Borgs are giant genetic elements with potential to expand metabolic capacity

https://doi.org/10.1038/s41586-022-05256-1

Received: 21 July 2021

Accepted: 22 August 2022

Published online: 19 October 2022

Open access

Check for updates

Basem Al-Shayeb[1,2], Marie C. Schoelmerich[1], Jacob West-Roberts[3], Luis E. Valentin-Alvarado[1,2], Rohan Sachdeva[1,4], Susan Mullen[4], Alexander Crits-Christoph[1,2], Michael J. Wilkins[5], Kenneth H. Williams[6,7], Jennifer A. Doudna[1,8] & Jillian F. Banfield[1,3,4,6,9] ✉

Anaerobic methane oxidation exerts a key control on greenhouse gas emissions[1], yet factors that modulate the activity of microorganisms performing this function remain poorly understood. Here we discovered extraordinarily large, diverse DNA sequences that primarily encode hypothetical proteins through studying groundwater, sediments and wetland soil where methane production and oxidation occur. Four curated, complete genomes are linear, up to approximately 1 Mb in length and share genome organization, including replichore structure, long inverted terminal repeats and genome-wide unique perfect tandem direct repeats that are intergenic or generate amino acid repeats. We infer that these are highly divergent archaeal extrachromosomal elements with a distinct evolutionary origin. Gene sequence similarity, phylogeny and local divergence of sequence composition indicate that many of their genes were assimilated from methane-oxidizing *Methanoperedens* archaea. We refer to these elements as 'Borgs'. We identified at least 19 different Borg types coexisting with *Methanoperedens* spp. in four distinct ecosystems. Borgs provide methane-oxidizing *Methanoperedens* archaea access to genes encoding proteins involved in redox reactions and energy conservation (for example, clusters of multihaem cytochromes and methyl coenzyme M reductase). These data suggest that Borgs might have previously unrecognized roles in the metabolism of this group of archaea, which are known to modulate greenhouse gas emissions, but further studies are now needed to establish their functional relevance.

Of all of the biogeochemical cycles on Earth, the methane cycle may be most tightly linked to climate. Methane (CH_4) is a greenhouse gas roughly 30 times more potent than carbon dioxide (CO_2), and approximately 1 gigatonne is produced annually by methanogenic (methane-producing) archaea that inhabit anoxic environments[2]. The efflux of methane into the atmosphere is mitigated by methane-oxidizing microorganisms (methanotrophs). In oxic environments, CH_4 is consumed by aerobic bacteria that use methane monooxygenase (MMO) and O_2 as a terminal electron acceptor[3], whereas in anoxic environments, anaerobic methanotrophic archaea (ANME) use a reverse methanogenesis pathway to oxidize CH_4, the key enzyme of which is methyl-CoM reductase (MCR)[4,5]. Some ANMEs rely on a syntrophic partner to couple CH_4 oxidation to the reduction of terminal electron acceptors, yet *Methanoperedens* (ANME-2d, phylum Euryarchaeota) can directly couple CH_4 oxidation to the reduction of iron, nitrate or manganese[6,7]. Some phenomena have been suggested to modulate rates of methane oxidation. For example, some phages can decrease rates of methane oxidation by infection and lysis of methane-oxidizing bacteria[8], and others with the critical subunit of MMO[9] probably increase the ability of their host bacteria to conserve energy during phage replication. Here we report the discovery of novel extrachromosomal elements (ECEs) that are inferred to replicate within *Methanoperedens* spp. Their numerous and diverse metabolism-relevant genes, huge size and distinctive genomic architecture distinguish these archaeal ECEs from all previously reported elements associated with archaea[10–12] and from bacteriophages, which typically have one or a few biogeochemically relevant genes[13,14]. We hypothesize that these novel ECEs may substantially impact the capacity of *Methanoperedens* spp. to oxidize methane.

