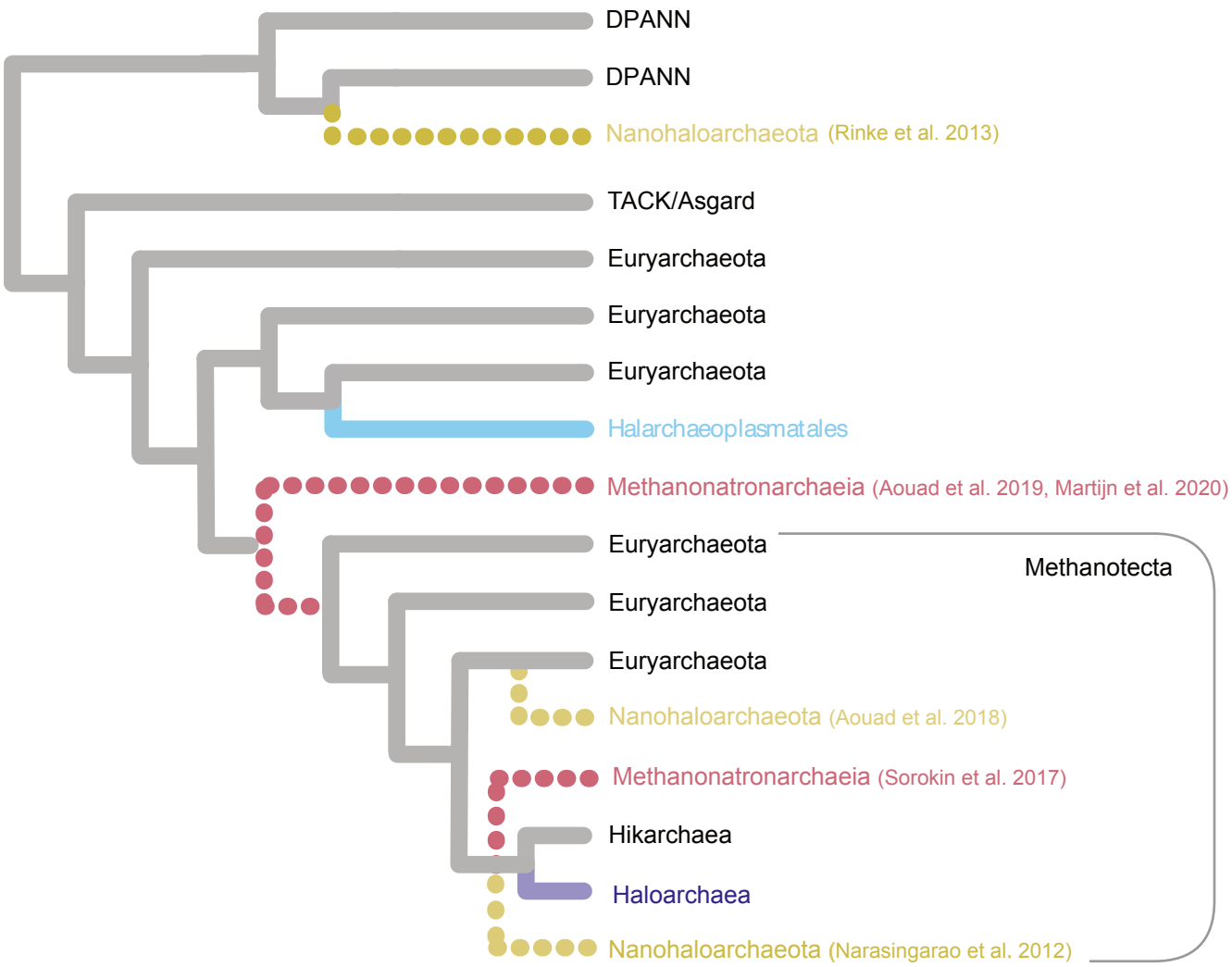
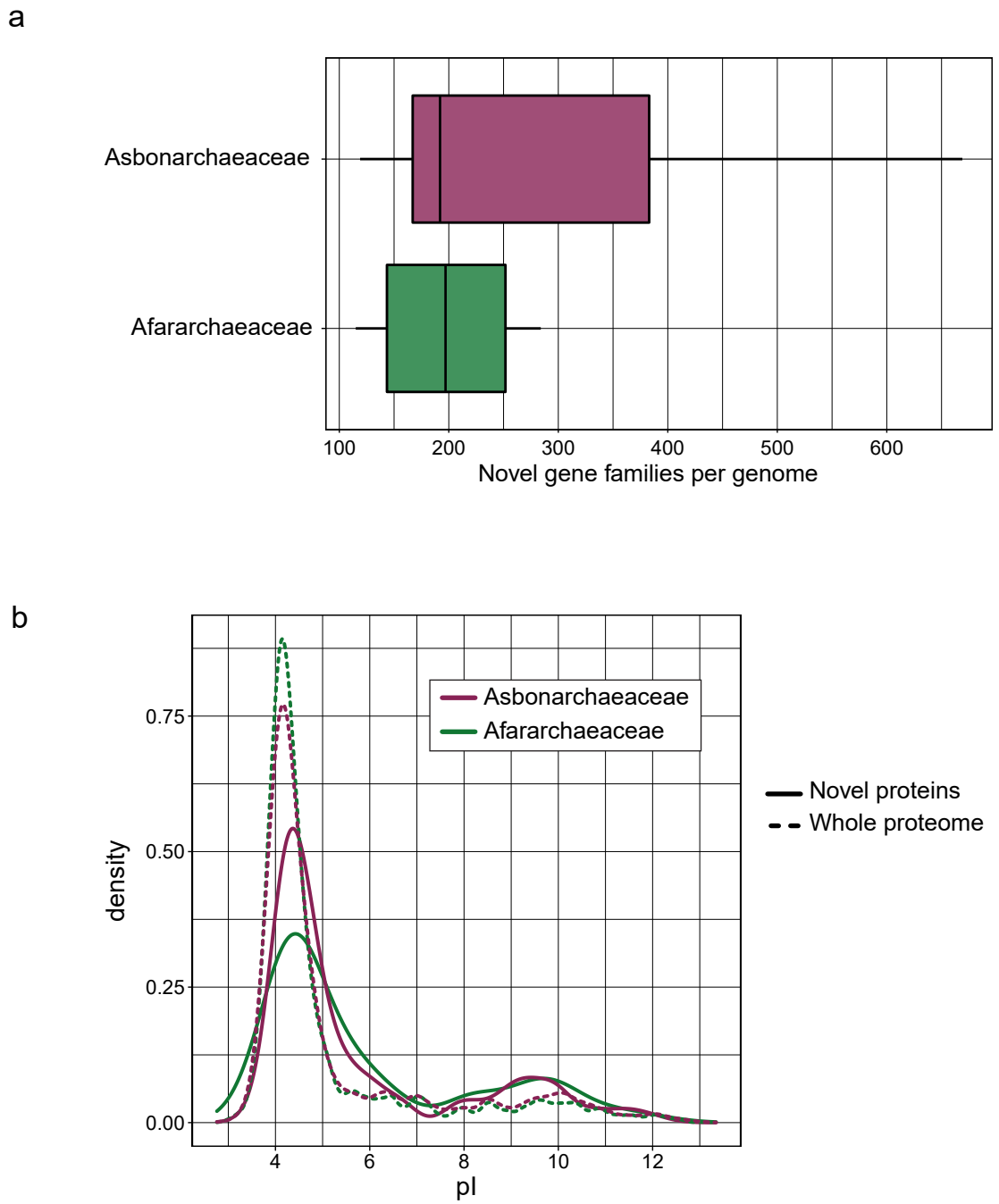
