**Materials and methods**

***Study area and datasets***

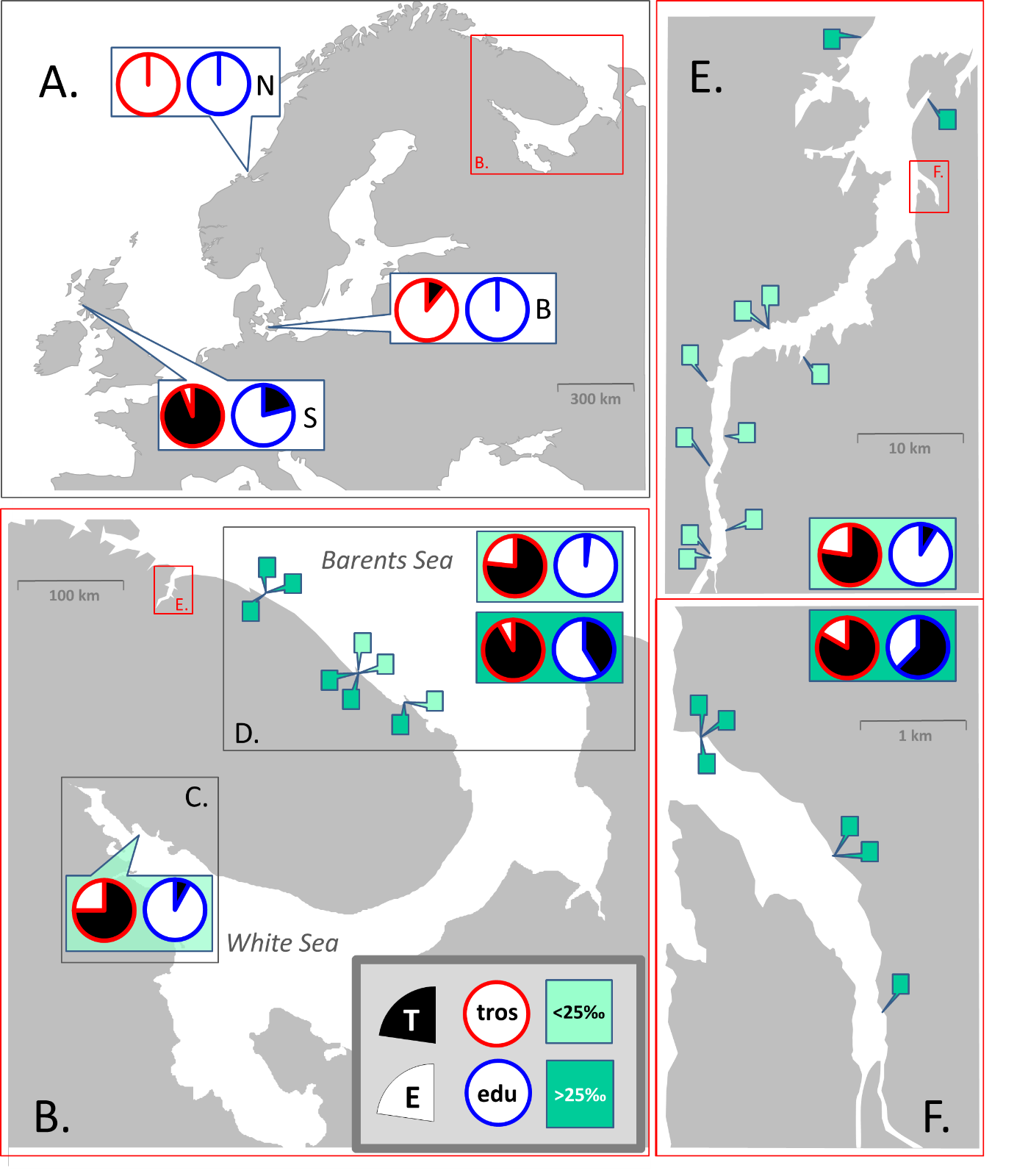
Altogether, 58 samples were included in the analyses: 24 previously studied samples from Kandalaksha Bay of the White Sea (Katolikova et al., 2016, total sample size N=1100), 2 from Western Norway (Vainola, Strelkov, 2011, N=120), 4 from the entrance to the Baltic Sea (Strelkov et al., 2017, N=340) and 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 1987–2019. Mussels were taken..я хз, как там заморских мидий собирали. Но кажется, что точно про наших надо написать про литораль-сублитораль, и про плавучие субстраты (водоросли, буи и тд). Случайные сборы без количественных привязок. Тут ещё не доделала

