A Validation Approach of an End-to-End Whole Genome Sequencing Workflow for Source Tracking of *Listeria monocytogenes* and *Salmonella enterica*

Anne-Catherine Portmann 1, Coralie Fournier 2, Johan Gimonet 1, Catherine Ngom-Bru 1, Caroline Barretto 1 and Leen Baert 1*

1 Nestec Ltd., Nestle Research Center, Lausanne, Switzerland. 2 Nestle Institute of Health Sciences SA, EPFL Innovation Park, Lausanne, Switzerland

Whole genome sequencing (WGS), using high throughput sequencing technology, reveals the complete sequence of the bacterial genome in a few days. WGS is increasingly being used for source tracking, pathogen surveillance and outbreak investigation due to its high discriminatory power. In the food industry, WGS used for source tracking is beneficial to support contamination investigations. Despite its increased use, no standards or guidelines are available today for the use of WGS in outbreak and/or trace-back investigations. Here we present a validation of our complete (end-to-end) WGS workflow for *Listeria monocytogenes* and *Salmonella enterica* including: subculture of isolates, DNA extraction, sequencing and bioinformatics analysis. This end-to-end WGS workflow was evaluated according to the following performance criteria: stability, repeatability, reproducibility, discriminatory power, and epidemiological concordance. The current study showed that few single nucleotide polymorphism (SNPs) were observed for *L. monocytogenes* and *S. enterica* when comparing genome sequences from five independent colonies from the first subculture and five independent colonies after the tenth subculture. Consequently, the stability of the WGS workflow for *L. monocytogenes* and *S. enterica* was demonstrated despite the few genomic variations that can occur during subculturing steps. Repeatability and reproducibility were also demonstrated. The WGS workflow was shown to have a high discriminatory power and has the ability to show genetic relatedness. Additionally, the WGS workflow was able to reproduce published outbreak investigation results, illustrating its capability of showing epidemiological concordance. The current study proposes a validation approach comprising all steps of a WGS workflow and demonstrates that the workflow can be applied to *L. monocytogenes* or *S. enterica*.

Keywords: validation, SNP, WGS, *Listeria monocytogenes*, *Salmonella enterica*, phylogeny, source tracking, outbreak
INTRODUCTION

The use of Whole genome sequencing (WGS) for outbreak investigation and pathogen source tracking has been described in several publications since a few years (Leekitcharoenphon et al., 2014; Schmid et al., 2014; Octavia et al., 2015; Wuys et al., 2015; Dallman et al., 2016; Jackson et al., 2016; Wilson et al., 2016; Inns et al., 2017). Current sequencing platforms (Illumina, Ion torrent, Oxford, PacBio) can be used and many bioinformatics approaches are available such as whole genome Multi Locus Sequence Typing (wgMLST), core genome MLST (cgMLST) or high quality Single Nucleotide Polymorphism (hqSNP). Several public health agencies and food authorities are using WGS for outbreak investigations or pathogen source tracking (Chen et al., 2017; Inns et al., 2017; Katz et al., 2017; Moran-Gilad, 2017), and food industries are gradually evaluating or implementing the WGS technology. In the food industry, WGS used for source tracking is beneficial to provide leads on the true cause of a contamination event, to find patterns and eventually prevent re-occurring issues. No standards or guidelines are available today on the use of WGS applied to outbreak and/or traceback investigations. Working groups have been set up (e.g., International Life Sciences Institute, ILSI and Global Microbial Identifier, GMI) to address harmonization of methodologies and results interpretation. In addition, the development of a standard has started within the International Organization for Standardization (ISO TC34-SC9-WG25) and aims at developing an internationally harmonized WGS methodology for source tracking.

The objective of the current study was to validate the end-to-end WGS workflow used for source tracking of Listeria monocytogenes and Salmonella enterica. Validation is crucial as WGS is used as a typing method. Thus, it is important to ensure the reliability when similarities or differences between genomes are observed. Since WGS for source tracking is an analytical approach often carried out by multiple groups/laboratories of different expertise/technologies (microbiology, sequencing technologies, bioinformatics, and genomics), it is important to emphasize the validation should cover the end-to-end workflow including all steps. The end-to-end WGS workflow in the current study consisted of four steps: subculturing isolates to obtain pure colonies, extracting DNA, performing short read sequencing with MiSeq Illumina and carrying out bioinformatics analysis based on read mapping allowing identification of hqSNPs with the CFSAN SNP Pipeline. Even though the CFSAN SNP Pipeline was previously evaluated on its robustness and accuracy (Davis et al., 2015), the end-to-end workflow has never been fully validated. The validation was done following the recommendations of the members of the European Study Group on Epidemiological Markers (ESGEM) and of the European Society for Clinical Microbiology and Infectious Diseases (ESCMID) (Streuels, 1996). In order to validate the end-to-end workflow from isolate subculture to bioinformatics analysis, the following performance criteria were tested: stability, repeatability, reproducibility and discriminatory power. The bioinformatics analysis pipeline was used to validate the epidemiological concordance. The evaluation of the stability assesses that there is no impact of genomic variation that might occur during bacterial replications, on the complete WGS workflow ability to recognize the clonal relatedness of isolates. Repeatability evaluates the impact of reiterating all the steps of the workflow by the same operators (sequencing of isolates extracted during the same DNA extraction, and sequencing run). Reproducibility evaluates the impact of different operators on DNA extraction and on sequencing, with different sequencing instruments and in different laboratories. The purpose of assessing the discriminatory power is to ensure that unrelated strains can be differentiated while revealing relatedness of clonal isolates. The epidemiological concordance evaluates the capability of the bioinformatics part of the WGS workflow to reproduce results from previous published outbreak investigations that used different bioinformatics software. In addition, the impact of using PacBio contigs or Illumina contigs as reference for the read mapping was evaluated.

The current study proposes a validation approach comprising all steps of a WGS workflow. This study presents results on performance characteristics applied to our complete WGS workflow used for source tracking of L. monocytogenes and S. enterica.

