Supporting Information

Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning

Tristan Cordier*[1], Philippe Esling[2], Franck Lejzerowicz[1], Joana Visco[3], Amine Ouadahi[1], Catarina Martins[4], Tomas Cedhagen[5], Jan Pawlowski[1,3]

[1] Department of Genetics and Evolution, University of Geneva, Boulevard d'Yvoy 4, CH 1205 Geneva, Switzerland

[2] IRCAM, UMR 9912, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France

[3] ID-Gene ecodiagnostics, Ltd, chemin des Aulx 14, 1228 Plan-les-Ouates, Switzerland

[4] Marine Harvest ASA, Sandviksboder 77AB, Bergen, 5035 Bergen, Norway

[5] Department of Bioscience, Section of Aquatic Biology, University of Aarhus, Building 1135, Ole Worms allé 1, DK-8000 Aarhus, Denmark

* corresponding author e-mail: tristan.cordier@gmail.com

Excel file:

Table S2: Tag to sample reference table following a Latin Square Design and tagged primers sequences for successful PCR amplified samples

Table S3: Quality filtering and assembly parameters

Table S4: Morpho-taxonomic dataset

Table S6: OTU-to-sample matrix with taxonomic assignments

Table S9: Pair-wise spearman correlation matrix of OTUs represented by more than 1000 reads and assigned to the same species.

Table S1: Sampling site coordinates and distribution of collected samples

Locality	Sampling date	Station	Grab/Station	Sample/Grab	Longitude	Latitude
Bjørnsvik	03/06/2015	4	2	3	67°31.969 N	15°23.262 E
Nedre Kvarv	04/06/2015	5	2	3	67°27.604 N	15°30.548 E
Beitveitnes	12/10/2015	5	2	3	62°08.388 N	5°19.659 E
Storvika	14/10/2015	5	2	3	62°48.159 N	6°58.808 E
Aukrasanden	15/10/2015	5	2	3	62°46.935 N	6°55.399 E

Table S5: Results of non-linear models testing the effect of farming site, distance from the cages and their interaction on four biotic indices values

	Df	Sum Sq	Mean Sq	F value	P > F
AMBI index					
Farming site	4	17.863	4.466	10.9296	**0.001*** **
Distance from cages	2	123.225	61.612	150.7938	**0.001*** **
Site x Distance	8	15.224	1.903	4.6575	**0.001*** **
ISI index					
Farming Site	4	22.794	5.698	11.028	**0.001*** **
Distance from cages	2	305.399	152.700	295.525	**0.001*** **
Site x Distance	8	61.584	7.698	14.898	**0.001*** **
NSI index					
Farming Site	4	556.5	139.13	24.327	**0.001*** **
Distance from cages	2	3725.7	1862.85	325.726	**0.001*** **
Site x Distance	8	653.6	81.70	14.286	**0.001*** **
NQI1 index					
Farming Site	4	0.27140	0.06785	20.559	**0.001*** **
Distance from cages	2	2.27750	1.13875	345.048	**0.001*** **
Site x Distance	8	0.41745	0.05218	15.811	**0.001*** **

*:P < 0.05; **:P < 0.01; ***:P < 0.001

Table S7: Taxonomic composition of foraminiferal communities

Taxa	OTUs	Reads	%
Rotaliida	778	3543496	42
Bulimina marginata	37	863584	10,2
Stainforthia fusiformis	21	649131	7,7
Cibicidoides lobatulus	39	588188	7
Monothamids	976	1680306	19,9
Bathysiphon argenteus	18	628010	7,4
Environmenal sp. (ENFOR9)	16	180642	2,1
Tinogullmia sp. (cladeY)	5	166875	2
Textulariida	562	683951	8,1
Reophax sp.	105	482838	5,7
Textularia gramen	17	82984	1
Spiroplectammina sp.	32	26487	0,3
Miliolida	53	2394	<0.1%
Globigerinida	4	29	<0.1%
Spirillinida	4	46	<0.1%
Robertinida	1	1	<0.1%
Uncultured foraminifera	190	360781	4,3
Unassigned	6602	2170118	25,7
Total	9170	8441122	

Table S8: Accuracy of BI values predictions from composition data, as a function of abundance filtering on the OTU-to-sample matrix. R^2 values are indicated and kappa statistics are between brackets. All R^2 and kappa statistics were significant ($p < 0.001$)

BI - Algorithm	0	10	100
AMBI - RF	0.662 (0.555)	0.65 (0.548)	0.652 (0.615)
AMBI - SOM	0.669 (0.711)	0.664 (0.694)	0.643 (0.656)
ISI - RF	0.56 (0.631)	0.548 (0.685)	0.57 (0.669)
ISI - SOM	0.615 (0.774)	0.653 (0.798)	0.663 (0.825)
NSI - RF	0.827 (0.832)	0.83 (0.851)	0.836 (0.85)
NSI - SOM	0.794 (0.871)	0.804 (0.881)	0.763 (0.874)
NQI1 - RF	0.81 (0.856)	0.809 (0.856)	0.82 (0.856)
NQI1 - SOM	0.803 (0.873)	0.774 (0.855)	0.819 (0.877)

Figure S1: Non-linear relationship between BI reference values and the distance to the cages. Dashed-lines represent the 95% confidence interval. In the titles are reported the R^2 and significance (***: p < 0.001, **: p <0.01, *: p < 0.05).

