

Seqenv: linking sequences to environments through text mining (#12292)

1

First submission

Please read the **Important notes** below, and the **Review guidance** on the next page. When ready [submit online](#). The manuscript starts on page 3.

Important notes

Editor and deadline

Timothy Read / 11 Aug 2016

Files

Please visit the overview page to [download and review](#) the files not included in this review pdf.

Declarations

One or more DNA sequences were reported.




Please in full read before you begin

How to review






When ready [submit your review online](#). The review form is divided into 5 sections. Please consider these when composing your review:

- 1. BASIC REPORTING**
- 2. EXPERIMENTAL DESIGN**
- 3. VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor



 You can also annotate this **pdf** and upload it as part of your review

To finish, enter your editorial recommendation (accept, revise or reject) and submit.







BASIC REPORTING

-  Clear, unambiguous, professional English language used throughout.
-  Intro & background to show context. Literature well referenced & relevant.
-  Structure conforms to [PeerJ standard](#), discipline norm, or improved for clarity.
-  Figures are relevant, high quality, well labelled & described.
-  Raw data supplied (See [PeerJ policy](#)).

VALIDITY OF THE FINDINGS

-  Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
-  Data is robust, statistically sound, & controlled.

EXPERIMENTAL DESIGN

-  Original primary research within [Scope of the journal](#).
-  Research question well defined, relevant & meaningful. It is stated how research fills an identified knowledge gap.
-  Rigorous investigation performed to a high technical & ethical standard.
-  Methods described with sufficient detail & information to replicate.
-  Conclusion well stated, linked to original research question & limited to supporting results.
-  Speculation is welcome, but should be identified as such.

The above is the editorial criteria summary. To view in full visit <https://peerj.com/about/editorial-criteria/>

Seqenv: linking sequences to environments through text mining

Lucas Sinclair^{1,2}, Umer Ijaz³, Lars Jensen⁴, Marco Coolen⁵, Cecile Gubry-Rangin⁶, Alica Chroňáková⁷, Anastasis Oulas⁸, Christina Pavloudi⁸, Julia Schnetzer⁹, Aaron Weimann¹⁰, Ali Ijaz¹¹, Alexander Eiler^{1,2}, Christopher Quince^{Corresp., 12}, Evangelos Pafilis^{Corresp. 8}

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

² Environmental bioinformatics consultants, Envonautics Ltd., Göteborg, Sweden

³ Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, Glasgow, United Kingdom

⁴ The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

⁵ Western Australia Organic and Isotope Geochemistry Centre (WA-OIGC), Department of Chemistry, Curtin University of Technology, Curtin, Australia

⁶ Institute of Biological & Environmental Sciences, University of Aberdeen, Aberdeen, United Kingdom

⁷ Institute of Soil Biology, The Czech Academy of Sciences, Prague, Czech Republic

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

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

¹⁰ Department of Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, 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, go to: **<https://github.com/xapple/seqenv>**

Seqenv: linking sequences to environments through text mining

2 Lucas Sinclair^{1,2,†}, Umer Zeeshan Ijaz^{3,†}, Lars Juhl Jensen⁴, Marco Coolen⁵, Cecile Gubry-Rangin⁶,
Alica Chroňáková⁷, Anastasis Oulas⁸, Christina Pavloudi⁸, Julia Schnetzer⁹, Aaron Weimann¹⁰, Ali
4 Zeeshan Ijaz¹¹, Alexander Eiler^{1,2}, Christopher Quince^{12,*}, Evangelos Pafilis^{8,*}.

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

6 2 Envonautics.com - environmental bioinformatics consultants

3 Infrastructure and Environment Research Division, School of Engineering, University of Glasgow,

8 United-Kingdom

4 The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences,

10 University of Copenhagen, Denmark

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

12 Bentley Campus, Curtin University, Australia

6 Institute of Biological & Environmental Sciences, University of Aberdeen, Scotland, United Kingdom

14 7 Biology Centre of the Czech Academy of Sciences, Institute of Soil Biology, Czech Republic

8 Institute of Marine Biology Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine

16 Research (HCMR), Crete, Greece

9 Max Planck Institute for Marine Microbiology, Department of Molecular Ecology, Microbial

18 Genomics and Bioinformatics Group, Bremen, Germany

10 Department of Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany

20 11 Hawkesbury Institute for the Environment, Western Sydney University, Australia

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

22 † These authors contributed jointly to this work.

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

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


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

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

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

28 *Version: July 25, 2016*

Abstract

30 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 gen-
32 erating 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
34 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
36 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
38 of every hit, extracts – if it is available – the <isolation source> textual metadata field. After collecting
all the isolation sources from all the search results, we run a text mining algorithm to identify and parse
40 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
42 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
44 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
46 biogeography.  install, go to: <https://github.com/xapple/seqenv>.

Introduction

48 The annotation of DNA sequences, *i.e.* attaching meaningful labels to them, is key to the interpretation of
genomics data. In essence, this process gives context to a sequence. For instance, annotation reveals the
50 taxon from which the sequence was derived [1] and/or gene families potential functions [2]. However,


one type of annotation for which no automated bioinformatics pipeline currently exists is the annotation
52 to the environmental source. In other words, determining the types of environment in which a given
sequence has previously been found. We introduce a new program titled “seqenv” which addresses
54 this gap, automatically labeling sequences to the Environmental Ontology (EnvO) [3]. We apply this
bioinformatics pipeline to two datasets of environmental marker genes derived from terrestrial archaeal
56 ammonia oxidizers (AOA) [4] and the Black Sea plankton paleome [5]. This method reveals, h^{to},
unknown patterns in AOA diversity, and adds to our understanding of the geological history of the Black
58 Sea.

