**Materials and methods**

***Study area and datasets***

Altogether, 58 без Америки samples were included in the analyses: 24 previously studied samples from Kandalaksha Bay of the White Sea (Katolikova et al., 2016, total sample size N=1100), 2 from Western Norway (Vainola, Strelkov, 2011, N=120), 4 from the entrance to the Baltic Sea (Strelkov et al., 2017, N=340) and new samples - 26 from the Barents Sea coast of Kola Peninsula (N=1650) and 2 sample from Loch Etive, Western Scotland (N=160). The material was collected in 1987?–2016. Samples were taken at the intertidal and subtidal zone, each sampling area was about 5 m2 (see ESM table 1 for details).

Samples from the Kola Peninsula coast were grouped into regional subsets based on the following considerations. The Kola Peninsula is washed by the Barents Sea and the White Sea. *M. edulis* and *M. trossulus* co-exist and hybridize in Kandalaksha Bay of the White Sea, Kola Bay in the Barents Sea and few localities? along the Barents Sea open coast (REF). However, these water areas have different salinity ranges: salinity of surface water in Kandalaksha Bay of the White Sea is low (below 25‰, REF) while in the Barents Sea salinity reaches 34‰ (REF). Kola bay of the Barents Sea is characterized by variable salinity conditions: in the top of the bay salinity is low (as in the White Sea), but in the mouth of the bay salinity is high (above 25‰). Salinity in sampling localities was either taken from literature (REF) or, in case of few open coast localities was predicted basing on the presence/absence of large rivers nearby (see ESM table 1 for details). And, conventionally, in our study we divided the habitats of mussels into low- and high saline.

For the further staistical analysis we divided samples into two groups: modelling data set and testing data set. The testing data set included 6 samples from the Barents Sea open coast outer of Kola bay (3 samples from low-saline areas and 6 samples from high-saline areas) and 4 samples from the White Sea. The later were selected from total set of White Sea samples as populations with proportion of M.trossulus close to conditionally chosen values 20, 40, 60 and 80%. All these samples were not used in any statistical analysis except of assessment of the congruence between predictions of the regression models and the data observed in testing data-set.

The modelling data set was used for most statistical analysis and consisted of samples from the White Sea (20 populations, hereafter “W”-subset), from low-saline areas in the top of Kola Bay (9 populations, “BL”-subset) and samples from high-saline areas in the entrance of Kola Bay (8 populations, “BH”-subset)

