International Journal of Food Microbiology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

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

Annemarie Pielaat [a,*], Martin P. Boer [b], Lucas M. Wijnands [a], Angela H.A.M. van Hoek [a], El Bouw [a], Gary C. Barker [c], Peter F.M. Teunis [a,d], Henk J.M. Aarts [a], Eelco Franz [a]

[a] National Institute for Public Health and the Environment (RIVM), Centre for Infectious Disease Control, A. van Leeuwenhoeklaan 9, 3720 BA Bilthoven, The Netherlands
[b] Wageningen UR Biometris, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands
[c] IFR, Institute of Food Research, Norwich Research Park, Norwich, UK
[d] Rollins School of Public Health, Emory University, Atlanta, GA, USA

ARTICLE INFO

Available online xxxx

Keywords:
Microbiology
Risk assessment
GWAS
SNP
STEC

ABSTRACT

The potential for using whole genome sequencing (WGS) data in microbiological risk assessment (MRA) has been discussed on several occasions since the beginning of this century. Still, the proposed heuristic approaches have never been applied in a practical framework. This is due to the non-trivial problem of mapping microbial information consisting of thousands of loci onto a probabilistic scale for risks. The paradigm change for MRA involves translation of multidimensional microbial genotypic information to much reduced (integrated) phenotypic information and onwards to a single measure of human risk (i.e. probability of illness).

In this paper a first approach in methodology development is described for the application of WGS data in MRA; this is supported by a practical example. That is, combining genetic data (single nucleotide polymorphisms; SNPs) for Shiga toxin-producing *Escherichia coli* (STEC) O157 with phenotypic data (in vitro adherence to epithelial cells as a proxy for virulence) leads to hazard identification in a Genome Wide Association Study (GWAS).

This application revealed practical implications when using SNP data for MRA. These can be summarized by considering the following main issues: 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, anchoring and integration of results into a systems biology approach for the translation of molecular studies to human health risk.

Future developments in genetic data analysis for MRA should aim at resolving the mapping problem of processing genetic sequences to come to a quantitative description of risk. The development of a clustering scheme focusing on biologically relevant information of the microbe involved would be a useful approach in molecular data reduction for risk assessment.

1. Introduction

Microbiological risk assessment is part of an established framework for risk analysis that consists of the following steps: statement of purpose, hazard identification, hazard characterization, exposure assessment and risk characterization (CAC, 1999). In the field of food safety, a 'farm to fork' quantitative risk assessment (QMRA, Quantitative Microbial Risk Assessment) approach is often applied to assess the public health risk for a particular pathogen/matrix combination (e.g. Romero-Barrios et al., 2013). Transmission of the pathogen (e.g. *Campylobacter* spp.) through a specific food production chain (e.g. poultry) may be quantified using a probabilistic QMRA model (e.g. Nauta et al., 2005). Variability and/or uncertainty in the pathogen prevalence, concentrations and food production process properties are included as model parameters. Monte Carlo simulations, or other probabilistic techniques, are used to predict public health risk and the effect of different intervention strategies can then be calculated to support industrial or governmental decision making (e.g. Pielaat et al., 2014). Systematic sensitivity analyses can be used to indicate the value of new evidence but only at the level of detail that was used during the model construction.

Since the introduction of high-throughput DNA sequencing technologies, however, food microbiology has moved beyond the assessment of microbial behavior in different food processes for agents classified at (sub)species and serovar level. Moreover, with the rapidly dropping costs of sequencing, whole genome sequencing (WGS) will soon

* Corresponding author. Tel.: +31 30 274 3711.
E-mail address: annemarie.pielaat@rivm.nl (A. Pielaat).

http://dx.doi.org/10.1016/j.ijfoodmicro.2015.04.009

become a standard surveillance technique for the subtyping of isolates for epidemiological purposes. Although the use of molecular data has proved to be a powerful tool in decision making during outbreak investigations (Dallman et al., 2014; Underwood et al., 2013), the application of this data in microbiological risk assessment is currently an unexplored area in the public health domain. In recent years, a number of reviews and opinions have been published exploring the potentials of 'omics techniques' for MRA (Abee et al., 2004; Brul et al., 2012; Carriço et al., 2013; Havelaar et al., 2010; Pielaat et al., 2013a,b) but, evidence based research, as a first step to convert these heuristic approaches into normative tools for practical use, is still needed.

The difficulties associated with using molecular data for food safety risk assessment are complex but are related to the prescribed framework and the current methodology which generally expresses a large (but closed) joint probability to represent a 'farm to fork' hazard domain. For example, where the variability and/or uncertainty of concentration and prevalence data are relevant in QMRA these can be described by probability distributions but it is not clear how to use this approach when the data consists of a genome sequence. Firstly, new technologies provide information at a completely different level of description (genes or their products) that makes their joint probability, in its simplest form, unmanageable. Secondly, the new description does not, in the first instance, provide a clear connection between the observed quantities and the output measures, such as survival or health impacts, that are the object of risk assessments. So, for decision support, the biggest challenge facing genomics is the prediction of phenotypic properties of a particular pathogen within a food chain based on genotypic data. An understanding of systems biology is needed, as the organizational principle in pathophysiology, to describe the relation between the new level of genetic sequence data and the health end points of concern. Whereas in the established framework for risk assessment the elements of a joint probability are considered to be known, or knowable, the introduction of a new level of description and a systems property leads to elements of a joint probability that cannot easily be formulated and to dependencies that are not easy to identify. To reduce the numbers of possible relations and translate genetic sequence into phenotypic properties, an understanding of pathogen physiology is needed. Currently, such understanding is incomplete and consequently the mapping of genetic sequences onto a quantitative description of risk is problematic: the number of genes outweighs the number of strain samples by many orders of magnitude. It should be clear that statistical analysis of WGS data is non-trivial, and that reproducible and meaningful associations between gene variability and phenotypic properties need to be established before genetic data can be used for decision making in food safety.

As indicated during EFSA's 20th scientific colloquium (EFSA, 2014), a diversity of exemplary data analyses need to be developed and shared within the scientific community to allow for the identification and appreciation of "best practices" in moving forward from current methodology. A (theoretical) methodology for hazard identification is proposed that uses WGS data analysis to link genomic sequences with phenotypic behavior for Shiga toxin-producing *Escherichia coli* O157 (STEC O157) as a case study. This is achieved by the integration of genomic (single nucleotide polymorphism (SNP) genotypes) data with phenotypic (attachment to epithelial cells) information and with epidemiological data (outbreak strains and the epidemiological relationship with sporadic cases) in a Genome Wide Association Study (GWAS).

The aim of this study is to introduce a method for hazard identification that links WGS data with results on in vitro adherence to epithelial cells as a proxy for virulence using a subset of STEC O157 isolates as a case study. An explanation of the concepts, identification of the value of the methodology and a relationship with the public health domain are supported by a thorough discussion of further research needs. This paper identifies a necessary paradigm change in public health microbiological risk assessments.