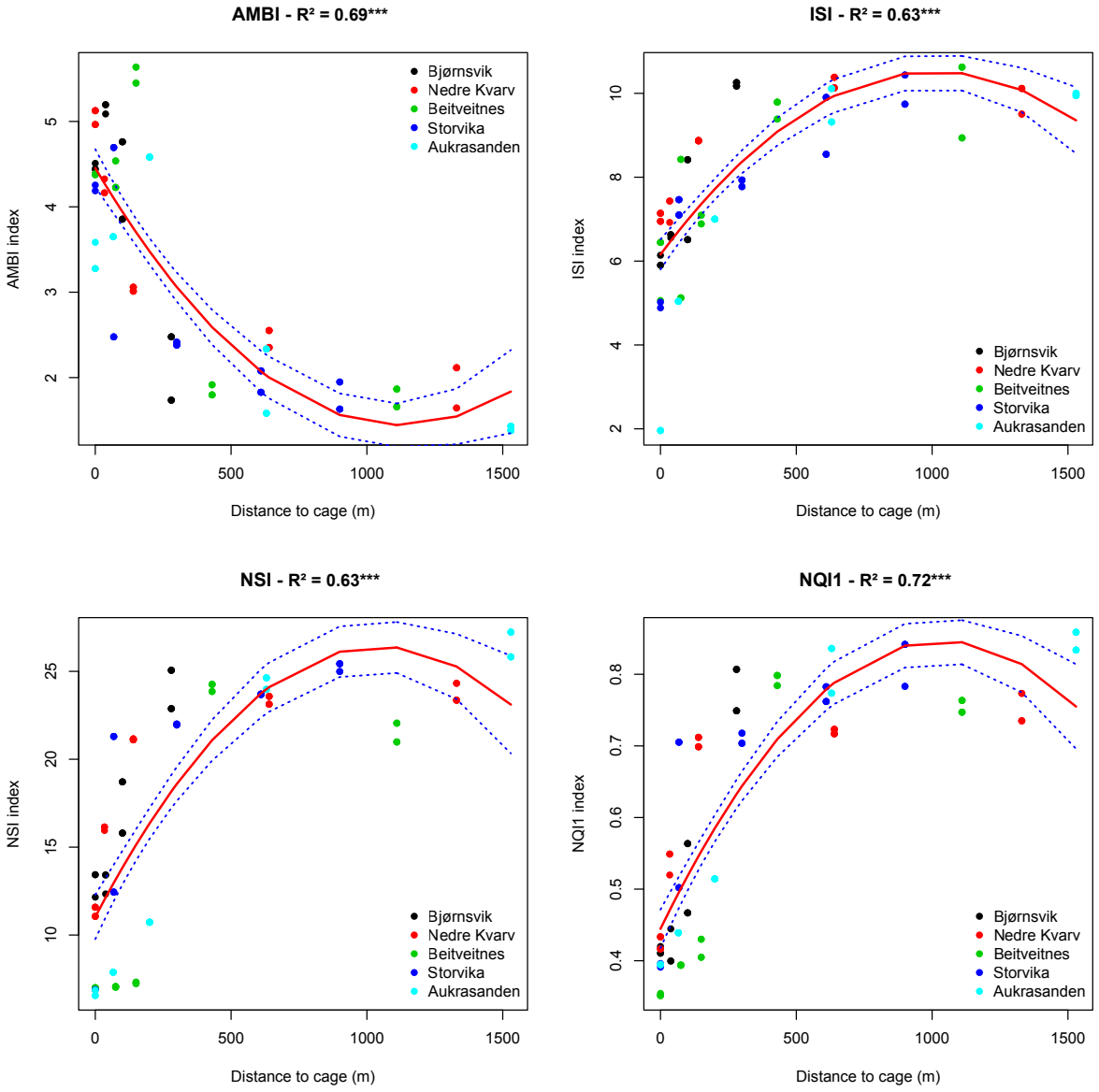

Figure S2: Correlation between alpha-diversity metrics and the distance from the cages. Diversity metrics were averaged over the 100 rarefied datasets and correlation with the distance from the cage were analyzed with non-linear models. Significant correlations are plotted, and dashed-lines represent the 95% confidence intervals. In the titles are reported the R^2 and significance (***: p < 0.001, **: p <0.01, *: p < 0.05, ns: non-significant).