Annotating sequences to environments has become increasingly relevant as a result of the growing
60 application of environmental genomics to microbiology. In environmental genomics, microbial DNA
is extracted directly from an environment and then sequenced, possibly following PCR amplification
62 of target marker genes such as the 16S rRNA gene [6]. The result is a catalog of the microorganisms
present in a particular sample. One of the first interrogations concerning such samples is to know what
64 other environments have these organisms been found. The answer can reveal ecologically relevant
insight about those organisms and may provide evidence for contamination from other environments.
66 There exists a wealth of information in available databases (most notably the ones provided by NCBI)
which can be used to gain a detailed overview of the biogeography of a particular sequence varieties.
68 The strategy adopted in seqenv is to take input sequences and match them against the NCBI’s database
using the time-tested BLAST search algorithm [7].

70 All hits within a level of identity approximating to species are kept and either the text field “iso-
lation source” extracted or the PubMed abstracts associated with the submission obtained. In general,
72 we have found the isolation source metadata to be the most dependable source of environmental infor-
mation and the results presented here are restricted to that field. A custom named entity recognition
74 (NER) system based on [8] is then used to label the resulting text with terms from the EnvO ontology
[3]. An ontology is a formal specifications of the terms in a particular knowledge domain and the rela-
76 tions among them. Ontologies are often represented as an acyclic directed graph. The Environmental
Ontology (<http://environmentontology.org/>) (or EnvO) provides an ontology for this concise, controlled

78 vocabulary for the description of environments. EnvO also has the appeal of having been adopted by the
Genomics Standards Consortium for metadata associated with environmental sequence submission [9].
80 The terms found associated with each sequence are then collated together to provide its environmental
context.

82 We can apply this environmental annotation scheme to any type of sequence, protein coding or ri-
bosomal rRNA. The sequences can be derived from a particular taxonomic grouping but they can also
84 correspond to operational taxonomic units (OTUs) used as proxies for taxa in environmental sequencing
studies [10].

86 In either case, the nature and diversity of environments associated with a particular microorganism can
elucidate and bring light to its ecology. Additionally, if OTUs are used, `seqenv` can also incorporate their
88 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 in the sample.
90  The scientist can then use these tables as a basis for multivariate statistics that contrast communities in
terms of the environmental terms associated with their constituent organisms. This novel approach is a
92 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 articles
[11]. Our approach is distinguished from this in two ways. Firstly, we start from sequences rather than
96 records, allowing us examine the distribution of environmental contexts within a certain level of sequence
similarity, secondly we associate to EnvO terms rather than geographic coordinates. This makes `seqenv`
98 more relevant to exploring the ecology of microbes, determining the distribution of OTUs across envi-
ronment types, as opposed to tracking viral outbreaks which was the focus in [11]. The information that
100 `seqenv` automatically generates can answer similar questions to those addressed in [12], where they
examined co-occurrence of OTUs across sampling sites, and classified isolation sources to EnvO terms
102 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.

112 **Materials and Methods**

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 input FASTA file. All the sequence
120 names 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
128 “nt” (nucleotide) database provided by NCBI is used and the BLAST algorithm is chosen to carry

o  the similarity search [7]. This step is the most costly computationally. It can, however, be
130 automatically parallel  ed by `seqenv` on multi-core systems.

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.

6. Then, we collect the “isolation source” text entries associated to all of the GI numbers recorded in
138 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 [8] and placing results into an SQL  database that is automatically
downloaded on the first run of `seqenv`.

8. With all the computed information, we are now able to describe each input sequence by a set of
150 EnvO terms and their associated frequency forming a term-frequency vector. Across the whole
dataset, this forms a sequence-term matrix. This matrix \mathbf{S} has elements $s_{j,k}$ given the weight of the
152 k th EnvO term associated with the j th sequence. These weights are calculated according to three

- 154 different normalization strategies. The first is named “flat” and consist of using the raw occurrence
counts. The second is termed “unique isolation” and will count every identical isolation source
156 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 of the
158 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, the rows
160 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”.
- 162 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
164 of the term vectors over all sequences, weighted by the abundance of that sequence in the sample.
Equivalently, the sample term matrix \mathbf{N} elements $n_{i,k}$, is the matrix product of the frequency matrix
166 \mathbf{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,
168 translated to english: “Sample Z is 25% brackish estuary, 25% river and 50% wetland.”
- 170 10. Other options are available to the user to further modify and filter the results. The “backtracking”
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,
172 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
174 through the given node in the ontology graph are removed.
- 176 11. The first output that is produced is a table serialized in the format of a tab-delimited plain text file
(TSV) representing the composition of each input sequence according to the EnvO terms associ-
ated to them, *i.e.* the matrix \mathbf{S} . The columns represents input sequence and rows represent the
178 normalized weight of EnvO terms.
12. If the user provided an frequency matrix (as described in step 2), the program can produce a simi-

180 lar TSV table representing the composition of each biological sample according to the EnvO terms
associated to them, *i.e.* the matrix \mathbf{N} . In this case, columns represents samples and rows represent
182 EnvO terms. Each value corresponds to the normalized weight of the EnvO term in the correspond-
ing sample.

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

188 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.


190 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:
192 <https://github.com/xapple/seqenv>. It can be installed on any computer with python by
simply typing: “`pip install seqenv`” in your shell.


194 Results

Earlier versions of the `seqenv` pipeline have already been used in a number of published studies includ-
196 ing 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
198 genes in lake sediments [15].

Here, to further illustrate its utility, we will apply it to two published datasets and demonstrate that
200 it provides additional insights into the processes that structure microbial communities not evident in the
original analyses. These two examples comprise:

202 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-

204 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
206 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
208  [?], 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)
210 as described in [4].

2. The Black Sea  me. This study included 454 pyrosequencing of 18S rRNA gene amplicons
212 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
214 described in [5]. Reads were denoised with AmpliconNoise and OTUs constructed at 3% sequence
divergence using average linkage hierarchical clustering as species proxies [16]. A total of 1'748
216 OTUs were obtained.

Patterns of ammonia oxidizing archaea (AOA) habitat usage

218 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
220 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 ,
222 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
224 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
226 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 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 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 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 positive association is observed between sample pH and total OTU diversity (adjusted R-squared: 0.131, p-value: 0.00922).

