Nestlé's state-of-the-art technologies to keep Food Safe and improve global Public Health

C. Ngom-Bru[1], L. Baert[1], N. Wagner[1], D. Moine[1], C. Fournier[1], M. Kassam[1], J. Gimonet[1], C. Barretto [1]

[1] *Nestlé Research Center, Lausanne, Switzerland*
E-mail : *catherine.ngombru@rdls.nestle.com*

Abstract
Whole-Genome Sequencing (WGS) for the purpose of food safety, more specifically for bacterial pathogen source tracking analysis is more and more used worldwide. WGS analysis consists of four main stages, (i) DNA extraction from pure cultures of bacteria contained in samples, (ii) whole genome sequencing of the complete chromosome and plasmids, (iii) bioinformatics analyses consisting in identifying Single Nucleotide Polymorphisms (SNP) and (iv) interpretation of the results in the context of each case. The bioinformatics analysis is a complex but nevertheless crucial step in the workflow. By comparing samples with clinical and environmental public data, the scientific IT infrastructure is impacted. Not only consequent data storage and sufficient network band to speed-up the analysis are needed but also server memory should be appropriate. The amount of data generated by the Next Generation Sequencing technology (NGS) falls into the Big Data category.
The Center for Food Safety and Applied Nutrition (CFSAN) of the US Food Drug Administration (FDA) is routinely using this type of analysis for Salmonella and Listeria foodborne pathogens isolated from raw material, final products and food manufacturing plants.
As the world's leading Nutrition, Health, and Wellness Company, Nestlé places great importance upon the application of scientific research to the interests of its consumers, especially regarding health and food safety. One of the roles of a large food manufacturer is to act in a protective fashion by actively screening in order to prevent biological and chemical contaminants from entering the food stream. The ability to consistently deliver safe and high quality products is essential and Nestle makes regular samplings for control of non-conformities in order to ensure safety of the food processing chain. Root-causes of non-conformities are systematically investigated, analyzed and corrective actions are applied.
Nestlé has adopted state-of-the-art technologies from which the DNA sequence of a detected pathogen undergoes an in-depth investigation in order to help identify the source of the contaminating organism.
Nestlé and FDA are collaborating to improve pathogen tracking methodology using WGS. This collaboration will not only promote the acceptance and the use of this approach by other stakeholders (industries, regulators, etc.) but also will improve public health by enhancing early detection and prevention of food-borne outbreaks and will contribute to the worldwide pathogen research.
The aim of this poster is to illustrate why the usage of WGS for pathogen tracking is generating Big Data and how it is managed.

Keywords: Big data, whole genome sequencing, Single Nucleotide Polymorphism, pathogen monitoring, data repository, IT infrastructure

1.   Introduction

As the world's leading Nutrition, Health, and Wellness Company, Nestlé places great importance upon the application of scientific research to the interests of its consumers, especially regarding health and food safety. One of the roles of a large food manufacturer is to act in a protective fashion by actively screening in order to prevent biological and chemical contaminants from entering the food stream. The ability to consistently deliver safe and high quality products is essential and Nestle makes regular samplings for control of non-conformities in order to ensure safety of the food processing chain. Root-causes of non-conformities are systematically investigated, analyzed and corrective actions are applied.

Nestlé has adopted state-of-the-art technologies from which the DNA sequence of a detected pathogen undergoes an in-depth investigation in order to help identify the source of the contaminating organism.

Whole-Genome Sequencing (WGS) for the purpose of food safety, more specifically for bacterial pathogen source tracking analysis is more and more used worldwide. The Center for Food Safety and Applied Nutrition (CFSAN) of the US Food Drug Administration (FDA) is routinely using this type of approach for Salmonella and Listeria foodborne pathogens isolated from raw material, final products and food manufacturing plants.

Nestlé and FDA are collaborating to improve pathogen tracking methodology using WGS. This collaboration will not only promote the acceptance and the use of this approach by other stakeholders (industries, regulators, etc.) but also will improve public health by enhancing early detection and prevention of food-borne outbreaks and will contribute to the worldwide pathogen research.

2.   Whole Genome Sequencing analysis generates big data

In a factory, pathogens such as *Salmonella* or *Listeria monocytogenes* can be introduced in different ways into the production areas: through the raw materials, on dust coming from the outside, through water infiltration, on pallets or other material coming from a lower hygiene level zone, and through operators for example on footwear or if personal hygiene rules are not followed. The presence of water residues will allow growth of the pathogen. From these multiplication sites, contamination can spread to other areas via dust or droplets in air currents, by workers, on utensils, by forklifts or water. Contaminated residues will find their way into interfaces, hollow bodies and holes on structures and equipment, leading to the permanent establishment of the pathogen as a house strain.

If the pathogen reaches the first priority rating environment, close to the line, the risk of line contamination becomes very high. When the line is contaminated, the risk of finished product contamination becomes almost inevitable. Therefore routine and investigative sampling of the process area is crucial to verify that pathogens are kept away from the line and will not contaminate products.

