Seqenv: linking sequences to environments through text mining

Lucas Sinclair [1] , **Umer Z Ijaz** [2] , **Lars Jensen** [3] , **Marco JL Coolen** [4] , **Cecile Gubry-Rangin** [5] , **Alica Chroňáková** [6] , **Anastasis Oulas** [7, 8] , **Christina Pavloudi** [8] , **Julia Schnetzer** [9] , **Aaron Weimann** [10] , **Ali Ijaz** [11] , **Alexander Eiler** [1] , **Christopher Quince** [Corresp., 12] , **Evangelos Pafilis** [Corresp. 8]

[1] Department of Ecology and Genetics, Limnology, Uppsala University, Uppsala, Sweden

[2] Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, Glasgow, United Kingdom

[3] The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[4] Western Australia Organic and Isotope Geochemistry Centre (WA-OIGC), Department of Chemistry, Curtin University of Technology, Bentley, WA, Australia

[5] Institute of Biological & Environmental Sciences, University of Aberdeen, Aberdeen, United Kingdom

[6] Institute of Soil Biology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic

[7] Bioinformatics Group, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus

[8] Institute of Marine Biology Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), Heraklion Crete, Greece

[9] Department of Molecular Ecology, Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

[10] Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Braunschweig, Germany

[11] Hawkesbury Institute for the Environment, University of Western Sydney, Hawkesbury, Sydney, Australia

[12] Warwick Medical School, University of Warwick, Warwick, United Kingdom

Corresponding Authors: Christopher Quince, Evangelos Pafilis
Email address: c.quince@warwick.ac.uk, pafilis@hcmr.gr

Understanding the distribution of taxa and associated traits across different environments is one of the central questions in microbial ecology. High-throughput sequencing (HTS) studies are presently generating huge volumes of data to address this biogeographical topic. However, these studies are often focused on specific environment types or processes leading to the production of individual, unconnected datasets. The large amounts of legacy sequence data with associated metadata that exist can be harnessed to better place the genetic information found in these surveys into a wider environmental context. Here we introduce a software program, *seqenv*, to carry out precisely such a task. It automatically performs similarity searches of short sequences against the "nt" nucleotide database provided by NCBI and, out of every hit, extracts – if it is available – the textual metadata field. After collecting all the isolation sources from all the search results, we run a text mining algorithm to identify and parse words that are associated with the Environmental Ontology (EnvO) controlled vocabulary. This, in turn, enables us to determine both in which environments individual sequences or taxa have previously been observed and, by weighted summation of those results, to summarize complete samples. We present two demonstrative applications of *seqenv* to a survey of ammonia oxidizing archaea as well as to a plankton paleome dataset from the Black Sea. These demonstrate

the ability of the tool to reveal novel patterns in HTS and its utility in the fields of environmental source tracking, paleontology, and studies of microbial biogeography. To install *seqenv*, go to: **https://github.com/xapple/seqenv**

Seqenv: linking sequences to environments through text mining

2   Lucas Sinclair[1,†], Umer Z Ijaz[2,†], Lars Juhl Jensen[3], Marco Coolen[4], Cecile Gubry-Rangin[5], Alica

Chroňáková[6], Anastasis Oulas[7,8], Christina Pavloudi[7], Julia Schnetzer[9], Aaron Weimann[10], Ali

4   Zeeshan Ijaz[11], Alexander Eiler[1], Christopher Quince[12,*], Evangelos Pafilis[7,*].

   1 Department of Ecology and Genetics, Limnology, Uppsala University, Sweden

6   **2** Infrastructure and Environment Research Division, School of Engineering, University of Glasgow,

   United-Kingdom

8   **3** The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences,

   University of Copenhagen, Denmark

10 **4** Western Australia Organic and Isotope Geochemistry Centre (WA-OIGC), Department of Chemistry,

   Bentley Campus, Curtin University, Australia

12 **5** Institute of Biological & Environmental Sciences, University of Aberdeen, Scotland, United Kingdom

   6 Biology Centre of the Czech Academy of Sciences, Institute of Soil Biology, Czech Republic

14 **7** Institute of Marine Biology Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine

   Research (HCMR), Heraklion Crete, Greece

16 **8** Bioinformatics Group, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus **9** Max

   Planck Institute for Marine Microbiology, Department of Molecular Ecology, Microbial Genomics and

18 Bioinformatics Group, Bremen, Germany

   10 Computational Biology of Infection Research, Helmholtz Centre for Infection Research,

20 Braunschweig, Germany

   11 Hawkesbury Institute for the Environment, Western Sydney University, Hawkesbury, Australia

22 **12** Warwick Medical School, University of Warwick, Coventry, United-Kingdom

   † These authors contributed jointly to this work.

24 * Corresponding authors: C. Quince, `c.quince@warwick.ac.uk`

   * Address: Warwick Medical School, University of Warwick, Coventry CV4 7AL, U.K.

26 * Corresponding authors: E. Pafilis, `pafilis@hcmr.gr`

∗ Address: Institute of Marine Biology Biotechnology and Aquaculture, Hellenic Centre for Marine

28 Research, P.O. Box 2214, Heraklion, 71003, Crete, Greece.

Version: October 8, 2016

Abstract

Understanding the distribution of taxa and associated traits across different environments is one of the

32 central questions in microbial ecology. High-throughput sequencing (HTS) studies are presently generating huge volumes of data to address this biogeographical topic. However, these studies are often

34 focused on specific environment types or processes, leading to the production of individual unconnected datasets. The large amounts of legacy sequence data with associated metadata that exist can be harnessed

36 to better place the genetic information found in these surveys into a wider environmental context. Here we introduce a software program, `seqenv`, to carry out precisely such a task. It automatically performs

38 similarity searches of short sequences against the "nt" nucleotide database provided by NCBI and, out of every hit, extracts – if it is available – the <isolation source> textual metadata field. After collecting

40 all the isolation sources from all the search results, we run a text mining algorithm to identify and parse words that are associated with the Environmental Ontology (EnvO) controlled vocabulary. This, in turn,

42 enables us to determine both in which environments individual sequences or taxa have previously been observed and, by weighted summation of those results, to summarize complete samples. We present two

44 demonstrative applications of `seqenv` to a survey of ammonia oxidizing archaea as well as to a plankton paleome dataset from the Black Sea. These demonstrate the ability of the tool to reveal novel patterns in

46 HTS and its utility in the fields of environmental source tracking, paleontology, and studies of microbial biogeography. To install `seqenv`, go to: `https://github.com/xapple/seqenv`.

