Chapter 10
Whole Genome Sequencing in Food Outbreak Investigation and Microbial Risk Analysis

Henk Aarts and Eelco Franz

Abstract Next Generation Sequencing (NGS) is a huge technological advance in the molecular typing of micro-organisms. This chapter describes the use of NGS in food-borne outbreak investigation, source attribution and also describes the first steps in using Whole Genome Sequencing (WGS) data in molecular risk assessment. The rapidly decreasing costs and operational time make that WGS will likely replace currently used molecular typing methods, such as MLST, MLVA and PFGE, in outbreak investigation and source attribution in the near future. Because of the superior level of resolution and because all genetic information (including virulence and antimicrobial resistance genes) of the organisms can be revealed, WGS can be considered as the ultimate "all-in-one" typing method.

Keywords Food-borne pathogens · Whole genome sequencing · Outbreaks · Typing · Microbial risk analysis

10.1 Introduction

Within the field of genomics, Whole Genome Sequencing (WGS) is becoming "the" method of choice for addressing foodborne outbreaks and for assessing the potential risks of pathogens. Next Generation Sequencing (NGS) platforms produce a large volume of data in a relatively short time and have turned out to be a powerful tool for WGS of viral and bacterial genomes. This was demonstrated

H. Aarts (✉)
Centre for Zoonoses and Environmental Microbiology (Z&O),
National Institute for Public Health and the Environment (RIVM),
Bilthoven, The Netherlands
e-mail: henk.aarts@rivm.nl

E. Franz
Centre for Infectious Diseases, Epidemiology and Surveillance,
National Institute for Public Health en the Environment (RIVM),
Bilthoven, The Netherlands

E. van Pelt-Verkuil et al. (eds.), *Molecular Diagnostics*,
DOI 10.1007/978-981-10-4511-0_10

during the recent entero-haemorrhagic *Escherichia coli* (EHEC) outbreak in Europe. Sequencing of the outbreak strain of *E. coli* O104 during the outbreak in Germany in 2011 took only 3 days, with the first assembly being released two days after completion of sequencing. Within one week, the data became available, showing a novel entero-aggregative *E. coli* O104 variant, which had acquired a prophage encoding Shiga toxin 2. The concept behind NGS technology is not different from Sanger sequencing performed in combination with capillary electrophoresis. However, the huge difference is that the number of reactions, which are run in parallel (1×10^6 times), is hugely increased in NGS, enabling the rapid sequencing of an entire genome of a microorganism in a single run at (relatively) low cost.

10.2 WGS in Relation to Food Safety

Because of its superior discriminatory power, WGS has huge potential for the detection and investigation of outbreaks, source attribution (i.e. the relative contribution of different reservoirs to the disease burden) and risk assessment. WGS will likely replace serological and molecular typing methods, such as serotyping, multilocus sequence typing (MLST) and probably also MLVA (multilocus variable number of tandem repeats analysis) and Pulse Field Gel Electrophoresis (PFGE). Furthermore, the characterization of the genetic content of pathogens will enable the identification of health related properties like virulence and antibiotic resistance. WGS will also provide the necessary genomic information for the development of diagnostic tests based on significant markers and for the detection of new variants. However, several challenges have still to be solved in order to make full use of the potential of WGS for food safety (EFSA 2013). These include storage and sharing of the data, harmonization of bioinformatics pipelines and standardized nomenclature and interpretation.

10.2.1 Molecular Typing Methods

Methods used for the characterization of bacterial or viral strains below the (sub-) species level, are by definition named “typing methods”. Important performance criteria for these methods are:

(i) Typeability: the proportion of strains that are assigned to a type by the typing method
(ii) Discriminatory power: the likelihood that strains can be distinguished by the typing method
(iii) Repeatability: the variation in the correct assignment performed under the same conditions (person, system, time etc.)

(iv) Reproducibility: the variation in the correct assignment performed under different conditions (person, system, time etc.)

(v) Stability: the ability of a typing system to recognize the clonal relatedness of strains derived from a common ancestor strain.