To determine which EnvO terms were most associated with the pH preference of the AOA OTUs, we performed a Random Forest regression of pH preference against the weighted EnvO terms. Random Forest uses an ensemble of decision trees constructed from artificial data sets generated by bootstrap aggregation or <bagging>, *i.e.* sampling with replacement across samples. This is combined with random selection of features. Since not all samples are used in each data set, a robust estimate of model accuracy is possible using the left out samples. Additionally, estimates of variable importance can be obtained by 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 Forest using the `randomForest` R package. The model explained 34.55% of the variation in pH preference. In figure 4, we visualize the weights of the top ten most important terms as determined by *%IncMSE* across OTUs ordered by their pH preference.

Environmental stages of the Black Sea paleome

The 48 sediment samples form a series from a Black Sea core spanning the last 11'400 years. In Coolen *et al.* [5], they defined four “Environmental Stages” (ES) in the geological evolution of the Black Sea that apply to this depth series on the basis of fossil evidence and isotope ratios:

- ES4: Lacustrine interval (~9.0 thousand years (ky) before present B.P.). During this lacustrine phase the Black Sea was disconnected 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
254 11'700 years ago, resulted in the connection of the Black Sea to the Mediterranean.

- 256 • ES3: A period of increasing salinity (~9.0-5.2 ky B.P.) corresponding to the warm and moist mid-Holocene climatic optimum.
- 258 • ES2: Establishment of modern environmental conditions (~5.2-2.5 ky B.P.) and further increasing salinity associated with the onset of the dry Subboreal.
- 260 • ES1: Freshening (~2.5 ky B.P.-present) with onset of the cool and wet Subatlantic climate and recent anthropogenic perturbations.

In figure 5 we visualise the community structures of these samples, in terms of the 18S rRNA OTU
262 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
264 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
266 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
268 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-
270 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,
272 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
274 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
276 indicated.

Discussion

278 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
280 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
282 diversity of habitats. As in the original study, a statistically significant relationship between sample OTU
richness and pH was evidenced. That is, as the pH of a sample increases, more species are observed.
284 We propose that these two observations may be connected: the fact that more environments appear ac-
cessible to the OTUs as the pH increases may generate the diversification of species that is reflected in
286 the 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.

288 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
290 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
292 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
294 to the IMI comprised organisms associated with freshwater habitats, important evidence that the IMI was
associated with a substantial increase in salinity.

296 Conclusion

The two studies described in this paper are not intended to be exhaustive, but present convincing vignettes
298 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
300 data. Hopefully, in the future, `seqenv` will contribute crucial insights and advances to the field of
environmental metagenomics.

302 **Acknowledgements**

Seqenv was originally conceived in a series of <hackathons> supported by the European Union's Earth
304 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
306 tools" and can be viewed at http://www.cost.eu/domains_actions/essem/Actions/ES1103.

308 We would like to thank the LifeWatchGreece project (<http://www.lifewatchgreece.eu/>) for covering
the participant's coffee breaks.

310 Authors LS and AE were funded by the Swedish Foundation for strategic research (ICA10-0015).
Author UI was funded by NERC IRF (NE/L011956/1). Author LJ was funded by the Novo Nordisk Foun-
312 dation (NNF14CC0001). Author EP was supported by the European Commission FP7-REGPOT project
MARBIGEN (grant agreement #264089) and the LifeWatchGreece Research Infrastructure (384676-
314 94/GSRT/NSRF C&E). CQ is funded through the MRC Cloud Infrastructure for Microbial Bioinformat-
ics (CLIMB) project (MR/L015080/1) through fellowship (MR/M50161X/1). Author CG was funded by
316 the Environment Research Council Fellowship (NE/J019151/1).

Author contributions

318 **LS** : Wrote the software product `seqenv` in python in its entirety. Contributed to writing this manuscript.

CQ : Participated in the development of the original idea. Wrote the majority of this manuscript and
320 performed analyses. Tested the software.

UI : Participated in the development of the original idea. Wrote the bash-based original version of
322 `seqenv` until version 0.8.0, after which, LS took over the implementation. Tested the software.

EP : Participated in the development of the original idea, tested software and organized hackathons.
324 Contributed to developing the NER software.

LJ : Developed the NER software that `seqenv` relies on, as well as helped with using and installing
326 it.

AC : Participated in the first hackathon, commented and suggested revisions to manuscript.

328 **AO** : Helped test the software.

CP : Participated in the hackathons, helped test the software, commented and suggested revisions to
330 manuscript.

JS : Participated in the first hackathon, commented and suggested revisions to manuscript.

332 **AW** : Participated in the second hackathon.

AI : Participated in the second hackathon.

334 **MC** : Commented and suggested revisions to manuscript. Supplied the Black Sea dataset.

AE : Commented and suggested revisions to manuscript.

336 **CG** : Provided AOA data and expertise, commented and suggested revisions to manuscript.

References

- 338 [1] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole (2007) **Naive Bayesian clas-**
340 **sifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.** *Applied and*
Environmental Microbiology, volume 73 (16), 5261
- [2] Agnieszka S Juncker, Lars J Jensen, Andrea Pierleoni, Andreas Bernsel, Michael L Tress, Peer Bork
342 et al. (2009) **Sequence-based feature prediction and annotation of proteins.** *Genome biology*,
volume 10 (206)
- 344 [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-**
346 **ical entities.** *Journal of Biomedical Semantics*, volume 4 (1), 43
- [4] Cécile Gubry-Rangin, Brigitte Hai, Christopher Quince, Marion Engel, Bruce C Thomson, Phillip
348 James et al. (2011) **Niche specialization of terrestrial archaeal ammonia oxidizers.** *Proceedings*
of the National Academy of Sciences of the United States of America, volume 108 (52), 21206
- 350 [5] Marco J L Coolen, William D Orsi, Chere Balkema, Christopher Quince, Keith Harris, Sean P
Sylva et al. (2013) **Evolution of the plankton paleome in the Black Sea from the Deglacial to**
352 **Anthropocene.** *Proceedings of the National Academy of Sciences of the United States of America*,
volume 110 (21), 8609
- 354 [6] Ramiro Logares, Thomas H A Haverkamp, Surendra Kumar, Anders Lanzén, Alexander J Neder-
bragt, Christopher Quince et al. (2012) **Environmental microbiology through the lens of high-**
356 **throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches.**
Journal of Microbiological Methods, volume 91 (1), 106
- 358 [7] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman (1990) **Basic local alignment search**
tool. *Journal of molecular biology*, volume 215 (3), 403

