International Journal of Food Microbiology 410 (2024) 110491

Contents lists available at ScienceDirect

International Journal of Food Microbiology

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

Predicting *Listeria monocytogenes* virulence potential using whole genome sequencing and machine learning

Alexander Gmeiner [a,*], Patrick Murigu Kamau Njage [a], Lisbeth Truelstrup Hansen [b], Frank M. Aarestrup [a], Pimlapas Leekitcharoenphon [a]

[a] National Food Institute, Technical University of Denmark, Research Group for Genomic Epidemiology, Kgs. Lyngby, Denmark
[b] National Food Institute, Technical University of Denmark, Research Group for Food Microbiology and Hygiene, Kgs. Lyngby, Denmark

ARTICLE INFO

Keywords:
Hazard characterization
Clinical frequency
National surveillance
Pan-genome
Virulence genes
Food safety

ABSTRACT

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 study, 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 WGS datasets used in this study for ML model training consisted of i) national surveillance isolates (n = 169, covering 38 MLST types) and ii) publicly available isolates acquired through the GenomeTrakr network (n = 2880, spanning 80 MLST types). We used the clinical frequency, i.e., ratio of the number of clinical isolates to total amount of isolates, as estimate for virulence potential. 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.

Our machine learning models predicted sub-species virulence potential with nested cross-validation F1-scores up to 0.88 for the majority voting classifier trained on national surveillance data and using pan-genome genes as input features. 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 a majority voting classifier 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).

1. Introduction

Food-borne zoonotic diseases continue to pose an immense threat to public health. The European Food Safety Authority (EFSA) reports over 350,000 cases of food-borne diseases annually in the European Union alone. Nevertheless, the actual number is estimated to be even higher since not all infections need medical attention (EFSA, 2023). *Listeria monocytogenes* is a concerning food-borne pathogen that can cause severe and diverse pathogenesis (Listeriosis), especially in elderly or immunocompromised individuals and foetuses (McLauchlin, 1996). Even though, the incidence of listeriosis is relatively low (0.1 to 10 cases per 1 million people per year) (WHO, 2023), its high case fatality rate

* Corresponding author at: National Food Institute, Technical University of Denmark, Kemitorvet, Building 204, DK-2800 Kgs. Lyngby, Denmark.
E-mail address: algm@food.dtu.dk (A. Gmeiner).

https://doi.org/10.1016/j.ijfoodmicro.2023.110491
Received 31 March 2023; Received in revised form 6 October 2023; Accepted 12 November 2023
Available online 17 November 2023

(17.6 %) makes listeriosis one of the most concerning food-borne diseases in the EU (EFSA and ECDC, 2021).

L. monocytogenes is widely distributed in the environment and in farm animals from where it quickly enters into food processing facilities, regardless of the high food safety standards in place (Kasalica et al., 2011). It is resilient to extreme environments and can withstand high salt concentrations, sustain a broad pH range, and grow at low (i.e., refrigeration) temperatures (Liu et al., 2005).

Due to the possible severity of *L. monocytogenes* infections, many authorities have implemented thorough regulations on *L. monocytogenes* in food (Neri et al., 2019). Most of these food safety policies assess the risk of *L. monocytogenes* on a species level and do not consider within species variation. However, recent studies show substantial differences on a sub-species level. In particular, the distribution of *L. monocytogenes* subtypes differs significantly between different niches making it possible to associate some subtypes with either food or clinical environments (Gray et al., 2004; Maury et al., 2016, 2019; Orsi et al., 2011). Within these niches, isolates can also demonstrate different pathogenic potentials. Predominantly, subtypes more frequently associated with the clinical cases are thought to be more virulent than those associated with food environments. Interestingly, recent studies also show that there are not only differences in virulence potential within different niches but that there are vast differences in pathogenicity, even within subtypes (Maury et al., 2016). These findings suggest that *L. monocytoges* virulence might be strain specific (Muchaamba et al., 2022). Even though consideration of sub-species information in risk management might be challenging, the exploration of more detailed risk characterization of *L. monocytogenes* in the food industry remains of interest.

In recent years, there has been an increasing interest in using Whole Genome Sequencing (WGS) techniques in the food industry (Ireland et al., 2018; Jagadeesan et al., 2018). The analysis of WGS data enables a very fine resolution of strains even to single nucleotide alternations. The benefits of this fine resolution can already be seen in other areas, such as pathogen surveillance and outbreak tracing (Rantsiou et al., 2018). Recent studies have also shown that WGS data can be used to predict important microbial traits such as virulence and antimicrobial resistance (Camp et al., 2020; Collineau et al., 2019; Njage et al., 2019; Pincus et al., 2020). A better understanding and prediction of certain microbial traits would allow food safety authorities to perform a more detailed characterization of microbial hazard. This refined hazard characterization could then aid risk management decisions in case of *L. monocytogenes* contamination in the food industry.

Hence, within this study, we further explore the use of Whole Genome Sequencing for microbial hazard characterization on sub-species level. In particular, we are focussing on the prediction of *L. monocytogenes* virulence potential (for simplicity, also referred to as 'virulence' throughout the study) on a sub-species level into three different classes (i.e., lower, medium, and higher risk). We analysed WGS data from two national surveillance programs using Machine Learning (ML). Our study aimed to i) develop a well-performing machine learning predictor for *L. monocytogenes* virulence and ii) explore predictive differences of three genomic levels as input features for our ML models and iii) evaluate feature extraction by raw read alignment as a more rapid alternative to assembly based methods. The results of this study facilitate a more granular understanding of *L. monocytogenes* risk potential on a sub-species level.

2. Materials and methods

2.1. Data acquisition

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 (Maury et al., 2016). For the Danish isolates, genomic DNA was extracted using Invitrogen Easy DNA. The sequencing libraries were prepared with the Nextera XT DNA library preparation kit according to the manufacturer's protocol (Illumina, Inc., San Diego, CA, United States). The DNA samples were sequenced using Illumina MiSeq. We deposited the raw sequencing data in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no.: PRJEB59720. Both the French and Danish surveillance systems exhaustively collect all *L. monocytogenes* isolates found in the clinic and food industry. Since the systems are strictly controlled, we expect them to give a good overview of their respective country's epidemiological *L. monocytogenes* landscape.

