SLUTRAPPORT/FINAL REPORT
NR. 2024-193

Pimlapas Leekitcharoenphon:

Bedre fødevaresikkerhed med helgenomsekventering og maskinlæring

Improving food safety using WGS and machine learning

Mejeribrugets
ForskningsFond

December 2024

Final report

for collaborative projects funded via the Danish Dairy Research Foundation (DDRF)

1. Title of the project

Danish: Bedre fødevaresikkerhed med helgenomsekventering og maskinlæring

English: Improving food safety using WGS and machine learning

2. Project manager

Pimlapas Leekitcharoenphon, DTU Food, Technical University of Denmark, Henrik Dams Allé, Bygning 202, 2800 Kgs Lyngby, +45 35 88 71 83, pile@food.dtu.dk

3. Other project staff

Patrick Murigu Kamau Njage, DTU Food, Technical University of Denmark, panj@food.dtu.dk

Lisbeth Truelstrup Hansen, DTU Food, Technical University of Denmark, litr@food.dtu.dk

Yinghua Xiao, Arla Foods, yixia@arlafoods.com

4. Sources of funding

MFF, Karl Pedersen og Hustrus Industrifond, and Arla Foods

5. Project period

Project period with DDRF funding:	January 2020 to December 2022
Revised, if necessary:	-
Total project period, if sub-project within a larger project:	[January 2020 to December 2024]
Revised, if necessary:	-

6. Project summary

Danish:

Identifikation af bakterier er afgørende, da forskellige beslutninger vil blive truffet såfremt en fødevarepatogen påvises på slægts- eller artsniveau. Imidlertid er de identifikationsteknikker (f.eks. konventionel typning) der anvendes i dag, tidskrævende, og endnu vigtigere mangler opløsning til at skelne imellem stammer af samme art. En sådan mangfoldighed på stammeniveau kan påvirke erkendelsen af risikoen og dermed beslutningen om de fremadrettede handlinger. Disse begrænsninger kan løses ved hjælp af WGS-teknologi (helgenomsekventering), men mangel på genetisk indsigt og adgang til avancerede computerfaciliteter er en vigtig hindring for mejeriselskaber i at anvende WGS-teknologi i

praksis. Et smart værktøj til at forudsige virulensniveauet og modstand mod desinfektionsmidler er nødvendigt for udviklingen af et sikkerheds- og kvalitetssikringssystem i den tidlige produktionsfase.

I betragtning af betydningen af *Listeria monocytogenes* i mejeriindustrien, sigter dette projekt mod at indarbejde WGS-data fra mere end 1600 kliniske isolationer og mejeriprodukter i machine learning værktøj for at identificere virulensniveauet og desinfektionsresistensen af *L. monocytogenes* stammer i realtid. De genetikbaserede resultater vil blive valideret eksperimentelt. Med det beskrevede machine learning værktøj behøver de industrielle brugere kun at uploade rå datasekvenser til serveren, som er hostet på DTU via et frit tilgængeligt webbaseret interface, og derefter modtage et letforståeligt resultat i løbet få minutter pr. isolat. Dette projekt vil forbedre og fremskynde beslutningsprocessen og fødevaresikkerhedshåndteringen af mejeriprodukter.

English:

Identification of bacteria is crucial, as different measures and actions will be taken whether food pathogens are detected, mostly at genus or species level. However, the identification techniques used today are time-consuming, and more importantly lack the resolution to differentiate between strains of the same species. Such strain-level diversity could affect the recognition of the risk level and decision of consequent actions. Certainly, these limitations can be resolved by whole-genome sequencing (WGS) technology. The lack of genetic insight and access to advanced computing facilities is a major hurdle for dairy companies to apply WGS technology in practice. A smart tool to predict virulence level and resistance to disinfectants is needed for frontline safety/quality assurance practice.

Considering the importance of *Listeria monocytogenes* to the dairy industry, this project aims to incorporate WGS data of more than 1600 clinical and dairy isolates in machine-learning predictors to identify the virulence level and disinfectant-resistance of *L. monocytogenes* strains in real-time. The genetics-based results have been validated. With the developed tool, the industrial users only need to upload raw read sequences data to the server hosted at DTU via a freely accessible web-based interface, in order to receive easily interpretable output within a few minutes. This project can improve and speed up the decision-making and food safety management of the dairy industry.

7. Project aim

Danish:

Hovedformålet med projektet er at bruge WGS og maskinlæring (ML) til at forbedre fødevaresikkerhedsstyringen af *L. monocytogenes* med følgende delmål:

- At bestemme prædiktorerne for virulensniveauet af *L. monocytogenes* i realtid ved hjælp af WGS og ML.
- At identificere markører for desinfektionsresistens i *L. monocytogenes* ved hjælp af WGS og ML.
- At konstruere en hurtig og brugervenlig genomisk pipeline (inkl. værtsserver på DTU og webbaseret interface) til påvisning af *L. monocytogenes* virulensniveau og desinfektionsresistens.