- 360 [8] Evangelos Pafilis, Sune P Frankild, Julia Schnetzer, Lucia Fanini, Sarah Faulwetter, Christina
Pavloudi et al. (2015) **ENVIRONMENTS and EOL: identification of Environment Ontology**
362 **terms in text and the annotation of the Encyclopedia of Life**. *Bioinformatics*, volume 31 (11),
btv045
- 364 [9] Dawn Field, Peter Sterk, Renzo Kottmann, J Wim De Smet, Linda Amaral-Zettler, Guy Cochrane
et al. (2014) **Genomic standards consortium projects**. *Standards in Genomic Sciences*, vol-
366 *ume 9* (3), 599
- [10] Mark Blaxter, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd et al. (2005)
368 **Defining operational taxonomic units using DNA barcode data**. *Philosophical transactions of*
the Royal Society of London. Series B, Biological sciences, volume 360 (1462), 1935
- 370 [11] Tasnia Tahsin, Davy Weissenbacher, Robert Rivera, Rachel Beard, Mari Firago, Garrick Wallstrom
et al. (2016) **A high-precision rule-based extraction system for expanding geospatial metadata**
372 **in GenBank records**. *Journal of the American Medical Informatics Association*, ocv172
- [12] S Chaffron, H Rehrauer, J Pernthaler, and C von Mering (2010) **A global network of coexisting mi-**
374 **crobes from environmental and whole-genome sequence data**. *Genome Research*, volume 20 (7),
947
- 376 [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),
378 2423
- [14] Domenico Savio, Lucas Sinclair, Umer Z Ijaz, Juraj Parajka, Georg H Reischer, Philipp Stadler
380 et al. (2015) **Bacterial diversity along a 2600 km river continuum**. *Environmental Microbiology*,
volume 17 (12), 4994
- 382 [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**

- 384 **O₂-Tolerant NiFe Membrane-Bound Hydrogenases in Lacustrine Sediment.** *Current microbiology*, volume 71 (2), 296
- 386 [16] Christopher Quince, Anders Lanzén, Russell J Davenport, and Peter J Turnbaugh (2011) **Removing noise from pyrosequenced amplicons.** *BMC Bioinformatics*, volume 12 (1), 38

