**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.

##############

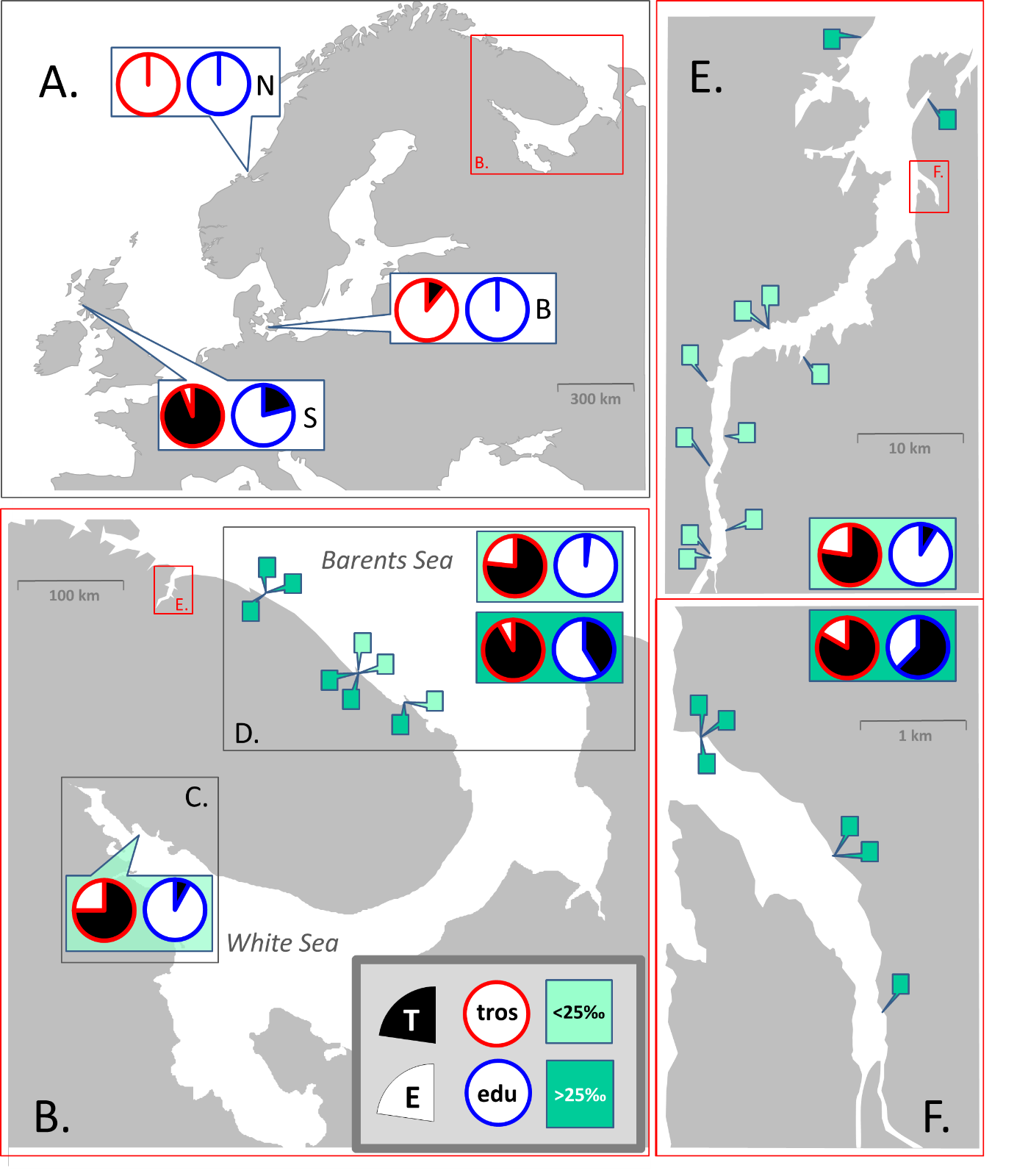
For the further staistical analysis we divided samples into two groups: modelling data set and testing data set. The testing data set included 6 samples from the Barents Sea open coast outer of Kola bay (3 samples from low-saline areas and 6 samples from high-saline areas) and 4 samples from the White 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)

European and American samples were considered as additional, sighting data ~~geographical dataset~~. This data-set was used to assess the possibility of morphological marker applyment 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 using 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) was performed with STRUCTURE software (REF, procedure of Bayesian STRUCTURE analysis as in Katolikova et al., 2016). Using the q-values calculated all individuals 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 was applied, as well, in Khaitov et al., 2018. Briefly, mussels represent two distinct forms: T-morphotype (mussel has underdeveolped nacre under the ligament nympha and dark strip of the prismatic layer could be recognised) and E-morphotype (mussel has well developed nacreous layer under ligment nympha and stripe of naked prismatic layer could not be recognised, 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. This approach is quite similar to the methodology of clinical practice where the presence of one or the other syndrome could be revealed by a set of indirect tests (REF). That is why we use in our work some terminology borrowed from the medical practice the explanation of which is given in the Table ++. Continuing the clinical analogousness we can accept that M. trossulus is a “sick” 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), whereas M. edulis could be considered as "healthy" mussel.

