|  |  |
| --- | --- |
| Finally, we tested how well the *Ptros,* *P(edu|E)* and *P(tros|T)* could be predicted using formulas Eq 1-3 and the data on morphotype proportions among species (*P(T|tros)*, *P(T|edu)*) in only two genotyped samples. We concede that the assumption that sensitivity and specificity do not depend on the prevalence can be violated in the morphotype test, as it often is in clinical tests (Leeflang et al. 2009, 2013). Therefore our interest was focused on finding out which samples are better suited for prediction on the basis of Eq. 1-3 and consequently could be used as “calibrating” ones: the most mixed samples (*Ptros*~0.5) or the combination of the two most pure samples of each species (Section “Prediction of taxonomic structure of populations and predictive values of the morphotype test based on calibrating samples”). |  |
| **Prediction of taxonomic structure of populations and predictive values of morphotype test basing on calibrating samples**  We applied Eq. 1-3 to predict *Ptros*, *P(edu|E)* and *P(tros|T)* for samples from each data set (*GOM*, *BALT*, *NORW*, *BH*, *WSBL*, *SCOT*) using estimates of morphotype proportions among species (*P(T|tros*), *P(T|edu)*) obtained from ~~pooled samples from each set and whenever possible, from~~ combinations of two calibrating samples selected based on the results of the following analysis.  To work out the strategy of calibrating samples selection, we considered the *WSBL* (36 samples) as a reference dataset. All 630 possible pair combinations of samples were considered. Each pair was characterized by an index of taxonomic similarity between the samples:  Delta = (*Ptros1)* \* (1 - *Ptros2*) + (*Ptros2*) \* (1 - *Ptros1*) [Eq. 4],  where *Ptros1* and *Ptros2* – higher and lower estimates of prevalence in samples. The index varies in a range [0; 1] and takes the value Delta=0 when both samples are pure *M. edulis* (*Ptros1* = *Ptros2* = 0) or pure *M. trossulus* (*Ptros1* = *Ptros2* = 1), Delta=0.5 when both samples are equivalent mixtures of two species (*Ptros1* = *Ptros2* = 0.5) and Delta=1 when one sample represent pure *M. trossulus* (*Ptros1* = 1) and another pure *M. edulis* (*Ptros2* = 0).  Estimates of *P(T|tros)*, *P(E|edu)* and *PT* were obtained from pooled data on each pair of samples and used for calculation of predicted values of *P(edu|E)* and *P(tros|T)* basing on Eq.1,2 for the range of *Ptros* [0;1] with the step 0.01 (“genotype by morphotype calculator”) and predicted values of *Ptros* basing on Eq.3 for the range of *PT* [0;1] with the step 0.01 (“*Ptros* by *PT* calculator”). Values of *P(edu|E)* and *P(tros|T)* obtained by the Eq. 1, 2 were contrasted with those ones predicted by the Model 6 and values of *Ptros* obtained by Eq. 3 were compared with predictions of the Model 4 using of correspondence statistics:  Goodness = 1 / Σ(Regression prediction - Equation prediction) 2 [Eq.5] |  |
| Goodness indices for each pair were plotted against the corresponding Delta values and the LOESS regression curve was fitted to find associations between them. Depending on the results of the analyses, we determined which samples could be used for predictions with best results: the most mixed samples (*Ptros1* ≈ *Ptros2* ≈ 0.5) or the combination of two most pure samples of each species (*Ptros1* ≈ 1; *Ptros2* ≈ 0).  For illustrative purposes and for the convenience of users of “morphotype test” or any similar semi-diagnostic tests we provide the online “*Ptros* by *PT*” and “genotype by morphotype” calculators implementing Eq. 1-3 at +++++. |  |
| **results** |  |
| **Prediction of taxonomic structure of populations and predictive values of the morphotype test basing on calibrating samples**  Using the coefficients of the regression models Model 4 and Model 6 (**Table 1**), we constructed a set of formulas predicting the taxonomic structure (*Ptros*) and the probability of correct species identification () using the morphotype test (Table 3). These formulas were used for the comparison of predictions made with these regression models and the predictions proposed by Eq. 1, 2 and 3.  Table 3. **Formulas used for taxonomic and individual assignment using morphotype tests in different sample sets** | **Prediction of taxonomic structure of populations and predictive values of the morphotype test basing on probability calculators**  Using the coefficients of the regression models Model 4 and Model 6 (Table 1), we constructed a set of formulas predicting the taxonomic structure (Ptros) and the probability of correct species identification () using the morphotype test (Table 3).  **Table 3.** Formulas used for taxonomic and individual assignment using morphotype tests in different sample sets accordingly to the regression model’s coefficients represented in Table 1.  Это не последний вариант таблицы. Здесь плохо выводятся формуля с отрцательными коэффициентами. Надо вставить таблицу из Figures\_final\_version.html  Using this formulas one can predict *Ptros*, *P(tros|T)* and *P(edu|E)* for any new sample taken from the corresponding area if proportion of T-morphotype, *PT* in the sample is known. |
