EXTERNAL SCIENTIFIC REPORT

APPROVED: 30 October 2018

doi:10.2903/sp.efsa.2018.EN-1498

# INNUENDO: A cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens

Ann-Katrin Llarena[1], Bruno Filipe Ribeiro-Gonçalves[2], Diogo Nuno Silva[2], Jani Halkilahti[3], Miguel Paulo Machado[2], Mickael Santos Da Silva[2], Anniina Jaakkonen[4], Joana Isidro[5], Crista Hämäläinen[1], Jasmin Joenperä[1], Vítor Borges[5], Luìs Viera[5], João Paulo Gomes[5], Cristina Correia[5], Janne Lunden[1], Riikka Laukkanen-Ninios[1], Maria Fredriksson-Ahomaa[1], Joseba Bikandi[6], Rosario San Millan[6], Ilargi Martinez-Ballesteros[6], Lorena Laorden[6], Mihael Mäesaar[7], Lelde Grantiņa-Ieviņa[8], Friederike Hilbert[9], Javier Garaizar[6], Mónica Oleastro[4], Mari Nevas[1], Saara Salmenlinna[3], Marjaana Hakkinen[4], João André Carriço[2] and Mirko Rossi[1]

[1]Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland; [2]Instituto de Microbiologia and Instituto de Medicina Molecular, Faculty of Medicine, University of Lisbon, Lisbon, Portugal; [3]Finnish National Institute for Health and Welfare, Helsinki, Finland; [4]Finnish Food Safety Authority, Evira, Helsinki, Finland; [5]National Institute of Health, Lisboa, Portugal; [6]Department of Immunology, Microbiology and Parasitology, Faculty of Pharmacy, University of the Basque Country, Vitoria-Gasteiz, Spain; [7]Veterinary and Food Laboratory, Tartu, Estonia; [8]Latvian Institute of Food Safety, Animal Health and Environment (BIOR), Riga, Latvia; [9]Institute of Meat Hygiene, Meat Technology and Food Science, University of Veterinary Medicine, Vienna, Austria

## Abstract

In response to the EFSA call "New approaches in identifying and characterizing microbial and chemical hazards", the project INNUENDO (https://sites.google.com/site/theinnuendoproject/) aimed to design an analytical platform and standard procedures for the use of whole-genome sequencing in surveillance and outbreak investigation of food-borne pathogens. The project firstly attempted to identify existing flaws and needs, and then to provide applicable cross-sectorial solutions. The project focused in developing a platform for small countries with limited economical and personnel resources. To achieve these goals, we applied a user-centered design strategy involving the end-users, such as microbiologists in public health and veterinary authorities, in every step of the design, development and implementation phases. As a result, we delivered the INNUENDO Platform V1.0 (https://innuendo.readthedocs.io/en/latest/), a stand-alone, portable, open-source, end-to-end system for the management, analysis, and sharing of bacterial genomic data. The platform uses Nextflow workflow manager to assemble analytical software modules in species-specific protocols that can be run using a user-friendly interface. The reproducibility of the process is ensured by using Docker containers and throught the annotation of the whole process using an ontology. Several modules, available at https://github.com/TheInnuendoProject, have been developed including: genome assembly and species confirmation; fast genome clustering; *in silico* typing; standardized species-specific phylogenetic frameworks for *Campylobacter jejuni, Yersinia enterocolitica, Salmonella enterica* and *Escherichia coli* based on an innovative gene-by-gene methodology; quality control measures from raw reads to allele calling; reporting system; a built-in communication protocols and a strain classification system enabling smooth communication during outbreak investigation. As proof-of-concepts, the proposed solutions have been thoroughly tested in simulated outbreak conditions by several public health and veterinary agencies across Europe. The results have been widely disseminated through several channels (web-sites, scientific publications, organization of workshops). The INNUENDO Platform V1.0 is effectively one of the models for the usage of open-source software in genomic epidemiology.

**Key words:** bioinformatics, whole genome sequencing, public health, outbreak investigation, communication, food safety

**Question number:** EFSA-Q-2015-00733

**Correspondence:** biocontam@efsa.europa.eu

**Disclaimer:** The present document has been produced and adopted by the bodies identified above as authors. In accordance with Article 36 of Regulation (EC) No 178/2002, this task has been carried out exclusively by the authors in the context of a grant agreement between the European Food Safety Authority and the authors. The present document is published complying with the transparency principle to which the Authority is subject. It cannot be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

**Acknowledgements:** We would like to thank the members of the advisory board: Professor Tom Humphrey (University of Swansea, the UK) for the insightful comments during proposal submission and the first half of the project, and Dr. Eduardo N. Taboada (Public Health Agency of Canada, Canada) for the encouragement and his guidance during the second half of the project. Without their precious support this project would not have been possible. A very special gratitude goes out to Professor Mario Ramirez (Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina Universidade de Lisboa, Lisboa, Portugal) for the guidance and encouragement throughout the project. We also thank Catarina Silva (National Institute of Health, Lisboa, Portugal) for her support on WGS wet-lab activities. We would like to express our gratitude to Wolfgang Hermann (Austrian Agency for Health and Food Safety Limited, Austria) for providing the project with isolates and accompanying information, and to Dr. Giuliano Garofolo, Dr. Cesare Cammà and Dr. Elisabetta di Giannatale (Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "G. Caporale", Teramo, Italy) for their cooperation. We would like to thank Professor Rene S. Hendriksen (Technical University of Denmark, Copenhagen, Denmark) and the members of the ENGAGE consortium (http://www.engage-europe.eu/) for the collaboration. We thank all the participants of the international workshop held at the University of the Basque Country Vitoria-Gasteiz, July 2017 (https://www.uik.eus/en/genomics-in-foodborne-pathogen-surveillance-and-outbreak-investigation) for their valuble comments on the very first prototype of the INNUENDO Platform. A special mention goes to all the participants of the international simulation for dedicating their time and for the insightful suggestions for improving the INNUENDO Platform: Outi Nyholm (Finnish National Institute for Health and Welfare, Helsinki, Finland), Rönnqvist Maria (Finnish Food Safety Authority, Evira, Helsinki, Finland), Taavi Riit (Veterinary and Food Laboratory, Tartu, Estonia), Juris Ķibilds (Latvian Institute of Food Safety, Animal Health and Environment "BIOR", Riga, Latvia), Catia S. Pacifico (University of Veterinary Medicine, Vienna, Austria), Carlus Deneke, Maria Borowiak and Burkhard Malorny (Bundesinstitut für Risikobewertung, Berlin, Germany), Dr. Joël Mossong (Laboratoire national de santé, Luxemburg), Iolanda Mangone (European Food Safety Authority & Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "G. Caporale", Teramo, Italy), Daniel Thomas, Alessandra Papanikolau, Federica Barucci and Beatriz Guerra (European Food Safety Authority, Parma, Italy), Ivo Van Walle (European Centre for Disease Prevention and Control, Stockholm, Sweden), Valeria Michelacci and Stefano Morabito (EURL-VTEC, Istituto Superiore di Sanità, Rome, Italy). We thank Jesus Maria Santaolalla for performing the streaming and edition of the videos of the Summer Course at University of the Basque Country (Vitoria-Gasteiz, July 2017). We would like to thank all the institutions that sponsored the Summer Course at the University of the Basque Country (Vitoria-Gasteiz, July 2017): the Basque Government, ELIKA, Vitoria-Gasteiz municipality, and the EFWISG study group of ESCMID. The international simulation was made possible with the infrastructure support of INCD - Infraestrutura Nacional de Computação Distribuida (http://www.incd.pt) funded by FCT and FEDER under the project 22153-01/SAICT/2016. The project was possible thanks the support also of CSC - Tieteen tietotekniikan keskus Oy (https://www.csc.fi) for providing access to cloud computing resources for all the analysis performed during the project, for hosting the INNUENDO platform during the INNUENDO international workshops and courses, and for hosting the INNUENDO platform during the national simulation. Finally, we would like to thank the Basque Government, Spain (Eusko Jaurlaritza/Gobierno Vasco) for co-funding the project.

**Suggested citation**: Llarena A-K, Ribeiro-Gonçalves BF, Nuno Silva D, Halkilahti J, Machado MP, Da Silva MS, Jaakkonen A, Isidro J, Hämäläinen C, Joenperä J, Borges V, Viera L, Gomes JP, Correia C,

Lunden J, Laukkanen-Ninios R, Fredriksson-Ahomaa M, Bikandi J, San Millan R, Martinez-Ballesteros I, Laorden L, Mäesaar M, Grantiņa-Ieviņa L, Hilbert F, Garaizar J, Oleastro M, Nevas M, Salmenlinna S, Hakkinen M, Carriço JA and Rossi M, 2018. INNUENDO: A cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens. EFSA supporting publication 2018:EN-1498. 142 pp. doi:10.2903/sp.efsa.2018.EN-1498

**ISSN:** 2397-8325

## Summary

This report presents the results of the project "A cross-sectorial platform for the integration of genomics in surveillance of food-borne pathogens". The project acronym, INNUENDO, will be used in this report.

The project INNUENDO (https://sites.google.com/site/theinnuendoproject/) aimed to design an analytical platform and standard procedures for the use of whole-genome sequencing (WGS) in the surveillance, outbreak detection and investigation of foodborne pathogens in the context of small countries with limited resources. The objectives of the projects were: (i) to identify the functionalities, flaws and needs in data flow during outbreak investigations and routine implementation of WGS in the molecular epidemiology of foodborne pathogens; (ii) to develop bioinformatics solutions for analyzing WGS raw data, including a species-specific evolutionary framework for *Campylobacter jejuni*, *Yersinia enterocolitica, Salmonella enterica* and *Escherichia coli*, to assess the epidemiological relationship among bacterial isolates in the food chain; (iii) to design a flexible software platform adapted to distinct IT infrastructures; (iv) to develop a standard reactive framework to assess the effectiveness of using WGS in food-borne pathogen surveillance and outbreak investigation, and to evaluate the possibilities of efficient utilization of WGS based information in solving outbreaks; (v) to enhance scientific cooperation between the food, veterinary and human health sectors to use WGS in food safety and public health protection.

As result we delivered the INNUENDO Platform V1.0 (https://innuendo.readthedocs.io/en/latest/) and associated protocols for its use in surveillance and outbreak detection and investigation for the target food-borne bacterial species. The platform includes several components needed for the correct implementation of WGS considering the limitations observed in several EU Member States (MS) regarding bioinformatics infrastructure and expertise: built-in standardized communication protocols, a legacy dataset of bacterial genomic information with minimal metadata, a defined species-specific phylogenetic framework including strain nomenclature, bioinformatics solutions for WGS data analysis, and a quality control measures from raw data (reads) to the final allele calling step using a gene-by-gene methodology.

To define a standard communication protocol related to food-borne outbreak investigations, an important aspect was the identification of the functionalities and flaws in national data sharing from local to central national authorities and on the same level of governance. An electronic questionnaire, aimed at the local and regional level officials and designed to focus on the efficiency of the data flow within an outbreak investigation process was formulated to evaluate the outbreak investigation protocols within three MS (i.e. Finland, Estonia, Latvia) participating in our project. Several factors in the current outbreak investigation process and infrastructure promoting successful outbreak investigation have been identified. However, factors hindering the investigation process and certain discrepancies between the perceptions of the local and central authorities were detected. Particularly, the survey highlighted the need for improving communication during outbreak investigation. A proposed solution was the creation of a decentralized, shared digital system allowing secure communication between all stakeholders in an outbreak setting and limiting the dependency for informal contact during outbreak investigations, facilitating coordination by central national authorities. To this goal, we have developed within the INNUENDO Platform V1.0 an internal messaging system aimed to simplify the use of high-resolution typing and standardize communication between users during outbreak investigation, especially between different sectors.

Public databases are frequently biased for time of sampling and/or geographical origin of the bacterial strains. This can lead to an uncorrected estimation of the genomic diversity in a population. Therefore, the genome of a total of 607 strains of the four species of interest have been sequenced (279 *Campylobacter jejuni*, 80 *Yersinia enterocolitica*, 129 *Salmonella enterica* serovar Enteritidis and 119 Shiga-toxin producing *Escherichia coli*) and submitted to EMBL-EBI European Nucleotide Archive (ENA). These genomes are part of the INNUENDO Legacy Dataset together with a selection of high-quality, publicly available genomes, and other genomes made available by the consortium members.

In total, the Legacy Dataset consists in 13,783 genomes. With this action we aimed to obtain enough genetic diversity to define an efficient strain nomenclature system for food-borne pathogen surveillance and outbreak investigations.

Due to its portability, easily exchangeable nomenclature and independence from a reference strain, the gene-by-gene approach was chosen as the phylogenetic framework in the INNUENDO project. To overcome the limitation of using a single schema, which might be unable to account the needs of long-term surveillance and outbreak investigations simultaneously, we proposed a dynamic shared-genome based methodology. Starting from a single schema composed of a curated set of core and accessory loci, this approach allows users firstly to classify samples in types based on a defined set of core loci in combination with strain nomenclature and then to perform cluster analysis based on a larger set of loci (i.e. including accessory genes). The first analysis uses a static core genome schema and three different levels (L) of strain nomenclature: L1 for outbreak detection and investigation, defined empirically analysing inter-strains variability from several outbreaks and sporadic cases; L2 for longitudinal long-term surveillance, defined based on analysis of cluster stability using Neighbourhood Adjusted Wallace Coefficient (AWC); L3 defined as threshold with the highest concordance with the classical 7 gene MLST classification using AWC. In the second analysis, by selecting interactively a sub-set of closed related strains (i.e. using a Graphical User Interface - GUI), users can increase the resolution of the analysis by including data from accessory loci shared specifically by the selected samples. This analysis allows the discrimination of cases during outbreak investigation. We have implemented this new methodology in a novel version of the bioinformatic tool PHYLOVIZ Online 2.0 which is integrated within the Platform. Curated whole genome MLST schemas have been developed and validated for all the four species of interested: they have been designed *ad hoc* for *C. jejuni* and *Y. enterocolitica* or by adapting available wgMLST schemas from EnteroBase (http://enterobase.warwick.ac.uk/) for *S. enterica* and *E. coli*. More details concerning the wgMLST schema implemented in the INNUENDO Platform V1.0 are available in GithHub (https://github.com/theInnuendoProject/chewBBACA_schemas) and in Zenodo (https://zenodo.org/communities/innuendo).

The different bioinformatics components of the INNUENDO Platform V1.0 have been developed in order to be computational efficient in high-end core laptops, transparent, flexible, automatic and accreditable. The Platform is composed by several working environments, each specific for a bacterial species, and operates through the creation of working spaces defined as "Project". In each "Project" user aggregates samples based on specific needs (e.g. all samples from an outbreak investigation or specific area of surveillance) and utilizes the Nextflow workflow manager to assemble a predefined set of analytical modules in an easy-to-use species-specific reproducible protocols and workflows to be applied to all the samples in the "Project". The platform guarantees the traceability and reproducibility of all the analysis processes through the use of the NextFlow pipeline description files and the docker images for all the software tools. Furthermore, each analysis associated data is stored using a specific ontology aiming at capturing the workflow of all the processes involved in next-generation sequencing (NGS) data analysis. The use of container technology facilitates the software versioning and the distribution of the analytical modules. This architecture allows flexibility to accommodate the different characteristics of the food-borne pathogens under investigation. Each protocol runs individually and is associated with the selected samples and user submitting the analysis. Currently, from the predefined modules, the user can perform two type of analyses: reference-based read mapping, used for performing *in silico* typing (i.e. rapid prediction of *Y. enterocolitica* and *E. coli* patho- and serotypes); and reference-free *de novo* assembly-based analysis, at the basis of the defined phylogenetic framework that uses a gene-by-gene approach and for rapid characterization of resistance and virulence genes. Recent versions of existing open-source validated methodologies, such as SISTR (https://lfz.corefacility.ca/sistr-app/) for *S. enterica* serotyping and ABRIcate (https://github.com/tseemann/abricate) for rapid characterization of resistance and virulence genes, have been implemented. However, to improve accuracy and efficiency, for several analytical modules novel bioinformatics solutions have been designed (available at https://github.com/theInnuendoProject/ and at https://github.com/B-UMMI). Therefore, during the course of the project the following pipelines for bacteria genome analysis have been developed: an

innovative approach for reference-based read mapping (ReMatCh), a pipeline for bacterial genome assembling and quality control of assemblies (INNUca), a novel allele calling engine for gene-by-gene analysis (chewBBACA), a new fast preliminary clustering method for bacterial genomes based on oligonucleotide frequencies (GScompare), novel methodologies for strain classification and querying databases of allelic profiles.

In order to investigate the INNUENDO Platform usability in the target user group of microbiologists and epidemiologists in the field of food safety and public health, we ran three separate usability tests during the developing phases. Through these, we measured the user's ability to complete one or more tasks using prototype versions of the INNUENDO Platform while we evaluate the efficiency, user-friendliness and satisfaction with all aspects of the platform, with special interest in sequence upload, graphics, the interface and communication protocols. The usability tests acted as proof-of-concept studies and consisted of observing how well the phylogenetic framework worked to identify clusters of possible epidemiological linked cases and how the add-on software tools were able to predict *in silico* pathotype and serotype, and to predict the presence of resistance and virulence genes.

Finally, recommendations and procedures were compiled as general guidelines for helping central national authorities to establish an effective WGS-based laboratory surveillance of food-borne pathogens using the INNUENDO Platform V1.0.

# Table of contents

Abstract ..... 1
Summary ..... 5
1. Introduction ..... 11
1.1. Background and Terms of Reference as provided by the requestor ..... 11
1.2. Interpretation of the Terms of Reference ..... 12
1.3. Overview of the project structure ..... 13
2. Identification of the flaws and needs in surveillance and outbreak responses ..... 15
2.1. Communication during outbreak investigation: functionalities and needs ..... 15
2.2. The inadequacy of available genome collections for four food-borne pathogens ..... 18
2.3. Identification of needs and current challenges in the implementation of WGS in routine surveillances ..... 19
2.3.1. Needs for quality control/optimization of wet lab procedures ..... 20
2.3.2. Need for quality control/optimization of dry lab procedures ..... 20
2.3.3. Needs for standardized bioinformatic analysis ..... 20
2.3.4. Needs for data sharing ..... 21
2.3.5. Needs for data storage and portable bioinformatic solutions ..... 22
3. The INNUENDO Platform V1.0: addressing the needs for harmonization and standardization in genome-based surveillance of food-borne pathogens ..... 22
3.1. Overview of the INNUENDO Platform V1.0 ..... 23
3.1.1. The Frontend server ..... 25
3.1.2. The Process controller/Calculation server ..... 25
3.1.3. Complete description of usage ..... 26
3.2. The legacy dataset ..... 31
3.3. Quality control measures ..... 32
3.3.1. Recommendations for WGS-related wet lab ..... 32
3.3.2. Recommendations for genome assembly: the INNUca pipeline ..... 32
3.4. Fast preliminary clustering method based on oligonucleotide frequencies (GSCompare) ..... 35
3.5. *In silico* prediction of pathotype, serotype, virulence and antibiotic resistance ..... 35
3.5.1. *In silico* typing using read-mapping ..... 35
3.5.2. Assembly based *in silico* typing ..... 36
3.6. The phylogenetic framework ..... 36
3.6.1. chewBBACA: a new suite for gene-by-gene methodology ..... 36
3.6.2. Species specific wgMLST and cgMLST schemas in the INNUENDO Platform ..... 38
3.6.3. Dynamic shared-genome based approach ..... 38
3.6.4. General guidelines for cluster analysis using the dynamic core-genome analysis ..... 42
3.7. Reporting and communication within the INNUENDO platform ..... 43
3.8. Keeping the database relevant and automatic upload to public repositories ..... 47
4. Implementation of the INNUENDO platform V 1.0 ..... 50
4.1. Different possibilities for data sharing ..... 50
4.2. Storage and computational requirements ..... 50
4.2.1. Storage ..... 51
4.2.2. Network connectivity ..... 52
5. Usability test and proof-of-concept studies ..... 53
5.1. Hands-on Workshop ..... 53
5.2. National simulation study performed between the two Finnish authorities (described in Appendix F) ..... 54
5.3. The international simulation study ..... 54
5.4. Summary of the proof-of-concept studies performed during the INNUENDO project ..... 57
5.4.1. Species determination using GScompare ..... 57
5.4.2. Pathotyping and serotyping of *E. coli* and *Y. enterocolitica* ..... 57

5.4.3. Cluster investigation using wgMLST....59
6. Recommendations for implementing WGS-based laboratory surveillance and outbreak investigation of food-borne pathogens using the INNUENDO Platform V1.0 ....61
6.1. Common recommendations ....61
6.2. A model protocol for rapid response to food-borne outbreak using the INNUENDO Platform V1.0 ....62
6.2.1. Identification of outbreak, collection of strains and WGS analysis....62
6.2.2. Communication between users ....64
7. Dissemination ....65
7.1. Publications in international peer-reviewed journals ....65
7.2. International conferences ....66
7.3. Press releases ,dissemination through the internet and citations ....66
7.4. Courses and capacity building activities ....67
7.4.1. "*Genomics in food-borne pathogen surveillance and outbreak investigation*" Workshop (Vitoria-Gasteiz, July 2017)....67
8. Contribution and feedback from consortium members ....68
8.1. Leader/coordinator: University of Helsinki (UH)....68
8.2. Partner 1: Universidade de Lisboa (UL) ....68
8.3. Partner 2: Universidad del Pais Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)....69
8.4. Partner 3: University of Veterinary Medicine, Vienna (VETMEDUNI)....70
8.5. Partner 4: National Institute for Health and Welfare (THL) ....71
8.6. Partner 5: Finnish Food Safety Authority (EVIRA)....72
8.7. Partner 6: Instituto Nacional de Saúde Dr. Ricardo Jorge (INSA)....74
8.8. Partner 7: Veterinary and Food Laboratory (VFL) ....75
8.9. Partner 8: Institute of Food Safety, Animal Health and Environment (BIOR) ....76
9. Discussion and conclusions....77
10. Additional Supporting Information ....79
References....80
Glossary ....86
Abbreviations....89
Appendix A – Whole genome sequencing activities of the INNUENDO project ....91
Appendix B – The INNUCA V3.1 modules....97
Appendix C – Fast preliminary clustering method based on oligonucleotide frequencies: GSCOMPARE....101
Appendix D – *In silico* typing using read-mapping: patho_typing and seq_typing tools .... ....111
Appendix E – The classification system within the INNUENDO Platform V1.0 ....117
Appendix F – Report of the national and international usability tests of the INNUENDO Platform125
Appendix G – "Genomics in food-borne pathogen surveillance and outbreak investigation" Workshop (Vitoria-Gasteiz, July 2017) ....139

# 1. Introduction

An effective indicator-based surveillance system of priority-listed pathogens is fundamental in combating and controlling food-borne disease. Such a system is most powerful when able to monitor the geographical location, spread, type and genomic variation of pathogens to rapidly detect the emergence of food-borne outbreaks. To efficiently do so, the system must separate epidemiologically linked cases (e.g. common source of infection) from baseline sporadic incidences, and microbial typing can support the traditional epidemiological investigations in such task (WHO, 2008).

Reduced costs of sequencing and availability of bench-top sequencers promote the implementation of whole genome sequencing (WGS) as the molecular typing technique of choice and enhance laboratory-based surveillance of food-borne diseases at local, national and international level (Llarena et al., 2017; ECDC, 2015). As a 'one-stop-shop' for rapid pathogen characterization, a full and functional implementation of WGS in public health and food safety microbiology allows a significant simplification of the analytical framework with a consequent reduction in human intervention and overall costs (WHO, 2008). Moreover, by being compatible with machine-to-machine communication and other eHealth solutions, WGS has a great potential to be interoperable across disciplines and laboratories. Therefore, WGS-based typing is replacing traditional analyses for certain microbial pathogens in several countries, revolutionizing outbreak detection and investigation, and gradually becoming a relevant tool for control-oriented surveillance (van Panhuis et al., 2014). Genomics also introduces new opportunities for more efficient use of isolate information, defined as contextual data (Griffiths et al., 2017), especially when combining high-resolution typing with the epidemiological and clinical data (such as exposures or clinical symptoms and outcomes) in real-time.

The value of WGS data extends well beyond a single laboratory or laboratory network, since WGS data is useful for answering various scientific questions with the ultimate aim of reducing the burden of disease worldwide. However, this is only possible if molecular typing data is shared, which facilitates immediate public health actions, strengthens long-term studies and has a strong impact on applied research (van Panhuis et al., 2014; Griffiths et al., 2017). However, there are country-specific legal barriers as well as several technical issues and diffuse political and ethical skepticism concerning sharing sensitive data from human or food products (van Panhuis et al., 2014).

To identify, assess and communicate current and emerging threats to human health posed by infectious diseases, the timely availability and sharing of genomic and epidemiological data will be critical in the forthcoming years for public health and food safety authorities across the European Union (EU). However, the road towards the EU-wide WGS implementation in pathogen surveillance is not free from challenges. Despite the current wide distribution of next-generation sequencing (NGS) technology, there are noticeable variations between and within EU member states (MS) regarding their capacity to translate genomic data to valuable information for its use in public health and food safety decisions (Revez et al., 2017; EFSA, 2018). In other words, although the technology is available and widely used, know-how and expertise are still underdeveloped in several EU MS, especially in less-resourced countries. In addition, the lack of a common language coupled with different points of view between the involved fields of expertise (i.e. bioinformaticians, microbiologists, epidemiologists, and practitioners) imposes a particular challenge for the efficient exploitation of WGS in public health actions.

With these critical aspects in mind, the INNUENDO project, co-funded by the European food Safety Authority (EFSA), was launched in 2016 with the aim at establishing a common framework to guarantee harmonized, quality controlled and validated bioinformatics tools and guidelines. The goal was to assure effective access to strategic application of WGS in surveillance of food-borne diseases for all stakeholders and end-users in the field, with special focus on smaller and/or less resourced stakeholders as target user. This report summarizes the achievements of the INNUENDO project. It clarifies the general strategy of the project, the scenario for WGS implementation in public health actions, and the flow of data and information during outbreak detection and investigation. It discusses the challenges in integrating contextual data into WGS analytical framework and communication during outbreak detection and investigation. It explains the rationale of the INNUENDO Platform V1.0

and each analytical component, presents the genomic legacy dataset provided within the Platform, and describes the gene-by-gene phylogenetic framework and its integration with the different software solutions. Finally, it proposes recommendations for implementing WGS-based laboratory surveillance of food-borne pathogens using the INNUENDO Platform V1.0.

The project aimed also in enhancing scientific cooperation between the food, veterinary and human health sectors and the report herein includes the feedback from each participating institutions concerning the overall experience in term of collaboration and capacity building.

## 1.1. Background and Terms of Reference as provided by the requestor

This grant was awarded by EFSA to: University of Helsinki (UH)

Beneficiary: University of Helsinki (UH), Universidade de Lisboa (UL), Universidad del Pais Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), University of Veterinary Medicine Vienna (VMU), National Institute for Health and Welfare (THL), Finnish Food Safety Authority (EVIRA), Instituto Nacional de Saúde Dr. Ricardo Jorge (INSA), Veterinary and Food Laboratory (VFL), Pārtikas drošības, dzīvnieku veselības un vides zinātniskais institūts (BIOR)

Grant title: New approaches in identifying and characterizing microbial and chemical hazards

Grant number: GP/EFSA/AFSCO/2015/01/CT2

**Main objective of the call**

The main objective of the grant agreement was to facilitate a scientific cooperation framework, the development and implementations of joint projects, and the exchange of expertise and best practises in the field of the European Food Safety Authority (EFSA) mission. In particular, the action financed by the EFSA grant to be awarded following the call for proposal GP/EFSA/AFSCO/2015/01/CT2 shall contribute to the objective of boosting scientific cooperation between scientists and research organizations with a competence in the development and validation of new approaches in the area of microbiological and chemical hazard assessment. It is of paramount importance to coordinate efforts between the food, veterinary and human health sectors in order to obtain maximum benefits from the use of WGS and read across methodologies for microbial and chemical food safety, respectively.

**Specific objective of the call**

Making use of molecular approaches to identify and characterize microbial food-borne pathogens, specifically using WGS analysis, to enhance the understanding, the traceability and the spread of the disease in human that these bacteria population may cause.

Molecular approaches to identify and characterize microbial food-borne pathogens, specifically using WGS analysis, provide a golden opportunity to (i) explore the bacterial genetic diversity within and between compartments in the food chain; (ii) to assess the epidemiological relationship of isolates from different compartments; and (iii) to identify the presence of putative markers conferring the potential to survive/multiply in the food chain and /or cause disease in humans (e.g. virulence and antimicrobial resistance). The methodology is very promising, and the technology is still evolving quickly. However, it is still unclear when and how this technology will be ready to be applied to routing activities and "proof of concept" projects for application in a public health context are needed.

There is currently limited experience in the use of WGS methods for microbial food safety in EU. The application of WGS to generate new data may provide risk assessors with a powerful tool. However, full integration of routing WGS of food-borne pathogens in food safety will only be possible after successful translational collaboration and coordination among scientists paying common pathways to overcome key challenges. The coordination of efforts between the food, veterinary and human health sectors is of paramount importance in order to obtain maximum benefits form the use of WGS for food safety and public health protection. In addition, novel means of analysing data and translating these into ´plain language´ reports that can be used for public health action need to be developed.

The project funded should concentrate on the applicability and integration of WGS methods for identification and characterization of microbial food-borne pathogens.

## 1.2. Interpretation of the Terms of Reference

Through a cross-sectorial collaboration which includes governmental organizations, authorities and research institutes from food, veterinary and human sectors, our goal was to create a series of standardized protocols and build a software platform to strengthen infectious disease surveillance for all actors within public health and food safety. Specifically, we aimed to develop user-centered species-specific analytical frameworks providing methods and harmonized nomenclatures for routine application of WGS in surveillance and outbreak responses.

The **specific objectives** are listed below.

1. To identify the functionalities, flaws and needs in data flow during outbreak investigations and routine implementation of WGS in the molecular epidemiology of food-borne pathogens (results concerning this objective are presented in Sections 2.1, 2.2 and 2.3).
2. To develop bioinformatics solutions for analyzing WGS raw data, including an evolutionary framework to assess the epidemiological relationship among bacterial isolates in the food chain (results concerning this objective are presented in Sections 3.2, 3.3, 3.4, 3.5 and 3.6, and in Appendices A, B, C, D and E).
3. To design a flexible software platform adapted to distinct IT infrastructures (results concerning this objective are presented in Sections 3.1, 3.7, 3.8, 4.1 and 4.2)
4. To develop a standard reactive framework to assess the effectiveness of using WGS in food-borne pathogen surveillance and outbreak investigation, and to evaluate the possibilities of efficient utilization of WGS based information in solving outbreaks (results concerning this objective are presented in Sections 5.1, 5.2, 5.3, 5.4, 6.1 and 6.2 and in Appendix F).
5. To enhance scientific cooperation between the food, veterinary and human health sectors to use WGS in food safety and public health protection (results concerning this objective are presented in Sections 7 and 8, and in Appendix G)

From our project we expected to secure the efficient application of WGS in food safety actions for public health and veterinary microbiologists and epidemiologists with limited IT resources. While our actions focused on four relevant food-borne pathogens (*Campylobacter jejuni*, *Yersinia enterocolitica, Salmonella enterica* serovar Enteritidis and Shiga-toxin producing *Escherichia coli* - STEC), the proposed methods and workflows are suitable for other relevant bacterial pathogens, such as *Listeria monocytogenes*.

Table 1 summarizes how the proposal responded to the objectives stated in the Call text.

**Table 1:** Answers to the Call objectives

| Objectives of the Call | How INNUENDO project responded to the Call |
|---|---|
| • ***Coordinate efforts between the food, veterinary and human health sectors in order to obtain maximum benefits from the use of WGS for microbial food safety***<br>• ***Full integration of routine whole genome sequencing of food-borne pathogens in food safety*** | Through a cross-sectorial collaboration, we have created a software platform allowing all actors who are assessing and managing risks of food-borne diseases to use standardised protocols and harmonized nomenclature for routine application of WGS in surveillance and responses to food-borne outbreak. Specifically, the project focused on the following relevant pathogens: *C. jejuni, Y. enterocolitica, S.* Enteritidis and STEC |
| • ***"Proof of concept" projects for application in a public health context are needed*** | The platform and the procedures have been tested in several proof-of-concept studies simulating real-time surveillance and fast outbreak response at national and transnational level. These actions allowed the identification of technical and legal issues that might interfere with the |

| Objectives of the Call | How INNUENDO project responded to the Call |
|---|---|
| | successful application of routine WGS in food safety and possible limitation in sharing pathogen specific genomic data and metadata. A particular point of interest was the evaluation of the efficacy of information flows at local and national level in food-borne outbreaks, with special emphasis on utilizing sequence-based information in solving outbreaks. |
| • ***WGS provides a golden opportunity for identification and characterization of microbial food-borne pathogens***<br>• ***Novel means of analysing data and translation of these in 'plain language' useful in public health actions*** | Exploring bacterial genetic diversity in order to assess epidemiological relationship of isolates was a central point which our proposal addressed. We focused in designing a nomenclature defining, specifically for each species, types and clones. This nomenclature fostered the establishment of a useful and simple common language to be used in public health and food safety actions. |

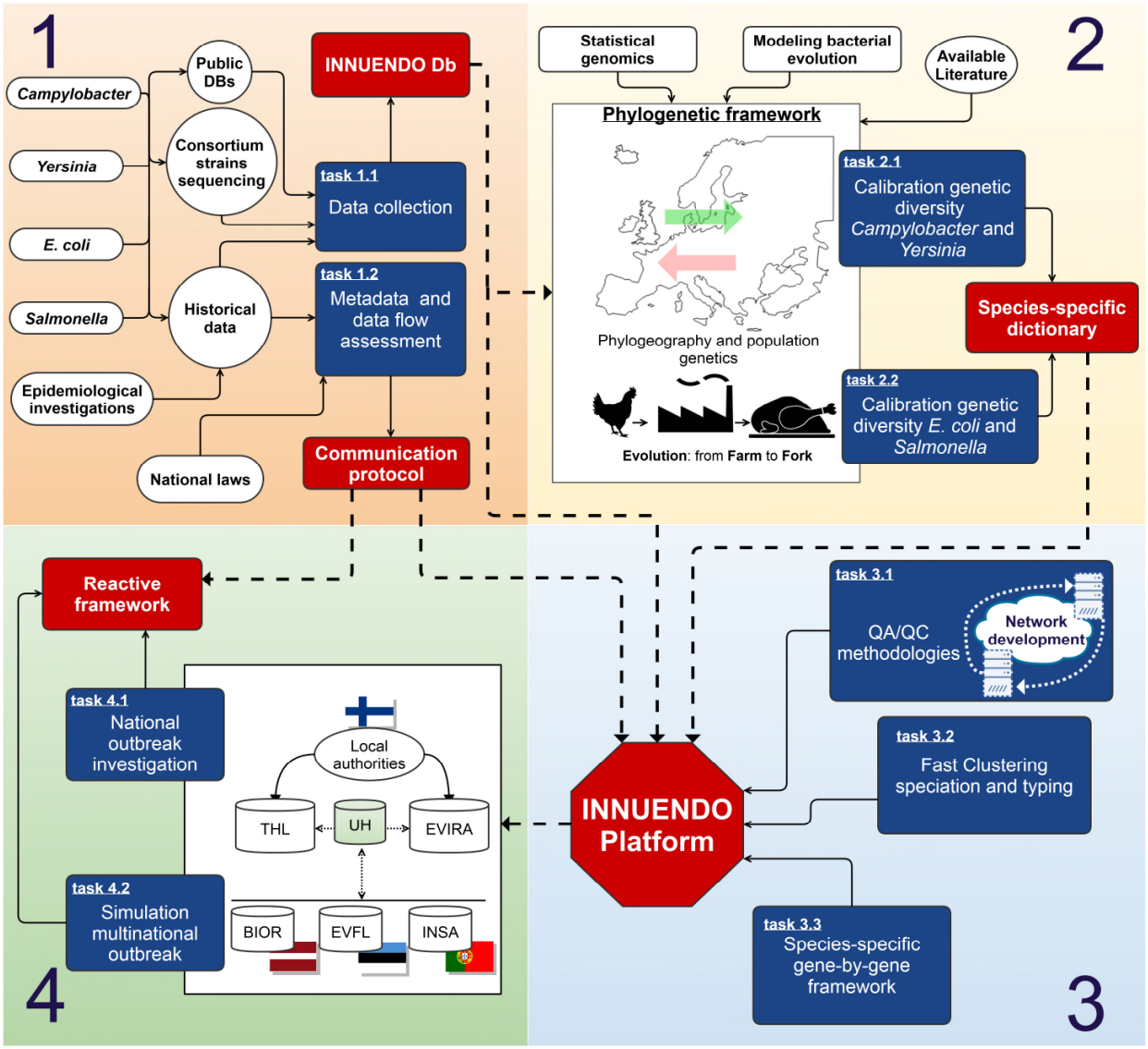
## 1.3. Overview of the project structure

The project was structured in four scientific (WP1-4) and one administrative work packages (WP5). Each WP contained two or three tasks. An overview of WPs and tasks and how they are linked to the specific objectives of the project listed above is summarised in Table 2.

**Table 2:** Overview of work packages and tasks

| Work packages (Leader) | Tasks | Task leader | Specific objectives |
|---|---|---|---|
| WP1:Data flow calibration Leader: UH | 1.1 Data collection and Sequencing | INSA | 2 |
| | 1.2 Metadata and data flow assessment | UH | 1 |
| WP2:Phylogenetic calibration Leader: UH | 2.1 *Campylobacter* and *Yersinia* | UH | 2 |
| | 2.2 *E. coli* and *Salmonella* | VMU | 2 |
| WP3:Infrastructure development Leader: UL | 3.1 QA/QC methodologies and Network development | INSA | 1, 2 & 3 |
| | 3.2 Fast Clustering speciation and typing (FCST) | UPV/EHU | 2 |
| | 3.3 Species-specific gene-by-gene framework | UL | 2 & 3 |
| WP4:Proof-of-concept actions Leader: THL | 4.1 National outbreak investigation | THL | 4 |
| | 4.2 Simulation multinational outbreak | EVIRA | 4 |
| WP5:Dissemination communication management Leader: UH | 5.1 Dissemination and training | UPV/EHU | 1-5 |
| | 5.2 Management | UH | 5 |

The interaction between the four scientific WPs of the project is illustrated in Figure 1.




Squares symbolise working packages. Blue boxes represent the tasks within each working package. Red boxes or hexagon indicate the expected outcomes for each working package. Lines denote the interaction within each work package while dotted lines show the expected interaction between work packages.

**Figure 1:** Graphical representation of the project phases

# 2. Identification of the flaws and needs in surveillance and outbreak responses

## 2.1. Communication during outbreak investigation: functionalities and needs

Rapid and accurate communication, i.e. data flow, is of utmost importance for successful outbreak investigation. Data collected from outbreak reports in the EU and USA show large variations in operational routines during outbreak investigations and reporting, use of epidemiological analysis and laboratory services (Gossner et al., 2015; CSPI, 2011). This might result in incongruent information and inefficient and unequal data collection on food-borne pathogens, which in turn compromise data sharing and ultimately public health actions (Murphree et al., 2012; Jones et al., 2004; Jones et al., 2013). As WGS is being implemented as the subtyping tool of choice in public health services, sufficient and well-defined contextual data describing the WGS in an epidemiologically, clinically and technically way (Griffiths et al., 2017) are very important for WGS to be a valuable tool in genomic epidemiology. Barriers of legal, political and technical nature challenge the sharing of data and integration between agencies, especially during multi-jurisdiction outbreaks (van Panhuis et al., 2014).

Therefore, harmonization of outbreak procedures on all levels and in different member states is needed to secure detection of outbreaks and success of outbreak investigations. Well established guidance documents on outbreak investigations are available from World Health Organization (WHO) (WHO, 2008) and a toolkit for investigation and response to food and water-borne outbreaks are accessible at the European Centre for Disease Prevention and Control (ECDC) website, but it is important to identify inconsistencies and deviations from these guidance documents between and within MS, with special focus on practical outbreak procedures and data sharing. Only after identification of such inconsistencies can harmonization between MS be achieved.

Two small (Latvia and Estonia) and on middle sized (Finland) EU MSs were available as model countries to assess communication flow during food-borne outbreak investigations. The two Baltic countries, Latvia and Estonia, with a population of 1,950,116 and 1,315,636, respectively, have low populations compared to the remainder of the EU MSs, while the Nordic country Finland (population 5,503,297) has an average population in this regard (http://ec.europa.eu/eurostat/). The governmental bodies of food hygiene, veterinary and public health at local, regional and national level in the three countries is presented in Table 3.

We evaluated legislation and guidelines on food-borne outbreak investigations and actual outbreak reports as background, and developed and conducted an electronic questionnaire (Annex A), targeted at local officials participating in the investigation of food-borne outbreak. Additionally, representatives of the central level authorities were interviewed for their perceptions on outbreak investigation procedures.

We found several factors in the current outbreak investigation process and infrastructure promoting successful outbreak investigation. In the majority of cases, the information flow between the local authorities as well as between the local and central level was considered as good. However, we also detected factors hindering the onset and slowing down the efficient progress of investigation process. There were also certain discrepancies between the perceptions of the local and central authorities, e.g. concerning the availability of consultation or the existene of detailed instuctions for standard operating procedures in outbreak cases.

**Table 3:** Overview of the governmental bodies in the three sample countries Estonia, Latvia and Finland with the laboratories these health services answer to

| Governmental body | Latvia[(a)] | Estonia[(a)] | Finland[(a)] |
|---|---|---|---|
| **Public health agencies** | The Centre for Disease Prevention and Control *Slimību profilakses un kontroles centrs* | The Health Board *Terviseamet* | National Institute for Health and Welfare *Terveyden ja hyvinvoinnin laitos /Institutet för hälsa och välfärd* (THL) |
| **Food and veterinary agencies** | The Food and Veterinary Service *Partikas un veterinarais dienests* | The Veterinary and Food Board *Veterinaar- ja toiduamet* | Finnish Food Safety Authority *Elintarviketurvallisuusvirasto /Livsmedelssäkerhetsverket* (EVIRA) |
| **Laboratories** | Institute of Food Safety, Animal Health and Environment "BIOR" *Pārtikas drošības, dzīvnieku veselības un vides zinātniskais institūts "BIOR"* | Three laboratories under the Health Board and several under the Veterinary and Food Board | Accredited laboratories by regional governments (AVI); National laboratories located at the National Institute for Health and Welfare and at the Finnish Food Safety Authority |

(a): Local names given in italics.

According to central governments, factors that were promoting efficient communication were (but not limited to): 1) the presence of well organized governmental structures for public health and food safety area, 2) acknowledged and detailed guidance documents for the execution of an outbreak investigation, 3) good communication and trust within and between governmental bodies centrally, and 4) good knowledge and know-how on outbreak investigation and motivated employees. The presence of a multidisciplinary outbreak control team (OCT) (WHO, 2008) in a jurisdiction varied between countries, but an OCT equipped with a predetermined chair was found to decrease the amount and severity of communication problems. For instance, when the chair was determined on a case-to-case basis, more problems related to outdated or faulty contact information and inadequate human resources were experienced. It can therefore be expected that a prepared OCT with a stable chair enhance the probability of solving a food-borne outbreak. Moreover, a prepared OCT is more likely to sample, communicate and execute studies more correctly as their experience collectively increase.

Of special importance for application of WGS-based methods in outbreak investigation were the findings that the number of samples collected is usually low. The lack of sufficient food, patient and environmental samples has been found to reduce the likelihood of solving outbreaks (Murphree et al., 2012; Jones et al., 2004). Our findings, based on questionnaire and interview responses, showed that late and infrequent notification of outbreaks by health centres and medical doctors of the local public health authorities complicated the detection and identification of cases and hampered the unraveling of the food consumption histories, reducing the chance of obtaining food samples and perform a food trace-back. Moreover, it was clear from the responses obtained from the survey that it was not completely understood the value of performing additional sampling during outbreak investigation. The proper use of laboratory capacities and knowledge varied between the countries, reducing the capacity to implement and exploit laboratory-based surveillance. For instance, species determination of bacteria was rarely done in Latvia and Estonia and typing (any methodology) was not routinely performed due to lack of resources and available methodology, e.g. when concerning virus diagnostics. On the contrary, the species determination was a standard procedure in Finland and with certain pathogens, such as *L. monocythogenes*, the WGS methodology is already applied (Revez et al., 2017).

Obtaining sufficient amounts of isolates is the Achilles heel of laboratory-based surveillance. Sampling, isolation and species identification of food-borne pathogens must therefore be intensified and routinely performed according to central legislation in all suspected food-borne outbreaks, and competence of the central laboratories must be utilized to maximize the outcome of sampling. Resources should be allocated to train participants in sampling and analytical epidemiology. Simultaneously, public central laboratories must be capable of performing sequencing and downstream bioinformatics analyses. Molecular typing increases the likelihood of detecting case clusters, tracing outbreak sources and confirming outbreak cases only when analysis of food and human samples are done simultaneously and the sequences are made available to all relevant parties. The establishment of a digital computer system available between authorities (described below) containing an isolate database with the isolates' associated contextual data, molecular typing data and other relevant characteristics could facilitate such a data exchange. Patient privacy legislation and IT security legislation hamper with the construction and the use of such a database. In outbreak situation, however, authorities may have access to and use personal information. Adaption of a patient privacy legislation that is flexible could be a possible solution. In addition, the implementation of specific legal framework on provision of isolates to public health laboratories, as already in place in some countries, may be an important driving force to guarantee that national reference laboratories receive a sufficient selection of both positive human and food samples and/or derived WGS data for further typing (ECDC, 2015). Such legislation will only work if sufficient financial resources follow.

We also identified factors unrelated to laboratory-based surveillance hindering communication and successful outbreak investigations shared by the three countries. The most relevant challenges were listed below.

- IT systems with varying accessibility for different officials, as well as strict patient privacy security legislation were perceived as hindering efficient data flow in all three countries. This is a well-described barrier against data sharing, and difficult to change (van Panhuis et al., 2014). In addition to already existing legal barriers, the adoption of the EU General Data Protection Regulation (GDPR) (www.eugdpr.org) creates further uncertainties on what is allowed to share, and might increase the negative impact this factor has on communication.
- Late outbreak notifications (as mentioned above) and a low use of analytical epidemiology studies were noted through the study. According to the interviews, the local authorities contacted central authorities for advice on how to design and interpret analytical epidemiological studies. However, the local officers experienced that access to consultation from central authorities vary, and was sometimes difficult to achieve. It could be that this contributed to an infrequent use of analytical epidemiological studies together with a tardy outbreak notification.
- The use of a logbook varied. While central authorities perceived that most or all outbreak investigations kept a logbook, some local officials used it infrequently, or shared it too rarely.
- Weak communication between food safety and public health authorities on the local level. Factors contributing to communication problems overall were hurry, inadequate humane resources, lack of routines and knowhow, difference in opinions and mistrust between participants, and difficulty getting hold of different participants for various reasons. Of special importance was that there is a lack of practical communication system or network, resulting in that most of the communication reliant on informal contact between officers.

Overall, we conclude that there is a need to improve communication and outbreak investigation in a number of ways in these countries. One is to create a decentralized, shared digital computer system and database allowing secure communication between all stakeholders in an outbreak setting, with implemented functions for submission of samples, outbreak notification and reporting. Currently, communication between laboratories and other national and local stakeholders is usually done by E-mail (at least in two of the studies countries), which is badly suited to preserve patient privacy security and IT-security (van Panhuis et al., 2014; Griffiths et al., 2017). A digital communication system would aid in the harmonization and storing of contextual information, using controlled

vocabulary and defined ontology. This system should be accessible to all relevant parties within the limits of the jurisdiction of their authority and rank. Such computer systems also limit the dependency for informal contact during outbreak investigations, and could facilitate coordination of outbreaks by central authorities. The facilitation of the local and national level outbreak investigation would further strengthen the possibilities to share the accurate data through the European centralized systems, such as EPIS-FWD, RASFF and EWRS.

Due to a small sample size of only three countries, care should be taken when extrapolating our results to bigger MSs with a different governmental structure. Even so, many of the findings described here resonate well with the recommendations from WHO on outbreak investigations (WHO, 2008) For instance, WHO recommends keeping records during the outbreak investigation (logbooks) and searching for advice from laboratories, which we also found to benefit the local outbreak investigations. Furthermore, we would like to expand on the WHO advice on establishing an OCT and recommend that the OCT has a predetermined leader, as this improves the communication. So, despite our study material being small, it concurs with available advice, such as the one from ECDC-toolkits (https://ecdc.europa.eu/en/publications-data/toolkit-investigation-and-response-food-and-waterborne-disease-outbreaks-eu) and research, (WHO, 2008; Jones et al., 2004; Murphree et al., 2012) indicating that it might be a suitable way of action in other countries as well.

## 2.2. The inadequacy of available genome collections for four food-borne pathogens

For several reasons, mainly associated with the fact that available bacterial genomes are largely provided by single research projects, public databases are frequently biased for time of sampling, source and/or geographical origin of the bacterial strains. This increases the risk of incorrectly estimating the genomic diversity of bacterial populations. The species-specific inadequacies of the available genome collections are summarized below.

- The availability of *C. jejuni* sequences on public databases is biased by geography, time and genotype, based on the 7-gene multilocus sequence typing (MLST typing). At the beginning of the project INNUENDO (January 2016), approximately 6,000 genomes were available in public repositories – such as EMBL-EBI European Nucleotide Archive (ENA), National Centre for Biotechnology Information (NCBI) Sequence Read Archive (SRA), pubMLST (https://pubmlst.org/). At that time, samples were mainly part of the sentinel surveillance study for human campylobacteriosis in Oxfordshire, UK (Cody et al., 2013). At the time of writing this report (June 2018), the amount of genomic data has increased substantially to approximately 18,000 submissions from 17 countries, as available in ENA. The main contributor is currently USA (GenomeTrakr Network) https://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/ucm363134.htm) with nearly 9,000 submissions, so geographical bias is still a concern. Relevant metadata is often lacking in the genomic databases; for instance, of the 18,000 submissions in ENA, 16,000 and 7,500 of these lack data on year or country of collection, respectively. The epidemiological context of sampling of the isolates is often unavailable, creating heavy biases in the calculations of genomic diversity when, for instance, clonal isolates are analyzed as sporadic cases. In addition, some genomic lineages are overrepresented: ST-21 CC is by far the most common lineage available. This makes estimation of genomic diversity within other less sampled lineages difficult and unreliable, representing a challenge in areas in which such lineages are more common.
- Although *Yersinia enterocolitica* bio-serotype 4/O:3 is the main cause of human yersiniosis in EU, (EFSA, ECDC, 2017) at the beginning of the project (January 2016) genomic data of this bio-serotype was limited to only 20 strains collected from New Zealand, Australia, UK and France (Reuter et al., 2015). Whole-genome alignment analysis of these strains showed that these are part of a monophyletic clade with very limited genetic diversity (8 to 882 pairwise Single Nucleotide Polymorphisms (SNPs) over 4.6 million base pairs) indicating a recent clonal expansion (Reuter et al., 2015). However, there is a lack of data on the background

population to ensure a correct estimation of genomic diversity of this pathogen and, especially, to estimate the effect of the monomorphic nature of this linage in long-term surveillance and outbreak investigation. Therefore, there is a clear need for enlarging the genome dataset for *Y. enterocolitica* bio-serotype 4/O:3.

- Both *Salmonella* and STEC are main targets of genomic studies across the globe both in research and public health settings. Therefore, for these two species public repositories are populated by a quite diverged set of samples, in terms of time and geographical origin. At the time of writing of this report, EnteroBase (Alikhan et al., 2018) includes approximately 161,000 and 83,000 genomes for *Salmonella* and *E. coli/Shigella*, respectively. Although the databases offer enough information for baseline diversity of the populations of the main pathogenic genotypes of both species, only 399 and 706 *Salmonella* spp. and *E.coli,* respectively, are collected from the INNUENDO participants' countries (i.e. Finland, Estonia, Latvia, Austria, Portugal, and Spain). For *E.coli,* Spain is heavily overrepresented among the participants' countries, as 524 strains were from this MS. Therefore, Enterobase may not correctly represent the genetic diversity of the bacteria circulating within the countries participating to the INNUENDO project, limiting the effectiveness of our pilot-studies.

## 2.3. Identification of needs and current challenges in the implementation of WGS in routine surveillance

To achieve full comparability of molecular data of strains at international, national and regional levels, the development of efficient, standardized and molecular-guided laboratory surveillance is necessary and of high priority. Extensive research has been done on the calibration of bacterial genetic evolution (Achtman, 2008; Llarena et al., 2014; Sheppard et al., 2014; Reuter et al., 2015; Alikhan et al., 2018), but there is still the need to translate these results to laboratory routines (WHO, 2008; ECDC, 2016, Llarena et al., 2016). The main goal of using WGS in molecular surveillance is, indeed, the detection of phylogenetically informative genetic variation that may indicate a common exposure, leading to public health actions. Therefore, WGS-guided epidemiology relies on applying a set of tools that maximize the detection of all possible epidemiologically significant variation between microorganisms to aid in the investigation of food-borne outbreaks (ECDC, 2016). Hence, standardization, calibration of the process, and simplification of data analysis and its subsequent interpretation are basic conditions that must be achieved to ensure WGS traceability, reliability and accuracy.

Although implementing WGS in routine surveillance is a strategic goal for many public health authorities all over the world, the transition from the old diagnostic paradigm to a full WGS consolidation is not free of barriers (ECDC, 2015; ECDC, 2016; Nadon et al., 2017; Revez et al., 2017; EFSA, 2018). Ongoing initiatives worldwide such as GenomeTrkr program in the USA (goo.gl/nnvqXc), FoodNet in Canada (goo.gl/e93GwV) PulseNet International (http://www.pulsenetinternational.org/protocols/wgs/) and COMPARE (https://compare.cbs.dtu.dk/), just to name a few, are examples of the complexity of the problem, the multidisciplinary competences involved and the investment needed. In different countries or regions, resource limitations in terms of budget and/or specific scientific knowledge to properly handle WGS-derived data are challenging the implementation of wet and dry laboratory procedures for WGS, leading to an uncertainty whether this cutting-edge approach will be effectively available (EFSA, 2015; Revez et al., 2017; Llarena et al., 2017). In this section we summarized the main needs (see below) for the implementation of WGS in routine surveillance and outbreak investigation identified during the project INNUENDO, which form the basis for the development of the solutions presented in Section 3. The identified needs are aligned to what international organizations such as EFSA, ECDC, Association of Public Health Laboratories (APHL), Food and Agriculture Organization of the United Nations (FAO) and WHO already published in several reports to ensure traceability and facilitate future accreditation (EFSA, 2015, ECDC, 2015; Gargis et al., 2016; Nadon et al., 2017), including: reduced costs and turnover of the process, simplified data analysis and interpretation, established quality control measures, increased portability of standardized bioinformatic pipelines, defined genome-based typing nomenclature for national and international comparison, and integration of WGS data into public health risk assessment methods.