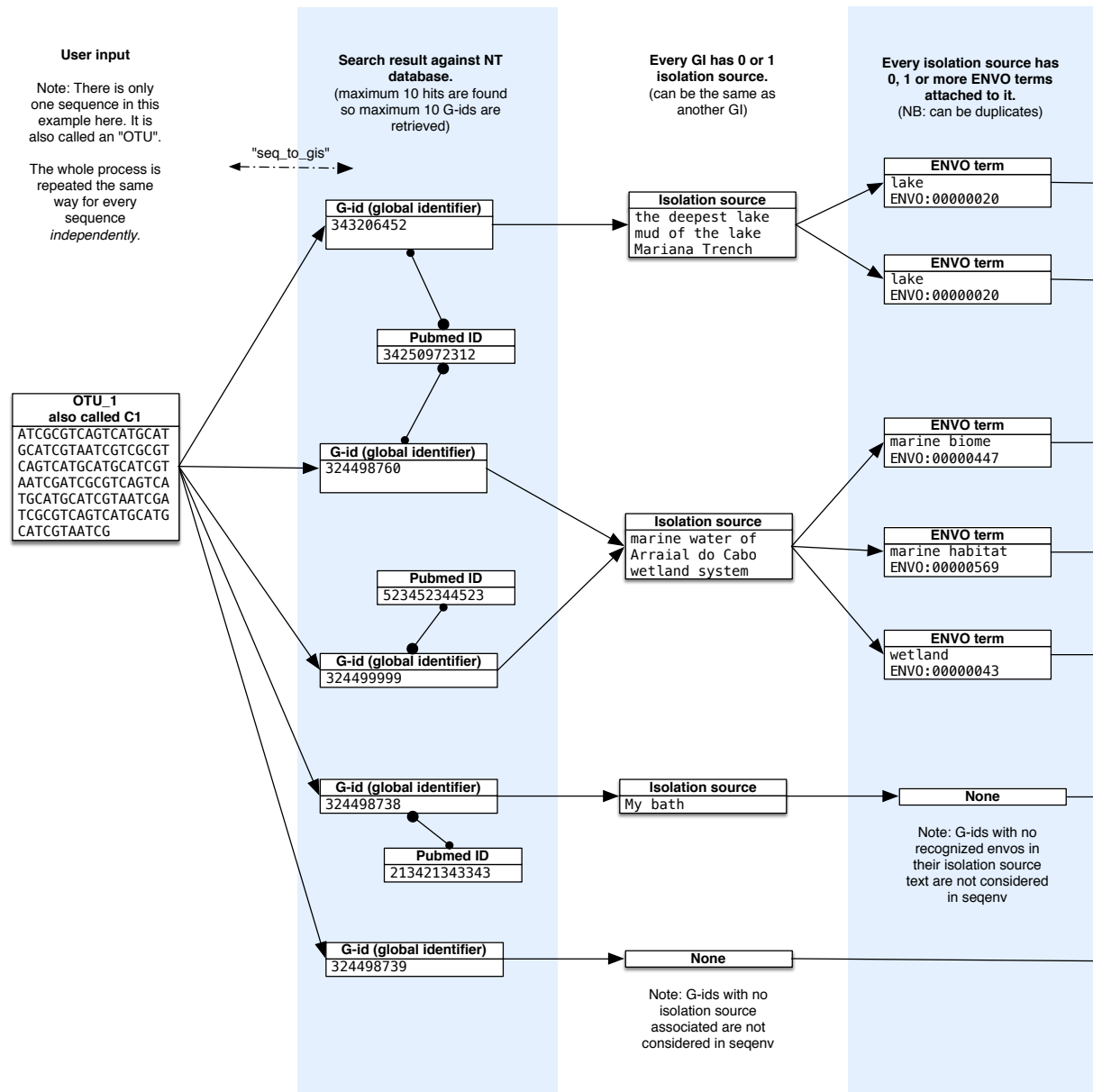
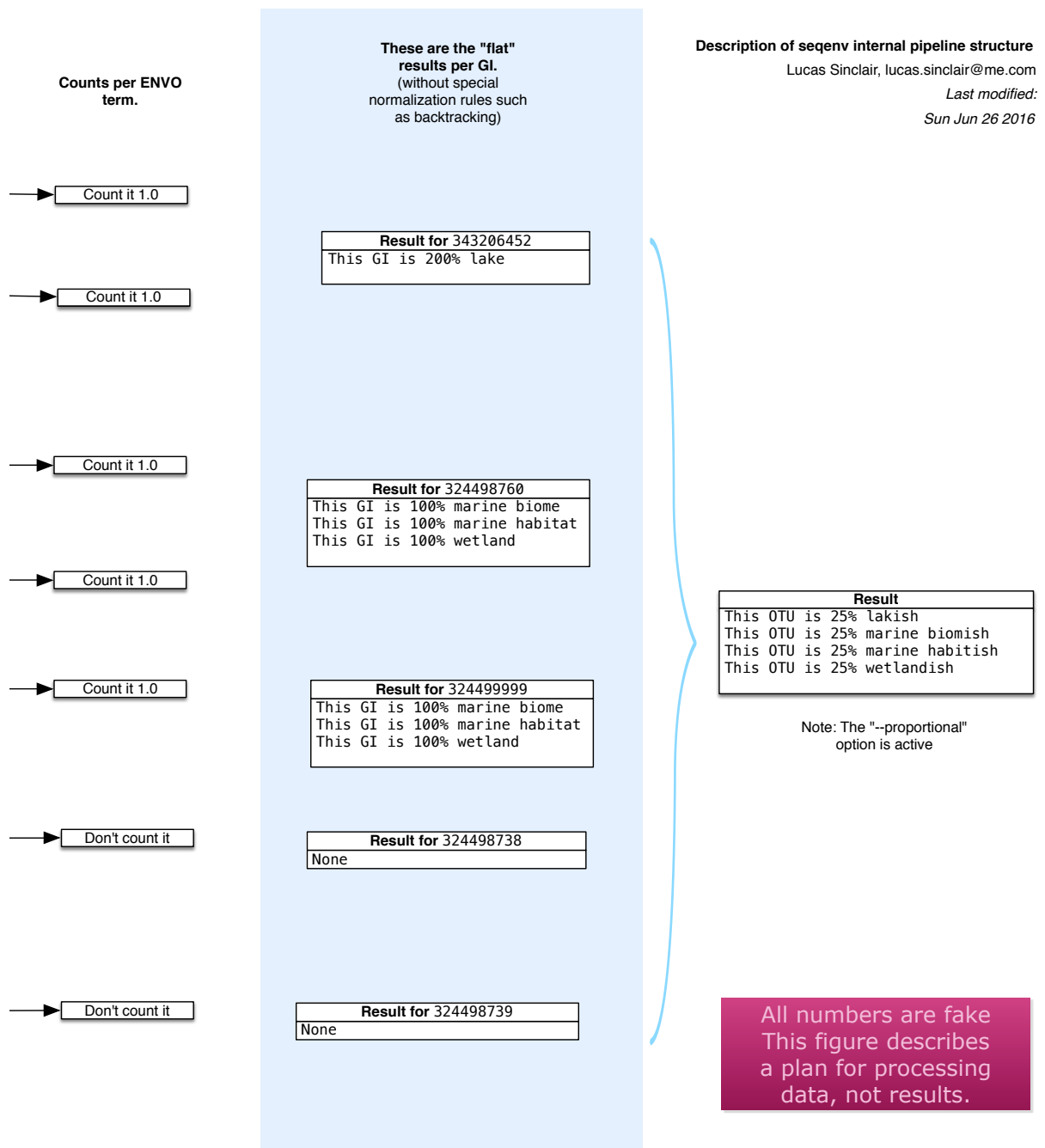
388 **Figures**

Figure 1: Schematic of the internal functioning of the seqenv pipeline.

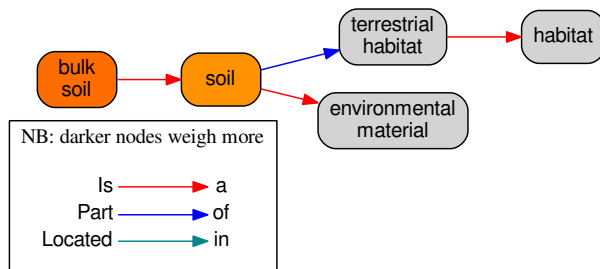
Continued on next page.



390

This figure details how EnvO term frequencies are computed. The numbers provided are fictional
 392 as the schematic focuses on representing the internal functioning of the pipeline and does not illustrate
 a concrete case. As each inputted short DNA sequence is processed independently in the all but the last
 394 stages of *seqenv*, only one input sequence is shown here.