MATERIALS AND METHODS

Bacteria and Subculture Conditions

Listeria monocytogenes and Salmonella enterica strains (Table 1) were stored on cryobeads (TSC, Lancashire, UK) at −80°C. Strains were streaked on Trycase Soy Agar (TSA) (BioMérieux, Marcy l’Etoile, France) to obtain single colonies. After 24 ± 2 h incubation at 37°C, one colony was used to inoculate 4 mL Brain Heart Infusion (BHI) (Thermo Scientific™ Oxoid™, Hampshire, UK) and was incubated at 37°C for 6–8 h. One mL of inoculated BHI was taken and centrifuged (5,000 × g, 5 min). The pellet was stored at −20°C until the DNA extraction was performed.

DNA Extraction

DNA Extraction for Short Read Sequencing: Illumina MiSeq and HiSeq

Listeria monocytogenes

The bacterial pellet was suspended in 160 μL buffer P1 (Qiagen, Hilden, Germany). Twenty microliter of lysozyme from chicken egg solution (Sigma, 200 mg/mL in DNA free water) was added and incubated at 37°C with agitation at 900 rpm for 2 h (Eppendorf Thermomixer C). Proteinase K (20 μL; Qiagen) was added and the solution was incubated 1 h at 56°C. The QIAamp DNA Mini Kit was used according to the supplier’s instructions. DNA was eluted in 100 μL Elution buffer (Qiagen) and stored at −20°C until further analysis.

Abbreviations: CDC, Center for Disease Control and Prevention; CFSAN, Center for Food Safety and Applied Nutrition; ESGEM, European Study Group on Epidemiological Markers; FDA/USFDA, Food and Drug Administration; GMI, Global Microbial Identifier; ILSI, International Life Sciences Institute; ISO, International Organization for Standardization; NCBI, National Center for Biotechnology Information; PHAC, Public Health Agency of Canada; PHE, Public Health England; UK, United Kingdom; USA, United States of America.
**Salmonella enterica**

The bacterial pellet was suspended in 180 μL Buffer ATL (Qiagen). Proteinase K (20 μL; Qiagen) was added and the solution was incubated 1 h at 56°C. The QIAamp DNA Mini kit was used according to the supplier’s instructions. DNA was eluted in 100 μL Elution buffer (Qiagen) and stored at −20°C until further analysis.

The Qubit™ dsDNA HS Assay kit (Invitrogen) was used to measure the DNA concentration with the Qubit® 2.0 Fluorometer (Invitrogen) according to the supplier's instructions. DNA was standardized to 10 ng/μL with Elution buffer (Qiagen) and stored at −20°C until used for sequencing.

**DNA Extraction for Long Read Sequencing: Pacific Biosciences (PacBio)**

The Gentra Puregene Yeast/Bact. kit (Qiagen) was used for the DNA purification of bacterial cultures.

**Listeria monocytogenes**

One colony (see section Bacteria and Subculture Conditions) was used to inoculate BHI broth and grown at 37°C overnight. Five mL of broth were centrifuged (3,500 rpm, 5 min) and supernatant was discarded. The pellet was washed three times using 1 mL of Phosphate-Buffered Saline (PBS). DNA extraction was carried out following supplier’s instructions indicated in the following protocol: DNA purification from Gram-Negative bacteria.

DNA was stored at −20°C until further analysis. The Qubit™ dsDNA BR Assay kit (Invitrogen) was used to measure the DNA concentration with the Qubit® 2.0 Fluorometer (Invitrogen) according to the supplier’s instructions.

**Sequencing**

**Short Read Sequencing (Illumina MiSeq and HiSeq)**

DNA was normalized at 0.2 ng/μL in order to start with 1 ng to perform a sequencing library preparation using Nextera XT kit (Illumina) following the supplier’s instructions. A final AMPure beads purification at ratio 0.6 was performed on a Sciclone robotic platform from Perkin Elmer. The quality and quantity of each library were evaluated using a capillary electrophoresis method (LabChip GX Touch from Perkin Elmer). Libraries were pooled based on molarity calculated by the LabChip GX Touch. The equimolar pool was assembled using a Hamilton robotic platform.

**MiSeq**

The sequencing was performed on a MiSeq platform (Illumina) using v2 chemistry, for a 2 × 250 cycles run. The pool was spiked with 2% PhiX, loaded at 12 pM and 16 to 19 samples were loaded per MiSeq flow cell (corresponding to ∼2 million reads pass filter per sample).

**HiSeq**

To ensure each library was present in the pool before sequencing, the equimolar pool was controlled by a MiSeq run v2 chemistry for 2 × 20 cycles. Then the pool was sequenced on a HiSeq 2500 platform (Illumina) using Rapid v2 chemistry. The pool was spiked with 2% PhiX, loaded between 9 or 12 pM and respectively 190 or 84 samples were loaded per HiSeq flow cell (corresponding to ∼3 or ∼9 million reads pass filter per sample).

---

**TABLE 1 | Stability and Repeatability dataset for L. monocytogenes and S. enterica.**

| Strain ID | Biosample | PacBio accession number | Genus/species | Serotype |
|----------|-----------|------------------------|---------------|----------|
| PIR00542 | SAMN08125517 | CP025222 | L. monocytogenes | 1/2a |
| PIR00543 | SAMN08125518 | CP025221 | L. monocytogenes | 1/2a |
| PIR00545 | SAMN08125520 | CP025560, CP025561 | L. monocytogenes | 1/2b |
| PIR00540 | SAMN08125515 | CP025568, CP025569 | L. monocytogenes | 1/2c |
| PIR00541 | SAMN08125516 | CP025566, CP025567 | L. monocytogenes | 1/2c |
| PIR00544 | SAMN08125519 | CP025562, CP025563, CP025564, CP025565 | L. monocytogenes | 4b |
| PIR00546 | SAMN08125521 | CP025220 | L. monocytogenes | 4b |
| PIR00547 | SAMN08125522 | CP025219 | L. monocytogenes | 4b |
| PIR00532 | SAMN08125525 | CP025557, CP025558, CP025559 | S. enterica | Enteritidis |
| PIR00558 | SAMN08125532 | CP025553, CP025554 | S. enterica | Enteritidis |
| PIR00533 | SAMN08125526 | PKPH00000000 | S. enterica | Hadar |
| PIR00534 | SAMN08125527 | PKPG00000000 | S. enterica | Hadar |
| PIR00537 | SAMN08125530 | CP025217 | S. enterica | Tennessee |
| PIR00535 | SAMN08125528 | CP025218 | S. enterica | Tennessee |
| PIR00536 | SAMN08125529 | CP009102 | S. enterica | Typhimurium |
| PIR00538 | SAMN08125531 | CP025555, CP025556 | S. enterica | Typhimurium |
All sequences have been submitted to the National Center for Biotechnology Information (NCBI) in BioProject: PRJNA420913 and all sequence read archive numbers (SRR) are available in Tables 2–6.