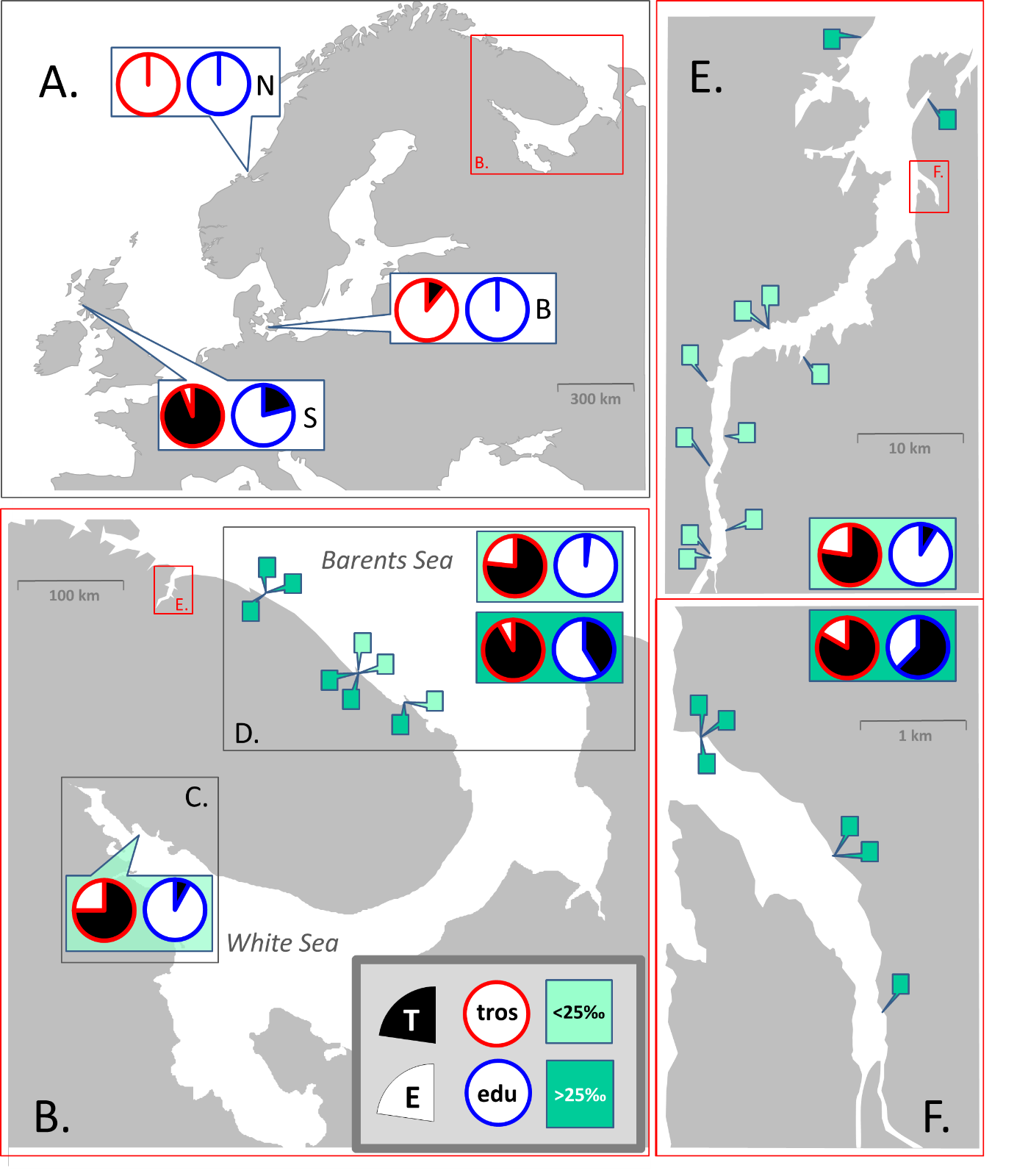
materials ofwere sepearated from all other samples and denoted as a *modelling dataset*. The *modelling dataset* was aimed to evaluate the morphotype-test in the White and Barents Seas materials given the variable salinity and different taxonomic structure of populations.

All other samples from the Barents Sea open coast (3 samples from low-saline areas and 6 samples from high-saline areas, BO-subset) and the White Sea (4 как их изымали? samples from Katolikova et al, 2016 with different taxonomic structure of populations) which were not included into *modelling dataset* were used as *testing dataset*. The aimed to test whether patterns in the *modelling dataset* are validity and general in populations along the Barents Sea coast.

European and American samples were considered as geographical dataset, which was used to assess the possibility to use morphological character for identification *M. edulis* and *M. trossulus* in areas out of the Barents-White Seas hybrid zone.

***Species? identification***

All individuals (N=3370) were genotyped at three or four taxonomically informative allozyme loci: Est-D, Gpi, Pgm, Odh. Estimation of the contribution of *M. edulis* and *M. trossulus* genes into individual genotypes (q-values) using the program STRUCTURE (REF, procedure of Bayesian STRUCTURE analysis as in Katolikova et al., 2016). Genotypes were classified into two categories: mussels with genotypes dominated by *M. trossulus* genes (q-value > 0.5, hereafter *M. trossulus*) and mussels with genotypes dominated by *M. edulis* genes (q-value ≤ 0.5, hereafter *M. edulis*). Hence, potential hybrids were not considered as separate categories but included into *M. edulis* and *M. trossulus*.



**Fig 1. Map of study area and sampling sites.** A. Sampled areas in Europe: Scotland (S), the Baltic Sea (B), Norway (N). Location of the Kola Peninsula is indicated. B. The Kola Peninsula. Location of Kola Bay is indicated. Inserts depict sampled areas in the White Sea (C) and along the open Barents Sea coast (D). E. Kola Bay. Location of Tyuva Inlet is indicated. F. Tyuva Inlet. Pie diagrams depict proportions of T-morphotypes (black sector) and E-morphotypes (white sector) among *M. edulis* (diagrams with blue borders) and *M. trossulus* (with red borders) in regional subsets and geographical dataset. For the Barents Sea, data on low-saline and high-saline areas are presented separately, on the upper and bottom diagrams, correspondingly. Pins depict sampling sites in the Barents Sea. Detailed data are in ESM table 1.