English:

The main objective of the project was to use WGS and machine learning (ML) to improve food safety management of *L. monocytogenes* with the following sub-objectives:

- To determine the predictors of virulence level of *L. monocytogenes* in real-time using WGS and ML.
- To identify markers of disinfectant-resistance in *L. monocytogenes* using WGS and ML.
- To construct a rapid and user-friendly genomic pipeline (incl. host server at DTU and web-based interface) for detecting the virulence level and disinfectant-resistance of *L. monocytogenes.*

8. Background for the project

Currently, the dairy companies are relying on conventional microbiological methods to manage food safety at their sites. This has been the standard for the past decades and enables dairies to detect and identify potential pathogenic bacteria in the environment and products. Identification of bacteria is crucial, as different measures and actions will be taken whether *L. monocytogenes*, *Salmonella*, etc. is detected, mostly at genus or species level. More and more studies in microbial physiology have demonstrated the diversity of various strains in virulence, stress tolerance, damage repair, which could affect the recognition of the risk level and decision of consequent actions. However, the identification techniques used today are time-consuming, and more importantly lack the resolution to differentiate between strains of the same species. Certainly, these limitations can be resolved by WGS technology.

With the extreme and continuous drop in costs of DNA sequencing (one complete bacterial genome can be sequenced for less than 100 EUR), analytical microbiology has entered a new era; the era of genomes. It is clear that the dairy industry, heavily fighting food pathogens and spoilers in order to successfully ensure product safety and quality, needs to harness and understand WGS technology. Within the next decade, WGS will gradually substitute other microbiological tests and become the core methodology in routine testing and trouble shooting. This development will result in faster and more precise identification than seen before. In addition to this, it will generate massive amounts of sequencing data that represents a treasure box of extra information about the microorganisms. Today, many features of bacteria, such as resistance to antibiotics and virulence, can be predicted by their genetic elements for experts within the field. Meanwhile, the lack of genetic insight and access to advanced computing facilities is a major hurdle for the employees of dairy companies in safety assurance to apply WGS technology in practice.

L. monocytogenes is the most well-known food pathogen and constantly challenges the current quality assurance systems of dairy companies. This project established and delivered a smart tool for frontline QA to predict virulence level and resistance to disinfectants, where the virulence part as a PhD project was funded by Karl Pedersen og Hustrus Industrifond. A predictor of the resistance level to disinfectants used at dairies is highly relevant, as it will directly facilitate the site to choose disinfectants depending on the specific bacterial isolate from the environment or products. The predictor is accessible online to all dairies, big and small. By simply processing the raw sequence datasets (locally generated or received from outsourcing laboratory) using standard computers, the quality manager at the site can obtain an easy-to-interpret output for making decisions.

9. Sub-activities in the entire project period

This project was conducted by 1 PhD student and 1 Postdoc and consists of the following work packages (Grantt chart).

WP1: Strain collection and characterization

L. monocytogenes genomes with known illness level from an exhaustive and epidemiologically balanced surveillance from our networks from French and Danish national surveillance were collected. Disinfectant- resistant/susceptible isolates from our networks were collected in this project. Lab experiments to test susceptibility of the isolates were performed. A total of more than 1600 isolates were used as training/testing sets in the ML model development.

- Deliverable 1.1: More than 1600 *L. monocytogenes* isolates

- Deliverable 1.2: Susceptibility results

- Milestone 1: Collection of isolates for machine learning model development and validation

WP2: Whole genome sequencing

The collection of the 1649 *L. monocytogenes* isolates was WGS using an Illumina HiSeq. The quality of the raw read sequences was assessed by DTU QC pipeline and *de novo* assembled using the SPAdes program. The assembled genomes (contigs) were further analyzed in WP3.

- **Deliverable 2.1**: WGS of 1649 isolates
- **Deliverable 2.2**: Quality checked raw reads and assembled genomes
- **Milestone 2**: High quality genomic data for machine learning development and validation

WP3: Pan-genome analysis and sequence alignment

Assembled genomes from all isolates were annotated using Prokka. The annotated genomes were used in Roary pipeline for pan-genome analysis. The set of all genes (pan-genome) from all isolates were received from Roary result. The set of pan-genome was used for alignment against all of the *L. monocytogenes* genomes to create a gene profile with percent similarity. The gene profile was used as input for machine learning.

- **Deliverable 3.1**: Annotated genomes
- **Deliverable 3.2**: Set of pan-genome (all genes) sequences
- **Deliverable 3.3**: Gene profile with percent identity
- **Milestone 3**: Gene profile of percent identity against all isolates

WP4: Machine learning

The gene profile data was partitioned into training, testing and validation datasets and ran through available ensemble classification techniques, which was followed by model evaluation and selection. The two best models was selected to predict the virulence/illness level and predict disinfectant-resistance of unknown *L. monocytogenes* genomes.

- **Deliverable 4.1**: Machine learning models
- **Deliverable 4.2**: The best predictive machine learning models
- **Milestone 4**: Predictive ML models for identification of virulence level and disinfectant-resistance

WP5: Validation

Public genomes of *L. monocytogenes* from previous studies on virulence/illness level as well as outbreaks and disinfectant resistance were used for validation of virulence prediction model. Benchmarking of different sequencing platform and type of sequence input were assessed and compared using the developed ML models.