Introduction

The annotation of DNA sequences, *i.e.* attaching meaningful labels to them, is key to the interpretation

50 of genomics data. In essence, this process gives context to a sequence. For instance, annotation reveals

the taxon from which the sequence was derived [1] and/or the gene families and potential functions

52   [2]. However, one type of annotation for which no automated bioinformatics pipeline currently exists

is the annotation to the environmental source. In other words, determining the types of environment in

54   which a given sequence has previously been found. We introduce a new program titled "`seqenv`" which

addresses this gap, automatically labeling sequences to the Environmental Ontology (EnvO) [3]. We

56   apply this bioinformatics pipeline to two datasets of environmental marker genes derived from terrestrial

archaeal ammonia oxidizers (AOA) [4] and the Black Sea plankton paleome [5]. This method reveals

58   hitherto unknown patterns in AOA diversity, and adds to our understanding of the geological history of

the Black Sea.

60   Annotating sequences to environments has become increasingly relevant as a result of the growing

application of environmental genomics to microbiology. In environmental genomics, microbial DNA

62   is extracted directly from an environment and then sequenced, possibly following PCR amplification of

target marker genes such as the 16S rRNA gene [6]. The result is a catalog of the microorganisms present

64   in a particular sample. One of the first interrogations concerning such samples is to know what other

environments these organisms have been found in. The answer can reveal ecologically relevant insight

66   about those organisms and may provide evidence for contamination from other environments. There

exists a wealth of information in available databases (most notably the ones provided by NCBI) which

68   can be used to gain a detailed overview of the biogeography of particular sequence varieties. The strategy

adopted in `seqenv` is to take input sequences and match them against the NCBI's database using the

70   time-tested BLAST search algorithm [7].

       All hits within a level of identity approximating to species are kept and either the text field "iso-

72   lation source" extracted or the PubMed abstracts associated with the submission obtained. In general,

we have found the isolation source metadata to be the most dependable source of environmental infor-

74   mation and the results presented here are restricted to that field. A custom named entity recognition

(NER) system based on [8] is then used to label the resulting text with terms from the EnvO ontology

76   [3]. An ontology is a formal specifications of the terms in a particular knowledge domain and the rela-

tions among them. Ontologies are often represented as an acyclic directed graph. The Environmental

78   Ontology (http://environmentontology.org/) (or EnvO) provides an ontology for this concise, controlled

vocabulary for the description of environments. EnvO also has the appeal of having been adopted by the

80   Genomics Standards Consortium for metadata associated with environmental sequence submission [9].

The terms found associated with each sequence are then collated together to provide its environmental

82   context.

        This environmental annotation scheme can be applied to any type of sequence, protein coding or ri-

84   bosomal RNA. The sequences can be derived from a particular taxonomic grouping but they can also

correspond to operational taxonomic units (OTUs) used as proxies for taxa in environmental sequencing

86   studies [10]. In either case, the nature and diversity of environments associated with a particular microor-

ganism can elucidate and bring light to its ecology. Additionally, if OTUs are used, `seqenv` can also

88   incorporate their abundances across samples. This furnishes a sample-level description of the EnvO terms

produced by simply summing the terms associated with each OTU weighted by their relative abundance

90   in the sample. These tables can then be used as a basis for multivariate statistics that contrast communities

in terms of the environmental terms associated with their constituent organisms. This novel approach is

92   a powerful means for exploring sample level differences in the origin of community constituents.

        Recently, a method has been developed for automatically associating geographic longitude and lati-

94   tude coordinates to Genbank records through rule based text mining of associated PubMed Central arti-

cles [11]. Our approach is distinguished from this in two ways. Firstly, we start from sequences rather

96   than records, allowing us to examine the distribution of environmental contexts within a certain level

of sequence similarity. Secondly, we associate to EnvO terms rather than to geographic coordinates.

98   This makes `seqenv` more relevant to exploring the ecology of microbes, determining the distribution of

OTUs across environment types, as opposed to tracking viral outbreaks which was a previous focus [11].

100 The information that `seqenv` automatically generates can answer similar questions to those addressed

previously [12], where co-occurrence of OTUs across sampling sites was examined and isolation sources

102 were classified to EnvO terms through text matching. We provide this functionality in a single coherent

software pipeline and promote user-friendliness.

104   To illustrate the usefulness of our pipeline, we apply it to two different datasets. The first is a previ-

ously published study of AOA derived from 45 soils [4]. These soils present a range of pHs, enabling us

106   to uncover how the spectrum of environments from which these organisms derive varies with changing

pH. The second dataset is from sediment cores deriving from the Black Sea [5]. Here the 18S rRNA

108   gene was sequenced by targeted-metagenomics, determining the eukaryotic plankton community struc-

ture over the last twelve thousand years. We can use `seqenv` to relate the environmental preferences

110   of these organisms to changes in Black Sea geology, most notably the initial Mediterranean sea influx

(IMI), hence, providing insight into the Black Sea environment prior to the IMI event.

Materials and Methods

112

The `seqenv` pipeline proceeds through the following steps, as illustrated diagrammatically in figure

114   1. The input is a user-supplied FASTA file containing thousands of DNA sequences and, optionally, a

frequency file containing the frequency counts of the sequences across multiple samples. This file takes

116   the form of a tab delimited text file containing the count matrix. In typical usage, the sequences would

correspond to the consensus sequences of OTUs and the matrix would represent their frequencies across

118   samples. After the following procedure, multiple outputs are generated:

1. The first step that `seqenv` executes is the parsing of the FASTA input file. All the sequence names

120   are removed and replaced by a place-holder title following the sequence "C1", "C2", "C3", etc. In

this fashion, problems caused by odd encodings or ambiguous characters are circumvented.

122   2. The second step consists of an optional filtering of the sequences to include only the most abun-

dant *i.e.* highest total frequency across samples. As the computation time scales with the number

124   of inputs, this filtering can greatly increase performance while leaving results statistically unaf-

fected. The number of selected sequences is a customizable parameter and the default is to use all

126   sequences. If no frequency matrix is provided this step is skipped.

3. Next, every remaining sequence is compared to a database of the user's choice. By default, the "nt"

128   (nucleotide) database provided by NCBI is used and the BLAST algorithm is chosen to carry out the

similarity search [7]. This step is the most costly computationally. It can, however, be parallelized

130  by `seqenv` on multi-core systems by performing a simple and automatic input-chopping strategy.

4. Taking all the results from the sequence similarity search, the best hits are selected by filtering

132  them according to the e-value of the comparison, the coverage of one sequence against the other, the identity between one sequence and the other, and a maximum number of targets for each input

134  sequence. These parameters default to 0.0001, 0.97, 0.97, and 10 respectively.

5. For every search hit from every input sequence, the corresponding GenInfo Identifier (GI) of the

136  homologous target within the database is recorded. This creates a table that links every input sequence to zero, one or more GI numbers.

