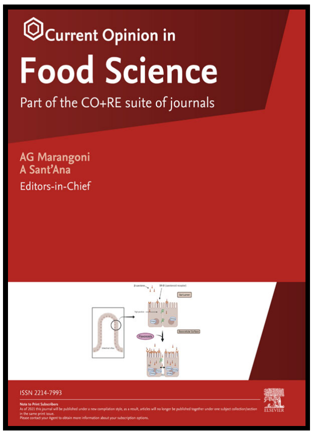

Taking account of genomics in quantitative microbial risk assessment: what methods? what issues?

Laurent Guillier, Federica Palma, Lena Fritsch

PII: S2214-7993(22)00124-2

DOI: https://doi.org/10.1016/j.cofs.2022.100922

Reference: COFS100922

To appear in: *Current Opinion in Food Science*

Please cite this article as: Laurent Guillier, Federica Palma and Lena Fritsch, Taking account of genomics in quantitative microbial risk assessment: what methods? what issues?, *Current Opinion in Food Science*, () doi:https://doi.org/10.1016/j.cofs.2022.100922

Title

Taking account of genomics in quantitative microbial risk assessment: what methods? what issues?

Authorship

Laurent Guillier[1], Federica Palma[2], Lena Fritsch[3]

Author names and affiliations

Laurent Guillier[1] (Laurent.GUILLIER@anses.fr)

Federica Palma[2] (federica.palma@pasteur.fr)

Lena Fritsch[3] (lena.fritsch@agroscope.admin.ch)

[1] ANSES – Université Paris-Est, French Agency for Food, Environmental and Occupational Health and Safety, Ris assessment Department, Maisons-Alfort, France

[2] Institut Pasteur - Université de Paris, Biological Resource Center of Institut Pasteur – CRBIP, Paris, France

[3] Agroscope, Research Division Food Microbial Systems, Bern, Switzerland

Corresponding author

Laurent Guillier (laurent.guillier@anses.fr)

CRediT author statement

Laurent Guillier: Conceptualization, Visualisation, Writing - Original Draft
Federica Palma: Writing- Reviewing and Editing
Lena Fritsch: Writing- Reviewing and Editing

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of interest

None

Abstract

The application of whole-genome sequencing (WGS) to the risk assessment of foodborne pathogens is a key challenge. WGS offers the highest level of strain discrimination for more precise hazard identification, hazard characterization and exposure assessment leading to deeper risk characterization.

Genome-wide association studies represent today powerful tools for the identification of associations between genomic elements and microbial phenotypic properties. Other cutting-edge tools include machine learning or statistical methods to characterize phenotype distribution on a phylogenetic tree. A panorama of the available methods is presented as well as the specific issues associated with the application of these methods to phenotypes of interest for risk assessment.

Keywords:

GWAS, quantitative microbial risk assessment, phenotypic marker

Highlights

- Numerous methods are available to identify markers of foodborne pathogen phenotypes,
- The complexity of acquiring phenotypes challenges the application of these methods,
- First applications of these methods in MRA show the concept of precision food safety.

1. Introduction

By the late 1990s, concepts of risk and risk assessment have been employed to inform decision-making for the management of food safety risks [1]. The framework for carrying out risk assessments of foodborne pathogens is well established and relies on four components including hazard identification, hazard characterization, exposure assessment and risk characterization [2].

A few years ago, novel scientific achievements demonstrated increasing evidence that food safety advances will require improved implementation of precision food safety approaches [3-5]. EFSA recently explored the role of whole-genome sequencing (WGS) and metagenomics to produce new information for food/feed risk assessments and which can contribute to better preparedness for risk management [6].

The characterization of uncertainty and variability is at the heart of the concerns of risk assessors in microbiological food safety [7]. Indeed, risk situations are often associated with strains presenting atypical characteristics (the most virulent strains, the most thermoresistant strains, etc.). Understanding the processes by which the foodborne pathogens adapt and evolve leading to these different phenotypes is, therefore, of major importance for establishing risk-based control measures. Beyond foodborne pathogens, this point also applies today to the attribution to antibiotic resistance [8].

The genetic determinants of these particular behaviours are often not identified. If these determinants are known, a simple search for their presence in strains is sufficient to predict their phenotype. If no markers are known, the only solution is to characterize the phenotypes of strains by experimentation.

Considering the need for an improved implementation of precision food safety and the availability of genomes and of the relevant methods for identifying markers of variability in pathogen behavior, the stage is now set for the incorporation of omics technologies into risk assessment [9]. This review aims to present the latest methodological advances for the identification of phenotype determinants and to identify the difficulties that need to be overcome in order to routinely use genomics in microbiological risk assessment of foods.