-**Deliverable 5**: Validation and benchmarking results of the predicted virulence and predicted disinfectant-resistance

- **Milestone 5**: Improving confidence of the ML models

WP6: Web-based tool construction

The ML pipeline was constructed as a web-based tool called 'LisPred' (http://genepi.food.dtu.dk/listpred) with Arla as an end user to test and validate how user-friendly the tool was. The tool has been maintained and updated by our Center for Genomic Epidemiology (CGE) at DTU, who have introduced and maintained several globally open access genomic sequence data analysis tools. The web-based tool is platform independent, needs minimal computer resources and is usable regardless of location. The tool has a user-friendly interface. A user can upload raw read or assembled genomes to the tool. The tool will report simple output with virulence level and disinfectant-resistance of an unknown *L. monocytogenes* genome within a few minutes per genome. The submitted genome is confidential and deleted after analysis. The LisPred also has stand-alone version in GitHub (https://github.com/genomicepidemiology/ListPred/tree/master/workflow) and Docker (https://hub.docker.com/r/genomicepidemiology/listpred).

- **Deliverable 6.1**: Pipeline for identification of virulence level and disinfectant-resistance
- **Deliverable 6.2**: Web-based tool and stand-alone version for identification of virulence level and disinfectant-resistance
- **Milestone 6**: Genomic tool for rapid identification of virulence/illness level and disinfectant-resistance of *L. monocytogenes*

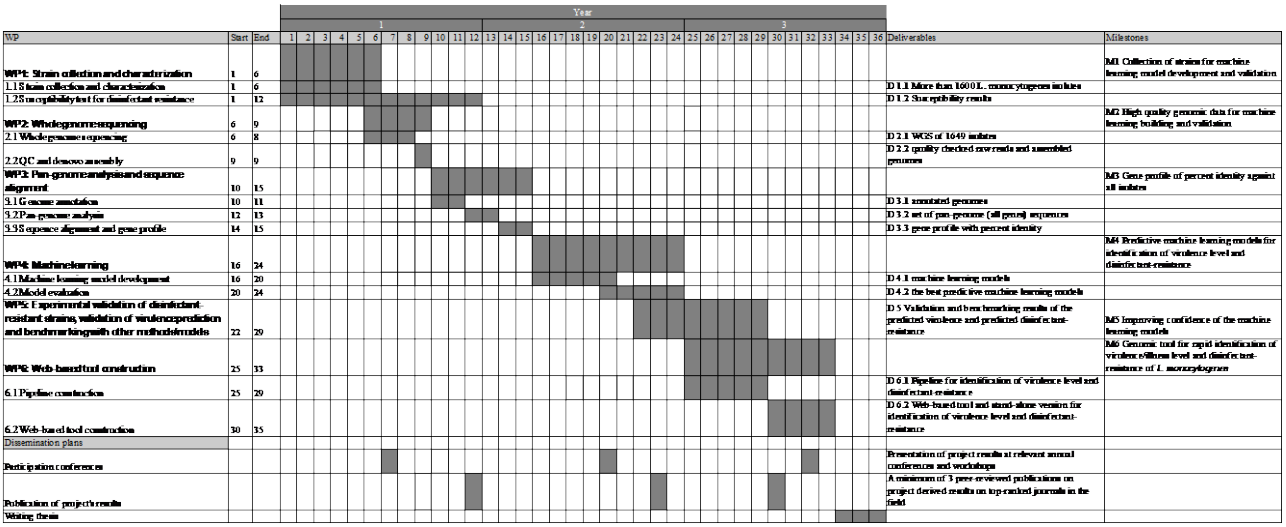


Table 1. Project Grant chart

10. Deviations

The entire project got one-year extension until the end of 2024. All lab and computer experiments and activities were completed within three years. The last year (2024), there was a PhD defense, submitting manuscripts to peer-review journals and revised the manuscripts according to reviewers' comments.

11. Project results

The development of predictive ML models for virulence potential of *L. monocytogenes* using WGS surveillance data

Contamination with food-borne pathogens, such as *Listeria monocytogenes*, remains a big concern for food safety. Hence, rigorous and continuous microbial surveillance is a standard procedure. At this point, however, the food industry and authorities only focus on detection of *Listeria monocytogenes* without characterization of individual strains into groups of more or less concern. As whole genome sequencing (WGS) gains increasing interest in the industry, this methodology presents an opportunity to obtain finer resolution of microbial traits such as virulence. Within this project, we therefore aimed to explore the use of WGS in combination with Machine Learning (ML) to predict *L. monocytogenes* virulence potential on a sub-species level.

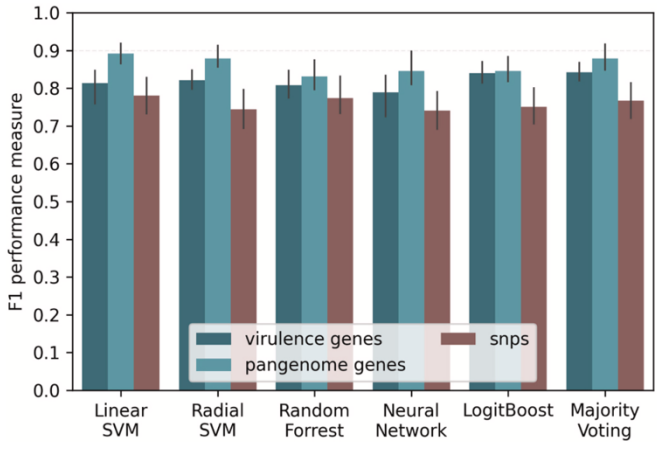
The 169 Whole Genome Sequencing isolates for this study were obtained from national surveillance programs in France (n = 60) and Denmark (n = 109). The genomic data from France were collected from a previous study [1]. Both the French and Danish surveillance systems exhaustively collect all *L. monocytogenes* isolates found in the clinic and food. Since the systems are strictly controlled, we expect them to give a good overview of their respective country's epidemiological *L. monocytogenes* landscape. To this date, only a few countries have implemented similar WGS surveillance systems for *L. monocytogenes*. Hence, the availability of epidemiologically exhaustive high-quality data remains sparse. Publicly available databases store a massive amount of sequencing data that could be used. However, it is important to remember that the deposition of sequencing data is often biased and might not be fully representative of the epidemiological