2. Materials and methods

2.1. STEC O157 as a case study

STEC is of public health concern because of its ability to cause outbreaks and severe disease such as hemorrhagic colitis (HC) or hemolytic–uremic syndrome (HUS). Currently, different STEC serogroups are placed in different risk classes (i.e. seropathotypes) based on their epidemiological association with severe disease and outbreaks (Karmali et al., 2003). However, this system is of limited use for two reasons. First, it is retrospective, only including known types. Secondly, the pathogenicity of STEC cannot be predicted from the serotype alone. Numerous (putative) virulence genes have been associated with increased disease severity and individual strains of STEC can differ considerably in their virulence profile and, consequently, in their pathogenic potential (Delannoy et al., 2013). Gene association studies are normally conducted by linking the genetic content of the strain to the seropathotype or more specific to the clinical symptoms it caused (Andersson et al., 2011; Persson et al., 2007). However, these association studies might be confounded by food and host effects. Within serogroup O157 considerable attention has been given to the non-random distribution of genotypes among bovine and human clinical isolates, showing considerable genome divergence (Franz et al., 2014). However, observed non-random distribution of clades and lineages among bovine and human clinical isolates might be the result of a differentiation in virulence, transmission capacity and survival, or some combination (Franz et al., 2012). For a better understanding of STEC O157 risks, these (or other) potential causes should be investigated separately. Recently it was shown that the environmental exposure route selects for strains characterized by the absence of mutations in the general stress response system *rpoS*, which are subsequently more likely to survive the human gastric barrier (Franz et al., 2011; van Hoek et al., 2012).

The evaluation of intrinsic differences in virulence requires a standardized model system. Several animal models for STEC disease exist and their value is clearly recognized (Melton-Celsa and O'Brien, 2003). However, for technical, economic, and ethical reasons, in vitro models offer a relevant alternative to in vivo studies. Although more distinct from a human system, in vitro models offer more stability in terms of reproducibility (Berk, 2008). The combination with WGS information subsequently allows for genotype–phenotype matching and comparative genomics of strains in order to identify genetic elements that differentiate highly virulent strains from less virulent ones. Analysis at the SNP level is a straightforward approach to extracting elementary information on genotype, and will be used in this study.

2.2. E. coli O157 strains

In an earlier study the frequency of *E. coli* O157 genotypes among 73 bovine, 29 food, and 85 human clinical isolates was determined in The Netherlands (Franz et al., 2012). The results demonstrated that O157 lineages (as defined by the lineage specific polymorphism, or LSPA, assay) were non-randomly distributed among isolates of bovine and clinical human origin. A selection of in total 38 human and animals strains from different LSPA lineages was selected for further investigation in this study (Table 1).

2.3. Genotypic data

Whole genome sequences of the 38 *E. coli* O157 were obtained using the Illumina MiSeq platform with 2 × 150 (human isolates) and 2 × 250 (animal isolates) paired end runs.

The genomic sequence of a human Shiga toxin-producing *E. coli* O157:H7 strain isolated during the Sakai outbreak which occurred in Japan during 1996 was used as a reference. The short sequencing reads were mapped onto the reference chromosome (accession

Please cite this article as: Pielaat, A., et al., First step in using molecular data for microbial food safety risk assessment; hazard identification of *Escherichia coli* O157:H7..., Int. J. Food Microbiol. (2015), http://dx.doi.org/10.1016/j.ijfoodmicro.2015.04.009

Table 1
E. coli O157 strains (n = 38) used in this study and some of their genetic characteristics.

Strain	Source	Year	LSPA[a]	*stx* gene(s)	Intimin (*eae*)	*tir* (A255T)[b]	clade 8[c]	SBI[d]
H06	Human	2005	I	stx_{2a}	+	T	−	na[e]
H07	Human	2005	I	stx_{2a}	+	T	−	na[e]
H09	Human	2005	I	stx_{2a}	+	T	−	na[e]
H13	Human	2006	I	stx_{2a}	+	T	−	3
H15	Human	2006	I	stx_{2a}	+	T	−	3
A42	Bovine	2002	I	stx_{2a}	+	T	−	3
H25	Human	2006	I/II	stx_{2c}	+	T	−	1
H27	Human	2006	I/II	stx_{2a}	+	T	−	1
H42	Human	2007	I/II	stx_{2a}, stx_{2c}	+	T	+	1
H44	Human	2007	I/II	stx_{2a}	+	T	−	21
H48	Human	2008	I/II	stx_1, stx_{2c}	+	T	−	16
H49	Human	2008	I/II	stx_1, stx_{2c}	+	T	−	6
H83	Human	2009	I/II	stx_1, stx_{2c}	+	T	−	16
A25	Bovine	2008	I/II	stx_{2c}	+	T	−	21
A37	Bovine	2007	I/II	stx_1, stx_{2c}	+	T	−	6
A40	Bovine	2002	I/II	stx_{2a}, stx_{2c}	+	T	+	1
A45	Bovine	2007	I/II	stx_1, stx_{2c}	+	T	−	16
A48	Bovine	2008	I/II	stx_1, stx_{2c}	+	T	−	6
A51	Bovine	2002	I/II	stx_1, stx_{2c}	+	T	−	6
A60	Bovine	2003	I/II	stx_{2c}	+	T	−	1
A62	Bovine	2008	I/II	stx_1, stx_{2c}	+	T	−	6
A63	Bovine	2008	I/II	stx_1, stx_{2c}	+	T	−	6
A69	Bovine	2002	I/II	stx_{2a}	+	T	−	1
A72	Bovine	2002	I/II	stx_{2a}	+	T	−	1
A76	Bovine	2003	I/II	stx_{2c}	+	T	−	5
H02	Human	2003	II	stx_{2a}	+	T	−	11
H17	Human	2006	II	stx_{2a}, stx_{2c}	+	A	−	1
H19	Human	2006	II	stx_{2c}	+	A	−	1
H24	Human	2006	II	stx_{2c}	+	A	−	5
H32	Human	2006	II	stx_{2c}	+	A	−	5
H51	Human	2008	II	stx_{2c}	+	A	−	1
A12	Bovine	2004	II	stx_{2c}	+	T	−	5
A13	Bovine	2008	II	stx_{2c}	+	A	−	5
A16	Bovine	2006	II	stx_{2c}	+	A	−	5
A29	Bovine	2009	II	stx_1, stx_{2a}, stx_{2c}	+	A	−	16
A30	Bovine	2009	II	stx_{2c}	+	T	−	5
A32	Bovine	2009	II	stx_1, stx_{2c}	+	A	−	6
A34	Bovine	2009	II	stx_{2c}	+	A	−	5

Note:
[a] Lineage-specific polymorphism assay (Yang et al., 2004).
[b] *tir* (*A255T*) polymorphism assay (Bono, 2009).
[c] Clade 8 status of isolates assessed by SNP analysis of ECs2357 (Riordan et al., 2008).
[d] Shiga toxin-encoding bacteriophage insertion site assay (Shaikh and Tarr, 2003; Besser et al., 2007).
[e] Not applicable, the insertion site assay did not result in a SBI genotype.

number: NC_002695) and its large plasmid pO157 (accession number: NC_002128) using the alignment tool *BWA* (Li and Durbin, 2010). The *SAMtools* software package converted the SAM format files to BAM format and sorted the BAM files (Li et al., 2009). SAMtools mpileup generated BCF format files and bcftools was used to call the SNPs (between reference and samples) in VCF format. SNPs were filtered with the quality threshold set at a minimum read depth of five. SNP data were transformed to binary, 0/1, data. Those sites where a SNP was identified in the STEC strains under study (test strains) compared to Sakai (reference strain) received a 1. A 0 was placed accordingly for identical nucleotides on the genome of each test strain compared to the reference strain Sakai. This resulted in a binary *m* by *n* matrix, where *m* is the number of SNPs and *n* the number of test strains ($n = 38$).