Typing methods can be divided in methods based on phenotypic properties (for example serotyping, phage typing, biotyping, antimicrobial resistance and protein profiling) and in methods based on the genetic content, such as DNA fingerprinting (PFGE, RAPD, AFLP, ERIC, MLVA) and DNA sequencing (MLST, full genome) techniques. All these methods have their own advantages and disadvantages. PFGE, for instance, is expensive and time consuming (4–5 days). On the other hand, PFGE is a method that can be used for all bacteria, that has a high discriminative power and that shows a high level of typeability. Table 10.1 shows a comparison (discriminative power, number of markers, reproducibility etc.) of commonly used bacterial typing techniques, which also includes the cost per isolate analysis. Sabat et al. (2013) have created a comprehensive overview, including the pros and cons, of the existing molecular typing methods that are used for outbreak investigations and epidemiological surveillance. The most advanced molecular typing method is WGS. For this method, the genome has to be fragmented and subsequently sequenced. The latter is done in specialized sequencers like the 454 Life Sciences (GS FLX Titanium) from Roche, the Ion Torrent & ABI Solid from Life Technologies, the HiSeq, MiSeq and GenomeAnalyzer from Illumina or the PacBio RS from Pacific Biosciences. In using WGS/NGS, huge amounts of data are generated and analysis of these data requires massive bioinformatics support and storage capacity (see also Volume 1, Chaps. 8 and 9).

An alternative upcoming method for microbial typing is Matrix-Assisted Laser Desorption/Ionization-Time Of Flight (MALDI-TOF). This mass spectrometry method has great advantages in that it is fast, easy to use, high through-put and low cost. However, a limitation of MALDI-TOF is that it does not reliably identify isolates at the subspecies and clonal level, it does not give sufficient information on different virulence and resistance genes and it is difficult to standardize in a way that enables the global exchange of data.

10.2.2 Genomic Subtyping of STEC O157

Shiga-toxin producing *E. coli* (STEC) is a zoonotic pathogen that causes diarrheal disease in humans and is of public health concern because of its ability to cause outbreaks and severe disease, such as haemorrhagic colitis (HC) or haemolytic-uremic syndrome (HUS). Numerous cases of HC and HUS have been attributed to EHEC serotype O157:H7 strains. It can be considered paradoxical that the number of human STEC O157 infections is rather low for a pathogen with a low infectious dose and a relatively high prevalence in the ruminant reservoir. This could, at least partly, be explained if only a subset of STEC O157 isolates present in

Table 10.1 Comparison of most common bacterial typing techniques (adapted from Foxman et al. 2005) by Pérez-Losada et al. (2013)

Typing method	Method description	No of markers	Temporal scale	Variation source	Discriminatory power	Reproducibility	Equipment/Time	Equipment/consumables-Reaction costs (per isolate)	Available databases
MLST	PCR amplification of housekeeping genes to create an allelic profile	7	Macroepidemiological Microepidemiological	DNA sequence	Moderate to high	High	Thermal cycler/Moderate	\$30K–45K High–\$80	Pubmist.org www.mist.net mlst.ucc.ie www.pasteur.fr/mlst
MLEE	Phenotypic characterization of the elcotrophoretic mobility of housekeeping enzymes	10–20	Macroepidemiological Microepidemiological	Electrophoretic mobility	Moderate	Moderate	Gel box, switching unit cooler, power supply/Moderate	\$10K–20K Moderate–\$20	NA
PFGE	Comparison of large genomic DNA fragments after digestion with rare restriction enzyme	NA	Microepidemiological	Banding pattern	Moderate to high	High	Gel box, switching unit cooler, power supply/High	\$10K–20K Moderate–\$22	NA
AFLP	Digestion of genomic DNA with two restriction enzymes, ligation of restriction fragments and selective amplification	NA	Microepidemiological	Banding pattern	Moderate to high	Low	Thermal cycler/Moderate	\$8K–12K Moderate–\$20	NA
MLVA	PCR amplification of VNTR loci followed by sizing of the PCR products to create an allelic profile	10–80	Microepidemiological	DNA sequence	Moderate to high	High	Thermal cycler/Low	\$30–45K Moderate–\$20	Minisatellites.u-psusd.fr www.mlva.net www.pasteur.fr/mlst
HRM	PCR amplification followed by characterization of amplicon melting curves	NA	Macroepidemiological Microepidemiological	Melting temperature	High	High	Real time thermal cycler/Very low[a]	\$30K–45K Very low[a]	NA
RFLP	Digestion of genomic DNA with restriction enzymes to produce	NA	Microepidemiological	Banding pattern	Low	Low	Southern transfer/High	\$8K–12K Low–\$14	NA