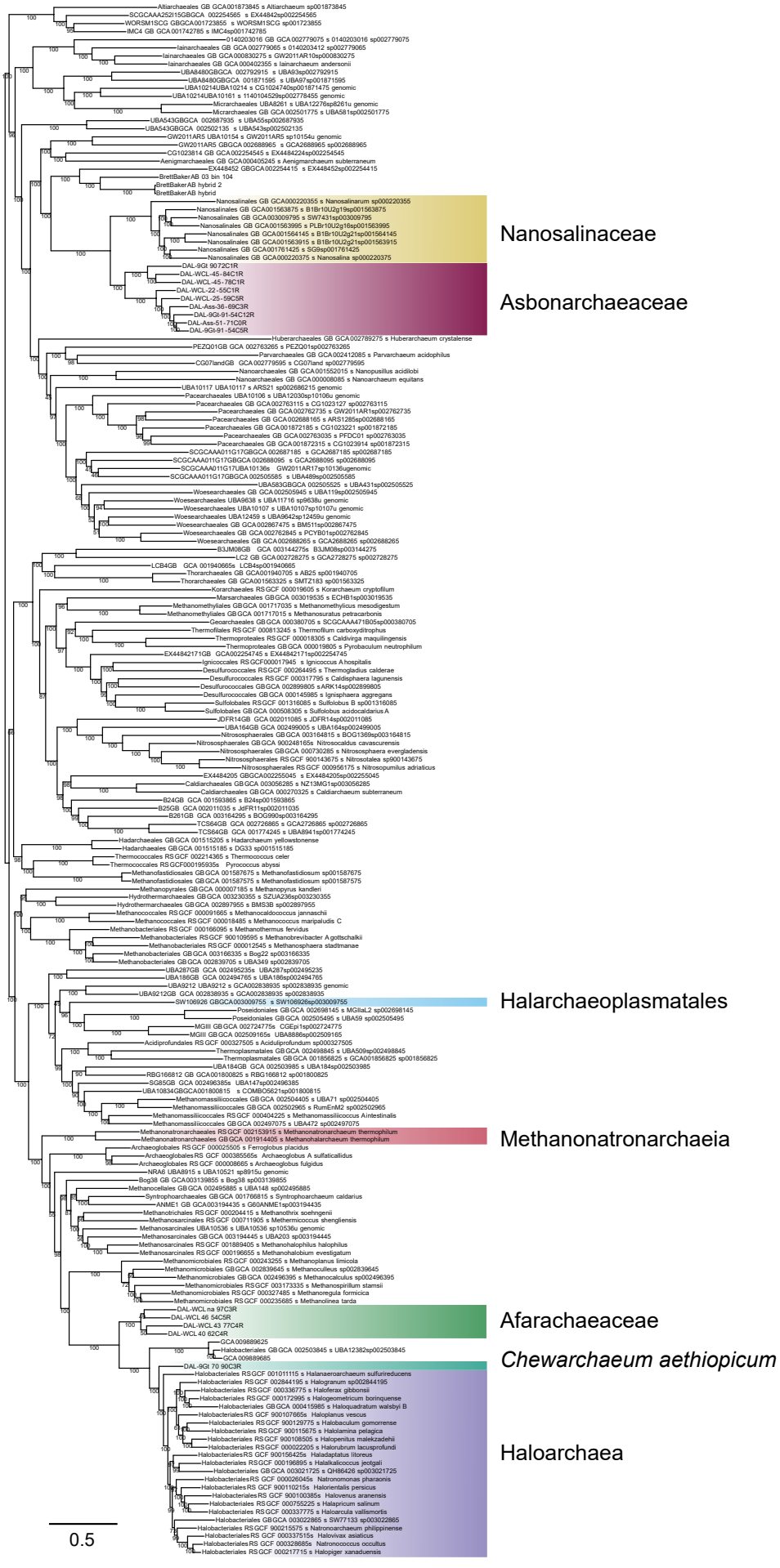