### 2.3.1. Needs for quality control/optimization of wet lab procedures

Although the increasing application of WGS as the standard genotyping method for routine surveillance and outbreak investigation of food-borne pathogens is expected to progressively lead to standardized and accredited WGS-related wet-lab practices, there is a general lack of information about individual and collaborative efforts currently being done towards the improvement and harmonization of such practices at both intra- and inter-laboratory levels. As a consequence, there might be an erroneous impression that WGS-related wet-lab procedures should not be subjected to quality control measures as rigorous as the ones applied to downstream bioinformatics. Particularly, there is an assumption that the transition to the application of WGS for routine surveillance and outbreak investigation of food-borne will not require substantial changes on the laboratories practices regarding DNA extraction, since obtaining the required amount of DNA is not challenging for food-borne pathogens, which can be easily cultured, and currently available commercial kits generally yield good-quality DNA. Likewise, downstream steps (i.e., library preparation and sequencing) are often performed in centralized sequencing facilities, or in external service providers, which might underestimate the gain that could come from applying highly controlled procedures based on preliminary proof-of-concept assays, continuous monitoring and circumstantial technical adjustment. This rationale might be a key driving force towards a pathogen-specific long-term, large-scale, routine WGS-based surveillance system intended to be reproducible, cost-effective and of high-quality. In summary, efforts are needed to understand to which extent the application of such rationale on both DNA extraction and sequencing-related procedures could positively impact the performance of the overall WGS process (e.g., reproducibility, quality, cost), which, again, may be of utmost importance for small and/or less-resourced countries or regions with less flexibility in available resources.

### 2.3.2. Need for quality control/optimization of dry lab procedures

In the epidemiological surveillance of food-borne pathogens using WGS data, raw read sequencing data (and downstream derived data) needs to be properly handled and interpreted in order to be useful for public health action, i.e., allow meaningful linkage of cases, and subsequently timely detection of clusters and outbreaks. Therefore, as in any diagnostic and laboratory procedures, a key factor necessary for ensuring the quality of WGS-based analyses is the establishment of quality control (QC) measures. QC procedures are defined to verify if the performance specifications are met for each run, monitoring whether or not every part of an analysis executes properly and delivers correct results. Using these procedures, the operator ensures that no sequence data move forward in the process without meeting the minimum quality standards. Few QC matrixes have been defined for NGS-based testing and several software applications have been developed for this scope (Gargis et al., 2016). However, these tools are either too challenging for non-bioinformaticians to use or simplified down to a one-button-click black box. When software are too difficult to use, the user often is overwhelmed by the number of parameters to choose/evaluate, while in a black box the user has no knowledge on what is actually being analyzed or done to the sequences. Therefore, a transparent easy-to-use pipeline for quality control measurement from raw sequences to phylogenetic tree is needed. It is important to establish appropriate QC procedures for the entire testing process, including all the dry-lab components, for validating the quality of raw reads (affecting both genome assembly and assembly-free typing methods), assemblies (affecting allele calling and *in silico* prediction), and allele callings (affecting clustering and WGS-based classification).

### 2.3.3. Needs for standardized bioinformatic analysis

To make WGS useful in public health services, sequence data should be translated to biologically relevant and communicable subtypes. The definition of subtypes requires a standardized methodological approach and a nomenclature to describe higher-order relationship between isolates (ECDC, 2015). Named subtypes enable rapid data-analysis, contextualized results, and efficient exchange of information, facilitating a swift public health response to infectious disease and ultimately an improved disease prevention and control.

The two epidemiological settings of long-term surveillance and outbreak investigations have different requirements that need to be considered when defining WGS-subtypes. The goal of molecular surveillance, expecially long-term surveillance, is usually to continuously record the types of bacterial pathogens circulating in a specific geographic area, while the genomic analysis of an outbreak investigation aims at identifying patterns of shared variation to infer transmission and identify a common source. ECDC advises that typing in an outbreak investigation is done with sufficient resolution to discriminate between outbreak isolates and sporadic cases, and therefore the used typing resolution must be able to define all possible diversity within a cluster of closely related strains (see ECDC toolkit for food-borne outbreaks; https://ecdc.europa.eu/en/publications-data/toolkit-investigation-and-response-food-and-waterborne-disease-outbreaks-eu). Such a high level of resolution hampers effective surveillance, as the establishment of too many subtypes may make it difficult to associate strains with each other and detect trends. Since the whole data analysis process affects the discrimination level of the strains, the methodologies applied used in WGS-based analysis (e.g. reference-based variant calling, gene-by-gene methodology, distance based clustering, etc.) and their parameters (e.g. inclusion criteria for variant presence, definition of reference locus or genome, etc.) have important implications for the interpretations of results. In addition, the process of defining clusters composed of strains likely to be related by the use of thresholds is an important consideration in the application of any subtyping scheme, since even small adjustments in these thresholds can have a dramatic impact on cluster composition and stability. The optimization of these thresholds/parameters, such as number of variant differences for clonal definition either by SNP or gene-by-gene analysis, has been a recurring challenge in the field of molecular epidemiology. An additional challenge of WGS-based subtyping is nomenclature, as WGS is sensitive to the addition of novel genome sequences. Clearly, a systematic methodology to set a robust nomenclature and thresholds for both surveillance and outbreak situations is needed.

### 2.3.4. Needs for data sharing

Sufficient and well-defined metadata and an efficient and accurate communication between public health, food and veterinary authorities, laboratories, medical practitioners, food industries, media and the public are important when solving an outbreak. For WGS subtyping to be a valuable tool in genomic epidemiology, it must be combined with epidemiological, clinical, laboratory, genomic and other health care data ("contextual data") (Griffiths et al., 2017). Availability of such data increase the utility of genomic information in outbreak investigations, but barriers of legal, political and technical nature challenge the sharing of data and integration between agencies, especially during multi-states outbreaks (van Panhuis et al., 2014). Today, the scientific community and authorities encourage fast, global sharing of raw sequence data, as real-time sharing of WGS data propels basic science research and diagnostics even during ongoing outbreaks (Ruppitsch et al., 2015). Especially sharing of international and national WGS data in (quasi) real-time is important for efficient cross-border and/ or multijurisdictional outbreak detection and investigation. Such real-time sharing of genome sequences marks an important departure from the traditional model of keeping data and analyzing in secure systems, which has been typical in public health surveillance up-to-now. However, bioinformatic solutions supporting simple, secure and standardize ways to share data and analysis need yet to be implemented.

In Europe, the *General Data Protection Regulation* (GDPR), which came into effect May 2018 (https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en), challenges WGS data sharing as identification codes for bacterial isolates (and sequences) obtained from a human patient are considered to be personal information. Therefore, anonymization of data was initially required in some countries if WGS data was to be shared in public repositories. However, insecurities remain in how to share WGS data in light of the different interpretation of the GDPR among MS, and international guidelines on the interpretation of GDPR for public health, food and veterinary authorities are needed.

Regardless recent improvements, there are still certain technical obstacles in batch submission of raw data (i.e. *fastq* reads) and assemblies (i.e. *fasta* files) to public databases. Moreover, the minimal

metadata fields required specifically for pathogen samples are still under discussion. The current ones might not be compatible with the information allowed to be provided to the public databases. For example, the sharing of complete collection date and collection site using *Global Positioning System* (GPS) coordinates is still not allowed by certain institutions. A few methods to circumvent several of these obstacles already exist, especially in regard to ENA submission (https://github.com/phe-bioinformatics/ena_submission), but improvements and a stabler API are still needed in order to allow a more generalized use.

### 2.3.5. Needs for data storage and portable bioinformatic solutions

Analysis of high throughput sequences has two distinct needs for data storage: storage of raw data (e.g. compressed files in *fastq* format) and disk space required for the bioinformatic analyses (intermediary and the final files generated by each softwaretool). Although it is only temporary, the storage needed for the analysis can be several times larger than the space needed to save raw data.

Raw data storage per strain increases with both genome size and depth of coverage and can scale up to the terabyte level for a laboratory analyzing approximately 200 strains per week. This represents considerable storage costs imposed by an increasing need for local storage and maintenance by IT support. Furthermore, bioinformatics software has disk space requirements to operate, and although transient, these requirements must be taken into consideration for the software to run smoothly. Also, bioinformatic tools have different computational requirements, in terms of the number of central processing unit (CPUs) and memory used, dependent on genome size and depth of coverage of analyzed strains: larger genome sizes and higher coverage take longer to compute due to the proportional increase of the number of reads in the raw data.

Depending of the resources available, institutions might wish for a software solution able to handle several analyses a day, with multiple accesses and available in High-Performance computers (HPC) or cloud-based Virtual Machines (VM). On the contrary, other institutions would be interested in an analytical platform which can run in a single laptop, regardless specific operating system (OS) requirements. These two examples show the challenges related to implementation of standardized bioinformatic solutions for public health microbiology. Therefore, portability, the ability to use an application in different computer environments, has enormous implications on software development for bioinformatics solutions in molecular epidemiology. Computer environments mean hardware, operating systems and interfaces with other software, users and programmers. Development of portable bioinformatics solutions able to account all different needs, network and IT-specifications of different institutions are currently high priority goals.

## 3. The INNUENDO Platform V1.0: addressing the needs for harmonization and standardization in genome-based surveillance of food-borne pathogens

A lack of standardized bioinformatics infrastructures for data processing and integration, together with still limited bioinformatics skills, continues to be some of the major hurdles towards routine implementation of WGS analysis in the service of public health and food safety authorities. A more efficient way to communicate epidemiological data, share sequences with strains information and their attached contextual data between stakeholders and a wider global community, should be encouraged and is needed. These needed attributes can be offered, at some degree, through web platforms which can operate globally. One of the most promising examples of such a platform is IRIDA (www.irida.ca), which uses open-source software to provide a distributed platform for genomic data analysis, but focused on SNP analysis and does not allow creating a nomenclature for strain types. However, there is still a need for a more comprehensive tool or platform to integrate epidemiology and sequence data together while maintaining the ability to share the data without the loss of patient privacy rights.

We therefore developed the INNUENDO Platform V1.0 as part of the INNUENDO project to tackle the needs of public health and food safety authorities, while addressing the functionalities and flaws in

data sharing and WGS analysis. A detailed description of the INNUENDO Platform V1.0 and its features is presented in the following Sections.

The INNUENDO Platform V1.0 is licensed under the GPLv3 license. The source code of INNUENDO Platform V1.0 and the documentation are available at https://innuendo.readthedocs.io.

## 3.1. Overview of the INNUENDO Platform V1.0

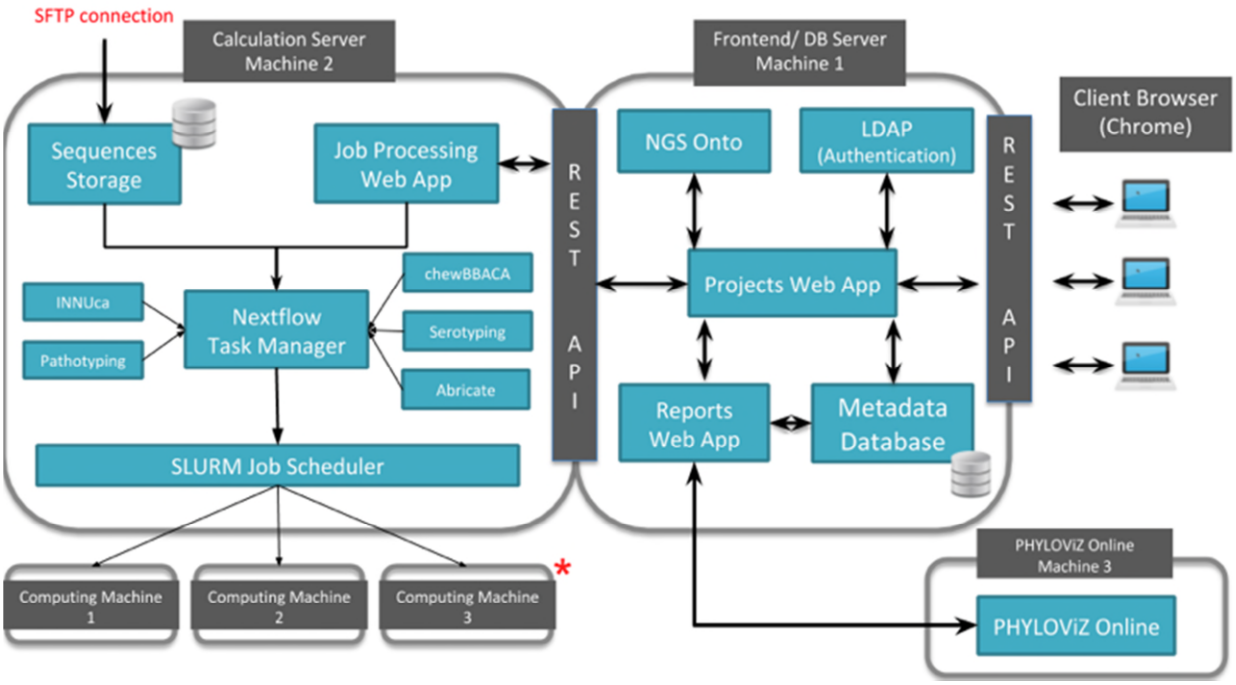
The INNUENDO Platform V1.0 is an open source software that provides a user-friendly interface and the required framework for WGS data analysis for use in public health and food safety laboratories. From raw data quality assurance to integration of epidemiological data and visualization of the final analyses, INNUENDO provides all necessary tools for the use of High Throughput Sequencing (HTS) techniques in everyday surveillance and outbreak investigation.

The Platform has been designed specifically for fulfilling the following requisites:

- Procedures (including software environments) for raw sequencing data quality control;
- An analytical framework for using WGS in phenotypic prediction and cluster analysis for surveillance and outbreak investigation;
- Web-based publicly accessible species-specific hybrid analytical pipelines (reads- and assembly-based analysis);
- Species-specific genomic-based nomenclature for pathogen surveillance and use of ontologies to annotate the analytical processes;
- A standard communication protocol to be used for exchanging pathogen specific data and metadata at multi-country level.

Due to requirements for future accreditation of analytical procedures, INNUENDO Platform V1.0 is made as a transparent box, meaning that each component of the INNUENDO Platform V1.0 is open source and the implementation process is clear in all phases. It has been designed based on a modular framework which is easy to upgrade and allows the integration of different software components required for the tasks described above. Therefore, the designed infrastructure allows automatic standardized analyses while being flexible enough to deal with the developing fields of technology and data analysis algorithms. It allows an easy incorporation of different bioinformatics tools needed for the characterization of different bacterial pathogens. Moreover, users can choose predefined set of analytical modules containing procedures that can be combined in species-specific workflows by a system administrator.

Moreover, portability is a key component of the INNUENDO Platform V1.0 and, therefore, it can run on a variety of computing resources in research labs, reference laboratories and health agencies. The system implemented is both scalable and elastic enough to fit the individual needs of different institutions and countries. INNUENDO Platform V1.0 is developed with a computationally efficient modular design, runnable on both high-end laptops (making use of Docker compose tool) and HPC or cloud-based VMs (for multiple configuration where multiple users are needed). All information about possible configurations can be found at the INNUENDO Platform V1.0 documentation page (https://innuendo.readthedocs.io/). Since the platform interaction is done through the web browser, the INNUENDO Platform V1.0 can be deployed, if resources are available, in a web server with internet access.
The INNUENDO Platform V1.0 is divided into two distinct components that communicate between each other and with the client web-browser through REST (Representational State Transfer) - API. The two applications are the INNUENDO Frontend Server and the INNUENDO Process controller/Calculation server (Figure 2). These two applications can therefore run simultaneously in different virtual or physical machines.




The Platform structure is divided into two applications (Calculation server and Frontend server) interacting between each other and with the client web-browser. The user uploads sequence data to the Storage component through Secure File Transfer Protocol (SFTP). All the components can be located on a single machine (in case of a high-end laptop) or distributed along several machines (HPCs).

*The calculation server through the workflow manager engine can distribute the computation of procedures between multiple machines.

**Figure 2:** Flowchart of the interaction between the different components of the INNUENDO Platform

Several analytical components have been included in the INNUENDO Platform V1.0 which are described in details in the following sections: *de novo* assemblies using the INNUca pipeline (https://github.com/B-UMMI/INNUca), Multi-Locus Sequence Typing (MLST) determination with mlst2.10 (https://github.com/tseemann/mlst) and allelic profiles assignment done by chewBBACA (https://github.com/B-UMMI/chewBBACA) (Silva et al., 2018). Antimicrobial and virulence factor detection is done by the ABRicate software (https://github.com/tseemann/abricate), while ReMatCh software (https://github.com/B-UMMI/ReMatCh) is used as engine for assembly-free serotyping of *E. coli* (using seq_typing - https://github.com/B-UMMI/seq_typing) and pathotyping of *E. coli* and *Y. enterocolitica* strains (using patho_typing - https://github.com/B-UMMI/patho_typing). The software SISTR (https://github.com/peterk87/sistr_cmd) is used for serotyping of *Salmonella enterica*. The results are visualized through a custom-built report and with PHYLOViZ Online 2.0 (http://online2.phyloviz.net). All these processes are summarized for *E. coli* in Figure 3.

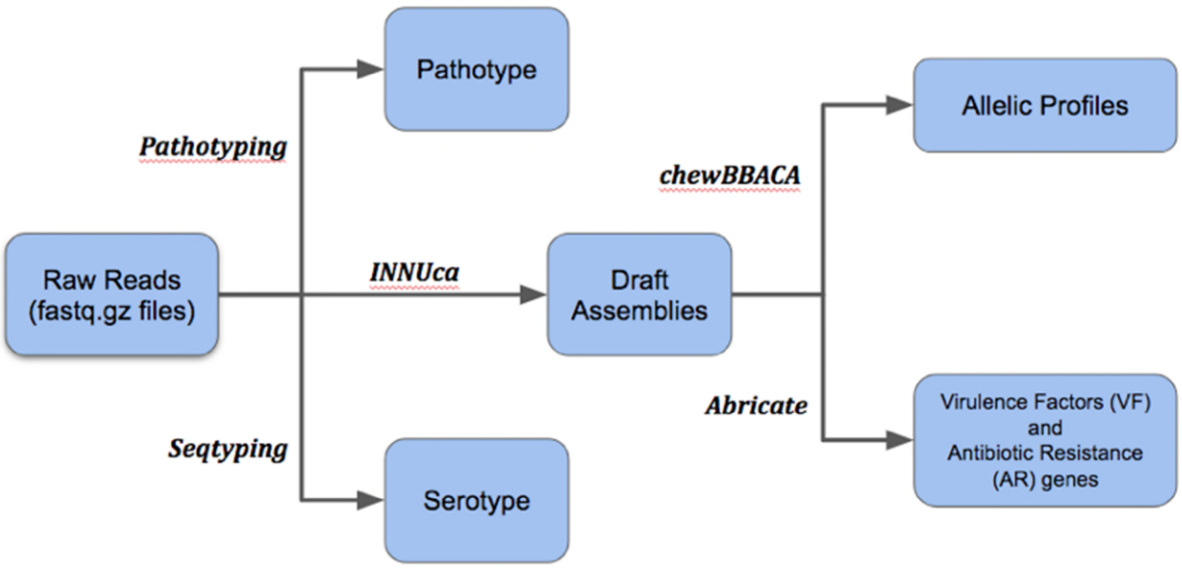



**Figure 3:** General representation of the INNUENDO Platform V1.0 bioinformatics analysis pipeline for *E. coli*

### 3.1.1. The Frontend server

The INNUENDO Frontend Server is the application that interacts directly with the user. It comprises a user-friendly web-interface available after login through a secure user authentication with Lightweight Directory Access Protocol (LDAP), previously configured by a system administrator. LDAP is not implemented in the laptop version as such versions are targeted for individual users.

The application uses a PostgreSQL metadata database for storage of sample laboratory information (i.e. DNA extraction, library preparation and sequencing information), epidemiological data of isolates and results from the bioinformatics procedures. All this data is then combined to allow linking epidemiological data and bioinformatics analyses results into highly customizable minimum spanning tree visualizations using PHYLOViZ Online software (Ribeiro-Gonçalves et al., 2016).

In the Reports web application, users can analyze the results and quality control measurements from the analytical procedures through tables and interactive charts provided. These reports can then be saved in a file and shared with other INNUENDO Platform V1.0 users or connected with third party software platforms.

The INNUENDO Frontend server also uses NGSOnto (Silva et al., 2013), an ontology that aims at capturing the workflow of all the processes involved in NGS data analysis in order to ensure the reproducibility of the process through the use of a controlled and specific vocabulary (https://github.com/mickaelsilva/NGSOnto). NGSOnto acts as the backbone for establishing relationships between user projects and their respective isolates, and keeps track of procedures running on those projects and their status. The data is stored in AllegroGraph® (https://franz.com/agraph/allegrograph/), a highly efficient triple store system.

### 3.1.2. The Process controller/Calculation server

The INNUENDO process controller/calculation server is the second application of the INNUENDO Platform V1.0 developed with the aim of working as a bridge to run analytical "Procedures" (pre-defined bioinformatics analyses) on a laptop or in a HPC, with the help of SLURM (https://slurm.schedmd.com/) process manager (Yoo et al., 2003).

The application interacts with the Nextflow workflow manager (https://www.nextflow.io) to run a series of different modules requested by the INNUENDO Frontend server (Di Tommaso et al., 2017), which are built using FlowCraft pipeline assembler (https://flowcraft.readthedocs.io/). Nextflow assembles the jobs precedence accordingly with the requested workflow and submit the jobs execution to SLURM that will manage the available resources in order to optimize the workload. Also, Nextflow communicates with the Frontend server to control and update the analysis status and store the analyses results into the PostgreSQL database. In order to run these jobs, the process controller server and all the computation machines must have access to the available shared storage to read and write software results, and to the raw data uploaded by the users to the INNUENDO Platform V1.0.

All modules ran by Nextflow are defined and stored as Docker images (Boettiger, 2014) which allows a better version for control and software update, allowing to maintain full reproducibility of analysis done with previous software versions when needed.

The Docker images can be found at https://hub.docker.com/u/ummidock/.

### 3.1.3. Complete description of usage

The INNUENDO Platform V1.0 requires two types of users that can be defined on the LDAP server: *admin* and *innuendo-users*. On the laptop version there is a single default user (*innuendo_user* by default but can be changed in the configuration file of the Platform) which act as *admin*.

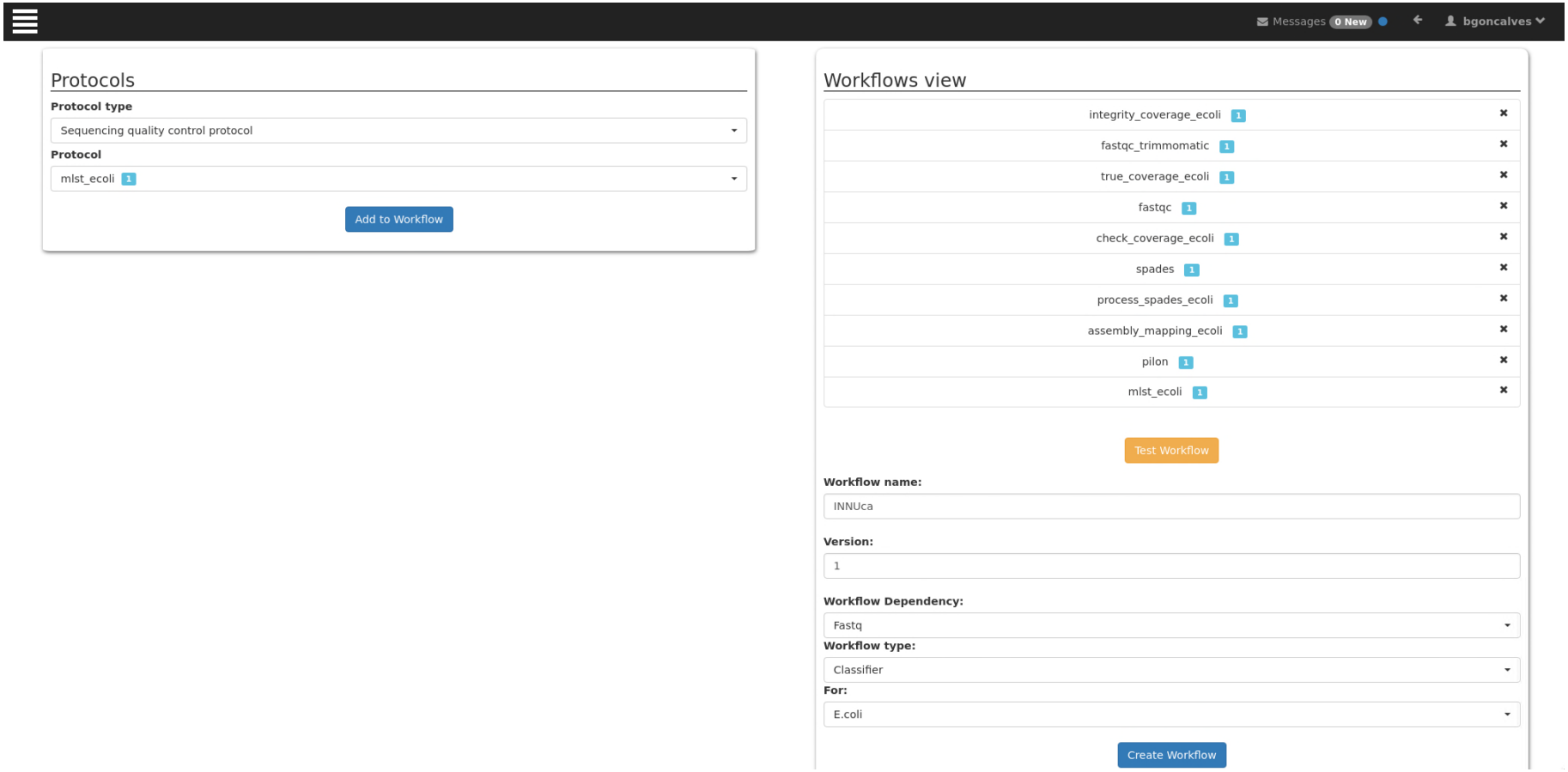
*Admin* users have the rights to create new *Protocols* and *Workflows* directly on the user web-interface. *Protocols* are given procedures with different sets of parameters. Those protocols can then be merged into a single *Workflow* and multiple *Workflows* can be applied directly to an isolate to construct customized *Pipelines.* After the job submission, the applied *Protocols* are passed to FlowCraft which builds the required files structure to run customized Nextflow jobs for the desired pipeline (Figure 4). The Platform operates through the creation of *Projects* that aggregate samples (Figure 5). These samples can be attached to a *Project* by filling a form and submitting to the database or by using some of the strains already available on the INNUENDO database.

*Fastq* files need to be uploaded *à priori* to the Platform through SSH File Transfer Protocol (SFTP) to a predefined IP address. Accession numbers can also be assigned to a strain if no *fastq* files are available in order to download the *fastq* data from ENA or SRA public databases.

After adding strains to a *Project* the user can select predefined *Procedures/Workflows* and apply them to the selected strains.

After job submission, a notification system is triggered that allows checking the status of each job by using a color code on each of the *Procedures/Workflows* associated to a given sample (Figure 6):

- White – Job not submitted.
- Orange – Pending job.
- Blue – Running job.
- Green – Job ran successfully.
- Yellow – Job ran with a warning message.
- Red – Job failed.




INNUca is composed of a series of Protocols that can be merged to build a Workflow to be applied to isolates. Information of the used Protocols is then passed to FlowCraft to build the Nexflow files required to run the jobs

**Figure 4:** Example of the creation of the INNUca Workflow based on its Protocols from *Admin*

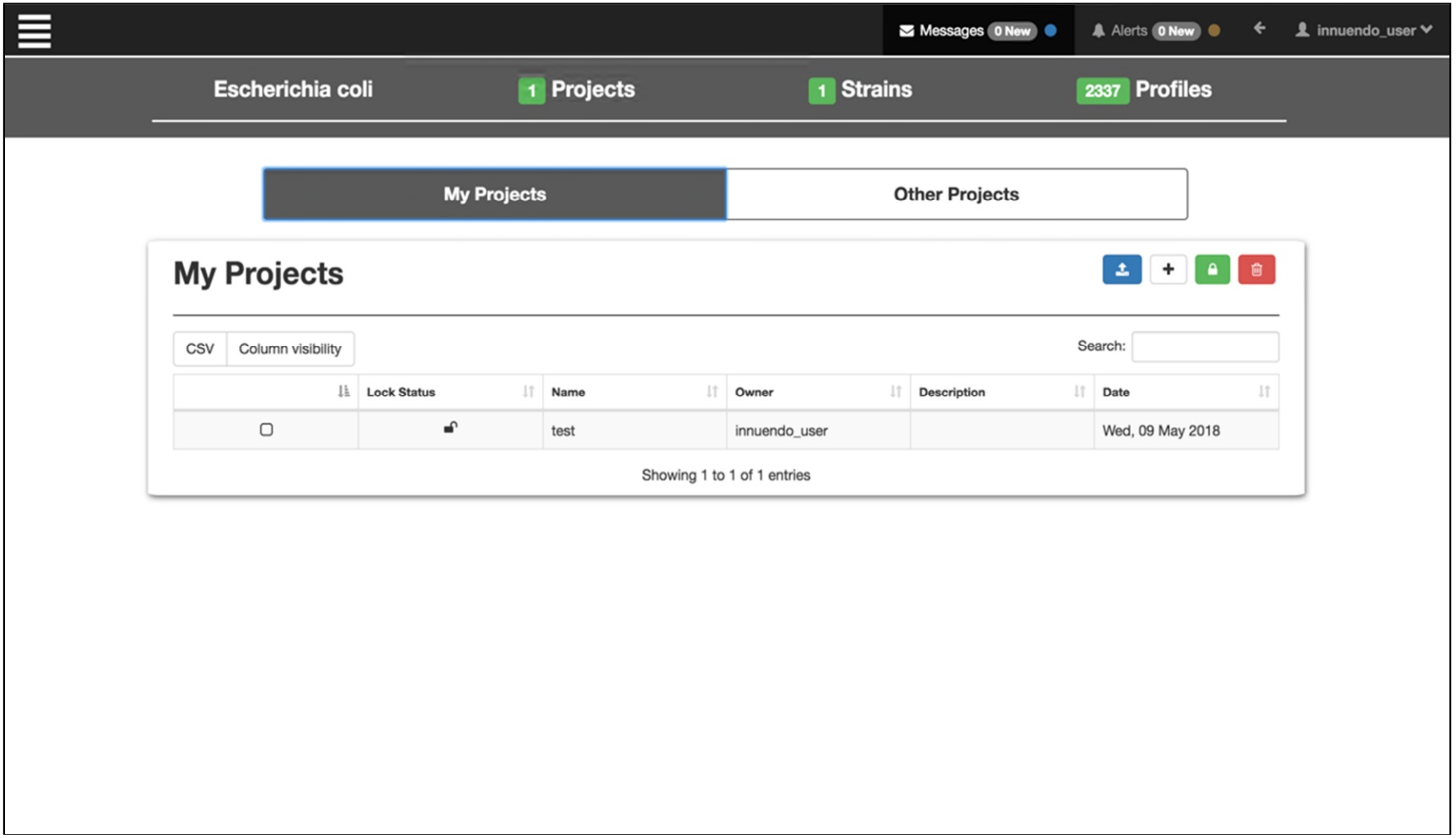
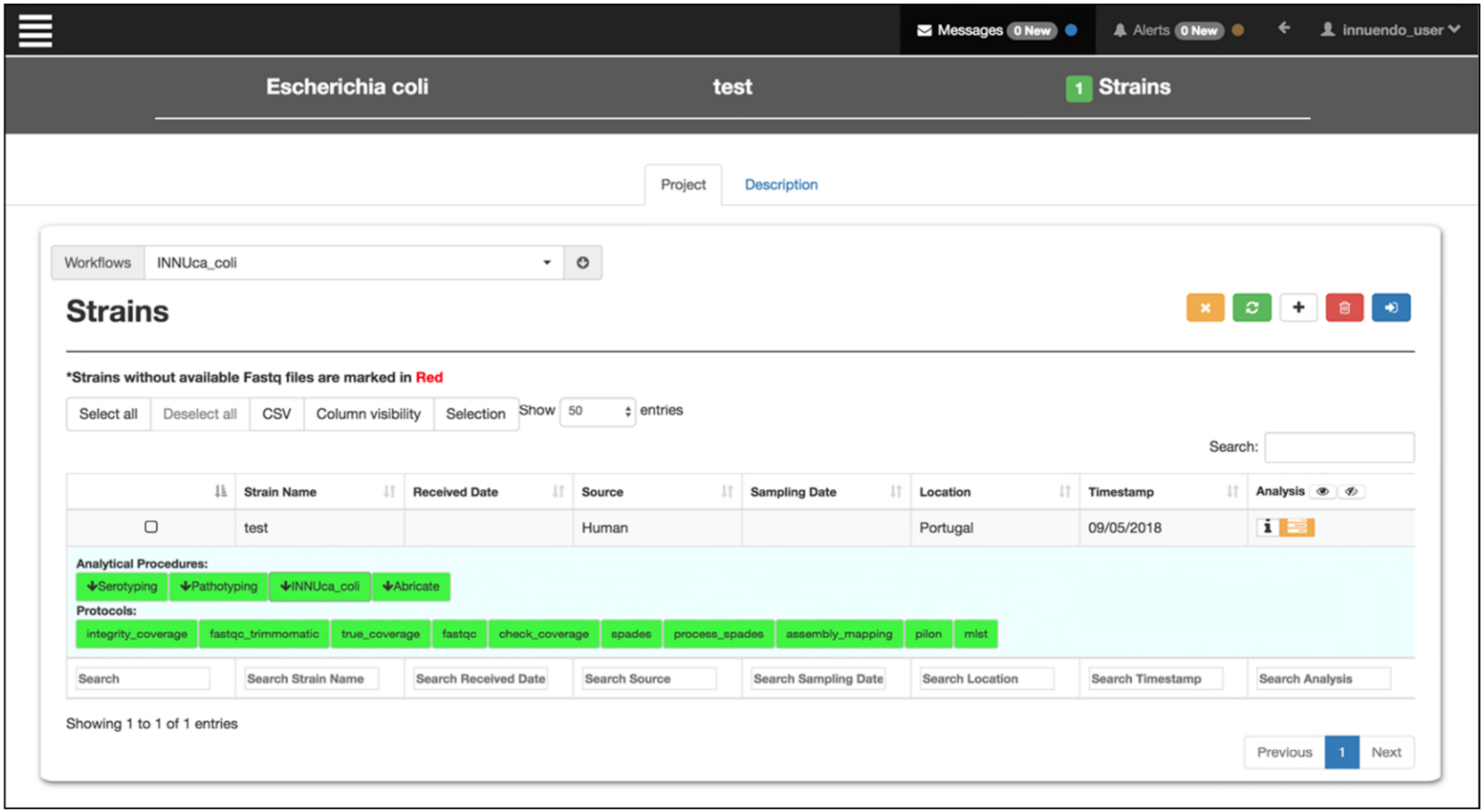



**Figure 5:** Screenshot of the INNUENDO Platform V1.0 showing the list of accessible projects




Everything is colour coded in accordance with the corresponding job status (in this case all ran successfully)

**Figure 6:** Screenshot of the INNUENDO Platform V1.0 showing a *Pipeline* with different *Workflows* (i.e. Serotyping, Pathotyping, INNUca and ABRIcate) and INNUca's the list of *Procedures*

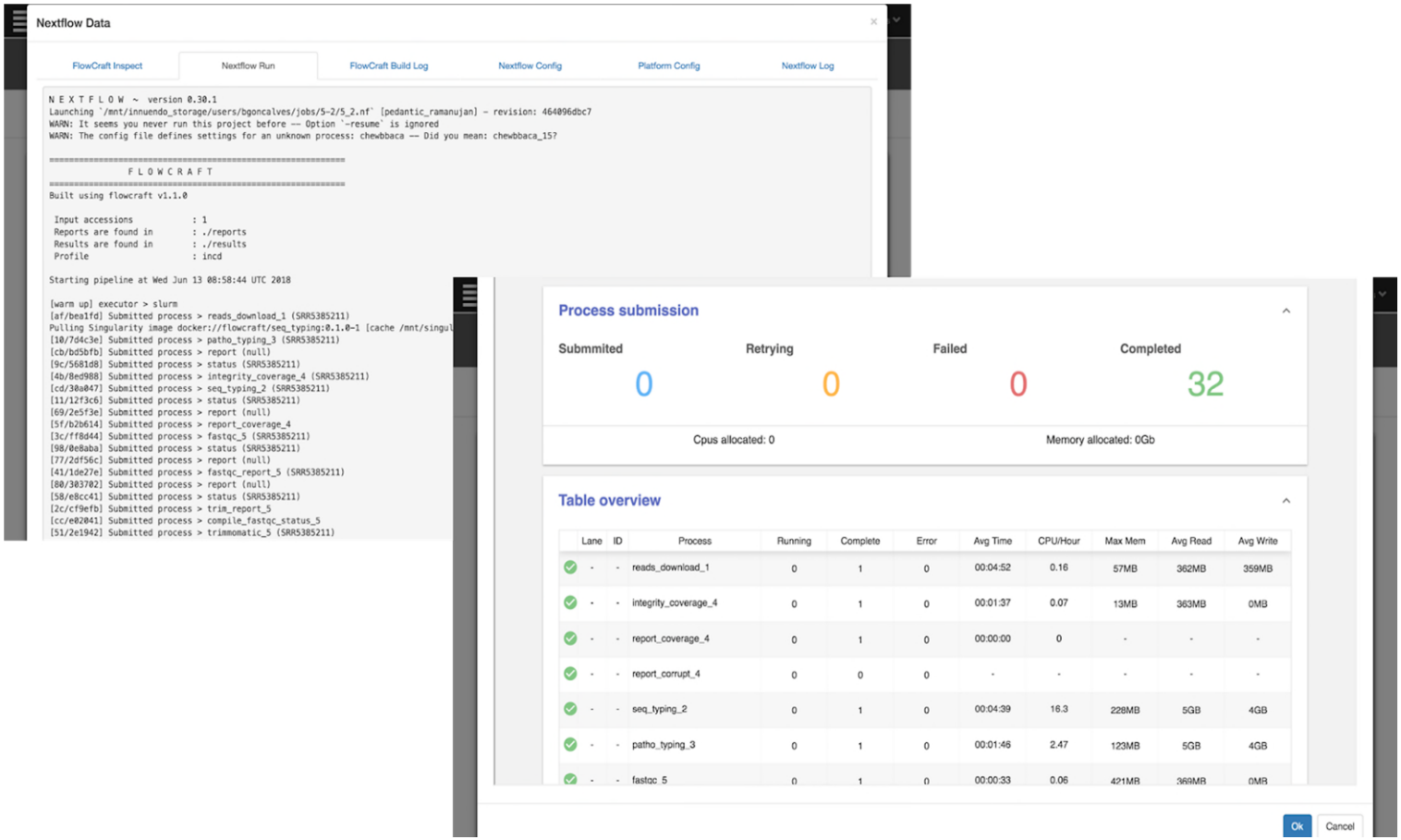



**Figure 7:** Screenshot of the INNUENDO Platform V1.0 showing *Admin* tools for job submission tracking for each strain (in the front) and single strain Nextflow log file (in the back)

After completing the job, a log report and a predefined output are available for each procedure. Additional results can then be retrieved at the *Reports* section of the web-interface, where a user can select strains from different projects to construct a report. In order to control status of submitted jobs and track possible errors, platform administrators have some additional features on the web application that allows them to visualize the generated files upon pipeline creation and job submission, and also have access to a service to keep track of all the pipeline steps in real time. This is possible through the use of a web application provided by FlowCraft (i.e. FlowCraft Inspect) (https://github.com/assemblerflow/flowcraft-webapp) that was merged into the platform (Figure 7).

## 3.2. The legacy dataset

The INNUENDO Platform V1.0 contains a ready analyzed legacy dataset composed by genomes collected from public repositories, sequenced within the project and provided by a partner organization.

At the start of the project, the amount of WGS data within the consortium was limited. So, with the aim of improving the capability to detect outbreaks using WGS, the strains selected were essentially intended to provide high-quality genome sequences to enlarge the public repositories from well characterized strains of *C. jejuni*, *Y. enterocolitica*, *S. enterica* and STEC. The consortium also focused on collecting genomic information of isolates obtained from outbreak and sporadic human infections, food and animal sources, both from public repositories and partner organizations to increase the knowledge of background variability. The database contains enough genetic diversity to be used in the gene-by-gene schema definition, and is relevant for other EU MS.

Isolates fulfilling the following requirement were considered for genome sequencing.

- *C. jejuni:* strains with known MLST sequence type and isolated between 1990 and 2016 from all possible sources.
- *Y. enterocolitica*: strains of serotype O:3 and/or biotype 4 isolated between 1990 and 2016 from all possible sources.
- *S. enterica*: strains from serovar Enteritidis for which subtyping and epidemiological information was available, isolated between 1990 and 2016 from all possible sources.
- STEC: strains from known serovars for which subtyping and epidemiological information was available, isolated between 1990 and 2016 from all possible sources.

Table 4 summarizes the composition of the INNUENDO Legacy Dataset for each of the four target species. Annex B include the list of submitted sequences to ENA including the minimum metadata. Details on the INNUENDO sequencing project and the selection of the genomes for the INNUENDO Legacy Dataset are available in Appendix A.

**Table 4:** Genomes which form the INNUENDO Legacy Dataset

| | INNUENDO sequencing project[a] | Genomes provided by partners or beneficiaries [c] | Genomes collected from public repositories | Total[d] (INNUENDO Legacy Dataset) |
|---|---|---|---|---|
| *C. jejuni* | 269 (279)[b] | 566 | 5,691 | 6,526 |
| *Y. enterocolitica* | 79 (80) [b] | 0 | 252 | 331 |
| *S. enterica* | 129 | 153 | 4,307 | 4,589 |
| *E. coli* | 119 | 0 | 2,218 | 2,337 |

a) The number of genomes sequenced by INSA (Instituto Nacional de Saúde Dr. Ricardo Jorge, Lisbon, Portugal) on behalf of the INNUENDO consortium included in the INNUENDO Legacy Dataset and shared publically in ENA under the project accession number PRJEB27020 (Annex B); the sequencing of the genomes were co-funded by EFSA and The Basque Government (Eusko Jaurlaritza/Gobierno Vasco).

b) A total of 10 and 1 genomes for *C. jejuni* and *Y. enterocolitica*, respectively, were submitted to ENA and listed in Annex B but not included in the Legacy Dataset since they did not pass the quality check for allele calling (see Section 3.6.3.2). The total amount of raw reads submitted to ENA for these two species are indicated between brackets.

c) Raw reads were produced for scopes other than the INNUENDO project, and shared by partners/beneficiaries after the submission of respective reports or publications; some of the raw reads were not publicaly available at the time of writing of this report (June 2018).
d) The sum of the genomes produced within the INNUENDO project, the genomes shared by the partner organizations and the genomes collected from public repositories (ENA and SRA) at the time of the assembly of the dataset. All together the genomes form the INNUENDO legacy dataset.

## 3.3. Quality control measures

### 3.3.1. Recommendations for WGS-related wet lab

DNA extraction is not assumed to be a main obstacle for WGS (especially for Gram-negative bacteria), but we observed that variables in this step (e.g., use of distinct methods, lack of species-oriented lysis, lack of fluorometric-based quantification and DNA integrity control) could impact the reproducibility of the results. Therefore, more efforts should be put on the importance of optimizing and validating species-specific WGS-oriented DNA extraction protocols.

Optimizing and validating the preliminary steps of sequencing-related procedures (i.e., library preparation and sequencing run) was done together with continuous monitoring of quality indicators. Furthermore, replicate aliquots of the same DNA were included in every NGS run as an "internal indicator" of reproducibility for each pathogen. These points ensured the success of this wet lab task marked by the following outcomes: i) WGS of all selected isolates were of high-quality and above the initially agreed depth of coverage; ii) highly balanced depth of coverage between samples and, consequently, less need for re-sequencing; and, iii) progressively higher outputs of sequencing runs, while keeping quality levels above manufacturer specifications. Taken together, these outcomes had a beneficial impact on downstream analysis (e.g., less inter-sample discrepancies in the total number of cgMLST loci called; more balanced number of contigs), while contributing to testing the robustness and reproducibility of both the wet- and dry-lab workflows.

In this context and following our approach, it is recommended that large-scale WGS for surveillance should be preceded by preliminary protocol testing/refinement relying on "control" samples and that the validation of technical adjustments throughout the time always aim to ensure that sequencing yield and quality levels are equivalent to those recommended by manufacturers. In addition, the same rational should be applied when other changes to manufacturer's recommendations are needed, in particular those imposed by constraints of material and equipment availability. This might be especially relevant for public laboratories, where the acquisition of resources is subjected to specific legislation.

Other less stringent recommendations that arose from the INNUENDO experience include: 1) to perform species-exclusive WGS runs is advisable, as sequencing genomes with similar size and GC content highly buffer the yield/coverage fluctuations; 2) to have dedicated human resources may result in a more controlled "start-to-end" WGS procedures, from DNA QC to the sequencing run; and, finally, 3) to promote collaborative efforts and inter-laboratory exchange of information on this subject is desirable towards the standardization and harmonization of WGS-related wet lab practices.

### 3.3.2. Recommendations for genome assembly: the INNUca pipeline

Evaluating sequence quality and the existence of possible technical errors can be done by setting up careful QC measures on all components of the WGS-based typing. Without these measures, complete automatization would not be possible and the accreditation of the proposed analytical framework would be at risk. Assessing genome assembly quality is a cornerstone in this process, as poor quality assemblies hamper downstream analysis resulting in incorrect interpretations. As such, it is critical to identify, evaluate and minimize technical errors occurring during sample isolation, DNA preparation sequencing and genome assembly.

Thus, the INNUENDO Platform V1.0 includes the INNUca pipeline: a standardized, fully automated, flexible, portable and pathogen-independent bioinformatics pipeline for bacterial genome assembly

and quality control to produce high quality assemblies. It also provides researchers with limited bioinformatics expertise a stable, but adjustable, pipeline to work with. The INNUca pipeline consists of several modules analyzing and processing raw sequencing data to *de novo* assembly, species confirmation and MLST determination. All the INNUca steps are subject to quality control using clearly defined thresholds to ensure data quality, resulting in a simple "FAIL/WARNING/PASS" flag for each module, compiled in a report at the end. To achieve high quality standards, INNUca makes use of already available tools in *de novo* bacteria genome assembly production (Figure 8).

First, INNUca calculates if the samples raw data fulfill a minimum expected coverage. Then, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) performs a read quality analysis and Trimmomatic (Bolger et al., 2014) (http://www.usadellab.org/cms/?page=trimmomatic) trims the reads.

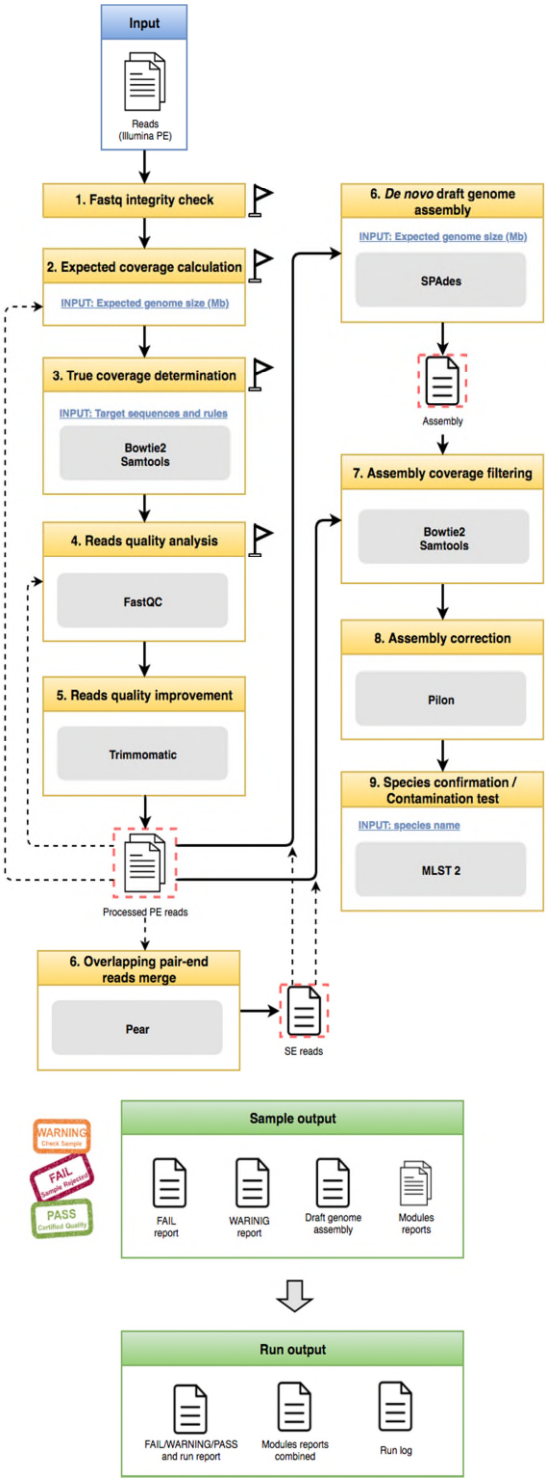
Trimmed reads' quality is again inspected with FastQC. The *de novo* draft genome assembly is performed with SPAdes (Bankevich et al., 2012) (http://cab.spbu.ru/software/spades/), and then subsequently coverage filtering using Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) and Samtools (http://www.htslib.org/doc/samtools.html) before being corrected using Pilon (Walker et al., 2014) (https://github.com/broadinstitute/pilon) in order to significantly improve the draft genome by removing very low represented sequences, correcting bases, fixing misassemblies and filling gaps. If required, overlapping pair-end reads can be merged using Pear (Zhang et al., 2014) (https://sco.h-its.org/exelixis/web/software/pear/doc.html) prior to assembly.

Additionally, for a pre-defined set of bacterial species (although it can be customized for any bacteria), at a very early stage of the workflow, INNUca estimates the bacterial chromosome depth of coverage. This module (named *TrueCoverage*) uses ReMatCh software (https://github.com/B-UMMI/ReMatCh) to map the reads against a set of reference core loci (approximately 20 for the four species of interest) distributed throughout the genome. This module can also detect contamination with different strains or species and will stop the process for a given sample if minimal criteria are unfulfilled.

Since, a critical step in QC of assemblies is the correct species determination, the INNUca workflow ends with species confirmation and MLST prediction using mlst2 (https://github.com/tseemann/mlst).

The INNUca pipeline allows saving all the intermediate files which can then be explored for further data exploration. Although INNUca is standardized and fully automated, it can be easly and extensively adjustable. As a part of a transparent box philosophy, INNUca provides flexibility to change the set QC parameters for each step if required, thus enabling it to be utilized as assembly pipeline independently of the bacterial species.

Details on the INNUca mocules can be found in Appendix B and the software is available at https://github.com/B-UMMI/INNUca.




Blue box and texts denote input data and information for INNUca to run. Yellow boxes describe the different modules run by INNUca. The external software required by INNUca to run is indicated inside the grey boxes. Green boxes show the different outputs produced by this workflow. Black flags mark quality checkpoints for INNUca workflow to proceed. Dashed arrows indicate quality control reassessments after reads quality improvement. Red dashed lines highlight intermediate outputs required for subsequent modules.

**Figure 8:** INNUca workflow

## 3.4. Fast preliminary clustering method based on oligonucleotide frequencies (GSCompare)

Alignment-free k-mer-based clustering methodologies are powerful tools for rapid reconstruction of phylogenetic relationships of draft bacterial genome sequences giving useful information for comparative and clinical genomic and molecular epidemiology applications (Bonham-Carter et al., 2014). These methodologies are particularly useful for a very fast and accurate species and subspecies confirmation. During the INNUENDO project an online service (http://gscompare.ehu.eus/) has been generated as support for users who want to perform species and subspecies classification, especially in cases of mixed samples. GScompare performs a comparison of a genomic signature (e.g. characteristic frequency of k-mer in a genome or sequence) of the sample of interest against the ones found for the Ensembl Genomes Release 38 including over 44,000 genomes from 8,244 bacterial species (http://ensemblgenomes.org/info/release-notes/38). This website has been optimized to compare octanucleotide composition of sequences by computing the Genomic Signature Distance (Campbell et al., 1999). More details on the methods and its application for *E. coli, S. enterica, Y. enterocolitica* and *C. jejuni* species determination and typing are available in Appendix C.

## 3.5. *In silico* prediction of pathotype, serotype, virulence and antibiotic resistance

To improve the response to public health events and facilitate communication between different stakeholders, WGS data need to be linked to traditional typing results (such as MLST, serotype, pathotype etc.). This action is important not only to place the isolates in a historical context, but also to predict their virulence potential.

### 3.5.1. *In silico* typing using read-mapping

While gene-by-gene methodology can provide a reference-free approach for comparing genomic content of multiple strains, approaches using reference mapping offer an efficient way to compare multiple samples without requiring assembly, thereby utilizing all information present in raw reads. Additionally, reference mapping can also assess variation in non-coding regions. In these approaches the reads are mapped against a reference sequence/genome and single nucleotide variations (SNVs) or small insertions and deletions (INDELs) are called throughout the genome for downstream phylogenetic analysis (Gardy et al., 2011). Mapping approaches can be used to achieve different goals. Instead of only quantifying the variability in the entire genome, such methods can efficiently assess the presence or absence of a given locus and possible genetic variations within these loci. This can be particularly important when the presence of a given gene or allele may indicate an important phenotype such as antimicrobial resistance, serotype or pathotype. Such data can be missing from the assembly due to, for example, contamination or intergenomic repeats. The bacterial ability to cause disease is strain dependent. It is therefore vital to quickly distinguish between pathogenic and non-pathogenic strains, especially in a context of a rapid response to food-borne outbreaks. Although several databases for bacterial virulence factors are accessible, no single computationally efficient method for fast *in silico* typing from raw reads is available. With this in mind, we developed a novel software, 'ReMatCh - REad MApping against Target sequences and consensus CHecking' and used it not only to quickly identify specific pathogenic *Y. enterocolitica* or diarrhoeagenic *E. coli* strains directly from raw Illumina reads. It also provides a framework for serotyping *E. coli* and for *stx* genes typing in STEC. ReMatCh provides users the ability to quickly and precisely query large collections of sequence reads for presence/absence and sequence variation of pre-specified target loci.

