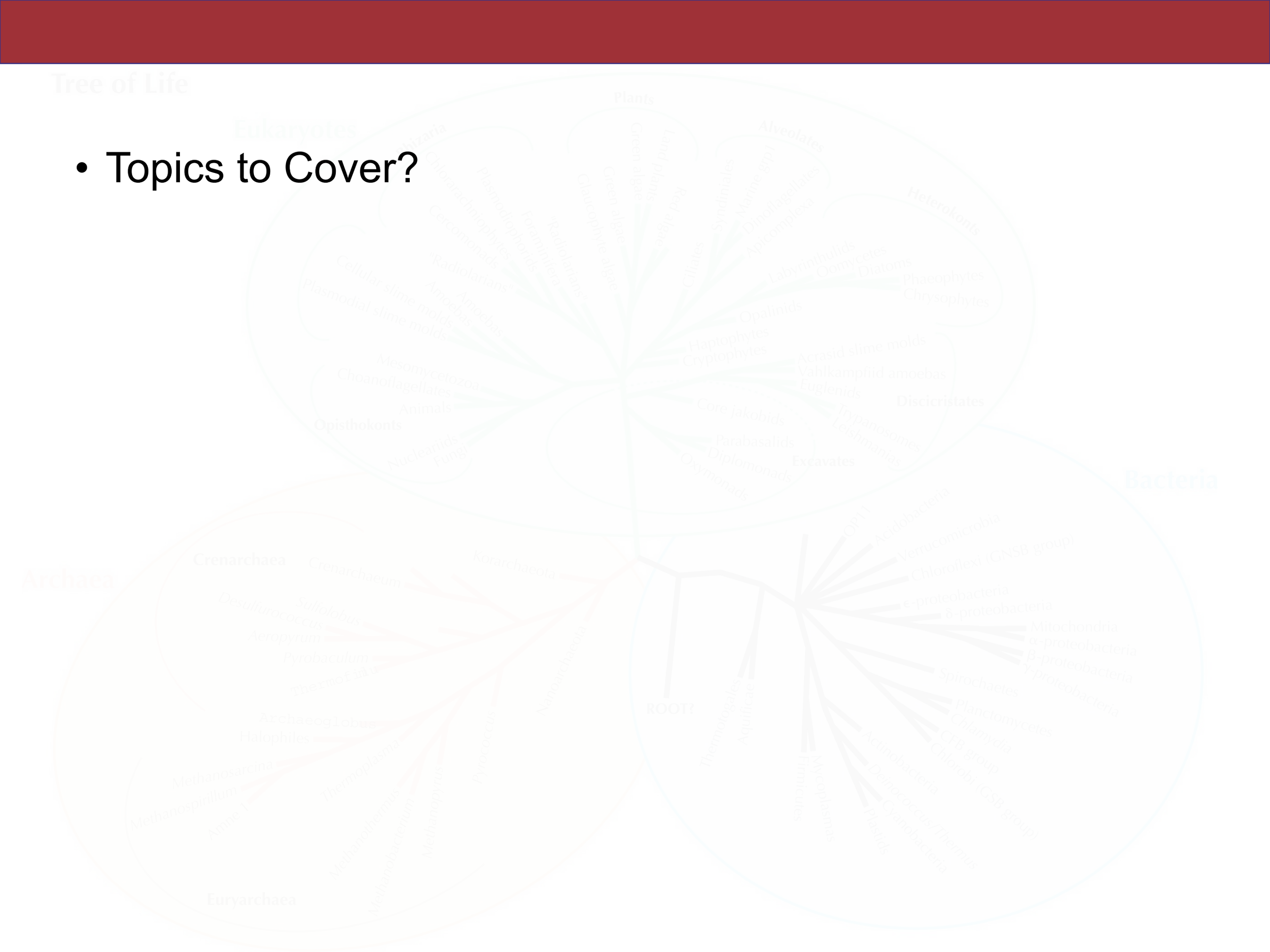
EVE 161:
Microbial Phylogenomics

Case Study
1000s of MAGs

UC Davis, Winter 2018
Instructor: Jonathan Eisen
Teaching Assistant: Cassie Ettinger

- Presenters?

- Topics to Cover?



nature microbiology

ARTICLES

DOI: 10.1038/s41564-017-0012-7

OPEN

Corrected: Author correction

Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz* and Gene W. Tyson*

Challenges in cultivating microorganisms have limited the phylogenetic diversity of currently available microbial genomes. This is being addressed by advances in sequencing throughput and computational techniques that allow for the cultivation-independent recovery of genomes from metagenomes. Here, we report the reconstruction of 7,903 bacterial and archaeal genomes from >1,500 public metagenomes. All genomes are estimated to be ≥50% complete and nearly half are ≥90% complete with ≤5% contamination. These genomes increase the phylogenetic diversity of bacterial and archaeal genome trees by >30% and provide the first representatives of 17 bacterial and three archaeal candidate phyla. We also recovered 245 genomes from the Patescibacteria superphylum (also known as the Candidate Phyla Radiation) and find that the relative diversity of this group varies substantially with different protein marker sets. The scale and quality of this data set demonstrate that recovering genomes from metagenomes provides an expedient path forward to exploring microbial dark matter.

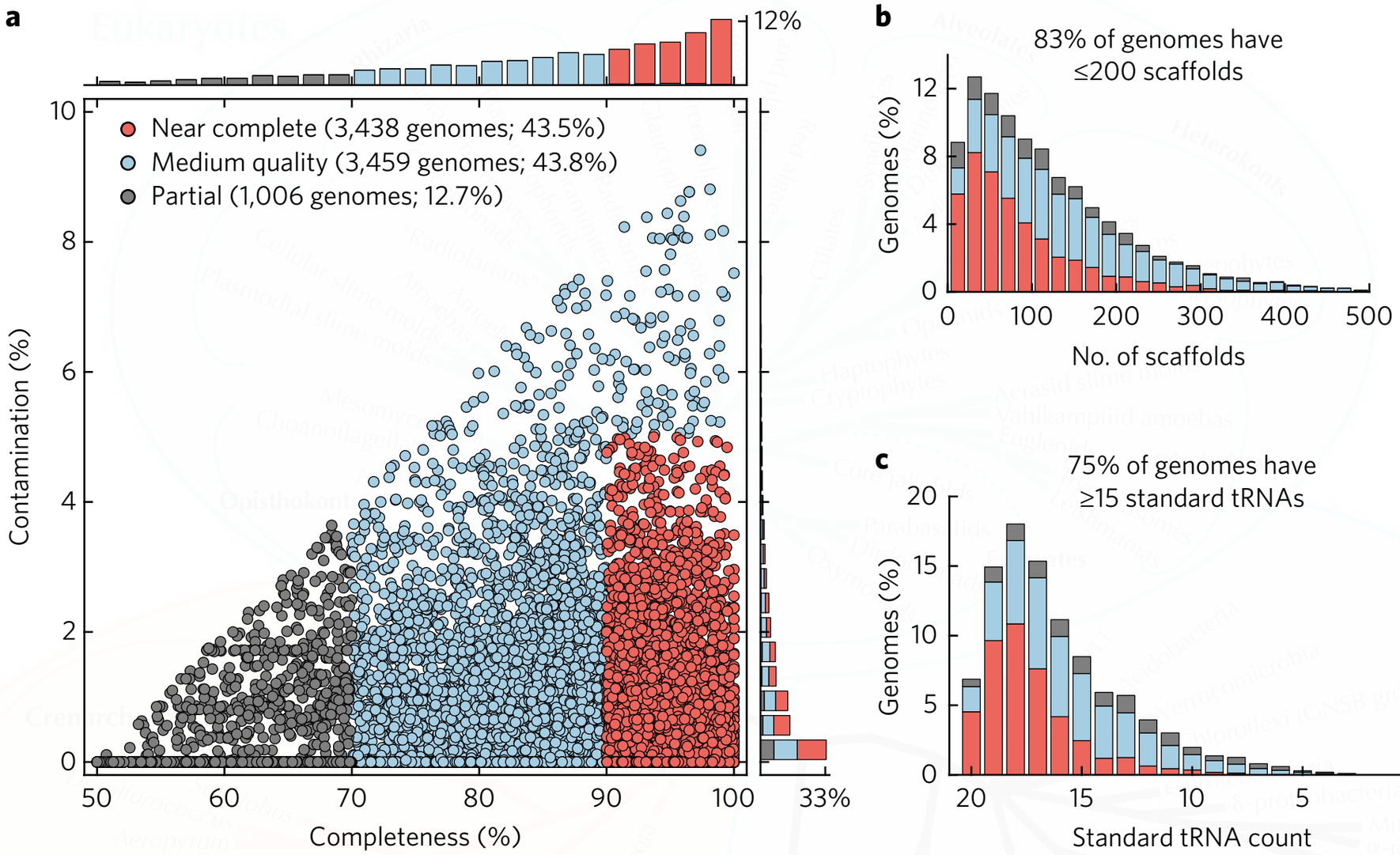


Fig. 1 | Assessment of genome quality. a, Estimated completeness and contamination of 7,903 genomes recovered from public metagenomes. Genome quality was defined as completeness − 5 × contamination, and only genomes with a quality of ≥50 were retained. Near-complete genomes (completeness ≥90%; contamination ≤5%) are shown in red, medium-quality genomes (completeness ≥70%; contamination ≤10%) in blue, and partial genomes (completeness ≥50%; contamination ≤4%) in grey. Histograms along the *x* and *y* axes show the percentage of genomes at varying levels of completeness and contamination, respectively. Notably, only 171 of the 7,903 (2.2%) UBA genomes have >5% contamination. **b**, Number of scaffolds comprising each of the 7,903 genomes with colours indicating genome quality. **c**, Number of tRNAs for each of the 20 standard amino acids identified within each of the 7,903 genomes with colours indicating genome quality.

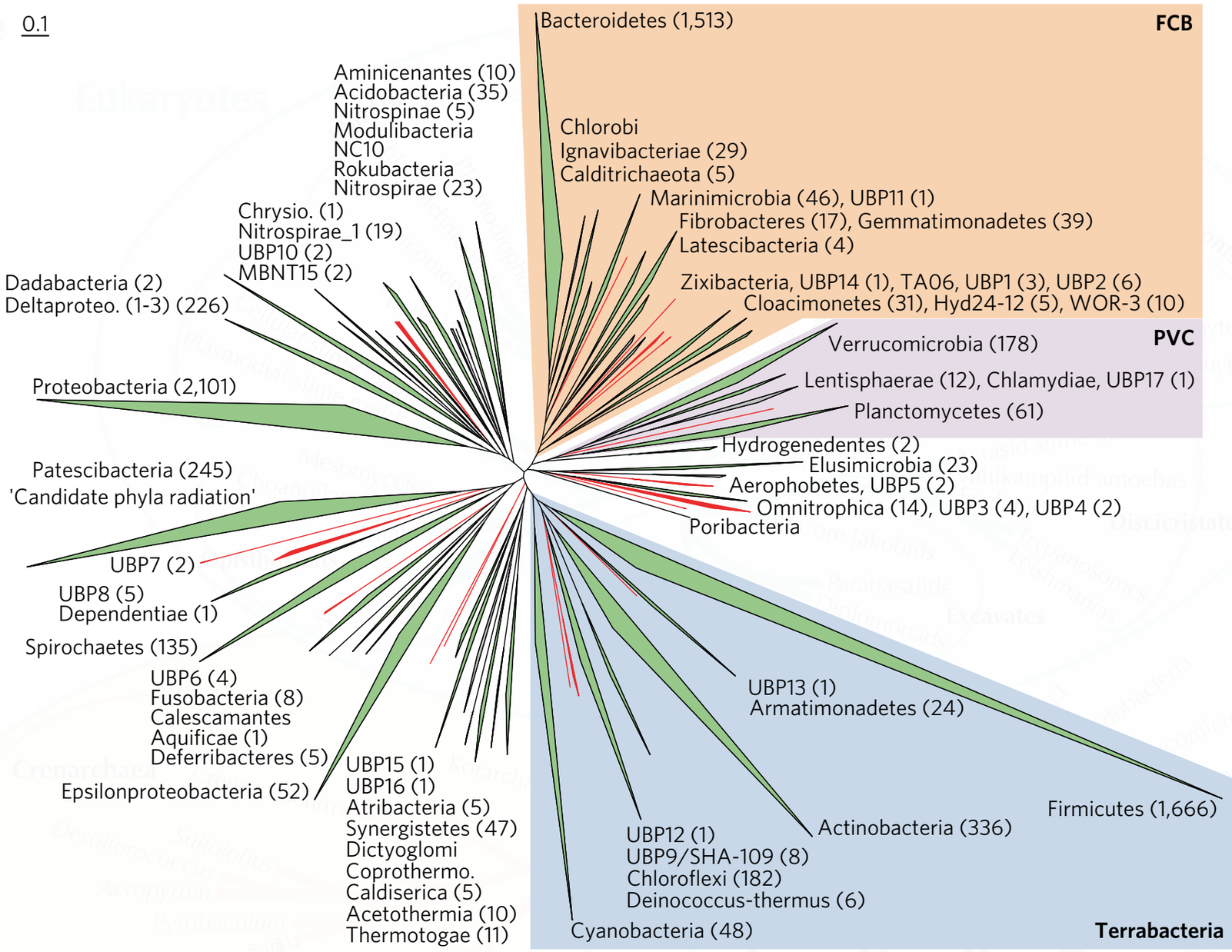


Fig. 2 | Distribution of uBA genomes across 76 bacterial phyla. The maximum likelihood tree was inferred from the concatenation of 120 proteins and spans a dereplicated set of 5,273 UBA and 14,304 NCBI genomes. Phyla containing UBA genomes are shown in green with the number of UBA genomes indicated in parentheses. Candidate phyla consisting only of UBA genomes are shown in red and have been named Uncultured Bacterial Phylum 1 to 17 (UBP1–UBP17). The Tenericutes and Thermodesulfobacteria lineages are not shown as they branch within the Firmicutes and Deltaproteobacteria, respectively. The Patescibacteria/CPR is shown as a single lineage for clarity and to illustrate its similarity to well-characterized phyla. The Nitrospirae and Deltaproteobacteria were found to be polyphyletic and an underscore and numerical identifier is used to distinguish between lineages. The Deltaproteobacteria are polyphyletic due to the placement of the Dadabacteria. All named lineages have bootstrap support of ≥75%, with the exception of the Chrysiogenetes (55%), Firmicutes (1%) and Verrucomicrobia (52%). The complete tree with support values is available in Newick format (Supplementary File 1).

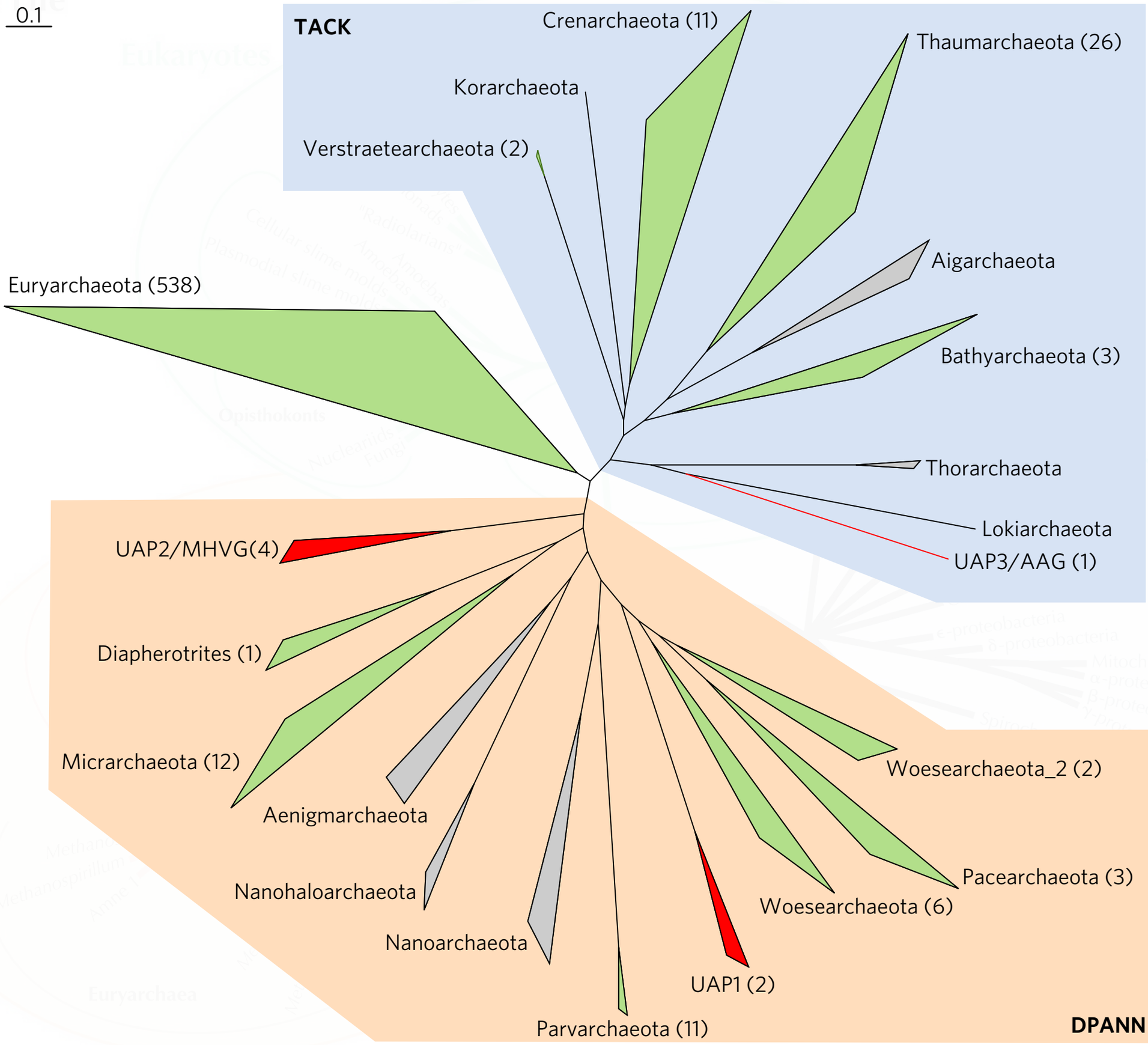

0.1
TACK
Crenarchaeota (11)
Thaumarchaeota (26)
Korarchaeota
Verstraetearchaeota (2)
Aigarchaeota
Euryarchaeota (538)
Bathyarchaeota (3)
Thorarchaeota
Lokiarchaeota
UAP3/AAG (1)
UAP2/MHVG(4)
Diapherotrites (1)
Micrarchaeota (12)
Aenigmarchaeota
Woesearchaeota_2 (2)
Pacearchaeota (3)
Nanohaloarchaeota
Woesearchaeota (6)
Nanoarchaeota
UAP1 (2)
Parvarchaeota (11)
DPANN

Fig. 3 | Distribution of UBA genomes across 21 archaeal phyla. The maximum likelihood tree was inferred from the concatenation of 122 proteins and spans a dereplicated set of 453 UBA and 630 NCBI genomes. Phyla containing UBA genomes are shown in green with the number of UBA genomes indicated in parentheses. Candidate phyla consisting only of UBA genomes are shown in red and have been named Uncultured Archaeal Phylum 1 to 3 (UAP1–UAP3). The phylum Woesearchaeota was found to be polyphyletic and an underscore and numerical identifier is used to distinguish between lineages. All named lineages have bootstrap support ≥90%, with the exception of the Crenarchaeota (73%). The complete tree with support values is available in Newick format (Supplementary File 2). MHVG, Marine Hydrothermal Vent Group; AAG, Ancient Archaeal Group.

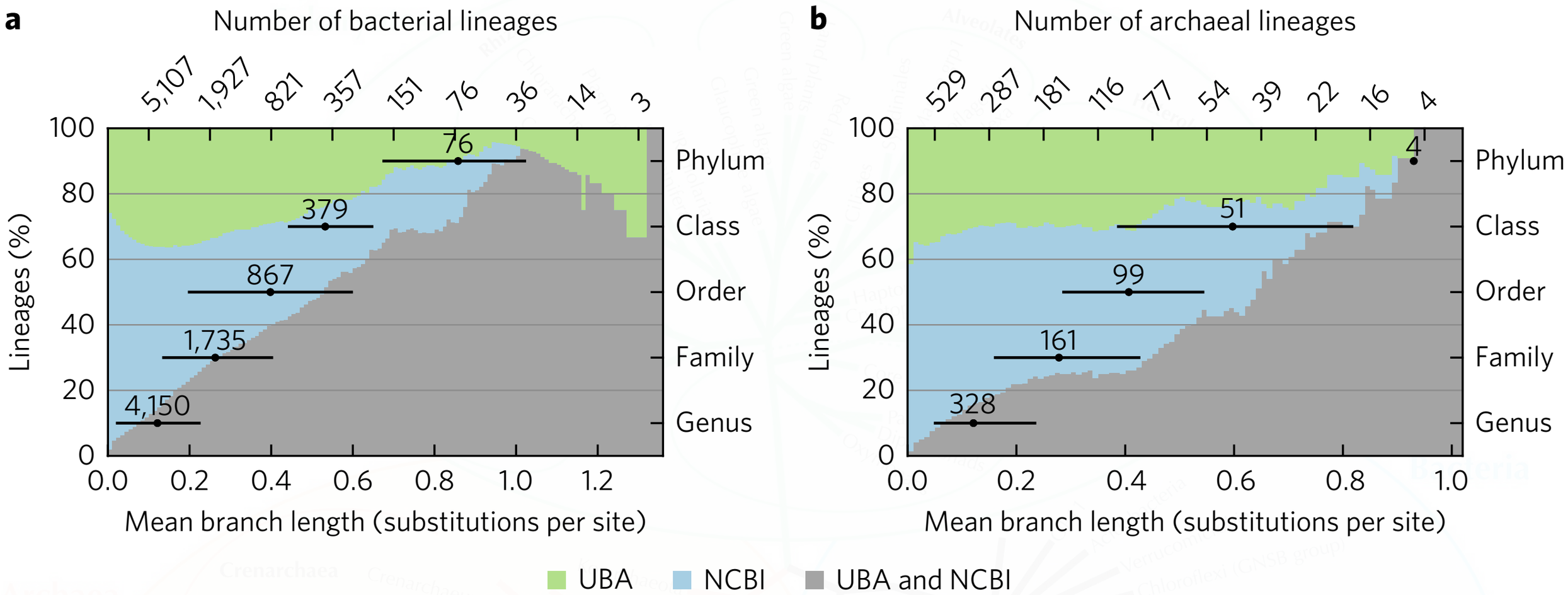
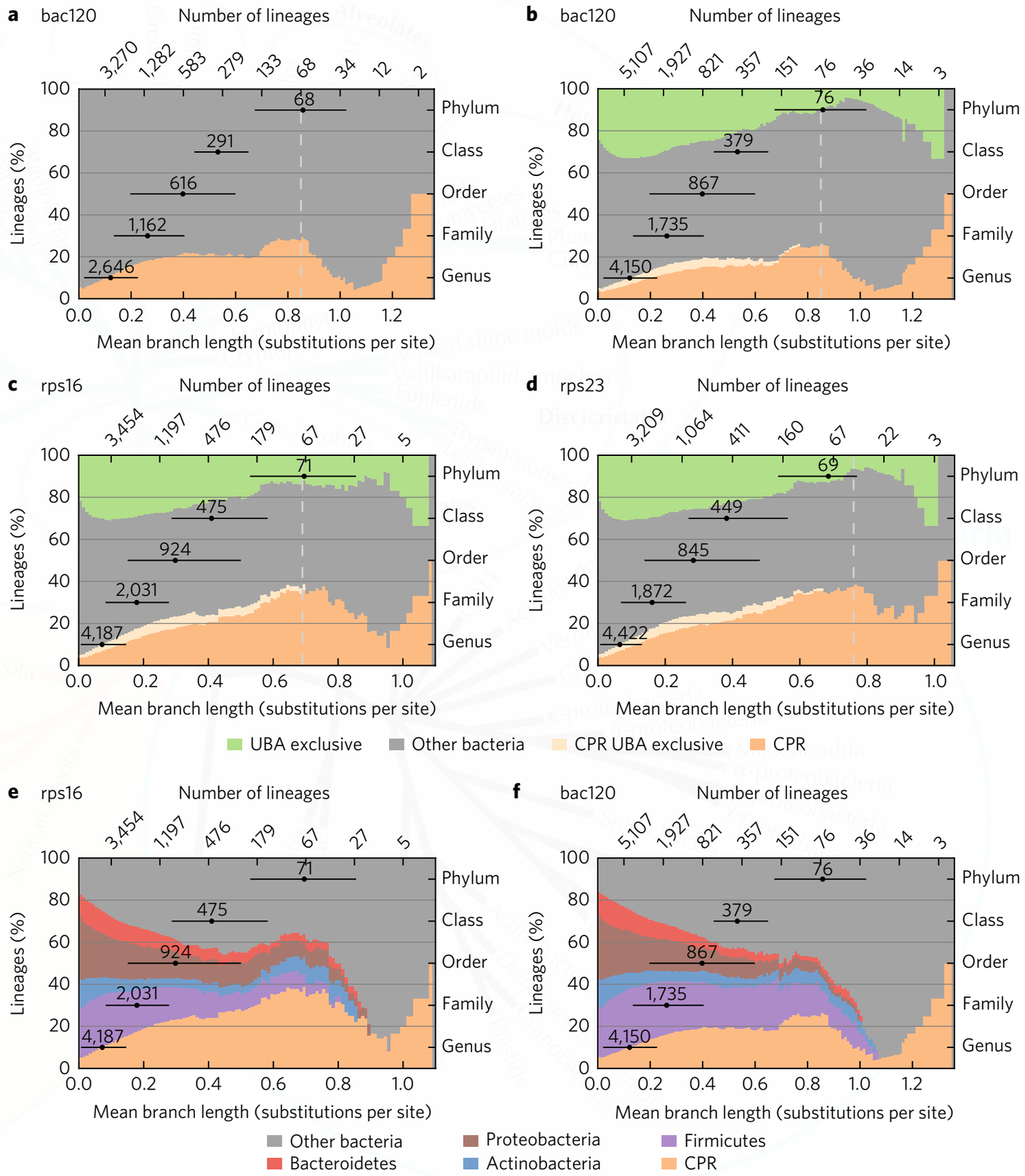


Fig. 4 | Percentage of lineages of equal evolutionary distance represented exclusively by UBA genomes. **a**, Percentage of bacterial lineages represented exclusively by the UBA genomes, genomes at NCBI, or both data sets as defined on the basis of mean branch length to extant taxa. Mean branch length to extant taxa was calculated for all internal nodes and lineages defined at a specific threshold to establish lineages of equal evolutionary distance. The bottom *x* axis indicates the threshold used to define lineages (0 = extant taxa, far right = root of tree), the top *x* axis indicates the number of lineages at different mean branch length thresholds, and the left *y* axis indicates the percentage of lineages. The right *y* axis indicates different taxonomic ranks, with the 5th and 90th percentiles of the distribution of mean branch length values for these ranks shown by black lines. The mean of each distribution is shown as a black circle along with the number of lineages at this value. **b**, Analogous plot to **a**, but for archaeal lineages. The phylum distribution is a single point, as the Euryarchaeota is the only archaeal phylum with phylum-level resolution (that is, containing multiple classes). Plots were determined across the domain-specific trees inferred from the concatenation of 120 bacterial and 122 archaeal proteins, respectively.

	Total		Dereplicated			Relative	
	NCBI	UBA	NCBI	UBA	Total PD	Taxon PD	UBA PG
Bacteria	67,034	7,244	14,304	5,273	100.0%		33.0%
UBP (all)	0	46	0	46	1.1%		1.1%
Acetothermia	1	10	1	6	0.1%		71.8%
Aminicenantes	4	10	4	10	0.2%		81.0%
Armatimonadetes	12	24	6	13	0.4%		67.2%
Bacteroidetes	1,324	1,513	941	1,057	9.5%		47.4%
Caldiserica	1	5	1	4	0.2%		74.3%
Chloroflexi	61	182	50	169	2.7%		73.1%
Dadabacteria	1	2	1	2	0.1%		74.8%
Elusimicrobia	3	23	3	14	0.3%		70.5%
Firmicutes	24,326	1,666	3,433	1,296	25.1%		32.2%
Gemmatimonadetes	4	39	3	37	0.2%		78.6%
Hyd24-12	3	5	3	5	0.1%		61.4%
Latescibacteria	1	4	1	3	0.1%		64.7%
Lentisphaerae	2	12	2	12	0.3%		75.8%
Omnitrophica	1	14	1	14	0.3%		89.3%
Patescibacteria / CPR	811	245	811	245	11.2%		25.8%
Proteobacteria	28,860	2,101	4,889	1,272	20.5%		25.0%
Verrucomicrobia	45	178	44	166	1.9%		65.5%
Archaea	825	623	630	453	100.0%		30.7%
UAP (all)	0	7	0	7	1.1%		100.0%
Euryarchaeota	534	538	422	369	52.7%		34.8%
Marine Group II	12	206	12	206	9.2%		85.9%
Micrarchaeota	1	12	1	12	2.4%		73.4%
Pacearchaeota	5	3	5	3	2.7%		37.9%
Thaumarchaeota	49	26	44	26	6.4%		40.8%
Woesearchaeota	7	6	7	6	4.4%		42.9%

Fig. 5 | Phylogenetic diversity and gain for select bacterial and archaeal phyla. The 7,903 UBA genomes cover considerable phylogenetic diversity (PD) and contribute substantial phylogenetic gain (PG) relative to genomes from RefSeq/GenBank release 76. PD and PG were assessed on the domain-specific trees inferred from the concatenation of 120 bacterial and 122 archaeal proteins. The bar charts give (1) the total PD for each lineage relative to the PD of the entire domain, (2) the PD of the NCBI and UBA genomes relative to the PD of the lineage, and (3) the PG contributed by the UBA genomes within the lineage. The 17 UBP and 3 UAP lineages were collapsed together to show their combined phylogenetic diversity and gain. Marine Group II, a lineage within the Euryarchaeota, has been included because it comprises the majority of euryarchaeotal UBA genomes.

Fig. 6 | Percentage of lineages of equal evolutionary distance within the CPR. a, Percentage of CPR lineages within the bac120 tree without the addition of the UBA genomes. The maximum percentage of phylum-level lineages within the CPR occurs at a mean branch length of 0.85, as denoted by the dashed grey line. The remaining figure elements are the same as in Fig. 4. **b**, Analogous plot to **a**, with the addition of the UBA genomes (maximum = 0.85).
c,**d**, Percentage of CPR lineages within the rp1 and rp2 trees, respectively, with the maximum percentage of phylum-level lineages within the CPR denoted with a dashed grey line (rp1 = 0.69; rp2 = 0.76)
e,**f**, Percentage of CPR, Firmicutes, Actinobacteria, Proteobacteria and Bacteroidetes lineages within the rp1 and bac120 trees, respectively. The trees span 14,290 (**a**), 19,198 (**b** and **f**), 18,081 (**c** and **e**) and 18,226 (**d**) genomes.



Recently, the diversity of the CPR was explored in the context of a genome tree inferred from 16 ribosomal proteins where it was divided into 36 named phyla and shown to represent approximately 50% of bacterial lineages of equal phylum-level evolutionary distance[30]. Our analyses using a 120 concatenated proteins contrast with this view, as the CPR is shown to comprise ~25% of phylum- level lineages under the same criterion (Fig. 6b and Supplementary Fig. 7). This suggests that ribosomal proteins within CPR organisms may be evolving atypically relative to other proteins, perhaps as a result of their unusual ribosome composition and the presence of self-splicing introns and proteins being encoded within their rRNA genes[10]. These contrasting views of the diversity of the CPR are equally valid and probably reflect the unique biology of the organ- isms within this group.