2. Methods for identifying markers of bacterial phenotypes

2.1. Requirements to identify markers of interest

The data required to identify genomic or phylogenetic markers for a qualitative or quantitative trait are the same in the three presented categories of methods (Figure 1A). The first step is to obtain a large collection of isolates. Although techniques such as the Ewens sampling formula [10] can be used to sample the diversity of a population, it is difficult to recommend a unique sample size [11]. The power of a method to identify causal variants or phylogeny clade associated to a phenotype is influenced by several other factors, including effect sizes, population structure, phenotype distribution, recombination rate [12, 13]. The diversity of the strains included in a study should be representative of the diversity of the pathogen or one of its sub-type of interest (e.g. ST, serovar) in the foods considered by the risk assessment. Another criterion to constitute the dataset is to consider the distribution of phenotypes for the trait of interest. The set of strains may be chosen in order to have a balanced distribution of phenotypes. This could improve the statistical power of the applied method to identify the markers. This criterion based on the balance of phenotypes may be in contradiction with the criterion of the representativeness of strains isolated from a food sector. Several sets of strains can thus be considered for different objectives: marker determination, validation (for these two sets, both genotypes and phenotypes will be available) and prediction based on the GWAS results. For this last set of strains, it is not necessary to have the phenotypes, only the genomes of strains representative of the food sector are necessary.

Then the acquisition of the genomic data using the recent developments in high-throughput sequencing technology and the phenotypic data of the whole set of bacterial isolates have to be carried out [14]. The genetic features that will be used by the marker identification methods must be determined. Four main types of genetic features have been used so far. A first type is any SNV (Single Nucleotide Variant) or small insertion/deletion found in the alignment of the set of genomes. Since this alignment usually focuses on the core regions shared by all genomes, testing only SNVs misses the identification of non-core markers. The second type of genomic feature that could be tested is therefore the matrix of presence or absence of the accessory genes [15]. Accessory genes are often acquired by lateral transfer between strains through mobile genetic elements bringing new trait combinations [16]. A different approach is to use the presence or absence in each of the genomes of short sequences of DNA also called k-mers. For each genome, the presence or absence of each unique k-mer is recorded. As this approach does not require an alignment, k-mers capture both core and accessory genome events. More recently, the concept of unitigs, which are compacted De Bruijn graphs of k-mers, has been proposed as the genetic feature for marker search items [17, 18]. As a unitig sequence is longer than a k-mer, its use in GWAS presents the advantage to reduce the redundancy present in k-mer counts and to generally easier interpretation of results [18]. MLST types, which are the consequence of the 4 mutations mentioned above, can also be used. Finally, a phylogeny that accounts for recombination can be determined generally based on of the alignment of core genome determinants.

The input data used by the three methods differ. In case of the GWAS (Genome Wide Association Study) approach, the input data includes (continuous or binary) phenotypes, genetic determinants and recombination-aware phylogenies (Figure 1B). Methods based on Machine learning (ML) require genetic determinants and phenotypes. Phylogenetic methods for markers identification rely on phenotypes and population structure-controlled phylogeny [14].

2.2 Genome-Wide Association Studies

GWAS is based on a simple principle: a genome-wide set of genetic variants in different individuals is statistically observed to see if any variant is causally associated with a trait. The phenotypic traits and the genomic sequences of a strain collection are therefore necessary to assess genetic variant candidates explaining the phenotypic trait of interest. Statistical tests can thus assess whether certain genetic elements are more frequent in strains with a specific trait than in those without [19]. GWAS can be made at different genomic levels. Most of the applications of bacterial GWAS have been carried out with SNVs, small insertions and deletions (INDELs), k-mers or differences in presence/absence of genes. More recently, unitigs have been proposed as the genetic feature for marker search items for the GWAS [17, 18].

An important challenge using GWAS is to take into account multiple statistical testing. To avoid high false-positive rates, a test correction must be performed considering the large number of tests being performed and to distinguish the significant results from those that will be observed by chance. Additional factors negatively affecting the performance of bacterial GWAS [20] are linkage disequilibrium (i.e., the nonrandom association of genetic alleles) and bacterial population structure (i.e., the presence of subpopulations that present large differences in the prevalence of both the allelic and phenotype frequencies). Feeding GWAS with a recombination-aware phylogenetic tree of the observed dataset helps to account for the population structure.

