**Material and Methods**

*Mussel field samples*

*Mussel genotype identification*

*Morphological marker identification*

The morphotype identification was described in details in Katolikova et al (2016). Shortly ++++.

*Statistical methods*

All analyses were performed with functions of R3.5.5 statistic programming language (+++).

*Этот кусок в какое-то другое место*

*Datasets*. We used two data-sets for analyses. The first one, which was denoted as “modelling data-set”, included mussels collected in the freshened area (samples from 9 local populations) and in the area with normal oceanic salinity (8 populations) of the Kola Bay as well mussels collected in the White Sea (24 populations). All mussels from this set were divided into three subsets accordingly to their origin and denoted as “Barents Fresh”, “Barents Normal” and “White” respectively. These three subsets were used for regression models constructing and other statistical comparisons. The second set denoted as “testing data-set” included mussels sampled in other regions of the Barents Sea (Fig ++) with hydrological conditions comparable with conditions of areas where mussels from modelling data set were sampled (3 populations fro freshened area associated with mouth of rivers and 6 populations from areas with normal oceanic salinity). Mussels from the testing data-set were used to check the predictions of regression models.

*Assessment of accuracy of identification in different regions.* Accordingly to logic of our approach we tested the reliability of using the mussel’s morphotype as a test to reveal the presence of M.trossulus (positive outcome, “disease”) in populations mixed with M.edulis (negative outcome, “health”). In practice of medical diagnostics the most powerful approach to the solution of similar tasks was developed in the framework of receiver operating characteristics, or ROC-analysis (+++).

To perform the analysis we coded mussels with E-mrpotype (a marker of negative outcome, i.e M.edulis) as 0 and mussels with T-morphotype (a marker of positive outcome) as 1. Then the ROC-curves were constructed separately for three Subsets of modelling data-set. The ROC-curve is the line in the ROC-space where OY-axis reflected the true positive rate (Sensitivity) and the OX-axis - the false positive rate (1-Specificity). The area under curve (AUC), as a good assessment of the test accuracy (Hajian-Tilaki, 2013), was calculated for each subset using function roc() from the package “pROC” (+++). Actually the AUC in the case of binary test-outcome is the value equal to 1/2(Sensitivity plus Specificity, see table ++ for terms explanation). However this value is not equal to the true AUC, which is calculated when test-outcome is continuous (see Muschelli, 2019 for critic analysis), there is no difficulty to use this value to assess the accuracy in the case of three identically treated subsets. Confidence intervals for AUC were calculated with the ci.auc() function and pairwise comparisons of AUCs were performed with the roc.test() function from the same package basing on the methods proposed in DeLong et al. (1988). Bonferroni correction of p-values were applied in the case of multiple pairwise comparisons.

*Association between frequency of T-morphotype and M.trossulus prevalence in populations (Moel 1).* We calculated the proportion of mussel with T-morphtype (“PT”) and proportion of M.trossulus (“MTprev”) among mussels collected in each population included in the modelling data-set. The “MTprev” value was modeled as a function of PT (continuous predictor) and “Subset” (discrete predictor with three levels) and interaction between them. The binomial distribution of MTprev was supposed and model was fitted as logistic regression one with logit link-function. The function glm() from the package “stats” (++) was used for the model fitting. Here and thereafter we checked the method assumptions by the visual analysis of residual-plots and checking overdispersion presence. After the full model (included all predictors and their interactions) was constructed it was simplified accordingly to stepwise backward model selection protocol (++). The model with lowest Akaike information criterion (AIC) was considered as the final one. The function drop1() from the package “stats” was used for the model simplification.

*Assessment of positive prediction value (PPV) and negative prediction value (NPV).* *PPV* in terms of probability theory is the conditional probability to identify a randomly taken mussel of T-morphotype as M.trossulus (*P(MT|T)*). Correspondingly *NPV* is the conditional probability to identify a randomly taken mussel of E-morphotype as M.edulis (*P(ME|E)*). Accordingly to Bayes theorem *PPV* in a population is dependent on *MT*-prevalence in the population (analogously for *NPV*). The dependence is reflected by formulas as follow.

