Aquaculture Reports 20 (2021) 100661

Contents lists available at ScienceDirect

Aquaculture Reports

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

Data imputation and machine learning improve association analysis and genomic prediction for resistance to fish photobacteriosis in the gilthead sea bream

Luca Bargelloni *, Oronzo Tassiello, Massimiliano Babbucci, Serena Ferraresso, Rafaella Franch, Ludovica Montanucci, Paolo Carnier

Department of Comparative Biomedicine and Food Science, School of Agriculture and Veterinary Medicine, University of Padova, 35020, Legnaro, Italy

ARTICLE INFO

Keywords:
Disease resistance
Genomic prediction
Data imputation
Machine learning
Sea bream

ABSTRACT

Disease resistance represents a key trait for breeding programs in aquaculture species. Here we re-analysed 2bRAD sequence data from two experimental challenges of gilthead sea bream with *Photobacterium damsealae piscicida*. Using a high quality reference genome, we carried out variant calling and data imputation with Beagle to obtain a large set of SNPs (80,744). This allowed the identification of eight novel QTLs for resistance to photobacteriosis across different chromosomes and revealed a highly polygenic genetic architecture.

Bayesian regression approaches and machine learning methods (support vector machines and linear bagging) were compared to evaluate relative performance to classify susceptible-resistant individuals. Both data sets showed higher Matthew Correlation Coefficient (MCC) and accuracy values for machine learning methods, particularly linear bagging, with 20–70 % increase in prediction performance. Overall, machine learning methods should be explored in parallel with parametric regression approaches to increase the chances of highly effective genomic prediction.

1. Introduction

Disease resistance is rapidly becoming one of the key traits to be selected for in most advanced breeding programs of aquaculture species (Chavanne et al., 2016; Houston et al., 2020) as the impact of infectious diseases on fish farming is recognized to be extremely relevant in terms of economic losses and animal welfare. Options to mitigate the impact of infections via vaccination, biosecurity, and pharmaceutical interventions are often limited in farmed fish (Houston et al., 2020). Genetic solutions might either complement or substitute such practices, which is particularly important for bacterial diseases, as pharmacological treatments in farmed fish have dramatically increased the presence of antibiotic resistant strains with great environmental risks (Limbu et al., 2020).

Photobacteriosis or fish pasteurellosis is a septicemia caused by the gram negative bacterium *Photobacterium damselae* subsp. *piscicida*, and is considered one of the most dangerous bacterial diseases in farmed fish species due to its wide host range, high mortality rate, and ubiquitous distribution (Andreoni and Magnani, 2014). In the gilthead sea bream, one of the most important species for aquaculture in Europe, photobacteriosis recurrently causes massive mortalities especially in juvenile fish. Breeding for resistance to this disease has long been proposed as a possible strategy to reduce its impact as the trait was shown to have moderate heritability (Antonello et al. 2009). Two recent studies (Palaiokostas et al. 2106, Aslam et al., 2018) have reported that using either genomic Best Linear Unbiased Prediction (GBLUP) or Bayesian methods based on 2bRAD genotyping-by-sequencing data significantly improves the accuracy of predicting phenotypes for resistance to photobacteriosis over pedigree BLUP. Here, we re-analysed these genomic data to test whether data imputation and machine learning (ML) might further increase prediction accuracy. Genotyping-by-sequencing (GBS) data often present missing data randomly distributed across loci and individuals (Robledo et al., 2018) and might particularly benefit from data imputation. The peculiar population structure in aquaculture species, with large families might provide an additional advantage in data imputation (Tsairidou et al., 2020). Finally, the recent availability of the first gilthead sea bream genome assembly (Pauletto et al., 2018) offered the opportunity to further improve both SNP calling and data imputation,

* Corresponding author.
E-mail address: luca.bargelloni@unipd.it (L. Bargelloni).

https://doi.org/10.1016/j.aqrep.2021.100661
Received 31 December 2020; Received in revised form 20 February 2021; Accepted 3 March 2021
Available online 10 March 2021

using the novel genome assembly to guide both processes (Pook et al., 2020).

