Supplementary Materials to:

Clouds influence the functioning of airborne microorganisms

Raphaëlle Péguilhan[1†], Florent Rossi[1‡], Muriel Joly[1], Engy Nasr[2], Bérénice Batut[2⁋], François Enault[3], Barbara Ervens[1], Pierre Amato[1]*

[1] Université Clermont Auvergne, Clermont Auvergne INP, CNRS, Institut de Chimie de Clermont-Ferrand, F-63000 Clermont–Ferrand, France.
[2] Department of Computer Science, University of Freiburg; Freiburg, 79110, Germany.
[3] Université Clermont Auvergne, CNRS, Laboratoire Microorganismes : Génome et Environnement (LMGE), F-63000 Clermont-Ferrand, France.
[†] Now at: Department of Chemical and Biochemical Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.
[‡] Now at: Département de Biochimie, de Microbiologie et de Bio-informatique, Faculté des Sciences et de Génie, Université Laval ; Québec, Canada.
[⁋] Now at: Institut Français de Bioinformatique, CNRS UAR 3601, France & Mésocentre Clermont-Auvergne, Université Clermont Auvergne, Aubiere, France.

**Correspondence to*: Pierre Amato (pierre.amato@uca.fr)

This supplement includes:

Supplementary text and references (Materials and methods)
Figures S1 to S16
Tables S1 to S4
Captions for Data S1 to S7 (separate electronic files)

Supplementary text and references (Materials and methods)

Bioinformatic workflow, differential expression analyses and data visualization

The bioinformatics workflow was elaborated by assembling relevant existing tools on a Galaxy instance (The Galaxy Community, 2022) deployed by the University Clermont Auvergne bioinformatics facility.

The preprocessing step consisted of (i) FastQC (v 0.72) (Andrews, 2010) for quality control analysis, (ii) Trimmomatic (v 0.36.6) (Bolger et al., 2014) to filter and trim erroneous reads, with an initial ILLUMINACLIP step for removing remaining Nextera paired-reads adapters, with a sliding window of 10:30, a minimum read length of 100 bp and the leading and trailing parameters with a quality threshold at 30. This step removed between 26% and 37 % of the reads in MGs (**Table S2**) and between 20 to 49 % of the reads in MTs (**Table S3**). Taxonomic affiliations were obtained from whole MG and MT datasets using Kraken2 (v 2.1.1) (Wood and Salzberg, 2014) against the "PlusPF" Kraken database (as of 2021-1-27) including known archaeal, bacterial, viral plasmid, human, protozoan, and fungal genomes, with a confidence score threshold of 0.1 and with the "report" and "report-zero-counts" options on. The trimmed reads were screened with SortMeRNA (v 2.1b.6) (Kopylova et al., 2012) to filter and recover the rRNA gene reads in separated files, with the default parameters, "paired-out" option, and all the available databases. Non-rRNA gene reads were processed for functional analyses. The proportions of rRNA gene reads in MG and MT datasets were generally between 1-2 % and 80-94%, respectively. Sample 20201124AIR's MT consisted of only 12% rRNA gene reads and was therefore excluded from further analysis (**Tables S2-S3**). Human reads were filtered from the non-rRNA gene reads using Bowtie2 (v 2.4.2) (Langmead and Salzberg, 2012), against the NCBI *Homo sapiens* genome "hg38_2021-5-18" with default parameters (**Tables S2-S3**). Human reads were excluded from further analyses.

To create a reference for all data, a non-redundant gene catalog was built using MGs (**Fig. S2**). Each individual dataset of non-RNA reads in MGs (each sample) was first *de novo* assembled using MEGAHIT (v 1.1.3.5) (Li et al., 2015), with default parameters and a minimum contig length of 500 bp. Depending on the sample, the number of assembled contigs ranged from ~43 000 to ~495 000 (**Table S2**), leading to 2 832 534 contigs with a maximal length of 21 000 to 200 000 bp, and a mean size of ~750 to 1 010 bp depending on samples. Genes were then predicted from these contigs using MetaGeneAnnotator (v 1.0.0) (Noguchi et al., 2008), with the "MetaGenomic" option and BED format as the output file. From the 2 832 534 contigs, 3 168 750 genes were predicted. Finally, to prevent redundancy, gene sequences were clustered at 95 % identity using CD-HIT (v 4.8.1) (Fu et al., 2012; Li and Godzik, 2006) and only the representative sequences for each cluster were kept. Only sequences >100 bp, and exhibiting >90 % identity with a reference sequence were kept. In total, 1 067 351 non-redundant genes were finally kept, with sequence length ranging from 100 to 22 065 bp and an average size of 330 bp. Functional annotation of the gene catalog was performed using DIAMOND (v 2.0.8.0) (Buchfink et al., 2015) in the "blastx" mode, with default parameters and the UniProtKB Swiss-Prot functional gene database (as of 2021_03) (The UniProt Consortium, 2019): 163 057 genes could be annotated, representing 40

264 unique UniProtKB entries, *i.e.*, ~7 % of the total number of entries in this database, illustrating the high biological diversity circulating in the atmosphere. Most (91.5 %) annotated genes are related to eukaryotes (~60 % to fungi), 23.6 % to viridiplantae, and 14.3 % to metazoa. Bacteria, viruses and archaea contributed 7.6 %, 0.6 %, and 0.3 % of the annotated genes, respectively.

Non-rRNA gene sequences from all MGs and MTs were finally mapped to the gene catalog to obtain read counts per gene using BWA-MEM (v 0.7.17.1) (Li and Durbin, 2009) with default parameters. Percentages of properly mapped reads against the gene catalog ranged between ~4 % and ~17 % for MGs and between ~3% and ~10% for MTs (**Tables S2-S3**). Only genes with >10 mapped sequences in MGs were considered, and the count tables for MGs and MTs were filtered in order to remove genes affiliated with "Embryophytes" and "Metazoa" and focus on microbial genes. Finally, unique 21 046 unique microbial genes (over 40 264 in total) were retained for downstream analyses. Annotated sequences were grouped according to their respective Gene Ontology terms (GOs) (Ashburner et al., 2000; The Gene Ontology Consortium, 2021).

Data normalization and differential expression analysis (DEA) were performed using the R package MTXmodel (R v4.0.3; MTXmodel v1.5.1) (Zhang et al., 2021). To highlight overall functional expression patterns independently from atmospheric conditions, this was run using the following options, with MT and MG datasets as the input files consisted of: no transformation, clr (centered-log ratio) normalization, LM analysis method, BH correction method, min abundance at 0.0001, min prevalence at 0.5, max significance at 0.25, and "DataType" (*i.e.*, MTs or MGs) as a fixed effect. DEA provides relative expression coefficients based on the selected fixed effect. Positive coefficients (coeff) here indicate features (taxonomic groups or genes) significantly more represented in MTs than in MGs, while negative coefficients indicate overrepresentation in MGs (i.e., no significant expression), with higher absolute values for higher representations.

To identify the functions and genes whose expression was related to atmospheric conditions, "EnvType" was selected as the fixed effect (cloud or clear condition), with MTs as input files, and MGs as "DNA data" for normalization. The environment type "cloud" was used as the reference, so positive coefficients (coeff) here indicate the features significantly more expressed in clouds compared with clear conditions, and the opposite for negative values.

Data visualizations were designed with the following R packages: *ggplot2* (v3.4.1), *ggrepel* (v0.9.3), *ggsignif* (v0.6.4), *ggdendro* (v0.1.23), *factoextra* (v1.0.7), *gridExtra* (v2.3), *vegan* (v2.6-4), *pheatmap* (v1.0.12). Relationship networks for GOs were generated using OLSVis (Vercruysse et al., 2012) and Cytoscape (v3.9.1). Metabolic pathways were generated from KEGG database (Kanehisa et al., 2023).

References cited:

Andrews, S.: FastQC: a quality control tool for high throughput sequence data, 2010.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nature Genetics, 25, 25–29, https://doi.org/10.1038/75556, 2000.

Bolger, A. M., Lohse, M., and Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics, 30, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170, 2014.

Buchfink, B., Xie, C., and Huson, D. H.: Fast and sensitive protein alignment using DIAMOND, Nature Methods, 12, 59–60, https://doi.org/10.1038/nmeth.3176, 2015.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W.: CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics, 28, 3150–3152, https://doi.org/10.1093/bioinformatics/bts565, 2012.

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M.: KEGG for taxonomy-based analysis of pathways and genomes, Nucleic Acids Research, 51, D587–D592, https://doi.org/10.1093/nar/gkac963, 2023.

Kopylova, E., Noé, L., and Touzet, H.: SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data, Bioinformatics, 28, 3211–3217, https://doi.org/10.1093/bioinformatics/bts611, 2012.

Langmead, B. and Salzberg, S. L.: Fast gapped-read alignment with Bowtie 2, Nature Methods, 9, 357–359, https://doi.org/10.1038/nmeth.1923, 2012.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W.: MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinformatics, 31, 1674–1676, https://doi.org/10.1093/bioinformatics/btv033, 2015.

Li, H. and Durbin, R.: Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics, 25, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324, 2009.

Li, W. and Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics, 22, 1658–1659, https://doi.org/10.1093/bioinformatics/btl158, 2006.

Noguchi, H., Taniguchi, T., and Itoh, T.: MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes, DNA Research, 15, 387–396, https://doi.org/10.1093/dnares/dsn027, 2008.

The Galaxy Community: The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, Nucleic Acids Research, 50, W345–W351, https://doi.org/10.1093/nar/gkac247, 2022.

The Gene Ontology Consortium: The Gene Ontology resource: enriching a GOld mine, Nucleic Acids Research, 49, D325–D334, https://doi.org/10.1093/nar/gkaa1113, 2021.

The UniProt Consortium: UniProt: a worldwide hub of protein knowledge, Nucleic Acids Research, 47, D506–D515, https://doi.org/10.1093/nar/gky1049, 2019.

Vercruysse, S., Venkatesan, A., and Kuiper, M.: OLSVis: an animated, interactive visual browser for bio-ontologies, BMC Bioinformatics, 13, 116, https://doi.org/10.1186/1471-2105-13-116, 2012.

Wood, D. E. and Salzberg, S. L.: Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biology, 15, R46, https://doi.org/10.1186/gb-2014-15-3-r46, 2014.

Zhang, Y., Thompson, K. N., Huttenhower, C., and Franzosa, E. A.: Statistical approaches for differential expression analysis in metatranscriptomics, Bioinformatics, 37, i34–i41, https://doi.org/10.1093/bioinformatics/btab327, 2021.

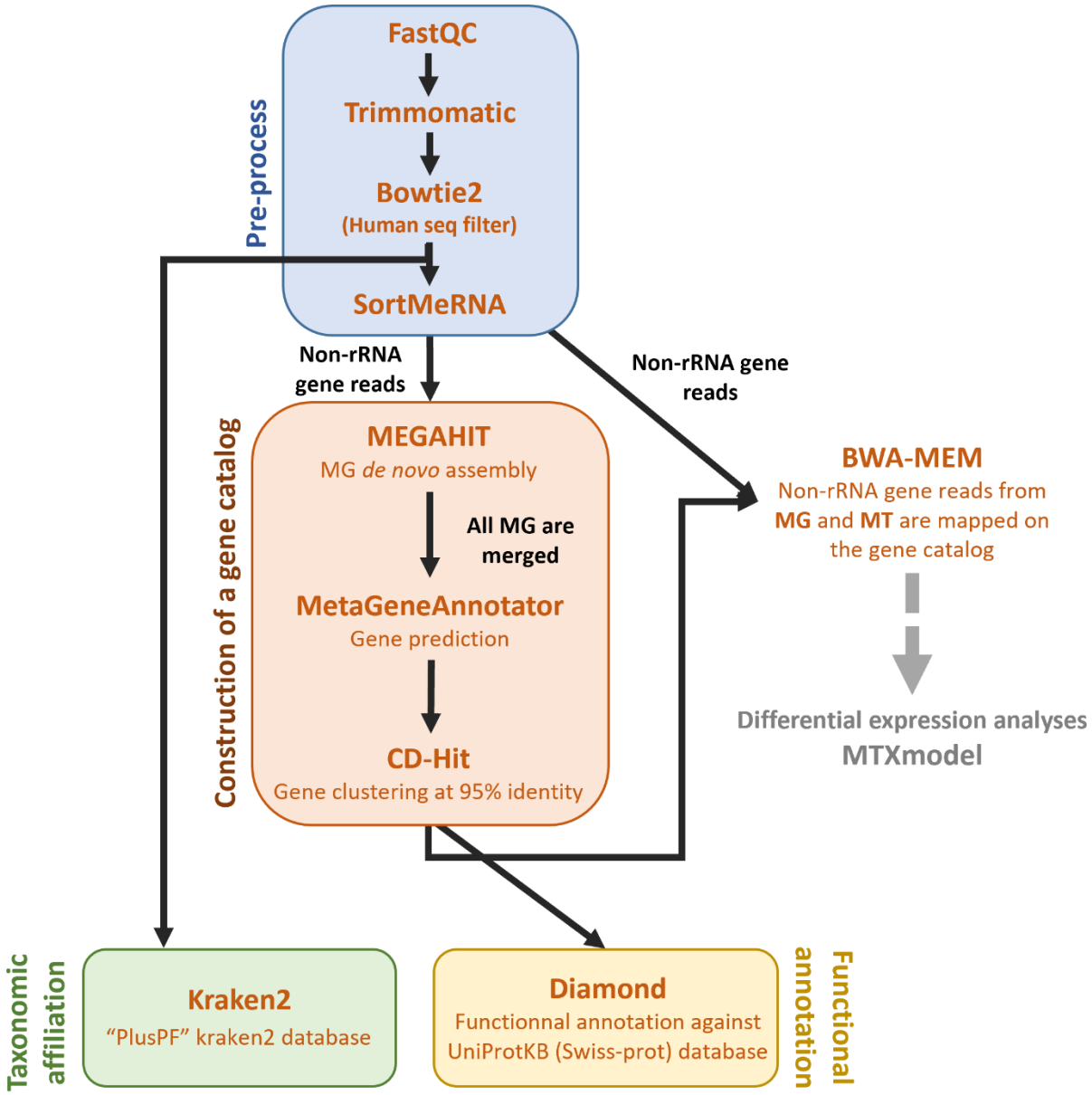

Fig S1.

Main steps of the bioinformatics workflow elaborated for this study. This uses the following tools, on a Galaxy environment: FastQC (v0.72), Trimmomatic (v0.36.6), Bowtie2 (v2.4.2), SortMeRNA (v2.1b.6), MEGAHIT (v1.1.3.5), MetaGeneAnnotator (v1.0.0), CD-Hit (v4.8.1), Kraken2 (v2.11), Diamond (v2.0.8.0), BWA-MEM (v0.7.17.1), and the MTXmodel R package (v1.5.1) (see supplementary text for references).

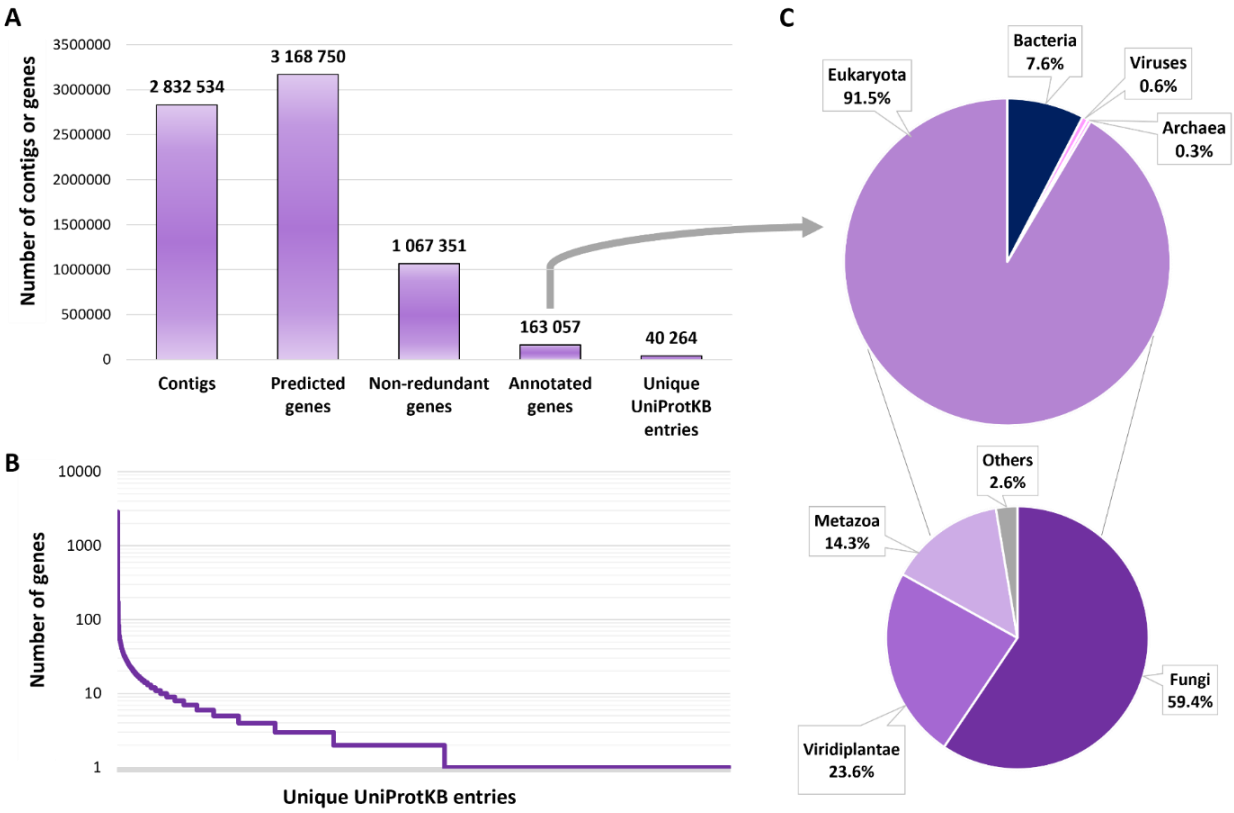

Fig S2.

Content of the gene catalog used as a common reference to all datasets, elaborated by merging all the metagenomes of the study and curated for redundancy. **A**. Contig or gene numbers at each step of the gene catalog construction, from left to right; **B.** Rank-abundance plot representing the number of genes associated with unique UniProtKB entries; **C**. Taxonomic affiliations associated with annotated genes; the taxa distribution in Eukaryotes is specified in the lower pie-chart.

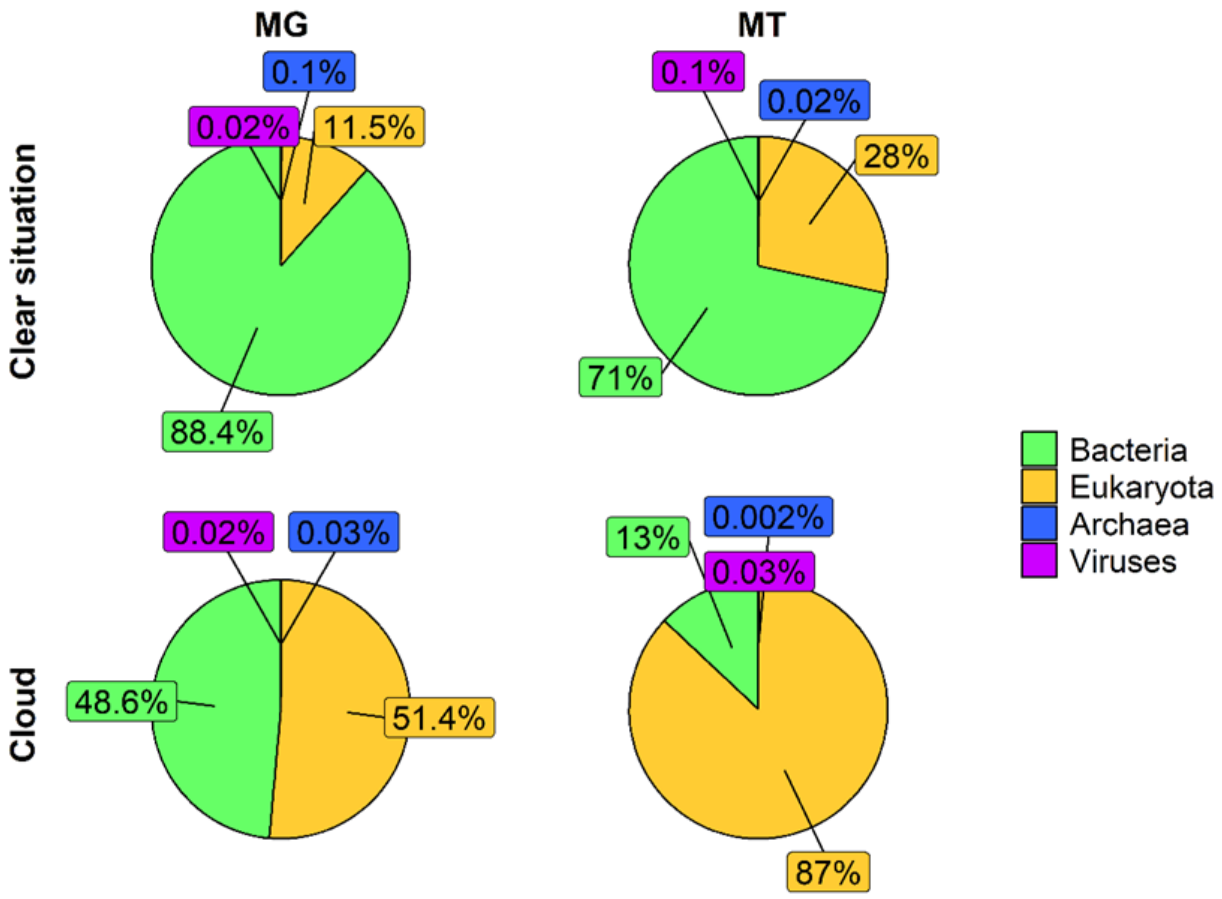

Fig S3.