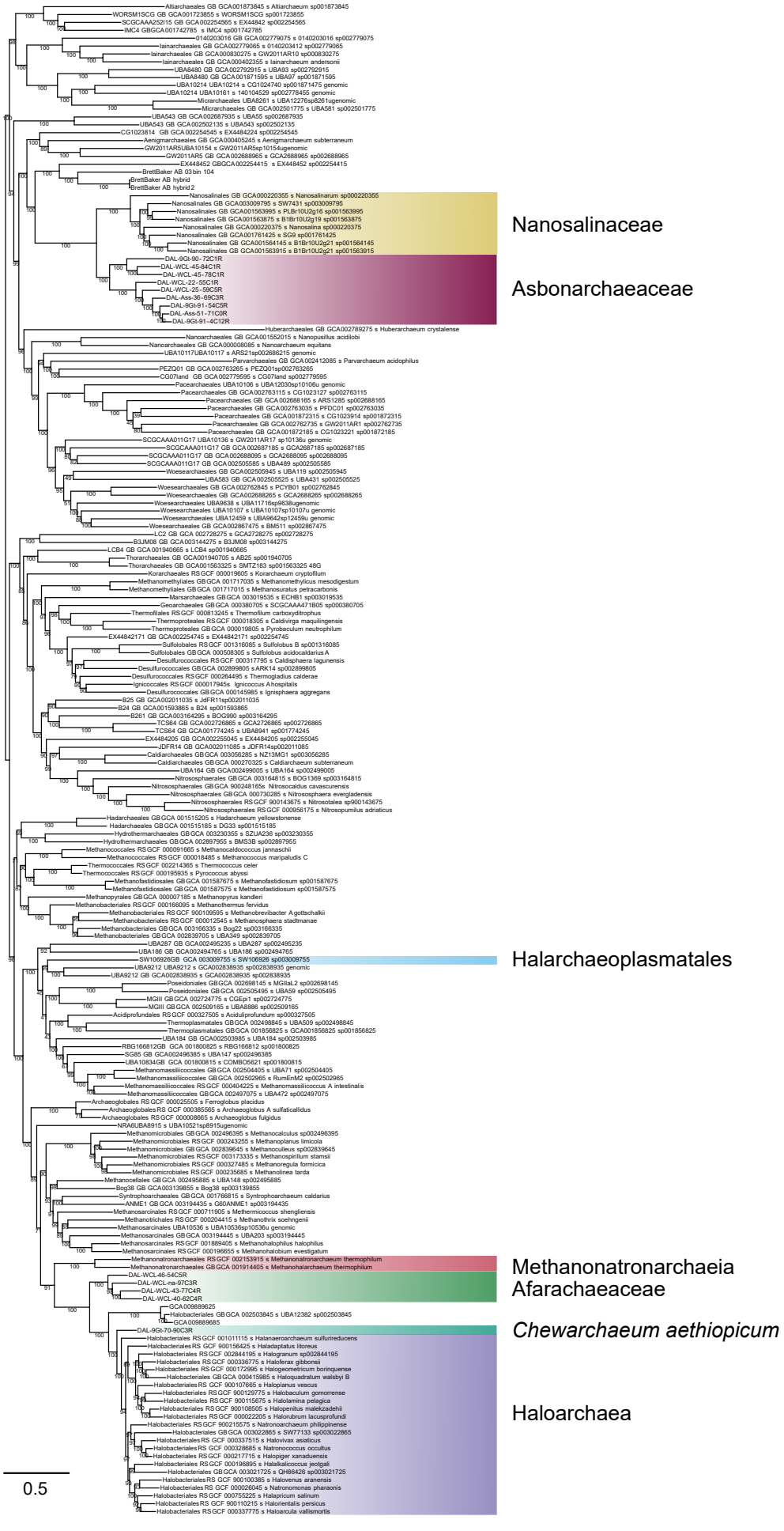
Extended Data Fig. 1 | Schematic tree showing the phylogenetic position of extremely halophilic archaeal groups (colored branches) proposed in previous articles. Branches that have been found at different places in the tree of archaea are indicated with dashed lines (Narasingarao et al. 2012, Rinke et al. 2013, Sorokin et al. 2017, Aouad et al. 2018, Aouad et al. 2019, Martijn et al. 2020).