Machine learning algorithms, which use data to model complex relationships and which improve their prediction performances as data increase, are thought to hold the promise to boost the analysis of genetic and genomic data (Libbrecht and Noble, 2015). The application of ML to genomic prediction in plant and livestock genetics is increasingly popular (Nayeri et al., 2019), albeit it has not been used in aquaculture species yet. Several ML methods have been implemented for predicting either binary or categorical or continuous traits in several farmed species (Nayeri et al., 2019). We tested here on binary survival data a classical supervised method, support vector machine (SVM), which has been extensively used for classification problems and relies on construction of multidimensional hyperplanes that separate similarly labelled objects into linearly separable sets. The second implemented method is linear bagging (LB), an ensemble approach that combines the prediction of multiple ML algorithms, in this case linear models. Results of ML prediction were compared with parametric regression approaches based on Bayesian inference (BayesB, BayesC, and Bayes Ridge Regression).

2. Methods

2.1. Data imputation and Bayesian analysis

Two 2bRAD data sets, which were already reported in Palaiokostas et al., 2016 and Aslam et al., 2018, were re- analysed here. Raw sequence data (PRJNA338774, PRJNA416847) were trimmed and filtered for low sequence quality. Burrows-Wheeler Aligner (Li and Durbin, 2009) was used to map all filtered reads against the sea bream genome (Pauletto et al., 2018). Variant calling was carried out using samtools-mpileup following a standard pipeline (http://www.htslib.org/workflow/#mapping_to_variant). Only single nucleotide polymorphism (SNP) mapping to the 24 sea bream linkage groups/scaffolds were included in all subsequent analyses. Data imputation was performed using Beagle 4.1 with default options (Browning and Browning, 2016). Two vcf files, one for each data set, were obtained, missing data that we were unable to predict and data with a minor allele frequency (MAF) of 1% were filtered out using vcftools (https://vcftools.github.io/).

Cervus version 3.0.7 (Kalinowski et al., 2007) was used to assign parentage of a total of 798 offspring. Cervus uses a likelihood-based approach to assign parental origin combined with simulation of parentage analysis to determine the confidence of parentage assignments (Cervus uses simulation of parentage analysis to evaluate the confidence in assignment of parentage to the most likely candidate parent. As well as using observed allele frequencies the simulation takes account of the number of candidate parents, the proportion of candidate parents sampled, completeness of genetic typing and estimated frequency of typing error when generating genotypes). Likelihood ratios are calculated allowing for the possibility that the genotypes of parents and offspring may be mistyped. Due to the maximum limit of loci that can be analysed by the software, four small datasets of 1500 randomly selected SNPs were created. Each Cervus run consisted of completing an allele frequency analysis, followed by a simulation of parentage analysis where the proportion of candidate parents sampled was set to 98 % and 99 % loci typed with a 1% error rate. A minimum of 500 typed loci were required for progeny to be analysed, and the number of progeny simulated was set to 100,000.

The results of the four independent runs were then compared with a custom script to assess the concordance across independent SNP sets.

For all individuals correctly assigned to a parental pair, we obtained mortality data from Palaiokostas et al., 2016 (PAL16) and Aslam et al., 2018 (ASL18). For PAL16, mortality at day 10 was considered as the threshold for dividing individuals into two classes. For ASL18, mortality at day 10 was already discriminant between susceptible and resistant animals.

Genome-wide association analysis (GWAS) was implemented in GCTA (Yang et al., 2011), using the module GCTA-MLMA with default options. Imputed 2bRAD data were first converted into PLINK (Purcell et al., 2007) format to be processed with GCTA.

The software GCTB (Zeng et al., 2018) was implemented to estimate the genetic architecture (polygenicity) of the trait. Default options were selected with the following MCMC settings: –chain-length 25000, –burn-in 5000, and initial $p = 0.1$.