Genome structure and features

By analysis of whole-community metagenomic data from wetland soils in California, USA (Extended Data Fig. 1), we discovered enigmatic genetic elements, the genomes for three of which were carefully manually curated to completion (Methods). From sediment samples from the Rifle, Colorado aquifer[15], we recovered partial genomes

[1]Innovative Genomics Institute, University of California, Berkeley, CA, USA. [2]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA. [3]Environmental Science, Policy and Management, University of California, Berkeley, CA, USA. [4]Earth and Planetary Science, University of California, Berkeley, CA, USA. [5]Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA. [6]Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [7]Rocky Mountain Biological Lab, Gothic, CO, USA. [8]Department of Chemistry, University of California, Berkeley, CA, USA. [9]The University of Melbourne, Melbourne, Victoria, Australia. ✉e-mail: jbanfield@berkeley.edu

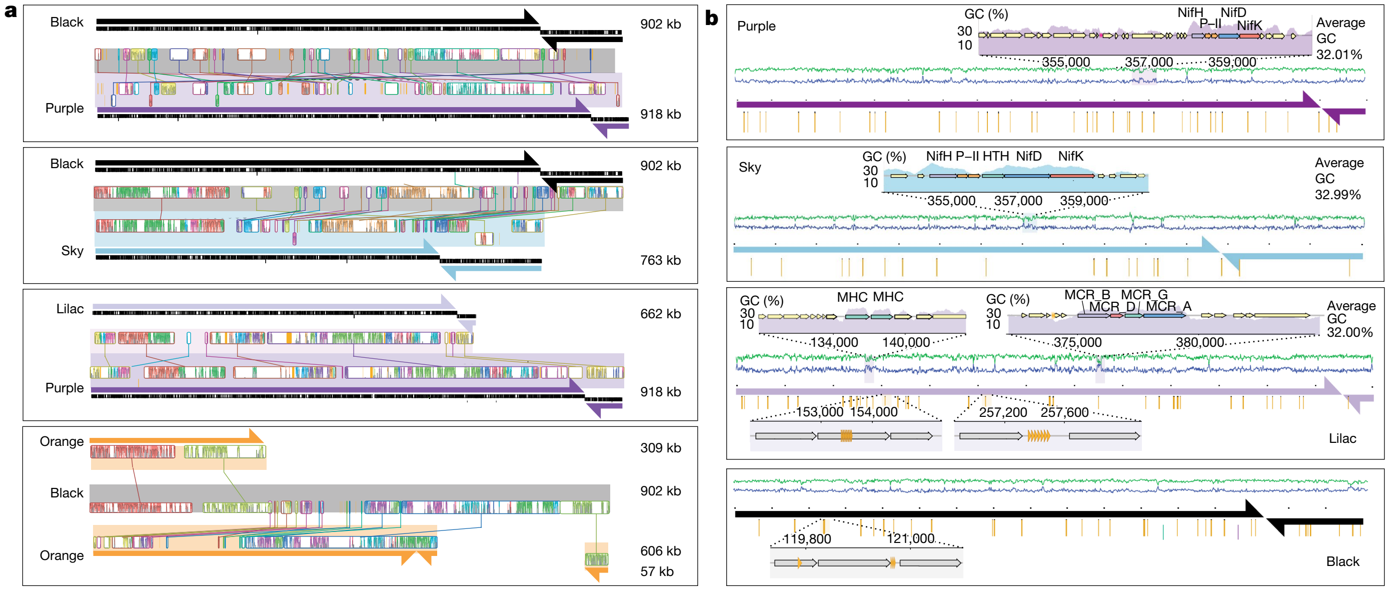


Fig. 1 | Borgs share overall genomic features. a, Genome replichores (arrows) and coding strands (black bars) for aligned pairs of the four complete (Black, Purple, Sky and Lilac) and one near-complete (Orange) Borg. Blocks of sequence with identifiable nucleotide similarity are shown in between each pair (coloured graphs linked by lines; *y* axes show similarity). **b**, Genome overviews showing the distribution of three or more perfect tandem direct repeats (gold rods) along the complete genomes. Insets provide examples of local elevated GC content associated with certain gene clusters and within gene and intergenic tandem direct repeats (gold arrows).