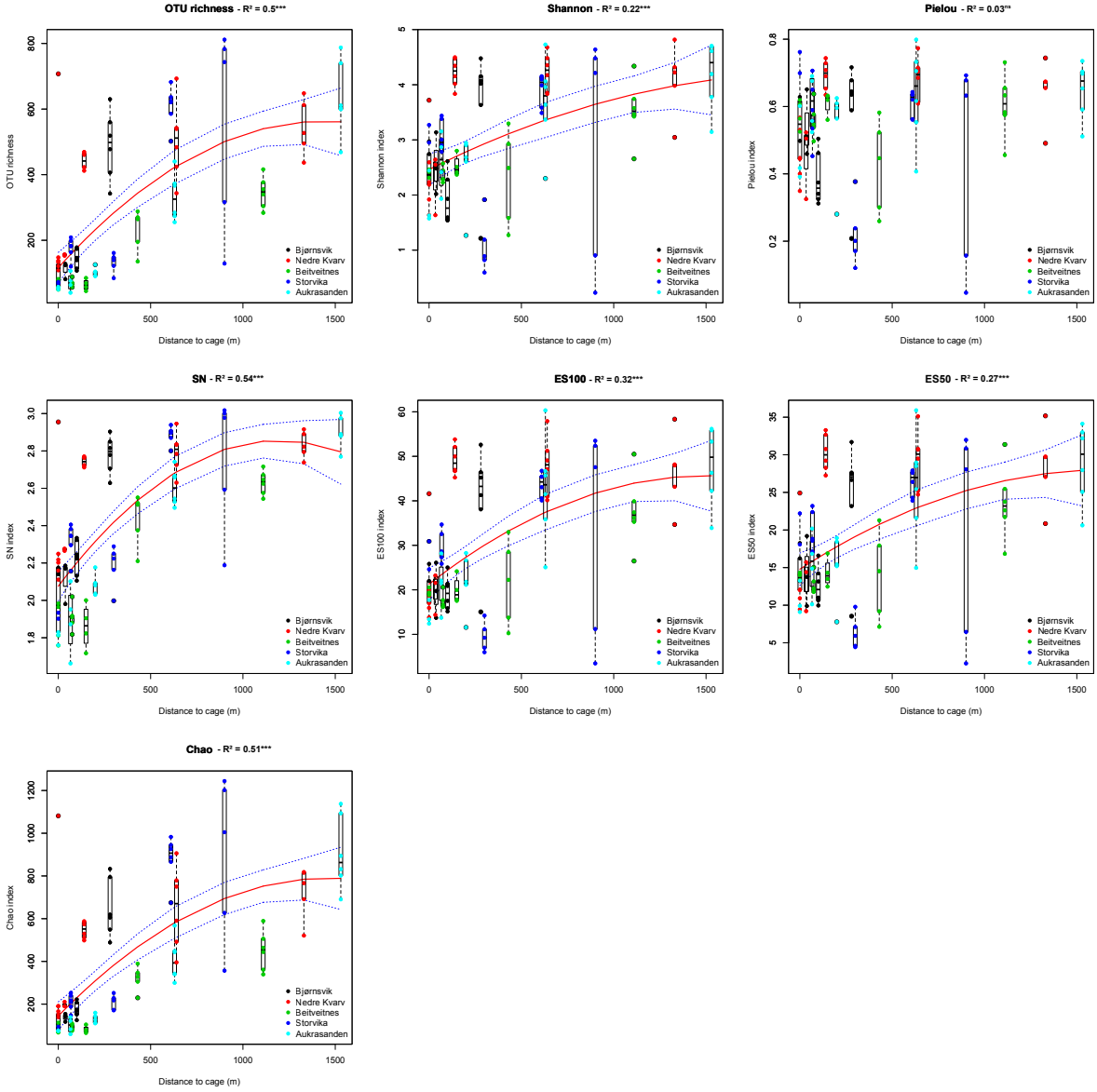

Figure S3: NMDS analysis of the Bray-Curtis beta-diversity matrix calculated from the CSS normalized foraminifera eDNA dataset. Black lines and associated numbers represent the distance from the cage (in meters). Red arrows indicate significant correlation between BI values and the foraminifera community dissimilarities. The stress value of the ordination is indicated on the plot.