138 6. Then, we collect the "isolation source" text entries associated to all of the GI numbers recorded in the previous step, provided the GI number was associated with such a field in NCBI's database,

140  failing which it is discarded. No internet connection is required as all text entries are stored in an SQLite3 database and can be accessed locally by `seqenv`. This database links every GI number

142  to its PubMed identifier, along with its isolation source text.

7. Using all the isolation source texts collected in the previous step and a text mining module, we

144  proceed to identify all terms that contain some type of environmental information. Words such "glacier", "pelagic" or "forest" are extracted and connected to the controlled EnvO vocabulary.

146  This consists of a hierarchically organized network of descriptive terms. In particular, the frequency of occurrence of each word is noted. Concretely, this is done offline by using a named entity

148  recognition (NER) system. The NER algorithm is an optimized dictionary-based tagger, it searches for keywords associated with each ENVO term but also using a stop-list of problematic words [8].

150  The ability of the NER engine to tag text with ENVO terms was evaluated in [8] through comparison to a manually curated corpus this resulted in 87.8% precision and 77.0% recall, corresponding to

152  an F1 score of 82.0%. The results were placed into an SQLite3 database that is automatically downloaded on the first run of `seqenv`.

154 8. With all the computed information, we are now able to describe each input sequence by a set of

EnvO terms and their associated frequency forming a term-frequency vector. Across the whole
156   dataset, this forms a sequence-term matrix. This matrix **S** has elements $s_{j,k}$ given the weight of
the k*th* EnvO term associated with the j*th* sequence. These weights are calculated according to
158   three different normalization strategies. The first is named "flat" and consists of using the raw
occurrence counts. The second is termed "unique isolation" and will count every identical isolation
160   source only once within the same input sequence, removing duplicated entries. The third is titled
"unique pubmed unique isolation" and will uniquify the frequency counts based on the text entry
162   of the isolation sources, as well as on the PubMed identifiers from which the GIs are obtained,
removing all but one matching sequence in the event they pertain to the same study. In all cases,
164   the rows of the matrix are normalized to 1.0, such that $s_{j,k} = s'_{j,k} / \sum_l s'_{j,l}$, where we are denoting
the raw counts by $s'_{j,k}$. The default normalization strategy is "flat".

166   9. If the user supplied a frequency matrix (c.f. second step), we are able to describe every one of
the original biological samples by a set of EnvO terms and frequencies that are simply the sum
168   of the term vectors over all sequences, weighted by the abundance of that sequence in the sample.
Equivalently, the sample term matrix **N** elements $n_{i,k}$, is the matrix product of the frequency matrix
170   **F** elements $f_{i,j}$ and the sequence-term matrix, *i.e.* $n'_{i,k} = \sum_j f_{i,j} s_{j,k}$. Normalizing by the total
frequency in the sample, such that $n_{i,k} = n'_{i,k} / \sum_l n'_{i,l}$, we obtain sample term vectors such as,
172   translated to English: "Sample Z is 25% brackish estuary, 25% river and 50% wetland.".

   10. Other options are available to the user to further modify and filter the results. The "backtracking"
174   option, when activated, will propagate frequency counts up the acyclic directed graph described
by the ontology for every EnvO term identified by the text mining module. The "restrict" option,
176   when specified by passing a given EnvO identifier (e.g. ENVO:00010483), will force the output to
contain only descendants from a single EnvO term. In effect, all other terms that are not reachable
178   through the given node in the ontology graph are removed.

   11. The first output that is produced is a table serialized in the format of a tab-delimited plain text file
180   (TSV) representing the composition of each input sequence according to the EnvO terms associated

to them, *i.e.* the matrix **S**. The columns represent input sequence and rows represent the normalized weight of EnvO terms.

12. If the user provided a frequency matrix (as described in step 2), the program can produce a similar TSV table representing the composition of each biological sample according to the EnvO terms associated to them, *i.e.* the matrix **N**. In this case, columns represent samples and rows represent EnvO terms. Each value corresponds to the normalized weight of the EnvO term in the corresponding sample.

13. For each sample, a visual representation of the hierarchy of the EnvO terms occurring in the isolation source of its imputed close relatives can be made. A PDF file is generated for each sequence and, if the user provided an abundance table, for each sample. In addition, every PDF has a corresponding DOT file which can be viewed and manipulated with the Graphviz software.

14. Other intermediary outputs are available as well, such as the output of the similarity search and a precise list of every EnvO term found in each input sequence.

The `seqenv` package is written in Python. The code follows a clean architecture, is commented and object-oriented. It is free and open-source carrying an MIT license. It is available on github here: `https://github.com/xapple/seqenv`. It can be installed on any computer with Python by simply typing: "`pip install seqenv`" in your shell.

Results

Earlier versions of the `seqenv` pipeline have already been used in a number of published studies including an analysis of the degree of recruitment of marine bacteria from freshwater sources and the air [13], as well as a survey of bacterial diversity along a 2'600 km river continuum [14], and a study of hydrogenase genes in lake sediments [15].

Here, to further illustrate its utility, we will apply it to two published datasets and demonstrate that it provides additional insights into the processes that structure microbial communities not evident in the

original analyses. These two examples comprise:

206 1. A survey of archaeal *amoA* gene data from 45 British soils, originating from a broad range of pH (min. 3.5, max. 8.7, median 6.2) [4]. The sequences were generated by bidirectional 454 pyrose-

208 quencing of part of the *amoA* gene, reads were denoised with AmpliconNoise [16], overlapped and further error checked by removing those with stop codons when translated into amino acids. For

210 this part of the analysis, we generated operational taxonomic units (OTUs) at 5% sequence diver-gence using average linkage hierarchical clustering. This will be higher resolution than species

212 [17], corresponding to ecotypes with well defined environmental preferences. This procedure re-sulted in just 67 OTU sequences. All sequences were from archaeal ammonia oxidizers (AOAs)

214 as described previously [4].

2. The Black Sea Paleome. This study included 454 pyrosequencing of 18S rRNA gene amplicons

216 from 48 deep sediment samples collected from the Black Sea enabling the reconstruction of mi-crobial eukaryote populations up to 11'400 years in the past. The V1-V3 region was sequenced as

218 described previously [5]. Reads were denoised with AmpliconNoise and OTUs constructed at 3% sequence divergence using average linkage hierarchical clustering as species proxies [16]. A total

220 of 1'748 OTUs were obtained.

Patterns of ammonia oxidizing archaea (AOA) habitat usage

222 In total 67 OTUs were observed across the 45 samples. These OTUs have been previously demonstrated as having well defined pH preferences [4]. For each OTU, we calculated the mean of its pH range as the

224 weighted averaged of the samples it was observed in, *i.e.*:

$$\bar{Y}_s = \sum_{n=1}^{N} x_{n,s} Y_n$$

where $\bar{Y}_s$ is the mean pH range for OTU s and $x_{n,s}$ is the relative abundance of s in sample n,

226 which has pH Y_n. We ran `seqenv` on the 95% OTU centroid nucleotide sequences considering up to 100 matches with 95% overlap and 95% identity to the query. Once again, this procedure should

228   return all sequences within approximate species boundaries. The default "flat" normalization was used.

