon.com for temporary computing of big data
- Azure (Microsoft)
 - <https://azure.microsoft.com/en-us/>
 - Multiple services divided into the following categories: AI + Machine Learning, Analytics, Compute, Containers, Databases, Developer Tools, DevOps, Identity, Integration, Internet of Things, Management Tools, Media, Migration, Mobile, Networking, Security, Storage, Web

Commercial software

- Bionumerics Seven
 - <http://www.applied-maths.com/applications>
 - Offers a range of tools to analyse sequence data including MLST, wgMLST, AMR profiling, wgSNPs.
- Ridom SeqSphere +
 - <http://www.ridom.de/seqsphere/index.shtml>
 - Software design to analyse NGS data by using MLST/cgMLST

Blogs and Twitter

A lot of useful information in the rapidly evolving field of bioinformatics can be gained by following bioinformaticians on twitter or reading their blogs.

- Blogs
 - Bits and bugs <https://bitsandbugs.org/>
 - Loman Labs <http://lab.loman.net/page3/>
 - Opinionomics <http://www.opiniomics.org/>
 - The genome factory <http://thegenomefactory.blogspot.co.uk/>
 - Simpson Lab Blog <http://simpsonlab.github.io/2016/08/23/R9/>
 - Jonathon Eisen's Lab <https://phylogenomics.wordpress.com/>
 - Living in an Ivory Basement <http://ivory.idyll.org/blog/>
 - Holt Lab <https://holtlab.net/>
 - Heng Li's blog <https://lh3.github.io/>
 - The Darling lab <http://darlinglab.org/blog/>
 - The Quinlan Lab <http://quinlanlab.org/>

- Help pages
 - <https://www.biostars.org>
 - <http://stackoverflow.com/>
- Bioinformaticians to follow on Twitter
 - @pathogenomenick @BioMickWatson @flashton2003 @WvSchaik @mattloose
@torstenseemann @tomrconnor @MikeyJ @jaredtsimpson @aphillipy @BillHanage
@happy_khan @daanensen @jennifergardy @genomiss @Becctococcus @phylogenomics
@ctitusbrown @DrKatHolt @ZaminIqbal @TimDallman @bioinformant @LaurenCowley4
@gkapatai @keithajolley @froggleston @lexnederbragt @jacarrico @biocomputerist
@mjpallen @Bio_mscook @bawee @lh3lh3 @andrewjpage @aaronquinlan @koadman
@Maxi_Zu

--- --- ---

Getting started: how to run Bio-Linux as a VM

Here is a brief guide on how to set up a Virtual Machine on your PC to simulate a Linux environment with several bioinformatics tools.

Downloading VirtualBox

VirtualBox is a free and powerful cross-platform VM manager found at <https://www.virtualbox.org/>.

1. Ensure you have at least 40GB of free disk space.
2. Download and install the appropriate version of VirtualBox using the link above.
3. Follow the installation instructions.
4. Wait before starting any new VM.