***Morphological marker identification***

The morphotype identification of the White Sea mussels was described in details in Katolikova et al., 2016, and after was applied in Khaitov et al., 2018. We used discrete morphotype classification: T-morphotype (mussel has an uninterrupted dark strip of the prismatic layer under the ligament on the inner side of the shell) and E-morphotype (mussel has an interrupted dark strip or lack dark strip under the ligament due to well developed nacreous layer, see ESM Fig. 1 for details это будут новые фотографии).

***Data analyses***

*Descriptive values*

Taking into account revealed morphological differences between species in the White Sea, we will use term morphotype-test which implies T-morphotype corresponds to *M. trossulus* and E-morphotype to *M. edulis*. Morphotype-test inherently is semi-diagnostic test, so we evaluate its results by comparision with genetically estimated reference results.

Assuming that **a** is a number of *M. trossulus* with T-morphotype in a certain population, **b** - of *M. trossulus* with E-morphotype, **c** – of *M. edulis* with T-morphotype, **d** – of *M. edulis* with E-morphotype, we calculated the values for each sample as follow (see Table 1).

|  |  |  |  |
| --- | --- | --- | --- |
| **abbreviation** | **formula** | **description** | **comments** |
| **population characteristics** | | | |
| **PT** | (a+c)/(a+b+c+d) | Proportion of mussels with  T-morphotype | morphotype proportion in population? |
| **Ptros** | (a+b)/(a+b+c+d) | Proportion of *M. trossulus* | taxonomic structure of population |
| **perfomance mesaures of morphotype-test** | | | |
| **P(T|tros)** | a/(a+b) | Proportion of mussels with  T-morphotype among *M. trossulus* | the ability of morphotype-test to correctly identify *M. trossulus* or  *M. edulis* |
| **P(E|edu)** | d/(с+d) | Proportion of mussels with  E-morphotype among *M. edulis* |
| **Pcorrect** | (a+d)/(a+b+c+d) | Joint proportion of *M. trossulus* with T-morphotype and *M. edulis* with E-morphotype | the ability of morphotype-test to correctly identify *M. trossulus* among  T-morphotype and *M. edulis* among  E-morphotype |
| **P(tros|T)** | a/(a+c) | Proportion of *M. trossulus* among mussels with  T-morphotype |
| **P(edu|E)** | d/(b+d) | Proportion of *M. edulis* among mussels with E-morphotype |

It is worth to mention that these proportions are used in clinical medicine for evaluation of diagnostic tests. If we accept to the view, conditionally, that *M. trossulus* is a “seek” or “bad” mussel (a reasonable assumption taking into account it’s putative invasive nature in some of European seas and its postulated threat to aquaculture, REF) and can be defined as a mussel with disease, and *M. edulis* as mussel without disease. The above mentioned proportions have the next names and properties: **Ptros** is named **prevalence**. **P(T|tros)** and **P(Е|edu)** are named **sensitivity** and **specificity**, **P(tros|T)** and **P(edu|E)** are named **positive predictive value (PPV)** and **negative predictive value (NPV), Pcorrect** is named **accuracy**. Ниже по тексту я пишу Ptros, a не prevalence

*Statistical analysis of modelling dataset*

All analyses were performed with functions of R3.6.1 statistic programming language (REF). Four regression models were fitted for the data obtained from the modelling dataset. We used generalized linear (mixed) models, GL(M)Ms, with binomial distribution and a logit link-function. For each analysis we first constructed the full models (included all predictors and their interactions) and after they were simplified accordingly to stepwise backward model selection protocol (REF). 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. The validity of the final models was visually checked by analysis of residual plots. The goodness of fit for the final models was assessed by the means of pseudo-R2 (REF) using the function r.squaredGLMM() from the package “MuMIn” (REF).

*Model 1: Morphotype proportions as function of taxonomic structure of populations.* All mussels from modelling dataset possessing T-morphotype were coded as 1 and those one with E-morphotype as 0. This data were used as a dependent variable which was regressed against **Ptros** (continuous predictor) and **Subset** (discrete predictor with three levels) and interaction between them. GLM was fitted with glm() function from the package “stats” (REF).