from a single population related to those from the wetland soils; the sequences were combined and manually curated to ultimately yield a fourth complete genome (Methods). All four curated genomes are linear and terminated by more than 1-kb inverted repeats. The genome sizes range from 661,708 to 918,293 kb (Fig. 1a, Extended Data Table 1 and Supplementary Table 1). Prominent features of all genomes are 25–54 regions composed of perfect tandem direct repeats (Fig. 1b and Supplementary Table 2) that are novel (Extended Data Fig. 2) and occur both in intergenic regions and in genes where they usually introduce perfect amino acid repeats (Supplementary Table 2). All genomes have two replichores of unequal lengths and initiate replication at the chromosome ends (Extended Data Fig. 3). Each replichore carries essentially all genes on one strand (Fig. 1a). Although the majority of genes are novel, approximately 21% of the predicted proteins have best matches to proteins of Archaea (Extended Data Fig. 4a), and the largest group of these have best matches to proteins of *Methanoperedens* spp. (Extended Data Fig. 4b). Of note, the GC contents of the four genomes are approximately 10% lower than those of previously reported and coexisting *Methanoperedens* species (Fig. 2a). We rule out the possibility that these sequences represent genomes of novel Archaea, as they lack almost all of the single-copy genes found in archaeal genomes and sets of ribosomal proteins that are present even in obligate symbionts (Extended Data Figs. 5 and 6a and Supplementary Tables 3–6). There are no additional sequences in the datasets that could comprise additional portions of these genomes. Thus, they are clearly neither part of *Methanoperedens* spp. genomes nor parts of the genomes of other archaea.

Abundances of *Methanoperedens* spp. and some ECEs are tightly correlated over a set of 46 different wetland soil samples (43 genomes were included in the analysis; Extended Data Fig. 6b). This observation supports other indications that these ECEs associate with *Methanoperedens* and suggests that specific ECEs have distinct *Methanoperedens* spp. hosts (Fig. 2b). This is true for one ECE whose abundances correlate reasonably well with a specific host group, in which ECE to *Methanoperedens* spp. abundance ratios range from 2:1 to 8:1. Given their up to approximately 1-Mb length, there may be more ECE DNA in some host cells than host DNA. The Borg sequences are much more abundant in deep, anoxic soil samples (Extended Data Fig. 7a,b).

A few percent of the genes in the genomes have locally elevated GC contents that approach, and in some cases match, those of coexisting *Methanoperedens* spp. (Fig. 1b). This, and the very high similarity of some protein sequences to those of *Methanoperedens* spp., indicates that these genes were acquired by lateral gene transfer from *Methanoperedens* spp. Other genes with best matches to *Methanoperedens* spp. genes have lower GC contents (closer to those of these ECEs at approximately 33%), suggesting that their DNA composition has partly or completely ameliorated since acquisition[16].

Archaeal ECEs include viruses[17], plasmids[18] and minichromosomes, sometimes also referred to as megaplasmids[10–12]. The genomes reported here are much larger than those of all known archaeal viruses, some of which have small, linear genomes[12], and at least three are larger than any known bacteriophage[19]. These linear elements are larger than all of the reported circular plasmids that affiliate with halophiles, methanogens and archaeal thermophiles. We did not detect genes for plasmid partitioning or conjugative systems, rRNA loci or encoded viral proteins (Supplementary Table 3), and the genomes were markedly different from recently reported *Methanoperedens* spp. plasmids[20]. The distinctly lower GC content and variable copy number argue against their classification as archaeal minichromosomes[12,21]. Thus, we cannot confidently classify the ECEs as viruses, plasmids or minichromosomes. Moreover, the protein family profiles are quite distinct from those of archaeal and bacterial ECEs (Fig. 2d and Extended Data Fig. 5). Some bacterial megaplasmids have been reported to be very large and linear, but they typically encode few or no essential genes[22], and if they contain repeats, they are interspaced (that is, not tandem)[23]. Each distinctive feature of the ECEs has been reported in microbial genomes, plasmids or viruses, but the combination of these features in these huge ECEs is unique. Thus, we conclude that the genomes represent novel archaeal ECEs that occur in association with, but not as part of, *Methanoperedens* spp. genomes. We refer to these as Borgs, a name that reflects their propensity to assimilate genes from organisms, most notably *Methanoperedens* spp.