The national surveillance isolates originate from five different source locations: Denmark ($n = 105$), France ($n = 38$), Poland ($n = 4$), Tunisia (n = 1), and N/A ($n = 21$). There are five different isolation sources in total. Apart from the source label "Other" (42 %), most isolates are isolated from food (24 %), followed by human (20 %), animal (9 %), and laboratory (5 %) origin. The isolates were distributed over 38 different MLST types, with ST121 (15 %), ST8 (14 %), and ST6 (13 %) being the most abundant (see Table S1).

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 (FDA, 2022), which can be accessed through NCBI's Pathogen Detection Portal (NCBI, 2022). Even though different international institutions are contributing to the GenomeTrakr network, in this study, we focussed 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.

Additionally, we only used WGS isolates from a clinical or food/food processing environment setting (i.e., environmental isolates are excluded). For more detailed information about the inclusion criteria and isolation sources, see the supplementary metadata in Table S2. The subset of isolates was subtyped using MLST (v2.0) (Clausen et al., 2018; Larsen et al., 2012) and filtered for isolates of multi-locus sequence type (MLST) clusters for which there were at least one clinical and one food-related sample ($n = 2880$).

In summary, the source of isolation for the GenomeTrakr isolates is either clinical (60 %) or environmental/other (40 %), mainly consisting of food and food processing environment isolates. Further, the dataset included 80 MLST types, with ST5 (17 %) and ST1 (12 %) being the most abundant.

2.2. Sample pre-processing

To confirm that all isolates are *L. monocytogenes*, we conducted species identification using KmerFinder (v3.0.2) (Clausen et al., 2018; Hasman et al., 2014; Larsen et al., 2014). Further, all raw sequencing reads were processed using the in-house FoodQC pipeline. The pipeline trims the raw reads with bbduk2 from BBTools (v36.49) (Bushnell, 2022) using only reads $\geq$50 bp in length, having a Phred score per base $\geq$20 from right to left, and filtering the institution-specific adapters. The trimmed reads were quality checked using FastQC (v0.11.5) (Andrews, 2010) and assembled using SPAdes (v3.11.0) (Prjibelski et al., 2020) with a kmer coverage of two and excluding contigs that are smaller than 500 bp.

2.3. Clinical frequency as a measure of virulence

In this study, 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). For the French isolates, the clinical frequency for individual clonal complexes (CC) was extracted from Maury et al. (2019). The clinical frequencies for each MLST group of the Danish data were derived from the annual report on zoonoses (Anonymous, 2019). Looking at the distribution of the isolates over the different clinical frequency classes, most of the isolates were in the lower-risk class (47 %), followed by the mid-risk (33 %) and higher-risk class (20 %) (Fig. 1a).

For the GenomeTrakr dataset, the isolates were grouped according to their MLST types and the group's respective clinical frequency was calculated according to Eq. (1). Where #clinical and #food represent the number of clinical and food isolates in each MLST, respectively.

$$clinical\ frequency = \frac{\#clinical}{(\#clinical + \#food)} \quad (1)$$

Each MLST or CC group has its individual clinical frequency value, which was used as an outcome variable for supervised Machine Learning. To ensure that the derived clinical frequency describes the subtype virulence as closely as possible, the data must represent the underlying epidemiological pattern as exhaustively as possible. We assume this to be true for the national surveillance data. However, this is a known limitation for the GenomeTrakr dataset as specific sampling events might lead to overrepresentation of certain isolates that would bias the clinical frequency calculation.

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 underrepresentation 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. Nevertheless, this is a fairly straight forward attempt to categorize virulence that might not be optimal. Looking at the distribution of the GenomeTrakr isolates over the clinical frequency classes, we find the biggest proportion in the higher-risk class (46 %), followed by the lower-risk class (37 %) and the medium-risk class (17 %) (Fig. 1b).

2.4. Feature description

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)). For the virulence level, we used a database of 136 virulence-associated genes of *L. monocytogenes*. This database resulted from an exhaustive literature review of virulence genes, virulence factors, virulence-associated factors, and environmental stress tolerance genes (Njage et al., 2019). In order to obtain a set of genes for the pan-genome level, the rapid prokaryotic annotation tool Prokka (v1.12) (Seemann, 2014) was used to extract relevant genomic features from all the assembled genomes. The Prokka output was subsequently processed with Roary (v3.13.0) (Page et al., 2015) to identify core and accessory genes.

The different isolate genomes were aligned to the virulence and pan-genome reference gene sets using two different methods. For assembled genomes, tblastn (NCBI-BLAST+ v2.11.0+) (Camacho et al., 2009) was used to align the genes and the database. Tblastn was run with the reference genes as a query, the genome as a subject, an *E*-value cut-off of 0.001, and the number of best query/target hits limited to one. For raw reads, KMA (v1.3.15) (Clausen et al., 2018) with the −1t1 flag was used to align the raw reads to the reference genes. To derive information about SNPs, all genomes were globally aligned to the EGD-e reference genome (Accession nr.: GCF_000196035.1). These global alignments were created with KMA using the same parameters described by Aytan-Aktug et al. (2021). The nucleotides of the individual isolates were compared to the nucleotides of the EGD-e reference and binary encoded (1 for matching nucleotides; −1 for mismatching nucleotides). We chose binary encoding as described by Aytan-Aktug et al. (2020) as the study has found no significant performance increases for different encodings.

2.5. Machine learning input matrix

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.

