Deliverable

2.4 Evaluated and documented data analysis pipeline

Version: 1

Due: Month 24 and onwards

Completed: Month 24

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 643476.

Contents

Deliverable Description

The overall aim of deliverable D2.4 (*Evaluated and documented data analysis pipeline*) is to summarize data analysis pipelines (toolbox) for metagenomics and to provide a software catalogue. The context of all deliverables in work package 2, and the overall workflow as well as the scope of deliverable 2.4 is given in Figure 1.

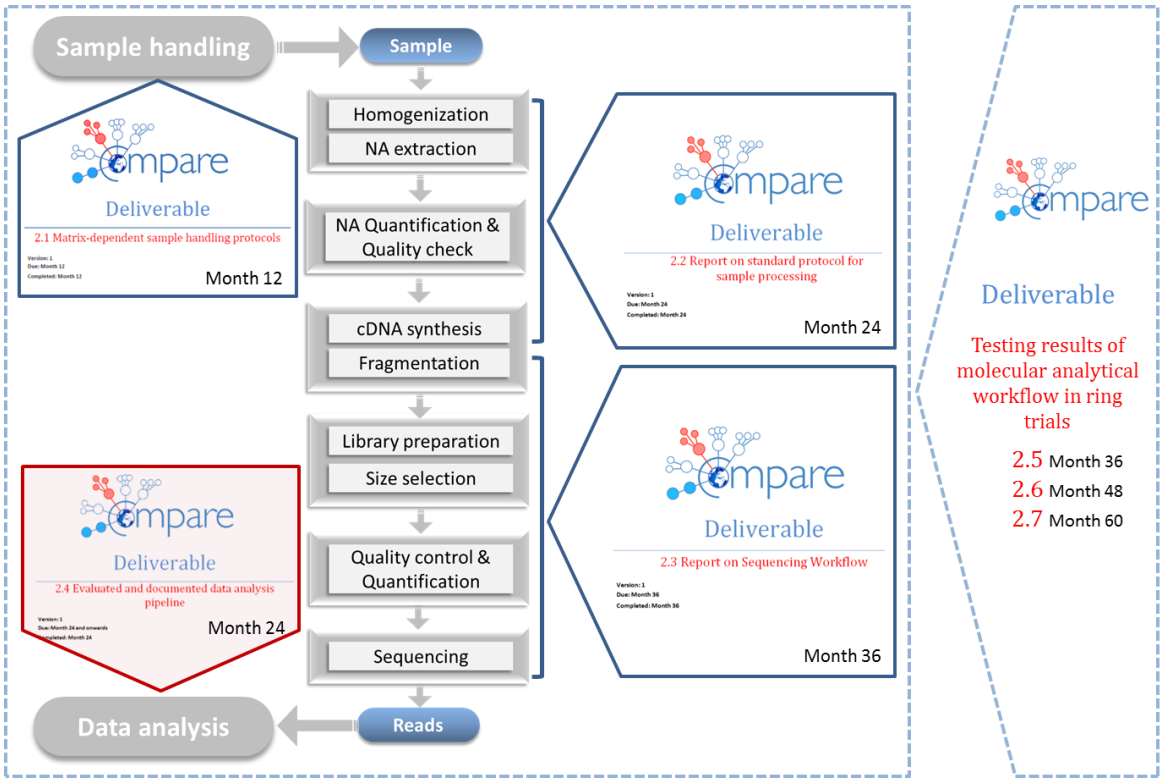

FIGURE 1: OVERVIEW OF DELIVERABLES (2.1 – 2.7) RELATED TO COMPARE WORK PACKAGE 2 WITH THEIR DUE-DATES AND SCOPES AS OUTLINED IN THE GRANT AGREEMENT. THE PRESENT DOCUMENT DEALS WITH D2.4 (MARKED RED).

Introduction

COMPARE Toolbox is a selection of useful COMPARE developed and COMPARE identified tools that will be used by COMPARE community for pathogen data analysis. These tools can be installed in any machines and be used by community while we have a Linux based virtual machine (COMPARE-VM) that we have installed the COMPARE Toolbox software. COMPARE-VM is an environment that shall be used by software engineers for development, adaptation, and tuning of workflows and tools. Some of the COMPARE Toolbox software only will be installed into COMPARE-VM by request to save space in the COMPARE-VM. A README file is available for COMPARE Bioinformatics Toolbox in GitHub and as we install new software we can update the file to always have the most up-to-date documentation.

COMPARE Toolbox minimal requirement

To add software to the toolbox, it needs to be requested by the community and also the software needs to satisfy the following requirements:

1) The software require to have a repository that contain the code
2) Dependencies needs to be up-to-date
3) Installation/deployment script/procedure that install your software, dependencies, databases, and required software.
4) Documentation (README) for installation and community usage
5) If it uses databases, a means/script/links to pull the database need to be provided. The options are to either provide a script that can build the database or you build the database, deposit it in somewhere, and provide the download link to download the already made database.
6) The output of the pipeline can be different file format but it needs to generate at least one tab delimitated file that summaries its findings (Only for relevant COMPARE developed softwares).
7) Providing Docker image and its required Docker files can be an alternative to source code in case that is a preferred option

Software list

System related software

UPDATE:

This is all security and non-security Linux (Ubuntu) updates that need to be run routinely or when specific software requires it.

CPAN:

This is a, regularly updated, installation of CPAN (The Comprehensive Perl Archive Network).

PIP:

Pip is a package management system used to install and manage software packages written in Python and it needs to be updated from time to time.

DOCKER:

This is to check to see if the docker has been installed and also adding users to the docker group, in order to be able to run the docker images on using non-root user accounts. Docker is a software container application.

Quality control applications

FASTQC:

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses, which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ files (any variant)

- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application
- A quality control tool for high throughput sequence data.

Full documentation is available in http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

TRIMMOMATIC:

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data. The selection of trimming steps and their associated parameters are supplied on the command line.

The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

It works with FASTQ (using phred + 33 or phred + 64 quality scores, depending on the Illumina pipeline used), either uncompressed or gzipp'ed FASTQ. Use of gzip format is determined based on the .gz extension.

For single-ended data, one input and one output file are specified, plus the processing steps. For paired-end data, two input files are specified, and 4 output files, 2 for the 'paired' output where both reads survived the processing, and 2 for corresponding 'unpaired' output where a read survived, but the partner read did not. Full documentation is available in
http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf

FLASH:

FLASH (Fast Length Adjustment of SHort reads) is a very fast and accurate software tool to merge paired-end reads from next-generation sequencing experiments. FLASH is designed to merge pairs of reads when the original DNA fragments are shorter than twice the length of reads. The resulting longer reads can significantly improve genome assemblies. They can also improve transcriptome assembly when FLASH is used to merge RNA-seq data.

Full documentation is available in https://ccb.jhu.edu/software/FLASH/

FASTX:

The FASTX-toolkit is a set of command line tools for processing of FASTA/FASTQ data files. It can be useful for any preprocessing of data, such as filtering by quality, base trimming etc. The FASTX-Toolkit tools

perform some of these preprocessing tasks:

- FASTQ-to-FASTA converter Convert FASTQ files to FASTA files.
- FASTQ Information Chart Quality Statistics and Nucleotide Distribution
- FASTQ/A Collapser Collapsing identical sequences in a FASTQ/A file into a single sequence (while maintaining reads counts)
- FASTQ/A Trimmer Shortening reads in a FASTQ or FASTQ files (removing barcodes or noise).
- FASTQ/A Renamer Renames the sequence identifiers in FASTQ/A file.
- FASTQ/A Clipper Removing sequencing adapters / linkers
- FASTQ/A Reverse-Complement Producing the Reverse-complement of each sequence in a FASTQ/FASTA file.
- FASTQ/A Barcode splitter Splitting a FASTQ/FASTA files containing multiple samples
- FASTA Formatter changes the width of sequences line in a FASTA file
- FASTA Nucleotide Changer Convets FASTA sequences from/to RNA/DNA
- FASTQ Quality Filter Filters sequences based on quality
- FASTQ Quality Trimmer Trims (cuts) sequences based on quality
- FASTQ Masker Masks nucleotides with 'N' (or other character) based on quality

Full documentation is available in http://hannonlab.cshl.edu/fastx_toolkit/

Genome assembly applications

VELVET:

Velvet is a de novo genomic assembler specially designed for short read sequencing technologies using de Bruijn graphs. Velvet currently takes in short read sequences, removes errors then produces high quality unique contigs. It then uses paired-end read and long read information, when available, to retrieve the repeated areas between contigs.
Full documentation is available in http://www.ebi.ac.uk/~zerbino/velvet/

VELVETOPTIMISER:

VelvetOptimiser is a multi-threaded Perl script for automatically optimising the three primary parameter options (K, -exp_cov, -cov_cutoff) for the Velvet de novo sequence assembler. Full documentation is available in http://bioinformatics.net.au/software.velvetoptimiser.shtml

IMPROVE_ASSEMBY:

This software takes in an assembly in FASTA format, reads in FASTQ format, and makes the assembly better by scaffolding and gap filling.
Full documentation is available in https://github.com/sanger-pathogens/assembly_improvement

SPADES:

SPAdes – St. Petersburg genome assembler – is intended for both standard isolates and single-cell MDA bacteria assemblies.
The current version of SPAdes works with Illumina or IonTorrent reads and is capable of providing hybrid assemblies using PacBio, Oxford Nanopore and Sanger reads. You can also provide additional contigs that will be used as long reads.

Version 3.9.0 of SPAdes supports paired-end reads, mate-pairs and unpaired reads. SPAdes can take as input several paired-end and mate-pair libraries simultaneously. Note, that SPAdes was initially designed for small genomes. It was tested on single-cell and standard bacterial and fungal data sets. SPAdes is not intended for larger genomes (e.g. mammalian size genomes). For such purposes you can use it at your own risk.

SPAdes 3.9.0 includes the following additional pipelines:

- dipSPAdes – a module for assembling highly polymorphic diploid genomes (see dipSPAdes manual).
- metaSPAdes – a pipeline for metagenomic data sets (see metaSPAdes options).
- plasmidSPAdes – a pipeline for extracting and assembling plasmids from WGS data sets (see plasmidSPAdes options).
- rnaSPAdes – a de novo transcriptome assembler from RNA-Seq data (see rnaSPAdes manual).
- truSPAdes – a module for TruSeq barcode assembly (see truSPAdes manual).

Full documentation is available in http://bioinf.spbau.ru/spades

RAY:

Ray is used for scalable de novo metagenome assembly and profiling.
Included:

- Ray de novo assembly of single genomes
- RayMéta de novo assembly of metagenomes
- RayCommunities microbe abundance + taxonomic profiling
- RayOntologies gene ontology profiling
- RaySurveyor compare genomic content between samples
- Ray -run-surveyor documentation code

Full documentation is available in http://denovoassembler.sourceforge.net

SGA:

SGA is a de novo genome assembler based on the concept of string graphs. The major goal of SGA is to be very memory efficient, which is achieved by using a compressed representation of DNA sequence reads.
Full documentation is available in https://github.com/jts/sga

ABYSS:

ABySS is a de novo, parallel, paired-end sequence assembler that is designed for short reads. The single-processor version is useful for assembling genomes up to 100 Mbases in size. The parallel version is implemented using MPI and is capable of assembling larger genomes.
Full documentation is available in https://github.com/bcgsc/abyss

SOAPDENOVO:

SOAPdenovo is a novel short-read assembly method that can build a de novo draft assembly for the human-sized genomes. The program is specially designed to assemble Illumina GA short reads. It creates new opportunities for building reference sequences and carrying out accurate analyses of unexplored genomes in a cost effective way. Now the new version is available. SOAPdenovo2, which has the advantage of a new algorithm design that reduces memory consumption in graph construction, resolves more repeat regions in contig assembly, increases coverage and length in scaffold construction, improves gap closing,

and optimizes for large genome. Full documentation is available in
http://soap.genomics.org.cn/soapdenovo.html

Alignment, sequence search, variation

BOWTIE:

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Full documentation is available in http://bowtie-bio.sourceforge.net/index.shtml

SAMTOOLS:

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.
Full documentation is available in http://samtools.sourceforge.net

BCFTOOLS:

BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF. All commands work transparently with both VCFs and BCFs, both uncompressed and BGZF-compressed.
Most commands accept VCF, bgzipped VCF and BCF with file type detected automatically even when streaming from a pipe. Indexed VCF and BCF will work in all situations. Un-indexed VCF and BCF and streams will work in most, but not all situations. In general, whenever multiple VCFs are read simultaneously, they must be indexed and therefore also compressed.
BCFtools is designed to work on a stream. It regards an input file "-" as the standard input (stdin) and outputs to the standard output (stdout). Several commands can thus be combined with Unix pipes.
Full documentation is available in https://samtools.github.io/bcftools/bcftools.html

PICARD:

Picard is a set of Java command line tools for manipulating high-throughput sequencing (HTS) data and formats. Picard is implemented using the HTSJDK Java library HTSJDK to support accessing file formats that are commonly used for high-throughput sequencing data such as SAM and VCF.
Full documentation is available in https://github.com/broadinstitute/picard

BWA:

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads. Full documentation is available in http://bio-bwa.sourceforge.net

BAMTOOLS:

BamTools provides both a programmer's API and an end-user's toolkit for handling BAM files. Full documentation is available in https://github.com/pezmaster31/bamtools

VCFTOOLS:

VCFtools is a program package designed for working with VCF files, such as those generated by the 1000 Genomes Project. The aim of VCFtools is to provide easily accessible methods for working with complex genetic variation data in the form of VCF files.
This toolset can be used to perform the following operations on VCF files:

- Filter out specific variants
- Compare files
- Summarize variants
- Convert to different file types
- Validate and merge files
- Create intersections and subsets of variants

VCFtools consists of two parts, a Perl module and a binary executable. The Perl module is a general Perl API for manipulating VCF files, whereas the binary executable provides general analysis routines.
Full documentation is available in http://vcftools.sourceforge.net

BAMUTILS:

BamUtil is a repository that contains several programs that perform operations on SAM/BAM files. All of these programs are built into a single executable, bam.
Full documentation is available in http://genome.sph.umich.edu/wiki/BamUtil

BLAST:

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.
Full documentation is available in https://blast.ncbi.nlm.nih.gov/Blast.cgi

Bioinformatics software suite

IMETAMOS:

iMetAMOS is an automated ensemble assembly pipeline; iMetAMOS encapsulates the process of running, validating, and selecting a single assembly from multiple assemblies. iMetAMOS packages several leading open-source tools into a single binary that automates parameter selection and execution of multiple assemblers, scores the resulting assemblies based on multiple validation metrics, and annotates the assemblies for genes and contaminants. iMetAMOS is available as a workflow within the metAMOS package starting with version 1.5.
Full documentation is available in https://www.cbcb.umd.edu/software/imetamos-0

EMBOSS:

EMBOSS hasn't yet been included into the COMPARE Toolbox but will be included shortly. EMBOSS is "The European Molecular Biology Open Software Suite". EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology (e.g. EMBnet) user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web. Also, as extensive libraries are provided with the package, it is a platform to allow other scientists to develop and release software in true open source spirit. EMBOSS also integrates a range of currently available packages and tools for sequence analysis into a seamless whole. EMBOSS breaks the historical trend towards commercial software packages.
Within EMBOSS you will find around hundreds of programs (applications) covering areas such as:

- Sequence alignment,
- Rapid database searching with sequence patterns,
- Protein motif identification, including domain analysis,
- Nucleotide sequence pattern analysis---for example to identify CpG islands or repeats,
- Codon usage analysis for small genomes,
- Rapid identification of sequence patterns in large scale sequence sets,
- Presentation tools for publication, and much more.

Popular applications include:

- prophet Gapped alignment for profiles.
- infoseq Displays some simple information about sequences.
- water Smith-Waterman local alignment.
- pepstats Protein statistics.
- showfeat Show features of a sequence.
- palindrome Looks for inverted repeats in a nucleotide sequence.
- eprimer3 Picks PCR primers and hybridization oligos.
- profit Scan a sequence or database with a matrix or profile.
- extractseq Extract regions from a sequence.
- marscan Finds MAR/SAR sites in nucleic sequences.
- tfscan Scans DNA sequences for transcription factors.
- patmatmotifs Compares a protein sequence to the PROSITE motif database.
- showdb Displays information on the currently available databases.
- wossname Finds programs by keywords in their one-line documentation.
- abiview Reads ABI file and display the trace.
- tranalign Align nucleic coding regions given the aligned proteins.

Full documentation is available in http://emboss.sourceforge.net/what/#Overview

Genome functional annotation

GFF3TOEMBL:

This software converts GFF3 files from the most commonly used prokaryote annotation tool Prokka into a format that is suitable for submission to EMBL. This implements some EMBL specific conventions and is not a generic conversion tool. It is also not a validator, so you need to pass in parameters, which are acceptable to EMBL. Full documentation is available in https://github.com/sanger-pathogens/gff3toembl

PROKKA:

Prokka is a software tool for the rapid annotation of prokaryotic genomes. A typical 4 Mbp genome can be fully annotated in less than 10 minutes on a quad-core computer, and scales well to 32 core SMP systems. It produces GFF3, GBK and SQN files that are ready for editing in Sequin and ultimately submitted to Genbank/DDJB/ENA.
Full documentation is available in http://www.vicbioinformatics.com/software.prokka.shtml

Compare developed analysis pipelines

DTU_CGE

The Center for Genomic Epidemiology (CGE) Pipeline, which runs in a Docker container and performs a series of analysis on pathogen genomics data.
Full documentation is available in https://bitbucket.org/genomicepidemiology/cge-pipeline-docker