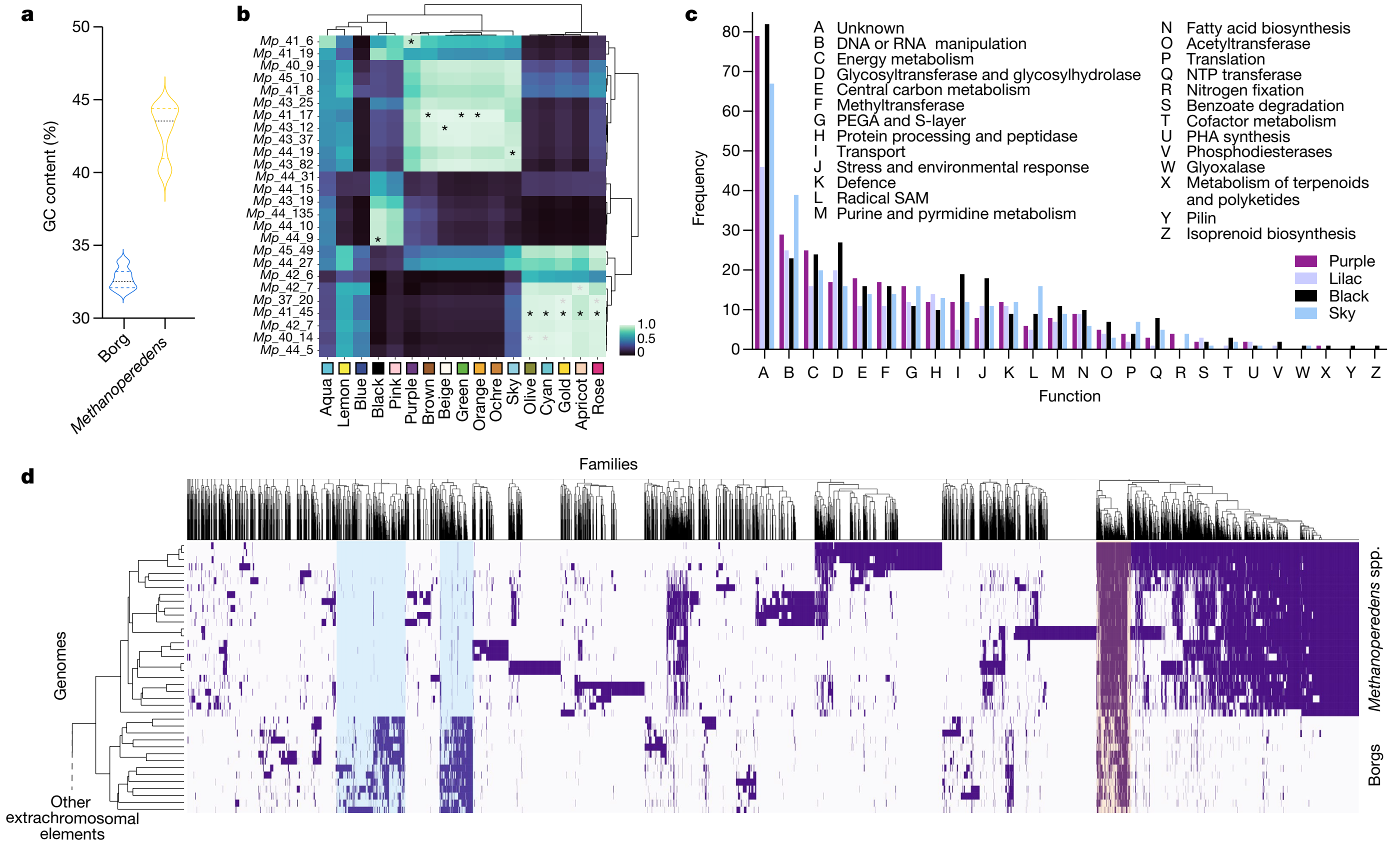


Fig. 2 | Borg and *Methanoperedens* spp. genomic features and abundance patterns. a, The average genome GC contents of Borgs and *Methanoperedens* spp. are distinct. The black line denotes the median, and the dashed lines show the interquartile range. **b**, Groups of related *Methanoperedens* spp. (rows) correlate with groups of Borgs (columns) across a set of 50 samples. The asterisks indicate two-sided Pearson correlations above 0.92 with FDR-corrected *P* values below 2.0 × 10^{-20} that suggest that Brown, Green, Orange, Beige and Ochre Borgs associate with one group of *Methanoperedens* spp., Olive, Cyan, Gold, Apricot and Rose associate with a second group, and Black associate with a third group.Black asterisks indicate best association with a *Methanoperedens* genome (correlation ≥ 0.92, $P \leq 1 \times 10^{-20}$); grey asterisks indicate association with a scaffold containing the *Methanoperedens* L11 marker gene (correlation ≥ 0.92, $P \leq 1 \times 10^{-20}$). **c**, Frequency of genes in different functional groups in the four complete Borg genomes. **d**, Comparison of the protein family composition of Borgs and *Methanoperedens* spp. Clustering on the basis of shared protein family content highlights groups of Borg-specific protein families (blue shading) and protein families shared with their hosts (orange shading). The full clustering, including diverse archaeal mobile elements, is shown in Extended Data Fig. 5. PEGA, surface layer protein; PHA, polyhydroxyalkanoate.

Using criteria based on the features of the four complete Borgs, we searched for additional Borgs in our metagenomic datasets from a wide diversity of environment types. From the wetland soil, we constructed bins for 11 additional Borgs, some of which exceed 1 Mb in length (Extended Data Table 1 and Supplementary Table 1). Other Borgs were sampled from the Rifle, Colorado aquifer, discharge from an abandoned Corona mercury mine in Napa County, California, and from shallow riverbed pore fluids in the East River, Colorado. In total, we recovered genome bins for 19 different Borgs, each of which was assigned a colour-based name. We found no Borgs in some samples, despite the presence of *Methanoperedens* spp. at very high abundance levels (Extended Data Fig. 7). Thus, it appears that these ECEs do not associate with all *Methanoperedens* spp.