(continued)

Table 10.1 (continued)

Typing method	Method description	No of markers	Temporal scale	Variation source	Discriminatory power	Reproducibility	Equipment/Time	Equipment/consumables-Reaction costs (per isolate)	Available databases
	multiple short restriction fragments								
rMLST	PCR amplification of *rps* genes to creat an allelic profile	53	Macroepidemiological Microepidemiological	DNA sequence	High	High	Thermal cycler/Moderate	$30K–45K High–$600 (if WGS is not needed)	http://pubmlst.org/software/database/bigsdb/
Pan-genome	Detection of similarities/differences in the pangenomic or distributed genes	>1000	Macroepidemiological Microepidemiological	Presence/absence of genes	High	High	NGS platforms or Microarrays/Moderate to high	$80K–130K Very high–$1K–20K per run depending on the NGS platform used	www.francisella.org

[a]If new melting profiles are not detected

the bovine reservoir are characterized by unique transmissibility and/or virulence characteristics. Distinguishing STEC O157 isolates that constitute a high risk to human health from isolates that have a lower association with clinical symptoms is an important aspect of risk-based monitoring and surveillance.

Franz et al. (2012) genetically characterized and compared bovine, food, and human clinical STEC O157 isolates from the Netherlands using various genotyping methods and multivariate statistics. The goal was to determine whether different genotypes occur at different frequencies among isolates from these different sources and to identify the most differentiating genetic features of the isolates from these sources. A set of 73 bovine, 29 food, and 85 human clinical isolates was used. The results clearly showed non-random distributions of genotypes among bovine, food, and human clinical isolates. In addition, comparison with published data revealed significant differences in the distribution of genetic lineages among bovine and human isolates between the Netherlands versus North America and Japan. As a follow-up to get more insight in the genomic and phenotypic differences between STEC O157 isolates from different geographic regions, strains from The Netherlands (n = 39) (low *E. coli* O157 disease incidence) and Scotland (n = 145) (high *E. coli* O157 disease incidence) were subjected to whole genome sequencing. Significant differences in genetic distance were observed between isolates from different countries. It was therefore concluded that after successful transmission, which occurred several times, local genetic variation occurs, resulting in a geographical patchwork of phylogeographical clades.

10.2.3 Whole Genome Sequencing of Food Borne Pathogens

Sequencing the whole genome of a pathogen results in huge amounts of data, which consists of many individual reads. In recent years, we have seen a rapid increase in the raw daily output of the different platforms and a decline in turnaround times for bacterial genomes and in cost per Mb assembled sequence.

The most common way to analyse WGS data is the reference-based single-nucleotide-polymorphism (SNP) approach. Here, the reads are mapped to a complete and reliable sequenced reference genome, which should be as closely related as possible to the strains under investigation. Subsequently, SNPs are called in the reads that map to the reference genome. Drawbacks of this reference method are that a closely related reference genome is necessary to map as much reads as possible and that a sequences absent in the reference genome will not be mapped. The latter may introduce a bias when the strains under investigation are diverse. For these reasons, researchers often choose the approach of reference-free '*de novo*' assembly. Hereby, reads are concatenated into longer stretches of reads, so-called contigs. This is done by specific algorithms and based on overlaps between reads. In turn, overlapping contigs can be assembled into scaffolds. A final step may be the closure of the entire genome, but this is often difficult due to gaps in the genome and the presence of repetitive elements. The quality of this de novo assembly will increase with the use of sequence platforms that produce longer reads, since these leave less uncertainty in the assembly procedure. The assembled genomes can subsequently be used for SNP detection, phylogeny (based on all genetic information, rather than the SNPs only, as with the mapping approach) or presence/absence detection of specific genes.

