Deliverable

D3.3 Standardised protocols for the detection of clusters of healthcare associated infections

Dissemination level: PU

Version: 1.0

Due: Month 54

Completed: Month 60

Authors: Sander van Boheemen (EMC), Victoria Janes (EMC), Anne Pohlmann (FLI), Britto Xavier Basil (UA)

Contributing Partners: AMC, EMC, FLI, UA

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

Contents

Deliverable Description

Clusters of healthcare-associated infections (HAIs) can increase morbidity and healthcare costs; hence, the detection of clusters of HAIs is crucial in preventing disease transmissions and in infection control. Rapid identification and characterisation of infectious pathogens are essential to guide therapy, to predict disease outcomes or treatment failures, and to detect outbreaks. HAIs are traditionally detected based on case reports. On the contrary, next-generation sequencing can combine identification, molecular typing and (potentially) prediction of antimicrobial susceptibility and virulence. Therefore, it can theoretically reduce the time-to-result from 1-2 days – the typical turn-around-time of conventional culture methods - to around 12 hours or less. In order to make this possible, a harmony of standardised protocols is needed for the identification, characterisation, and typing of pathogens, specifically for hospital clusters and nosocomial transmission.

In this deliverable, we described protocols for metagenomics workflow and bioinformatics analytical pipeline that can be applied in clinical diagnosis for detecting of HAIs. On the wet-lab side, the deliverable includes the protocol for the characterisation of the viral metagenomes and the use of *Thermus thermophilus* DNA as an internal control for metagenomics sequencing. On the dry-lab side, the deliverable described two computational pipelines: RIEMS (Reliable Information Extraction from Metagenomic Sequence) and BacPipe for analysing sequencing data from metagenomics sequencing and whole-genome sequencing, respectively. The studies mentioned in this deliverable resulted in four publications and a pilot study.

Background

1. Protocols for characterisation of the viral metagenome using random amplification combined with deep-sequencing (EMC)

Introduction

Detection and characterisation of viruses by metagenomics take advantage of the sensitivity of next-generation sequencing while being non-specific for viruses present in a given sample. Viromes are structurally and functionally diverse and distinct to habitats in hosts and environments. Elimination of host nucleic acids while retaining the viral nucleic acids, is essential to detect viral sequences in clinical samples.

Protocols

Characterisation of the viral metagenome using random amplification combined with deep-sequencing.

Virus-containing supernatant was centrifuged for 10 min at 3,000 rpm to remove cellular debris. The supernatant was filtered through a 0.45-µm pore centrifugal filter unit (Millipore) to remove bacterial contamination. To remove free DNA and RNA from the sample, 2,5 µl Omnicleave endonuclease (Epicenter Biotechnologies) and 50 µL MgCl2 (25mM) was added and incubated at 37°C for 1 hour.

Viral RNA was extracted from the sample using a High Pure RNA isolation kit (Roche).

To obtain cDNA, 20 µl of RNA, 2 µl dNTPs and 4 µl random hexamers were mixed and incubated for 5 minutes at 65°C. Next, 8µl SSIV Buffer, 2µl DTT, 2µl RNaseOUT and 2µl SuperScript IV were added to the mixture. The reverse transcription mixture was sequentially incubated 10 minutes at 23°C, 10 minutes at 50°C, 10 minutes at 80°C, and 3 minutes at 95°C. Subsequently, 1 µl Klenow DNA polymerase (5 U) (New England BioLabs) was added, and the mixture was sequentially incubated at for 1 hour at 37°C, and 10 minutes at 75°C to obtain double-stranded cDNA.

The KAPA HyperPlus Kit (Roche) was used to prepare the sequencing library according to the manufacturer's instructions. Adaptions to the protocol consisted of dilution of the adapters ten times and subsequent double purification of the ligation reaction using AMPure beads.

Libraries were sequenced using the MiSeq sequencing platform, using the MiSeq Reagent Kit v2 300-cycles (Illumina).

2. Protocols for metagenomics (AMC)

Introduction

Metagenomic sequencing is still mainly done in research settings. Methods often differ and process control is lacking, while there are many technical and biological factors that can influence the sequencing result. Should metagenomics be applied in routine clinical microbiology diagnostics, careful process controls should be in place.

Publication

Janes VA, Van der Laan JS, Mende DR, Matamoros S, Schultsz C. *Thermus thermophilus* DNA internal control for process monitoring of diagnostic metagenomic sequencing. Manuscript under preparation.