We restricted the analysis to all EnvO terms that inherit from the term "environmental material" which

230   is identified by the number ENVO:00010483. Thereby, the redundancy across different terms in our

analysis was reduced. In figure 2, we show the EnvO terms associated with two OTUs deriving from

232   the extremes of the observed pH ranges (C46 - 3.5) and (C66 - 8.5). These OTUs had two and fifteen

EnvO terms associated with them in total respectively. In all, we obtained EnvO terms for 66 OTUs. The

234   67th OTU did not match to any sequences carrying environmental information in the database. In the top

panel of figure 3, the total number of terms found for each OTU as a function of its preferred pH range is

236   plotted. A significant positive correlation between the diversity of habitats and the pH of the samples the

organism was found (adjusted R-squared: 0.274, p-value: 3.85e-06). Another weaker but still significant

238   positive association is observed between sample pH and total OTU diversity (adjusted R-squared: 0.131,

p-value: 0.00922).

240   To determine which EnvO terms were most associated with the pH preference of the AOA OTUs, we

performed a Random Forest regression [18] of pH preference against the weighted EnvO terms. Random

242   Forest uses an ensemble of decision trees constructed from artificial datasets generated by bootstrap

aggregation or <bagging>, *i.e.* sampling with replacement across samples. This is combined with random

244   selection of features. Since not all samples are used in each dataset, a robust estimate of model accuracy

is possible using the left out samples. Additionally, estimates of variable importance can be obtained by

246   comparing accuracy of prediction with and without randomly permuting the variable of interest. This

is measured by the statistic: percentage mean decrease of accuracy ($\%IncMSE$). We fitted a Random

248   Forest using the `randomForest` R package [19] [20]. The model explained 34.6% of the variation in

pH preference. In figure 4, we visualize the weights of the top ten most important terms as determined

250   by $\%IncMSE$ across OTUs ordered by their pH preference.

Environmental stages of the Black Sea paleome

252   The 48 sediment samples form a series from a Black Sea core spanning the last 11.4 thousand years before

present (kyBP). In Coolen *et al.* [5], they defined four "Environmental Stages" (ES) in the geological

254   evolution of the Black Sea that apply to this depth series on the basis of fossil evidence and isotope ratios:

- ES4: Lacustrine interval ($\sim$11.4–9.0 kyBP). During this lacustrine phase the Black Sea was discon-
256   nected from the Mediterranean Sea due to low sea levels. This phase ends with the initial marine
inflow (IMI) as rising sea levels, due to the end of the ice age 11'700 years ago, resulted in the
258   connection of the Black Sea to the Mediterranean.

- ES3: A period of increasing salinity ($\sim$9.0-5.2 kyBP) corresponding to the warm and moist mid-
260   Holocene climatic optimum.

- ES2: Establishment of modern environmental conditions ($\sim$5.2-2.5 kyBP) and further increasing
262   salinity associated with the onset of the dry Subboreal.

- ES1: Freshening ($\sim$2.5 kyBP to present) with onset of the cool and wet Subatlantic climate and
264   recent anthropogenic perturbations.

     In figure 5, we visualise the community structures of these samples, in terms of the 18S rRNA OTU
266   proportions using a 2D non-metric multi-dimensional scaling (NMDS). This is very similar to Figure
2A of Coolen *et al.* [5], except that the OTUs in our study were constructed differently, but we include
268   it here for the sake of completeness. The trajectory through time of the samples together with their
Environmental Stages are shown. From this it is clear that there is a coherent change in structure during
270   the geological history of the Black Sea and that the samples cluster according to ES.

     We next ran `seqenv` on the 1'748 18S rRNA OTU centroid sequences taking into account up to 100
272   matches with 97% overlap and 97% identity to the query. As above, we restricted the analysis to those
terms that inherit from the term ENVO:00010483 "environmental material" and used the "flat" normal-
274   ization option. The normalized term vectors for each OTU were then combined with the relative OTU
frequencies across the 48 sediment samples to obtain the weighted frequency of terms across samples,
276   as described above. In total we observed 99 separate EnvO terms across the 48 samples. As above, we
used a random forest classifier to predict these environmental stages from the EnvO terms associated with
278   each sample. This classifier had an error rate of 12.5%. In figure 6 we show the relative frequency of

the ten most important terms in this classifier across the samples, ordered by age and with the ES groups

280   indicated.

Discussion

282   The two analyses presented above demonstrate the value of using `seqenv` to associate EnvO terms with

both individual OTUs and whole samples. In the analysis of AOA OTUs, we demonstrated a significant

284   association between the pH that an OTU is adapted to and the diversity of environments that it is found.

These results indicate that, as their optimum pH increases, the AOA OTUs are present across a greater

286   diversity of habitats. As in the original study, a statistically significant relationship between sample OTU

richness and pH was observed. That is, as the pH of a sample increases, more species are observed. We

288   propose that these two observations may be connected: the fact that more environments appear accessi-

ble to the OTUs as the pH increases may generate the diversification of species that is reflected in the

290   increasing sample richness with pH. At higher pHs, we might expect both of these relationships to be

reversed due to increased competition with bacterial ammonia oxidizers.

292   In the geological history of the Black Sea, one of the key questions is the nature of that environment

prior to the initial Mediterranean sea influx (IMI). For example, was it a Brackish environment, or was it

294   akin to a freshwater lake landscape? In our Black Sea dataset analysis, we can note a discrete change in

the EnvO terms associated with the samples at this event when we transition from ES4 to ES3. Prior to

296   this point, terms such as "freshwater lake" and "river" are frequent, afterwards the samples are dominated

by organisms associated with "sea water", "ocean water" and "estuary". The microbial community prior

298   to the IMI comprised organisms associated with freshwater habitats, important evidence that the IMI was

associated with a substantial increase in salinity.

300 ## Conclusion

The two studies described in this paper are not intended to be exhaustive, but present convincing vignettes

302   of the usefulness of `seqenv`. We believe the methods presented here will prove to be an effective and

extremely valuable tool to the community for distilling, analyzing and adding context to DNA sequence

304   data. There may be areas in which our methodology could be improved for example by weighting terms in

OTUs or samples using more sophisticated approaches from information retrieval. In any case, we believe

306   that, in the future, `seqenv` will contribute crucial insights and advances to the field of environmental

metagenomics.

308 Acknowledgements

Seqenv was originally conceived in a series of <hackathons> supported by the European Union's Earth

310   System Science and Environmental Management COST Action. This project was titled "Microbial ecol-

ogy & the earth system: collaborating for insight and success with the new generation of sequencing

312   tools" and can be viewed at `http://www.cost.eu/domains_actions/essem/Actions/`
`ES1103`.