Parametric regression methods based on Bayesian procedures performing numerical integration through the Gibbs sampler as implemented in the R package BGLR (Pérez and de Los Campos, 2014) were used for the prediction of phenotypes for mortality. Three probit models differing in the prior density used for marker genotype effects (BayesB, BayesC, and Bayes Ridge Regression; Pérez and de Los Campos, 2014) were implemented. In a probit model, the probability of mortality is linked to the linear predictor of a latent variable, the liability, according to the probit function. The liability is modelled through a linear regression on marker genotypes. It is assumed that mortality is observed when the liability exceeds a given threshold. In a Bayesian analysis, a prior probability density for marker genotype effects needs to be specified. Such prior density was a mixture of a point of mass at zero and a scaled-t slab, a mixture of a point of mass at zero and a Gaussian slab or a Gaussian prior for BayesB, BayesC or Bayes Ridge Regression, respectively. Each Bayesian analysis was performed by generating a single Gibbs chain of 200,000 iterations. One Gibbs sample was saved every 100 iterations and the initial 1 000 samples of the saved Gibbs chain were discarded.

2.2. Machine learning analysis

We build two classifiers based on two different machine learning methods: support vector machines (SVM) and linear bagging (LB) classifier. The prediction performance of Bayesian and machine learning methods was assessed with a stratified 12-fold cross-validation procedure for ASL18 and a stratified 10-fold cross-validation procedure for PAL16. In both cases individuals of the same family were assigned either to training, validation or test set. The procedure of assigning all the individuals of the same family to the same cross-validation set was adopted to avoid the risk of overfitting. This has to be taken into account when comparing the results of this study with previous studies which did not adopt this caution and thus might be affected by overfitting. For ASL18 ten training subsets were created (>800 individuals in total), whereas for PAL16 eight training sets were generated (>600 individuals in total). For both data sets, 100 additional individuals formed the blind set. These non-overlapping training and validation sets were used to develop the Bayesian and machine learning classifiers and to evaluate their performance in prediction, respectively. Matthew Correlation Coefficient (MCC) was used to evaluate the quality of the prediction:

- Matthew Correlation Coefficient

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

With
TP=number of true positive
TN=number of true negative
FP=number of false postive
FN=number of false negative

Kernel is the function that transforms input data into a space in which the data is separable. Here we adopted a linear kernel, which is the most appropriate for high-dimensional input spaces. The ML methods that we used are also characterized by the following parameters whose values need to be set. These parameters are: *C* for SVM and *ne* (number of estimators) and *mf* (max feature) and *ms* (max sample) for

LB. The *ms* parameter was set to default value ($ms = 1$), while we chose to set the remaining parameters to the values which maximize the MCC value of the classification on the validation set.

After training and parameter optimization, a final classification step was performed on a third set (test set) of approximately 100 non-overlapping samples for each data set (PAL16 and ASL18) to evaluate the performances of the trained classifiers on data which was not previously seen by the classifier. The performance of the binary classifiers were evaluated using also the following metrics:

- Accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

- True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

- Precision or Positive Predictive Value

$$PPV = \frac{TP}{TP + FP}$$

Both SVM and LB classifiers were implemented in the python library sci-kit learn, using respectively sklearn.svm.SVC and sklearn.ensemble.BaggingClassifier methods (Pedregosa et al., 2011). Data analysis was executed on a server with the following specifics: 12 CPU 2 thread per core, processor Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40 GHz, RAM 252Gb, Hard Disk 16Tb running on a GNU/Linux CentOS system version 7. GNU Parallel was used to parallelize computing procedures for the machine learning methods.

3. Results

The pipeline for variant calling and data imputation yielded a final data set (quality, MAF, imputation) of 80,714 SNPs after all filtering steps (quality, MAF, imputation), across 798 samples for the PAL16 experiment and 1060 individuals for ALS18. Parentage assignment provided highly consistent results across random SNP subsets and confirmed the presence of a highly skewed distribution of family size for the PAL16 experiment (data not shown). Mortality data were re-calculated from the original experiments to obtain binary survival data for all 798 (PAL16) and 1060 (ASL18) individuals that were re-analysed in the present study (Fig. 1). Cumulative mortality at day 10 post-challenge was over 50 % for PAL16 and 38 % for ALS18.

Genome-wide association analysis using GCTA showed no genome-wide significant loci associated to survival after day 10 for PAL16 (Fig. 2), while 15 SNPs showed genome-wide significance for ASL18 (Fig. 3) distributed across nine chromosomes.