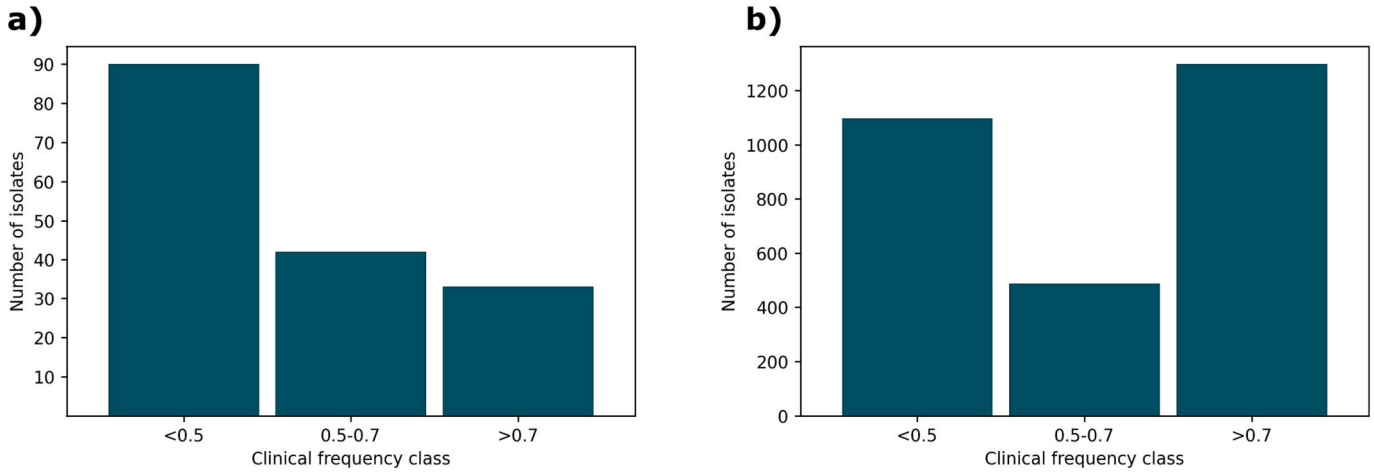

Fig. 1. a) Clinical frequency barplot for the national surveillance isolates. The isolates were grouped into three distinct categories, i.e., lower (<0.5), medium (0.5–0.7), and higher (>0.7), according to their clinical frequency value. **b)** clinical frequency barplot for the GenomeTrakr isolates. The isolates were grouped into three distinct categories, i.e., lower (<0.5), medium (0.5–0.7), and higher (>0.7), according to their clinical frequency value.

2.6. Feature filtering

The number of features compared to the number of samples significantly affects Machine Learning and is often referred to as the "curse of dimensionality". If the feature space increases, it is also becoming more sparse. More information (i.e., samples) needs to be added to find the desired pattern or signals in this increasingly sparse space. However, in many cases obtaining more data is not feasible. Another solution is to reduce the feature space. This can be done in two ways: through dimensionality reduction algorithms or feature selection. For our study, we chose a feature selection method rather than dimensionality reduction to preserve the original features (i.e., %-identity of a gene, absence/presence of a SNP). We applied this feature reduction step for the large pan-gene and SNPs feature space. For the pan-genome genes, we calculated the variance of the individual feature columns and sorted them in decreasing order.

Subsequently, we explored the effect of different numbers of features on the model's predictive performance to find an optimal value. When the pangenome genes were used as input, the model benefited from an increased number of features up until about 120 features (see Fig. S1). Beyond that, the performance increase levelled off, which means that additional features do not seem to benefit the model's performance considerably. For the SNPs level, the same method was used. However, since the encoding for SNPs (i.e., presence and absence) was categorical rather than continuous, we used entropy as a measure for feature variation instead of variance. A similar trend was seen as for the pan-genes level, i.e., the performance increased with additional number of features (see Fig. S2) followed by levelling off to a constant at 140 SNPs. Therefore, the models did not benefit considerably from adding more features beyond this number. In summary, we chose 120 and 140 as the optimal number of features for the pan-gene and SNPs levels, respectively.

2.7. Machine learning

For the Machine Learning analysis, we used the sci-kit learn library (v0.23.1) (Pedregosa et al., 2011) in Python (v3.8.3). To find a model that is best suited to capture the underlying signal of the data we compared the performance of six different algorithms: Support Vector Machine with a linear kernel (SVC-lin), Support Vector Machine with a radial kernel (SVC-rad), Random Forrest (RF), Neural Networks (NN), LogitBoost, and Majority Voting. In addition, some of these models have been used in similar studies and showed promising results (Njage et al., 2019). To prevent information leakage, which can bias the generalization performance, we ensured that very closely related isolates were kept together in either the train, test, or validation set. This method is commonly referred to as "blocking" and has been suggested to account for the relatedness of biological samples such as phylogenetic relations (Arning et al., 2021; Lees et al., 2020; Whalen et al., 2022). Even though some dependence structures might be too complex to be fully addressed by blocking, it will still prevent the inflation of generalization error by limiting data leakage (Whalen et al., 2022). Nevertheless, blocking will, at least to some extent, hinder the model from prioritizing features that correlate with the outcome due to phylogenetic relations (Lees et al., 2020). To mimic the distribution of our training set in the testing set, we are stratifying our split in regards to the clinical frequency. This should result in a closer approximation of the generalization error. We clustered the isolates using KMA as described by Aytan-Aktug et al. (2021) with a template and query coverage threshold of 98 %.

The workflow for training the different algorithms was structured as follows: The input data with corresponding clinical frequency labels were split into a training and test set using stratified and grouped 5-fold cross-validation (i.e., stratified by clinical frequency, grouped by clusters). The training set was subsequently used for training the algorithms. For each algorithm, the hyperparameters were tuned using repeated (n = 8) and grouped train-/validation (90 %/10 %) splits and random search cross-validation except for SVC-lin, for which grid search was used. The hyperparameter tuning splits were filtered to ensure that the validation set contained at least one isolate from each of the clinical frequency classes. The actual model training consisted of a pipeline that included the pre-processing steps. In the first pre-processing step, the input features were filtered according to the genomic level described in the method section. The number of samples for the underrepresented clinical frequency classes was increased using SMOTE (Chawla et al., 2002) to improve the prediction performance of the underpopulated classes. Lastly, the performance on the test set was evaluated for the individual algorithms using the best-performing hyperparameters. We reported the performances using six measures: accuracy, precision, recall, F1-score, ROC-AUC, and MCC. The described workflow involving random splitting, model training, and performance evaluation was repeated 30 times. This resulted in 30 different values for each performance measure and algorithm (i.e., 30 values times six performance measures times six algorithms). The mean performance and the 95 % confidence interval (CI) were calculated using Bootstrapping (n_datapoints = 30; n_reps = 100) and plotted in a bar diagram. The final machine learning training pipeline was deposited on GitHub (https://github.com/agmei/LmonoVirulenceML).

2.8. Model validation

For an independent validation set, data was collected from 12 in vitro *L. monocytogenes* virulence studies for which WGS data was available (den Bakker et al., 2012; Briers et al., 2011; Chen et al., 2011; Hurley et al., 2019; Jensen et al., 2016; McMullen et al., 2012; Muchaamba et al., 2022; Steele et al., 2011; Wagner et al., 2020, 2022; Yin et al., 2019) The final validation dataset consisted of 101 isolates. A detailed list of accession numbers and virulence levels can be found in supplementary material Table S3. To standardize the output of the studies, we grouped the laboratory results into two categories, i.e., putatively hypervirulent ($n = 33$) and putatively hypovirulent ($n = 68$). We excluded the prediction on the SNPs level since this has been found to perform considerably worse.