314   We would like to thank the LifeWatchGreece project (http://www.lifewatchgreece.eu/) for their gen-

erous support in the organization of these meetings.

Abbreviations

316

16S	16 Svedberg sedimentation mark (non-SI unit)
ANOVA	Analysis of variance
AOA	Ammonia oxidising archaea
BLAST	Basic local alignment search tool
BP	Base pair (of nucleotides)
EnvO	Environmental ontology
GI	GenInfo identifier
HTS	High-throughput (genetic) sequencing
MSE	Mean squared error
NCBI	National center for biotechnology information (in the US)
NER	Named entity recognition
NMDS	Non-parametric multidimensional scaling (ordination plot)
OTU	Operational Taxonomic Unit
PCR	Polymerase chain reaction
rRNA	Ribosomal ribonucleic acid
TSV	Tab separated values

References

318

[1] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole (2007) **Naive Bayesian clas-**
320   **sifier for rapid assignment of rRNA sequences into the new bacterial taxonomy**. *Applied and Environmental Microbiology*, *volume 73* (16), 5261

322 [2] Agnieszka S Juncker, Lars J Jensen, Andrea Pierleoni, Andreas Bernsel, Michael L Tress, Peer Bork, Gunnar von Heijne, Alfonso Valencia, Christos A Ouzounis, Rita Casadio, and Søren Brunak

324   (2009) **Sequence-based feature prediction and annotation of proteins**. *Genome biology*, *volume 10* (206)

326   [3] Pier Buttigieg, Norman Morrison, Barry Smith, Christopher J Mungall, Suzanna E Lewis, and the ENVO Consortium (2013) **The environment ontology: contextualising biological and biomed-**

328   **ical entities**. *Journal of Biomedical Semantics*, *volume 4* (1), 43

       [4] Cécile Gubry-Rangin, Brigitte Hai, Christopher Quince, Marion Engel, Bruce C Thomson, Phillip

330   James, Michael Schloter, Robert I Griffiths, James I Prosser, and Graeme W Nicol (2011) **Niche**

       specialization of terrestrial archaeal ammonia oxidizers. *Proceedings of the National Academy*

332   *of Sciences of the United States of America*, *volume 108* (52), 21206

       [5] Marco J L Coolen, William D Orsi, Cherel Balkema, Christopher Quince, Keith Harris, Sean P

334   Sylva, Mariana Filipova-Marinova, and Liviu Giosan (2013) **Evolution of the plankton paleome**

       in the Black Sea from the Deglacial to Anthropocene. *Proceedings of the National Academy of*

336   *Sciences of the United States of America*, *volume 110* (21), 8609

       [6] Ramiro Logares, Thomas H A Haverkamp, Surendra Kumar, Anders Lanzén, Alexander J Neder-

338   bragt, Christopher Quince, and Håvard Kauserud (2012) **Environmental microbiology through**

       the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinfor-

340   **matics approaches**. *Journal of Microbiological Methods*, *volume 91* (1), 106

       [7] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman (1990) **Basic local alignment search**

342   **tool**. *Journal of molecular biology*, *volume 215* (3), 403

       [8] Evangelos Pafilis, Sune P Frankild, Julia Schnetzer, Lucia Fanini, Sarah Faulwetter, Christina

344   Pavloudi, Aikaterini Vasileiadou, Patrick Leary, Jennifer Hammock, Katja Schulz, Cynthia Sims

       Parr, Christos Arvanitidis, and Lars Juhl Jensen (2015) **ENVIRONMENTS and EOL: identifica-**

346   **tion of Environment Ontology terms in text and the annotation of the Encyclopedia of Life**.

       Bioinformatics, *volume 31* (11), 045

348 [9] Dawn Field, Peter Sterk, Renzo Kottmann, J Wim De Smet, Linda Amaral-Zettler, Guy Cochrane, James R Cole, Neil Davies, Peter Dawyndt, George M Garrity, Jack A Gilbert, Frank Oliver

350 Glöckner, Lynette Hirschman, Hans-Peter Klenk, Rob Knight, Nikos Kyrpides, Folker Meyer, Ilene Karsch-Mizrachi, Norman Morrison, Robert Robbins, Inigo San Gil, Susanna Sansone, Lynn

352 Schriml, Tatiana Tatusova, Dave Ussery, Pelin Yilmaz, Owen White, John Wooley, and Gregory Caporaso (2014) **Genomic standards consortium projects**. *Standards in Genomic Sciences*, *vol-*

354 *ume 9* (3), 599

[10] Mark Blaxter, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd, and

356 Eyualem Abebe (2005) **Defining operational taxonomic units using DNA barcode data**. *Philo-sophical transactions of the Royal Society of London. Series B, Biological sciences*, *volume*

358 *360* (1462), 1935

[11] Tasnia Tahsin, Davy Weissenbacher, Robert Rivera, Rachel Beard, Mari Firago, Garrick Wallstrom,

360 Matthew Scotch, and Graciela Gonzalez (2016) **A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records**. *Journal of the American Medical In-*

362 *formatics Association*

[12] S Chaffron, H Rehrauer, J Pernthaler, and C von Mering (2010) **A global network of coexisting mi-**

364 **crobes from environmental and whole-genome sequence data**. *Genome Research*, *volume 20* (7), 947

366 [13] Jérôme Comte, Eva S Lindström, Alexander Eiler, and Silke Langenheder (2014) **Can marine bacteria be recruited from freshwater sources and the air?**. *The ISME Journal*, *volume 8* (12),

368 2423

[14] Domenico Savio, Lucas Sinclair, Umer Z Ijaz, Juraj Parajka, Georg H Reischer, Philipp Stadler, Al-

370 fred P Blaschke, Günter Blöschl, Robert L Mach, Alexander K T Kirschner, Andreas H Farnleitner, and Alexander Eiler (2015) **Bacterial diversity along a 2600 km river continuum**. *Environmental*

372 *Microbiology*, *volume 17* (12), 4994

[15] Jillian M Couto, Umer Zeeshan Ijaz, Vernon R Phoenix, Melanie Schirmer, and William T Sloan (2015) **Metagenomic Sequencing Unravels Gene Fragments with Phylogenetic Signatures of O2-Tolerant NiFe Membrane-Bound Hydrogenases in Lacustrine Sediment**. *Current microbiology*, *volume 71* (2), 296

[16] Christopher Quince, Anders Lanzén, Russell J Davenport, and Peter J Turnbaugh (2011) **Removing noise from pyrosequenced amplicons**. *BMC Bioinformatics*, *volume 12* (1), 38