***ReMatCh: REad MApping against Target sequences and consensus CHecking***

ReMatCh was designed to map raw reads onto a set of reference sequences in order to determine the presence or absence of those sequences in a sample strain and to identify any variation compared to the reference. ReMatCh uses locally stored sequence data, but it can also directly interact with ENA or SRA databases, downloading the read files from sample/run accession numbers provided by the user,

or all data associated with a given taxon name. ReMatCh determines if a sequence is present or absent based on: 1) the proportion of reference sequence length covered by at least a pre-defined number of reads; and 2) the sequence similarity. ReMatCh relies on the strength of high read numbers to correctly identify two types of variants: SNVs and short INDELs. When a position does not meet the criteria for being unambiguously called, ReMatCh will designate it as a potential heterozygous position. ReMatCh software dependencies are: Bowtie2 (Langmead and Salzberg, 2012) for read mapping, Samtools (Li et al., 2009) for SAM/BAM manipulation and variant calling and Bcftools (Li, 2011) or consensus sequence production. Besides the parallelization implemented within Bowtie2 and Samtools, ReMatCh assigns one sequence variant analysis and coverage determination to each available thread. Details concerning the implementation of ReMatCh for pathotyping and serotyping of *E. coli* and *Y. enterocolitica* are available in Appendix D.

***Use in QC measures for assembly-free in silico typing***

To avoid jeopardizing the *in silico* typing results by low quality sequences, we implemented dedicated QC measures for these modules. We saw that low sequencing depth represented the main obstacle against correct identification of target genes using raw sequencing read mapping approaches. Therefore, the reads are controlled using the *TrueCoverage* module in the INNUca pipeline prior to *in silico* typing (see Section 3.3). Reads will be considered suitable for downstream procedures only if they will fulfill the following QC measures:

- maximum number of missing genes = 1 or 2 (for *Y. enterocolitica* and *E. coli*, respectively)
- maximum number genes with multiple alleles (with heterozygous positions) = 1 or 2 (for *Y. enterocolitica* and *E. coli*, respectively)
- minimum read coverage = 25

### 3.5.2. Assembly based *In silico* typing

In addition to the assembly-free *in silico* typing, the INNUENDO Platform V 1.0 offers the user a set of tools for *in silico* typing directly from assembly. MLST is calculated for all four species using mlst2.10 (https://github.com/tseemann/mlst). Serotype of *S. enterica* is performed using SISTR (https://lfz.corefacility.ca/sistr-app/). Annotation of the assembly in terms of presence of antibiotic resistance genes, virulence factors and plasmids is performed using the ABRicate software (https://github.com/tseemann/abricate). The ABRicate software use BLAStn to search for matches in the Resfinder, CARD, VFDB and PlasmidFinder databases (Lihong et al., 2005; Zankari et al., 2012; McArthur et al., 2013; Carattoli et al.; 2014). ABRicate produces a tab-separated output file including the following information: product, %identity and %coverage. The INNUENDO Platform V1.0 includes the raw output file of ABRicate without applying any thresholds regarding minimum coverage or identity for defining a gene present in the report, leaving to the user complete autonomy in interpreting the results and in using them in the analyses.

## 3.6. The phylogenetic framework

### 3.6.1. chewBBACA: a new suite for gene-by-gene methodology

A critical point for implementing global pathogen surveillance using WGS is the translation of WGS sequence data in 'plain language' (i.e. sub-types). Named subtypes enable rapid data-analysis and efficient exchange of information, thereby contributing to a rapid response to infectious disease, promoting disease prevention and control. Defining such WGS-subtypes is not trivial and requires the use of standardized methodological approaches and a nomenclature to describe the relationship between isolates. The phylogenetic framework in the INNUENDO Platform V1.0 makes use of the gene-by-gene (GbG) approach (Maiden et al., 2013). This method compares genomes (complete or draft) against a predefined set of loci collected in a schema composed by all possible known variation of those loci (Maiden et al., 2013). If the schema consists of core loci, i.e. loci present in all (100%) or the great majority of the bacterial population (e.g. >95%), the schema is referred to as a core

genome MLST schema (cgMLST). Alternatively, the schema can include loci which are part of the accessory genome, i.e. present in only a fraction (<95%) of the strain population. A schema which includes loci from both core and accessory genomes can either be called pangenome MLST (pgMLST) or whole genome MLST (wgMLST). The definitions of cgMLST, and pgMLST or wgMLST are inherently changing due to the natural evolution of a bacterial species and are operational in nature since they are based on the number of isolates analyzed to date. GbG methodology for subtyping has great appeal due to its portable nomenclature and independence from a reference strain. As such, PulseNet International adopted this approach for WGS-typing of food-borne pathogens (Nadon et al., 2017).

A number of software packages are available for GbG allele calling. Among them there are two commercial software packages and six open-source platforms (Table 5).

**Table 5:** Software packages for gene-by-gene allele calling

| Software | C/OS | Link | Reference |
|---|---|---|---|
| Ridom SeqSphere+ | C | http://ridom.de/seqsphere/ | - |
| BioNumerics | C | http://www.applied-maths.com/applications/wgmlst | - |
| BIGSdb | OS | https://pubmlst.org/ | Jolley and Maiden (2010) |
| MIST | OS | https://bitbucket.org/peterk87/microbialinsilicotyper | Kruczkiewicz et al. (2013) |
| GeP | OS | https://sourceforge.net/projects/genomeprofiler/ | Zhang et al. (2015) |
| FastGeP | OS | https://github.com/jizhang-nz/fast-GeP | Zhang et al. (2018) |
| EnteroBase | OS | https://enterobase.warwick.ac.uk/ | Alikhan et al. (2018) |
| MentaLiST | OS | https://github.com/WGS-TB/MentaLiST | Feijao et al. (2018) |

C = commercial software; OS = Open source software.

A drawback of the commercial platforms is the lack of description of their allele-calling algorithms, which jeopardizes the transparency and flexibility of the software. BIGSdb and EnteroBase do not offer a stand-alone version of its allele calling algorithm, and therefore requires the user to submit their reads to either the EnteroBase website or another public repository and is such dependent of the website infrastructure and computational resource. GeP and its faster version FastGep are a stand-alone GbG allele-calling algorithm, however they are unsuitable for large-scale analyses as they run solely on a single CPU core. MIST is designed for *in silico* prediction of genomic information and relies on existing schema. MentaLiST performs allele calling directly from reads, but relies on existing schemas and allele definitions. To our knowledge, no GbG allele calling algorithm offer schema creation, modification and validation of their wgMLST and cgMLST schema.

We therefore developed chewBBACA (comprehensive and highly efficient workflow for a Blast Score Ratio Based Allele Calling Algorithm) (Silva et al., 2018), a gene-by-gene typing schema offering an open-source, freely available computational solution for the creation, evaluation and use of wgMLST and cgMLST schemas. The chewBBACA is a suite written in python3 and it is the first algorithm to provide integrated schema creation and validation tools, thereby allowing the user to develop wg/cgMLST schemes for any bacterial species from a set of genomes of interest. In addition, the allele definition in chewBBACA is unique: only alleles that correspond to potential coding sequence (CDS) are identified as alleles. This definition offers potential insight into the genetic and phenotypic variability observed, for instance the identification of potential mechanisms underlying the ecological success or the virulence potential of particular clones. In addition, a subset consisting of the most distinct alleles are used as reference for allele calling instead of a single sequence for each locus. This secures that even fast evolving loci will be identified. Furthermore, chewBBACA is as scalable as the INNUENDO platform, executable in everything from high-end Unix-based laptop to HPC, facilitating its adoption into large-scale automated analysis pipelines. chewBBACA has a specific way to annotate loci and alleles which allows automatic curation of the schema used. More details are available at https://github.com/B-UMMI/chewBBACA/wiki.

The ability of chewBBACA to run locally removes the need for uploading raw data to central repositories or web services and offers therefore a beneficial independence from third party servers. Data protection policies, ethical or legal concerns would not hinder the use of chewBBACA, making it a

suitable option for public health and food safety authorities. Also, chewBBACA can use any cgMLST or wgMLST schema according to the user's preference, as long as each locus is a CDS, being a suitable tool for target-based *in silico* phenotypic prediction. Noteworthy, chewBBACA can therefore use and perform allele calls on BIGSdb and EnteroBase cgMLST/wgMLST schemas, as the great majority of alleles code for CDSs.

### 3.6.2. Species specific wgMLST and cgMLST schemas in the INNUENDO Platform

The INNUENDO Platform V1.0 contains curated wgMLST schema for four different pathogens (*S. enterica*, *E. coli*, *Y. enterocolitica* and *C. jejuni*). The chewBBACA suite was used for validating all schemas. If the original wgMLST schema was obtained from a third party (i.e. EnteroBase for *S. enterica* and *E. coli*), loci were initially curated using *AutoAlleleCDSCuration* for removing non-CDS alleles. The *de novo* schemas were based either on pangenome analysis defined by *Roary* (Page et al., 2015) with default setting (i.e. *Campylobacter*) or using *SchemaCreation* function of chewBBACA (i.e. *Y. enterocolitica*). For all schemas, the quality of the loci was assessed using *SchemaEvaluation*, wherein loci with single alleles and high length variability (i.e. more than one allele outside the mode +/- 0.05 size) have been removed. The schema was further curated by excluding "Repeated Loci" and loci annotated as "non-informative paralogous hit (NIPH/ NIPHEM)" or "Allele Larger/ Smaller than length mode (ALM/ ASM)" by the *AlleleCalling* engine present in more than 1% of the respective genomes dataset (details on chewBBACA allele calling https://github.com/B-UMMI/chewBBACA/wiki/2.-Allele-Calling).

Finally, the set of loci defining the cgMLST schema have been extracted. We defined as static cgMLST schema the set of loci present in ≥ 99% of the samples contained in the INNUENDO Legacy Dataset, allowing up to 2% missing loci per single genome. A higher cut-off was set for *C. jejuni*: loci present ≥ 99.9% of the samples. This higher cut off was needed for avoiding the exclusion of too many genomes which did not satisfy the 2% missing loci limit.

For details and rationales on the schema creation and validation please visit the Github page at https://github.com/TheInnuendoProject/chewBBACA_schemas. Schemas are deposited in Zenodo (Rossi et al., 2018a, b, c, d).

### 3.6.3. Dynamic shared-genome based approach

As described in Section 2.3.2, the type of schema used for a gene-by-gene approach has important implications whether it is to be applied in long-term surveillance or outbreak investigation. We therefore developed an interactive and innovative multi-step way to cluster strains, based on higher and lower resolution of genomic diversity relative to the need, achievable in one single operation. We here present these possibilities implemented within the INNUENDO Platform V1.0 generated by the GbG analysis using chewBBACA as allele calling algorithm and the schemas described in Section 3.5.2.

This advantageous approach, named dynamic shared-genome based approach, limits allele calling to a single curated wgMLST schema, and two separate levels of analysis can be extracted for their intended use in both long-term surveillance and outbreak investigations. For the first analysis (called *Classification*), allelic designation for the static cgMLST is extracted (see Section 3.5.2) and the sample is assigned a specific type based on a defined strain nomenclature. At this level, up to 2% of missing loci are allowed and based on the number of missing loci quality control measures have been established to define the quality of the allele calling for the second analysis. The use of a harmonized nomenclature (see Section 3.6.3.1) at this level secures the communication between laboratories and other stakeholders. The proposed nomenclature is adjustable to all needs from outbreak investigation to pathogen surveillance. The second analysis consists in the actual cluster investigation. Firstly, the INNUENDO Platform V1.0 will automatic search the most similar samples (k-closest) in the database among the one under analysis, then it will send the complete set of wgMLST allelic profiles to a novel implementation of PHYLOVIZ Online 2.0 which will construct minimum spanning tree (MST) based on

100% of shared loci between the selected strains. By directly interacting with the tree, the user can then increase resolution of the analysis (i.e. increase the number of shared loci under evaluation) by re-calculating the MST for a restricted set of strains. This approach allows the user to interactively increase resolution, which might be relevant for discriminating cases during outbreak investigation.

The process of clustering is automatized as much as possible, but still requires the user to make conscious choices based on the strains and pathogen in question and the local epidemiology, the chewBACCA quality control step and epidemiological data available. Cluster analysis needs an active participation from several actors (e.g. genomic specialists, bacteriologists specialized in the species of interest, epidemiologists, practitioners) to achieve biological relevant results. Therefore, one of the most important aspects in successful cluster identification remains a timely and effective communication between stakeholders.

#### 3.6.3.1. The three levels of strain nomenclature within the INNUENDO Platform V1.0

As the two epidemiological settings of surveillance and outbreak investigations have different goals, they have different needs for resolution. Within the INNUENDO Platform V1.0, three different levels of strain nomenclature have been specified: L1 for outbreak detection and investigation, L2 for longitudinal surveillance and L3 for congruence to other relevant subtyping (MLST). This classification system is hierarchical (i.e. L3 $\in$ L2 $\in$ L1) and based on goeBURST clustering methodology (Francisco et al., 2009) (Table 6). Strains are labeled in the INNUENDO Platform V1.0 as L1:L2:L3, and new types are added in the database after discovery.

**Table 6:** Strain nomenclature and allele calling quality matrices

| Species | wgMLST loci[(a)] | core loci[(b)] | L1 (%)[(c)] | L2 (%)[(c)] | L3 (%)[(c)] | QC[(d)] PASS | QC[(e)] WARNING | QC[(f)] FAIL |
|---|---|---|---|---|---|---|---|---|
| *E. coli* | 7601 | 2360 | 8 (0.34) | 112 (4.7) | 793 (33.6) | ≤ 8 | >8 ≤ 47 | >47 |
| *S. enterica* | 8558 | 3255 | 14 (0.43) | 338 (10.4) | 997 (30.6) | ≤ 14 | > 14 ≤ 65 | >65 |
| *Y. enterocolitica* | 6344 | 2406 | 9 (0.37) | 133 (5.5) | 1189 (49.4) | ≤ 9 | > 9 ≤ 48 | >48 |
| *C. jejuni* | 2795 | 678 | 4 (0.59) | 59 (8.7) | 292 (43.1) | ≤ 4 | > 4 ≤ 13 | >13 |

(a): number of loci included in the wgMLST schema;
(b): number of core loci on which the nomenclature have been designed;
(c): three different levels of strain nomenclature defined based on the core loci: L1 for outbreak detection and investigation, L2 for longitudinal surveillance and L3 for congruence to classical 7 genes MLST; between brackets the corresponding percentage of core loci;
(d): upper limit of the number of allowed missing loci for passing quality check of the allele calling; see Section 3.6.3.2;
(e): range of number of allowed missing loci for passing quality check of the allele calling with a warning message; see Section 3.6.3.2;
(f): any genomes showing > 2% of missing loci will fail the allele calling quality check; the column indicates for each species the number of the missing loci corresponding to 2%; see Section 3.6.3.2.

The nomenclature representing the highest resolution, L1, was set by investigating the concordance between genomic clustering at different thresholds of similarity in sets of epidemiologically verified outbreak isolates (i.e. cluster efficiency): four *E. coli* outbreaks, seven *S. enterica* serovar Enteritidis outbreaks, four *C. jejuni* outbreaks and three *Y. enterocolitica* outbreaks. We discovered that a similarity threshold of 0.3-0.6% of allele differences between strains subtyped with static cgMLST is concordant with the epidemiological information in all the four species of interest.

To define the L2 nomenclature we implemented the methodology called Neighborhood Adjusted Wallace Coefficient (nAWC) (Barker et al., 2018). Briefly, goeBURST (Francisco et al., 2009) was used to examine cluster membership for cgMLST profiles of the four species through a continuous range of similarity thresholds and nAWC was calculated to assess cluster consolidation dynamics. This method robustly sets similarity thresholds that generate quasi-stable clusters. We found that allele differences in a range of 4.7-10.4% of cgMLST schemas were the lowest threshold producing stable clusters for the four pathogens of interest.

The L3 was defined as the cgMLST goeBURST threshold with higher concordance with MLST definition using Adjusted Wallace Coefficient (AWC) as described in Carriço et al. (2006). The partitions produced by goeBURST were compared with those produced by MLST.

Details on how strain nomenclature has been defined are available in Appendix E.

#### 3.6.3.2. Quality control of the allele calling

Up to 2% missing cgMLST loci is allowed in a genome for the sample to be included in the platform dataset and assigned a three-letter nomenclature code. There is three reasons for this 2% limit for missing loci: 1) genomes with more than 2% missing loci have a tendency to create MST with low numbers of shared-loci; 2) genomes with more than 2% missing loci often fail INNUca QC or pass with severe warnings, typically due to an excess of contigs, and 3) a disproportionate number of missing cgMLST loci might be a sign that the sample actually is another species or subspecies. Therefore, genomes with > 2% missing loci are considered to be of too low quality for the INNUENDO Platform V1.0 and are marked as "FAIL" in the allele calling QC and will not be added to the database.

The total number of missing cgMLST loci is designated by a label from the allele calling QC; samples labeled "PASS" contain less missing cgMLST loci than the similarity threshold of L1, while samples labeled "WARNING" contain missing cgMLST loci between the L1 similarity threshold and the 2% missing cgMLST loci limit for inclusion (Table 6). This simple QC designation enhances communication by quickly directing the user to misclassified samples, which is especially beneficial during outbreak identification. However, even for samples labeled "PASS", missing loci might cause trouble when assigning the sample to a specific type, possibly affecting a higher order classification.

The effect of missing loci on overall cgMLST allelic profile divergence is a critical assessment for those samples passing the allele calling QC (either with "PASS" or "WARNING").

Assuming a scenario where a missing loci is defined as either the allele with 50% probability in the population or other allele (or alleles) summing up to 50% frequency in the population, the probability of a sample $x$ to be correctly classified in a type $n$ can be calculated with the formula (1):

$$(1) \qquad P(x,n) = \frac{\sum_{i=0}^{t-d+1} \binom{m}{i}}{2^m}$$

where $t$ is the threshold for the classification at $n$, $d$ is the lowest observed number of allelic differences with any sample of cluster $n$ and $m$ is the number of missing cgMLST loci. For example, assuming a cut-off for defining $n$ as $t$=136 and a sample $x$ with $d$=122 and $m$=15, the probability of $x$ to be typed as $n$ is 99.63%. In case $d$=127 but the same missing loci $m$=15, the probability to correctly assign the strain $x$ to $n$ drops to 69.3%.

Due to the fact that the frequency of alleles in the population for a given loci is most likely not as assumed above, formula (1) is probably overestimating the probability of $x$ for $n$. Nevertheless, this formula is a good approximation of the true probability for $d + m \to t$ (since $P$ drops quickly for $d + m \gg t$).

#### 3.6.3.3. Mathematical framework for sample classification and querying databases

Each time a new profile is generated by chewBBACA, the INNUENDO Platform V1.0 performs the *Classification* step (i.e. the process for giving to new sample the L3:L2:L1 type) by extracting the cgMLST profile of the sample from the wgMLST schema and by identifying the closest samples in the database based on an algorithm described in Carriço et al. (2018) and briefly described below.

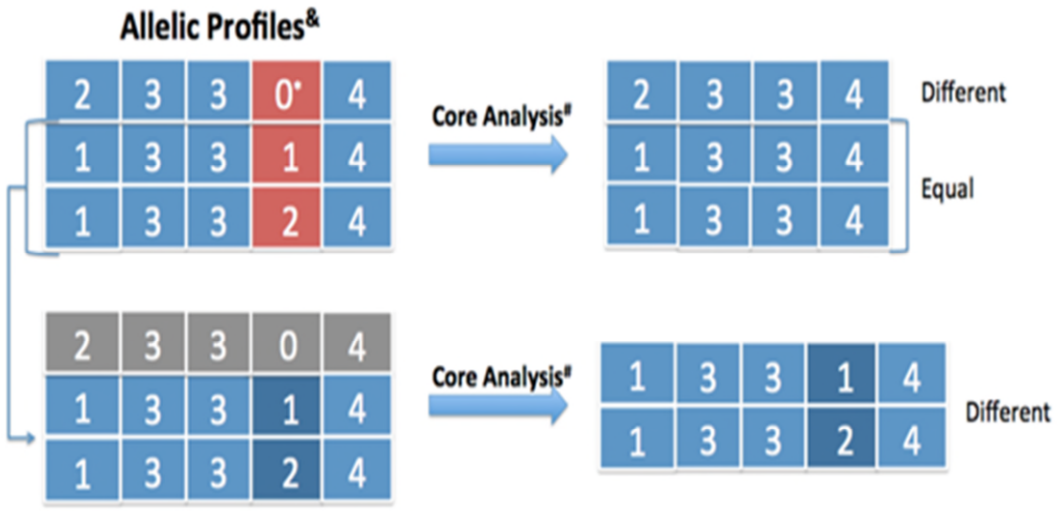
Assuming a set of profiles $P$ with a given number of loci (i.e. the schema in the INNUENDO platform) the algorithm finds all profiles $v \in P$ that are at Hamming distance (i.e. the number alleles at which two profiles differ) at most $k$ alleles differences from a query profile $u$, in short $H(u,v) \leq k$.

The INNUENDO Platform V1.0 uses the cgMLST index file during the classification step to perform the search and pre-calculate the goeBURST clustering of the database based on the defined nomenclature thresholds at all three levels. If the method returns at least one match, it classifies the new profile with the classification of the closest. If not, a new classification is assigned. A new entry is then added to the INNUENDO database as well as to the cgMLST and wgMLST profiles files and the index files are updated.

The *Classification* step is performed automatically after any allele calling and the results are available in the report page of the Platform. Based on the result of the *Classification* step (in addition to all the other information available in the platform for each sample), users can then select the strains pertinent for the cluster analysis. Since the analysis aims to identify clonal relationship between strains within the databases and the user-selected samples, only the closest strains of the database are relevant. To define the input data for visualization methods according to a defined number of differences on closest strains, the platform applied the algorithm as described in Carriço et al. (2018) to identify the so called *k-closest*. The method searches for the most similar strains while considering the most differences possible among all wgMLST loci for each profile used as input for the search (i.e. the samples selected by the users to be sent to PHYLOViZ Online 2.0). Duplicate matches can occur between the profiles used for each search. Therefore, the final file used as input for the visualization methods is the intersection of the results of the *k-closest* profiles between each input strain. The set of strains identifiers are then used to query the INNUENDO database to get the profiles and ancillary data to be used in the phylogenetic analysis.

#### 3.6.3.4. Interactive dynamic core genome analysis in PHYLOViZ Online 2.0

After selecting the k-closest, the user is directed to PHYLOViZ Online 2.0 to perform the actual cluster analysis with the goeBURST algorithm. To reduce the impact of missing data and its influence in strain clustering when analyzing closely related strains, we developed a new approach to dynamically increase the discriminatory power of the comparison between profiles obtained with curated wgMLST schema. From the uploaded wgMLST profile, the application constructs a profile based on shared loci for the entire set of isolates (Figure 9). Therefore, depending of the total allelic differences among the selected strains and the number of k-closest searched, the first MST showed in PHYLOViZ Online 2.0 might be based on very different set of shared-loci. The application then allows interactive selection of subsets of interest (i.e. the suspected outbreak cluster) that can be automatically reanalysed by constructing a tree from a new wgMLST profile that maximizes the shared loci in that particular subset (Figure 10). This process can then be repeated for further discrimination of a novel subset. Since closely related isolates are expected to share an expanded set of loci when comparing to more distantly related ones, this approach reduces the impact of missing data when analyzing closely related isolates, allowing the user to make the most of the available data. A demonstration video is available at goo.gl/t5q6HF.




0* = missing data identifier. Core Analysis = only shared loci are used; loci with missing data are removed from the analysis before constructing the tree. Allelic Profiles = wg MLST allelic profile.

**Figure 9:** Example of effect of missing data on the classification of strains during the wg/cgMLST analysis

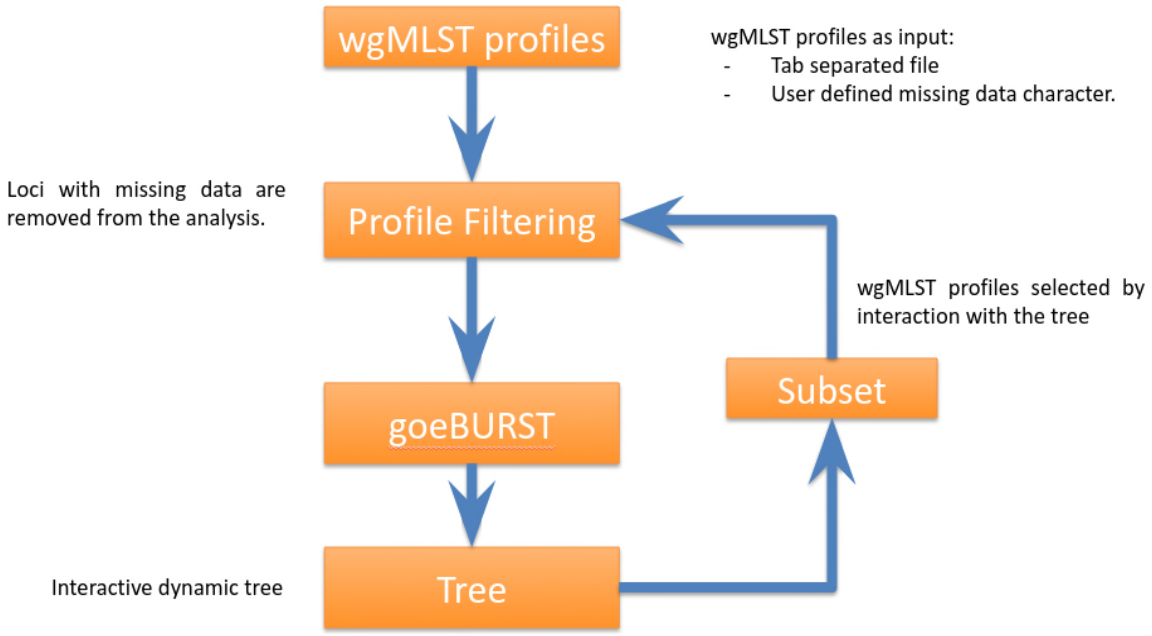



**Figure 10:** Flowchart of the interactive dynamic shared-genome analysis using PHYLOViZ Online v2 (beta version)

### 3.6.4. General guidelines for cluster analysis using the dynamic core-genome analysis

There is no universal cut-off for identifying epidemiological significant clusters applying wgMLST based analysis, being significantly dependent on the species of interest and, within species, population lineages. Moreover, the epidemiology of a specific food-borne disease varies between regions and countries, resulting in different genomic diversity of circulating strains over space and time, affecting the limits for defining genomic clusters. The operator's know-how of the pathogen in question and

available epidemiological data influence the value of the cut-off for genomic clustering. We therefore developed working procedures on how to perform dynamic core genome analysis in PHYLOViZ Online 2.0 for cluster investigation. As a general guideline, the identification of relevant genomic clustering starts with defining goeBURST groups at 0.5 to 1% allelic differences based on the profile calculated by PHYLOViZ Online 2.0 after searching for the *k-closest*. Mapping the strains metadata and cgMLST nomenclature on the tree can verify the cluster composition and aid in deciding the correct percentage of allelic differences. The user then proceeds to manually investigate the clusters by increase resolution of the analysis selecting a subset of strains of interest, as depicted in Figure 10. This can help the user resolving ambiguous positions of strains, for instance if they cluster together in a way that is not supported by the epidemiological information. The operator can refine the cluster thresholds on the tree and/or investigate pairwise allelic distance in detail to identify possible outliers. All possible information available for each sample must be taken in consideration for validating any clusters. In addition to epidemiological information, the gene content concerning resistance and virulence gene pull, the presence of plasmids, patho- and serotyping as well as any phenotypic data should and could be taken in consideration during the analysis, and the platform allows the user to interactively define the granularity of the information needed visualized in PHYLOViZ Online 2.0 to facilitate cluster identification.

## 3.7. Reporting and communication within the INNUENDO platform

Correct reporting of laboratory results is important for efficient communication between different stakeholders partaking in outbreak detection. The type of the report should take in consideration the different sensibilities and the perceptions of all the actors involved in the process, and, especially, should guarantee a unique and clear interpretation of the genomic analysis for both microbiologists and epidemiologists. In addition, the e-technician/microbiologist directly involved in the analysis should have the opportunity to delve into the outcomes of the analyses and the QC assessments produce by the Platform. The INNUENDO Platform V1.0 has a dedicated web application, *Reports*, which was implemented to interactively explore the results from both assembly-free (i.e. *in silico* typing) and assembly-based analysis (i.e. allele calling, cgMLST classification, and annotation of resistance and virulence genes) and the QC measurements from reads, assembly and allele call. User can select strains from different projects to construct a report, which can be saved on the platform or downloaded locally as JSON file (Figure 11). The JSON file can be reloaded for further visualizations.

The report is divided in *Components*, which contains all the results of the analyses and the QC assessments, genomic typing and *Trees* (which contain the links of all the users PHYLOViZ MSTs). The users can here explore: strain metadata; *in silico* typing results (e.g. for *E. coli* Serotyping, Pathotyping and stx-subtyping); assembly status, statistics and QC; annotation of the genomes and presence of resistance or virulence associated genes or plasmid; chewBBACA results, statistics and QC. The user can also delve into interactive charts and graphs base quality, sequence quality, GC content, sequence length, coverage, base N distribution and annotation collectively trough aggregated reporting tool or for each sample individually (Figures 12 and 13). The web application reports the QC measurements using colour and name codes. Assembly quality is reported with colours such as light green when warnings are not issued, yellow when moderate and high severity warnings are issued, and red if the assembly quality failed. By clicking on the button user can explore the reason of the warning/failure.

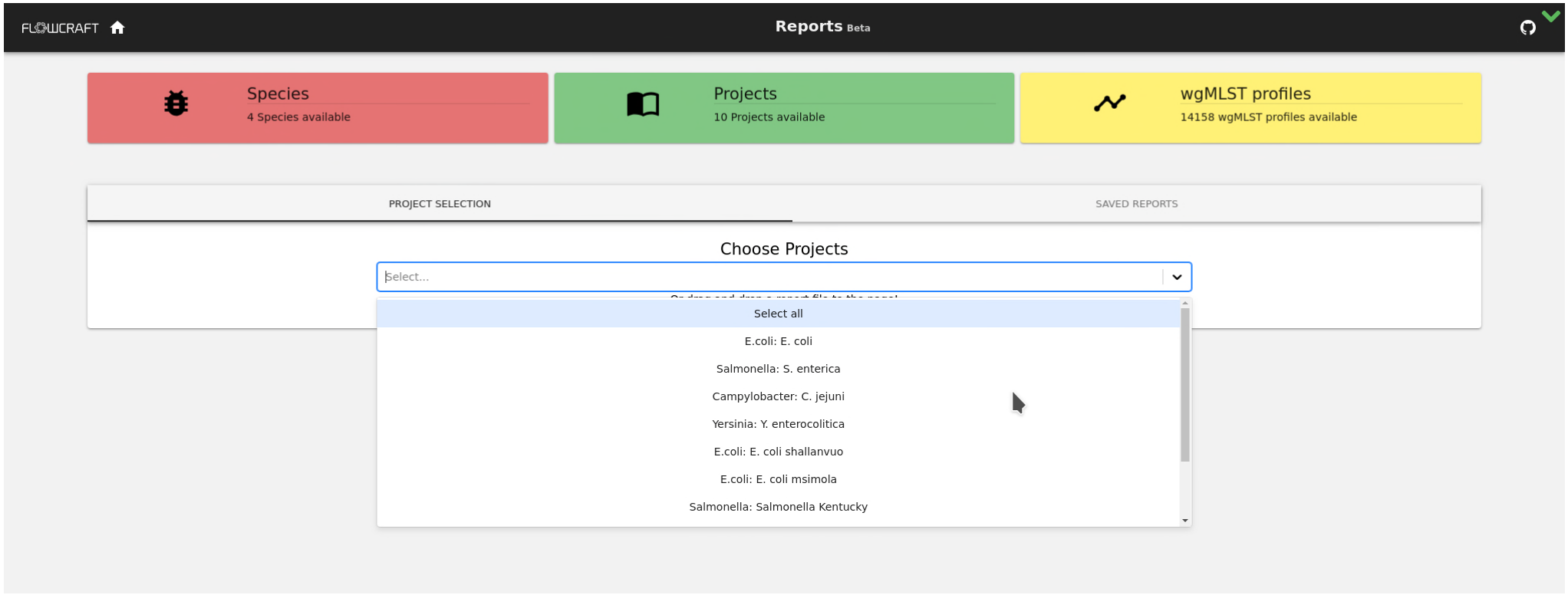



**Figure 11:** Front-page of the Reports web application, based on the FlowCraft reports page

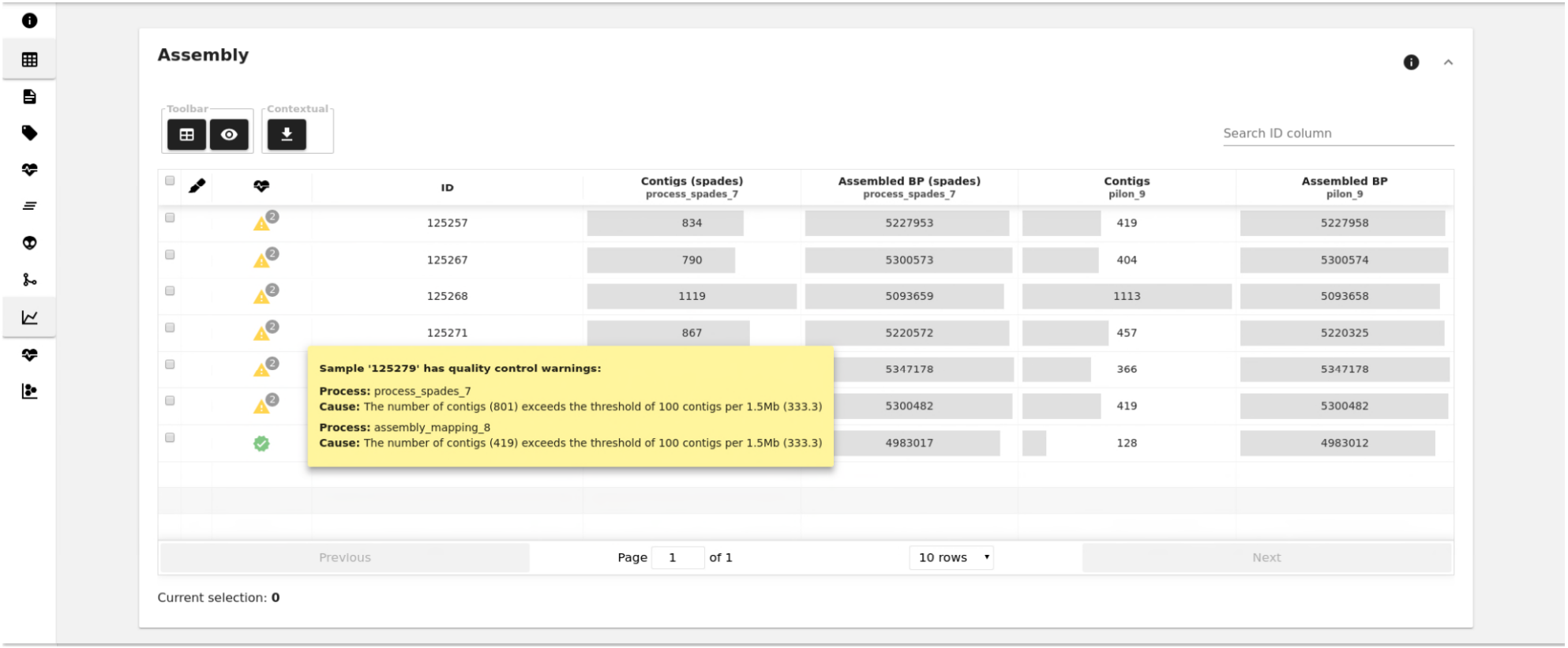



**Figure 12:** Screenshots of the INNUENDO Platform V1.0 Reports page, showing the assembly results table including QC codes

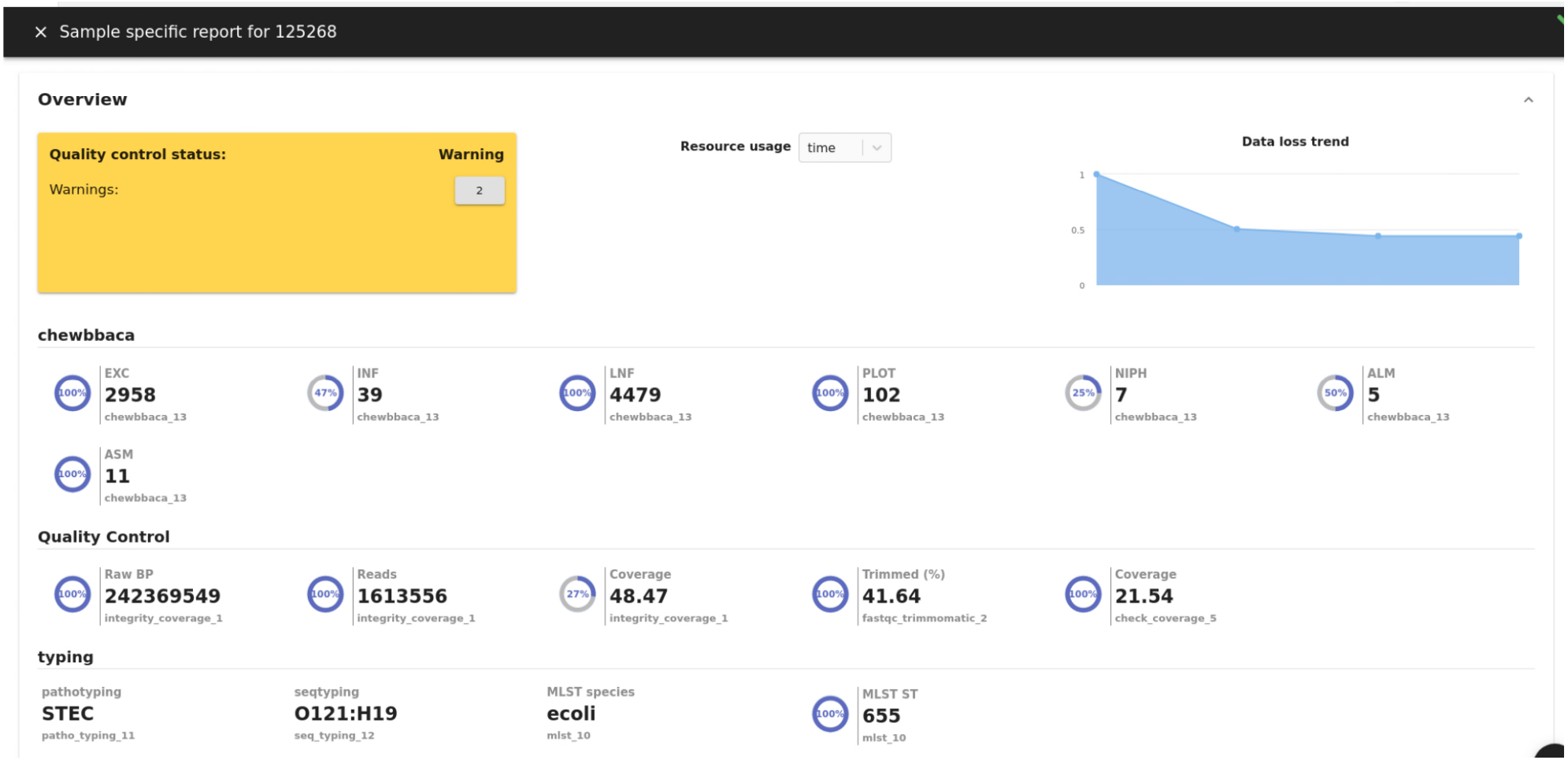



**Figure 13:** Screenshots of the INNUENDO Platform V1.0 Reports page, showing a single sample summary

The allele calling QC (defined as *chewBBACA status*) are reported as "pass" (green square), "warning" (yellow square) and "fail" (red square) depending of the percentage of missing loci (which can be visualize by moving the cursor on top of the square). From the Reports web application, users can select the strains for which the cluster analysis is needed and send to PHYLOViZ online 2.0 along with the *k-closest* from the database and a defined set of metadata. The unique URL of the tree is then stored in *Trees* session of the application and can be shared with other users.

Communication between different users of the Platform has important implications especially when the Platform is shared between different authorities (such as Public Health and Food and Veterinary authorities). The INNUENDO Platform V1.0 contains an internal messaging system aimed to simplify and standardize communication between users during outbreak investigation. The submitter saves reports and sends internal notifications of the results from the web application using specific templates (Figure 14). Also, an alert system to notify users on clustering is currently being implemented, both when the user is doing the analysis, and when the study was launched by another user somewhere else.

Therefore, three different communication tools are, or will be, in use in the INNUENDO platform (Figure 15):

- Internal messaging system for free text between users
- Sending automatic generated reports to other users/owners of strains
- Alert system for notification of clustering upon launching the INNUENDO platform

## 3.8. Keeping the database relevant and automatic upload to public repositories

Public repositories are constantly growing with several thousand submissions a week for certain relevant pathogens (such as the ones included in the INNUENDO Platform V1.0). Therefore, there is the need to have a mechanism to keep the INNUENDO Platform V1.0 strain database relevant with the addition of the most recent submissions. For this scope the INNUENDO Platform v. 1.0 also contains a module allowing the user to improve the available dataset through the download of selected *fastq* files deposited in SRA or ENA for subsequent analysis in the platform.

Since global response to food-borne diseases requires (quasi)real-time sharing of data (especially HTS data), the possibility for the user to submit the raw data to public repositories is of equal importance. Regardless the difficulties experienced up to now, the API of EMBL-EBI ENA repository is stabilizing, allowing the developing of a procedure for the automatic submission of raw *fastq* files, simplifying the process. This module is not implemented in the platform yet, but user can utilize available easy software tools such as ena_submission tool from Public Health England (https://github.com/phe-bioinformatics/ena_submission).

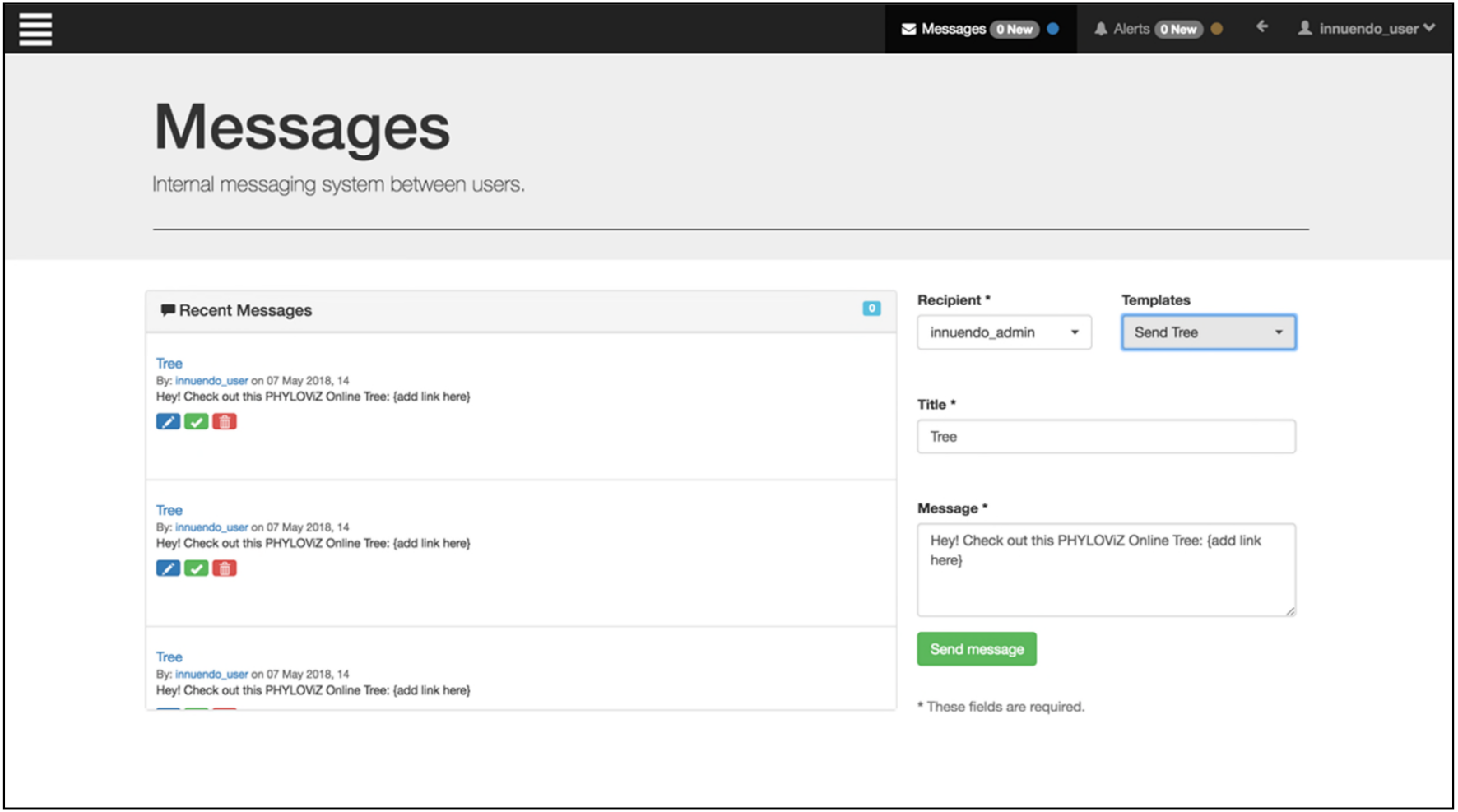



**Figure 14:** Screenshots of the INNUENDO Platform V 1.0.0 messaging system page

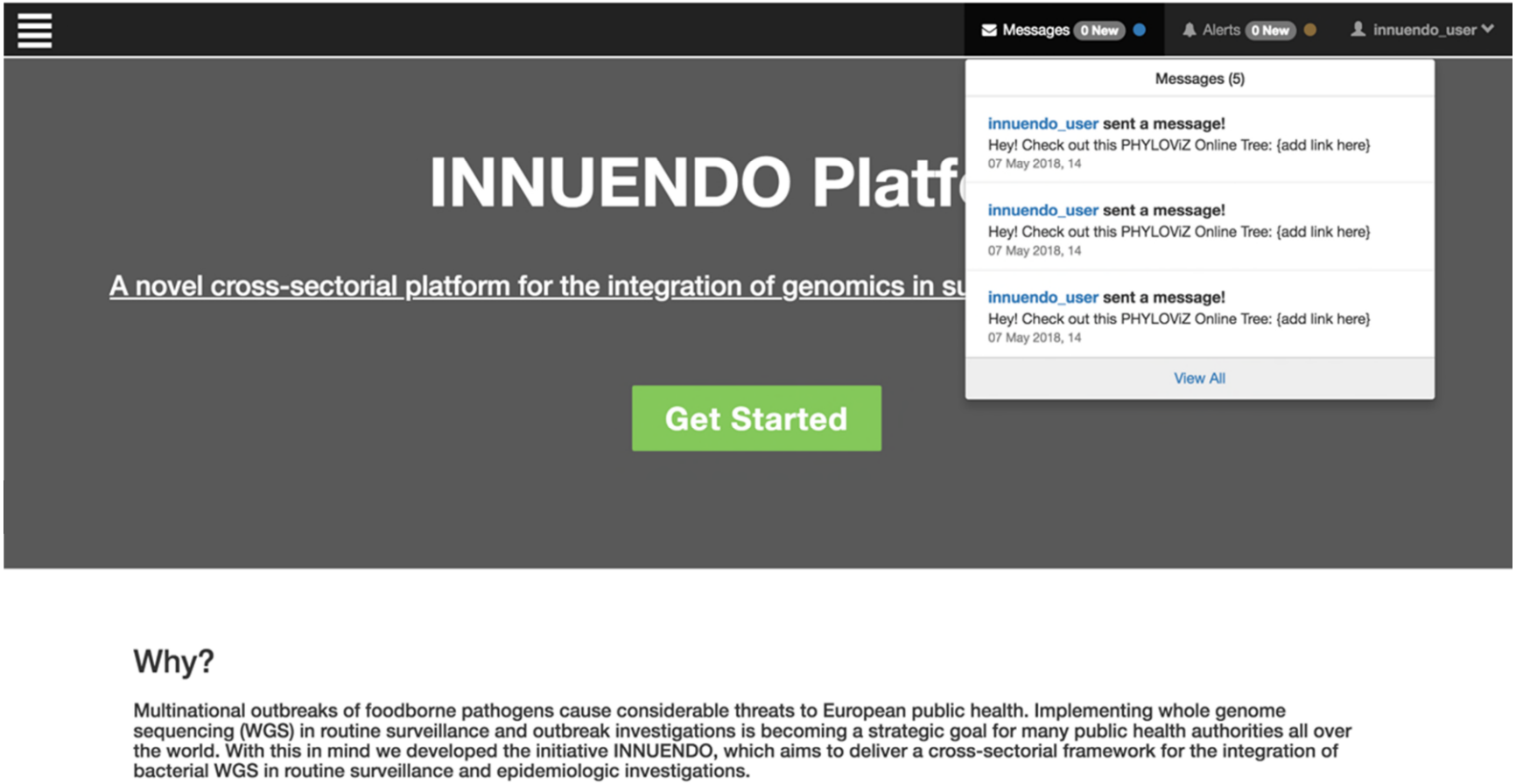



**Figure 15:** Screenshots of the INNUENDO Platform V 1.0 starting page with messaging system

# 4. Implementation of the INNUENDO platform V 1.0

## 4.1. Different possibilities for data sharing

One of the challenges to implement the INNUENDO Platform V1.0 is to define how data can be shared between users and platforms.

Four different scenarios were considered:

1. Multiple platforms closed for each user
2. Shared platform with different access levels for accessing metadata or analysis results
3. Shared platform with shared analysis and minimum metadata shared
4. Multiple platforms with the ability to share data between them

The first scenario is similar to already available software such as Applied Maths Bionumerics™ or Ridom Seqsphere+™. In this case, the platform is contained on the user's own machines, and sharing raw data and/or associated metadata must be done actively on request to become available to other users. For this scenario to run on a laptop, the INNUENDO Platform V 1.0 can be installed as a Docker compose instance, which builds all the required software components automatically and retrieves the required data, facilitating its installation and its use.

The second scenario requires the definition and configuration of different user-types for accessing metadata and results. Usually such systems are cumbersome to make, as different levels of authorization need to be available for each datum in the system. Only a well-defined ontology with defined user roles with different authorization rights to the epidemiological data can make such a scenario possible. Future collaborative works with the Genomic Epidemiology Ontology (GeneEpiO - http://genepio.org/) developers of the IRIDA consortium (www.irida.ca) will be done to implement metadata fields annotated with GeneEPiO ontology in the INNUENDO platform, resulting in epidemiological metadata descriptions and definitions of classes shareable at different authorization levels.

The third scenario is the prototype implemented in the INNUENDO Platform V 1.0 currently in use by the public health and food authorities to survey food-borne pathogens in Finland. Both agencies share one platform where common analysis results and a minimum level of strain metadata are accessible for all users. A common incremental database is generated in the platform for each bacterial species, accessible for all users upon query of the platform for similar strains to the ones being analyzed using chewBBACA and PHYLOViZ Online 2.0. This has the benefit of quick detection outbreaks and stimulates cross-sectorial communication between users of both agencies.

Scenarios 2) and 3) are centralized, meaning that the platform scale-up in terms of number of simultaneous users is limited to the available computing facilities and storage space. Centralized systems also require high maintenance since a single event such as network failure or server breakdown will automatically stop the system until repaired, impairing all users of the system.

The fourth scenario would present the ideal situation: a decentralized network of INNUENDO platforms running the same bioinformatics pipelines with the ability to share and query data between any platform in a peer-to-peer communication protocol. This is the ideal scenario for future versions of the INNUENDO Platform.

## 4.2. Storage and computational requirements

The HTS analysis of microbial strains for genomic epidemiology is a highly computational demanding process, and therefore costly in terms of required software, trained bioinformatics personnel and computational infrastructure.

The overall analysis of one strain from raw data (i.e. compressed fastq files with reads) to identification of profiles takes about 20 to 40 minutes in a high-end laptop. However, scaling up to

analyze hundreds to thousands of strains at the same time requires a significant higher computing capacity to finish in a timely manner. Since public databases increase in size (deposited strains) and global data sharing between health agencies is essential for efficient tracing and resolution of outbreaks, the system that aims to analysis these ever-increasing datasets should be scalable and elastic. System scalability is required at two levels: data storage and computational power for data processing. System elasticity, which in this case, is borrowed from "Elastic computing concepts", refers to the ability for the same system to run in different computational setups: from servers to HPC or even cloud-based systems. In the latter, "elastic computing" is to provide computational resources on demand, scaling up or down the systems processing capability to achieve effective management of computational costs.

The INNUENDO Platform was developed with these principles in mind and is adaptable to different realities of computing resources in various research laboratories, reference laboratories and health agencies. The current version of the platform (V 1.0) can be installed in a high-end server, HPC or in a Cloud-Based environment such as OpenStack (https://www.openstack.org/).

### 4.2.1. Storage

The computational needs of INNUENDO include storage of raw reads (i.e. reads as obtained from an Illumina™ sequencer), storage capacity needed for processing raw reads and the computational requirements for the different software used in the analysis.

#### 4.2.1.1. Storing raw sequences

Several options for storage of raw data are possible. Most of the raw data is only analyzed once through the complete INNUENDO pipeline, so raw data should be removed and stored on the outside the platform servers to save storage space required for novel data once the pipeline is done. The INNUENDO Platform V 1.0 supports this automatic removal of the raw data on request from the user.

If storage is local, it needs to be maintained by IT support, and if this is unwanted, cloud-based solutions could be an alternative. In addition, it is highly desirable that raw data, not subject to ethical issues, is stored in the International Nucleotide Sequence Database Collaboration (INSDC databases). There are three online database services that provide synchronized repositories of biological datasets: SRA, ENA and the DNA Data Bank of Japan (DDBJ). Sharing data on these repositories stimulates an Open Data for Science policy and provides data for researchers all over the world to analyze. As more researchers and health agencies share their genomic raw data in a timely fashion in these repositories, the field can move "towards a genomics-informed, real-time, global pathogen surveillance system" (quoted from Gardy and Loman, 2018). An added benefit is avoided storage costs for raw data and provides a reliable off-site data backup.

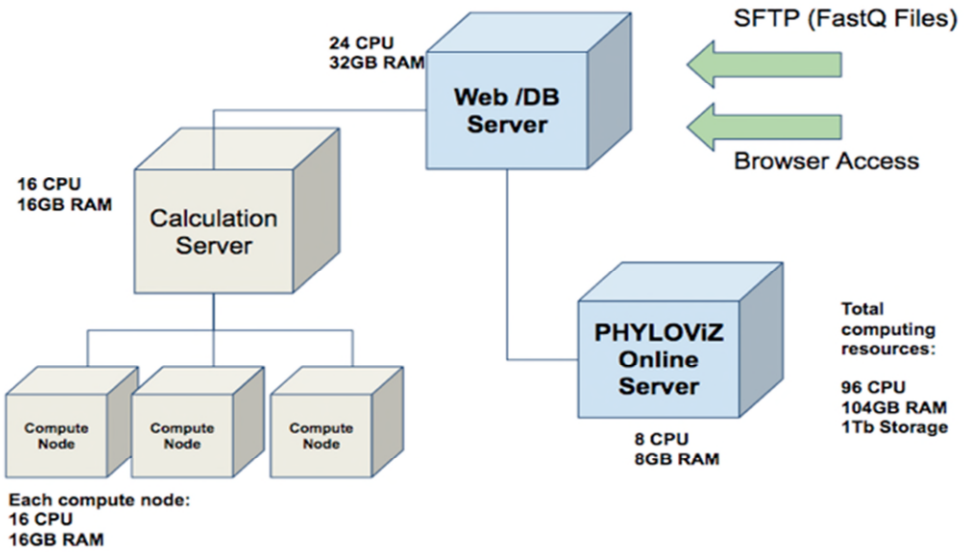
#### 4.2.1.2. Storing data from bioinformatic analysis

The bioinformatics software packages have a transient disk space requirement for them to operate, and these requirements must be taken into consideration for the software to run smoothly. Furthermore, the storage space needed for this intermediate data from analysis is bigger than that for storage of raw data. The INNUENDO Platform V 1.0 is therefore equipped with a feature to reduce the disk space needed for analysis, that we named "project locking". In a "locked project", the temporary data of the workflow is removed. This reduces the required space per strain to e.g. for *E. coli* ~12.5 MB since only final assemblies, allele call results and annotation results for antibiotic resistance and virulence genes are permanently stored. However, the user looses the ability to re-run part of the pipeline in that project, and a new project needs to be created for that purpose.

#### 4.2.1.3. Computational requirements for data analysis

Software applications in the INNUENDO Platform V 1.0 have different computational requirements (in CPUs and RAMs), and parameters can be optimized for maximum pipeline performance in different processes. One of these systems is OpenStack, a cloud operating system that controls large pools of

compute, storage and networking resources (virtual machines, VMs) through a datacenter. The current version of the INNUENDO Platform has been thoroughly tested using two different OpenStack systems. One is hosted on "Infraestrutura Nacional de Computação Distribuida" (INCD) (http://www.incd.pt/) and the other is hosted by the IT Center for Science (CSC) (www.csc.fi) in Finland. The latter was used for hosting the prototype of the INNUENDO Platform during the national usability test and it will be the host for the Platform the Finnish agencies will use for food-borne pathogen surveillance. INCD was used for the international usability test. Figure 16 shows the implementation of the prototype of the INNUENDO Platform used in the international simulation hosted at INCD.




The boxes represent different VMs configured in the OpenStack environment and the links between boxes represent allowed information flow between the VMs.

**Figure 16:** INNUENDO Platform installed at INCD OpenStack Cloud platform

The performance of the platforms processes can be fine-tuned by the use of NextFlow (https://www.nextflow.io) and the Slurm scheduling engine (https://slurm.schedmd.com). The configuration parameters limit the number of simultaneous processes that can be run on the compute nodes, as multiple processes running simultaneously is usually a performance bottleneck, leading to increased computing time. To validate this, we analyzed eight *E. coli* strains using the INNUENDO Platform simultaneously, and with a setup at INCD it took 30–40 min. The number of simultaneous strains being analyzed can be automatically increased by using more compute nodes, demonstrating the elasticity of the platform. Furthermore, the platform can be installed in a high-end laptop running Linux or a multi-CPU high performance server (usually 4/8 CPUS and 8-16GB).

### 4.2.2. Network connectivity

As the INNUENDO Platform V1.0 uses a web-application for accessing the Frontend server, a secure file and metadata transfer is done by secure Hypertex Transfer Protocol (*https*) access through a web-browser to a hosting system-server. The users use an *sftp* software to upload the *fastq* files to the system. Therefore, the web server hosting the platform must be configured for accessibility to the adequate network zone and the users must have permission from their institutions IT teams to access the platform web address and use *sftp* to the file transfer. The network zones can be configured to limit access within a private network (for instance the intranet for the relevant institute) or accessed by internet through a public IP-address.

For single server or laptop versions only local access is possible, thereby removing the need for *sftp* to transfer files.

# 5. Usability test and proof-of-concept studies

To devise how well the different proposed solutions work in the hands of microbiologists and epidemiologists in the field of food safety and public health (user group), we ran three separate usability tests. A first usability test was performed during the hands-on workshop in Vitoria-Gasteiz (July 2017) where approximately forty students acted as central national authorities. The second was a usability test conducted at national level in Finland (Finnish public health and veterinary authorities) and the third an asynchronous remote usability evaluation (as described by Dray, 2004) with twelve authorities across the EU (11 in the international simulation study, Appendix F). Through these, we measured the user's ability to complete one or more tasks using the platform (i.e. proof-of-concept studies) while we evaluated the efficiency, user-friendliness and satisfaction with all aspects of the platform, with special interest in sequence upload, graphics, the interface and communication protocols.

The three proof-of-concept studies performed in the context of the usability tests (i.e. the Vitoria-Gasteiz workshop hold in July 2017, the national simulation hold in October 2017 and the international simulation hold on February 2018) consisted of observing how well the phylogenetic framework worked to identify clusters and how the add-on software is able to predict *in silico* pathotype and serotype, and to predict the presence of resistance and virulence genes.

## 5.1. Hands-on Workshop

The INNUENDO consortium and the University of the Basque Country organized a Summer Course in July 2017 in Vitoria-Gasteiz. The course included a practical hands-on course on bioinformatics solution for public health microbiology. During the hands-on exercise, students learned the bioinformatics tools imbedded within the INNUENDO platform.

Students were divided in 10 groups of four which performed a simulation experiment on a set of well-characterized STEC strains from an US outbreak (Rusconi et al., 2016) using the first prototype version of the INNUENDO platform. Thanks to the support of CSC (www.csc.fi), students accessed, through a web browser, a total of ten virtual machines (VMs) with 16 CPUs and 80 Gb of RAM each hosted in cPOUTA cluster (https://research.csc.fi/pouta-user-guide) where a prototype version of the INNUENDO Platform was installed. Each group was further divided in two subgroups; one playing the role of human health authority (HHA) and the second the role for food and animal health authority (FAHA). In each group, both HHA and FAHA used the same VM (meaning that both had access to the same data), but through different computers, using different user identifications and performing analyses on a different set of strains.

The objective of the day was to identify which of the given strains belonged to the onging proposed outbreak investigation using the Platform and stimulate the discussion among the students of the same group to indentify possible flaws in communication between different authorities. A secondary objective for the day was to identify all possible flaws or limitations of the INNUENDO platform prototype. Students were asked to perform a "stress test" on the platform and report to the developers any errors found. In addition, developers interviewed the students, asking for feedback on the functionality of the platform itself. More information on the feedback form and the answers reported are presented in Appendix G

All the groups were able to correctly identity the simulated STEC outbreak strains. User pre-knowledge on STEC was found to be relevant in the ability of correctly interpreting the clustering data. Several disfunctionalities of the platform interface were identified by the users and corrected prior to the national usability study. Overall, the experience of the platform was graded as very good by the students, as resulted from a satisfaction survey performed at the end of the workshop (see Appendix G).

## 5.2. National simulation study performed between the two Finnish authorities (described in Appendix F)

Since 2016, the two Finnish authorities responsible for outbreak investigation of food-borne pathogens, THL and Evira, are moving away from traditional typing (PFGE) to WGS for outbreak investigations and HTS technics are being introduced in routine surveillance for *L. momocytogenes*, STEC and *Salmonella* spp., and procedures for exchanging molecular typing data (i.e. essentially PFGE) from outbreak investigations are in place between the two. For the national simulation study hold in October 2017, a single prototype of the INNUENDO Platform was installed at the CSC (www.csc.fi) and shared between the two institutions with one user each. One naïve microbiologist from each institution (i.e. a person not involved in the development of the tools or the interface) was instructed in the use of the platform and attached protocols. The proof-of-concept study consisted of retrospectively analyzing a nation-wide *E. coli* STEC outbreak.

The outbreak consisted of 26 isolates of human, food, animals and environmental origin, and the following INNUENDO modules were used in the analysis: pathotyping, INNUca (assembly and QC), chewBBACA with cluster analysis in PHYLOViZ Online 2.0. Clusters were defined as strains with a maximum of 0.7% pairwise differences from each other (independent from resolution used), and three outbreak clusters A, B and C were detectable in the provided dataset (more details in Annex D). The users were told to communicate cluster-detections to opposite authority by E-mail and deliver outbreak-reports to the study organizers. A logbook was kept concurrently by the study organizers to document the investigation process and evaluate the data flow and functionalities of the platform. As a result, a list of actions was made to remediate the weaknesses detected (Appendix F) and accommodate the user experiences collected during the national simulation. An improved version of the INNUENDO Platform and the definition of a different set of procedures to be used in the simulation were produced.

Importantly, as for the hands-on worshop hold in July 2017, the national simulation study provided a proof-of-concept of the INNUENDO platform's ability to pathotype *E. coli* and cluster STEC. A total of 21 of 26 isolates were pathotyped correctly, and reasons for failure included inferior sequence quality (flagged/failed in INNUca) and the presence of an atypical Enteropathogenic *E. coli* (EPEC) strain. Limited epidemiological data were provided for the strains, making the users to base their decision of clustering solely on sequence data. The users found that almost all except for a few strains of low sequence quality -6 sequences failed INNUca assembly pipeline due to: low coverage below mean read coverage, 25x, (three strains), or assembled coverage, 30x, (three strains)- clustered correctly (Appendix F). Therefore, the INNUENDO Platform and PHYLOViZ Online 2.0 provided a solution for the identification of suspected outbreaks by comparison against the entire database.

Issues in pathotyping functions were raised during the national simulation. The users wished for a more exhaustive list of pathotype genes in the final report generated by the platform, and improvement of a misleading pathotype nomenclature. These issues were addressed in the INNUENDO Platform prior to the international simulation.

## 5.3. The international simulation study

A usability study was conducted to assess the workflow, efficiency and satisfaction with the INNUENDO Platform and associated communication protocols in Ferbruary 2018. To be able to evaluate the sharing of data and results across borders, we simulated the occurrence of several international food-borne outbreaks with STEC with 12 central laboratories throughout Europe as participating investigators (Table 7).

**Table 7:** Participants of the international simulation

| Institution name | Country |
|---|---|
| Austrian Agency for Health and Food Safety | Austria |
| German Federal Institute for Risk Assessment | Germany |
| Institute of Food Safety, Animal Health and Environment | Latvia |
| European Centre for Disease Prevention and Control | EU/Sweden |
| European Food Safety Authority | EU/Italy |
| European Reference Laboratory for *E. coli*, Istituto Superiore di Sanità | EU/Italy |
| Finnish Food Safety Authority | Finland |
| National Institute of Health Dr. Ricardo Jorge | Portugal |
| Laboratoire National de Santé | Luxembourg |
| National Institute for Health and Welfare | Finland |
| University of Veterinary Medicine Vienna | Austria |
| Estonian Veterinary and Food Laboratory | Estonia |

As the platform was used in the participants´ home or offices in their own chosen time, the simulation exercise was classified as an asynchronous remote usability study as defined by Dray, 2004. A renewed beta version of the platform was installed in INCD as described in Section 4.2.3. The participants were provided a communication protocol (Appendix F) in case of outbreak detection and instructions on how to formulate an outbreak investigation report, while the participants playing the central European authority kept logbooks on the investigation process. After the international simulation, a survey was conducted to get feedback of the user's needs and experiences with the INNUENDO platform. As in the national simulation study, the SOPs and INNUENDO Platform were modified post-simulation taking the participants' feedback in consideration.

The twelve participants received each data in zipped archives containing paired-end Illumina sequences (.fastq.gz) with metadata (.csv) in two batches of 5 genomes (120 genomes in total), of which several belonged to one of the three clusters A, B or C as described in the national outbreak simulation (Appendix F). It was ensured that each participant received at least one cluster isolate. The participants uploaded the sequences to the INNUENDO Platform and performed all analyses available on the beta version of the platform: pathotyping, serotyping, INNUca (assembly and QC) with MLST determination, chewBBACA allele calling, with cluster analysis in PHYLOViZ Online 2.0 and ABRIcate (*in silico* typing using ResFinder, CARD, VFDB and PlasmidFinder databases).

If clusters containing samples from several participants were detected, the participants were instructed to contact each other as soon as possible using E-mail, with relevant "central authorities" as cc (Appendix F). The title and content of the E-mail was predetermined, emphasizing the use of nomenclature (L1, see Section 3.5.3) and trace-back to source if possible, but no fill-out-form was provided. They were also instructed to use a predefined title of the E-mail including the following items: INNUENDO *E. coli* cluster <classifier(s) of the strains in the cluster> (according to the L1 level) <mlst(ST)>. In addition, contents of the E-mail were instructed as follows: sample ID, submitter, project-name, and sample.

The international simulation study provided and strengthened the proof-of-concept achieved in the national simulation study. Most clusters were identified and communicated between the participants, as all samples expected to cluster were reported to do so. The participants used several results from the platform to conclude on clusters definition and adequate sample quality. They based decisions on 7 genes MLST sequence type definition, the classification of the strains based on the static cgMLST implemented in the platform, dynamic clustering using wgMLST and visualizing as MST in PHYLOViZ Online 2.0, and, finally, the results from quality control: INNUca assembly quality and allele quality check. . If not enough confidence could be obtained from these, results on resistance and virulence genes, and plasmids found by ABRIcate were evaluated to support the clustering hypothesis, demonstrating an use of add-on software that was not foreseen prior to the usability test. In addition to clustering, the usability test provided proof-of-concept that overall the INNUENDO *in silico* typing

workflow performed satisfactorily, as the pipeline was able to correctly predict both serotype and pathotype for the majority of strains.

Feedback on the participant's experience of the INNUENDO Platform was collected during the simulation on a digital channel and through a feedback form (See Appendix F). All the institutions participating in the simulation responded to the questionnaire rating their user experience of the INNUENDO Platform overall as very good. The users were particularly satisfied with the general organization of the platform, the running procedures, the color coding for run progresses, the report page, the nomenclature and the modules within INNUENDO Platform, of which INNUca (the assembly pipeline) received the highest score. Sample submission, the presentation of chewBBACA and ABRicate results and cluster analysis received average ratings and some improvements were suggested. The participants enjoyed the user-friendliness and provision of complete analysis solution with visualizations for *E. coli* offered by the INNUENDO Platform. In addition, they appreciated the ability to compare the results with other users (i.e. the real-time sharing of genomic data with minimum metadata).

On the downside, the participants wished for even more automatization in pre-analysis steps, suggesting a functional file upload of metadata and ability to select all workflows at once. Also, the interface between the INNUENDO Platform and PHYLOViZ Online 2.0 received some critics. For what concerns the user experience, the participants who fit INNUENDO's user group evaluated the platform with higher ratings (very good or excellent) and found it visually appealing and user-friendly, after only one short webinar training. This suggests that INNUENDO Platform met the needs of potential users and was easy to learn for them.

It was clear that during the simulation the communication and reporting were perceived as laboursome and poorly functioning by the majority of participants. Without a standardized format for cluster alarm notification, reports and comments on these directly from the INNUENDO/PHYLOViZ, the participants experienced communicating results time-consuming and subjected to errors. As a result, substantial variation was observed in the organization and format of E-mail communication, with various ways to notify on clustering, lack of requested information or extra non-requested data provided. E-mails were also used for other types of communication, like reflecting over results and constructing hypotheses, i.e. not just reporting of clusters, indicating that a secure communication pathway within the platform might be benefitial.

In summary, the feedback reflected that the INNUENDO Platform provided a lot of information and suggested that the platform serves users as intended by providing a quick judgment of data quality and sufficient information to support the decisions on borderline samples. Heeding the adjustments suggested in below would make the INNUENDO Platform a complete analysis solution of *E. coli* for public health and food safety reference laboratories.

Main findings of the usability studies are listed below and more details are available in Appendix F.

Factors achieving positive feedback:

- The platform is easy and intuitive to use, especially by microbiologists.
- The platform provides sufficient information for outbreak investigations, especially data quality metrics.
- The platform provides results that are useful in outbreak investigations.
- The platform provides a useful nomenclature to identify clusters and facilitate communication.
- The platform has an appealing visual look and easy-to-grasp visualizations.

Points for improvements:

- The platform needs functions that explain names, labels and provides help to guide new users and allow more intuitive use.
- The platform needs an optimization of serotyping and the upload of metadata-files.

- The platform needs to implement the VirulenceFinder database in ABRicate to complete the analysis of *E. coli* needed for reference laboratories.
- The platform needs to develop the interface between the INNUENDO platform and PHYLOViZ.
- The platform needs to develop a smoother transfer from "Projects" to "Reports" to provide a better user experience.
- The platform needs to create automatic summary reports and E-mail notifications of the results for use in communication between laboratories and other stakeholders (cautiously).

## 5.4. Summary of the proof-of-concept studies performed during the INNUENDO project

The national and international simulation studies represent not only usability tests of the INNUENDO Platform V1.0, but also proof-of-concepts of the use of WGS for cluster detection and investigation. In both simulations' users were clearly able to identify clusters of genetically close *E. coli* strains belonging to well characterized outbreaks. In addition to the simulations, during the project several additional proof-of-concept studies have been done with aim of validating the procedures implemented in the INNUENDO Platform V1.0 including: 1) species determination using GScompare 2) pathotyping and serotyping of *Y. enterocolitica* and *E. coli*; 3) cluster investigation using wgMLST. Each partner performed smaller studies and these are summarized in Section 8.

### 5.4.1. Species determination using GScompare

The aim of this proof-of-concept study was to test if GScompare was able to correctly determine the species of the four food-borne pathogens of interest (*C. jejuni, E. coli, Y. enterocolitica* and *S. enterica*). The analysis was performed by UPV/EHU. GScompare determines species by comparing oligonucleotide content of the sample to a database containing a wide range of bacterial species (http://gscompare.ehu.eus/) as described in Section 3.4 and Appendix C.

The Ensembl Genomes Release 34 database containing 41,610 genomes (41,198 bacteria and 412 *Archaea*) was obtained from the INSDC and used in this proof-of-concept study (http://ensemblgenomes.org/; December 2016).

Briefly, octanucleotide content of genomes of *C. jejuni, E. coli, Y. enterocolitica* and *S. enterica* were compared to all genomes within the Ensembl Genomes Release 34 database and distances were computed. If the most similar genome upon query belonged to the same or different species, the result was noted as a positive or negative assignment, respectively. So, in case of a 100% positive assignment, all genomes of that species were assigned a correct species. For the four species of interest, the GScompare software achieved assignment scores of 98.7, 97.8, 100.0 and 99.7%, and for *C. jejuni, Y. enterocolitica, S. enterica* and *E. coli,* respectively. Therefore, comparison of octanucleotide content of a query genome against a database may be used with high confidence as a fast speciation method for these four pathogens, as each genome is computed and compared within 0.4 seconds on one core.

This methodology is complementary to the platform and the online resources can be used as supplementary methodology for species confirmation, especially in the case of failing the quality check of assembly or allele calling.

### 5.4.2. Pathotyping and serotyping of *E. coli* and *Y. enterocolitica*

Numerous studies have defined important virulence determinants in *E. coli* (as reviewed in Rivas et al., 2015) and *Y. enterocolitica* (Reuter et al., 2015). As the majority of the subpopulations of these two bacteria are nonpathogenic commensal or environmental bacteria, the differentiation between which subpopulation cause disease (i.e. define the pathotype) and which do not is an important task in public health actions. This type of classification must be considered operational in nature, since it is

vague and potentially deceptive due to the continuous evolution of bacterial populations resulting in hybrid strains or in new strains that do not comply with known categories. Nevertheless, the actual subdivision of these pathogens in pathotypes was shown to be vital both clinically and epidemiologically, guiding clinical management and public health interventions (Robins-Browne et al., 2016).

Although modern genomic phylogenetic framework will become the future standard for one-for-all typing system for both organisms, there are still needs to contextualize the results of novel epidemiological investigations within historical data. Therefore, as described for pathotyping, the serotype prediction of pathogenic bacteria is relevant information for epidemiologists and clinicians for a correct response to food-borne outbreaks.

Several studies have shown the potential to predict patho- and serotype of bacterial pathogens directly from genomic data by searching for the presence of specific set of genes. With this in mind we have developed two modules for *in silico* prediction of pathotypes and serotypes of *E. coli* and *Y. enterocolitica* based on read mapping. ReMatCh software (https://github.com/B-UMMI/ReMatCh) is used as engine for assembly-free pathotyping of *E. coli* and patho_serotyping of *Y. enterocolitica* strains (using patho_typing - https://github.com/B-UMMI/patho_typing) and serotyping of *E. coli* (using seq_typing - https://github.com/B-UMMI/seq_typing), for details of the implementation please see 3.5.1 and Appendix D.

#### 5.4.2.1. Patho_typing: prediction of pathotype of *E. coli* and of *Y. enterocolitica*

In order to validate the efficiency of *patho_typing* module, raw reads passing the QC protocol (as defined in Section 3.5.1) of 655 *E. coli* strains belonging to different pathotypes were selected from the available literature (Dallman et al., 2013; von Mentzer et al., 2014, Grande et al., 2016; Ingle et al., 2016; Pettengill et al., 2016): 20 Enteroaggregative *E. coli* (EAEC), 26 Enteroinvasive *E. coli* (EIEC), 198 EPEC, 268 Enterotoxigenic *E. coli* (ETEC), 55 *Shigella* spp. and 98 STEC. For *Y. enterocolitica*, a total of 114 pathogenic and non-pathogenic strains from Reuter et al., 2015 and Reuter et al., 2014 were selected belonging to different serotypes and phylotypes. For *Y. enterocolitica* this method predicts also phylotype and certain serotypes since pathotype, serotype and phylotype are (at least, partially) correlated (Reuter et al., 2015, Reuter et al., 2014). For *E. coli* a strain classified as Avian Pathogenic *E. coli* (APEC) was selected (Ronco et al., 2017) as a negative control.

This methodology had a sensitivity and specificity of 99.46% (CI: 98.44%, 99.89%) and 97.78% (CI: 88.23%, 99.94%), respectively, to correctly predict pathotype for *E. coli* and 100% specificity and sensibility for *Y. enterocolitica*. For further details see Appendix D.

#### 5.4.2.2. Seq_typing: serotype prediction of *E. coli*

Using EnteroBase prediction as reference methodology, the ability of *seq_typing* module in predicting *E. coli* serotype using SerotypeFinder (https://cge.cbs.dtu.dk/services/SerotypeFinder/) database has been evaluated on a large set of public available genomes. To sample several times each O and H type, up to two strains from each available O/H combination type have been selected. Raw *fastq* reads for a total of 2,719 samples have been downloaded from ENA or SRA using *getSeqENA* (https://github.com/B-UMMI/getSeqENA) and passed the QC as defined above.

*Seq_typing* was highly concordant with EnteroBase prediction being able to find 96% of the O-types and 98% of the H-types. For a total of 65 and 46 over 2,719 samples O-type and H-type predicted by *seq_typing*, respectively, was different from the prediction of EnteroBase. In 55 and 13 cases the O- and H-type was not predicted, respectively.

To validate the serotyping prediction, a set of 279 *E. coli* with web-lab validated serotype was used. *Seq_typing* was able to predict 94.98% of the samples correctly. False results might occur due to several reasons including mix culture, laboratory mistakes or mapping problem. Although we found that *seq_typing* showed enough sensibility and specificity and it is able to predict equally as EnteroBase the serotype of *E.coli*, the user should consider possible false results and need to consider

it as "possible" type, validating it with all results from other *in silico* typing (MLST, resistome, etc.) and especially cluster analysis. The user needs to consider critically the biological implication of "possible" types as defined by the method. For further details see Appendix D.

### 5.4.3. Cluster investigation using wgMLST

#### 5.4.3.1. Two STEC outbreaks

To validate wgMLST clustering approach, UH and THL applied the guidelines described in 3.6.4 to perform cluster analysis on a *E. coli* epi-validated outbreak occurring in Finland during the summer of 2016 and a second one happening in US in 2014 (the latter described in Timme et al., 2017). The outbreaks include STEC of two different serotypes (O157:H7 and O121:H19) and two MLST types (ST-11 and ST-655). In addition, the dataset contained 10 isolates collected from sporadic cases of *E. coli* infections. In both analyses for each sample the 25 closely related strains belonging to the legacy dataset of the INNUENDO Platform V1.0 were extracted using the *k-closest* algorithm (Carriço et al., 2018).

Outbreak 1 (Kinnula et al., 2018). Fifteen of the strains were of the O157:H7 serotype, of which 11 were part of an outbreak and four were sporadic cases collected during the same time period. The initial tree was based on a goeBURST size of 2,795 loci. At approximately 1% distance the outbreak strains clustered together and separated clearly from all cases defined as sporadic but one. One sporadic case, the *E. coli* IN_STEC_FI_111 strain, was 0.25% distant from an O157:H7 outbreak strain. This strain shared the same L1 classification of the outbreak strains and all the strains showed same antibiotic resistance and virulence profile. We then performed a subset analysis including only the outbreak strains and IN_STEC_FI_111 producing a goeBURST cgMLST profile of 3,345 loci, resulting in a MST with 10 allelic distance (~0.3%) as the longest branch separating IN_STEC_FI_111 from the rest of the cluster. Pairwise distance analysis support the hypothesis that the sporadic strain is an outlier showing a median of 16 (0.48%) allelic differences with the outbreak strains which were 8-13 alleles from each other.

Outbreak 2 (Kinnula et al., 2018). Of the serotype O121:H19, three outbreak strains and six sporadic cases were available. The initial tree was based on a goeBURST size of 3,233 loci. At approximately 1% distance (32 allele differences), the outbreak strains formed a large cluster together with several sporadic cases and from strains of the legacy dataset. The cluster was composed by several cgMLST L1 types. Decreasing the tree cut-off to ~0.5% (15 allele differences), the three outbreak strains clustered together with a sample from legacy dataset (ESC_CA6748AA) with a different L1 classification. This sample, although being an American ST-655 strain as also the outbreak strains, was isolated in 2009 and, therefore, clearly unrelated to the 2016 outbreak. A subset analysis of the three outbreak strains and ESC_CA6748AA produced a goeBURST size of 3,365 loci. Now, the three outbreak strains showed one to two allele differences between each other, while the unrelated case was at 16, 17 and 18 allele differences from the outbreak cases.

In both outbreak analyses the wgMLST dynamic core-genome approach was able to discriminate outbreak from non-outbreak strains. However, in both cases epidemiological information was essential for correctly interpreting the clustering analysis. These two small examples clearly showed that clonal association between samples is not per se able to forecast epidemiological relationships between cases, especially for monomorphic lineages such as for certain *E. coli* STEC.

#### 5.4.3.2. Effect of re-sequencing and coverage on cluster analysis

Due to the high resolution the dynamic core-genome methodology can achieve, this method is sensitive to artificial differences between strains due to errors accumulated during the entire process from DNA extraction to sequencing, assembly and allele calling. We investigated the effect of resequencing on cluster analysis by performing in different MiSeq runs replicates of one strain for each species (Table 6). Starting from the same DNA extract, the replicates were subject to the same library preparation and MiSeq running protocol. The assumption was that the maximum allele differences between replicates should be of the same order of magnitude of what observed between

clonal isolates. The replicates showed different level of depth of coverage: *E. coli* replicates ranged between 60x and 100x, *Y. enterocolitica* between 52x and 97x, *S. enterica* between 48x and 92x and *C. jejuni* between 59x and 120x. All the assemblies passed the INNUca QC. A warning message for the allele calling was issued for one *E. coli* and one *C. jejuni* replicate. Regardless the warning message, all replicates were classified in the same L1:L2:L3 types based on the static cgMLST schema. The results of the allele calling and the cluster analysis are summarized in Table 8.

As expected, overall all the replicates were or identical or at few allelic differences from each other. The main differences between replicates were the total number of loci called which varied between 17 and 26 and which might affect the resolution of the wgMLST analysis.

**Table 8:** Results of chewBBACA allele calling and PHYLOViZ cluster analysis of the replicates for each species included in the INNUENDO Legacy Dataset

| Species | N. of replicates | Range total loci called (Median) | goeBURST profile size | Max distance | Max pairwise distance |
|---|---|---|---|---|---|
| ***E. coli*** | 9 | 3,446–3,658 (3,552) | 3,432 | 2 | 8 |
| ***Y. enterocolitica*** | 6 | 3,353–3,376 (3,375) | 3,318 | 2 | 3 |
| ***S. enterica*** | 9 | 3,753–3,770 (3,768) | 3,724 | 4 | 6 |
| ***C. jejuni*** | 7 | 925–950 (938) | 894 | 1 | 1 |

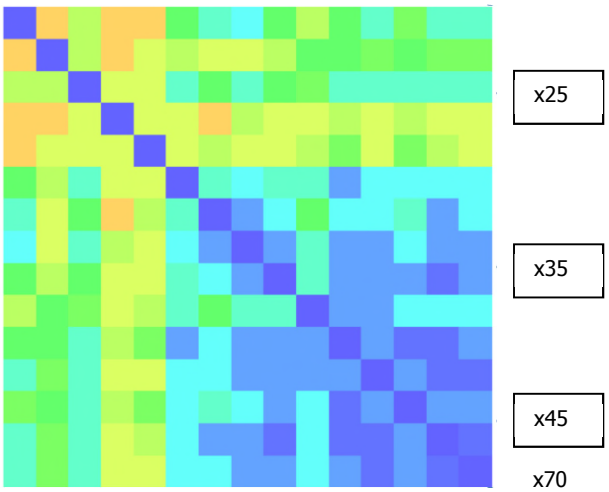
We further evaluated the effect of coverage on cluster analysis by randomly subsampling the raw reads of two *E. coli* samples named ERR163841 and SRR341556 at 25x, 35x, 45x and 70x coverage (five replicates each). All samples passed the INNUca and chewBBACA QC, and each of them clustered independently, calling each time the ten k-closest. The subsampling of ERR163841 resulted in a MST based on a goeBURST profile size of 3,544 loci. The pairwise distance matrix (Figure 17) shows a clear relationship between coverage and allelic distance. The five x70 coverage replicates were able to produce identical profiles, while reduced coverage from 70x to 25x coverage gradually increased the allelic distance to a maximum of three at 45x coverage, seven at 30X coverage and 13 at 25x coverage.

The subsampling of SRR341556 resulted in a MST based on a smaller goeBURST profile size of 3297 loci and all replicates but one produced an identical profile. The only different replicate showed a single allelic difference from the other replicates.

These analyses point out that artificial differences due to resequencing or difference in coverage might cause genomic diversity that, in certain cases, could be relevant to the cluster analysis. However, they also showed that this effect is stochastic and depends on several combinations of variables difficult to control. The only systematic effect registered is the effect of low coverage on cluster analysis as a low number of loci were called for these genomes (x25 and x35, Table 9).

**Table 9:** Averages and standard deviations of the total number of allele called for each coverage for two strains of *E. coli* test

| | Coverage | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | x25 | | x35 | | x45 | | x70 | |
| ***E .coli*** | AV | SD | AV | SD | AV | SD | AV | SD |
| **ERR163841** | 3,594 | 3 | 3,659 | 2 | 3,664 | 1 | 3,664 | 2 |
| **SRR341556** | 3,340 | 2 | 3,412 | 2 | 3,443 | 2 | 3,444 | 1 |




AV=Average; SD=Standard deviation;
Colours: Dark blue 0-1 alleles; Orange 12-13 alleles.

**Figure 17:** Heatmap showing the pairwise allelic distance among the 20 replicates of ERR163841

# 6. Recommendations for implementing WGS-based laboratory surveillance and outbreak investigation of food-borne pathogens using the INNUENDO Platform V1.0

WGS-based laboratory surveillance and outbreak investigation of food-borne pathogens require multidisciplinary knowledge and know-how. The information should flow fluently between the different authorities and professionals. These recommendations are written to overcome the noticed impediments highlighted in Section 2 and the three usability studies (Section 5) in order to reach an efficient way of delivering and sharing the information despite of different processes and organization structures within different countries.

More specifically, we here present the recommendations on how samples and their metadata should be handled from the specimen to final results and their interpretation. We provide a model of communication nationally, internationally and between different jurisdictions. In this report a general model for WGS-based surveillance and communication of the results is presented.

Certain prerequisites for the procedures to function correctly are presented in Section 6.1 - and were already highlighted in Section 2.1 - and mainly focuse on the formation of a local OCT, trained regularly to respond to food-borne outbreaks, efficient outbreak notification, and collection of sufficient samples analyzed correctly and shipped to reference laboratories for WGS analysis.

The following recommendations and best practices are meant for helping central national authorities to establish an effective WGS-based laboratory surveillance of food-borne pathogens using the INNUENDO Platform V1.0. They are not proposed as substitutes of the guidelines for outbreak detection and investigation provided by international organizations such as WHO (WHO, 2008) or ECDC (ECDC toolkit for food- and waterborne outbreak). Moreover, the following recommendations should always be applied within the framework of national and international legislations.

## 6.1. Common recommendations

Below a summary of the prerequisites for a correct function of any procedures for performing WGS-based laboratory surveillance and outbreak investigation of food-borne pathogens. They are compiled from the needs highlighted in Section 2.1 and the general recommendations from WHO (WHO, 2008) and ECDC (ECDC toolkit for food- and waterborne outbreak). Clearly defined and detailed instructions for performing WGS-based laboratory surveillance and outbreak investigation lessen the need for informal personal communications and decrease possible problems related to personal relations or

new employees. The instructions should be formulated at national level and applied locally, if possible, only with minor changes to ensure coherent structures. The areas of responsibility should be clearly defined and communicated and an updated flow chart for the actions to take during outbreak investigation must be available. For each action in the flow chart, a more detailed written instruction should be drawn up for instance for notification, logbook, sampling, and reporting. More specifically, the notification of a suspected outbreak should be described in detail, meaning which method is to be used for communication (datasystem, E-mail, phone), within which timeframe should the notification have been made and what the notification should consist of. All directions of communication must be defined within and between authorities/health professionals. For the logbook, kept to enhance communication and keep all the stakeholders informed, the instructions should describe how the logbook is held, how it is kept and how often and with whom it is to be shared. With regards to sampling, the instructions should state who is coordinating the collection of different specimens, and how the communication with the laboratory is maintained, including when and how the laboratory should be consulted for expert advice during the investigation. Instructions on how to write an outbreak report should be available, including information on what it should contain, who should usually write the report, to whom and when should it be submitted after the outbreak is solved.

## 6.2. A model protocol for rapid response to food-borne outbreak using the INNUENDO Platform V1.0

The INNUENDO project followed the implementation of WGS-based typing of food-borne bacterial pathogens for surveillance and outbreak investigation in Finland and Portugal. Based on the experiences with the prototype of the INNUENDO Platform implemented in Finland, a model protocol for surveillance and outbreak detection was drafted. The protocol includes steps from the handling of specimens to the communication of the final results. The protocol relies on the presene of two reference laboratories supporting surveillance and outbreak investigation with molecular typing of food-borne pathogens: one for human health and the other for food/veterinary analysis. Furthermore, this protocol presumes that the INNUENDO Platform V1.0 is used in shared analyses and that a minimum epidemiological data (Tables 10 and 11) is provided. However, the protocol can be adjusted to suit a different reality as well, but that goes beyond the scope of this report.

In the example below the INNUENDO Platform V1.0 is implemented in cloud and hosted in the national computing infrastructure, and it is accessible only by the two Finnish authorities through a secured login (i.e. the Platform is not accessible from other stakeholders). The Platform is administrated by the public health authority, who defined the protocols and workflows together with the food/veterinary authority, and the "users" are defined as the operators of the public health and food/veterinary authority involved in the surveillance and outbreak investigation of food-borne pathogens.

### 6.2.1. Identification of outbreak, collection of strains and WGS analysis

#### 6.2.1.1. The outbreak notification - how do the laboratories get knowledge of an ongoing outbreak?

Municipal health authorities are responsible for local outbreak investigations and notification of suspected food and waterborne outbreaks to national authorities preferably via online reporting system (if available). The notifications submitted through the online reporting systems must be followed up by the epidemiologists and laboratory microbiologists at the public health authorities and food safety and veterinary authorities, respectively.

#### 6.2.1.2. Role of laboratories at central authorities

The reference laboratories of public health and food safety both identify suspected outbreaks by cluster analysis (as defined in Section 3.6.4) and communicate their findings to each other, public health epidemiologists and food control authorities. Usually, the central public health authority will coordinate outbreak investigation if the outbreak occurs over several health care districts or includes

several municipal outbreak investigation groups. In this case, the public health authority and the food safety authority will contact the municipal health and environmental health authorities to start investigation in their respective area.

#### 6.2.1.3. Collection of human specimens

Clinical microbiology laboratories and physicians report all notifiable disease cases to a type of national infectious disease register. Isolates or samples collected from certain notifiable diseases specified by law should be sent to the public health reference laboratory for typing. For some of these WGS is performed immediately, for other samples other typing (such as serotyping) is performed in order to identify possible outbreaks. Available metadata for each strain is stored to a secure laboratory information system and a *line* list (i.e. the list of cases suspected or confirmed to be related to an outbreak) shared between the laboratory and epidemiologists of public health authorities locally. The *line* list links each strain to a patient with case ID and contains epidemiological data. Epidemiological data is collected at the public health authority by interviewing persons with laboratory confirmed infection.

#### 6.2.1.4. Collection of food specimens

Food safety and veterinary health authorities and municipal authorities agree on food sampling, usually in response to an outbreak notification. If a relevant pathogen is isolated, it should be subjected to WGS. Available metadata for each strain is stored to the secure laboratory information system (Tables 10 and 11).

#### 6.2.1.5. Upload sequence data

Several options are here available for storage of raw sequence data (see Section 4.2.1). The Finnish solution consists of uploading raw data to private data storage volumes on an INNUENDO Platform V1.0 hosted by a cloud service provider. Each submitting user (submitter) logs in to the platform with own credentials. The submitter creates a new project as follows: *<institution_acronym> <date [yyyymmdd]> <running number> <operator_initials>*. The public health authority reference laboratory fills in the metadata (if available) for each strain on the platform as listed in Table 10. The food and veterinary safety authority fills in the metadata (if available) for each strain on the platform as listed in Table 11.

**Table 10:** Metadata for human samples

| Type of metadata | Optional or mandatory | Fill in always or if available |
|---|---|---|
| *Primary Identifier* | Mandatory | Always |
| *Case ID* | Optional | Always |
| *Sampling date* | Optional | If available |
| *Sample received* | Optional | If available |
| *Submitter* | Mandatory - automatic | Always |
| *Owner* | Mandatory - automatic | Always |
| *Source* | Mandatory - drop down meny | Human |
| *Location* | Optional | Awaits approval |
| *Additional information* | Optional - free text | Fill in RYMY id |

**Table 11:** Metadata for food and environmental samples

| Type of metadata | Optional or mandatory | Fill in always or if available |
|---|---|---|
| *Primary Identifier* | Mandatory | Always |
| *Case ID* | Optional | Leave blank |
| *Sampling date* | Optional | If available |
| *Sample received* | Optional | If available |
| *Submitter* | Mandatory - automatic | Always |

| Type of metadata | Optional or mandatory | Fill in always or if available |
|---|---|---|
| *Owner* | Mandatory - automatic | Always |
| *Source* | Mandatory - drop down menu | Food; Animal, cattle; Animal, poultry; Animal, swine; Animal, other; Environment; Water; Feed; Unknown |
| *Location* | Optional | Awaits approval |
| *Additional information* | Optional - free text | Fill in RYMY id, case ID of suspected patient(s), specified source |

#### 6.2.1.6. Analyses to be performed on INNUENDO Platform V1.0 and in PHYLOViZ Online 2.0

The submitter performs all the analyses available on the platform: PathoTyping, serotyping, INNUca_plus_mlst (assembly, QC and MLST), chewBBACA (allele calling), Abricate (*in silico* typing using ResFinder, CARD, VFDB, VirulenceFinder, PlasmidFinder databases). The submitter performs cluster analysis by sending the results to PHYLOViZ Online 2.0 with the dataset name as shown in Table 12.

**Table 12:** Data to be submitted to PHYLOViZ Online 2.0

| | |
|---|---|
| **Dataset name** | *<your_institution_acronym> <date [yyyymmdd]> <running number> <your_name_initials>* |
| **Description** | Optional |
| **Number of closest strains** | 20 |
| **Additional data** | Optional (species-specific procedure) |
| **Make dataset publicly available to other users** | Yes |

"users": operators of the public health and food/veterinary authority involved in the surveillance and outbreak investigation of food-borne pathogens

#### 6.2.1.7. Interpretation of cluster analysis

The submitter interprets the results of cluster analysis in PHYLOViZ Online 2.0. Clustering could be suspected around "Tree cut-off" 0.5% of the "goeBURST Profile Size" (depending on the cgMLST profile it might vary approx. between 10-18 allele difference), but the submitter defines the final cut-off for interpretation (see Section 3.6.3). Borderline samples should be included in a subset analysis on higher resolution. Samples that were investigated as a suspected outbreak should be included in a subset analysis although they were excluded from the cluster in interpretation (select tree nodes with shift + left mouse click). The submitter creates a subset based on the interpretation with the settings listed in Table 13.

**Table 13:** Subset settings

| | |
|---|---|
| **Dataset name** | *<institution_acronym> <date [yyyymmdd] of the original tree> <running number of the original tree > <subset_running number> <operator_initials>* |
| **Analysis method** | Core analysis |
| **Has missing loci** | Yes |
| **Missing characters** | 0 |

### 6.2.2. Communication between users

To achieve traceability, the submitter downloads the full report as a reference for the traceability required by accreditation. All the results of the INNUENDO analysis that are saved to the laboratory information systems can be traced to this report. Furthermore, for detected clusters, the submitter saves the epidemiological report as .xlsx logbook to, for instance, an extranet workspace. All later analysis results (epi reports) related to this cluster are saved to the same file.

For each human strain, the epi-report contents are saved to a line list of cases (.xlsx table containing cases in one year) both at:

- Public health authority, containing the full epidemiological information of the patients
- On an extranet workspace in big .xlsx file per each year, containing only the epi report contents, organized by case IDs. Summary reports of the outbreak investigations are saved to this workspace after the investigation has been finished and the results are final.

Below we present a list which summarizes the communication of results at central level:

- E-mail notifications from PHYLOViZ handle communications between the laboratories of public health and food safety authorities and to epidemiologists at public health and food control authorities.
- The laboratory information system feeds the National Infectious Disease Register (ie. an information system storing disease notifications from clinical laboratories) with microbiological information.
- Line list of cases between the laboratory and epidemiologists contains the epidemiological data of patients at the public health authorities.
- Extranet workspace for outbreak investigations to which reference laboratories, epidemiologists and food control authorities have access.
- Extranet workspace for notifications of detection ofpathogens in humans with suspected link to food or farm
- Certificate of analyses from laboratory information systems to municipal authorities and clinical microbiology laboratories

Further communication to external instances should be done following the established set procedures e.g. informing ECDC (Tessy), EFSA, EURLs, RASFF system, etc if appropriate.

# 7. Dissemination

Activities and the results of our project have been disseminated through several channels including publications in international peer-reviewed journals, presentations in international conferences, publication of open-source software in a public repository, and organization of workshops and courses. Some parts of the present report have been already published in international peer-review journals and references are included where required. The publications are listed in Section 7.1.

## 7.1. Publications in international peer-reviewed journals

Below we present the list of the scientific manuscripts published in peer-reviewed journals involving the activities within the remit of the project INNUENDO.

- Llarena et al., 2016. Monomorphic genotypes within a generalist lineage of *Campylobacter jejuni* show signs of global dispersion. Microb. Genomics. 2(10):e000088. doi: 10.1099/mgen.0.000088 1: Fast phylogenetic inference from typing data
- Llarena et al., 2017. Whole-genome sequencing in the epidemiology of *Campylobacter jejuni* infections. J. Clin. Microbiol. JCM.00017-17. doi:10.1128/JCM.00017-17
- Carriço et al., 2018. Fast phylogenetic inference from typing data. Algorithms Mol Biol., 13: 4.
- Silva et al., 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification Microb Genom. Mar; 4(3): e000166.
- Palma et al., 2018. Genome-wide identification of geographical segregated genetic markers in *Salmonella enterica* serovar Typhimurium variant 4,[5],12:i:-. Scientific Report, 8(1):15251.

Several publications are submitted or forseen to be submitted in the coming months.

Submitted and available in Preprint

- Barker et al., 2018. Rapid identification of stable clusters in bacterial populations using the adjusted Wallace coefficient. Biorxiv; Preprint:doi: https://doi.org/10.1101/2993.

Publication planned to be submitted in 2018 and 2019 will include "The INNUENDO Platform", "INNUca, a standardized pipeline for bacteria genome assembly and quality control", "chewBBACA nomenclature server", "GS Compare", "Communication during outbreak investigation: functionalities and needs", "Phylogeography and population structure of *Yersinia enterocolitica* PG3". In addition, other publications will make use of the strains and the tools developed within the INNUENDO project are forseen.

## 7.2. International conferences

The project was publicly presented during the 11th International Meeting on Microbial Epidemiological Markers (IMMEM XI) 9th to 12th March 2016, Estoril, Portugal with a poster. Similarly, a poster illustrating the project framework and consortium was presented at the ESCMID Networking Corner at the European Congress of Clinical Microbiology and Infectious Diseases 27th annual congress, 22nd to 25th April 2017 Vienna, Austria.

Results from the development of the bioinformatics solutions for public heath microbiology were presented at the following conferences:

- 6th Applied Bioinformatics and Public Health Microbiology conference, 17th to 19th May 2017, Hinxton, UK
- Bioinformatics Open Day, 22nd to 24th February 2017 Braga, Portugal
- Bioinformatics Open Day, 14th to 16th March 2018 Braga, Portugal
- 27th European Congress of Clinical Microbiology and Infectious Diseases, 21st to 24th April 2018, Madrid, Spain

Moerover the INNUENDO Platform V1.0 has been presented at the Pulsenet International Meeting, Atlanta, 12th and 13th June, 2018 (https://www.cdc.gov/pulsenet/index.html).

## 7.3. Press releases, dissemination through the internet and citations

The main communication platform with the stakeholders was the web site of the project which can be reached at this address: https://sites.google.com/site/theinnuendoproject. At the beginning of the project, national and international press releases (including one in the Nature Microbiology Community blog) were published in several languages. Please note that links concerning the press releases are available in the side bar of the INNUENDO Project webpage.

In order to fulfil our policy of transparency and openness, all the bioinformatics solutions developed during the project (including all the versioning) are available at a dedicated github account https://github.com/TheInnuendoProject/ and also at https://github.com/B-UMMI/. All the software tools were shared during developing phase. That allows the scientific community to use and validate the tools in real-time.

Below we present a list of publications in international peer-reviewed journals citing or using tools developed within the INNUENDO project.

- Ribeiro et al., 2018. *Citrobacter portucalensis* sp. nov., isolated from an aquatic sample. Int J Syst Evol Microbiol. 2017 Sep;67(9):3513-3517.
- Motro and Moran-Gilad, 2017. Next-generation sequencing applications in clinical bacteriology. Biomol Detect Quantif. Oct 23;14:1-6.

- Carriço et al., 2018. A primer on microbial bioinformatics for nonbioinformaticians. Clin Microbiol Infect. Apr;24(4):342-349.
- Culebro et al., 2018. Origin, evolution, and distribution of the molecular machinery for biosynthesis of sialylated lipooligosaccharide structures in *Campylobacter coli*. Sci Rep. 2018; 8: 3028.
- Pasquali et al., 2018. Listeria monocytogenes sequence types 121 and 14 repeatedly isolated within one year of sampling in a rabbit meat processing plant: persistence and ecophysiology. Front Microbiol. Mar 29;9:596.
- Pardos de la Gandara et al., 2018. Genetic Determinants of High-Level Oxacillin Resistance in Methicillin-Resistant *Staphylococcus aureus*. Antimicrob Agents Chemother. May 25;62(6).
- Chung et al., 2018. Phenotypic signatures and genetic determinants of oxacillin tolerance in a laboratory mutant of *Staphylococcus aureus*. PLoS One. 2018 Jul 3;13(7):e0199707.

Lecture materials from team members are available in Slideshare and Figshare personal accounts (i.e. Mirko Rossi and João André Carriço) as well as developments in the field were communicated through Twitter (@innuendocon in addition to those from Mirko Rossi @happygipsy and João André Carriço @jaCarriço).

## 7.4. Courses and capacity building activities

The consortium organized two courses in May 2016 and in December 2017 in collaboration with the Doctoral School in Environmental, Food and Biological Sciences (YEB) of the University of Helsinki. The course was tailored for PhD students. During the one week course the students learnt and applied bioinformatic techniques to study microbial communities, metagenomics and performing population genetics. Online tutorials are available in a dedicated github account: https://github.com/mirossilabcourses

Moreover, members of the consortium collaborated as lecturers in an official Master of the University of the Basque Country entitled "Microbiology and health" at the subject "Epidemiology and Infections" in 2017 and 2018. Eleven master students received the newest information about WGS at epidemiology of Infectious Diseases.

In addition to the above course, the consortium organized a specific workshop intended as capacity building activity for stakeholders around Europe. The workshop was called:"Genomics in food-borne pathogen surveillance and outbreak investigation". It was held at the University of the Basque Country, in Vitoria-Gasteiz (Basque Country, Spain) 12$^{th}$ to 13$^{th}$ July 2017.

### 7.4.1. "Genomics in food-borne pathogen surveillance and outbreak investigation" Workshop (Vitoria-Gasteiz, July 2017)

The workshop "Genomics in food-borne pathogen surveillance and outbreak investigation" (Appendix G) contributed to the professional continuous development of public health stakeholders, providing expert advice on the analysis and interpretation of WGS-based typing data and giving the opportunity to access to the bioinformatic data analysis methodologies under development within the INNUENDO project. With this course we aimed to contribute to structure research training at the European level and to strengthen European public health capacity in this new genomic era.

The Summer Course was divided in two parts. The first part was a one day of Scientific Symposium during which invited speakers and speakers within Innuendo consortium presented recent advances in the field. The symposium was broadcasted online through a webinar platform.

The symposium day was recorded and it is fully available at the following link: https://ehutb.ehu.es/series/59b66d7bf82b2b150d8b468e.

The second part included a practical hands-on course on bioinformatics solution for public health microbiology. During the hands-on activities, the students learned bioinformatics tools developed within INNUENDO project through the use of a prototype of the INNUENDO platform. Details of the hands-on activities are presented in Section 5.1.

# 8. Contribution and feedback from consortium members

## 8.1. Leader/coordinator: University of Helsinki (UH)

The University of Helsinki team was composed by eight researches working at the Department of Food Hygiene and Environmental Health, Faculty of Veterinary Medicine. Associate Professor Mirko Rossi acted as coordinator of the consortium and leader of the UH team. In WP1 the UH team was involved in the selection of the *Campylobacter jejuni* and *Yersinia enterocolitica* strains to be sequenced. Moreover, within the first WP UH team (leaded by Docent Dr. Mari Nevas) performed all the activities regarding Task 1.2 "Metadata and data flow assessment": design the questionnaire, contact the local and national authorities in Finland, Estonia and Latvia, performed the interview and the analysis. The UH team coordinated the WP2 by defining, in close collaboration with UL, THL and EVIRA, the phylogenetic framework to be implemented in the INNUENDO Platform, designing and curating the wg/cgMLST schemas for the four species, validating the protocols and defining the allele and strain nomenclature (in close collaboration with Eduardo Taboada's team at Public Health Agency of Canada). For what concerns the WP3, the UH team participated in all the phases of design, implementation and testing of the bioinformatic solutions developed by Univeristy of Lisbon (UL) team. Furthermore, the team supported THL and EVIRA in the organization of the international simulation within WP4 by contacting the participants, hosting the training session and chairing the feedback session. In addition to managing the project, the UH team actively participated in the WP5 by ensuring a good communication between participants and between the consortium members and stakeholder throught the curation of the website and public repositories. As scientific dissemination activities, several articles were published and the UH team had several invited presentations concerning the activity of the project.

## 8.2. Partner 1: Universidade de Lisboa (UL)

In the INNUENDO project, the UL team was leading the WP3 (Infrastructure development) and had participation in all other planned work packages. This allowed the development of the INNUENDO Platform V1.0 and its modules: INNUca, chewBBACA, seq_typing/pathotyping for *Y. enterocolitica* and *E. coli* and PHYLOViZ Online 2.0. The interaction with all the other partners and end-users was fundamental for the software development, which allowed not only the definition of the necessary requirements for the platform, but also for the modules. This resulted in much better usability of the software modules by providing the desired outputs for the end-users. The INNUENDO project also allowed UL team to explore several computational aspects needed for the effectiveness of the platform, such as the use of container technology (Docker images) and the deployment of software on cloud-based systems such as the OpenStack. For the WP2 (Phylogenetic calibration), the UL team collaborated closely with the UH team, for the development of needed software and in the resulting analysis. The UL team also had a major contribution on the WP4 and WP5, by setting up the platforms needed for the simulation studies as well as for the training workshops, where they also played a role on the teaching and presenting the platform modules. As scientific dissemination activities (see Section 7), several articles were published, and the UL team had several invited presentations concerning the platform or its modules in the following conferences or institutions. The INNUENDO Platform V1.0 will also form the basis for other ongoing projects that focus on nosocomial infections and the creation of user-friendly platforms for WGS analysis in these settings.

Abstracts in conferences:

6th Applied Bioinformatics and Public Health Microbiology conference, 17-19 May 2017, Hinxton, UK

- Ribeiro-Gonçalves B, Rossi M, Ramirez M, Carriço JA. Dynamic cgMLST analysis: making the most out of your gene-by-gene data. (Poster Presentation)
- Halkilahti J., Machado M.P., Salmenlinna S., Nyholm O., Mendes C.I., Nalbantoglu Y., Jaakkonen A., Borges V., Ramirez M., Rossi M., Carriço J.A. INNUca, a standardized pipeline for bacteria genome assembly and quality control. (Poster Presentation)
- Carriço J.A., Silva M., Rossi M., Ramirez M. chewBBACA – a comprehensive and highly efficient workflow for a Blast Score Ratio based allele calling algorithm. (e-poster presentation)

Bioinformatics Open Day, 22-24 February 2017 Braga, Portugal

- Machado, M.P., Rossi, M., Mendes, C.I., Nalbantoglu, Y., Ramirez, M., Borges, V., Carriço, J.A.. INNUca, a standardized pipeline for bacteria genome assembly and quality control. (Poster Presentation)
- Mickael Silva, Mirko Rossi, Mário Ramirez and João André Carriço. chewBBACA – an efficient framework for large-scale prokaryote whole genome/core genome MultiLocus Sequence Typing analysis. (Poster Presentation)

Bioinformatics Open Day, 14-16 March 2018 Braga, Portugal

- Ribeiro-Gonçalves B., Silva D., Machado M. P., Silva M., Halkilahti J., Jaakkonen A., Ramirez M., Rossi M., Carriço J. A.. Implementing High-Throughput Sequencing in bacterial food-borne pathogen surveillance: The INNUENDO Platform. (Poster Presentation)

27th European Congress of Clinical Microbiology and Infectious Diseases, 21 - 24 April 2018, Madrid, Spain

- Ribeiro-Gonçalves B., Silva D., Machado M. P., Silva M., Halkilahti J., Jaakkonen A., Ramirez M., Rossi M., Carriço J. A.. The INNUENDO platform: a user-friendly platform for the integration of high-throughput sequencing in bacterial food-borne pathogen surveillance (Mini-oral ePoster sessions #O0760).

