Терминологические проблемы estuarine, full saline, “freshed” – есть такое слово? т.пр.

М.б. формально brackish<30, saline >30 ppt; high saline/low saline?

Я голосую за brackish и saline, если все наши выборки из «полносоленых» мест формально попадают в категорию >30 ppt

**Reliability of Mytilus edulis and M. trossulus taxonomic identification by ~~semi-diagnostic~~ simple conchological character varies between saline and brackish water habitats across Kola Peninsula (White Sea, Barents Sea) the same as macrogeographically.**

**Abstract.** Cryptic blue mussel species (unambiguous identification by multilocus genotypes only) Mytilus edulis (МЕ) and M. trossulus (МТ) co-occur along coasts of Kola Peninsula (the White Sea, salinity up to 25 ppt, the Barents Sea, salinity up to 35 ppt). Previous studies revealed morphological differences between species in the White Sea: the majority of MT bear an uninterrupted nacreous layer under the ligament nympha (Т-morphotype) while the majority of ME lack this character (E-morphotype). The main objectives of our study was to find out whether morphological differences between ME and MT persist in populations along the Barents Sea coast, and to develop practical recommendations for mussel taxonomic identification by morphotype. 2750 *(белое + баренцево)* mussels classified into ME and MT by their multilocus genotypes from 50 *(белое + баренцево)* populations were included into analysis. We also analyzed congruence between species genotype and morphotype in limited collections (645 *mussels,* 9 *samples,* включая две аллопатрические из Зап Атлантики, LE – эдулисная выборка из Шотландии, ещё не добавлена, и надо решить, нужна ли она вообще) from other regions of the Ocean where two species co-occur. In the areas of the Barents Sea with the salinity as in the Whites Sea, differences in morphotype frequencies between МЕ and МТ were founded to be nearly as large as in the White Sea (70 % and 77% это если вместе с тестингом, correspondingly). In more oceanic habitats differences were much lower (31% это если вместе с тестингом on average ), due to the increased frequency of Т-morphotypes among ME. For the brackish water areas where the conchological character was in a good agreement with species genotype, functions relating frequencies of T-morphotype and expected frequencies of МЕ and МТ in a sample, and probabilities of correct taxonomic identification of individual mussel by morphotype are provided. Analyses of overseas material revealed that there are prospects for use of the character for species identification in Western Atlantics and Scotland, but not in Western Norway and the Baltic Sea. =277

**Introduction.**

…

The main objectives of our study was to find out whether strong morphological differences between Mytilus edulis and M. trossulus revealed in the White Sea persist in populations of other regions where two species co-occur, first of all along the adjacent oceanic Barents Sea coast, and to develop practical recommendations for mussel taxonomic identification by morphotype. Also, using broad material from the Barents Sea we verify the “ecological” hypothesis that the discriminative power of morphological method varies between areas with low salinity (as in the White Sea), and normal oceanic salinity.

**Materials and methods.**

***Samples.***

Altogether, 59 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=1103), 2 from Western Norway (N=120), 4 from the entrance to the Baltic Sea (Strelkov et al. 2017, N=341), 2 from the NW Atlantic (N=84) and new samples - 26 from the Barents Sea coast of Kola Peninsular (N=1646) and one sample from Loch Etive, Western Scotland (N=100) (Fig. 1, see co-ordinates and other sampling details in the ESM table 1).

~~For the purposes of analyses~~ samples from the White and Barents Seas were grouped into four regional sets. The WS set included all samples from the White Sea (salinity up to 25 ppt), the set KF - 9 samples from permanently freshed areas in the top of Kola Bay (salinity up to 25 ppt), the set KN - 8 samples from the areas with high salinity in the entrance of Kola Bay (salinity above 30 ppt), the last set BO included samples from localities along the open Barents Sea coast among which three were permanently freshed while six had normal oceanic salinity (Fig. 1, ESM table 1). Salinity in sampling localities was either taken from literature (Кольский залив, Зеленцы, Ивановская REF) or, in case of few open coast localities predicted basing on the presence/absence of large rivers nearby (see ESM table 1… for details).