2.4. Phenotypic data

The adhesive properties to human intestinal cells were used as a proxy for virulence in this study. In total, 18 human O157 isolates (H-numbers) and 20 animal O157 isolates (A-numbers) from lineage I, I/II and II (see Table 1 for strain properties) were investigated and compared. Differentiated Caco-2 cells were used as a representative model system for human intestinal cells. They were produced, treated and seeded as previously described by Oliveira et al. (2011). For adhesion experiments the cells were seeded in 12-well plates at a concentration of 1.6×10^5 cells/well. Each STEC strain was inoculated in BHI broth and incubated overnight (ON) at 37 °C to obtain a culture consisting of approximately 10^9 CFU/ml. Subsequently, three decimal dilutions resulting in a STEC suspension of each strain of approximately 10^6 CFU/ml were made. From these last suspensions, Caco-2 cells in 12-well plates were inoculated with 40 μl per well, per STEC strain six wells. Plates were centrifuged (1 min at 175 ×*g*) and incubated at 37 °C in a humidified atmosphere of 95% air and 5% CO_2 for 1 h.

After incubation and three washings with pre-warmed sterile phosphate-buffered saline (PBS), 1 ml 1% Triton-X100 in PBS (pre-warmed) was added to each well to detach the cells.

The detached cells were collected and the contents of three wells were combined. Decimal dilutions were prepared in peptone physiological salt solution, and the dilutions were plated on Brilliance™ *E. coli*/coliforms Selective Agar (Oxoid, Badhoevedorp, The Netherlands).

3. Statistical data analysis

3.1. Phenotypic data

The fractional adherence was calculated by dividing the number of STECs after the adhesion assay by the number of STECs added to the Caco-2 cells. If cells are assumed to be homogeneously distributed in the ON culture, then the number of cell counts per dilution is Poisson distributed with a parameter λ. Based on this assumption, the best estimate for λ, the expected number of cells in a sample, can be estimated from counts in serial dilutions of the original sample according to $\frac{\sum_{i=j}^{k} n_i}{\sum_{i=j}^{k} 10^{-i}}$, where *n* is the number of counts in the *i*th dilution. The point estimate for the fraction of bacteria attached to the Caco-2 cells is $\frac{\lambda_2}{\lambda_1}$, where λ_1 is the expected number of bacteria in the overnight culture and λ_2 is the expected number of bacteria attached to the Caco-2 cells.

3.2. Simple linear regression

The basic idea in this study is to identify SNPs in the test strains that could be associated with an increased virulence behavior (represented by relatively high Caco-2 cell attachment fractions compared to other test strains without these SNPs).

The strength of this association can, in the most basic form, be estimated using the following simple linear regression model for each SNP:

$$y_i = \mu + \beta x_i + \varepsilon_i ,$$

where y_i is the fractional Caco-2 adhesion for strain *i*, μ is the mean response, β is the SNP effect, x_i is an indicator variable with

$$x_i = \begin{cases} 0 & \text{if the marker (here, SNP) score of test strain i is equal to the reference strain} \\ 1 & \text{otherwise,} \end{cases}$$

and the residual errors, ε_i , have independent normal distributions with variance σ_e^2.

For each SNP the null hypothesis $H_0 : \beta = 0$ is tested against an alternative hypothesis $H_a : \beta \neq 0$ and the P-values and effects β are calculated and transformed to a $-\log 10$ scale.

In Genome Wide Association Studies (GWAS) a corresponding test is applied for all markers. In terms of hazard identification this simplified model would assign strains to one of two classes of hazards, with different rates of adhesion, depending on the presence of a single SNP.

There are several issues in GWAS studies that are more complex than in standard linear regression. First of all, as the number of SNPs (m) outweighs the number of test strains (here, $n = 38$) one has to correct for multiple testing. If a standard 5% significance threshold per

marker is used, there will be many false positives. One solution is to use a Bonferroni correction (Kuehl, 2000), using a genome wide significance threshold of α/m. A second problem in GWAS is that the minor allele frequency (MAF) has to be high enough, i.e. we cannot use markers where almost all the markers are equal to the reference strain (score 0), or where almost all the markers are different from the reference strain (score 1). A third problem with a simple linear regression model is more complicated: there is no correction for population structure. In the next section a solution for this will be presented by using a more complicated statistical model called a mixed model.

3.3. Correcting for population structure using mixed effect models

The problem with population structure can be explained with a simple example: suppose there are two groups of strains, A and B, with small genetic differences within groups, and a large genetic difference between groups. Then most SNP scores will follow the same pattern. That is, SNP scores will be attributable to differences in the groups rather than the difference in adherence capacity. As a consequence, the linear regressions will also be quite similar. In other words, in the case where one has two clearly separated groups one is mainly testing for differences between groups, not within groups. In such an example with two clear groups A and B, an extra term in the regression model can be used to correct for group effect, that is

$$y_i = \mu + \beta x_i + g z_i + \varepsilon_i ,$$

where the parameter g is the group effect, and z_i indicates whether strain i is in group A or B. With this correction for group effect one can test for SNP effects along the genome, i.e. testing for each SNP the significance of parameter β. A similar approach can be used in situations where there are strains which can be subdivided in several groups (Kraakman et al., 2004; Pritchard et al., 2000).

In many cases, however, there is not such a clear separation into different groups. A way to check this is by calculating the similarity between the strains. For each pair of strains, the fraction of SNPs that have the same marker score can be calculated, resulting in an $n \times n$ similarity matrix K.

The similarity matrix may be used to correct for population structure (Malosetti et al., 2007; Patterson et al., 2006; Yu et al., 2005). A solution that is often used is a mixed effect model approach:

$$y_i = \mu + \beta x_i + G_i + \varepsilon_i ,$$

where $(G_1, G_2, \ldots, G_n)$ follows a multinormal distribution, $G_i \sim N(0, \sigma_g^2 K)$, and σ_g^2 is the genetic variance. This model is used to test for the significance of the SNP effects, β, along the genome.

4. Results

The fractional adhesion of the STEC O157 strains to Caco-2 cells is highly variable with a frequency distribution resembling an overall skewed distribution (average 0.16, median 0.11). Higher fractions are more rare and concentrated in the human strains (Fig. 1). Mapping of the individual sequences to the reference genome resulted in the identification of 27,980 SNPs among the total set of 38 test strains. When the totality of identified SNPs was used to infer the population structure, at least three separate populations could be identified (Fig. 2). Fig. 2 shows the principal coordinate plot of the 38 strains, with 12.4% of the variance explained by the first axis, and 8.9% by the second axis. The three identified populations in Fig. 2 reflect the LSPA lineages within STEC O157 (LI, LI/II and LII) as described by Franz et al. (2012).

After application of the most basic linear regression model (without correcting for population structure), 17 SNPs appeared to be significantly associated with increased adherence to Caco-2 cells (Fig. 3, Table 2). That is, using a MAF ≥ 0.05 and a significance level $\alpha \sim 10^{-4}$, 17 positions

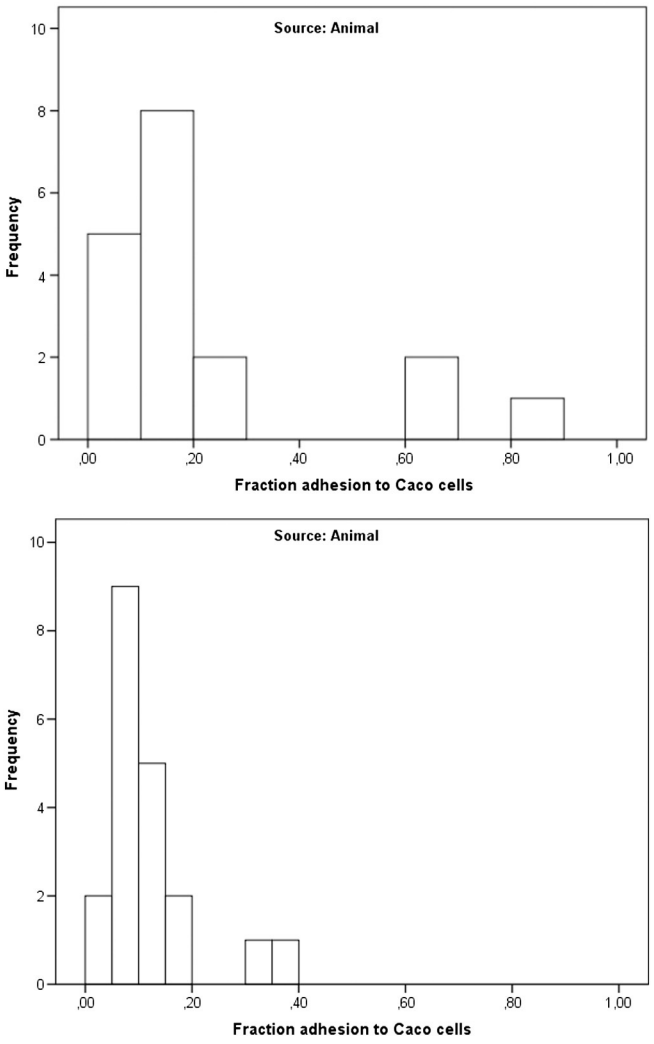

Fig. 1. Frequency distributions of the fractional adhesion for the different STEC strains to Caco-2 cells (separated for human and animal strains).

on the chromosome of the reference strain show a positive linear relation between SNPs identified in the test strains when comparing fractional adhesion to Caco-2 cells with that for the Sakai reference.

Table 3 shows the loci and, if known, information on biological function, associated with the significant SNPs (SNPs having a MAF ≥ 0.05 and, in bold face, MAF ≥ 0.1, i.e. ID 8, 9 and 15) as presented in Table 2. As explained above, instead of having any biological relevance, the results in this SNP analysis could also be the product of a type I error. That is, identifying significant association between genotypic (SNP) and phenotypic (fraction adhesion) information, where there is none.

Having said this, the 17 SNPs identified warrant further investigation with respect to their role in virulence and their use as risk markers for hazard identification. Of these 17 SNPs, eight were non-synonymous in protein-coding regions (Table 3). These SNPs change the protein-sequence and thereby potentially the function of the product.

Table 4 shows which test strains were responsible for the significant effects (identified in Table 2, Fig. 3) with the corresponding fractional adhesion to Caco-2 cells. Here the problem associated with a low MAF (identified in the Statistical data analysis section) becomes visible. Setting the MAF threshold at 0.05 will result in a significant effect when two (or more) test strains appear to share a SNP and are associated with a relatively high (or low) fraction of attachment compared to the test strains that do not differ from the reference Sakai strain for that

Please cite this article as: Pielaat, A., et al., First step in using molecular data for microbial food safety risk assessment; hazard identification of *Escherichia coli* O157:H7..., Int. J. Food Microbiol. (2015), http://dx.doi.org/10.1016/j.ijfoodmicro.2015.04.009

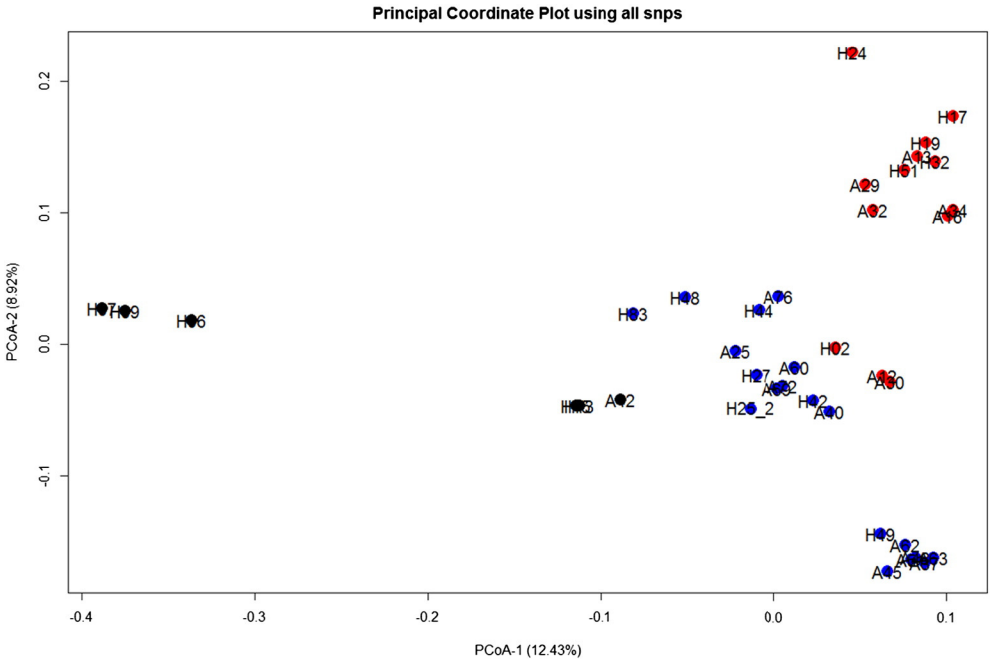

Fig. 2. Principal coordinate plot of the similarity matrix *K*. For each pair of strains the similarity was calculated as the fraction of SNPs that have the same score. Black, blue and red dots represent lineage I, I/II and II STEC O157 strains respectively.

site. The result of a type I error in multiple testing (here 27,980 tests) and/or not correcting for population structure may be the cause of this effect.

When correcting for population structure one SNP appears to have a significant effect in this GWAS (Fig. 4). The identified SNP position is 3,115,507 which is close to ID 11 in Table 3 and also has an intergenic position, which in many cases might be irrelevant, but could still be related to promoter sequences.

5. Discussion

Microbiological risk assessment is intended to support decision making in the farm to fork food production chain. Currently food safety criteria are implemented at many levels down to serovar/serotype level for some pathogens (e.g. absence of *Salmonella* Typhimurium and *Salmonella* Enteritidis in fresh poultry or STEC O157, O26, O103, O111, O145 and O104:H4 in sprouts). From this perspective it is difficult to assess how the increasing amount of new (molecular) data will influence decision making. Does a 'new' genotype identify a new hazard and thus require a change in policy? And, how should the presence/absence of a virulence gene influence the hazard characterization? There are many such questions and probably even more possible answers related to this problem which can, in general, be referred to as disaggregation. For example, if for a previous 'generalized' pathogen (e.g. a serotype) additional information can be specified to describe two 'less general' pathogens there is an increase in the number of hazards. This increase

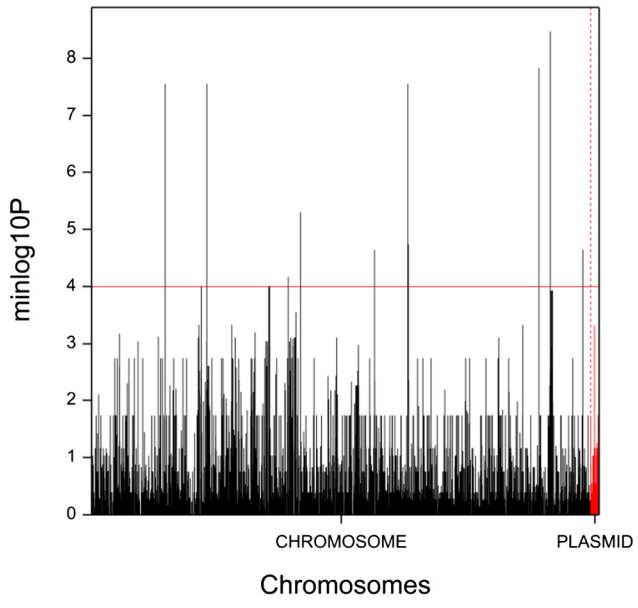

Fig. 3. GWAS plot for trait fractional adhesion to Caco-2 cells. In this plot there is no correction for population structure, i.e. highly overestimating the number of true significant SNP effects.