Table +. Terms and values used in the analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

Denotes: a - number of M. trossulus with T-morphotype in a certain population, b - M. trossulus with E-morphotype, c – of M. edulis with T-morphotype, d – of M. edulis with E-morphotype.

The values mentioned have the properties as follow: **Ptros** is **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 R statistic programming language (REF). We constructed four generalized linear models (GLM, but in some cases mixed effect models, GLMM, with binomial distribution and logit link-function) on the base of the modelling dataset. 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 and test for overdispersion. 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).

*GLM-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-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**, **Subset** and **Species**? (discrete predictor with two levels) and interaction between terms. **Samples** was included into the model as random factor.

*GLM-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 identified as *M. trossulus* and mussels with E-morphotype correctly identified as *M. edulis)* and as 0 (an all other cases). The dependent variable was modelled as a function of **Ptros**, **Subset**  and interaction between the terms.

*GLMM-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**, **Morphotype** (discrete predictor with two levels), **Subset** and interaction between the terms. **Samples** was included into model as random factor.

*GLM-Model 5: Taxonomic structure as a function of morphotype proportions in populations.* The dependent variable was codedanalogiously to *Model 1* and was modelled as function of **PT** (continuous predictor) and **Subset** and interaction between them.

In all cases GLMs was fitted with glm() function from the package “stats” (REF) and random intercept GLMMs - with glmer() function from the package “lme4” (REF).

*#################################*

*Analysis of data from other regions*

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..

*Theoretical probabilistic models development*

The only correct way to assess the association between morphotype and genotype and apply morphotype-test in any given particular geographical area is to construct regression models similar to models 4 and 5 proposed above. However we belive that applying of some approaches basing on probability theory could facilitate this work. To use this approach a researcher ought to take some calibrating samples of mussels (30-50 individuals) from the region of interest and genotype them. After genotypes in calibrating samples were identified, two key values should be assessed: the proportion of T-morphotypes among M.trossulus P(T|MT)c and the proportion of E-morphotypes among M.edulis P(E|ME)c. If these values are assessed accurately two theoretical models could be constructed. The Theoretical model 1 (Eq. 1) predicts the probability of M.trossulus presence if we know the proportion of T-morphotype in any sample from the region of interest. The Theoretical model 2 (Eq 2a, b) predicts the probability of correct identification of given specimens as M.trossulus if we know that this specimen possess the T-morphotype (Eq. 2 a) or the probability of correct identification as M.edulis in the case of a mussel with E-morphotype (Eq. 2 b). The derivation of these formulas is given in SEM.

**= (1)**

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

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

The question, however, arise: how to perform rationally the samples for calibrating? To answer this question we pooled together W and BL parts of modelling data set (the rationale for such pooling is given in SEM здесь, в приложении, мы дадим все выкладки про то, почему такое объединение возможно) and used this data to construct two new regression models (hereafter denoting as “reference regression models”).

*Reference model 1* aimed to predict taxonomic structure based on morphotype proportions in populations. *Reference model 2*  aimed to predict the probability of correct species identification by morphotype*.* The structures of these models were analogous to *Model 5* and Model 4 correspondingly but predictor “Subset” was removed from the reference models.

We assumed that acceptable calibrating samples are those one using of which the theoretical models produce the predictions most similar to predictions of reference regression models. Additionally we assumed that calibrating samples should be taken from at least two populations. Two scenario of choosing of calibrating population could be considered. The first one - both populations could be maximally differ in their genetic structure (one population is dominated by M.edulis but another one by M.trossulus). The second scenario - both populations are maximally mixed (proportion of M.trossulus in both populations is close to 0.5). To assess the similarity of genetic structure we used the index represented by Eq. 3. The Similarity index approaches to 1 if proportion of M.trossulus in both populations is differ. When the proportion of M.trossulus is close to 0.5 then the similarity index approaches to 0.25.

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

We found all possible pairs of populations from combined W and BL dataset (++++ possible pairs). For each pair we calculated P(T|MT) and P(E|ME). Next, using these values we calculated predictions of Theoretical model 1 for all possible PT. We represented the possible PT as a vector of numbers from 0 to 1with standard step equal to 0.01. However the number of elements in this vector (N) could be lesser than 101 for particular combinations of populations, since some values of PT may be deprecated due to Eq.1 limitations (see SEM ++++ не знаю как короче сказать, возможно более подробное объяснение надо выводить в SEM). Using the same vector of PT we calculated predictions of Reference model 1. After the predictions of both models were calculated we assessed the goodness of correspondence between Reference model 1 and Theoretical model 1 using the Eq 4. The algorithm described above was applied for each possible pairs of populations. The results were graphically represented as a plot of Goodness against Similarity index.

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

Algorithm analogous to described above was applied for Theoretical model 2 and Reference model 2, but instead of possible PT we considered the vector of possible Ptros.