Official presentations of the INNUENDO Platform (João Andrè Carriço):

- University of Antwerp, 22nd May 2018
- Pulsenet International Meeting, Atlanta, 12th and 13th June 2018
- US Center for Disease Control and Prevention, Atlanta, 14th June 2018

## 8.3. Partner 2: Universidad del Pais Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)

The participation at INNUENDO Project allowed us to increase highly our knowledge about the use of WGS in epidemiology of infections, not only in our research group but also at the University of the Basque Country (UPV/EHU) level and beyond. The Basque Government co-funded economically the INNUENDO project due their high interest and responsibility in food-safety and its interest in collaborating with the research projects developed by UPV/EHU, as the only public university at the Basque Country.

Apart of the general contribution of the UPV/EHU group of part of their *Salmonella enterica* strains collection to the INNUENDO Sequencing Project and the development of the GScompare software (See Section 3.4. and Appendix C), the members of our research group were also involved in a parallel project aiming to study the genetic diversity of 70 *Salmonella* spp. Typhimurium strains and its monophasic variant obtained from human and pigs samples in Spain. These serovars are very

frequently responsible of food-borne outbreaks at national and international level. We used our sequencing facility at the UPV/EHU, followed the instructions and recommendations from INSA (Portugal), and analyzed the data using GSCompare, INNUca and chewBBACA software programs developed by the INNUENDO team. This study was performed in collaboration with members of the Spanish Research Council (CSIC), Pamplona.

The group was involved in the dissemination activity:

- attending ESGEM and ECCMID scientific meetings
- organizing the Summer Course in "Genomics in food-borne pathogen surveillance and outbreak investigation" (See Section 7.4.1)
- organizing a meeting at the Agriculture Department of the Basque Country Government In March 2018 involving more than 100 food-safety researchers of the Basque Country area.

## 8.4. Partner 3: University of Veterinary Medicine, Vienna (VETMEDUNI)

In the WP1, The team of VETMEDUNI, Vienna was engaged in the selection of INNUENDO's *Salmonella* and *Escherichia coli* isolates. Additionally, the selection of isolates from Austria was managed and distributed by the team. Discussion and cooperation with the National Reference Centre in Austria was effective in dedicating national outbreak isolates and isolates of *Salmonella*, *Yersinia, E. coli* and *Campylobacter* to the INNUENDO project. *Salmonella, Yersinia and Campylobacter* isolates from slaughter carcasses, food and animals were included, too. Isolates came from the National Reference Centre and from the strain collection of the Institute of Meat Hygiene. All Austrian isolates were either sent to the VETMEDUNI Vienna of streaked from the strain collection of the Institute for DNA isolation except isolates of STEC which were sent directly to partner 6 (INSA). DNA was purified as specified by the INNUENDO protocol for DNA isolation. Further DNA was sent for WGS. For the calibration of the database VETMEDUNI contributed WGS of *Campylobacter* isolates of the total food chain from 14 days old broiler to the packaged products at no costs to EFSA. From one flock Campylobacter isolates were sampled on weekly bases until slaughter. At each slaughter step at two different slaughter days (fractionated slaughter) to the packaged product. The sequences were used to study genomic variation along the food chain and horizontal gene transfer was considered by analysing *C. jejuni* and *C. coli* isolates from the same flock. For what concerns WP2, in Austria only less than 13% of reported outbreaks in 2016 could be linked by confirmed evidence to a certain food vehicle (https://www.ages.at/download/0/0/b97d0bc3a76aefc5ffc43bfe3e1011ea100f0220/fileadmin/AGES2015/Themen/Krankheitserreger_Dateien/Zoonosen/LM_bedingte_Ausbr%C3%BCche/lebensmittelbedingte_krankheitsausbrueche_2016.pdf). Selected isolates of *Salmonella* and *E. coli* were divided into sero-groups. Sequences were obtained from outbreak isolates and sporadic cases and along the food-chain.

A workshop on WGS introducing the platform of INNUENDO was held in Vienna during the German-French summer school in 2018 in the first two weeks of July. Twenty students attended the course and followed a theoretical and practical training on the INNUENDO Platform and on WGS for outbreak investigation of food-borne pathogens.

To further analyse the usefulness of the INNUENDO Platform as a research tool we approached part of the platform and the total platform for two scientific projects.

- Firstly, the INNUENDO Platform was used to describe the relation of *Campylobacter* isolates. As a result, a new tetracycline resistance gene was identified in two of these isolates, conferring low level tetracycline resistance (below the epidemiological breakpoint defined by EUCAST for *Campylobacter jejuni*). Investigations will follow beyond the end of the INNUENDO project and will be published when finalized.
- Additionally, we used the INNUENDO Platform, particularly the implementation of ABRicate, for identifying antimicrobial and virlulence genes from *Salmonella* in a coliphage (i.e. phage with the primary host *Escherichia coli*). Part of this gene material was transduced into a

standard host of *E. coli* as a part of a larger project. Investigations will follow beyond the end of the INNUENDO project and will be published when finalized.

## 8.5. Partner 4: National Institute for Health and Welfare (THL)

THL acted as WP4 leader and contributed to the simulation studies performed nationally and internationally (Sections 5.2 and 5.3). These simulations were used to test the different prototypes of the INNUENDO Platform and provide feedback for further development. THL also contributed to the microbiological bases of analysis performed by the Platform, such as defining the genes needed for *E. coli* pathotyping (Section 3.5.1 and Appendix D) largely according to a thesis work on diarrheagenic *E. coli* (http://urn.fi/URN:ISBN:978-951-51-2625-2).

The collaboration and communication between THL and EVIRA (Partner 5) has long tradition. During the INNUENDO project, the collaboration was boosted even further. THL and Evira drafted a preliminary procedure (summarized in Section 6.2), which will form the basis of a technical standard operating procedure (SOP) in preparation, for the use of INNUENDO Platform V1.0 by researchers from both Institutes. This will be the first shared technical SOP for the laboratories situated in different Institutes. The preparation of the SOP required review of communication methods/practices and identified needs to improve communication. Several meetings were held between the laboratories where experiences and technical issues on sequencing and other typing methods were shared, and data security, sampling frames regarding surveillance, and future project plans were discussed.

THL provided 84 isolates to be sequenced by Partner 6 (INSA). These sequences can now be used as baseline information in national genomic surveillance. The STEC sequences are already used as material in another research project focusing on STEC-infections in children with or without Hemolytic Uremic Syndrome (HUS). INSA also shared their protocol for library preparation which helped us in reducing sequencing costs.

At the beginning of the project, THL had sequenced bacterial isolates for approximately one year. During the INNUENDO project sequencing was established as the main tool for outbreak investigation. The project required us to solve issues related to IT-infrastructure, data security policy, and establishment of cloud service (cPouta) at CSC (Finnish center of expertise in ICT; www.csc.fi).

The project researcher from THL also participated actively and directly to the INNUENDO Platform development. Each prototype version of the Platform, and the modules available through github were used. Different tasks for testing the functionality of the Platform were asked by project leader in UH and by Partner 1 (UL). These tasks included selection and testing of target genes for determination of true coverage (Section 3.3.2 and Appendix B), development, assessment and testing of *different in silico* typing methods (Section 3.5.1 and Appendix D), participation of executing and assessing the effects of coverage by downsampling to allele calling (Section 5.4), general usability testing and designing of the platform (Sections 5.2 and 5.3). The discussion between the project researcher in THL and consortium members in UL and UH were frequent (at least weekly, sometimes more often). The project researcher received training for use of INNUENDO Platform V1.0 as the administrative user, but he also acted as trainer and provided technical support at the simulation studies. The trainings were organized on line as well as during project meetings in Vitoria-Gasteiz (Section 5.1) and in Parma during the final meeting of the project. At the end of the project, one advanced training session was organized for the THL project researcher in Lisbon (visit 2-8.7.2018). The intense collaboration regarding the development of the Platform to serve as much as possible the needs of THL was a key element also in increasing the capability in bioinformatics. The involvement of the THL project researcher in the different tasks requiring skills in bioinformatics provided hands on training during the entire project. Improvement of skills included especially quality assessment, *in silico* typing by using existing open source software, and critical assessment of different software. These skills will and have already benefited also all other persons working in this field at THL.

Because of limited funding of the project, there was no possibility to maintain one "Test" version and another "Development" version of the Platform. If this would have been possible, the continuous use of "Test" version for analysis of routine surveillance samples would have allowed faster learning and

adoption of the Platform to routine use before. Instead, each development phase required upload of the previously uploaded sequences again. Also, the Platform could not be available for the basic users during development phase. Therefore, persons other than the project researcher could not access the Platform (development versions) as much as initially estimated. However, the end result was that the Platform contains several functionalities, which were not even anticipated at the beginning of the project and is as such a complete solution for surveillance and outbreak investigation.

One of the development versions of the Platform was used in real time for investigating a *Yersinia enterocolitica* outbreak investigation. Increase of *Y. enterocolitica* bio/serotype 4/O3 was observed in several health care districts in 2017. Sequencing and analysis of patient isolates by INNUENDO Platform showed that there was no single strain causing the increase. Instead, there were several smaller clusters and sporadic isolates. Suspected food samples did not contain *Y. enterocolitica* bioserotype 4/O3, and therefore testing Platform together with Evira was not yet possible in real time outbreak setting. However, the MST trees generated by the Platform were shared with Evira and local outbreak investigation teams.

THL participated in the dissemination of the project by presenting one poster in collaboration with coordinator and Partner 1 at the 6th Applied Bioinformatics and Public Health Microbiology conference, 17-19 May 2017, Hinxton, UK (Halkilahti et al., INNUca, a standardized pipeline for bacteria genome assembly and quality control) and by presenting the implementation phases of the INNUENDO Platform at the Vitoria-Gasteiz workshop (Section 7.4.1) and at the Nordic Zoonosis meeting in Oslo 19th of October 2018.

In conclusion, this project proved extremely valuable for THL, and will set an example for future projects. The main achievements of the project from THL's point of view are:

- Establishment of a shared Platform for THL and Evira for genomic surveillance and outbreak investigation of food-borne pathogens (INNUENDO Platform V1.0, installed and available for use, September 10th 2018);
- Increased capability in bioinformatics, which will benefit preparedness for genomic pathogen surveillance in Finland and different research projects (other than INNUENDO);
- Networking with researchers with state-of-the-art knowledge in bioinformatics;
- Review of communication protocols of sequence data in outbreak investigation and surveillance.

## 8.6. Partner 5: Finnish Food Safety Authority (EVIRA)

EVIRA contributed to all WPs of the INNUENDO project, namely Tasks 1.1, 1.2, 2.1., 2.2, 3.1, 4.1 and 4.2 (see Section 1 for details). EVIRA participated in the collection of the INNUENDO dataset (Task 1.1; Section 3.2) by donating and choosing of isolates and extracting and dispatching their DNA to INSA. EVIRA donated the DNA of 44 STEC isolates and 13 *Y. enterocolitica* isolates. In addition, 14 *S*. Enteritidis and 67 *C. jejuni* isolates were donated through collaborators at THL and UH. Altogether, 155 isolates of the INNUEDO Legacy Dataset originated from EVIRA, containing isolates from food, animal and environmental samples related to outbreaks and human cases, longitudinal study on dairy farms and monitoring programs for STEC and *Campylobacter*.

In addition, Evira contributed to the assessment of data flow and the development of communication protocol for outbreak investigations (Task 1.2; Section 2.1) by providing feedback, comments on national practices and results from the INNUENDO proof-of-concept studies (Section 5). Furthermore, efforts were made to interpret the *General Data Protection Regulation* (https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en) and find solutions for opening the data when publishing in ENA.

Within the scope of WP2, EVIRA and collaborators conducted research related to the calibration of genetic diversity of *Y. enterocolitica* (Task 2.1; Section 3.6), *C. jejuni* (Task 2.1; Section 3.6) and STEC (Task 2.2; Section 3.6). These studies provided insights on genomic variation and persistence of

bacterial strains on dairy farms in the course of time. Furthermore, strain characteristics affecting persistence were investigated. EVIRA also participated in the development of the true coverage module in INNUca (Task 3.1; Section 3.3.2 and Appendix B) by exploring diversity of *Salmonella* spp. genomes in order to find suitable gene targets. Development of the INNUENDO components (WP3) was also supported by testing the tools under development, defining user needs and providing feedback.

As the major responsibility, however, EVIRA conducted two proof-of-concept studies (the national and international simulation, Sections 5.2 and 5.3) together with THL (Partner 4). In the national study, EVIRA participated in the simulation and coordination of the simulation while THL took the main responsibility on planning and reporting of the study. In the international simulation, EVIRA was responsible for writing the plan, instructions for the participants, and the report. Furthermore, EVIRA participated in organizing the training, dispatching the sequences to the participants, collection of the results and keeping of logbooks as a central authority. EVIRA also participated in the study as an investigator. Unlike originally planned, the international simulation was opened for participants outside of the consortium, which expanded the organization work. This work was tackled with the help of UH (coordinator) and UL (Partner 1) teams.

When starting the INNUENDO project in January 2016, EVIRA had only utilized WGS for few isolates through research collaboration. However, establishment of the technology had been fostered with high priority by the executive board and efforts were made to build up knowledge. One project researcher was recruited to the INNUENDO project with some prior knowledge on bioinformatics and experience on molecular methods and microbiological outbreak investigations. The training of the researcher in bioinformatics was further supported during the project by PhD courses, conferences and collaborations. The researcher was later appointed to a permanent post.

License of the Ridom SeqSphere+ software was purchased in January 2016 and schemas that had been developed or established at THL were readily adopted to prepare for collaborating in WGS-based outbreak investigations. From then on, outbreak isolates were sent to THL for sequencing and data analysis, but THL shared the raw data with Evira to allow repeating the analyses in training purposes. In addition, some isolates were subjected to WGS in research projects by outsourcing. Both EVIRA and THL used their own Ridom SeqSphere+ instances, which did not allow direct sharing of data and results. In October 2017, EVIRA and THL purchased computing resources together from CSC – IT Center for Science Ltd. (Espoo, Finland) to set up a test version of the INNUENDO Platform for easier sharing of results.

In autumn 2017, EVIRA invested in in-house sequencing facility by purchasing Illumina MiSeq and equipment for library preparation. The sequencer was shared among the laboratory department: food and feed microbiology; veterinary bacteriology, pathology and virology; chemistry and plant diseases. However, other units except microbiology and bacteriology had only minor sequencing plans at the time. Library preparation facilities were shared between microbiology and bacteriology in Helsinki. To establish and validate the wet-laboratory methods for WGS and train the staff, a working group was set, involving altogether five researchers and four laboratory technicians from microbiology and bacteriology units. The group was led by the INNUENDO project researcher who took the main responsibility on training the group members and other researchers.

Training consisted of small-group lectures and laboratory demos and peer-learning was encouraged. Training sessions (1–2 hours) were held almost weekly in autumn 2017, later more seldom. All group participants, both researchers and technicians, attended training on both wet-laboratory methods and data analyses. The emphasis was on understanding how sample quality affects data quality, and subsequently analysis results, and how to ensure quality through the whole process. Validation plan was written, including a plan for data management, and responsibilities were divided within the group. Proficiency tests on WGS were participated, organized by the European reference laboratories.

To further disseminate the fruits of the INNUENDO project, the project researcher held a seminar presentation in May and October 2017 and a lecture in May 2018 that were open to the entire personnel at EVIRA. The first seminar was organized in collaboration with other research units at

EVIRA and included use-case topics on WGS. The second seminar was EVIRA's annual Science Day that was a public event, unlike the other two in-house events. The lecture concentrated on phylogenomics and interpretation of trees and received plenty of positive feedback, revealing the need for this type of training. The lecture was attended by risk assessors and food control authorities who work on outbreak investigations, in addition to research and laboratory personnel.

During the INNUENDO project, EVIRA build capacity for performing WGS in wet laboratory and analyzing the data almost from scratch. Support received from the INNUENDO collaborators had an utmost impact on the implementation of this new technology. At the end of the project, in July 2018, EVIRA had not yet developed routine for the sequencing of outbreak-related isolates in-house or for analyzing data on the shared INNUENDO Platform with THL but had all the capacity and will to do so by the time of the deployment of the INNUENDO V1.0 by September 2018. Validation of the methods and training of the staff continued to be an ongoing effort. Dissemination of knowledge on WGS-based methods was further planned to be extended to the local laboratories (i.e. EVIRA's customers) in autumn 2018.

During the capacity building effort, problems were recognized related to governmental IT solutions, data protection regulation and funding of the sequencing activities. Solving of the remaining issues was planned as the next step after the INNUENDO project, aiming to further enhance the flow, management and sharing of the data. The ultimate goal was set towards well-maintained and representative national data resources that can be used in surveillance and research within the authorities and science community.

## 8.7. Partner 6: Instituto Nacional de Saúde Dr. Ricardo Jorge (INSA)

INSA contributed to the project at different levels. INSA lead the data collection and INNUENDO sequencing project (Task 1.1; Section 3.2) by performing the following activities:

- Selection of strains
- Setting guidelines for DNA samples preparation and shipment
- DNA samples reception and assessment of DNA quantity and quality
- Optimization and validation of sequencing-related procedures (i.e., library preparation and sequencing run)
- Whole genome sequencing
- Setting guidelines for WGS data submission to public repositories

Moreover, INSA participated in the design and/or testing of the key bioinformatics tools developed/optimized on behalf of INNUENDO (Tasks 3.1, 3.3, 4.2, 5.1), such as INNUca (Section 3.3), chewBBACA (Section 3.6.1) and patho_typing (Section 3.5.1 and Appendix D). Furthermore, INSA participated to the Summer Course "Genomics in food-borne pathogen surveillance and outbreak investigation" (Section 5.1) and the international simulation study focused on testing the workflow and efficiency of the INNUENDO Platform and associated communication protocols (Section 5.3).

INSA also contribute in disseminate results from the INNUENDO project (Task 5.1)

- Divulgation of INNUENDO in Nature Microbiology section entitled "Nature Microbiology Community" (https://naturemicrobiologycommunity.nature.com/users/19645-joao-paulo-gomes/posts/14787-innuendo-one-size-fits-all-approach-for-bacterial-wgs-integration-in-routine-surveillance-and-epidemiological-investigation).
- Organization of a seminar at INSA focused on promoting INNUENDO as a framework for the integration of genomics in surveillance across sectors (food, veterinary and human health) (http://www.insa.min-saude.pt/forum-de-discussao-virar-de-pagina-na-vigilancia-laboratorial-de-doencas-infeciosas-aplicacao-da-sequenciacao-total-do-genoma-wgs/)

- Participation in the divulgation of the INNUca bioinformatics suite (through poster presentations) in two scientific congresses: 1) the 6th Applied Bioinformatics and Public Health Microbiology (ABPHM), Cambridge, UK, 17-19 May, 2017; 2) the Bioinformatics Open Day 2017, Braga, 22-24 February, 2017.

The activities underwent by INSA during the project timeframe strongly contributed to the current status regarding INSA capacity to perform large-scale WGS-based routine surveillance of FWD pathogens, on behalf of its role as National Reference Laboratory. Indeed, for example, the optimized and validated sequencing-related procedures (i.e., library preparation and sequencing run) and the continuous monitoring of quality indicators constitute nowadays the standard procedures in INSA. Finally, the fellowship hired by INNUENDO, who was fully dedicated to the development of these activities, recently integrated the INSA staff, reinforcing the INSA capacity in WGS. The bioinformatics solutions developed by INNUENDO (see Sections 3.3, 3.4 and 3.6) were fully adopted by the Bioinformatics Unit of INSA, revealing to be key for the improvement of their analytical capacity by ensuring reliability of data analysis while promoting standardization and traceability. For example, INNUca (Section 3.3.2), by allowing automate pathogen-independent bacterial *de novo* assembly and quality control, largely increased INSA capacity to perform large-scale routine analysis. Also, chewBBACA (Section 3.6.1) opened the possibility of using freely available software for cgMLST/wgMLST, overcoming the need of relying on pay-per-use closed source software. It must be highlighted that the training activities promoted by INNUENDO, namely the Summer Course (Section 5.1) and the International outbreak simulation (Section 5.3), which allowed testing the functionality of the INNUENDO bioinformatics modules and attest their user-friendly usability, consolidated the perspective that the INNUENDO Platform V1.0 will be the key tool for WGS-based surveillance of food-borne diseases at the Portuguese National Institute of Health. In summary, the integrative application of INNUENDO modules enables a more comprehensive and efficient integration of the high-resolution typing data provided by WGS with the epidemiological and clinical data, strengthening the ability of the us as Portuguese Reference Laboratory to detect phylogenetically informative data leading to public health actions (i.e. the main goal of using WGS as the genotyping method for surveillance). As future perspectives, based on our current application of both the optimized wet-lab procedures and bioinformatics tools as a means to large-scale long-term WGS-based routine surveillance, it is anticipated that the INNUENDO Platform V1.0 will be a major facilitator: i) for a short-term integration of WGS as the gold standard typing method for FWD pathogens in sectors other than Human Health; and, ii) for an effective interoperability between the food, veterinary and human health sectors with expected long-term benefits for food safety and public health protection on behalf of the "One Health" concept.

## 8.8. Partner 7: Veterinary and Food Laboratory (VFL)

Prior to taking part in the INNUENDO Project, the Estonian Veterinary and Food Laboratory (VFL) had no experience with next generation sequencing (NGS) or with bioinformatics involved in subsequent data analysis. As taking part in the INNUENDO project was our first contact with the workflow of analysing WGS data, this was a very useful experience for us and gave us the knowledge and experience to undertake WGS sequencing in our laboratory for the first time.

Partly due to our inexperience in WGS, we were a suitable candidate for testing the usability of the INNUENDO Platform by someone who has little or no prior experience. As our contribution we took part in the Summer Course "Genomics in food-borne pathogen surveillance and outbreak investigation" (Section 5.1). During the meeting, we were able to test different aspects of the platform and were asked to highlight any issues or inconveniences, which we believed should be corrected.

In September 2017, VFL organized an INNUENDO project related closed round-table discussion in Tallinn, Estonia. The main speaker on behalf of INNUENDO consortium was Jani Halkilahti from THL (Partner 4). The INNUENDO project was introduced, and the bioinformatics platform capabilities were demonstrated to the Estonian authorities, which included representatives form the Ministry of Rural Affairs, Veterinary and Food Board, Health Board, Veterinary and Food Laboratory and Estonian

University of Life Sciences. The event fulfilled its purpose as participants learned about the platform's capabilities and the benefits of using it.

After this, in February 2018, we took part in the online INNUENDO international simulation training session (Section 5.3), which focused on typing and outbreak investigation of Verocytotoxin-producing *Escherichia coli*. This meeting gave us the necessary experience with the platform to take part in the soon to follow international simulation aimed at validating the usability of the platform. During the simulation, we again were able to highlight any issues regarding the performance of the platform and gained further experience in handling WGS data for typing in the context of surveillance and outbreak investigation. In conclusion, we have found the developed platform to be very user friendly and easy to learn, for instance in comparison to command line tools. It incorporates all necessary tools for routine WGS data analysis and allows quickly determine visualize and determine possible outbreaks.

## 8.9. Partner 8: Institute of Food Safety, Animal Health and Environment (BIOR)

Institute of Food Safety, Animal Health and Environment "BIOR" acquired its Illumina MiSeq instrument in 2016 and its staff have been adopting next generation sequencing (NGS) technology since. Any NGS workflow consists of two major parts – sample preparation and sequencing (so called wet-lab) and sequence data analysis (dry-lab). Any person who has received education and training in molecular biology can learn the principles of NGS technologies and master the various wet-lab protocols quite easily. The principles how whole genome sequence data from pathogenic organisms can be used in clinical and epidemiological settings are also relatively easy to grasp. However, even with understanding of the underlying principles, it takes a lot of effort and additional training to become truly proficient in bioinformatic analysis of sequence data (dry-lab part). Therefore, it is a common situation that a lab is capable of efficiently generating high quality sequence data but data analysis is the main bottleneck in NGS workflows and research projects. This has also been the case in BIOR for the first years of implementing NGS.

Participation in this project has given our staff some valuable insights about sequencing and bioinformatics during workshops and communication throughout the project. As a food safety institution, BIOR sees immediate benefit in the main result of this project – the software platform for integration of genomic data in food-borne pathogen surveillance. Through an international simulation of surveillance of *Escherichia coli* and outbreak detection (Section 5.3) it was demonstrated that this platform is fit for the intended purpose.

Overall, the project has provided additional knowledge and skills regarding the application of NGS in food-borne infection outbreak situations. A lot was learned during the simulation and workshops that will help to ensure more successful operations in BIOR during future outbreaks. As mentioned earlier, sequence data analysis has been the bottleneck for NGS based workflows in BIOR. Even though the INNUENDO Platform is not applicable for many pathogenic organisms yet, it can streamline the genomic outbreak investigation process in many cases. As the awareness of other public health and food safety related institutions about the potential of NGS technology is increasing while not fully understanding the limitations, they expect rapid and comprehensive answers from an NGS-capable laboratory in outbreak situations. Application of the INNUENDO Platform in such cases will provide labs with the tool to meet most of these expectations in a relatively straightforward way.

As a participant of the project, BIOR took part in the course "Genomics in food-borne pathogen surveillance and outbreak investigation" (Section 5.1) and the platform test simulation and provided useful feedback on the usability of the platform as well as reported some errors that could then be fixed in order to further improve the reliability and stability of the platform.

After the final project meeting which included a training course on administration and setup of the software platform, BIOR staff have just started to set up the platform on a local server. After that, validation and adjustment of analysis parameters with own datasets would follow to enable application in future scenarios where rapid response would be needed. As BIOR is one of the main national food security laboratories in Latvia, it is expected that our use of the INNUENDO Platform will familiarize

national food security and public health officials with the possibilities of it and will facilitate the adoption of NGS technology throughout the network of institutions involved in outbreak investigation and control, thus unlocking the true potential and benefit of NGS to public health.

Lastly, open accessibility of the software platform itself and its documentation should facilitate the dissemination and wider application of project results not only among the "insiders" (participants of the project) but among food safety laboratories and officials worldwide.

# 9. Discussion and conclusions

INNUENDO project's goals were to create standardized protocols and build a software platform to be used by all actors within public health and food safety sectors to strengthen food-borne disease surveillance and outbreak investigation. As a result, we delivered the INNUENDO Platform V1.0, a stand-alone, portable, open-source, end-to-end system for the management, analysis, and sharing of bacterial genomic data. The platform allows users to assemble analytical modules in simple-to-run species-specific protocols and ensures the reproducibility of the process through the use of a specific ontology and container technology. The modules include: genome assembly and species confirmation (INNUca, Section 3.3 and Appendix B); *in silico* typing (e.g. prediction of serotype, pathotype and resistance gene content) (Seq_typing and pathotyping, Section 3.5.1 and Appendix D); standardized species-specific phylogenetic frameworks for *S.enterica*, *E. coli*, *C. jejuni* and *Y. enterocolitica* based on an innovative gene-by-gene methodology (chewBBACA and PHYLOViZ Online 2.0, Section 3.6 and Appendix E); quality control measures from raw reads to allele calling (implemented in INNUca and chewBBACA, Sections 3.3 and 3.6); an effective standard reporting system based on FlowCraft (Section 3.7); built-in standardized communication protocols and a strain classification system enabling smooth communication during outbreak investigation (Section 3.7).

Several project tasks contributed to the development of the platform:

- The assessment of the communication flow during outbreak investigations underlined the necessary reporting implementation for the platform. Furthermore, the communication assessment helped the compilation of model protocols for outbreak investigations using the INNUENDO Platform V1.0 for molecular typing (Section 6.2).
- Also of great importance was the development of the species-specific phylogenetic framework provided users with guidelines on how to assess the epidemiological relationship between strains and the nomenclature made consistent communication of results between different stakeholders easy and possible (Section 3.6.3).
- The simulation exercises supplied proof-of-concept studies for the phylogenetic framework, nomenclature, communication framework and the platform usability, as a whole (Section 5).
- The sequencing of the 607 strains was also a critical action for the successful development of the platform. The interest for applying WGS in public health services have increased considerably since the launch of the INNUENDO project in 2016, fuelled by big sequencing and surveillance programs resulting in thousands of sequenced genomes of food-borne pathogens deposited to public databases. Although, the number of strains sequenced within the INNUENDO project may seem limited, the availability of genomes with verified epidemiological metadata has been important in the execution of the proof-of-concept studies and the sequenced strains have been used in numerous validations of the software programs designed for the platform. In addition, the strains sequenced make up a valuable proportion of the legacy dataset on which the allele call algorithm rests its functionality on. Therefore, the sequenced strains have had a significant impact in validation, schema construction and proof-of-concept studies within the INNUENDO project.

Overall, the project promoted cooperation between research, public health and food and veterinary safety sectors and gave them novel tools for supporting the sharing and the analysis of the strains, strengthening the public health and food safety through a One-Health approach. Through the use of effective classification system and high-resolution typing methodologies, and the sharing of genomic

data alongside epidemiological information, the platform is useful to promote early detection of outbreaks (as demonstrated during the simulation studies), to improve surveillance, to conduct trace-back investigations and ultimately to provide material for source attribution. The INNUENDO Platform V1.0 is an effective model for the usage of open-source software in genomic epidemiology being one of the first available open-source platform explicitly constructed to conduct standardized shared analysis in real-time between multiple users in both public health and food safety sectors.

The development of the INNUENDO Platform V1.0 as a freely available, end-to-end solution configurable with open-source software modules facilitates the use of WGS in molecular epidemiology approachable for smaller players with limited resources and bioinformatic skills. As mentioned in Section 2, in different countries or regions, resource limitations in terms of budget and/or specific scientific knowledge are challenging the implementation of wet and dry laboratory procedures for WGS (EFSA, 2015; Revez et al., 2017; Llarena et al., 2017, EFSA 2018.). Although the most obvious solution is improving funding allocation for these new technologies, this is not happening in several cases. The INNUENDO project had taken this problem in serious consideration during design and development phases. Starting from the needs of the target users, the project succeeded in providing a portable solution able to account all different needs, network and IT-specifications of different institutions. Developing any platform or product must be an iterative process consisting of platform development, platform testing, and platform development again as response to feedback from the test phase. This cycle should repeat itself several times to ensure the robustness of, in this case, the bioinformatic solutions with adjourned communication protocols. Therefore, we held several usability studies within the project; the international workshop, and the national and international simulation studies. For each test stage, the users experienced problems that were corrected pronto or prior to the next usability test. These activities not only allowed a user-centered design of the platform, but were also important training events aimed to improve the capacity in public health and food safety authorities in the use NGS technologies for surveillance and outbreak investigation of food-borne pathogens.

In addition to the Platform itself, the trainings, simulations and communication protocols added to the breadth and strength of the project, establishing INNUENDO as a full proof-of-concept achievement. The thorough work performed under the INNUENDO umbrella will hopefully influence the international community in their surveillance of food-borne pathogens and introduce a standard where every step implemented in the analysis chain and communication should be knowledge-based, validated and robust. This is especially relevant considering that it is highly unlikely that every stakeholder in this field will use the exact same platform or tool. Also, other platforms will emerge to address the gaps left by the available platforms, as was the case for the establishment of the INNUENDO Platform V1.0. Therefore, the workflow and methods used during the development phase of INNUENDO should and could be used in the creation of other platform solutions as well. Furthermore, future work will have to find ways to establish communication between platforms, not only by sharing nomenclature and databases, but also translating analysis methods between platforms in an easy and deployable way.

As we currently are closer to a validated, evidence-based, full solution for analysis of WGS in outbreak investigations and surveillance of food-borne pathogens, the concept of culture-independent diagnostic tests (CIDT) arise. Such diagnostics do not yield isolates, on which public health officials have up to now relied on to perform WGS for surveillance purposes for instance for monitoring trends related to antimicrobial resistance. One solution is to perform so-called reflex culture, i.e. culturing of a specimen in case of a positive CIDT result. However, as the INNUENDO Platform V1.0 is flexible enough to include novel analytical pipelines such as those needed by CIDT, this might not even be necessary as new analysis methods will secure that we can perform molecular epidemiology even in the absence of isolates.

To conclude, the INNUENDO project provided novel, thoroughly validated and tested pipelines for application of WGS in public health and food safety authorities, complete with a novel and communication protocol and function, significantly improving and assisting the distribution and sharing of WGS results. Based on the national and multinational simulations performed during the project, the novel phylogenetic framework, completed with a robust nomenclature followed by a set of guidelines

and methods for interpreting the epidemiological relationship between strains, was found to work very well for outbreak detection and investigation. The novel methodologies developed within the INNUENDO projects are particularly useful for those National agencies in the process of migrating from traditional typing to full WGS implementation for laboratory-based surveillance of foodborne pathogens as demonstrated for the public health and food safety agency participating in the project. Moreover, the ability to share data, analyses and nomenclature between users make the INNUENDO Platform V1.0 workflow a good candidate for future harmonization between different countries.

In addition to all these advances within the field of bioinformatics and application of useful communication tools, the experiences with the INNUENDO Platform V1.0 and WGS analysis by the different participants represents a necessary capacity building in the laboratories abilities to use NGS technologies in public health and food safety actions (as summarized in Section 7). Moreover, some INNUENDO partners have started to use the platform or its separate modules in research as well (e.g. in the area of antimicrobial resistance, phylogeography, population structure analysis, pangenome analysis) adapting it to their needs and surpassing their tasks within the project (as summarize in Section 7).

Although we have strived to provide a complete and end-to-end analysis solution for WGS application in public health surveillance and outbreak investigation, the concept of a transparent box still leaves the user with considerable power to assess and adjust the analysis done. This makes accreditation possible, as all components and processes are openly available. This also means that for the analysis to be fulfiled, there is a need for judgment and assessment by the user, which must incorporate skills such as microbiology, statistics, bioinformatics and clinical medicine, to mention a few. As most users do not harbour all these disciplines themselves, proper cooperation and communication are much needed to achieve plausible and biological relevant results.

As any available broad scope computational system, the platform itself will need maintenance and extension to novel species and bioinformatic analysis, which raises the question about future sustainability of our initiatitive. As for other open-source software platforms, tools and databases, the sustaninability of the INNUENDO Platform and its components lies in the hands of the public through the creation of a community of users. The tools developed within the remit of the project are already in use in different public health and food safety agengies, and research institutions across the globe. Therefore, an active community of users will contribute to the development of new modules for the platform as new issues, knowledge or technological developments arise in the future. We foresee that the enhancement of a massive interest from the public users could stimulate other stakeholders (such as governmental organizations and funding agencies) to allocate more funding for future long-term sustainability of this kind of platforms.

# 10. Additional Supporting Information

**Annex A** – Word file: English translation of the questionnaire submitted to Latvian, Estonian and Finnish local and central authorities.

Annex A can be found in the online version of this output ('Supporting information' section): https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1498

**Annex B** - Excel file: Sequences produced within the INNUENDO project and submitted to ENA under the accession number PRJEB27020. This table includes ENA submission numbers, alias_OR_sample_name taxon, organism, serovar, pathotype, collection_date, geographic location (country and/or sea), host scientific name.

Annex B can be found in the online version of this output ('Supporting information' section): https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1498

**Annex C** – Excel file: results of the positive assignments to species by GScompare. The table shows percentage of positive assignments for each species and the corresponding ratio (positive assignments and total number of genomes within the species).

Annex C can be found in the online version of this output ('Supporting information' section): https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1498

**Annex D** – Excel file: results of the positive assignments to genera by GScompare. The table shows percentage of positive assignments for each genus and the corresponding ratio (positive assignments and total number of genomes within the genus).

Annex D can be found in the online version of this output ('Supporting information' section): https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1498

**Annex E** – Excel file: reporting form used by the participants of the international simulation study. The file includes detailed instructions and the form used for the reporting of the 1st and 2nd batches of genomes analysed during the simulation.

Annex E can be found in the online version of this output ('Supporting information' section): https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1498

**Annex F** – pdf file: Feedback form filled by the participants of the international simulation on the INNUENDO Platform usability study.

Annex F can be found in the online version of this output ('Supporting information' section): https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1498

## References

Achtman M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol 2008; 62:53-70.

Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. PLoS Genet 2018; 14:e1007261.

Ashton PM, Nair S, Peters TM, et al. Identification of salmonella for public health surveillance using whole genome sequencing. PeerJ. 2016;4:e1752. doi: 10.7717/peerj.1752

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012; 19:455-477.

Barker D, Carriço JA, Kruczkiewicz P, Palma F, Rossi M, Taboada E. Rapid identification of stable clusters in bacterial populations using the adjusted wallace coefficient. Biorxiv 2018; Preprint:doi: https://doi.org/10.1101/2993.

Boettiger C. Docker for reproducible research, with examples from the R environment. ACM SIGOP Oper Syst Rev 2014; 49:doi: 10.1145/2723872.2723882.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30:2114-2120.

Bonham-Carter O, Steele J, Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. Brief Bioinform 2014; 15:890-905.

Campbell A, Mrazek J, Karlin S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc Natl Acad Sci U S A 1999; 96:9184-9189.

Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, Moller Aarestrup F, Hasman H. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 2014; 58:3895-3903.

Carriço JA, Crochemore M, Francisco AP, Pissis SP, Ribeiro-Goncalves B, Vaz C. Fast phylogenetic inference from typing data. Algorithms Mol Biol 2018; 13:4-017-0119-7. eCollection 2018.

Carriço JA, Silva-Costa C, Melo-Cristino J, Pinto FR, de Lencastre H, Almeida JS, Ramirez M. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant Streptococcus pyogenes. J Clin Microbiol 2006; 44:2524-2532.

CSPI (Center for Science in the Public Interest). All over the map: A 10-year review of state outbreak reporting, 2011. 2011; https://cspinet.org/resource/all-over-map-10-year-review-state-outbreak-reporting-2011.

Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, Dingle KE, Bowler IC, Jolley KA, Maiden MC. Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol 2013; 51:2526-2534.

Dallman T, Cross L, Bishop C, Perry N, Olesen B, Grant KA, Jenkins C. Whole genome sequencing of an unusual serotype of Shiga toxin-producing Escherichia coli. Emerg Infect Dis 2013; 19:1302-1304.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017; 35:316-319.

Dray S. Remote possibilities? International usability testing at a distance. Interactions 2004; 11:10-17.

ECDC (European Centre for Disease Prevention and Control). Expert Opinion on the introduction of next-generation typing methods for food- and waterborne diseases in the EU and EEA. ECDC 2015, (available at https://ecdc.europa.eu/en/publications-data/expert-opinion-introduction-next-generation-typing-methods-food-and-waterborne).

ECDC (European Centre for Disease Prevention and Control). Expert opinion on whole genome sequencing for public health surveillance. ECDC 2016 (available at https://ecdc.europa.eu/en/publications-data/expert-opinion-whole-genome-sequencing-public-health-surveillance).

EFSA (European Food Safety Authority). EFSAs 20th Scientific Colloquium on Whole Genome Sequencing of food-borne pathogens for public health protection. EFSA Supporting Publications 2015; 12. https://doi.org/10.2903/sp.efsa.2015.EN-743

EFSA (European Food Safety Authority) and ECDC (European Centre for Disease Prevention and Control). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2016. EFSA Journal 2017; 15 (12):10.2903/j.efsa.2017.5077.

EFSA (European Food Safety Authority), García Fierro R, Thomas-Lopez D, Deserio D, Liebana E, Rizzi V and Guerra B, 2018. Outcome of EC/EFSA questionnaire (2016) on use of Whole Genome Sequencing (WGS) for food- and waterborne pathogens isolated from animals, food, feed and related environmental samples in EU/EFTA countries. EFSA supporting publication 2018:EN-1432, 49 pp. doi:10.2903/sp.efsa.2018.EN-1432

FAO (Food and Agriculture Organization of the United Nations). Applications of Whole Genome Sequencing in food safety management. 2016. I5619E/1/05.16. (available at http://www.fao.org/documents/card/en/c/61e44b34-b328-4239-b59c-a9e926e327b4/)

Feijao P, Yao HT, Fornika D, Gardy J, Hsiao W, Chauve C, Chindelevitch L. MentaLiST - A fast MLST caller for large MLST schemes. Microb Genom 2018; .

Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics 2009; 10:152-2105-10-152.

Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 2011; 364:730-739.

Gardy JL and Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 2018; 19:9-20.

Gargis AS, Kalman L, Lubin IM. Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. J Clin Microbiol. 2016;54(12):2857-2865.

Gossner CM, de Jong B, Hoebe CJ, Coulombier D, European Food and Waterborne Diseases Study Group. Event-based surveillance of food- and waterborne diseases in Europe: urgent inquiries (outbreak alerts) during 2008 to 2013. Euro Surveill 2015; 20:19-28.

Grande, L., Michelacci, V., Bondì, R., Gigliucci, F., Franz, E., Badouei, M.A., Schlager, S., Minelli, F., Tozzoli, R., Caprioli, A., Morabito, S., 2016. Whole-Genome Characterization and Strain Comparison of VT2f-Producing Escherichia coli Causing Hemolytic Uremic Syndrome. Emerg. Infect. Dis. 22, 2078–2086. doi:10.3201/eid2212.160017

Griffiths E, Dooley D, Graham M, Van Domselaar G, Brinkman FSL, Hsiao WWL. Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance. Front Microbiol 2017; 8:1068.

Ingle, D.J., Tauschek, M., Edwards, D.J., Hocking, D.M., Pickard, D.J., Azzopardi, K.I., Amarasena, T., Bennett-Wood, V., Pearson, J.S., Tamboura, B., Antonio, M., Ochieng, J.B., Oundo, J., Mandomando, I., Qureshi, S., Ramamurthy, T., Hossain, A., Kotloff, K.L., Nataro, J.P., Dougan, G., Levine, M.M., Robins-Browne, R.M., Holt, K.E., 2016. Evolution of atypical enteropathogenic E. coli by repeated acquisition of LEE pathogenicity island variants. Nat. Microbiol. 1, 15010. doi:10.1038/nmicrobiol.2015.10

Inns T, Ashton PM, Herrera-Leon S, et al. Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of Salmonella enteritidis. Epidemiol Infect. 2017;145(2):289-298. doi: S0950268816001941 [pii].

Jolley KA and Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 2010; 11:595-2105-11-595.

Jones TF, Imhoff B, Samuel M, Mshar P, McCombs KG, Hawkins M, Deneen V, Cambridge M, Olsen SJ, Emerging Infections Program FoodNet Working Group. Limitations to successful investigation and reporting of food-borne outbreaks: an analysis of food-borne disease outbreaks in FoodNet catchment areas, 1998-1999. Clin Infect Dis 2004; 38 Suppl 3:S297-302.

Jones TF, Rosenberg L, Kubota K, Ingram LA. Variability among states in investigating food-borne disease outbreaks. Food-borne Pathog Dis 2013; 10:69-73.

Kinnula S, Hemminki K, Kotilainen H, Ruotsalainen E, Tarkka E, Salmenlinna S, Hallanvuo S, Leinonen E, Jukka O, Rimhanen-Finne R. 2018. Outbreak of multiple strains of non-O157 Shiga toxin-producing and enteropathogenic Escherichia coli associated with rocket salad, Finland, autumn 2016. Euro Surveill. 23(35).

Kovanen S, Kivistö R, Llarena A-, Zhang J, Kärkkäinen U-, Tuuminen T, Uksila J, Hakkinen M, Rossi M, Hänninen ML. Tracing isolates from domestic human Campylobacter jejuni infections to chicken slaughter batches using whole-genome multilocus sequence typing. Int J Food Microbiol. 2;226:53-60.

Kovanen S, Kivistö R, Llarena A-, Zhang J, Kärkkäinen U-, Tuuminen T, Uksila J, Hakkinen M, Rossi M, Hänninen ML. Tracing isolates from domestic human Campylobacter jejuni infections to chicken slaughter batches and swimming water using whole-genome multilocus sequence typing. Int J Food Microbiol 2016; 226:53-60-10.1016/j.ijfoodmicro.2016.03.009.

Kovanen SM, Kivisto RI, Rossi M, Schott T, Karkkainen UM, Tuuminen T, Uksila J, Rautelin H, Hanninen ML. Multilocus sequence typing (MLST) and whole-genome MLST of Campylobacter jejuni isolates from human infections in three districts during a seasonal peak in Finland. J Clin Microbiol 2014; 52:4147-4154.

Kruczkiewicz P, Mutschall S, Barker D, Thomas JE, Domselaar GVH, Gannon VP, Carrillo CD, Taboada E. MIST: A Tool for Rapid in silico Generation of Molecular Data from Bacterial Genome Sequences. Bioinformatic 2013; 316-323.

Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012; 9:357-359.

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011; 27:2987-2993.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078-2079.

Lihong C, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. VFDB: a reference database fo rbacterial virulence factors. Nucleic Acids Res 2005; 33:D325-D328-doi: 10.1093/nar/gki008.

Llarena A-, Zhang J, Vehkala M, Välimäki N, Hakkinen M, Hänninen ML, Roasto M, Mäesaar M, Taboada E, Barker D, Garfolo G, Camma C, Di Giannatale E, Corander J, Rossi M. Monomorphic genotypes within a generalist lineage of Campylobacter jejuni show signs of global dispersion. Microb Genom 2016; Published Ahead of Print: 14 September, doi: 10.1099/mgen.0.000088Doi 10.1099/mgen.0.000088.

Llarena AK, Sivonen K, Hänninen ML. Campylobacter jejuni prevalence and hygienic quality of retail bovine ground meat in Finland. Lett Appl Microbiol 2014; 58:408-413.

Llarena AK, Taboada E, Rossi M. Whole-Genome Sequencing in Epidemiology of Campylobacter jejuni Infections. J Clin Microbiol 2017; 55:1269-1275.

Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 2013; 11:728-736.

McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O´Brian JS, Pawlowski AC, Piddock LJV, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. The Comprehensive Antibiotic Resistance Database. Antimicrob Agents Chemother 2013; 7:3348-3357-doi: 10.1128/AAC.00419-13.

Murphree R, Garman K, Phan Q, Everstine K, Gould LH, Jones TF. Characteristics of food-borne disease outbreak investigations conducted by Food-borne Diseases Active Surveillance Network (FoodNet) sites, 2003-2008. Clin Infect Dis 2012; 54 Suppl 5:S498-503.

Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, Gilpin B, Smith AM, Man Kam K, Perez E, Trees E, Kubota K, Takkinen J, Nielsen EM, Carleton H, FWD-NEXT Expert Panel. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill 2017; 22:10.2807/1560-7917.ES.2017.22.23.30544.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015; 31:3691-3693.

Palma F, Manfreda G, Silva M, Parisi A, Barker D, Taboada E, Pasquali F, Rossi M. Genome-wide identification of geographical segregated genetic markers in Salmonella enterica serovar Typhimurium 4,[5],12:i:- Sci Rep Accepted.

Pettengill, E.A., Pettengill, J.B., Binet, R., 2016. Phylogenetic Analyses of Shigella and Enteroinvasive Escherichia coli for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation. Front. Microbiol. 6. doi:10.3389/fmicb.2015.01573

Reuter S, Corander J, deBeen M, Harris S, Cheng L, Hall M, Thomson WM, McNally A. Directional gene flow and ecological separation in Yersinia enterocolitica. MGen 2015; 1:10.1099/mgen.0.000030.

Reuter S, Connor TR, Barquist L, Walker D, Feltwell T, Harris SR, Fookes M, Hall ME, Petty NK, Fuchs TM, Corander J, Dufour M, Ringwood T, Savin C, Bouchier C, Martin L, Miettinen M, Shubin M, Riehm JM, Laukkanen-Ninios R, Sihvonen LM, Siitonen A, Skurnik M, Falcao JP, Fukushima H, Scholz HC, Prentice MB, Wren BW, Parkhill J, Carniel E, Achtman M, McNally A, Thomson NR. Parallel independent evolution of pathogenicity within the genus Yersinia. Proc Natl Acad Sci U S A 2014; 111:6768-6773.

Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ, ECDC National Microbiology Focal Points and Experts Group. Survey on the Use of Whole-Genome Sequencing for Infectious Diseases Surveillance: Rapid Expansion of European National Capacities, 2015-2016. Front Public Health 2017; 5:347.

Ribeiro-Gonçalves B, Francisco AP, Vaz C, Ramirez M, Carriço JA. 2016. PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic Acids Res. 44(W1):W246-51.

Rivas L, Mellor G, Gobius K, Fegan N. Introduction to Pathogenic Escherichia coli. . Hartel RW(eds.). New York: Springer, New York, NY, 2015, 1-38.

Robins-Browne RM, Holt KE, Ingle DJ, Hocking DM, Yang J, Tauschek M. Are Escherichia coli Pathotypes Still Relevant in the Era of Whole-Genome Sequencing? Front Cell Infect Microbiol 2016; 6:141.

Rossi M, Santos Da Silva M, Ribeiro-Gonçalves BF, Silva DN, Machado MP, Oleastro M, Borges V, Isidro J, Viera L, Halkilahti J, Jaakkonen A, Palma F, Salmenlinna S, Hakkinen M, Garaizar J, Bikandi J, Hilbert F, Carriço JA. 2018a. INNUENDO whole genome and core genome MLST schemas and datasets for *Salmonella enterica*. (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1323684

Rossi M, Santos Da Silva M, Ribeiro-Gonçalves BF, Silva DN, Machado MP, Oleastro M, Borges V, Isidro J, Viera L, Halkilahti J, Jaakkonen A, Palma F, Salmenlinna S, Hakkinen M, Garaizar J, Bikandi J, Hilbert F, Carriço JA. 2018b. INNUENDO whole genome and core genome MLST schemas and datasets for *Escherichia coli*. (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1323690

Rossi M, Santos Da Silva M, Ribeiro-Gonçalves BF, Silva DN, Machado MP, Oleastro M, Borges V, Isidro J, Viera L, Barker DOR, Llarena AK, Halkilahti J, Jaakkonen A, Kivistö R, Kovanen S, Nieminen T, Hänninen ML, Salmenlinna S, Hakkinen M, Garaizar J, Bikandi J, Hilbert F, Taboada EN, Carriço JA. 2018c. INNUENDO whole genome and core genome MLST schemas and datasets for *Campylobacter jejuni*. (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1322564

Rossi M, Santos Da Silva M, Ribeiro-Gonçalves BF, Silva DN, Machado MP, Oleastro M, Borges V, Isidro J, Viera L, Halkilahti J, Jaakkonen A, Laukkanen-Ninios R, Fredriksson-Ahomaa M, Salmenlinna S, Hakkinen M, Garaizar J, Bikandi J, Hilbert F, Carriço JA. 2018d. INNUENDO whole genome and core genome MLST schemas and datasets for *Yersinia enterocolitica*. (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1421262

Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, Harmsen D, Mellmann A. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes. J Clin Microbiol 2015; 53:2869-2876.

Rusconi B, Sanjar F, Koenig SS, Mammel MK, Tarr PI, Eppinger M. Whole Genome Sequencing for Genomics-Guided Investigations of Escherichia coli O157:H7 Outbreaks. Front Microbiol 2016; 7:985.

Sheppard SK, Cheng L, Meric G, de Haan CP, Llarena AK, Marttinen P, Vidal A, Ridley A, Clifton-Hadley F, Connor TR, Strachan NJ, Forbes K, Colles FM, Jolley KA, Bentley SD, Maiden MC, Hänninen ML, Parkhill J, Hanage WP, Corander J. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. Mol Ecol 2014; 23:2442-2451.

Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microb Genom 2018; .

Silva MJ, Vaz C, Francisco AP, Ramirez M, Couto FM, Carriço JA. NGSOnto - keeping track of the NGS pipeline using an ontological approach and its application on molecular epidemiology. Anonymous Portugal: FCUL, 2013, .

van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, Heymann D, Burke DS. A systematic review of barriers to data sharing in public health. BMC Public Health 2014; 14:1144-2458-14-1144.

von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, Rasko DA, Joffre E, Corander J, Pickard D, Wiklund G, Svennerholm AM, Sjoling A, Dougan G. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. Nat Genet 2014; 46:1321-1326.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014; 9:e112963.

World Health Organization (WHO). Food-borne disease outbreaks : Guidelines for investigation and control. 2008; .

Yoo AB, Morris AJ, Grondona M. SLURM: Simple Linux Utility for Resource Managment. Anonymous USA: 2003, 44-60.

Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH, Taboada EN. The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. PLoS One. 2016 Jan 22;11(1):e0147101. doi: 10.1371/journal.pone.0147101

Zankari E, Hasman H, Consentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 2012; doi: 10.1093/jac/dks261.

Zhang J, Halkilahti J, Hanninen ML, Rossi M. Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy. J Clin Microbiol 2015; 53:1765-1767.

Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 2014; 30:614-620.

Zhang J, Xiong Y, Rogers L, Carter GP, French N. Genome-by-genome approach for fast bacterial genealogical relationship evaluation. Bioinformatics 2018; .