Extended Data Fig. 2 | Number and isoelectric point of novel gene families identified in the Asbonarchaeaceae and Afararchaeaceae MAGs. (a) The average number of novel genes in the nine asbonarchaeal and four afarachaeal MAGs described in this study (see Methods). **(b)** The isoelectric point of these novel proteins (solid lines) compared to the average isoelectric point of the whole proteomes (dashed lines).

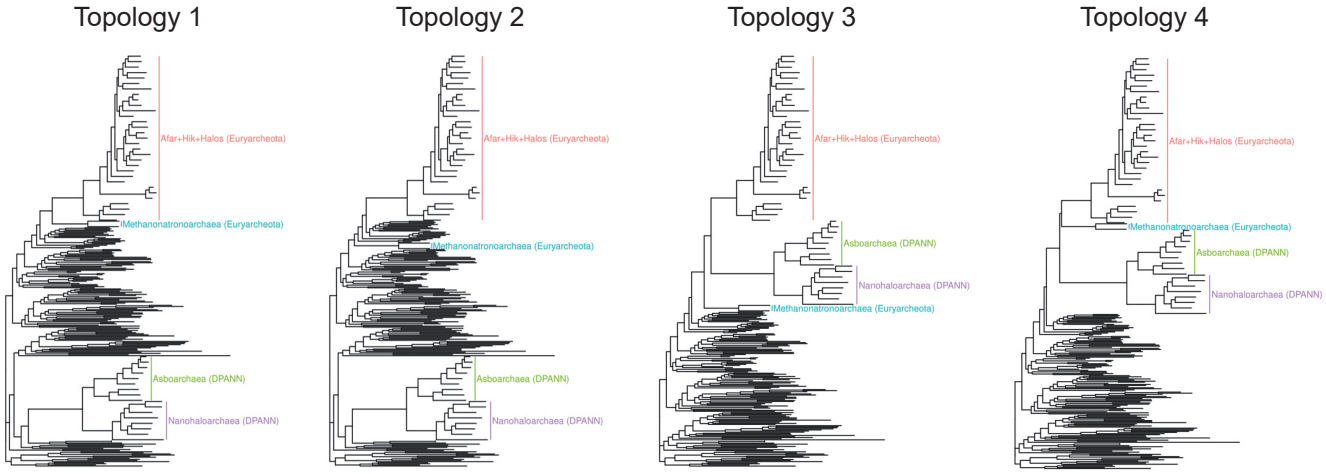


Extended Data Fig. 3 | Maximum likelihood phylogeny of 192 archaea based on the NM dataset. The ML tree was inferred with the LG+C60+F+Γ4 model of sequence evolution with 1,000 ultrafast bootstraps as implemented via IQ-TREE. The scale bar indicates the expected average number of substitutions per site. Extremely halophilic archaea are indicated in color.



