**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). Просто перечисление имеющихся данных

Может лучше в тексте просто написать из каких морей взяты выборки, а всю более подробную информацию свести в таблицу.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Area | Number of genotyped mussels / Number of populations sampled | Reference for initial publication of data |
| W | White Sea | 1100/24 | Katolikova et al., 2016 |
| Norv | Western Norway | 120/2 | Vainola, Strelkov, 2011 |
| Balt | Baltic Sea | 240/4 | Strelkov et al., 2017 |
| LEtiv | Loch Etive, Western Scotland | 160/2 | Denovo |
| BF | Freshened area of the Barents Sea |  | Denovo |
| BN | Area of the Barents Sea with normal salinity |  | Denovo |

Здесь нужна некоторая вводная фраза. Например так.

Since previous investigations (REF) revealed the high role of salinity in regulation of ME and MT distribution all populations sampled in the Barents Sea denovo were divvided into two groups. The first group included populations situated in area (Fig. ++) associated with river mouths with low salinity (in fact compareble with salinity in the freshened White Sea). The second one - the group of populations situated in areas with normal oceanic salinity.

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 in the north and the White Sea in the east and southeast. *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: 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.

Здесь я пыталась объяснить, почему мы так делим наш Кольский материал на сабсеты

По мне так лучше перестроить на манер того, что я описал выше.

Здесь надо еще описать технику сбора мидий.

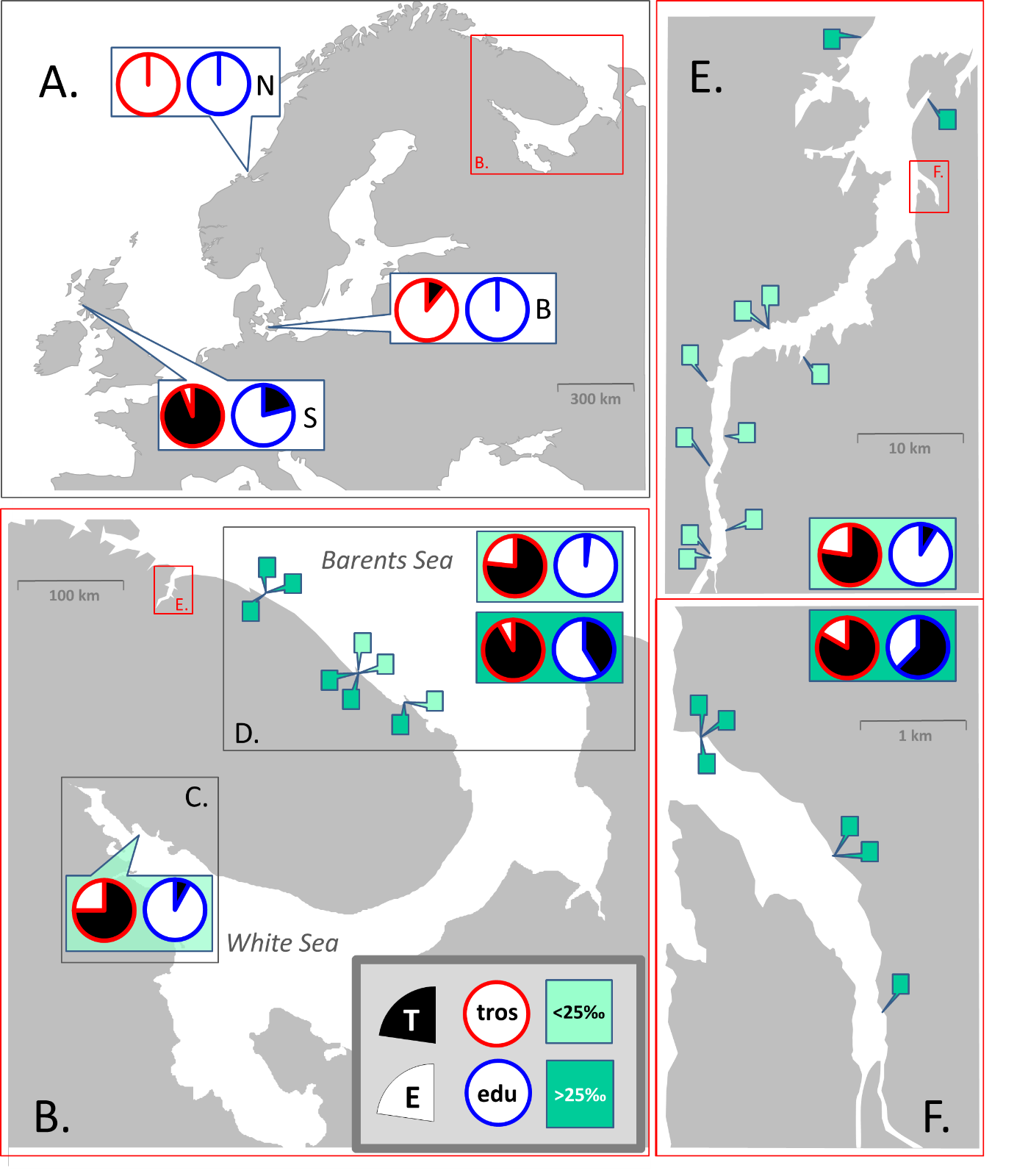
For the further staistical analysis materials of samples from the White Sea (24, W-subset), from low-saline areas in the top of Kola Bay (9, BL-subset) and samples from high-saline areas in the entrance of Kola Bay (8, BH-subset) were sepearated from all other samples and denoted as a *modelling dataset*. 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 which were not included into modelling dataset were used as *testing dataset. The later aimed* to test whether patterns in the modelling set are general in populations along the Barents Sea coast. Тут я типа стараюсь придать тестингу роль не только как стат проверки.