An alternative to the reference-based or reference-free SNP approach is the gene-by-gene approach. In this case, allelic variation (which can be caused by SNPs, recombination, and/or insertion/deletion events) in pre-defined loci are identified. This approach is already applied in the classical MLST, where Sanger sequencing determines the sequence variation of a predefined set of household genes. The resulting allelic variations are translated into a numeric nomenclature, which is curated. With WGS, this classical MLST can be performed in silico, but it can also be extended to the all known loci in the core-genome (i.e. the part of the genome shared by all strains under investigation) or even to the pan-genome (i.e. unique genes). The development of a core-genome or genome-wide MLST scheme, however, does require a closed reference genome and preferably a set of query genomes (which can be in contig or scaffold stage) in order to define a more stable scheme.

The reads of a single bacterial genome, at a level of a 100-fold coverage will take 200 MB of disk space. Handling the information of a large number of strains requires a complex IT infrastructure, which generates computational challenges that must be addressed by specialized software. Many platforms incorporate superseded versions of algorithms. Various algorithms exist, for instance for '*de novo*' assembly and RNA-sequencing. Besides commercial packages, there are also freely available "open-resource" algorithms from academic institutions. The available algorithms make it possible to assemble genomes in many instances, meaning that the assembly can be performed using the existing computer resources in the laboratory. For instance, an *E. coli* genome can be assembled within a time frame of 15 min on a 32 bit Windows desktop computer with 32 GB of RAM.

Recently, several bioinformatics tools have been published, such as "RESfinder". This software package is developed by the Technical University of Denmark (DTU) and subtracts antibiotic resistance gene (or related sequences) information from whole genome sequences. They showed that there was a good concordance between the "in silico" data and data obtained from phenotypic analysis (personal communication). Besides RESfinder, a number of other software packages, like PathogenFinder, KmerFinder, VirulenceFinder, are available on their website (http://www.genomicepidemiology.org). Although good concordance between the detected genes and phenotypic properties has been found, this does not mean that genome sequences always predict the potentially conditional expression of particular genes, nor their level of expression. Consequently, it will not always be possible to match phenotypic tests.

It is required that databases are public and that the data from food, (food-producing) animals, environment and humans are closely linked, in such a way that public health can profit the most from the strength of WGS. In Europe, both EFSA and ECDC have taken the initiative to encourage the collection of molecular data to ensure a better linkage of molecular typing data from humans to similar data from food and animals. A popular database is the Bacterial Isolate Genome Sequence database (BIGSdb) developed by Jolley and Maiden (2010). BIGSdb is able to handle NGS data of microbial genomes and can perform extended MLST.

The currently available databases suffer from several drawbacks. They lack harmonized protocols for submission, curation and storage of the data and interpretation, annotation and standardized data sharing formats are still missing.

10.3 Whole Genome Sequencing in Outbreak Investigation

Typing of the isolated organisms can aid in the detection of outbreak cases, can lead to a more specific case definition and is required to definitively link a source to the outbreak. Serotyping, PFGE, MLVA or any combination of these methods are currently used (PulseNet, www.cdc.gov/pulsenet/about/index.html). Serotyping does not have the resolution required for consistent detection of outbreaks. PFGE data is of fairly high resolution and is harmonized and shared in PulseNet, but it suffers from a certain lack in repeatability. In addition, for very monomorphic bacteria, like *Salmonella enterica* serovar Enteritidis, PFGE often does not have the required discriminatory power for outbreak investigation. As a result of the increased speed and decreased costs of acquisition and the high resolution of WGS, comparative genomic analysis of bacterial pathogens is rapidly becoming part of many outbreak investigations. However, the challenge for WGS is the large amount of data to be transferred and stored, as well as development of internationally harmonized analysis methods. While the development of tools for analysis and interpretation of WGS data is rapidly progressing, there is presently much uncertainty and certainly no consensus on how to translate these data into practically useful information for public health purposes. Nomenclature is, in its essence, a technique to reduce the amount of available information by assigning a short, yet still epidemiological informative code to isolates. For WGS data, one or more agreed nomenclature schemes are also required, since efficient communication between organizations is a prerequisite due to the international nature of infectious diseases. The currently used approach for nomenclature is under discussion. A nomenclature based on the earlier mentioned extended MLST approach is a good candidate, since it builds on the nomenclature of the already applied classical MLST. A prerequisite, however, is that all involved stakeholders need to agree on a specific scheme. Unique identifiers are given to unique alleles, which are subsequently curated in a global database. As an example of the potential of wgMLST for outbreak investigation, Schmid and colleagues showed in 2014 (Schmid 2014) that WGS is capable of discriminating *Listeria monocytogenes* SV1/2b clones that are not distinguishable by PFGE and fluorescent amplified fragment length polymorphism (fAFLP).

They investigated a cluster of seven human cases of listeriosis that occurred in Austria and in Germany between April 2011 and July 2013. In January 2013, the Austrian Food Authority mandated the Austrian Agency for Health and Food Safety to investigate the source and health related issues of an in 2012 dominant PFGE clone. Active case finding in Germany and Austria resulted in seven cases, all elderly women. Strains from food producers were also included in this study. The human and food related strains of *L. monocytogenes* SV1/2b could not be distinguished by PFGE and fAFLP. Epidemiological investigation, however, suggested that the Austrian cases were linked to one of the two Austrian food producers. To further analyse this listeriosis cluster, the seven human isolates, a control strain with a different PFGE/fAFLP profile and ten food isolates were subjected to whole genome sequencing (WGS). The investigators used MLST+, a whole genome wide MLST scheme based on 2298 genes that where present in all strains. MLST+ analysis suggested that the outbreak cases were linked with either one of the two Austrian (cheese or meat) food producers. It is worthwhile to mention that the products of the Austrian food producers appeared on the grocery bills collected from the outbreak cases. In this paper it was clearly shown that WGS was capable of discriminating *L. monocytogenes* SV1/2b clones that were not distinguishable by PFGE and fAFLP, hereby clearly showing the potential of WGS in outbreak investigations.

10.4 Viral Food-Borne Pathogens

The detection of viral food-borne illness relies on a combination of laboratory diagnosis, epidemiological investigation, pathogen typing and food trace-back investigations. There is a great challenge in reliable detection of viruses in food, a practice that is an essential part of outbreak investigations. Virus detection, in contrary to the detection of bacteria, cannot rely on a culture step. Food borne viruses are difficult to extract form food matrixes and the limit of detection is high. Another approach is the application of metagenomics based on deep sequencing, either by targeted sequencing of specific viruses or in an unbiased manner to determine the viral population. The detection and characterization of (novel) viruses are of paramount importance in outbreak investigation and the forecasting of future outbreaks of viral diseases in humans. Metagenomics studies based on deep sequencing in the analysis of clinical samples, natural reservoirs and food will help to control or prevent major outbreaks.

Due to the rapid rate of evolution of viruses, virologists have been using genome-wide sequencing for a long time. The interplay of epidemiological and evolutionary patterns, based on WGS, is emerging as a new field in public health microbiology. The availability of software such as BEAST (Drummond et al. 2012) has facilitated this strongly.

One important food borne virus is norovirus. Norovirus is a highly contagious particle which is easily transmitted via the faeces and vomit of infected people, but also through direct contact or through contaminated objects or via food and water. A network (NoroNet) of scientists working in public health institutes and universities share virological, epidemiological and molecular data on this virus. In order to type Noroviruses, several new tools and databases have been developed during the last couple of years. The database and tools are hosted by the RIVM (National Institute of Health and the Environment, The Netherlands).

10.5 Whole Genome Sequencing in Microbial Risk Assessment

The sequence data obtained by WGS contains all the genetic information of the strain. These data can be used for what is called 'comparative genomics', which is a term used for when the genetic content of two strains is compared. By doing this, short and long term evolutionary relationships and pathogenicity of food-borne pathogens, such as virulence, severity of disease, host specificity, ecological niche and mechanisms to adapt to particular (environmental) stress situations, can be determined. In addition, strain-, lineage- and niche-specific regions can be found andevents that resulted in the loss or acquisition of DNA can be identified. Comparison between the genome sequences of commensal (harmless) bacteria and those of pathogenic bacteria will help to identify genes that are involved in host specificity and to identify mechanisms of host-microorganism interaction.