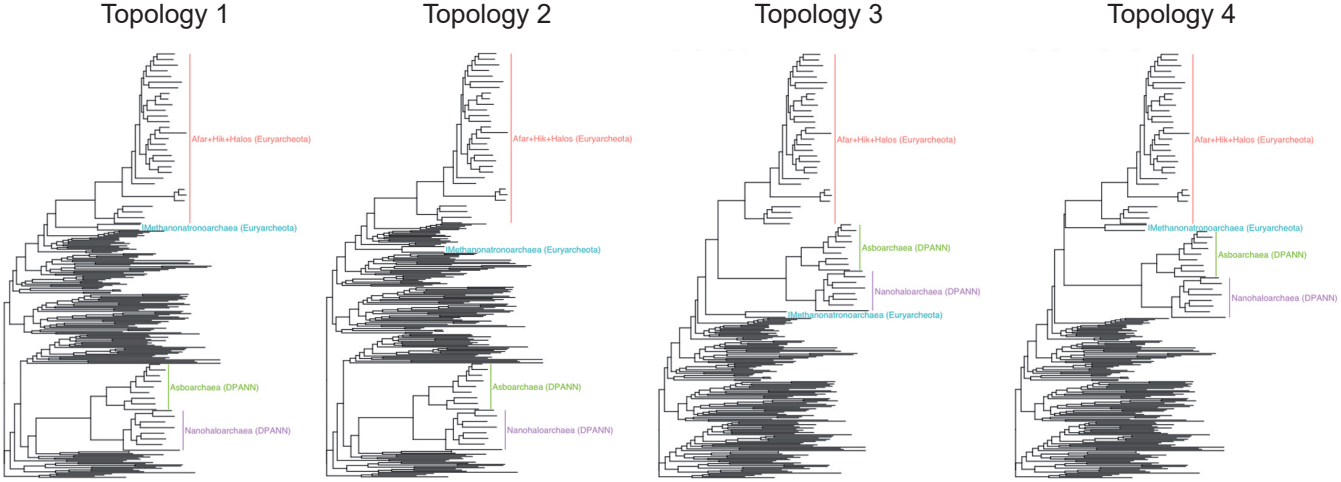
Extended Data Fig. 4 | Maximum likelihood phylogeny of 192 archaea based on the RP dataset. The ML tree was inferred with the LG+C60+F+Γ4 model of sequence evolution with 1,000 ultrafast bootstraps as implemented via IQ-TREE. The scale bar indicates the expected average number of substitutions per site. Extremely halophilic archaea are indicated in color.

a



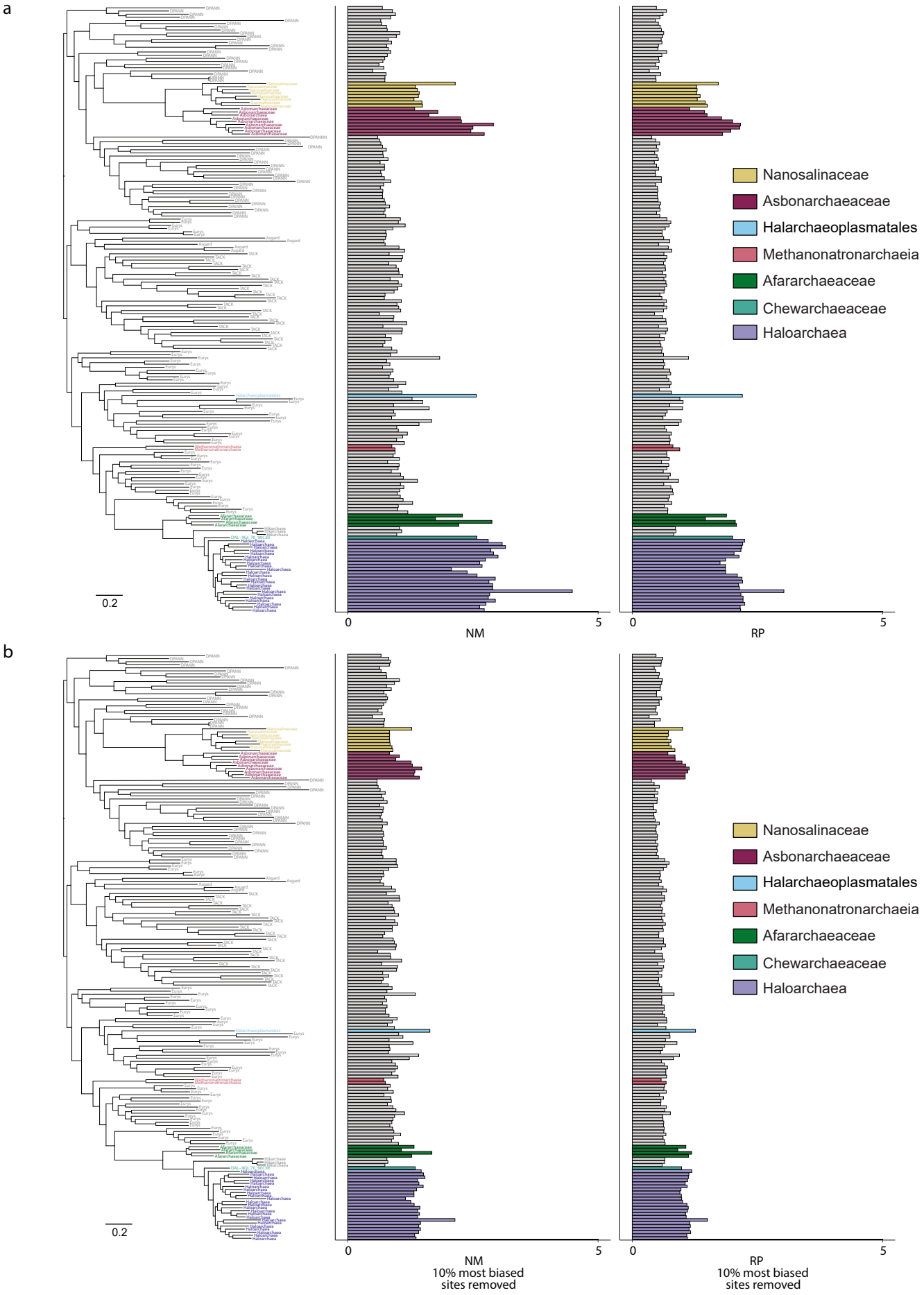
192-taxa RP dataset						
Model	**Bin 1**	**Bin 2**	Topology 1	**Topology 2**	**Topology 3**	**Topology 4**
LG+C60+G+F	None	None	Likelihood=-1317976,09 (Total di ff.=0)	Likelihood= -1318026,107 (Total diff.= -50,078)	Likelihood= -1318569,945 (Total diff.= -593,916)	Likelihood= -1318599,398 (Total diff.= -623,369)
LG+C60+G+F+GFmix	DE	IK	Likelihood=-1314937,111 (Total di ff.=0)	Likelihood= -1314977,65 (Total diff.= -40,539)	Likelihood= -1315670,918 (Total diff.= -733,808)	Likelihood= -1315685,576 (Total diff.= -748,465)
LG+C60+G+F+GFmix	DEQTAVG	KILCMFYWS	Likelihood=-1315018,218 (Total di ff.=0)	Likelihood= -1315065,115 (Total diff.= -46,898)	Likelihood= -1315704,916 (Total diff.= --686,698)	Likelihood= -1315727,147 (Total diff.= -708,93)

