

Diverse recruitment to a taxonomically structured global atmospheric microbiota

Stephen Archer

Auckland University of Technology

Kevin Lee

Auckland University of Technology

Tancredi Caruso

University College Dublin

Marcus Leung

City University of Hong Kong <https://orcid.org/0000-0002-6342-8181>

Xinzhao Tong

City University of Hong Kong

Susannah J. Salter

University of Cambridge

Graham Hinchliffe

Auckland University of Technology

Teruya Maki

Kindai University

Tina Santl-Temkiv

Aarhus University

Kimberley Warren-Rhodes

NASA Ames Research Center

Benito Gomez-Silva

Universidad de Antofagasta

Kevin Hyde

Mae Fah Luang University

Celine Liu

Yale-NUS College

Antonio Alcamí

Centro de Biología Molecular Severo Ochoa <https://orcid.org/0000-0002-3333-6016>

Dina Al-Mailem

Kuwait University

Jonathan Araya

Universidad de Antofagasta

Stephen Cary

University of Waikato

Don Cowan

University of Pretoria

Jessica Dempsey

East Asian Observatory

Claudia Etchebehere

Biological Research Institute Clemente Estable

Batdelger Gantsetseg

Institute of Meteorology, Hydrology and Environment

Sean Hartery

University of Canterbury

Mike Harvey

National Institution for Water and Atmosphere

Kazuichi Hayakawa

Kanazawa University

Ian Hogg

Canadian High Arctic Research Station, Polar Knowledge Canada, Cambridge Bay, Nunavut, Canada;
and School of Science, University of Waikato, Hamilton, New Zealand

Mutsoe Inoue

Kanazawa University

Mayada Kansour

Kuwait University

Tim Lawrence

Auckland University of Technology

Charles Lee

University of Waikato <https://orcid.org/0000-0002-6562-4733>

Matthius Leopold

University of Western Australia

Christopher McKay

NASA Ames Research center

Seiya Nagao

Kanazawa University

Yan Hong Poh

Yale-NUS College

Jean-Baptiste Ramond

Pontificia Universidad Católica de Chile <https://orcid.org/0000-0003-4790-6232>

Alberto Rastrojo

Universidad Autonoma de Madrid

Toshio Sekiguchi

Kanazawa University

Joo Huang Sim

Yale-NUS College

William Stahm

East Asian Observatory

Henry Sun

Desert Research Institute

Ning Tang

Kanazawa University <https://orcid.org/0000-0002-3106-6534>

Bryan Vandenbrink

Canadian High Arctic Research Station

Craig Walther

East Asian Observatory

Patrick Lee

City University of Hong Kong <https://orcid.org/0000-0003-0911-5317>

Stephen Brian Pointing (✉ stephen.pointing@yale-nus.edu.sg)

National University of Singapore <https://orcid.org/0000-0002-7547-7714>

Biological Sciences - Article

Keywords: atmospheric microbiology, microbial ecology, global ecosystem

Posted Date: November 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-244923/v3>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Diverse recruitment to a taxonomically structured global atmospheric microbiota

Stephen D.J. Archer¹, Kevin C. Lee¹, Tancredi Caruso², Marcus H.Y. Leung³, Xinzhao Tong³, Susannah J Salter⁴, Graham Hinchliffe¹, Teruya Maki⁵, Tina Santl-Temkiv⁶, Kimberley A. Warren-Rhodes^{7,8}, Benito Gomez-Silva⁹, Kevin D. Hyde¹⁰, Celine J.N. Liu¹¹, Antonio Alcamí¹², Dina M. Al Mailem¹³, Jonathan G.Araya¹⁴, S. Craig Cary¹⁵, Don A. Cowan¹⁶, Jessica Dempsey¹⁷, Claudia Etchebehere¹⁸, Batdelger Gantsetseg¹⁹, Sean Hartery²⁰, Mike Harvey²¹, Kazuichi Hayakawa²², Ian Hogg²³, Mutsoe Inoue²², Mayada K. Kansour¹³, Timothy Lawrence¹, Charles K. Lee¹⁵, Matthias Leopold²⁴, Christopher P. McKay⁷, Seiya Nagao²², Yan Hong Poh¹¹, Jean-Baptiste Ramond²⁵, Alberto Rastrojo¹², Toshio Sekiguchi²², Joo Huang Sim¹¹, William Stahm¹⁷, Henry J. Sun²⁶, Ning Tang²², Bryan Vandenbrink²³, Craig Walther¹⁷, Patrick K.H. Lee³, Stephen B. Pointing^{11,22,27*}

¹ School of Science, Auckland University of Technology, Auckland, New Zealand

² School of Biology and Environmental Science, University College Dublin, Dublin, Ireland

³ School of Energy and Environment, City University of Hong Kong, Hong Kong, China

⁴ Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom

⁵ Department of Life Sciences, Kindai University, Osaka, Japan

⁶ Department of Biology, Aarhus University, Aarhus, Denmark

⁷ NASA Ames Research Center, Mountain View, California, USA

⁸ SETI Institute, Mountain View, California, USA

⁹ Departamento Biomédico, Universidad de Antofagasta, Antofagasta, Chile

¹⁰ Center of Excellence in Fungal Diversity, Mae Fah Luang University, Chiang Rai, Thailand

¹¹ Yale-NUS College, National University of Singapore, Singapore

26 ¹² Centro de Biología Molecular Severo Ochoa, Consejo Superior de Investigaciones
 27 Científicas (CSIC), Universidad Autónoma de Madrid, Madrid, Spain
 28 ¹³ Department of Biological Sciences, Kuwait University, Kuwait City, Kuwait
 29 ¹⁴ Instituto Antofagasta, Universidad de Antofagasta, Antofagasta, Chile
 30 ¹⁵ School of Science, University of Waikato, Hamilton, New Zealand
 31 ¹⁶ Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria,
 32 South Africa
 33 ¹⁷ East Asian Observatory, Hilo, Hawaii, USA
 34 ¹⁸ Biological Research Institute Clemente Estable, Ministry of Education, Montevideo,
 35 Uruguay
 36 ¹⁹ Institute of Meteorology, Hydrology and Environment, Ulan Bator, Tuv, Mongolia
 37 ²⁰ School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New
 38 Zealand
 39 ²¹ National Institute of Water and Atmospheric Research, Wellington, New Zealand
 40 ²² Institute of Nature and Environmental Technology, Kanazawa University, Kanazawa,
 41 Japan
 42 ²³ Canadian High Arctic Research Station, Cambridge Bay, Nunavut, Canada
 43 ²⁴ UWA School of Agriculture and Environment, University of Western Australia, Perth,
 44 Australia
 45 ²⁵ Departamento de Genética Molecular y Microbiología, Pontificia Universidad Católica de
 46 Chile, Santiago, Chile
 47 ²⁶ Desert Research Institute, Las Vegas, Nevada, USA
 48 ²⁷ Department of Biological Sciences, National University of Singapore, Singapore
 49 *e-mail: S.B.P. stephen.pointing@yale-nus.edu.sg
 50

Abstract

Atmospheric transport is critical to dispersal of microorganisms between habitats and this underpins resilience in terrestrial and marine ecosystems globally. A key unresolved question is whether microorganisms assemble to form a taxonomically distinct, geographically variable, and functionally adapted atmospheric microbiota. Here we characterised inter-continental patterns of microbial taxonomic and functional diversity in air within and above the atmospheric boundary layer and in underlying soils for 596 globally sourced samples. Bacterial and fungal assemblages in air were taxonomically structured and deviated significantly from purely stochastic assembly. Patterns differed with location and reflected underlying surface cover and environmental filtering. Source-tracking indicated a complex recruitment process involving local soils plus globally distributed inputs from drylands and the phyllosphere. Assemblages displayed stress-response and metabolic traits relevant to survival in air, and taxonomic and functional diversity were correlated with macroclimate and atmospheric variables. Our findings highlight complexity in the atmospheric microbiota that is key to understanding regional and global ecosystem connectivity.

Introduction

Microorganisms occupy central roles in terrestrial and marine ecosystems globally [1, 2]. Movement of viable cells and propagules between habitats occurs largely through the troposphere, which is the atmospheric layer closest to Earth [3]. This is critical to recruitment and turnover that drive ecological resilience of these systems [2–4], as well as influencing dispersal of pathogens and invasive taxa [5]. There is also a growing awareness that microorganisms suspended in the atmosphere are potentially capable of *in situ* metabolic and biophysical activities that can influence climatic processes [6]. However, despite the central importance of the atmosphere to these ecological outcomes, assessments of microbial

diversity in air at broad geographic scales remain limited [7, 8]. As a result, there is little understanding of how variable the overall microbial composition of the atmosphere may be on a global scale, the extent to which it may be decoupled from underlying local surface communities that are the sources and sinks for atmospheric microorganisms, or the importance of environmental or biotic factors in shaping diversity. The unique role of the atmosphere as a transport medium for microorganisms has also obscured the question of whether it supports a functionally adapted microbiome with the potential for metabolic transformations and cell proliferation [3].

Previous research has been hampered by lack of consensus for community structures of atmospheric microbiota due to the different experimental approaches, lack of ecologically relevant scaling and taxonomic resolution, and the confounding effect of contamination in the ultra-low biomass atmospheric habitat [9, 10]. Nonetheless, inferred community structure for air at various locales within the near-ground atmospheric boundary layer where the bulk of surface-atmosphere interactions occur have described bacterial and fungal communities that were correlated with local abiotic variables such as temperature and humidity [11] or land use [8, 12]. Several studies have related variation in communities to different history of sampled air masses and this suggests combined influence of the different sources and conditions to which microorganisms are exposed during transit [13–16]. Indirect surveys from ground-deposited desert dust [17] or precipitation [18] have yielded valuable insight on long-range dispersal across inter-continental scales although they reflect deposition and differ somewhat to direct estimates from air [15]. Sampling in the free troposphere at higher altitudes above the atmospheric boundary layer is challenging and scarce data indicates a more restricted microbial occurrence[19]. The conventional dogma that atmospheric transport is a neutral process involving ubiquitous distribution of taxa has been challenged by recent theoretical [20, 21] and experimental advances [18, 22, 23]. Adaptive traits have generally been inferred

from taxonomy, although laboratory estimates of metabolic activity by atmospheric bacterial isolates [24], and recovery of RNA from air and cloud water [25, 26], indicate that atmospheric microorganisms are potentially active *in situ*.

Here we report findings testing our hypothesis that atmospheric microbial diversity is distinct from that in underlying surface habitats, is non-randomly assembled, and environmental filtering and diverse recruitment explain observed patterns. We characterised taxonomic and functional diversity in a large globally sourced original dataset ($n = 596$) for air within the atmospheric boundary layer that delineates the majority of physical interactions with the Earth's surface [27] (near-ground air), as well as aircraft sampling of free tropospheric air at higher altitudes above the atmospheric boundary layer (high-altitude air) (Fig. 1; Supplementary Information: Fig. S1, Fig. S2, Table S1). We combined this with concurrent sampling of underlying surface soils and sediments to allow direct air-surface connectivity comparisons. Importantly we report the first study to conduct and report extensive decontamination of sequence data in order to provide a confident diversity estimate where unavoidable sampling and reagent contamination due to the ultra-low biomass atmospheric system must be mitigated [9, 10]. We provide multiple lines of evidence for a taxonomically distinct, non-randomly assembled, altitudinally, geographically and functionally variable atmospheric microbiota that is influenced by a complex suite of biotic and abiotic drivers.

Methods

Sample recovery

The sampling campaign retrieved 596 air and soil samples from 18 locations spanning all major climatic regions and continents plus two oceans. Several dryland (desert) locations were incorporated because they comprise the most abundant terrestrial biome on earth and are

also the single largest natural source of particulate emissions to the atmosphere [62]. The Southern Hemisphere was sampled during April-May 2019 and the Northern Hemisphere during June-July 2019 (Supplementary Information: Table S1). The two circumnavigations encompassed every major climatic biome and included high and low growing seasons. Oceanic and remote land samples were retrieved during independent voyages (May 2017 – June 2018) using the same sampling methodology and samples from the two previously interrogated locations were re-sequenced for this study [23, 63]. Bulk phase boundary layer air was sampled at 1.5m above the surface (near-ground air, $n = 501$) using tripod-mounted air samplers and also above the boundary layer for surface interactions at 2,000 m above local surface level using aircraft mounted-air samplers (high-altitude air, $n = 11$) [27]. Concurrent sampling of underlying soil immediately after each air sampling was conducted within a 25 m radius of air sampling devices (soil, $n = 84$). Ship-board sampling was conducted at 25m above the ocean surface to avoid sea-spray contamination. Logistical challenges limited high-altitude air sampling to six locations although these were nonetheless able to capture a broad geographic and climatic range for both hemispheres.

Recovery of bulk phase air was achieved using three Coriolis μ high-volume impingement devices (Bertin Instruments, France) operated concurrently. This device has been shown to perform well against other samplers [64]. All equipment was transported between locations in sterile containers and bags. Each device was dis-assembled and contact surfaces soaked for one hour with 1.5% v/v sodium hypochlorite (NaClO) followed by three washes of Milli-Q H₂O prior to and after each sampling in order to minimise contamination from cells or nucleic acids. All apparatus and work surfaces used during sampling and sample processing were also cleaned in this way prior to use. All operators wore surface sterilised nitrile gloves during field collections. Randomised collection cones were assembled into the devices without activating the air pump, and these were used as the negative sampling

controls at each location. Additional control samples for potential human contamination were provided via swabs from the inside of anonymised used nitrile gloves (human operator controls).

Samplers were located 3m apart from each other at each sampling location and all inlets were aligned facing prevalent local wind direction. Bulk air was recovered at 300 L/min⁻¹ and particulates recovered after cyclonic deceleration into a sterile phosphate-buffered saline (PBS) impingement medium in each collection cone. Samplers were only approached from downwind during operation. Each device was used to collect discrete 18 m³ air samples as this volume has been shown to result in recoverable environmental DNA [23]. Samples were recovered hourly between 10:00 – 16:00 hrs daily, and then processed immediately by syringe filtration onto a 25 mm polycarbonate filter with 0.2 µm pore size and preserved in 0.5 mL of DNA/RNA Shield (Zymo Research, USA) at ambient temperature during transit and then frozen at -20° C until processed for DNA extraction in the laboratory.

At each location undisturbed surface soil or sediment samples (upper 2cm soil captured in sterile 50 mL screw-cap tubes) were also collected from the base of each Coriolis µ device as well as 25 m away in 120° intervals from the point of sampling. In recognition of inherent soil heterogeneity each sample comprised five subsamples that were mixed and then resampled to yield each sample for analysis. For each sample 0.5 g was preserved for DNA extraction in 0.5 mL of DNA/RNA Shield (Zymo Research, USA) at ambient temperature during transit and then frozen at -20° C until processed for DNA extraction in the laboratory. The remaining sample fraction was archived. It was recognised that soil is not the primary reservoir for terrestrial fungal diversity but in the absence of a practical means to globally sample the diversity of other fungal substrates we accepted this limitation to the study.

DNA extractions from samples were performed under strict microbiological biosafety conditions in randomised sample batches, and each with discrete laboratory controls to assess

potential laboratory or reagent contamination. Each sample tube was processed individually in order to avoid potential cross-contamination between samples due to micro-droplet transfer. All sequencing outputs were evaluated for contamination and spatial and temporal auto-correlation as described in statistical treatments and ecological modelling (Supplementary Information: Figs S5-S14). Environmental DNA was recovered from filtered air and soil samples using a CTAB-based manual extraction protocol optimised for low biomass samples [23]. DNA yield was quantified using the Qubit 2.0 Fluorometer (Invitrogen, USA) and samples were then stored at -20 °C until processed.

Environmental and climate metadata and modelling

Local climate metadata for each sampling location were retrieved from public databases: Mean annual precipitation and mean annual temperature [65], Köppen-Geiger climate classifications [66], growing season that defined the time period when photoautotrophy can occur [67]. Other local variables were recorded using handheld devices: Temperature (°C), relative humidity (%), wind speed (m/S) and wind direction (Kestrel Meters, USA); Particulate matter (PM 2.5 and PM 10.0) (HoldPeak, China).

Back trajectories and metadata relevant to the *in situ* conditions that airborne microorganisms were exposed to during their transit towards each sampling location was modelled from National Oceanic and Atmospheric Administration (NOAA) atmospheric transport and dispersion models [68]. Fourteen day back trajectories of air masses for each sampling time were generated because this is the estimated maximum residence time for microorganisms in the troposphere [69]. Data was obtained from the NOAA HYSPLIT-model and long-range trajectories were estimated using the GDAS database (<https://ready.arl.noaa.gov/HYSPLIT.php>). Data was processed using ArcGIS Pro, version 2.6 (<https://www.esri.com>). The following variables were calculated at intervals along each tropospheric transport path from the HYSPLIT models: altitude (m above ground level,

AGL), wind speed (m/S), direction, temperature (°C), Relative humidity (%), solar irradiance (W/m²), precipitation events, and transit duration over land or ocean surface. Points were plotted on WGS 1984 Web Mercator coordinate system with date line wrapping and climate mapping at 1km resolution [66].

DNA recovery and gene copy number estimation

DNA yield per m³ air or per gram of soil were used as a proxy for total biomass [70], although estimates between soil and air are not directly comparable due to different substrate volumes and composition. Taxonomic assignment of reads in metagenomes was used to approximate relative abundance between samples within each kingdom. An additional and commonly-used estimate of relative abundance using real-time quantitative PCR (qPCR) was also employed for the most abundant microbial groups (bacteria and fungi) [71]. Primers used for bacteria targeted the 16S rRNA gene V3-V4 hypervariable region (Fwd 341 5'-CCTACGGGNGGCWGCAG-3' and Rev 785 5'-GACTACHVGGGTATCTAATCC-3') [72, 73], with LightCycler 480 SYBR Green I Master mix (Roche Holding, Switzerland). Primers for fungi targeted the 18S rRNA gene (FungiQuant-F 5'-GGRAAACTCACCAGGTCCAG-3' and FungiQuant-R 5'-GSWCTATCCCCAKCACGA-3') with TaqMan probe FungiQuant-PrbLNA 6FAM-5'-TGGTGCATGGCCGTT-3'-BBQ [74]. TaqMan Fast Advanced Master Mix was used for qPCR reactions with the following conditions: denaturing step: 95 °C for 20 s; cycling step: 35 cycles of 95 °C for 1 s and 60 °C for 20 s [39]. A qPCR standard for each target sequence was developed to estimate gene copy number using pooled samples. These were amplified using TaqMan Fast Advanced Master Mix as described above but without fluorescent markers (Applied biosystems, USA) and quantified using a Bioanalyzer (Agilent Technologies, USA). Serial dilutions of the template were used to generate standard curves.

Amplicon sequencing

Targeted high-throughput amplicon sequencing was employed to gain insight into taxonomic diversity of bacteria and fungi because they comprised the majority of microbial reads in our metagenomic libraries and have been estimated as the most abundant microorganisms in aerosols [22]. Amplicon sequence libraries were prepared using Illumina MiSeq v3 600 cycle chemistry as per manufacturer's protocol with PhiX positive spike-in controls. All samples were sequenced to near-asymptote (Supplementary Information). Template DNA in samples was normalised to 2.5ng/ul prior to two-step PCR amplification for the bacterial 16S rRNA gene V3-V4 hypervariable region (Fwd 341 5'- CCTACGGGNGGCWGCAG-3' and Rev 785 5'- GACTACHVGGGTATCTAATCC -3') [72, 73], and fungal ITS1 region (Fwd ITS1 5'- CTTGGTCATTTAGAGGAAGTAA -3' and Rev ITS2 5'-GCTGCGTTCTTCATCGATGC -3') [75, 76]. The amplicon sequence libraries were first processed with cutadapt v2.7 [77] to remove primer sequences. Amplicon sequence variants (ASVs [78]) were generated for 16S rRNA amplicons (truncLen=c(230,220), maxN=0, maxEE=c(2,5), truncQ=2) and ITS amplicons (minLen=50, maxN=0, maxEE=c(5,8), truncQ=2) dada2 v1.14 [79]. Pseudo-pooling was used in ASV calling to increase sensitivity and accuracy in alpha diversity estimation. Taxonomic classification was conducted in dada2 with SILVA v138 [80] and UNITE v7.2 [81] as references. Overall the amplicon sequencing generated 19.5 million bacterial reads and 1.7 million fungal reads and these resolved to over 200,000 genuine ASVs. After the decontamination steps, true samples with > 1,000 reads (16S rRNA $n = 529$, ITS $n = 444$) were used for all subsequent analyses.

The use of high-throughput DNA sequencing for samples from low biomass habitats such as air raises the issue of confounding signal due to contaminants that are otherwise indistinguishable in higher biomass samples. We employed an experimental design for sample recovery and quality filtering of sequence data that embraced recommended best practice for minimising contaminant signal [10] (Supplementary Information). Diversity

estimation occurred only after an aggressive decontamination protocol to mitigate putative contamination in sequence libraries from our ultra-low biomass air and soil samples (Supplementary Information: Figs S5-S14, Table S2) [9, 10, 82]. This comprised subtractive filtering steps for ASVs as follows: 1) removal of non-target sequences; 2) Removal of ASVs with suspicious frequency and/or prevalence; 3) Removal of all ASVs encountered in any of the sampling, human operator or laboratory controls from all samples (i.e. not just from controls specific to a given location); 4) A highly aggressive subtractive filtering at genus level based on a meta-analysis of putative contaminants from other studies of low biomass samples [9, 10] regardless of whether or not they were also encountered in our controls. Suspiciously frequent and/or prevalent ASV were identified using the [isContaminant] function of the R package decontam [83] and removed if they met the stringent statistical threshold for frequency (0.1) or prevalence (0.5). The prevalence test was used as a further check on the step for removal of ASV from controls. The genus-level filtering targeted human-associated bacterial and fungal genera. Overall, the decontamination pipeline for our ultra-low biomass samples resulted in the removal of 16% soil, 43% NG air and 38% HA air reads from 16S rRNA gene libraries, and 28% soil, 55% NG air and 61% HA air reads from ITS libraries, and these values were in line with recommended best practice and encounter expectations for ultra-low biomass samples [9, 10, 82].

Downstream analyses were performed on datasets with and without the decontamination steps for comparison and the community composition was compared at ASV and genus level as a check for ASV inflation of diversity estimation. Post-hoc analysis of the pre- and post-decontaminated datasets were employed to estimate the effectiveness of the multi-step subtractive filtering process and identify any remaining ASVs that may represent potential residual contaminants, as well as identify any evidence for cross-contamination or auto-correlation between sample types. Data were subsampled and rare taxa removed prior to

analysis performed on genus level taxa. Taxa were assigned pairwise correlation scores using FastSpar v1.0.0 [84] and heatmaps of correlation scores generated using gplots [85]. Groups of tightly co-associated taxa with consistent relative abundances were proposed as candidate artefacts: samples containing these artefacts were checked for correlation to available metadata such as DNA extraction or sequencing batches, and spatial or temporal sampling effort. The artefacts identified using this approach in the pre-decontaminated data were confirmed to have been removed by the decontamination pipeline. The small number of potential residual artefacts in the post-decontamination dataset are reported in the Supplementary Information.

Shotgun metagenomics

Independent replicates were pooled by sampling day and device to yield 120 pooled samples and 3 pooled controls for metagenomics sequencing. Libraries were prepared using a low-input preparation protocol where required [86] and using the Nextera XT library kit and sequenced (2×150 bp paired-end) on an Illumina NextSeq 500 (Illumina, USA). Kneaddata (v0.7.4, default settings, <https://github.com/biobakery/kneaddata>) was used to remove low-quality reads and human DNA using the human genome hG37 as reference from raw fastq files.

Similar to the steps adopted for amplicon sequencing, filtered metagenomics reads were further processed in a multi-step fashion to systematically identify and remove potential contaminating nearest taxonomic units (NTU). Filtered reads from the controls were co-assembled into contigs using the “assembly” module of MetaWRAP (v1.2.1) [87]. Reads in the samples that mapped to the contigs constructed ($\geq 1,000$ bp) in the controls were removed using Kneaddata. Next, taxonomic classification for NTU was performed using Kraken (v2.0.9-beta) [88] based on the PlusPFP database (Dec 2nd, 2020 update) and species-level NTU classification was optimised using Bracken (v2.6.0) [89]. Fungal species were

identified using FindFungi (v0.23.3) [89]. Species-level information from Kraken2 and FindFungi were processed to identify potential contaminating taxa using the same steps applied to our ASV data [83]. Putative contaminants were subsequently removed using the “extract_kraken_reads.py” command (option --exclude and --include-children) from KrakenTools (v2.0.8-beta, <https://github.com/jenniferlu717/KrakenTools>). The percentage of unassigned reads in metagenomes was 70.34% (s.d. 14.61%), and this compares favourably with other recent studies of metagenomes from air [39]. Reads cleared of bacterial contaminants were subjected to another round of contaminant removal using Kneaddata to discard reads that mapped to representative genomes of the fungal contaminants. Genus-level subtractive filtering was not applied to archaea or protists, and viruses were poorly represented in our metagenome libraries although methodological limitations may have reduced their detection. After all the decontamination steps, a total of 1,498,558,646 high-quality paired-end reads were retained across the entire dataset of 120 metagenomes, averaging 12,487,988 reads per sample. The aggressive decontamination pipeline resulted in removal of 6% soil, 8% NG air and 8% HA air filtered reads. Taxonomic profiles (phylum and species) of the high-quality reads were generated with Kraken2 and Bracken. In addition, FindFungi was used to identify fungal phyla.

Functional potentials of the metagenomes were queried using HUMAnN (v3.0.0.alpha.3) [90], generating a total of 1,855,295 unique features, which were subsequently converted to 10,440 protein families (Pfams). Pfams corresponding to genes encoding carbon fixation, cold shock response, nitrogen cycle, oxidative stress, phototrophy, respiration, sporulation, starvation, trace gas metabolism, and UV repair proteins (Supplementary Information: Table S5) were examined to understand factors shaping the abundance of metabolic and stress response potentials. Gene abundance data were expressed in terms of copies per million reads from HUMAnN, which takes into account library size

differences between samples. For functional gene analysis, differentially more abundant genes in air versus soil were inversely correlated with biomass and taxonomic richness, they affiliated with taxa observed as enriched in air, and all values were averaged by location. We therefore interpreted the elevated abundance of genes in air as reflective of assemblage composition rather than an artefact of sampling effort.

Statistical treatments and ecological modelling

Statistical analysis: General processing of the community data including the calculation of relative abundance and estimates of alpha diversity were conducted using the R package phyloseq [91] and visualised using ggplot2 [92]. Comparative statistical analyses were performed using R: ANOVA, Kruskal-Wallis, Mann-Whitney test, Mantel test, PERMANOVA, and Procrustes analysis using vegan [93]; lmPerm (permutation test for ANOVA) (<https://cran.r-project.org/web/packages/lmPerm/index.html>), dunn.test (Dunn's test for post hoc analysis, *P*-values were adjusted by the Holm–Bonferroni method) (<https://cran.r-project.org/web/packages/dunn.test/index.html>), ANCOMBC (ancombc differential abundance analysis, *P*-values were adjusted by the Holm–Bonferroni method) [94]. Correlations used in compositional analysis of taxonomic data to determine potential residual contaminants were calculated using FastSpar [84]. Sequence data were rarefied as appropriate for each analysis. Calculation of geographic distances were performed using R package geosphere [95] function distGeo with WGS84 ellipsoid. Source tracking was conducted by fast expectation-maximization using FEAST [44] with data from other studies (processed using dada2 following the same parameters as this study) as additional sources/sinks [96–102] and NCBI BioProject PRJEB42801. For correlation analysis between abiotic and biotic variables the Pearson correlation coefficient for multiple pairwise combinations were calculated using the R package corrplot [103], with *P*-value cut-off of 0.05 corrected for multiple tests using Bonferroni correction. To visualise patterns of

community dissimilarity, two methods were used. Hellinger distances were ordinated with t-distributed stochastic neighbour embedding (tSNE) using R package Rtsne [104]. We also calculated Jaccard sample pair-wise distances based on the 10,000 most abundant and frequent ASVs using the R package vegan [93], and tested for locations and habitat (i.e. soil, near ground air and high elevation air) using PERMANOVA [105]. A preliminary analysis based on all reads and ASVs showed qualitatively similar patterns but higher noise (i.e. the amount of variance accounted for by the major ordination axes was relatively low due to a very high number of ASVs found only at one or two locations). We decomposed the Jaccard matrix with Principal Coordinate Analysis (PCoA) which provided a quantification of the variance accounted by each ordination axis [106].