landscape. For our study, we used WGS data collected through the GenomeTrakr network [2], which can be accessed through NCBI's Pathogen Detection Portal [3]. Even though different international institutions are contributing to the GenomeTrakr network, in this study, we focused only on US American isolates as they were the most abundant. To mimic the national surveillance data as much as possible, we limited our study to isolates collected between 2014 and 2018, as this corresponds to the timeframe of the Danish isolates.

We used the frequency of clinical cases to estimate *L. monocytogenes* virulence (i.e., harmfulness). The clinical frequency describes the ratio of the number of samples found in a clinical setting to the total amount of samples (i.e., isolates found in a clinical and food industry setting). To make predictions more easily interpretable, we binned the outcome variable (i.e., clinical frequency) into three clinical concern categories: lower (<0.5), medium (0.5–0.7), and higher (>0.7). These thresholds were chosen considering over-/underrepresentation of clinical samples in comparison to food and food processing samples. For example, for a clinical frequency of <0.5, fewer clinical samples are found in comparison to food-related isolates. This suggests an under-representation of clinical samples. Similarly, a clinical frequency between 0.5 and 0.7 suggests a mild overrepresentation of clinical samples, and a clinical frequency > 0.7 suggests a clear overrepresentation, and hence a higher virulence potential.

In order to find the best predictive features for Machine Learning, we compared three different genomic levels (i.e., virulence genes, pan- genome genes, and single nucleotide polymorphisms (SNPs)). The individual isolates' level-specific features and associated clinical frequencies were used to create an input matrix for the ML model training. This matrix consists of the different genome isolates as rows, the different features (i.e., gene identities and binary absence/presence encoding) as columns, and an extra column containing the respective clinical frequency values. For the virulence and pan-genome level, the alignment identities were transformed into absolute values (i.e., 90 % ➔ 0.9), which brings the input values into a numerical space between 0 and 1 suitable for many machine learning algorithms. The SNPs features are already in a binary feature space of 1 and −1, which can be used directly for machine learning. The predictive performance of input features from three different genomic levels (i.e., virulence genes, pan-genome genes, and single nucleotide polymorphisms (SNPs)) and six machine learning algorithms (i.e., Support Vector Machine with a linear kernel, Support Vector Machine with a radial kernel, Random Forrest, Neural Networks, LogitBoost, and Majority Voting) were compared.

In this part of the project, we used isolate WGS data from two national surveillance programs to train multiple supervised machine learning algorithms and compared their ability to predict clinical frequency. Looking at the results of the virulence gene level (Figure 1), we can see that all chosen ML models perform similarly well. In particular, there is an overlap in the 95 % - CI interval for all six models. A comparable pattern is seen when we use absence and presence of SNPs as input features. However, the pan-genome gene level results show more variation between the performances of different models. This becomes especially clear for the Support Vector Machine with a linear kernel and the Random Forest classifier, as their 95 %-CI do not overlap.



Comparing the different input levels to each other, we found that models trained at the pan-genome level generally yielded an increased performance. In particular, using pan-genome features over SNPs features resulted in higher performances for all models except Random Forrest. Similarly, the use of pan-genome features in comparison to virulence features resulted in higher performances for four out of the six models.

Figure 1 National surveillance dataset F1-score comparison plot for three different genomic levels. The F1-scores represent the bootstrapping results and the 95 %-CI of the 30 repeated nested-cross validation performance values.

The validation of the pre-trained ML models based on 101 previously in vivo studied isolates resulted in F1-scores up to 0.76. Furthermore, we found that the more rapid and less computationally intensive raw read alignment yields comparably accurate models as de novo assembly.

The results of our study suggest that machine learning trained on pan-genome genes is the best and most robust choice for the prediction of clinical frequency. Our study contributes to more rapid and precise characterization of *L. monocytogenes* virulence and its variation on a sub-species level. We further demonstrated a possible application of WGS data in the context of microbial hazard characterization for food safety. In the future, predictive models may assist case-specific microbial risk management in the food industry. The python code, pre-trained models, and prediction pipeline are deposited at (https://github.com/agmei/LmonoVirulenceML).