Pairs of the four complete Borg genomes (Purple, Black, Sky and Lilac) and three fragments of the Orange Borg are alignable over much of their lengths (Fig. 1a). The Rose and Sky Borg genomes are also largely syntenous (Extended Data Fig. 8a) and were reconstructed from different samples that contain these Borgs at very different levels of abundance. Despite only sharing a less than 50% average nucleotide identity across most of their genomes, the genomes have multiple regions that share 100% nucleotide identity, one of which is approximately 11 kb in length (Extended Data Fig. 8b,c). This suggests that these two Borgs recombined, indicating that they recently coexisted within the same host cell.

Borg gene inventories

Many Borg genomes encode mobile element defence systems, including RNA-targeting type III-A CRISPR–Cas systems that lack spacer acquisition machinery, a feature previously noted in huge bacterial viruses[19]. An Orange Borg CRISPR spacer targets a gene in a mobile region in a coexisting *Methanoperedens* spp. (Extended Data Fig. 8d), further supporting the conclusion that *Methanoperedens* spp. are the Borg hosts.

The four complete genomes and almost all of the near-complete and partial genomes encode ribosomal protein L11 (rpL11), and some have one or two other ribosomal proteins (Extended Data Fig. 6a). The rpL11 protein sequences form a group that places phylogenetically sibling to those of *Methanoperedens* spp. (Extended Data Fig. 9), further reinforcing the link between Borgs and *Methanoperedens* spp. Four additional rpL11 sequences were identified on short contigs from the wetland group with the Borg sequences and probably represent additional Borgs (Supplementary Table 1). The topology of the rpL11 tree, and similar topologies observed for phylogenetic trees constructed using other ribosomal proteins, MCR proteins, electron transfer flavoproteins and aconitase, may indicate the presence of translation-related genes in the Borg ancestor (Extended Data Fig. 9 and Supplementary Information).

The most highly represented Borg genes encode glycosyltransferases, which are proteins involved in DNA and RNA manipulation,

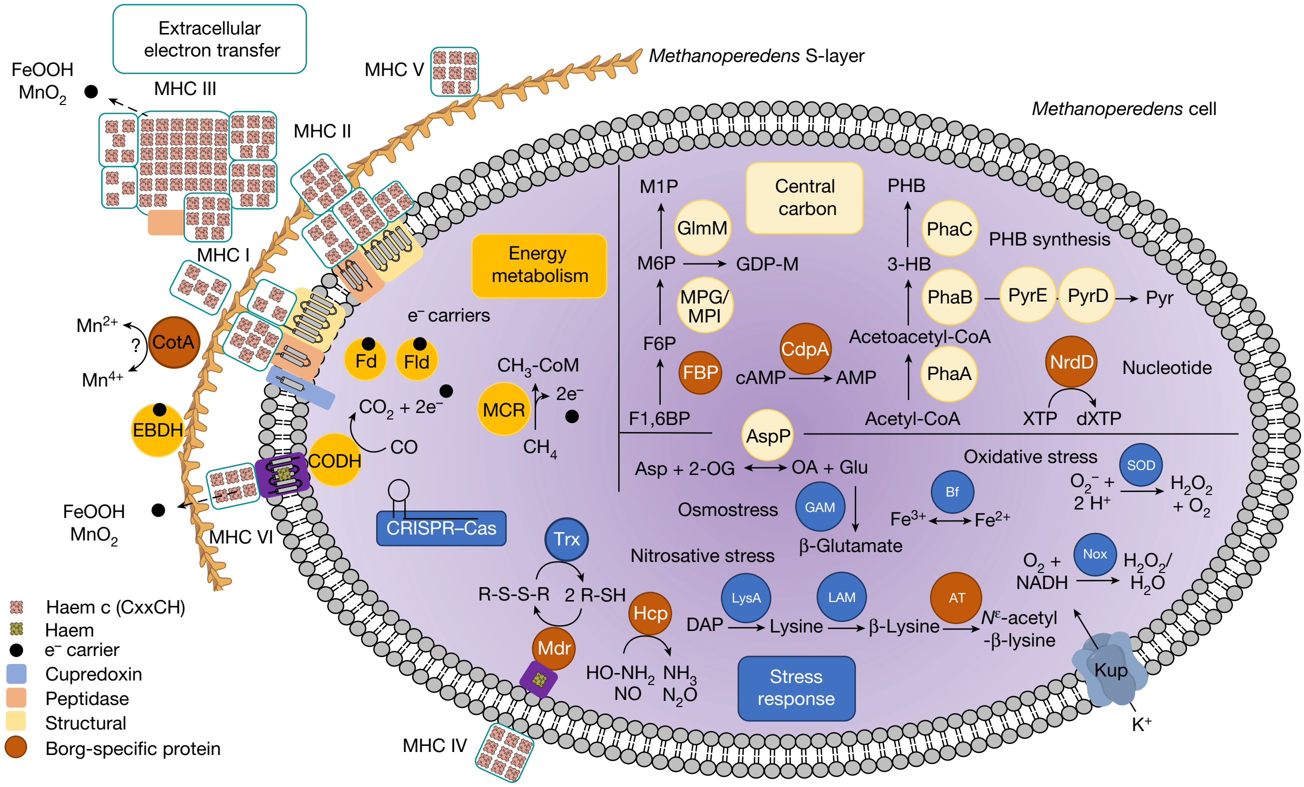


Fig. 3 | Cell cartoon illustrating capacities inferred to be provided to *Methanoperedens* spp. by the coexisting Lilac Borg. Like all Borgs, this Borg lacks the capacity for independent existence, and we infer that it replicates within host *Methanoperedens* spp. cells. Borg-specific proteins are those that were not identified in the genome of coexisting *Methanoperedens* spp. Borg-encoded capacities are grouped into the major categories of energy metabolism (including the MCR complex involved in methane oxidation), extracellular electron transfer (including MHCs) involved in electron transport to external electron acceptors, central carbon metabolism (including genes that enable production of polyhydroxybutyrate (PHB)) and stress response/defence (including production of compatible solutes). Locus codes are listed in Supplementary Table 7.