Nowadays, a multiplicity of methods and complete pipelines, e.g. Scorary or pyseer, to conduct microbial GWAS are freely accessible. Two recently published reviews [20, 21] present in detail the specificity of these tools based on their ability to consider categorical and continuous phenotypes, the core and accessory genomic features, and their strategy to handle issues related to multiple-testing, linkage disequilibrium and population structure.

2.3 Machine learning methods

By definition, machine learning methods rely on computer systems that are able to learn and adapt, by using algorithms and statistical methods to analyze and draw inferences from patterns in data. Numerous algorithms have been trained to recognize patterns in bacterial genomic data [22]. The general process of the identification of genomic markers is presented in Figure 1B. The dataset is split in a training and a testing dataset. The training is usually carried out with different algorithms which are then sorted based on their performance on the testing dataset [23]. After sufficient repetitions and selection of the algorithm, the trained machine can take as input the genome features of other strains to predict their phenotype.

The genomic features used in machine learning methods are the same as applied in GWAS to predict bacterial phenotypes. The k-mers based approaches are often preferred [24, 25]. But genes presence/absence could be used as well [26]. Recently, as for GWAS, unitigs have been used for machine-learning identification of phenotype markers [22]. An advantage of these methods is that they allow the relative importance of each input variable to be specified and thus reduce the size of the data to be considered. Moreover, they allow to predict the phenotype of strains for which only the genome is known. The combination of ML methods with GWAS has also been shown by several authors to be promising for prioritizing loci, though this application is still in its infancy [27, 28].

Contrary to GWAS, few dedicated software tools are available [22, 24]. Most ML methods rely on generic ML Python libraries or R packages and user-defined scripts [23, 29-31].

2.4 Phylogenetic methods

Phylogenetic methods do not account directly on genomic data but only through the phylogenetic trees derived from genomic data. Two classes of modelling approaches are using phylogeny. In the first, the evolution of the phenotype over time is modelled. In the second, the evolution of the phenotype distribution along the phylogenetic branches is modelled.

The first approach using phylogenies is to study the evolution of the phenotype with time. A researcher may be interested in whether, a bacterial phenotypic trait (e.g. lower temperature limit for growth) has evolved in association with environmental conditions (e.g. temperature) or in association with other traits (e.g. minimal pH limit). For solving this issue, Phylogenetic Comparative Methods (PCM) are an active field, that has shown many developments in the last few years [32]. Several methods have been specifically developed to study adaptive evolution. They rely on different models, such as Brownian motion or Ornstein–Uhlenbeck models that are implemented in R packages (see e.g. phytools, [33]; or PhylogeneticEM, [32]). All these modelling approaches have been mainly applied to eukaryotic organisms [34], and only a few articles described their application to prokaryotes yet [35].

In the second class of phylogenetic approach, statistical methods are proposed to test for association between the phenotypic trait and a fixed tree structure across all levels of the tree hierarchy. Two R packages have been developed, treeBreaker [36] and treeSeg [37]. Contrary to the first category of phylogenetic method, they consider the evolution of the phenotypic distribution itself rather than the phenotype.

3. The challenge of applying genomics in risk assessment

3.1 How to deal with continuous phenotypes in GWAS

The input parameters used for risk assessment present some particularities. They do not usually correspond with the single measurement of a simple phenotype under a single set of in vitro experimental conditions. The parameters of interest are more complex and require extensive data acquisition and a modelling step [see e.g. 38]. An illustrative example is the minimum growth temperature that is used in secondary models for assessing the growth of the pathogen in the exposure assessment step. About ten kinetics with about ten points per kinetic must be collected. Then, models have to be fitted to kinetics in order to retrieve the growth rate. Finally, another model has to be fitted in order to determine the minimum growth temperature. It thus requires a lot of experimental work and the estimated parameter used as the phenotypic trait is prone to uncertainty. So far very few studies have attempted to address the issue of performing marker searches on such parameters [39]. Most of the current applications of marker research methods are applied to simple phenotypes (e.g. minimum inhibitory concentration) of food pathogen behavior directly derived from experiments [40-42]. The same can be established for the characterisation of the hazard. The parameters of the dose-response relationship are not easily accessible. Their estimation is complex and may involve epidemiological data and exposure assessment [26, 43].