Network null models were employed to detect non-random structure in the microbial assemblages. A statistical mechanics approach was employed for network construction [31], and defined our networks as bipartite matrices with two layers: location and taxa. Analyses were performed at multiple taxonomic ranks: Phylum, Class, Order, Family, Genus, and ASV to test the expectation that the higher the taxonomic rank the more widespread major taxa are likely to be, which may result in a random distribution of taxa across locations (i.e. most phyla are found everywhere, with difference between location due to random sampling errors and difference in richness between locations). To fully test our hypothesis, we employed degree sequence constraints to enforce that for each taxon, the total number of locations in which the taxon was found was a constraint, and for each location the total number of taxa found in that location, disregarding location or taxa identity, was also a constraint. Specifically, we used the canonical bipartite configuration model, where the constraint was enforced on average and we thus we used maximum-likelihood [107, 108] to estimate the probability distribution that maximised entropy for the canonical ensembles. This approach resolved the issue encountered with random permutation for very heterogeneous and sparse

matrices, which cannot be generated and sampled in a statistically unbiased way [109, 110]. We sampled the resulting probability distribution [107, 108], to obtain 999 null matrices representing an unbiased sample of the canonical ensemble of our location by taxa matrices using the MatLab routine Max&Sam [111], and imported the null matrices in R for downstream analyses. We used the R packages bipartite [112] and vegan [93] to calculate the nestedness metric of NODF and Jaccard dissimilarity on the observed and null matrices. We then used the standard definition of effect size [113] to quantify the difference between observed metrics and the null distribution of the metrics. Since the distribution of the 999 null metrics were approximately normal, an effect size larger than 2 standard errors corresponded to taxonomic composition that diverged more than expected under purely random assembly with an approximate P -value < 0.05 . We calculated Z-scores for nestedness to indicate the number of standard deviations a given data point lay from the mean using the commonly employed NODF metric [114], and also using Jaccard distance to estimate pair-wise assemblage dissimilarity and test with the null model if the average dissimilarity deviated from that expected under random assembly.

Results

An overview of inter-domain diversity from our metagenomic libraries indicated that composition was more variable in air than soil (average Bray-Curtis dissimilarity within air samples = 0.256 vs. within soil samples = 0.038, Mann Whitney $U = 8.7 \times 10^7$, $P = < 2.2 \times 10^{-16}$, Wendt effect size $r = 0.493$) (Fig. 1). Bacteria were the most abundant component of soil and air metagenomes and fungi were the second highest microbial category in near-ground air. Relatively low and patchy contribution was observed for archaea and microbial eukaryotes. We therefore focused further community profiling effort on bacteria and fungi with shotgun metagenomics and targeted amplicon sequencing. Sequencing of 596 samples to

near-asymptote (Supplementary Information: Fig. S3, Fig. S4) was combined with an
 unprecedented effort to mitigate against the occurrence of putative contaminant taxa and
 cross-contamination that have plagued low-biomass microbiological studies [9, 10]
 (Supplementary Information: Figs S5-S16, Table S2). Bacterial and fungal gene copy number
 indicated a conserved pattern globally where values were highest per g soil > per m³ near-
 ground air > per m³ high-altitude air (Kruskal-Wallis $H(2) = 93.019$, $P < 0.05$ and $H(2) =$
 98.825 , $P < 0.05$ respectively, pairwise comparison with post-hoc Dunn's test (between all
 comparisons $P < 0.05$) (Fig. 2; Supplementary Information: Fig. S17). Although soil and air
 within and above the atmospheric boundary layer are not directly comparable in terms of
 habitable characteristics, magnitude differences in gene copy number and estimated biomass
 occurred between soil and air habitats at all but the most extreme Atacama Desert location
 where soils are microbiologically depauperate (Supplementary Information: Fig. S17) [28].
 Our community composition estimation using amplicon sequencing and shotgun
 metagenomics were positively correlated (Procrustes: Bacteria $m^2 = 0.76$, correlation = 0.49,
 $P = 0.001$; Fungi $m^2 = 0.56$, correlation = 0.66, $P = 0.001$) (Supplementary Information: Fig.
 S18), and so we focused our fine scale phylogenetic interrogation on amplicon sequence data
 because this approach allowed better ecological representation of the targeted assemblages in
 terms of sampling depth and taxonomic resolution [29]. We employed the Jaccard distance
 matrix to visualise the bacterial and fungal assemblages at sampling locations in three non-
 overlapping and highly significant two-dimensional (Fig. 2) and three-dimensional
 (Supplementary Information: Fig. S19) clusters (PERMANOVA both $P = < 0.01$, $R^2 = 0.157$
 and $R^2 = 0.364$ respectively). The clusters corresponded to soil, near-ground air and high-
 altitude air habitats and this dominated community patterns. We also used Hellinger
 transformed Bray-Curtis distances to visualise bacterial and fungal communities grouped by
 habitat (soil, near-ground air, high-altitude air) and locations (Supplementary Information:

Fig. S20). The interaction terms for bacteria and fungi were significant, however, the dispersions were not homogeneous (betadisper $P < 0.05$) and so this may also influence observed patterns. Significant linear distance decay relationships for diversity were observed for bacterial assemblages (Mantel test, Soil: $r = 0.357$, $P = 0.001$; NG air: $r = 0.353$, $P = 0.001$; HA air: $r = 0.433$, $P = 0.002$) and fungal (Soil: $r = 0.507$, $P = 0.001$; NG air: $r = 0.482$, $P = 0.001$; HA air: $r = 0.535$, $P = 0.004$) (Supplementary Information: Fig. S21). However, there was no evidence for a latitudinal gradient in richness and this mirrored observations for global soil bacterial diversity [30], and also reflected the inclusion of diverse habitats including deserts, mountains, high latitude and ocean locations in our study.

We then constructed a general null model of the taxa occurrence matrices by habitat and location to validate our hypothesis that microbial diversity in air is non-randomly assembled (Supplementary Information: Fig. S22). We used the statistical mechanics of networks [31] to formulate correct null models for our data set subject to the constraints of observed taxonomic richness at each location. We calculated two metrics of community structure on the observed matrix and the ensemble of null model matrices: NODF for nestedness [32], and the classical Jaccard index for taxonomic compositional dissimilarity which is a proxy to beta diversity and turnover in taxonomic composition. Bacterial and fungal assemblages in both near-ground and high-altitude air were significantly less nested than null models when compared to the confidence interval of the baseline provided by the models, and therefore identified as taxonomically structured and non-randomly assembled (Fig. 2) [33]. The specific non-random patterns (i.e. significantly less nested than expected under random taxonomic composition) implied taxa specificity to habitat and location, and we interpreted these as indicative of strong filtering for taxa. Bacterial assemblages in soil were more similar than expected under the null model and this reflects observed global diversity patterns for soil bacteria [30]. The highly significant deviations from the null

models (Null model overall $P = <0.01$), also at higher altitudes above the atmospheric boundary layer where abiotic stressors are more pronounced, suggested that the communities were selected towards non-random taxonomic compositions. The pattern was corroborated by Jaccard index estimates that showed observed bacterial and fungal assemblages in air were more dissimilar between locations than expected under the null model (Null model with overall $P = <0.01$) (Supplementary Information: Fig. S15, S16). Nestedness patterns converged towards the null model for all habitats at broader taxonomic ranks and this pattern has been interpreted as indicative of conserved traits among soil microbial groups, rather than due to diversification and dispersal over short time scales [33]. The pattern persisted between hemispheres sampled at peak and low growing season and across major climatic boundaries and land use types. The non-random distribution of taxa across habitat and locations suggested that some form of ecological selection (*sensu* [34]) was operating on the microbial assemblages. We propose that this was environmental filtering in both near-ground and high-altitude air, combined with dispersal limitation, which most likely operated in terms of local surface emissions to air. We conclude that this resulted in structured and biogeographically predictable patterns for bacteria and fungi (i.e., different environmental matrices such as soil and air, and different locations, display specific, non-random taxonomic compositions).

A detailed analysis of the taxonomic composition of assemblages further confirmed the macroecological patterns quantified with our null model approach. At broad taxonomic ranks (phylum-class) a relatively consistent diversity was observed in air globally regardless of underlying biome or growing season (Fig. 2; Supplementary Information: Fig. S23). A comparison of Hellinger distances between ASV and Genus defined communities revealed observations to be highly congruent between the classification methods for bacteria ($m^2 = 0.239$, correlation = 0.872, $P = <0.01$) and fungi ($m^2 = 0.187$, correlation = 0.902, $P = <0.01$). Our amplicon sequence variant (ASV) approach to diversity analysis revealed that at finer

taxonomic scale (genus-ASV) and after extensive decontamination effort there were 13% of
 bacterial and 10% of fungal genera co-occurring among $\geq 50\%$ of globally distributed air
 samples (Supplementary Information: Table S3, Table S4). The only genus with ubiquitous
 representation in all air samples was *Sphingomonas*, a diverse group linked with emissions
 from the phyllosphere [12, 35]. Evidence from taxonomic data for environmental filtering of
 assemblages in air supported the conclusions of our nestedness analysis. Bacteria enriched in
 near-ground air compared to soil were largely accounted for by classes encompassing taxa
 with known tolerance to environmental stress (Actinobacteria, Firmicutes [Bacilli, Clostridia,
 Limmnochordia, Negativicutes] and Gammaproteobacteria) (ANCOM-BC Holm Adjusted P
 $= < 0.05$; Effect sizes $W = -2.00; 0.05; -7.89; -9.59; -3.76; -5.66; -4.09$ respectively), although
 it cannot be ruled out that this also indicates taxa that possess adaptive traits that favour
 aerosolization [36]. At higher altitudes where environmental stress is exacerbated the spore-
 forming Actinobacteria and Firmicutes were more abundant (ANCOM-BC Holm Adjusted P
 $= < 0.05$; Effect sizes $W = \text{NG air-HA air } -3.35, \text{ NG air-Soil: } -2.39; \text{ NG air-HA air: } -4.67,$
 $\text{NG-Soil: } -8.96$) (Fig. 2; Supplementary Information: Fig. S23), suggesting selection towards
 survival as passive resting stages. Significantly elevated abundance of gammaproteobacterial
 taxa at the farm location in South Africa (ANCOM-BC Holm Adjusted $P = < 0.05$; Effect
 size $W = < 0$ for all comparisons) was consistent with emissions of this group from agricultural
 surfaces [37]. In the absence of observed mycelia in air samples we concluded that spores
 accounted for much of the fungal signature in air (Supplementary Information: Fig. S23).
 This is corroborated by the elevated relative abundance of macrofungi (Agaricomycetes) and
 prolific spore-formers (Dothidiomycetes) (Fig. 2). The Agaricomycetes were significantly
 more abundant in tropical near-ground air than all other locations globally (ANCOMBC
 Holm Adjusted $P = < 0.05$, Effect size $W = 12.83$) and this likely reflected global patterns for
 terrestrial fungi [38]. In near-ground air the abundance of common fungal agricultural

pathogens (Ustilaginomycetes) was significantly elevated in temperate Northern Hemisphere locations sampled during peak growing season (ANCOMBC $P = < 0.05$, Holm Adjusted $P = < 0.05$; Effect size $W = 8.66$), as opposed to reduced abundance in Southern Hemisphere samples collected at the end of the growing season. Based on this observation, we suggest that seasonality in land use on a global scale is among decisive factors impacting diversity of atmospheric fungi. Previous studies at individual near-ground locales have concluded that inter-seasonal variation may variously be absent [39], weak [13], pronounced for some taxa [14] or stochastic [15]. Elevated fungal diversity in ultra-low biomass high-altitude air was indicative of persistent fungal propagules that are tolerant to extreme, prolonged UV and thermal stress. This is consistent with typically extended residence time for airborne cells at high altitudes, and this necessitates effective tolerance to these stressors during potentially long-distance dispersal [40]. Overall, our combined ecological and taxonomic data provided strong evidence that contrary to long-held dogma in microbial ecology that microbial transport in air is ubiquitous and neutral to dispersal outcomes [11, 23, 40], instead the atmospheric microbiota exhibit a pronounced biogeographic distribution.

In order to further interrogate possible explanations for the observed patterns, we conducted source tracking analysis to assess the likely origin of bacteria and fungi encountered in the air. First, a connectivity analysis revealed that near-ground air displayed greatest taxonomic connectivity with local soil at any given location and less connectivity with soil from different locations (two-way ANOVA with permutation test [5,000 iterations] $P = < 2.2 \times 10^{-16}$) (Fig. 3). Assemblages in high-altitude air displayed significantly fewer shared taxa with underlying near-ground air or soil (two-way ANOVA with permutation test [5,000 iterations] $P = < 2.2 \times 10^{-16}$). The shapes of the curves demonstrate the ASVs shared among all samples in a habitat type, with shared ASVs co-occurring the most in soil>near-ground air>high-altitude air (Supplementary Information: Fig. S24). Both bacterial and fungal

communities in high-altitude air showed much lower overall ASVs than other sample types, and displayed an obvious proximity to their lower asymptotes, suggesting more distinct and less connected communities at higher altitudes. Aerosolization of microorganisms not only occurs from soil but also from different terrestrial and aquatic surfaces, e.g. ocean surface waters [16, 22], the phyllosphere [35, 41] and desert dust events [42, 43]. We therefore employed fast expectation-maximization source tracking (FEAST) [44], to estimate recruitment to air microbiota from the surface habitats of different climatic regions (Fig. 3; Supplementary Information: Fig. S25). We matched exact ASV taxa rather than a more general operational taxonomic unit (OTU) approach based on 97-99% sequence similarity that has been previously applied to atmospheric source-tracking, and this resulted in a large volume of taxa with unexplained source but may also reflect that it is impossible to exhaustively sample potential sources. For most locations local soil was the major explained source of bacteria and fungi in air, and bacteria were sourced in a more cosmopolitan manner than fungi (Supplementary Information: Fig. S25). Many sampled air masses had significant transit over oceans and yet marine sources were a relatively minor contributor to observed diversity in air above terrestrial locations. This reflects that fewer microorganisms occur above the oceans than over land [22], and also the limited number of oceanically sourced sequence libraries for comparison. Clear patterns for terrestrial sources were apparent. Dryland soils (dry deserts, polar/alpine and dry continental locations) were pronounced sources for bacteria globally (Mann-Whitney $U = 73$, $P = 2.937 \times 10^{-5}$; Wendt effect size $r = 0.70$) and this may reflect the more readily aerosolised non-cohesive soils typical of these biomes [43]. This expands the influence of deserts to global-scale atmospheric microbiota beyond the well-defined intercontinental desert dust transit routes for microbial dispersal [42]. For the fungi, polar and alpine soils were major sources and this is congruent with the notion that permanently cold surface substrates in these environments have been proposed to act as

long-term reservoirs for inactive fungal propagules [23]. The phyllosphere was a pervasive contributor to bacterial diversity globally, and the relatively minor contribution to fungal sources likely reflects the lack of available comparative data and broad diversity in host surfaces. This may emerge as a more significant source as the inventory of phyllosphere microbiomes increases. For high-altitude air, significant major sources of bacteria were dry deserts and polar/alpine sources (Mann-Whitney $U = 1$, $P = 0.018$, Wendt effect size $r = 0.93$), and this likely reflects in part the adaptive advantages that taxa from these habitats have in air, e.g. UV repair and desiccation tolerance [43]. The ability to become aerosolised may vary between taxa in marine [36] and terrestrial [45] systems and so deterministic biotic drivers may also be relevant to recruitment from sources, as well as selective deposition during transit [46]. Overall, the source tracking demonstrated that atmospheric diversity is driven by a complex recruitment process involving cell emissions from local soils and transport from more distant sources, and particularly from drylands and the phyllosphere.

To generate further insight into possible biotic drivers of the observed diversity patterns we conducted a functional metagenomic analysis for 120 metagenomes of selected metabolic and stress-response genes relevant to the atmospheric habitat [47, 48] (Fig. 4; Supplementary Information: Fig. S10). We targeted bacteria because they likely comprise any active fraction of the atmospheric microbiota [25]. Distribution of marker genes in air broadly reflected that for underlying soil at terrestrial locations and this supported our identification of soil as a major source for atmospheric bacteria. Traits were widely distributed globally and those for stress tolerance were notably more abundant in bacterial assemblages in air above dry and polar/alpine regions (Mann Whitney $U = 2.1 \times 10^4$ $P = 0.02$, Wendt effect size $r = 0.12$), thus further supporting our hypothesis that microorganisms from these surface environments are adapted to survival of the stressors encountered in air [49]. Compared to soil, air communities possessed higher abundance of the oxidative stress gene *msrQ* (Mann

Whitney $U = 2,061$, $P = 0.02$, Wendt effect size $r = 0.22$), UV-repair gene *phrB* (Mann Whitney $U = 2,114$, $P = 0.01$, Wendt effect size $r = 0.25$), and starvation gene *slp* (Mann Whitney $U = 1916.5$, $P = 0.01$, Wendt effect size $r = 0.24$) (Extended Data Fig. 10). High abundance for stress-response genes in air above ocean locations may indicate that the low biomass and taxonomic richness above marine surfaces reflects strong environmental filtering. This may arise during long-distance transport from largely terrestrial sources, as well as during recruitment of bacteria from the sea surface micro-layer [50]. Metabolic marker genes for respiration were widespread, and notably for the *ccoN* proteobacterial cytochrome oxidase that correlated with elevated proteobacteria in air versus soil. Markers for the metabolism and fixation of a variety of gaseous atmospheric substrates including carbon dioxide, hydrogen, methane, nitrogen and isoprene, as well as phototrophy were also abundant in air. Elevated occurrence in air of the *coxL* gene associated with carbon monoxide metabolism was indicative of the potential for interaction with anthropogenic emissions [51]. This limited functional interrogation provided a much-needed glimpse into the potential for an active and stress-adapted atmospheric microbiome. Our data indicates that there is capacity for greater metabolic plasticity than the existing inventory from molecular genetics [26, 45] and transformation of substrates by atmospheric isolates under laboratory conditions currently suggests [52, 53].

We examined possible interactions between the taxonomic and functional diversity of assemblages and abiotic variables relevant to survival in air and soil (Fig. 5). These included both location-specific macroclimate variables, and a novel geospatial analysis approach to capture environmental conditions encountered by microorganisms during transit in air (Supplementary Information: Fig. S2). Significant correlations were revealed between both local macroclimate and transit abiotic variables and community metrics of taxonomic and functional diversity in air ($P = 0.005$ after Bonferroni correction). Relatively strong negative

correlations for bacterial and fungal richness and relative abundance with solar radiation and altitude provided further evidence for UV exposure as a strong selective force on global bacterial and fungal diversity. Functional genes were most strongly correlated with mean annual precipitation, and this likely reflects niche differentiation of source communities in underlying soil at different climatic locations since we have shown they are coupled to diversity in local air. Transit variables were less influential on functional diversity and this was consistent with our hypothesis that most microorganisms in air are inactive. The correlations between occurrence of phototrophy and carbon fixation genes and several abiotic variables suggested photoautotrophic bacteria may be subject to greater selective pressure than other groups, but also likely reflects source climate because emissive dryland surfaces are typically dominated by photoautotrophic microbial soil crusts compared with plant cover in temperate and tropical climates [43]. For soil communities the correlations with macroclimate variables were broadly congruent with those observed for other global studies of soil microbial diversity [54], and this provided triangulation for our approach. These data highlight that although variables at a specific location are important, the conditions to which microorganisms are exposed during transit are a significant but previously overlooked factor affecting with taxonomic and functional diversity and are influential to dispersal outcomes.

Discussion

Overall we have demonstrated that atmospheric microbiota from different continents and climatic zones are non-randomly assembled, display geographic and altitudinal biogeography across large spatial scales, are recruited from a complex combination of local and distant sources, and display functional attributes favourable to survival in the atmosphere. A major strength of our study is the unprecedented attention to mitigation of confounding factors associated with sampling and bioinformatics processing of microbial diversity data from the

ultra-low biomass atmospheric environment. We have demonstrated that standardised sampling and careful attention to decontamination of environmental sequence data from ultra-low biomass atmospheric samples can reveal clear biogeographic patterns.

Based upon these findings, we envisage a global system where for any ecological region highly filtered and taxonomically structured microbial communities assemble across multiple spatial and temporal scales, which are fundamentally affected by cell survival due to environmental filtering as well as flux from source habitats. Underlying surface habitats serve as key local sources although it is clear from our findings that aerosolization and residence in air exert strong environmental filtering. It is not possible to clearly delineate between these two drivers given current understanding in microbial ecology, but we envisage selective emissions from different microbial habitats are important over short timescales whereas for longer residence times in air environmental filtering becomes more influential. We identified that atmospheric diversity is punctuated by long-distance transport of taxa primarily sourced from drylands and from the phyllosphere. Drylands support microorganisms that can be regarded as pre-adapted to atmospheric survival in view of the similar environmental stressors, namely xeric and osmotic stress and photo-oxidative stress [55]. The phyllosphere is a major source of microorganisms [41], and a bioprecipitation feedback has been proposed that links atmospheric and phyllosphere microbiota via their involvement in ice nucleation that influences precipitation and vegetation patterns [56]. Our data suggests the influence of this feedback may extend across a broad geographic range. Future research that integrates spatial and temporal scales will yield further insight on atmospheric microbial biogeography.

Our findings also point towards an atmospheric microbiota that is enriched in stress adaptation and metabolic traits that may allow microbial activity in the atmosphere. This has important implications for consideration of the atmosphere as a true microbial habitat as opposed to a transit medium, and several laboratory studies have demonstrated primary

metabolism by atmospheric isolates, e.g. [53, 57], as well as evidence for potential metabolic activity in clouds, e.g. [58, 59]. It is important to recognise that opportunity for metabolic transformations by bacteria or fungi in air are likely to be very limited due to short residence times in air and highly heterogeneous conditions. Whether sufficient moisture, temperature, substrate availability and stress avoidance for cell homeostasis and reproduction are energetically feasible, or only quasi-dormancy within the short timeframe of favourable conditions offered during atmospheric transport remains unexplored.

Given the physicochemical and dynamic complexity of the atmosphere and the broad range of correlations we observed between taxonomic and functional diversity and abiotic factors, a chaotic system of interplay may emerge that influences atmospheric microbial ecology as envisaged for highly dispersed marine larvae [60]. Taken together we anticipate these findings will be valuable in future hypothesis-driven research both to identify interactions between surface habitats across multiple ecological scales which are mediated by the atmospheric microbiota, and to test models of recruitment, turnover, functionality and resilience. Given that the atmosphere is also a sink for a large fraction of anthropogenic emissions [51], it is timely that an accurate global inventory of microbial diversity is provided in order to present a baseline for measuring future responses to change. Finally, the study complements efforts to inventory global soil [30, 33, 54] and oceanic microbiomes [61] and expands the scope of the pan-global microbiota.

References

1. Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, et al. Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* 2019; **17**: 569–586.
2. Barberan A, Casamayor EO, Fierer N. The microbial contribution to macroecology.

Front Microbiol 2014; **5**: 1–8.

3. Womack AM, Bohannan BJM, Green JL. Biodiversity and biogeography of the atmosphere. *Philos Trans R Soc Lond B Biol Sci* 2010; **365**: 3645–3653.
4. Lighthart B. The ecology of bacteria in the alfresco atmosphere. *FEMS Microbiol Ecol* 1997; **23**: 263–274.
5. Griffin DW. Atmospheric movement of microorganisms in clouds of desert dust and implications for human health. *Clin Microbiol Rev* 2007; **20**: 459–477.
6. Deguillaume L, Leriche M, Amato P, Ariya P a., Delort AM, Pöschl U, et al. Microbiology and atmospheric processes: Chemical interactions of primary biological aerosols. *Biogeosciences* 2008; **5**: 1073–1084.
7. Fröhlich-Nowoisky J, Burrows SM, Xie Z, Engling G, Solomon PA, Fraser MP, et al. Biogeography in the air: Fungal diversity over land and oceans. *Biogeosciences* 2012; **9**: 1125–1136.
8. Tignat-Perrier R, Dommergue A, Thollot A, Keuschnig C, Magand O, Vogel TM, et al. Global airborne microbial communities controlled by surrounding landscapes and wind conditions. *Sci Rep* 2019; **9**: 1–11.
9. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014; **12**: 87.
10. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol* 2019; **27**: 105–117.
11. Šantl-Temkiv T, Gosewinkel U, Starnawski P, Lever M, Finster K. Aeolian dispersal of bacteria in southwest Greenland: Their sources, abundance, diversity and physiological states. *FEMS Microbiol Ecol* 2018; **94**: fty031.

12. Bowers RM, McLetchie S, Knight R, Fierer N. Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments. *ISME J* 2011; **5**: 601–612.
13. Tignat-Perrier R, Dommergue A, Thollot A, Magand O, Amato P, Joly M, et al. Seasonal shift in airborne microbial communities. *Sci Total Environ* 2020; **716**: 137129.
14. Bowers RM, McCubbin IB, Hallar AG, Fierer N. Seasonal variability in airborne bacterial communities at a high-elevation site. *Atmos Environ* 2012; **50**: 41–49.
15. Els N, Larose C, Baumann-Stanzer K, Tignat-Perrier R, Keuschnig C, Vogel TM, et al. Microbial composition in seasonal time series of free tropospheric air and precipitation reveals community separation. *Aerobiologia (Bologna)* 2019; **35**: 671–701.
16. Uetake J, Tobo Y, Uji Y, Hill TCJ, DeMott PJ, Kreidenweis SM, et al. Seasonal changes of airborne bacterial communities over Tokyo and influence of local meteorology. *Front Microbiol* 2019; **10**: 1572.
17. Favet J, Lapanje A, Giongo A, Kennedy S, Aung Y-Y, Cattaneo A, et al. Microbial hitchhikers on intercontinental dust: catching a lift in Chad. *ISME J* 2013; **7**: 850–867.
18. Cáliz J, Triadó-Margarit X, Camarero L, Casamayor EO. A long-term survey unveils strong seasonal patterns in the airborne microbiome coupled to general and regional atmospheric circulations. *Proc Natl Acad Sci U S A* 2018; **115**: 12229–12234.
19. DeLeon-Rodriguez N, Latham TL, Rodriguez-R LM, Barazesh JM, Anderson BE, Beyersdorf AJ, et al. Microbiome of the upper troposphere: Species composition and prevalence, effects of tropical storms, and atmospheric implications. *Proc Natl Acad Sci U S A* 2013; **110**: 2575–2580.
20. Lowe WH, McPeck MA. Is dispersal neutral? *Trends Ecol Evol* 2014; **29**: 444–450.
21. Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBHH. Beyond biogeographic

patterns: Processes shaping the microbial landscape. *Nat Rev Microbiol* 2012; **10**: 497–506.

22. Mayol E, Arrieta JM, Jiménez MA, Martínez-Asensio A, Garcias-Bonet N, Dachs J, et al. Long-range transport of airborne microbes over the global tropical and subtropical ocean. *Nat Commun* 2017; **8**: 201.