Table 2
Number of positions (1–17) on the chromosome of STEC O157 strains, locus on the reference genome (position (bp)), minor allele frequency (MAF), point estimate for the regression coefficient (β) and $-\log10$ P-value for the SNPs on the test strains compared to the reference strain having a positive relation with the fraction adhesion to Caco-2 cells. This table shows the results for the regression model without correction for population structure, i.e. overestimating the effects.

ID	Position (bp)	MAF	β	$-\log10$ (P)
1	808,227	0.05	0.31	7.55
2	1,204,977	0.08	0.20	4.01
3	1,265,758	0.05	0.31	7.55
4	1,265,760	0.05	0.31	7.55
5	1,955,401	0.08	0.20	4.01
6	1,963,016	0.08	0.20	4.01
7	1,965,259	0.08	0.20	4.01
8	2,168,378	0.11	0.18	4.17
9	2,168,379	0.11	0.18	4.17
10	2,303,672	0.08	0.22	5.29
11	3,115,509	0.08	0.21	4.64
12	3,480,394	0.05	0.31	7.55
13	3,486,443	0.08	0.21	4.73
14	3,486,494	0.08	0.20	4.14
15	4,929,010	0.11	0.23	7.83
16	5,054,140	0.05	0.32	8.47
17	5,409,931	0.08	0.21	4.64

Table 3
Biological information regarding the 17 significant SNPs, obtained using a model without correction for population structure (in Table 2); locus on the reference genome (position (bp)), function of this locus and description of the SNP.

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-933 V 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:
SNPs having a MAF ≥ 0.05 and, in bold face, MAF ≥ 0.1, i.e. ID 8, 9 and 15.
[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.

is exponential because all possible combinations may be relevant. Obviously, when this argument is stretched to the full sequence specification, the disaggregation catastrophe is clear. There will never be enough risk assessments to quantify and predict the full spectrum of all risks. On the other hand, two different subtypes of a 'generalized' pathogen may have identical virulence, and their distinction is irrelevant for public health. So, an 'organization principle' as a basis for priority setting of highly pathogenic strains is a necessary prerequisite for risk assessment developed from WGS information. This is why linking genotypic with phenotypic properties is essential as it will help to identify *high risk* isolates from the full spectrum of strains obtained by WGS. In addition, methods for clustering may further reduce the total number of hazards to manageable proportions, for both risk scientists and risk managers. A decisive factor in specifying the clustering level of WGS information is having a good definition of the statement of purpose for the risk analysis. This and insight in the main biological process underlying the risk (e.g. surviving process conditions, the time to initiate growth or invasion of gut epithelium) will guide the data requirements (e.g. expression data, (single) cell growth or SNP data for marker identification).

For this case study, the purpose was to improve hazard identification for STEC O157. The first obvious step in the 'organization principle' was to identify a measure for virulence, here adhesion to the gut epithelium. The next step in identifying high risk isolates was discovery of an association between this important virulence factor and genome sequences (i.e. SNP data) for different STEC O157 strains. A further biological analysis of the identified associations may attribute some part of the biological variation (here, fractional attachment) to the activity of the specific genes corresponding to the 'significant' SNPs.

5.1. Biological relevance of identified SNPs

A potential biological relevance can be ascribed to the eight non-synonymous SNPs that are significantly (MAF ≥ 0.05 and a significance level, $\alpha \sim 10^{-4}$) associated with higher in vitro adherence to the epithelial cells using the simple linear regression model (Table 3). ECs2332

Table 4
Test strains (second row) causing a 'significant' effect (MAF ≥ 0.05 in Table 3) with accompanying fractional adhesion to Caco-2 cells (first row). e.g, for ID 1 the contrast of strains H24 and H19 compared to all 36 other strains gives a significant effect. For each SNP, only the major frequent allele is shown, the empty cells refer to the minor frequent allele.

ID	Fraction adhesion	0.66	0.12	0.07	0.18	0.87	0.10	0.11	0.35	0.17	0.11	0.09	0.62
	Strain	H13	H15	H25	H44	H24	H32	H51	A13	A16	H02	H17	H19
1						1							1
2						1						1	1
3						1							1
4						1							1
5						1						1	1
6						1						1	1
7						1						1	1
8				1		1			1				1
9				1		1			1				1
10		1			1	1							
11		1				1		1					
12						1							1
13						1				1			1
14						1	1						1
15		0	0			0							0
16		1				1							
17		1				1					1		

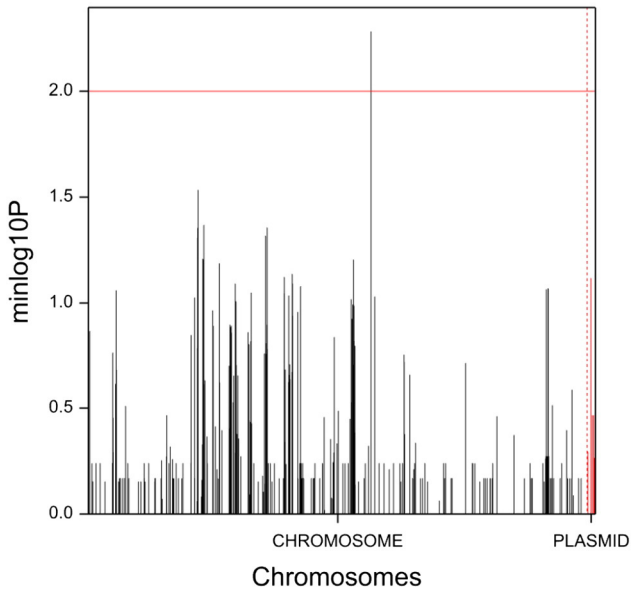

Fig. 4. GWAS plot for trait fractional adhesion to Caco-2 cells, with correction for population structure using a mixed model. There is one significant SNP, at position 3,115,507, with effect $\beta = 0.09$, and $-\log10$ (P-value) $= 2.28$.

encodes for a L-arabinose 1-dehydrogenase (family of oxidoreductases). L-Arabinose is generally well utilized as a sole carbon source by *E. coli* O157 strains (Franz et al., 2011) and the intestinal niche occupied by pathogenic *E. coli* O157 strain EDL933 has been shown to be largely defined by the utilization of arabinose for colonization (Fabich et al., 2008; Maltby et al., 2013). Phenotypic characterization revealed that mutations in the stringent response system resulted in defective utilization of L-arabinose (Oh and Cho, 2014). The ECs5283 locus represents the DNA-binding transcriptional repressor UxuR which plays a role in the D-glucuronate metabolism of *E. coli* and has been shown to be necessary for maximum ability of *E. coli* to colonize the intestine (Chang et al., 2004). ECs4969 is a putative phage portal protein, which forms a key component during viral chromosome packaging. Although the same locus was found to be significantly upregulated upon exposure of *E. coli* O157 to apple juice (Bergholz et al., 2009) the role in pathogenicity remains elusive. Sequence divergence of the phage borne antitermination gene *Q* (ECs1203), located upstream of stx_2, has been associated with variation in transcription of stx_2 and with a nonrandom distribution among bovine and human isolates (Franz et al., 2012; Lejeune et al., 2004). The same locus was shown to be significantly up-regulated in a lineage (clade 8) of *E. coli* O157:H7 commonly associated with human infections (Abu-Ali et al., 2010). Recently, Xu et al. (2012) proposed a model in which Stx2 promotes epithelial cell colonization. Since the antitermination gene *Q* controls the level of Stx2 production, this gene is a potential candidate for an increased attachment marker. The *rhs* genes are rearrangement hot spots within *E. coli* and appear to be under strong positive selection (Petersen et al., 2007). Their extracellular nature may result in a strong positive selective pressure from the host's immune system and may therefore be involved in O157 host specificity (Liu et al., 2009). They have been reported to promote intestinal colonization of calves (van Diemen et al., 2005).