The development of predictive ML modes for disinfectant tolerance of *L. monocytogenes* using WGS data and laboratory-analysed phenotype data

Listeria monocytogenes is a potentially severe disease-causing bacteria mainly transmitted through food. This pathogen is of great concern for public health and the food industry in particular. Many countries have implemented thorough regulations, and some have even set 'zero-tolerance' thresholds for particular food products to minimize the risk of *L. monocytogenes* outbreaks. This emphasizes that proper sanitation of food processing plants is of utmost importance. Consequently, in recent years, there has been an increased interest in *L. monocytogenes* tolerance to disinfectants used in the food industry. Even though many studies are focusing on laboratory quantification of *L. monocytogenes* tolerance, the possibility of predictive models remains poorly studied.

Within this project, we explored the prediction of tolerance and minimum inhibitory concentrations (MIC) using whole genome sequencing (WGS) and machine learning (ML). We used WGS data and MIC values to quaternary ammonium compound (QAC) disinfectants from 1649 *L. monocytogenes* isolates to train different ML predictors.

To see the effect of genomic input on the ML prediction, we compared the predictive performance of several ML models using pan-genome features (n=14531) or SNP features (n=316084) as input. For classification into tolerant and sensitive classes, using pan-genome features resulted in higher cross-validated BA scores.

Our project shows promising results for predicting tolerance to QAC disinfectants using WGS and machine learning. We were able to train high-performing ML classifiers to predict tolerance with balanced accuracy scores up to 0.97±0.02. For the prediction of MIC values, we were able to train ML regressors with mean squared error as low as 0.07±0.02 (Figure 2). We also identified several new genes related to cell wall anchor domains, plasmids, and phages, putatively associated with disinfectant tolerance in *L. monocytogenes*.

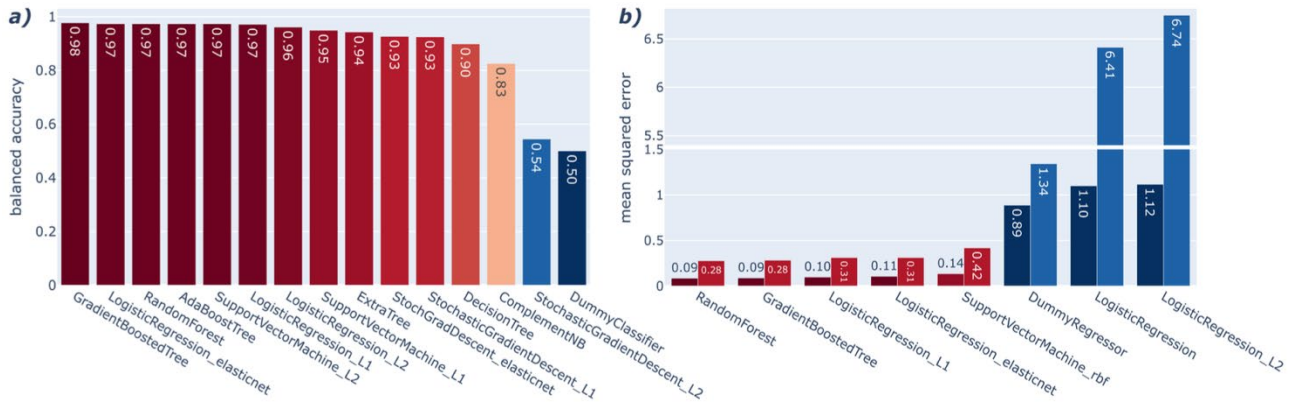


Figure 2 Performances of machine learning for disinfectant resistance. a) Classification performance comparison of the cross-validated balanced accuracy results (predict susceptible or resistance to disinfectance). The colour scale describes the model's performance from red (good) to blue (poor). b) Regression performance comparison of the cross-validated

mean squared error results (predict MIC value). The colour scale describes the model's performance from red (good) to blue (poor). The darker shades (right columns) correspond to the regression of non-transformed MIC values, and the lighter shades (left columns) correspond to the regression of log2 transformed MIC values.

The findings of this study are a first step towards prediction of *L. monocytogenes* tolerance to QAC disinfectants used in the food industry. In the future, predictive models might be used to monitor disinfectant tolerance in food production and might support the conceptualization of more nuanced sanitation programs.

As WGS analysis finds more popularity in the food industry, there is a great variety of application possibilities. For example, predictive models like the one proposed in this project might be beneficial to monitor disinfectant tolerance in food production sites and to aid the development of more effective sanitation programs. In addition, this study serves as a proof- of-concept for predicting disinfectant tolerance using WGS and ML. This methodology could be extended to other disinfectant agents, such as peracetic acid, for which tolerance mechanisms are mostly unknown.

However, there are still some limitations that minimize the transferability of the current models to an industry setting. In practice, many factors, such as biofilm formation, mediate the survival of sensitive *L. monocytogenes* isolates that are not considered in this project. Nevertheless, this project is an essential first step towards predicting disinfectant tolerance from WGS data for application in a food industry setting and sets priority directions for further research.

The development of WGS-ML tool

Despite current surveillance and sanitation strategies, foodborne pathogens continue to threaten the food industry and public health. Whole genome sequencing (WGS) has reached an unprecedented resolution to analyze and compare pathogenic bacterial isolates. The increased resolution significantly enhances the possibility of tracing transmission routes and contamination sources of foodborne pathogens. In addition, machine learning (ML) on WGS data has shown promising applications for predicting important microbial traits such as virulence, growth potential, and resistance to antimicrobials. Many regulatory agencies have already adapted WGS and ML methods. However, the food industry hasn't followed a similarly enthusiastic implementation. Some possible reasons for this might be the lack of computational resources and limited expertise to analyze WGS and ML data and interpret the results. Here, we present ListPred, a ML tool to analyze WGS data of *Listeria monocytogenes*, a very concerning foodborne pathogen. ListPred is able to predict two important bacterial traits, namely virulence potential and disinfectant tolerance, and only requires limited-computational resources and practically no bioinformatic expertise, which is essential for a broad application in the food industry.