23. Archer SDJ, Lee KC, Caruso T, Maki T, Lee CK, Cary SC, et al. Airborne microbial transport limitation to isolated Antarctic soil habitats. *Nat Microbiol* 2019; **4**: 925–932.

24. Khaled A, Zhang M, Amato P, Delort A-M, Ervens B. Biodegradation by bacteria in clouds: An underestimated sink for some organics in the atmospheric multiphase system. *Atmos Chem Phys Discuss* 2020; **2020**: 1–32.

25. Klein AM, Bohannon BJM, Jaffe DA, Levin DA, Green JL. Molecular evidence for metabolically active bacteria in the atmosphere. *Front Microbiol* 2016; **7**: 1–11.

26. Amato P, Besaury L, Joly M, Penaud B, Deguillaume L, Delort A-M. Metatranscriptomic exploration of microbial functioning in clouds. *Sci Rep* 2019; **9**: 4383.

27. Stull RB. A boundary layer definition. *Introduction to Boundary Layer Meteorology*. 1988. Kluwer Academic Publishers, Dordrecht, pp 1–27.

28. Warren-Rhodes K, Lee K, Archer S, Cabrol N, Ng-Boyle L, Wettergreen D, et al. Subsurface microbial habitats in an extreme desert Mars-analogue environment. *Front Microbiol* 2019; **10**: 10.3389/fmicb.2019.00069.

29. Rausch P, Rühlemann M, Hermes BM, Doms S, Dagan T, Dierking K, et al. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome* 2019; **7**: 133.

30. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, et al. A global atlas of the dominant bacteria found in soil. *Science* (80-

2018; **359**: 320–325.

31. Cimini G, Squartini T, Saracco F, Garlaschelli D, Gabrielli A, Caldarelli G. The statistical physics of real-world networks. *Nat Rev Phys* 2019; **1**: 58–71.
32. Ulrich W, Almeida-Neto M, Gotelli NJ. A consumer's guide to nestedness analysis. *Oikos* 2009; **118**: 3–17.
33. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017; **551**: 457–463.
34. Vellund M. Conceptual synthesis in community ecology. *Q Rev Biol* 2010; **385**: 183–206.
35. Lymperopoulou DS, Adams RI, Lindow SE. Contribution of vegetation to the microbial composition of nearby outdoor air. *Appl Environ Microbiol* 2016; **82**: 3822–3833.
36. Michaud JM, Thompson LR, Kaul D, Espinoza JL, Richter RA, Xu ZZ, et al. Taxon-specific aerosolization of bacteria and viruses in an experimental ocean-atmosphere mesocosm. *Nat Commun* 2018; **9**: 2017.
37. Zhao Y, Aarnink A, De Jong M, Groot Koerkamp PWG (Peter). Airborne Microorganisms From Livestock Production Systems and Their Relation to Dust. *Crit Rev Environ Sci Technol* 2014; **44**: 1071–1128.
38. Hawksworth DL. The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycol Res* 2001; **105**: 1422–1432.
39. Gusareva ES, Acerbi E, Lau KJX, Luhung I, Premkrishnan BN V, Kolundžija S, et al. Microbial communities in the tropical air ecosystem follow a precise diel cycle. *Proc Natl Acad Sci U S A* 2019; **116**: 23299–23308.
40. Bryan NC, Christner BC, Guzik TG, Granger DJ, Stewart MF. Abundance and

- survival of microbial aerosols in the troposphere and stratosphere. *ISME J* 2019; **13**: 2789–2799.
41. Vorholt JA. Microbial life in the phyllosphere. *Nat Rev Microbiol* 2012; **10**: 828–840.
42. Kellogg CA, Griffin DW. Aerobiology and the global transport of desert dust. *Trends Ecol Evol* 2006; **21**: 638–644.
43. Pointing SBSB, Belnap J. Microbial colonization and controls in dryland systems. *Nat Rev Microbiol* 2012; **10**: 551–562.
44. Shenhav L, Thompson M, Joseph TA, Briscoe L, Furman O, Bogumil D, et al. FEAST: fast expectation-maximization for microbial source tracking. *Nat Methods* 2019; **16**: 627–632.
45. Aalismail NA, Ngugi DK, Díaz-Rúa R, Alam I, Cusack M, Duarte CM. Functional metagenomic analysis of dust-associated microbiomes above the Red Sea. *Sci Rep* 2019; **9**: 1–12.
46. Reche I, D’Orta G, Mladenov N, Winget DM, Suttle CA, Gaetano ●, et al. Deposition rates of viruses and bacteria above the atmospheric boundary layer. *ISME J* 2018; **12**: 1154–1162.
47. Fröhlich-Nowoisky J, Kampf CJ, Weber B, Huffman JA, Pöhlker C, Andreae MO, et al. Bioaerosols in the Earth system: Climate, health, and ecosystem interactions. *Atmos Res* . 2016. Elsevier. , **182**: 346–376
48. Després VR, Alex Huffman J, Burrows SM, Hoose C, Safatov AS, Buryak G, et al. Primary biological aerosol particles in the atmosphere: A review. *Tellus, Ser B Chem Phys Meteorol* . 2012. , **64**: doi: 10.3402/tellusb.v64i0.15598
49. Chan Y, Van Nostrand JD, Zhou J, Pointing SB, Farrell RL. Functional ecology of an Antarctic Dry Valley. *Proc Natl Acad Sci U S A* 2013; **110**: 8990–8995.
50. Hartery S, Toohey D, Revell L, Sellegri K, Kuma P, Harvey M, et al. Constraining the

- Surface Flux of Sea Spray Particles From the Southern Ocean. *J Geophys Res Atmos* 2020; **125**: e2019JD032026.
51. Archer SDJJ, Pointing SB. Anthropogenic impact on the atmospheric microbiome. *Nat Microbiol* 2020; **5**: 229–231.
 52. Matulová M, Husárová S, Capek P, Sancelme M, Delort AM. Biotransformation of various saccharides and production of exopolymeric substances by cloud-borne *Bacillus* sp. 3B6. *Environ Sci Technol* 2014; **48**: 14238–14247.
 53. Šantl-Temkiv T, Finster K, Hansen BM, Pašić L, Karlson UG. Viable methanotrophic bacteria enriched from air and rain can oxidize methane at cloud-like conditions. *Aerobiologia (Bologna)* 2013; **29**: 373–384.
 54. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. *Nature* 2018; **560**: 233–237.
 55. Lebre PH, De Maayer P, Cowan DA. Xerotolerant bacteria: surviving through a dry spell. *Nat Rev Microbiol* 2017; **15**: 285–296.
 56. Morris CE, Conen F, Alex Huffman J, Phillips V, Pöschl U, Sands DC. Bioprecipitation: a feedback cycle linking Earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere. *Glob Chang Biol* 2014; **20**: 341–351.
 57. Amato P, Parazols M, Sancelme M, Laj P, Mailhot G, Delort AM. Microorganisms isolated from the water phase of tropospheric clouds at the Puy de Dôme: Major groups and growth abilities at low temperatures. *FEMS Microbiol Ecol* 2007; **59**: 242–254.
 58. Bianco A, Deguillaume L, Chaumerliac N, Vařtilingom M, Wang M, Delort A-M, et al. Effect of endogenous microbiota on the molecular composition of cloud water: a

study by Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR MS).
Sci Rep 2019; **9**: 7663.

59. Vaitilingom M, Deguillaume L, Vinatier V, Sancelme M, Amato P, Chaumerliac N, et al. Potential impact of microbial activity on the oxidant capacity and organic carbon budget in clouds. *Proc Natl Acad Sci* 2013; **110**: 559–564.

60. Álvarez-Noriega M, Burgess SC, Byers JE, Pringle JM, Wares JP, Marshall DJ. Global biogeography of marine dispersal potential. *Nat Ecol Evol* 2020; **4**: 1196–1203.

61. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. *Science* (80-) 2015; **348**: 1–10.

62. Pointing SBSB, Belnap J. Disturbance to desert soil ecosystems contributes to dust-mediated impacts at regional scales. *Biodivers Conserv* 2014; **23**: 1659–1667.

63. Archer SDJ, Lee KC, Caruso T, King-Miaow K, Harvey M, Huang D, et al. Air mass source determines airborne microbial diversity at the ocean–atmosphere interface of the Great Barrier Reef marine ecosystem. *ISME J* 2020; **14**: 871–876.

64. Dybwad M, Skogan G, Blatny JM. Comparative testing and evaluation of nine different air samplers: End-to-end sampling efficiencies as specific performance measurements for bioaerosol applications. *Aerosol Sci Technol* 2014; **48**: 282–295.

65. University of East Anglia Climatic Research Unit; Harris, I.C.; Jones, P.D.; Osborn T. No Title. *CRU TS4.04: Climatic Research Unit (CRU) Time-Series (TS) version 4.04 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901-Dec. 2019)*. .

66. Beck HE, Zimmermann NE, McVicar TR, Vergopolan N, Berg A, Wood EF. Present and future köppen-geiger climate classification maps at 1-km resolution. *Sci Data* 2018; **5**: 1–12.

67. Fischer G. World Food and Agriculture to 2030/50: How do climate change and bioenergy alter the long-term outlook for food, agriculture and resource availability? *Expert Meet. How to Feed World 2050*. 2009. pp 1–49.
68. Stein AF, Draxler RR, Rolph GD, Stunder BJB, Cohen MD, Ngan F. NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System. *Bull Am Meteorol Soc* 2015; **96**: 2059–2077.
69. Burrows SM, Butler T, Jöckel P, Tost H, Kerkweg A, Pöschl U, et al. Bacteria in the global atmosphere - Part 2: Modeling of emissions and transport between different ecosystems. *Atmos Chem Phys* 2009; **9**: 9281–9297.
70. Luhung I, Uchida A, Lim SBY, Gaultier NE, Kee C, Lau KJX, et al. Experimental parameters defining ultra-low biomass bioaerosol analysis. *npj Biofilms Microbiomes* 2021; **7**: 37.
71. Hospodsky D, Yamamoto N, Peccia J. Accuracy, precision, and method detection limits of quantitative PCR for airborne bacteria and fungi. *Appl Environ Microbiol* 2010; **76**: 7004–7012.
72. Herlemann DPR, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J* 2011; **5**: 1571–1579.
73. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 2013; **41**: e1.
74. Liu CM, Kachur S, Dwan MG, Abraham AG, Aziz M, Hsueh PR, et al. FungiQuant: a broad-coverage fungal quantitative real-time PCR assay. *BMC Microbiol* 2012; **12**: 255.
75. Gardes M, Bruns TD. ITS primers with enhanced specificity for basidiomycetes -

application to the identification of mycorrhizae and rusts. *Mol Ecol* 1993; **2**: 113–118.

76. White T, Burns T, Lee S, Taylor J. No Title. In: Innis M, Gelfand D, Sninsky J, White T (eds). *PCR protocols: a guide to methods and applications*. 1990. Academic Press, New York, USA, pp 315–322.

77. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 2011; **17**: 10.

78. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 2017; **11**: 2639–2643.

79. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, A AJ. DADA2: High resolution sample inference from Illumina amplicon data. *Nat Methods* 2016; **13**: 581–583.

80. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* 2013; **41**: 590–596.

81. Nilsson RH, Larsson K-H, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 2019; **47**: D259–D264.

82. Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, et al. Controlling for contaminants in low biomass 16S rRNA gene sequencing experiments. *mSystems* 2019; **4**: e00290-19.

83. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 2018; **6**: 226.

84. Watts SC, Ritchie SC, Inouye M, Holt KE. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics* 2019; **35**: 1064–1066.

901 85. Warnes G, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots:
902 Various R programming tools for plotting data. *R package version* . 2005.

903 86. Rinke C, Low S, Woodcroft BJ, Raina J-B, Skarshewski A, Le XH, et al. Validation of
904 picogram- and femtogram-input DNA libraries for microscale metagenomics. *PeerJ*
905 2016; **4**: e2486.

906 87. Uritskiy G V, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-
907 resolved metagenomic data analysis. *Microbiome* 2018; **6**: 158.

908 88. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2.
909 *Genome Biol* 2019; **20**: 257.

910 89. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance
911 in metagenomics data. *PeerJ Comput Sci* 2017; **3**: e104.

912 90. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al.
913 Species-level functional profiling of metagenomes and metatranscriptomes. *Nat*
914 *Methods* 2018; **15**: 962–968.

915 91. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive
916 Analysis and Graphics of Microbiome Census Data. *PLoS One* 2013; **8**: e61217.

917 92. Wickham H. ggplot2. 2009. Springer New York, New York, NY.

918 93. Oksanen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, et al.
919 R package: Vegan. 2017.

920 94. Lin H, Peddada S Das. Analysis of compositions of microbiomes with bias correction.
921 *Nat Commun* 2020; **11**: 3514.

922 95. Hijmans RJ. R package: Geosphere: spherical trigonometry. 2019. R package.

923 96. Ueki T, Fujie M, Romaidi, Satoh N. Symbiotic bacteria associated with ascidian
924 vanadium accumulation identified by 16S rRNA amplicon sequencing. *Mar Genomics*
925 2019; **43**: 33–42.

97. Purahong W, Orrù L, Donati I, Perpetuini G, Cellini A, Lamontanara A, et al. Plant Microbiome and Its Link to Plant Health: Host Species, Organs and *Pseudomonas syringae* pv. *actinidiae* Infection Shaping Bacterial Phyllosphere Communities of Kiwifruit Plants. *Front Plant Sci* 2018; **9**: 1563.
98. Hermans SM, Buckley HL, Case BS, Lear G. Connecting through space and time: catchment-scale distributions of bacteria in soil, stream water and sediment. *Environ Microbiol* 2020; **22**: 1000–1010.
99. Yang J, Wang Y, Cui X, Xue K, Zhang Y, Yu Z. Habitat filtering shapes the differential structure of microbial communities in the Xilingol grassland. *Sci Rep* 2019; **9**: 19326.
100. Wang G, Bei S, Li J, Bao X, Zhang J, Schultz PA, et al. Soil microbial legacy drives crop diversity advantage: Linking ecological plant–soil feedback with agricultural intercropping. *J Appl Ecol* 2021.
101. Carini P, Delgado-Baquerizo M, Hinckley E-LS, Holland-Moritz H, Brewer TE, Rue G, et al. Effects of Spatial Variability and Relic DNA Removal on the Detection of Temporal Dynamics in Soil Microbial Communities. *MBio* 2020; **11**: 10.1128/mBio.02776-19.
102. Szoboszlay M, Näther A, Mullins E, Tebbe CC. Annual replication is essential in evaluating the response of the soil microbiome to the genetic modification of maize in different biogeographical regions. *PLoS One* 2019; **14**: e0222737.
103. Wei T, Simco V. R package: Corrplot: Visualisation of a correlation matrix. 2017. R package.
104. Krijthe JE. Rtsne: t-distributed stochastic neighbor embedding using a Barnes-Hut implementation. 2015. R package.
105. Anderson MJ. Permutational Multivariate Analysis of Variance (PERMANOVA).

951 *Wiley StatsRef: Statistics Reference Online*. 2017. American Cancer Society, pp 1–15.

- 952 106. Legendre P, Borcard D, Peres-Neto PR. Analyzing beta diversity: Partitioning the
953 spatial variation of community composition data. *Ecol Monogr* 2005; **75**: 435–450.
- 954 107. Squartini T, Garlaschelli D. Maximum-entropy networks: Pattern detection, network
955 reconstruction and graph combinatorics. 2017. Springer.
- 956 108. Squartini T, Garlaschelli D. Analytical maximum-likelihood method to detect patterns
957 in real networks. *New J Phys* 2011; **13**: 83001.
- 958 109. Artzy-Randrup Y, Stone L. Generating uniformly distributed random networks. *Phys*
959 *Rev E* 2005; **72**: 56708.
- 960 110. Roberts ES, Coolen ACC. Unbiased degree-preserving randomization of directed
961 binary networks. *Phys Rev E* 2012; **85**: 46103.
- 962 111. Rossana. MATLAB package: MAX&SAM. 2021. MATLAB.
- 963 112. Dormann CF, Frund J, Bluthgen N, Gruber B. Indices, Graphs and Null Models:
964 Analyzing Bipartite Ecological Networks. *Open Ecol J* 2009; **2**: 7–24.
- 965 113. Gotelli NJ, Ellison AM. A primer of ecological statistics, second edi. 2012. Oxford
966 University Press, Oxford.
- 967 114. Almeida-Neto M, Guimarães P, Guimarães PR, Loyola RD, Ulrich W. A consistent
968 metric for nestedness analysis in ecological systems: reconciling concept and
969 measurement. *Oikos* 2008; **117**: 1227–1239.

970

971 **Acknowledgements**

972 The research was funded by the Singapore Ministry of Education and Yale-NUS College,
973 grant number R-607-265-331-121.

974

975 **Author contributions**

976 S.B.P. designed the study, secured funding and led the research; S.D.J.A was field team
977 leader and contributed to experimental design; S.D.J.A., K.C.L., T.M., S.B.P. and K.A.W-R
978 conducted sampling; A.A., D.A.M., J.G.A., S.C.C., C.E., S.H., M.I., M.K.K., C.K.L.,
979 C.J.N.L., J.B.R., A.R., T.S., W.S., H.S., B.V. and C.W. assisted with fieldwork; S.C.C.,
980 D.A.C., J.D., B.G., B.G-S., M.H., K.H., I.H., M.L., C.P.M., S.N. and N.T. facilitated access
981 to remote field locations; T.L., Y.H.P. and J.H.S. provided scientific assistance for laboratory
982 and field experiments; S.D.J.A. and G.H. collected metadata; S.D.J.A. performed sample
983 processing and laboratory experiments; M.H.Y.L., K.C.L., S.J.S. and X.T. performed
984 bioinformatics analysis; T.C. and K.C.L. performed ecological analysis and modelling; G.H.
985 performed geospatial analysis and modelling; P.K.H.L. and S.B.P. supervised data analysis;
986 S.D.J.A., T.C., B.G-S., K.C.L., P.K.H.L., M.H.Y.L., K.D.H., T.M., S.B.P., S.J.S., T.S-T.,
987 X.T. and K.A.W-R. interpreted the findings; S.B.P. led the data interpretation and wrote the
988 paper.

989

990 **Data availability**

991 Sequencing data are accessible in the NCBI Sequence Read Archive (SRA) under BioProject
992 numbers PRJEB42754 (<https://www.ncbi.nlm.nih.gov/search/all/?term=PRJEB42754>) for
993 amplicon sequencing and PRJNA694999
994 (<https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA694999>) for metagenomes.

995

996 **Competing Interests**

997 The authors declare no competing interests.

998

999 **Additional Information**

1000 **Supplementary Information** is available for this paper.

Correspondence and requests for materials should be addressed to S.B.P., e-mail:

stephen.pointing@yale-nus.edu.sg

Figure Legends

Fig 1 | A globally distributed survey of microbial communities in the atmospheric

boundary layer, free troposphere and underlying soil. a) Locations are indicated by green

boxes where: 1, Canada; 2, Mongolia; 3, Spain; 4, Japan; 5, California, USA; 6, Kuwait; 7,

Hilo, Hawaii, USA; 8; Mauna Kea, Hawaii, USA; 9, Singapore; 10, Coral Sea; 11, Namibia;

12, Chile; 13, South Africa; 14, Australia; 15, Uruguay; 16, New Zealand; 17, Southern

Ocean; 18, Antarctica. Meta-data for each location are shown in Supplementary Information.

Back trajectories are shown for near-ground air (blue lines) and high-altitude air (red lines),

The survey comprised 596 biologically independent replicates. **b)** Inter-domain abundance of

reads for metagenomes ($n=120$). Other eukaryotes included all microbial eukaryotes, viruses

were not a specific target of our study and so this likely includes only cell-associated viruses,

HA air denotes high-altitude air; NG air denotes near-ground air.

Fig. 2 | Bacterial and fungal assemblages are taxonomically structured. a) Relative

abundance of bacterial classes in soil ($n = 79$), near-ground air (NG air) ($n = 437$) and high-

altitude air (HA air) ($n = 13$). **b)** Alpha diversity metrics for bacteria. **c)** Bacterial assemblage

dissimilarity (Jaccard Index) by location (3-D visualizations are presented in Supplementary

Information). **d)** Modelled nestedness estimates for bacteria were based upon networks

constructed for each habitat and location (Fig. S22). **e)** Relative abundance of fungal classes

in soil ($n = 79$), near-ground air (NG air) ($n = 437$) and high-altitude air (HA air) ($n = 13$). **f)**

Alpha diversity metrics for fungi. **g)** Fungal assemblage dissimilarity (Jaccard Index) by

location (3-D visualizations are presented in Fig. S19). **h)** Modelled nestedness estimates for fungi were based upon networks constructed for each habitat and location (Fig. S22).

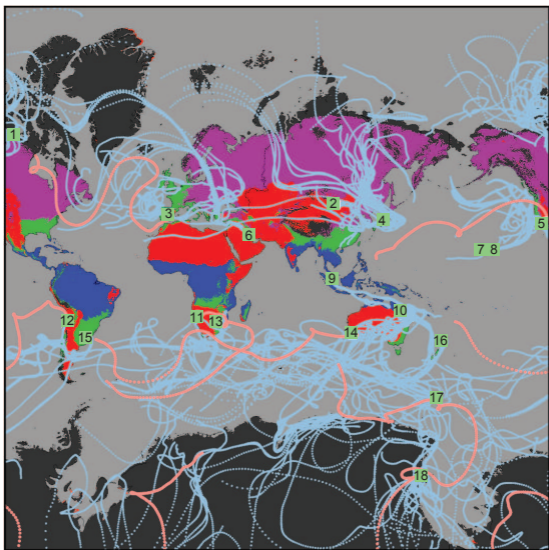
Fig. 3 | Bacterial and fungal assemblages are recruited from local and global sources. a) Shared bacterial and fungal taxa among habitat types within and between locations. **b)** Source contribution by climate for observed bacterial and fungal diversity in air, values averaged for each source to mitigate sample size effects (Bacteria $n=529$, Fungi $n=444$). Locations: 1, Canada; 2, Mongolia; 3, Spain; 4, Japan; 5, California, USA; 6, Kuwait; 7, Hilo, Hawaii, USA; 8, Mauna Kea, Hawaii, USA; 9, Singapore; 10, Coral Sea; 11, Namibia; 12, Chile; 13, South Africa; 14, Australia; 15, Uruguay; 16, New Zealand; 17, Southern Ocean; 18, Antarctica. HA air denotes high-altitude air; NG air denotes near-ground air.

Fig. 4 | Assemblages possessed stress response and metabolic genes relevant to survival in the atmosphere. a) Heatmap summarising functional metagenomic profiling of targeted stress-response and metabolic genes by habitat type ($n = 120$). Full data for all replicates at each sample-location combination is shown in Fig. S26. **b)** Distribution of stress-response and metabolic genes by climatic region, with all locations globally pooled by climate ($n = 120$). Oxid. Stress denotes oxidative stress; Trace gas met. denotes trace gas metabolism. HA air, high-altitude air; NG air, near-ground air.

Fig. 5 | Atmospheric taxonomic and functional diversity are correlated with macroclimate and atmospheric transit abiotic variables. a) Correlation of location-specific macroclimate factors and abiotic stressors during transit (Supplementary Information) with biotic traits of atmospheric assemblages. **b)** Correlation of location-

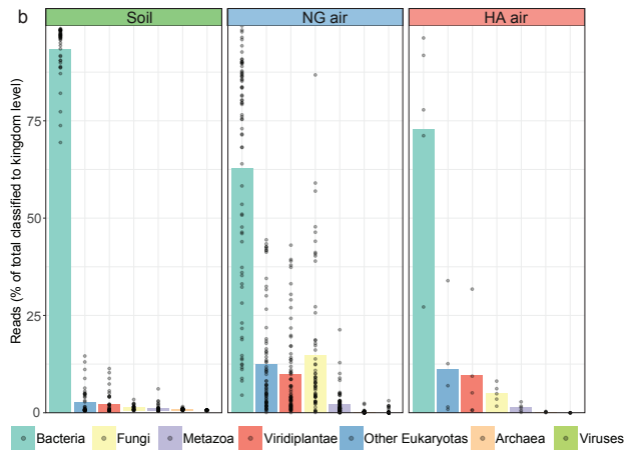
1049 specific macroclimate factors with biotic traits of soil assemblages. Blue circles denote
1050 positive correlations and red circles denote negative correlations. Circle colour intensity and
1051 size denote magnitude of correlation. Correlations were significant at $P = <0.05$. Asterisks
1052 denote correlations that were significant after Bonferroni Correction (single asterisk $P = 0.05$,
1053 double asterisk $P = 0.003$). MAT, mean annual temperature; MAP, mean annual
1054 precipitation; RH, relative humidity; UV, ultraviolet radiation. Abundance, qPCR estimated
1055 gene copy number; Richness, Chao1 estimation from rRNA gene diversity.

a

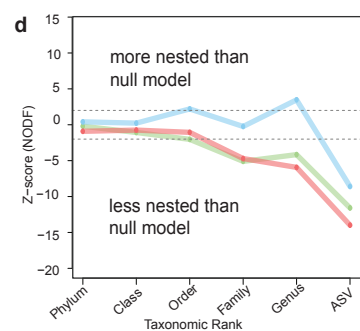
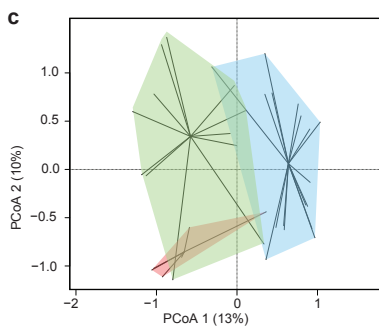
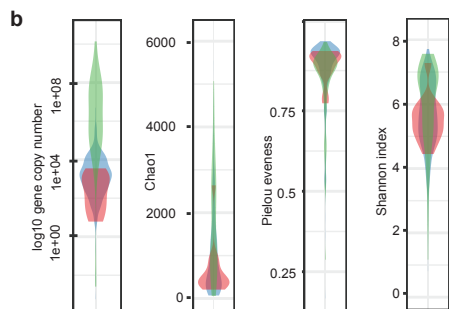
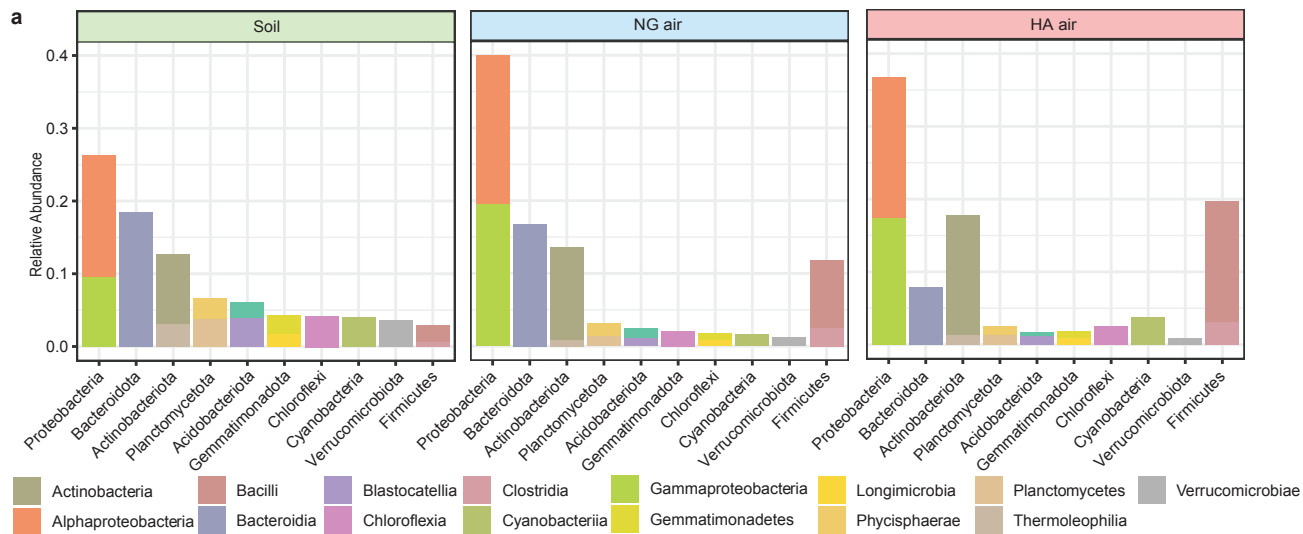