The GCTB implementation of a Bayesian method to assess trait genetic architecture failed to reach convergence for PAL16, while for ASL18 a pi value of 0.0017 was obtained, which corresponds to 139 SNPs (SD ± 35) significantly contributing to the trait.

Accuracy and MCC values for prediction of survival after day 10 using either parametric Bayesian methods or ML ones for both data sets are reported in Tables 1 and 2. Overall, higher accuracy and MCC values were achieved using ML methods: Performance of both parametric and non-parametric methods was consistently higher for the ASL18 data set. Optimal values for parameters c (SVM) and ne and mf (LB) were c = 0.01, ne = 500, mf = 0.8 and c = 0.01, ne = 100, mf = 0.8 respectively for PAL16 and ASL18.

4. Discussion

The implementation of a pipeline for variant calling that relies on the use of a high quality reference genome yielded a much higher number of loci (80,714) than previously reported for the same experiments. The final data set in Palaiokostas et al. (2016) counted 12,085 SNPs and the total number of analysed loci in Aslam et al. (2018) was 22,544. In addition to the use of a reference genome, the much larger number of identified SNPs might be explained with the less stringent settings for variant calling and the merging of two sets of data. More stringent options for SNP calling were enforced in the previous analyses because of the lack of a reference genome and the specific nature of 2bRAD data, which are characterized by very short sequence reads. Merging data from different studies becomes increasingly feasible as GBS is routinely applied for genetic analysis of various traits in the same species. Meta-analysis is generally deemed difficult and cumbersome, as it often is, but it might provide additional information at limited cost when properly implemented. The higher number of available loci, however, would have been rather useless in itself, had not be possible to significantly increase the quality of genetic data using imputation. GBS generally provides a large set of genetic variants, without ascertainment bias, because SNP discovery and SNP genotyping occur simultaneously in the same population (Andrews et al., 2016). SNP arrays, on the other hand, are generally developed on one or more reference populations and once designed cannot be easily adapted to new populations. However, due to its intrinsic characteristics, GBS suffers of high rates of missing data. To mitigate such a problem, GBS data imputation has been successfully used in several plant and animal species and data imputation significantly improved genomic prediction (Wang et al., 2020). As

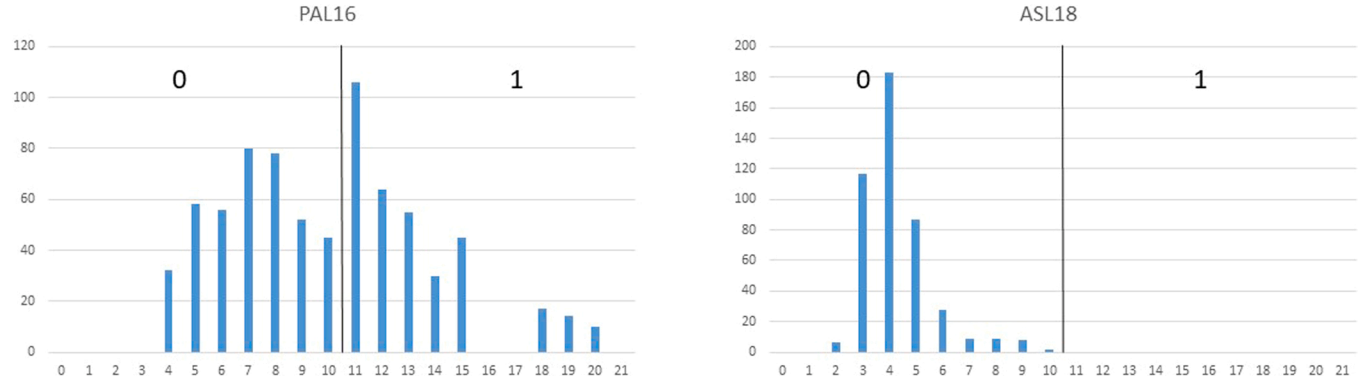

Fig. 1. Daily mortality data for PAL16 (left) and ASL18 (right). On the x-axis days post challenge, on the y-axis, number of dead fish on that day. A vertical line describes the threshold for considering susceptible (0) and resistant (1) individuals. For ASL18 no mortality was observed after day 10.