[17] Michael Pester, Thomas Rattei, Stefan Flechl, Alexander Gröngröft, Andreas Richter, Jörg Overmann, Barbara Reinhold-Hurek, Alexander Loy, and Michael Wagner (2011) **amoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA genes from soils of four different geographic regions**. *Environmental Microbiology*, *volume 14* (2), 525

[18] Leo Breiman (2001) **Random Forests**. *Machine learning*, *volume 45* (1), 5

[19] R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

[20] A Liaw and M Wiener (2002) **Classification and regression by randomForest**. *R news*, *volume 2* (3), 18

388 **Figures**
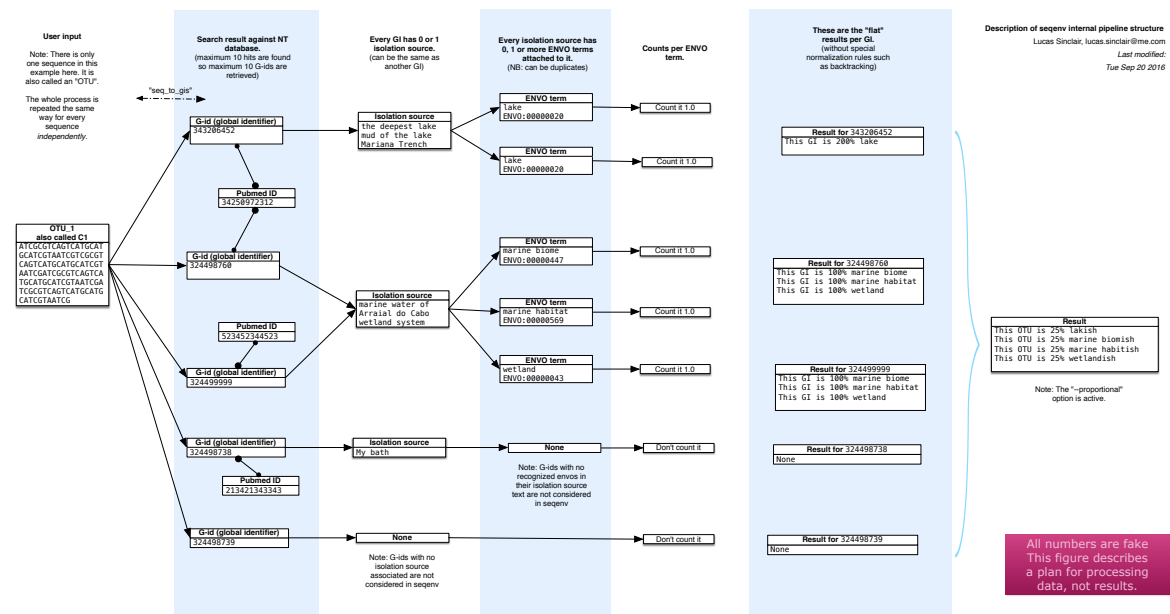
Figure 1: Schematic of the internal functioning of the seqenv pipeline.

This figure details how EnvO term frequencies are computed. The numbers provided are fictional

390 as the schematic focuses on representing the internal functioning of the pipeline and does not illustrate a

concrete case. As each inputed short DNA sequence is processed independently in all but the last stages

392 of seqenv, only one input sequence is shown here.
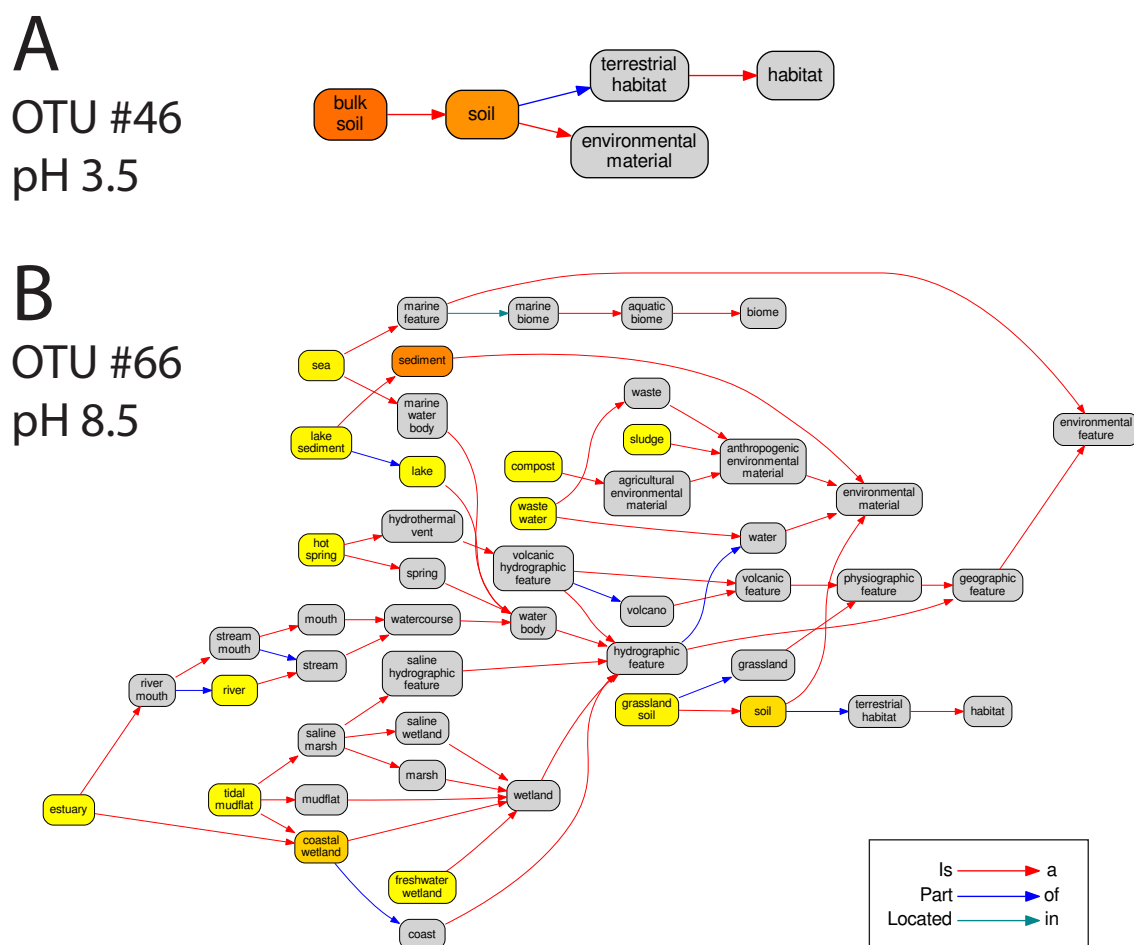
Figure 2: The EnvO terms associated with two AOA OTUs.