**Long Read Sequencing (Pacific Biosciences)**

High molecular weight DNA was sheared using g-TUBE (Covaris) to obtain around 20 kb DNA fragments. After shearing, the DNA size distribution was checked using the TapeStation 2500 system (Agilent) or Fragment Analyzer (Advanced Analytical).

DNA was quantified using the Qubit system (Qubit 2.0 Fluorometer and dsDNA Assay HS) and around 5 µg of the sheared DNA was used to prepare a SMRTbell library following the protocol Procedure and Checklist – 20 kb Template Preparation Using BluePippin™ Size-Selection System. (Pacific Biosciences). The library was size selected using a BluePippin system (Sage Science). Due to low DNA quantity after shearing for isolates PIR00544 and PIR00547, the 10–20 kb Template Preparation and Sequencing with Low-Input DNA protocol (Pacific Biosciences) was applied without size selection.

The library was sequenced using RSII platform (Pacific Biosciences) on one or two SMRT cells with P6-C4 chemistry and MagBeads loading (Pacific Biosciences). Sequencing time was between 240 and 380 min.

All sequences accession numbers are available in Table 1 (BioProject: PRJNA420913).

**Bioinformatics Analysis**

**Raw Data Quality Check**

In order to evaluate the sequencing run quality, fastq files were assessed with the FastQC software (v.011.5) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). To choose the best isolate for genome assembly that will serve as reference during hqSNP analysis, the reads should pass the “per base sequence quality” FASTQC threshold.

**Genome Assembly**

Illumina reads were first trimmed with Trimmomatic software (version 0.36) (Bolger et al., 2014). Bases at both ends of reads generated with Illumina were removed within a sliding window of 10 base pairs when the average quality in this window was lower than Q20 score (equivalent Phred score). Then, SPAdes (v3.9.0) (Bankevich et al., 2012) de novo assembled the trimmed reads and generated a multi contigs genome.

The PacBio genomes were built with a de novo assembly carried out using the PacBio Hierarchical Genome Assembly Process (HGAP) version 3 (Pacific Biosciences), closure of the genome by Circulator (Hunt et al., 2015) and final polishing by Quiver v1 (Pacific Biosciences).

**Single Nucleotide Polymorphism (SNP) Analysis**

High quality SNP pipeline developed by the Center for Food Safety and Applied Nutrition (CFSAN SNP Pipeline v.1.0.0/FDA) was used for SNP calling on L. monocytogenes and S. enterica isolates (Davis et al., 2015). The reference genome was either a genome generated by PacBio assembly, a public available genome or a genome assembly from Illumina reads.
TABLE 3  |  L. monocytogenes sequence data used for assessing the reproducibility by hqSNP analysis.

| Strain ID | Sequence Read Archive number | DNA extraction operator (DEO) | Sequencing operator (SO) | Sequencing laboratories (SL) | Accession number for PacBio sequence used as reference |
|-----------|-----------------------------|-------------------------------|-------------------------|-----------------------------|--------------------------------------------------------|
| PIR00542  | *SRR6347440                 | DEO1                          | SO1                     | SL1                         | CP025222                                               |
| SRR6347433| DEO1                        |                               | SO2                     | SL1                         |                                                        |
| SRR6347434| DEO1                        |                               | SO3                     | SL2                         |                                                        |
| SRR6347435| DEO2                        | SO1                           |                         | SL1                         |                                                        |
| SRR6347436| DEO2                        | SO2                           |                         | SL1                         |                                                        |
| SRR6347430| DEO2                        | SO3                           |                         | SL2                         |                                                        |
| PIR00544  | *SRR6347431                 | DEO1                          | SO1                     | SL1                         | CP025582, CP025583, CP025564, CP025565                |
| SRR6347459| DEO1                        |                               | SO2                     | SL1                         |                                                        |
| SRR6347458| DEO1                        |                               | SO3                     | SL2                         |                                                        |
| SRR6347457| DEO2                        | SO1                           |                         | SL1                         |                                                        |
| SRR6347456| DEO2                        | SO2                           |                         | SL1                         |                                                        |
| SRR6347463| DEO2                        | SO3                           |                         | SL2                         |                                                        |
| PIR00547  | *SRR6347462                 | DEO1                          | SO1                     | SL1                         | CP025219                                               |
| SRR6347461| DEO1                        |                               | SO2                     | SL1                         |                                                        |
| SRR6347460| DEO1                        |                               | SO3                     | SL2                         |                                                        |
| SRR6347455| DEO2                        | SO1                           |                         | SL1                         |                                                        |
| SRR6347454| DEO2                        | SO2                           |                         | SL1                         |                                                        |
| SRR6347526| DEO2                        | SO3                           |                         | SL2                         |                                                        |

*Two different operators performed the DNA extractions (DEO1 and DEO2).

Three different sequencing operators (SO1, SO2, and SO3) sequenced the DNA at two different sequencing laboratories (SL1 and SL2).

For each analysis two references were used and evaluated: de novo assembled Illumina reads of isolate (*) and PacBio contigs.

Phylogenetic Analysis
Maximum-likelihood phylogenetic trees were built with GARLI (Version 2.01.1067) on the SNP analysis result and visualized with Figtree (version 1.4.3) (http://tree.bio.ed.ac.uk/software/figtree/). All trees shown in this paper are midpoint rooted.