|  | ~~However Eq. 1, 2 (“genotype by morphotype calculator”) and Eq 3 (“~~*~~Ptros~~* ~~by PT calculator”) also could be used for predictions of values~~ *~~Ptros~~*~~,~~ *~~P(tros|T)~~* ~~and~~ *~~P(edu|E)~~* ~~on the base of~~ *~~PT~~* ~~if information on the~~ *~~P(T|tros)~~* ~~and~~ *~~P(T|edu)~~* ~~is available for any area (one can use the applett presented at +++++ to make the calculations accordingly to these equations). The key components of these equations are the values of~~ *~~P(T|tros)~~* ~~and~~ *~~P(T|edu)~~* ~~which could be obtained only after samples were genotyped and mussel’s mrphotype assessed. We will denote such samples as calibrating ones.~~  ~~To work out the strategy of searching of such best calibrating samples we compare the predictions of Eq 3 with predictions of the regression Model 4 (Table 3) and Eq 1, 2 with the regression Model 6 (Table 3) using as an input the data from WSBL.~~ |
| We applied Eq.1 and Eq. 2 (predictive values as a function of *Ptros*, *P(T|tros)* and *P(E|edu)*, “genotype by morphotype calculator”) and Eq. 3 (“*Ptros* by *PT* calculator”) using as an input the data on all possible pairs of populations from *WSBL* and compared the values predicted by these equations with those predicted by regression models 6 and 4, respectively (**Table 3**). | To work out the strategy of searching of ~~such~~ best “calibrating” samples we compared the predictions of Eq 3 with predictions of the regression Model 4 (Table 3) and Eq 1, 2 with the regression Model 6 (Table 3) using as an input the data from WSBL.  For all possible pairs of samples from *WSBL* we calculated the values of P(T|tros) and P(T|edu) for the pooled data to adjust the Eq 1, 2 and 3. Then we compared the values predicted by these equations with those calculated by regression models. |
|  |  |
| Fig. 4 illustrates the goodness of correspondence of the two predictions depending on the genetic constitution of the paired samples as expressed by the Delta index.  The best predictions of *Ptros* were obtained when the most dissimilar samples consisting of nearly pure *M. edulis* and *M. trossulus* were used, while the best predictions of P(edu|E) and P(tros|T) values were obtained when the most mixed samples (*Ptros* of both samples close to 0.5) were taken for calibration.  We applied the “calculators” to all five geographical sets using, where possible, two most dissimilar samples for the calculation of *Ptros* and two most mixed samples for the calculation of predictive values (Fig. 3; note that only two samples were available for *Scot*). Visual inspection of Fig. 3 shows good correspondence between the predicted by the “genotype by morphotype calculator” and regression lines не поняла, между чем и чем? in all cases but in *NORW*. The latter was due to the formal choice of the only outlier sample with an extremely low *P(tros|T)* as a calibrating one. In its turn, the “*Ptros* by *PT* calculator” was inaccurate for *BH*, *NORW* and *BALT* but nearly ideal for *WSBS* and *GOM*.  Fig. +. Correspondence between “genotype by morphotype calculator” (Eq. 1-2 , left graph) and “*Ptros* by *PT* calculator” (Eq. 3, right graph) and regression models (Model 6 and Model 4, respectively). Each point corresponds to a unique pair combination of samples from WSBL. . OX axis reflects dissimilarity of genetic structure in each pair (Delta) (for pure conspecific samples Delta takes a value of zero, for equally mixed samples – 0.5, for two pure heterospecific samples - 1). OY: goodness of correspondence between assessment of predictive values by equations and regression models. | Fig. 4 illustrates the goodness of correspondence of the two predictions depending on the genetic constitution of the paired samples as expressed by the Delta index.  The comparison of predictions from Eq 3 and Model 4 revealed the best correspondence of both when Delta was close to one (Fig. 4 a). It means that the best results of “Ptros by PT calculator” is obtained if samples with maximally differ taxonomic structure would be used as calibrating ones to assess P(T|tros) and P(T|edu)*.*  Another result was obtained in comparison of predictions from Eq 1, 2 and regression Model 6. The highest correspondence between these predictions was observed for Delta distributed between 0.25 and 0.5. Thus the best results of “genotype by morphotype calculator” , i.e. assessment of P(edu|E) and P(tros|T) values, were obtained when P(T|tros) and P(T|edu) were assessed in the most mixed calibrating samples (*Ptros* of both samples close to 0.5). |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |