

Deliverable

3.1 Analytical workflow for clinical diagnostic application

Version: 3

Due: Month 18

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	2
NGS Workflow – From raw data to results	3
Reads quality filtering and adapter trimming.....	3
Genome assembly	3
Pathogen typing and detection of plasmids, virulence genes, and resistance genes	4
Genome annotation (optional).....	4
Resfams.....	4
Graphical User interface (GUI).....	4
Pipeline with different starting points.....	5
Outbreak Module	5
Testing	5
Future development	5
Publication	6

Deliverable Description

The rapid development of next-generation sequencing (NGS) technology enabled us to produce a large amount of data for the study of pathogen genomes as well as for soil-, marine-, and human-associated metagenomes, giving us unprecedented insights into hidden reservoirs and novel antibiotic resistance genes. As costs of sequencing are steadily decreasing and response times getting shorter, its utility as a tool for tracking MDR pathogens in real time for routine hospital epidemiology or as an early warning system for outbreak detection is steadily increasing. Currently, depending on the pathogen, the identification and characterization process may take 1 to 7 days for culture, an additional 1 to 2 days for species identification and susceptibility testing, and weeks for molecular typing. As WGS of primary isolates combines identification, molecular typing, and prediction of antimicrobial susceptibility and virulence, it can theoretically reduce this time to 1-2 days for culture and around 12 h or less for sequencing and analysis.

The overall aim of Deliverable D3.1 (Analytical workflow for clinical diagnostic application) is to develop an analytical NGS workflow to rapidly detect and characterize relevant pathogens in the appropriate clinical samples and in a timeframe, that permits clinical decision-making. The goal is to come up with an entirely automated user-friendly pipeline for the NGS workflow presented in Figure 1. To date, there are already available alternatives such as web based the CGE Bacterial Analysis Pipeline (<https://cge.cbs.dtu.dk/services/cge/>), but it requires the upload of the sequencing data and a metadata file to a server. In this way, there is a loss of time for the uploading (depending on the speed of the internet connection) and the response of the server based on its load. Moreover, it could happen that the access to the server is not possible due to technical issues, causing a delay on the pathogen characterization that could be critical in few cases.

To address these problems, University of Antwerp (UA, WP3) has developed '**BacPipe**', a bacterial whole genome sequencing pipeline that does not require an internet connection to function post-installation (except for version updates) or to upload the sequencing data to perform the analysis. Ideally, it will require minimal input from the user and can be run directly by clinical microbiologists or even by clinicians without the need for any bioinformatics knowledge.

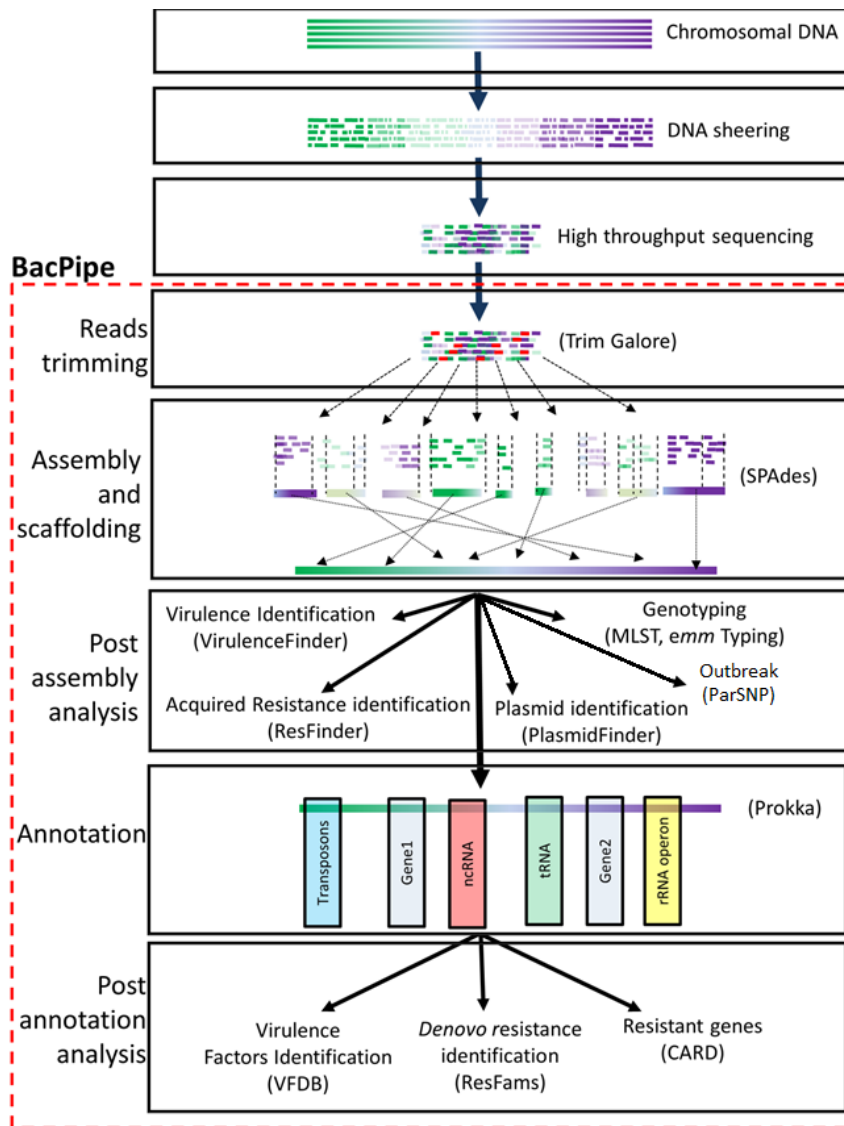


Figure 1: Automated modules in BacPipe

NGS Workflow – From raw data to results

Reads quality filtering and adapter trimming

The first step is filtering out the low-quality reads and removing the Illumina adapter sequences from the sequencing data. This is done by using Trim Galore.

(http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), that is capable of automatically detecting which adapters were used.

Genome assembly

Reads filtered by Trim Galore are used for *de-novo* genome assembly with SPAdes v 3.10.0 (<http://bioinf.spbau.ru/spades>), and the resulting contigs in FASTA format are used for further analysis.

Pathogen typing and detection of plasmids, virulence genes, and resistance genes

Using the assembled genome as input, pathogen typing is performed using the tool MLST from the Center for Genomic Epidemiology (<https://bitbucket.org/genomicepidemiology/mlst>). Other tools from CGE are then used for identification of plasmids (PlasmidFinder, <https://bitbucket.org/genomicepidemiology/plasmidfinder>), virulence genes (VirulenceFinder, <https://bitbucket.org/genomicepidemiology/virulencefinder>) and acquired resistance genes (ResFinder, <https://bitbucket.org/genomicepidemiology/resfinder>).

Genome annotation (optional)

Genome annotation is done with Prokka (<https://github.com/tseemann/prokka>), a tool to annotate bacterial, archaeal and viral genomes and produce standards-compliant output files (GenBank, fasta) this annotation tool is one of the most comprehensive annotation tools as existing ones.

Resfams

As most of the antibiotic resistance genes were originated from environmental (soil) or soil act as a reservoir for most of the resistance genes. So, we thought including this database as optional will help in detecting novel resistance genes. Also, resfams are build data from Beta-lactamase, soil metagenome, and CARD also it's well curated. So, that it will cover all the genes missed by resfinder since it is restricted to acquired resistance genes.

Graphical User interface (GUI)

The pipeline comes with GUI which helps people with no knowledge of bioinformatics also can easily utilize this tool on their personal computer without any assistance. Like this program doesn't require any powerful hardware. As they also it doesn't need an internet connection so it can be used at anywhere and anytime. Moreover, this tool is compatible is for Mac and Linux.

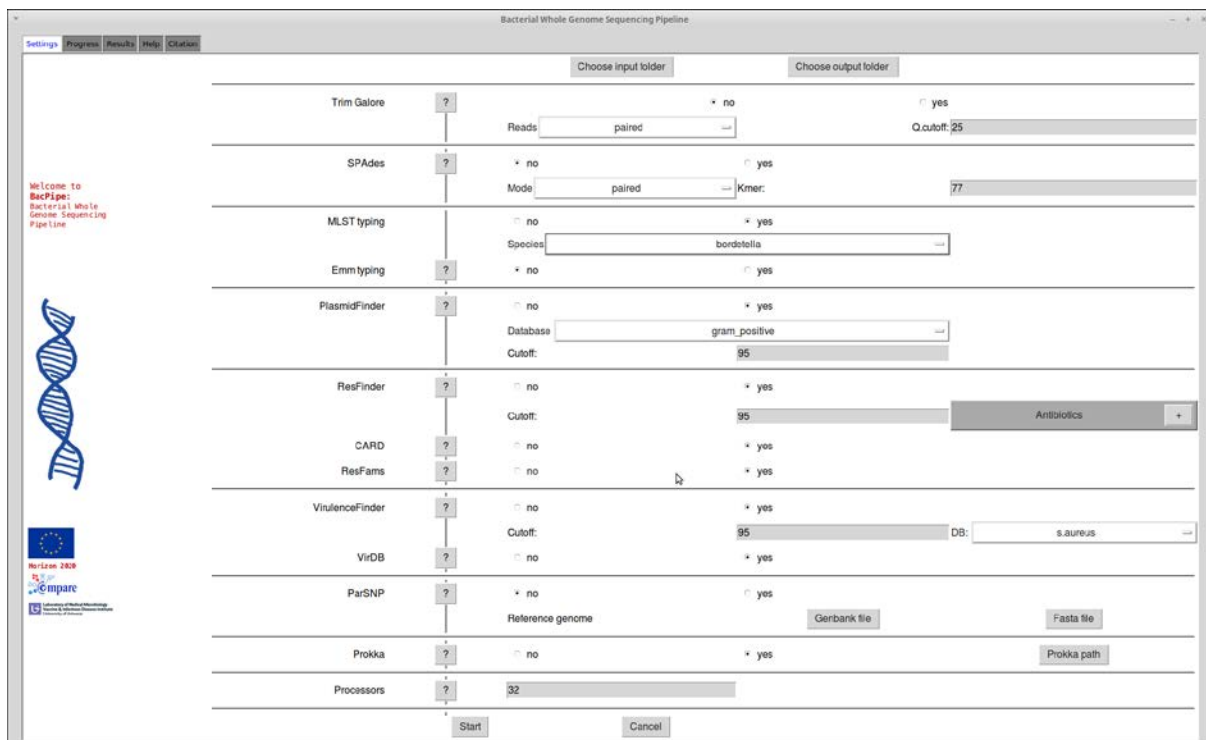


Figure 2: Screen shot of Graphical user interface (GUI) window of BacPipe.

Pipeline with different starting points

This pipeline can be used to particular tool only with deselecting the options or results which you don't want in that way results can be generated in a few minutes. You will have a report for the respective tool.

Outbreak Module

A quick SNP-based phylogeny implemented in the pipeline to generate core genome based SNPs using ParSNP and for the region for recombination sensitive area using 'Phipack' (<http://harvest.readthedocs.io/en/latest/content/parsnp.html>).

