

1 **Antarctic biodiversity predictions through substrate qualities and** 2 **environmental DNA**

3 **Authors:**

4 Paul Czechowski^{1*}, Michel de Lange^{2,3}, Micheal Knapp¹, Aleks Terauds⁴, and Mark I.
5 Stevens^{5,6}

6 **Affiliations:**

7 ¹ University of Otago, Department of Anatomy, 270 Great King Street, Dunedin, OTA 9016,
8 New Zealand.

9 ² University of Otago, Department of Biostatistics, 18 Frederick Street, Dunedin, OTA 9016,
10 New Zealand.

11 ³ Pacific Edge Ltd, Centre for Innovation, 87 St David Street, Dunedin, OTA 9016, New
12 Zealand.

13 ⁴ Australian Antarctic Division, Department of Agriculture, Water and the Environment, 203
14 Channel Highway, Kingston, TAS 7050, Australia.

15 ⁵ South Australian Museum, 61-68 North Terrace, Adelaide, SA 5000, Australia.

16 ⁶ School of Biological Sciences, The University of Adelaide, SA 5005, Australia.

17 *Correspondence to Paul Czechowski, paul.czechowski@otago.ac.nz

18 **Open Research Statement:**

19 Parts of the data are already published, with those publications cited in this article. All data
20 were provided as in-confidence for peer review and have been revised during peer review to
21 accompany this article. All versions are available via <https://doi.org/10.5281/zenodo.4579841>
22 and github.com/OldMortality/eukaryotes.

23 **Abstract:**

24 Antarctic conservation science is important to enhance Antarctic policy and to understand
25 alterations of terrestrial Antarctic biodiversity. Antarctic conservation will have limited long-
26 term effect in the absence of large-scale biodiversity data, but if such data were available, it is
27 likely to improve environmental protection regimes. To enable Antarctic biodiversity
28 prediction across continental spatial scales through proxy variables, in the absence of baseline
29 surveys, we link Antarctic substrate-derived environmental DNA (eDNA) sequence data

30 from the remote Antarctic Prince Charles Mountains to a selected range of concomitantly
31 collected measurements of substrate properties. We achieve this using a statistical method
32 commonly used in machine learning. We find neutral substrate pH, low conductivity, and
33 some substrate minerals to be important predictors of presence for basidiomycetes,
34 chlorophytes, ciliophorans, nematodes, or tardigrades. Our bootstrapped regression reveals
35 how variations of the identified substrate parameters influence probabilities of detecting
36 eukaryote phyla across vast and remote areas of Antarctica. We believe that our work may
37 improve future taxon distribution modelling and aid targeting logistically challenging
38 biodiversity surveys.

39 **Introduction:**

40 Although only 0.3% of continental Antarctica is ice-free, many organisms including bacteria,
41 unicellular eukaryotes, fungi, lichen, cryptogamic plants and invertebrates are scattered
42 across the continent in extremely isolated, remote, island-like terrestrial habitats, for example
43 in soil-like substrates, lakes, and cryoconite holes (Convey *et al.* 2014; Chown *et al.* 2015).
44 Threats to this Antarctic biodiversity are posed by human activity, climate change, pollution,
45 and invasive species. It is becoming increasingly clear that mitigation of these threats and
46 further alterations to the Antarctic biosphere rely on well-tailored management strategies
47 across the continent's bioregions (eg Coetzee *et al.* 2017).

48 Effective continental-scale conservation management requires continental-scale data (eg
49 Wauchope *et al.* 2019). However, knowledge of terrestrial Antarctic biodiversity is still
50 limited because most of Antarctica's ice-free areas remain unstudied due to logistic
51 difficulties exacerbated by the harsh environmental conditions, and funding constraints.
52 Environmental DNA (eDNA) analysis, despite shortcomings, is arguably one of the most
53 practical and economical options for continental-wide baseline surveys of terrestrial Antarctic
54 biodiversity, especially when facing logistical challenges typical for work on the Antarctic
55 continent (reviewed in Czechowski *et al.* 2017). Comparable large-scale systematic
56 approaches to protect soil diversity are recognized as required globally, but often are limited
57 to charismatic groups such as those found in the Arctic (Gillespie *et al.* 2020).

58 Here, we link commonly measured substrate properties to the cryptic eukaryotic biodiversity
59 of terrestrial Antarctic ice-free regions. Soil nutrient status is the most important attribute of
60 biodiverse soils (Geisen *et al.* 2019), and corresponding key variables can be, and are,
61 routinely measured economically. We analyzed molecular data (eDNA) from an extremely

62 remote Antarctic terrestrial region to clarify relationships between substrate properties and
63 eukaryote phylum presence. We envisage our approach to be useful in predicting biodiversity
64 across a wide taxonomic spectrum across large areas of Antarctica, especially to identify
65 regions worthy of lower-level taxonomic biodiversity surveys, then possibly realized with
66 “barcoding” using mitochondrial DNA (such as with the Cytochrome Oxidase 1) or
67 logistically more challenging morphological biodiversity assessments.