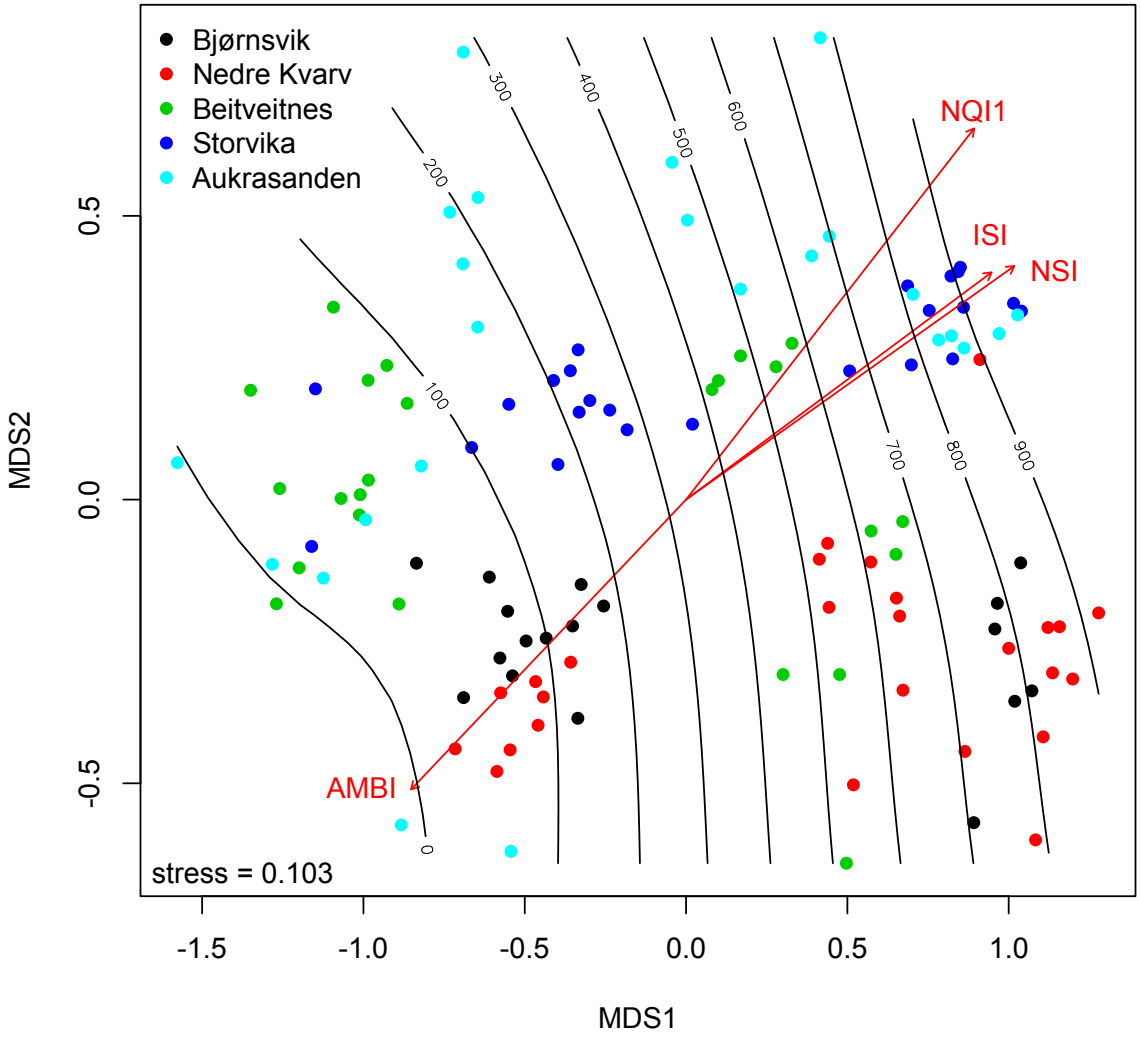

Figure S4: Effect of rarefying the OTU-to-sample matrix on the variation of inferred NSI values per sample. Each dot represents the average of the NSI values predicted for each sample and error bars represent the standard deviation over the 100 rarefied datasets. These standard deviations range from 0.09 to 1.67, with an average of 0.6. The top right barplot indicates the amount (number over bars) and percentage (y axis) of correct classifications (0 on x axis) and misclassifications. The bottom right boxplot indicates the median and quartile values (black numbers) and the mean value (red number) of the differences between the predicted and the reference NSI values.