Bioinformatic analyses support the conclusion that bacterial Rhs proteins commonly carry toxin domains and it is proposed that contact-dependent growth inhibition is the primary function of these proteins (Hayes et al., 2014). However, Rhs also appears to coordinate multicellular behavior and biofilm formation (Poole et al., 2011). Interesting however is the fact that sequence diversity in *rhs* is also linked to the differentiation of STEC O157 into different clades which in turn have different associations with epidemiology (Liu et al., 2009; Manning et al., 2008). The *rhs* should be further investigated with respect to combined relevance for biological functioning, phylogeny and epidemiological relevance.

5.2. Advantages and limitations of the approach

Traditionally, STEC hazard identification is based on the seropathotype (SPT) concept of classifying STEC serogroups into different risk classes based on their epidemiological relevance (i.e. severity of disease and involvement in outbreaks) (Karmali et al., 2003). Although informative as an ex post facto determinant of virulence potential, the dynamic nature of STEC virulence in time and place exposes a limitation of SPT classification as a predictive indicator of microbial risk. The approach described here, matching in vitro adherence to epithelial cells as a proxy for virulence with SNP data offers a standardized, reproducible, serogroup independent method for identifying potential candidate genes to be included in epidemiological association studies or more refined hazard identification. Additionally the analysis does not become rapidly unmanageable as more strains are included. SNP analysis of a broad spectrum of isolates (from different sources that are not necessarily associated with human cases) may lead to a less biased association between genotypic and phenotypic strain characteristics. Still, the identification of significant SNPs associated with in vitro virulence should be treated with some caution for several reasons.

First, the in vitro adherence of O157 to human epithelial cells was used as a proxy for in vivo virulence. However, it should be noted that adherence to epithelial cells is only one aspect of the etiology of O157 infection in humans. Although the production of Shiga toxins is the main virulence factor responsible for the more severe symptoms, adherence to the gut epithelium is an important first step in the etiology of STEC infections. The production of Shiga toxins (on expression level, produced toxin level, and/or effect on Vero cells) could be added to this framework and can be analyzed in the same way (with a focus on the prophage regions). Second, SNP analysis based on comparing test strains with one reference strain (here, *E. coli* O157 strain Sakai) cannot identify SNPs that might be of relevance for hazard identification, but have sequences that are not present in the reference strain. Reference to the pan genome of STEC O157 or even STEC in general could account for this omission (Laing et al., 2010). Alternatively, a "gene-by-gene" approach could be adopted in which the presence/absence of all genes is scored as well as the allelic variation within each gene (Maiden et al., 2013).

Third, instead of having any biological relevance, the results in this SNP analysis could also be the product of a type I error. That is, identifying a false positive association between genotypic (SNP) and phenotypic (fractional adhesion) information, where there is none. This problem will often occur in SNP analysis for MRA where the number of SNPs usually outweighs the number of phenotypic observations. Setting the MAF threshold to a higher level, e.g. 0.1, could alleviate the problem of the small population size. Table 2 shows that 'only' three positions on the chromosome of Sakai will be identified as being significant when a MAF ≥ 0.1 would be applied in this study. These represent two loci on the reference genome, i.e. ECs2164 and ECs4864 associated with 'minor tail protein encoded by prophage CP-933O' and 'RhsH protein' respectively (Table 3). These are, however, non-synonymous SNPs. A further correction for population structure reveals one significant SNP position (3,115,507 which is intergenic) of potential relevance for further biological investigation.

Third, although the loci in which SNPs occur are potentially candidate marker genes for increased adherence to epithelial cells (and maybe virulence in general), the molecular postulates have to be fulfilled in order to prove causal relations (Falkow, 1988).

Finally, the presence of a biomarker (e.g. SNP, gene, metabolite, protein) may by itself not always be a good predictor for risk, since the expression is influenced by a large variety of (dependent biological) factors.

Please cite this article as: Pielaat, A., et al., First step in using molecular data for microbial food safety risk assessment; hazard identification of *Escherichia coli* O157:H7..., Int. J. Food Microbiol. (2015), http://dx.doi.org/10.1016/j.ijfoodmicro.2015.04.009

Beside practical implications, this case study also gives insight into which basic statistical elements need to be considered before an association can be a subject for further analysis. That is,

1. The power for analysis needs to be sufficient in order to be able to detect an effect.
2. An optimal number of strains need to be established to get a representative view of the whole (molecular) population.
3. One should correct for population structure before proceeding with further molecular data analysis.

For now, the number of strain samples is very limited for GWAS purposes, and the aim is to continue combining not outbreak biased genotypic with phenotypic STEC O157 data to build a valid statistical model.

Biological confirmation is an essential step before 'significant' SNPs can be identified as the cause of hazardous strains. This process consists of,

1. Deletion and complementation studies (according to the principle of Koch's molecular postulates) to establish the causal relationships.
2. Quantification and calibration of how many differences (e.g. SNP variants) between genomes will lead to treating them as separate categories needs to be established.
3. Reproducibility of the analysis (both experimental and statistical) is a prerequisite for assigning any SNP with a biological relevance. This involves both biological and technical replicates to address variability in the phenotypic response during the statistical analysis.
4. Linking epidemiological data to show the study how disease outcomes following infection are determined by the pathogen or by host factors. Human cases caused by any of the isolates with known phenotype can be used to study the association between phenotype (e.g. adherence to the gut epithelium, and growth rate) and disease outcome in humans. The inclusion of outbreak strains facilitates this construction.

Finally, applying the outcomes of statistical associations (confirmed by reproducible data) to microbiological risk assessment, raises the following issues:

1. Integration. A systems biology approach is probably the best way forward to make a link from single scale in vitro testing to a multiple scale interpretation of effects.
2. Anchoring. Translation of molecular investigations to human health risk is still a challenge.
3. Communication. Current policy strategies (e.g. target setting) are based on serovar/serotype level research. This will only change when (statistically) validated model systems can be applied in the domain of public health food safety.
4. Technology. Although not fully expressed in this study, there is a need for development of 'open data' systems to support rapid progression. The increasing generation of high throughput data on a global scale needs a central (sentinel) organization to facilitate comparison of risk relevant outputs from individual investigations.

6. Concluding remarks

The bottleneck in the application of molecular tools for microbiological risk assessment has shifted from data acquisition (costs, time) to data analysis and systems approaches. The inability to draw systematic conclusions from this study to STEC in general, represents a bottleneck in the flow of large volumes of WGS data into food safety knowledge. This has, in more general terms, been described by Bromberg (2013). The paradigm change involves the translation of multidimensional information on genotype level (in the order of over 10^3 genes, 10^4 SNPs, etc.) via reduced information on phenotype level (in the order of 10^1 biologically relevant characteristics for MRA, like growth rate, survival, attachment to the gut epithelium, and acid tolerance) to a single measure of risk, such as the number of human cases of illness. Future computational assessments of genetic data should aim at solving this mapping problem without losing biologically relevant information for MRA. Not until then can risk assessors provide reliable answers from WGS data to the posed health questions of a policy maker in an accessible manner.