Я бы упростил вот так.

European and American samples were considered as additional *geographical dataset*, which was used to assess wether the patterns revealed in our work can be applied to areas out of Baents-White Sea hybride zone.

А это собственно рассказ, что включают датасеты и зачем они нам нужны



**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. Пока такая подпись..надеюсь, что ещё добавится Америка

*Species identification Надо бы определиться мы говорим о видах или генотипах!*

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*. ПП это ещё не смотрел, может всё обосрёт и заставит переписывать

*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 ваши новые прекрасные фотографии).

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

According to the above-mentioned studies, we assume that T-morphotype is an attribute of *M. trossulus* while E-morphotype of *M. edulis*. НЕТ! Так нельзя! Собственно цель работы показать, что это так.

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

***Data analyses***

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

Все это надо убрать из материалов и методов

The ultimate goal of the analysis was to answer questions about the capacity of a simple morphological test to discriminate *M. edulis* and *M. trossulus* taking into account the possible effect of geography and ecology (salinity conditions?) on morphotype proportions among *M. edulis* and *M. trossulus*, and to develop practical recommendations for mussel taxonomic identification by morphotypes. Ну это типо краткое саммари того, что будет потом описано..

Ниже сформулированы вопросы (в основном из дальневосточной презентации, последний сама сочинила), на которые мы должны ответить в статье, как кажется. Наверное, надо их лучше в конце введения задавать? ДА УЖ точно не здесь

1. Are morphological differences between *M. edulis* and *M. trossulus* persist in populations along the Barents Sea coast of Kola Peninsula?
2. Can taxonomic structure of populations be predicted basing on morphotype proportions in a sample?
3. Is there a variability of morphotype proportions among subsamples of *M. edulis* and *M. trossulus*?
4. How likely is that a randomly taken mussel with T-morphotype is *M. trossulus* and mussel with E-morphotype is *M. edulis*?
5. Is there a possibility of using this method in other geographical regions of Europe?
6. How to calibrate the method of mussel taxonomic identification by morphotypes to study of other hybrid zones taking into minimal expenses on genotyping?

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

Discriptive 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)**; Я уже не помню, почему я назвала это Pfit, можно переименовать ДА, это ведь congr лучше бы какое-то иное название.

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* ~~among~~ *~~M. trossulus,~~* or *M. edulis* ~~among~~ *~~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. Я так и не поняла, о чём мы там вчера договорились насчёт терминов. В формате статьи это будет маленький абзац, пусть остаётся? Сюда надо написать, что в дальнейшем мы именуем долю МТ как prevalence? Ниже по тексту я пока пишу Ptros

*Statistical analysis of modelling dataset*

~~We analyzed and compared distributions of~~ **~~PT~~**~~,~~ **~~P(T|tros)~~** ~~and~~ **~~P(Е|edu), P(tros|T)~~** ~~and~~ **~~P(edu|E)~~**~~, and~~ **~~Pfit~~** ~~as functions of~~**~~Ptros~~** ~~among regional subsets (W-, BL- and BH-subset).~~

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.