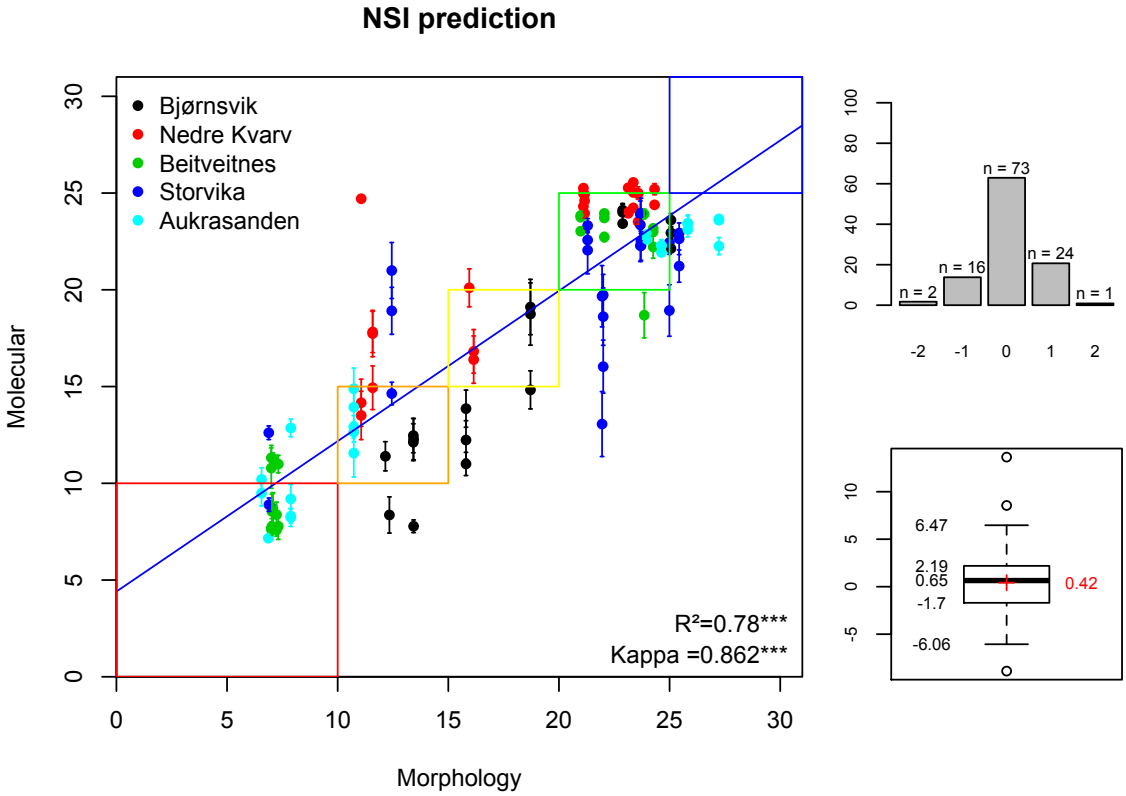

Figure S5: Effect of rarefying the OTU-to-sample dataset on the variation of inferred NQI1 values per sample. Each dot represents the average of the NSI values predicted for each sample and error bars represent the standard deviation over the 100 rarefied datasets. These standard deviations range from 0.009 to 0.035, with an average of 0.014. The top right barplot indicates the amount (number over bars) and percentage (y axis) of correct classifications (0 on x axis) and misclassifications. The bottom right boxplot indicates the median and quartile values (black numbers) and the mean value (red number) of the differences between the predicted and the reference NQI1 values.

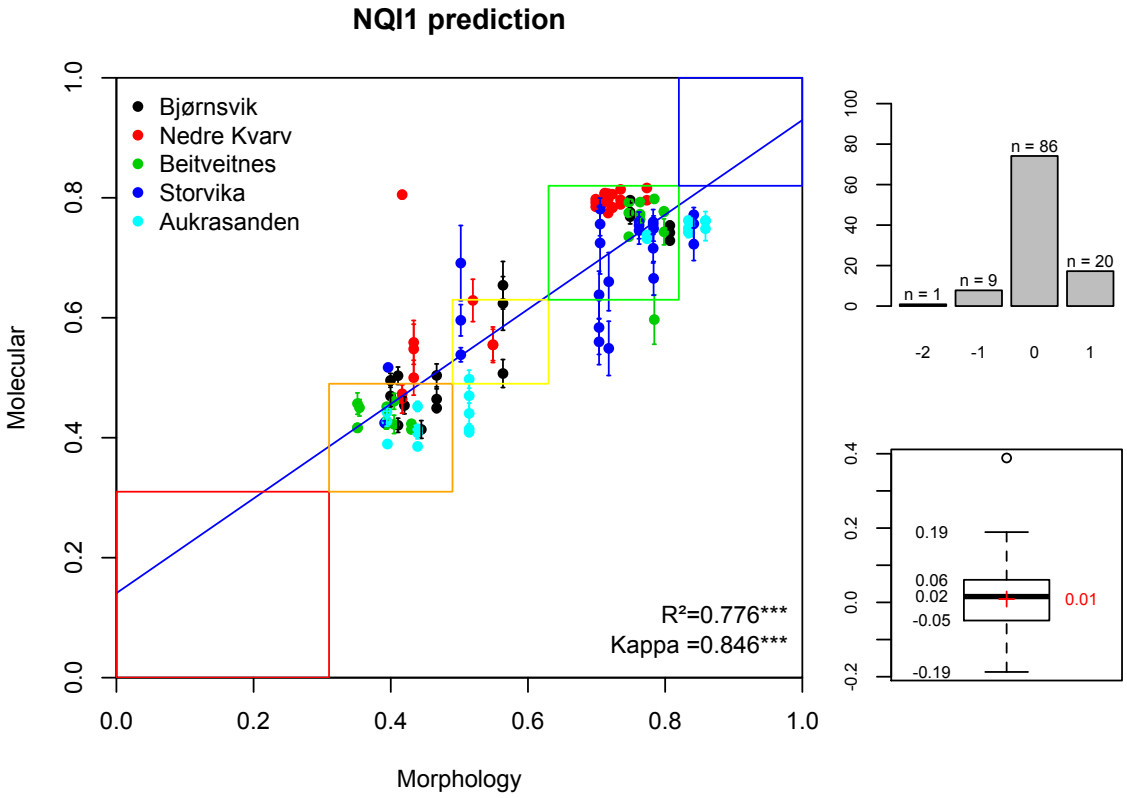

Figure S6: Diversity metrics importance in the prediction of the NSI index from the Random Forest models built from diversity metrics computed over 100 rarefied OTU-to-sample matrix. Error bars represent the standard deviation of importance over the 100 rarefied datasets.