ListPred is accessible through different channels and has three usage variations to accommodate varying user expertise and computational resources. We conducted a small-scale benchmarking study assessing the prediction consistency across different sequencing platforms, i.e., short- and long-read sequencing, and different sequencing data input types, i.e., raw reads or assemblies.

We found a discordance of 8% - 10% when comparing raw read and assembled inputs of both sequencing methods (i.e., short and long reads) and prediction tasks (i.e., virulence potential and disinfectant tolerance). The comparison of predictions for virulence potential from raw and assembled long reads yielded the highest concordance of 92.8%. Predictions of virulence potential and disinfectant resistance comparing raw and assembled short reads showed the lowest concordance of 89.7%. Generally, we observed that using long reads resulted in higher prediction overlaps than short reads when comparing raw reads versus assemblies as input. We could observe a great overlap between predictions from short-read raw reads and assemblies of 99.4%.

Short-read sequencing is still the most widely used sequencing method, probably due to the high accuracy of the reads produced, the reasonably low cost, and the availability of third-party provided services. Nevertheless, with advances in long-read sequencing, the accuracy of long-read sequencers continues to improve, and the costs of sequencing in general has been continuously dropping [4, 5]. As WGS is slowly gaining more traction in the food industry, it will be necessary for individual producers to assess the best fit for them [6].

Consequently, food producers will most likely use both methods, especially in the current transitional phase to a broader implementation of WGS in the food industry. Overall, we observed relatively small differences of maximum 11.3% for predicting virulence potential and 3.1% for disinfectant tolerance. This indicates that ListPred predictions are pretty stable across input from different sequencing platforms.

The need for quick and user-friendly analysis tools increase with the rise of sequencing methods, such as WGS, in the food industry. Initially, there is a big difference in computational and bioinformatic expertise across food production companies. Bigger producers might have the economic resources for sequencing and computing facilities, whereas medium and smaller-sized producers might struggle to allocate the funding for the implementation of these methods. Hence, we developed ListPred to have a point-and-click web application that alleviates computational resources and can be used with basic computational skills. The output of ListPred is clear and descriptive to enhance interpretability. Microbial data collected by food-producing companies, especially from pathogenic bacteria, is generally confidential, which might limit the willingness to upload data to a web application, even if data is not stored. To ease concerns and facilitate maximum privacy of sensitive data, we additionally distributed ListPred in two stand-alone versions that can be run entirely offline and in-house.

In conclusion, ListPred aims to enable quick and accessible WGS analysis in the food industry. The increasing implementation of sequencing technologies in the food industry will raise the need for easy-to-use tools that don't require computational or bioinformatic expertise. The phenotypic traits predicted by ListPred could give risk managers important insight into the potential threat of *L. monocytogenes* isolates found in the production environment and how to eliminate them more effectively. This information might even be considered in the design of risk mitigation strategies and sanitation plans. Apart from this, ListPred is an important example of possible applications and possible benefits of WGS implementation in the food industry. Considering further application showcases, lowering sequencing prices, and the development of novel end-to-end analysis tools might shift current cost-benefit evaluations in favour of a more routine implementation of WGS in the food industry.

Conclusion

L. monocytogenes is a concerning foodborne pathogen that can cause severe health implications for at-risk individuals. However, not all *L. monocytogenes* isolates are the same, as they can display a diverse set of phenotypes on a subspecies level. Hence, in case of contamination of food or the food production environment, it is of utmost importance to thoroughly characterize the found *L. monocytogenes* isolates to evaluate their potential risk. Due to its unprecedented resolution, WGS has gained a lot of interest for the characterization of pathogens in the food industry. The use of ML in combination with WGS to predict a variety of phenotypic traits has been extensively studied. However, most of the research focuses on public health aspects and only a few studies explore the application of WGS and ML in the context of food safety.

This project presents valuable research to improve food safety using whole genome sequencing and machine learning. The predictive AI models developed in this project might help to evaluate *L. monocytogenes* risk on a subspecies level and to inform risk management. For example, the virulence potential prediction model could be a valuable addition to risk ranking schemes. Additionally, the model predicting disinfectant tolerance could aid the effective mitigation of *L. monocytogenes* contamination in food production environments and help to conceptualize cleaning procedures. To facilitate the translation of this academic research into practice, all of the applications are easily accessible through the distribution of ListPred.

The potential benefit of the research in this project vastly depends on the implementation of WGS in the food industry. The accessibility to sequencing either in-house or through outsourcing, the decrease in WGS sequencing and analysis cost, the adaptation of food legislation and regulations, and the further development of easy-to-use analysis tools will lead the way toward a more routine implementation of WGS in the food industry. In the future, more research is needed to address some of the current barriers and concerns and to highlight additional applications which could shift cost-benefit evaluations in favor of a more routine WGS implementation.