Experimental Design for Performance Criteria
Stability
Listeria monocytogenes and S. enterica strains from −80°C were streaked on TSA and Columbia agar (BioMérieux) respectively, for 24 ± 2 h at 37°C to obtain single colonies. One colony was dissolved in 0.5 mL tryptone (1 g/L; Oxoid) salt (8.5 g/L; Merck, Darmstadt, Germany) solution streaked on Columbia agar (L. monocytogenes) or TSA (S. enterica), and incubated for 24 ± 2 h at 37°C to ensure colonies were pure. This procedure was repeated three times. Afterwards, five colonies (A, B, C, D, and E) were picked from the same agar plate and each colony was grown in 4 mL BHI broth as indicated in section Bacteria and Subculture Conditions. DNA was extracted as described in section DNA Extraction and sequenced as described in section Sequencing. The dataset used to evaluate stability is presented in Tables 1, 2. SNP analysis was carried out using for each isolate two reference genomes, one created by PacBio sequencing and one by using the contigs after assembling HiSeq sequence reads obtained from colony A.

Repeatability
The repeatability was evaluated using the sequences obtained from the eight L. monocytogenes and eight S. enterica from the stability experiment (colonies A, B, C, D, and E as described in Table 2). DNA was extracted the same day by the same operator and isolates were sequenced on the same HiSeq run with the same library preparation.

SNP analysis was carried out using for each isolate two reference genomes, one created by PacBio sequencing and one by using the contigs after assembling HiSeq sequence reads obtained from colony A.

Reproducibility
Listeria monocytogenes strains PIR00542, PIR00544 and PIR00547 and S. enterica strains PIR00534, PIR00536 and PIR00503 were selected to perform the reproducibility experiment. DNA extractions of each set of isolates were carried out by two operators. DNA obtained by each operator was split in three tubes per isolate and sequenced on different Illumina MiSeq instruments using different batches of V2 chemistry by...
three different operators in two different laboratories. A total of
18 genomes sequences per species was analyzed.

SNP analysis was carried out using for each isolate two
reference genomes, one created by PacBio sequencing and one
by using the contigs after assembling MiSeq sequence reads
(Supplementary Materials reproducibility table). The dataset
used to evaluate reproducibility is presented in Tables 3, 4.

Discriminatory Power
Since the WGS method has a high resolution capability, not only
isolates known to be unrelated (derived from the information
provided by bacterial reference collections of the selected strains),
but also isolates having clonal relatedness were included in the
analysis in order to show that the end-to-end workflow was able
discriminate correctly while not missing similarities.

The choice of L. monocytogenes strains to be included in the
discriminatory power experiment was based on serotype
selection (4b, 1/2a, and 1/2c). For serotype 4b, two strains known
to be unrelated were selected (Table 5). With this design, four
distinct groups of genomes were supposed to be discriminated
while not missing similarities within groups. Table 5 summarizes
the different groups of related L. monocytogenes isolates.

Groups 1, 2, and 3 were respectively composed of sequences
from independent colonies A, B, and C coming from three
different strains of the stability experiment. Group 4 was composed
of two NCTC 11994 isolates bought from BioMérieux
at two different dates (2011: PIR00492 and 2015: PIR00548).

For S. enterica, three serovars (Enteritidis, Typhimurium and
Tennessee) were selected to perform three distinct discriminatory
experiments. For each serovar, strains were selected based
on the epidemiological relationships information, in order to
create three to four groups to be discriminated while not
missing similarities within groups. Table 6 summarizes the
groups per serovar. For the three serovars, groups 1 and 2 were
both composed of sequences obtained from three independent
colonies (A, B, and C) coming from two different strains of the
stability experiment.

For S. enteritidis, group 3 was composed of three related
isolates from an European egg related outbreak (Dallman et al.,
2016). Group 4 was composed of two sequenced isolates from a
Belgian egg related outbreak published by Wuyts and coworkers
(Wuyts et al., 2015).

For S. Tennessee, group 3 was composed of two sequenced
isolates from a peanut butter outbreak in USA (2006–2007)
published in 2016 by CFSAN FDA (Wilson et al., 2016).

For S. Typhimurium, group 3 was composed of two sequenced
isolates from a large Denmark outbreak (Leekitcharoenphon
et al., 2014). Group 4 was composed of two sequenced isolates
from an Australian chocolate mousse outbreak (Octavia et al.,
2015).

Epidemiological Concordance
Struelens and coworkers indicated that the epidemiological
concordance should be used to verify that the method is able

---

**Table 4 | S. enterica sequence data used for assessing the reproducibility by hqSNP analysis.**

| Isolate ID | Sequence Read Archive number | DNA extraction operator (DEO) | Sequencing operator (SO) | Sequencing laboratories (SL) | Accession number for PacBio sequence used as reference |
|------------|------------------------------|-------------------------------|--------------------------|-------------------------------|-----------------------------------------------------|
| PIR00534   | *SRR6347416                  | DEO1                          | SO1                      | SL1                           | PKPG0000000000                                       |
| SRR6347417 | SRR6347414                  | DEO1                          | SO2                      | SL1                           |                                                     |
| SRR6347415 | DEO2                          | SO3                           | SL2                      |                               |                                                     |
| SRR6347420 | SRR6347421                  | DEO2                          | SO2                      | SL1                           |                                                     |
| SRR6347423 | SRR6347445                  | DEO1                          | SO3                      | SL1                           |                                                     |
| SRR6347447 | SRR6347444                  | DEO2                          | SO3                      | SL2                           |                                                     |

---

Table 4 displays the sequence data used for assessing the reproducibility by hqSNP analysis for S. enterica isolates. The data includes isolate IDs, sequence read archive numbers, DNA extraction operators, sequencing operators, sequencing laboratories, and accession numbers for PacBio sequence used as reference. The columns indicate the specific details for each isolate, such as the DNA extraction operator (DEO), sequencing operator (SO), sequencing laboratories (SL), and the accession number. The table highlights the reproducibility assessment by comparing different operators and laboratories to ensure consistent results. The table entries are structured to facilitate easy comparison and analysis of the data across different conditions. The table entries include various identifiers and values relevant to the study's methodology. The table is formatted to clearly display these details in a structured manner, aiding in the analysis and understanding of the reproducibility assessment.