b

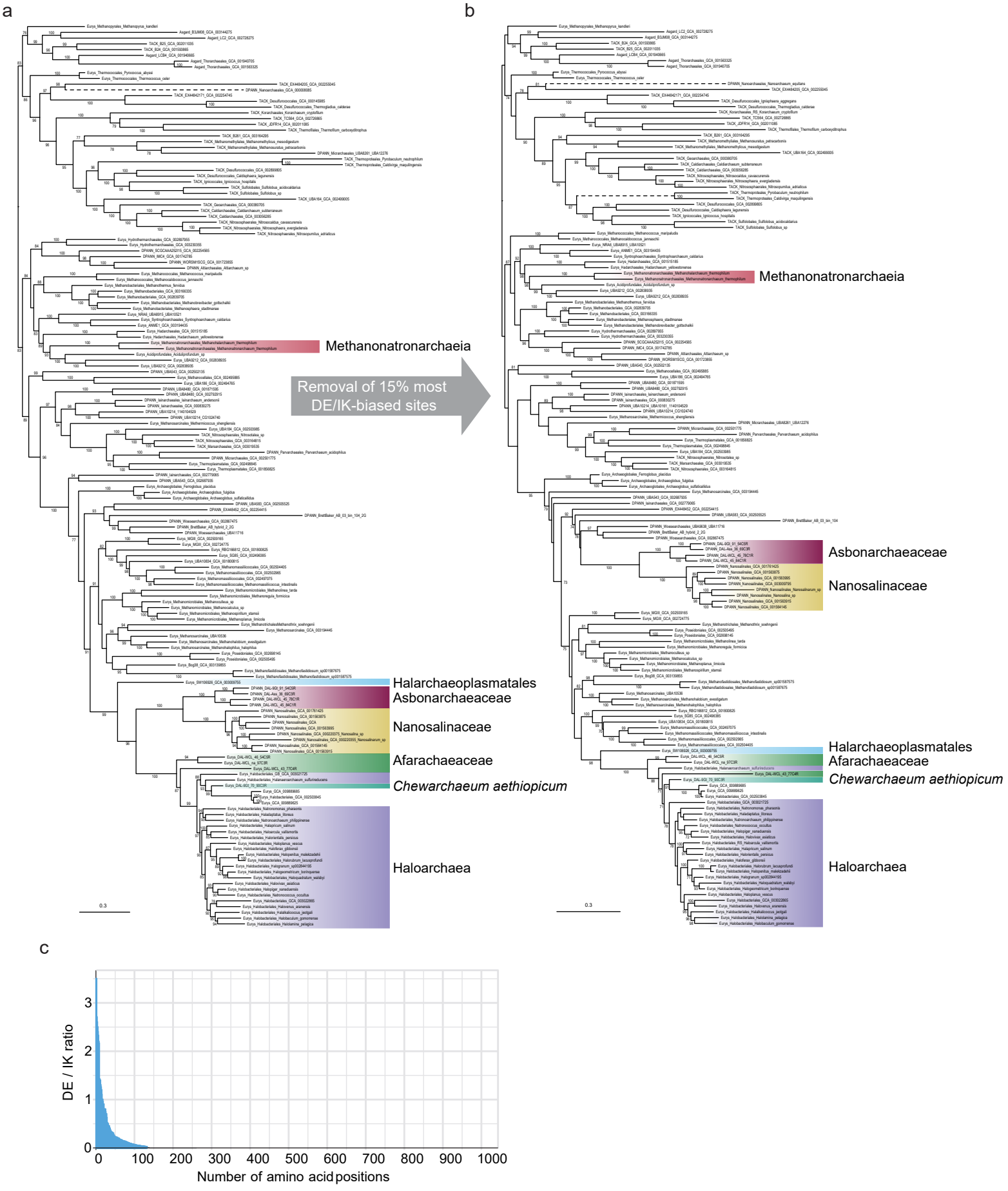


192-taxa NM dataset						
Model	**Bin 1**	**Bin 2**	**Topology 1**	Topology 2	**Topology 3**	**Topology 4**
LG+C60+G+F	None	None	Likelihood= -6557969,994 (Total diff.= -330,418)	Likelihood=-6557639,577 (Total di ff.=0)	Likelihood= -6561223,122 (Total diff.= -3583,545)	Likelihood= -6561620,115 (Total diff.= -3980,538)
LG+C60+G+F+GFmix	DE	IK	Likelihood= -6547076,576 (Total diff.= -364,896)	Likelihood=-6546711,68 (Total di ff.=0)	Likelihood= -6550656,783 (Total diff.= -3945,103)	Likelihood= -6550979,577 (Total diff.= -4267,898)
LG+C60+G+F+GFmix	DEAVTQGRHP	KINFSCLYM	Likelihood= -6541714,646 (Total diff.= -309,156)	Likelihood=-6541405,489 (Total di ff.=0)	Likelihood= -6545093,421 (Total diff.= -3687,931)	Likelihood= -6545459,741 (Total diff.= -4054,252)

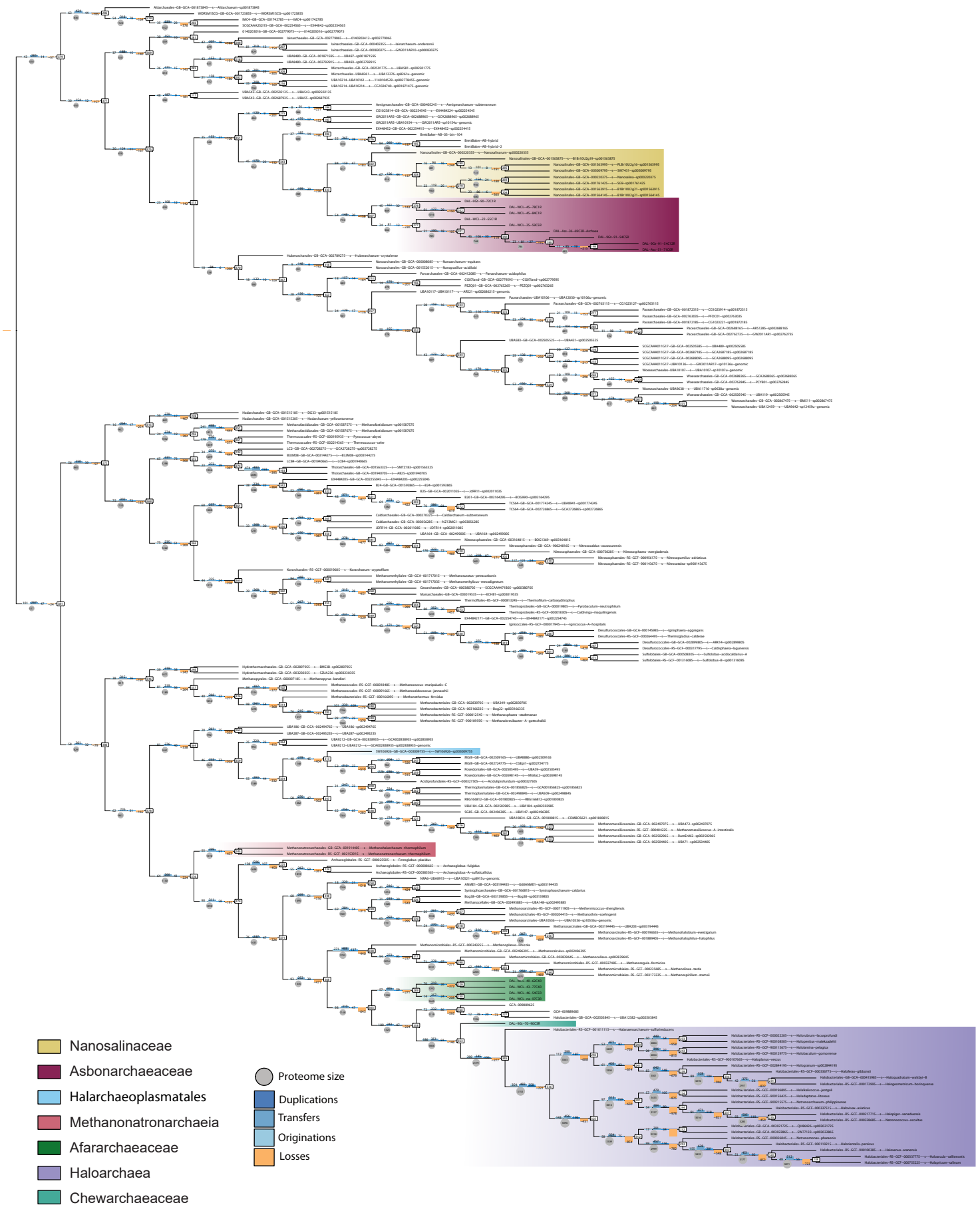
Extended Data Fig. 5 | Likelihood values for alternative positions of the extremely halophilic archaeal lineages. Likelihoods are calculated using IQ-TREE with the LG+C60+F+Γ4 model alone or combined with the new GFmix model (taking into account all significantly enriched (Bin 1) or depleted (Bin 2) amino acids in halophiles or only the most extremely biased ones (D+E and I+K). The highest-scoring topology is indicated with a red rectangle for the **(a)** RP and **(b)** NM datasets. Likelihood differences between a given topology and the highest-scoring topology per model are given in parentheses.



Extended Data Fig. 6 | Halophilic-specific amino acid compositional biases along the phylogeny of 192 archaeal taxa. **(a)** The ratio of [D+E/I+K] amino acids of 192 archaeal taxa was calculated along the untreated NM and RP alignments (39,385 and 6,792 amino acid positions, respectively). **(b)** 10% of the most biased sites (i.e., those with the highest ratio) were removed from the NM and RP alignments. Distinct halophilic clades are indicated in color, including the Nanosalinaceae (sand), Asbonarchaeaceae (wine), Halarchaeoplasmatales (cyan), Methanonatronarchaeia (rose), Afarachaeaceae (green), and Haloarchaea (indigo). The scale bar indicates the expected average number of substitutions per site.