## Glossary

| Term | Definitions |
|---|---|
| ABRIcate | Mass screening of contigs for antimicrobial resistance or virulence genes (https://github.com/tseemann/abricate). It is implemented in the INNUENDO Platform V1.0 (Section 3.5.2). |
| Accessory genome | The subset of loci that is present in only a fraction of the strains of a given species. |
| Allele calling | Bioinformatics process involving the allele sequence extraction (from either reads or assemblies) by comparison with a database of possible alleles for several loci and subsequent assignment of allele identifiers. |
| Allele sequence | Total or partial sequence of a particular locus (usually a open reading frame). |
| Allelic distance | Number of allele differences between two isolates calculated from their allelic profiles (usually applied as a measure of genetic relatedness). |
| Allelic profile | The set of allele identifiers detected for a particular isolate using MLST-based schemas (e.g., traditional seven-loci MLST, cgMLST or wgMLST schemas). |
| Assembly (or assembled genome) | One or more contigs that together constitute the partial or complete genome sequence of a given strain (usually in FASTA file format). |
| chewBBACA | The allele calling engine for gene-by-gene analysis implemented in the INNUENDO Platform V1.0 (Section 3.6.1). |
| Clustering | Bioinformatics analysis for reconstructing phylogenetic relationships among isolates (from either reads or assemblies) as a means to provide useful information for clinical genomic and molecular epidemiology applications. |
| Contig | Contiguous sequence. One or more non-overlapping contigs constitute an assembled (complete or partial) genome. |
| Core genome | The subset of loci that is common to all or the large majority of the strains of a given species. |
| Core genome MLST (cgMLST) | Gene-by-gene analysis relying on allele calling of a set of loci from the core-genome. |
| Contextual information | In genomic epidemiology it refers to all data associated to a sample including: sample provenance (e.g. specimen types and sources), sample processing (e.g. DNA extraction and sequencing library construction), quality control (e.g. sequence quality and contamination detection), data analysis (i.e. bioinformatic pipelines), laboratory testing (e.g. antimicrobial susceptibility), epidemiological data (e.g. sources of exposure and risk, geographical distribution, time), clinical data. |
| Depth of coverage | The average number of times each position in the genome is represented in a read. Usually estimated by the number of nucleotides in the reads (usually after QC control) divided by the length of the assembled genome reconstructed from the same reads. |
| Docker container and image | A Docker container is a standard unit of software that packages up code and all its dependencies, so the application runs quickly and reliably from one computing environment to another. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system |

| | |
|---|---|
| | libraries and settings (from https://www.docker.com/resources/what-container). |
| FlowCraft | It is an assembler of pipelines written in nextflow for analyses of genomic data. Using the FlowCraft the administrator of the INNUENDO Platform is able to define Protocols and Workflows (https://flowcraft.readthedocs.io/). |
| Gene-by-gene analysis/methodology | Extension of the concept MultiLocus Sequence Typing (MLST) applied to whole-genome sequencing data for phylogenetic reconstruction. |
| GScompare | A fast-preliminary clustering method for bacterial genomes based on oligonucleotide frequencies (Section 3.4). |
| *de novo* assembly | Bioinformatics process of constructing one or more contigs from sequence reads in order to build the original partial or complete genome sequence. The output assembled genome is usually stored as FASTA format. |
| INNUCA | The pipeline for bacterial genome assembling and quality control of assemblies implemented in the INNUENDO Platform V1.0 (Section 3.3.2). |
| *In silico* typing | Bioinformatics processes involving the computational screening of whole-genome data (reads or assemblies) to extract/infer traditional geno- or phenotyping data (e.g., MLST types, genetic resistance determinants). |
| K-mer | Short sequences (typically below 20 bases) derived *in silico* from raw reads or contigs. |
| Metadata | In the context of genome epidemiology, it refers to a set of data that describes and gives information about sequencing data. A minimum set of metadata for a given sample have been defined for the INNUENDO Platform to include: source of the sample (i.e. Human, Animal, Food, Environment), time and place of isolation, owner of the sample, submitter information. |
| Minimum spanning tree | Tree-like representation of relationships between strains based on the distance between allelic profiles, where all the internal on the nodes also represent the strains under study. |
| Nextflow | Nextflow is a reactive workflow framework and a programming Domain-specific language (DSL) that simplifies the writing of computational pipelines with complex data (https://www.nextflow.io/). It is the workflow manager implemented in the INNUENDO Platform V1.0. |
| Pan genome | The set of all loci observed in all strains of a given species, i.e. the core genome plus the accessory genome. |
| Paralogous | Homologous sequences within the genome typically arising from a duplication genetic event. |
| Pipeline | The combination of several bioinformatics steps (including software and parameter settings) performed sequentially. |
| PostgreSQL database | It is an open source relational database management system (https://www.postgresql.org/about/). |
| Quality Control (QC) | Set of procedures aiming at identifying and correcting defects towards quality improvement. |
| Raw sequence reads (reads) | Sequence data output generated by a sequencer equipment. With short read technology, reads currently have from 100 base pairs to 300 base pairs and they are usually represented in usually in fastq file format, that represent the read sequence and the quality of each base pair in the read. |

| | |
|---|---|
| Reference-based read mapping | Bioinformatics process of aligning reads (usually after QC control) against a pre-defined reference (or set of reference) sequences. |
| ReMatCh | The reference-based read mapping tool implemented in the INNUENDO Platform V1.0 (Section 3.5). |
| SISTR | Salmonella *In Silico* Typing Resource. It the tool implemented in the INNUENDO Platform V1.0 for the prediction of *Salmonella enterica* serotype (https://figshare.com/articles/sistr_cmd_v1_0_2_serotyping_databases/6615938 ). |
| Schema | A fixed set of genome loci, including one or more reference allele sequences per locus used for gene-by-gene analysis. Schemas can enroll loci from the core (for cgMLST analysis) or pan genome (for wgMLST analysis). |
| Shared genome | The subset of loci shared by two or more strains. |
| Single nucleotide polymorphism | A variation in a single nucleotide which is found fixed in a portion of the population (usually 1%). |
| Single nucleotide variation | A variation in a single nucleotide without any limitations of frequency in the population. |
| Slurm | The Slurm Workload Manager is a free and open-source job scheduler for Linux and Unix-like kernels (https://github.com/SchedMD/slurm). |
| SFTP (SSH File Transfer Protocol) | It is a secure file transfer protocol which runs over the SSH protocol which is a method for secure remote login from one computer to another (https://www.ssh.com). |
| Strain nomenclature | Strain nomenclature is a construct devised to classify and accordingly label an isolate, placing it into a designated category within the diversity of the species (Nadon et al., 2017). |
| Trimming | Bioinformatics process of improving the quality of the raw reads by removal of adaptors, exclusion of reads below a read length threshold, and exclusion of read positions (usually the start and end of a read) below a given quality threshold. |
| Whole genome MLST (wgMLST) | Gene-by-gene analysis relying on allele calling of a set of loci from the pan genome (also called as Pan genome MLST; pgMLST). |

## Abbreviations

Abbreviations in order of appearance

| | |
|---|---|
| WGS | Whole genome sequencing |
| IT | Information Technology |
| MS | EU member state |
| ENA | European Nucleotide Archive |
| GUI | Graphical User Interface |
| NGS | Next-Generation Sequencing |
| EU | European Union |
| EFSA | European Food Safety Authority |
| STEC | Shiga-toxin producing *Escherichia coli* |
| WHO | World Health Organization |
| ECDC | European Centre for Disease Prevention and Control |
| OCT | Outbreak control team |
| EPIS-FWD | Epidemic Intelligence Information System – Food- and Water-borne diseases |
| RASFF | Rapid Alert System for Food and Feed |
| EWRS | Early Warning Response System |
| NCBI | National Centre for Biotechnology Information |
| SRA | Sequence Read Archive |
| APHL | Association of Public Health Laboratories |
| FAO | Food and Agriculture Organization of the United Nations |
| QC | Quality control |
| GDPR | General Data Protection Regulation |
| API | Application Programming Interface |
| GPS | Global Positioning System |
| HPC | High-Performance computers |
| VMs | Virtual Machines |
| OS | Operating System |
| HTS | High Throughput Sequencing |
| REST | Representational State Transfer |
| MLST | Multi-Locus Sequence Typing |
| SFTP | Secure File Transfer Protocol |
| LDAP | Lightweight Directory Access Protocol |
| EMBL-EBI | The European Bioinformatics Institute |
| rST | ribosomal Sequence Type |

| | |
|---|---|
| SNVS | Single nucleotide variations (SNVs) |
| INDELs | Small insertions and deletions |
| GbG | Gene-by-gene approach |
| CDS | Coding sequence |
| cgMLST | Core genome MLST |
| pgMLST | Pan genome MLST |
| wgMLST | Whole genome MLST |
| MST | Minimum spanning tree |
| nAWC | Neighborhood Adjusted Wallace Coefficient (nAWC) |
| AWC | Adjusted Wallace Coefficient |
| GeneEpiO | Genomic Epidemiology Ontology |
| INSDC | International Nucleotide Sequence Database Collaboration |
| DDBJ | DNA Data Bank of Japan |
| INCD | Infraestrutura Nacional de Computação Distribuida |
| CSC | IT Center for Science |
| https | Secure Hypertex Transfer Protocol |
| HHA | Human health authority |
| FAHA | Food and animal health authority |
| EPEC | Enteropathogenic *E. coli* |
| EAEC | Enteroaggregative *E. coli* |
| EIEC | Enteroinvasive *E. coli* |
| ETEC | Enterotoxigenic *E. coli* |
| HUS | Hemolytic Uremic Syndrome |

# Appendix A – Whole genome sequencing activities of the INNUENDO project

Authors: Ann-Katrin Llarena[1], Joana Isidro[2], Miguel Paulo Machado[3], Anniina Jaakkonen[4], Luis Viera[2], João Paulo Gomes[2], Cristina Correia[2], Riikka Laukkanen-Ninios[1], Maria Fredriksson-Ahomaa[1], Joseba Bikandi[5], Rosario San Millan[5], Ilargi Martinez-Ballesteros[5], Lorena Laorden[5], Javier Garaizar[5], Friederike Hilbert[6], Saara Salmenlinna[7], Marjaana Hakkinen[4], João André Carriço[3], Mirko Rossi[1], Vítor Borges[2], Mónica Oleastro[2]

[1]Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland; [2]National Institute of Health, Lisboa, Portugal;[3]Instituto de Microbiologia and Instituto de Medicina Molecular, Faculty of Medicine, University of Lisbon, Lisbon, Portugal; [4]Finnish Food Safety Authority, Evira, Helsinki, Finland; [5]Department of Immunology, Microbiology and Parasitology, Faculty of Pharmacy, University of the Basque Country, Vitoria-Gasteiz, Spain; [6]Institute of Meat Hygiene, Meat Technology and Food Science, University of Veterinary Medicine, Vienna, Austria; [7]Finnish National Institute for Health and Welfare, Helsinki, Finland

## A.1. Summary

This Appendix describes the process of data collection and genomes sequencing of four foodborne pathogens. The specific objectives were:

- Collect genomic data and associated metadata on *Campylobacter jejuni, Yersinia enterocolitica, Salmonella enterica* serovar Enteritidis, and shiga-toxin producing *E. coli* available from partner organizations
- Sequence of the genome of strains from the consortium, from outbreak and sporadic human infections, food and animal sources;
- Perform quality check and submission of the genomes to genomes databases

In addition, an overview of the genomes selected from public repositories is included.

## A.2. Introduction

This document contains the description of the activities performed within task 1.1: rationale of the selection of the strains, Whole Genome sequencing (WGS) and data submission. It describes the logic underlying the selection of the strains, the sequencing procedure and the implementation of the quality check (QC) developed (as presented in Section 3.3). Finally, it summarizes the strategy concerning genome submission to public databases and the selection of corresponding metadata.

## A.3. The INNUENDO sequencing project

The first step was the inventory of strains available from beneficiaries and third partners. A survey has been sent to all beneficiaries at the beginning of the task. Each beneficiary has been asked to list all possible strains of each selected FBP which fulfil the requirement listed in Table A1.

**Table A1** Survey of FBP strains information from beneficiaries

| | |
|---|---|
| *Campylobacter jejuni* | strains for which MLST sequence type was available isolated between 1990 and 2016 from all possible sources. |
| *Yersinia enterocolitica* | strains of serotype O:3 and/or biotype 4 isolated between 1990 and 2016 from all possible sources. |
| *Salmonella enterica* | strains from serovar Enteritidis for which subtyping (e.g. PFGE, MLVA) and epidemiological information were available, isolated between 1990 and 2016 from all possible sources. |
| Shiga-toxin producing *Escherichia coli* | strains from known serovar for which subtyping (e.g. PFGE, MLVA) and epidemiological information were available, isolated between 1990 and 2016 from all possible sources. |

If possible, information from third parties was included. For each strain the beneficiary needed to include the following metadata: country, year of isolation, source (human - assuming is gastroenteritis; animal species; food); if it is known, beneficiary should indicate if human sample was from domestic case or not. Beneficiaries have been asked to include information on available genome sequences for each species with required metadata.

After the survey was completed, strain information was used for selecting a set of strains considering also the available genome sequence in public repository. Therefore, the strain selection aimed to capture enough background diversity to ensure a correct estimation of genomic diversity within the populations of the various pathogens and, when it was possible, to secure that enough epi-linked isolates were available.

Information on publicly available genomes was retrieved during March 2016 – 2017 for *Campylobacter jejuni* from PubMLST database (https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates) and for *Enterobacteriaceae* from EnteroBase (https://enterobase.warwick.ac.uk/).

### A.3.1. *Campylobacter jejuni*

The main objective in selecting *Campylobacter jejuni* strains to be sequenced was to secure a sufficient number of isolates collected over a satisfactory wide geographical and temporal space and to increase diversity of available genomes based on the classical 7 gene MLST typing method. From each 7 gene MLST sequence types available in PubMLST database we identified the ones with less diversity in term of place and time of isolation (< 30 strains from < 2 countries collected in < 5 years) and we focused our selection based on information retrieved from partners.

Based on the availability at parner level, a total of 284 strains have been selected. However, only 279 strains were sussefully sequenced (five strains failed to be cultured or the sequencing was unsuccessful after several attempts). The data includes strains from 37 sequence types (STs) belonging to 17 clonal complexes (CCs). Samples were isolated from Finland, Austria and Portugal in the period between 1995 and 2015. The sources were feces from poultry, bovine and wild birds (135), from human afflicted with sporadic campylobacteriosis (118), and retail meat (26).

### *A.3.2. Yersinia enterocolitica*

*Yersinia enterocolitica* biotype 4 - serotype O:3 is the main cause of human yersiniosis in EU and pig seems to be the primary source (Fredriksson-Ahomaa et al., 2006). Outbreaks of yersiniosis (mainly family clusters) are rare, with reported cases being mainly sporadic. However, genomic information of this important genotype is limited to 20 strains collected from New Zealand, Australia, UK and France, and core genome analysis revealed limited genetic diversity (8 – 882 variants; Reuter et al., 2015). There is a clear lack of data on background diversity to ensure a correct estimation of genomic diversity within this pathogen. Information from a total of 155 strains of *Y. enterocolitica* biotype 4 - serotype O:3. Selection was done to ensure wider sampling possible regarding geographical origin, sample time and source host.

Based on the availability of strains at parner level, a total of 80 *Y. enterocolitica* biotype 4 - serotype O:3 strains have selected and successfully sequenced. They were collected from 11 EU countries in the period between 1999 and 2016. The sources were feces from pigs, wild boar and dogs (11), from human afflicted with sporadic yersiniosis (42) or from three outbreaks from Austria and Portugal (11), from swine retail meat (8) or other foods and water (7).

### *A.3.3. Salmonella enterica and Escherichia coli*

Due to the enormous amount of available genomes sequences for both *S. enterica* and *E. coli* (source ENTEROBASE: 76,129 and 43,590, respectively at the time of selection), the selection of the strains aimed essentially to include samples for the dominant types in Finland, Portugal, Spain and Austria from sporadic cases (including both human and food isolates representing these types). This selection

was coupled with representative isolates from historical outbreaks. The project focused in sequencing *S.* Enteritidis and Shiga-toxin producing *E. coli* (STEC).

#### A.3.3.1. *Salmonella enterica* serovar Enteritidis

A total of 129 *S.* Enteritidis strains have been selected for genome sequencing. They were collected from Finland, Portugal, Spain and Austria. In addition to sporadic human and food samples from Finland, Spain and Portugal, the following outbreaks have been selected:

- A Finnish outbreak traced back to Chinese chicken cubes, 2012
- An Austrian outbreak associated with a Chinese restaurant, 2012
- An Austrian outbreak associated with football camp, 2010
- Three local Austrian outbreaks from 2008, 2007 and 2009

#### *A.3.3.2.* Shiga-toxin producing *Escherichia coli* (STEC)

A total of 119 STEC strains have been selected for genome sequencing. Strains originated from Finland, isolated in 2002 and 2014, and from Austria, isolated in 2013 and 2015. The selection criteria for the 100 Finnish strains were: known PFGE type in cattle or environment in Finland matching with cases regardless contact with farm; Hemolytic-Uremic Syndrome (HUS) cases; cases linked to a day care and school epidemic.

Most of the Finnish strains (77.3%) were from serovar O157 and only a minority were of serotype O104, O121, O146 and O26. The collection of 22 strains from Austria was obtained from sporadic human cases and food in 2015 and food-producing animals in 2013 and includes strains of the following serotypes: O27, O91, O75, O146, O48, O5 and O26.

## A.4. DNA sample preparation, quality assessment and genome sequences

### A.4.1. Setting guidelines for DNA samples preparation and shipment

The sequence of the selected strains was performed at Partner 6 (INSA). DNA quality is of utmost importance to obtain good quality reads. Therefore, guidelines for DNA preparation and shipment to Partner 6 (INSA) were prepared and distributed to all beneficiaries. Briefly, DNA samples must fulfil the following criteria: minimum double stranded DNA quantity of 400 ng or 200 ng when shipped frozen; minimum double stranded DNA concentration of 20 ng/µL or 10 ng/µL when shipped frozen; minimum final volume of 20 µl; required absorbance ratios of A260/A280 > 1.8 and A260/A230 ≥ 2.0. The DNA concentration must be determined using a fluorometric method (e.g. Quibt) and the absorbance ratios should be measured by spectrophotometric instruments. The integrity of DNA must be checked by agarose gel electrophoresis.

### A.4.2. DNA samples reception and assessment of DNA quantity and quality

Upon receipt, all DNA samples are re-evaluated for DNA quantity by fluorometric method (Qubit dsDNA BR Assay Kit) and for quality by agarose gel electrophoresis.

### A.4.3. Whole genome sequencing

High-quality DNA samples were used to prepare sequencing libraries using the Nextera XT DNA Sample Preparation Kit. Library samples were subjected to cluster generation and paired-end sequencing (2x250 bp) on a MiSeq (Illumina Inc., San Diego, CA, USA), according to the Nextera XT DNA Library Prep Reference Guide (15031942 v01) (http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-library-prep-guide-15031942-01.pdf).

For each run, a report is prepared and comprises data from library preparations and library quality control, run data, analysis data and global appreciation on yield and quality of the reads.

## A.5. Quality check and quality assurance of the sequenced sample and submission to public repository

Quality of raw fastQ sequences are analyzed using Illumina primary quality check analysis. Successful sequences are then subjected to the INNUca pipeline (Section 3.3.2). If the sample failed primary or following INNUca analysis, it was subjected to re-sequencing.

## A.6. Overview of the genomes belonging to the INNUENDO Legacy Dataset

In addition to the genomes sequenced within the INNUENDO Sequencing Project, genomes donated by partner organizations and public available in Sequence Read Archive (SRA) or the European Nucleotide Archive (ENA) were included in the INNUENDO Legacy Dataset (see Section 3.2).

### *A.6.1. Campylobacter jejuni*

#### A.6.1.1. Additional strains from beneficiaries and partner organizations

A total of 566 genomes from benecifiary (UH and EVIRA) were included: 447 *C. jejuni* strains previously published in Kovanen et al. (2014a, b, 2016), Revez et al. (2014a, b), Zhang et al. (2015), Llarena et al. (2016), and Gacia-Sanchez et al. (2017); 92 *C. jejuni* strains isolated during the 2014 summer peak (June to October) in Finland from human cases collected in the Pori region (western Finland) and from all the positive chicken batches slaughtered in Finland in the same period (strains are idendified as DS6691283- DS6691376 in Rossi et al., 2018); 27 strains collected from Finnish fur animals (i.e. Nyctereutes procyonoides, Vulpes lagopus, Neovison) collected between 2014 and 2016 and kindly donated by Timo Nieminen (previously Ruralia-instituutti, Helsingin yliopisto). The data includes strains from 96 STs belonging to 25 CCs. Strains were isolated from Finland and Spain in the period between 1996 and 2016. The sources were feces from poultry, bovine, fur animals and wild birds (369), from human afflicted with sporadic campylobacteriosis (154), retail meat (27) and environment (i.e. water from rivers and lakes) (16).

#### A.6.1.2. Selection from public repositories

Raw sequencing reads for *C. jejuni* were retrieved from Sequence Read Archive (SRA) or the European Nucleic acid archive (ENA) (n = 7,126) in November 2016. The INNUca pipeline (Section 3.3.1) was employed to produce high-quality assembly and to perform quality check (QC). A total of 5691 samples (79% of available genomes) were sussefully assembled and fulfilled the INNUca QC check.

The dataset includes genomes from 552 STs belonging to 38 CCs. A total 564 samples were not assigned to any CCs, of which 167 samples were not typed at the ST level. Apart from ST, metadata (source PubMLST: country of origin and year of isolation) was available for only 2683 samples (47.1%). A total of 2010 samples originated from UK and were isolated between 1997 and 2014, 655 were isolated from USA between 2000 and 2016, and 18 from Malawi in 2012 and 2013. In addition, 155 samples originated from Denmark and one from South Africa but their date of isolation was not available.

### *A.6.2. Yersinia enterocolitica*

#### A.6.2.1. Additional strains from beneficiaries and partner organizations

For *Y. enterocolitica* there were no additional strains available from beneficiaries or partner organizations.

#### A.6.2.2. Selection from public repositories

Illumina raw reads from a total of 252 sequences deposited in SRA/ENA as *Y. enterocolitica* were retrieved in August 2018 using getSeqENA (https://github.com/B-UMMI/getSeqENA) and successfully assembled with INNUca v3.1 (Section 3.3). These genomes include strains which were previously classified in the phylogroups PG1-6 (Reuter et al., 2015) and classified by patho_typing (see Section 3.5.1 and Appendix D) as: non-pathogenic *Y. enterocolitica* PG1 (119), non-O:8 high pathogenic PG2 (6), O:8 high pathogenic PG2 (11), O:3 low pathogenic PG3 (31), O:5,27 low pathogenic PG4 (22), O:9 low pathogenic PG5 (36), O:1,2/O:1,2,3 low pathogenic PG6 (4). For 23 strains deposited as biotypes 1A (5), 1B (2), 3 (3), 4 (3) or unknown (10), patho_typing was unable to identity the phylogroup. Strains were isolated from 14 countries (three continents) between 1934 and 2018. For 37 strains information of either year or country of isolation was not available. The sources were feces from pet animals, ruminants, pig and wild boar (23), from human afflicted with sporadic yersiniosis (185), from swine retail meat (22) or other foods and water (22).

### A.6.3. *Salmonella enterica* serovar Enteritidis

#### A.6.3.1. Additional strains from beneficiaries and partner organizations

A total of 153 *S. enterica* serovar Typhymurium 4,[5],12:i:- were kindly donated by University of Bologna. The stains were part of a larger study recent published (Palma et al., 2018).

#### A.6.3.2. Selection from public repositories

At the time of assemblying of the INNUENDO Legacy Dataset, among the 76,129 strains available in EnteroBase, only 7,074 have been submitted with complete metadata (country, year, source or/and host). Only a small number of strains (< 10) were available from consortium countries (Finland, Austria, Spain and Portugal). Overall, a total of 4,307 publicly available draft or complete genome assemblies along with available metadata have been downloaded from public repositories (i.e. EnteroBase -https://enterobase.warwick.ac.uk/, National Center for Biotechnology Information NCBI - https://www.ncbi.nlm.nih.gov/ and The European Bioinformatics Institute EMBL-EBI - https://www.ebi.ac.uk/; accessed April 2017). The reference collection includes 1465 ser. Enteritidis, 2442 ser. Typhimurium (including all available 4,[5],12:i:- variants), and 400 of other frequently isolated serovars in Europe (EFSA and ECDC, 2016). For each of the other serovars, genomes have been selected to maintain the same proportions of genetic diversity representatives of all the diversity revealed by ribosomal MLST (rMLST; Alikhan et al., 2018) as existing in EnteroBase at the date of collection (April 2017). Strains were isolated from 108 countries between 1900 and 2017. For 987 strains information of either year or country of isolation was not available. A total of 990 strains were isolated from animal, 1887 from human patient and 509 from food, feed or environment. For 921 strains source of isolation was not available.

### A.6.4. Shiga-toxin producing *Escherichia coli*

#### A.6.4.1. Additional strains from beneficiaries and partner organizations

For *E. coli* there were no additional strains available from beneficiaries or partner organizations.

#### A.6.4.2. Selection from public repositories

At the time of assemblying of the INNUENDO Legacy Dataset, among the 43,590 strains available in EnteroBase, 10,939 have been submitted with complete metadata (country, year, source or/and host). Only a small number of strains (< 10) were available from consortium countries (Finland, Austria, Spain and Portugal). Overall, 2,218 public drafts or complete genome assemblies have been downloaded from EnteroBase in April 2017. Genomes have been selected on the basis of rMLST classification available in EnteroBase: from the same rMLST type, genomes have been randomly selected and downloaded. The number of samples for each rMLST type in the final dataset is

proportional to those available in EnteroBase in April 2017. Strains were isolated from 58 countries between 1800 and 2017. For 1,127 strains information of either year or country of isolation was not available. A total of 201 strains were isolated from animal, 835 from human patient and 141 from food or environment. For 1041 strains source of isolation was not available.

## References

Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. 2018. A genomic overview of the population structure of *Salmonella*. PLOS Genet. 14, e1007261.

European Food Safety Authority & European Centre for Disease Prevention and Control. 2016. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2015. EFSA J. 14.

Fredriksson-Ahomaa M, Stolle A, Korkeala H. 2006. Molecular epidemiology of *Yersinia enterocolitica* infections. FEMS Immunol Med Microbiol. 47(3):315-29.

García-Sánchez L, Melero B, Jaime I, Hänninen ML, Rossi M, Rovira J. 2017. *Campylobacter jejuni* survival in a poultry processing plant environment. Food Microbiol. 65:185-192.

Kovanen S, Kivistö R, Llarena AK, Zhang J, Kärkkäinen UM, Tuuminen T, Uksila J, Hakkinen M, Rossi M, Hänninen ML. 2016. Tracing isolates from domestic human *Campylobacter jejuni* infections to chicken slaughter batches and swimming water using whole-genome multilocus sequence typing. Int J Food Microbiol. 226:53-60.

Kovanen SM, Kivistö RI, Rossi M, Hänninen ML. 2014a. A combination of MLST and CRISPR typing reveals dominant *Campylobacter jejuni* types in organically farmed laying hens. J Appl Microbiol. 117(1):249-57.

Kovanen SM, Kivistö RI, Rossi M, Schott T, Kärkkäinen UM, Tuuminen T, Uksila J, Rautelin H, Hänninen ML. 2014b. Multilocus sequence typing (MLST) and whole-genome MLST of *Campylobacter jejuni* isolates from human infections in three districts during a seasonal peak in Finland. J Clin Microbiol. 52(12):4147-54.

Llarena AK, Zhang J, Vehkala M, Välimäki N, Hakkinen M, Hänninen ML, Roasto M, Mäesaar M, Taboada E, Barker D, Garofolo G, Cammà C, Di Giannatale E, Corander J, Rossi M. 2016. Monomorphic genotypes within a generalist lineage of *Campylobacter jejuni* show signs of global dispersion. Microb Genom. 2(10):e000088.

Reuter S, Corander J, de Been M, Harris S, Cheng L, Hall M, Thomson NR, McNally A. 2015. Directional gene flow and ecological separation in *Yersinia enterocolitica*. Microb Genom 1: 1-9.

Revez J, Llarena AK, Schott T, Kuusi M, Hakkinen M, Kivistö R, Hänninen ML, Rossi M. 2014a. Genome analysis of *Campylobacter jejuni* strains isolated from a waterborne outbreak. BMC Genomics. 15:768.

Revez J, Zhang J, Schott T, Kivistö R, Rossi M, Hänninen ML. 2014b. Genomic variation between *Campylobacter jejuni* isolates associated with milk-borne-disease outbreaks. J Clin Microbiol. 52(8):2782-6.

Zhang J, Halkilahti J, Hänninen ML, Rossi M. 2015. Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy. J Clin Microbiol. 53(5):1765-7.

# Appendix B – The INNUCA V3.1 modules

Authors: Miguel Paulo Machado[1], Jani Halkilahti[2], Anniina Jaakkonen[3], Diogo Nuno Silva[1], Inês Mendes[1], Yucel Nalbantoglu[4], Vitor Borges[5], Mario Ramirez[1], Mirko Rossi[4], João A Carriço[1]

[1]Instituto de Microbiologia and Instituto de Medicina Molecular, Faculty of Medicine, University of Lisbon, Lisbon, Portugal; [2]Finnish National Institute for Health and Welfare, Helsinki, Finland; [3]Finnish Food Safety Authority, Evira, Helsinki, Finland; [4]Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland; [5]National Institute of Health, Lisboa, Portugal

## B.1. Summary

This Appendix includes description, measurable outcome and criteria of acceptance for the quality check of each module belonging to INNUca V3.1.

## B.2. Description of the INNUca V3.1 modules

### B.2.1. FastQ integrity check

Occasionally transfer errors can occur, resulting in a partial transfer of the files. This module assesses file integrity before proceeding to avoid spending running time with partial/incomplete data.

#### B.2.1.1. Measurable outcome

The module also reports the Phred score used to code reads nucleotide quality and the maximum reads length.

#### B.2.1.2. Criteria of acceptance (default settings)

Only uncorrupted files (meaning complete files) are allowed to proceed.

### B.2.2. Expected coverage calculation

This module aims estimate the theoretical value of depth of coverage per sample, and it will be calculated twice during the QA/QC process (before and after trimming).

#### B.2.2.1. Measurable outcome

The number of sequenced nucleotides (raw or processed reads) divided by the expected genome size (bps).

#### B.2.2.2. Criteria of acceptance (default settings)

Expected coverage must be equal or greater than 15 time (15x).

### B.2.3. True coverage determination

Estimation of the true bacterial chromosome coverage *via* read mapping against reference core gene sequences distributed throughout the genome. Alternatively, historical 7 gene MLST are good proxy for true bacterial chromosome coverage assessment. User can customize the module by selecting the reference gene lists. For *E. coli*, *S. enterica*, *Y. entercolitica* and *C. jejuni* the reference core genes are predefined. The module use ReMatCh (https://github.com/B-UMMI/ReMatCh) as engine for the read mapping. If several heterozygous positions are found for the target schemas, it is used as an indication of presence of multiple strains of the same species in the read set.

#### B.2.3.1. Measurable outcome

Number of genes absent. Number of genes with heterozygous positions. Mean sample coverage depth of the genes present.

#### B.2.3.2. Criteria of acceptance (default settings)

The acceptance parameters for the raw sequencing reads are species dependent. Nevertheless, some thresholds are provided as reference values based on the already analyzed and tested samples.

Maximum number of absent genes: 2
Maximum number of genes with heterozygosities: 1
Minimum sample coverage depth: 25x

### B.2.4. Read quality analysis

The module use FastQC to check reads quality before (raw reads) and after quality improvement (processed read) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Since sample QC pass/fail assessment will be influenced by reads quality, the processed reads QC pass/fail assessment will take priority over the one determined using raw reads. All quality scores are reported in PHRED scale.

#### B.2.4.1. Measurable outcome

Per Base Sequence Quality, Per Sequence GC Content, Per Base N Content, Sequence Length Distribution, Overrepresented Sequences, Adapter Content, Maximum reads length and cropping positions.

#### B.2.4.2. Criteria of acceptance (default settings)

The samples will fail based on the following basis:
Per Base Sequence Quality: the value of the lower quartile of sequence quality for any base is greater than or equal to 5, and the median for any base is greater than or equal to 20.
Per Sequence GC Content: the sum of the deviations from the modeled normal distribution of GC content represents less or equal 30% of the reads.
Per Base N Content: any base position shows an N content 5% or less.
Sequence Length Distribution: any of the sequences do not have a zero length.
Overrepresented Sequences: any sequence is not found to represent more than 1% of the total.
Adapter Content: any adapter is not present in more than 5% of all reads.

### B.2.5. Read quality improvement

This module uses Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) for performing the following actions: trimming the 3' and 5' end of the reads; adapter removal; remove of low quality reads using a sliding window approach; to remove reads below specific length.

#### B.2.5.1. Measurable outcome

Trimmed reads.

#### B.2.5.2. Criteria of acceptance (default settings)

If for any reason Trimmomatic does not run successfully, or zero read pairs survived Trimmomatic cleaning, the original fastq files will be used in subsequent INNUca steps and only the first FastQC assessment will be used.

## B.2.6. Assembly module

This module uses SPAdes assembler v3.11.1 (default) and it performed a filtering step after the assembly is finished. The filtering is based on the following parameters (as default settings):

- Minimum contigs length: contigs smaller than 200 bps are excluded.
- Minimum k-mer coverage: contigs with less than 2x k-mer coverage are excluded.
- GC content: contigs with a GC content lower than 5% or greater than 95% are excluded.

### B.2.6.1. Measurable outcome

The module automatically extracts statistics such as number of contigs and number of assembled nucleotides in contigs. Two assessments are performed: one using the SPAdes raw assembly, and another one for the filtered assembly.

### B.2.6.2. Criteria of acceptance (default settings)

The module raises WARNINGS whenever a deviation to the values below is found.

Number of assembled nucleotides: the number of assembled nucleotides must be between 80% and 150% of the expected genome size. The expected genome size can be calculated based on the mean sequence length of complete chromosome sequences available for a given species, i.e., plasmid length should not be included in the "expected genome size".

Number of contigs: by default, the number of contigs must be smaller than 100 contigs per 1.5 Mbp of assembled nucleotides. However, species specific tuning is advisable.

## B.2.7. Assembly coverage filtering

The module uses a reference-mapping approach with Bowtie2 to map the reads against the draft genome assembly to calculate the depth of coverage for each contig and subsequently remove those sequences that are supported by lower mean depth of coverage. This filtering step validates the final assembly by removing possible contaminant sequences, while accurately determining the final assembly depth of coverage.

### B.2.7.1. Measurable outcome

Contig mean depth of coverage. The module removes contigs for which contig mean depth of coverage is lower than one third of the entire draft genome mean depth of coverage, or lower than 10x if the one third value is lower than 10x.

### B.2.7.2. Criteria of acceptance (default settings)

A sample will fail if minimum draft genome mean coverage depth is lower than 30x. The module raises WARNINGS whenever a deviation to the values below is found.

Percentage of reads mapped: equal or higher than 95%.

Number of assembled nucleotides: the number of assembled nucleotides, after applying the depth of coverage filter, should be between 80% and 150% of the expected genome size.

Number of contigs: by default, the number of contigs must be smaller than 100 contigs per 1.5 Mbp of assembled nucleotides. However, species specific tuning is advisable.

## B.2.8. Assembly correction

This module uses the read mapping results from module "Assembly coverage filtering" and Pilon software to identify inconsistencies between the input genome and the evidence in the reads and to correct the final assembly.

#### B.2.8.1. Measurable outcome

Number of changes made by Pilon. Number of contigs in which Pilon had made changes to the SPAdes assembly. Corrected assembly.

#### B.2.8.2. Criteria of acceptance (default settings)

Although some metrics can be obtained with this module, it is not possible to determine clear thresholds for assessing assembly quality.

### B.2.9. Species confirmation and Contamination test

This module uses MLST 2.0 software (https://github.com/tseemann/mlst) for scanning the sequence files against PubMLST MLST schemes using NCBI BLAST+ blastn software allowing the determination of a strain MLST sequence type (ST). MLST software is run with auto-detection mode enabled to return the scheme from which the SPAdes contigs are most likely to belong to.

#### B.2.9.1. Measurable outcome

MLST scheme found using the assembly sequences. Depending of the species provided and the MLST scheme found, different quality control status can be obtained. If a contamination with different species occurs, multiple MLST alleles of different schemas are commonly found which impairs the assignment to a unique schema.

#### B.2.9.2. Criteria of acceptance (default settings)

If there is a scheme for the expected species and a different MLST scheme is found, that sample fails quality control. That sample should be inspected for mislabeling (for example). The module raises WARNINGS whenever a MLST scheme is found for an expected species with unknown MLST scheme.

# Appendix C – Fast preliminary clustering method based on oligonucleotide frequencies: GSCOMPARE

Authors: Joseba Bikandi, Rosario San Millan, Ilargi Martinez-Ballesteros, Lorena Laorden, Javier Garaizar

Department of Immunology, Microbiology and Parasitology, Faculty of Pharmacy, University of the Basque Country, Vitoria-Gasteiz, Spain

## C.1. Summary

This Appendix describes a method for fast speciation and clustering based on analysis of k-mer (octanucleotides) content of genomes.

The content of the Appendix is divided into 4 sections:

a. Introduction.

b. Objectives. This section describes the oligonucleotide based comparison of genomes, and it presents the rationale of using k-mer distance for fast clustering of bacterial species.

c. The Genomic Signature Difference. This section describes the algorithm developed to compute k-mer distance, its optimization and implementation.

d. Comparison of Genomic Signature Difference among Prokaryotic genomes. In this section the algorithm is used to confirm its value to determinate the species of query genomes and the capacity of the proposed method to group together member of the same *Escherichia coli* and *Salmonella enterica* serotypes.

## C.2. Introduction

This document describes the methods developed to compute and compare oligonucleotide (also known as a k-mer) content of assembled genomes. The aim is to design a powerful method to assign a genome at the species level. The tool is named GScompare. We have generated an online service at http://gscompare.ehu.eus/ to demonstrate the utility of the method. A video (http://gscompare.ehu.eus/?video) is available to check the features accessible in this service.

## C.3. Objectives

### C.3.1. Basic concepts of k-mer analyses

It is widely known that a bacterial genome is not a random sequence of bases. It is also known that a specific oligonucleotide may be present in the genome in higher or lower frequencies than expected by chance alone from its nucleotide composition. The comparison of oligonucleotide frequencies began quite long ago (Burge et al., 1992) with two basic purposes: to determine phylogenetic relationships (Deschavanne et al. 1999; Fértil et al., 2005, Teeling et al., 2004, Takahashi et al., 2009) and to search Horizontal Gene Transfer events (Dufraigne et al., 2005). In most cases, very short oligonucleotides were searched (often tetranucleotides). Similar frequencies were observed in phylogenetically related genomes and this similarity was especially important for genomes within the same species, genera or family (Wang et al., 2005, Karamichalis et al. 2015).

We found that oligonucleotides over four bases long could be used to cluster phylogenetically related genomes, and we hypnotized that k-mer based clustering might be a good methodology to perform microbial speciation (i.e. classify a strain within a valid described bacterial species). Particularly, 8 bases long oligonucleotides showed to be good candidates for fast clustering at species level when comparing publicly available genomic sequences. Longer k-mers might result in increased computations, compromising performance and speed of the analysis.

### C.3.2. Searching for a method to compare oligonucleotide frequencies

Several methods are available to compute and compare oligonucleotides. To find the most suitable method to compare and cluster prokaryotic genomes using these strategies, different oligonucleotide frequencies were computed for different k-mer lengths. We aimed to identify which frequency normalization method and which comparison method for computing distance was more efficient to correctly assign randomly selected subsequences from 1,024 genomes to the correct type or species. The following oligonucleotide frequencies were computed: i) raw frequencies; ii) standardized frequencies (computed from raw frequencies as described by Wang et al., 2005); iii) Zero'th Order Marchov Chain frequencies (ZOM), First Order Marchov Chain frequencies (FOM), and Second Order Marchov Chain frequencies (SOM) for tetranucleotides (by using the notation by Bohlin & Skjerve, 2009); iv) z-scores of tetranucleotides as described by Teeling et al. (2004). When comparing the frequencies, four statistical procedures were used: i) Pearson's distance; ii) Euclidean distance; iii) the Genomic Signature Difference as described by Campbell et al. (1999); iv) Weighted Pearson's distance as described by Almeida et al. (2001). Clustering was performed with UPGMA. We used metagenomic data from Rich et al. (2011) in our experiments, which allowed us to determine whether the different methods were able to cluster the samples as theoretically expected.

After those experiments we concluded that:

- The Genomic Signature Difference is the best method to compare oligonucleotide frequencies,
- Oligonucleotide frequencies must be calculated for both DNA strands for better assignment,
- The longer the subsequence and the length of the oligonucleotide, the better the assignment of sub-sequences to their genomes

Further information related to the selection of The Genomic Signature Difference as the best statistical method may be obtained at http://gscompare.ehu.eus/docs/03_Assingment_of_sequences.pdf (25 pages document).

## C.4. The Genomic Signature Difference

This section will explain the algorithm and the optimization steps followed to speed up the computing.

### C.4.1. Algorithm description

The Genomic Signature Difference *d* between array $(X_i)_{(1\le i\le n)}$ and array $(Y_i)_{(1\le i\le n)}$ was defined by Campbell et al. (1999) only for dinucleotides with the following formula:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|f'_{xi} - f'_{yi}\right|$$

Where $f'_{xi}$ and $f'_{yi}$ are the standardized frequencies of *i*-th oligonucleotide in arrays X and Y, and n is equal to $4^k$ (number of k bases long oligonucleotides).

Wang et al. (2005) applied the same formula for longer oligonucleotides and they named it as the Hamming distance, but as the Hamming distance is usually defined for strings or vectors, and only accounts for the number of positions they differ, we will name the statistical as Genomic Signature Difference.

The oligonucleotide occurrences obtained from DNA sequences must be standardized by using the following formula:

$$f'_{xi} = \frac{\mathrm{n}\, f_{x\,i}}{\sum_{i=1}^{n} f_{x\,i}} \qquad f'_{yi} = \frac{\mathrm{n}\, f_{y_i}}{\sum_{i=1}^{n} f_{y_i}}$$

where $f_{xi}$ and $f_{yi}$ are the number of occurrences of *i*-th and *j*-th oligonucleotides within the arrays $X_i$ and $Y_i$. For standardized data, the sum of all values equals the number of elements.

We may merge and simplified the formulas above as:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|f'_{x\,i} - f'_{y\,i}\right| = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\mathrm{n}.fx_i}{\sum_{i=1}^{n} fx_i} - \frac{\mathrm{n}.fx_i}{\sum_{i=1}^{n} fy_i}\right|$$

$$d = \sum_{i=1}^{n}\left|\frac{fx_i}{\sum_{i=1}^{n} fx_i} - \frac{fy_i}{\sum_{i=1}^{n} fy_i}\right|$$

The new formula allows computing the Genomic Signature Difference without needing to standardize the oligonucleotide frequencies.

### C.4.2. Optimization step 1: computing and storage of oligonucleotide occurrences

Occurrences of all oligonucleotides of length k in a sequence are computed by using a sliding window of length k to the end of the sequence. Due to complementarity of the two DNA strands, occurrences of oligonucleotides in one strand may be used to compute the occurrences of oligonucleotides in both strands. To illustrate this, dinucleotide occurrences are computed in Figure C1.

In the example two types of dinucleotides are separated:

- Type I dinucleotides: the dinucleotide and its reverse complement are different, (for example AA and TT). For dinucleotides, 12 dinucleotides are type I. The number of occurrences for this type of dinucleotides in both DNA strands (i.e.: AA or TT), is equal to occurrences of the oligonucleotide and its reverse complement in one strand (p.e.: occurrences of AA and TT in one strand).
- Type II dinucleotides; the dinucleotide and its reverse complement are identical (for example, AT). For dinucleotides, 4 dinucleotides are type II. For type II dinucleotides (dinucleotides AT, CG, GC and TA), the number of occurrences in both strands is twice the number of occurrences in one strand.

**a)DNA sequence**

**Strand A: 5'-GACTCAGGCGTTTAGCCTGGAAGCCGCATCGCCTATCACC-3'**
**Strand B: 3'-CTGAGTCCGCAAATCGGACCTTCGGCGTAGCGGATAGTGG-5'**

| b) Dinucleotide | Strand | | |
|---|---|---|---|
| | A | B | A+B |
| AA | 1 | 2 | 3 |
| AC | 2 | 1 | 3 |
| AG | 3 | 3 | 6 |
| AT | 2 | 2 | 4 |
| CA | 3 | 1 | 4 |
| CC | 4 | 2 | 6 |
| CG | 3 | 3 | 6 |
| CT | 3 | 3 | 6 |
| GA | 2 | 3 | 5 |
| GC | 5 | 5 | 10 |
| GG | 2 | 4 | 6 |
| GT | 1 | 2 | 3 |
| TA | 2 | 2 | 4 |
| TC | 3 | 2 | 5 |
| TG | 1 | 3 | 4 |
| TT | 2 | 1 | 3 |

**c)Occurrence of dinucleotides in both DNA strands:**

| | |
|---|---|
| **Type I** | |
| **AA/TT** | 3 |
| **AC/GT** | 3 |
| **AG/CT** | 6 |
| **CA/TG** | 4 |
| **CC/GG** | 6 |
| **GA/TC** | 5 |
| **Type II** | |
| **AT** | 4 |
| **CG** | 6 |
| **GC** | 10 |
| **TA** | 4 |

**Figure C1:** Example of DNA sequences (a), dinucleotide occurrences in the DNA strands (b), and summarized dinucleotide occurrences discerning Type I and Type II oligonucleotides (c)

When this strategy is applied to longer oligonucleotides, the storage requirements are reduced according to Table C1. This approach requires controlling properly the order of the oligonucleotides in the database, but is also allows fast computing of the Genomic Signature Difference (or other distances, as for example Pearson`s distance or Euclidean) as described in Table C1.

**Table C1**: Storage requirements to save k bases long oligonucleotide

| Oligonucleotide length (k) | All oligonucleotides ($4^k$) | Type I oligonucleotides | Type II oligonucleotides | Types I+II |
|---|---|---|---|---|
| 2 | 16 | 6 | 4 | 10 |
| 3 | 64 | 32 | - | 32 |
| 4 | 256 | 120 | 16 | 136 |
| 5 | 1,024 | 512 | - | 512 |
| 6 | 4,096 | 2,016 | 64 | 2,080 |
| 7 | 16,384 | 8,192 | - | 8,192 |
| 8 | 65,536 | 32,640 | 256 | 32,896 |

Note: When k is an odd number, no type II oligonucleotides exist. The occurrences are shown in last column when data from type I and type II oligonucleotides are discerned.

### C.4.3. Optimization step 2: fast computing of the Genomic Signature Difference

To optimize computation, two approaches were applied: precomputing of the sum of all oligonucleotide occurrences, and adaptation of the formula to be used with type I and II oligonucleotides. The sum of all oligonucleotide occurrences in array X ($\Sigma f_{xi}$) and array Y ($\Sigma f_{yi}$) may be considered constants. Those values may be computed while computing oligonucleotide occurrences, and they may be stored in a database.

A new formula for the Genomic Signature Difference:

$$d = \sum_{i=1}^{n} \left| \frac{fx_i}{sumX} - \frac{fy_i}{SumY} \right|$$

where sumX and sumY are precomputed constant values for Σfxi and Σfyi that are not computed each time they are required.

Additionally, the Genomic Signature Difference is modified so that type I and type II oligonucleotides described above are used separately:

$$d = 2\sum_{i=1}^{m} \left| \frac{fx_i}{sumX} - \frac{fy_i}{SumY} \right| - \sum_{i=1}^{p} \left| \frac{fx_i}{sumX} - \frac{fy_i}{SumY} \right|$$

where m is the number of type I oligonucleotides, and p is the number of type II oligonucleotides.

For example, for octa nucleotides, m will correspond to 32,640 octa nucleotides and p to 256. The first part of the formula is multiplied by two to include in the formula the occurrences of the reverse complement oligonucleotides that were not saved to the database (but values are identical to the saved ones). Consequently, to compute octa nucleotide-based distances, 32,896 oligonucleotides occurrence pairs will be used in the formula, and this is an important saving comparing to usage of 65,536 pairs of values.

### C.4.4. Implementation c programs for fast computing

It is generally accepted that interpreted scripting languages are slower than compiled languages using binary data. Additionally, data stored in the computer as binary often requires less storage space and manipulation, which also allows reducing computing time and requirements. Due to those reasons two basic scripts were computed in c:

A c script that computes oligonucleotide content from fasta files containing genomic information (contigs, scaffolds or closed chromosomes). The program reads the fasta file containing the sequences, goes along the sequences ones to compute frequencies, computes the frequencies in both strands and saves the data as described above to save storage space. The programs computes oligonucleotide frequencies for a 5 MB genomes in <0.1 seconds.

A c script that computes genomic signature distance between selected genomes is <0.002 seconds per comparison.

## C.5. Speciation experiments

The aim of these experiments was to compare genomes belonging to the four species of interest to a database with genomes belonging to a wide number of species and to determinate if query genomes are similar to genomes from the same species based on oligonucleotide content.

The database used in this experiment was Ensembl Genomes Release 34 (http://ensemblgenomes.org/; December 2016), which was obtained from The International Nucleotide Sequence Database Collaboration (INSDC) archives. Release 34 included 41,610 genomes (41,198 bacteria and 412 archaea).

Briefly, octanucleotide content of genomes of *Escherichia coli, Salmonella enterica, Yersinia enterocolitica* and *Campylobacter jejuni* were compared to all genomes within the Ensembl Genomes Release 34 database and distances were computed. Distances were sorted by the most similar genomes to the query searched. In case the most similar genome to query belongs to the same species, result was accounted as positive assignment, and in case the query genome was more similar to members of a different species, the result was accounted as negative assignment.

A 100% percentage of positive assignment means that when comparing all genomes from the specified species, for each query genome the most similar genome in the database belongs to the same species.

According to this experiment, comparison of octa nucleotide content of a query genome against a database may be used with high confidence as a fast speciation method for *C. jejuni, E.coli S. enterica* and *Y. enterocolitica* (Table C2).

Our results pointed out the possible presence of misidentified genomes in the database.

Speciation with this method requires being fast in order to be potentially useful. Just as reference, the scripts developed allows computing octanucleotide content for each genome in <0.1 seconds (1 core) and comparison of each genome to all genomes in the database is performed in <0.4 seconds (1 core).

**Table C2:** Positive assignment of tested genomes

| Species | No. genomes | Positive assignment | Positive Assignment Percentage |
|---|---|---|---|
| *Campylobacter jejuni* | 152 | 150 | 98.7 |
| *Escherichia coli* | 2,653 | 2,645 | 99.7 |
| *Salmonella enterica* | 3,974 | 3,974 | 100.0 |
| *Yersinia enterocolitica* | 136 | 133 | 97.8 |

### C.5.1. Speciation experiments with *Escherichia coli*

Ensembl Genomes Release 34 includes 2,653 *E. coli*, 3 *E. albertii* and 4 *E. fergusonii* genomes. Octanucleotide content all *E. coli* genomes were compared to all genomes within the Ensembl Genomes Release 34 database and distances were computed. The sorted list of distances was recorded so that query genomes could be related to the most similar genomes in the database.

The experiment showed that for 2,645 genomes out of 2,653 genomes identified as *E. coli* in the database (99,7%), the most similar genomes were other *E. coli* genomes.

Seven genomes showed non-expected behaviour:

- Genome GCA_000935475.1 was similar to the three *E. albertii* included in the database. When comparing the *E. albertii* genomes to the database results showed that the three *E. albertii* genomes and genome GCA_000935475.1 constitute a differentiated group of genomes based on their oligonucleotide content.

- Genomes GCA_000935475.1 and GCA_000617165.2 were similar to *Shigella sonnei* genomes in the database.
- Genomes GCA_000714305.1 and GCA_000529265.1 were similar to each other and to genomes belonging to *Citrobacter freundii*.
- NOTE: genome GCA_000714305.1 was renamed as *Citrobacter freundii* in Ensembl Genomes Release 35 (April 2017)
- Genome GCA_000529815.1 was similar to genomes belonging to *Klebsiella pneumonia*.
- Genome GCA_001443095.1 was similar to genomes belonging to *Enterobacter cloacae*.

### C.5.2. Speciation experiments with *Yersinia enterocolitica*

Ensembl Genomes Release 34 includes 402 genomes from 14 *Yersinia* species, including 136 genomes identified as *Y. enterocolitica*.

The experiment showed that for 133 genomes out of 136 genomes identified as *Y. enterocolitica* in the database (97,8%), the most similar genomes are other *Y. enterocolitica* genomes.

Three genomes showed non-expected behaviour:

- Genome GCA_001106265.1 was more similar to the nine *Y. moralletii* genomes included in the database than to other *Y. enterocolitica* genomes.
- Genome GCA_001319955.1 was more similar to the three *Y. bercovieri* genomes included in the database than to other *Y. enterocolitica* genomes.
- Genome GCA_000597945.1 was similar to 12 out of 13 *Y. kristensenii* genomes included in the database. The discordant *Y. kristensenii* genome (GCA_001115185.1) was more similar based in octanucleotide content to other non-Y. *kristensenii* genomes. In the top positions of the list was also included genome GCA_000834215.1, which is identified in the database as *Yersinia frederiksenii*, but our method shown this genome was similar to the other *Y. kristensenii* genomes.

### C.5.3. Speciation experiments with *Campylobacter jejuni*

Ensembl Genomes Release 34 includes 331 genomes from 8 *Campylobacter* species, including 152 genomes identified as *C. jejuni*.

The experiment showed that for 150 genomes out of 152 genomes identified as *C. jejuni* in the database (98,7%), the most similar genomes are other *C. jejuni* genomes.

- Genome GCA_000686425.1 was very similar to *Campylobacter coli* genomes included in the database.
- Genome GCA_000172355.1 was similar to other non-jejuni genomes included in the database.

### C.5.4. Speciation experiments with *Salmonella enterica*

Ensembl Genomes Release 34 includes 3,974 *Salmonella enterica* and 3 *Salmonella bongori*.

The experiment showed that for all *S. enterica* genomes in the database the most similar genomes are other *S. enterica* genomes (100%).

### C.5.5. Speciation experiments with all genera in the database

This proposed method was applied to all genomes in Ensembl Genomes Release 34 database.

Annex C shows for each bacterial species in Ensembl Genomes Release 34 the percentage of genomes for which the octanucleotide based distance assigns the genome to another member of the same species. Only results for species with a minimum of two genomes are shown. The overall correct speciation for these four-target species was 96.98%.

Important factors that influence the results were detected:

- As suggested above, for some genomes the species identification may by incorrect.
- The list includes genomes which are not identified at the species level. Although those genomes are not used as query genomes in the experiment, they query genome is sometimes similar to them. Those cases were considered as incorrect assignments in this experiment.
- All sequences included in the database are not good quality ones. Just an example: for some enterobacteria nearly half the genomic information provided contains unresolved nucleotides.
- For same species it has been previously described that species identification is very problematic, for example for *Pseudomonas* spp., and for many members or uncommon and less searched genera with just two or a limited number of sequenced genomes, distances among members of the species is quite bigger than the ones computed for members of deeply searched species.

In Annex D the same procedure was applied at the genera level, and the overall positive genus assignment was 99.2%.

## C.6. Typing experiments

The aim of next experiments was to evaluate the power of the method to separate samples at sub-species level. To do so, we performed experiments based in the same method used for speciation, but search was performed with specific *E. coli* and *Salmonella enterica* serotypes.

### C.6.1. Typing experiments with *Escherichia coli* O157 genomes

Some genome names in Ensembl Genomes Release 34 include serotype information. For example, serotype O157 is identified for 122 out of 2,653 *E. coli* genomes. In order to know whether octanucleotide content of genomes from a specific serotype are more similar to each other than to genomes to other serotypes and species, octanucleotide content of *E. coli* serotype O157 genomes was compared to the complete Ensembl Genomes database. Distances were sorted by the most similar genomes to the query searched. In case the most similar genome to query belongs to the same *E. coli* serotype, result was accounted as positive assignment, and in case the query genome was more similar to members of a different serotypes or species, the result was accounted as negative assignment. Our hypothesis was that members of the same *E. coli* serotype will be similar and that the assignment will be positive.

The preliminary data showed that the more similar genome for 109 out of 122 *E. coli* serotype O157 genomes were genomes from the same serotype. The remaining 13 genomes were in most cases similar to genomes which do not contained serotype information in the name. After deep search in databases we discovered some of those genomes belong to strains identified as serotype O157. After introducing lost serotype information in our experiment, 120 out of 122 *E. coli* serotype O157 genomes were correctly identified.