**Notes:**
- Two different operators performed the DNA extractions (DEO1 and DEO2).
- Three different sequencing operators (SO1, SO2, and SO3) sequence the DNA at two different sequencing laboratories (SL1 and SL2).
- For each analysis two references were used and evaluated: de novo assembled Illumina reads of isolate (*) and PacBio contigs.
TABLE 5 | L. monocytogenes sequence dataset for assessing the discriminatory power by hqSNP analysis.

| Serotype | Epidemiological relationship | Isolate ID | SRR number | Groups with clonal relationship |
|----------|-----------------------------|------------|------------|--------------------------------|
| 1/2c     | ATCC 51779 (Belgium) Independent colonies from the first subculture | PIR00541 (A) | SRR6347510 | Group 1                        |
|          |                             | PIR00541 (B) | SRR6347509 |                                |
|          |                             | PIR00541 (C) | SRR6347512 |                                |
| 1/2a     | ATCC 51775 (Belgium) Independent colonies from the first subculture | PIR00542 (A) | SRR6347487 | Group 2                        |
|          |                             | PIR00542 (B) | SRR6347488 |                                |
|          |                             | PIR00542 (C) | SRR6347489 |                                |
| 4b       | ATCC 13932 (Germany) Independent colonies from the first subculture | PIR00547 (A) | SRR6347353 | Group 3                        |
|          |                             | PIR00547 (B) | SRR6347402 |                                |
|          |                             | PIR00547 (C) | SRR6347351 |                                |
|          | NCTC 11994 (UK)             | PIR00492    | SRR6347453 | Group 4                        |
|          |                             | PIR00548a   | SRR6347452 |                                |
|          |                             | PIR00548b   | SRR6347437 |                                |
|          |                             |             | SRR6347438 |                                |
|          |                             |             | SRR6347439 |                                |

a For this isolate two genome sequences were generated from independent subcultures.
b For this isolate three genome sequences were generated from independent subcultures.

to obtain the same conclusions as the ones from well-defined outbreak studies (Struelens, 1996). In this study, one published outbreak of L. monocytogenes and one of S. enterica were analyzed and the outcomes of the bioinformatics analysis were compared to the original findings. For each outbreak, the analysis was done following two approaches. In the first approach, the analysis was performed using the isolate reference of the authors with the CFSAN SNP Pipeline which is part of our workflow. Then, in a second approach, the analysis was reproduced using the food isolate as reference.

RESULTS
Stability
The stability of L. monocytogenes and S. enterica was evaluated by the selection of eight L. monocytogenes and eight S. enterica strains. The selection of strains covered the major serotypes related to food. For each strain ten genome sequences were generated: five from independent colonies of the first subculture (A, B, C, D, and E) and five from independent colonies after the tenth subculture (A1, A2, A3, A4, and A5). The number of SNPs was determined between the 10 generated genome sequences per strain.

Table 7 summarizes the results of the hqSNP analysis for L. monocytogenes. For L. monocytogenes PIR00540, PIR00542, PIR00543, PIR00545, PIR00546, and PIR00547, no SNP was found between colonies of the first subculture (A, B, C, D, and E) and after the tenth subculture (A1, A2, A3, A4, and A5). For two strains, PIR00541 and PIR00544, few SNPs were identified. For PIR00541, colony B had 1 SNP compared to colonies A, C, D, and E and after the tenth subculture (A1, A2, A3, A4, and A5). For PIR00544, colony B had 1 SNP compared to all others colonies (A, B, C, D, A1, A3, A4, and A5). The number of SNPs was determined between the 10 generated genome sequences per strain.

Table 8 summarizes the results of the hqSNP analysis for S. enterica per serovar. For S. Hadar, no SNP was found between all colonies for the two strains PIR00533 and PIR00534. No SNP was found between all colonies for S. Typhimurium strain PIR00536. The hqSNP analysis of sequences obtained from strain PIR00538 revealed that colony A2 had 1 SNP compared to all others colonies (A, B, C, D, E, A1, A3, A4, and A5). For S. Enteritidis, no SNP was found between all colonies for strain PIR00532. Colonies A1, A2, A3, A4, and A5 of strain PIR00558 had 1 SNP compared to the colonies from the first subculture (A, B, C, D, and E). Finally, for S. Tennessee, no SNP was found between all colonies for strain PIR00537, while PIR00535 colonies A1, A2, A3, A4, and A5, after the tenth subculture, had 1 SNP compared to the colonies from the first subculture (A, B, C, D, and E).

Repeatability
The sequence data used were selected from the stability experiment (colonies A, B, C, D, and E as described in Tables 1, 2) were used to assess the repeatability. In total 80 sequences from eight L. monocytogenes strains and eight S. enterica strains were analyzed.

Table 7 summarizes the results obtained for L. monocytogenes. No SNP was found between the five independent colonies (A, B, C, D, and E) for seven strains out of eight. For PIR00541, 1 SNP was observed in colony B compared to colonies A, C, D, and E after the tenth subculture (A1, A2, A3, A4, and A5). The number of SNPs was determined between the 10 generated genome sequences per strain.
**Table 6 |** S. enterica sequence dataset for assessing the discriminatory power by hqSNP analysis.