Здесь или где сказать, как эти сеты выступали в стат.анализе и определить, кто тестинг, а кто моделинг? Почему Белое море моделинг если BO вроде как сопоставляется с KF и KN?

Нужны эти аббревиатурные названия для сетов, вообще?

Аббревиатуры, думаю, обязательно нужны. Моделлинг – это мы сравниваем Белое с Баренцевым. Если тема с тремя уровнями фактора годится в математическом и идеологическом плане, то давайте так и оставлять.

Тестинг – тестируем гипотезу о «повсеместном» влиянии солености в Баренцевом.

Вадим Михайлович, ведь так?

Если про модели и прочее, что ниже, можно обойтись и без таблицы, то для описаня материала точно нужна таблица, в котрой все выборки (может быть группы выборок) будут описаны (даны им обозначения) и сказано для чего они были использованы.

***Species identification.*** … were genotyped at three or four taxonomically informative allozyme loci … . Structure (). Hence potential hybrids were not considered as separate category but included into … .Как это влияет на данные?

Я бы провела отдельный анализ, где посмотрела бы, достоверно ли уподобляются МТ-гибриды чистым МТ (и с МЕ аналогично). Для Белого этого опубликовано, а для Баренцева – нет. Это отметет лишний вопрос рецензентов.

***Это лишнее! Скорее, надо обосновать выбор границы structure = 0.5.***

***Morphological marker identification.***The morphotype identification was described in details in Katolikova et al. (2016). In brief, ++++. **After Katolikova .. and Khaitov, we assume that T-morphotype is an a ttribute of M. trossulus while E-morphotype, of M. edulis.**

***Data analyses*.**

The ultimate goal of the analysis was to answer the questions what is a capacity of conchological marker to discriminate *M. edulis* and *M. trossulus*, how *M. trossulus* frequencies in populations could be predicted basing of morphotype frequencies and how likely is it that a random mussel with T-morphotype is *M. trossulus* and a mussel with E-morphotype is *M. edulis*, correspondingly, taking into account the possible effect of geography and ecology on morphotype frequencies among *M. edulis* and *M. trossulus*.

Assuming that ***a*** is a frequency of *M. trossulus* with T-morphotype, ***b*** - of *M. trossulus* with E-morphotype, ***c*** – of *M. edulis* with T-morphotype, ***d*** – of *M. edulis* with E-morphotype, we calculated the next proportions for each sample:

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

Proportion of M. trossulus (***MTprev***), ***MTprev*** = (***a***+***b***)/( ***a***+***b***+***c***+***d***); Зачем нам MTprev, если мы отказываемся от терминологии ROC-анализа? Пусть уж тогда *PMT* или как-то так

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

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

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

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

Proportion of mussels, which phenotype fits the expected genotype (i.e. joint proportion of *M. trossulus* with T-morphotypes and *M. edulis* with E-morphotypes) (***AUC***), ***AUC*** = (***a***+***d***)/(***a***+***b***+***c***+***d***) = 0.5\*(***P(MT|T)* + *P(ME|E)***). Это справедливо только для вычисления AUC в случае бинарного теста, и это надо указать.

Можно ли придумать для статистик более «говорящие» обозначения, не требующие запоминания терминологии? А использовать то, что в скобках, в тексте статьи – это плохой тон?

По мне так лучше использовать и то и другое. В формулах удобнее писать обозачения из теории веротноястей, а в тексте их словесные аналоги. Только словесные аналоги где-то надо сопрячь с теорверскими обозначениями.