Average proportions of reads associated with Bacteria, Eukaryota, Archaea, and Viruses in metagenomes (MG, left) and metatranscriptomes (MT, right), during cloudy (lower charts) and clear conditions (upper charts).

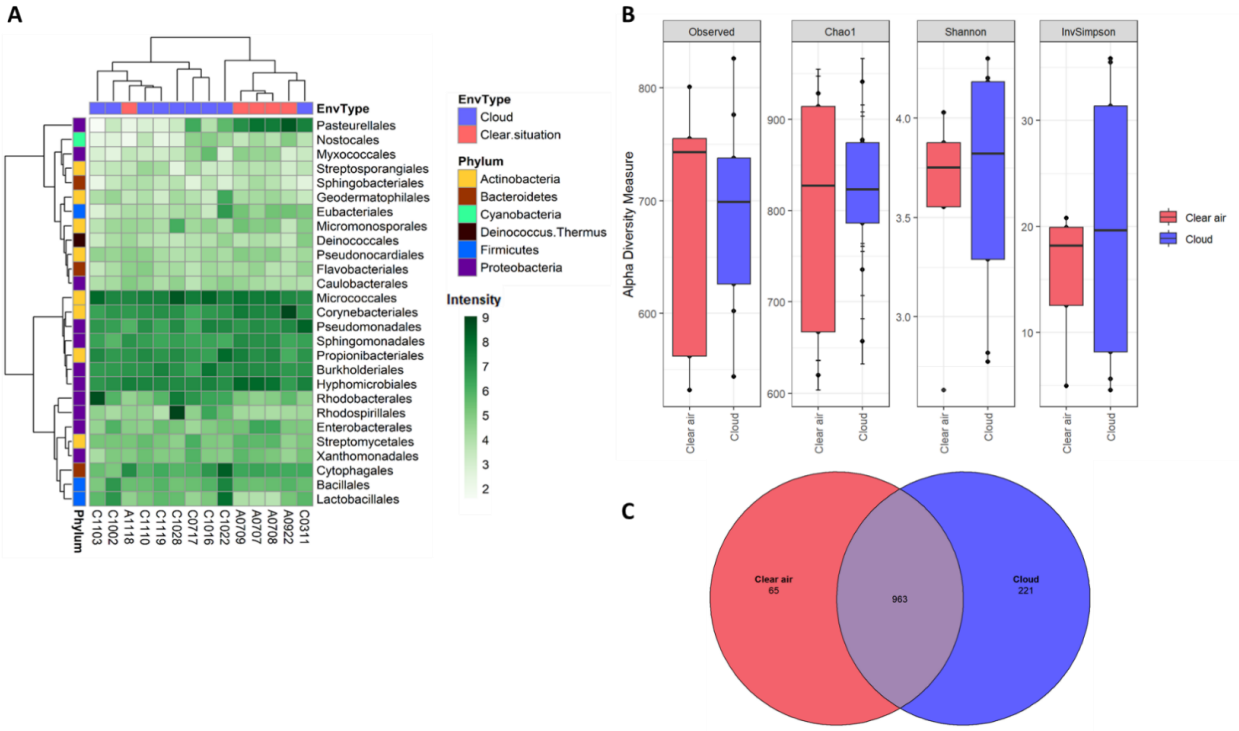

Fig S4.

Bacteria diversity. **(A)** Distribution of the most abundant bacterial orders in the metagenomes, and corresponding hierarchical clusterings (Ward's method, "ward.D2"). The intensity scale depicts centered-log ratio (clr) abundances. EnvType: environment type. The samples are named as follows: "A" for clear conditions (air) or "C" for cloud, followed by the sampling date in the format "mmdd" (month and day); **(B)** Alpha diversity indexes (observed and estimated richness, Shannon's diversity and Inverse Simpson evenness) in clouds and clear atmosphere samples; **(C)** Venn diagram depicting the distribution of richness between clouds and clear atmosphere samples at the genus level.

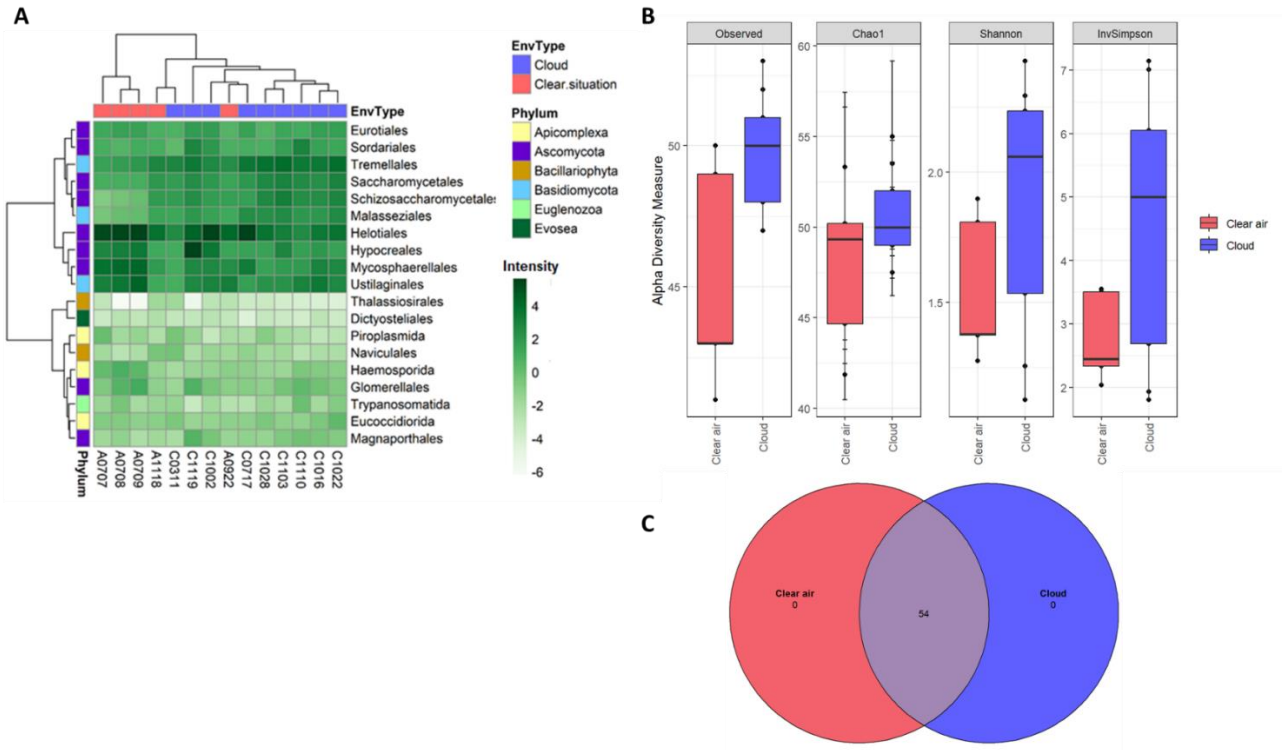

Fig S5.

Eukaryotic diversity. **(A)** Distribution of eukaryotic orders in the metagenomes, and corresponding hierarchical clusterings (Ward's method, "ward.D2"). The intensity scale depicts centered-log ratio (clr) abundances. EnvType: environment type. The samples are named as follows: "A" for clear conditions (air) or "C" for cloud, followed by the sampling date in the format "mmdd" (month and day); **(B)** Alpha diversity indexes (observed and estimated richness, Shannon's diversity and Inverse Simpson evenness) in clouds and clear atmosphere samples; **(C)** Venn diagram depicting the distribution of richness between clouds and clear atmosphere samples at the genus level.

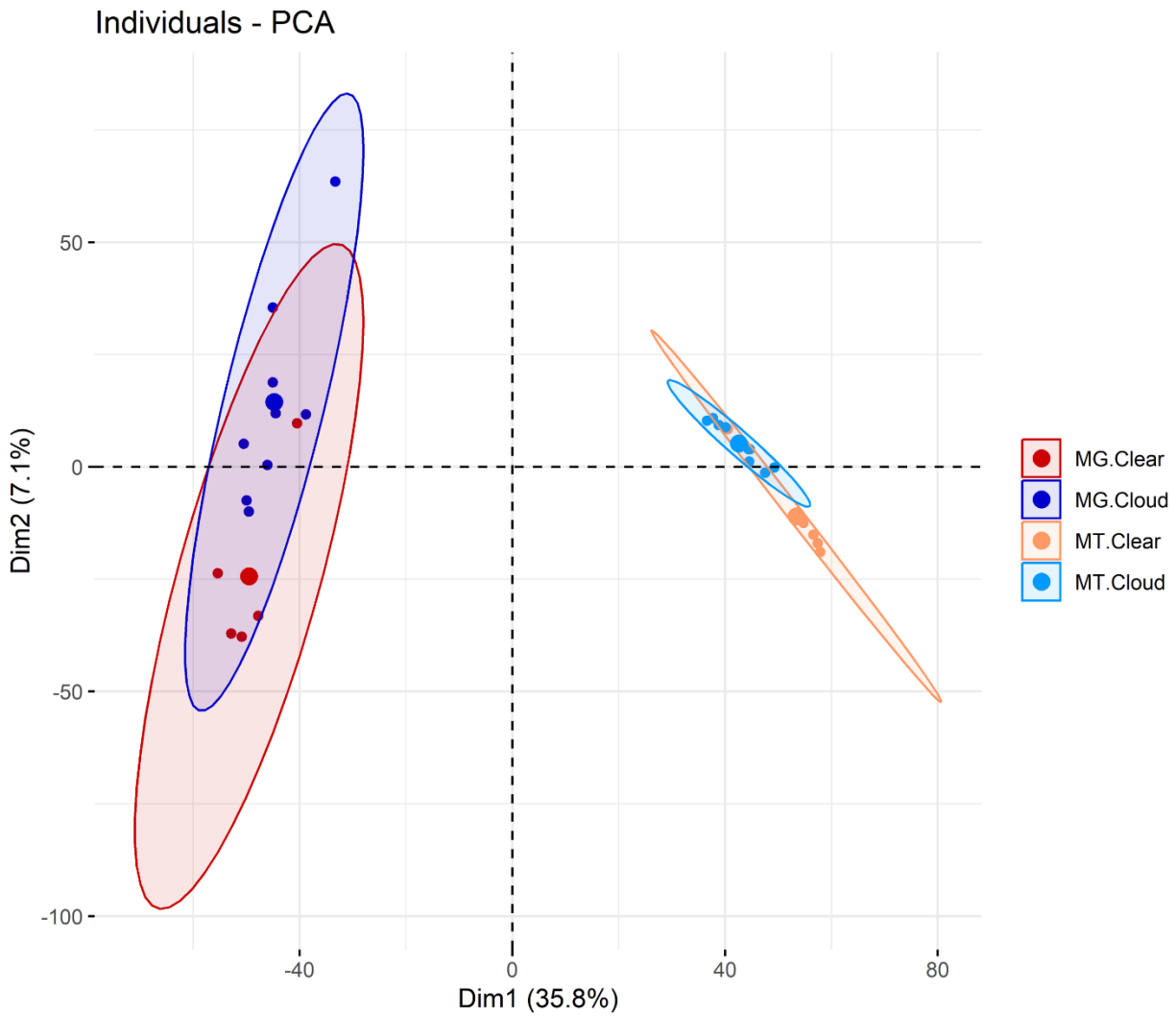

Fig S6.