In order to make predictions from the trained ML models, the genomic data of the independent validation isolates was translated into the same feature space as for the training dataset (i.e., the individual genomes were screened against the reference databases, and the percent identities of the genes were reported). The input features were submitted to the selected pre-trained ML models, which returned an estimated clinical frequency class.

Our pre-trained models predicted three classes (i.e., lower, medium, and higher clinical frequency). However, the in vitro studies reported a binary outcome (putatively hyper/hypo virulent). To compare the ML predictions with the laboratory results, we regrouped the ML results from the validation dataset. The lower-risk clinical frequency class now corresponds to the putatively hypovirulent laboratory phenotype, while the medium and higher-risk classes corresponds to the putatively hypervirulent laboratory phenotype.

In this experiment, we explored the performance of four different Majority Voting models on independent validation data. These models were trained using two different datasets (i.e., national and GenomeTrakr) and input feature levels (i.e., virulence and pan-genome genes). The predicted clinical frequency classes were then validated against the laboratory phenotypes. The results were presented in confusion matrices, and the F1 performance measures were reported.

3. Results

3.1. Machine learning model selection based on national surveillance isolate data

In this part of the study, 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 (see Table 1, Table S4), 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 (see Fig. 2). 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 (i.e., no overlap of the 95 %-CI (see Fig. 2)) 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. Only for the two tree-based classifiers (i.e., Random Forrest and LogitBoost) we didn't observe a clear difference in 95 %-CI (see Table S4).

3.2. Assessing direct feature extraction from raw reads as faster alternative to assembly based methods

As de novo assembly of entire genomes is time-consuming and computationally expensive, hence, we explored direct raw read alignment against the reference gene databases as a possible alternative. We compared the predictive performance of the different ML algorithms trained on the national isolate WGS data. We aligned the assemblies or raw reads to the virulence gene databases using BLAST and KMA respectively. As expected, the ML results yield similar performances between the two alignment methods, i.e., the 95 %-CI are overlapping for all of the six different ML models (see Fig. S3).

3.3. Machine Learning model selection based on extended surveillance data from the GenomeTrakr network

To see how more data affects the prediction performance of our ML and to explore if we can use public datasets to predict *L. monocytogenes*, we used isolate WGS data from the GenomeTrakr network. Similarly, as for the national surveillance data, we are comparing the predictive performance of different ML classification algorithms in combination with different genomic input features. There were marked performance differences across different ML algorithms (see Fig. 3). In particular, the tree-based classifiers (i.e., Random Forrest and LogitBoost) seem to perform better than the others. For the virulence level, the confidence intervals of the tree based classifiers do not overlap with those of the other classifiers (Random Forrest F1 CI [0.95, 0.96], LogitBoost F1 CI [0.93, 0.95]). Looking at the pan-genome level, only the majority voting classifier (F1 CI [0.92, 0.94]) has an overlapping confidence interval with the tree based classifiers (Random Forrest F1 CI [0.94, 0.96], LogitBoost F1 CI [0.94, 0.95]).

Table 1
Nested cross-validation F1 performances for the national surveillance and GenomeTrakr datasets.

	National surveillance			GenomeTrakr	
	Virulence level	Pan-genome level	SNPs level	Virulence level	Pan-genome level
Linear SVC	0.81	0.89	0.77	0.83	0.82
Radial SVC	0.82	0.88	0.74	0.87	0.90
Random forrest	0.81	0.83	0.77	0.95	0.95
Neural network	0.79	0.84	0.74	0.91	0.89
LogitBoost	0.84	0.85	0.75	0.94	0.95
Majority voting	0.84	0.88	0.76	0.92	0.93

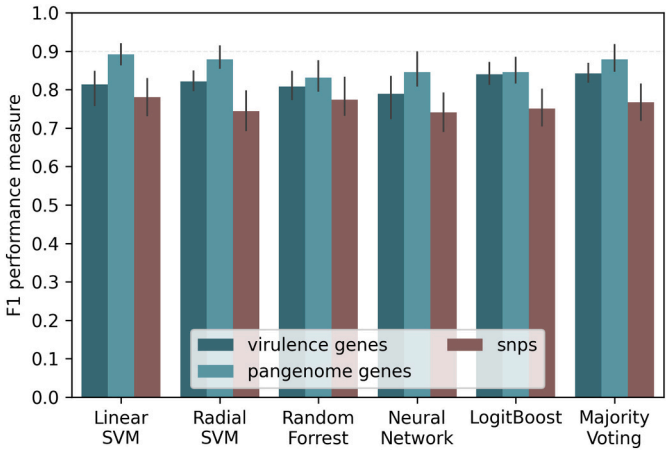

Fig. 2. National surveillance dataset F1-score comparison plot for three different genomic levels. The F1-scores represent the bootstrapping results and the 95 %-CI of the 30 repeated nested-cross validation performance values.

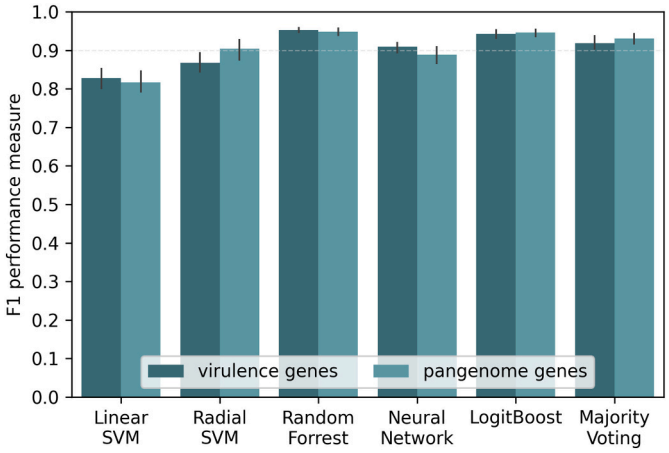

Fig. 3. GenomeTrakr dataset F1-score comparison plot for two different genomic levels. The F1-scores represent the bootstrapping results and the 95 %-CI of the 30 repeated nested-cross validation performance values.

Interestingly and in contrast to the results from the national dataset, there is only a little difference in performance between the virulence gene and the pan-genome level for most of the models. Only the Support Vector Machine with a radial kernel appears to benefit from using pan-genome features as input for the model training (virulence F1 CI [0.84, 0.89], pan-genome F1 CI [0.88, 0.92]). Further, the results show that the overall performance variation is less than for the national data set, which can be seen by smaller confidence intervals.