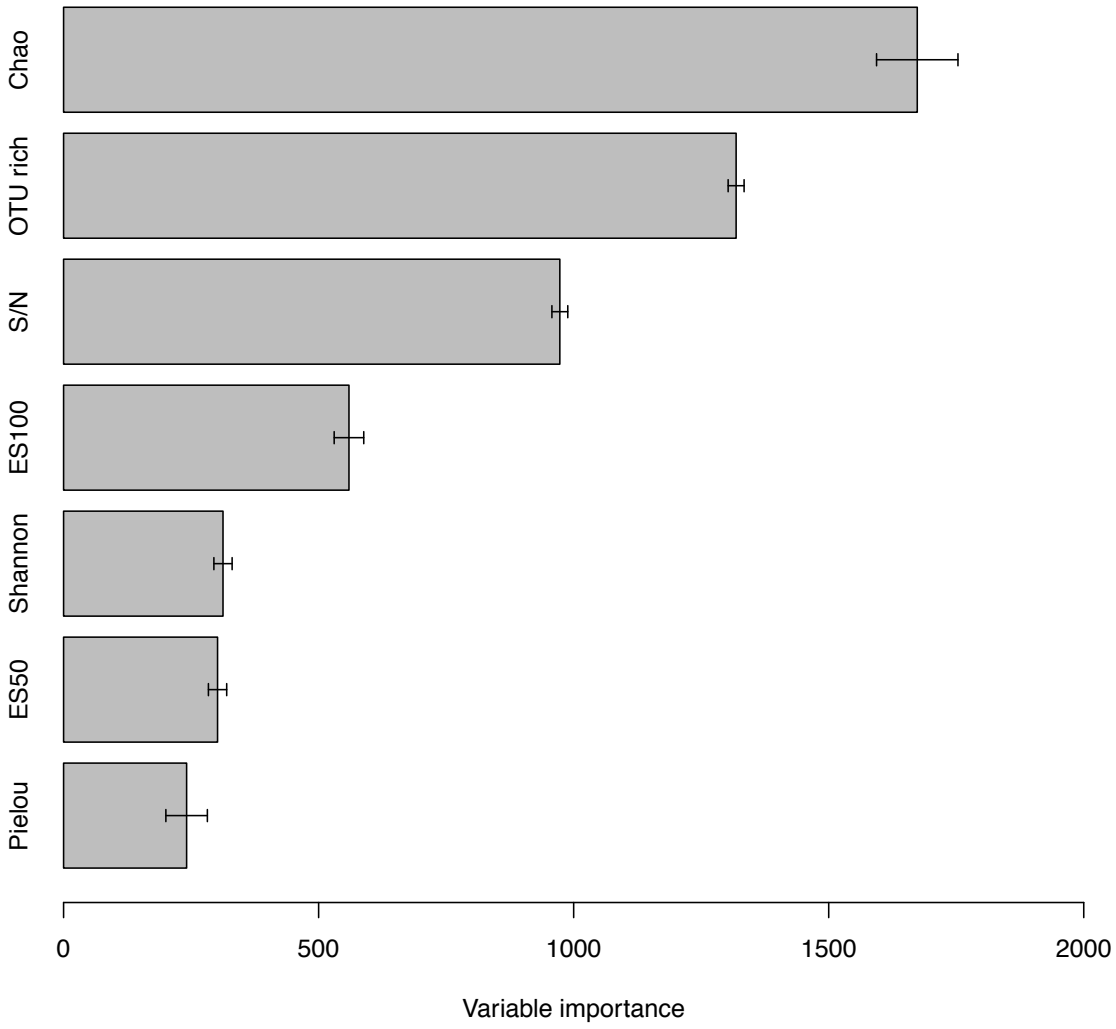

Figure S7: Diversity metrics importance in the prediction of the NQI1 index from the Random Forest models built from diversity metrics computed over 100 rarefied OTU-to-sample matrix. Error bars represent the standard deviation of importance over the 100 rarefied datasets.