PPV = MTprev Sen/[MTprev Sen + (1-MTprev)(1-Spe)]

NPV = (1-MTprev)Spe/[(1-MTprev)Spe + MTprev (1-Sen)]

We assessed this PPV and NPV by two ways. The first approach was based on the construction of theoretically expected curves, reflecting dependency of PPV and NPV on MTprev (free varying from 0 up to 1) on the basis of assessment of sensitivity and specificity calculated for each Subset of modelling data-set.

The second approach was based on the regression analysis (Model 2). We coded all mussels with congruence of T-morphotipe and MT-genotype or E-morphotype and ME-genotype as 1, and as 0 in all other cases. So the “one” means correct identification of mussel species using its morphotype (we denoted this variable as “Congr”). Further we constructed the logistic regression model with mixed effects (GLMM, Zuur et al, +++) which described the association between the probability of correct identification (i.e. positive outcome of Congr) and predictors: morphotype (“Morph”, discrete predictor with two levels), M.trossulus prevalence observed in each population (MTprev, continuous predictor), Subset (discrete predictor with three levels) and all possible interactions between predictors. Population (“Pop”) was included into model as random factor influencing the model intercept. The model was fitted using glmer() function from the package “lme4” (++++). After the full model was fitted it was simplified accordingly to backward selection protocol. Model with minimal AIC was considered as final model.

*Testing of models predictions.* We used testing data-set to compare predictions of the regression model constructed vs data observed. Using the data on PT in populations from testing data-set we calculated values of MTprev predicted by the Model 1 for each testing population and compared them with observed data by visual analysis of scatter diagram. The correspondence between predicted and observed values was considered as appropriate if points on the diagram were scattered along Y=X line.

For testing of Model 2 prediction we took the MTprev predicted by the Model 1 for each testing population. Then using these values along with data on Morph and Subset from testing data and coefficients from the Model 2 we calculated the probability of correct identification for each mussel. Then for each mussels of E-morphotype we substituted their predicted probability of correct identification as M.edulis by the value equal to 1 minus the predicted value. For mussels with T-morphotype we left the predicted value unchanged. Thus the new value obtained reflects the probability of being identified as M.trossulus both for mussels of T- and E-morphotypes. After all we calculated Spearman correlation coefficient between the values calculated on the previous step and structure-indexes reflecting the probabability of M.trossulus alleles presence. The coefficients were calculated separately for Barents Fresh and Barens Normal subsets. The p-value for correlation coefficients obtained were calculated using permutation test performed with perm.cor.test() function from the package “jmuOutlier” (+++). Bonferroni correction was applied for the two correlation analysis performed for two subsets.

Results

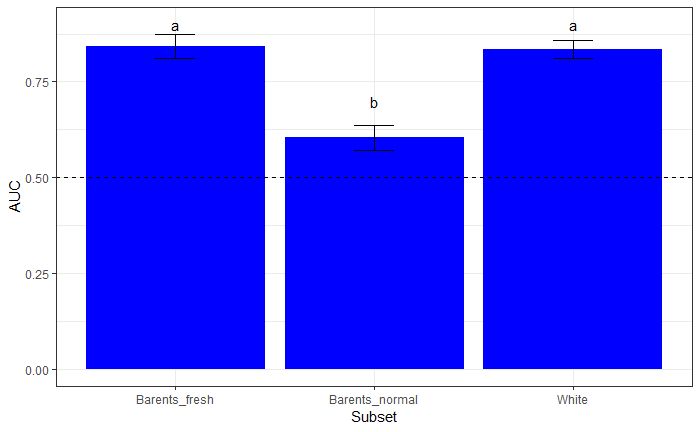


Fig. +. Area under ROC-curve for different mussels subsets. Different letters above bars indicate significant difference between AUC. Whiskers represent 95% confidential interval.