68 The Prince Charles Mountains (PCMs), the most remote terrestrial areas in eastern
69 Antarctica, were first sighted by US Operation Highjump (1946/47) and mapped in more
70 detail by Australian (1954–1961) and Russian (1983–1991) expeditioners. In 2011 we
71 obtained environmental DNA samples from substrates throughout the PCMs and measured
72 various geochemical and mineral properties. Previously, Czechowski *et al.* (2016b) focused
73 on invertebrates as the primary substrate-inhabiting metazoans and discovered major changes
74 in their distribution over salinity gradients, as known from other areas and taxa of Antarctica
75 (eg Bottos *et al.* 2020). Here, we expand our analyses of environmental variables using a
76 predictive approach to the full spectrum of eukaryote phyla, and thereby explore approaches
77 of inferring biodiversity presence that could be applied across the entirety of ice-free
78 terrestrial Antarctica. Beyond phylum-level surveys, our technique may be applied using
79 other genetic markers and predictors to link future smaller-scaled conservation projects
80 anywhere in terrestrial Antarctica, aid taxon distribution modelling, and thus contributes
81 towards improving conservation management strategies across the Antarctic bioregions.

82 **Methods:**

83 Fieldwork took place in the Prince Charles Mountains (PCMs; East Antarctica, Figure 1)
84 from 26 November 2011 to 21 January 2012 close to Mount Menzies (MM; 73°25'29.38"S,
85 62°0'37.61"E), Mawson Escarpment (ME; 73°19'16.91"S, 68°19'31.20"E) and Lake
86 Terrasovoje (LT; 70°32'23.58"S, 67°57'28.05"E) as described earlier (Czechowski *et al.*
87 2016a, b). 154 field samples were considered for this study (26 MM, 70 ME, 58 LT; Web
88 Table 1).

89 To infer climatic conditions in the PCMs, we used rasters from Quantarctica 3 (Matsuoka *et al.*
90 2021) encoding annual mean precipitation (mm), wind speed (m s^{-1} 10m above ground)
91 and mean annual temperature ($^{\circ}\text{C}$ 2 m above ground, as only temperature data distributed via
92 Quantarctica). We disaggregated the layer rasterization from 35 km px^{-1} to 1 km px^{-1} through

93 bilinear interpolation. We then extracted median values for the three variables from a 20 km
94 buffer surrounding each sampling location (Web Figure 1).

95 As predictor data for eukaryote phylum presence in substrates, geochemical composition
96 (NH_4^+ , C, ρ , NO_3^- , $\text{pH}_{\text{H}_2\text{O}}$, $\text{pH}_{\text{CaCl}_2}$, P, K, S, texture) was analyzed by agricultural soil testing
97 service APAL (www.apal.com.au). Many measurements below detection level needed to be
98 excluded to yield data completeness of at least 96.7% (Web Table 2). The final analysis
99 included K, S, ρ , and $\text{pH}_{\text{CaCl}_2}$ ($\text{pH}_{\text{H}_2\text{O}}$ excluded as co-linear, texture excluded as categorical).

100 As additional predictors, the substrate mineral composition was considered through
101 integration of X-ray diffraction spectra of the minerals quartz, calcite, feldspar, titanite,
102 pyroxene / amphibole / garnet, micas, dolomite and kaolin / chlorite, and chlorite (see
103 Czechowski *et al.* 2016b). We handled the sum-to-unity constraint of our mineral
104 compositions by excluding quartz as the most common mineral from further analysis. As
105 further predictors for most locations (MM: n=26, ME: n=69, LT: n=57), we included hitherto
106 unpublished measurements of soil-substrate ATP (eg Conklin and Macgregor 1972), obtained
107 with a Clean-Trace Luminometer (3M, Maplewood, US-MN), and slope measurements. Prior
108 to regression, all predictors were standardized to mean of 0 and unit variance. Predictor
109 densities are provided in Web Figure 2.

110 Biological response data were prepared in QIIME 2020-2 (Bolyen *et al.* 2019) and R 4.0.0 (R
111 Core Development Team 2019) from raw sequence data generated as described elsewhere
112 (Czechowski *et al.* 2016b, 2017). In summary, 125 bp eukaryotic 18S rDNA PCR products
113 (yielding an 85 bp target region) had been amplified using primers ‘Euk1391f’ and ‘EukBr’
114 (Caporaso *et al.* 2012), as established for eukaryotic microbial surveying (Thompson *et al.*
115 2017). As recommended, PCRs had been carried out in triplicates, each replicate carrying
116 identical barcodes. The resulting eDNA libraries had been combined for sequencing across
117 two MiSeq runs (Web Figure 3). We re-defined Amplicon Sequence Variants (ASVs; *sensu*
118 Callahan *et al.* 2017) from those data with Qiime: after pre-filtering (Phred score ≥ 25), we
119 trimmed read pairs with Cutadapt v1.18 (Martin 2011), and denoised using DADA2 (v1.6.0;
120 Callahan *et al.* 2016). We retained merged reads with an expected error value less than 3, that
121 we not deemed chimeric.