3.4. Validation of pre-trained ML models on independent WGS data with laboratory described virulence potential

To assess how well our ML predictors perform on independent data, we collected a set of WGS sequences from laboratory *L. monocytogenes* virulence studies. The results found in these studies were categorized into putatively hypervirulent and putatively hypovirulent categories. The laboratory phenotypes were compared to the ML predicted clinical frequency classes of four pre-trained Majority Voting models. All of the models could predict reasonably well on the validation test set. Looking at Table 2, we can see that the model trained on the national surveillance isolate data using the pan-genome genes has the best overall

Table 2
Validation dataset performance evaluation of the Majority Voting classifier pre-trained on the national surveillance and GenomeTrakr datasets and different genomic input levels.

	National surveillance		GenomeTrakr	
	Virulence level	Pan-genome level	Virulence level	Pan-genome level
F1	0.71	0.76	0.68	0.73
Correctly predicted (out of 101 isolates)	71	76	68	73
False negative (FN)	24	16	18	17

performance (F1 = 0.76). It was able to classify 76 out of 101 isolates correctly. This model is followed by the model trained on GenomeTrakr data using pan-genome features (F1 = 0.73), national data and virulence genes (F1 = 0.71), and lastly, GenomeTrakr data using virulence genes (F1 = 0.68) with 73, 71, and 68 correctly classified isolates respectively. Similarly to the national dataset, models trained using pan-genome features seem to have increased performance over the models trained on the same dataset but using virulence features.

For many ML applications, it is favourable to classify different classes equally well. However, in a food safety context, false negatives (FN), i.e., isolates that are highly virulent but predicted to be hypovirulent are of more prominent concern. Looking at the false negative values for the four models, we obtain a similar ranking as for the F1 performance measure (see Table 2, Table S5). The model trained on national data using pan-genome features still performs best (FN = 16), followed by the model trained on GenomeTrakr data using pan-genome features (FN = 17). The only difference is that the model trained on GenomeTrakr data on a virulence gene level (FN = 18) outperforms the model trained on national data using virulence genes (FN = 24).

4. Discussion

In this study, we aimed to build a well-performing Machine Learning model to predict the clinical frequency (i.e., virulence potential) of *L. monocytogenes* on a sub-species level. To do so, we compared the predictive performance of six different ML algorithms on two distinct WGS datasets from i) national surveillance programs and ii) a publicly available database. In addition, we explored the influence of different genomic levels (i.e., virulence genes, pan-genome genes, and SNPs) and direct read raw read alignment against the reference gene databases on the ML models.

Overall, all six ML models yield good performances on both datasets. For the national surveillance data, the results showed only little differences in the performance of the different ML models. This indicates that all models could capture parts of the underlying connection between the selected genomic features and the clinical frequency. Interestingly, we found increased performance when predicting with pan-genome information compared to using known virulence genes or SNPs to train the models. We can actually see that several of the virulence genes are present in all of our isolates regardless of their clinical frequency, hence reducing the value of these genes as predictive features. Another possible explanation might be that, by using pan-genome genes, we are capturing more complex patterns of the underlying relationships between gene features and virulence. Additionally, this could potentially indicate the presence of genes with currently unknown relationships to virulence in the pan-genome.

The pan-genome genes selected in the feature filtering process can vary between the repeated training and test splits. Extracting the filtered genes from the 30 repeated splits, we obtained 282 pan-genome genes in total. A detailed distribution of these genes over the 30 splits can be found in Table S6. To further investigate which pan-genome genes were preferentially used by the ML to predict the clinical frequency, we extracted the pan-genome genes used by the best performing model for each algorithm. This resulted in a subset of 224 pan-genome genes. Around 90 % of the genes in the subset were classified as hypothetical genes without a known function. However, for the remaining 10 % (n = 22), Prokka returned an annotation. Interestingly, some of these genes were possibly linked to virulence in *L. monocytogenes*. For example, *efeN*, *efeM*, and *efeU* are genes related to iron uptake, which plays a vital role in many cellular processes (Begg, 2019; Jesse et al., 2014). Additionally, metal ions were also found to play an active role during microbial pathogen infection and in the host defence mechanism (Hood and Skaar, 2012). Recent studies suggest that there is a link between antimicrobial resistance and virulence (Guillard et al., 2016; Scortti et al., 2018). Within the gene set, we could find multiple genes related to drug resistance, e.g., *blaR1*, *blaI*, *marA*, *fosB*. The filtering also selected *rmlD* and *rmlC* genes. It has been hypothesised that *ramlD* might be co-expressed with *fosX*, a homolog of *fosB*. This particular gene arrangement has been described as unique to *Listeria* (Scortti et al., 2018). Further, the *rmlABCD* locus in *L. monocytogenes* is involved in L-rhamnosylation of wall teichoic acids (WTAs), which has been described to contribute to in vivo virulence in mice (Carvalho et al., 2015). Apart from the possible link between antimicrobial genes and virulence, increased antimicrobial tolerance can also be beneficial for overall *L. monocytogenes* persistence, which increases the chances of a subsequent infection. Some of the filtered pan-genome genes with known annotation we could not link directly to pathogenicity, i.e., *essD*, *essD_1*, *queC*, *esxB*, *wapA*, *wapA_4*, *TatAy*, *tatC2*, and *ybiA*. However, a comparative genomic analysis of a highly hypervirulent isolate XYSN and EGD-e identified *wapA* as one of the genes that have previously not been observed in *L. monocytogenes* (Yin et al., 2019). Surprisingly, *arnT* was also found in the filtered set. The *arnT* has been described to be involved in the modification of lipid A which is suggested to infer polymyxin resistance (Needham and Trent, 2013; Tavares-Carreón et al., 2016). Lipid A is a significant component of Gram-negative bacteria's outer cell wall. However, *L. monocytogenes* is a gram-positive bacteria. Additionally, we aligned the 224 preferentially used pan-genomes to *L. monocytogenes* pathogenicity islands LIPI-1, LIPI-3, and LIPI-4. We were able to find five of the eight genes that comprise LIPI-3 (i.e., *llsB*, *llsD*, *llsH*, *llsX*, *llsY*).