It is worth mentioning that these proportions are used in clinical medicine for evaluation of diagnostic tests and has the next names, utilities and properties (after … REF). ***MTprev*** is named prevalence. ***P(T|MT)*** and ***P(Е|MЕ)*** are named *sensitivity* and *specificity* and evaluates the ability of the test to identify correctly *M. trossulus* among *M. trossulus* and *M. edulis* among *M. edulis*, correspondingly. ***P(MT|T)***and ***P(ME|E)*** are named *positive predictive value* and *negative predictive value* and evaluates the ability of the test to identify correctly *M. trossulus* among T-morphotypes and *M. edulis* among E-morphotypes ~~(answers the questions how likely is it that a mussel with T-morphotype is~~ *~~M. trossulus~~* ~~and a mussel with E-morphotype is~~ *~~M. edulis~~*~~)~~, correspondingly. ***AUC*** is named *accuracy* ну вообще наоборот, AUC в ROC-анализе употребляется чаще, чем accuracy, как мне казалось and measures the ability of a test to differentiate *M. edulis* and *M. trossulus*. Надо прописать, что тест – это морфотип, в самом начале. While ***P(T|MT)***, ***P(Е|MЕ)*** and ***AUC*** are independent on the ***MTprev*** in case when the test is reliable and universal, ***P(MT|T)***and ***P(ME|E)*** are inherently dependent on the taxonomic structure of a population, and are negatively correlated. Если писать «negatively correlated» сейчас, то совсем непонятно, о чём речь идёт. Remembering that neither *M. edulis* no *M. trossulus* are fixed by their species-specific morphotypes (Katolikova et al. 2016; i.e. P(T|MT) and P(Е|MЕ) are < 100%), one can expect that in “pure” *M. edulis* population, a rare mussel with T-morphotype would be no way *M. edulis* м.б. *M. trossulus?* (***P(MT|T)****=0*), the same as any mussel with E-morphotype (***P(ME|E)***=1). In mixed populations, with the increasing ***MTprev***, ***P(MT|T)***would increase while ***P(ME|E)***would decrease.

In “pure” *M. trossulus* population, mussel of any morphotype from would be *M. trossulus* (***P(MT|T)***=1; ***P(ME|E)***=0). По сути, это краткая формулировка того, что выше написано подробно

If we ~~imagine~~  accept 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), in terms of clinical medicine the tasks of the analyses would be: what is an accuracy of the test про тест ничего не написано, нужно выше пояснить, how the “bad” mussel prevalence could be predicted basing on the results of the “morphotype test”, and what are positive and negative predictive values of a test taking into account the possible effect of geography and ecology on its sensitivity and specificity.

*Вот этот абзац было бы хорошо поместить во введение, так как он задает фреймворк в котором мы далее строим все повествование.*

***Statistical testing and modeling ну тогда уж modeling and testing***

Разными параграфами что делали с беломорско-баренцевоморскими и с заморскими выборками, или как? Пока не знаю

We analyzed and compared distributions of ***PT***, ***P(T|MT)*** , ***P(Е|MЕ), P(MT|T),******P(ME|E)*** and ***AUC*** against ***MTprev*** among samples and also among sample sets. In the analysis of material from the White and Barents Seas, the BO set comprised of samples from both low- and high-saline localities was treated as a testing data set while ~~WS (? Да?),~~ KF and KN composed of samples from low- and high-saline areas of Kola Bay - as a modeling data set. Логика д.б. обратная, сначала надо W, KF и KN, а потом BO.**That is … сперва находили регрессионные зависимости для KN и KF, а потом с помощью этих моделей предсказывали значения статистик для выборок из соленых и опресненных мест из BO; эмпирические и предсказанные значения для выборок из BO сравнивали на глаз?** Если я правильно поняла, то вы Спирмена делали, правильно? All или какие-то? Да, все analyses were performed with functions of R3.5.5 statistic programming language (+++).

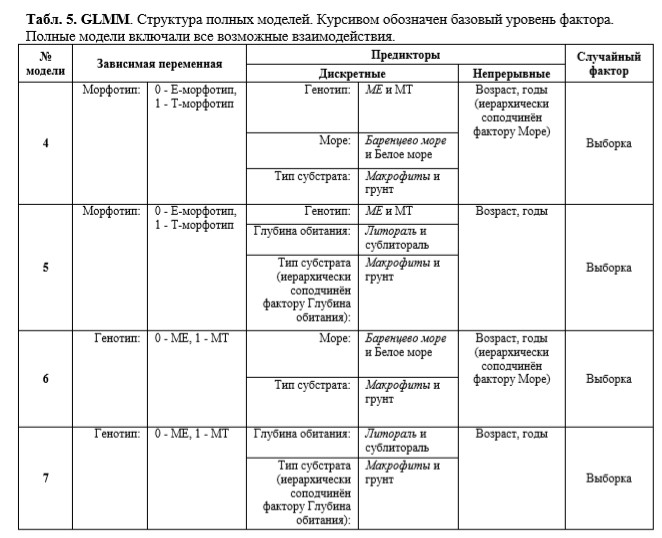
Однозначно, в каждом пункте, где описываются какие-то анализы, должны быть ссылки на программное обеспечение (R пкеты и лучше бы и функции, это принято и облегчает понимане тем, кто в теме). Я бешусь, когда не вижу этой информации в статьях.