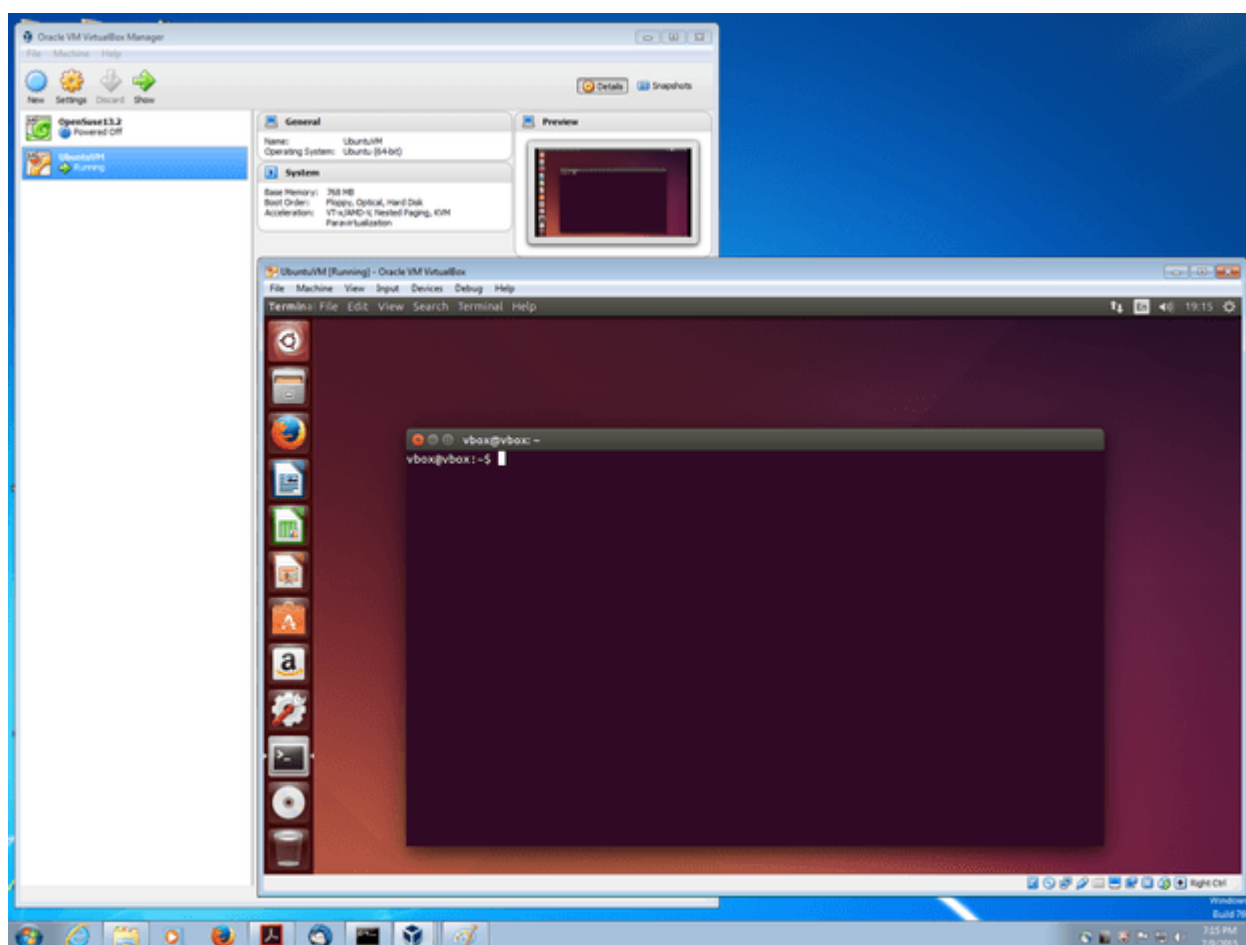


Figure 1 VirtualBox 5.0 for Windows. Within VirtualBox Ubuntu 14.04 is running.

For further info on how to setup a VM on/with whichever OS you like, please refer to the manual (also enclosed to this email).

Downloadin Bio-Linux 8 as an OVA file

In order to minimize the number of tools we need to manually set up for our training, we choose to work with Bio-Linux 8, a free bioinformatics workstation platform that can be installed on anything from a laptop to a large server, or run as a virtual machine. Bio-Linux 8 adds more than 250

bioinformatics packages to an Ubuntu Linux 14.04 LTS base, providing around 50 graphical applications and several hundred command line tools. You can find more information on it [here](#)².

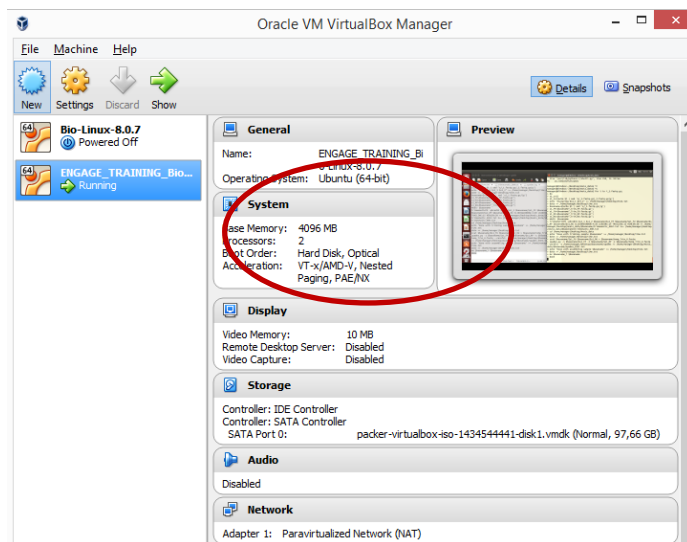
△ Bio-Linux is a 64-bit operating system. Virtually all modern PC processors support 64-bits, even if you have 32-bit Windows installed. As a rule of thumb, if you have more than 1 processor core you will have 64-bit support. See: <https://www.virtualbox.org/manual/ch03.html#intro-64bitguests>. For our purposes, you should download the Bio-Linux 8 OVA file from <http://nebc.nerc.ac.uk/downloads/bio-linux-8-latest.ova>. The OVA file is designed for use with VirtualBox but should also work with similar systems like VMWare and Parallels.

Setting up your VM instance

To setup Bio-Linux 8 for VirtualBox:

1. Start VirtualBox
2. Select Import Appliance from the File menu and import the .ova file (don't worry that it says you need an OVF file) [NOTE: this step may take several minutes to perform...]
 - a. When importing the appliance, select the option to reinitialize the MAC addresses of network cards.
3. Start the VM
4. If you see a log-in screen, log in as user **manager** with password **manager**.

Once this is working, you can delete the .ova file to save space. See the VirtualBox docs for more details including how to share folders (also detailed in the next paragraph) and hardware. You will also want to adjust hardware settings such as CPU, RAM and video acceleration settings to suit your hardware, by tuning the parameters of the "System" tab of your VM (when it's not running).



For example, on a Windows 8.1 machine with

- Intel i5-5200U CPU @ 2.20GHz 2.20GHz processor
- 8.00 GB RAM memory
- 64bit operating system, x64 processor

we suggest the following settings:

² Note, however, that this project is no more funded/developed and therefore there might be a better long-term choice to setup a Linux/Ubuntu based machine where you can install all the tools you need.

- Memory: 4096 MB
- 2 CPUs
- 10 MB video memory

or, more generally, we suggest to set both memory and CPUs values at half the value of your actual system and never below 2GB of memory.

Now you're ready to start your Biolinux VM!

For a list of all the tools included in this release of Bio-Linux, see [this page](#).

NOTE: You should treat the VM as a real machine for security purposes and apply all system security updates in a timely manner. The default manager password is, clearly, not secure. This might not be a problem because by default nobody can access the Linux VM unless they have direct access to your computer, but if you open up the network settings (eg. by adding port forwarding rules) then you must secure the account with a strong password or else take other steps to limit remote access. Ideally enforce key-only access via SSH.

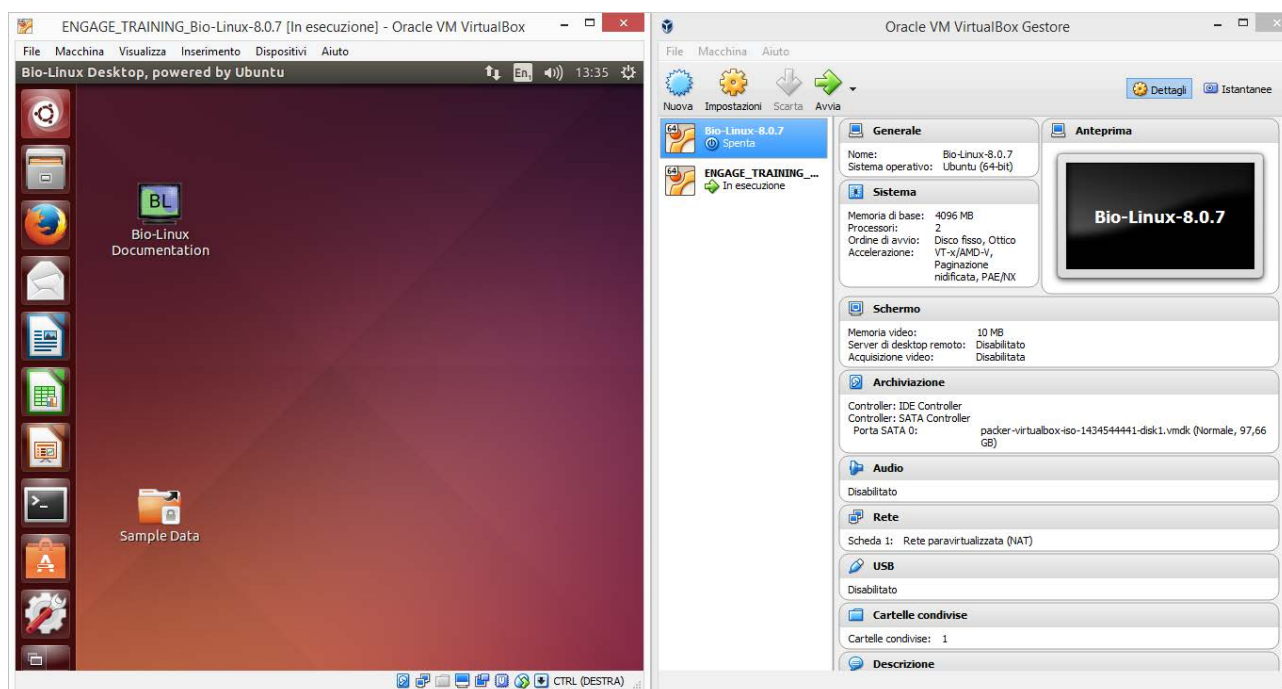


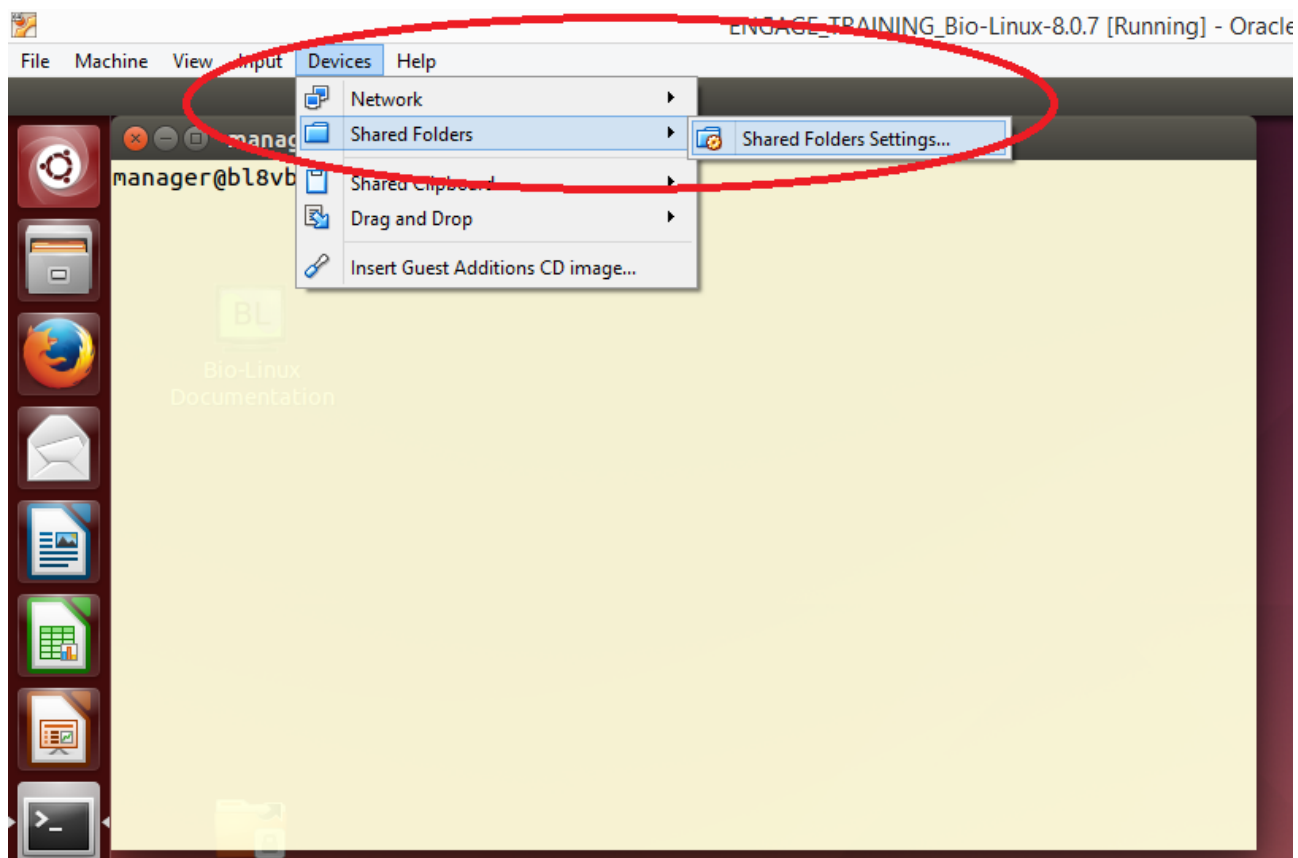
Figure 2 A screenshot of my Bio-Linux VM instance


Setting up a shared folder between your real machine and the VM

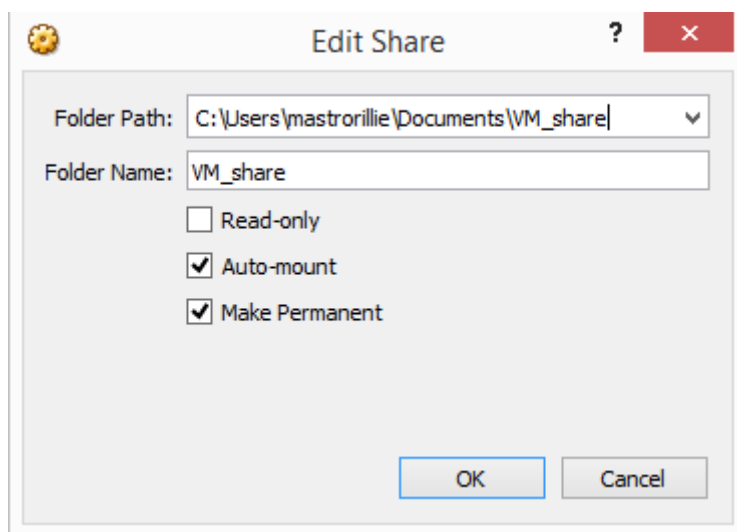
It is sometimes very useful to have the chance of sharing files between your real machine and the VM. With the "shared folders" feature of VirtualBox, you can access files of your host system from within the guest system. This is similar how you would use network shares in Windows networks – except that shared folders do not need require networking, only the Guest Additions. Shared Folders are supported with Windows (2000 or newer), Linux and Solaris guests. Shared folders must physically reside on the host and are then shared with the guest, which uses a special file system driver in the Guest Addition to talk to the host. For Windows guests, shared folders are implemented as a pseudo-network redirector; for Linux and Solaris guests, the Guest Additions provide a virtual file system. To share a host folder with a virtual machine in VirtualBox, you must specify the path of that

folder and choose for it a “share name” that the guest can use to access it. Hence, **first create the shared folder on the host** (e.g. we will refer to a folder called VM_share that I have in the Documents folder on my Win machine); then, within the guest, connect to it. In order to set an existing folder (on the host) as shared (with the VM)

- Start your VM
- Go to Devices > Shared Folders > Shared Folders Settings...



- Use  to add a shared folder
- Navigate to the folder path
- Tick the options “Auto-mount” and “Make Permanent”
- Restart your virtual machine to see the changes.



