bettercallsal: better calling of *Salmonella* serotypes from enrichment cultures using shotgun metagenomic profiling and its application in an outbreak setting

Kranti Konganti[1], Elizabeth Reed[1], Mark Mammel[1], Tunc Kayikcioglu[1], Rachel Binet[1], Karen Jarvis[1], Christina M. Ferreira[1], Rebecca L. Bell[1], Jie Zheng[1], Amanda M. Windsor[1], Andrea Ottesen[2], Christopher Grim[1], and Padmini Ramachandran[1]

1. Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, MD.
2. Center for Veterinary Medicine, U.S. Food and Drug Administration, Laurel, MD.

Introduction

- Recent foodborne outbreaks that have been attributed to multiple *Salmonella* serotypes force us to question whether these are rare events or if previous methods simply did not have the throughput to provide an accurate picture of the complex ecology that is connected to outbreak etiologies.
- Most of the metagenomic profiling tools using either marker- or *k-mer* based approaches for classification are sensitive down to the species rank and cannot accurately discern between highly clonal *Salmonella* spp. serotypes.
- Here we introduce bioinformatics innovations primarily based on DNA sketching algorithms for a metagenomic outbreak response workflow through the software tool called bettercallsal which is one of the first analysis tools that can accurately identify multiple *Salmonella* serovars at the same time in a much higher throughput approach.
- We leverage the NCBI Pathogen Detection (PD) project and provide hyperlinks to isolate genome(s) hits via the NCBI Isolates Browser, which in turn allows visualization within the NCBI SNP Tree Viewer if that genome hit is a member of a clonally related cluster (Sayers et al., 2021).
- The workflow is publicly available for download and use at https://github.com/CFSAN-Biostatistics/bettercallsal.

Materials and Methods

The main workflow uses a custom database generated via the automated Nextflow workflow called bettercallsal_db (Figure 1).

- First, the metadata for *Salmonella* spp. is downloaded from the NCBI Pathogen Detection project.
- In the next step, all the GenBank (GCA_) and RefSeq (GCF_) accessions are used to create an accession catalog to query NCBI and to retrieve assembly statistics, such as contig N50 and scaffold N50.
- For "*per_snp_cluster*" database, a single longest genome by N50 size is retained and for the "*per_computed_serotype*" database, up to 10 longest genomes by N50 size are retained for each of the "*computed_serotype*".
- Finally, for both database types, the contigs are joined by 10 N's and a MASH sketch is created. Certain pre-formatted flat files are also created which are used during the main analysis workflow.

The main analysis workflow is a single label metagenomic classification, wherein each genome assembly/accession match is mapped to the corresponding pre-indexed metadata (Figure 1).

- bettercallsal is also automated using Nextflow and starts with a "screen" command from MASH to generate an initial hit list followed by further genome fraction filtering using sourmash.
- Finally, kma and salmon tools are used to generate alignments and the final serotype calls with relative abundance levels of each of the possible serotypes within each sample.
- A brief stand-alone MultiQC HTML report generated at the end of workflow with the call results can be shared (Figure 2).

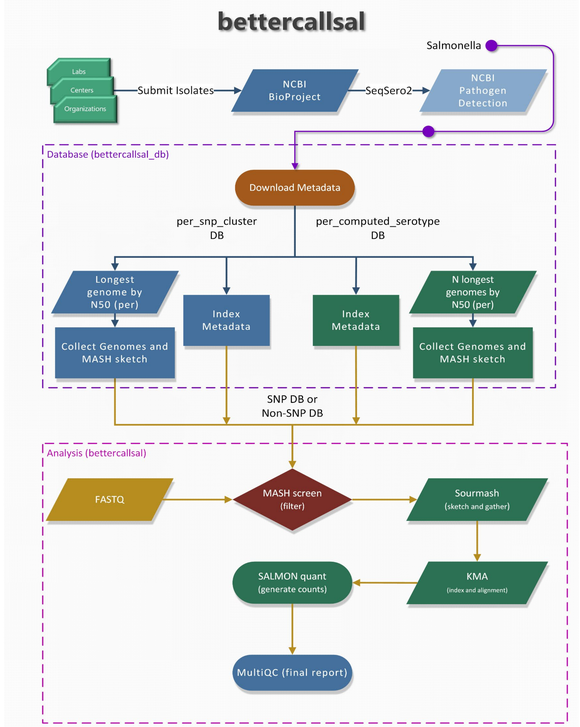


Figure 1. Overview of the bettercallsal and bettercallsal_db workflows.