Во всех случаях, где описываются регрессионные модели схема их описания общая:

1. Что является зависимой переменной
2. Что является предикторами
3. Какое тип модели строили (в нашем случае логистическая регрессионная модель с логит-связывающей функцией и случайным эффектом популяции)
4. Какой тип оценки значений парамтеров использовался ( внашем случае ML).
5. Какие функции из каких пакетов применялись.

The “MTprev” was modeled as a function of PT (continuous predictor) and “Set” (discrete predictor with three levels) and interaction between them. Sample (“Pop”) was included into model as a random factor influencing the model intercept? ДА, НЕТ? ДА The binomial distribution of MTprev was supposed and model was fitted as logistic regression one with logit link-function. The function glmer() from the package “lme4” (++) was used for the model fitting. Here and thereafter (Где? Когда? – when employing regression analysis? Лучше GLMMs) 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.

Мне кажется, что должна быть таблица с GLMMами, которые мы будем использовать. Типо такой, как была у меня в диссере (см. картинку ниже)



Потом у меня возникает вопрос: мы ещё должны обсудить возможное влияние на язычковость и других факторов: размера (от него ПП отказываться не хочет, да и я тоже), мб субстрата (почему бы не проверить, больше язычковых на фукусах в Тюве и Зеленцах или нет? И сюда можно добавить вывод о том, что МТ и МЕ субстраты в Бар море не различают) и тд и тп. Если всё это анализировать, то об этих моделях надо тут же писать, в мат-методах, правильно?

Блин, не надо плодить сущности! Размер? А почему не возраст? А почему субстрат, а не высота над уровнем моря? Или не плотность поселения? Последнюю ведь не учитывали? Проблема с размером всплывает из-за смещенности и (по большому счету) неслучайности выборок мидий, использованных для генотипирования. Ведь при сборах мидий брали не весь размерный ряд, пропорционально представленности в размеров в популяции. Хапали то, что бросалось в глаза (где-то покрупнее, потому что водолазу легче и т.п.). Для обсуждения связи с размерами и субстратами требуются систематические сборы, направленные именно на проверку данной гипотезы. Я бы, будь рецензентом, обосрал материал, в котором полученные связи (или их отсутствие, как в случае с субстратами) являются следствием некорректно собранного материала. Хороший тому пример - отскок частот троссулусов, когда моллюсков хапнули с антропогенного буя. А в других местах такие буи не исследовали. Потому и не получили всплесков троссулусов. Вот генотипируете сборы этого года, увидите, что в абсолютно неопресненной бухте Могильной на буе сидят троссулусы с прекрасными язычками, а эдулисы без оных. Это не будет укладываться в картину мира, где в неопресненных участках язычок перестает работать.

Короче, то что размер играет роль для меня совсем не факт. Скорее это результат того, что в какой-то выборке, где много крупных, оказалось что частоты одного из видов выше. К тому же я не вижу никаких явных паттернов в остатках в связи с размерами (но зато вижу их в связи со значением structure, что по мне так гораздо больший криминал, но его можно объяснить за счет формального введениия границы structure = 0.5). ПП ни хера в регрессионном анализе не понимает и потому боится всего, что сложнее коэффициента корреляции Спирмена. Но зато уж если дорвался, то давай клепать множественные сравнения... Ладно, это я злюсь...

Про субстраты вообще смех. Пока это все показано по очень малым выборкам в немногих точках это не убеждает. А маленькие значения пи-вельюс , как мы знаем, бывают и случайными. Короче не надо дразнить никого и вводить лишние рассуждения, на основе плохо собранного материала.