Ocean
 Polar/alpine
 Dry
 Continental
 Temperate
 Tropical

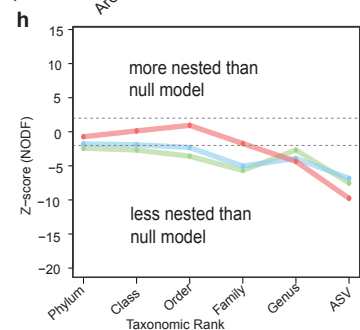
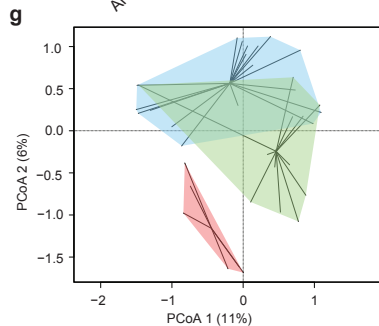
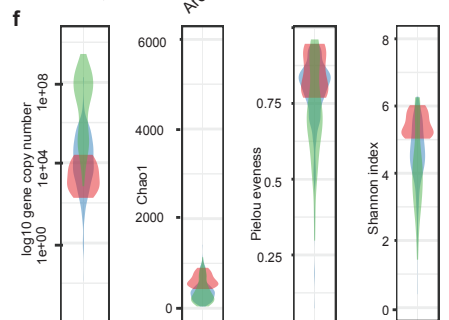
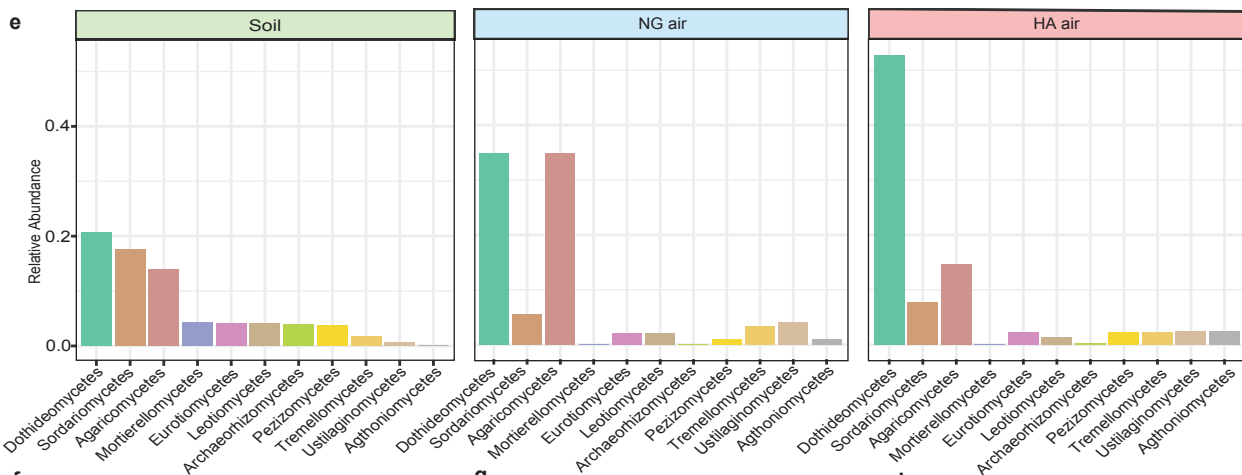
b



Bacteria

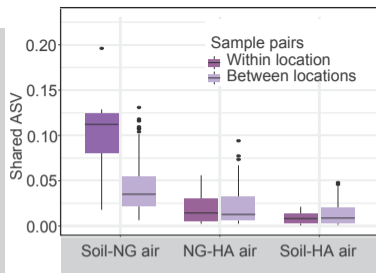


Fungi

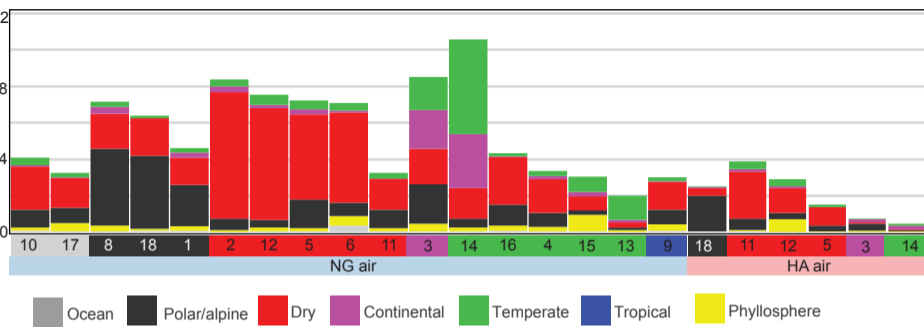


a

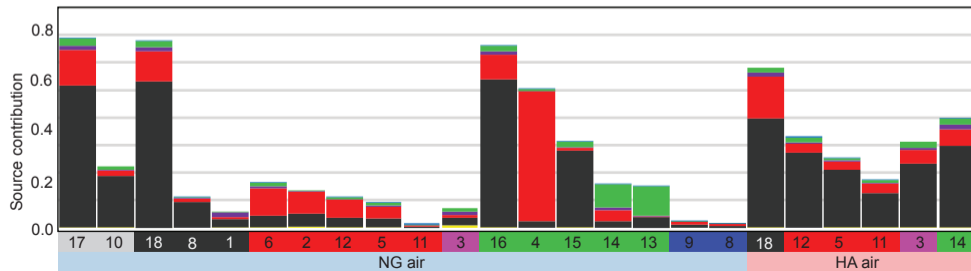
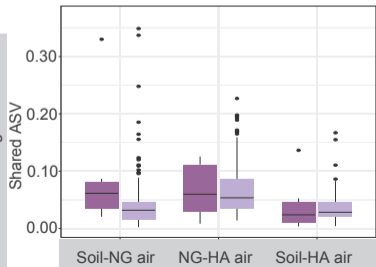
Bacteria

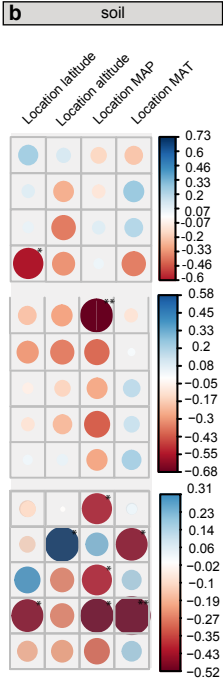
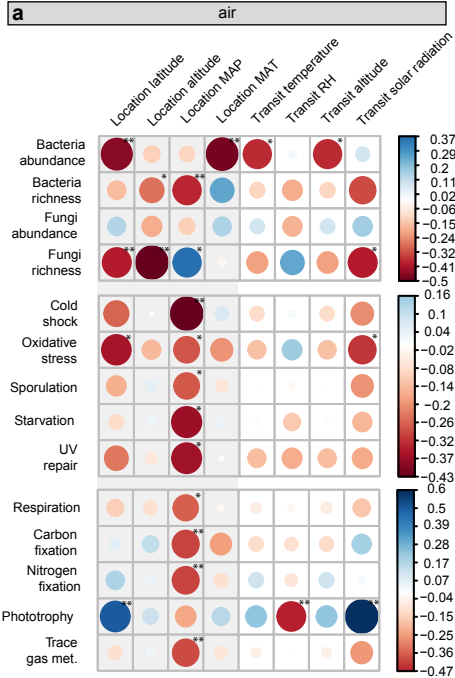
**b**

Source contribution



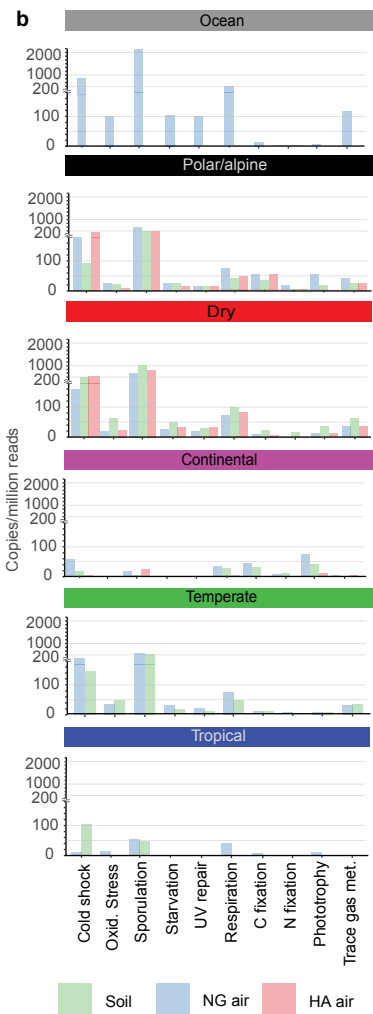
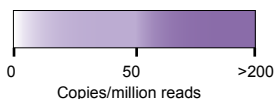
Fungi





a

| | | Soil | NG air | HA air |
|----------------------|-------------------|------|--------|--------|
| Cold shock | <i>cspA/B/G/I</i> | | | |
| Oxidative stress | <i>msrP</i> | | | |
| | <i>msrQ</i> | | | |
| Sporulation | <i>spo0A</i> | | | |
| Starvation | <i>dps</i> | | | |
| | <i>slp</i> | | | |
| UV repair | <i>phrB</i> | | | |
| | <i>uvrB</i> | | | |
| | <i>uvrC</i> | | | |
| Respiration | <i>atpA</i> | | | |
| | <i>ccoN/coxA</i> | | | |
| | <i>cydA</i> | | | |
| | <i>cyoA</i> | | | |
| | <i>nuoF</i> | | | |
| | <i>sdhA/frdA</i> | | | |
| Carbon fixation | <i>rbcL</i> | | | |
| Nitrogen fixation | <i>nifH</i> | | | |
| Phototrophy | <i>psaA</i> | | | |
| | <i>psbA</i> | | | |
| | <i>RHO</i> | | | |
| Trace gas metabolism | <i>coxL</i> | | | |
| | <i>Fe/FeFe</i> | | | |
| | <i>isoA</i> | | | |
| | <i>isoA/mmoX</i> | | | |
| | <i>pmoA</i> | | | |



Diverse recruitment to a taxonomically structured global atmospheric microbiota

Supplementary Information

Stephen D.J. Archer, Kevin C. Lee, Tancredi Caruso, Marcus H.Y. Leung, Xinzhao Tong, Susannah J. Salter, Graham Hinchliffe, Teruya Maki, Tina Santl-Temkiv, Kimberley A. Warren-Rhodes, Benito Gomez-Silva, Kevin D. Hyde, Celine J.N. Liu, Antonio Alcamí, Dina M. Al Mailem, Jonathan G. Araya, S. Craig Cary, Don A. Cowan, Jessica Dempsey, Claudia Etchebehere, Batdelger Gantsetseg, Sean Hartery, Mike Harvey, Kazuichi Hayakawa, Ian Hogg, Mutsoe Inoue, Mayada K. Kansour, Timothy Lawrence, Charles K. Lee, Matthias Leopold, Christopher P. McKay, Seiya Nagao, Yan Hong Poh, Jean-Baptiste Ramond, Alberto Rastrojo, Toshio Sekiguchi, Joo Huang Sim, William Stahm, Henry J. Sun, Ning Tang, Bryan Vandenbrink, Craig Walther, Patrick K.H. Lee, Stephen B. Pointing*

*e-mail: S.B.P. stephen.pointing@yale-nus.edu.sg

Contents

Field sampling (page 4)

Table S1 | Globally distributed sampling locations included in the study (page 4)

Fig. S1 | Mean transit altitudes for sampled air (page 5)

Fig. S2 | Atmospheric variables encountered by microorganisms during transit to each sampling location (page 6)

Sampling effort for amplicon and metagenome sequencing (page 7)

Fig S3 | Rarefaction curves for amplicon sequencing (page 7)

Fig S4 | Rarefaction curves for metagenomes (page 8)

Decontamination of environmental sequence data (page 9)

Fig S5 | Summary of steps taken to minimise contamination during environmental DNA recovery and sequencing from low biomass air and soil samples (page 9)

Fig. S6 | Sequencing read depth for controls and environmental samples (page 10)

Table S2 | Total percentage of sequenced reads removed from each step of the decontamination process. (page 11)

Fig S7 | Taxonomic identity of reads removed during the decontamination process (page 12)

Fig. S8 | Post-decontamination check for residual contaminants (page 13)

Fig. S9 | Taxonomic composition of potential residual contaminants (page 14)

Fig S10 | Evaluation for batch effects and temporal auto-correlation in samples (page 16)

Fig. S11 | Taxonomic identity of reads removed during the decontamination process by location (page 17)

Fig. S12 | Comparison of bacterial diversity estimation pre- and post-decontamination (page 18)

Fig. S13 | Comparison of fungal diversity estimation pre- and post-decontamination (page 18)

Fig. S14 | Comparison of community resemblance pre- and post-decontamination (page 19)

Supplementary discussion on use of Null models in the study (page 20)

Fig. S15 | Jaccard network null model for bacterial ASV (page 21)

Fig. S16 | Jaccard network null model for fungal ASV (page 22)

Inventory of taxa (page 22)

Table S3 | List of bacterial genera prevalent in $\geq 50\%$ of soil and/or air samples (page 23)

Table S4 | List of fungal genera prevalent in $\geq 50\%$ of soil and/or air samples (page 24)

Metagenomics functional gene targets (page 25)

Table S5 | Functional genes targeted in the metagenomics inquiry of air and soil (page 25)

Supplementary data analysis (page 26)

Fig. S17 | Global patterns in alpha diversity for bacteria and fungi in air and soil (page 26)

Fig. S18 | Comparison of diversity estimation using shotgun metagenomics and amplicon sequencing of air and soil (page 27)

Fig. S19 | Dissimilarity of assemblages by habitat type and location (page 28)

Fig. S20 | Phylogenetic distance of assemblages by habitat type and location (page 29)

Fig. S21 | Global-scale distance decay plots for bacteria and fungi in air and soil (page 30)

Fig. S22 | Null model networks for bacterial and fungal assemblages in air and soil (page 31)

Fig. S23 | Taxonomic composition of bacterial and fungal assemblages in air and underlying soil (page 32)

Fig. S24 | Co-occurrence of bacterial and fungal taxa in globally distributed air and soil. Fig. S25 | Source contribution to observed diversity in air (page 33)

Fig. S26 | Functional metagenomics profiling of targeted stress-response and metabolic genes by habitat type (page 35)

Supplementary references (page 36)

Field sampling

Table S1 | Globally distributed sampling locations included in the study. Locations are ordered by latitude from north to south. Full metadata for each sampling location and event are available upon request at <http://atmospheric-microbiome.com/>.

| Location No | GPS (decimal degrees) | Atmospheric cell [^] | Climate ⁺ | Altitude (m AMSL) | Country/Ocean | Surface cover |
|-------------|-----------------------|-------------------------------|----------------------|-------------------|-------------------------|--|
| 1 | 69.131, -105.057 | N Polar | Polar (EF) | 50 | Canada | Arctic tundra |
| 2 | 44.573, 105.648 | N Ferrel | Dry (BWk) | 1,235 | Mongolia | Steppe |
| 3 | 40.825, -3.961 | N Ferrel | Continental (Dsb) | 1,830 | Spain | Grassland, Sierra de Guadarrama National Park* |
| 4 | 37.308, 137.232 | N Ferrel | Temperate (Cfa) | 6 | Japan | Coastal forest |
| 5 | 35.142, -116.104 | N Ferrel | Dry (BWk) | 300 | USA, California | Mojave Desert |
| 6 | 28.951, 48.192 | N Hadley | Dry (BWh) | 0 | Kuwait | Kuwait Desert |
| 7 | 19.703, -155.090 | N Hadley | Tropical (Af) | 120 | USA, Hawaii (Hilo) | Coastal forest |
| 8 | 19.823, -155.478 | N Hadley | Polar (ET) | 4,200 | USA, Hawaii (Mauna Kea) | Mountain |
| 9 | 1.306, 103.772 | Equatorial | Tropical (Af) | 19 | Singapore | Tropical forest, suburban |
| 10 | -20.827, 153.067 | S Hadley | Tropical ocean | 10 | Coral Sea | Ocean |
| 11 | -23.603, 15.038 | S Hadley | Dry (BWh) | 380 | Namibia | Namib Desert |
| 12 | -24.105, -70.016 | S Hadley | Dry (BWh) | 90 | Chile | Atacama Desert |
| 13 | -25.753, -28.258 | S Hadley | Temperate (Cwb) | 1,380 | South Africa | Livestock and arable farmland |
| 14 | -32.898, 116.906 | S Ferrel | Temperate (Csa) | 320 | Australia | Grassland and arable |
| 15 | -34.354, -57.235 | S Ferrel | Temperate (Cfa) | 10 | Uruguay | Rural wooded grassland |
| 16 | -36.916, 174.646 | S Ferrel | Temperate (Cfb) | 61 | New Zealand | Wooded suburban |
| 17 | -60.288, 171.424 | S Polar | Polar ocean | 10 | Southern Ocean | Ocean |
| 18 | -75.520, 163.949 | S Polar | Polar (EF) | 98 | Antarctica | Polar desert |

[^] Atmospheric cells describe the prevailing air movement in latitude-defined cells: Polar Cells occur at 60° and higher latitudes, Ferrel Cells at mid-latitudes between 30-60°, and Hadley Cells at latitudes of 30° and below.

⁺ Climate codes follow the Köppen climate classification, major delineations were: A, tropical; B, dry; C, temperate; D, continental, E, polar[1].

*This location was affected by a Sahara Desert atmospheric dust intrusion during sampling.

Fig. S1 | Mean transit altitudes for sampled air. Blue denotes ground sampling, red denotes aircraft samples, bar indicates altitude of sampling location. Values were estimated based upon the NOAA HYSPLIT model. Locations: 1, Canada; 2, Mongolia; 3, Spain; 4, Japan; 5, California, USA; 6, Kuwait; 7, Hilo, Hawaii, USA; 8; Mauna Kea, Hawaii, USA; 9, Singapore; 10, Coral Sea; 11, Namibia; 12, Chile; 13, South Africa; 14, Australia; 15, Uruguay; 16, New Zealand; 17, Southern Ocean; 18, Antarctica. HA air denotes high-altitude air; NG air denotes near-ground air.

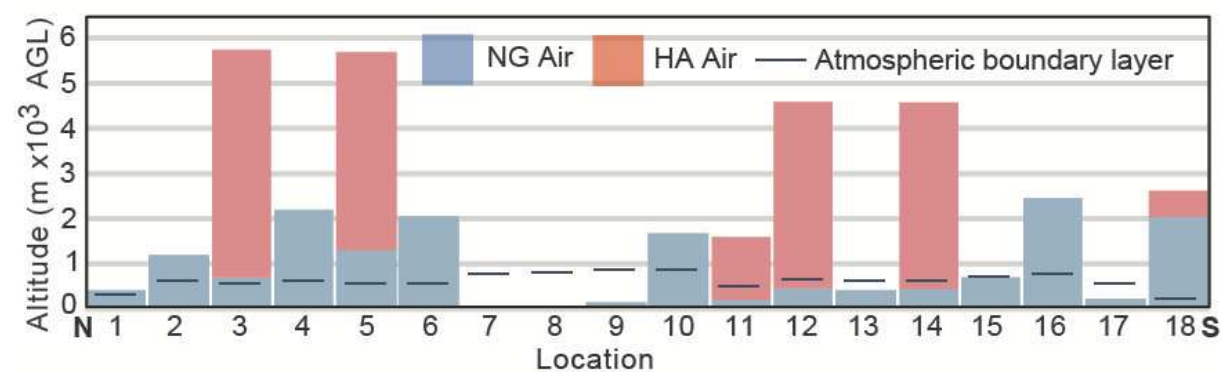
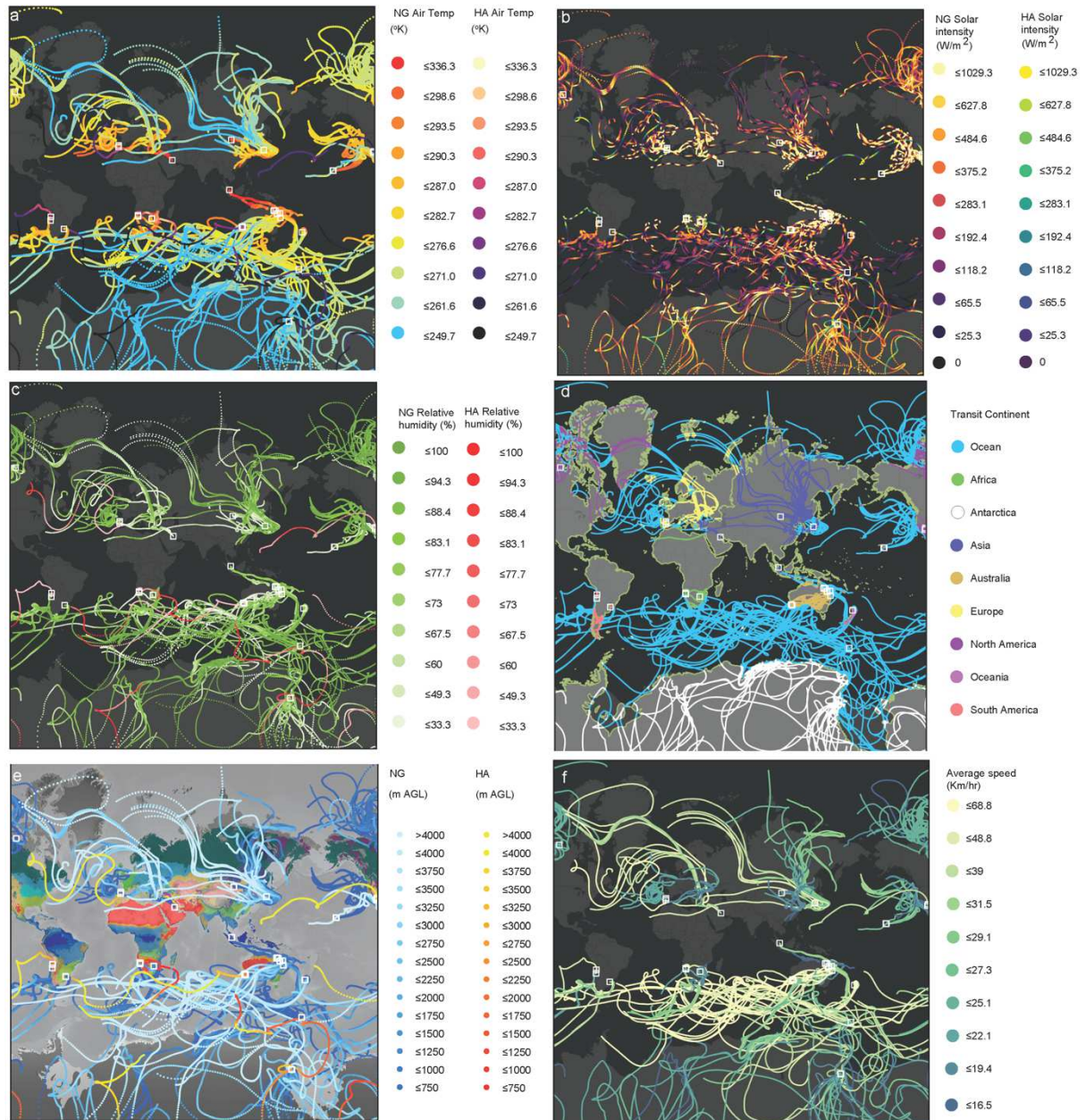


Fig. S2 | Atmospheric variables encountered by microorganisms during transit to each sampling location.
a, Temperature. **b**, Solar flux. **c**, Relative humidity. **d**, Transit over land/ocean. **e**, Transit altitude. **f**, Transit velocity. Data was obtained from the NOAA HYSPLIT-model, and long-range trajectories and abiotic variables were estimated using the GDAS database.



Sampling effort for amplicon and metagenome sequencing

Fig S3 | Rarefaction curves for amplicon sequencing. a, Bacteria (soil ($n = 79$), near-ground air (NG air) ($n = 437$) and high-altitude air (HA air)). **b,** Fungi (soil ($n = 70$), near-ground air (NG air, $n = 363$) and high-altitude air (HA air, $n = 11$)).