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

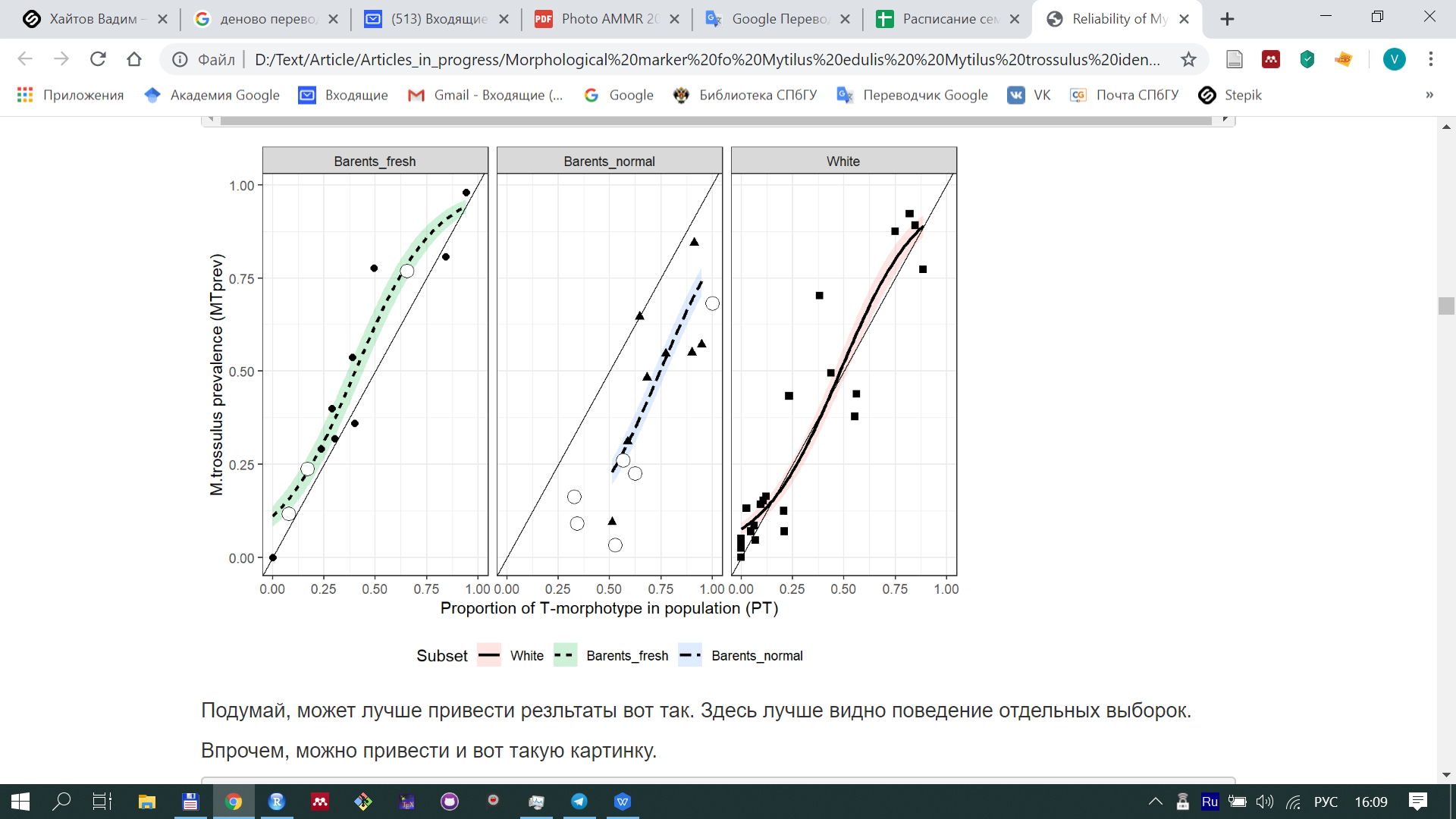
We used a generalized linear models (GLM, the function betareg() from the package “betareg” (REF) was used for the models fitting) and a generalized linear mixed-effect models (GLMM, the function glmer() from the package “lme4” (REF) was used for the models fitting). For each analysis we first constructed 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. Нужны ли нам marginal and conditional pseudo-R2? А также ICC? Здесь я пыталась какие-то общие процедуры вынести в шапку. Если накосячила – поправьте, пожалуйста

*Эту шапку лучше вообще убрать!*

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

The **PT** was modelled as a function of **Ptros** (continuous predictor) and **Subset** (discrete predictor with three levels) and interaction between them to test for the correspondence between morphotype proportions in a samles and taxonomic structure of populations. The dependent variable had a beta distribution (a value from 0 to 1), therefore, we used a GLM with a logit link function.