Где анализ, призванный показать, что у M. edulis частоты язычков варьируются от сета к сету, а у M. trossulus ХЗ, но по крайней мере меньше. где? Нет его ещё в том виде, в каком хочет ПП

Кажется, будто методика анализы P(T|MT) и P(Е|MЕ) относительно MTprev – сходная с предыдущей. Я не понимаю, о чём это. А вы?

The P(MT|T) and P(ME|E) were modelled /simultaneously? Одна модель, две разные модели? Если одна, то это про AUC…Не поняла, о чём речь / as a function of “MTprev”, “Set” and interaction between them. In each? case we coded all mussels whose phenotype matched the expected genotype (i.e. *M. trossulus* with T-morphotype or *M. edulis* with E-morphotype) as 1, others as 0. Sample (“Pop”) was included into models as a random factor. We constructed the logistic regression models with mixed effects (GLMM, Zuur et al, +++) which described the association between the probability of correct identification and predictors: morphotype (“Morph”, discrete predictor with two levels), “MTprev” (continuous predictor), “Set” (discrete predictor with three levels) and all possible interactions between predictors. Sample (“Pop”) was included into the model as a random factor. The model were 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. НУ вот чтобы 10 раз одно и то же не писать, я предлагаю делать таблицу с информацией о структуре моделей

We also calculated the expected distributions of ***P(MT|T)*** and ***P(ME|E)*** against ***MTprev*** basing on the assumption that ***P(T|MT)***and ***P(E|ME)*** does not vary among samples within individual sets, hence ***PT*** and ***P(MT|T)*** and ***P(ME|E)*** could be predicted basing solely of the genotypic structure of a sample, expressed as ***MTprev***, using formulas:

P(MT|T) = (***MTprev\*PT***)/***PT\_exp***, where ***PT\_exp***=(***MTprev\*P(T|MT))+((1- MTprev)\* (1-P(Е|MЕ))***, P(T|MT) and P(E|ME) – estimates based on pooled samples from particular sets

Почему ***MTprev****,* если мы сами задаем долю **МТ (от 0 до 1)?** Надо по-другому назвать.

Почему в числителе \***РТ,** если д.б. ***P(T|MT)*** (т.е. бублик), или я чего-то не понимаю?

P(ME|E) = …

СРЕДНИЙ? AUC and confidence intervals for AUC for each set were calculated with the … – так ли нужна эта программа? Почему нельзя единообразно?

Я бы прописала это в тексте, раз мы теперь любим AUC, это ведь норм стат оценка различий между выделами. Но это уже после формального описания точек на графике, о котором пишет ПП ниже.