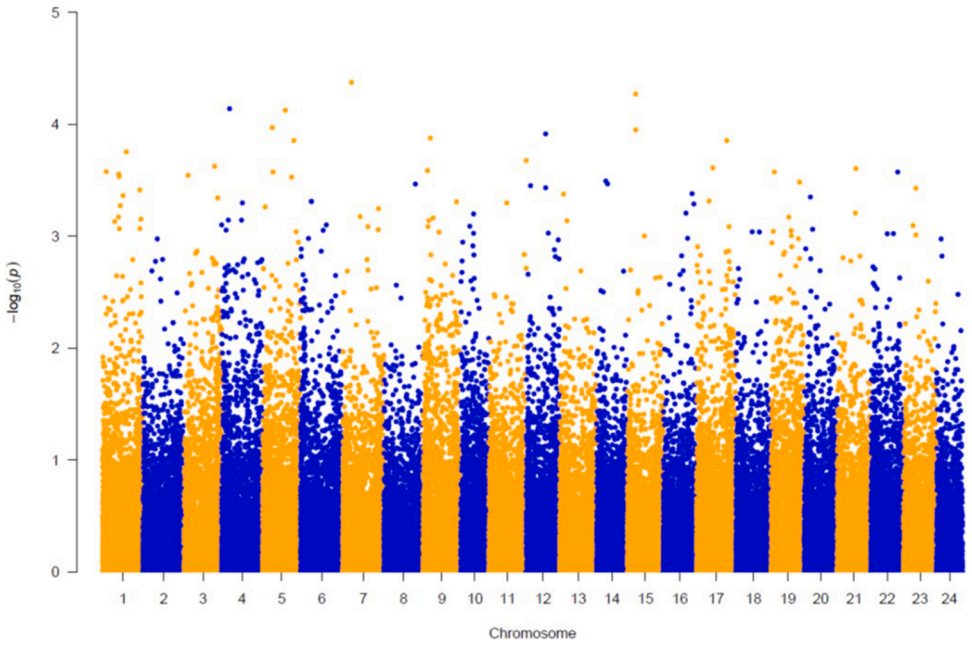

Fig. 2. Manhattan plot describing the results of GWAS for photobacteriosis resistance for PAL16 data. On the x-axis, sea bream chromosomes (Pauletto et al., 2018), on the y-axis the $-\log_{10}$ value of p associated to each SNP.