Abstract

Thermus thermophilus DNA internal control for process monitoring of diagnostic metagenomic sequencing

Janes VA, Van der Laan JS, Mende DR, Matamoros S, Schultsz C.

Academic Medical Center, Department of Medical Microbiology and Department of Global Health

Background Metagenomic sequencing for clinical diagnostics is still mainly done in research settings. Methods often differ and process control is lacking, while there are many technical and biological factors that can influence the sequencing result. To monitor the sequencing process, exogenous DNA which functions as an internal control (IC) can be added to the sample. Here we report the use of *Thermus thermophilus* DNA to monitor the diagnostic metagenomic sequencing of urine samples.

Methods We developed a semi-quantitative IC for process monitoring of diagnostic urine sample sequencing on the Ion Torrent Proton and PGM. Ten urine samples were sequenced in the absence of IC. Aliquots of these ten samples were spiked with an IC concentration of 0.5%, 2%, or 5% of the sample total DNA concentration. We also sequenced a positive control (only IC DNA) and negative control (sample DNA without IC). We compared the relative abundance (RA) of pathogens between the aliquots from the same sample and the RA of IC in aliquots from different samples spiked with the same IC concentration. The optimal spike-in concentration was defined as the concentration giving the smallest difference in relative abundance of pathogens compared to the sample that was sequenced in the absence of IC. Further we visualised the Bray-Curtis distance of aliquots sequenced in the presence and absence of IC, and before and after *in silico* removal of IC reads with non-metric multidimensional scaling (NMDS) to assess the effect of adding an IC on the microbial composition of the sample.

Results The positive control contained 1.53% non-IC sequence reads while the negative control contained 0.09% IC reads, indicating minimal cross-contamination of (IC) DNA. The RA of the IC reads decreased when lower IC concentrations were spiked into the sample. IC added in a

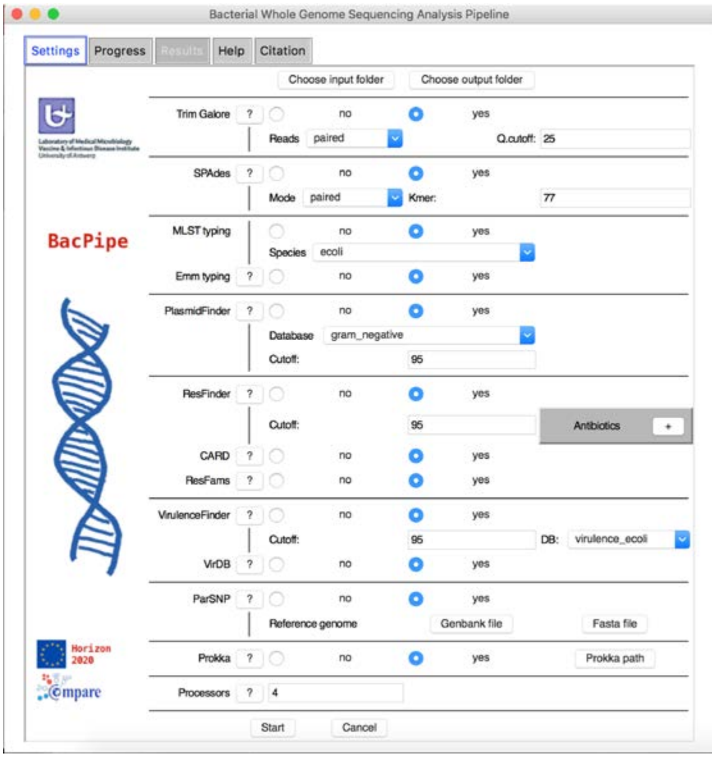
concentration of 0.5% of total DNA was detected in all samples and gave the smallest difference in RA of pathogens, relative to all bacterial reads, compared to the samples sequenced in the absence of IC. After *in silico* removal of IC reads, aliquots from the same sample formed distinct overlapping clusters in NMDS analysis, indicating the addition of IC did not introduce additional variation in bacterial composition than can be expected based on sequencing of replicates.

Conclusion *T. thermophilus* DNA IC spiked in a concentration of 0.5% of total DNA can be used for process monitoring of library preparation and sequencing of clinical urine samples. We developed a flow chart for the interpretation of IC RA in different sample types.

3. E-learning module for BacPipe (UA)

Background

Despite rapid advances in whole-genome sequencing (WGS) technologies, their integration into routine microbiological diagnostics and infection control has been hampered by the need for downstream bioinformatics analyses that require considerable expertise. We have developed a comprehensive, rapid, and computationally low-resource bioinformatics pipeline (BacPipe) for the analysis of bacterial whole genome sequences obtained from second and third-generation sequencing technologies. Users can choose to directly analyse raw sequencing reads or contigs or scaffolds in BacPipe. The pipeline is an ensemble of state-of-the-art, open-access bioinformatics tools for quality verification, genome assembly and Annotation, and identification of the bacterial genotype (MLST and *emm* typing), resistance genes, plasmid(s), virulence genes, and single nucleotide polymorphisms (SNPs). The outbreak module in BacPipe can be used, along with the SNPs and patient metadata, to simultaneously analyse many strains to understand evolutionary relationships and rapidly construct bacterial transmission routes. Importantly, BacPipe is designed to run multiple tools simultaneously which considerably reduces the time-to-result. We validated BacPipe using prior published WGS datasets from confirmed outbreaks of MRSA, carbapenem-resistant *Klebsiella pneumoniae*, and *Salmonella enterica*, and from transmission studies of *Clostridium difficile* and *Mycobacterium tuberculosis* where BacPipe helped build the same analyses and conclusions within a few hours. We believe this fully automated pipeline will contribute to overcoming one of the primary hurdles faced by microbiologists for analysing and interpreting WGS data, facilitating its direct application for routine patient care in hospitals and public health and infection control monitoring.



Trim Galore ? ○ no ● yes
Reads: paired Q.cutoff: 25
SPAdes ? ○ no ● yes
Mode: paired Kmer: 77
MLST typing ○ no ● yes
Species: ecoli
Emm typing ? ○ no ● yes
PlasmidFinder ? ○ no ● yes
Database: gram_negative
Cutoff: 95
ResFinder ? ○ no ● yes
Cutoff: 95
CARD ? ○ no ● yes
ResFams ? ○ no ● yes
VirulenceFinder ? ○ no ● yes
Cutoff: 95 DB: virulence_ecoli
VirDB ? ○ no ● yes
ParSNP ? ○ no ● yes
Reference genome
Prokka ? ○ no ● yes
Processors ? 4