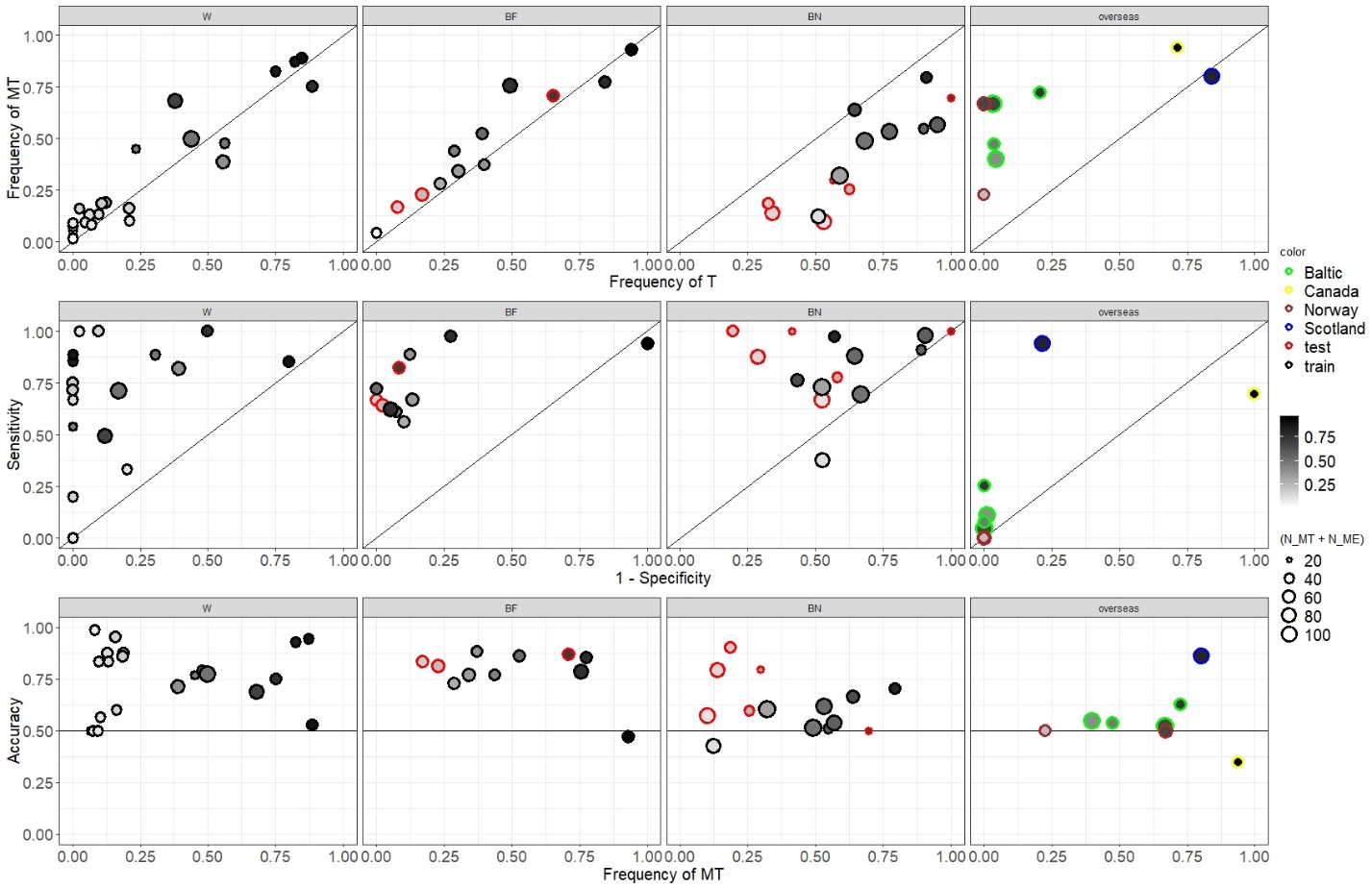
Блин! А где модель для congr. Это вообще ключевая модель, без которой все рушится.

Короче! Здесь пока какая-то несистематическая мозаика. Завтра поробую написать план, такой как он мне видится.

И никаких размеров и субстратов!!!!

***Результаты*** начинать с карты и вербального описания паттернов на картинках типа этих.

Дальше я удаляю текст ПП и старую картинку, а вставляю новую (код я присылала немного на другую, там добавлены американские эдулисы, но суть одна и та же)



Variability of morphotype frequencies among samples and subsamples of *M. edulis* and *M. trossulus* within samples from different regions. A. PT against MTprev. Diagonal line is at PT=MTprev. B. P(T|MT) against 1-P(E|ME). Diagonal line is at P(T|MT) =1-P(E|ME). C. AUC against MTprev. AUC>0.5 reflects predominance of mussels with T-morphotype among M. trossulus. AUC<0.5 – predominance of mussels with T-morphotype among M. edulis. AUC=0.5 (horizontal line) reflects random association between genotype and morphotype. Diagrams, from left to right: samples from the White Sea, from low and high salinity areas of the Barents Sea, from other geographical regions. Dots representing samples from the White Sea and Kola Bay of the Barents Sea are depicted by black borders, from the open coast by red, from Scotland by blue, from the Baltic Sea by green, from the NW Atlantic (на этой картинке подписано Canada) by yellow, from Norway by brown borders. The size of dots is proportional to sample sizes, the saturation of the dot filling – to proportions of M. trossulus in samples).

Есть идея просто их все заливать разными цветами (а не просто делать границы разного цвета), и делать границу от светлого к темному в зависимости от стракчи (как сейчас сделано с заливкой от белого к черному во всех точках).