The most commonly studied phenotypes in risk assessment are quantitative, such as the probability of illness for one cell, the maximal growth rate, or the cardinal values. The first software used in bacterial GWAS or for phylogenetic methods concentrates on qualitative phenotypes. Dividing the continuous phenotypes of strains into well-defined categories is often tricky, even in a priori discriminatory conditions close to growth limits, due to minor differences between strains and experimental uncertainty [44]. In such a situation, a solution is to carry out the analysis on the most extreme phenotypes and to proceed with the identification of markers excluding strains that present phenotypes around the median, e.g. by excluding values between plus or minus one standard deviation around the median [42]. If the whole dataset is kept, hierarchical clustering could help to objectively define the phenotypic groups [40]. Software able to overcome issues are now available for taking into account the continuous nature of some phenotypes [18, 45].

3.2 Drawbacks associated with the nature of the phenotypic trait of interest

The phenotypes of interest for risk assessment in foods (e.g. temperature adaptation, ability to induce infection or to colonize animal reservoirs…) have a complex multifactorial nature as the adaptation of bacterial strains may involve different genes or metabolic functions. The experience shows that GWAS and ML methods could return several hundred to thousands of genetic elements associated with complex traits [46, 47]. Recent genome-wide association studies focusing on risk assessment phenotypes showed a higher number of candidate markers and lower statistical association values than association studies on microbial phenotypes of medical interest such as antibiotic resistance [40, 41]. This represents a challenge for GWAS methods, as it makes difficult to detect less prevalent adaptation mechanisms through simple statistical associations. Thus, it is complicated to identify the role of individual genes and look for (epistatic) interactions between them. While most applications of GWAS to date have used the single-locus testing framework, recent innovations seek to expand upon this paradigm to elucidate more complex genotype-phenotype associations. The introduction of the concept of unitigs in ML and GWAS methods may help to decipher complex association as unitigs considerably reduce the number of potential causal markers and improve interpretability. Even though the use of unitigs has practical implications for MRA, two main issues to be solved for the deployment of GWAS are anchoring and integration of results into biological systems approach for translating molecular studies into risk [48]. Taking into account the homoplasy (the occurrence of multiple independent mutations at the same site) in the identification of mutations is also expected to improve the association [49].

3.3 From identification of markers to their use in quantitative microbial risk assessment

The question of validation of the markers of the phenotypic traits is crucial for their use in risk assessment models. Meanwhile, the approach could be different according to the scientific fields [46]. For data analysts, markers are considered validated if robust statistical associations are proven. For laboratory-based researchers, validation is only considered valid when effects can be reproduced using complementary experimental approaches (see e.g. [50]). Experimental validation by reverse genetics of the many markers associated with phenotypes of interest for microbial risk assessment is probably not feasible, at least in the short term. It is likely that risk assessors will be satisfied with a statistical validation (a p-value below the corrected significance threshold) of markers.

Validation of phenotype predictions on strains other than those used for marker identification also remains an achievable goal, even if it does not provide functional genomics justifications to scientists.

An important issue for the application of genomics to describe phenotype variability in risk assessment models is also the ability of the methods to predict the phenotype of sequenced uncharacterized strains. Indeed, the collection of strains used to identify markers may represent either a fraction or a totally different set of strains for which prediction is needed. Here, ML methods present a clear advantage compared to most of GWAS or phylogenetic methods. The objective of these methods is to predict after the model is trained. The advantage of ML methods is that the uncertainty in their predictions can be easily incorporated into the uncertainty dimension of the QMRA models. Most GWAS and phylogenetic methods usually have a different objective, the central objective is mainly to establish an association between genomic features or clades in the phylogeny and the phenotype rather than to predict the unknown phenotype of a strain. Relative to that drawback recent development in GWAS was proposed by [18] where a penalized regression model completes the approach to predict the phenotype from the presence/absence of significant markers identified by GWAS. In their original phylogenetic approach, treeSeg, [37] proposes to predict the phenotype of a branch of the phylogeny.

Regarding QMRA models, a high number of sources of variability are modelled. Not all these sources have the same importance concerning the outcome of QMRA models [51], thus the implementation of biomarkers is only meaningful for the highest priority sources of microbiological variability according to uncertainty or sensitivity analysis methods. It’s worth to mention that hazard characterization for strain virulence is usually an important source of variability [44] and that there are high expectations for the application of these methods to improve the performance of risk assessments [5].