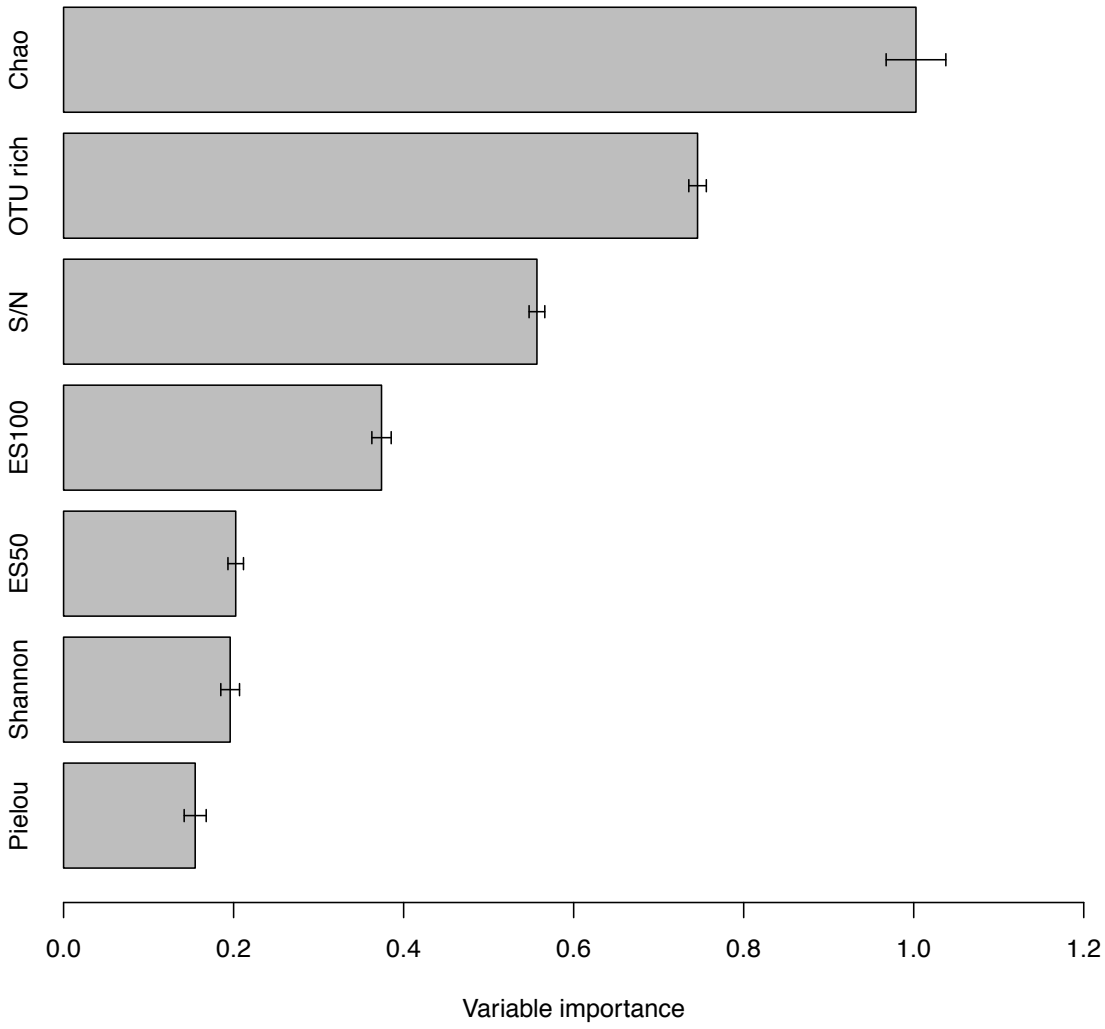

Figure S8: OTU importance in the prediction of the AMBI index from the Random Forest model built from composition data. The percentage of the total number of reads is indicated under the OTU identification number. Taxonomic assignment is indicated in the bars.