Acknowledgments

We are very grateful to Alex Bossers and Frank Harders (CVI, Wageningen University, The Netherlands) and Jim Bono (U.S. Meat Animal Research Center, USDA, USA) for performing the whole genome sequencing of the human and animal *E. coli* O157 strains.

The authors thank Eric Evers for his valuable comments during the preparation of this manuscript. This research is funded by the strategic program of the RIVM (SPR) (S/114001). With this program RIVM is contributing to the development of expertise and innovative research projects, to prepare RIVM for questions that may arise in the future. Gary Barker is supported by the Biotechnology and Biological Sciences Research Council (BB/J004529/1), UK.

References

Abee, T., van Schaik, W., Siezen, R.J., 2004. Impact of genomics on microbial food safety. Trends Biotechnol. 22, 653–660.

Abu-Ali, G.S., Ouellette, L.M., Henderson, S.T., Lacher, D.W., Riordan, J.T., Whittam, T.S., Manning, S.D., 2010. Increased adherence and expression of virulence genes in a lineage of *Escherichia coli* O157:H7 commonly associated with human infections. PLoS ONE 5, e10167.

Andersson, T., Nilsson, C., Kjellin, E., Toljander, J., Welinder-Olsson, C., Lindmark, H., 2011. Modeling gene associations for virulence classification of verocytotoxin-producing *E. coli* (VTEC) from patients and beef. Virulence 2, 41–53.

Bergholz, T.M., Vanaja, S.K., Whittam, T.S., 2009. Gene expression induced in *Escherichia coli* O157:H7 upon exposure to model apple juice. Appl. Environ. Microbiol. 75, 3542–3553.

Berk, P.A., 2008. In Vitro and In Vivo Virulence of *Salmonella* Typhimurium DT104: A Parallelogram Approach. Thesis Wageningen University, Wageningen, The Netherlands.

Besser, T.E., Shaikh, N., Holt, N.J., Tarr, P.I., Konkel, M.E., Malik-Kale, P., Walsh, C.W., Whittam, T.S., Bono, J.L., 2007. Greater diversity of Shiga toxin-encoding bacteriophage insertion sites among *Escherichia coli* O157:H7 isolates from cattle than in those from humans. Appl. Environ. Microbiol. 73, 671–679.

Bono, J.L., 2009. Genotyping *Escherichia coli* O157:H7 for its ability to cause disease in humans. Curr. Protoc. Microbiol. 14 (5A.3.1–5A.3.10).

Bromberg, Y., 2013. Building a genome analysis pipeline to predict disease risk and prevent disease. J. Mol. Biol. 425, 3993–4005.

Brul, S., Bassett, J., Cook, P., Kathariou, S., McClure, P., Jasti, P.R., Betts, R., 2012. Omics' technologies in quantitative microbial risk assessment. Trends Food Sci. Technol. 27, 12–24.

CAC (Codex Alimentarius Commission), 1999. Principles and Guidelines for the Conduct of Microbiological Risk Assessment.

Carriço, J.A., Sabat, A.J., Friedrich, A.W., Ramirez, M., 2013. Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. Euro Surveill. 18 (pii = 20382).

Chang, D.E., Smalley, D.J., Tucker, D.L., Leatham, M.P., Norris, W.E., Stevenson, S.J., Anderson, A.B., Grissom, J.E., Laux, D.C., Cohen, P.S., Conway, T., 2004. Carbon nutrition of *Escherichia coli* in the mouse intestine. Proc. Natl. Acad. Sci. U. S. A. 101, 7427–7432.

Dallman, T.J., Byrne, L., Launders, N., Glen, K., Grant, K.A., Jenkins, C., 2014. The utility and public health implications of PCR and whole genome sequencing for the detection and investigation of an outbreak of Shiga toxin-producing *Escherichia coli* serogroup O26:H11. Epidemiol. Infect. http://dx.doi.org/10.1017/S0950268814002696.

Delannoy, S., Beutin, L., Fach, P., 2013. Towards a molecular definition of enterohemorrhagic *Escherichia coli* (EHEC): detection of genes located on O island 57 as markers to distinguish EHEC from closely related enteropathogenic *E. coli* strains. J. Clin. Microbiol. 51, 1083–1088.

EFSA (European Food Safety Authority), 2014. Summary report on the use of Whole Genome Sequencing (WGS) of food-borne pathogens for public health protection. EFSA Scientific Colloquium 20 (Available online at: http//www.efsa.europa.eu).

Fabich, A.J., Jones, S.A., Chowdhury, F.Z., Cernosek, A., Anderson, A., Smalley, D., McHargue, J.W., Hightower, G.A., Smith, J.T., Autieri, S.M., Leatham, M.P., Lins, J.J., Allen, R.L., Laux, D.C., Cohen, P.S., Conway, T., 2008. Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine. Infect. Immun. 76, 1143–1152.

Falkow, S., 1988. Molecular Koch's postulates applied to microbial pathogenicity. Rev. Infect. Dis. 10, S274–S276.

Franz, E., van Hoek, A.H.A.M., Bouw, E., Aarts, H.J.M., 2011. *E. coli* O157 strain variability in manure-amended soil survival in relation to strain origin, virulence profile and carbon nutrition profile. Appl. Environ. Microbiol. 77, 8088–8096.

Franz, E., van Hoek, A.H.A.M., van der Wal, F.J., de Boer, A., Zwartkruis-Nahuis, A., van der Zwaluw, K., Aarts, H.J.M., Heuvelink, A.E., 2012. Genetic features differentiating