3.4 What can we expect in the coming years?

There are various strategies for the application of genomics in quantitative risk assessment in the longer and shorter term. In the short term, genomics can be used to remove source of uncertainty in risk assessment. Strains’ variability will be better grasped and modelled by identifying molecular markers of adaptation (in connection with predictive microbiology) or virulence markers (for the parameters of the dose-response relationship) and their identification in a collection of strains representative of a food sector. This better understanding of the intraspecific variability of the strains will, in particular, make it possible to test the commonly used hypothesis that the variability observed in laboratory strains is the same as that the one of strains present in the food production chain. In addition to reducing uncertainty, the use of genomics paves the way for easier validation of QMRA models by comparing the genomic diversity in patients predicted by the model with that observed by the epidemiologists. In the longer term, it is conceivable that genomic markers could be used to establish management measures better adapted to the different potentials of the strains [48]. But this implementation requires the development of rapid microbiological methods for the identification of markers on isolated clones or the systematic sequencing of strains. The standardization and the validation of microbiological method is a lengthy process and systematic sequencing of strains is difficult to envisage given that many labs cannot afford the costs and time required to obtain genomic sequence information yet.

4. Conclusion

The application of GWAS, Machine-Learning and phylogenetic methods will probably considerably improve the identification of relevant markers of phenotypes of interest for foodborne pathogens in the next few years. Despite further efforts of biologists and computer scientists being needed to improve and validate comparative genomic methods, several publications have successfully used genomics in risk assessment [52, 53].

Although the degree of correlation between genotypic and phenotypic profiles still shows some uncertainty, genomics has a clear potential to improve model predictions, allow a link between QMRA and epidemiological observations and pave the way for precision food safety. The application of the three methods here presented is straightforward and easy when considering a phenotypic trait measured in a specific condition. Beyond the tools, the sharing of genomic data by risk assessment bodies is being achieved [54, 55]. In this context, the bottleneck in the application of genomics for microbiological risk assessment is no longer the acquisition of genomic data or their analysis. Today, the difficulty lies more in acquiring the parameters of risk assessment models on a large number of strains than in sequencing or determining phenotype markers. One solution would be to set up ambitious research projects that would allow the characterization of these phenotypes at high throughput. Another possible solution is data sharing between the scientific actors of the predictive microbiology community. The application of standardized methods for experimental data acquisition and model fitting would be essential for a full exploitation of phenotype and identification of shared markers.

Acknowledgments

This work was supported by funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 773830: One Health European Joint Programme (project CARE: Cross-sectoral framework for quality Assurance Resources for countries in the European Union). The research leading to this article has also received funding from the French National Research Agency (Opticold Project, ANR-15-CE21-0011-02).

References

1. LeJeune JT, Zhou K, Kopko C, Igarashi H: **FAO/WHO Joint Expert Meeting on Microbiological Risk Assessment (JEMRA): Twenty Years of International Microbiological Risk Assessment**. *Foods* 2021, **10**:1873.

2. FAO, WHO: *Microbiological Risk Assessment Guidelines for food*. FAO and WHO; 2021.

3. Kovac J, Bakker H den, Carroll LM, Wiedmann M: **Precision food safety: A systems approach to food safety facilitated by genomics tools**. *TrAC Trends in Analytical Chemistry* 2017, **96**:52–61.

4. Besten HMW den, Amézquita A, Bover-Cid S, Dagnas S, Ellouze M, Guillou S, Nychas G, OMahony C, Pérez-Rodriguez F, Membré J-M: **Next generation of**

microbiological risk assessment: Potential of omics data for exposure assessment. *International Journal of Food Microbiology* 2018, **287**:18–27.

5. Haddad N, Johnson N, Kathariou S, Métris A, Phister T, Pielaat A, Tassou C, Wells-Bennik MHJ, Zwietering MH: **Next generation microbiological risk assessment Potential of omics data for hazard characterisation**. *International Journal of Food Microbiology* 2018, **287**:28–39.

6. Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, Chemaly M, Davies R, Cesare AD, Hilbert F, Lindqvist R, et al. **Whole genome sequencing and metagenomics for outbreak investigation source attribution and risk assessment of food-borne microorganisms**. *EFSA Journal* 2019, **17**.

7. Pouillot R, Guillier L: **Understanding Uncertainty and Variability in Risk Assessment**. In *Risk Assessment Methods for Biological and Chemical Hazards in Food*. . CRC Press; 2020:165–190.

8. Collineau L, Boerlin P, Carson CA, Chapman B, Fazil A, Hetman B, McEwen SA, Parmley EJ, Reid-Smith RJ, Taboada E, Smith BA: **Integrating whole-genome sequencing data into quantitative risk assessment of foodborne antimicrobial resistance: a review of opportunities and challenges**. *Front Microbiol* 2019, **10**:1107.

9. Pasquali F, Remondini D, Snary EL, Hald T, Guillier L: **Editorial: Integrating Whole Genome Sequencing Into Source Attribution and Risk Assessment of Foodborne Bacterial Pathogens.** *Front Microbiol* 2021, **12**:795098.