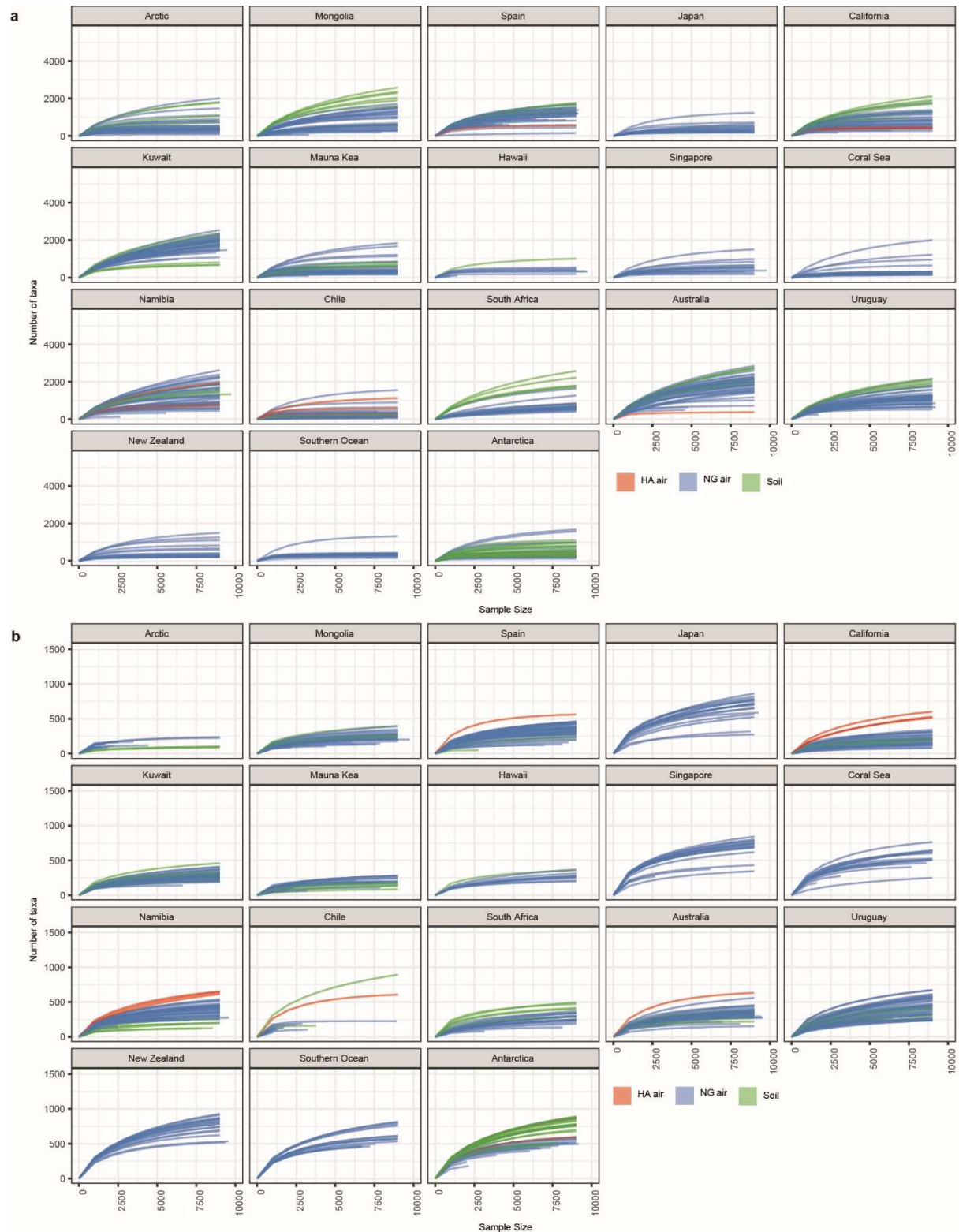
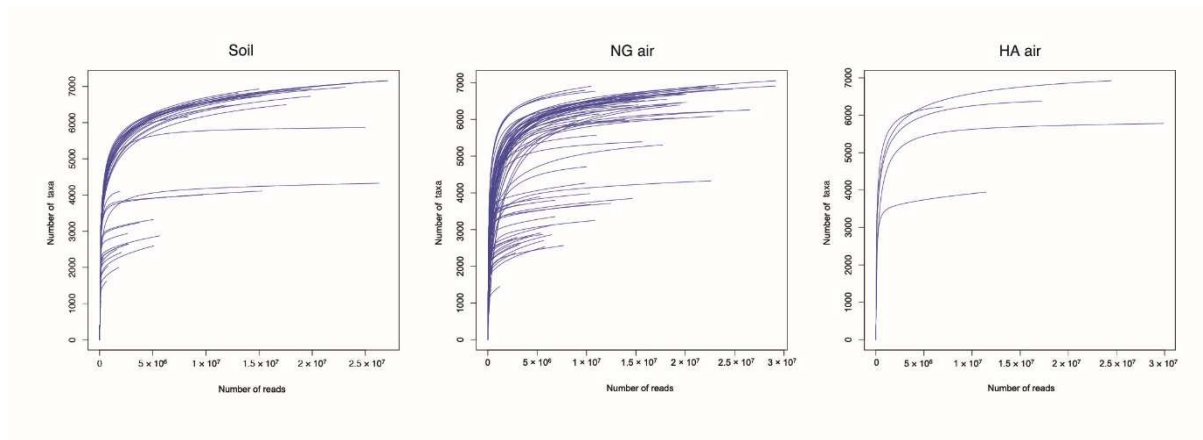


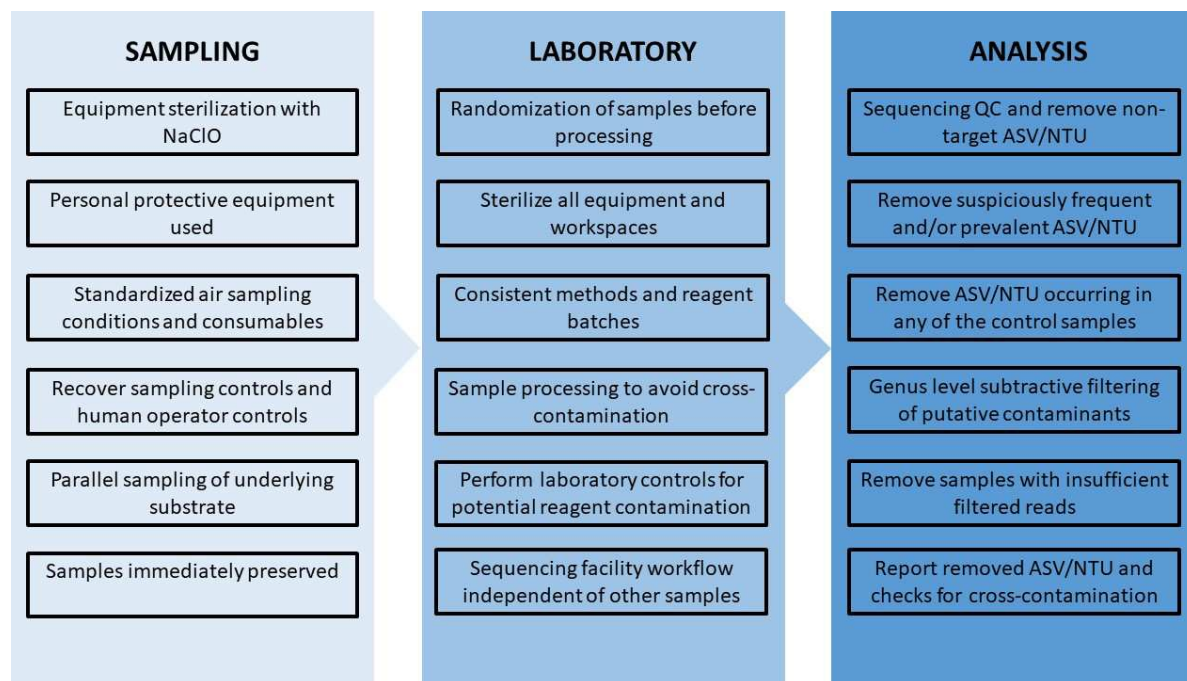
Fig S4 | Rarefaction curves for Metagenomes ($n = 120$). NG air denotes near-ground air and HA air denotes high-altitude air.



Decontamination of environmental sequence data

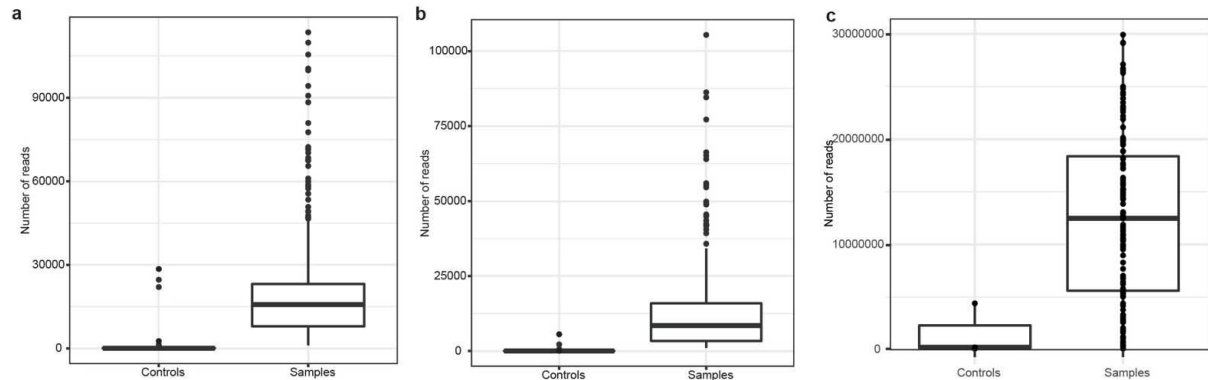
Here we present supporting evidence to demonstrate our decontamination pipeline resulted in effective removal of contaminants and did not adversely impact the observed ecological patterns for our data. Our approach to decontamination adopted and fully reports a thorough multi-step decontamination protocol as recommended in landmark meta-analyses and methods papers [2, 3]. The decontamination of our sequence data was an integral part of our wider approach to careful sampling of ultra-low biomass atmospheric samples and soils (Fig. S5).

Fig S5 | Summary of steps taken to minimise contamination during environmental DNA recovery and sequencing from low biomass air and soil samples. QC denotes quality control, ASV denotes amplicon sequence variant, NTU denotes nearest taxonomic unit used for taxa assignment from metagenomes.



Our close attention to careful field sampling and laboratory workflow resulted in very low read numbers for our control samples compared to environmental samples (Fig. S6). A very small number of 16S rRNA gene sequencing controls ($n = 3$) with higher reads did not identify significantly more ASVs for removal, rather they largely comprised higher read numbers of ASVs that also occurred in other controls. We regard this as an important part of the heterogeneity encountered in environmental microbiology and so we retained these controls in our decontamination pipeline. Due to our careful attention to randomization and replication the higher reads in the small number of controls did not impact the ecological patterns presented for the filtered data.

Fig. S6 | Sequencing read depth for controls and environmental samples. a, Bacteria read depth (controls: n = 35; samples: n = 529). **b,** Fungi read depth (controls: n = 35; samples: n = 444). **c,** Metagenomes read depth for controls (n = 3, pooled by type from 35 independent control samples) and samples (n = 120). Boxplots indicate median (line) and interquartile range (boxes).



We report the removal of reads at each of the aggressive decontamination steps for our amplicon sequence data as per recommended best practice (Table S2) [3]. Sequencing studies of dilution series for mock communities have shown that there is an unavoidable increase in contaminant signal as starting template decreases [4]. This and other studies, e.g. [2], indicate that the percentage of contaminants detected during dilution series to obtain template DNA levels similar to those in our study were comparable or higher (approx. 50-80% in mock communities) to those we observed for our environmental samples, thus supporting that the level of contaminants we reported was not unexpectedly high for the ultra-low biomass habitats we interrogated. Furthermore, the proportion of reads removed from our dataset is consistent with the range in published reports for other ultra-low biomass environmental studies where numbers have been disclosed, i.e. Kiledal *et al.* (2021) removed 85% of amplicon reads during decontamination [5]; Els *et al.* (2019) removed 40% of amplicon OTUs during decontamination [6].

We report the diversity of ASV removed at each stage of our decontamination process (Fig. S7). Untargeted sequence removal included reads affiliated with chloroplasts, mitochondria and other non-target organisms. The R package decontam was applied using a stringent statistical threshold for frequency (0.1) and/or prevalence (0.5) as described in the Methods. Control removal included any taxa occurring in any of the field blanks, laboratory blanks or swabs of human operator gloves being filtered from all samples. Finally, the filtering of named genera targeted any remaining taxa commonly reported as human-associated (n = 21 genera) contaminants in published meta-analyses of ultra-low biomass microbiomes [2, 3].

Table S2 | Total percentage of sequenced reads removed from each step of the decontamination process.

| 16S rRNA gene sequencing decontamination (% of total reads removed from previous step) | | | |
|---|-------------|---------------|---------------|
| | Soil | NG air | HA air |
| Untargeted (%) | 0.71 | 10.84 | 6.35 |
| Decontam (%) ⁺ | 9.67 | 21.65 | 21.24 |
| Control removal (%) [#] | 4.99 | 9.57 | 8.43 |
| Genus-level filtering (%) [^] | 0.22 | 0.9 | 1.58 |
| Total removed (%) | 15.58 | 42.98 | 37.60 |
| ITS sequencing decontamination (% of total reads removed from previous step) | | | |
| | Soil | NG air | HA air |
| Untargeted (%) | 1.0 | 0.07 | 0.05 |
| Decontam (%) ⁺ | 23.28 | 42.98 | 51.25 |
| Control removal (%) [#] | 3.24 | 12.39 | 10.09 |
| Genus-level filtering (%) [^] | 0.13 | 0.06 | 0.03 |
| Total removed (%) | 27.55 | 55.49 | 61.44 |

⁺ The removal of suspiciously frequent and/or prevalent ASVs using the R package decontam resulted in higher numbers of removed ASVs from near-ground air and high-altitude air samples due to the ultra-low biomass of samples and unavoidable increase in contaminant signal [2, 3].

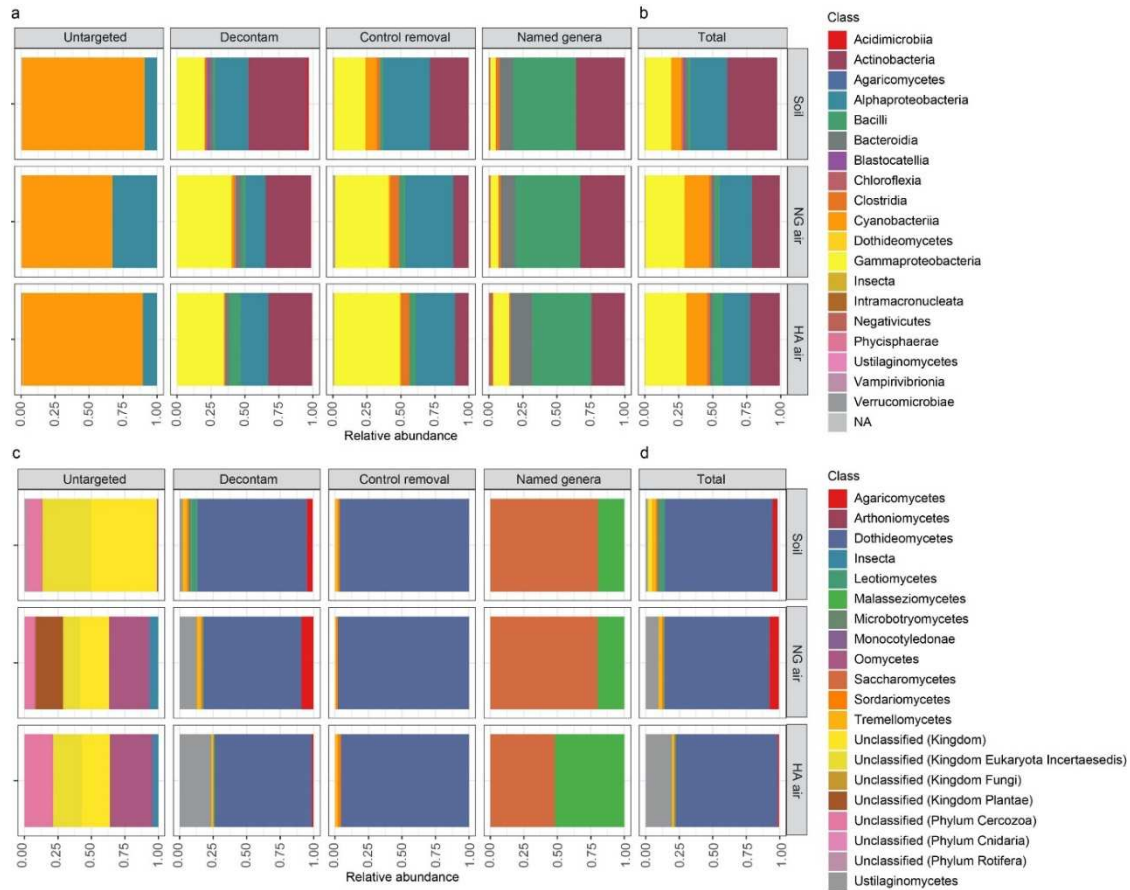
[#] ASVs occurring in any of the field sampling controls, human operator controls or laboratory reagent controls were removed from all samples regardless of whether they were encountered in location-specific or habitat-specific controls where a given sample was recovered.

[^] Whilst our decontam and control removal steps removed most common human and reagent contaminants, we also applied an additional subtractive filter to remove any remaining suspected human-associated contaminants. We targeted the following bacterial genera: *Bacteroides*, *Bifidobacterium*, *Corynebacterium*, *Cutibacterium/Propionibacterium*, *Escherichia*, *Faecalibacterium*, *Haemophilus*, *Klebsiella*, *Lactobacillus*, *Listeria*, *Moraxella*, *Neisseria*, *Porphyromonas*, *Prevotella*, *Shigella*, *Salmonella*, *Staphylococcus*, *Streptococcus* and *Veillonella*. We also removed ASV affiliating with the human-associated fungal genera *Candida* and *Malassezia*. We acknowledge that some of these genera, e.g. *Klebsiella*, *Lactobacillus*, *Listeria*, *Candida* (together approx. 0.1% reads in our study) may also support environmental taxa but we chose to take a cautious approach and remove all ASVs in the genus because most in our study affiliated to human-associated taxa or it was uncertain if they were genuine environmental taxa.

A similarly rigorous process was applied to our metagenomes:

| Shotgun metagenome decontamination (% of total reads removed from previous step) | | | |
|---|-------------|---------------|---------------|
| | Soil | NG air | HA air |
| Contaminant contig mapping (%) | 4.86 | 5.33 | 4.21 |
| Decontam (Bacteria) (%) | 0.29 | 1.03 | 1.5 |
| Genus-level filtering (Bacteria) (%) | 0.61 | 1.56 | 1.13 |
| Decontam (Fungi) (%) | 0.069 | 0.47 | 1.52 |
| Genus-level filtering (Fungi) (%) | 0.0002 | 0.0008 | 0.014 |
| Total removed(%) | 5.83 | 8.39 | 8.38 |

Fig. S7 | Taxonomic identity of reads removed during the decontamination process. Taxonomic composition of removed reads for the 10 most abundant classes at each decontamination step (a) and overall (b) bacterial amplicon data; Taxonomic composition of removed reads for the 10 most abundant classes at each decontamination step (c) and overall (d) fungal amplicon data. Note: Cyanobacteria removed reads affiliated to chloroplasts, a portion of Alphaproteobacteria reads removed affiliated to mitochondria.

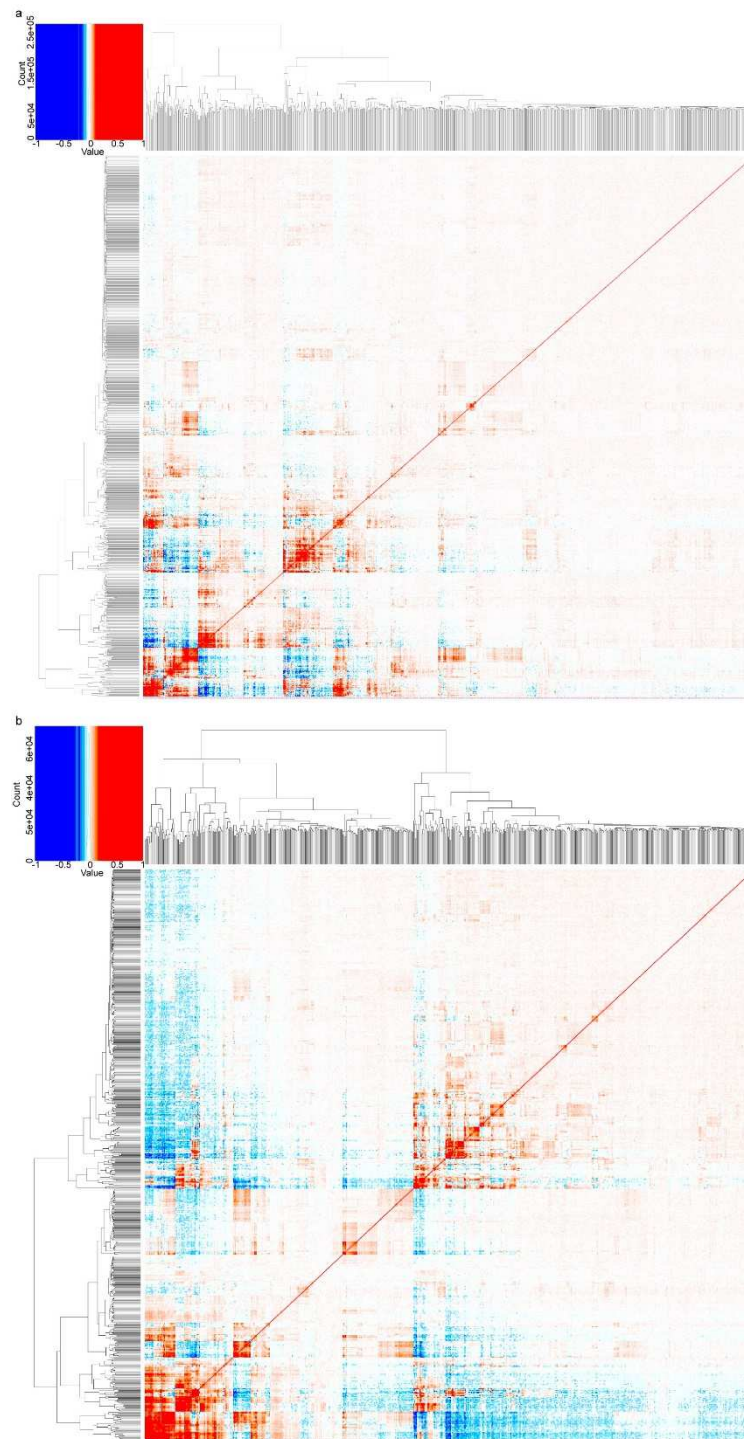


We next performed a post-hoc analysis of our filtered dataset after decontamination to estimate the effectiveness of the multi-step subtractive filtering process and identify any remaining ASVs that may represent potential residual contaminants, as well as identify any evidence for cross-contamination between sample types. In order to focus on unequivocal contamination events, we subsampled the data to 1000 reads per sample and removed taxa that were present in fewer than 10 samples. FastSpar correlation scores and their P values were determined for genus-level taxa and then heatmaps were used to visualise clusters indicative of potential artefacts (Fig S8). Next, taxonomic composition of each cluster were checked for frequency and prevalence of the taxa throughout the dataset. For a cluster to be identified as an artefact suitable for subtraction we applied the following criteria: i) It should align with metadata such as processing batch, sequencing batch, or reagent lot numbers, ii) It should be implausibly consistent between samples, for example if it spans multiple ecological locations, iii) it should be extremely rare beyond the artefact itself.

The analysis revealed a small number of clusters in the bacterial and fungal datasets and weak correlation blocks overall. This indicates a low possibility of exogenous contaminant since taxa arising from a common source would correlate more strongly than real ecological associations. There was no evidence for batch effects (i.e. correlation blocks did not match up with processing batch, sequencing batch, or reagent lot numbers) but a small number of genera displayed clustering indicative of potential residual contamination. A cluster comprising *Ampullimonas*, *Rhizobacter*, *Tardiphaga*, *Variovorax*, and unclassified *Methylophilaceae* appeared to form a tightly correlating pattern that was not tied to geographic origin and so they are likely to be residual contaminants. Additional weak clusters yielded inconclusive evidence for potential residual

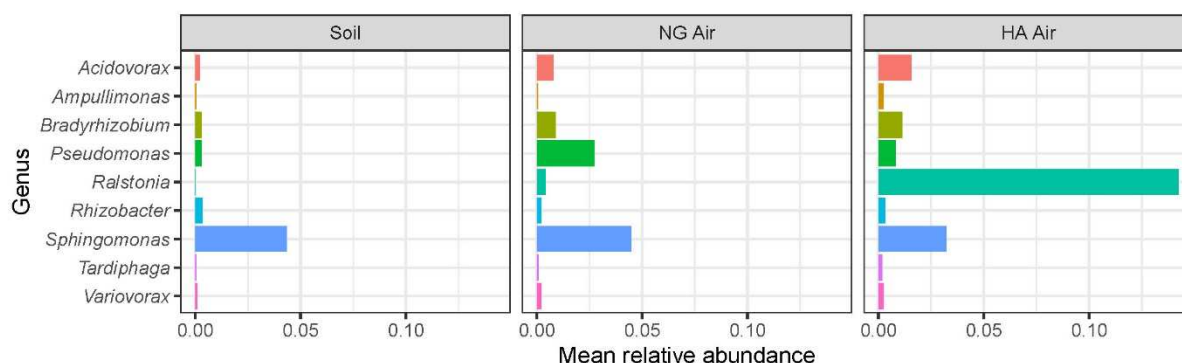
contamination from *Acidovorax*, *Bradyrhizobium*, *Pseudomonas*, *Ralstonia*, and *Sphingomonas*. The latter genera contain ASVs that were removed during our decontamination steps but they also contain known environmental taxa and so removing them at genus level was deemed imprudent. For the fungi a single correlation block affecting 88 samples was identified as worthy of further scrutiny, but since these samples all originated from marine and maritime locations in the southern hemisphere it was concluded this likely represented a valid separation from other samples geographically and by ocean-land site separation.

Fig. S8 | Post-decontamination check for residual contaminants. Heatmaps visualise taxon correlation clustering for a) bacterial and b) fungal genera.



We identified all ASVs in our filtered bacterial dataset that affiliated within the genera identified as potentially suspicious from this analysis (Fig. S9). Most of these formed a very low percentage of overall diversity with the exception of *Ralstonia*, where a small number of ASVs occurred with higher frequency in some high-altitude air samples. In view of the overall low number of potential residual contaminants, that they affiliated with genera that have known environmental taxa as well as commonly encountered contaminants, and in order to preserve the clear step-wise decontamination process was applied identically across samples; we chose not to perform further removal of these ASVs, although we report their taxonomic affiliation so that this can be used to nuance our data. The filtered fungal dataset revealed no potential residual contaminants.

Fig. S9 | Taxonomic composition of potential residual contaminants. Data shown as a fraction of the filtered bacterial datasets from different habitats. NG Air = near-ground air, HA Air = high-altitude air.



We further analysed our data to examine the potential for batch effects (cross contamination) because this is of particular importance when sampling habitats with highly differing biomass such as air and soil. We also examined our data for potential signals of temporal or spatial auto-correlation.

Several steps were taken to minimise the potential for cross contamination:

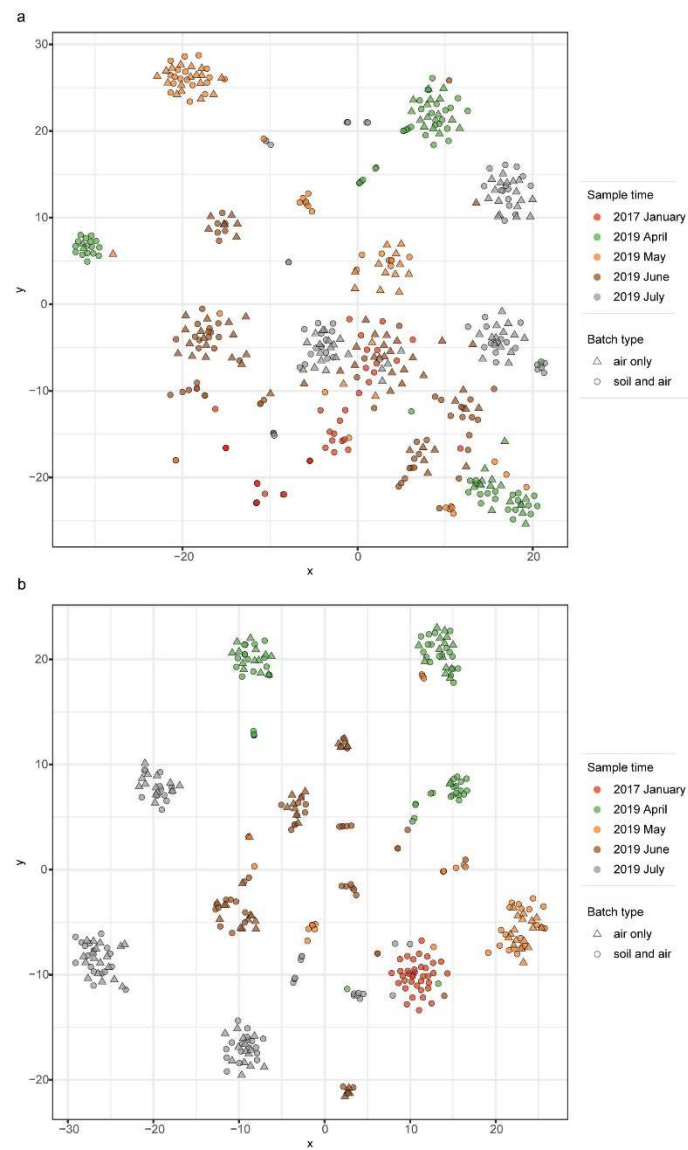
1. Soil and air were sampled separately in time (i.e. soil was collected immediately after each air sampling) and in terms of potential cross-contact (i.e. separate field equipment and processing for soil and air). All samples (i.e. both soil and air) entered the laboratory in the same state in separate sealed and surface-sterilised sample tubes as a mixture in nucleic acid preservative solution (i.e. no free soil particulates were introduced to the laboratory).
2. The processing of all samples was fully randomised for location and habitat type. The randomised workflow involved some batches where both air and soil samples were processed in the same batch and others where air samples only were processed, but we conducted two very important steps to check for potential cross contamination: i) We adopted a workflow where for every batch, each sample tube was processed individually within a BSL2 flow hood. This means that at any given time during laboratory work only a single sample tube was opened during processing to reduce the chance of cross-contamination (the main source of cross-contamination is thought to be micro-fluid splashes between samples). No multiplexing was employed during pipetting. We believe this was a necessary although extremely laborious step to avoid cross-contamination. ii) We conducted statistical tests to determine batch-type did not result in significantly different estimates, i.e. there was no evidence for cross-contamination between soil and air samples. Matching samples (from same location and same sample types but differing in sequencing batches which had either air samples only or with air and soil samples) showed significant concordance and low sum of squared errors, thus high goodness of fit (Bacteria: $n = 250$, $m^2 = 0.043333$, $P = 0.001$; Fungi $n = 186$, $m^2 = 0.049571$, $P = 0.001$) between their ordination coordinates (PCoA with Hellinger distance) (Fig. S10). This suggests low influence within each sequencing batch (the dissimilarities of the communities were preserved regardless of other

samples in the sequencing batches). This was also supported by a Mantel test of the community distances (Bacteria: $n = 250$, $r = 0.818$, $P = 0.001$; Fungi $n = 186$, $r = 0.915$, $P = 0.001$). These findings were also supported by our post-decontamination correlation analysis above (Fig. S8), where there it was clear that cross contamination between air and soil samples had been successfully avoided.