For each original inputted sequence, `seqenv` outputs a network representing the EnvO terms identified. Two examples of such hierarchical ontologies are shown. The two OTUs chosen had a mean pH of 3.5 and 8.5. The intensity of the node's background color reflects the frequency of that term within hits. Gray indicates the lowest frequency recorded and darker shades of yellow to orange indicate higher frequencies.
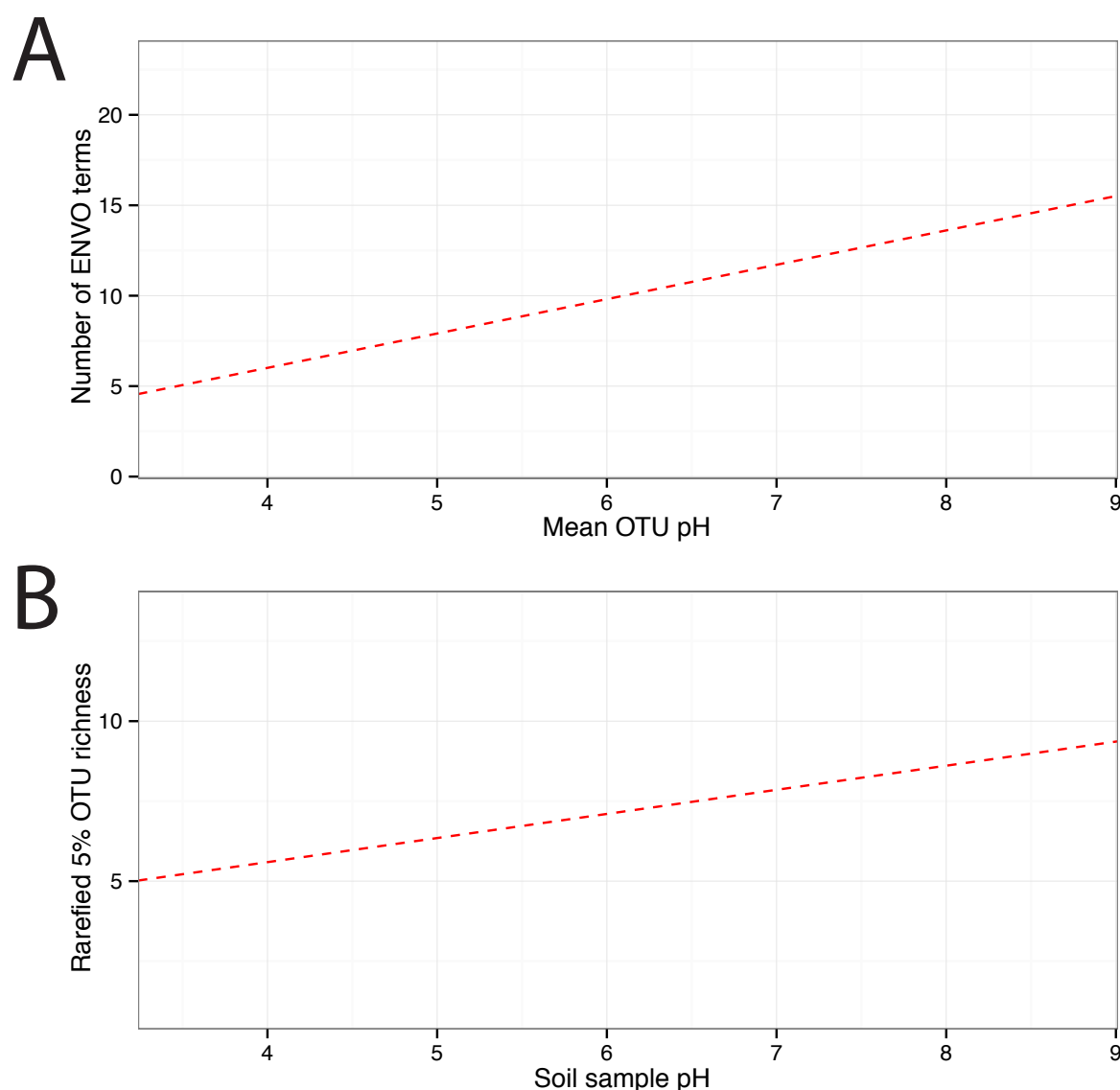
Figure 3: EnvO terms and OTU richness against mean OTU pH.

398   Figure 3A shows the total number of EnvO terms against OTU pH. The dashed red line indicates a linear regression of number of EnvO terms with OTU pH (adjusted R-squared: 0.2742, p-value: 3.85e-

400 06). Figure 3B shows the community OTU diversity against sample pH for the AOA dataset. OTU richness was calculated after rarefying to 1'000 reads. Linear regression of sample diversity against pH

402 (adjusted R-squared: 0.1305, p-value: 0.009217).
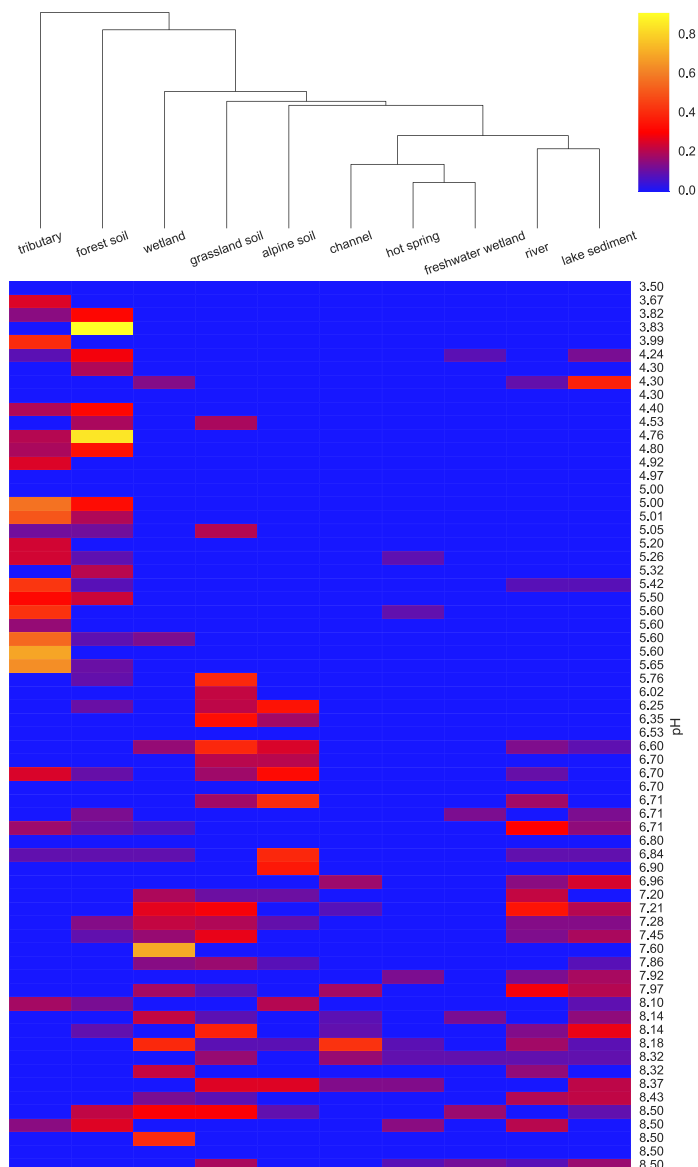
Figure 4: Heatmap of top ten EnvO terms for determining OTU pH.

