Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Food Science

Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens

Eelco Franz[1], Lapo Mughini Gras[1] and Tim Dallman[2]

As the whole genome sequencing (WGS) revolution is rapidly gaining momentum, it is essential to understand the significance of this technology and its future applications in food safety. This review discusses the recent advances concerning the use of WGS for outbreak detection and surveillance, microbial source attribution and microbial risk assessment. Although the WGS is mainly being applied for surveillance and outbreak investigation purposes, there is still, a strong need for harmonization of methods (sample preparation, sequence quality and case-definition, analysis) and consensus on suitable nomenclature (SNP versus allele level)and case definition. The application of WGS in source attribution and microbial risk assessment is largely unexplored. The use of WGS in source attribution requires the development of new modelling approaches that can handle the large amount of data and the high discriminatory power associated with WGS. For microbial risk assessment, the link with phenotypic features is crucial, but the short-comings regarding the reproducibility of genome-wide-association studies and the link to epidemiology need innovative statistical approaches. Overall, WGS data alone are of limited use without a sound public health or biological context. Defining hypotheses and research questions beforehand are crucial for the correct analysis and interpretation of WGS data.

Addresses
[1] National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
[2] Public Health England, London, United Kingdom

Corresponding author: Franz, Eelco (eelco.franz@rivm.nl)

Current Opinion in Food Science 2016, **8**:74–79

This review comes from a themed issue on **Food microbiology**

Edited by **Luca Cocolin** and **Kalliopi Rantsiou**

For a complete overview see the Issue and the Editorial

Available online 22nd April 2016

http://dx.doi.org/10.1016/j.cofs.2016.04.004

Introduction

Whole-genome sequencing (WGS) reveals the complete DNA make-up of an organism. The most significant advantage of WGS for pathogenic microorganisms is that their typing can be conducted at a much greater resolution than with traditional molecular typing methods. These include banding pattern-based methods (e.g., pulse field gel electrophoresis — PFGE, ribotyping, etc.) as well as locus-based sequencing methods (e.g., multi-locus sequence typing — MLST, multiple-locus variable number tandem repeat analysis — MLVA, etc.). In addition, WGS allows for the reconstruction of evolutionary relationships (including transmission and host distribution reconstructions) between microorganisms at a level which was not previously possible to achieve using phenotypic methods.

Because of the increasing speed and decreasing operational and acquisition cost of high-throughput sequencing, comparative genomic analysis of foodborne pathogens is increasingly integrated into surveillance, control, and research activities. In addition, it provides promising public health benefits by increasing the understanding of pathogen ecology and epidemiology. This review aimed at providing an overview of different applications of WGS relevant for food safety.

WGS in outbreak investigation and surveillance

WGS has emerged as a powerful tool for outbreak investigations. Current gold-standard subtyping methods including PFGE do not often provide the resolution needed to discriminate between outbreak-related and sporadic cases. This is especially relevant for monomorphic microorganisms like *Salmonella* serovars Enteritidis [1] and Heidelberg [2]. In some cases, for example with *Listeria monocytogenes*, WGS can address the issue of over-discrimination by PFGE caused by the gain or loss of mobile genetic elements [3,4••]. For outbreak investigation, having a highly specific and sensitive microbiological case definition allows for a more robust epidemiological analysis, increasing the chances of detecting the causative source. An increasing number of studies are being published showing various applications of WGS in *retrospective* foodborne disease outbreak investigations [5–11]. However, few studies report on the successful use of WGS in *prospective* surveillance [4••,9,12].

Routine surveillance with WGS provides the opportunity to assess the total genetic diversity of a specific pathogen at a high genetic resolution. The impact of the use of WGS in surveillance can essentially be appreciated in the detection of a higher number of (small and/or diffuse) outbreaks, some of which would probably pass unnoticed with the sole use of the 'traditional' subtyping methods.

The major advantage of using WGS for surveillance is therefore inherent in the higher resolution of the WGS output itself, which allows for improvements in the ability to detect temporal and spatial clusters of genetically related pathogens.

High throughput machines that produce short read (<300 bp) sequences have predominantly been adopted for WGS surveillance of foodborne pathogens, currently fully implemented in routine typing at several institutes and authorities (Food and Drug Administration Genome TrakR network, Public Health England, and Staten Serum Institute). The large variety by which WGS data can be analyzed represents a significant hurdle in standardization and harmonization and in providing epidemiologists and decision makers with interpretable information for action. These approaches differ in using the nucleotide or the allele as unit of interest and whether the method is reference-based or reference-free (Table 1). One of the quickest and simplest methods is the k-mer approach by which phylogenetic trees can be constructed from the frequency profile of k-mers across the selected genomes [13]. Perhaps the most common analytical approach is to align the sequenced reads to a common reference genome as to identify single nucleotide polymorphisms (SNPs) [14]. The reference genome will ideally be as related as possible to the sequenced organism(s). The 'core' genome, i.e. those DNA sequences shared between the sequenced isolates and the reference genome, can then be analyzed based on differing SNPs. A slightly different approach is the nucleotide difference approach in which the number of nucleotide differences between a pair of raw read mapped reference genomes is identified rather than identify such difference as SNPs [15]. A fundamentally different method is to base the analysis on protein-encoding alleles. Core-genome and whole-genome Multi Locus Sequence Typing (cgMLST/wgMLST) usually take an assembled genome to their input and the alleles are subsequently identified based on nucleotide identity to a defined scheme [10,16••,17,18]. Downstream analysis is performed either on the sequence of the shared alleles or using a distance

Table 1

Overview of different approaches to WGS data analysis.

Approach	Alignment	Reference	Example	Description	Pro's	Con's
k-mer based	No	No	k-mer tree [13]	k-mer based grouping of closest genome matches by comparison across very short sequences. A tree can be constructed from the frequency profile of k-mers across the selected genomes	Fast; handles diverse genomes well since it is independent from reference genome	Loss of resolution due to condensation of sequence data into a vector of k-mer counts, and neglecting the order of k-mers
SNP-based	Yes	Yes	SNPtree [14]	Sequence reads are aligned to a reference genome in order to identify SNPs	From raw reads (no assembly bias but slow) or contigs (fast)	Strongly depends on a closely related (finished) reference genome; sequences not present in the reference will be ignored; limited value when comparing diverse genomes
SNP-based	No	No	kSNP [19]	Detects SNPs on the basis of k-mers of odd-number length that differ at the central base but are identical at all bases flanking that central base	From raw reads, contigs or finished genomes; fast, high-throughput	Accuracy is sensitive to the level of sequence diversity and level of recombination
Nucleotide difference	Yes	Yes	NDtree [15]	Identifies the number of nucleotide difference between a pair of raw read mapped reference genomes rather than identify the difference as a SNP	Does not require concatenated sequence for alignment; also SNPs not present in reference genome are identified	Somewhat sensitive to settings; from raw reads only (slow)
Allele-based	Yes	Yes	Genome-wide MLST	Sequence variation of predefined loci are indexed	Not requiring closely related reference genome; allows diverse genomes; included; nomenclature similar to traditional MLST; curated; scalable	Requires assemblies (bias, difficulties with repeats); only detects variation in predefined loci