- This will link your VM_share folder between the real and virtual machine, by putting it into the /media folder on Bio-Linux. Note that all shared folders will have “sf_” as a prefix.
- Now you can move to that directory (either from command line or from file explorer GUI) and copy files from it to have them locally on the VM.

```

manager@bl8vbox: /media/sf_VM_share
manager@bl8vbox:~$ ls /media/
sf_VM_share
manager@bl8vbox:~$ cd /media/sf_VM_share/
manager@bl8vbox:/media/sf_VM_share$ cp Install_SPADES.txt /home/manager/Desktop/
manager@bl8vbox:/media/sf_VM_share$ █
  
```

Installing some tools from command line

Now all is set up to start working on your VM. If you want, you can try installing these two tools (which are not included in the Bio-Linux 8 release and that we will be using a lot during the training) directly from the command line. Please make sure you have internet connection available and open a terminal window to follow the instructions below

Trimmomatic

Trimmomatic is a flexible read trimming tool for Illumina NGS data. It is a Java-based tool, so first of all check if you have it installed on your VM by typing

```
which java
```

(default output should be /usr/bin/java). Then get trimmomatic by typing

```
sudo apt-get install trimmomatic
```

and **insert the password “manager”**. Once the installation is completed, you should be able to find it by typing

```
which TrimmomaticPE
```

To get usage information, just type

```
man TrimmomaticPE
```

on the command line.

To use Trimmomatic, you need to retrieve the ADAPTERS files (fasta format).

Run

```
#### GET THE ADAPATERS FOR TRIMMOMATIC
cd /usr/local/bioinf
```

```
sudo wget \ http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.36.zip
```

(note that the last three lines is actually one command only).

Then type

```
sudo unzip Trimmomatic-0.36.zip
```

to extract the files. A usage example for Trimmomatic would be

```
#### RUN TRIMMOMATIC ON ONE SAMPLE
```

#!/\ you should de locate one folder above the sample!

training_set

|

| __ sample1

| __ sample2

| __ sample3

| __ sample4

|

| __ ...

| __ sample9

```
TrimmomaticPE -phred33 sample1/sample1.raw.R1.fastq.gz \ sample1/sample1.raw.R2.fastq.gz
```

```
sample1/sample1.raw.process.R1.fastq.gz \ sample1/sample1.raw.orphans.R1.fastq.gz \
```

```
sample1/sample1.raw.process.R2.fastq.gz \ sample1/sample1.raw.orphans.R2.fastq.gz \
```

```
ILLUMINACLIP:/usr/local/bioinf/Trimmomatic-0.36/adapters/NexteraPE-PE.fa:2:30:10:8:true
```

```
LEADING:30 TRAILING:30 SLIDINGWINDOW:10:20 \ MINLEN:50
```

NOTE: Remember that your output files should always be in the format:

```
sample1_R1_processed
```

```
sample1_R1_orphans
```

```
sample2_R2_processed
```

```
sample2_R2_orphans
```

You can retrieve more information (all the explanation for options meaning and why/how to set them) from Anais's presentation.

Spades

SPAdes – St. Petersburg genome assembler – is an assembly toolkit containing various assembly pipelines.

To get it, open the terminal and type

```
wget http://cab.spbu.ru/files/release3.11.0/SPAdes-3.11.0-Linux.tar.gz
```

Move it to the bin folder

```
sudo cp SPAdes-3.11.0-Linux.tar.gz /usr/local/bin
```

if password is required, type "manager". Move to the selected folder and uncompress the file

```
cd /usr/local/bin
```

```
sudo tar -xzf SPAdes-3.11.0-Linux.tar.gz
```

[Optional] Create a soft link to the folder, so you don't have to change much if you install a newer version later on:

```
sudo ln -s SPAdes-3.11.0-Linux/ spades
```

Add the folder to the path by modifying the .zshrc file (if your command line interpreter is zsh) or your /etc/profile file (if your command line interpreter is bash)³

```
sudo nano ~/.zshrc
```

add the line in the header of the file (see next screenshot).

```
PATH="$PATH:/usr/local/bin/spades/bin"
```

Save (Ctrl^O+Enter) and exit (Ctrl^X+Enter).

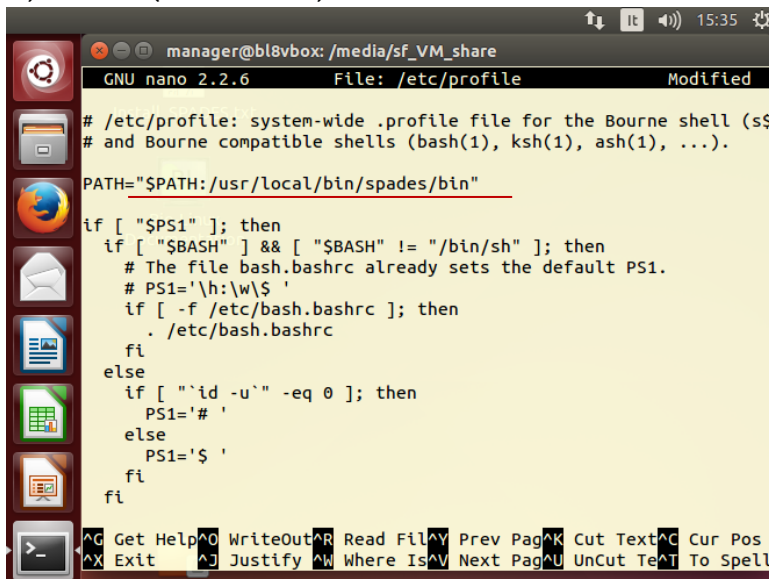


Figure 3 Screenshot of the .zshrc file. Please insert the PATH command right after the comments (lines starting with #) and ignore the rest of the file content.

In order to see the changes to the path without restarting the VM, re-type in the command line

```
export PATH="$PATH:/usr/local/bin/spades/bin"
```

For testing purposes, SPAdes comes with a toy data set (reads that align to first 1000 bp of E. coli). To try SPAdes on this data set, run from command line:

```
spades.py --test
```

If the installation is successful, you will find the following information at the end of the log:

```

===== Assembling finished. Used k-mer sizes: 21, 33, 55
* Corrected reads are in spades_test/corrected/
* Assembled contigs are in spades_test/contigs.fasta
* Assembled scaffolds are in spades_test/scaffolds.fasta
* Assembly graph is in spades_test/assembly_graph.fastg
* Assembly graph in GFA format is in spades_test/assembly_graph.gfa
* Paths in the assembly graph corresponding to the contigs are in spades_test/contigs.paths
* Paths in the assembly graph corresponding to the scaffolds are in spades_test/scaffolds.paths
===== SPAdes pipeline finished.
===== TEST PASSED CORRECTLY.
  
```

³ In order to test which interpreter you are using, write echo \$0 on the command line. If your result is zsh and you wish to change it to bash, just type "chsh -s /bin/bash" on the command line and restart the VM.

SPAdes log can be found here: spades_test/spades.log
Thank you for using SPAdes!

Quast

Quast is a quality assessment tool for measuring the quality of your genome assembly. It is particularly useful because it can generate a table comparing different metrics of your genome assemblies.

To download the tool, run:

```
wget https://downloads.sourceforge.net/project/quast/quast-4.5.tar.gz
sudo cp quast-4.5.tar.gz /usr/local/bin
cd /usr/local/bin
sudo tar -xzf quast-4.5.tar.gz
echo "PATH=\"$PATH:/usr/local/bin/quast-4.5\" >> ~/.zshrc
export PATH=\"$PATH:/usr/local/bin/quast-4.5\""
```

Let's analyze what these lines are doing:

1. get the compressed installation file from internet
2. copy the compressed file into the /usr/local/bin folder; you have to use sudo to have the administrator permissions to copy into this folder
3. change directory to /usr/local/bin
4. uncompress your compressed file
5. update your config file so to add the quast folder to your PATH variable
6. update your PATH on the fly to avoid rebooting your machine.

Now that you have quast at hand, you can use it with a list of contig files to compare their qualities.

References

<https://www.virtualbox.org/>
<http://environmentalomics.org/whats-new-in-bio-linux-8/>
<http://www.usadellab.org/cms/index.php?page=trimmomatic>
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/>
<http://cab.spbu.ru/software/spades/>
<http://quast.sourceforge.net/>
<https://www.ncbi.nlm.nih.gov/pubmed/22506599>