OTU - C46 (pH 3.5):



OTU - C66 (pH 8.5):

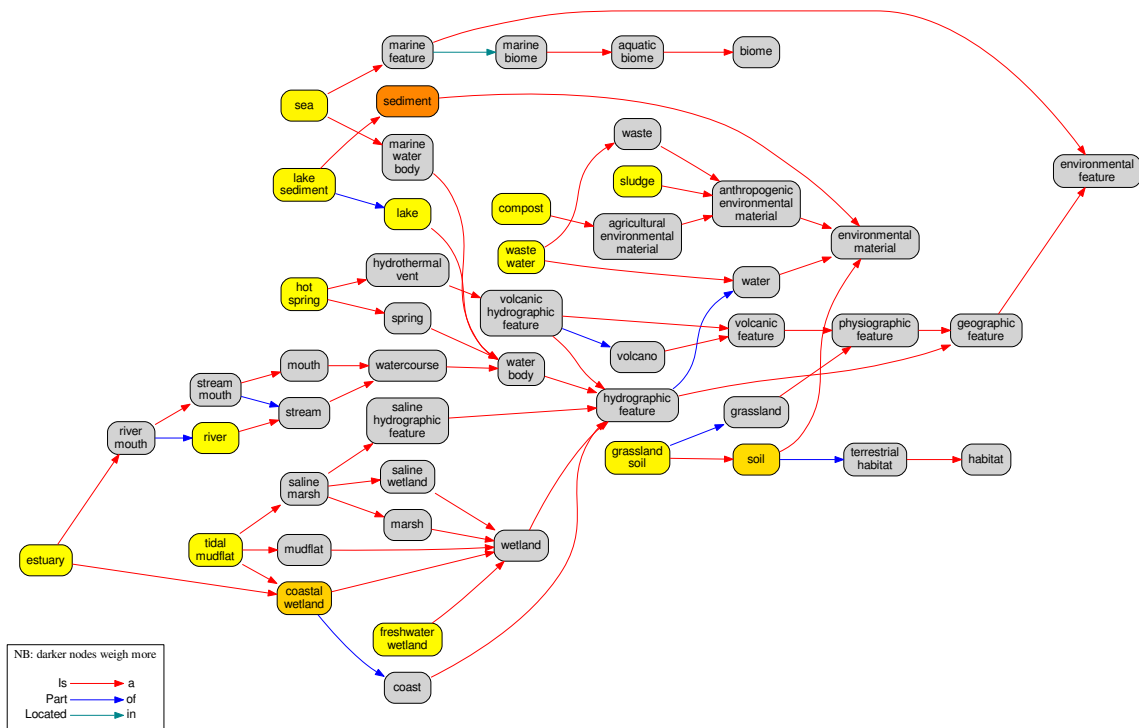


Figure 2: The EnvO terms associated with two AOA OTUs.

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

398 hits. Gray indicates the lowest frequency recorded and darker shades of yellow to orange indicate higher frequencies.

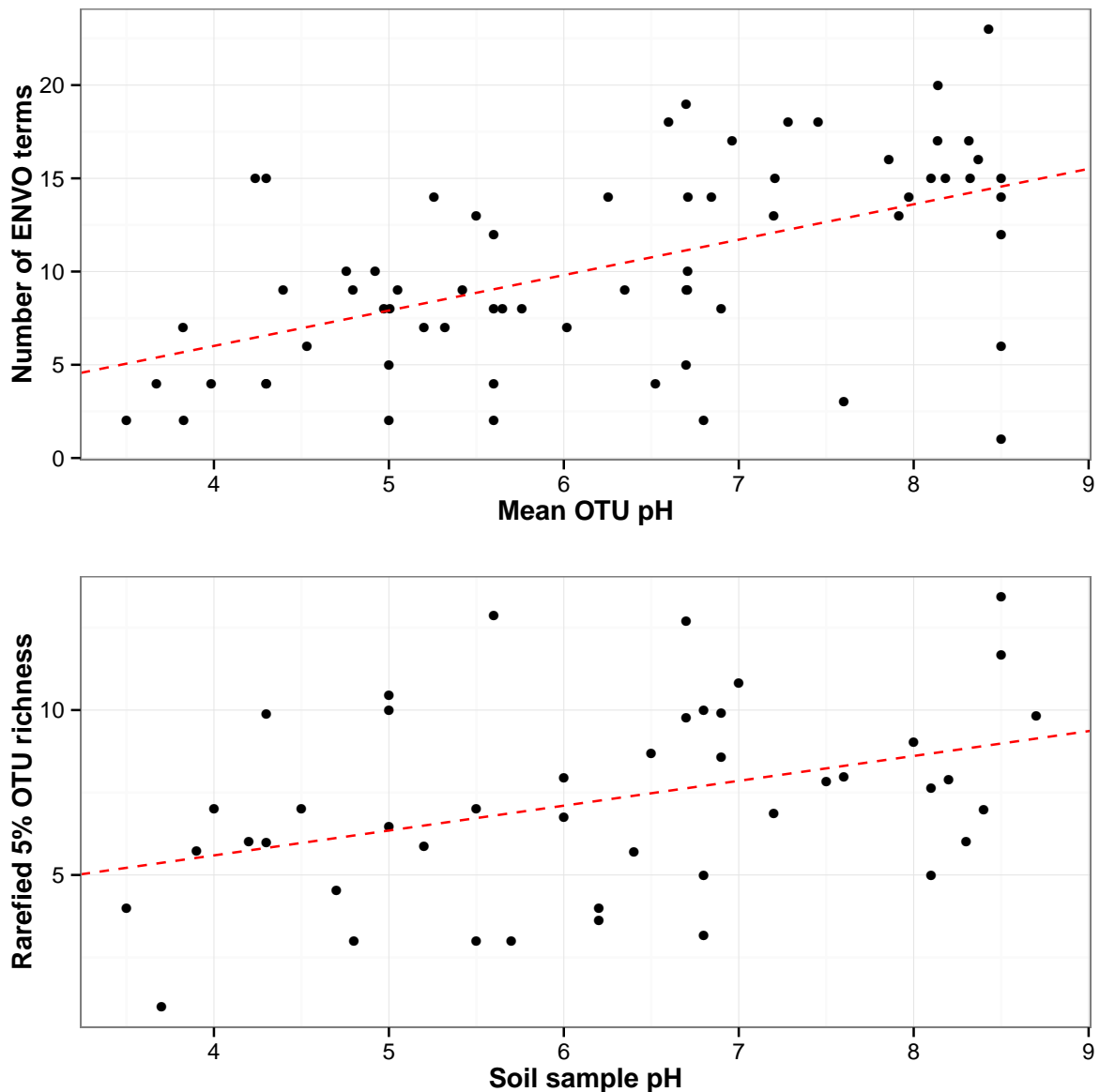


Figure 3: ENVO terms and OTU richness against mean OTU pH.