122 Due to the shortness and slow evolution of the employed 18S marker, we set out to conduct
123 our analyses on the phylum level, and to use species level assignments solely to verify data
124 credibility. Accordingly, we designed the retrieval of taxonomic annotations for our Antarctic
125 DNA sequences in such fashion so as to yield reliable species identifications in cases where

126 Antarctic reference data were available, while still retuning higher taxonomic (eg phylum
127 level) identifications in cases where closely matching reference data were not available.
128 Doing so, we were able include a larger amount of Antarctic sequences into our statistical
129 analysis on phylum level, but needed to consider species level identifications as potentially
130 unreliable, and verify them on alignment level. We identified eukaryotic sequences among
131 our reads with a recent local copy (April 2020) of the entire NCBI nucleotide collection in
132 conjunction with Blast 2.10.0+. Taxonomic assignments were retrieved from reference
133 sequences *at least* 50% identical to queries, with an assignment significance threshold (*e*
134 value) of 10^{-10} , considering only matches with at least 90% coverage, and excluding
135 environmental sequences (*evaluate* $1e^{-10}$, *max_hsps* 5, *max_target_seqs* 5, *qcov_hsp_perc* 90
136 and *perc_identity* 50). For each Antarctic sequence search query, we used the highest Bit
137 score among all returned sequences from the NCBI database for that query to choose the final
138 taxonomic assignment. Subsequently, we used R package *decontam* (Davis *et al.* 2018) to
139 remove putatively contaminating reads, and likewise subtracted all sequences and taxa in
140 negative controls from field samples. Focusing on eukaryotes, we discarded all non-
141 eukaryote reads (Web Figure 4).

142 With the Lasso (Tibshirani 1996) of R package *glmnet* (Friedman *et al.* 2010) we regressed
143 each phylum present in at least 12 samples against the aforementioned predictors (Web
144 Figure 5). In regressions, we disregarded sequence read abundances as meaningless due to
145 inherent constraints of amplicon sequencing (eg Czechowski *et al.* 2017), analyzed presences
146 instead, and used the most biodiverse of all locations (LT; Czechowski *et al.* 2016b; also
147 Figure 2) as a reference location, so that we report predictor effects at MM and ME as
148 relative to LT. We initially retrieved the active set (variables not set to 0) estimated by Lasso,
149 repeated the regression of phylum presence against 1,000 randomly chosen sample-sets of
150 predictors, calculated the number of times each variable was estimated to be non-zero, and
151 report variables non-zero more than 950 times as significant. Accordingly, we calculated
152 95% non-parametric bootstrap confidence intervals for our estimates. We did not adjust for
153 multiple comparisons.

154 Furthermore, we explored the global distribution of the obtained putative species level
155 assignments among phyla significantly influenced by environmental predictors (see below)
156 by querying BISON (bison.usgs.gov), GBIF (www.gbif.org) and iNaturalist
157 (www.inaturalist.org; see Web Text 1 for detailed methods) with R package *spocc*.

158 **Results:**

159 Keeping in mind the coarse raster resolution and model-like character of the climate data,
160 annual mean climate at MM was coldest (-32 ± 0.3 °C), windiest (10.2 ± 0.05 ms⁻¹) and with
161 an intermediate amount of precipitation (86 ± 1 mm), when compared to the other two
162 locations (Web Figure 1). ME exhibited the least amount of precipitation (55.3 ± 7 mm),
163 comparatively low wind speeds (5.4 ± 0.5 ms⁻¹), and slightly higher temperatures than MM ($-$
164 28.4 ± 0.6 mm). Closest to the coast, and exposed, LT appeared influenced by the highest
165 precipitation (136 ± 16 mm), variable but moderate wind speeds (5.5 ± 1.7 ms⁻¹) and the
166 highest temperature in the surveyed area (-24.1 ± 1.6 °C). We found our chosen climatic
167 variables strongly correlated with the sampling locations, and to improve predictive power
168 excluded the former from further considerations. Instead, we interpreted the statistical effect
169 of location (below) to be a function of annual mean climatic variables.

170 Retention of eukaryotes in field-derived samples after filtering yielded 2,285,773 reads across
171 145 samples, derived from 16,524,031 unfiltered sequences (Web Table 3). Per-sample mean
172 coverage was 9,450 reads (min: 2, median: 2,379, max: 86,804). ASV mean coverage after
173 filtering was 2,984 reads (min: 2, median: 132, max: 207,718; Web Figure 6). Collectively
174 after filtering, 766 ASVs were assigned to 495 species across 25 phyla (Web Table 4). Most
175 prevalent phyla (and among those: most prevalent species) by coverage were Ascomycota
176 (*Acanthothesia fontana*), Chlorophyta (*Coccomyxa* sp.), Basidiomycota (*Mrakia frigida*),
177 Ciliophora (*Pseudochilodonopsis quadrivacuolata*), Nematoda (*Scottinema lindsayae*),
178 Rotifera (*Embata laticeps*), and Tardigrada (*Mesobiotus furciger*). (All taxonomic
179 assignments listed here aligned with reference data without gaps at full coverage, and a bit
180 score of 154.6, apart from bit score of 145.6 for *P. quadrivacuolata*)

181 We found the distribution of five phyla (26 classes, 59 orders, 100 families, 173 species)
182 across the PCMs to be significantly correlated with the considered soil predictors (Figure 2,
183 Web Tables 5 and 6, Web Figure 7). Those taxa were defined by 265 ASVs across 1,210,855
184 sequences and 142 samples (23 MM, 64 ME, 55 LT). Per-sample mean coverage was 9,460
185 (min: 2, med: 3863, max: 84,892), per-ASV mean coverage was 4,596, (min: 2, median: 157,
186 max: 128,358; Web Figure 6).