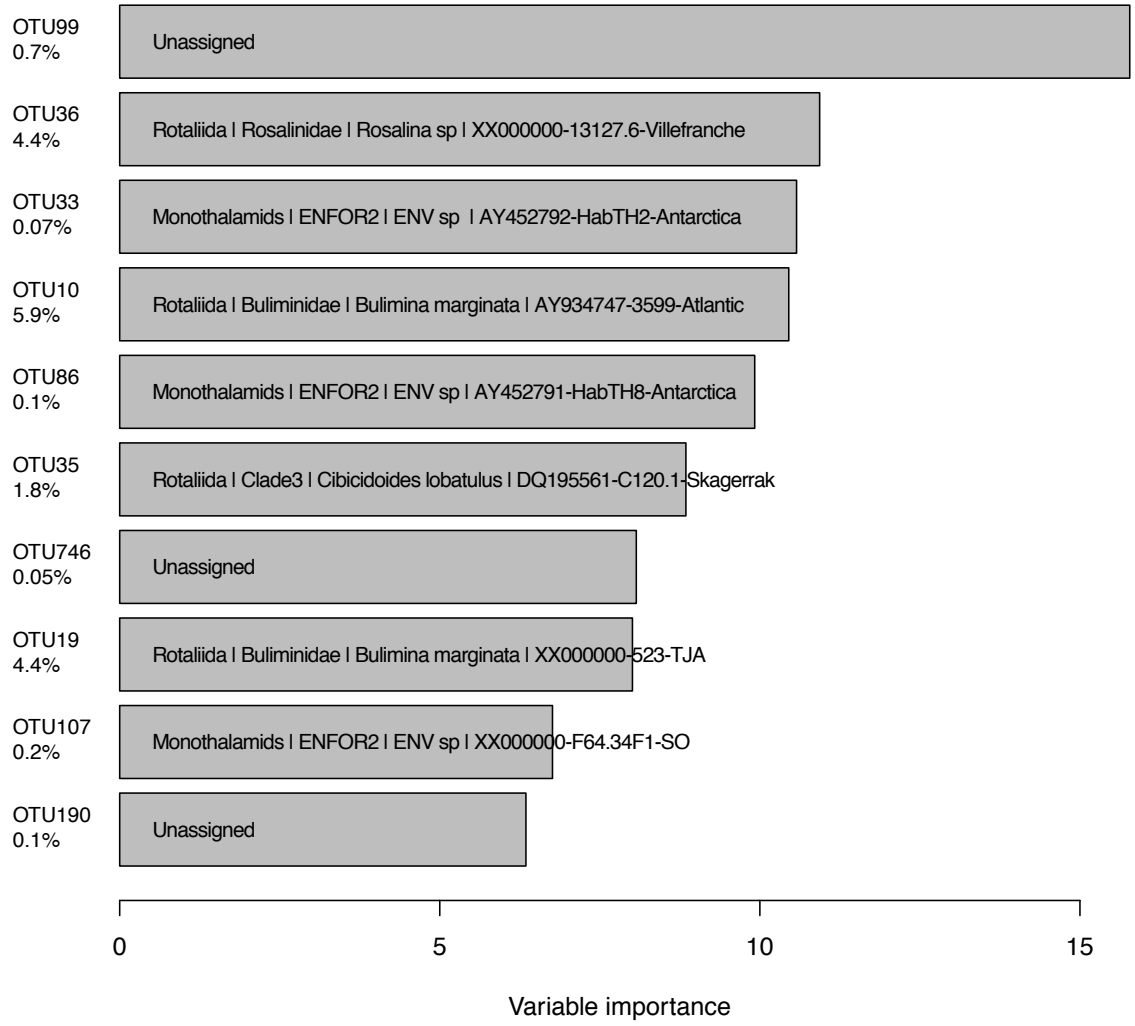

Figure S9: OTU importance in the prediction of the ISI index from the Random Forest model built from composition data. The percentage of the total number of reads is indicated under the OTU identification number. Taxonomic assignment is indicated in the bars

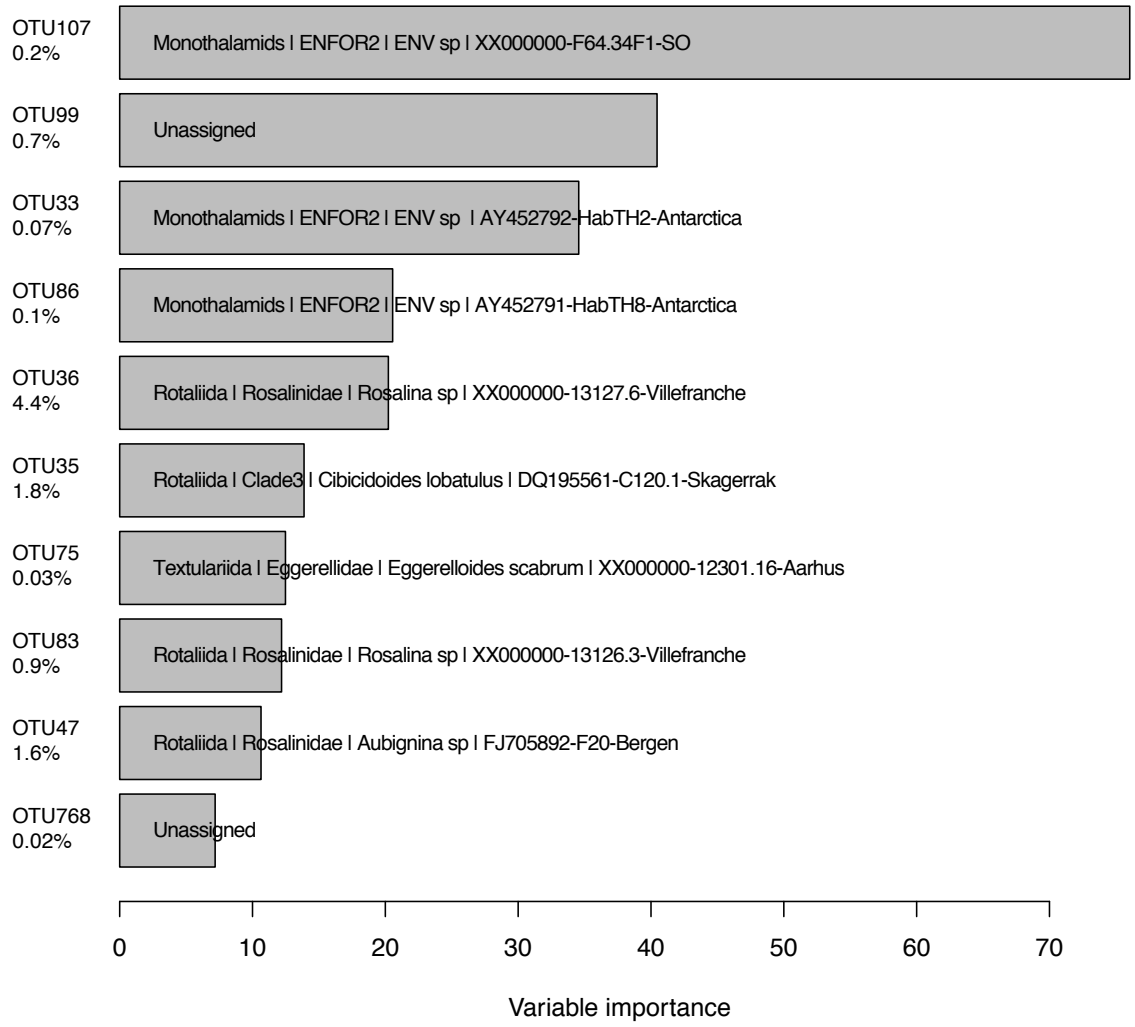

Figure S10: Boxplot and Mann-Whitney tests for difference between predicted BI values with the best predictive models and reference BI values for each combination of farm and BI. Results of the test are indicated on each plot, and ns is for non-significant (p-value > 0.05). Numbers on boxplots indicate the median and quartiles values (black numbers) and the mean value (red number).