Figure: Screenshot of BacPipe

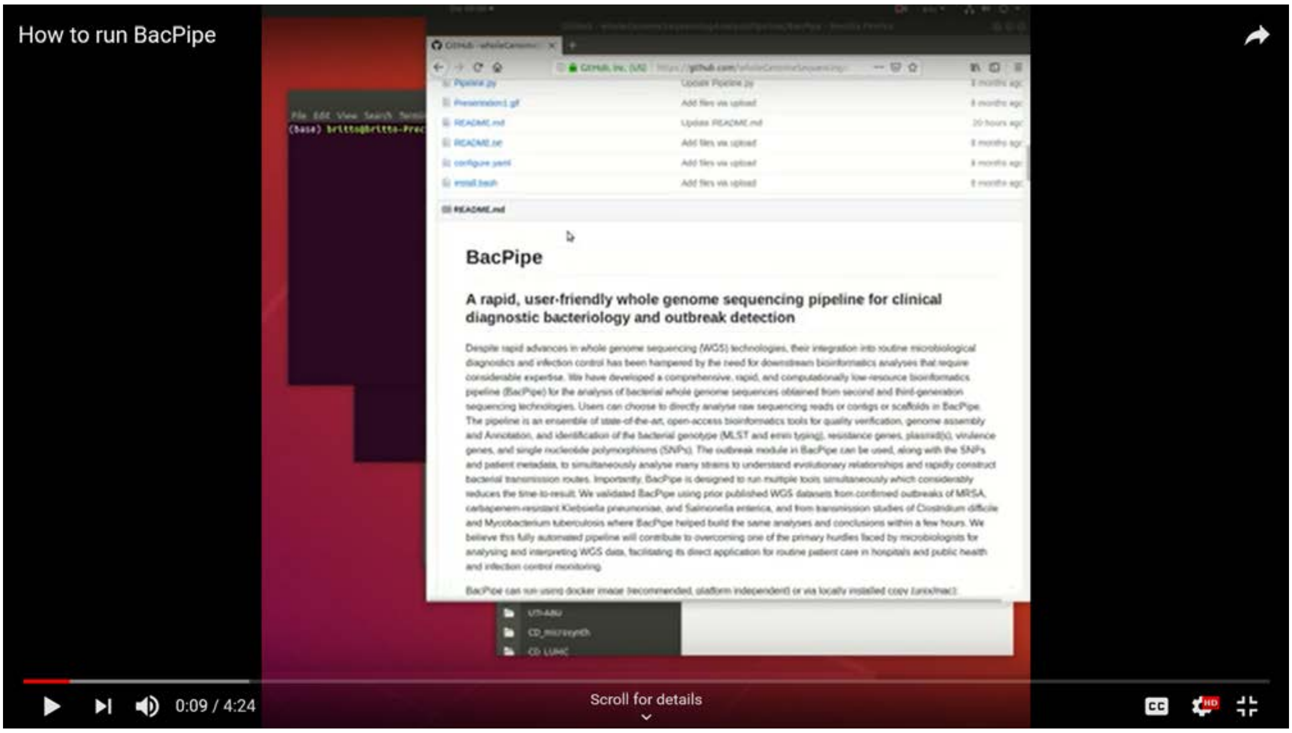
Publication

Xavier, Basil Britto and Mysara, Mohamed and Bolzan, Mattia and Ribeiro-Gonçalves, Bruno and T.F Alako, Blaise and Harrison, Peter and Lammens, Christine and Kumar-Singh, Samir and Goossens, Herman and A Carriço, João and Cochrane, Guy and Malhotra-Kumar, Surbhi, BacPipe: A Rapid, User-Friendly Whole Genome Sequencing Pipeline for Clinical Diagnostic Bacteriology (June 25, 2019). **ISCIENCE-D-19-00569 CellPress**. Available at SSRN: https://ssrn.com/abstract=3409459 or http://dx.doi.org/10.2139/ssrn.3409459

BacPipe v.1.2.6 can be downloaded from the GitHub link below:
https://github.com/wholeGenomeSequencingAnalysisPipeline/BacPipe

We also made E-learning video on how to install and run BacPipe v.1.26

https://youtu.be/kBPyknBMttY



BacPipe

A rapid, user-friendly whole genome sequencing pipeline for clinical diagnostic bacteriology and outbreak detection

Despite rapid advances in whole genome sequencing (WGS) technologies, their integration into routine microbiological diagnostics and infection control has been hampered by the need for downstream bioinformatics analyses that require considerable expertise. We have developed a comprehensive, rapid, and computationally low-resource bioinformatics pipeline (BacPipe) for the analysis of bacterial whole genome sequences obtained from second and third-generation sequencing technologies. Users can choose to directly analyse raw sequencing reads or contigs or scaffolds in BacPipe. The pipeline is an ensemble of state-of-the-art, open-access bioinformatics tools for quality verification, genome assembly and Annotation, and identification of the bacterial genotype (MLST and emm typing), resistance genes, plasmid(s), virulence genes, and single nucleotide polymorphisms (SNPs). The outbreak module in BacPipe can be used, along with the SNPs and patient metadata, to simultaneously analyse many strains to understand evolutionary relationships and rapidly construct bacterial transmission routes. Importantly, BacPipe is designed to run multiple tools simultaneously which considerably reduces the time-to-result. We validated BacPipe using prior published WGS datasets from confirmed outbreaks of MRSA, carbapenem-resistant Klebsiella pneumoniae, and Salmonella enterica, and from transmission studies of Clostridium difficile and Mycobacterium tuberculosis where BacPipe helped build the same analyses and conclusions within a few hours. We believe this fully automated pipeline will contribute to overcoming one of the primary hurdles faced by microbiologists for analysing and interpreting WGS data, facilitating its direct application for routine patient care in hospitals and public health and infection control monitoring.

BacPipe can run using docker image (recommended, platform independent) or via locally installed copy (unix/mac):

Abstract

Objectives: Despite rapid advances in whole-genome sequencing (WGS) technologies, their integration into routine microbiological diagnostics has been hampered by the need for standardised downstream bioinformatics analysis. Here we developed a comprehensive and computationally low-resource bioinformatics pipeline (BacPipe) enabling direct analyses of bacterial whole-genome sequences (raw reads or contigs) obtained from second or third-generation sequencing technologies.

Methods: Open-access tools for quality verification, de novo assembly (SPAdes), annotation (Prokka), bacterial typing (MLST, emm typing), and for identification of resistance genes (Resfams), plasmids, virulence genes, single nucleotide polymorphisms (SNPs) and core genome phylogeny were integrated into a single Python script. A graphical user interface (GUI) was developed to allow real-time progression of the analysis. The scalability and speed of BacPipe in handling large data-sets was further demonstrated using 4139 Illumina paired-end sequence files of publicly-available bacterial genomes (2.9‒5.4 Mb) from the European Nucleotide Archive (ENA).