187 For each predictor significantly correlating with a phylum's presence (Web Figure 8) we
188 report the expected effect on phylum presence corresponding to one standard deviation (σ)
189 increase of the predictor from its mean (μ), with all other variables held at mean μ . Key
190 significant results included:

- 191 i) Low levels of Basidiomycota (62 putative species assignments, Figure 2a) in high pH
192 environments ($\mu = 7.15$, $\sigma = 0.88$, $E[\text{present } \mu] = 0.6$ and $E[\text{present } \mu + 1\sigma] = 0.4$), and a
193 strong positive relationship of this phylum with dolomite ($\mu = 0.025$ %, $\sigma = 0.05$ %,
194 $E[\text{present } \mu + 1\sigma] = 0.7$).
- 195
- 196 ii) Very low levels of Chlorophytes (47 species, Figure 2b) at MM plausibly attributable
197 to harsh environmental conditions encountered there (see Supplemental Materials;
198 $E[\text{present }_{LT}] = 0.61$ and $E[\text{present }_{MM}] = 0.32$, including more alkaline substrates
199 ($E[\text{present } \mu + 1\sigma] = 0.46$)
- 200
- 201 iii) Very low levels of Ciliophorans (47 species, Figure 2c) at MM ($E[\text{present }_{LT}] = 0.70$
202 and $E[\text{present }_{MM}] = 0.39$), in Sulphur-rich substrates ($\mu = 528$ mg kg⁻¹, $\sigma = 1410$ mg
203 kg⁻¹, $E[\text{present } \mu + 1\sigma] = 0.61$), and in areas relatively rich in pyroxene, amphibole or
204 garnet ($\mu = 4$ %, $\sigma = 4$ %, $E[\text{present } \mu + 1\sigma] = 0.52$)
- 205
- 206 iv) Very low levels of nematodes (8 species, Figure 2d) at MM ($E[\text{present }_{LT}] = 0.47$ and
207 $E[\text{present }_{MM}] = 0.28$), and in highly conductive substrates ($\mu = 0.55$ dSm⁻¹, $\sigma = 1.07$
208 dSm⁻¹, $E[\text{present } \mu + 1\sigma] = 0.35$)
- 209
- 210 v) Very low levels of tardigrades (9 species, Figure 2e) in alkaline substrates ($E[\text{present}$
211 $\mu] = 0.22$, $E[\text{present } \mu + 1\sigma] = 0.14$)

212 Observed fractions of non-zero coefficients are shown Table 1 and Web Figure 8. (95% non-
213 parametric bootstrap confidence intervals for non-0 estimates also provided in Web Figure 8.)
214 Directions of all predictor effects on all analyzed taxa presences, including insignificant
215 effects, are listed in Web Table 7.

216 For 66 of our 173 putative species assignments 778 georeferenced records could be obtained
217 (of those 65% from GBIF, 27% iNaturalist, 7% BISON). Of the obtained 123 locations 4%
218 were in Africa, 1.6% in Antarctica, 13% Asia, 32% Europe, 21% North America and 10% in
219 South America (Web Figures 9 and 10, Web Table 7). The sole species recorded for
220 Antarctica (here: south of 66.56°) was the nematode *Scottinema lindsayae*. Observations north
221 of the polar circle (likewise 66.56°) included Basidiomycota (*Gloiocephala aquatica*,
222 *Stereum rugosum*, *Mrakia frigida*, *Rhodotorula mucilaginosa*), Chlorophyta (*Haematococcus*
223 *lacustris*, *Oophila amblystomatis*), and Ciliophora (*Furgasonia blochmanni*, *Chilodonella*

224 *acuta* (Ciliophora), *Tachysoma pellationellum*). Refer to Web Tables 4 and 5 for alignment
225 qualities.

226 **Discussion:**

227 Our Antarctic case study demonstrates two key technologies to be useful for baseline
228 biodiversity surveys across large spatial scales in extremely remote environments – robust
229 predictive statistics, such as the Lasso, now often used in machine learning algorithms
230 (Muthukrishnan and Rohini 2016), as well as biodiversity information derived from
231 environmental DNA (Czechowski *et al.* 2017). To the best of our knowledge, our work is the
232 first in associating environmental DNA data to environmental predictors by means of the
233 Lasso to yield accurate detection probabilities for taxonomic groups, also in Antarctica. Thus,
234 we present an analytical framework to identify areas for targeted species-level biodiversity
235 surveys, using other markers, or predictors for Antarctica, and possible other hardly
236 accessible locations.

237 Our expanded analyses of the original raw data (Czechowski *et al.* 2016b) made use of new
238 algorithms for processing environmental DNA sequences (eg Callahan *et al.* 2016, 2017),
239 along with more extensive reference databases for taxonomic assignment, and new
240 algorithms available with R (R Core Development Team 2019). While our results are in line
241 with earlier findings relating eukaryote distribution to their environment in the PCMs and
242 Antarctica (eg Czechowski *et al.* 2016a, b; Bottos *et al.* 2020), our approach adds accuracy to
243 those findings with respect to five phyla.