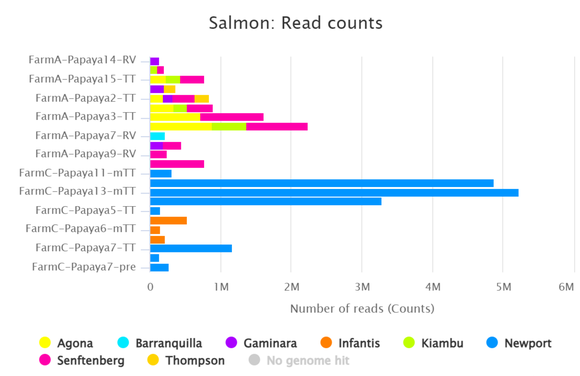


Figure 2. Salmon read counts plot exported from MultiQC HTML report from Papaya outbreak. The stand-alone MultiQC HTML report contains multiple, relevant sequence quality metrics and visualizations, ANI matrix between samples and genomes, an aggregated results table of serotype calls with integrated hyperlinks to NCBI PD Isolates Browser and Salmon read counts plot showing proportions of identified serotype(s) within each sample. Most of the visualizations are interactive and the result tables can be downloaded.

Results and Discussion

- An in-silico benchmark dataset, comprising 29 unique *Salmonella*, 46 non-*Salmonella* bacterial and 10 viral genomes was generated using InSilicoSeq with read depths from 0.5 million to 10 million read pairs using both MiSeq and NextSeq 500 error models.
- The in-silico dataset analyzed with bettercallsal revealed that precision, recall and accuracy increased as read depth increased for single-end and concatenated reads (R1+R2), to 100%, 90% and 95% respectively (Figure 3).
- The performance of the workflow was similar on multiple Illumina sequencing platforms but required more depth as the read lengths decreased based on the sequencing chemistries (Figure 3).

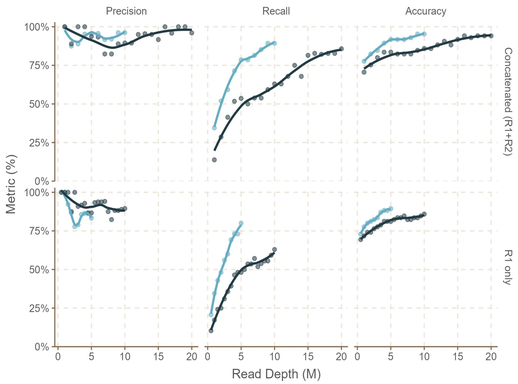
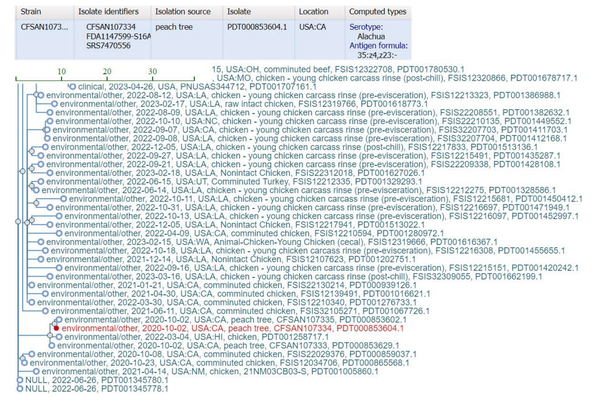


Figure 3. Beyond 9 to 10 million read depth (R1+R2), there are diminishing returns for MiSeq (2x300 bp) reads whereas similar performance or better is achieved between 16 to 20 million read depth (R1+R2) for NextSeq (2x150 bp) reads. The maximum precision, recall and accuracy achieved were 96.1%, 89.2% and 95.2% for 9M (R1 + R2) and 10M (R1 + R2) MiSeq reads compared to 100%, 82.7% and 94.1% for 16M (R1+R2) NextSeq reads.

- bettercallsal was run on previously sequenced quasi-metagenomics data sets (enrichment cultures) from Papaya and Peach outbreaks which resulted in identification of multiple serotypes from a single sample and traceback to the actual isolate(s) (Figure 2, Figure 4 and Figure 5).