3. In addition, controls were included at all sampling and lab processing stages and these were included in the randomised workflow to check for sampling and laboratory reagent contamination (i.e. controls were randomly incorporated into each batch). We also employed other stringent measures to avoid contamination such as full PPE for operators, bleach-sterilization of all laboratory surfaces and equipment contact surfaces, and UV illumination of workstations when not in use.
4. Our amplicon and metagenome sequencing devices were located in different laboratories and yielded comparable diversity trends across all samples, which we confirmed statistically using Procrustes analysis (Bacteria $M^2 = 0.76$, correlation = 0.49, $p = 0.001$; Fungi $M^2 = 56$, correlation = 0.66, $p = 0.001$). This provided evidence that cross-contamination during sequencing was highly unlikely due to our randomised workflow.

Matching samples (from same location and same sample types but differ in sequencing batches which had either air sample only or with air and soil samples) showed significant concordance and low sum of squared errors, thus high goodness of fit (Bacteria: $n = 250$, $m^2 = 0.043333$, $P = 0.001$; Fungi $n = 186$, $m^2 = 0.049571$, $P = 0.001$) between their ordination coordinates (PCoA with Hellinger distance). This suggests low influence within each sequencing batch (the dissimilarities of the communities were preserved regardless of other samples in the sequencing batches). This was also supported by a Mantel test of the community distances (Bacteria: $n = 250$, $r = 0.818$, $P = 0.001$; Fungi $n = 186$, $r = 0.915$, $P = 0.001$). It was clear from ordinations that sampling date did not affect location and habitat-specific clustering and so temporal auto-correlation was discounted as a confounding factor (Bacteria $n = 449$, Fungi $n = 365$) (Fig. S10). We also performed checks using PCNM to discount the effects of spatial auto-correlation on our findings [7, 8]. Our sampling design was dominated by very large distances that separated the main locations across the major biomes and continents we sampled across the globe. Due to the lack of hierarchical structure this design was not intended to account for smaller spatial scales such as within continents or between locations within the same continent. Because of this reason, when we applied PCNMs to our dataset, the results indicated that latitude and longitude were the major spatial correlates of the multivariate distribution displayed, as shown in the Results where samples clearly clustered by geographic location. The simple linear effect of latitude and longitude thus dominated as a simple expression of the large distance between our sampling locations, and the resulting PCNMs (eigenvectors) extractable from our distance matrix accounted for such a small amount of variance that we eventually excluded these vectors after our preliminary check for auto-correlation. We therefore also calculated and reported distance decay and co-occurrence analysis in the main manuscript to add another layer of support that spatial auto-correlation was not a significant confounding factor on our observations. Our post-decontamination correlation plot for location at genus level (Fig. S8) also revealed no patterns that support the existence of strong spatial auto-correlation independent of the differences expected and observed due to different climate and habitat.

Fig. S10 | Evaluation for batch effects and temporal auto-correlation in samples. Hellinger distance for global bacterial (a) and fungal (b) assemblages, decomposed as weighted community diversity-abundance using tSNE. Processing batch type (batches with air samples only versus batches with air and soil samples) and sampling date (temporal variation) are visualised.



We then plotted visualisations of our data to illustrate ASV removed from each location and habitat type (Fig. S11), and the diversity of filtered ASV in all samples pre- and post-decontamination (Fig. S12, Fig. S13).

Fig. S11 | Taxonomic identity of reads removed during the decontamination process by location and habitat type. Taxonomic profile of putative contaminants removed from bacterial amplicon (a) and metagenome (c) data averaged by location. Taxonomic profile of putative contaminants removed from fungal amplicon (b) and metagenome (d) data averaged by location. NG air = near-ground air, HA air = high-altitude air. Relative abundance indicates occurrence as a percentage of all contaminant ASV and not the occurrence of legitimate sample ASV. Locations: 1, Canada; 2, Mongolia; 3, Spain; 4, Japan; 5, California, USA; 6, Kuwait; 7, Hilo, Hawaii, USA; 8, Mauna Kea, Hawaii, USA; 9, Singapore; 10, Coral Sea; 11, Namibia; 12, Chile; 13, South Africa; 14, Australia; 15, Uruguay; 16, New Zealand; 17, Southern Ocean; 18, Antarctica. Location metadata are shown in Supplementary Information.

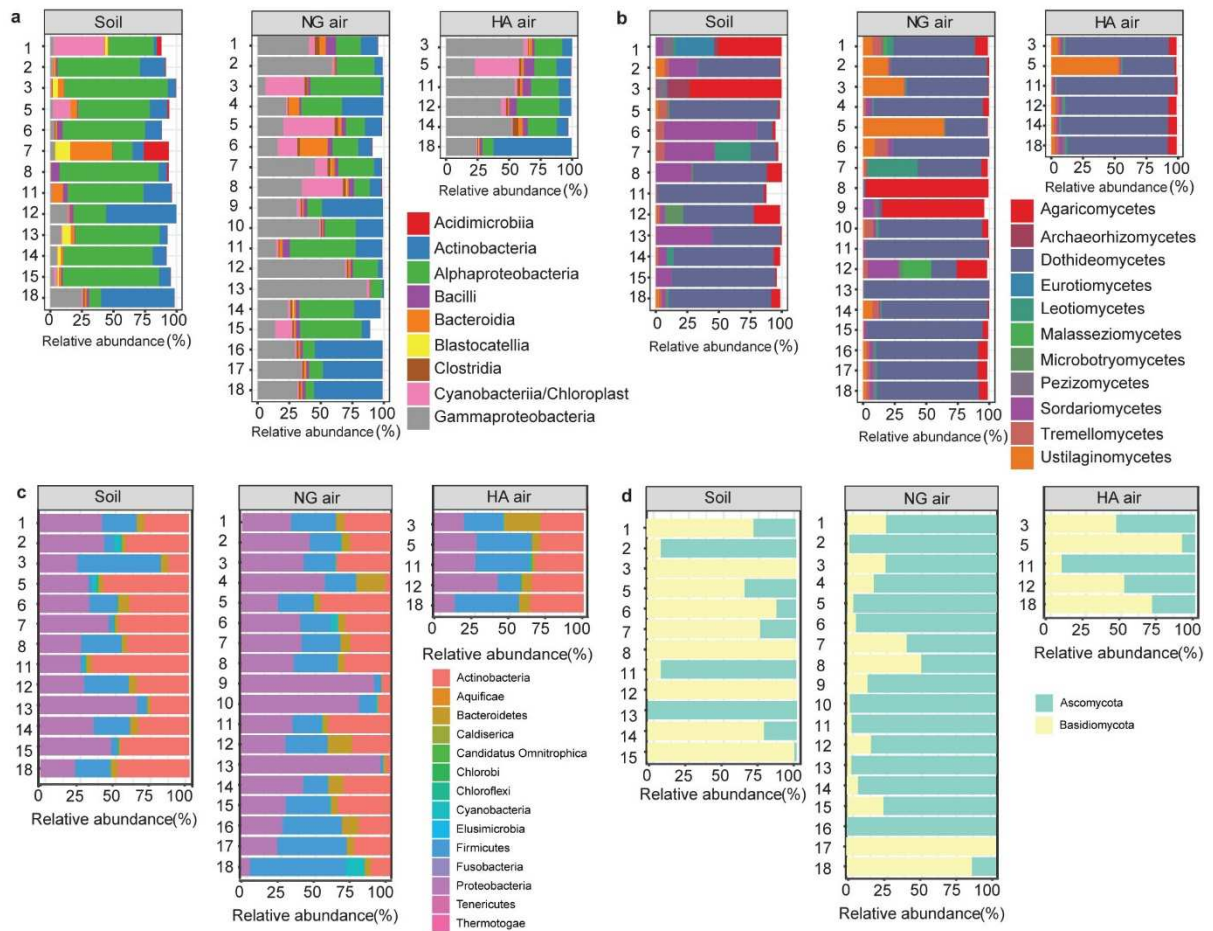


Fig. S12 | Comparison of bacterial diversity estimation pre- and post-decontamination. Taxonomic composition for bacteria (for classes $\geq 0.1\%$ mean relative abundance) for raw data (a) and post decontamination pipeline (b).

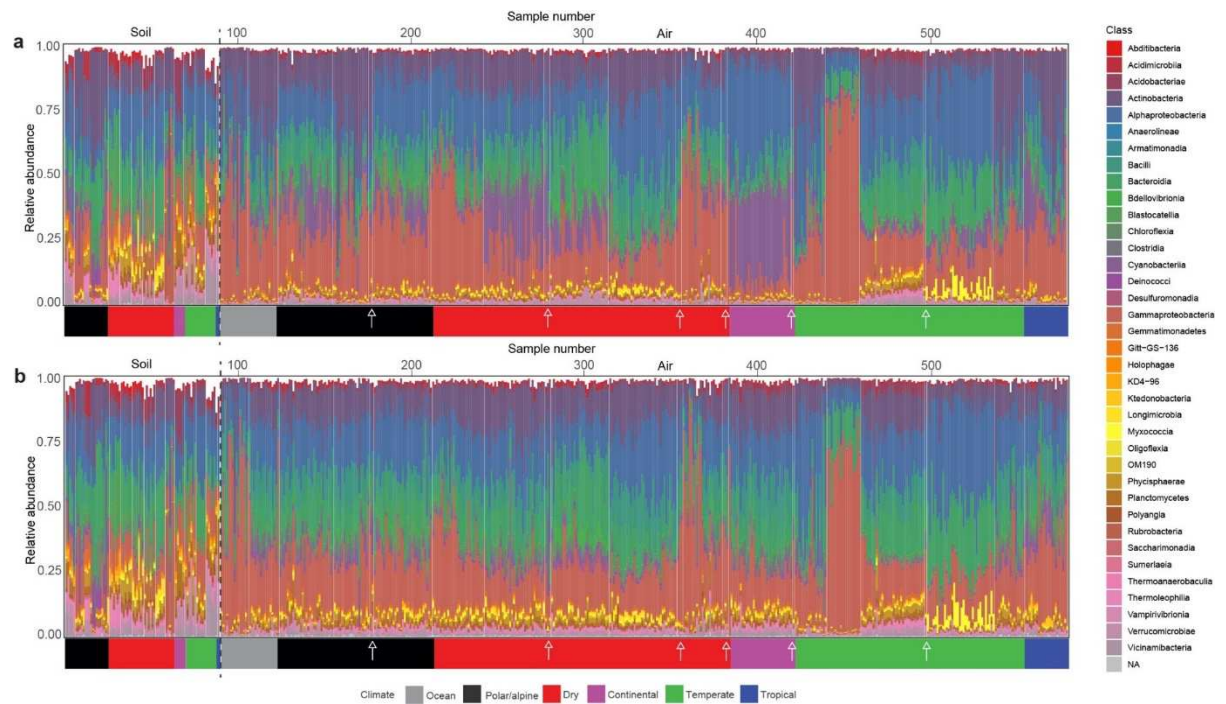
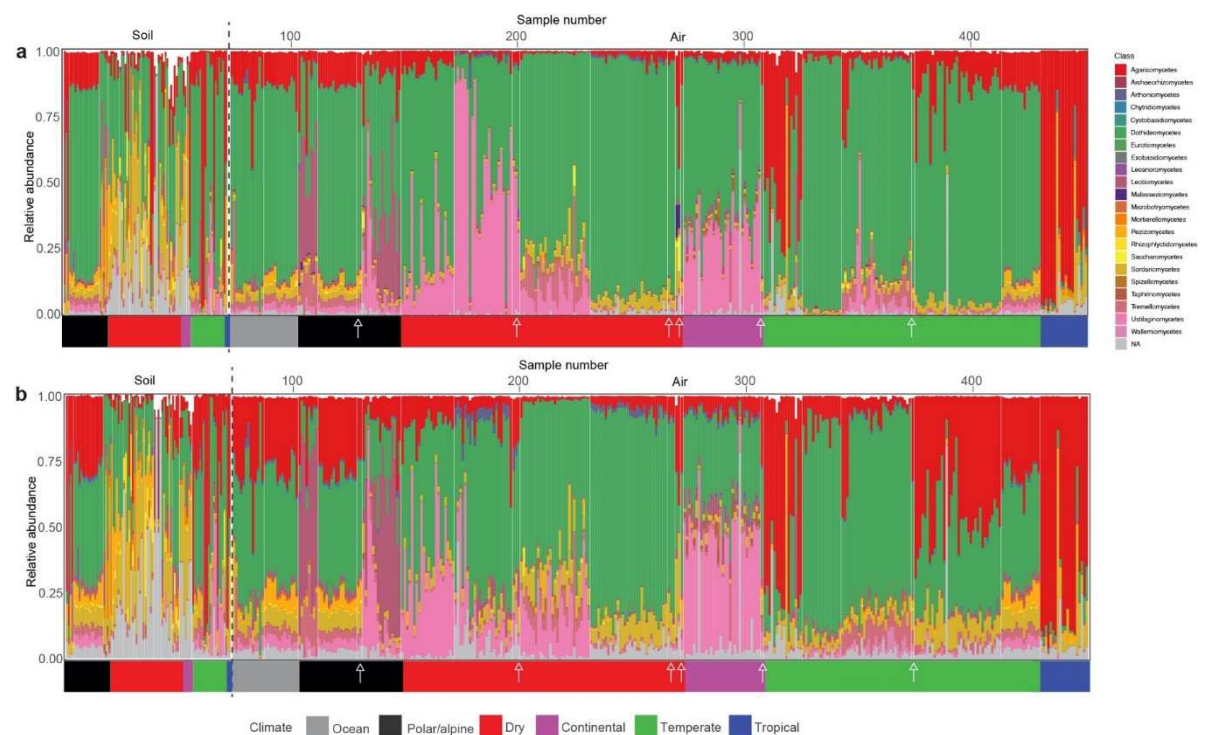
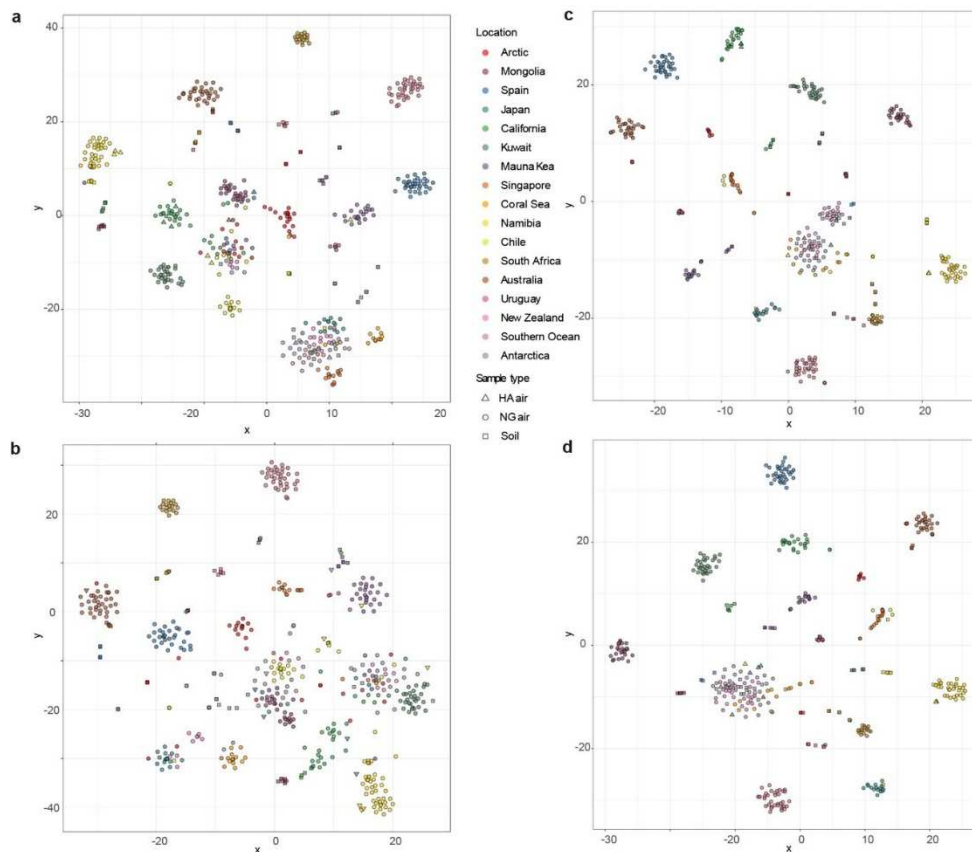


Fig. S13 | Comparison of fungal diversity estimation pre- and post-decontamination. Taxonomic composition for fungi (for classes $\geq 0.1\%$ mean relative abundance) for raw data (a) and post decontamination pipeline (b).



We then tested that our robust decontamination protocol did not have a significant impact on the strong ecological patterns by habitat type and location as highlighted in our manuscript (Fig. 2, Extended Data Figs 3 & 4). We performed ordinations (tSNE based on Hellinger Distances) for the community separately for raw and filtered data (i.e. pre- and post-decontamination) (Fig. S14). This was confirmed with strong and significant correlations for both Procrustes and Mantel tests based on Pearson's product-moment correlation for congruence for bacteria (Procrustes $m^2 = 0.089$, correlation = 0.955, $P = <0.01$ and confirmed by Mantel statistic $r = 0.744$, $P = <0.01$); and fungi (Procrustes $m^2 = 0.083$, correlation = 0.958, $P = <0.01$ and confirmed by Mantel statistic $r = 0.901$, $P = <0.01$).

Fig. S14 | Comparison of community clustering pre- and post-decontamination. Ordinations (tSNE based on Hellinger Distances) for bacteria raw data (a), bacteria post-decontamination pipeline (b); Fungi raw data (c) and fungi post-decontamination pipeline (d). We tested that our robust decontamination protocol did not have a significant impact on the strong ecological patterns by habitat type and location as highlighted in the manuscript (Fig. 2, Extended Data Figs 3 & 4). We performed ordinations (tSNE based on Hellinger Distances) for the community separately for raw and filtered data (Fig. 6). This was confirmed with strong and significant correlations for both Procrustes and Mantel tests based on Pearson's product-moment correlation for congruence for bacteria (Procrustes $m^2 = 0.089$, correlation = 0.955, $P = <0.01$ and confirmed by Mantel statistic $r = 0.744$, $P = <0.01$); and fungi (Procrustes $m^2 = 0.083$, correlation = 0.958, $P = <0.01$ and confirmed by Mantel statistic $r = 0.901$, $P = <0.01$).



Supplementary discussion on use of Null models in the study

Models

Null models are widely used by ecologists to detect non-random patterns in data matrices. These models generally require a randomization of observed data subjected to some constraints. The constraints should reflect the hypotheses under investigation [9–11]. The typical data matrix analysed by ecologists are species by site tables and tables that represent interactions between different groups of species (for example plant-pollinator or plant-root symbiont networks). Classical ecological null models are constructed by constrained permutations, which usually fix some general property of the data matrix such as the column or row margins [11–13]. A metric quantifying the pattern under investigation is then calculated both on the observed data matrix and the randomly generated matrices. Using the logic of null models, the difference between the observed metrics and the metrics in the randomised null matrices is expressed as the z -score of the metric, say metric X , as:

$$z_X = \frac{X(\mathbf{O}) - \langle X \rangle}{\sigma(X)} = \frac{X^* - \langle X \rangle}{\sigma(X)}$$

Where \mathbf{O} is the observed data matrix, X^* is the observed value of the metric and $\langle X \rangle$ and $\sigma(X)$ are the expected value and standard deviation in the ensemble of the permuted matrices. If $\langle X \rangle$ and $\sigma(X)$ describe a normal distribution, the probability of observing a difference beyond two standard deviations just by chance would be roughly 0.05, but if the distribution is not normal an operational p -value < 0.05 can be calculated following classical null modelling [9]. As explained in the main methods, we did not use permutations of the raw data to construct our null models. The reason is that producing a large number of randomly rewired matrices using permutations leads to a set of biased null model matrices, if the original matrix is heterogeneous (e.g. some taxa are much more widespread than others) and sparse [14, 15]. Local randomization algorithms risk sampling the set of null, random matrices non-uniformly, which means that the estimates of the metrics measured on the data matrix is not guaranteed to correspond to the correct theoretical expectation. Given all these issues, which would likely affect a classical null model analysis of our data set, we used the so-called network canonical ensemble, a model belonging to the set of methods known as the statistical mechanics of network. The most complete reference for our approach is in the book by Squartini & Garlaschelli 2017 [16]. Briefly, we interpreted the taxon by location data matrix as a binary bipartite network that describes the occurrence of each taxon at each location. We used the number of taxa at each location or, conversely, the number of locations in which each taxon was found as the constraining vector for the construction of the null model ensemble. The null model matrices will thus be fully random in terms of which taxon is found at each location, which is central to our hypothesis of a non-randomly structured air microbiome. At the same time, the number of taxa per location and also the number of locations in which each taxon is found is, on average, the same in the null models and in the observed matrices. This is important, because it means that deviations of observed data from the null models will be affected solely by the taxonomic composition of the assemblage. As it can be easily shown following the derivations [16], the Hamiltonian of the graph corresponding to our model would then be

$$H(\mathbf{A}, \boldsymbol{\theta}) = \sum_i \theta_i k_i(\mathbf{A}) = \sum_i \sum_{j < i} (\theta_i + \theta_j) a_{ij}$$

for a general binary matrix, and the probability distribution of each occurrence in the matrix

$$P(\mathbf{A}|\boldsymbol{\theta}) = \prod_i \prod_{j < i} p_{ij}^{a_{ij}} (1 - p_{ij})^{1-a_{ij}}.$$

This solution can easily be reparametrized to describe a binary bipartite graph [17]. The probability distribution $P(\mathbf{A}|\boldsymbol{\theta})$ is the correct (i.e. it maximises entropy) probability distribution for each taxon randomly occurring in each location, subject to the constraints given by the number of locations in which that taxon can be found, and the number of taxa found in each location. The parameters of the distribution can be estimated using maximum likelihood methods [18] (described in the Methods). The probability distribution is then sampled [19], to generate the desired number of randomised matrices (we sampled 999 matrices) and construct the null model. As the probability distribution is analytically derived for each of the three types of samples we analysed (soil, near ground and high elevation air), each type of sample (soil, NG air and HA air) has its own baseline or null model.

Metrics

The metrics we tested in the null models were NODF, a metric of nestedness, and the Jaccard index. There is a vast, and often lively debated literature on the property of nestedness. Nested pattern have been defined as (quote) “[nested patterns] are those in which the species composition of small assemblages is a nested subset of larger assemblage” [20]. This applies also to randomly assembled communities and so a null model is needed to test whether nestedness pattern are not random. Assume the microbial composition at each location is completely random and just reflects recruitment of the most abundant taxa from the regional pool. Then, any local community that randomly recruits a relatively small number of taxa will form a subset of the local communities that randomly happen to recruit a relatively large number of taxa. However, if different groups of local communities or, also, individual local communities, are characterised by a particular and unique taxonomic composition, with some taxa that specifically occur only at certain locations and others that occur at other locations, observed nestedness will be much lower than expected under a null model that fully randomise taxa composition.

Our second metrics, the Jaccard index on presence/absence data, quantifies pairwise community dissimilarity. The higher the index the lower the number of taxa shared by the two compared locations. One can thus calculate the average dissimilarity of a taxon by location matrix, which we did both for the observed and null model matrices. Our results (Fig. 2) showed that bacterial (Fig. S15) and fungal communities (Fig. S16) in air were more dissimilar than expected under the null model. Fungi in soil showed the same difference. Instead, bacteria in soil were more similar than expected under the null model.

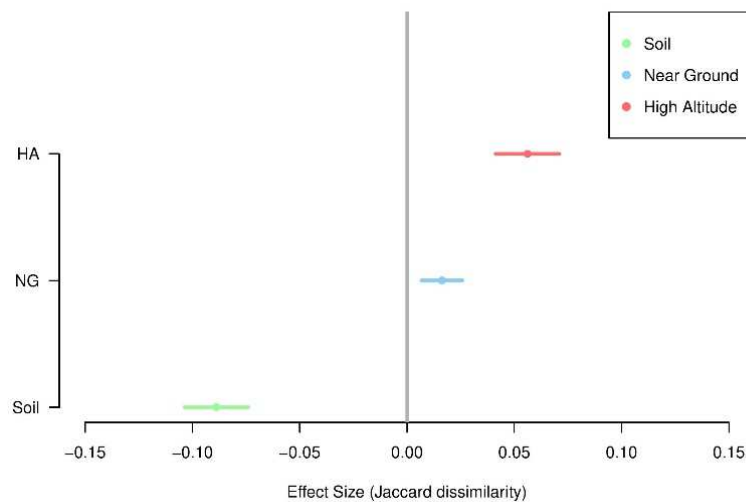


Fig. S15 | Jaccard network null model for bacterial ASV.

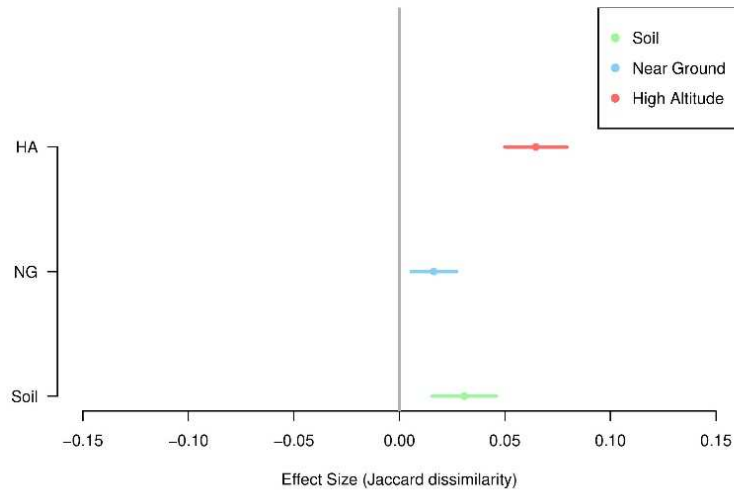


Fig. S16 | Jaccard network null model for fungal ASV.