244 A strength of our analyses is the relatively easy retrieval of biological survey data
245 encompassing many phyla (probably including many cryptic and unknown species) across
246 many samples. The weakness of the employed 18S marker is its limited ability to discern
247 many distinct sequence variants on a low taxonomic (eg species) level. Regardless,
248 identification of species with likely Antarctic occurrence such as the known Antarctic
249 nematode *Scottinema lindsayae* and tardigrade *Mesobiotus furciger* by means of a relatively
250 short and highly conserved primer pair highlights the ability of environmental DNA to
251 retrieve species occurrence records, provided that sufficient sequence data is available for
252 taxonomic assignment. Consequently, we believe that environmental DNA analysis should be
253 the method of choice to obtain biodiversity data from Antarctica, particularly when many
254 samples are to be analyzed, but other markers are needed to investigate fine scaled
255 endemism, and to obtain better taxonomic resolution.

256 Georeferencing our putative species assignments by means of publicly accessible databases
257 had limited success. The limitations of reference databases became obvious when known
258 Antarctic species, such as *Acutuncus antarcticus* (Web text 1), the latter identified among our
259 data through a perfect alignment with bit score 154.6, were not found, and only 38% of all
260 putative Antarctic species assigned by us were georeferenced at all. High occurrence
261 prevalence in North America, and Europe indicates sampling bias in GBIF, iNaturalist and
262 BISON and highlight a substantial weaknesses of publicly accessible global biodiversity data
263 concerning cryptic eukaryotes.

264 Eukaryotic distribution patterns reported in related Antarctic studies provide context for our
265 observations from the PCMs. The rarity of Chlorophytes, Ciliophorans, and the otherwise
266 ubiquitous nematodes at MM in relation to the two other lower altitude and more northerly
267 locations (ME, LT) seem to confirm trends of increasing eukaryotic richness and diversity
268 with decreasing latitude and altitude (Czechowski *et al.* 2016a; Thompson *et al.* 2020; Zhang
269 *et al.* 2020), but such patterns are not always evident at the scales investigated here. Rather,
270 Antarctic biodiversity can be surprisingly regionalized (Convey *et al.* 2014) and
271 correspondingly, our study finds surprisingly high eukaryotic diversity to unexpectedly occur
272 even in the harshest environments, such as local ice-soil substrate boundaries at Mount
273 Menzies (Figures 1a, 2). The absence of ciliophorans from Sulphur-rich substrates, and of
274 nematodes from highly conductive soil interstices matches findings of distribution patterns
275 being shaped by age-related salt accumulation at the surface-air interface of frozen soils
276 described with other analytical approaches (Velasco-Castrillón *et al.* 2014b; Lee *et al.* 2019).

277 In absence of other predictors, our study highlights the importance of neutral substrate pH,
278 low conductivity, and key minerals (dolomite, pyroxene, amphibole, or garnet) to predict
279 high eukaryote density in Antarctic substrates. We corroborate the negative influence of
280 substrate alkalinity on Antarctic Basidiomycota (Arenz and Blanchette 2011).
281 Bioregionalization notwithstanding, distance to coast once more appears as suitable proxy
282 variable negatively related to the presence of chlorophytes and ciliophorans (Thompson *et al.*
283 2020). Additionally, we find soil alkalinity, Sulphur content and substrates pyroxene,
284 amphibole, or garnets to constrain distribution of the former. Among nematodes, our results
285 (i.e. perfect alignment between our Antarctic 18S sequence from Mount Menzies and an
286 annotated reference sequence) indicate that *Scottnema lindsayae* could likely occur in high
287 altitude and high latitude environments such as MM, but then would be influenced by the
288 species' general indifference (rather than affinity, compare Zawierucha *et al.* 2019) to

289 alkaline substrates, and must be highly localized (at least at MM) if encountered at high
290 abundance (Smykla *et al.* 2018; Zawierucha *et al.* 2019). Lastly, we confirm the negative
291 association between tardigrade occurrence and alkaline substrates observed in Victoria Land
292 (eg Smykla *et al.* 2018).

293 Based on our findings, ice-free areas with high annual mean precipitation, low wind speeds
294 and relatively high temperatures, exhibiting substrates with a neutral pH and low
295 conductivity, which are rich in dolomite but poor in pyroxene, amphibole, or garnets, are
296 likely to be highly biodiverse in the Antarctic and should harbor candidates for more focused
297 conservation management and higher resolution DNA markers with morphological species
298 level investigations. Furthermore, locations with more extreme environmental conditions may
299 harbor endemic relic fauna equally warranting protection (Convey *et al.* 2014). Our results
300 are in line with observations in other (including polar and alpine) ecosystems, where soil pH
301 was found to be an important factor determining bacterial and fungal community (Siciliano *et al.*
302 *et al.* 2014; Bottos *et al.* 2020). At the same time, Antarctic soil ecosystems are relatively
303 simple and are assumed to mostly lack complex biotic interactions, although such interactions
304 may be more present in coastal terrestrial ecosystems (Velasco-Castrillón *et al.* 2014b; Lee *et al.*
305 *et al.* 2019). Consequently, the soil eukaryote distribution patterns observed especially at Mount
306 Menzies are likely predominantly shaped by abiotic factors and would be gradually more
307 influenced by limited biotic interactions, lower latitude substrates or more coastal substrates
308 (ME, LT).