As such, WGS has huge potential in the characterization of the genetic content of pathogens, in risk estimation of virulence potential, in antibiotic resistance profiles and in other health related properties. This information, in combination with phenotypic testing of bacterial behaviour (growth, survival, stress resistance, metabolism and adhesion to epithelial cells), provides a powerful combination to understand and control pathogens in the food supply chain. The outbreak strain causing the major outbreak in Germany in 2011 was characterized as an enteroaggregative *E. coli* (EAEC) that acquired toxin producing *stx*-genes by horizontal gene transfer. This resulted in a highly unusual combination of virulence traits, resulting in severe illness and death. The outbreak caused a paradigm change with respect to human pathogenicity of STEC and revealed substantial knowledge gaps in the emergence of hybrid types of pathogenic *E. coli*. A major obstacle for efficient monitoring, justifiable public health actions and efficient clinical management is the current inability to discriminate STEC strains posing a serious risk to humans (i.e. EHEC) from STEC strains that are not associated with severe and/or epidemic disease. This is primarily due to the huge variety in the presence/absence of virulence genes. Different classes of pathogenic strains are divided in five different seropathotypes (A–E), which is shown in Table 10.2. This model is based on the presence/absence of Shiga-toxin encoding genes (stx_1 and stx_2) and the intimin gene (*eae*) in relation to incidence in humans (reported frequency in human disease), association with outbreaks and the association with severe disease. To find

Table 10.2 Virulence markers in the seropathotype concept as proposed by Karmali et al. (2003)

Seropathotype	Incidence in human disease[a]	Outbreaks	Association with severe disease[b]	Virulence markers		Serotypes
				Vtx	Eae	
A	High	Common	Yes	Vtx2 (but may in addition also carry vtx1)	+	O157:H7, O157:NM
B	Moderate	Uncommon	Yes	Vtx1 and/or vtx2	+	O26:H11, O103:H2, O111:NM, O121:H19, O145NM
C	Low	Rare	Yes	Vtx1 and/or vtx2	±	O91:H21, O104:H21, O113:H21, O5:NM, O121:NM, O165:H25
D	Low	Rare	No	Vtx1 and/or vtx2	±	Multiple
E	Non-human only	NA[c]	NA[c]	Vtx1 and/or vtx2	±	Multiple

[a] Reported frequency in human disease

[b] Haemolytic uraemic syndrome (HUS) or haemorrhagic colitis (HC)

[c] NA = not applicable

out whether virulence genes or clusters of virulence genes could be linked to these different classes, the presence (or absence) of these genes was determined for a set of nearly 225 STEC strains that were isolated from humans, food and cattle by Franz and colleagues. The number of virulence genes in seropathotype A (with high incidence in humans and associated with severe disease, in which currently only STEC O157 and stx-producing EAEC O104:H4 are grouped) and seropathotype B (with moderate incidence in humans and associated with severe disease, in which among others O26 and O103 are grouped) were significantly higher than in STEC belonging to seropathotype C, D and E (Fig. 10.1). However, the situation was more complicated, as they also identified STEC serotypes with high numbers of virulence factors that were originally (based on epidemiological association) classified as seropathotype C (low incidence, associated with severe disease; O76:H19, O84:H-, O5:H-, O165:H-, O55:H7) and D (low incidence, not associated with severe disease; O6:H25, O80:H-, O101:H9). This shows that enforcement of food safety relying only on a predefined set of serotypes, those with a known historical epidemiological association with severe disease and outbreaks, is not sufficient and will not protect consumers from emerging pathogenic STEC.

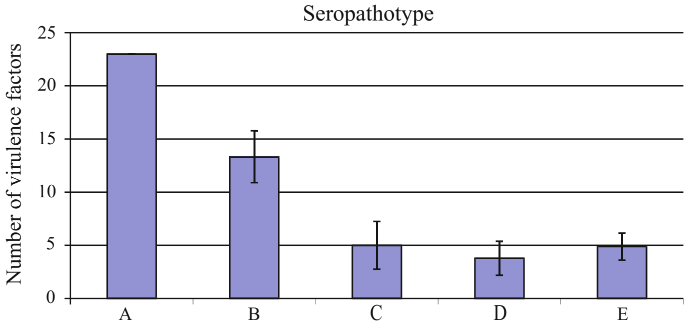

Fig. 10.1 Relation between the average numbers of virulence factors present in STEC isolates grouped in different seropathotypes (based on epidemiological association with severe disease and HUS)