Generally, it is very interesting to look at the features which guide the learning of ML models. However, at this point, it is essential to emphasize that one should be cautious when concluding from the filtered features reported in this study and their importance to virulence. The set of features may seem reasonable for virulence prediction; however, we only work with a limited amount of data, making it difficult to generalize. Additionally, the low number of samples limits the choice of feature selection methods. In general, there is a lack of consensus about how to properly address the feature filtering for this type of studies. Possible alternative filtering methods would be according to multicollinearity between features or correlation of features with the outcome (i.e. clinical frequency). However, further research is needed to assess the difference between methods. Our feature selection method is driven by variation in the absence/presence of genes which does not necessarily concord with the importance of particular genes to virulence mechanisms. Hence, to be able to draw more conclusive and possibly casual relationships, the feature candidates would need to be further tested in vivo/vitro.

Even though we were able to obtain ML models with great predictive performance, there are some limitations to our study. One of the major limitations is the availability of epidemiologically balanced data. Such data is needed to train an accurate ML model using the described methodology. To tackle this issue, we explored publicly available data in the framework of the GenomeTrakr network. Our study shows that the predictive performances of our Machine Learning algorithms benefit from an increased amount of data. This becomes apparent through the greatly increased performance across five out of six algorithms and overall narrower 95 % confidence intervals. Further, it can also be seen

that the tree-based algorithms seem to immensely benefit from additional data. The Support Vector Machine with a linear kernel was the only algorithm for which we did not observe a similar trend (see Table 1).

The ability to generalize to new data is an important aspect when evaluating the performance of trained Machine Learning models. This assessment is often done on a subset of samples held out during model training or preferably on an independent validation set. All our pre-trained machine learning models reported good performances on the independent validation set. However, we can see that there are differences in the generalized nested cross-validation performance of the models based on the training datasets and the model performance on the independent validation dataset. Different factors could cause this divergence. Even though a few countries have already implemented thorough *L. monocytogenes* WGS surveillance initiatives, acquiring the data remains to be an obstacle. Therefore, our study is limited to data from three countries (i.e., Denmark, France, and the USA). This may limit the applicability of our models to different geographical contexts. Our model might have issues correctly predicting virulence in isolates from other geographical regions as they were not included during training. Another possible source of disagreement might be the different ways of assessing virulence. The laboratory studies used for validation generally studied the effect of isolated mechanisms or factors, e.g., cell invasion or cell adhesion, as estimators for virulence. The clinical frequency used in our study encompasses virulence in a broader context. It describes the overall clinical risk, which can be influenced by many factors such as infectious dose, persistence in the food processing environment, and invasion ability. Hence, it might be challenging to compare these measures of virulence directly. All of these factors might affect the prediction ability of our models on the validation set resulting in relatively high false negative rates (i.e., prediction of isolates with high virulence as low virulent). These miss classifications could have major implications on public health. Hence, further assessment of the factors that lead to these misclassifications and additional refinement of the models is needed in the future.

WGS analysis already finds broad application in surveillance and source tracking and could be extended to sub-species hazard characterization. However, these techniques bring new challenges such as an increased demand for computational resources and the need for specially trained workforce to analyse the results which is generally not available in the industry. Hence, there would be a need for user-friendly tools/pipelines that can process the data and output easily interpretable results.

Currently, FAO and WHO recommend the global control of *L. monocytogenes* without consideration of sub-species variation. However, they allow the use of subtype information for risk management in some countries (FAO and WHO, 2022). As more knowledge is gathered on hazard differences of *L. monocytogenes*, it might become more interesting for future risk assessments and risk management procedures to also consider within species variation.

For example in a recent meeting report, FAO and WHO (2022) proposed how sub-species information could be used to rank *L. monocytogenes* isolates regarding their virulence. We can imagine that predictive ML models, similar to those described in our study, might be a beneficial addition to the proposed multifactorial characterization methods. As more research is conducted, similar hazard characterization models could maybe guide risk mitigation actions in case of *L. monocytogenes* contamination in the future. A possible scenario might be that risk management authorities could evaluate, in consideration of hazard characterization, food matrix, and consumer target, if a recall is necessary or not. Nevertheless, further research will be needed to show clearer evidence and evaluate the actual benefit.

5. Conclusion

The findings described in this paper bring us a step closer to microbial hazard characterization by virulence prediction of *L. monocytogenes* on a sub-species level. We were able to build well-performing Machine Learning models which can predict the clinical frequency of our training isolate datasets and an independent validation set. The results did not show a clear performance benefit of one particular Machine Learning algorithm over the others. However, our study indicates an overall benefit of using the pan-genome level. Hence, we suggest that using the Majority Voting classifier trained on a subset of pan-genome genes as features is the best choice for future applications. Furthermore, we found that the more rapid and less computationally intensive raw read alignment yields comparably accurate models as de novo assembly.

As more research is highlighting the within species differences of *L. monocytogenes*, hazard characterization on sub-species level could gain more importance. The predictive models trained in this study could act as part of a multifactorial tool to characterize *L. monocytogenes* risk in the future. As more knowledge is gathered, such tools might be used as guidance during contamination mitigation procedures.

Nevertheless, many factors contribute to a successful *L. monocytogenes* infection of the host. These factors range from cellular mechanisms to the ability to tolerate environmental stress and the resulting persistence in a food context. To facilitate a more detailed risk characterization, it will be crucial to continue to explore virulence and other external factors to obtain a deeper understanding of *L. monocytogenes* risk on a sub-species level.

Declaration of competing interest

None.

Data availability

The French isolate data can be extracted from Maury et al., 2016. The Danish isolate data was deposited in ENA accession no.: PRJEB59720. Code is available on Bitbucket.

Acknowledgement

We would like to thank Mirena Ivanova for the contribution in this study on an independent WGS data with laboratory described virulence potential. In addition, this work was supported by Karl Pedersen og Hustrus Industrifond (DI-2019-07020), Danish Dairy Research Foundation, and the Milk Levy Fund.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijfoodmicro.2023.110491.

References

Andrews, S., 2010. FastQC: A quality control tool for high throughput sequence data [WWW Document]. URL. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 2.10.23).

Anonymous, 2019. Annual Report on Zoonoses in Denmark 2018. Technical University of Denmark, National Food Institute.

Arning, N., Sheppard, S.K., Bayliss, S., Clifton, D.A., Wilson, D.J., 2021. Machine learning to predict the source of campylobacteriosis using whole genome data. PLoS Genet 17 (10), e1009436.