12. The relevance of the results, including relevance for the dairy industry

An innovative, rapid and cost-effective omics tool for *L. monocytogenes* to determine virulence and resistance to disinfectants has been created. The user-friendly web-based tool, is freely accessible online and requires no prior knowledge of bioinformatics for interpretation of output. The tool is hosted and maintained by DTU via our CGE website[1]. If any industries would not like to upload sequence data to an online tool, they can access the tool via a standalone version hosted by our CGE bitbucket website[2]. The tool can be used for all kinds of sequencing technologies including Nanopore technology, which can sequence a *L. monocytogenes* genome within 6 – 7 hours[3]. The tool can be used for all food industries concerned with *L. monocytogenes* contamination. The result from this project can also be used for public authority in order to surveillance of *L. monocytogenes.*

Future work after this project can include an extension of the tool to cover other foodborne pathogens. In addition, the sequencing technology is already standardized and AI solution tool is also possible for food industry. The possible research area on this would be by making sequencing much more easily to access and to apply in practice with less time and less cost than the current routine work in food industry.

13. Communication and knowledge sharing about the project

Papers in international journals:

Gmeiner A, Njage PMK, Hansen LT, Aarestrup FM, Leekitcharoenphon P. "Predicting Listeria monocytogenes Virulence using Whole Genome Sequencing and Machine Learning". Int J Food Microbiol. 2024

Karlsmose AK, Ivanova M, Kragh ML, Kjeldgaard JS, Otani S, Svendsen CA, Papić B, Zdovc I, Tasara T, Stephan R, Heir E, Langsrud S, Møretrø T, Dalgaard P, Fagerlund A, Hansen LT, Aarestrup FM, Leekitcharoenphon P. "A novel metagenomic approach uncovers phage genes as markers for increased disinfectant tolerance in mixed *Listeria monocytogenes* communities". Infect Genet Evol. 2024

Brown P, Murray RGE, Galsworthy S, Ivanova M, Leekitcharoenphon P, Ward T, Kucerova Z, Chen Y, Elhanafi D, Siletzky R, Kathariou S. "Draft genome sequences of a historical collection of *Listeria monocytogenes* from humans and other sources, 1926-1964". Microbiol Resour Announc. 2023

Brown P, Kucerova Z, Gorski L, Chen Y, Ivanova M, Leekitcharoenphon P, Parsons C, Niedermeyer J, Jackson J, Kathariou S. "Horizontal Gene Transfer and Loss of Serotype-Specific Genes in *Listeria monocytogenes* Can Lead to Incorrect Serotype Designations with a Commonly-Employed Molecular Serotyping Scheme". Microbiol Spectr. 2023

Gmeiner, A., Ivanova, M., Njage, P.M.K. et al. Quantitative prediction of disinfectant tolerance in Listeria monocytogenes using whole genome sequencing and machine learning. Sci Rep 15, 10382 (2025).

[1] https://cge.cbs.dtu.dk/services/

[2] https://github.com/genomicepidemiology/ListPred/tree/master/workflow, https://hub.docker.com/r/genomicepidemiology/listpred

[3] https://orbit.dtu.dk/en/projects/udvikling-af-tidseffektiv-dnabaseret-metode-til-salmonella-serotypning(ac3e9f24-2aaf-42cc-a1cb-8a7a9651206a).html

Gmeiner A, Ivanova M, Kaas RS, Xiao Y, Otani S, Leekitcharoenphon P. ListPred: A predictive ML tool for virulence potential and disinfectant tolerance in Listeria monocytogenes. Infect Genet Evol. 2025 Mar 18;130:105739. doi: 10.1016/j.meegid.2025.105739.

Ivanova M, Laage Kragh M, Szarvas J, Tosun ES, Holmud NF, Gmeiner A, Amar C, Guldimann C, Huynh TN, Karpíšková R, Rota C, Gomez D, Aboagye E, Etter A, Centorame P, Torresi M, De Angelis ME, Pomilio F, Okholm AH, Xiao Y, Kleta S, Lüth S, Pietzka A, Kovacevic J, Pagotto F, Rychli K, Zdovc I, Papić B, Heir E, Langsrud S, Møretrø T, Brown P, Kathariou S, Stephan R, Tasara T, Dalgaard P, Njage PMK, Fagerlund A, Aarestrup F, Truelstrup Hansen L, Leekitcharoenphon P. Large-scale phenotypic and genomic analysis of Listeria monocytogenes reveals diversity in the sensitivity to quaternary ammonium compounds but not to peracetic acid. Appl Environ Microbiol. 2025 Mar 4:e0182924. doi: 10.1128/aem.01829-24.

Eaily read papers:

Pimlapas Leekitcharoenphon, Alexander Gmeiner, Mirena Ivanova, Patrick Murigu Kamau Njage, Lisbeth Truelstrup Hansen, Yinghua Xiao (2023). Developing omics tool for food safety. Mælkeritidende, 2, 1-2.

Pimlapas Leekitcharoenphon, Alexander Gmeiner, Mirena Ivanova, Patrick Murigu Kamau Njage, Lisbeth Truelstrup Hansen, Yinghua Xiao (2025). Improving food safety using AI. Mælkeritidende, 1, 18-19.