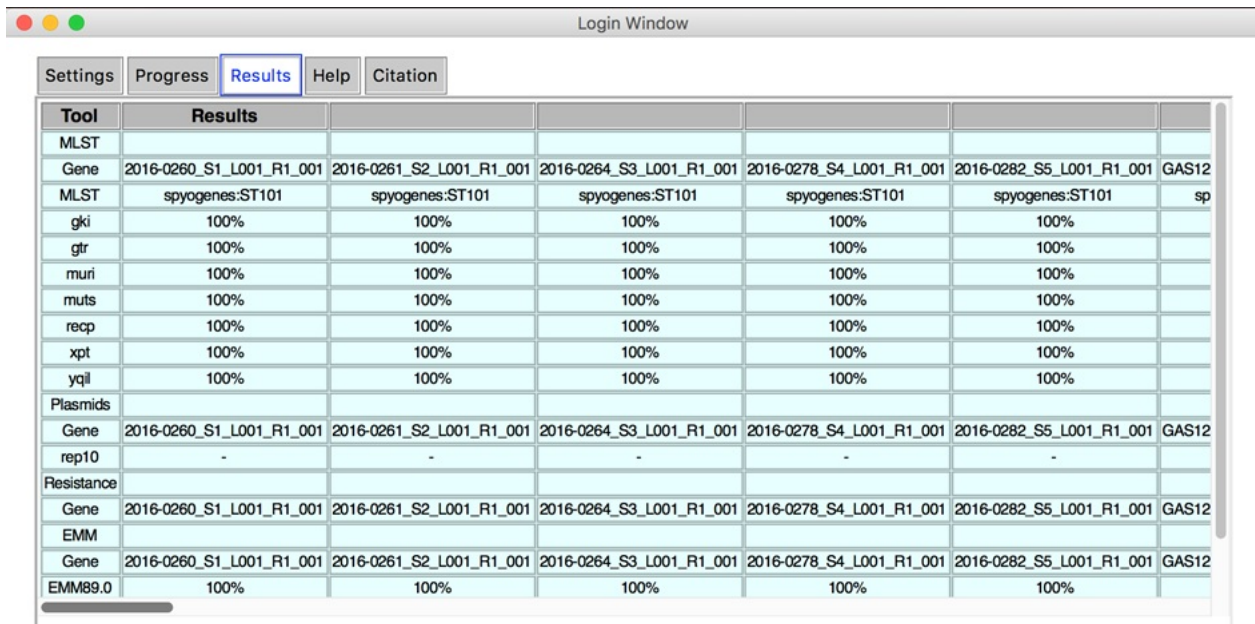
Testing

UA has tested the BacPipe pipeline with multiple pathogens (*E. coli* 32 minutes, *E. faecium* 20 minutes, *S. aureus* 20 mins) and with strains from the COMPARE GMI Proficiency Test 2016. The reads used as input were paired-end reads and the time necessary to get the results varied from 34 minutes (*Campylobacter jejuni* data from GMI PT 2016) to 90 minutes (*Klebsiella pneumoniae* data from GMI PT 2016) but it should be feasible to reduce this time further. All these strains have ~ 100-fold coverage raw data.

Future development

As we reported in the deliverable description, the goal is to come up with an entirely automated user-friendly tool that can be easily utilized in clinical diagnostics by people (clinicians, microbiologists) with no knowledge of bioinformatics. To achieve the objective, we developed a user-friendly graphical interface that permits the user to quickly run the pipeline without intervention on the multiple parameters that are required to run all the tools.

The results will be generated in an individual excel files in the separate folders. Also, a comprehensive overview will appear on the end of the analysis (Figure 3).



Tool	Results					
MLST						
Gene	2016-0260_S1_L001_R1_001	2016-0261_S2_L001_R1_001	2016-0264_S3_L001_R1_001	2016-0278_S4_L001_R1_001	2016-0282_S5_L001_R1_001	GAS12
MLST	spyogenes:ST101	spyogenes:ST101	spyogenes:ST101	spyogenes:ST101	spyogenes:ST101	sp
gki	100%	100%	100%	100%	100%	
gtr	100%	100%	100%	100%	100%	
muri	100%	100%	100%	100%	100%	
mutS	100%	100%	100%	100%	100%	
recP	100%	100%	100%	100%	100%	
xpt	100%	100%	100%	100%	100%	
yqiL	100%	100%	100%	100%	100%	
Plasmids						
Gene	2016-0260_S1_L001_R1_001	2016-0261_S2_L001_R1_001	2016-0264_S3_L001_R1_001	2016-0278_S4_L001_R1_001	2016-0282_S5_L001_R1_001	GAS12
rep10	-	-	-	-	-	
Resistance						
Gene	2016-0260_S1_L001_R1_001	2016-0261_S2_L001_R1_001	2016-0264_S3_L001_R1_001	2016-0278_S4_L001_R1_001	2016-0282_S5_L001_R1_001	GAS12
EMM						
Gene	2016-0260_S1_L001_R1_001	2016-0261_S2_L001_R1_001	2016-0264_S3_L001_R1_001	2016-0278_S4_L001_R1_001	2016-0282_S5_L001_R1_001	GAS12
EMM89.0	100%	100%	100%	100%	100%	

Figure 3: Screenshot of FINAL results output from 'BacPipe.'



COllaborative Management Platform for detection and Analyses
of (Re-) emerging and foodborne outbreaks in Europe

The next version of BacPipe will have additional tools for the analysis. To identify prophages, ISfinder, CRISPR, and wgMLST.

- It will be expanded to other organisms (viruses)
- The test will be done to run in metagenomic samples.
- It will come to compatible to Microsoft Windows without additional tools to run.

The 'BacPipe' is already being tested by work package members and it will be available in COMPARE share site and github <https://github.com/basilbritto/BacPipe>

Publication

Xavier BB, Mysara M, Bolzan M, Lammens C, Kumar-Singh S, Goossens H, Malhotra-Kumar S. (2017). BacPipe: A rapid, user-friendly whole genome sequencing pipeline for clinical diagnostic bacteriology and outbreak detection Euro Surveill. 2017 (manuscript preparation).