10. Babu GJ, Manstavicius E: Random permutations and the Ewens sampling formula in genetics. *Probability Theory and Mathematical Statistics* 2020, 33-42.

11. Lees J: **The background of bacterial GWAS**. Figshare 2017, https://figshare.com/articles/thesis/The_background_of_bacterial_GWAS/5550037 1-30.

12. Saber MM, Shapiro BJ: **Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes**. *Microbial genomics* 2020, **6(**3).

13. Saund K, Lapp Z, Thiede SN, Pirani A, Snitkin ES: **prewas: data pre-processing for more informative bacterial GWAS**. *Microbial genomics* 2020, **6**(5).

14. Didelot X: **Phylogenetic Methods for Genome-Wide Association Studies in Bacteria.** *Methods Mol Biol* 2021, **2242**:205–220.

15. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V: **Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary.** *Genome Biol* 2016, **17**:238.

16. Innamorati KA, Earl JP, Aggarwal SD, Ehrlich GD, Hiller NL: **The bacterial guide to designing a diversified gene portfolio**. *The Pangenome* 2020, 51-87.

17. Jaillard M, Lima L, Tournoud M, Mahé P, van BA, Lacroix V, Jacob L: **A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events.** *PLoS Genet* 2018, **14**:e1007758.

18. Lees JA, Mai TT, Galardini M, Wheeler NE, Horsfield ST, Parkhill J, Corander J: **Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions**. *mBio* 2020, **11**.

19. Falush D: **Bacterial genomics: Microbial GWAS coming of age.** *Nat Microbiol* 2016, **1**:16059.

20. Allen JP, Snitkin E, Pincus NB, Hauser AR: **Forest and Trees: Exploring Bacterial Virulence with Genome-wide Association Studies and Machine Learning**. *Trends in Microbiology* 2021, **29**:621–633.

21. San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, Fonseca V, Mogaka J, Power R, Oliveira T de: **Current Affairs of Microbial Genome-Wide Association Studies: Approaches Bottlenecks and Analytical Pitfalls**. *Frontiers in Microbiology* 2020, **10**.

22. Jaillard M, Palmieri M, van BA, Mahé P: **Interpreting k-mer-based signatures for antibiotic resistance prediction.** *Gigascience* 2020, **9**.

23. Njage PMK, Leekitcharoenphon P, Hald T: **Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: Predicting clinical outcomes in shigatoxigenic *Escherichia coli*.** *Int J Food Microbiol* 2019, **292**:72–82.

24. Drouin A, Letarte G, Raymond F, Marchand M, Corbeil J, Laviolette F: **Interpretable genotype-to-phenotype classifiers with performance guarantees.** *Sci Rep* 2019, **9**:4071.

25. Hicks AL, Wheeler N, Sánchez-Busó L, Rakeman JL, Harris SR, Grad YH: **Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data.** *PLoS Comput Biol* 2019, **15**:e1007349.

26. Njage PMK, Henri C, Leekitcharoenphon P, Mistou MY, Hendriksen RS, Hald T: **Machine Learning Methods as a Tool for Predicting Risk of Illness Applying Next-Generation Sequencing Data.** *Risk Anal* 2019, **39**:1397–1413.

27. Buckley SJ, Harvey RJ (2021). **Lessons learnt from using the machine learning random forest algorithm to predict virulence in *Streptococcus pyogenes***. *Frontiers Cell Inf Microbiol*, 2021, **1353**.

28. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP: **Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci**. *Front Genetics* 2020, **11**:350.

29. Karlsen ST, Vesth TC, Oregaard G, Poulsen VK, Lund O, Henderson G, Bælum J: **Machine learning predicts and provides insights into milk acidification rates of Lactococcus lactis.**. *PLoS One* 2021, **16**:e0246287.

30. Wheeler NE, Gardner PP, Barquist L: **Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*.** *PLoS Genet* 2018, **14**:e1007333.

31. Lupolova N, Dallman TJ, Holden NJ, Gally DL: **Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*.** *Microb Genom* 2017, **3**:e000135.

32. Bastide P, Ané C, Robin S, Mariadassou M: **Inference of Adaptive Shifts for Multivariate Correlated Traits**. *Systematic Biology* 2018, **67**:662–680.

33. Revell LJ: **A variable-rate quantitative trait evolution model using penalized-likelihood.** *PeerJ* 2021, **9**:e11997.

34. Lajoie G, Kembel SW: **Making the Most of Trait-Based Approaches for Microbial Ecology**. *Trends in Microbiology* 2019, **27**:814–823.