Random forests were used to perform a regression of pH against EnvO terms (Var. explained: 34.6%). The abundance of the top ten most important terms, as determined by percentage mean decrease of accuracy ($\%IncMSE$) are shown across OTUs ordered by their pH preference.
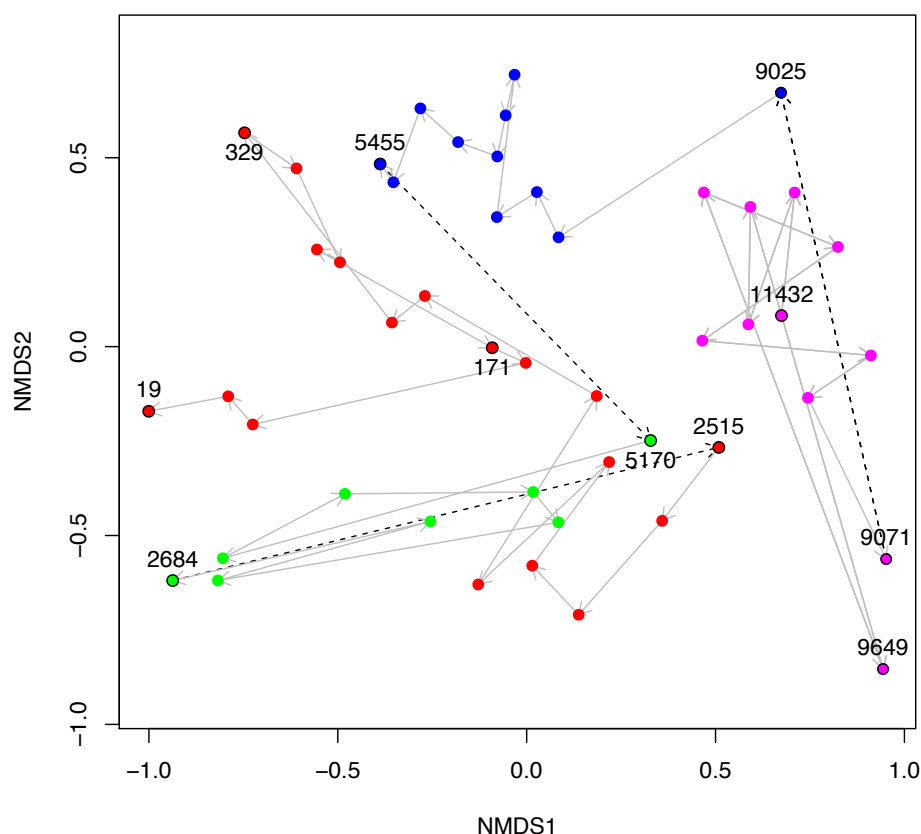
Figure 5: NMDS plot of Black Sea plankton 18S rRNA samples.

406   Non-metric multidimensional scaling (NMDS) of OTU relative abundances with Bray-Curtis dis-
tances were used to ordinate the 18S rRNA Black Sea plankton samples in two dimensions. The age

408 of key samples are indicated together with the Environmental Stage: ES4 (magenta), ES3 (blue), ES2
(green) and ES1 (red). Arrows indicate the temporal succession of samples, and dotted arrows represent

410 the transition between environmental stages.
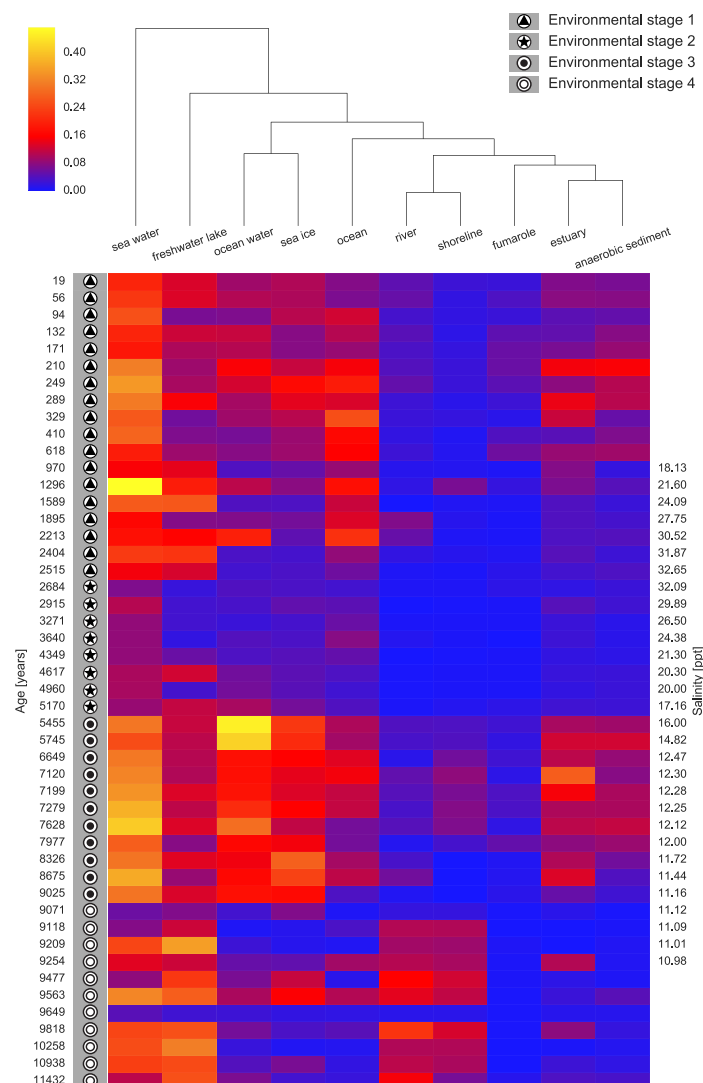
Figure 6: Heatmap of top ten EnvO terms for determining ES.

Random forests were used to perform a classification of the Black Sea Environmental Stage (ES) against EnvO terms (Error rate: 12.5%). The abundance of the top ten most important terms, as determined by the percentage mean decrease of error rate ($\%IncMSE$) are shown. Samples are labelled with ES and order by time before present, and salinity is given for the central part of the sediment core.