400 The top panel 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-
402 06). The bottom panel 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

404 (adjusted R-squared: 0.1305, p-value: 0.009217).

Random forests were used to perform a regression of pH against EnvO terms (Var. explained: 34.6%).

406 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.

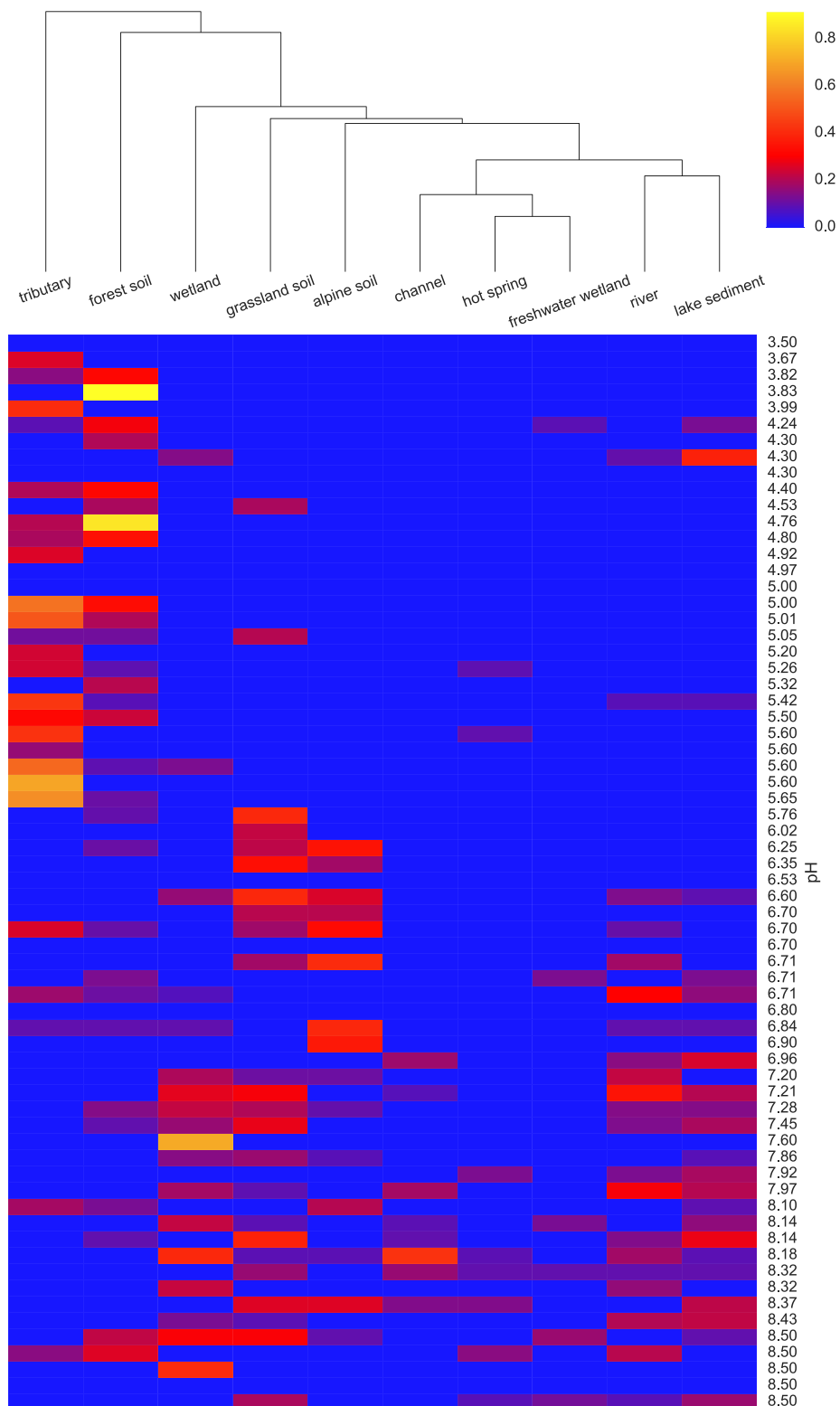


Figure 4: Heatmap of top ten EnvO terms for determining OTU pH.