In addition to PCR analysis, a subset of these strains was investigated by WGS and by studying the adherence to epithelial cells as an in vitro proxy for virulence. The sequences of the O157 strains were mapped to a reference genome (O157 Sakai) in order to identify SNPs. In total, 17 Single Nucleotide Polymorphisms (SNPs) could be identified that were significantly associated with an increased adherence to Caco-2 cells, a human gut epithelial cell line. These SNPs are shown in Table 10.3, together with the biological information about the genes involved (modified from Table 3 in Pielaat et al. 2015). As shown in this Table, 8 mutations lead to a change in the amino acid sequence of the protein and have a potential biological effect. Such mutations are called “nonsynonymous”. The other 9 SNPs are called “synonymous” and do not alter the amino acid sequences. As an example, locus ECs1203 (antitermination protein Q encoded by prophage CP-933R) contains two SNPs (one synonymous and one non-synonymous) and is located upstream of the toxin producing gene (causing the severe disease haemolytic-uremic syndrome (HUS)) stx_2. Since the antitermination gene Q controls the level of Stx2 production and as it has been proposed that Stx2 promotes epithelial cell colonization, this gene is a potential candidate for an increased attachment marker. Maybe the most important finding of this study was that it revealed practical implications when using SNP data for risk assessment in a genome-wide association approach. These implications include optimum sample size for valid inference on population level, correction for population structure, quantification and calibration of results, reproducibility of the analysis, links with epidemiological data and linking to human health. Another major result of the study was the identification of genes with potential biological significance, although biological confirmation has not yet been performed. And as such, this study provides a first approach in (statistical) methodology development for WGS application in risk assessment.

In conclusion, the technical advance in sequencing whole genomes has enormous impact on the identification, typing and characterization of foodborne pathogens. WGS is becoming “the” method of choice for addressing foodborne outbreaks, attribution studies and for assessing the risk of food-borne pathogens.

Table 10.3 Biological information regarding the 17 significant SNPs, obtained using a model without correction for population structure; locus on the reference genome [position (bp)], function of this locus and description of the SNP (from Pielaat et al. 2015)

ID	Position (bp)	Locus tag Sakai	Function	SNP description[a]
1	808,227	ECs0729	RhsC protein	Synonymous; C219T
2	1,204,977	ECs1121	Prophage CP-933R tail fiber protein; putative host specificity protein	Synonymous; C1741T
3	1,265,758	ECs1203	Antitermination protein Q	Synonymous; C12T
4	1,265,760		Encoded by prophage CP-933R	Non-synonymous; G14A (R5Q)
5	1,955,401	ECs1977	Phage capsid and scaffold protein	Synonymous; C156T
6	1,963,016	ECs1987	Tail assembly protein	Synonymous; G351C/T
7	1,965,259	ECs1990	Prophage CP-933V tail fiber protein; putative host specificity protein	Synonymous; C1062T
8	2,168,378	ECs2164	Minor tail protein encoded by	Non-synonymous
9	2,168,379		Prophage CP-933O	T424G, C425A (S142E)
10	2,303,672	ECs2332	L-Arabinose 1-dehydrogenase	Non-synonymous; C268A (H90N)
11	3,115,509		Intergenic	G → A
12	3,480,394	ECs3489	Phage tail fiber protein encoded by prophage CP-933P	Synonymous; G252A
13	3,486,443	ECs3499	Hypothetical protein	Non-synonymous; T98C (L33S)
14	3,486,494			Non-synonymous; T149C (I50T)
15	4,929,010	ECs4864	RhsH protein	Non-synonymous; T134C (F45S)
16	5,054,140	ECs4969	Putative portal protein	Non-synonymous; G190A (E64K)
17	5,409,931	ECs5283	DNA-binding transcriptional repressor UxuR	Non-synonymous; C534A (N178K)

Note

[a] SNPs are displayed by type and position in the locus, followed in parentheses by the effect on the amino acid sequence in case of a non-synonymous SNP

10.6 Impact of WGS/NGS

1. WGS represents a huge step in diagnostic microbiological practice by giving a boost in the development of tools for detection, identification and characterization of food-borne pathogens