309 **Conclusion:**

310 We provide a case study highlighting the utility of environmental molecular data and
311 predictive analysis algorithms to inform on the presence of eukaryote taxa by means of
312 relatively easily measured soil predictors, which can be combined with readily available
313 climate data. Rather than recognizing trends, our analytical technique provides accurate
314 detection probabilities for Basidiomycota, Chlorophytes, nematodes, and tardigrades in
315 relation to bedrock mineral composition, pH, conductivity, Sulphur contents, and arguably,
316 overall harshness of environmental conditions. These, here quantified, relationships enable
317 more precise distribution modeling of phylum presences over large spatial scales. Our
318 approach may be used identify regions worthy of species level biodiversity surveys, possibly
319 employing faster evolving molecular markers or logistically more challenging morphologic
320 biodiversity assessments. We believe our approach to be valuable to inform further
321 development and understanding of both Antarctic biogeography and conservation areas.

322 **Acknowledgments:**

323 P.C. was supported by The University of Adelaide through an International Post-Graduate
324 Research Scholarship, through the Royal Society of South Australia, and the University of
325 Otago. M.S. received AAS funding from The Australian Antarctic Division, science project
326 2355 and A.T. was supported by AAS 4296. Alan Cooper and M.S. received funding for this
327 project through Australian Research Council linkage grant LP0991985. M.S. and P.C.
328 received funding for this project from the Sir Mark Mitchell Foundation and the University of
329 Otago. M.K was funded by the University of Otago. We thank Robert McPhee for the artistic
330 contribution of Figure 2. We thank Diana Wall for providing helpful comments on an earlier
331 draft of the manuscript. Authors' contributions to this article conform with the CASRAI
332 Contributor Roles Taxonomy (<https://casrai.org/credit>).

333 **References:**

- 334 Arenz BE and Blanchette RA. 2011. Distribution and abundance of soil fungi in Antarctica at
335 sites on the Peninsula, Ross Sea Region and McMurdo Dry Valleys. *Soil Biol Biochem*
336 **43**: 308–15.
- 337 Bolyen E, Rideout JR, Dillon MR, *et al.* 2019. Reproducible, interactive, scalable and
338 extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**: 852–7.
- 339 Bottos EM, Laughlin DC, Herbold CW, *et al.* 2020. Abiotic factors influence patterns of
340 bacterial diversity and community composition in the Dry Valleys of Antarctica. *FEMS*
341 *Microbiol Ecol* **96**: 1–12.
- 342 Callahan BJ, McMurdie PJ, and Holmes SP. 2017. Exact sequence variants should replace
343 operational taxonomic units in marker-gene data analysis. *ISME J* **11**: 113597.
- 344 Callahan BJ, McMurdie PJ, Rosen MJ, *et al.* 2016. DADA2: High-resolution sample
345 inference from Illumina amplicon data. *Nat Methods* **13**: 581–3.
- 346 Caporaso JG, Lauber CL, Walters WA, *et al.* 2012. Ultra-high-throughput microbial
347 community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621–4.
- 348 Chown SL, Clarke A, Fraser CI, *et al.* 2015. The changing form of Antarctic biodiversity.
349 *Nature* **522**: 431–8.
- 350 Coetzee BWT, Convey P, and Chown SL. 2017. Expanding the protected area network in
351 Antarctica is urgent and readily achievable. *Conserv Lett* **10**: 670–80.

- 352 Conklin AR and Macgregor AN. 1972. Soil adenosine triphosphate: Extraction, recovery and
353 half-life. *Bull Environ Contam Toxicol* **7**: 296–300.
- 354 Convey P, Chown SL, Clarke A, *et al.* 2014. The spatial structure of Antarctic biodiversity.
355 *Ecol Monogr* **84**: 203–44.
- 356 Czechowski P, Clarke LJ, Breen J, *et al.* 2016a. Antarctic eukaryotic soil diversity of the
357 Prince Charles Mountains revealed by high-throughput sequencing. *Soil Biol Biochem*
358 **95**: 112–21.
- 359 Czechowski P, Clarke LJ, Cooper A, and Stevens MI. 2017. A primer to metabarcoding
360 surveys of Antarctic terrestrial biodiversity. *Antarct Sci* **29**: 3–15.
- 361 Czechowski P, White D, Clarke L, *et al.* 2016b. Age-related environmental gradients
362 influence invertebrate distribution in the Prince Charles Mountains, East Antarctica. *R*
363 *Soc Open Sci* **3**: 160296.
- 364 Darienko T, Lukešová A, and Pröschold T. 2018. The polyphasic approach revealed new
365 species of *Chloroidium* (Trebouxiophyceae, Chlorophyta). *Phytotaxa* **372**: 51.
- 366 Davis NM, Proctor DiM, Holmes SP, *et al.* 2018. Simple statistical identification and
367 removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*
368 **6**: 226.
- 369 Friedman J, Hastie T, and Tibshirani R. 2010. Regularization paths for generalized linear
370 models via coordinate descent. *J Stat Softw* **33**: 1–22.
- 371 Geisen S, Briones MJI, Gan H, *et al.* 2019. A methodological framework to embrace soil
372 biodiversity. *Soil Biol Biochem* **136**: 107536.
- 373 Gillespie MAK, Alfredsson M, Barrio IC, *et al.* 2020. Circumpolar terrestrial arthropod
374 monitoring: A review of ongoing activities, opportunities and challenges, with a focus
375 on spiders. *Ambio* **49**: 704–17.
- 376 Lee CK, Laughlin DC, Bottos EM, *et al.* 2019. Biotic interactions are an unexpected yet
377 critical control on the complexity of an abiotically driven polar ecosystem. *Commun*
378 *Biol* **2**: 62.
- 379 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing
380 reads. *EMBnet.journal* **17**: 10.
- 381 Matsuoka K, Skoglund A, Roth G, *et al.* 2021. Quantarctica, an integrated mapping