Student theses:

Gmeiner Alexander, “Whole genome sequencing and machine learning to improve food safety”, PhD thesis, Technical University of Denmark, 2024

https://backend.orbit.dtu.dk/ws/portalfiles/portal/364784712/PhD_thesis_GmeinerA_final.pdf

Balthasar Clemens Schlotmann on “Exploration of Whole Genome Sequencing-based Machine Learning (WGS-ML) in Food Safety Crisis Management”, Master thesis, Technical University of Denmark, 2020

Agnete Kirstine Karlsmose on “Metagenomic detection of genes associated with disinfectant resistance in mock Listeria monocytogenes communities”, Master thesis, Technical University of Denmark, 2023

Oral presentations at scientific conferences, symposiums etc.:

31st ECCMID, 9th -12th July 2021, “Improving Food Safety using Whole Genome Sequencing and Machine Learning”, (Oral presentation via online)

ELIXIR 6th Annual Danish Bioinformatics Conference, Aalborg, Denmark, 18th – 19th November 2021, “Improving Food Safety using Whole Genome Sequencing and Machine Learning”, (Poster)

IAFP 2022, Pittsburgh, USA, 31st July – 3rd August 2022, “Improving Food Safety Using Whole Genome Sequencing and Machine Learning” (Oral presentation)

Foodmicro, Athens, Greece, 28th – 31st August 2022, “Predicting the Virulence of *Listeria Monocytogenes* using Whole Genome Sequencing and Machine Learning”, (Oral presentation)

4th ICAHS, Copenhagen, Denmark, 3rd-5th May 2022, “Predicting Virulence of *Listeria Monocytogenes* Using Whole Genome Sequencing and Machine Learning”, (Poster)

ASM Microbe, Washington DC, USA, 9th – 13th June 2022, “Predicting Virulence of *Listeria Monocytogenes* Using Whole Genome Sequencing and Machine Learning”, (Poster)

21st ECCB, Sitges, Spain, 18th -21st September 2022, “Predicting Virulence of *Listeria Monocytogenes* Using Whole Genome Sequencing and Machine Learning”, (Poster)

ISMB/ECCB, Lyon, France, 23th – 27th July 2023, “Predicting Disinfectant Resistance in *L. monocytogenes* using Whole Genome Sequencing and Machine Learning”, (Poster)

33rd ECCMID, Copenhagen, Denmark, 15th - 18th April 2023, “Large-scale phenotypic and genomic analyses of *Listeria monocytogenes* susceptibility to benzalkonium chloride” (poster)

ASM Microbe, Houston, USA, 15th - 19th June 2023, “A Novel Metagenomic Approach Uncovers Phage Genes As Markers For Increased Disinfectant Tolerance In Mixed *Listeria Monocytogenes* Communities” (poster)

Dairy Research Day 2023, Herning, Denmark, 16th March 2023, “Improving food safety using whole genome sequencing and machine learning” (Oral presentation)

IAFP 2023, Toronto, Canada, 16th - 19th July 2023, “Global Distribution of Genes Conferring Increased Tolerance to Food Industry Disinfectants in *Listeria monocytogenes*” (Oral presentation)

IAFP 2024, Geneva, Switzerland, 30th April – 2nd May 2024, “LisPred: A predictive ML tool for virulence potential and disinfectant tolerance in *Listeria monocytogenes*” (Oral presentation)

14. Contribution to master and PhD education

Master thesis for Balthasar Clemens Schlotmann on “Exploration of Whole Genome Sequencing-based Machine Learning (WGS-ML) in Food Safety Crisis Management”, Technical University of Denmark, 2020

Master thesis for Agnete Kirstine Karlsmose on “Metagenomic detection of genes associated with disinfectant resistance in mock Listeria monocytogenes communities”, Technical University of Denmark, 2023

PhD research stay for PhD student Gmeiner Alexander at Imperial College in London for 6 months in 2023

PhD education for Gmeiner Alexander, “Whole genome sequencing and machine learning to improve food safety”, Technical University of Denmark, completed in 2024

15. New contacts/projects

This project is the first project to incorporate whole genome sequencing and AI for food safety in food industry. The results from this project has shed light on multiple possibilities to include or apply AI in food safety.

16. References

1. Maury MM, Tsai Y-H, Charlier C, et al (2016) Uncovering Listeria monocytogenes hypervirulence by harnessing its biodiversity. Nat Genet 48:308–313. https://doi.org/10.1038/ng.3501

2. FDA (2022) GenomeTrakr. In: https://www.fda.gov/food/whole-g enome-sequencing-wgs-program/genometrakr-network

3. NCBI (2022) NCBI’s Pathogen Detection Portal. In: https://www. ncbi.nlm.nih.gov/pathogens

4. Eisenstein M (2023) Innovative technologies crowd the short-read sequencing market. Nature 614:798–800.

https://doi.org/10.1038/d41586-023-00512-4

5. Muir P, Li S, Lou S, et al (2016) The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol 17:53. https://doi.org/10.1186/s13059-016-0917-0

6. Jagadeesan B, Gerner-Smidt P, Allard MW, et al (2019) The use of next generation sequencing for improving food safety: Translation into practice. Food Microbiol 79:96–115. https://doi.org/10.1016/j.fm.2018.11.005