Figure 4. External link from bettercallsal results table of the HTML report file to the NCBI pathogen detection website, which shows SNP cluster information and computed serotype information for the papaya outbreak. The closest genome hits (red) reported by bettercallsal are the isolate genomes from the outbreak investigation for both S. Newport and S. Senftenberg per NCBI's Isolate SNP Tree viewer.

Figure 4. SNP cluster information and computed serotype information for the peach outbreak via external link from bettercallsal results table of the HTML report file to the NCBI pathogen detection website. The closest genome hit (red) reported by bettercallsal clustered with the same isolate genome from chicken reported by the FDA investigation with the peach outbreak.



Strain	Isolate identifiers	Isolation source	Isolate	Location	Computed types
CFSAN1073...	CFSAN107334 FDA1147599-S16A SRS7470556	peach tree	PDT000853604.1	USA:CA	Serotype: Alachua Antigen formula: 35:z4,z23:-

Table 1. *Salmonella enterica* serovars detected in papaya outbreak samples from Farm A by bettercallsal, xMAP, custom *k*-mer, SeqSeroz, and Kallisto analyses. Luminex xMAP calls represent culture ground truth but bettercallsal identified additional serotypes like Agona, Gaminara and Kiambu which were detected in samples 2, 7 and 15 when compared to xMAP.

Papaya Farm A sample #	bettercallsal	k-mer	SeqSeroz	Kallisto	xMAP
1	No Call	Senftenberg	No Call	No call	Agona
2	Agona, Gaminara, Thompson, Senftenberg	Agona, Senftenberg	No Call	Agona, Thompson	Senftenberg, Thompson
3	Agona, Kiambu, Senftenberg	Agona, Senftenberg	No Call	Agona	Agona, Senftenberg
7	Barranquilla*, Gaminara, Senftenberg	Senftenberg	No Call	Gaminara	Senftenberg
9	Senftenberg	Senftenberg	Senftenberg	Senftenberg	Senftenberg
14	Gaminara	No Call	No Call	Gaminara	Kiambu
15	Agona, Kiambu, Senftenberg	Agona, Senftenberg	Kiambu	Agona	Senftenberg

Table 2. *Salmonella enterica* serovars detected in papaya outbreak samples from Farm C by bettercallsal, xMAP, *k*-mer, Kallisto and SeqSeroz analyses. *k*-mer analyses and Kallisto analyses also identified several other false positive *Salmonella* serotypes. Luminex xMAP calls represent culture ground truth.

Papaya Farm C Sample #	bettercallsal	k-mer	Seqseroz	Kallisto	xMAP
2	No Call	No Call	No Call	No call	No Call
5	Newport	Newport	Infantis	Newport	Infantis, Newport
6	Infantis	Infantis	Infantis	Infantis	Infantis
7	Newport	Newport	Newport	Newport	Salmonella group
10	No Call	No Call	No Call	No call	No Call
11	Newport	Newport	No Call	Newport	Salmonella group
12	No Call	No Call	No Call	No call	No Call
13	Newport	Newport	No Call	Newport	Newport

Conclusions

- We demonstrated that shotgun metagenomic sequencing of pre-enrichment and selective enrichments (quasi-metagenomic) along with a precision analysis tool such as bettercallsal facilitated the identification of multiple *Salmonella* serotypes and may provide equivalent trace-back utility as isolate WGS.
- To our knowledge, bettercallsal is one of the first analysis tools with the potential to identify multiple *Salmonella* spp. serotypes from a metagenomic or quasi-metagenomic data set with high accuracy and can provide rapid insights into the distribution, transmission, and source tracking of a foodborne pathogen.
- Use of Nextflow as workflow language enables reproducibility of the results along with platform agnostic process execution with an easy-to-share brief run report.

References

Konganti, Kranti, Elizabeth Reed, Mark Mammel, Tunc Kayikcioglu, Rachel Binet, Karen Jarvis, Christina Ferreira et al. "***bettercallsal: better calling of Salmonella serotypes from enrichment cultures using shotgun metagenomic profiling and its application in an outbreak setting.***" Frontiers in Microbiology https://www.frontiersin.org/articles/10.3389/fmicb.2023.1200983