*Model 2:* *Morphotype proportions among species as a function of taxonomic structure of populations.* The dependent variable was codedanalogously to *Model 1* described above and was modelled as function of **Ptros** (continuous predictor), **Genotype или Species**? (discrete predictor with two levels), **Subset** (discrete predictor with three levels) and interaction between them. **Samples** was included into model as random factor influencing the model intercept. GLMM was fitted with glmer() function from the package “lme4” (REF).

*Model 3:* *Accuracy of morphotype-test as a function of taxonomic structure of populations.* The dependent variable was coded as 1(mussels with T-morphotype correctly identify *M. trossulus* and mussels with E-morphotype correctly identify *M. edulis)* and as 0 (an all other cases), and was modelled as a function of **Ptros** (continuous predictor), **Subset** (discrete predictor with three levels) and interaction between them. GLM was fitted with glm() function from the package “stats” (REF).

*Model 4:* C*orrectness of species identification as a function of taxonomic structure of populations*. The dependent variable was codedanalogously to *Model 3* described above The set of predictors for the model was as follow: **Ptros** (continuous predictor), **Morphotype** (discrete predictor with two levels), **Subset** (discrete predictor with three levels) and interaction between them. **Samples** was included into model as random factor influencing the model intercept. GLMM was fitted with glmer() function from the package “lme4” (REF).

*Model 5: Taxonomic structure as a function of morphotype proportions in populations.* The dependent variable was codedanalogiously to *Model 1* described above and was modelled as function of **PT** (continuous predictor) and **Subset** (discrete predictor with three levels) and interaction between them. GLM was fitted with glm() function from the package “stats” (REF).

*Analysis of geographical dataset*

Given the limited data from European populations (incomplete range of taxonomic structure of populations) we described morphotype proportions among *M. trossulus* and *M. edulis* based on pooled samples for each regions. Ну пока мы только так и делали..

*Analysis of theoretical probabilistic models*

Since the abilities of morphotype-test are probabilistic, we believe that taxonomic structure of populations and the correctness of species identification can be predicted using the following formulas:

**= (1)**

**P(MT|T) = (2a)**

**P(ME|E) = (2b)**

Formula 1 derived from law of total probability and implies that Ptros may be predicted based on two conditional probability – P(T|MT)c and P(E|ME)c, and observed PT in population (hereafter, Theoretical model 1). And knowing Ptros, you can calculate next conditional probability – P(MT|T) and P(ME|E), based on Bayes’ theorem (formulas 2a and 2b, hereafter, Theoretical model 2).

The question arose, however, which control samples to use for calculating P(T|MT)c and P(E|ME)c. To answer that question we did next procedures:

1. we combined such Subsets from modelling dataset, where morphotype-test are reliable for identification of species (see Results) and constructed two new Reference models:

*Model 6: Reference model for predicting taxonomic structure based on morphotype proportions in populations.* The modelling approachwas analogous to *Model 5*

*Model 7:* *Reference model for predicting correctness of species identification by morphotype.* The modelling approachwas analogous to *Model 4*

*Зачем-то там..Model 6* vs. *Model 5* and *Model 7* vs. *Model 4* were compared by AIC and BIC

1. we assumed that acceptable control samples are those two samples by which a predictions of Theoretical models are most similar to predictions of Reference models
2. we found all of the possible pairs of samples from combined reference dataset and calculated similarity index of taxonomic structure of populations:

**Similarity index = max(Ptros1;Ptros2)(1 - min(Ptros1;Ptros2) (3)**

Similarity index tends to 1 in case of paired samples have maximal different taxonomic structure of populations (“pure” *M. edulis* population and “pure” *M. trossulus* population), tends to 0.25 in case of paired samples have maximal similar taxonomic structure (two mixed populations, Ptros=0.5), tends to 0 in case of paired samples represented two “pure” *M. edulis* or two “pure” *M. trossulus* populations.

1. for each paired of samples we calculated P(T|MT)c and P(E|ME)c and constructed Theoretical model 1 and Theoretical model 2
2. we evaluated the deviations of Theoretical models predictions from the Reference models using the formula:

**Goodness = 1 / (Σ(Reference prediction- Theoretical prediction)2) (4)**

Goodness tends to 0 if predictions strong deviate and tends to infinity if prediction match.

1. пункт про то, как мы оценивали результат. Точечные диаграммы и сглаживающие кривые?

Having conducted this analysis, we can give recommendations to users about maximal cheap and fast approach for identification species by single character.