СТОП! Мне казалось, что все наоборот! Мы знаем долю мидий Т-морфотипа и по ней мы выносим суждение о соотношении видов в популяции. Вот как это было в изначальных анализах. ПП что решил все переиграть? Нахрена такая модель?



По мне так это должно выглядеть вот так. Прочитай внимательно, я многое переделал.

*Model 1: Taxonomic structure of populations as a function of morphotype proportion.*

The **Ptros** was modelled as a function of **PT** (continuous variable) and **Subset** (discrete predictor with three levels) and interaction between the terms. We used beta-regression approach which has proven itself well in the case of proportions as dependent variables (REF). The model was fitted with ++++ function from the package ++++ (REF) with a logit as a link-function .

Model 2: *Morphotype proportions 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.

~~. To assess the morphological differences and potential? variability of morphotype proportions among of~~ *~~M. edulis~~* ~~and~~ *~~M. trossulus~~* ~~we used the logistic mixed-effect model (GLMM with a binomial distribution and a logit link function) was constructed to predict the probability of having T-morphotype?? among~~ *~~M. trossulus~~* ~~(conventionally,~~ **~~P(T|tros)~~** ~~for each samples) and among~~ *~~M. edulis~~* ~~(conventionally,~~ **~~1 - P(E|edu)~~** ~~for each samples) as a function of a number of predictors. The dependent variable was coded as 1 for mussels with T-morphotype and 0 for mussels with E-morphotype and was modeled as a function of~~

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 (“Sample”) was included into model as random factor influencing the model intercept.

*Statistical analysis of testing dataset*

Пока у тебя на картинках все данные из тестоваого датасета были просто нарисованы тех же осях, что и данные из моделингового датасета. Если это и дальше плнируется делать (что, по-моему, плохо) то это одно описание, если мы таки найдем тестовые данные из Белого моря (что очень хотелось бы), то тогда схема анализа будет совсем иной. Идеальная методика была бы такая.

All three (Это если из Белого что-то найдем) testing subsets were treated as follow. Using data on the proportion of T-morphtype in each population we calculated the predicted proportion of M.trossulus (Ptros) in each of them using the parameters of the Model 1. Then we compared the observed proportion with the predicted one. Analogiously we calculated the predicted proportion of correctly identified mussels using the parameters of fixed part of the Model 4. Then we compared the observed proportions with predicted one.

~~We used testing data-set to compare predictions of the regression models based on modelling dataset (low and high saline habitats from Kola Bay) vs data observed (low and high saline habitats from the Barents Sea open coast). Using the data on Ptros in populations from testing dataset we calculated predicted values by the Models 1-4 for each testing population and compared them with observed data first 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 diagonal (Y=X line). А потом я плохо поняла, что вы там делали (из вашего июньского файла Material and Methods). Предлагаю по возможности сделать для каждой из четырех моделей единообразную процедуру и написать об этом 2 предложения)~~

*Analysis of geographical dataset*

Given the limited data from European populations (different taxonomic structure wasn't represented) we only described morphotype proportions among *M. trossulus* and *M. edulis* based on pooled samples for each regions. Или не only?

*Practical recommendations*

Сейчас по-русски план подраздела: Задача: хотим понять, что предложить пользователям: сколько и каких выборок из гибридной зоны брать, чтобы по теореме байеса (формулы даем) сделать калькулятор.

Что мы делали: мы брали такие-то наборы выборок (по выборке из чистых популяций, пара выборок из мешаных популяций, может ещё что-то добавить?), делали по ним бублики и пользовались формулами. Потом смотрели, при использовании какого набора выборок, наши предсказанные по калькулятору закономерности лучше всего сходятся с нашими данными. Лучший результат и рекомендуем (в обсуждение пойдет присказка о связи с Ptros и гибридах).

Открытые вопросы: надо понять, барентс фреш и белое – одно и то же? Если да, то давайте сливать. Если тестинг та же фигня окажется, то давайте и его сливать. И в итоге мы опубликаем формулы с коэффициентами для всей нашей гибридной зоны (для мест с низкой соленостью), и типа кто хочет, может ей пользоваться в Белом и Баренцевом морях. А вот в других гибридных зонах, если верите, используйте наши рекомендации к использованию теоремы байеса.

**P(MT|T) = (Ptros\*** **P(T|MT)**)/**PT\_exp**, where

**PT\_exp =** (**Ptros\*P(T|MT))+((1 - MTprev)\* (1 - P(Е|MЕ))**,

**P(ME|E) = ((1 – Ptros) \*** **P(E|ME)**)/**PE\_exp**, where

**PE\_exp =** (**(1 – Ptros)\*P(E|ME))+(Ptros\* (1 - P(T|MT))**,

**P(T|MT)** and **P(E|ME)** – estimates based on pooled samples from particular sets,

**Ptros** might be varies from 0 to 1

Как мне кажется надо делать следующее.