Consequently, our results showed the overall correct typing for *E. coli* O157 genomes based in their oligonucleotide content was at least 98.4%.

### C.6.2. Typing experiments with *Salmonella* Enteritidis genomes

Identical procedure was applied to *Salmonella* Enteritidis. In Ensembl Genomes Release 34 database 189 genomes out of 3,977 *Salmonella enterica* genomes were classified as members of this serotype. Oligonucleotide content of each genome was searched against the complete database, and the more similar genomes recorded. The assignment experiment showed that in 183 out of 189 genomes serotype was correct based in the oligonucleotide approach.

- The 6 genomes for which our method was unable to identify as *S.* Enteritidis, were uploaded to SeqSero service (http://www.denglab.info/SeqSero; Access April 2017) to obtain a predicted serotype.
- Genome GCA_001102865.1 was identified as *S.* Typhi by SeqSero. Additonaly, 7 genes based MLST type was identical to many *S.* Typhi genomes in Ensembl Genomes. In fact that MLST type was only present in genomes identified as *Salmonella* Typhi. Oligonucleotide based comparison yielded the same result: Genome GCA_001102865.1 was similar to *Salmonella* Typhi genomes.
- Genome GCA_001479885.1 was identified by SeqSero as *S.* Hadar, and the same identification was provided by using our approach.
- Genome GCA_001448615.1 was identified by both methods as *S.* Dublin, and the same identification was provided by using our approach.
- NOTE: genome GCA_001448615.1 was renamed as *Salmonella* Dublin in Ensembl Genomes Release 35 (April 2017)
- Genome GCA_000505105.1, GCA_001448475.1 and GCA_000330445.1 were serotyped by Seqsero as serotype 3,10:-:- , serotype Berta (9:f,g,t:-) and serotype - 9:g,m:-. SeqSero did not recognized these genomes as Enteritidis. According to oligonucleotide approach *Salmonella* Schwarzengrund was the most similar genome to genome GCA_000505105.1, but distance was bigger than the normal distances detected between genomes belonging to the same serotype. For second and third genomes a non-serotyped genome and a *S.* Typhimurium were the most similar ones, but again, the distance was bigger than expected for members from the same serotype.
- NOTE: genome GCA_001448615.1 was renamed as *S.* Berta in Ensembl Genomes Release 35 (April 2017)

The overall correct assignment of *Salmonella* Enteritidis genomes was 100%

The results provided in this report points out the proposed method may be used for classifying a strain at species level or even at sub-type level (e.g. serotypes).

## References

Almeida JS, Carriço JA, Maretzek A, Noble PA, Fletcher M. 2001. Analysis of genomic sequences by Chaos Game Representation. Bioinformatics, 17, 429-437.

Bohlin,J. and Skjerve,E. 2009. Examination of genome homogeneity in prokaryotes using genomic signatures. PLoS One, 4, e8113.

Campbell, A., Mrazek, J., Karlin, S., 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America 96, 9184– 9189.

Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol., 16, 1391-9.

Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P. 2005. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res., 33, e6.

Karamichalis R, Kari L, Konstantinidis S, Kopecki S. 2015. An investigation into inter- and intragenomic variations of graphic genomic signatures. BMC Bioinformatics, 16, 246.

Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C, Humphrey J, Kerhornou A, Khobova J, Aranganathan NK, Langridge N, Lowy E, McDowall MD, Maheswari U, Nuhn M, Ong CK, Overduin B, Paulini M, Pedro H, Perry E, Spudich G, Tapanari E, Walts B, Williams G, Tello-Ruiz M, Stein J, Wei S, Ware D, Bolser DM, Howe KL, Kulesha E, Lawson D, Maslen G, Staines DM. 2016. Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res., 44, D574–D580.

Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO. 2004. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics, 5, 163.

Rich VI, Pham VD, Eppley J, Shi Y, DeLong EF. 2011. Time-series analyses of Monterey Bay coastal microbial picoplankton using a 'genome proxy' microarray. Environ. Microbiol., 13, 116-34.

Takahashi M, Kryukov K, Saitou N. 2009. Estimation of bacterial species phylogeny through oligonucleotide frequency distances. Genomics, 93, 525-533.

Wang Y, Hill K, Singh S, Kari L. 2005. The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene, 346, 173-85.

Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. 2015. Salmonella serotype determination utilizing high-throughput genome sequencing data. J. Clin. Microbiol., 53,1685-92.

# Appendix D – *In silico* typing using read-mapping: patho_typing and seq_typing tools

Authors: Miguel Paulo Machado[1], Jani Halkilahti[2], Mirko Rossi[3], João André Carriço[1]

[1]Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina Universidade de Lisboa, Lisboa, Portugal;

[2]Finnish National Institute for Health and Welfare, Helsinki, Finland;

[3]Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland

## D.1. Summary

This Appendix describes the implementation of ReMatCh to determine the pathotypes of *Escherichia coli* and *Yersinia enterocolitica* through patho_typing tool, and for serotype prediction of *E. coli* isolates using seq_typing tool, both read-based approaches. Both tools are implemented in the INNUENDO Platform V1.0.

## D.2. Patho_typing module: implementation of ReMatCh for pathotyping of *E. coli* and *Y. enterocolitica*

Numerous studies have defined important virulence determinants in *E. coli* (Robins-Browne et al., 2016) and *Y. enterocolitica* (Reuter et al., 2014). Although the majority of the populations are composed by nonpathogenic commensal or environmental bacteria, in both organisms there are several adapted subpopulations that have acquired specific virulence attributes and have developed the ability to cause several illnesses. For *E. coli*, the diseases caused by these subpopulations can range from gastrointestinal and urinary tract problems to central nervous system disorders, affecting even the healthiest individual (Kaper et al., 2004; Robins-Browne et al., 2016). These pathovars can be broadly classified as either diarrhoeagenic *E. coli* (DEC) or extraintestinal *E. coli* (ExPEC) (Kaper et al., 2004; Robins-Browne et al., 2016). DEC includes different groups of strains that, within each group, possess similar virulence factors (presence and/or absence of pathotype-specific virulence markers) and tend to cause similar diseases with similar pathology (Kaper et al., 2004; Robins-Browne et al., 2016). For *Y. enterocolitica* the classification in different pathogroups are based on the pathogenicity in a mouse infection model, however there is a partial congruence between pathogroups and sero-biotypes (Reuter et al., 2014). Recent work defining the phylogeny of the genus *Yersinia* subdivided *Y. enterocolitica* into six distinct phylogroups (Reuter et al., 2015, 2014) and it was clear that the phylogeny is largely congruent with serotypes and pathotypes.

In general, the classification into pathotypes must be considered operational in nature, since the definition of the different groups reflects the current knowledge and, therefore, is potentially deceptive due to the continuous evolution of bacterial populations resulting in hybrid strains or in new strains that do not comply with known categories. Nevertheless, the actual subdivision of these pathogens in pathotypes had shown to be important both clinically and epidemiologically, guiding clinical management and public health interventions (Robins-Browne et al., 2016). Moreover, although modern genomic phylogenetic framework will become the future standard in one-for-all typing system for both *E. coli* and *Y. enterocolitica* organisms, there are still needs to contextualize the results of novel epidemiological investigations within historical data.

Therefore, we developed a rapid methodology for *in silico* typing of DEC and *Y. enterocolitica* strains from raw sequence reads. For *Y. enterocolitica* this method predicts also certain serotypes and phylotypes since pathotype, serotype and phylotype are (at least, partially) correlated (Reuter et al., 2015, 2014). To our best knowledge, such methodology is still unavailable. The method is designed to accommodate the needs identified by the public health and food/veterinary authorities participating to the project INNUENDO and is focused on the types listed in Table D1 and D2. The definition of *E. coli* pathotypes is based on the classification used at the Finnish National Institute for Health and Welfare

(THL) and at the Finnish Food Safety Authority (EVIRA), largely according Nyholm, 2016. *Y. enterocolitica* classification is based on the phylogroup division as presented by Reuter and colleagues (Reuter et al., 2015, 2014).

**Table D1:** DEC types classification in use at THL and EVIRA and rules for patho_typing (Nyholm, 2016)

| Pathogroup | Acronym | Marker genes | Gene combination required |
|---|---|---|---|
| Shiga toxin-producing *E. coli* | STEC | *stx1A, stx1B, stx2A, stx2B, stx2fA, stx2fB, eae* | At least one *stx* type with/without *eae* |
| Enteropathogenic *E. coli*, typical | tEPEC | *eae, bfpA* | *eae* AND *bfpA*, NO *stx* |
| Enteropathogenic *E. coli*, atypical | aEPEC | *eae* | *eae*, NO *bfpA*, NO *stx* |
| Enterotoxigenic *E. coli* | ETEC | *eltA, eltB, sta1, sta2* | At least one |
| Enteroaggregative *E. coli* | EAEC | *aaiC, aap, aatA, aggR* | At least one |
| Enteroinvasive *E. coli* or *Shigella* spp. | EIEC | *ipaH, icsA* | At least one |
| Shigella dysenteriae serotype 1 | - | *ipaH, icsA, stx1* | At least one of *ipaH* and *icsA* AND *stx1* |

**Table D2:** Target genes, types and rules for *in silico* typing of *Y. enterocolitica* strains classification based on the phylogroups described by Reuter and colleagues (Reuter et al., 2015, 2014)

| Type definition | *fyuA*[(a)] | *spiA*(Ygt) | *myfA* | *ystA* | *ail* | *inv* | *ywrD* | *per* | *wbbU* | *wbcA* | *wzt* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| O:3 low pathogenic (PG3) | 0[(b)] | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| O:1,2/O:1,2,3 low pathogenic (PG6) | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| O:5,27 low pathogenic (PG4) | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| O:9 low pathogenic (PG5) | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| O:8 high pathogenic (PG2) | 1 | 0 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 1 | 0 |
| non-O:8 high pathogenic (PG2) | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| non-pathogenic (PG1) | 0 | 1 | 2 | 2 | 0 | 1 | 1 | 2 | 2 | 2 | 2 |

(a): f*yuA*, Pesticin receptor; s*piA*, Type III secretion system outer membrane protein; m*yfA*, Fimbrial protein; y*stA*, Heat-stable enterotoxin A; y*wrD*, Putative gamma-glutamyltransferase; a*il*, Attachment invasion locus protein; i*nvA*, Invasion protein; p*er*, perosamine synthetase; w*bbU*, O:3 specific dTDP-4-dehydrorhamnose 3,5-epimerase; w*bcA*, O:8 specific dTDP-4-dehydrorhamnose 3,5-epimerase; w*zt*, O-antigen/lipopolysaccharide ABC transporter ATP-binding protein.
(b): 0, absent gene; 1, gene present; 2, variable.

## D.2.1.Selection of the target genes

For *E. coli*, the target genes were selected based Nyholm, 2016 and available literature (Antikainen et al., 2009; Dallman et al., 2014; Grande et al., 2016; Ingle et al., 2016; Lima et al., 2013; Lluque et al., 2015; Pettengill et al., 2016; von Mentzer et al., 2014). Specifications for each target gene including target name, product annotation, accession number and locus_tag in reference genome are listed in Table D3.

For *Y. enterocolitica*, the selection of the target genes was based on pangenome analysis performed *ad hoc* with Roary (Page et al., 2015) based on the genomes published by Reuter and colleagues (Reuter et al., 2014), combined with the clustering analysis performed by Reuter and colleagues (Reuter et al., 2015) and the comparative genomic analysis performed by Garzetti and colleagues

(who identified specific genes for *Y. enterocolitica* serotypes) (Garzetti et al., 2014) and the target genes are listed in Table D2 together with the rules.

**Table D3:** List of target genes for DEC pathotyping included in the patho_typing module selected based on Nyholm (2016)

| Target name | Product | Reference genome | Accession | Locus_tag in reference genome |
|---|---|---|---|---|
| *stx1A* | Shiga toxin 1 subunit A | *Escherichia coli* O157:H7 str. Sakai chromosome | NC_002695 | ECs2974 |
| *stx1B* | Shiga toxin 1 subunit B | *Escherichia coli* O157:H7 str. Sakai chromosome | NC_002695 | ECs2973 |
| *stx2A* | Shiga toxin 2 subunit A | *Escherichia coli* O157:H7 str. Sakai chromosome | NC_002695 | ECs1205 |
| *stx2B* | Shiga toxin 2 subunit B | *Escherichia coli* O157:H7 str. Sakai chromosome | NC_002695 | ECs1206 |
| *stx2fA* | Shiga toxin 2f subunit A | *Escherichia coli* stx2fA, stx2fB genes for Shiga toxin 2f A subunit, Shiga toxin 2f B subunit, complete cds, serovar: O128:HNM, strain: O1-1 | AB499813.1 | - |
| *stx2fB* | Shiga toxin 2f subunit B | *Escherichia coli* stx2fA, stx2fB genes for Shiga toxin 2f A subunit, Shiga toxin 2f B subunit, complete cds, serovar: O128:HNM, strain: O1-1 | AB499813.1 | - |
| *eae* | Intimin in LEE PAI | *Escherichia coli* O26:H11 str. 11368 chromosome | NC_013361 | ECO26_5280 |
| *bfpA* | Bundle forming pilus (BFP) subunit A | *Escherichia coli* plasmid EAF RepI (repI), Rsv (rsv) genes and bundle formingpilus (BFP) locus, complete cds | U27184.1 | - |
| *eltA* | Heat-labile enterotoxin A chain precursor | *Escherichia coli* ETEC H10407 plasmid p666 | NC_017722 | ETEC_RS29400 |
| *eltB* | Heat-labile enterotoxin B chain precursor | *Escherichia coli* ETEC H10407 plasmid p666 | NC_017722 | ETEC_RS26420 |
| *sta1* | ST-IA family heat-stable enterotoxin | *Escherichia coli* ETEC H10407 plasmid p666 | NC_017722 | ETEC_RS29410 |
| *sta2* | Heat-stable enterotoxin ST-I group b | *Escherichia coli* ETEC H10407 plasmid p948 | NC_017724 | ETEC_RS29525 |
| *aaiC* | Type VI secretion system in aai PAI, secretion protein | *Escherichia coli* 042 chromosome | FN554766 | EC042_4564 |
| *aggR* | Transcriptional activator | *Escherichia coli* 042 plasmid pAA | FN554767 | EC042_pAA052 |
| *aatA* | AatA outermembrane protein, ABC transporter | *Escherichia coli* O104:H4 str. 2009EL-2050 plasmid pAA-09EL50 | CP003299.1 | O3M_26392 |
| *aap* | Dispersin | *Escherichia coli* 042 plasmid pAA | FN554767 | EC042_pAA055 |
| *ipaH* | Invasion plasmid antigen | *Shigella sonnei* Ss046, complete genome | CP000038.1 | SSON_0751 |
| *icsA* | Synonyme VirG; outermembrane protein | *Shigella dysenteriae* Sd197 plasmid pSD1_197 | CP000035.1 | SDY_P214 |

### D.2.2.Rules

Since ReMatCh behaves differently depending of the type and level of heterogenety of each single gene, thresholds for the percentage of gene sequence covered and sequence indentity were selected in a species-specific manner in order to adjust the different degrees of sequences diversity found between isolates of these organisms. For *E. coli* pathotyping a gene was defined as present if at least 60% of the gene length was covered and it has at least 70% of nucleotide identity, while for *Y. enterocolitica* pathotyping a gene was defined as present if at least 80% of the gene length was mapped with at least 70% of nucleotide identity. In both cases, a minimum depth of coverage of 25x is required to consider a gene as being present. These species-specific thresholds allowed to decrese the rate of false negatives, with lower values selected in case of higher population diversity.
Based on the mapping results in terms of presence/absence of the genes, a set of rules have been designed for both species and used for the classification (Tables 2 and 3). A software code was written in Python that compiles the ReMatCh run and the rules, and returns an unique classification for the pathotype. It is called *patho_typing* and is available at https://github.com/B-UMMI/patho_typing. Briefly, *patho_typing* uses ReMatCh to map reads to a set of reference sequences and, based on the length of the sequence covered, the nucleotide identity and the depth of coverage reported, it scores those for presence or absence. According to the combination of sequences present, a pathotype is returned following the matrix rule for sequences presence/absence. Some of the sequences can be either present or absent.

### D.2.3.Validation (Section 5.4.2)

In order to validate the efficiency of *patho_typing*, raw reads passing the QC protocol (as defined in Section 3.5.1) of 655 *E. coli* strains belonging to different pathotypes were selected from the available literature (Dallman et al., 2014; von Mentzer et al., 2014, Grande et al., 2016; Ingle et al., 2016; Pettengill et al., 2016): 20 Enteroaggregative *E. coli* (EAEC), 26 Enteroinvasive *E. coli* (EIEC), 198 EPEC, 268 Enterotoxigenic *E. coli* (ETEC), 55 *Shigella* spp. and 98 STEC.

For *Y. enterocolitica*, a total of 114 pathogenic and non-pathogenic strains from Reuter et al., 2015 and Reuter et al., 2014 were selected belonging to different serotypes. For *E. coli* a strain classified as Avian Pathogenic *E. coli* (APEC) was selected as a negative control.

This methodology had a sensitivity and specificity of 99.46% (CI: 98.44%, 99.89%) and 97.78% (CI: 88.23%, 99.94%), respectively, to correctly predict pathotype for *E. coli* and 100% specificity and sensibility for *Y. enterocolitica*.

## D.3. seq_typing module: implementation of ReMatCh for the serotyping of *E. coli*

As pathotyping, the serotype prediction of pathogenic bacteria is a relevant information for epidemiologists and clinicians for a correct response to foodborne outbreaks. Several studies have shown that, for many gram-negative bacteria, it is possible to predict the serotype by detecting specific genes associated to the biosynthesis of the O-chain of the lipopolysaccharides, for example in *E. coli* or in *Salmonella* (Zhang et al., 2015) or due to the congruence between serotype and population structure, for example in *Salmonella* (Joensen et al., 2014; Yoshida et al., 2016). Several tools are available for *in silico* serotyping of several bacterial species. Among the four species targeted by the INNUENDO project, the information of serotype is relevant for *S. enterica*, *Y. enterocolitica* and *E. coli*. For *S. enterica* the INNUENDO Platform V 1.0 applies the well validated methodology called SISTR (Yoshida et al., 2016), while for *Y. enterocolitica*, the serotype prediction has been merged (at least at certain extent) within the patho_typing tool. On the contrary, for *E. coli* the serotype prediction is limited to two tools: SerotypeFinder (Joensen et al., 2014) and a second tool built-in in EnteroBase. Both tools are not stand-alone and depend on third parties, hampering their implementation in the INNUENDO platform. These two tools are based on O-specific genes determination in *E. coli* assemblies using SerotypeFinder database or its modification.

The curated dataset of O and H genes from SerotypeFinder (Joensen et al., 2014) is freely available in https://bitbucket.org/account/user/genomicepidemiology/projects/DB.
To overcome the limitation of available methodologies we have designed a stand-alone software compatible with INNUENDO platform needs and specifications that is able to use the SerotypeFinder database to predict serotype in *E. coli* by reads' mapping. The software, written in Python, is called *seq_typing* and is available in https://github.com/B-UMMI/seq_typing. By mapping the reads to SerotypeFinder database using ReMatCh, *seq_typing* decides which reference sequence is more likely to be present based on the length of the sequence covered and its depth of coverage, and returns the type associated with such sequence.

### D.3.1.Validation (Section 5.4.2)

Using EnteroBase prediction as reference methodology, the ability of *seq_typing* in predicting *E. coli* serotype has been evaluated on a large set of public available genomes. To sample several times each O and H type, up to two strains from each available O/H combination type have been selected. Raw fastq reads for a total of 2,719 samples have been downloaded from ENA or SRA using getSeqENA (https://github.com/B-UMMI/getSeqENA) and passed the QC as defined in Section 3.5.1.

Seq_typing was highly concordant with EnteroBase prediction being able to find 96% of the O-types and 98% of the H-types. For a total of 65 and 46 over 2,719 samples O-type and H-type predicted by seq_typing, respectively, was different from the prediction of EnteroBase. In 55 and 13 cases the O- and H-type was not predicted, respectively.

To validate the serotyping prediction, a set of 279 *E. coli* with web-lab validated serotype was used. Seq_typing was able to predict 94.98% of the samples correctly.

## References

Antikainen, J., Tarkka, E., Haukka, K., Siitonen, A., Vaara, M., Kirveskari, J., 2009. New 16-plex PCR method for rapid detection of diarrheagenic Escherichia coli directly from stool samples. Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol. 28, 899–908. doi:10.1007/s10096-009-0720-x

Dallman, T.J., Chattaway, M.A., Cowley, L.A., Doumith, M., Tewolde, R., Wooldridge, D.J., Underwood, A., Ready, D., Wain, J., Foster, K., Grant, K.A., Jenkins, C., 2014. An Investigation of the Diversity of Strains of Enteroaggregative Escherichia coli Isolated from Cases Associated with a Large Multi-Pathogen Foodborne Outbreak in the UK. PLOS ONE 9, e98103. doi:10.1371/journal.pone.0098103

Garzetti, D., Susen, R., Fruth, A., Tietze, E., Heesemann, J., Rakin, A., 2014. A molecular scheme for Yersinia enterocolitica patho-serotyping derived from genome-wide analysis. Int. J. Med. Microbiol. 304, 275–283. doi:10.1016/j.ijmm.2013.10.007

Grande, L., Michelacci, V., Bondì, R., Gigliucci, F., Franz, E., Badouei, M.A., Schlager, S., Minelli, F., Tozzoli, R., Caprioli, A., Morabito, S., 2016. Whole-Genome Characterization and Strain Comparison of VT2f-Producing Escherichia coli Causing Hemolytic Uremic Syndrome. Emerg. Infect. Dis. 22, 2078–2086. doi:10.3201/eid2212.160017

Ingle, D.J., Tauschek, M., Edwards, D.J., Hocking, D.M., Pickard, D.J., Azzopardi, K.I., Amarasena, T., Bennett-Wood, V., Pearson, J.S., Tamboura, B., Antonio, M., Ochieng, J.B., Oundo, J., Mandomando, I., Qureshi, S., Ramamurthy, T., Hossain, A., Kotloff, K.L., Nataro, J.P., Dougan, G., Levine, M.M., Robins-Browne, R.M., Holt, K.E., 2016. Evolution of atypical enteropathogenic E. coli by repeated acquisition of LEE pathogenicity island variants. Nat. Microbiol. 1, 15010. doi:10.1038/nmicrobiol.2015.10

Kaper, J.B., Nataro, J.P., Mobley, H.L.T., 2004. Pathogenic Escherichia coli. Nat. Rev. Microbiol. 2, 123–140. doi:10.1038/nrmicro818

Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol 52:1501–1510. doi:10.1128/JCM.03617-13

Lima, I.F.N., Boisen, N., Silva, J. da Q., Havt, A., de Carvalho, E.B., Soares, A.M., Lima, N.L., Mota, R.M.S., Nataro, J.P., Guerrant, R.L., Lima, A.Â.M., 2013. Prevalence of enteroaggregative Escherichia coli and its virulence-related genes in a case–control study among children from north-eastern Brazil. J. Med. Microbiol. 62, 683–693. doi:10.1099/jmm.0.054262-0

Lluque, A., Mosquito, S., Gomes, C., Riveros, M., Durand, D., Tilley, D.H., Bernal, M., Prada, A., Ochoa, T.J., Ruiz, J., 2015. Virulence factors and mechanisms of antimicrobial resistance in Shigella strains from periurban areas of Lima (Peru). Int. J. Med. Microbiol. IJMM 305, 480–490. doi:10.1016/j.ijmm.2015.04.005

Nyholm O. 2016. Virulence variety and hybrid strains of diarrheagenic Escherichia coli in Finland and Burkina Faso. Univeristy of Helsinki. Doctoral dissertation (article-based). http://urn.fi/URN:ISBN:978-951-51-2625-2

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693. doi:10.1093/bioinformatics/btv421

Pettengill, E.A., Pettengill, J.B., Binet, R., 2016. Phylogenetic Analyses of Shigella and Enteroinvasive Escherichia coli for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation. Front. Microbiol. 6. doi:10.3389/fmicb.2015.01573

Reuter, S., Connor, T.R., Barquist, L., Walker, D., Feltwell, T., Harris, S.R., Fookes, M., Hall, M.E., Petty, N.K., Fuchs, T.M., Corander, J., Dufour, M., Ringwood, T., Savin, C., Bouchier, C., Martin, L., Miettinen, M., Shubin, M., Riehm, J.M., Laukkanen-Ninios, R., Sihvonen, L.M., Siitonen, A., Skurnik, M., Falcão, J.P., Fukushima, H., Scholz, H.C., Prentice, M.B., Wren, B.W., Parkhill, J., Carniel, E., Achtman, M., McNally, A., Thomson, N.R., 2014. Parallel independent evolution of pathogenicity within the genus Yersinia. Proc. Natl. Acad. Sci. 111, 6768–6773. doi:10.1073/pnas.1317161111

Reuter, S., Corander, J., de Been, M., Harris, S., Cheng, L., Hall, M., Thomson, N.R., McNally, A., 2015. Directional gene flow and ecological separation in Yersinia enterocolitica. Microb. Genomics 1. doi:10.1099/mgen.0.000030

Robins-Browne, R.M., Holt, K.E., Ingle, D.J., Hocking, D.M., Yang, J., Tauschek, M., 2016. Are Escherichia coli Pathotypes Still Relevant in the Era of Whole-Genome Sequencing? Front. Cell. Infect. Microbiol. 6. doi:10.3389/fcimb.2016.00141

von Mentzer, A., Connor, T.R., Wieler, L.H., Semmler, T., Iguchi, A., Thomson, N.R., Rasko, D.A., Joffre, E., Corander, J., Pickard, D., Wiklund, G., Svennerholm, A.-M., Sjöling, Å., Dougan, G., 2014. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. Nat. Genet. 46, 1321–1326. doi:10.1038/ng.3145

Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH, Taboada EN. The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. PLoS One. 2016 Jan 22;11(1):e0147101. doi: 10.1371/journal.pone.0147101

Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. 2015. Salmonella serotype determination utilizing high-throughput genome sequencing data. J. Clin. Microbiol., 53,1685-92.

# Appendix E – The classification system within the INNUENDO Platform V1.0

Authors: Ann-Katrin Llarena[1], Mickael Santos Da Silva[2], Jani Halkilahti[3], João André Carriço[2], Eduardo Taboada[4], Mirko Rossi[1]

[1]Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland;

[2]Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina Universidade de Lisboa, Lisboa, Portugal;

[3]Finnish National Institute for Health and Welfare, Helsinki, Finland;

[4]National Microbiology Laboratory at Lethbridge, Public Health Agency of Canada, Lethbridge, Canada

## E.1. Summary

Advances in next generation sequencing (NGS) methods have led to an adoption of whole-genome sequencing (WGS) as a primary method to characterize microbial pathogens in public health services. A critical point for implementing global pathogen surveillance with WGS is the transformation of WGS-sequences data into subtypes with an associated nomenclature for definition of clusters of different genetic diversity. The typing resolution needed varies with the purpose of sequencing: while a high-resolution benefits outbreak investigation, surveillance aims to identify shifts in the bacterial population over time and requires a lower resolution level. The goal was therefore to construct a way to subtype foodborne pathogens and classify these with a nomenclature by developing organism-specific thresholds for *Campylobacter* spp., *Yersinia enterocolitica, Salmonella enterica* and *E. coli* (with focus on STEC).

We chose the gene-by-gene method of subtyping implemented in the INNUENDO platform: the chewBACCA core genome (cgMLST) schema specifically designed for *E. coli*, *S. enterica*, *Y. enterocolitica* and *Campylobacter*.

Within the INNUENDO platform, three levels of nomenclature have been specified: L1 for outbreak detection and investigation, L2 for longitudinal surveillance and L3 for congruence to other relevant subtyping (MLST). L1 was defined by examining the genetic diversity among epi-linked strains of *S. enterica* serovar Enteritidis, *E.coli, Y. enterocolitica* and *C. jejuni* relative to genomic diversity between isolates from sporadic cases. We propose that a similarity threshold of 0.3-0.4% allele differences between strains subtyped with cgMLST launched by chewBACCA is concordant with epidemiological information in all the four species of interest.

To define the L2 nomenclature, we used Neighbourhood Adjusted Wallace Coefficient (nAWC), a novel methodology to define cluster stability. By this method, the robust L2 nomenclature was created, which delineated mid-sized, stable clusters representing major lineages in the bacterial population. L2 is therefore useful in both surveillance and source attribution.

L3 represents a goeBURST threshold of cgMLST with the highest concordance with the classical 7 gene MLST classification using Adjusted Rand and Adjusted Wallace coinficient.

This adaptable, portable and useful subtyping schema and nomenclature is based on solid methodology and represents a significant leap forward in the use of WGS in public health services.

## E.2. Introduction

Molecular typing, defined as the method for clustering isolates of the same species based on their genetic and/or phenotypic relatedness, is a central component of epidemiological investigations. Molecular typing plays a key role in public health actions to control foodborne pathogens and enhance efficient communication between laboratories and different actors (microbiologists, epidemiologists, practitioners).

Advances in next generation of high-throughput sequencing (NGS) methods have led to an adoption of whole genome sequencing (WGS) as the primary method to characterize and type microbial pathogens (Nadon et al., 2017). Therefore, a critical point for implementing global pathogen

surveillance using WGS is the translation of WGS sequence data in 'plain language' (i.e. sub-types) useful in public health actions. Named subtypes enable rapid data-analysis and efficient exchange of information, thereby contributing to a rapid response to infectious disease, promoting disease prevention and control.

The definition of WGS-subtypes requires the use of standardized methodological approaches and a nomenclature to describe the relationship between isolates. One WGS-based typing is achieved by the gene-by-gene approach. The method compares genomes (complete or draft) against a predefined set of loci collected in a schema composed by all possible known variation of those loci (alleles) (Maiden et al., 2013). If the schema consists of core loci, i.e. loci present in all (100%) or the great majority of the bacterial population, the schema is referred to as a core genome multilocus sequence typing schema (cgMLST). These definitions of cgMLST are inherently changing due to the natural evolution of a bacterial species, and are operational in nature since they are based on the number of isolates analysed to date. Gene-by-gene methodology for subtyping have a great appeal due to its portable nomenclature and independence from a reference strain. As such, PulseNet International (Nadon et al., 2017), US Center for Disease Control and Prevention and the European Centre for Disease Control use these gene-by-gene approaches for bacterial discrimination routinely.

The two epidemiological settings of surveillance and outbreak investigations have different objectives. The intention of molecular surveillance is to record the types of bacterial pathogens circulating in a specific geographical area in a continuous fashion, while genomic analyses within outbreak investigations aim to identify patterns of shared variation to infer transmission and find a common source. As consequence, in outbreak investigations, clusters must be defined to include and exclude isolates, while surveillance monitoring requires a cluster-definition practical for longitudinal tracking of strains of interest. The typing resolution needed is therefore different: while a high resolution makes allows the definition of all diversity within a cluster in outbreak investigations, such a high level of diversity might result in too many subtypes for an effective surveillance. Therefore, the type of schema used in a gene-by-gene approach has important implications for the efficiency of surveillance or outbreak investigation. In addition, the process of defining clusters composed of strains likely to be related by the use of thresholds is an important consideration in the application of any subtyping scheme. The optimization of this threshold-parameter has been a perennial challenge in the field of molecular epidemiology (Llarena et al., 2017), as even small adjustments in these thresholds can have a dramatic impact on cluster composition and stability. An additional challenge of WGS-based subtyping is calibration, as WGS is sensitive to the addition of novel genome sequences.

The goal was therefore to construct a way to subtype foodborne pathogens and classify these with a higher order nomenclature by developing organism-specific thresholds for *Campylobacter* spp., *Yersinia enterocolitica*, *Salmonella enterica* and shiga-toxin producing *Esherichia coli* (STEC).

## E.3. Materials and Methods

### E.3.1. Databases

As presented in Section 3.2 and Appendix A, the genomes included in the INNUENDO Legacy Datasets were either sequenced within the INNUENDO project provided by partner organization, or collected from public repositories. All draft genomes were assembled using INNUca V3, except for the downloaded *E. coli* and *S. enterica* genomes. Detailed information on the datasets used in the analysis and schema creation can be found at https://github.com/TheInnuendoProject/chewBBACA_schemas.

### E.3.2. Definition of the schemas

Below a summary of what already presented in Section 3.6. For the scope of the INNUENDO platform, we adopted the definition of loci as proposed by Silva et al. (2018) (Section 3.6.1): only loci defined as coding sequences (CDS) can be identified as an allele. Moreover, we define the core genome MLST schema as the set of loci present in ≥ 99% of the samples, allowing up to 2% missing loci per single

genome. A higher cut-off was set for *C. jejuni*: loci present ≥ 99.9% of the samples, allowing up to 2% missing loci per single genome, were included in the cgMLST schema. This higher cut off was needed for avoiding the exclusion too many genomes which did not satisfy the 2% missing loci limit. See GitHub link provided above for details and rationales (Section 3.6.3).

The chewBBACA suite (Silva et al., 2018) was used for validating all schemas. If the original schema was obtained from a third party (i.e. EnteroBase), loci were initially curated using chewBBACA AutoAlleleCDSCuration for removing non-CDS alleles. The *de novo* schemas were based either on pangenome analysis defined by Roary with default setting (i.e. *Campylobacter*) or using chewBBACA Schema creation (i.e. *Y. enterocolitica*). For all schemas, the quality of the loci was assessed using chewBBACA Schema Evaluation, wherein loci with single alleles and high length variability (i.e. more than one allele outside the mode +/- 0.05 size) have been removed. The schema was further curated by excluding "Repeated Loci" and loci annotated as "non-informative paralogous hit (NIPH/ NIPHEM)" or "Allele Larger/ Smaller than length mode (ALM/ ASM)" by the chewBBACA Allele Calling engine (Silva et al., 2018) present in more than 1% of the respective genome datasets. Finally, the set of loci defining the cgMLST schema have been extracted (Table E1).

### E.3.3. Defining the cgMLST nomenclature

Within the INNUENDO Platform, three different levels of strain nomenclature have been specified: L1, L2 and L3. This classification system is hierarchical (i.e. L3 → L2 → L1) and it is based on goeBURST clustering methodology (Francisco et al., 2009). The most discriminatory threshold, L1, was set by investigating the concordance between genomic clustering at different thresholds of similarity in sets of epidemiologically verified outbreak isolates (i.e. cluster efficiency): four *E. coli* outbreaks, seven *S. enterica* serovar Enteritidis outbreaks, four *C. jejuni* outbreaks and three *Y. enterocolitica* outbreaks. In addition, we investigated the effect of error in allele calling related to sequencing and genome coverage on this nomenclature and its thresholds.

To define the L2 nomenclature we implemented the methodology described in Barker et al., 2018. Briefly, goeBURST (Francisco et al., 2009) was used to examine cluster membership for cgMLST profiles of the *E. coli, S. enterica* serovar Enteritidis, *C. jejuni* and *Y. enterocolitica* in our dataset (see "Databases") through a continuous range of similarity thresholds. Neighbourhood Adjusted Wallace Coefficient (nAWC) was calculated to assess cluster consolidation dynamics. nAWC uses the AWC of Severiano et al. (22) to examine the partition congruence between adjacent similarity thresholds used for cluster definition ($T(n+1) \rightarrow T(n)$). This method sets similarity thresholds that generates quasi-stable cluster configuration by identifying nearly flat areas close to nAWS close to 1.0 in the nAWS graph (Figures C1a-d). In these areas of the graph, clustering is relatively stable independently of the similarity thresholds due to production of similar partitions. Peaks deviating from one, on the other hand, indicate that those thresholds yield considerable differences in partitioning and therefore unstable clustering. It follows that we search for areas of where nAWS remains close to one for several consecutive thresholds where smaller changes in the similarity threshold for cluster definition will insignificantly affect clustering, producing a robust L2 nomenclature.

L3 was defined as goeBURST threshold of cgMLST with the highest concordance with the classical 7 gene MLST classification using Adjusted Wallace coefficient (Carriço et al., 2006).

## E.4. Results

### E.4.1. Curated whole genome schema and the definition of cgMLST

Detailed information on schema creation, validation and the genome datasets used, as well as the schema and allele profiles for each species are available for consultation and download at https://github.com/TheInnuendoProject/chewBBACA_schemas. Table E1 (which is an extract of Table 6) shows the size of the static cgMLST schema and the selected thresholds defining the three level of the strain nomenclature.

**Table E1:** Number of selected core loci and threshold defining the strain nomenclature

| Species | core loci[(a)] | L1 (%)[(b)] | L2 (%)[(b)] | L3 (%)[(b)] |
|---|---|---|---|---|
| *E. coli* | 2,360 | 8 (0.34) | 112 (4.7) | 793 (33.6) |
| *S. enterica* | 3,255 | 14 (0.43) | 338 (10.4) | 997 (30.6) |
| *Y. enterocolitica* | 2,406 | 9 (0.37) | 133 (5.5) | 1,189 (49.4) |
| *C. jejuni* | 678 | 4 (0.59) | 59 (8.7) | 292 (43.1) |

(a): number of core loci on which the nomenclature have been designed;
(b): three different levels of strain nomenclature defined based on the core loci: L1 for outbreak detection and investigation, L2 for longitudinal surveillance and L3 for congruence to classical 7 genes MLST; between brackets: the corresponding percentage of core loci.

### E.4.2. Nomenclature definition based on defined cgMLST: Level 1

For defining L1, we first investigated the level of differences observed among *E. coli* isolated from three genomic clusters detected during two distinct Finnish outbreaks in 2016 and one outbreak from the *E. coli* benchmark dataset of Timme et al. (2017). Together, the dataset contained 136 *E. coli* strains, of which approximately 100 were from sporadic cases. The outbreak strains were of two different pathotypes, namely Shiga-toxin producing *E. coli* (STEC) and Enteropathogenic *E. coli* (EPEC) and four serotypes (O157:H7, NT:H11, O111:H8 and O121:H19) and 4 MLST types (ST-11, ST-295, ST-327 and ST-655, respectively). All strains were assembled with INNUca v3.1 and the cgMLST schema was called using chewBBACA. Excluding strain IN_STEC_FI_111, the maximum allelic distance observed among outbreak strains was 8/2360 core loci (0.34%). The strain showing more than eight allelic differences belongs to a NT:H11 outbreak and lacked nine loci versus an average of three missing loci observed in the remainder of the outbreak strains. This high number of missing loci results in an increased genetic difference. Therefore, the similarity threshold for clustering of *E. coli* according to the L1 nomenclature is set to eight allele differences.

To validate the similarity threshold, the diversity between the sporadic strains and outbreak strains were evaluated. With one exception (IN_STEC_FI_111), the sporadic *E. coli* separated from the outbreak strains with at least 24 allele differences. The IN_STEC_FI_111 *E. coli* strain had six allele differences compared with a strain in the O157:H7 outbreak, but showed between nine and 15 allele differences with the other strains in this outbreak. Since the cgMLST analysis supports its close relationship to the O157:H7 outbreak and the nearest sporadic strain was 43 alleles different from the IN_STEC_FI_111, it might be that this particular strain was wrongly classified as sporadic. However, there was not sufficient epidemiological information to confirm this theory.

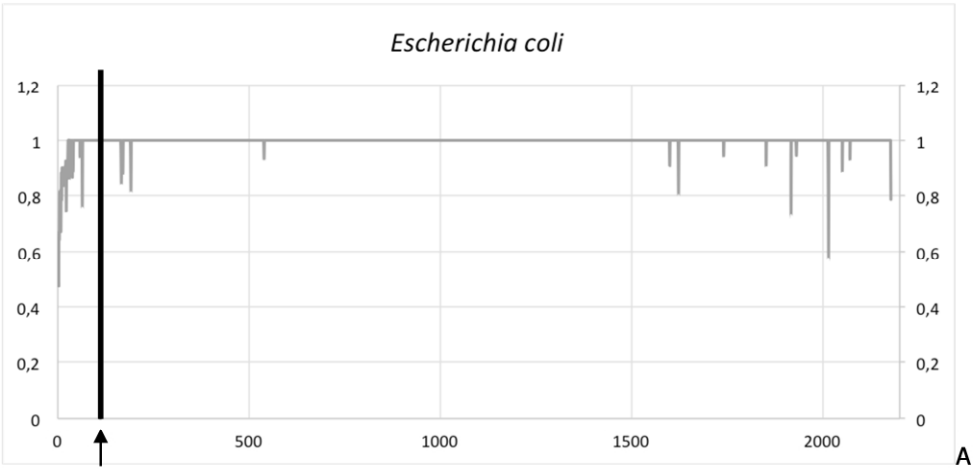
We further examined the robustness of the L1 nomenclature threshold by examining how errors in allele calling related to sequencing and genome coverage might affect clustering. First, we investigated the effect of allele calling by sequencing nine replica of the borderline O157:H7 outbreak strain (IN_STEC_FI_111) in MiSeq runs at 60x to 100x coverage. The effect of coverage on allele calling was also investigating by randomly subsampling the raw reads of two samples (SRR341556 and ERR163841) at 25x, 35x, 45x and 70x coverage (5 replicates each). Raw data were assembled using INNUca v3.1 and cgMLST schema was called using chewBBACA. A total of 8/9 IN_STEC_FI_111 replicates had no missing loci and identical cgMLST profiles, suggesting no effect of resequencing on cgMLST cluster definition. Only a single replicate with 15 missing loci resulted in increased genomic diversity. Among the subsamples themselves, replicates had up to four missing loci (median two) and three to six allelic differences, and all allelic differences were found in between the five strains of low coverage (25X). Regardless the observed differences, the replicates clustered together using the L1 threshold of 0.34% allele differences (including both missing loci and allele differences), corroborating that this similarity threshold set for *E. coli* is concordant with the epidemiological information and robust enough to be applied for cluster identification in outbreak investigations.

To verify if the cgMLST similarity threshold detected for *E. coli* (0.3-0.4% allele differences) was valid for other species, we examined the genetic diversity among epi-linked strains of *S. enterica* serovar Enteritidis*, Y. enterocolitica* and *C. jejuni*. For *Salmonella* (seven outbreaks), *C. jejuni* (four outbreaks) and *Y. enterocolitica* (three outbreaks) the genomic diversity of the outbreak strains was in

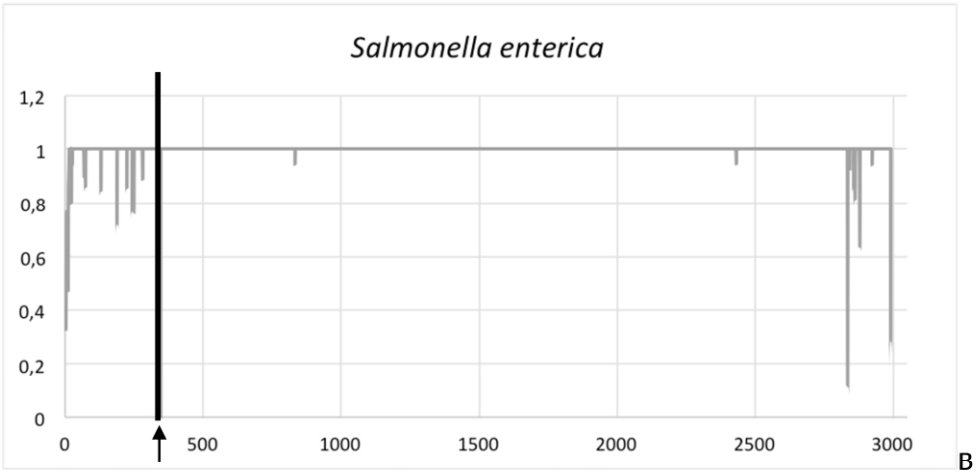
concordance with the epidemiological data at a similarity threshold of 0.43% (14/3255 loci), 0.59% (4/678 loci) and 0.37% (9/2403 loci) allele differences, respectively. Taken together, we propose that a similarity threshold of 0.3-0.4% allele differences between strains subtyped with cgMLST launched by chewBACCA are concordant with epidemiological information in all the four species of interest. Table 1 summaries the L1 cut-off proposed for each species.

### E.4.3. Nomenclature definition based on defined cgMLST: Level 2

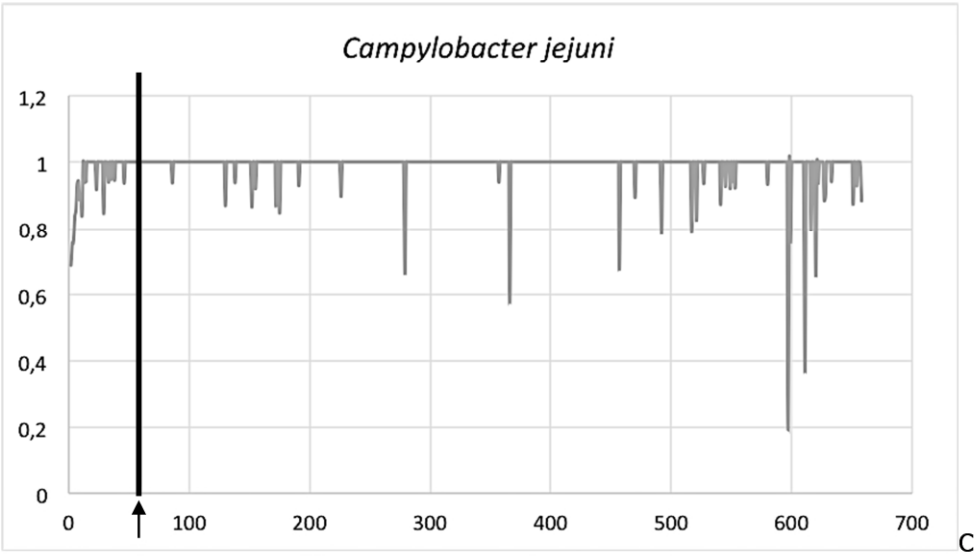
To select species-specific L2 nomenclature, the cluster stability of the strain allele profiles was investigated. Cluster stability can be thought of as the number of consecutive thresholds over which cluster congruence is high. We first examined how the number of clusters changed as a function of T (i.e. the similarity threshold), since this simple metric could be used to determine the rate of cluster consolidation (i.e. merging of clusters), a basic measure of cluster stability. For each organism we observed two distinct phases (I and II) characterized by differences in the rate of cluster consolidation (Figure E1). In the first (Phase I), an increase in the threshold T(n+1) → T(n) and to subsequent thresholds led to a steep decline in the number of clusters observed, as genomes with unique allelic profiles clustered into small groups of highly similar genomes and these further consolidated into fewer clusters of larger size. This region was characterized by successive thresholds of nAWC values deviating from unity, but converging towards this maximum value as the rate of cluster consolidation decreased. Therefore, definition of clusters in Phase I would lead to an unstable nomenclature. It is in this region we find our L1 nomenclature, but as elaborated on in the Discussion, this instability is acceptable for its intended use. We then observed a secondary phase (Phase II), in which cluster consolidation decreased dramatically and the number of overall clusters decreased gradually until all genomes collapsed into a single cluster. The quasi-stable cluster configuration of Phase II is characterized by nAWC plateaus, on which successive thresholds (minimum the number of allowed missing loci) generated nAWC values that remained at or near one. In this phase, mid-sized clusters represent major lineages in the population, and increase in size through merging with smaller clusters, periodically consolidating with other mid-sized clusters and producing concomitant drops in the nAWC, observed as valleys in the nAWC plot. Phase II is consistent with highly stable cluster configurations and reduced rates of cluster consolidation. We defined L2 nomenclature with its similarity thresholds in the start of Phase II of the first nAWC plateau where the number of allowed missing loci would not significantly affect clustering. This method assures the highest resolution possible while at the same time maintaining stable clustering and neutralizes the effect of the allowed missing loci (the cgMLST schema allows up to 2% missing loci in a genome).



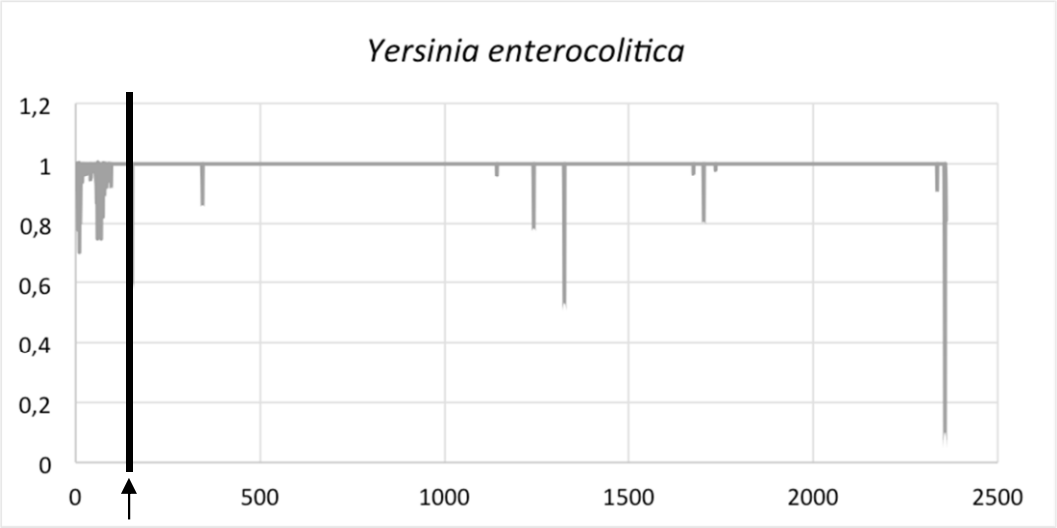

A




B




C




D

The L2 similarity threshold is illustrated with a black line maked with an arrow in the Phase II area of the graph. L1 similarity threshold lies within the area at the very left of the graph where nAWS deviates significantly from 1 and is very variable. The y-axis indicates AWC values. the x-axis indicates the goeBURST thresholds.

**Figure E1:** The nAWS graphs for *E. coli* (A), *S. enterica* (B), *C. jejuni* (C) and *Y. enterocolitica* (D)

To identify the correct plateau, we calculated the distance between two consecutive valleys at T(n) and T(n+1) (d = T(n+1) –T(n)) in the Phase II of the nAWC plot. If the distance d is lower than the allowed missing loci (ML), then L2 similarity threshold is defined as T(n) + ML.

### E.4.4. Nomenclature definition based on defined cgMLST: Level 3

The L3 was defined as the cgMLST goeBURST threshold with higher concordance with MLST definition using Adjusted Wallace as described in Carriço et al. (2006). The partitions produced by goeBURST were compared with those produced by MLST using http://www.comparingpartitions.info website. Table E1 summarizes the thresholds with maximum concordance.

## E.5. Discussion

By assigning a short, human readable code to isolates, nomenclature reduces the amount of information to be shared and allows the use of a common language between laboratories, clinicians, epidemiologists, researchers and governmental stakeholders, and reduces misconceptions and misunderstanding. Therefore, a stable typing nomenclature have a strong impact on surveillance processes by allowing an effective communication of molecular typing results to and between the public health, food safety, and research communities. A good example of this is the MLST or MLVA-associated nomenclature and its central role in modern molecular epidemiology.

Within the INNUENDO platform, three levels of nomenclature have been specified: Level 1 (L1) for outbreak detection and investigation, Level 2 (L2) for longitudinal surveillance monitoring and Level 3 (L3) for congruency with other commonly used nomenclature systems such MLST.

The first nomenclature (L1) is defined at low allelic profile divergence, employing a similarity threshold in Phase I of the nAWC graph. L1 is designed to support outbreak detection and investigation by classifying isolates under the same "type" if the samples share a high degree of genomic similarity, and, thus, can be considered clonally related. This level of our nomenclature is operational in nature, and is intended for short-term cluster identification, as occurs during outbreaks. L1 is not suited for longitudinal surveillance for several reasons: 1) the similarity threshold defined for L1 lies in an area of cluster instability on the goeBURST graph, meaning that L1 threshold is sensitive to changes in the

bacterial genetic pool and adjustments of the L1 similarity threshold, and 2) clusters defined by L1 nomenclature does not represent the major lineages in the population. Moreover, for certain species or certain lineages within a species, the designation of groups based on L1 might cause either false positive or false negative clustering, resulting in misleading and unnecessary follow-up epidemiological investigations. This is especially the case for *C. jejuni*, as the cgMLST schema is of only 678 loci and therefore offers lower resolution when compared to the number of loci obtain for the other species. The chance of wrongly assigning strains to a cluster can be higher, and therefore we suggest that it might need to be resolved on a higher level of resolution, such as using whole genome MLST approach (see Section 3.6.3). Also, the application of L1 for grouping of samples collected through surveillance programs can lead to false clustering. L1 threshold should therefore be used a guideline to support cluster analysis and to facilitate communicating between different actors (e.g. epi-team and other laboratories).

As L1 is not suitable for longitudinal surveillance, a second classification level (L2) was defined based on nAWC analysis. L2 will cluster strains according to stable lineages, creating groups characterized by short intra-cluster and large inter-cluster distances. This level is useful for surveillance purposes, evolutionary studies and division of isolates in subtypes for source attribution, as the thorough method of nAWC applied to define the similarity threshold is sure to be stable over prolonged periods of time. To our knowledge, this is the first attempt to assign a similarity threshold for a nomenclature with a systematic methodology, and the nAWC graph method could be exported to set cut-off values for other pathogens and subtyping schemes as well.

In conclusion, we present adaptable, portable and useful subtyping schema and nomenclature based on solid methodology, which represents a significant leap forward in the use of WGS in public health services. The applications are implemented in the INNUENDO platform, in which quality control and WGS analysis are extensively streamlined while maintaining the ultimate control with the user.

## References

Carriço JA, Silva-Costa C, Melo-Cristino J, Pinto FR, de Lencastre H, Almeida JS, Ramirez M. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant Streptococcus pyogenes. J ClinMicrobiol. 2006 Jul;44(7):2524-32.

Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics. 2009 May 18;10:152. doi: 10.1186/1471-2105-10-152.

Llarena AK, Taboada E, Rossi M. Whole-Genome Sequencing in Epidemiology of Campylobacter jejuni Infections. J Clin Microbiol. 2017 May;55(5):1269-1275. doi: 10.1128/JCM.00017-17.

Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013 Oct;11(10):728-36.

Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, Gilpin B, Smith AM, Man Kam K, Perez E, Trees E, Kubota K, Takkinen J, Nielsen EM, Carleton H; FWD-NEXT Expert Panel. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill. 2017 Jun 8;22(23).

Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microb Genom. 2018 Mar 15. doi: 10.1099/mgen.0.000166.

Timme RE, Rand H, Shumway M, Trees EK, Simmons M, Agarwala R, Davis S, Tillman GE, Defibaugh-Chavez S, Carleton HA, Klimke WA, Katz LS. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. PeerJ. 2017 Oct 6;5:e3893.