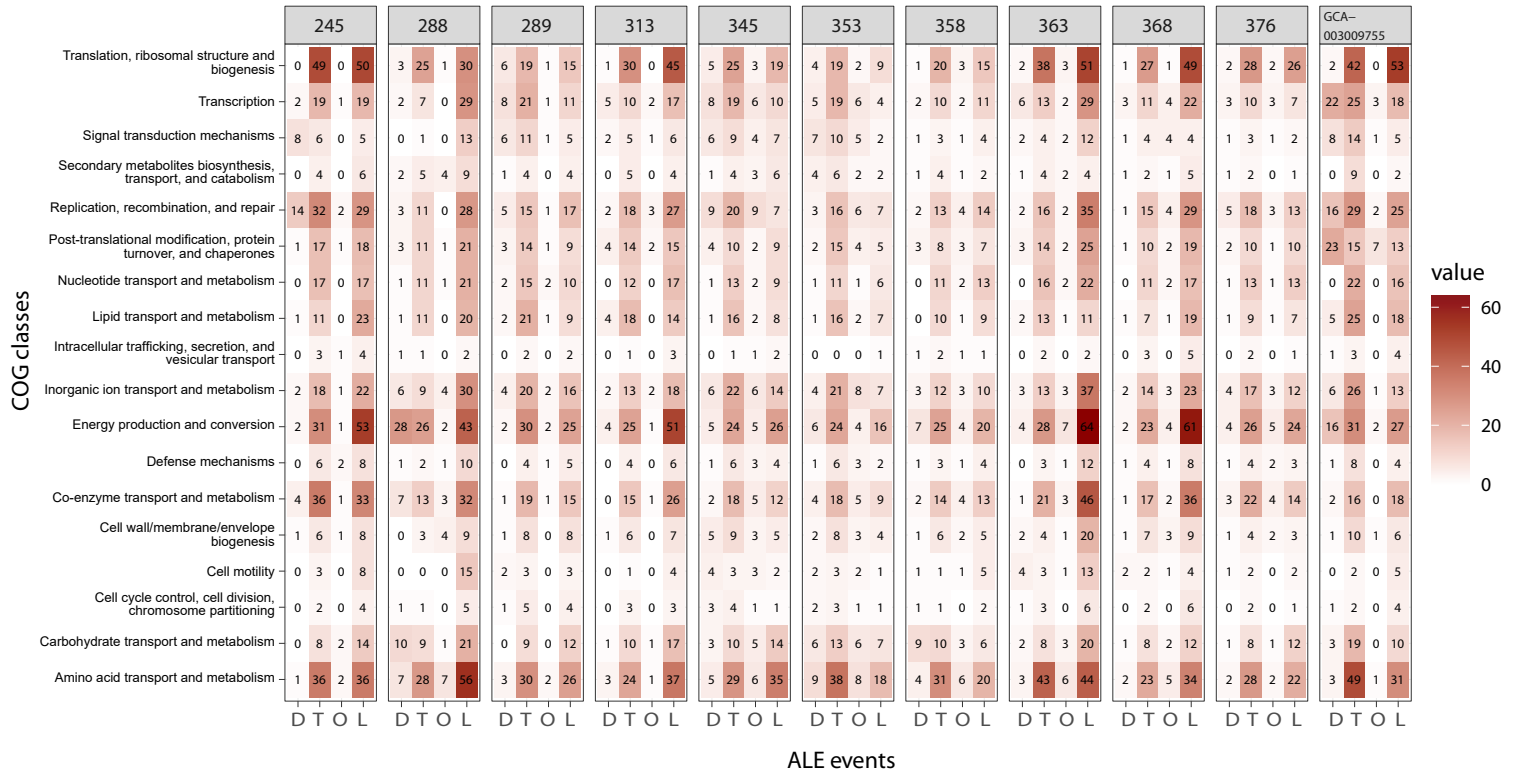


Extended Data Fig. 7 | Impact of compositional bias on the phylogeny of archaeal ATP synthase. Maximum likelihood phylogenetic trees based on the concatenation of ATP synthase subunits A and B **(a)** before and **(b)** after removal of 15% of sites with the highest D+E/I+K ratio. Notice the shift in the position of the Nanosalinaceae+Asbonarchaeaceae group. The trees were reconstructed using the LG+C60+F+Γ4 model of sequence evolution. Numbers at branches indicate 1,000 ultrafast bootstrap support values. Only values >70% are indicated. The scale bar indicates the expected average number of substitutions per site. **(c)** D+E/I+K ratio for all sites in the ATP synthase subunits A and B dataset ordered from highest to lowest values.