Results: Computational time on Bacpipe, demonstrated on a 8 Gb RAM personal computer, was 21, 25, 28 and 30 minutes for sequencing coverage of 50-, 70-, 100- and 120-folds of a 5.1 Mb bacterial genome, respectively. Compiled results of every individual genome/strain are saved as an Excel file. Up to 56% reduction in analysis time was achieved by a unique parallelisation of post-assembly and post-annotation tools in Bacpipe compared to running these tools in succession. On the 4139 Illumina paired-end sequence files, running time was on average 50 minutes/strain.

Bacpipe is also integrated into EBI-SELECTA, a project-specific portal (H2020 COMPARE), and is also available as an independent docker image that can be used across Windows- and Unix-based systems.

Conclusion: BacPipe offers a fully automated 'one-stop' bacterial WGS analysis pipeline with a user-friendly GUI which can contribute to overcoming the major hurdle of WGS data analysis in hospitals and public-health and for infection control monitoring.

4. RIEMS (FLI)

Background

Diagnostic metagenomics became a powerful tool for the analysis of microbial communities. The biggest challenge is the extraction of relevant information from the huge sequence datasets generated for metagenomics studies. The Software-Pipeline RIEMS assigns every individual read sequence within a dataset taxonomically by cascading different sequence analyses with decreasing stringency of the assignments using various software applications. After completion of the analyses, the results are summarized in a clearly structured result protocol organized taxonomically.

During the COMPARE project, RIEMS was integrated in EBI-SELECTA. Several adaptions and optimizations were made to the original version. In order to be able to use the results of the workflow in a database system, it was first necessary to fundamentally restructure the corresponding output format. As a result, the software process has been adjusted accordingly without changing the basic function. These adjustments also included optimizing larger datasets and making the best use of server capacity. Especially the improvements in terms of larger data sets were imperative because NGS sequencing is a rapidly evolving method. Finally, the improved RIEMS version was integrated into server structure including the new output formats. The new output format now includes both, a database-compatible machine-readable version, and a user-friendly file that compiles the results a clear way and provides necessary overviews and tables for a comprehensive assessment of the results.

Publication

The revised RIEMS workflow was already successfully used in the COMPARE proficiency tests which is in detail described here: Brinkmann A, Andrusch A, Belka A, Wylezich C, Höper D, Pohlmann A, Nordahl Petersen T, Lucas P, Blanchard Y, Papa A, Melidou A, Oude Munnink BB, Matthijnssens J, Deboutte W, Ellis RJ, Hansmann F, Baumgärtner W, van der Vries E, Osterhaus A, Camma C, Mangone I, Lorusso A, Marcacci M, Nunes A, Pinto M, Borges V, Kroneman A, Schmitz D, Corman VM, Drosten C, Jones TC, Hendriksen RS, Aarestrup FM, Koopmans M, Beer M, Nitsche A. *Proficiency Testing of Virus Diagnostics Based on Bioinformatics Analysis of Simulated In Silico High-Throughput Sequencing Data Sets*. J Clin Microbiol. 2019 Aug;57(8). doi: 10.1128/JCM.00466-19.

The structure and validation of the pipeline was initially described in:

Scheuch M, Höper D, Beer M. RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. BMC Bioinformatics. 2015 Mar 3;16:69. doi: 10.1186/s12859-015-0503-6.

RIEMS Software-Pipeline can be downloaded here:
https://www.fli.de/fileadmin/FLI/IVD/Microarray-Diagnostics/RIEMS.tar.gz

Abstract

Background:

Fuelled by the advent and subsequent development of next generation sequencing technologies, metagenomics became a powerful tool for the analysis of microbial communities both scientifically and diagnostically. The biggest challenge is the extraction of relevant information from the huge sequence datasets generated for metagenomics studies. Although a plethora of tools are available, data analysis is still a bottleneck.

Results:

To overcome the bottleneck of data analysis, we developed an automated computational workflow called RIEMS - Reliable Information Extraction from Metagenomic Sequence datasets. RIEMS assigns every individual read sequence within a dataset taxonomically by cascading different sequence analyses with decreasing stringency of the assignments using various software applications. After completion of the analyses, the results are summarised in a clearly structured result protocol organised taxonomically. The high accuracy and performance of RIEMS analyses were proven in comparison with other tools for metagenomics data analysis using simulated sequencing read datasets.

Conclusions:

RIEMS has the potential to fill the gap that still exists with regard to data analysis for metagenomics studies. The usefulness and power of RIEMS for the analysis of genuine sequencing datasets was demonstrated with an early version of RIEMS in 2011 when it was used to detect the orthobunyavirus sequences leading to the discovery of Schmallenberg virus.