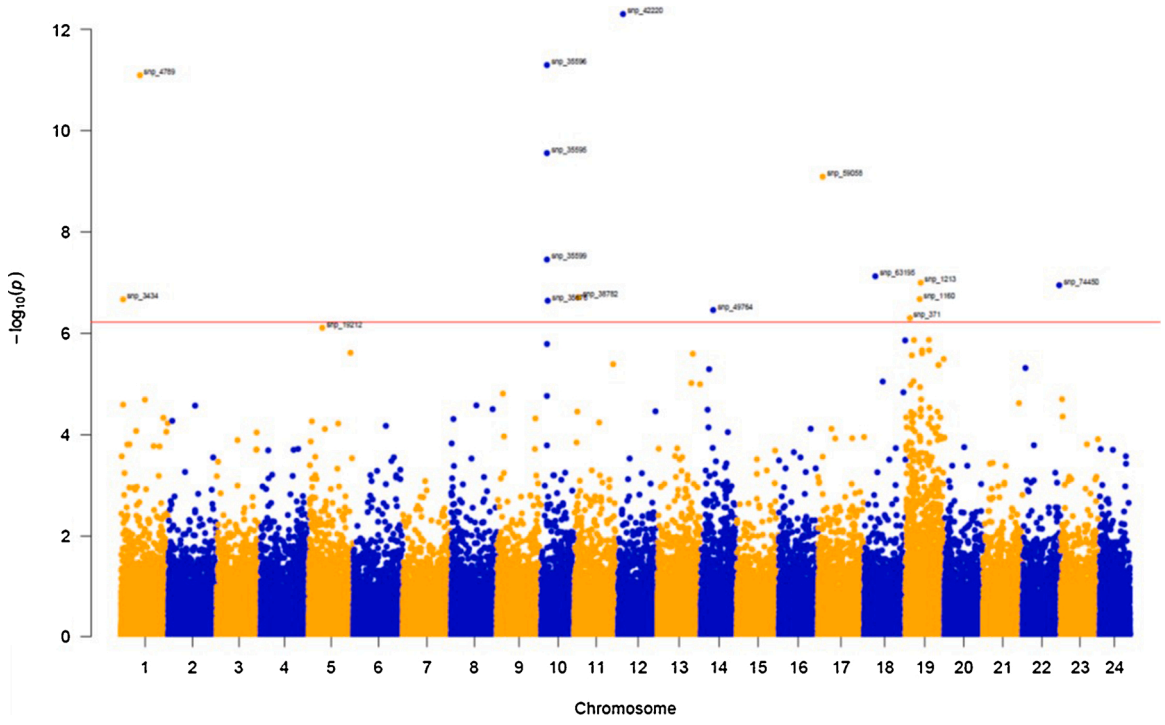

Fig. 3. Manhattan plot describing the results of GWAS for photobacteriosis resistance for ALS18 data. On the x-axis, sea bream chromosomes (Pauletto et al., 2018), on the y-axis the $-\log_{10}$ value of p associated to each SNP.

Table 1
MCC values for phenotype prediction using different methods.

Data set	BayesB	BayesC	BayesRR	SVM	LB
ASL18	0.39	0.38	0.39	0.47	0.48
PAL16	0.11	0.13	0.15	0.18	0.19

Table 2
Accuracy of phenotype prediction using different methods.

Data set	BayesB	BayesC	BayesRR	SVM	LB
ASL18	0.72	0.72	0.72	0.76	0.77
PAL16	0.56	0.56	0.57	0.59	0.60

already mentioned, (Houston et al., 2020; Tsairidou et al., 2020), the typical population structure in fish breeding programs with large sib/-halfsib families appears particularly favourable for genetic data imputation as large haplotype blocks are expected. On the other hand, in several fish species reproduction occur through mass spawning, where a large number of males and females breed in uncontrolled crosses, but often only very few animals significantly contribute to the progeny. This leads to highly skewed family size distribution, which might determine higher accuracy in data imputation for individuals belonging to large families. Here, this might be the case of the PAL16 data set, which has been already shown to have two dominating half-sib families (Palaiokostas et al., 2016), while the ASL18 population originated from a series of controlled crosses (Aslam et al., 2018). At least for parentage assignment, the potential bias in imputation accuracy for PAL16 was not evident as for both data sets all reconstructed families were fully concordant with previously determined families. The great improvement in the number of genotyped variants implementing genome-based variant calling and imputation for GBS data also suggests that in the future low coverage whole-genome sequencing might replace GBS as a valid alternative to array-based SNP genotyping. The sharp reduction in sequencing costs and increase in throughput (confront Table 1 in Logsdon et al., 2020 for an updated summary of sequencing costs), the optimization of cost- and labour-effectiveness of library preparation, and the improvement in data imputation algorithms might soon make possible to use whole-genome sequencing as a routine genotyping method.

GBS data imputation confirmed the absence of significant QTLs in the PAL16 data set as already reported (Palaiokostas et al., 2016). More importantly, it increased the number of significant QTLs for disease resistance in the gilthead sea bream in the ASL18 data set. While the single genome-wide significant QTL on chromosome 19 reported in Aslam et al., 2018 was confirmed, at least eight additional QTLs were identified. However, resistance to photobacteriosis appears to have a highly polygenic architecture as over 100 loci were estimated to contribute to the trait. This number is likely underestimated as the analysis was based on a relatively limited number of individuals and loci (Zeng et al., 2018). Indirect evidence of polygenicity comes from the results of genomic prediction using parametric methods. BayesB and BayesC showed similar prediction performance as Bayes Ridge Regression, which is equivalent to GBLUP (Goddard, 2009). It has been shown that variable selection models (BayesB and BayesC) perform better than GBLUP when large effect QTLs are present, but are comparable or less performant in case of a large number of small effect loci (Clark et al., 2011). In fact, BayesB and BayesC performed equally to BayesRR on PAL16 data and slightly worse on ASL18 ones (Table 1).