Aytan-Aktug, D., Clausen, P.T.L.C., Bortolaia, V., Aarestrup, F.M., Lund, O., 2020. Prediction of acquired antimicrobial resistance for multiple bacterial species using neural networks. Msystems 5, e00774-19.

Aytan-Aktug, D., Nguyen, M., Clausen, P.T.L.C., Stevens, R.L., Aarestrup, F.M., Lund, O., Davis, J.J., 2021. Predicting antimicrobial resistance using partial genome alignments. Msystems 6, e00185-21.

den Bakker, H.C., Bowen, B.M., Rodriguez-Rivera, L.D., Wiedmann, M., 2012. FSL J1-208, a virulent uncommon phylogenetic lineage IV *Listeria monocytogenes* strain with a small chromosome size and a putative virulence plasmid carrying internalin-like genes. Appl. Environ. Microbiol. 78, 1876–1889.

Begg, S.L., 2019. The role of metal ions in the virulence and viability of bacterial pathogens. Biochem Soc T 47, 77–87.
Briers, Y., Klumpp, J., Schuppler, M., Loessner, M.J., 2011. Genome sequence of *Listeria monocytogenes* Scott A, a clinical isolate from a food-borne listeriosis outbreak. J. Bacteriol. 193, 4284–4285.
Bushnell, B., 2022. BBtools [WWW Document]. URL. https://jgi.doe.gov/data-and-tools/software-tools/bbtools/ (accessed 2.10.23).
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. Bmc Bioinformatics 10, 421.
Camp, P.-J.V., Haslam, D.B., Porollo, A., 2020. Prediction of antimicrobial resistance in gram-negative bacteria from whole-genome sequencing data. Front Microbiol 11, 1013.
Carvalho, F., Atilano, M.L., Pombinho, R., Covas, G., Gallo, R.L., Filipe, S.R., Sousa, S., Cabanes, D., 2015. L-Rhamnosylation of *Listeria monocytogenes* wall teichoic acids promotes resistance to antimicrobial peptides by delaying interaction with the membrane. Plos Pathog 11, e1004919.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16, 321–357.
Chen, J., Xia, Y., Cheng, C., Fang, C., Shan, Y., Jin, G., Fang, W., 2011. Genome sequence of the nonpathogenic *Listeria monocytogenes* serovar 4a strain M7. J Bacteriol 193, 5019–5020.
Clausen, P.T.L.C., Aarestrup, F.M., Lund, O., 2018. Rapid and precise alignment of raw reads against redundant databases with KMA. Bmc Bioinformatics 19, 307.
Collineau, L., Boerlin, P., Carson, C.A., Chapman, B., Fazil, A., Hetman, B., McEwen, S.A., Parmley, E.J., Reid-Smith, R.J., Taboada, E.N., Smith, B.A., 2019. Integrating whole-genome sequencing data into quantitative risk assessment of foodborne antimicrobial resistance: a review of opportunities and challenges. Front Microbiol 10, 1107.
EFSA, 2023. Foodborne zoonotic diseases [WWW Document]. URL. https://www.efsa.europa.eu/en/topics/topic/foodborne-zoonotic-diseases (accessed 2.7.23).
EFSA and ECDC (European Food Safety Authority and European Centre for Disease Prevention and Control), 2021. The European Union one health 2019 zoonoses report. EFSA Journal 19 (2), 6406, 286 pp. https://doi.org/10.2903/j.efsa.2021.
FAO and WHO, 2022. Listeria monocytogenes in ready-to-eat (RTE) foods: attribution, characterization and monitoring – meeting report. In: Microbiological Risk Assessment Series No. 38.
FDA, 2022. GenomeTrakr [WWW Document]. URL. https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-network (accessed 3.1.22).
Gray, M.J., Zadoks, R.N., Fortes, E.D., Dogan, B., Cai, S., Chen, Y., Scott, V.N., Gombas, D.E., Boor, K.J., Wiedmann, M., 2004. *Listeria monocytogenes* isolates from foods and humans form distinct but overlapping populations. Appl Environ Microb 70, 5833–5841.
Guillard, T., Pons, S., Roux, D., Pier, G.B., Skurnik, D., 2016. Antibiotic resistance and virulence: understanding the link and its consequences for prophylaxis and therapy. Bioessays 38, 682–693.
Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C.A., Frimodt-Møller, N., Aarestrup, F.M., 2014. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol 52, 139–146.
Hood, M.I., Skaar, E.P., 2012. Nutritional immunity: transition metals at the pathogen–host interface. Nat. Rev. Microbiol. 10, 525–537.
Hurley, D., Luque-Sastre, L., Parker, C.T., Huynh, S., Eshwar, A.K., Nguyen, S.V., Andrews, N., Moura, A., Fox, E.M., Jordan, K., Lehner, A., Stephan, R., Fanning, S., 2019. Whole-genome sequencing-based characterization of 100 *Listeria monocytogenes* isolates collected from food processing environments over a four-year period. Msphere 4, e00252-19.
Ireland, U.C., for F.S., Dublin, Hoorde, K.V., Butler, F., 2018. Use of next-generation sequencing in microbial risk assessment. Efsa J 16, e16086.
Jagadeesan, B., Gerner-Smidt, P., Allard, M.W., Leuillet, S., Winkler, A., Xiao, Y., Chaffron, S., Vossen, J.V.D., Tang, S., Katase, M., Mcclure, P., Kimura, B., Chai, L.C., Chapman, J., Grant, K., 2018. The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol 79, 96–115.
Jensen, A., Willia, D., Irvin, E.A., Gram, L., Ith, M.A., 2016. A processing plant persistent strain of *Listeria monocytogenes* crosses the fetoplacental barrier in a pregnant guinea pig model. J Food Protect 71, 1028–1034.
Jesse, H.E., Roberts, I.S., Cavet, J.S., 2014. Chapter three metal ion homeostasis in *Listeria monocytogenes* and importance in host–pathogen interactions. Adv Microb Physiol 65, 83–123.
Kasalica, A., Vuković, V., Vranješ, A., Memiši, N., 2011. *Listeria monocytogenes* in milk and dairy products. Biotechnology Animal Husb 27, 1067–1082.
Larsen, M.V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D.W., Aarestrup, F.M., Lund, O., 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol 50, 1355–1361.