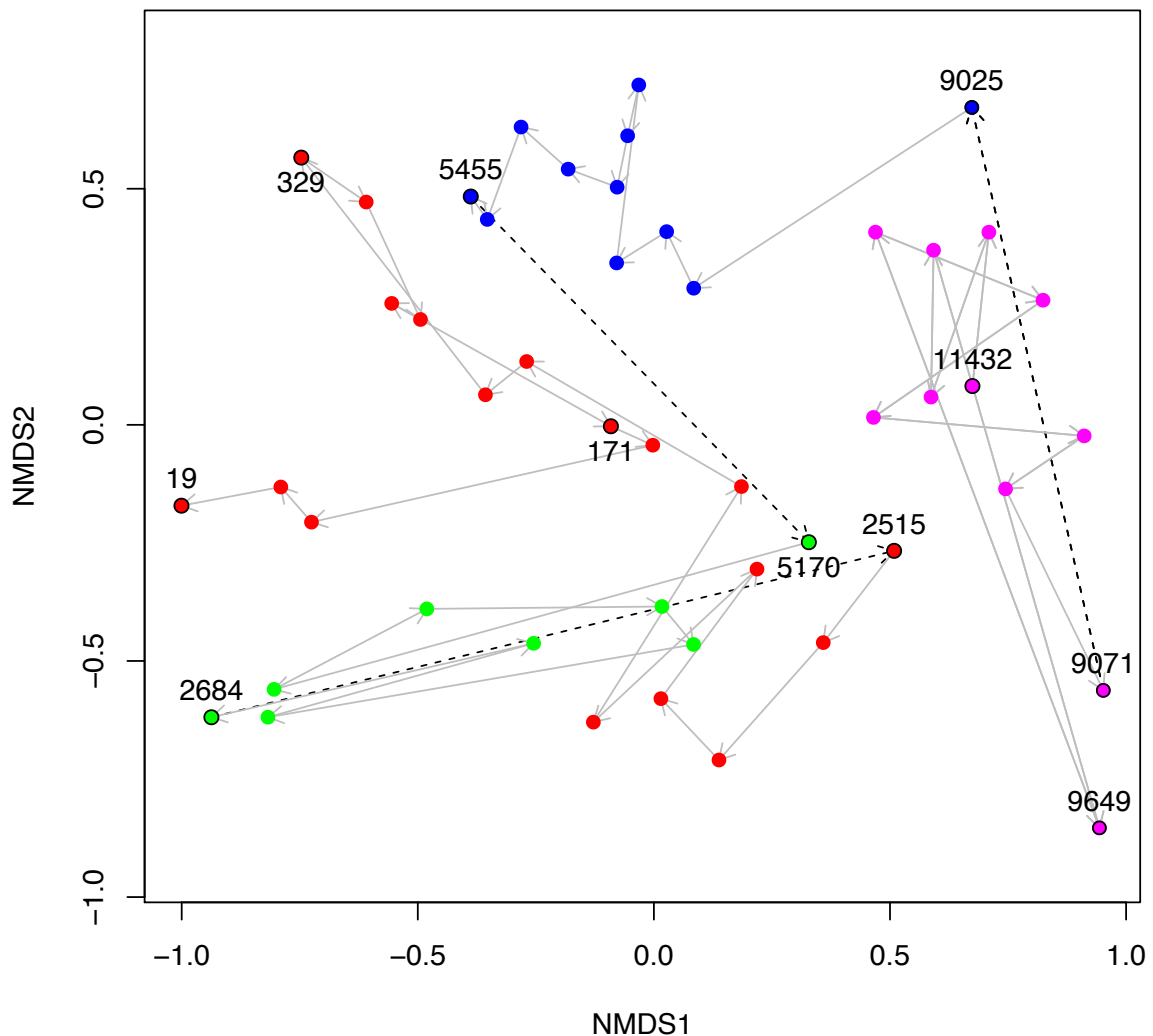


Figure 5: NMDS plot of Black Sea plankton 18S rRNA samples.

408 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
410 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

412 the transition between environmental stages.

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

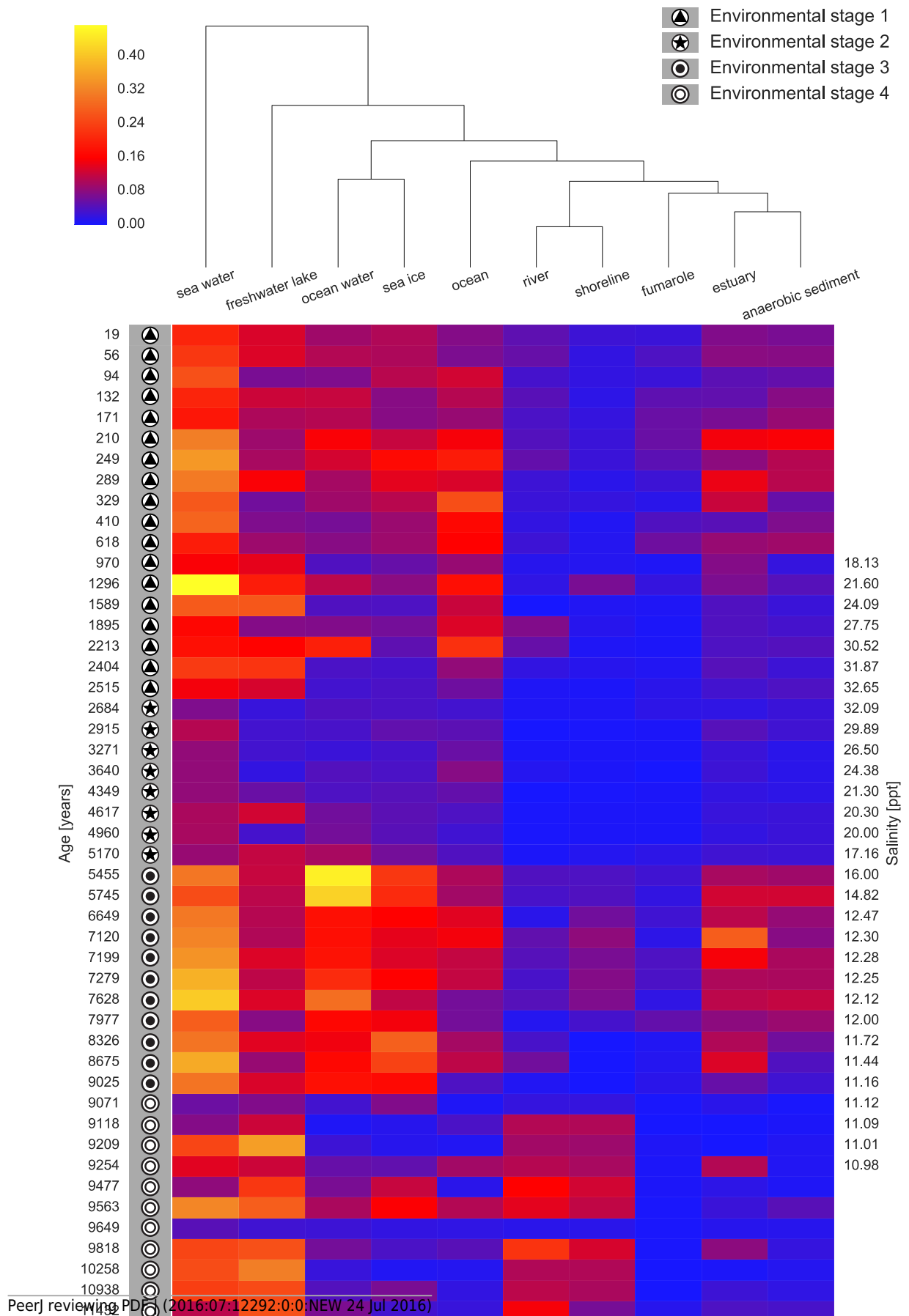


Figure 6: Heatmap of top ten EnvO terms for determining ES.