- 382 environment for Antarctica, the Southern Ocean, and sub-Antarctic islands. *Environ*
383 *Model Softw* **140**: 105015.
- 384 Muthukrishnan R and Rohini R. 2016. LASSO: A feature selection technique in predictive
385 modeling for machine learning. In: 2016 IEEE International Conference on Advances in
386 Computer Applications (ICACA). IEEE.
- 387 R Core Development Team. 2019. R: A language and environment for statistical computing.
388 R Foundation for Statistical Computing, Vienna, Austria. Available via [https://www.R-](https://www.R-project.org/)
389 [project.org/](https://www.R-project.org/).
- 390 Siciliano SD, Palmer AS, Winsley T, *et al.* 2014. Soil fertility is associated with fungal and
391 bacterial richness, whereas pH is associated with community composition in polar soil
392 microbial communities. *Soil Biol Biochem* **78**: 10–20.
- 393 Smykla J, Porazinska DL, Iakovenko NS, *et al.* 2018. Geochemical and biotic factors
394 influencing the diversity and distribution of soil microfauna across ice-free coastal
395 habitats in Victoria Land, Antarctica. *Soil Biol Biochem* **116**: 265–76.
- 396 Thompson AR, Geisen S, and Adams BJ. 2020. Shotgun metagenomics reveal a diverse
397 assemblage of protists in a model Antarctic soil ecosystem. *Environ Microbiol* **22**:
398 4620–32.
- 399 Thompson LR, Sanders JG, McDonald D, *et al.* 2017. A communal catalogue reveals Earth’s
400 multiscale microbial diversity. *Nature* **551**: 457–63.
- 401 Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *Society* **58**: 267–88.
- 402 Velasco-Castrillón A, Gibson JAE, and Stevens MI. 2014a. A review of current Antarctic
403 limno-terrestrial microfauna. *Polar Biol* **37**: 1517–31.
- 404 Velasco-Castrillón A, Schultz MB, Colombo F, *et al.* 2014b. Distribution and diversity of
405 soil microfauna from East Antarctica: Assessing the link between biotic and abiotic
406 factors (X Wang, Ed). *PLoS One* **9**: e87529.
- 407 Wauchope HS, Shaw JD, and Terauds A. 2019. A snapshot of biodiversity protection in
408 Antarctica. *Nat Commun* **10**: 946.
- 409 Xin M and Zhou P. 2007. *Mrakia psychrophila* sp. nov., a new species isolated from
410 Antarctic soil. *J Zhejiang Univ Sci B* **8**: 260–5.
- 411 Zawierucha K, Marshall CJ, Wharton D, and Janko K. 2019. A nematode in the mist:

412 *Scottnema lindsayae* is the only soil metazoan in remote antarctic deserts, at greater
413 densities with altitude. *Polar Res* **38**: 1–12.

414 Zhang E, Thibaut LM, Terauds A, *et al.* 2020. Lifting the veil on arid-to-hyperarid Antarctic
415 soil microbiomes: a tale of two oases. *Microbiome* **8**: 37.