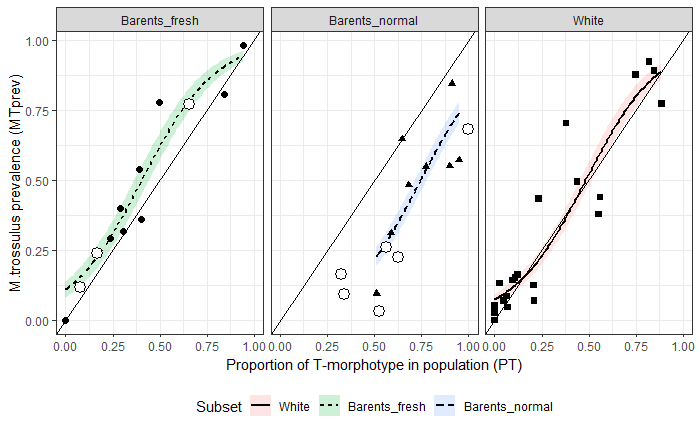


Fig. +. Association between proportion of T-morphotype (PT) and proportion of M.trossulus (MTprev) in three Subsets. Filled points represents observed values in populations from modelling data-set. Empty large points - populations from testing data-set. Diagonal represents Y=X line. Curves represent logistic regression lines with 95% confidence band accordingly to Model 1.

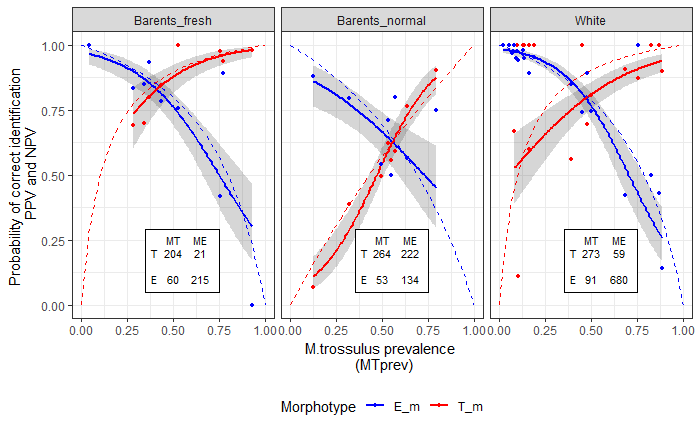


Fig. +. Probability of correct identification of individual mussels using their morphotype (mussels of T-morphotype identifiead as M.trossulus and E-morphotype as M.edulis). Blue solid curves - logistic regression fitted for E-morphotypes predicted by the Model 2, solid red curves - the same for T-morphotypes; Small red and blue dots - observed proportion of correct species identification for two morphotypes in particular populations from modelling data-set. Tables inside the graph represents distribution of mussels with positive (T) and negative (E) tests for M.trossulus (MT) and M.edulis (ME) genotypes. Dotted lines - theoretical curves for PPV (red) and NPV (blue) calculated on the basis of sensitivity and specificity for each subset as functions of M.trossulus prevalence.

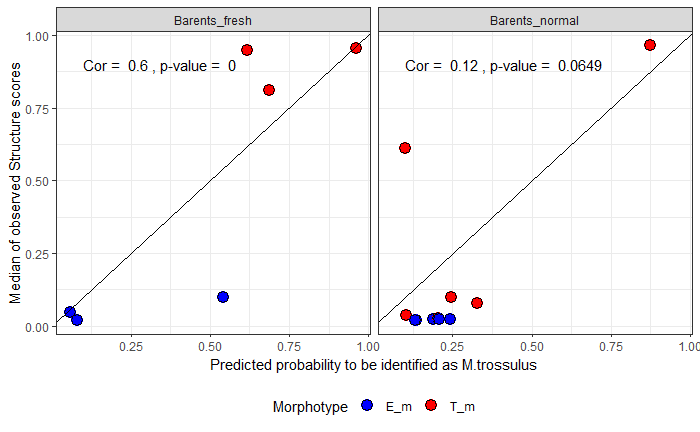


Fig. +. Correspondence between probability to be identified as M.trossulus predicted by Model 2 and observed structure scores which represented as median value for mussels with given predicted values. Prevalence of M.trossulus (MTprev) which is the predictor for Model 2 was calculated on the basis of T-morphotype proportion in testing populations accordingly to Model 1. Diagonal - Y=X line. Spearman correlation between predicted values and observed structure scores are given (p-values are obtained by permutation test with subsequent Bonnferroni correction).