| Serovars   | Isolate ID          | Sequence Read Archive numbers | Epidemiological relationship                                                                 | Groups with clonal relationship |
|------------|---------------------|-------------------------------|-------------------------------------------------------------------------------------------------|---------------------------------|
| S. Enteritidis | PIR00532 (A) | SRR6347380                  | Independent colonies from the first subculturea                                               | Group 1                         |
|            | PIR00532 (B)     | SRR6347381                  |                                                                                                 |                                 |
|            | PIR00532 (C)     | SRR6347378                  |                                                                                                 |                                 |
|            | PIR00558 (A)     | SRR6347399                  | ATCC BAA-708 Independent colonies from the first subculture                                     | Group 2                         |
|            | PIR00558 (B)     | SRR6347398                  |                                                                                                 |                                 |
|            | PIR00558 (C)     | SRR6347397                  |                                                                                                 |                                 |
| PHE: 32476 | SRR1965122       |                              | European egg outbreak                                                                           | Group 3                         |
| PHE: 21785 | SRR1965313       |                              |                                                                                                 |                                 |
| PHE: 32477 | SRR1966289       |                              |                                                                                                 |                                 |
| S14BD001753| SRR2088895       |                              | Belgian egg outbreak                                                                            | Group 4                         |
| S14FP01877 | SRR2088898       |                              |                                                                                                 |                                 |
| S. Tennessee | PIR00535 (A) | SRR6347498                  | ATCC 10722 Independent colonies from the first subculture                                      | Group 1                         |
|            | PIR00535 (B)     | SRR6347497                  |                                                                                                 |                                 |
|            | PIR00535 (C)     | SRR6347496                  |                                                                                                 |                                 |
|            | PIR00537 (A)     | SRR6347535                  | Independent colonies from the first subculture                                                  | Group 2                         |
|            | PIR00537 (B)     | SRR6347534                  |                                                                                                 |                                 |
|            | PIR00537 (C)     | SRR6347537                  |                                                                                                 |                                 |
| CFSAN001387| SRR965704        |                              | American peanut butter outbreaks                                                                | Group 3                         |
| CFSAN001349| SRR1177176       |                              |                                                                                                 |                                 |
| S. Typhimurium | PIR00536 (A) | SRR6347358                  | ATCC13311 Independent colonies from the first subculture                                       | Group 1                         |
|            | PIR00536 (B)     | SRR6347361                  |                                                                                                 |                                 |
|            | PIR00536 (C)     | SRR6347521                  |                                                                                                 |                                 |
|            | PIR00538 (A)     | SRR6347364                  | Independent colonies from the first subculture                                                  | Group 2                         |
|            | PIR00538 (B)     | SRR6347365                  |                                                                                                 |                                 |
|            | PIR00538 (C)     | SRR6347366                  |                                                                                                 |                                 |
| Styph-0803T7157| ERR277220     |                              | Large Denmark outbreak                                                                         | Group 3                         |
| Styph-0808S61603| ERR277226  |                              |                                                                                                 |                                 |
| 1687       | SRR1645473       |                              | Australian chocolate mousse outbreak                                                            | Group 4                         |
| 1700       | SRR1645486       |                              |                                                                                                 |                                 |

aSequence data obtained from the stability experiment described in Table 2.

Table 8 summarizes the results obtained for S. enterica. For the eight tested S. enterica strains, no SNP was found between the five independent colonies (A, B, C, D, and E).

**Reproducibility**

A total of 18 L. monocytogenes genome sequences and 18 S. enterica genome sequences were analyzed. For L. monocytogenes and S. Typhimurium and S. Hadar, no SNP was found between genome sequences obtained from different laboratories, operators and instruments.

**Discriminatory Power**

The results obtained for the discriminatory power evaluation of L. monocytogenes and Salmonella enterica, were in agreement with the design (Tables 5, 6).

Listeria monocytogenes strains of different serotypes were discriminated with more than 30,000 SNPs between groups, as well as strains of serotype 4b with more than 8,000 SNPs. Isolates within a group had a maximum of 1 SNP difference.

For S. Enteritidis, S. Tennessee, and S. Typhimurium, hqSNP analysis was performed per serotype. For the three serovars, results were in accordance with the design (Table 6). For S. Enteritidis, the groups were discriminated with more than 100 SNPs and isolates within a group had a maximum of 4 SNPs difference. The analysis of S. Tennessee isolates revealed more than 80 SNPs between the groups. Isolates within a group had a maximum of 5 SNPs difference. For the last serotype S. Typhimurium, the groups were discriminated with more than 490 SNPs and isolates within a group had a maximum of 3 SNPs.
TABLE 7 | Maximum SNPs observed during the stability test of *L. monocytogenes* between five independent colonies from the first subculture and after the tenth subculture.

| Isolate ID | Serotype | Max. SNPs between colonies within passage (A vs. B vs. C vs. D vs. E) (A1 vs. A2 vs. A3 vs. A4 vs. A5) | Max. SNPs after 10 subcultures (A, B, C, D, E) vs. (A1, A2, A3, A4, A5) |
|------------|----------|-------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| PIR00542   | 1/2a     | 0                                                                                                | 0                                                                                               |
| PIR00543   | 1/2a     | 0                                                                                                | 0                                                                                               |
| PIR00545   | 1/2b     | 0                                                                                                | 0                                                                                               |
| PIR00540   | 1/2c     | 0                                                                                                | 0                                                                                               |
| PIR00541   | 1/2c     | 1                                                                                                | 1                                                                                               |
| PIR00544   | 4b       | 2                                                                                                | 3                                                                                               |
| PIR00546   | 4b       | 0                                                                                                | 0                                                                                               |
| PIR00547   | 4b       | 0                                                                                                | 0                                                                                               |

TABLE 8 | Maximum SNPs observed during the stability test of *Salmonella enterica* between five independent colonies from the first subculture and after the tenth subculture.

| Isolate ID | Serotype | Max. SNPs between colonies within passage (A vs. B vs. C vs. D vs. E) (A1 vs. A2 vs. A3 vs. A4 vs. A5) | Max. SNPs after 10 subcultures (A, B, C, D, E) vs. (A1, A2, A3, A4, A5) |
|------------|----------|-------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| PIR00532   | Enteritidis | 0                                                                                               | 0                                                                                               |
| PIR00538   | Enteritidis | 0                                                                                                | 1                                                                                               |
| PIR00533   | Hadar     | 0                                                                                                | 1                                                                                               |
| PIR00534   | Hadar     | 0                                                                                                | 0                                                                                               |
| PIR00537   | Tennessee | 0                                                                                                | 0                                                                                               |
| PIR00535   | Tennessee | 0                                                                                                | 1                                                                                               |
| PIR00536   | Typhimurium | 0                                                | 0                                                                                               |
| PIR00538   | Typhimurium | 0                                                | 1                                                                                               |