Both parametric regressions methods and ML had consistently high performance in phenotype prediction for ASL18. Such evidence is rather simply explained with the lower trait heritability estimated from PAL16 data ($h^2 = 0.18$) than from ASL18 ($h^2 = 0.54$). In both data sets, however, prediction performance of ML methods was higher than that of parametric regression ones. Matthews Correlation Coefficient is generally considered the most informative index connecting all four measures in a confusion matrix, and it is particularly suited to measure the performance of a binary classifier, in particular when there is a significant size bias between classes. Comparison of MCC values clearly suggested that performance of ML methods are 20–70 % better than that of parametric regression methods for PAL16 data and 20–23 % higher for ASl18 ones. In a recent study (Abdollahi-Arpanahi et al., 2020), several parametric regression and ML methods, including deep learning were compared under different simulated scenarios (moderate heritability, either small or large set of QTN, either presence of only additive effects or presence of non-additive effects). Parametric methods outperformed ML when only additive gene interactions were present. However, if non-additive effects were included ML ensemble methods, in particular gradient boosting showed significantly greater performance, confirming evidence from previous studies (e.g. Howard et al., 2014). The greater performance observed for ML methods, especially LB, in predicting photobacteriosis resistance is likely due to non-additive gene interactions. As the presence of dominance or epistasis might be more frequent than generally thought, it could be useful to implement parametric regression approaches as well as ML and ensemble methods on the same data set and to empirically evaluate the relative performance of each method. While such an "extended" strategy might be more time-consuming, it would ensure a better chance for high prediction performance. In the present study, running times for SVM training were approximately five days per data set, and 10 days for LB. Therefore, it might be reasonable to explore several options. On the other hand, running time depended on the number of features, and for the ensemble method it depended also by the number of estimators used, therefore one should decide which is the right balance between exhaustiveness and time-effectiveness of the analysis.

Finally, neither Bayesian parametric regression methods nor ML classifiers were able to predict phenotypes across data sets. When one algorithm was trained on PAL16 and used to classify ASL18 samples or vice versa, classification performance was null (data not shown). The trait might appear the same, *i.e.* the same host species, the same pathogen, survival at day 10 post infection. However, animal age/size, infection dynamics, and genetic background were different, making accurate prediction across experiments quite unlikely.

Author statement

Luca Bargelloni, Ludovica Montanucci, and Paolo Carnier devised the study and supervised all the analyses. Oronzo Tassiello carried out all machine learning analyses. Paolo Carnier performed Bayesian genomic prediction. Serena Ferraresso, Rafaella Franch, and Massimiliano Babbucci processed all sequence data. Massimiliano Babbucci carried out GWAS. Luca Bargelloni wrote the manuscript. All authors read and commented on the text.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (KBBE.2013.1.2-10) under grant agreement n° 613611 (FISHBOOST).

References

Abdollahi-Arpanahi, R., Gianola, D., Peñagaricano, F., 2020. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet. Sel. Evol. 52 (February 24 (1)), 12.
Andreoni, F., Magnani, M., 2014. Photobacteriosis: prevention and diagnosis. J. Immunol. Res., ID793817
Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17 (February (2)), 81–92.
Aslam, M.L., Carraro, R., Bestin, A., Cariou, S., Sonesson, A.K., Bruant, J.S., Haffray, P., Bargelloni, L., Meuwissen, T.H.E., 2018. Genetics of resistance to photobacteriosis in gilthead sea bream (Sparus aurata) using 2b-RAD sequencing. BMC Genet. 19 (July 11 (1)), 43.
Browning, B.L., Browning, S.R., 2016. Genotype imputation with millions of reference samples. Am. J. Hum. Genet. 98 (January 7 (1)), 116–126.
Chavanne, H., Janssen, K.P.E., Hofherr, J., Contini, F., Haffray, P., Komen, J., Nielsen, E. E., Bargelloni, L., 2016. Aquac. Int. 24 (5), 1287–1307.
Clark, S.A., Hickey, J.M., van der Werf, J.H.J., 2011. Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. 43, 18-10.
Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257.
Houston, R.D., Bean, T.P., Macqueen, D.J., Gundappa, M.K., Jin, Y.H., Jenkins, T.L., Selly, S.L.C., Martin, S.A.M., Stevens, J.R., Santos, E.M., Davie, A., Robledo, D.,