35. Krause S, Bodegom PM van, Cornwell WK, Bodelier PLE: **Weak phylogenetic signal in physiological traits of methane-oxidizing bacteria**. *Journal of Evolutionary Biology* 2014, **27**:1240–1247.

36. Ansari MA, Didelot X: **Bayesian inference of the evolution of a phenotype distribution on a phylogenetic tree**. *Genetics* 2016, **204**:89–98.

37. Behr M, Ansari MA, Munk A, Holmes C: **Testing for dependence on tree structures.** *Proc Natl Acad Sci U S A* 2020, **117**:9787–9792.

38. Koukou I, Mejlholm O, Dalgaard P: Cardinal parameter growth and growth boundary model for non-proteolytic *Clostridium botulinum*–Effect of eight environmental factors. *Int J Food Microbiol* 2021, **346**: 109162.

39. Liang Y, Li B, Zhang Q, Zhang S, He X, Jiang L, Jin Y: **Interaction analyses based on growth parameters of GWAS between *Escherichia coli* and *Staphylococcus aureus*.** *AMB Express* 2021, **11**:34.

40. Lee B-H, Cole S, Badel-Berchoux S, Guillier L, Felix B, Krezdorn N, Hébraud M, Bernardi T, Sultan I, Piveteau P: **Biofilm Formation of *Listeria monocytogenes* Strains Under Food Processing Environments and Pan-Genome-Wide Association Study**. *Frontiers in Microbiology* 2019, **10**.

41. Fritsch L, Felten A, Palma F, Mariet J-F, Radomski N, Mistou M-Y, Augustin J-C, Guillier L: **Insights from genome-wide approaches to identify variants associated to phenotypes at pan-genome scale: Application to *L. monocytogenes* ability to grow in cold conditions**. *International Journal of Food Microbiology* 2019, **291**:181–188.

42. Hingston P, Chen J, Dhillon BK, Laing C, Bertelli C, Gannon V, Tasara T, Allen K, Brinkman FS, Truelstrup HL, et al.: **Genotypes Associated with *Listeria monocytogenes* Isolates Displaying Impaired or Enhanced Tolerances to Cold, Salt, Acid, or Desiccation Stress.**. *Front Microbiol* 2017, **8**:369.

43. Pouillot R, Hoelzer K, Chen Y, Dennis SB: ***Listeria monocytogenes* dose response revisited—incorporating adjustments for variability in strain virulence and host susceptibility**. *Risk Analysis* 2015, **35:** 90-108.

44. Kuijpers AF, Bonacic Marinovic AA, Wijnands LM, Delfgou-van Asch, EH, van Hoek, AH, Franz E, Pielaat, A. **Phenotypic prediction: linking in vitro virulence to the genomics of 59 *Salmonella enterica* strains**. *Frontiers in microbiology* 2019, **9**:3182.

45. Collins C, Didelot X: **A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination.** *PLoS Comput Biol* 2018, **14**:e1005958.

46. Kobras CM, Fenton AK, Sheppard SK: **Next-generation microbiology: from comparative genomics to gene function**. *Genome Biology* 2021, **22**.

47. Vila Nova M, Durimel K, La K, Felten A, Bessières P, Mistou MY, Mariadassou M, Radomski N. **Genetic and metabolic signatures of *Salmonella enterica subsp. enterica* associated with animal sources at the pangenomic scale.** *BMC Genomics* 2019, **20**:814.

48. Pielaat A, Boer MP, Wijnands LM, van Hoek AH, Bouw E, Barker GC, Teunis FFM, Aarts, HJM, Franz E: **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, **213**:130-138.

49. Lai YP, Ioerger TR: **Exploiting Homoplasy in Genome-Wide Association Studies to Enhance Identification of Antibiotic-Resistance Mutations in Bacterial Genomes.** *Evol Bioinform Online* 2020, **16**:1176934320944932.

50. Hwang W, Yong JH, Min KB, Lee KM, Pascoe B, Sheppard SK, Yoon SS: **Genome-wide association study of signature genetic alterations among *Pseudomonas aeruginosa* cystic fibrosis isolates.**. *PLoS Pathog* 2021, **17**:e1009681.

51. Ellouze M, Gauchi JP, Augustin JC: **Global sensitivity analysis applied to a contamination assessment model of *Listeria monocytogenes* in cold smoked salmon at consumption**. *Risk Analysis* 2010 **30**:841-852.