Epidemiological Concordance

Jackson and coworkers reported the analysis of a *L. monocytogenes* outbreak presumably involving one patient from Ohio who had consumed a contaminated prepackaged romaine lettuce (Jackson et al., 2016). The Public Health Agency of Canada (PHAC) used Single Nucleotide Variant Phylogenomic pipeline (Petkau et al., 2017). While the Center for Disease Control and Prevention (CDC) used hqSNP (Lyve-SET; Katz et al., 2017) and wgMLST (BioNumerics, Applied Maths) to analyze this case. Twenty-six additional patient isolates, indistinguishable by pulsed-field gel electrophoresis (PFGE) analysis, were added to this investigation. Only the patient isolate from Ohio was found closely related to the prepackaged romaine lettuce. SRR1451258 and SRR972392 had 28 and 51 SNPs and all the other isolates had ≥80 SNPs with the lettuce isolate, respectively (Figure 1). With this analysis, we reproduced the clustering of the Ohio patient isolate with the prepackaged lettuce isolate.

In case no PFGE profile data are available and to conclude if any patient isolate is related to the food, the lettuce isolate would be our preferred reference to use with our methodology. Therefore, in the second approach, analysis was performed on the 26 patients isolates, the Ohio isolate and the lettuce isolate. The Ohio patient isolate grouped with the lettuce isolate with 1 SNP difference. SRR1451258 and SRR972392 had 28 and 51 SNPs respectively, and all the other isolates had ≥75 SNPs with the lettuce isolate (Figure 2).

In England and Wales, the national reference laboratory uses WGS with hqSNP analysis routinely, since April 2015, on all cultures of *Salmonella* sp. referred by local laboratories (Inns et al., 2017). In May 2015, they found a first cluster of 29 *S. enteritidis* cases. In the following months, this cluster increased and Public Health England (PHE) declared an outbreak. The outbreak control team was in charge of finding the source of infection. A total of 136 cases were identified from a large outbreak over 20 months implicating UK patients, Spanish
The objective of this work was to validate the end-to-end WGS workflow used for source tracking analysis of *L. monocytogenes* and of *S. enterica*. Previously, molecular subtyping tools such as PFGE had been used for source tracking to differentiate related and nonrelated microbial strains to support epidemiological investigations of foodborne outbreaks (Ronholm et al., 2016). Nowadays WGS is replacing conventional molecular subtyping tools due to the higher resolution of the technique (Nadon et al., 2017). The implementation of a new typing technique requires an evaluation of its performance characteristics as previously described (Struelens, 1996). Validation of genomic typing tools is not evident because of the absence of standardized technical procedures, reference materials and criteria for results interpretation. In the current study, WGS applied to source tracking of *L. monocytogenes* and of *S. enterica* was validated using the following performance criteria: stability, repeatability, reproducibility, discriminatory power, and epidemiological concordance (Struelens, 1996). Although the bioinformatics CFSAN SNP Pipeline was previously evaluated on its robustness and accuracy on a dataset of 1,000 *in silico* mutated genomes (Davis et al., 2015), the end-to-end workflow including subculturing, DNA extraction, sequencing and bioinformatics analysis has to the best of our knowledge not been validated. This is the first study addressing the validation of the complete WGS workflow for the application of source tracking. The high resolution of WGS is known to be able to discriminate isolates that were previously shown to have the same PFGE profiles (Moura et al., 2016) or combined PFGE-MLVA (Multi Loci Variable number tandem repeat) profiles (den Bakker et al., 2014; Jackson et al., 2016). Consequently, the differences between genomes identified by WGS need to be trusted and a validation of all steps of the WGS workflow is recommended.

The evaluation of the stability assessed whether the WGS workflow was able to recognize the clonal relatedness of isolates despite the genomic variation that might occur with bacterial replication (Struelens, 1996). We performed a comparison of genomic sequences from five independent colonies from the first subculture and five independent colonies after the tenth subculture. In our experimental setup, few SNPs were observed for the eight *L. monocytogenes* and eight *S. enterica* strains.

**DISCUSSION**

### Comparison of hqSNP Results Using Reference Genomes Obtained From Illumina and PacBio Sequencing Technologies

HqSNP analysis is based on the mapping of Illumina reads on a reference genome. As a good practice, not only the reference genome needs to be genetically related to the genomes analyzed, but also needs to be the most complete as possible, since absent regions will not be analyzed and potential SNPs in these areas would not be detected. Even if there is an advantage of PacBio read length over Illumina, it is less costly and faster to use Illumina data to obtain a reference genome for hqSNP analysis. For that reason, in this study, the impact of the sequencing technology to obtain the reference genome was assessed by using as reference a genome obtained with de novo assembled Illumina contigs and a genome obtained with PacBio contigs for *L. monocytogenes* and *S. enterica* strains.

The evaluation was done by the SNP calling on sequences obtained from the stability, repeatability and reproducibility experiments. A total of 22 results obtained with PacBio reference genomes were compared to 22 results obtained with de novo assembled Illumina contigs. For all *L. monocytogenes* and *S. enterica* strains, the exact same SNPs were identified when using the PacBio genome or using the Illumina assembled contigs.
evaluated. Allard and coworkers had observed as well one to three substitutions between *S. Montevideo* isolates from different colonies or subcultures (four passages) (Allard et al., 2012). The few SNPs observed in our analysis might be due to biological variations introduced by the subculturing steps and/or to any of the subsequent steps of the WGS workflow (DNA extraction and sequencing). Orsi and coworkers had suggested the possibility of the introduction of SNPs during subculturing of *L. monocytogenes* isolates (Orsi et al., 2008). From the serotypes tested, we did not observe that a specific serotype was more prone to acquire additional SNPs.

The complete WGS workflow generated results with high repeatability and reproducibility which is in concordance with the study of Kozyreva and coworkers (Kozyreva et al., 2017).