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. 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: 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 materials of samples from the White Sea (24, W-subset), from low-saline areas in the top of Kola Bay (9, BL-subset) and samples from high-saline areas in the entrance of Kola Bay (8, BH-subset) were sepearated from all other samples and denoted as a *modelling dataset*. The *modelling dataset* was used in the analysis aimed to compare White and Barents Sea materials and to compare low- and high-saline areas. All other samples from the Barents Sea open coast (3 samples from low-saline areas and 6 samples from high-saline areas, BO-subset) which were not included into *modelling dataset* were used as *testing dataset*. The later aimed to test whether patterns in the *modelling dataset* are 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 Sea 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 category 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*

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.

Proportion of mussels with T-morphotype: **PT = (a+c)/(a+b+c+d)**;

Proportion of *M. trossulus*: **Ptros = (a+b)/( a+b+c+d)**;

Proportion of mussels with T-morphotype among *M. trossulus*: **P(T|tros) = a/(a+b)**;

Proportion of mussels with E-morphotype among *M. edulis*: **P(Е|edu) = d/(с+d)**;

Joint proportion of *M. trossulus* with T-morphotypes and *M. edulis* with E-morphotypes:   
**Pfit = (a+d)/(a+b+c+d)**;

Proportion of *M. trossulus* among mussels with T-morphotype: **P(tros|T) = a/(a+c)**;

Proportion of *M. edulis* among mussels with E-morphotype: **P(edu|E) = d/(b+d)**.

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** and evaluates the ability of the morphotype as a test to identify correctly *M. trossulus* or *M.*edulis, correspondingly. **P(tros|T)** and **P(edu|E)** are named **positive predictive value** and **negative predictive value** and evaluate the ability of the morphotype-test to identify correctly *M. trossulus* among T-morphotypes and *M. edulis* among E-morphotype. **Pfit** is named **accuracy** and evaluates the ability of the morphotype-test to identify correctly in both *M. trossulus* and *M. edulis* in a sample, but does not take into account incorrect identification in species subsample. Ниже по тексту я пока 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. 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) and function …….

*Model 1: Morphotype proportion in a samples as function of taxonomic structure of populations.* The **PT** was modelled as a function of **Ptros** (continuous predictor) and Subset (discrete predictor with three levels) and interaction between them. We used beta-regression approach which has proven itself well in the case of proportions as dependent variables (REF). The model was fitted with betareg() function from the package “betareg” (REF) with a logit as a link-function.

*Model 2:* *Morphotype proportions among species as a function of taxonomic structure of populations.* Generalized mixed effect model with binomial distribution of outcome (REF) was used for the analysis. All mussels from modelling dataset posessing T-morphotype were coded as 1 and those one with E-morphotype as 0. This data was used as a dependent variable which was regressed against **Ptros** (continuous predictor), **Genotype или все-таки** **Species**? (discrete predictor with two levels), **Subset** (discrete predictor with three levels) and interaction between terms. **Samples** was included into model as random factor influencing the model intercept.

*Model 3:* *Accuracy of morphotype-test as a function of taxonomic structure of populations.* The accuracy of morphotype-test (**Pfit**) was modelled as a function of **Ptros** (continuous predictor), **Subset** (discrete predictor with three levels) and interaction between terms. We used beta-regression approach analogiously to *Model 1* described above.

*Model 4:* C*orrectness of species identification as a function of taxonomic structure of populations*. All mussels of both T-morphotype identified as *M. trossulus* and E-morphotype identified as *M. edulis* were coded as 1 (correct species identification by morphotype) in all other cases mussels were coded as 0. This data was used as dependent variable for the logistic mixed-effect model (GLMM with a binomial distribution and a logit link-function). 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 terms. **Samples** was included into model as random factor influencing the model intercept.

*Statistical analysis of testing dataset*

All two testing subsets were treated as follow. Using **Ptros** in each population we calculated the predicted values in each of them using the parameters of fixed parts of the *Model 1-4*. Then we compared the observed proportion with the predicted one. Я забыла дописать, что мы с вами в последний раз обсуждали..что в моделях 2 и 4 мы брали средние значения. Точно не помню, как это формулировалось.

*Analysis of geographical dataset*

Given the limited data from European populations (different taxonomic structure wasn't represented) we described morphotype proportions among *M. trossulus* and *M. edulis* based on pooled samples for each regions. Ну пока мы только так и делали..

*Practical recommendations*