Principal component analysis (PCA) representing taxonomy distribution in metagenomes (MG) and metatranscriptomes (MT) for clouds and clear atmospheric conditions, based on 6 373 unique taxa. Count data were centered-log ratio (clr) transformed. Ellipses indicate 95% confidence levels.

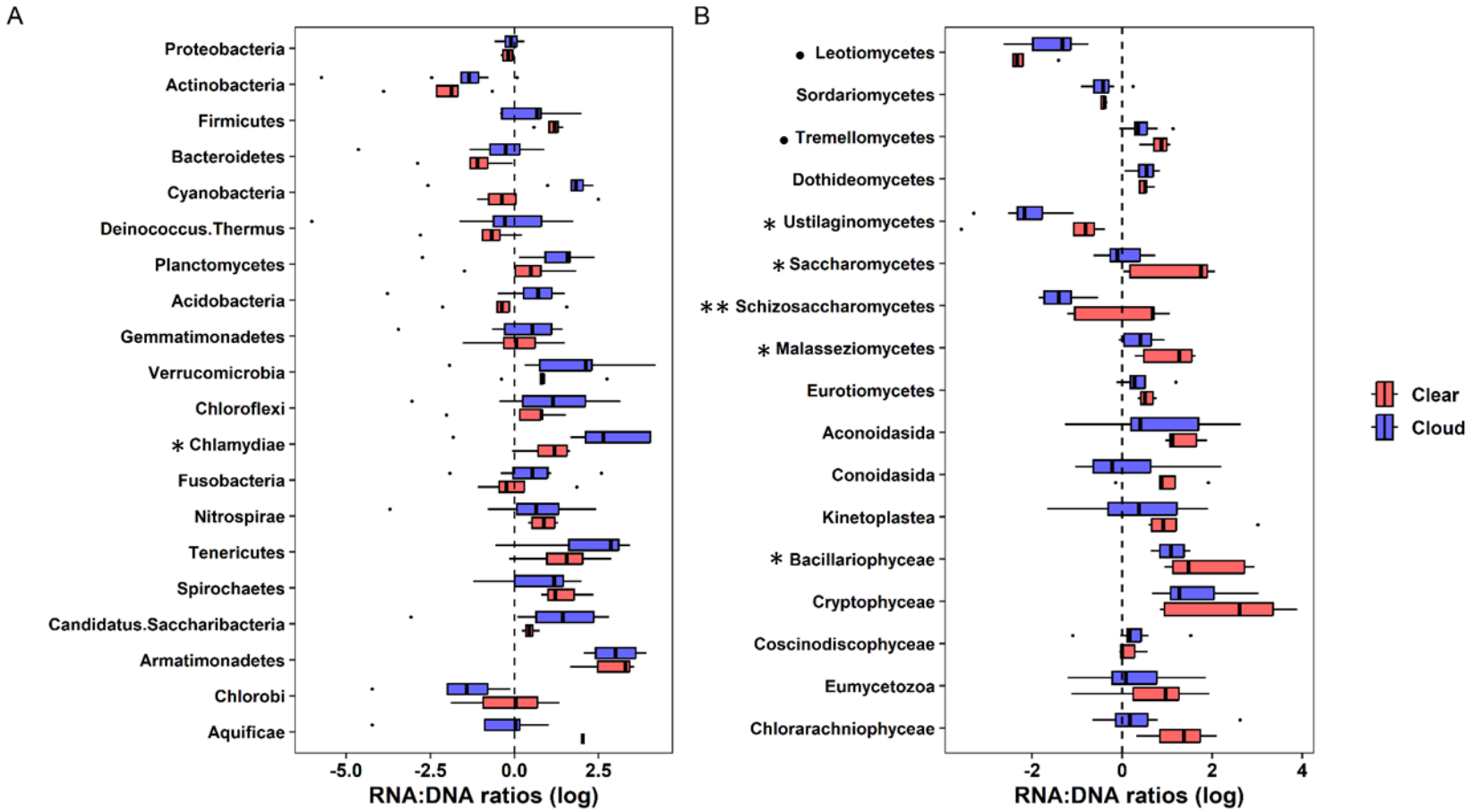

Fig S7.

Relative representation of microbial taxa in metatranscriptomes compared with metagenomes (termed as RNA:DNA ratio) in clouds (blue) and in clear atmospheric conditions (red) for **(A)** the 20 most represented phyla of bacteria and **(B)** classes of eukaryotes. Taxa are ordered in descending order of the number of reads in metagenomes, from top to bottom. P-value: $0 <$ *** $< 0.001 \leq$ ** $< 0.01 \leq$ * $< 0.05 \leq$ • < 0.1 (calculated using *indicspecies* R package, "r.g" method).

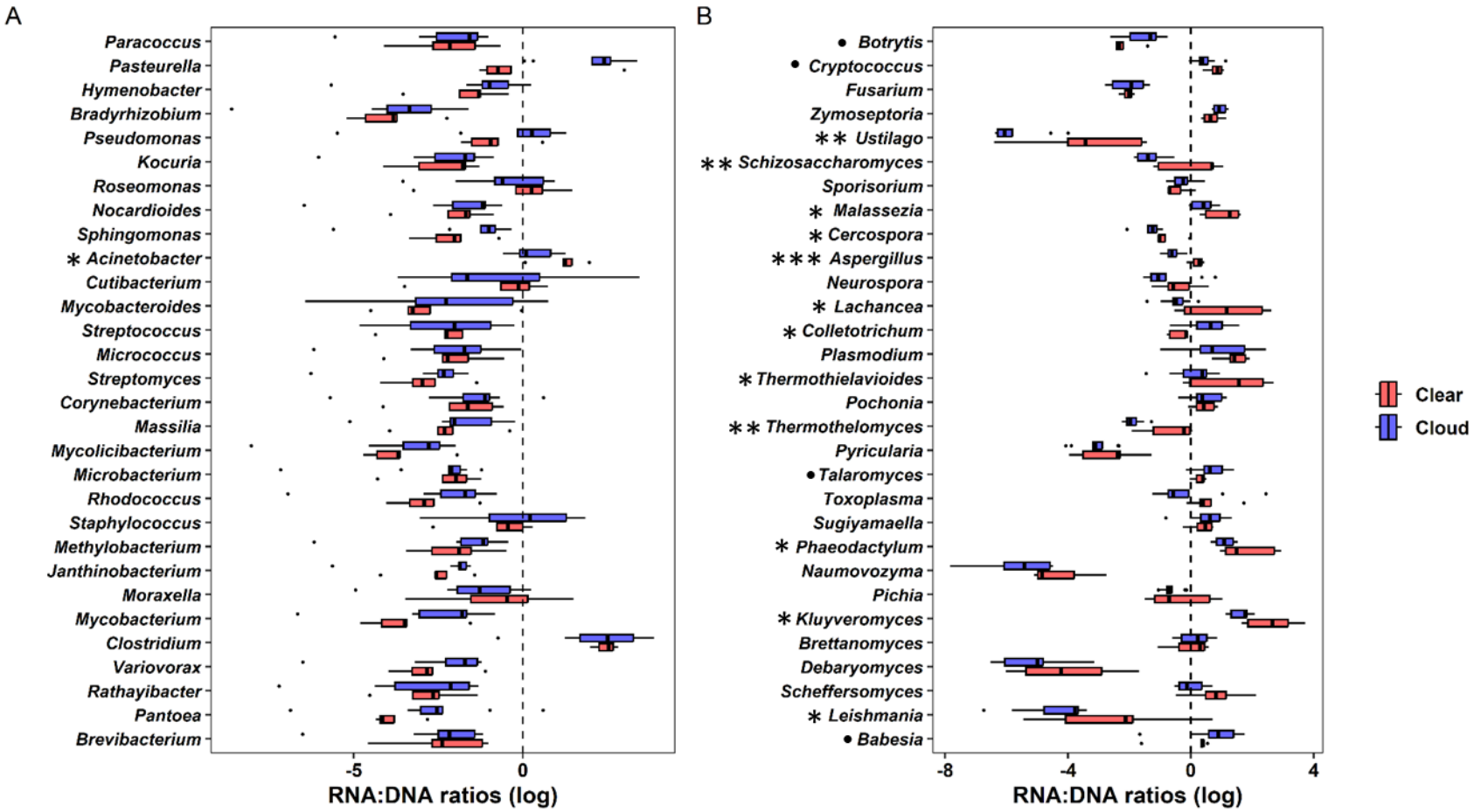

Fig S8.
Same as **Fig S7**, at the genus levels for **(A)** bacteria (30 most represented) and **(B)** eukaryotes.