transport, energy and the cell surface (PEGA and S-layer proteins). Also prevalent are many genes encoding membrane-associated proteins of unknown function that may impact the membrane profile of their host (Fig. 2c). At least seven Borgs carry a *nifHDK* operon for nitrogen fixation, also predicted in *Methanoperedens* spp. genomes, and may augment the influence of the host on nitrogen cycling (Fig. 1b, Supplementary Information and Supplementary Table 6). Potentially related to survival under resource limitation are genes in at least ten Borg genomes for synthesis of the carbon storage compound polyhydroxyalkanoate (PHA), a capacity also predicted for *Methanoperedens* spp.[24]. Other stress-related genes encode tellurium resistance proteins that do not occur in *Methanoperedens* spp. genomes (Supplementary Table 5). All Borgs carry large FtsZ-tubulin homologues that may be involved in cell division, and proteins with the TEP1-like TROVE domain protein that also do not occur in *Methanoperedens* spp. genomes (Supplementary Table 5). These may form a complex similar to Telomerase, Ro or Vault ribonucleoproteins, although their function remains unclear[25]. Several Borgs encode two genes of the tricarboxylic acid cycle (citrate synthase and aconitase; Supplementary Information).

Many Borg genes are predicted to have roles in redox and respiratory reactions. The Black Borg encodes *cfbB* and *cfbC*, genes involved in the biosynthesis of F430, which is the cofactor for MCR, the central enzyme involved in methane oxidation by *Methanoperedens* spp. The similarity in GC content of Borg *cfbB* and *cfbC* and protein sequences of coexisting *Methanoperedens* spp. suggests that these genes were acquired from *Methanoperedens* spp. recently. The Blue and Olive Borgs encode *cofE* (encoding coenzyme F420:L-glutamate ligase), which is involved in the biosynthesis of a precursor for F420. The Blue and Pink Borgs have an electron bifurcating complex (Supplementary Information) that includes D-lactate dehydrogenase. Eight Borgs encode genes for biosynthesis of tetrahydromethanopterin, a coenzyme used in methanogenesis, and ferredoxin proteins, which may serve as electron carriers. The Green and Sky Borgs also encode 5,6,7,8-tetrahydromethanopterin hydro-lyase (Fae), an enzyme responsible for formaldehyde detoxification and involved in pentose-phosphate synthesis. Also identified were genes encoding carbon monoxide dehydrogenase (CODH), plastocyanin, cupredoxins and many multihaem cytochromes (MHCs). These results indicate substantial Borg potential to augment the energy conservation by *Methanoperedens* spp. This is especially apparent for the Lilac Borg.

Host-relevant gene inventory of Lilac Borg

We analysed the genes of the complete Lilac Borg genome in detail as, unlike the other Borgs, the Lilac Borg co-occurs with a single group of *Methanoperedens* spp. that probably represent the host (Fig. 3 and Supplementary Table 7). Remarkably, this Borg genome encodes an MCR complex, which is central to methanogenesis and reverse methanogenesis. The *mcrBDGA* cluster shares high (75–88%) amino acid sequence identity with that of the coexisting *Methanoperedens* spp. genome. This complex is also encoded by a fragment of the Steel Borg. For both the Lilac and the Steel Borgs, the GC content of the region encoding this operon is elevated relative to the average Borg values. *Methanoperedens* spp. pass electrons from methane oxidation to terminal electron acceptors (Fe^{3+}, NO_3^- or Mn^{4+}) via MHCs[26–28]. The Lilac Borg genome encodes 16 MHCs with up to

32 haem-binding motifs within one protein. By analogy with experiments showing that cyanophages with a photosystem gene increase host fitness, we suggest that MHC genes may increase the capacity of *Methanoperedens* spp. to oxidize methane[9,29]. However, this needs to be tested experimentally. Membrane-bound and extracellular MHC may diversify the range of *Methanoperedens* spp. extracellular electron acceptors.

The Lilac Borg encodes a functional NiFe CODH, but this is fragmented in some genomes. Other genes for the acetyl-CoA decarbonylase–synthase complex are present only in *Methanoperedens* spp. The CODH is located in proximity to a cytochrome *b* and cytochrome *c*, so electrons from CO oxidation could be passed to an extracellular terminal acceptor such as Fe^{3+} in an energetically downhill reaction. This would allow the removal of toxic CO and may contribute to the formation of a proton gradient that can be harnessed for energy conservation.