1. Сливать Белое и Баренцево опресненное можно, но зачем? Ведь это надо еще дополнительный тест на различия проводить, его описывать в материалах и методах, отдельно описывать результаты, давать их трактовки.... Много и незачем. Лучше пусть это будет дробная модель. Люди работающие в Белом море возьмут формулы для него, а в Баренцевом море для него. Для Белого моря и опресненного Баренцева моря алгоритм будет такой:
   1. Взять выборку и подсчитать долю Т-морфотипа.
   2. Это даст возможность вычислить Ptros (Model 1). Для кого-то этого будет достаточно, так как интересует только генетическая структура.
   3. Для тех, кого интересует экспресс-оценка генетического статуса каждой отдельной мидии предлагаем следующий шаг. Например, это будет важно для экспериментов с массовым использованием разных видов мидий, когда генотипирование будет очень дорого (ну например, аля наши работы со звездами) или когда генотипирование невозможно (например анализ мертвых створок). Для этих целей надо учитывать не только морфотип мидии, но еще и Ptros . Поэтому, полученную на предыдущем шаге Ptros подставляем в Model 4. Это дает нам возможность сказать с какой вероятностью мы корректно определяем в данной выборке конкретную мидию Т-морфотипа как троссулус, а мидию Е-морфотипа, как эдулис.
   4. Для желающих можно для популяции посчитать соотношение корректно определенных троссулусов и эдулисов.

Мы уверены, что для других опресненных мест типа Лох-Этива такой метод тоже будет работать.

Для нормально соленых мест Баренцева моря, а также, похоже, для других подобных мест (типа Норвегии), а также для мест, где известно, что мидии очень сильно гибридизованы (Балтика, вот здесь бы картику с синими пятнами, из которой следует, что гибриды все портят) мы пока не рекомендуем использование морфотипа, как инструмента определения. Возможно, более подробные исследования популяций из таких мест позволят выявить дополнительные условия применения предложенного метода.

Однако для мест, где морфотипы **могут работать** можно предложить общий алгоритм анализа.

1. Найти в изучаемой акватории популяции относительно чистые, где Ptros~1 или Ptros~0. Это можно сделать, ориентируясь на долю мидий Т-морфотипа, которая в первом случае должна быть минимальна, а во втором случае максимальна.
2. Выборки из этих популяций надо генотипировать. Это будут калибровочные данные.
3. На основании полученных генотипов надо вычислить калибровочные **P(T|MT)с** и **P(Е|MЕ)с.** Если эти величины будут близки к тому, что получилось в наших сборах, то можно использовать наши модели. Если нет, то надо будет построить модели, специфичные для данного региона. Делается это так.
4. Мы можем построить «линейку» популяций, в которых МТ и МЕ будут смешаны в пропорциях от 0 до 1, что даст возможность определить вероятность правильного определения видов по морфотипам, согласно теореме Байеса .

**P(MT|T) = (Ptros\*** **P(T|MT)с**)/**PT\_exp**, where

**PT\_exp =** (**Ptros\*P(T|MT)с)+((1 - MTprev)\* (1 - P(Е|MЕ)с)**

аналогично

**P(ME|E) = ((1 – Ptros) \*** **P(E|ME)с**)/**PE\_exp**, where

**PE\_exp =** (**(1 – Ptros)\*P(E|ME)с)+(Ptros\* (1 - P(T|MT)с)**,

**Ptros** varies from 0 to 1

Это будет наш байесовский график (кстати, его бы привести, таки).

1. По этому графику (или по формулам) можно оценить вероятности правильного определения мидий по морфотипу в популяции с заданным **Ptros** .
2. **Ptros** должна быть пропорциональна величине **PT.**  Параметры этой связи можно оценить по величине **P(MT|T)с** и **P(ME|E)с** в калибровочных выборках.

Надо подумать, но что-то мне подсказывает, что это будет такая формула

**Ptros**  **= PT\*P(MT|T)с + (1-PT)\*(1-P(ME|E)с)**

Где PT - доля Т-морфотипа в данной, произвольно взятой, выборке.

Надо проверить, не ошибаюсь ли я.

1. Исходя из полученных оценок вероятности правильного определения видов по морфотипу, можно оценить соотношение видов в любой популяции данного места если мы знаем частоту Т-морфотипа или Е-морфотипа в данной популяции.