At Nestlé, in order to prevent microbiological contaminants from entering the food stream, systematic screening is done by swabbing many locations within the factories. *Salmonella* and *Listeria monocytogenes* are the two major pathogens that are systematically screened. In case of positive found, root-causes are investigated, analyzed and corrective actions are applied. For the purpose of bacterial pathogen source tracking, many samples are collected, purified, cultured and stored in a repository with the associated metadata. Among those samples, several are sent for

whole genome sequencing either with PacBio technology to obtain close to completion genomes, or with Illumina generating large amount of high quality short sequences.

WGS analysis consists of four main stages (Figure 1)
1. DNA extraction from pure cultures of bacteria contained in samples,
2. Whole genome sequencing of the complete chromosome and plasmids,
3. Bioinformatics analyses consisting in identifying Single Nucleotide Polymorphisms (SNP).
4. Interpretation of the results in the context of each case.

The Center for Food Safety and Applied Nutrition (CFSAN) of the US Food Drug Administration (FDA) is routinely using this type of approach for *Salmonella* and *Listeria* foodborne pathogens isolated from raw material, final products and food manufacturing plants (Davis S. et. al).
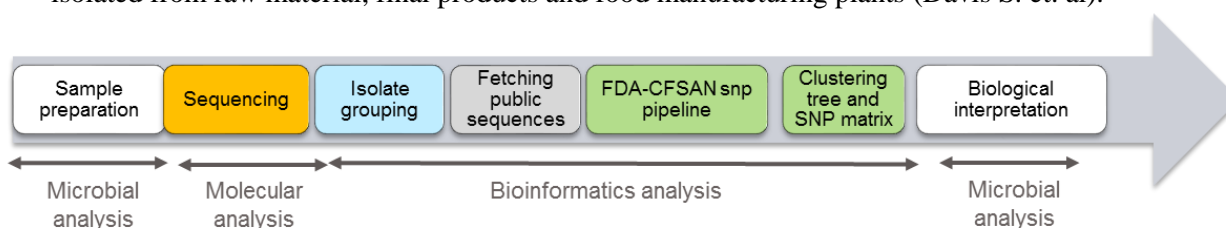

Figure 1: General workflow for pathogen characterization using whole genome mapping for Single Nucleotide Polymorphism detection.

The basic principle of Single Nucleotide Polymorphism (SNP) detection is to compare one or several genomes to a common reference genome, and identify each nucleotide position along the reference sequence where one or several of the compared genomes differ. The SNP detection accuracy is highly dependent on the similarity of the reference compared to the samples, and the sequencing quality of both the reference and the samples to be compared. The reference genome should preferably be a finished genome but can also be in an unfinished version with several contigs. The samples to be compared are sequenced with a technology generating large amount of high quality short sequences. At Nestlé, depending on the number of samples to be sequenced at once, the technology used may be Illumina HiSeq (> 80 samples) or Illumina MiSeq (< 20 samples). In general these sequencing technologies generate around 3 GB data per sample and currently one analysis comprises a minimum of 3 Nestlé samples and up to hundreds. In the current workflow of analysis, Nestlé needs to download raw sequencing data from public databases to be added to the analysis. Today raw sequencing data is publicly available for >76'000 *Salmonella* and >7'000 *Listeria monocytogenes*. The current analysis pipeline setup downloads the 30 closest raw sequencing genomes per analysis, hence 30 x 3 GB per analysis are included in addition to Nestlé samples. These raw sequencing data are further processed, generating several intermediate files of significant volume size.

As a consequence, these large amount of data require significant data storage capacity and sufficient network band to speed-up the data transfer. Calculation power and server memory must be scaled to the amount of cases to be analyzed.

3. WGS for pathogen monitoring: from prototype to commonly used technology

Nestlé and CFSAN FDA are currently working together to develop further the bioinformatics SNP pipeline and communicate on the advantages of using WGS for pathogen monitoring. From an industrial perspective, this approach is of great interest because it is more accurate than classical methods like PFGE, riboprinting or MLST. However, since this technology is still in its infancy, significant investments in terms of sequencing and bioinformatics competencies, equipment for sequencing, and all IT infrastructure are crucial.

From a regulatory perspective, it is key for FDA to embark Nestlé as the world leader food industry to advocate for this newly developed technology among other food companies.

From a Nestlé R&D perspective, storing data and accessing historical data will be key in the future. Not only the pipeline needs to answer appropriately the biological questions, but the complete IT infrastructure needs to be developed to be able to achieve recurrent and automatic analyses. Convincing other regulatory bodies and food companies on the relevance of WGS for pathogen monitoring will make this technology evolve from exclusive to popular, attracting software provider companies to develop user friendly wrappers of the current analysis pipelines. This will in turn facilitate transfer of the technology in a routine mode to operations.

References

Davis S., Pettengill J. B., Luo Y., Payne J., Shpuntoff A., Rand H., Strain E. (2015). CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ CompSci.