We assessed the discriminatory power to ensure that the typing system assigns a different type to two unrelated strains. Unlike conventional typing methods, the ability of WGS to differentiate strains is less of a concern as whole genome data offer the highest possible resolution to elucidate phylogenetic relations (Ronholm et al., 2016). Therefore, the approach to evaluate the discriminatory power was slightly adapted from Struelens and coworkers in order to illustrate the ability to discriminate unrelated strains while revealing relatedness of close isolates (Struelens, 1996). In the current study, epidemiologically unrelated ATCC or NCTC strains were used. For each unrelated strain, a set of known related sequences obtained from independent cultures of the strain or acquired at different dates were included in the analysis to ensure that the methodology still allowed to identify related isolates. For *L. monocytogenes*, three different serotypes including the serotype commonly associated with food (1/2a) and the serotype commonly related to outbreaks (4b) were evaluated (Nelson et al., 2004). In case of *S. enterica*, three different serovars were evaluated including a previously described clonal serovar such as *S. Enteritidis* (Taylor et al., 2015). For both *L. monocytogenes* and *S. enterica*, the complete WGS workflow allowed to differentiate unrelated strains without losing the ability to identify closely related strains.

The epidemiological concordance evaluated the capability of the CFSAN SNP Pipeline to reproduce results from previously published outbreak investigations using other bioinformatics pipelines. In this exercise, only the bioinformatics analysis was considered as the sequences were obtained from the public database at NCBI or directly from the authors of a published outbreak investigation. We selected one *L. monocytogenes*
and one *S. enterica* outbreak investigation in which other bioinformatics pipelines were used. For both studies, agreement between the results obtained with the CFSAN SNP Pipeline and the author’s pipelines was observed.

The ideal read mapping analysis would require the usage of a closed genome as reference for SNP calling. The impact of using a high quality Illumina assembly or PacBio assembly genome as reference was evaluated. To the best of our knowledge, no data are reported in the literature to show equivalence. In the present study, we report no impact in terms of SNP calling with these two sequencing technologies, and therefore we conclude that both alternatives to obtain the reference sequence are equivalent in the setup evaluated.

In conclusion, the current study is the first study proposing a validation approach including all steps of the end-to-end WGS workflow. We validated this WGS workflow, for the application of source tracking of *L. monocytogenes* and *S. enterica*, using the stability, repeatability, and reproducibility evaluations. Additionally, the complete WGS workflow was shown to have a high discriminatory power while enabling the ability to show genetic relatedness. For other bacteria, a validation would be required where the same performance criteria, as proposed in the current study, can be used. We believe that this work contributes as a first step to harmonize methods and thus obtain reliable results.
AUTHOR CONTRIBUTIONS

JG and LB performed the lab experiment steps including the DNA extraction step. CF performed the sequencing of all DNA

REFERENCES

Allard, M. W., Luo, Y., Strain, E., Li, C., Keys, C. E., Son, I., et al. (2012). High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics 13:32. doi: 10.1186/1471-2164-13-32

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. doi: 10.1089/cmb.2012.0021

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimming tool for illumina sequence data. Bioinformatics 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Chen, Y., Luo, Y., Curry, P., Timme, R., Melka, D., Doyle, M., et al. (2017). Assessing the genome level diversity of Listeria monocytogenes from contaminated ice cream and environmental samples linked to a listeriosis outbreak in the United States. PLoS ONE 12:e0171389. doi: 10.1371/journal.pone.0171389

Dallman, T., Inns, T., Jombart, T., Ashton, P., Loman, N., Chatt, C., et al. (2016). Delineating community outbreaks of Salmonella enterica serovar Typhimurium by use of whole-genome sequencing: insights into genomic variability within an outbreak. J. Clin. Microbiol. 53, 1063–1071. doi: 10.1128/JCM.03235-14

Orsi, R. H., Borowsky, M. L., Lauer, P., Young, S. K., Nusbaum, C., Galagan, J. E., et al. (2008). Short-term genome evolution of Listeria monocytogenes in a non-controlled environment. BMC Genomics 9:359. doi: 10.1186/1471-2164-9-359

Petkau, A., Mabon, P., Sieffert, C., Knox, N. C., Cabral, J., Iskander, M., et al. (2017). SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. Microbial. Genomics 3:e000116. doi: 10.1099/mgen.0.000116

Ronholm, J., Petronella, N., and Tamber, S. (2016). Draft genome sequences of two Salmonella enterica strains isolated from spouted chia and flax seed powders. Genome Announc. 4:e00963-16. doi: 10.1128/genomeA.00963-16

Schmid, D., Allerberger, F., Huhulescu, S., Pietzka, A., Amar, C., Kleta, S., et al. (2014). Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011–2013. Clin. Microbiol. Infect. 20, 431–436. doi: 10.1111/1469-0691.12638

Struelens, M. J. (1996). Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems. Clin. Microbiol. Infect. 2, 2–11. doi: 10.1111/j.1469-0691.1996.tb00193.x

Taylor, A. J., Lappi, V., Wolfgang, W. J., Lapiere, P., Palumbo, M. J., Medus, C., et al. (2015). Characterization of foodborne outbreaks of Salmonella enterica serovar Enteritidis with whole-genome single nucleotide polymorphism-based analysis for surveillance and outbreak detection. J. Clin. Microbiol. 53, 3334–3340. doi: 10.1128/JCM.01280-15

Wilson, M. R., Brown, E., Keys, C., Strain, E., Luo, Y., Murvuniva, T., et al. (2016). Whole genome DNA sequence analysis of Salmonella subspecies enterica serotype Tennessee obtained from related peanut butter foodborne outbreaks. PLoS ONE 11:e0146929. doi: 10.1371/journal.pone.0146929

Wyuts, V., Denayer, S., Roosens, N. H., Mattheus, W., Bertrand, S., Marchal, K., et al. (2015). Whole genome sequence analysis of Salmonella Enteritidis PT14 outbreaks from a national reference laboratory’s viewpoint. PLoS Curr. 7. doi:10.1371/currents.outbreaks.aa5372d90826e6cb0136ff66bb7a62fc

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Portmann, Fournier, Gimonet, Ngom-Bratu, Barretto and Baert. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.