2. NGS allows identifying rare(r) variants that are otherwise missed due to the relative low costs to obtain high sequence coverage during a single run
3. NGS holds great promise for improving surveillance, (dispersed) outbreak investigation and the detection of emerging foodborne diseases, as genetic variation can be investigated at a much higher resolution, resulting in higher discriminatory power
4. NGS provides the opportunity to create a global system of linked databases for the identification and detailed genetic characterization of all microorganisms in clinical (and other) settings. Such a global system will strengthen local, national, and international surveillance of infectious diseases (Aarestrup et al. 2012)
5. Aarestrup et al. (2012) also discussed the advantage of the use of a single technology applied in different disciplines (e.g., bacteriology, virology, parasitology) and domains (human, food, animal, environment). This would facilitate global cross-cutting collaboration and information exchange (integrated surveillance), enabling rapid and coordinated responses to novel and known health threats
6. Sequencing of genomes will lead to a better identification of the phylogenetic relationships between strains and WGS is likely to become the method of choice for monitoring pathogens in time and geographically
7. Specific to food safety and food-borne illness, WGS has the potential capability to predict the virulence properties (and thereby the disease development), antimicrobial resistance (and thereby treatment options), host specificity (and thereby potential sources), etc. of an (outbreak) strain.

However, a successful introduction of Whole Genome Sequencing as the routine molecular (typing) method in outbreak investigation and in molecular risk assessment will need:

1. An agreed and interpretable epidemiological nomenclature for WGS typing
2. Agreed threshold levels for the degree of similarity that strains must comply with in order to make inference on whether an isolate is part of an outbreak or is unrelated to the outbreak. Over-discrimination might be a pitfall
3. As the volume of information contained in a genome sequence is vast, policies and security measures should be available to maintain the privacy and safety of this information.

References

Aarestrup FM, et al. Integrating Genome-based Informatics to Modernize Global Disease Monitoring, Information Sharing, and Response. EID. 2012;18(11).

Drummond AJ, et al. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29:1969–73.

EFSA. Scientific Opinion on the evaluation of molecular typing methods for major food-borne microbiological hazards and their use for attribution modelling, outbreak investigation and scanning surveillance: Part 1 (evaluation of methods and applications. EFSA J. 2013;11:3502.

Franz E, et al. Genetic features differentiating bovine, food, and human isolates of Shiga toxin-producing *Escherichia coli* O157 in The Netherlands. J Clin Microbiol. 2012;50(3):772–80.

Foxman B, et al. Choosing an appropriate bacterial typing technique for epidemiologic studies. Epidemiol Perspect & Innov. 2005;2:10.

Jolley KA, Maiden MCJ. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinform. 2010;11:595–606.

Karmali MA, et al. Association of genomic O island 122 of Escherichia coli EDL 933 with verocytotoxin-producing *Escherichia coli* seropathotypes that are Linked to epidemic and/or serious disease. J Clin Microbiol. 2003;41(11):4930–40.

Pérez-Losada M, et al. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. Infect Genet Evol. 2013;3. http://dx.doi.org/10.1016/j.meegid.2013.01.009.4.

Pielaat A, et al. First step in using molecular data for microbial food safety risk assessment; hazard identification of *Escherichia coli* O157:H7 by coupling genomic data with in vitro adherence to human epithelial cells. Int J Food Microbiol. 2015. http://dx.doi.org/10.1016/j.ijfoodmicro.2015.04.009.

Sabat AJ, et al. Overview of molecular typing methods for outbreak detection and epidemiological surveillance (Review). Eurosurveillance. 2013;18(4).

Schmid D. Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011–2013. Clin Microbiol Infect. 2014;20:431–6.

Recommended Literature

Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, Thomas JE, Gannon VPJ. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinform. 2010;11.

Stasiewicz MJ, et al. Genomics tools in microbial food safety. Curr Opin Food Sci. 2015;4:105–10.

Strachan NJC, Rotariu O, Lopes B, Macrae M, Fairley S, Laing C, Gannon V, Allison LJ, Hanson MF, Dallman T, et al. Whole genome sequencing demonstrates that geographic variation of *Escherichia coli* O157 genotypes dominates host association. Sci Rep. 2015;5.

Struelens MJ, Brisse S. From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. Eurosurveillance. 2013;18(4).

Underwood AP, et al. Public health value of next-generation DNA sequencing of enterohemorrhagic *Escherichia coli* isolates from an outbreak. J Clin Microbiol. 2013;51:232–7.