# Appendix F – Report of the national and international usability tests of the INNUENDO Platform

Authors: Annina Jaakkonen[1], Bruno Filipe Gonçalves[2], Diogo Silva[2], Jani Halkilahti[3], João André Carriço[2], Mirko Rossi[4], Marjaana Hakkinen[1], Saara Salmenlinna[3]

[1]Finnish Food Safety Authority, Evira, Helsinki, Finland;

[2]Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina Universidade de Lisboa, Lisboa, Portugal;

[3]Finnish National Institute for Health and Welfare, Helsinki, Finland;

[4]Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland

## F.1. Summary

This Appendix describes the technical demonstration of the INNUENDO platform by using two proof-of-concept studies. The proof-of-concept consists of simulations of a national outbreak in Finland and an international outbreak among 12 participants representing seven countries and three European Institutions. Authorities of public health and food safety and veterinary medicine participated in these exercises.

The purpose of these simulations was to test the ability of the INNUENDO Platform to recognize clusters of Shiga-toxin producing *E. coli* (STEC). Twentysix samples including three known clusters and sporadic isolates were used as test material and analysed by the all participants. A clearer cluster definition was implemented, of which some aspects aided in the communication between authorities. Clusters were identified by most participants, but the communication as designed by email notification was suboptimal.

Participants of the international simulation generally experienced the INNUENDO Platform and PHYLOViZ as a very good tool for sequence analysis (average 4.06/5 points in feedback questionnaire) and suggested further useful developments of the platform.

Based on simulation studies, Finland has decided to implement INNUENDO platform as a One Health surveillance tool for STEC, and possibly also for other foodborne pathogens. Communication of INNUENDO results to epidemiologists, trace back authorities, municipalities and international community (such as restricted access EPIS Urgent Inquiries) still require revision and collaboration.

## F.2. Introduction

Pathogen surveillance and outbreak investigations are crucial tasks to prevent and control transmission of foodborne and environmentally transmitted diseases. In this regard, molecular and genomic typing is frequently used in epidemiological investigations to establish relationships between different isolates. For many authorities in European countries, sequencers are already available, but personnel with adequate bioinformatics skills are scarce. Also, comparing sequence data internationally in dedicated platforms, and sometimes by dedicated curators, is increasing, but still in development phase. International sequence comparison initiatives primarily aim to detect cross-border outbreak, while there may be other priorities and needs at national level. This emphasizes the importance of self-sustainability in bioinformatics in small countries and regions.
This document provides a detailed overview of the evaluation of the INNUENDO Platform as a tool to recognize simulated national and international outbreaks in One Health collaboration between authorities. The document also includes description of legal and IT infrastructure challenges encountered and solutions chosen by the public health and food safety authorities in Finland in order to get INNUENDO Platform functional. The document further provides suggestions for improvements

in order to adapt the Platform for practical requirements for routine surveillance of bacterial foodborne pathogens.

## F.3. Objectives

This exercise aims at testing the usability of the INNUENDO platform with associated communication protocols in outbreak investigation at national and international level by analyzing whole genome sequences (WGS) of E. coli collected from previously identified outbreaks and clusters.

## F.4. Simulation of a national outbreak

### F.4.1. Materials and methods

Two participants (later referred to as investigators), one from Evira and one from THL, were included in the blind, one week-long simulation study. The investigators were familiar with WGS analysis in general but were not involved in STEC surveillance. An introduction session to the INNUENDO platform was held by the administrators the week prior to the simulation study. The introduction was held by INNUENDO Administrators in THL and in Evira. The sequences were made available to the investigators on two occasions; Monday 23rd of October 2017 and Wednesday 25th of October 2017.

The sequences were uploaded to INNUENDO platform by the Administrator at THL (not by investigators themselves). The investigators received written instructions for the analysis, performing all analyses available on the INNUENDO platform version 1: Pathotyping, INNUca (assembly and quality control), and chewBBACA (allele calling). Cluster analysis was done in PHYLOViZ Online V2. *Possible E. coli* clusters in this study were defined as a maximum of 0.7% difference between two samples regardless of the resolution used in wgMLST analysis.

The investigators were asked to perform the analyses as soon as possible and notify their findings by email to the responsible person for STEC surveillance in both Evira and THL. The investigators were also asked to fill in a reporting form (Annex E). Communication between the two investigators was neither forbidden nor encouraged.

The investigators received 26 *E. coli* genomes collected from two confirmed outbreaks of human, food and animal origin. One outbreak (cluster A) concerned two independent families with children with confirmed STEC infections and history of contact with small ruminants. The second outbreak was supposedly due to consumption of rocket during several gatherings in one weekend in 2016 creating two clusters: STEC (cluster B) and EPEC (cluster C). This dataset was complemented with 94 *E. coli* genomes as background data from 2016 collected by THL through surveillance activities already added in the INNUENDO platform database (Table F1).

**Table F1:** Isolates analyzed in the national simulation

| Strain id | Strain characteristics | Typing information | Origin | Information for Investigator | Expected clustering |
|---|---|---|---|---|---|
| 1910149 | Mixed culture of EPEC and STEC | O111 (stx+, eae+) | roast beef on a bed of rocket | food | Ambiguous |
| 1910150 | Mixed culture of EPEC and STEC | O111 (stx+, eae+) | roast beef on a bed of rocket | food | Ambiguous |
| 1910072 | STEC | O157 (stx2+, eae+) | human, family has contact to sheep | human | Cluster A |
| 1910082 | STEC | O157 (stx2+, eae+) | human, family has contact to goats | human | Cluster A |
| 1910140 | STEC | O157 (stx2+, eae+) | sheep | animal | Cluster A |
| 1910141 | STEC | O157 (stx2+, eae+) | sheep | animal | Cluster A |
| 1910142 | STEC | O157 (stx2+, eae+) | sheep | animal | Cluster A |
| 1910100 | STEC | ONT:H11 (stx2+, eae-) | human, attended to one of the gatherings | human | Cluster B |
| 1910103 | STEC | ONT:H11 (stx2+, eae-) | human, attended to one of the gatherings | human | Cluster B |
| 1910108 | STEC | ONT:H11 (stx2+, eae-) | human, attended to one of the gatherings | human | Cluster B |
| 1910109 | STEC | ONT:H11 (stx2+, eae-) | human, attended to one of the gatherings | human | Cluster B |
| 1910119 | STEC | ONT:H11 (stx2+, eae-) | human, attended to one of the gatherings | human | Cluster B |
| 1910111 | STEC | ONT:H11 (stx2+, eae-) | roast beef on a bed of rocket | food | Cluster B |
| 1910152 | STEC purified from mixed culture | ONT:H11 (stx2+, eae-) | roast beef on a bed of rocket | food | Cluster B |
| 1910113 | STEC | ONT:H11 (stx2+, eae-) | broiler spiced with rocket | food | Cluster B |
| 1910152 | STEC | ONT:H11 (stx2+, eae-) | beef steak | food | Cluster B |
| 1910101 | STEC | O166:H28, (stx2+, eae-) | human, attended to one of the gatherings | human | Sporadic |
| 1910102 | EPEC | O111:H8 (stx-, eae+) | human, attended to one of the gatherings | human | Cluster C |
| 1910116 | EPEC | O111:H8 (stx-, eae+) | human, attended to one of the gatherings | human | Cluster C |
| 1910151 | EPEC | O111:H8 (stx-, eae+) | roast beef on a bed of rocket | food | Cluster C |
| 1910112 | EPEC purified from mixed culture | O111:H8 (stx-, eae+) | roast beef on a bed of rocket | food | Cluster C |
| 1910145 | STEC | O146 | human, EHEC infection within family, year 2017 | human | Sporadic |
| 1910144 | STEC | O111 | human , travel to Turkey, year 2017 | human | Sporadic |
| 1910143 | STEC | O8 | human, travel to Croatia, year 2017 | human | Sporadic |
| 1910146 | STEC | O157 (stx1+, stx2+, eae+) | bovine, feces | animal | Sporadic |
| 1910147 | STEC | O103 (stx+, eae-) | goat, feces | animal | Sporadic |

The INNUENDO platform was hosted at cPouta cloud service at CSC, Finland https://research.csc.fi/cloud-computing. The setup consisted of 5 virtual machines (VMs) with the following tasks: (i) frontend (web/database/PHYLOViZ Online V2) server (4 CPU, 8 GB RAM), (ii) calculation server (2 CPU, 2 GB RAM), (iii) storage server (2 CPU, 2 GB RAM) and (iv) 2 computing nodes (each with 24 CPU, 100 GB RAM). Altogether, 56 CPU, 212 GB RAM and 1.25 TB of storage were allocated for the virtual machines. The servers' specifications were standard flavours for frontend, calculation and storage servers and high-performance computing for the two computing nodes (https://research.csc.fi/pouta-flavours). Both participating institutes received their own authentication information and own data storage space on the platform.

## F.4.2. Results

### F.4.2.1. Pathotyping

Twenty one of 26 isolates were identified in situ with the correct pathotype. Of the five isolates with NA result in Pathotyping, four also failed the INNUca QA/QC, suggestive of inferior sequence quality. For one isolate (1910101), the reason for missing Pathotype remained unclear.

In one STEC+EPEC mixed culture, an atypical EPEC (aEPEC) was found. The expected result would have been to report only STEC, as stx-positivity should rule out EPEC. It might be that this sample actually was an aEPEC, and therefore the classification "error" is unrelated to the performance of the INNUENDO.

Some issues in misleading pathotype nomenclature were raised during the national simulation. The needs to improve pathotype reporting was issued by the participants and addressed before the international simulation.

#### F.4.2.2. INNUca

Six of the 26 sequences failed INNUca due to low coverage below mean read coverage (25x, three strains) or assembled coverage (30x, three strains). In comparison to a different pipeline used by the authorities (i.e. Ridom SeqSphereTM) with Velvet assembler, the assembled average coverage varied between 27 and 41, and the percentage of Good Targets (number of loci in the scheme identified in each sample) between 96.2 and 99.1%.

#### F.4.2.3. Cluster analysis

Depending on the succession of sequence analysis, the investigator performing the last analysis was able to detect all three expected clusters (i.e. clusters A, B, C). In addition, the investigators were able to discover isolates from the background material within the A, B and C clusters. The cluster analysis was dynamic, as one human and one food strain reported as part of a cluster (cluster B, see below) on Monday were no longer part of that cluster two days later. None of the sporadic isolates clustered with outbreak strains. Below a detailed summary of the analyses of the three clusters:

- Cluster A (STEC, O157, sorbitol positive) contained five strains; two human and three sheep strains. One human and sheep strain failed INNUca, resulting in THL not seeing their own samples cluster, but discovered clustering of its human sample with three background samples originating from sheep. Evira reported a match between the remaining two sheep strains and the human strain uploaded earlier by THL and four human strains in background material. One discrepancy between THL and Evira interpretation was evident, as Evira included a background isolate (1910018) in the cluster while THL did not.
- Cluster B (an epidemic related to consumption of food spiced with rocket, ONT:H11, ST 295) contained nine strains and two mixed cultures from food samples, of which both contained the cluster B isolates. On Monday during the simulation week, Evira analysed four food strains and one mixed culture. Of these, two strains did not pass INNUca, and the rest were reported to cluster together with themselves and with four background isolates. THL then analysed two human strains, and reported these to cluster together with the food strains and the four background strains previously analysed by Evira. On Wednesday, THL analysed three additional human strains and reported these to cluster together with all the samples mentioned above. However, one human strain (1910100) and the mixed culture sample (1910150), earlier reported as part of the cluster identified on Monday, were separated from the remainder of the cluster on Wednesday. Clustering of the mixed cultures was consistent with their pathotyping results, but it was not evident from the WGS analysis that these samples were mixed.
- Cluster C is related to the same epidemic as cluster B, EPEC, O111:H8, ST 327 and contained four outbreak strains and the two mixed samples as mentioned above. THL analysed two strains on Monday and reported these to cluster, while Evira analysed two food strains and one mixed culture on Wednesday and reported these to cluster together among themselves and with the human strains from THL.

#### F.4.2.4. Communication

E-mail was used to communicate discovered clusters between Evira and THL. Excel file for reporting results was used. During national simulation, a preliminary cgMLST-based nomenclature was

available, but since it was designed for clustering samples at high level of alleles differences (e.g. concordant with population genetics), it wasn't used for interpreting clusters. The need of a lower cut-off for cgMLST based nomenclature concordant with epidemiological information was issued and addressed in the international simulation.

## F.5. Simulation of an international outbreak

### F.5.1. Materials and methods

#### F.5.1.1. Participants, general organization and schedule

Twelve institutes (Table F2) participated with one or two investigators in the blinded testing of the INNUENDO platform. The majority of the institutes represented European or national central authorities or reference laboratories, equally divided between public health and food safety sector. The analyses were performed during two weeks, 5–19 Feb 2018.

As a prerequisite, the investigators assured to obtain a list of technical requirements and participated in a webinar training session, held on Friday, 2nd of February 2018. The webinar lasted for 2.5 hours and covered the following topics: (i) background information of the analytical modules available on the platform, (ii) platform demo and (iii) instructions for the simulation. Presentation slides, a recording of the webinar (https://connect.funet.fi/p46jx17nppj0/) and written instructions for the simulation were shared with the participants.

The investigators received their user account information for the INNUENDO platform and access to a test sample on Sunday, 4th of February. During the two weeks, each investigator received a batch of five samples per week, ten samples in total. The investigator was instructed to analyze the samples as soon as possible. They reported their findings of the first week samples by Monday, 12th of February and findings from both weeks by Monday, 19th of February on predetermined reporting forms.

During the simulation, the participants were able to receive real-time support and send feedback in a chat community hosted in Slack (https://simulation2018.slack.com). After the simulation, feedback on the user experience was collected from the participants on a feedback form (Annex F).

#### F.5.1.2. Upload and analyses of data

The participants received the data in zipped archives containing paired-end Illumina sequences (.fastq.gz) and their metadata (.csv) through Funet FileSender (https://filesender.funet.fi/). The participants downloaded, extracted the archives and uploaded the sequences into the INNUENDO platform through SFTP client. In the INNUENDO web application, the participants created a project, added samples to the project and filled in sample metadata including: sample ID, source, sample received date and submitter (submitting institute). Source was selected from a drop-down menu consisting of: human, food, animal cattle, animal poultry, animal swine, animal other, environment and water.

For each sample, the participants performed all analyses available on the INNUENDO platform version 2: Pathotyping, Serotyping, INNUca_plus_mlst (assembly, quality control and MLST), chewBBACA (allele calling) and ABRicate (*in silico* typing using ResFinder V2.0, CARD, VFDB and PlasmidFinder databases versions at February 2018). After receiving the results, they did a cluster analysis using PHYLOViZ Online v2.

**Table F2:** Participants and data dispatch schedule

| Participants | Institute name | Country | Field | 1st batch | 2nd batch |
|---|---|---|---|---|---|
| P1 | Austrian Agency for Health and Food Safety (AGES) | Austria | Public health | Mon, 5 Feb | Tue, 13 Feb |
| P2 | German Federal Institute for Risk Assessment (BfR) | Germany | Food safety | Tue, 6 Feb | Mon, 12 Feb |
| P3 | Institute of Food Safety, Animal Health and Environment (BIOR) | Latvia | Food safety | Tue, 6 Feb | Mon, 12 Feb |
| P4 | European Centre for Disease Prevention and Control (ECDC) | EU/Sweden | Public health | Wed, 7 Feb | Mon, 12 Feb |
| P5 | European Food Safety Authority (EFSA) | EU/Italy | Food safety | Thu, 8 Feb | Tue, 13 Feb |
| P6 | European Reference Laboratory for E. coli, Istituto Superiore di Sanità (EURL VTEC, ISS) | EU/Italy | Public health | Mon, 5 Feb | Tue, 13 Feb |
| P7 | Finnish Food Safety Authority (Evira) | Finland | Food safety | Wed, 7 Feb | Mon, 12 Feb |
| P8 | National Institute of Health Dr. Ricardo Jorge (INSA) | Portugal | Public health | Tue, 6 Feb | Mon, 12 Feb |
| P9 | Laboratoire National de Santé (LNS) | Luxembourg | Public health | Wed, 7 Feb | Mon, 12 Feb |
| P10 | National Institute for Health and Welfare (THL) | Finland | Public health | Mon, 5 Feb | Tue, 13 Feb |
| P11 | University of Veterinary Medicine, Vienna (VetMedUni) | Austria | Food safety | Thu, 8 Feb | Tue, 13 Feb |
| P12 | Estonian Veterinary and Food Laboratory (VFL) | Estonia | Food safety | Thu, 8 Feb | Tue, 13 Feb |

#### F.5.1.3. Communication

If the investigators found their samples to cluster with samples submitted by other investigators within the platform, they were instructed to contact the submitter of the other samples by email as soon as possible. In addition, they were instructed to send a copy of the e-mail to the central authorities: either public health, food safety or both depending on the source of the clustering samples. They were also instructed to use a predefined title of the email including the following items: INNUENDO *E. coli* cluster <classifier(s) of the strains in the cluster> (according to the L1 level) <mlst(ST)>. In addition, contents of the e-mail were instructed as follows: sample ID, submitter, project name, sample received date, source, pathotype, serotype, mlst(ST) and classifier of all isolates belonging to the cluster, PHYLOViZ dataset name (tree name) and link.

The central authorities, played by INNUENDO team members from THL (public health) and EVIRA (food safety), kept a real-time logbook of the outbreak investigations in an extranet workspace (Microsoft SharePoint 2010, https://eviranet.evira.fi/). For each outbreak, a logbook (.xlsx) was created with the file name including the following items: INNUENDO *E. coli* cluster <classifier(s) of the strains in the cluster> <mlst(ST)>. In the logbook, the following column headers were used: reporting time, reporting person, report recipients, PHYLOViZ dataset name, PHLYLOViZ link and report content.

The participants were instructed to trace back the source (food, animal or environmental) that possibly caused the outbreak through email communication. The participants playing food laboratories were given additional source information that was not visible to all on the INNUENDO platform. They were expected to communicate this information with other participants involved in outbreak investigation. However, no epidemiological information (such as person identification, symptoms, place of infection, date of sampling, or connection to farms) was given to public health laboratories.

#### F.5.1.4. Data

The outbreak and background isolates were the same as used in the national simulation. The 120 genomes were divided between the participants, ensuring that each participant received samples belonging to one of the clusters A, B or C. For some of the samples, source was included as metadata, while fake metadata was invented for some of the human samples from THL (claim to be of non-human origin, cluster C origin spinach instead of rocket). In addition, samples failing quality control were included of which six failed sequences were duplicated and divided to different participants. Some of these failed sequences clustered in clusters A and B.

#### F.5.1.5. Computing infrastructure

The INNUENDO platform was hosted at INCD (National Distributed Computing Infrastructure, Lisbon, Portugal). The setup consisted of six VMs with the following tasks: frontend (web/database) server (24 CPU, 32 GB RAM), calculation server (16 CPU, 16 GB RAM), PHYLOViZ Online server (8 CPU, 8 GB RAM) and three computing nodes (each with 16 CPU, 16 GB RAM). Altogether, 96 CPU, 104 GB RAM and 1 TB of storage were allocated for the virtual machines. Each participant received their own authentication information and own data storage space on the platform

### F.5.2. Results

#### F.5.2.1. Cluster analysis

Seventy of 120 samples were reported to cluster by at least one investigator, 35 were reported as sporadic and 15 were not analyzed for clusters due to several reasons. One of the twelve participants (P4) did not submit any report, and clustering of its samples was based on the results of other participants. Of the 15 samples not analyzed for clusters, six samples were ran by P4, eight samples failed the assembly quality and one sample (id 1910154) raised a warning in chewBBACA status and was not included in the cluster analysis by P8.

The participants reported 18 clusters in total (Table F3). Of these, 12 clusters (clusters 1–12) were reported by more than one participant and communicated by email. Six clusters (clusters 13–18) consisted of samples delivered to one or two participants, and were reported to the central authority by a single participant. Clusters 1, 2 and 3 corresponded respectively to the clusters A, B and C in the national simulation. Altogether, 31 samples were expected to belong in the three clusters or fail in quality. All of these samples were reported to cluster as expected, excluding two borderline samples (1910079 and 1910018) in Cluster 1.

**Table F3**: Clusters reported by the participants in the international outbreak simulation

| No | Classifier | MLST | Pathotype and serotype | Submitters of samples | No of samples | Source of the outbreak or contamination |
|---|---|---|---|---|---|---|
| 1 (A) | 2103, 2094 | ecoli (11) | STEC O157:H7 | P1, P2, P3, P9, P10, P11 | 11 | sheep/ovine cheese |
| 2 (B) | 2101, 2105, 2147 | ecoli (295) | STEC NT:H11 | P2, P4, P5, P6, P8 | 15 | rocket |
| 3 (C) | 2095 | ecoli (327) | aEPEC O111:H8 | P3, P7, P10, P12 | 5 | spinach |
| 4 | 2102, 2115, 2149 | ecoli (11) | STEC O157:H7 | P2, P6, P7, P12 | 5 | unresolved: spring water/cookie dough/ovine cheese |
| 5 | 2106, 2104 | ecoli (335) | STEC O55:H7 | P5, P6, P9 | 4 | spinach, organic |
| 6 | 2116 | ecoli (32) | STEC O145:H28 | P7, P12, P11 | 3 | unresolved: leafy greens/ chicken salad/beef tartar |
| 7 | 2099 | ecoli (11) | STEC O157:H7 | P2, P3 | 4 | unresolved: beef/cattle/sheep/swimming |

| No | Classifier | MLST | Pathotype and serotype | Submitters of samples | No of samples | Source of the outbreak or contamination |
|---|---|---|---|---|---|---|
| | | | | | | water |
| 8 | 2108 | ecoli (1833) | STEC O157:H7 | P1, P4 | 3 | human only |
| 9 | 2093 | ecoli (360) | NA O8:H19 | P6, P10 | 2 | human only |
| 10 | 2117 | ecoli (718) | - NT:H8 | P7, P12 | 2 | beef steak |
| 11 | 2145 | ecoli (21) | STEC O26:H11 | P1, P2 | 2 | cattle (sample from feeding table of cows) |
| 12 | 2148 | ecoli (6274) | STEC O181:H16 | P5, P8 | 2 | unresolved: sprouts, fenugreek |
| 13 | 2146 | ecoli (753) | STEC O27:H30 | P5, P8 | 2 | unresolved: irrigation water of sprouts |
| 14 | 2100, 2097 | ecoli (29) | STEC O26:H11 | P3, P10 | 2 | unresolved: flour |
| 15 | 2109 | ecoli (-) | STEC O157:H7 | P1 | 2 | human only |
| 16 | 2111 | ecoli (11) | STEC NT:NT | P8 | 2 | human only |
| 17 | 2122 | ecoli (11) | STEC O157:H7 | P9 | 2 | human only |
| 18 | 2137 | ecoli (5683) | tEPEC O171:H25 | P10 | 2 | human only |

Below a detailed summary of the analyses of the detected clusters:

- Cluster 1: it was reported to contain ten samples of classifier 2103 and one sample (id 1910079) of classifier 2094. Four of the six investigators involved in the cluster reported this differing sample (id 1910079) ambiguously belonging to the cluster. P10 reported the sample to be within the tree cut-off 0.55% (13 with the goeBURST profile size of 2385), and P9 reported this sample to be distant from the other samples without specifying the cut-off. All the eleven samples showed matching results in assembly quality (C) and chewBBACA status (pass). For P10, these quality metrics indicated enough loci for comparison and thus justified inclusion of the sample 1910079 in the cluster. Their decision was further supported by the concordance of ResFinder results between the clustering samples. In addition to the samples 1910079, P11 suspected ambiguities in the clustering of four samples (1910017, 1910082, 1910083 and 1910140), but without further reasoning. Sample 1910018 was excluded from the cluster by all participants, but represented a borderline sample with a slightly different PFGE type from the rest of the samples.
- Cluster 2: it contained eleven samples of classifier 2101, one sample of classifier 2105 (id 1910100) and one sample of classifier 2147 (id 1910150). Both samples of classifier 2105 and 2147 were of assembly quality C and raised a chewBBACA warning and thus, were suggested for resequencing by P6. Notably, the sample 1910150 represented a mixed culture of two *E. coli* strains. The other samples in the cluster passed their chewBBACA status and were graded as A in their assembly quality, except for two samples (1910113 and 1910155) that failed assembly. Despite failed quality, these samples passed their chewBBACA status. The failed samples 1910113 and 1910155 were reported as ambiguously clustering or suggested for resequencing by three participants involved in the cluster investigation. P2 detected assembly length of 8.5 Mb for the sample 1910155 and suspected misassembly or two concurrent strains. Likewise, P6 commented on too many assembled bp for both of the samples. P5 reported insufficient coverage for cluster analysis for both of the samples and suspected contamination by other organism based on the GC content. Although not satisfied with the assembly quality, P5 justified inclusion of the sample 1910113 in the cluster by the similar number of virulence factors, resistance genes and good chewBBACA status. Both 1910113 and 1910155 were duplicates of the same fastq data.
- Cluster 3: ambiguities were not reported.

- Cluster 4: it contained two samples of classifier 2102, two samples of classifier 2115 and one sample of classifier 2149. However, one of the four participants involved in the cluster investigation (P6) reported the samples of classifier 2115 to form one cluster and the other three samples to form another cluster. In addition, two participants (P2 and P12) reported the samples of classifiers 2115 ambiguously belong to the same cluster with the other three samples. P6 used the tree cut-off of 0.50% (13 with the goeBURST profile size of 2596) to make their decision, whereas P2 and P12 used a cut-off of 0.62% (15 with the goeBURST profile size of 2390). P7 reported the sample 1910089 (classifier 2115) as part of the cluster (with the samples of classifier 2102) after the first week, but not after the second week.
- Cluster 5: it was reported to contain three samples of classifier 2106 and one sample of classifier 2104 (id 1910097). Two of the three participants involved in the cluster investigation (P5 and P6) suggested resequencing of the sample 1910097 having a different classifier because of poor coverage. P9 reported the sample 1910097 and another sample (id 1910025) to be distant from the spinach isolate (id 1910093) and thus ambiguously belonging to the cluster.
- Cluster 6: it contained three samples of classifier 2116, submitted by three participants. All the three samples were reported as ambiguously clustering by P11 without further reasoning. No confidence on the clustering hypothesis could be obtained from the (fake) metadata: the contamination pattern remained unclear as the samples sourced from leafy greens, chicken salad and beef tartar.
- Cluster 7 and 8: ambiguities were not reported. However, the contamination route remained unresolved for Cluster 7 as the samples sourced from beef, cattle feces, sheep feces and swimming water. As the source of the contamination remained unclear, P3 reported to have used the ABRricate results to validate the clustering hypothesis: similar hit counts supported the hypothesis. Cluster 8 contained only human samples.
- Clusters 9 and 10: they originated from the duplication of fastq data. All the four samples raised a warning in chewBBACA status although the assembly quality was graded A or B. In the cluster 9, P6 commented on the poor coverage and suggested both samples for resequencing. P10 found confidence for the clustering hypothesis by the same results in PlasmidFinder and ResFinder. In cluster 10, P12 used all available information to confirm a cluster hypothesis in spite of chewBBACA warning, but did suggest resequencing.
- Cluster 11 and 12: ambiguities were not reported. In Cluster 12, P5 reported sample 1910120 to have a borderline coverage to be evaluated as "good" for clustering. Instead of suggesting resequencing, they requested more epidemiological information. P8 used VFDB database of ABRicate to confirm that the strain harbored *stx*-genes like the related strain, as Pathotyping failed to provide the result.
- Clusters 13 and 14: they were reported by a single participant each (P8 and P3 respectively), although the cluster-samples were submitted to two participants. In Cluster 13, P8 found an ambiguous match between two samples by using a cut-off of 0.93% (31 with the goeBURST profile size of 3327). The other sample (id 1910124) failed the chewBBACA status, but P8 decided to analyze it anyways because the only alert concerned an excessive number of contigs while other indicators were fine. In addition, the number of exact allele matches (EXC) was high. As no result was obtained by Pathotyping, P8 used VFDB database of ABRicate to confirm that the strain 1910124 harbored *stx* genes like the related strain. In Cluster 14, P3 used a cut-off of 1.43% (36 with the goeBURST profile size of 2515) and found that two strains of classifier 2100 and 2097 clustered together. Everything else showed high similarity except one plasmid that was present in one strain but not in the other.

The remaining of the clusters (15–18) were national clusters and ambiguities were not reported.

#### F.5.2.2. *In silico* typing

Pathotype was obtained for 93 of 120 (77%) samples, while 20 (17%) and seven (6%) were designated as having no pathotype (blank result) or "NA" (not available), respectively. Half of the 20 samples producing a blank pathotype failed in assembly quality, and when excluding these, 85% of

the remaining samples obtained a pathotype. However, *stx* genes were found from VFDB database by ABRicate for all of the failed assembly genomes (10), but not by Pathotyping.

Of 120 samples, complete O:H prediction was obtained for 68 (57%) samples. Of the 52 samples with missing O or H prediction, eight samples failed in assembly quality and 15 belonged to the Cluster 2 with serotype NT:H11, confirmed to be an O-nontypeable at the WHO Collaborating Centre for Reference and Research on *Escherichia* and *Klebsiella*, Statens Serum Institut (Copenhagen, Denmark). Of the remaining samples, the majority was graded C in assembly quality, so complete O:H type could be obtained for part of the samples in five clusters (1, 5, 6, 7 and 13).

### F.5.2.3. Quality control and annotation

Of 120 samples, 17 (14%) were graded as assembly quality A, 28 (23%) as B, 65 (54%) as C and 10 (8%) as failed. The failed samples included eight samples that were not analyzed further and two samples (id 1910113 and 1910155) that were reported to belong in Cluster 2. Half of the failed samples were not *E. coli*, two samples (1910136 and 1910139) had too low first coverage (<15.0x), and one sample (id 1910128) failed after trimming. The sample 1910128 showed a shifted average GC percentage around 30% instead of typical 50% for *E. coli*. Two samples grouping to Cluster 2 failed assembly quality due to a large genome size (8.5 Mb) in assembly mapping and a double peak in GC - these samples were actually duplicates of the same fastq data. True coverage module (i.e. exact coverage calculation of selected set of targets) in INNUca was not implemented in the INNUENDO platform version 2, due to limited time. This resulted in a less stringent QC compared with the national simulation.

ChewBBACA status raised a warning and fail for 16 samples (13%) and one sample (<1%), respectively. Altogether, genomes with warnings and failure had 0.34–1.02% and 2.12% missing core loci, respectively, and samples with a warning were graded as A (n=4), B (n=3) or C (n=9) in assembly quality. Eight samples with a chewBBACA warning belonged to clusters, although with a different classifier in the majority of cases (7/8), but the chewBACCA warning was detected for only two of the seven samples by the investigators.

Eight investigators used results from quality control to make decision on clustering. Satisfactory assembly quality and a passed chewBBACA status gave enough confidence for concluding clusters, especially if the allele differences were small. P6 reported to have used information on coverage and assembly length when evaluating the assembly quality. Some participants reported that if not enough confidence could be obtained from the assembly quality and chewBBACA status, annotation results by ABRicate were evaluated to support the clustering hypothesis. ABRicate results were also evaluated if the potential source of infection seemed unclear from the metadata or if Pathotyping failed to provide results on stx genes. All ABRicate databases were used to compare the hit counts of samples in the same cluster. If hit counts differed, P3 investigated virulence and resistance genes and found the same genes in the clustering samples, only with different copy numbers. P5, P10, P11 and P1 used antibiotic resistance profile (CARD or ResFinder) to validate their clustering hypothesis. In addition, P10 and P1 evaluated clustering based on PlasmidFinder results. VFDB database was used to investigate the presence of *stx* genes.

### F.5.2.4. Communication

Twelve clusters (clusters 1–12) were reported by more than one participant and communicated by email. The participants were able to trace back the source with the limited epidemiological information provided. In addition to source information, some of the participants shared their thoughts on excluding, including and resequencing samples in clusters (Cluster 1, 5, 10 and 11). Despite active communication, four clusters remained unsolved, namely Cluster 4, 6, 7 and 12 mainly due to insufficient information on available metadata.

The participants were instructed to use a predefined title and information content in the email, but no fixed format was given. As a result, substantial variation was observed in the organization and format of this information. Some of the participants chose to send the information as a table attachment

(.xlsx), sample IDs by rows and metadata/results by columns. Some participants reported the information as a table, sample IDs by rows in the email, whereas others preferred listing the metadata/results by rows in one column, underneath the sample ID. Even with the table format, the organization of headers varied. The information content also varied: some requested information was excluded or extra information (e.g. on quality metrics) was included in the email. The participants often excluded PHYLOViZ tree name and project name, but their counterparts in other institutes did not request this information. However, missing the classifier, MLST(ST) and PHYLOViZ tree link resulted in problems and were identified by the participants as an obstacle for a correct communication during the simulation.

In addition to the communication between the participants, the central authorities in this exercise (THL and Evira) followed the discussion on clusters and kept real-time logbook for each outbreak investigation separately. The predefined title enabled distinguishing of discussions on simultaneous outbreaks efficiently. However, the lack of standardized format for the information content and organization of the emails complicated filling in a table-format logbook and hindered its readability.

#### F.5.2.5. Feedback

Feedback on the participant's experience of the INNUENDO platform was collected during the simulation on a Slack channel and the week following the simulation through a feedback form (Annex F). Thirteen responded from ten institutes, resulting in a response rate of 100% among the participating institutes. The participants were asked to rate 29 aspects related to the usability of INNUENDO platform on a scale from 1 to 5 (1=poor, 2=fair, 3=good, 4=very good, 5=excellent). The global average of all questions was 4.06 (min. 2.85–max. 4.85), corresponding to very good. Only two of the 13 respondents rated the platform weaker than good (2.85 or 2.93). The participants were satisfied with the general organization of the platform, running procedures, color coding for run progress, report page, classification and individual tools (Pathotyping, Serotyping, INNUca, chewBBACA allele calling and schema, PHYLOViZ Online v2) (rating >4 (3–5), INNUca received the highest score (4.38).

Weaker ratings, but still in the range of good and very good, (range average 3.69–4.0) were noted for sample submission (including SFTP and available metadata), chewBBACA reports (including quality control and presentation of statistics) and cluster analysis. Rating of cluster analysis consisted of: sending chewBBACA profiles to PHYLOViZ Online, increase of analysis resolution in PHYLOViZ and searching for the closest strains in the database.

Presentation of ABRicate results was rated 3.85, as well as usefulness of sharing the results on the platform to promote communication and use of classifier in cluster analysis. Participants with bioinformatics experience generally rated the question "how well does INNUENDO platform meet your needs?" and "how easy was it to learn how to use the INNUENDO platform?" with lower values than other participants.

The participants enjoyed user-friendliness and provision of complete analysis solution with visualizations for *E. coli* offered by the INNUENDO platform. In addition, they appreciated the ability to compare the results with other user. The participants wished for even more automation in pre-analysis steps: functional file upload of metadata and ability to select all workflows at once.

The interface between the INNUENDO platform and PHYLOViZ Online received some critics:

- There was no possibility to move back and forth between the platform and PHYLOViZ and update existing trees.
- The information on L1 classification was only available in PHYLOViZ.
- Moving between the project (Projects page) and its results (Reports page) could have been more straightforward.
- The serotyping module should be improved, since the results were too often missing for meeting the needs of a reference laboratory
- Some names were misleading in their context.

- Communication of clusters proved difficult without a standardized email form and because no (user-defined) summary report could be generated from the platform or PHYLOViZ, requiring relevant information to be copied separately from different Report sections and PHYLOViZ.
- In addition to the tools already available on the platform, the participants wished for implementation of VirulenceFinder database because important genes were missing from the VFDB database (e.g. *stx* subtypes, *subA, iss, lpfA, iha*...), and implementation of tools involved in the detection of mobile genetic elements or prophage sequences

## F.6. Discussion

The INNUENDO project aimed at providing solutions for the implementation of whole genome sequencing in foodborne pathogen surveillance and outbreak investigations, especially in smaller European countries. In the project, an analysis platform was developed for the target group of microbiologists with limited bioinformatics skills. Usability of the INNUENDO platform was evaluated in a national and international simulation with participants from twelve European institutes. The background expertise of the participants varied: both bioinformaticians and microbiologists with more or less experience in genomics participated in the simulation study. Overall, the participants rated the INNUENDO platform very good with the average of 4.06 in their feedback. The most critical ratings were received from bioinformaticians and could be explained by their experience and preference for other analysis solutions or platforms. This was supported by their lower ratings to the question "how easy was it to learn how to use the INNUENDO platform?" despite their higher skill level. Importantly, the participants who fit the target group of INNUENDO evaluated the platform with higher ratings (very good or excellent) and found it visually appealing and user-friendly, after only a short webinar training. This suggests that INNUENDO platform met the needs of potential users very well and could be easily learned by them.

In the international simulation, the participants ran all the analysis modules available for *E. coli* on the INNUENDO platform and used all the results to conclude which strains were clustering to identify suspected outbreaks. The analysis modules for Pathotyping and Serotyping were based on read mapping. INNUca module provided assembly, quality control and MLST. ABRicate was used to screen assembled contigs for virulence genes (VFDB database), resistance genes (ResFinder and CARD) and plasmids (PlasmidFinder). Allele calling was performed by chewBBACA, followed by cluster analysis and visualization in PHYLOViZ Online v2. All the analysis modules and databases proved to be useful in the simulated outbreak investigations as the participants used the results in their decision making.

Pathotype was obtained for the majority (85%) of samples that passed the assembly quality. For 6% of the samples, pathotype was not available ("NA"), and for 9% missing (blank result). Missing result meant that pathotype was not determined because data quality was too low for reliable prediction. For the majority of samples with a missing pathotype, chewBBACA status warned about too many missing loci and thus indicated possible defects in data quality although the samples had passed assembly quality. Despite Pathotyping failed to provide result, *stx* genes could be found for these samples by ABRicate against VFDB database. This suggests that screening by ABRicate is useful and complementing the read-based classification in pathotype by the Pathotyping module and can confirm the presence of STEC when defects are suspected in data quality. However, need for more intuitive labels than blank and "NA" to indicate lack of predictive pathotyping due to low data quality or simply not detected, respectively, was recognized as the participants requested help in interpreting the results.

Serotyping provided complete O:H prediction for only 57% of samples. Some of the missing serotype predictions could be explained by strains that were confirmed to be O-untypeable or by failed assembly quality, but not for 24% of samples graded C in assembly quality. For 15 samples with a missing O or H serotype, clustering with other samples suggested existence of the serotype in the database. All of these 15 samples were graded as C in assembly quality, suggesting that serotype was not predicted because minimal data quality was not met for serotype prediction. The rate of serotype prediction was too low to meet the routine needs of a reference laboratory, as recognized by the participants. For further development, Serotyping is suggested to provide the user with serotype

prediction even if the minimal data quality for reliable prediction is not met (see Section 3.5). In this case, the user should be warned with a quality label.

In the feedback, INNUca received the highest score (4.38) of the analysis tools, highlighting the need of automatic pipeline for determination of quality of reads and assembly, and detection of contaminating species. INNUca was developed between the national and international simulation studies by introducing assembly quality grades A–C. In addition, the threshold for passed quality was lower in the international simulation than before because true coverage determination step was not implemented in the INNUENDO platform used for the study yet. This left more responsibility for the user to evaluate the need for resequencing in each case. In the platform version 2 however, FastQC and Assembly charts were made visible to the user to aid the decision on borderline samples. As expected, more samples passed the assembly quality in the international than in the national simulation. In the international simulation, only 8% of the samples failed assembly quality, majority because of contamination, suggesting appropriate automatic rejection. In addition to assembly quality, chewBBACA status could be used to evaluate data quality and reliability of clustering results in the international simulation. Warnings and failures of chewBBACA status were based on percentage of missing loci. In the international simulation, chewBBACA status raised a warning for 13% and failure for <1% of samples

As the participants reported, satisfactory assembly quality grade and passed chewBBACA status provided usually enough confidence for the decision on clustering, especially with small allele distances. However, if not enough confidence on clustering could be obtained by these automatic quality labels, the participants used other information available on the platform in their decision making, such as ABRicate. All ABRicate databases available in this study proved to be useful in the decision making on clustering. However, VFDB database was noticed to lack some virulence genes of *E. coli* that are relevant in public health decision making, especially stx subtypes. Therefore, implementation of VirulenceFinder database was suggested in feedback. The feedback reflected that the INNUENDO platform easily provided a lot of information, suggesting that the platform serves users as intended: providing quick determination of satisfactory data quality, and enough information to support the decision on borderline samples. Implementation of VirulenceFinder database would make INNUENDO a complete analysis solution of *E. coli* for public health and food safety reference laboratories.

After finishing the analyses on the INNUENDO platform, the results were sent to PHYLOViZ Online V2 for cluster analysis and visualization as minimum spanning trees. Clustering was based on shared loci from a wgMLST schema of 7,601 curated loci. PHYLOViZ allowed dynamic change of resolution by sub-setting of trees. In addition, nomenclature on level L1 (called *classifier*) was calculated based on a similarity threshold on cgMLST loci (2,360 loci). If a sample lacked more than L1 of core genes, it raised a warning in chewBBACA status. In the feedback, the interface between the INNUENDO platform and PHYLOViZ received some critics. When PHYLOViZ datasets were created in INNUENDO, the user had to determine parameters and auxiliary data to show and type in PHYLOViZ authentication information. Displaying information selected from PHYLOViZ back to INNUENDO Reports and updating of existing trees were impossible. As suggested, usability of the INNUENDO platform could be improved by further development of the interface. However, the participants were able to report the requested information and solve the three pre-known clusters, along with fifteen extra clusters, by the available cluster analysis approach. Some of the interpretations on strains belonging to a cluster varied between the participants and between the international and national simulation mainly because slight differences in used cut-offs. As limited epidemiological data were provided to the participants, they based their decision on clustering mainly on sequence data. In this regard, INNUENDO platform and PHYLOViZ Online V2 successfully provided a solution for the identification of suspected outbreaks by comparison against the entire database.

The nomenclature proved to be useful in the outbreak investigations, although the name "*classifier*" caused confusion and was suggested to be changed. Nomenclature aided in both communication and identification of clusters. Because cluster analysis was based on shared loci in wgMLST and nomenclature on cgMLST, some variation were expected between the clusters defined by the

nomenclature alone and the clusters defined by user interpretation, especially in *E. coli* with a wide accessory genome. Furthermore, instability of the L1 nomenclature was expected (see Appendix E) with the similarity threshold of seven loci, allowing only a few missing loci. As anticipated, two or three types were detected in five clusters. In these clusters, altogether seven samples represented different types from other samples and two of these samples raised a chewBBACA warning. ChewBBACA warnings were also raised by six clustering samples with the same type. Therefore, missing loci did not explain the observed variation in classifiers alone. However, seven samples represented only 10% of all seventy samples that were reported to belong in clusters. This suggests that L1 classification was useful in the automatic identification of clusters.

In the international simulation, the participants uploaded the sequencing reads in their private volume on the platform through SFTP, but the analysis results and metadata were visible to all twelve participants who shared the platform. This allowed control of the raw data, and simultaneous comparison of results for faster outbreak investigation and communication. SFTP transfer received lower ratings from the participants, which could be due to restrictions by data security policies in some institutes during the simulation. However, SFTP could offer easy data transfer with sufficient data security in routine use of INNUENDO where user institutes have control over the computing infrastructure. In addition, the participants wished to reduce the manual steps before running the analyses, which would likely be improved by fixing the file upload of sample metadata.

Ability to share the results between all users was appreciated in the open feedback. However, usability of the results in communication received weaker ratings, probably because communication issues were recognized overall. Although the contents of communication emails were pre-instructed, the contents and form of the emails varied extensively. Furthermore, the participants had to pick the information from different locations on the INNUENDO platform and PHYLOViZ. Therefore, automatic generation of a summary report was suggested to enhance communication and reduce manual work. Furthermore, the automatic format of the report should allow easy incorporation of the information in outbreak investigation logbooks, which would probably prefer table format. Despite that these communication issues were recognized, nomenclature proved to ease the communication on clusters and keeping of logbooks. Both the classifier and MLST(ST) were used in communication and successfully allowed distinguishing between different clusters, even when several classifiers were found in the same cluster. Therefore, hierarchical nomenclature is suggested to be used in communication on clusters.

These simulation studies provided proof of concept for the usability of the INNUENDO platform in outbreak investigations on *E. coli* and communication. Although needs for further development were recognized, the platform met the needs of the reference laboratories well.

Both national and international simulations showed that INNUENDO platform is a promising tool for detection of foodborne outbreaks between authorities. For this reason, in Finland INNUENDO platform will be set up to be used by public health and food safety/veterinary authorities as One Health shared analytical platform for outbreak investigation of foodborne pathogens (specifically for *E. coli*). In addition to functional technical tool, there are several issues, irrespective of INNUENDO, required to be solved before this kind of One Health system is operational. During the simulation studies, some of these issues were solved. These included in-house IT problems related to government IT-network, user restrictions, and purchase of calculation capacity for two institute users in CSC cloud service cPouta. The information security clarification for using cPouta service remains to be completed. Another issues related to information security is the ability to use the strain identification numbers from the laboratory information system as identifiers in the INNUENDO analysis. For the simulation analysis, alias numbers were created, but this is not practical for routine analysis.

Long term use of INNUENDO Platform also requires a sustainable system for supporting continuous uninterrupted use of the platform, version development, and preferably customer support. As well as for all other platforms available, there is no guarantee for INNUENDO to be functional and adaptable for new requirements in the future. This would require continuous financial support, commitment of the developers as well as adequate training of the in-house administrators of the platform.

# Appendix G – "Genomics in food-borne pathogen surveillance and outbreak investigation" Workshop (Vitoria-Gasteiz, July 2017)

## G.1. Overview of the workshop achievements

The course was organized by the INNUENDO consortium within the Summer School of the University of Basque Country (UPV/EHU; https://www.uik.eus/).

The course was sponsored also by: Gobierno Vasco/Eusko Jaurlaritza (http://www.euskadi.eus/), master and doctoral school of the University of Basque Country (https://www.ehu.eus/es/web/mde), Vice-rectorate of Campus of Álava, Elika (http://www.elika.eus/en/), Ayuntamiento de Vitoria-Gasteiz/Vitoria-Gasteisko Udala (https://www.vitoria-gasteiz.org/), ESCMID Food- and Water-borne Infections Study Group – EFWISG (https://www.escmid.org/research_projects/study_groups/foodwater_infections/) and the Finnish center of expertise in ICT (www.csc.fi).

The organization committee was composed by: Javier Garaizar, University of the Basque Country UPV/EHU, Spain; Mirko Rossi, University of Helsinki, Finland; Joseba Bikandi, University of the Basque Country UPV/EHU, Spain; João André Carriço, Universidade de Lisboa, Portugal.

Information concerning the course was widely disseminated using several media. The coordinator contacted the Food and Waterborne disease network and the Microbial coordination at the ECDC. The operators at ECDC forwarded the invitation of both the webinar and the course to their national focal points. In addition, the EURL for *Campylobacter*, Antibiotic Resistance, VTEC and *Salmonella* were informed about the course and they disseminated the information within the corresponding network.

Information concerning the course and the topic taught during the course was disseminated through local media. Below the links:

- El Confidencial: http://www.elconfidencial.com/ultima-hora-en-vivo/2017-07-07/expertos-europeos-ultiman-tecnicas-para-controlar-infecciones-alimentarias_1262989/
- El Correo farmacéutico: http://www.correofarmaceutico.com/2017/06/12/al-dia/medicina/la-seguridad-alimentaria-requiere-adentrarse-en-la-era-genomica
- La Vanguardia: http://www.lavanguardia.com/vida/20170707/423948766768/expertos-europeos-ultiman-tecnicas-para-controlar-infecciones-alimentarias.html
- Deia: http://m.deia.com/2017/07/08/sociedad/euskadi/la-genomica-lucha-contra-las-infecciones-alimentarias
- Madri+d: http://www.madrimasd.org/notiweb/noticias/expertos-europeos-ultiman-tecnicas-controlar-infecciones-alimentarias
- Radio Vitoria: http://www.eitb.eus/es/radio/radio-vitoria/programas/araba-gaur-8h/detalle/4961461/expertos-internacionales-seguridad-alimentaria-araba-gaur-1/
- Onda Vasca: https://www.ivoox.com/gasteiz-capital-seguridad-alimentaria-17-07-11-audios-mp3_rf_19731236_1.html

## G.2. Feedback from students

A total of 45 Summer Course students from different parts of Europe (mainly from public sector) and 10 Master and Doctorate from UPV/EHU have attended the course. In addition, several stakeholders followed the symposium day from a total of 52 IPs across Europe. After the course for collecting feedback, a survey has been submitted to the 45 Summer Course students and 5 persons who participated only to the webinar. The survey was performed using the EUsurvey platform and it is available at this link https://ec.europa.eu/eusurvey/runner/Vitoria-Workshop-2017. A total of 28 persons responded (56%). In general the course has been graded really well, especially for those students coming from public sector.

Tables G1 and G2 summarize the results from the feedback form.

**Table G1:** Feedback from the students, multiple choice questions

| How would you rate the | Very good | Good | Average | Poor | Very poor | I don't know |
|---|---|---|---|---|---|---|
| Workshop overall | 54% | 36% | 11% | 0% | 0% | 0% |
| The quality fo the presenations of the symposium day | 54% | 46% | 0% | 0% | 0% | 0% |
| The quality of the presentations during the hands-on activities | 46% | 21% | 18% | 0% | 0% | 14% |
| **Please indicate your level of agreement with the statements below** | **Strongly agree** | **Agree** | **Netrual** | **Disagree** | **Strongly disagree** | **I don't know** |
| The objectives of the workshop were clearly defined | 46% | 39% | 11% | 4% | 0% | 0% |
| The topics covered were relevant to me | 64% | 32% | 4% | 0% | 0% | 0% |
| The simulation of the outbreak investigation was useful and well organized | 43% | 39% | 4% | 4% | 0% | 11% |
| The instructors active and supportive | 68% | 18% | 0% | 0% | 0% | 14% |

**Table G2:** Feedback from the students, open questions

| **What did you like most about this workshop?** |
|---|
| • Great atmosphere of the group. Vitoria-Gasteiz is very interesting city. |
| • The enthusiasm of the organizers/presenters. |
| • All the speakers were very clear and knowledgeable, and gave us the possibility to explore and to be updated on the last advances on the Bioinformatics analysis of WGS data for pathogens typing. |
| • The thematic presented in a very clear way |
| • Excellent tutors and instructors, interactive atmosphere. |
| • The dedication and support of the tutors and instructions. Also, the networking possibilities that the course provided. |
| • The shown pipelines to analyse WGS data, free available and not only user friendly but also with the option of command line. And apart from the content, of course the city and the social activities |
| • I really liked the location, the friendly atmosphere, and the group activities and of course the opportunity to learn more about the Innuendo project. |
| • hand-on workshop for Innuendo tool |
| • personal contact |
| • Very good lectures the first day- not only about innuendo, but taking the opportunity to put things into a bigger context |
| • Interaction with leading researchers and other participants |
| • A combination of theoretical and practical classes, good and kind presenters (and their presentations). |
| • I like the effort of the tutors and instructor. They tried to help us as much as they could. |
| • Perfect tutors, optimal time divided for lectures and hands on training, very good location, great workshop participants. |
| • Everything was very well organized and the workshop contents were very useful. |
| • WGS has been largely been used as a research tool, including to guide therapeutic intervention. So this workshop was interesting with easy software presented which i would like to implement in my work |

| What aspects of the workshop could be improved? |
|---|
| • The room chosen for the first day (symposium) was not adequate for a screen presentation. Screen was placed too close to the floor. People seated at the back found hard to see the slides. |
| • The organization was perfect and the argumentations were a lot and very stimulating but concentrated in only two days. Maybe it could be useful to extend the workshop especially for hands-on training. |
| • More time for the hands-on activities. |
| • More time for the hands-on training. |
| • Division into groups should have accounted for previous experience and skill level. Some additional tasks could have been allocated to those who are already familiar with sequence analysis. |
| • It would be very useful to have the presentations, especially for the hands-on activities, to fully understand what the different pipelines do. |
| • It would be nice to have a print-out the task for the trainings session next time. |
| • Preparation of handouts of the talks and hand-on workshop. |
| • More time for discussing in more detail about technical details. |
| • It would be good to receive all presentations in pdf. |
| • Although some lectures were introducing topics also for non-expert users, some lectures would have needed to start slower introducing the topic rather than digging into details. |
| • Hands-on activities were a little hap-hazard. It would have been beneficial if participants could bring their own genome data. |
| • More hands-on time, more bioinformatics. I would be happy to learn to run all of the tools we were using at the workshop in a command line mode. |
| • I think that the course should include more basic contents to fully understand all the subjects explained during the oral and practical sessions. |
| **What did you take from this workshop and use in your current and/or future profession?** |
| • How to manage with big data and the different steps that should be done in order to analyse a complete genome sequence. Different techniques to compare ant type genomes. |
| • We will use the platform for epidemiological studies. |
| • I learnt about current issues on the use of WGS for the detection of food-borne pathogens in food/clinical samples and was a good atmosphere for networking. |
| • The addressed topics are essential for my PhD studies. The tools proposed during the workshop show reliability, feasibility and transparency, enough to be adopted to achieve the goals of my projects. |
| • I would really like to try and use presented tools. |
| • That WGS data analysis can be accessible to anyone; I'll use the platform in a near future in my profession. |
| • Use of INNUENDO platform in routine work (hopefully) |
| • Bioinformatics tools |
| • I would like to use the Innuendo platform when it becomes available to the public. |
| • All the pipelines. |
| • Especially the QA/QC aspects implemented in the INNUca pipeline are going to be a great help in order to access the quality and to determine thresholds for WGS data. |
| • the right knowledge to apply these new investigation methods for my job |
| • Innuendo tools will be installed and further tested and possibly used in routine for QC analysis, and outbreak investigation. |
| • Being able to download the databank, which (hopefully) will enable us to compare data with other European countries. |
| • WGS has been largely used as a research tool, including to therapeutic intervention. So this workshop was interesting with easy software presented which i would like to implemented in my lab. |
| • knowledge of Bioinformatics applied to microbiology |
| • The most useful was information about core genome WGS approach in WGS data analysis. |
| • It is useful to know what innuendo aims for, and I will also bring with me a deeper understanding of the complexity for clustering |
| • Everything |
| • I learnt quite a lot about technical aspects of Spades, core genome construction and allele calling as well as tree construction. |
| • An overview of NGS-based typing and the tools used for it. The major limitations I see in using the platform is the different chemistry (SE Ion Torrent Seq) and a limited number of pathogen species. |
| • In my opinion it is necessary to have previous skills to fully understand the contents of the course. I do not think I have the capacity to apply what I was taught in my current profession. |

- the tools are very suitable for my work, and will be helpful for future activities
- Presentations of the symposium day were very interesting for me. However, the course would be more useful if the software presented during the hands-on activity were already available.
- I am currently using and I will use all software presented in my current profession.
- In the future I hope to use INNUENDO for working with bacterial whole genome sequences.
- I started using the INNUca pipeline in command line environment.