bovine, food and human isolates of Shiga toxin-producing *Escherichia coli* O157 in The Netherlands. J. Clin. Microbiol. 50, 772–780.
Franz, E., Delaquis, P., Morabito, S., Beutin, L., Gobius, K., Rasko, D.A., Bono, J., French, N., Osek, J., Lindstedt, B.-A., Muniesa, M., Manning, S., LeJeune, J., Callaway, T., Beatson, S., Eppinger, M., Dallman, T., Forbes, K.J., Aarts, H., Pearl, D.L., Gannon, V.P.J., Laing, C.R., Strachan, N.J.C., 2014. Exploiting the explosion of information associated with whole genome sequencing to tackle Shiga toxin-producing *Escherichia coli* (STEC) in global food production systems. Int. J. Food Microbiol. 87, 57–72.
Havelaar, A.H., Brul, S., de Jong, A., de Jonge, R., Zwietering, M.H., ter Kuile, B.H., 2010. Future challenges to microbial food safety. Int. J. Food Microbiol. 139, S79–S94.
Hayes, C.S., Koskiniemi, S., Ruhe, Z.C., Poole, S.J., Low, D.A., 2014. Mechanisms and biological roles of contact-dependent growth inhibition systems. In: Cossart, P., Maloy, S. (Eds.), Cold Spring Harbor Perspectives in Medicine Vol. 4. Cold Spring Harbor Laboratory Press, pp. 1–12.
Karmali, M.A., Mascarenhas, M., Shen, S., Ziebell, K., Johnson, S., Reid-Smith, R., Isaac-Renton, J., Clark, C., Rahn, K., Kaper, J.B., 2003. 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. 41, 4930–4940.
Kraakman, A.T.W., Niks, R.E., Van den Berg, P.M.M.M., Stam, P., Van Eeuwijk, F.A., 2004. Linkage disequilibrium mapping of yield and yield stability in modern spring Barley cultivars. Genetics 168, 435–446.
Kuehl, R.O., 2000. Design of Experiments: Statistical Principles of Research Design and Analysis. Second ed. Brooks/Cole, Pacific Grove, pp. 95–96.
Laing, C., Buchanan, C., Taboada, E.N., Zhang, Y., Kropinski, A., Villegas, A., Thomas, J.E., Gannon, V.P.J., 2010. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinf. 11, 461.
Lejeune, J.T., Abedon, S.T., Takemura, K., Christie, N.P., Sreevatsan, S., 2004. Human *Escherichia coli* O157:H7 genetic marker in isolates of bovine origin. Emerg. Infect. Dis. 10, 1482–1485.
Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.
Liu, K., Knabel, S.J., Dudley, E.G., 2009. *rhs* genes are potential markers for multilocus sequence typing of *Escherichia coli* O157:H7 strains. Appl. Environ. Microbiol. 75, 5853–5862.
Maiden, M.C.J., Jansen van Rensburg, M.J., Bray, J.E., Earle, S.G., Ford, S.A., Jolley, K.A., McCarthy, N.D., 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat. Rev. Microbiol. 11, 728–736.
Malosetti, M., van der Linden, C.G., Vosman, B., van Eeuwijk, F.A., 2007. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics 175, 879–889.
Maltby, R., Leatham-Jensen, M.P., Gibson, T., Cohen, P.S., Conway, T., 2013. Nutritional basis for colonization resistance by human commensal *Escherichia coli* strains HS and Nissle 1917 against *E. coli* O157:H7 in the mouse intestine. PLoS ONE 8, e53957.
Manning, S.D., Motiwala, A.S., Springman, A.C., Qi, W., Lacher, D.W., Ouellette, L.M., Mladonicky, J.M., Somsel, P., Rudrik, J.T., Dietrich, S.E., Zhang, W., Swaminathan, B., Alland, D., Whittam, T.S., 2008. Variation in virulence among clades of *Escherichia coli* O157:H7, associated with disease outbreaks. Proc. Natl. Acad. Sci. U. S. A. 105, 4868–4873.
Melton-Celsa, A.R., O'Brien, A.D., 2003. Animal models for STEC-mediated disease. Methods Mol. Med. 73, 291–305.
Nauta, M., van der Fels-Klerx, I., Havelaar, A., 2005. A poultry-processing model for quantitative microbiological risk assessment. Risk Anal. 25, 85–98.
Oh, K.-H., Cho, S.-H., 2014. Interaction between the quorum sensing and stringent response regulation systems in the enterohemorrhagic *Escherichia coli* O157:H7 EDL933 strain. J. Microbiol. Biotechnol. 24, 401–407.
Oliveira, M., Wijnands, L., Abadias, I., Aarts, H., Franz, E., 2011. Pathogenic potential of *Salmonella* Typhimurium DT104 following sequential passage through soil, packaged fresh-cut lettuce and a model gastrointestinal tract. Int. J. Food Microbiol. 148, 149–155.
Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLoS Genet. 2, e190.
Persson, S., Olsen, K.E.P., Ethelberg, S., Scheutz, F., 2007. Subtyping method for *Escherichia coli* Shiga toxin (Verocytotoxin) 2 variants and correlations to clinical manifestations. J. Clin. Microbiol. 45, 2020–2024.
Petersen, L., Bollback, J.P., Dimmic, M., Hubisz, M., Nielsen, R., 2007. Genes under positive selection in *Escherichia coli*. Genome Res. 17, 1336–1343.
Pielaat, A., Barker, G.C., Hendriksen, P., Hollman, P., Peijnenburg, A., Ter Kuile, B.H., 2013a. A foresight study on emerging technologies: state of the art of omics technologies and potential applications in food and feed safety. REPORT 1: review on the state of art of omics technologies in risk assessment related to food and feed safety. EFSA Supporting Publication EN-495, pp. 1–83.
Pielaat, A., Barker, G.C., Hendriksen, P., Hollman, P., Peijnenburg, A., Ter Kuile, B.H., 2013b. A foresight study on emerging technologies: state of the art of omics technologies and potential applications in food and feed safety. REPORT 2: application of omics to hazard and emerging risks identification and foresight on potential future applications of omics in risk assessment. EFSA Supporting Publication EN-495, pp. 84–126.
Pielaat, A., van Leusden, F.M., Wijnands, L.M., 2014. Microbiological risk from minimally processed packaged salads in the Dutch food chain. J. Food Prot. 77, 395–403.
Poole, S.J., Diner, E.J., Aoki, S.K., Braaten, B.A., t'Kint de Roodenbeke, C., Low, D.A., Hayes, C.S., 2011. Identification of functional toxin/immunity genes linked to contact-dependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems. PLoS Genet. 7, e1002217.
Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P., 2000. Association mapping in structured populations. Am. J. Hum. Genet. 67, 170–181.
Riordan, J.T., Viswanath, S.B., Manning, S.D., Whittam, T.S., 2008. Genetic differentiation of *Escherichia coli* O157:H7 clades associated with human disease by real-time PCR. J. Clin. Microbiol. 46, 2070–2073.
Romero-Barrios, P., Hempen, M., Messens, W., Stella, P., Hugas, M., 2013. Quantitative microbiological risk assessment (QMRA) of food-borne zoonoses at the European level. Food Control 29, 343–349.
Shaikh, N., Tarr, P.I., 2003. *Escherichia coli* O157:H7 Shiga toxin-encoding bacteriophages: Integrations, excisions, truncations, and evolutionary implications. J. Bacteriol. 185, 3596–3605.
Underwood, A.P., Dallman, T., Thomson, N.R., Williams, M., Harker, K., Perry, N., Adak, B., Willshaw, G., Cheasty, T., Green, J., Dougan, G., Parkhill, J., Wain, J., 2013. Public health value of next-generation DNA sequencing of enterohemorrhagic *Escherichia coli isolates* from an outbreak. J. Clin. Microbiol. 51, 232–237.
van Diemen, P.M., Dziva, F., Stevens, M.P., Wallis, T.S., 2005. Identification of enterohemorrhagic *Escherichia coli* O26:H-genes required for intestinal colonization in calves. Infect. Immun. 73, 1735–1743.
van Hoek, A.H.A.M., Aarts, H.J.M., Bouw, E., van Overbeek, W.M., Franz, E., 2012. The role of *rpoS* in *Escherichia coli* O157 in manure-amended soil survival and distribution of allelic variants among bovine, food and clinical isolates. FEMS Microbiol. Lett. 338, 18–23.
Xu, X., McAteer, S.P., Tree, J.J., Shaw, D.J., Wolfson, E.B.K., Beatson, S.A., Roe, A.J., Allison, L.J., Chase-Topping, M.E., Mahajan, A., Tozzoli, R., Woolhouse, M.E.J., Morabito, S., Gally, D.L., 2012. Lysogeny with Shiga toxin 2-encoding bacteriophages represses type III secretion in enterohemorrhagic *Escherichia coli*. PLoS Pathog. 8, e1002672.
Yang, Z., Kovar, J., Kim, J., Nietfeldt, J., Smith, D.R., Moxley, R.A., Olson, M.E., Fey, P.D., Benson, A.K., 2004. Identification of common subpopulations of non-sorbitol-fermenting, betaglucuronidase-negative *Escherichia coli* O157:H7 from bovine production environments and human clinical samples. Appl. Environ. Microbiol. 70, 6846–6854.
Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al., 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203–208.

Please cite this article as: Pielaat, A., et al., First step in using molecular data for microbial food safety risk assessment; hazard identification of *Escherichia coli* O157:H7..., Int. J. Food Microbiol. (2015), http://dx.doi.org/10.1016/j.ijfoodmicro.2015.04.009