There are several possible explanations for convergence and divergence of taxonomic composition relative to the baseline provided by a stochastic null model (e.g. [21–25]). Soil appeared to have “homogenised” ASV and also at higher rank community composition (not shown) relative to the null model. This could be due to selection forces such as environmental filtering combined with limited dispersal, as discussed in the main text. In contrast, Air made ASV and also higher rank taxa composition more heterogeneous than in the null model. This type of compositional divergence, too, can be due to selective forces but with the environmental conditions of the locations varying to a large extent between locations, which diversifies the composition of communities across locations. The explanations we are offering for these patterns of compositional dissimilarity observed in air and soil remain hypotheses to be tested in the future but the Jaccard pattern, together with NODF demonstrates the non-random structure of the air microbiome in terms of the distribution of taxa across the network of locations of this study.

Inventory of taxa

Table S3 | List of bacterial genera prevalent in $\geq 50\%$ of soil and/or air samples.

| Prevalent in soil only | Prevalent in both | Prevalent in air only |
|--------------------------------|--|-------------------------|
| <i>Acidovorax</i> | <i>Abditibacterium</i> , <i>Actinoplanes</i> , <i>Adhaeribacter</i> , <i>Allorhizobium</i> - <i>Neorhizobium</i> - <i>Pararhizobium</i> - <i>Rhizobium</i> , <i>Altererythrobacter</i> , <i>Arthrobacter</i> , <i>Bacillus</i> , <i>Bdellovibrio</i> , <i>Belnapia</i> , <i>Blastocatella</i> , <i>Blastococcus</i> , <i>Bryobacter</i> , <i>Candidatus_Alysiosphaera</i> , <i>Candidatus_Udaeobacter</i> , <i>Cellulomonas</i> , <i>Chthoniobacter</i> , <i>Conexibacter</i> , <i>Devosia</i> , <i>Ellin6055</i> , <i>Flaviaesturariibacter</i> , <i>Flavisolibacter</i> , <i>Flavobacterium</i> , <i>Friedmanniella</i> , <i>Gaiella</i> , <i>Gemmata</i> , <i>Gemmatimonas</i> , <i>Geodermatophilus</i> , <i>Haliangium</i> , <i>Hymenobacter</i> , <i>Lautropia</i> , <i>Luteitalea</i> , <i>Lysobacter</i> , <i>Marmoricola</i> , <i>Massilia</i> , <i>Methylobacterium</i> - <i>Methylorubrum</i> , <i>Micromonospora</i> , <i>Microvirga</i> , <i>Modestobacter</i> , <i>Mycobacterium</i> , <i>Nocardioideis</i> , <i>Noviherbaspirillum</i> , <i>Novosphingobium</i> , <i>Oligoflexus</i> , <i>Pedobacter</i> , <i>Peredibacter</i> , <i>Pirellula</i> , <i>Pontibacter</i> , <i>Pseudomonas</i> , <i>Pseudonocardia</i> , <i>Psychroglaciecola</i> , <i>Quadrisphaera</i> , <i>Ramlibacter</i> , <i>RB41</i> , <i>Rhizobacter</i> , <i>Rhodocytophaga</i> , <i>Roseisolibacter</i> , <i>Roseomonas</i> , <i>Rubellimicrobium</i> , <i>Rubrobacter</i> , <i>Segetibacter</i> , <i>Skermanella</i> , <i>Solirubrobacter</i> , <i>Sphingomonas</i> , <i>Spirosoma</i> , <i>Stenotrophobacter</i> , <i>Steroidobacter</i> , <i>Streptomyces</i> , <i>Subgroup_10</i> , <i>Sumerlaea</i> , <i>Truepera</i> , <i>Variovorax</i> , <i>YC-ZSS-LKJ147</i> | <i>Acidibacter</i> |
| <i>Acinetobacter</i> | | <i>Amaricoccus</i> |
| <i>Ammoniphilus</i> | | <i>Archangium</i> |
| <i>Aquabacterium</i> | | <i>Aridibacter</i> |
| <i>Atopostipes</i> | | <i>Caenimonas</i> |
| <i>Bhargavaea</i> | | <i>Caulobacter</i> |
| <i>Chryseobacterium</i> | | CL500-29_marine_group |
| <i>Clostridium</i> | | <i>Crossiella</i> |
| <i>Cnuella</i> | | <i>Edaphobaculum</i> |
| <i>Craurococcus-Caldovatus</i> | | Ellin517 |
| <i>Deinococcus</i> | | Ellin6067 |
| <i>Domibacillus</i> | | <i>Ferruginibacter</i> |
| <i>Georgenia</i> | | <i>Fimbrioglobus</i> |
| <i>Kineococcus</i> | | <i>Flavitalea</i> |
| <i>Kocuria</i> | | <i>Herpetosiphon</i> |
| <i>Limnobacter</i> | | <i>Iamia</i> |
| <i>Longimicrobium</i> | | JGI_0001001-H03 |
| <i>Luteimonas</i> | | <i>Leptothrix</i> |
| <i>Lysinibacillus</i> | | <i>Mesorhizobium</i> |
| <i>Microbacterium</i> | | MND1 |
| <i>Oceanobacillus</i> | | <i>Nitrosospira</i> |
| <i>Ornithinimicrobium</i> | | <i>Nitrospira</i> |
| <i>Paenibacillus</i> | | <i>Opitutus</i> |
| <i>Paeniclostridium</i> | | <i>Pajaroellobacter</i> |
| <i>Pantoea</i> | | <i>Pedomicrobium</i> |
| <i>Paracoccus</i> | | <i>Phaselicystis</i> |
| <i>Planococcus</i> | | <i>Phenylobacterium</i> |
| <i>Planomicrobium</i> | | <i>Pir4_lineage</i> |
| <i>Pseudarthrobacter</i> | | <i>Reyranella</i> |
| <i>Rufibacter</i> | | <i>Rhodopirellula</i> |
| <i>Salinicoccus</i> | | <i>Rhodoplanes</i> |
| <i>Salinimicrobium</i> | | <i>Sporocytophaga</i> |
| <i>Solibacillus</i> | | <i>Tepidisphaera</i> |
| <i>Sporosarcina</i> | | TM7a |
| <i>Tumebacillus</i> | | |
| <i>Turicibacter</i> | | |
| UCG-005 | | |

Table S4 | List of fungal genera prevalent in $\geq 50\%$ of soil and/or air samples.

| Prevalent in soil only | Prevalent in both | Prevalent in air only |
|---------------------------|-------------------------|-------------------------|
| <i>Acremonium</i> | <i>Alternaria</i> | <i>Exophiala</i> |
| <i>Agaricus</i> | <i>Aspergillus</i> | <i>Knufia</i> |
| <i>Botrytis</i> | <i>Bipolaris</i> | <i>Mortierella</i> |
| <i>Ceriporia</i> | <i>Chaetomium</i> | <i>Oedocephalum</i> |
| <i>Cladosporium</i> | <i>Coprinellus</i> | <i>Powellomyces</i> |
| <i>Dioszegia</i> | <i>Coprinopsis</i> | <i>Rhizophlyctis</i> |
| <i>Ganoderma</i> | <i>Curvularia</i> | <i>Saitozyma</i> |
| <i>Gymnopus</i> | <i>Didymella</i> | <i>Stagonosporopsis</i> |
| <i>Mycosphaerella</i> | <i>Filobasidium</i> | <i>Westerdykella</i> |
| <i>Neoascochyta</i> | <i>Fomitopsis</i> | |
| <i>Nigrospora</i> | <i>Fusarium</i> | |
| <i>Paradendryphiella</i> | <i>Gibberella</i> | |
| <i>Paraphaeosphaeria</i> | <i>Naganishia</i> | |
| <i>Phellinus</i> | <i>Neocamarosporium</i> | |
| <i>Phlebiopsis</i> | <i>Papiliotrema</i> | |
| <i>Psathyrella</i> | <i>Penicillium</i> | |
| <i>Resinicium</i> | <i>Peniophora</i> | |
| <i>Sarocladium</i> | <i>Periconia</i> | |
| <i>Selenophoma</i> | <i>Phaeococcomyces</i> | |
| <i>Spegazzinia</i> | <i>Phaeosphaeria</i> | |
| <i>Torula</i> | <i>Phanerochaete</i> | |
| <i>Toxicocladosporium</i> | <i>Phlebia</i> | |
| <i>Trametes</i> | <i>Preussia</i> | |
| <i>Tranzscheliella</i> | <i>Pseudopithomyces</i> | |
| <i>Tulostoma</i> | <i>Vishniacozyma</i> | |
| <i>Udeniomyces</i> | | |
| <i>Ustilago</i> | | |
| <i>Wallemia</i> | | |

Metagenomics functional gene targets

Table S5 | Functional genes targeted in the metagenomics inquiry of air and soil. A suite of respiratory genes was used as a general marker of potential for metabolically active taxa, and targeted metabolic and stress response genes were selected based upon substrates and stressors encountered in the atmospheric habitat. No hits were recorded for *ina* genes and this likely reflects low homology between taxa.

| Trait | Protein | Gene |
|---|---|------------------|
| Respiration | ATP synthase | <i>atpA</i> |
| | Cytochrome cbb3 oxidase | <i>ccoN</i> |
| | Cytochrome aa3 oxidase | <i>coxA</i> |
| | Cytochrome bd oxidase | <i>cydA</i> |
| | Cytochrome bo3 oxidase | <i>cyoA</i> |
| | NADH-ubiquinone oxidoreductase subunit F | <i>nuoF</i> |
| | Succinate dehydrogenase/fumarate reductase | <i>sdhA/frdA</i> |
| Carbon fixation | Ribulose 1,5-bisphosphate carboxylase/oxygenase | <i>rbcL</i> |
| Nitrogen fixation | Nitrogenase | <i>nifH</i> |
| Phototrophy | Photosystem I reaction centre protein | <i>psaA</i> |
| | Photosystem II reaction centre protein | <i>psbA</i> |
| | Microbial rhodopsin | <i>RHO</i> |
| Atmospheric trace gas metabolism (carbon monoxide, hydrogen, methane, isoprene) | Aerobic carbon monoxide dehydrogenase | <i>coxL</i> |
| | Hydrogenase | <i>Fe</i> |
| | Hydrogenase | <i>FeFe</i> |
| | Isoprene oxidation gene | <i>isoA/mmoX</i> |
| | Soluble methane monooxygenase | <i>mmoX</i> |
| | Particulate methane monooxygenase | <i>pmoA</i> |
| cold-shock | Cold shock protein A | <i>cspA</i> |
| | Cold shock protein B | <i>cspB</i> |
| | Cold shock protein G | <i>cspG</i> |
| | Cold shock protein I | <i>cspI</i> |
| Oxidative stress | Protein-methionine sulfoxide reductase | <i>msrP</i> |
| | Protein-methionine sulfoxide reductase | <i>msrQ</i> |
| Sporulation | Sporulation protein 0A | <i>spo0A</i> |
| Starvation/stationary phase | DNA-protection during starvation protein | <i>dps</i> |
| | Outer membrane protein | <i>slp</i> |
| UV response/repair | Deoxyribodipyrimidine photo-lyase | <i>phrB</i> |
| | UvrABC system protein A | <i>uvrA</i> |
| | UvrABC system protein B | <i>uvrB</i> |
| | UvrABC system protein C | <i>uvrC</i> |
| Ice nucleation | Ice nucleation protein A | <i>inaA</i> |
| | Ice nucleation protein X | <i>inaX</i> |
| | Ice nucleation protein W | <i>inaW</i> |
| | Ice nucleation protein Z | <i>inaZ</i> |

Supplementary data analysis

Fig. S17 | Global patterns in alpha diversity for bacteria and fungi in air and soil. **a)** DNA yield from air and soil samples, note that values are not directly comparable between soil and air. **b)** Bacterial alpha diversity metrics ($n = 529$). **c)** Fungal alpha diversity metrics ($n = 444$). Abbreviations: MK, Mauna Kea; NZ, New Zealand; SA, South Africa; SO, Southern Ocean; HA air, high-altitude air; NG air, near-ground air.

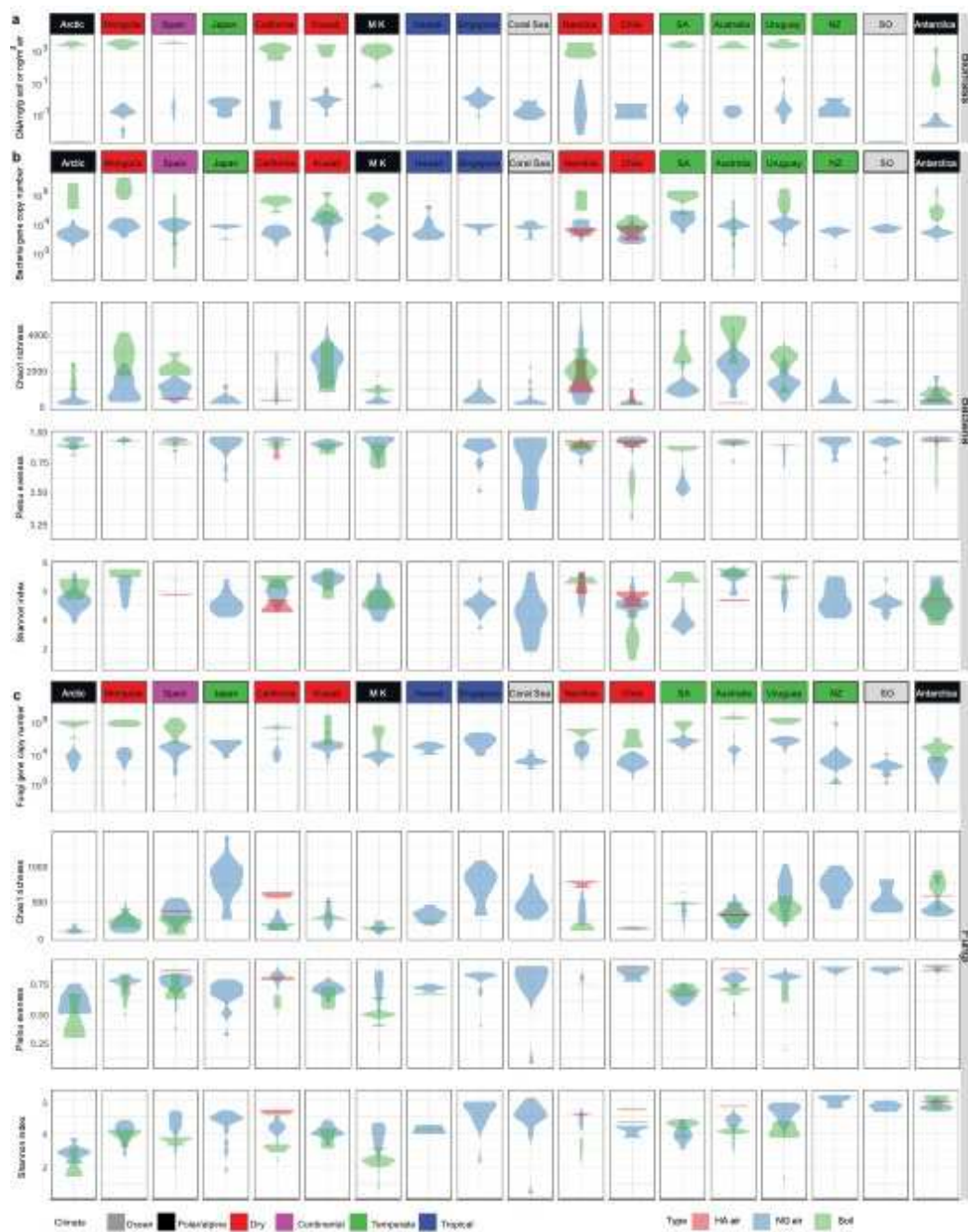


Fig. S18 | Comparison of diversity estimation using shotgun metagenomics and amplicon sequencing of air and soil. a) Relative abundance of twenty most abundant bacterial phyla estimated from metagenomes ($n = 120$). **b)** Relative abundance of twenty most abundant bacterial phyla estimated from amplicon sequencing ($n = 529$). **c)** Relative abundance of fungal phyla estimated from metagenomes ($n = 120$). **d)** Relative abundance of fungal phyla estimated from amplicon sequencing ($n = 444$). Community composition estimation using amplicon sequencing and shotgun metagenomics were positively correlated (Procrustes: Bacteria $m^2 = 0.76$, correlation = 0.49, $P = 0.001$; Fungi $m^2 = 0.56$, correlation = 0.66, $P = 0.001$), and so we focused our fine scale phylogenetic interrogation on amplicon sequence data because this approach allowed better ecological representation of the targeted assemblages in terms of sampling depth and taxonomic resolution. Locations: 1, Canada; 2, Mongolia; 3, Spain; 4, Japan; 5, California, USA; 6, Kuwait; 7, Hilo, Hawaii, USA; 8, Mauna Kea, Hawaii, USA; 9, Singapore; 10, Coral Sea; 11, Namibia; 12, Chile; 13, South Africa; 14, Australia; 15, Uruguay; 16, New Zealand; 17, Southern Ocean; 18, Antarctica. HA air, high-altitude air; NG air, near-ground air; NA, no taxonomy assigned.

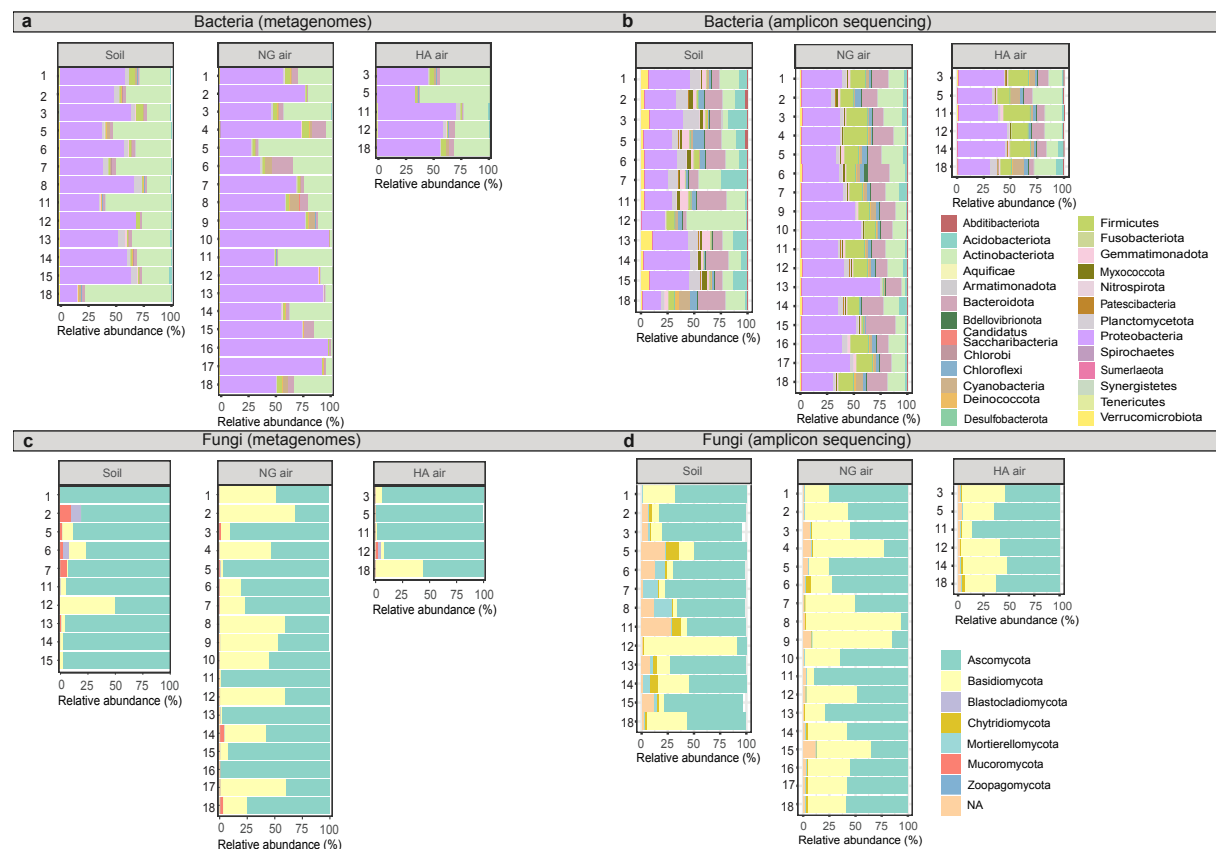


Fig. S19 | Dissimilarity of assemblages by habitat type and location. Jaccard Index for global bacterial (a) and fungal (b) assemblages, decomposed using PCoA. Clustering by habitat type (green, soil; blue, NG air; red, HA air) was also conserved when analysis was reiterated using only locations where concurrent sampling for all habitat types occurred to mitigate against potential sample-size effects.

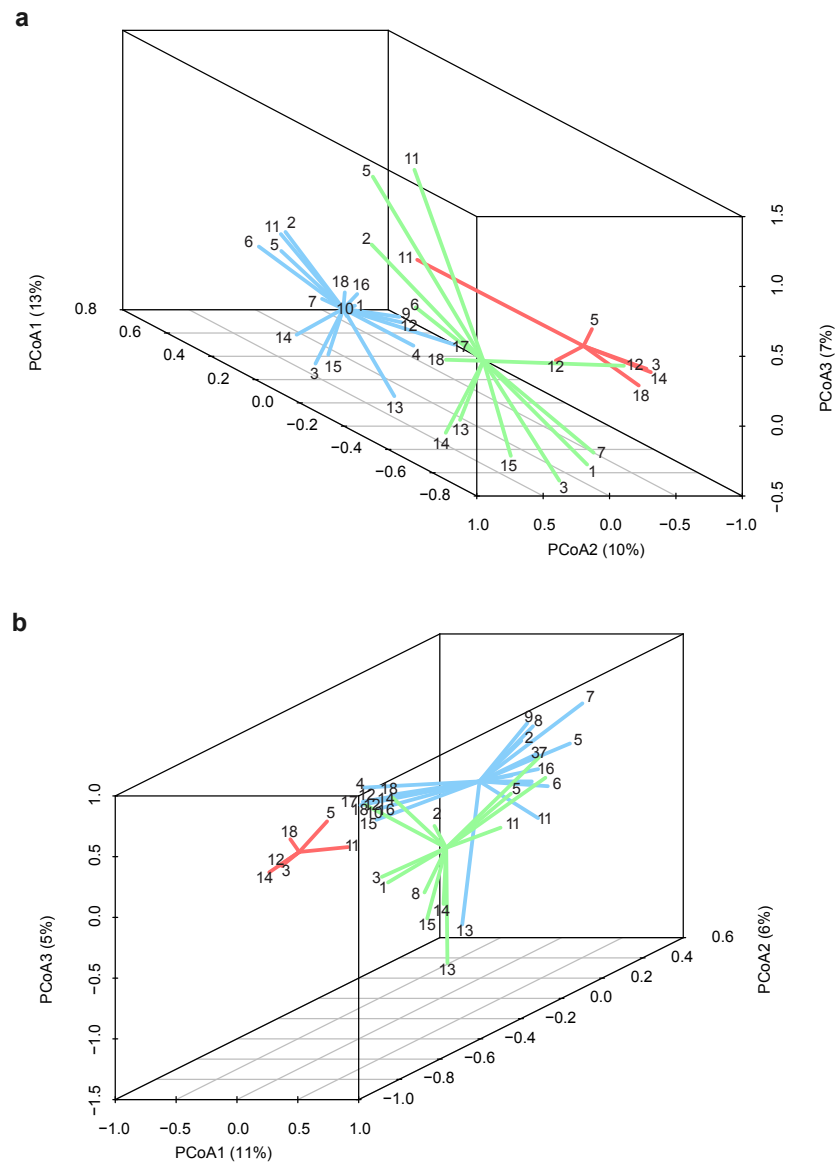


Fig. S20 | Phylogenetic distance of assemblages by habitat type and location. Hellinger Distance for global bacterial (a) and fungal (b) ASV-defined assemblages, decomposed as weighted community diversity-abundance using tSNE. HA air, high-altitude air; NG air, near-ground air. Separation by location is evident except for a cluster of the regionally proximal New Zealand, Southern Ocean and Antarctic locations (Bacteria $n = 529$; Fungi $n = 444$). To ensure that classifications were appropriate to provide accurate ecological insight we also conducted a Procrustes analyses comparing PCoA with Hellinger distances between ASV and Genus defined communities. This found the observations to be highly congruent between the classification methods for bacteria ($m^2 = 0.239$, correlation = 0.872, $P = <0.01$) and fungi ($m^2 = 0.187$, correlation = 0.902, $P = <0.01$).

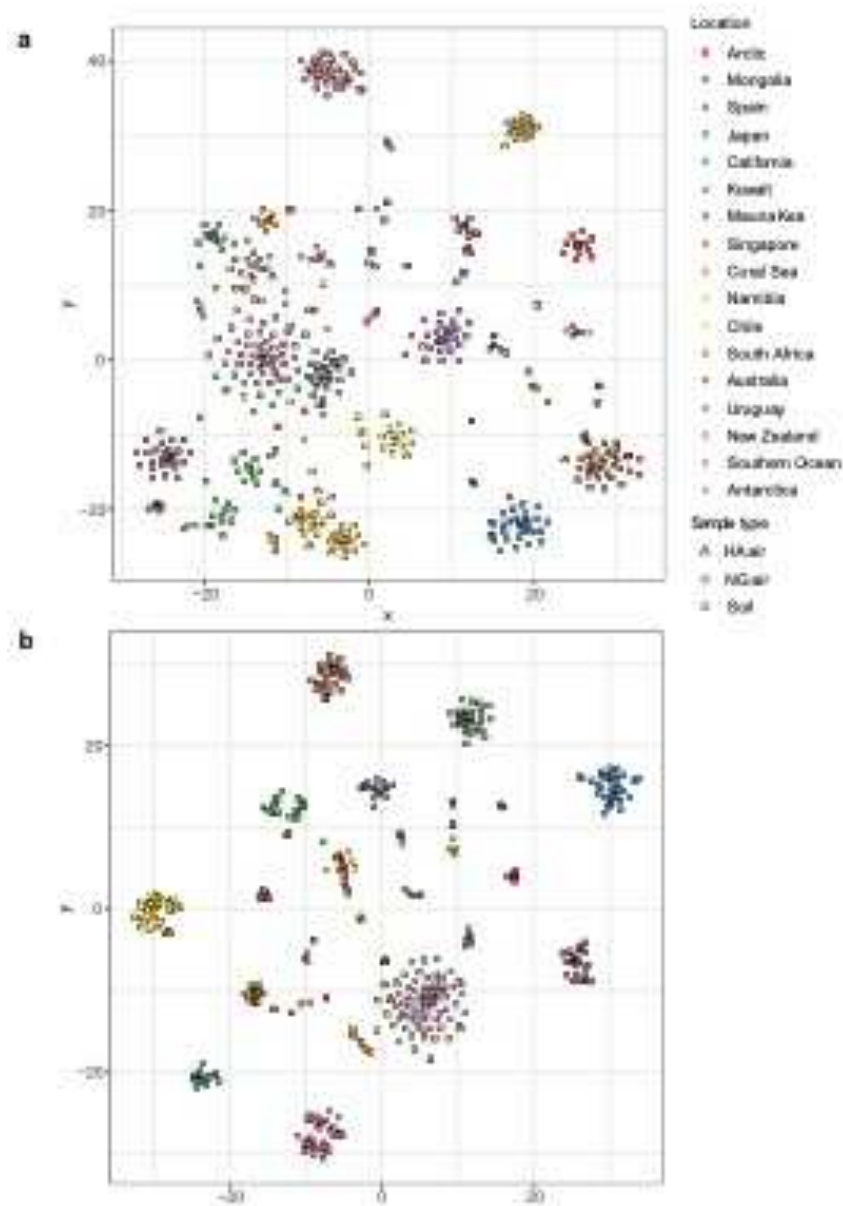


Fig. S21 | Global-scale distance decay plots for bacteria and fungi in air and soil. Shaded area surrounding line of best fit indicates 95% confidence intervals. HA air, high-altitude air; NG air, near-ground air. Significant linear distance decay relationships for diversity were observed for bacterial (Mantel test, Soil: $r = 0.357$, $P = 0.001$; NG air: $r = 0.353$, $P = 0.001$; HA air: $r = 0.433$, $P = 0.002$) and fungal (Mantel test, Soil: $r = 0.507$, $P = 0.001$; NG air: $r = 0.482$, $P = 0.001$; HA air: $r = 0.535$, $P = 0.004$) assemblages. The pattern for near-ground air was fairly pronounced compared to that for underlying soil and this reflects that soil habitats and their microbial diversity were markedly more diverse. Limited inference could be made for the pattern in high-altitude air due to lower sample numbers.