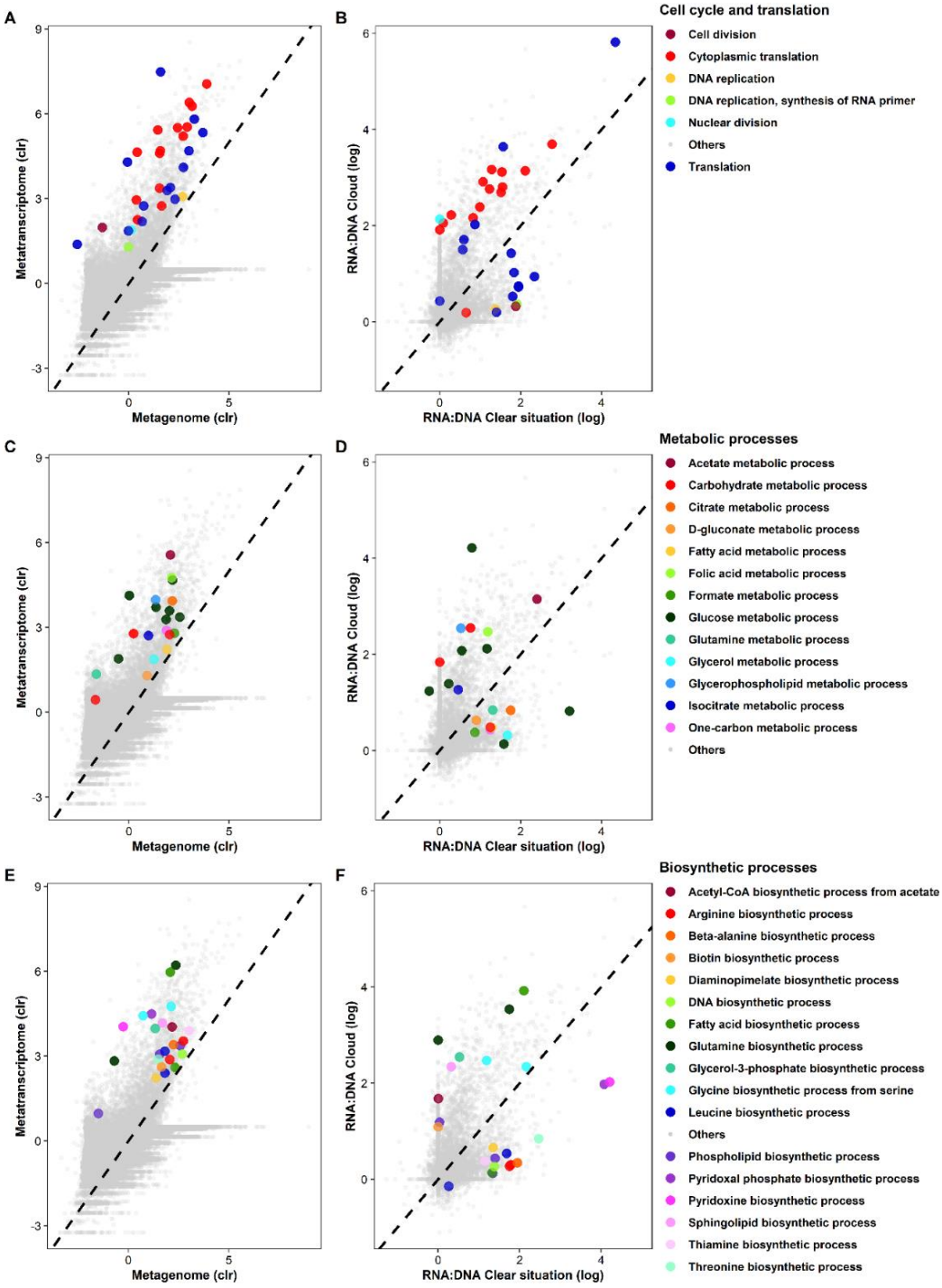

Fig S9.

GO terms representation in genes and transcripts in metagenomes and metatranscriptomes (**A; C; E**), and **(B; D; F)** relative representation (termed as RNA:DNA) in metatranscriptomes in clouds *versus* clear conditions as compared with the corresponding metagenomes, for Biological Processes related to: cell cycle and translation (**A; B**), metabolic processes (**C; D**), and biosynthetic processes (**E; F**); clr: centered-log ratio transformation.

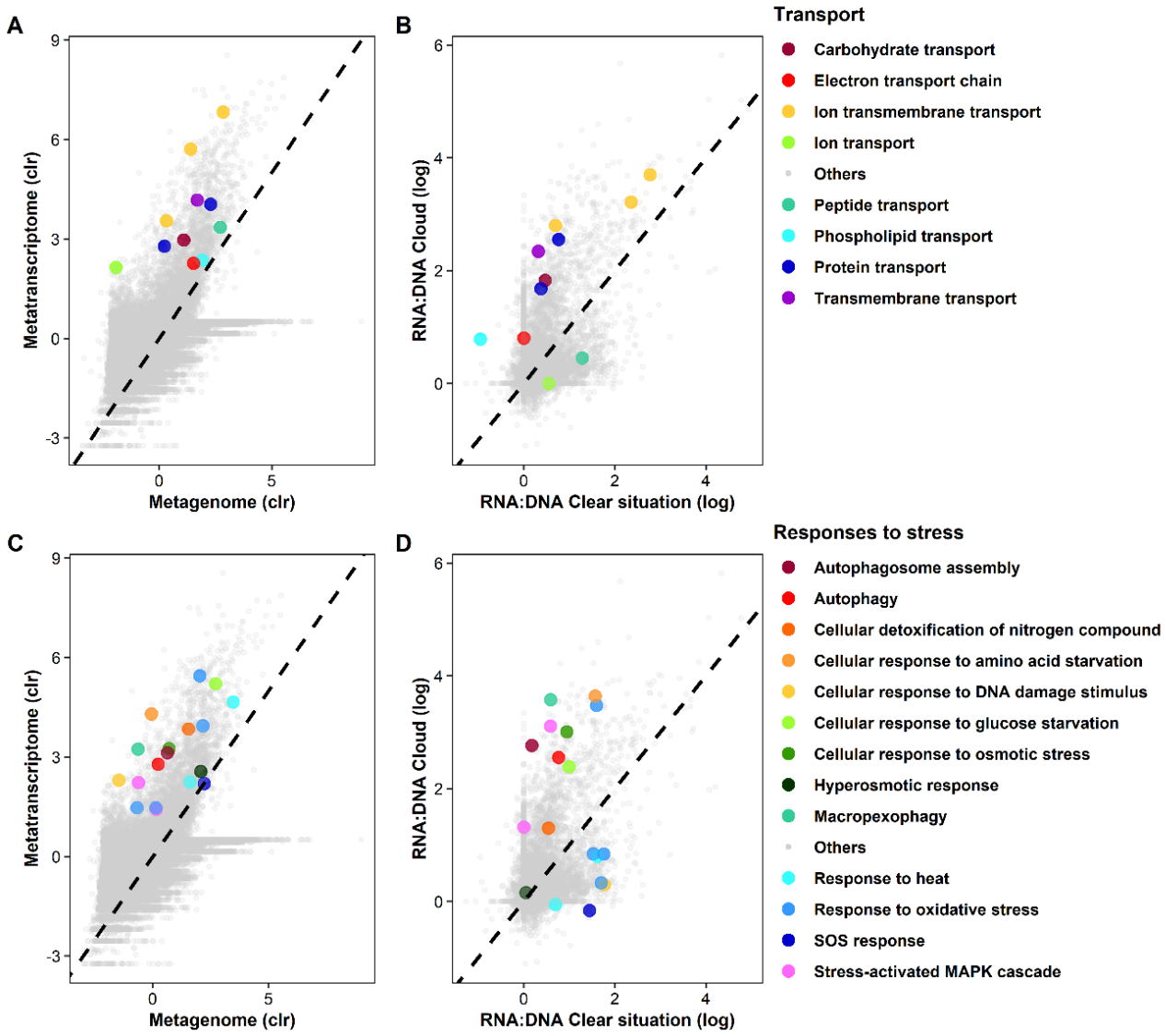

Fig S10.

Same as **Fig S9** for Biological Processes GO terms related to: transport (**A; B**) and responses to stress (**C; D**).

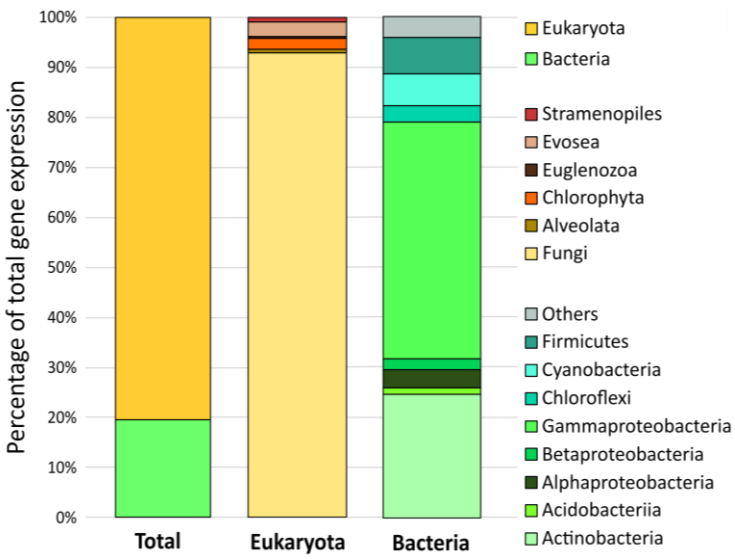

Fig S11.
Distribution of the taxonomic affiliations associated with the 488 transcripts detected overrepresented by differential expression analysis (DEA).

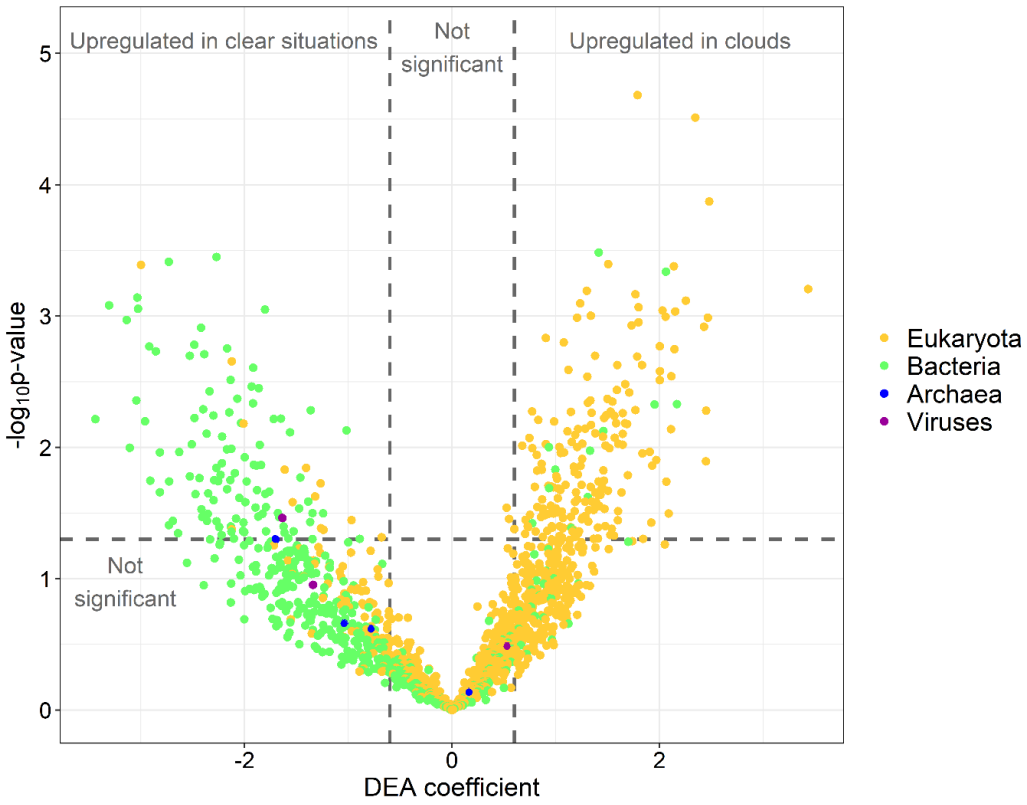

Fig S12.
Volcano plot representing the taxonomic affiliation at the kingdom level of differentially represented transcripts between clouds (positive coefficients) and clear conditions (negative coefficients); dashed lines indicate significance thresholds of the DEA analysis.

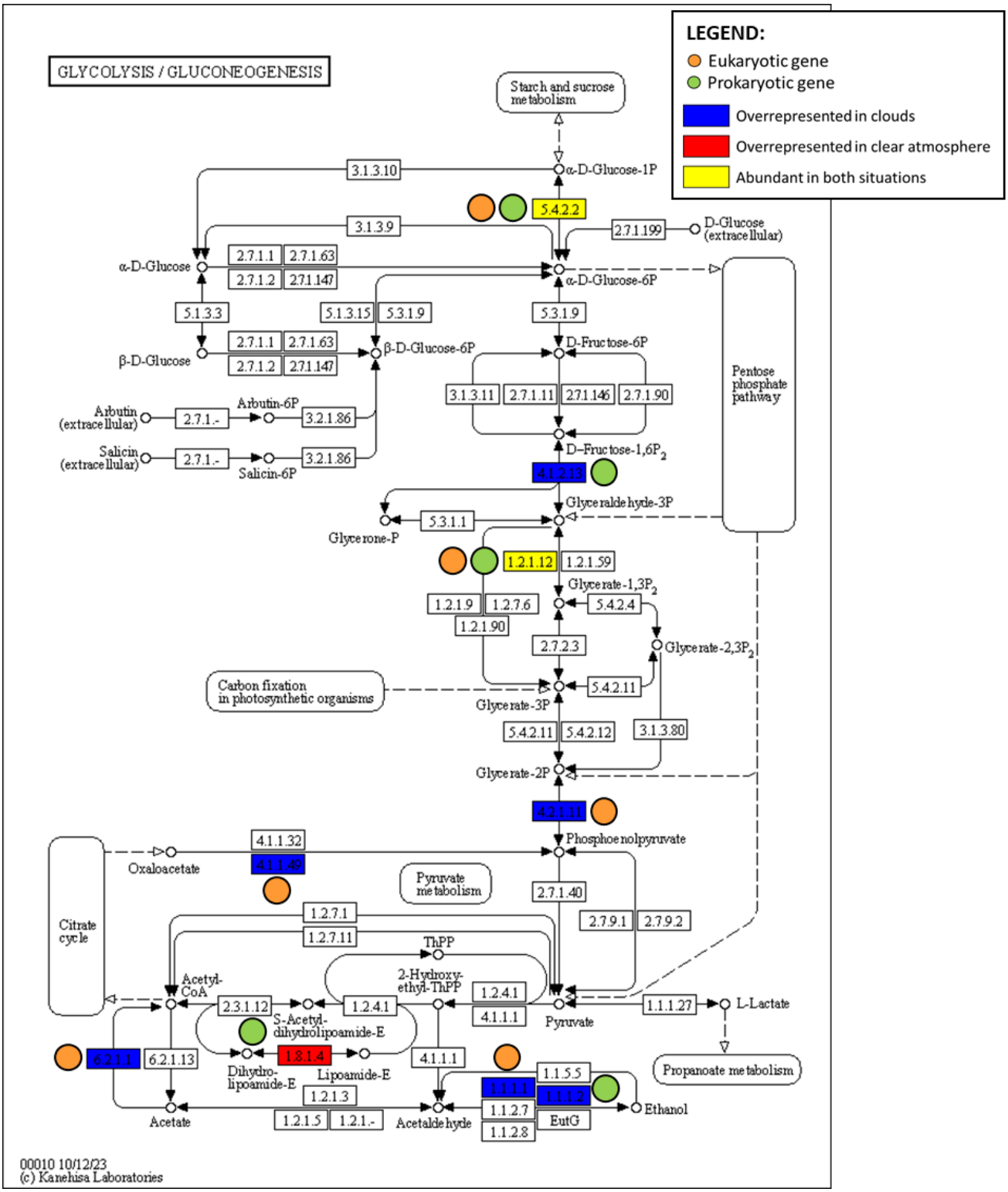

Fig S13.

Metabolic pathways related to glycolysis/gluconeogenesis, depicting overrepresented enzyme transcripts in clouds and/or in clear atmosphere by eukaryotes and/or by prokaryotes (from UniprotKB identifiers and KEGG database).

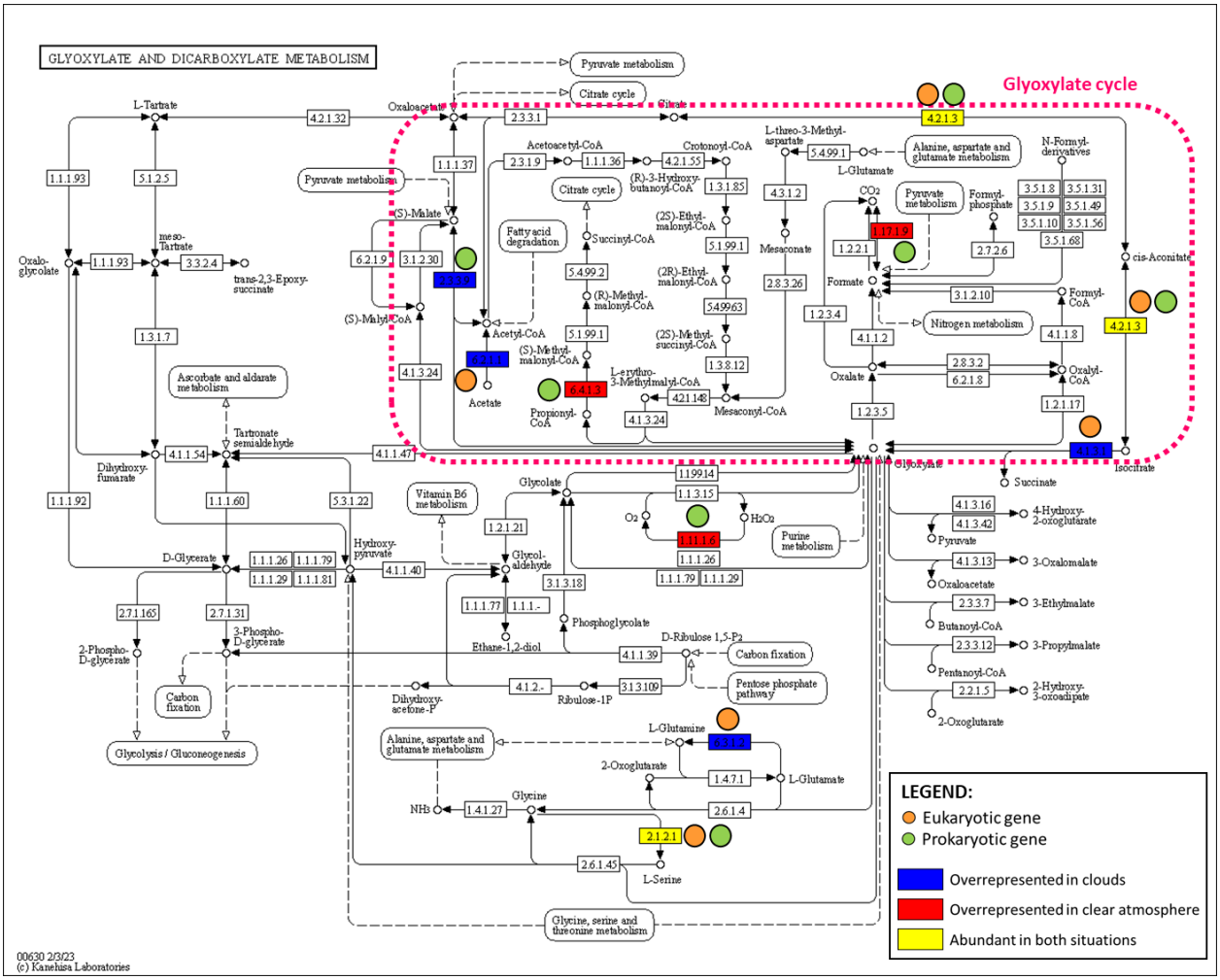

Fig S14.

Metabolic pathways related to glyoxylate and dicarboxylate metabolism, depicting overrepresented enzyme transcripts in clouds and/or in clear atmosphere by eukaryotes and/or by prokaryotes (from UniprotKB identifiers and KEGG database).

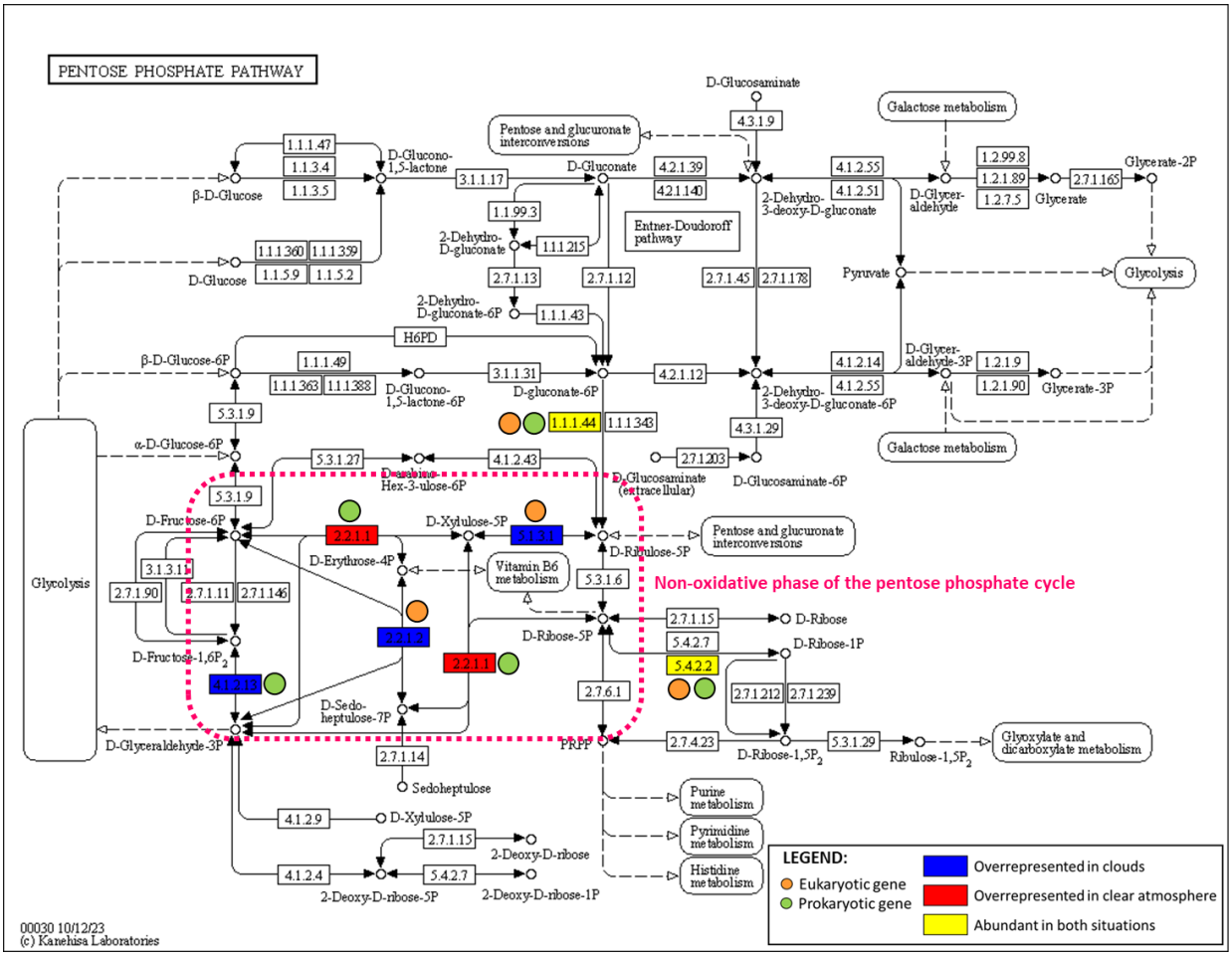

Fig S15.