2020. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat. Rev. Genet. 21 (July (7)), 389–409.
Howard, R., Carriquiry, A.L., Beavis, W.D., 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 (Bethesda). 4 (April 11 (6)), 1027–1046.
Kalinowski, S.T., Taper, M.L., Marshall, T.C., 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16, 1099–1106.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25 (July 15 (14)), 1754–1760.
Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16 (June (6)), 321–332.
Limbu, S.M., Chen, L.Q., Zhang, M.L., Du, Z.Y., 2020. A global analysis on the systemic effects of antibiotics in cultured fish and their potential human health risk: a review. Rev. Aquacult. 1–45.
Logsdon, G.A., Vollger, M.R., Eichler, E.E., 2020. Long-read human genome sequencing and its applications. Nat. Rev. Genet. 21 (October (10)), 597–614.
Nayeri, S., Sargolzaei, M., Tulpan, D., 2019. A review of traditional and machine learning methods applied to animal breeding. Anim. Health Res. Rev. 20 (June (1)), 31–46.
Palaiokostas, C., Ferraresso, S., Franch, R., Houston, R.D., Bargelloni, L., 2016. Genomic prediction of resistance to pasteurellosis in Gilthead Sea Bream (*Sparus aurata*) using 2b-RAD sequencing. G3 (Bethesda). 6 (November 8 (11)), 3693–3700.
Pauletto, M., Manousaki, T., Ferraresso, S., Babbucci, M., Tsakogiannis, A., Louro, B., Vitulo, N., Quoc, V.H., Carraro, R., Bertotto, D., Franch, R., Maroso, F., Aslam, M.L., Sonesson, A.K., Simionati, B., Malacrida, G., Cestaro, A., Caberlotto, S., Sarropoulou, E., Mylonas, C.C., Power, D.M., Patarnello, T., Canario, A.V.M., Tsigenopoulos, C., Bargelloni, L., 2018. Genomic analysis of *Sparus aurata* reveals the evolutionary dynamics of sex-biased genes in a sequential hermaphrodite fish. Commun Biol. 1 (August 17), 119.
Pedregosa, et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pérez, P., de Los Campos, G., 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198 (October 1), 483–495.
Pook, T., Mayer, M., Geibel, J., Weigend, S., Cavero, D., Schoen, C.C., Simianer, H., 2020. Improving imputation quality in BEAGLE for crop and livestock data. G3 (Bethesda). 10 (January 7 (1)), 177–188.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81 (September (3)), 559–575.
Robledo, D., Palaiokostas, C., Bargelloni, L., Martínez, P., Houston, R., 2018. Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev Aquac. 10 (August (3)), 670–682.
Tsairidou, S., Hamilton, A., Robledo, D., Bron, J.E., Houston, R.D., 2020. Optimizing low-cost genotyping and imputation strategies for genomic selection in Atlantic Salmon. G3 Bethesda (Bethesda) 10 (February 6 (2)), 581–590.
Wang, X., Su, G., Hao, D., Lund, M.S., Kadarmideen, H.N., 2020. Comparisons of improved genomic predictions generated by different imputation methods for genotyping by sequencing data in livestock populations. J. Anim. Sci. Biotechnol. 11 (January 7), 3.
Yang, J., Lee, S.H., Goddard, M.E., 2011. Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88 (January 7 (1)), 76–82.
Zeng, J., de Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., Powell, J.E., Montgomery, G.W., Metspalu, A., Esko, T., Gibson, G., Wray, N.R., Visscher, P.M., Yang, J., 2018. Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. 50 (May (5)), 746–753.