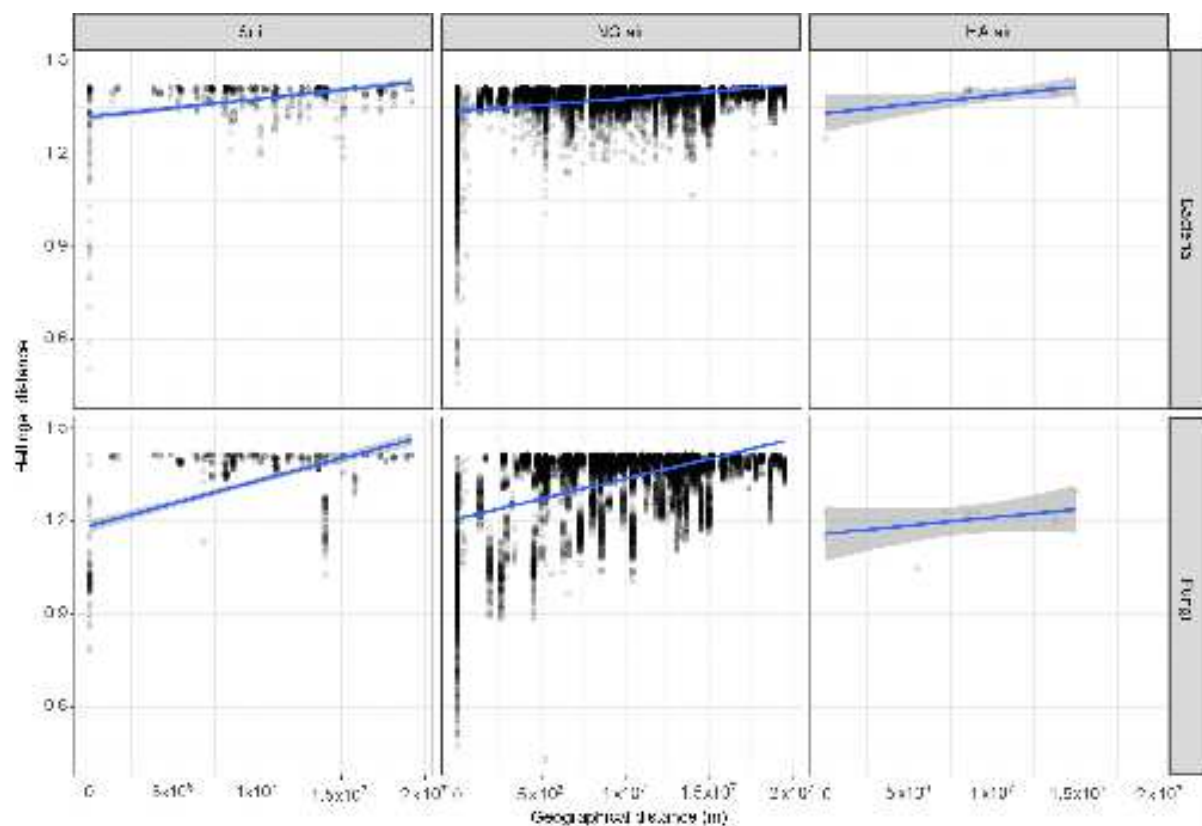


Fig. S22 | Null model networks for bacterial and fungal assemblages in air and soil. a) High-altitude air. b) Near-ground air. c) Soil. For nestedness estimates from phylum to family level all taxa were used, and for genus and ASV level the 1,000 most prevalent taxa. The null models were based upon statistical mechanics reconstruction of the taxa by location bipartite network. Maximum-likelihood was used to estimate the probability distribution that maximised the entropy function of the null network conditional on the constraint of the observed degree sequence, which was enforced as an average vector (Canonical ensemble).

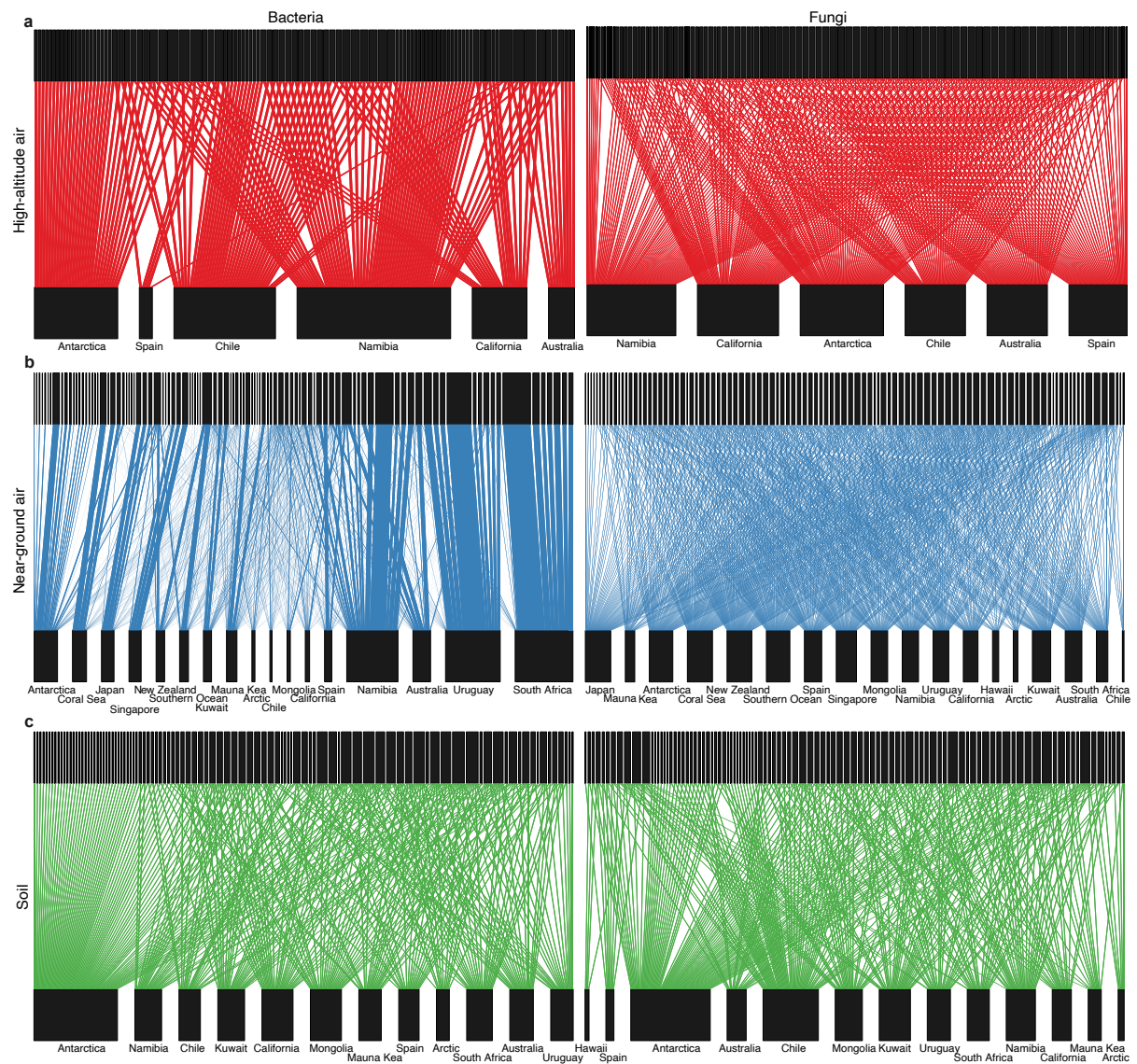


Fig. S23 | Taxonomic composition of bacterial and fungal assemblages in air and underlying soil.

Estimates shown are for ASV-defined assemblages for a) bacteria ($n = 529$) and b) fungi ($n = 444$). Arrows indicate locations where aircraft sampling of high-altitude air was undertaken. A taxa filtering criteria of $<0.1\%$ mean relative abundance was used for these plots and so incomplete bars for some samples indicates taxa with very low abundance rather than unidentified taxa.

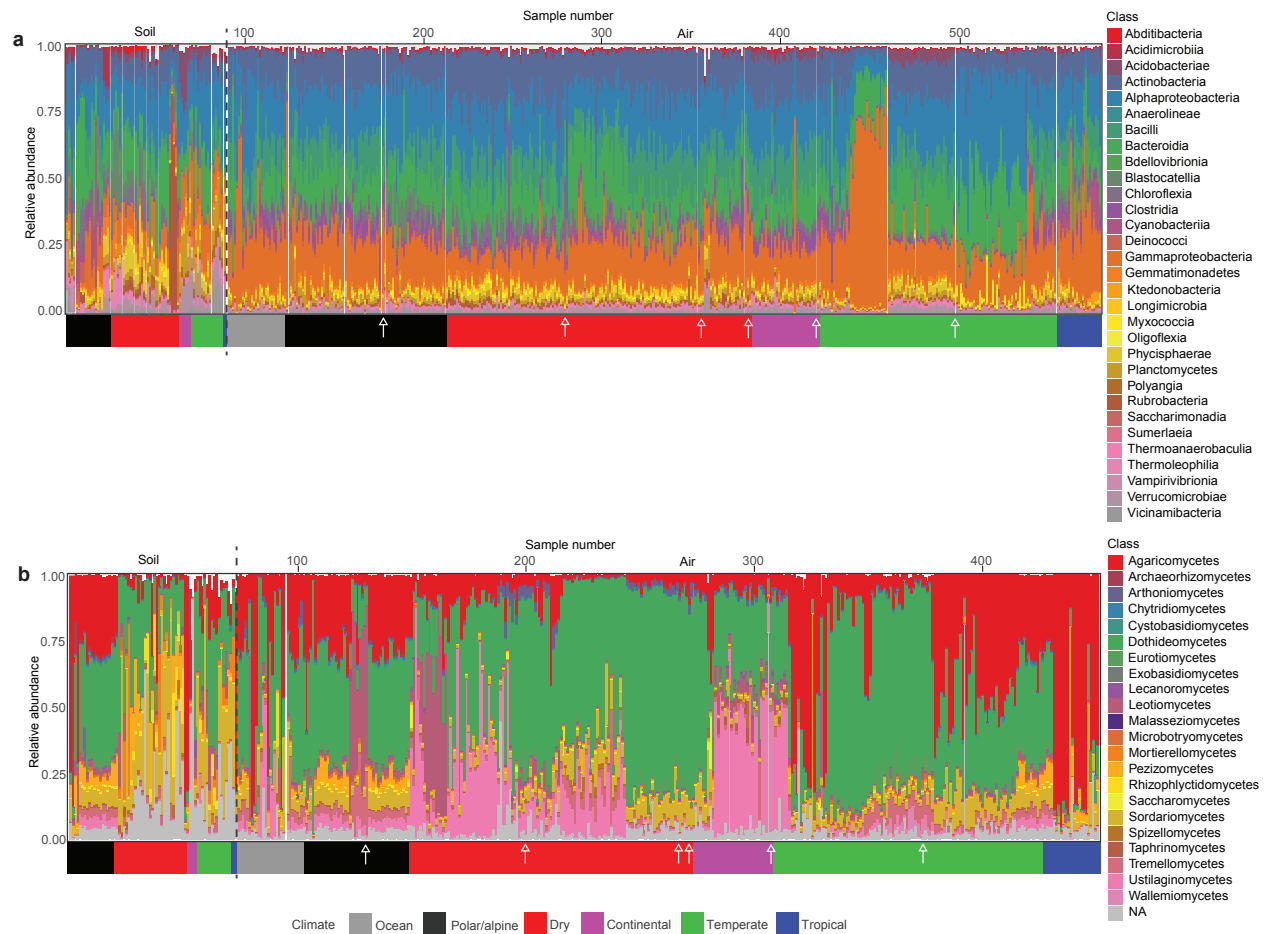


Fig. S24 | Co-occurrence of bacterial and fungal taxa in globally distributed air and soil. The scatterplots show number of ASVs (Y-axis) shared among \geq number of samples (X-axis). Generally, fewer ASVs were shared across larger number of samples. Shared ASVs were determined by simple presence/absence of a given ASV between two assemblages (Bacteria, $n = 529$; Fungi $n = 444$).

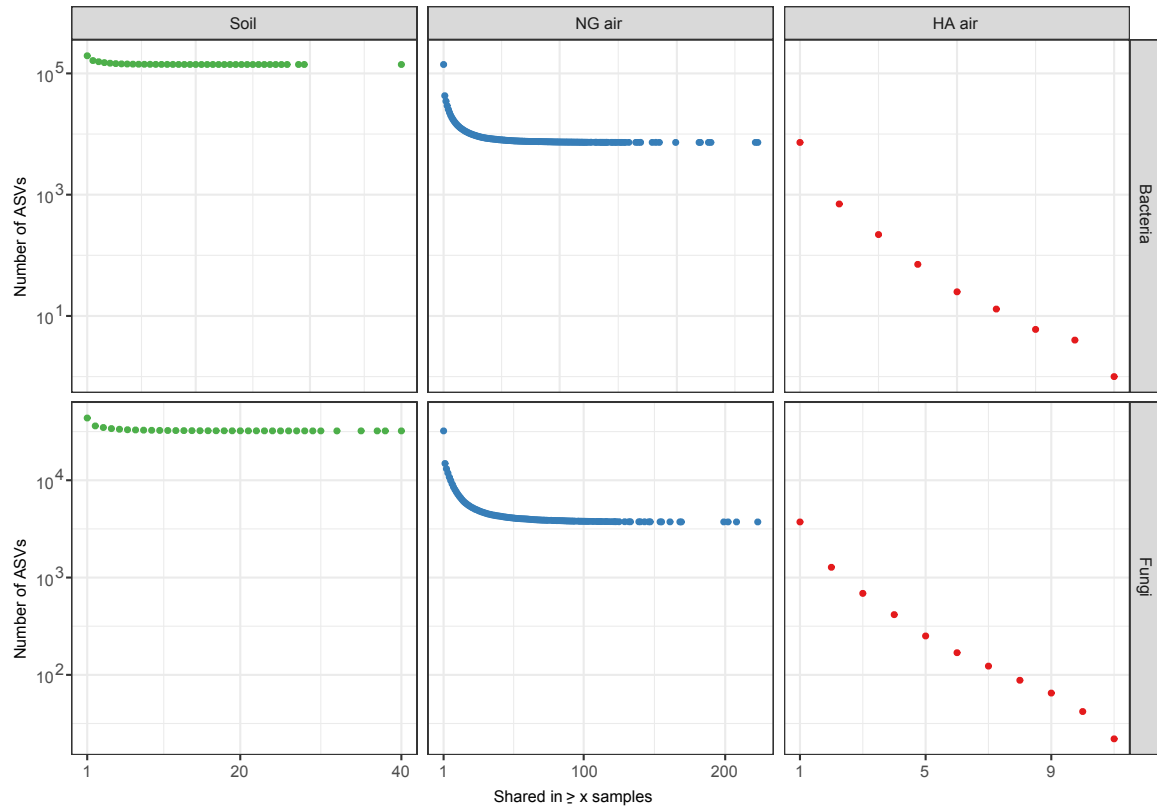


Fig. S25 | Source contribution to observed diversity in air. a) Estimated contribution of aquatic, phyllosphere and soil sources to bacterial diversity. **b)** Estimated contribution of aquatic, phyllosphere and soil sources to fungal diversity. Location 3 (Spain) was impacted by a minor intrusion of Sahara Desert atmospheric dust during sampling. HA air, high-altitude air; NG air, near-ground air. Top row: green shading indicates source locations sampled in this study, grey boxes indicate data from other studies employed in the meta-analysis. Bottom and right: coloured boxes correspond with climate as shown in Fig. 3. Numbers in each box indicate the estimate proportion of contribution from a given source (top row) in each sink community (left).

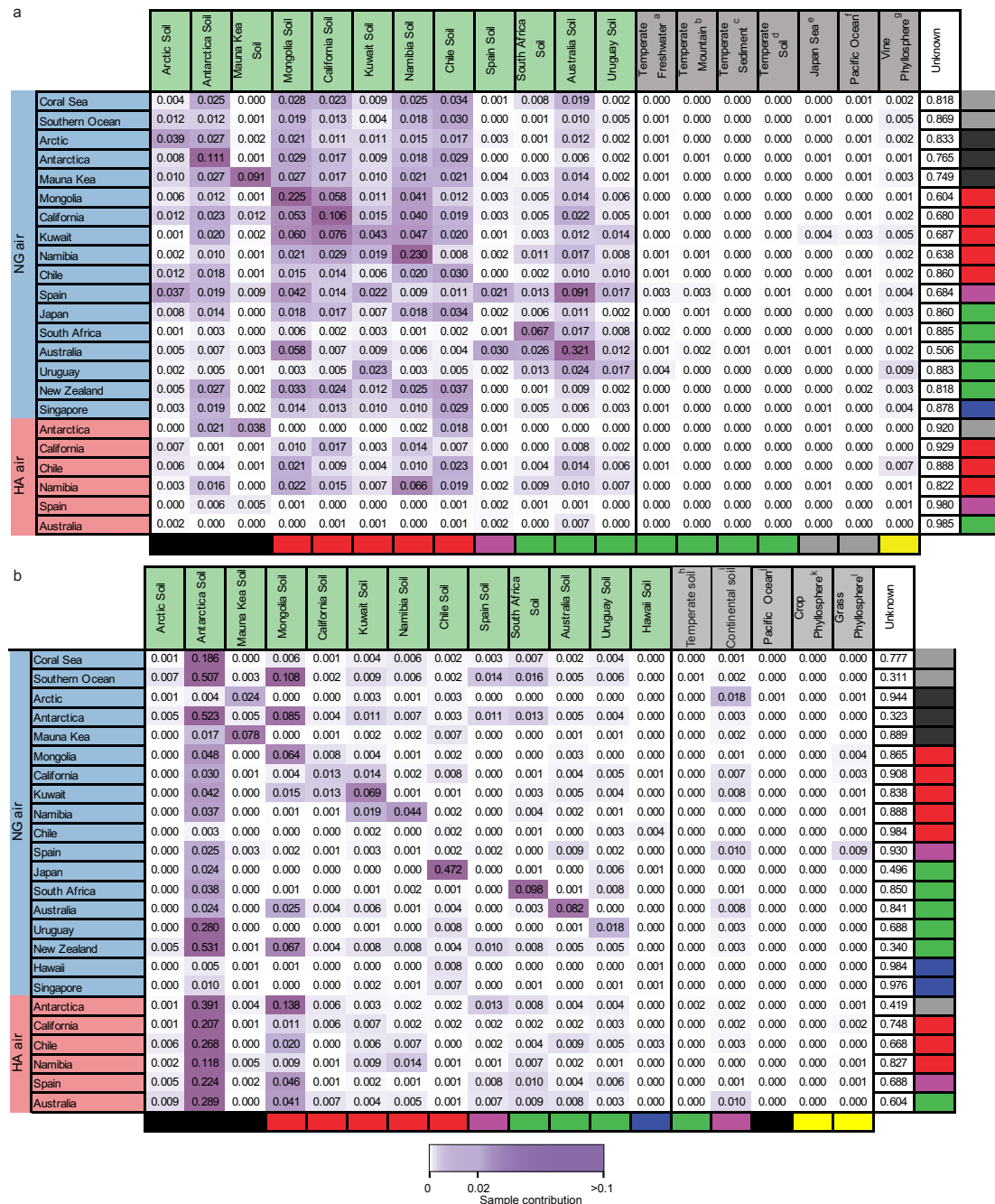
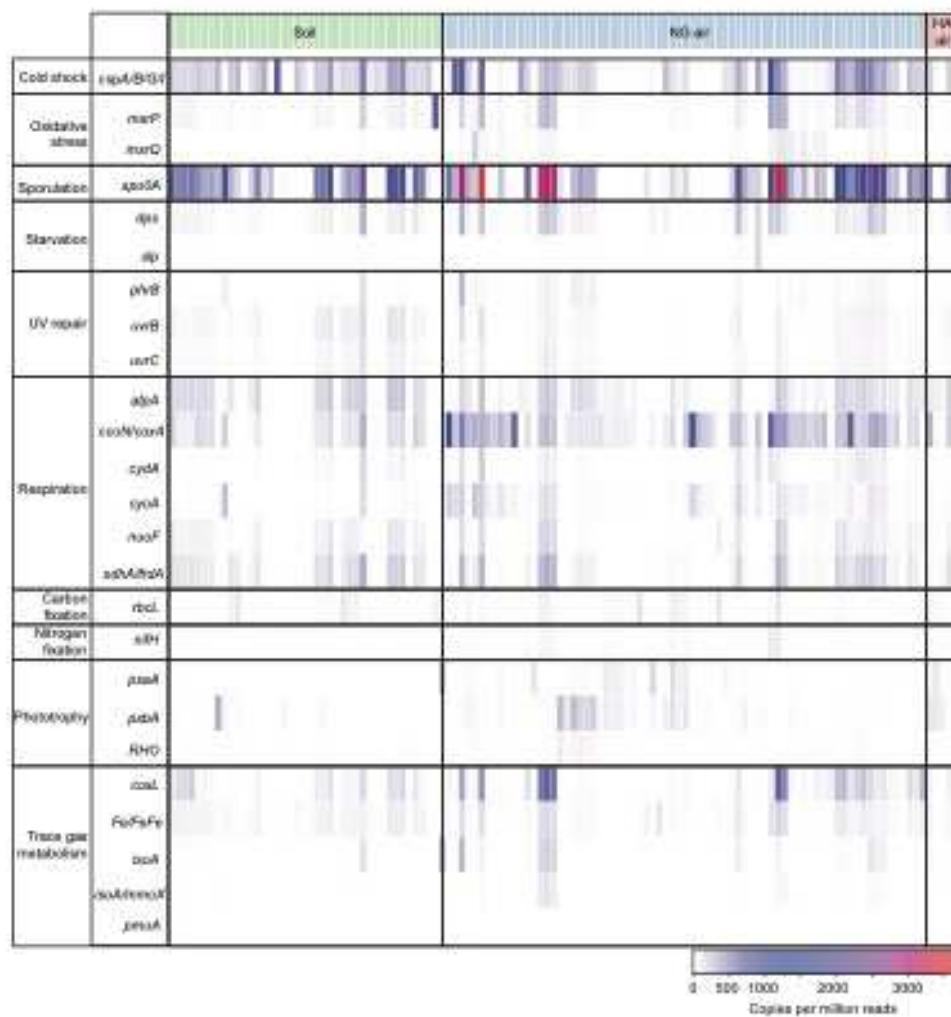


Fig. S26 | Functional metagenomics profiling of targeted stress-response and metabolic genes by habitat type. Abundance values for selected genes of the included stress-response and metabolic pathways are expressed on the heatmap as copies per million reads. The illustrated values represent all samples (total $n = 120$) within each air and soil sample group. HA air, high-altitude air; NG air, near-ground air.



Supplementary references

1. Peel MC, Finlayson BL. Updated world map of the Köppen-Geiger climate classification. *Hydrol Earth Syst Sci* 2007; **11**: 1633–1644.
2. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014; **12**: 87.
3. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol* 2019; **27**: 105–117.
4. Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, et al. Controlling for contaminants in low biomass 16S rRNA gene sequencing experiments. *mSystems* 2019; **4**: e00290-19.
5. Kiledal EA, Keffer JL, Maresca JA. Bacterial Communities in Concrete Reflect Its Composite Nature and Change with Weathering. *mSystems* 2021; **6**: e-01153-20.
6. Els N, Larose C, Baumann-Stanzer K, Tignat-Perrier R, Keuschnig C, Vogel TM, et al. Microbial composition in seasonal time series of free tropospheric air and precipitation reveals community separation. *Aerobiologia (Bologna)* 2019; **35**: 671–701.
7. Dray S, Legendre P, Peres-Neto PR. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol Modell* 2006; **196**: 483–493.
8. Rangel TF, Diniz-Filho JAF, Bini LM. SAM: a comprehensive application for Spatial Analysis in Macroecology. *Ecography (Cop)* 2010; **33**: 46–50.
9. Gotelli NJ. Null model analysis of species co-occurrence patterns. *Ecology* 2000; **81**: 2606–2621.
10. Gotelli NJ, Ulrich W. Statistical challenges in null model analysis. *Oikos* 2012; **121**: 171–180.
11. Dormann CF, Strauss R. A method for detecting modules in quantitative bipartite networks. *Methods Ecol Evol* 2014; **5**: 90–98.
12. Dormann CF, Frund J, Bluthgen N, Gruber B. Indices, Graphs and Null Models: Analyzing Bipartite Ecological Networks. *Open Ecol J* 2009; **2**: 7–24.
13. Payrató-Borràs C, Hernández L, Moreno Y. Breaking the Spell of Nestedness: The Entropic Origin of Nestedness in Mutualistic Systems. *Phys Rev X* 2019; **9**: 31024.
14. Artzy-Randrup Y, Stone L. Generating uniformly distributed random networks. *Phys Rev E* 2005; **72**: 56708.
15. Roberts ES, Coolen ACC. Unbiased degree-preserving randomization of directed binary networks. *Phys Rev E* 2012; **85**: 46103.
16. Squartini T, Garlaschelli D. Maximum-entropy networks: Pattern detection, network reconstruction and graph combinatorics. 2017. Springer.
17. Saracco F, Di Clemente R, Gabrielli A, Squartini T. Randomizing bipartite networks: the case of the World Trade Web. *Sci Rep* 2015; **5**: 10595.
18. Garlaschelli D, Loffredo MI. Maximum likelihood: Extracting unbiased information from complex networks. *Phys Rev E* 2008; **78**: 15101.
19. Squartini T, Mastrandrea R, Garlaschelli D. Unbiased sampling of network ensembles. *New J Phys* 2015; **17**: 23052.
20. Ulrich W, Almeida-Neto M, Gotelli NJ. A consumer's guide to nestedness analysis. *Oikos* 2009; **118**: 3–17.
21. Etienne RS. A neutral sampling formula for multiple samples and an 'exact' test of neutrality. *Ecol Lett* 2007; **10**: 608–618.
22. Kembel SW. Disentangling niche and neutral influences on community assembly : assessing the performance of community phylogenetic structure tests. *Ecol Lett* 2009; **12**: 949–960.
23. Caruso T, Chan Y, Lacap DC, Lau MCY, McKay CP, Pointing SB. Stochastic and deterministic processes interact in the assembly of desert microbial communities on a global scale. *ISME J* 2011; **5**: 1406–1413.
24. Caruso T, Hempel S, Powell JR, Barto EK, Rillig MC. Compositional divergence and convergence in arbuscular mycorrhizal fungal communities. *Ecology* 2012; **93**: 1115–1124.
25. Maaß S, Migliorini M, Rillig MC, Caruso T. Disturbance, neutral theory, and patterns of beta diversity in soil communities. *Ecol Evol* 2014; **4**: 4766–4774.