52. Fritsch L, Guillier L, Augustin J-C: **Next generation quantitative microbiological risk assessment: refinement of the cold smoked salmon-related listeriosis risk model by integrating genomic data**. *Microbial Risk Analysis* 2018, **10**:20–27.

53. Njage PMK, Leekitcharoenphon P, Hansen LT, Hendriksen RS, Faes C, Aerts M, Hald T: **Quantitative Microbial Risk Assessment Based on Whole Genome Sequencing Data: Case of *Listeria monocytogenes***. *Microorganisms* 2020, **8**.

54. European Centre for Disease Control, European Food Safety Authority: **EFSA and ECDC technical report on the collection and analysis of whole genome sequencing data from food-borne pathogens and other relevant microorganisms isolated from human, animal, food, feed and food/feed environmental samples in the joint ECDC-EFSA molecular typing database**. *Efsa supp publication*, 16: 1337E.

55. Stevens EL, Carleton HA, Beal J, Tillman GE, Lindsey RL, Lauer AC, ... & Braden, C: **Use of Whole Genome Sequencing by the Federal Interagency Collaboration for Genomics for Food and Feed Safety in the United States**. *Journal of Food Protection* 2022, **85:**755-772.

Annotation

Papers of particular interest, published within the period of review, have been highlighted as:
* of special interest
** of outstanding interest

- **Allen *et al*. 2021 [20]. This review presents recent advances in comparative genomic approaches to identify bacterial virulence determinants, with a focus on GWAS and ML. The list of methods identified is very large and the concept are very well presented.
- **San *et al.* 2020 [21]: This review provides an exhaustive review of the prominent tools in GWAS. It thoroughly discuss pitfalls and bottlenecks and provide insights into the selection of appropriate tools.
- *Didelot, 2021 [14] presented in this reference the practical aspects for carrying out bacterial genome-wide association studies.
- *Njage *et al*., 2021 [53]. This article provides a convincing application of genomics in quantitative microbial risk assessment.

Figure caption

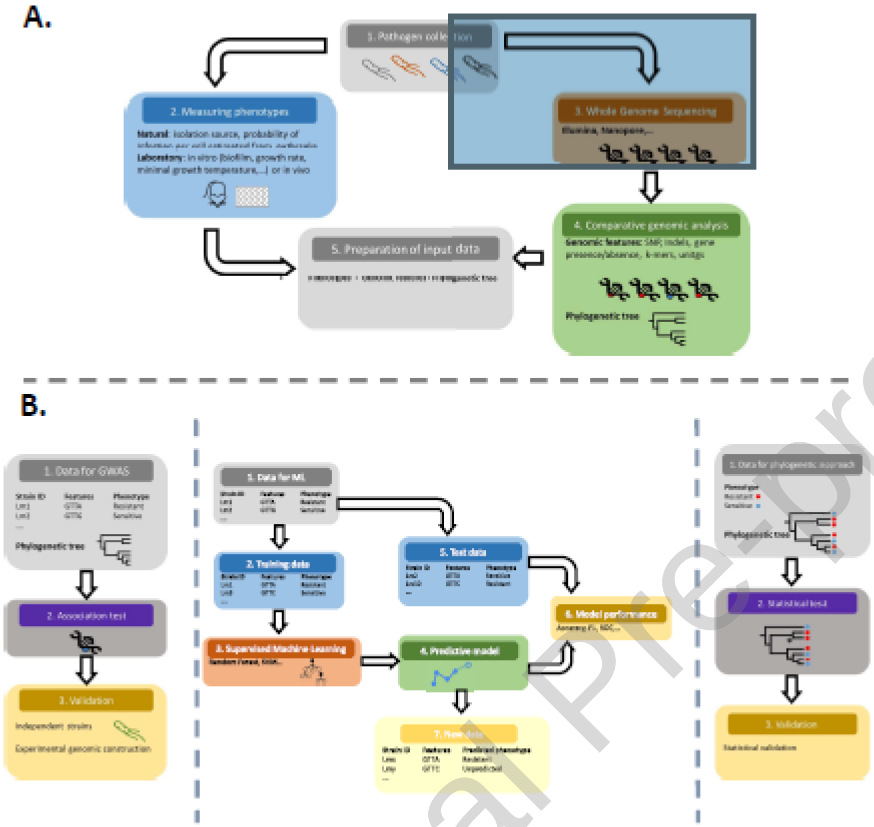

Figure 1. Dataset preparation (A) and general approaches (B) for identifying markers of phenotypes by using GWAS, machine learning or phylogenetic approaches.

Conflict of interest

The authors declare no conflict of interest