416

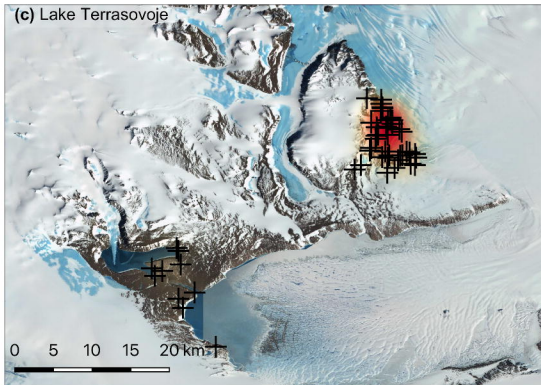
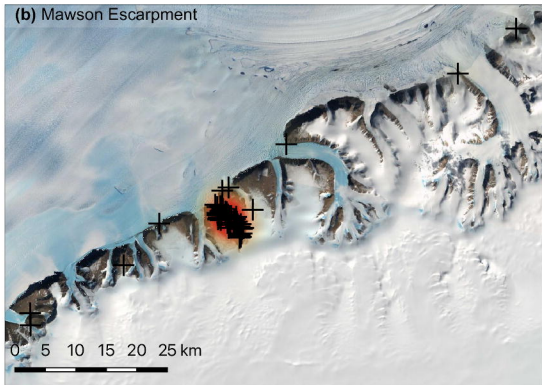
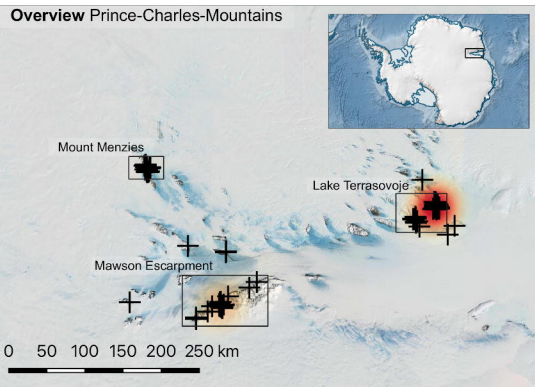
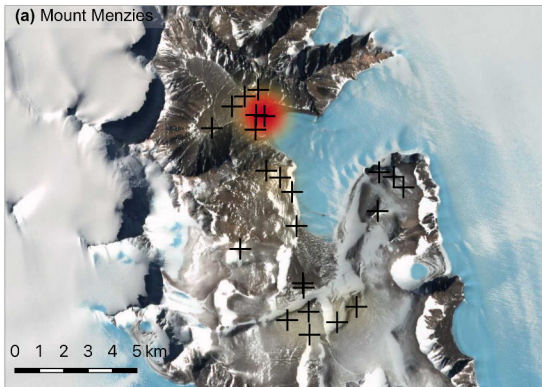
417 **Figure Captions:**

418 **Figure 1:** Sampling area. All sampling locations are marked with a crosshair. Heat shading
419 (at map scale) indicates density of 18S Amplicon Sequence Variants (*sensu* Callahan *et al.*
420 2017) determined to be significantly influenced by substrate qualities as available. Base
421 layers compiled by the Norwegian Polar Institute and distributed in the Quantarctica package.
422 Visit <http://www.quantarctica.org/>. Base layers courtesy of the SCAR Antarctic Digital
423 Database, © 1993–2015 Scientific Committee on Antarctic Research; The National Snow and
424 Ice Data Centre, University of Colorado, Boulder; NASA, Visible Earth Team,
425 <http://visibleearth.nasa.gov/>; Australian Antarctic Division, © Commonwealth of Australia
426 2006.

427 **Figure 2:** Counts of amplicon sequence variants for phyla deemed significantly influenced by
428 substrate composition (left) and examples of taxonomic assignments (right). The employed
429 relatively short primer pair resulted in survey data encompassing diverse soil life forms of
430 various phyla, at the expense of low-level taxonomic certainty, see Web Table 4 for
431 alignment qualities. (a) *Mrakia frigida* (Basidiomycota; perfect alignment) is closely related
432 to a recently described Antarctic species (Xin and Zhou 2007). (b) *Chloroidium*
433 *angustoellipsoideum* (Chlorophyta; perfect alignment) is in the same genus as the recently
434 described *Chloroidium antarcticum* (Dariencko *et al.* 2018). (c) For *Dileptus jonesi*
435 (Ciliophora; 97.6% identity) possible Antarctic distribution could not be confirmed. Both (d)
436 *Scottnema lindsayae* (Nematoda; perfect alignment) and (e) *Mesobiotus furciger* (Tardigrada;
437 perfect alignment) are known Antarctic species with good reference data coverage (Velasco-
438 Castrillón *et al.* 2014a). Base layers courtesy of the SCAR Antarctic Digital Database, ©
439 1993–2015 Scientific Committee on Antarctic Research; The National Snow and Ice Data
440 Centre, University of Colorado, Boulder; NASA, Visible Earth Team,
441 <http://visibleearth.nasa.gov/>; Australian Antarctic Division, © Commonwealth of Australia
442 2006.

443 **Table 1:** Numerical summary of significant coefficient estimates for each phylum as obtained
 444 through lasso logistic regression.

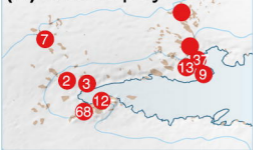
phylum	predictor	95% CI Coefficient		95% CI Odds ratio		Proportion of bootstrap replicates not zero
		lower	upper	lower	upper	
Basidiomycota	Dolomite	0	1.32	1	-3.70	0.93
	PH	-1.54	0.46	0.21	-0.63	1.00
Chlorophytes	MM*	-1.32	-0.10	0.26	0.90	0.99
	PH	-1.28	-0.10	0.28	0.91	0.99
Ciliophora	Garnets	-2.07	-0.11	0.13	0.90	0.99
	MM	-1.22	0.00	0.29	1.00	0.93
	Sulphur	-3.14	0.00	0.04	1.00	0.85
Nematodes	Cond	-2.17	0.00	0.11	1.00	0.99
	MM	-2.10	-0.26	0.12	0.77	0.99
Tardigrades	PH	-1.42	0.00	0.23	1.00	0.95



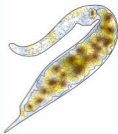
(a) Basidiomycota



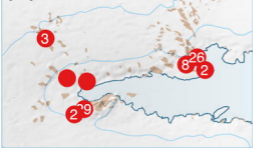
(b) Chlorophyta



(c) Ciliophora



(d) Nematoda



(e) Tardigrada