Metabolic pathways related to pentose phosphate metabolism, depicting overrepresented enzyme transcripts in clouds and/or in clear atmosphere by eukaryotes and/or by prokaryotes (from UniprotKB identifiers and KEGG database).

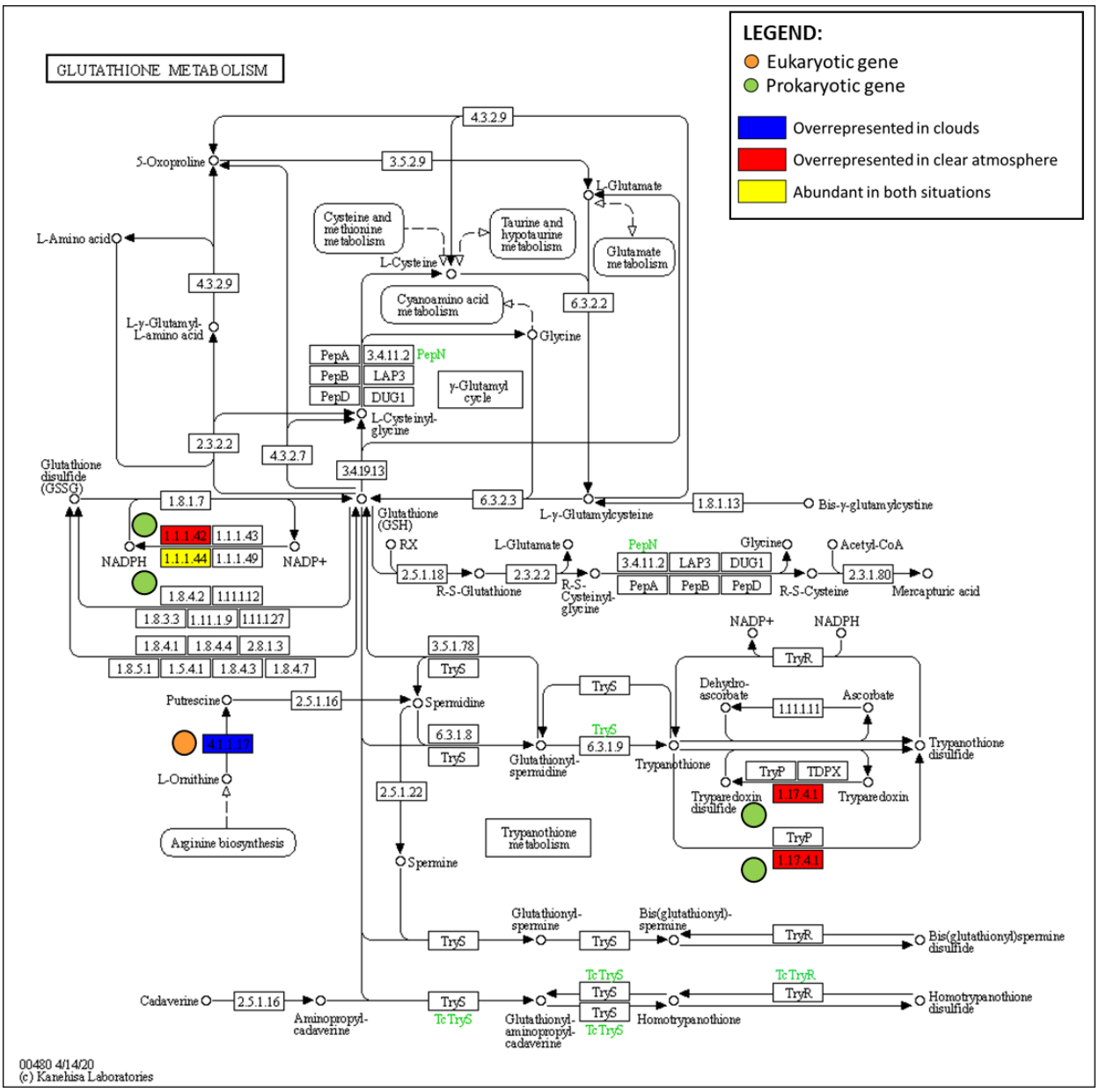

Fig S16.

Metabolic pathways related to glutathione metabolism, depicting overrepresented enzyme transcripts in clouds and/or in clear atmosphere by eukaryotes and/or by prokaryotes (from UniprotKB identifiers and KEGG database).

Table S1.

Nucleic acid concentrations in the air volumes sampled during cloudy and clear conditions, and corresponding numbers of annotated genes.

Sample ID	Total DNA concentration ($ng.m^{-3}$ of air) [#]	Total RNA concentration ($ng.m^{-3}$ of air) [#]	RNA to DNA concentration ratio	Number of annotated genes in MG	Number of annotated genes in MT
CLEAR CONDITIONS					
20200707AIR	2.19	0.47	0.21	5 463	2 237
20200708AIR	1.35	0.38	0.28	3 287	2 095
20200709AIR	0.87	0.33	0.38	6 649	1 676
20200922AIR	0.70	0.68	0.98	11 690	2 248
20201118AIR	0.14	0.23	1.64	15 431	4 471
20201124AIR	0.12	0.06	0.53	-	-
Minimum	**0.12**	**0.06**	**0.21**	**3 287**	**1 676**
Maximum	**2.19**	**0.68**	**1.64**	**15 431**	**4 471**
Median	**0.78**	**0.36**	**0.46**	**6 649**	**2 237**
Mean	**0.89**	**0.36**	**0.67**	**8 504**	**2 545**
Standard error	**0.78**	**0.21**	**0.55**	**4 951**	**1 101**
CLOUDS					
20191002CLOUD	0.18	0.32	1.77	19 010	2 219
20191022CLOUD	0.24	0.65	2.70	14 804	3 855
20200311CLOUD	0.31	0.62	1.98	14 220	985
20200717CLOUD	0.33	0.52	1.59	12 491	1 477
20201016CLOUD	0.16	0.23	1.49	17 912	6 134
20201028CLOUD	0.16	0.36	2.24	16 271	3 368
20201103CLOUD	0.37	0.97	2.65	15 958	2 737
20201110CLOUD	0.38	1.27	3.38	16 527	3 110
20201119CLOUD	0.28	1.00	3.62	16 064	3 406
Minimum	**0.16**	**0.23**	**1.49**	**12 491**	**985**
Maximum	**0.38**	**1.27**	**3.62**	**19 010**	**6 134**
Median	**0.28**	**0.62**	**2.24**	**16 064**	**3 110**
Mean	**0.27**	**0.66**	**2.38**	**15 917**	**3 032**
Standard error	**0.09**	**0.35**	**0.76**	**1 934**	**1 496**
P-value	**0.32**	**0.16**	**0.004****	**0.01***	**0.59**

[#]: as inferred from quantification in the extracts, based on sampling time and air flow rate; NA: no data available; *: significant p-value (< 0.05); **: highly significant p-value (< 0.01) (Mann-Whitney test; clouds *vs* clear conditions).

Table S2.

Processing information of sequences in MGs for (A) clouds and (B) clear conditions.

A)	CLOUD 20191002	CLOUD 20191022	CLOUD 20200311	CLOUD 20200717	CLOUD 20201016	CLOUD 20201028	CLOUD 20201103	CLOUD 20201110	CLOUD 20201119
Number of raw reads	65 812 666	259 998 456	58 746 330	43 757 944	97 184 944	66 400 646	54 975 726	60 932 548	77 768 050
After quality control (QC)	41 675 340	175 692 186	40 113 524	29 200 252	64 669 674	42 495 176	35 964 782	40 917 502	51 587 742
% removed	37	32	32	33	33	36	35	33	34
Number of rRNA gene reads	479 510	2 105 871	681 376	497 119	825 886	553 734	461 444	569 711	608 520
% of rRNA gene reads	1.15	1.20	1.70	1.70	1.28	1.30	1.28	1.39	1.18
Number of non-rRNA gene reads	41 195 830	173 586 315	39 432 148	28 703 133	63 843 788	41 941 442	35 503 338	40 347 791	50 979 222
% of non-rRNA gene reads	98.85	98.80	98.30	98.30	98.72	98.70	98.72	98.61	98.82
% of human reads in non-rRNA gene reads	0.11	0.71	0.02	0.04	0.08	0.14	0.07	0.08	0.09
Number of assembled contigs	194 547	495 663	275 249	129 639	316 632	213 954	193 399	241 377	207 227
Number of properly paired reads	2 060 240	17 067 006	3 109 058	1 853 656	5 065 598	3 131 000	3 707 712	3 154 518	2 198 260
% of properly mapped reads	5	9.9	7.9	6.4	7.9	7.5	10.4	7.8	4.3
Number of affiliated reads	554 018	2 271 135	480 815	506 200	769 566	469 334	337 543	374 346	597 562
% affiliated	2.66	2.59	2.40	3.47	2.38	2.21	1.18	1.83	2.32
Number of affiliated human reads	41 379	654 422	19 638	19 244	38 655	35 478	18 311	26 211	39 429
% of human reads	7.5	28.8	4.1	3.8	5	7.6	5.4	7	6.6

B)	CLEAR 20200707	CLEAR 20200708	CLEAR 20200709	CLEAR 20200922	CLEAR 20201118	CLEAR 20201124
Number of raw reads	47 409 014	30 435 818	44 295 022	41 909 288	40 730 342	41 086 722
After quality control (QC)	31 161 586	19 145 910	30 985 400	31 152 980	25 543 444	28 294 184
% removed	34	37	30	26	37	31
Number of rRNA gene reads	524 095	361 927	333 292	437 338	384 484	305 043
% of rRNA gene reads	1.68	1.89	1.08	1.40	1.51	1.08
Number of non-rRNA gene reads	30 637 491	18 783 983	30 652 108	30 715 642	25 158 960	27 989 141
% of non-rRNA gene reads	98.32	98.11	98.92	98.60	98.49	98.92
% of human reads in non-rRNA gene reads	0.02	0.02	0.01	0.02	0.14	-
Number of assembled contigs	140 392	43 627	99 268	179 815	101 745	-
Number of properly paired reads	4 511 490	2 149 274	3 511 940	5 361 278	1 450 314	-
% of properly mapped reads	14.6	11.3	11.4	17.4	5.8	-
Number of affiliated reads	628 977	361 163	614 940	303 720	499 417	-
% affiliated	4.04	3.77	3.97	1.95	3.91	-
Number of affiliated human reads	48 092	27 287	14 197	23 093	36 927	-
% of human reads	7.6	7.6	2.3	7.6	7.4	-

Table S3.

Processing information of sequences in MTs for (A) clouds and (B) clear conditions.

A)	CLOUD 20191002	CLOUD 20191022	CLOUD 20200311	CLOUD 20200717	CLOUD 20201016	CLOUD 20201028	CLOUD 20201103	CLOUD 20201110	CLOUD 20201019
Number of raw reads	93 499 990	186 010 124	82 131 152	81 247 584	79 489 694	85 009 094	110 129 198	69 916 188	195 503 428
After quality control (QC)	64 702 194	94 221 480	61 733 384	60 764 224	54 020 462	64 420 980	71 515 884	50 964 222	132 952 184
% removed	31	49	25	25	32	24	35	27	32
Number of rRNA reads	59 259 226	85 291 849	58 058 408	56 202 145	43 175 946	57 199 748	28 842 768	45 243 381	115 553 819
% of rRNA reads	91.59	90.52	94.05	92.49	79.93	88.79	40.33	88.77	86.91
Number of non-rRNA reads	5 442 968	8 929 631	3 674 976	4 562 079	10 844 516	7 221 232	42 673 116	5 720 841	17 398 365
% of non-rRNA reads	8.41	9.48	5.95	7.51	20.07	11.21	59.67	11.23	13.09
% of human reads in non-rRNA reads	0.86	0.05	0.32	0.42	0.82	0.46	89.47	0.29	0.31
Number de reads properly paired	275 906	655 764	262 850	275 216	833 402	697 632	464 440	423 202	1 007 468
% of properly mapped reads	4.6	7	7	5.7	7.5	8.9	3.9	7.1	5.6
Number of affiliated reads	24 817 282	37 049 964	26 829 087	22 589 666	18 755 564	24 842 334	32 332 229	17 680 329	52 249 268
% affiliated	76.71	78.64	86.92	74.35	69.44	77.12	90.42	69.38	78.60
Number of affiliated human reads	31 612	9 229	15 483	26 700	64 652	25 003	≈ 19 000 000	22 925	46 348
% of human reads	0.1	0.02	0.06	0.1	0.3	0.1	≈ 59	0.1	0.1

B)	CLEAR 20200707	CLEAR 20200708	CLEAR 20200709	CLEAR 20200922	CLEAR 20201118	CLEAR 20201124
Number of raw reads	71 487 464	116 985 074	96 661 644	65 554 022	76 813 684	68 355 146
After quality control (QC)	49 784 022	63 268 244	56 706 232	48 733 804	54 728 858	54 728 858
% removed	30	46	41	26	29	20
Number of rRNA reads	41 199 078	51 568 599	49 530 808	42 808 640	45 507 747	5 986 089
% of rRNA reads	82.76	81.51	87.35	87.84	83.15	12.06
Number of non-rRNA reads	8 584 944	11 699 645	7 175 424	5 925 164	9 221 111	43 642 337
% of non-rRNA reads	17.24	18.49	12.65	12.16	16.85	87.94
% of human reads in non-rRNA reads	1.95	1.37	12.02	0.2	0.7	-
Number of reads properly paired	648 418	383 000	190 608	667 098	467 706	-
% of properly mapped reads	7.6	3.3	2.7	10.8	4.9	-
Number of affiliated reads	18 903 493	22 757 792	22 429 335	19 290 161	17 432 686	-
% affiliated	75.94	71.94	79.11	79.17	63.71	-
Number of affiliated human reads	103 756	119 150	164 072	18 204	96 729	-
% of human reads	0.5	0.5	0.7	0.1	0.6	-

Table S4.
Distribution of the E.C. numbers corresponding to the 488 overrepresented transcripts in clouds or clear atmosphere.

		Number of related gene entries up-regulated during:	
		Clouds	**Clear conditions**
Oxidoreductases E.C. 1.-.-.-.			
E.C. 1.-.-.-	Others/ND	2	0
E.C. 1.1.-.-	Acting on the CH-OH group of donors (alcohol)	6	2
E.C. 1.11.-.-	Acting on a peroxide as acceptor	1	1
E.C. 1.17.-.-	Acting on CH or CH2 groups	0	2
E.C. 1.2.-.-	Acting on the aldehyde or oxo group of donors	3	3
E.C. 1.3.-.-	Acting on the CH-CH group of donors	1	1
E.C. 1.4.-.-	Acting on the CH-NH2 group of donors (aminoacids)	1	1
E.C. 1.8.-.-	Acting on a sulfur group of donors	0	1
	Total	**14**	**11**
Transferases E.C. 2.-.-.-.			
2.-.-.-	Others/ND	0	0
2.1.-.-	Transferring one-carbon groups (Methyltransferases)	1	1
2.2.-.-	Transferring aldehyde or ketonic groups	1	1
2.3.-.-	Acyltransferases	5	2
2.4.-.-	Glycosyltransferases	7	1
2.6.-.-	Transferring nitrogenous groups	0	1
2.7.-.-	Transferring phosphorus-containing groups	10	13
2.8.-.-	Transferring sulfur-containing groups	0	1
	Total	**24**	**20**
Hydrolases E.C. 3.-.-.-.			
3.-.-.-	Others/ND	3	0
3.1.-.-	Acting on ester bonds (Esterases)	4	0
3.2.-.-	Glycosylases	5	2
3.4.-.-	Acting on peptide bonds (peptidases)	9	2
3.5.-.-	Acting on carbon-nitrogen bonds, other than peptide bonds	1	0
3.6.-.-	Acting on acid anhydrides (Polyphosphatases)	18	7
3.7.-.-	Acting on carbon-carbon bonds	0	0
3.13.-.-	C-S hydrolases	0	1
	Total	**40**	**12**
Lyases E.C. 4.-.-.-.			
4.-.-.-	Others/ND	0	0
4.1.-.-	Carbon-carbon lyases (Decarboxylases)	4	3
4.2.-.-	Carbon-oxygen lyases (Dehydratases)	3	8
4.3.-.-	Carbon-nitrogen lyases	0	2
4.6.-.-	Phosphorus-oxygen lyases	1	0
	Total	**8**	**13**
Isomerases E.C. 5.-.-.-.			
5.-.-.-	Others/ND	0	0
5.1.-.-	Racemases and epimerases	1	0
5.4.-.-	Intramolecular transferases	2	2
5.5.-.-	Intramolecular lyases	1	1
5.6.-.-	Isomerases altering macromolecular conformation	0	3
	Total	**4**	**6**
Ligases E.C. 6.-.-.-.			
6.-.-.-	Others/ND	0	0
6.1.-.-	Forming carbon-oxygen bonds	0	3
6.2.-.-	Forming carbon-sulfur bonds	2	0
6.3.-.-	Forming carbon-nitrogen bonds	2	4
6.4.-.-	Forming carbon-carbon bonds	0	2
	Total	**4**	**9**
Translocases E.C. 7.-.-.-.			
7.-.-.-	Others/ND	0	0
7.1.-.-	Catalysing the translocation of hydrons	21	9
7.2.-.-	Catalysing the translocation of inorganic cations	0	1
7.4.-.-	Catalysing the translocation amino acids and peptides	0	4
	Total	**21**	**14**
	TOTAL	**115**	**85**

Other Supplementary Materials (separate electronic files):

- **Data S1.** Read counts affiliated with Bacteria in MGs and MTs based on taxonomy annotations in Kraken's plusPF database, at various taxonomic levels. The samples are named as "A" for clear conditions or "C" for clouds, respectively, followed by the sampling date as "yyyymmdd".
- **Data S2.** Read counts affiliated with Eukaryota in MGs and MTs based on taxonomy annotations in Kraken's plusPF database, at various taxonomic levels. The samples are named as "A" for clear conditions or "C" for clouds, respectively, followed by the sampling date as "yyyymmdd".
- **Data S3.** Taxon-based differential expression analysis (DEA), from bacterial and Eukaryotic families and genera representation in MTs *versus* MGs, all samples considered without distinction between atmospheric conditions. Value: factor of reference for which positive DEA coefficients indicate overrepresentation; coef: DEA coefficient from MTXmodel; stderr: standard deviation of the DEA coefficient; N: total number of samples considered; N.not.0: number of samples with corresponding reads number >0; pval: p-value associated with the DEA coefficient. The taxa significantly more represented in MTs than in MGs are highlighted in green (positive coefficient and p-value < 0.05).
- **Data S4.** List of the 488 overrepresented transcripts based on differential expression analysis (DEA), without distinction between atmospheric conditions, and proteins, genes, organisms, GO terms and E.C. numbers associated with them. Value: factor of reference for which positive DEA coefficients indicate overrepresentation; coef: DEA coefficient from MTXmodel; stderr: standard deviation of the DEA coefficient; N: total number of samples considered; N.not.0: number of samples with corresponding reads number >0; pval: p-value associated with the DEA coefficient.
- **Data S5.** Average DEA coefficients of GO terms associated with overrepresented transcripts, all samples considered without distinction between atmospheric conditions. Positive coefficients are highlighted in green and indicate overrepresented GO terms. Biological Processes are represented as **Fig 2**.
- **Data S6.** List of the 320 transcripts differentially represented between clouds and clear atmosphere, based on differential expression analysis (DEA), and proteins, genes, organisms, GO terms and E.C. numbers associated with them. Value: factor of reference for which positive DEA coefficients indicate overrepresentation; coef: DEA coefficient from MTXmodel; stderr: standard deviation of the DEA coefficient; N: total number of samples considered; N.not.0: number of samples with corresponding reads number >0; pval: p-value associated with the DEA coefficient.
- **Data S7.** Average DEA coefficients of GO terms associated with differentially represented transcripts between clouds and clear atmosphere. Positive coefficients are highlighted in green and indicate GO terms overrepresented in clouds. Biological Processes are represented in **Fig 4D**.