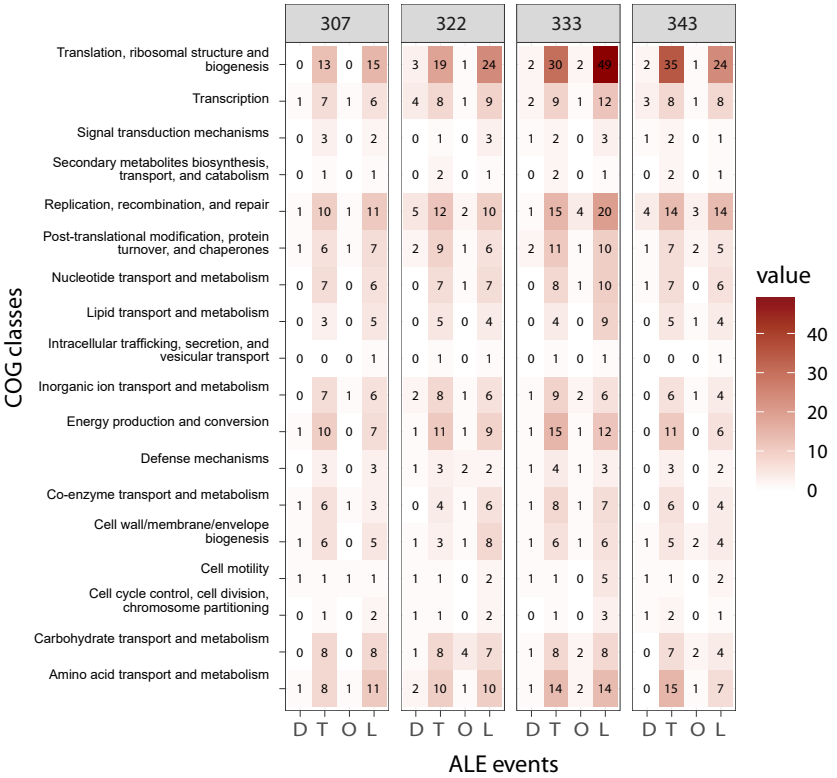


Extended Data Fig. 8 | Ancestral proteome sizes and numbers of gene loss, duplication, and gain inferred from reconciliation analyses across 192-NM archaeal taxa. Barplots at each branch indicate duplication, transfer, and origination events (blue bars; see legend) and loss events (orange bars) (see Methods). Each grey circle indicates the inferred proteome size (i.e., the number of protein-coding gene copies) for the ancestor to the branch's right. Numbers in rounded rectangles correspond to the node identifiers. Halophilic clades are colored.

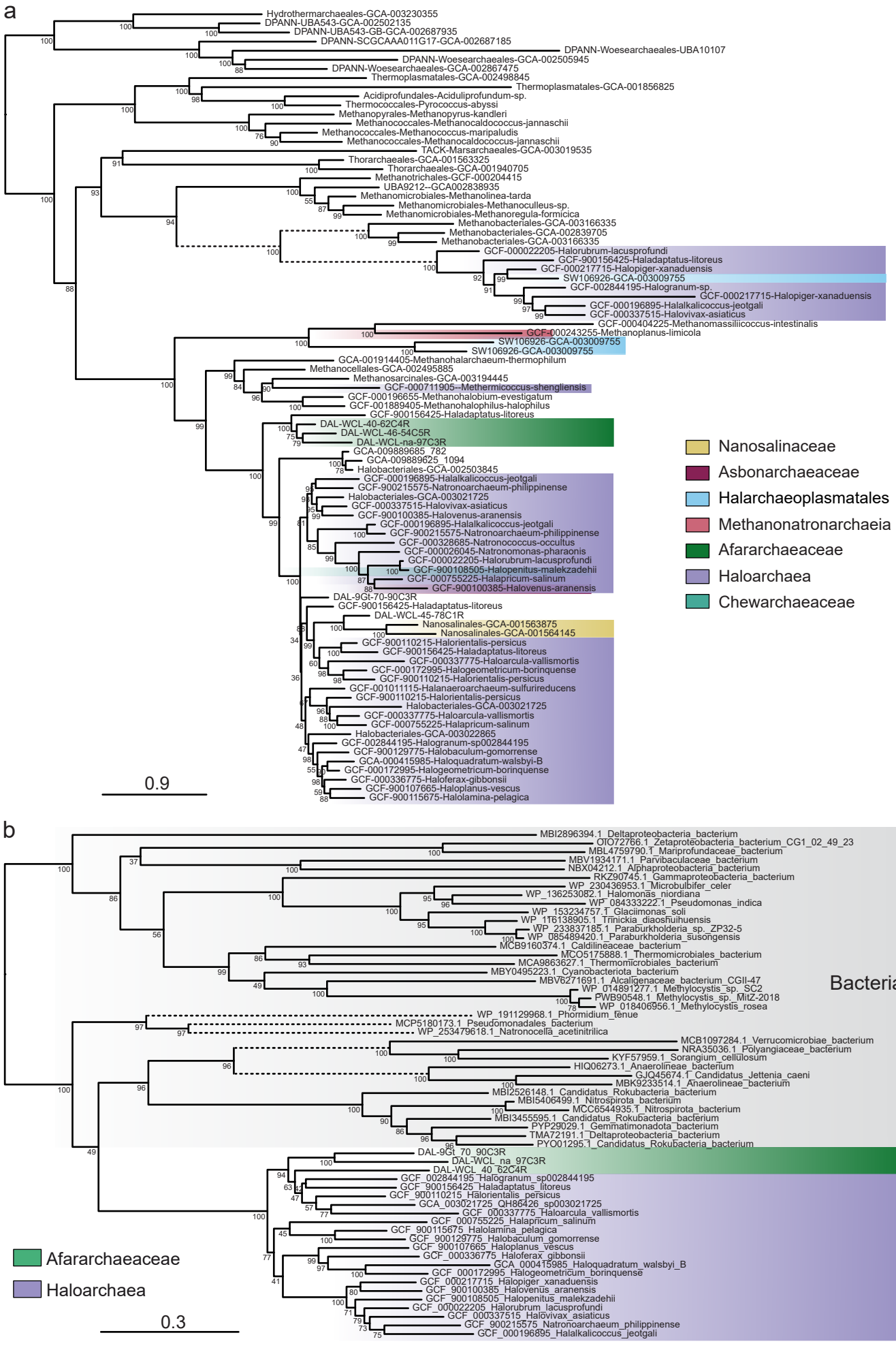
a



b



Extended Data Fig. 9 | Heat map of the number of gene duplications, transfers, originations, and losses in various archaeal halophilic lineages according to their COG classification. The counts were obtained using the amalgamated likelihood estimation (ALE) tree reconciliation method on the set of 17,288 orthologous genes present in the 192-taxa genomic dataset for several nodes within the **(a)** Euryarchaeota and the **(b)** DPANN archaea (see Methods). Node numbers correspond to the nodes in the complete tree shown in Extended Data Fig. 8.



Extended Data Fig. 10 | Maximum likelihood trees showing cases of horizontal gene transfer involving archaeal halophilic lineages. (a) NhaP-type Na+/H+ and K+/H+ antiporters. **(b)** choline dehydrogenase BetA. The trees were constructed with the LG+C60+F+Γ4 model. Dashed branches have been shortened to half of their actual length. The scale bar indicates the expected average number of substitutions per site.