Larsen, M.V., Cosentino, S., Lukjancenko, O., Saputra, D., Rasmussen, S., Hasman, H., Sicheritz-Pontén, T., Aarestrup, F.M., Ussery, D.W., Lund, O., 2014. Benchmarking of methods for genomic taxonomy. J Clin Microbiol 52, 1529–1539.
Lees, J.A., Mai, T.T., Galardini, M., Wheeler, N.E., Horsfield, S.T., Parkhill, J., Corander, J., 2020. Improved prediction of bacterial genotype phenotype associations using interpretable pangenome-spanning regressions. mBio 11, e01344-20. https://doi.org/10.1128/mBio.01344-20.
Liu, D., Lawrence, M.L., Ainsworth, A.J., Austin, F.W., 2005. Comparative assessment of acid, alkali and salt tolerance in *Listeria monocytogenes* virulent and avirulent strains. Fems Microbiol Lett 243, 373–378.
Maury, M.M., Tsai, Y.-H., Charlier, C., Touchon, M., Chenal-Francisque, V., Leclercq, A., Criscuolo, A., Gaultier, C., Roussel, S., Brisabois, A., Disson, O., Rocha, E.P.C., Brisse, S., Lecuit, M., 2016. Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. Nat Genet 48, 308–313.
Maury, M.M., Bracq-Dieye, H., Huang, L., Vales, G., Lavina, M., Thouvenot, P., Disson, O., Leclercq, A., Brisse, S., Lecuit, M., 2019. Hypervirulent *Listeria monocytogenes* clones' adaption to mammalian gut accounts for their association with dairy products. Nat Commun 10, 2488.
McLauchlin, J., 1996. The relationship between Listeria and listeriosis. Food Control 7, 187–193.
McMullen, P.D., Gillaspy, A.F., Gipson, J., Bobo, L.D., Skiest, D.J., Freitag, N.E., 2012. Genome sequence of *Listeria monocytogenes* 07PF0776, a cardiotropic serovar 4b strain. J Bacteriol 194, 3552.
Muchaamba, F., Eshwar, A.K., Stevens, M.J.A., Stephan, R., Tasara, T., 2022. Different shades of *Listeria monocytogenes*: Strain, serotype, and lineage-based variability in virulence and stress tolerance Profiles. Front Microbiol 12, 792162.
NCBI, 2022. NCBI's Pathogen Detection Portal [WWW Document]. URL. https://www.ncbi.nlm.nih.gov/pathogens.
Needham, B.D., Trent, M.S., 2013. Fortifying the barrier: the impact of lipid a remodelling on bacterial pathogenesis. Nat Rev Microbiol 11, 467–481.
Neri, D., Antoci, S., Iannetti, L., Ciorba, A.B., D'Aurelio, R., Matto, I.D., Leonardo, M.D., Giovannini, A., Prencipe, V.A., Pomilio, F., Santarelli, G.A., Migliorati, G., 2019. EU and US control measures on *Listeria monocytogenes* and *Salmonella spp.* in certain ready-to-eat meat products: an equivalence study. Food Control 96, 98–103.
Njage, P.M.K., Henri, C., Leekitcharoenphon, P., Mistou, M., Hendriksen, R.S., Hald, T., 2019. Machine learning methods as a tool for predicting risk of illness applying next-generation sequencing data. Risk Anal 39, 1397–1413.
Orsi, R.H., den Bakker, H.C., Wiedmann, M., 2011. *Listeria monocytogenes* lineages: Genomics, evolution, ecology, and phenotypic characteristics. Int J Med Microbiol 301, 79–96.
Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.
Pincus, N.B., Ozer, E.A., Allen, J.P., Nguyen, M., Davis, J.J., Winter, D.R., Chuang, C.-H., Chiu, C.-H., Zamorano, L., Oliver, A., Hauser, A.R., 2020. A genome-based model to predict the virulence of *Pseudomonas aeruginosa* isolates. Mbio 11, e01527-20.
Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., Korobeynikov, A., 2020. Using SPAdes de novo assembler. Curr Protoc Bioinform 70, e102.
Rantsiou, K., Kathariou, S., Winkler, A., Skandamis, P., Saint-Cyr, M.J., Rouzeau-Szynalski, K., Amézquita, A., 2018. Next generation microbiological risk assessment: opportunities of whole genome sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment. Int J Food Microbiol 287, 3–9.
Scortti, M., Han, L., Alvarez, S., Leclercq, A., Moura, A., Lecuit, M., Vazquez-Boland, J., 2018. Epistatic control of intrinsic resistance by virulence genes in Listeria. Plos Genet 14, e1007525.
Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069.
Steele, C.L., Donaldson, J.R., Paul, D., Banes, M.M., Arick, T., Bridges, S.M., Lawrence, M.L., 2011. Genome sequence of lineage III *Listeria monocytogenes* strain HCC23. J Bacteriol 193, 3679–3680.
Tavares-Carreón, F., Mohamed, Y.F., Andrade, A., Valvano, M.A., 2016. ArnT proteins that catalyze the glycosylation of lipopolysaccharide share common features with bacterial N-oligosaccharyltransferases. Glycobiology 26, 286–300.
Wagner, E., Zaiser, A., Leitner, R., Quijada, N.M., Pracser, N., Pietzka, A., Ruppitsch, W., Schmitz-Esser, S., Wagner, M., Rychli, K., 2020. Virulence characterization and comparative genomics of Listeria monocytogenes sequence type 155 strains. Bmc Genomics 21, 847.
Wagner, E., Fagerlund, A., Thalguter, S., Jensen, M.R., Heir, E., Møretrø, T., Moen, B., Langsrud, S., Rychli, K., 2022. Deciphering the virulence potential of Listeria monocytogenes in the Norwegian meat and salmon processing industry by combining whole genome sequencing and in vitro data. Int J Food Microbiol 383, 109962.
Whalen, S., Schreiber, J., Noble, W.S., et al., 2022. Navigating the pitfalls of applying machine learning in genomics. Nat Rev Genet 23, 169–181.
WHO, 2023. Listeriosis [WWW Document]. URL. https://www.who.int/news-room/fact-sheets/detail/listeriosis (accessed 2.7.23).
Yin, Y., Yao, H., Doijad, S., Kong, S., Shen, Y., Cai, X., Tan, W., Wang, Y., Feng, Y., Ling, Z., Wang, G., Hu, Y., Lian, K., Sun, X., Liu, Y., Wang, C., Jiao, K., Liu, G., Song, R., Chen, X., Pan, Z., Loessner, M.J., Chakraborty, T., Jiao, X., 2019. A hybrid sub-lineage of Listeria monocytogenes comprising hypervirulent isolates. Nat Commun 10, 4283.