The Lilac Borg has a gene resembling the γ-subunit of ethylbenzene dehydrogenase (EBDH), which is involved in transferring electrons liberated from the hydroxylation of ethylbenzene and propylbenzene[30]. This EBDH-like protein is located extracellularly, and given haem-binding and cohesin domains, it may be involved in electron transfer and attachment.

Although the Lilac Borg lacks genes for a nitrate reductase, it encodes a probable hydroxylamine reductase (Hcp) that may scavenge toxic NO and hydroxylamine byproducts of *Methanoperedens* spp. nitrate metabolism. As the *hcp* gene was not identified in coexisting *Methanoperedens* spp., the Borg gene may protect *Methanoperedens* spp. from nitrosative stress. Proteins such as H_2O_2-forming NADH oxidase (Nox) and superoxide dismutase (SOD) may protect against reactive oxygen species. An alkylhydroperoxidase, two probable disulfide reductases and a bacterioferritin all may detoxify the H_2O_2 byproduct of Nox and SOD. The Lilac Borg also encodes genes that probably augment osmotic stress tolerance. This Borg, but not *Methanoperedens* spp., provides genes to make N^{ε}-acetyl-β-lysine as an osmolyte. An aspartate aminotransferase links the tricarboxylic acid cycle and amino acid synthesis, producing glutamate that can be used for the production of the osmolyte β-glutamate. More importantly, perhaps, it has recently been established that a bacterial homologue of this single enzyme can produce methane from methylamine[31], raising the possibility of methane cycling within the Borg–*Methanoperedens* spp. system.

The Lilac Borg has three large clusters of genes. The first may be involved in cell wall modification, as it encodes large membrane-integral proteins with up to 17 transmembrane domains, proteins for polysaccharide synthesis, glycosyltransferases and probably carbohydrate-active proteins. The second contains key metabolic valves that connect gluconeogenesis with mannose metabolism for the production of glycans. One gene, encoding fructose 1,6-bisphosphatase (FBP), was not identified in the *Methanoperedens* spp. genomes and may regulate carbon flow from gluconeogenesis to mannose metabolism. In between these clusters are 12 genes with PEGA domains with similarity to S-layer proteins. Cell-surface proteins, along with these PEGA proteins, account for approximately 13% of all Lilac Borg genes. We conclude that functionalities related to cell wall architecture and modification are key to the effect of these ECEs on their host, perhaps triggering cell wall modification for adaptation to changing environmental conditions (Fig. 3).

Conclusions

Borgs are enigmatic ECEs that can approach (and probably exceed) 1 Mb in length (Extended Data Table 1). We can neither prove that they are archaeal viruses or plasmids or minichromosomes, nor prove that they are not. Although they may ultimately be classified as megaplasmids, they are clearly different from anything that has been previously reported. It is fascinating to ponder their possible evolutionary origins. Borg homologous recombination may indicate movement among hosts, thus their possible roles as gene transfer agents. It has been noted that *Methanoperedens* spp. have been particularly open to gene acquisition from diverse bacteria and archaea[6], and Borgs may have contributed to this. The existence of Borgs encoding MCR demonstrates for the first time (to our knowledge) that MCR and MCR-like proteins for metabolism of methane and short-chain hydrocarbons can exist on ECEs and thus could potentially be dispersed across lineages, as is inferred to have occurred several times over the course of archaeal evolution[17,32]. Borgs carry numerous metabolic genes, some of which produce variants of *Methanoperedens* spp. proteins that could have distinct biophysical and biochemical properties. Assuming that these genes either augment *Methanoperedens* spp. energy metabolism or extend the conditions under which they can function, Borgs may have far-reaching biogeochemical consequences, with important and unanticipated climate implications. Confirmation that Borgs impact the rate of oxidation of methane by *Methanoperedens* and extend the conditions under which these archaea can function will require experimental evidence. This could be pursued by establishing cultures that include *Methanoperedens* with and without Borgs and comparison of the methane oxidation rates, with testing performed under a range of geochemical conditions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05256-1.

1. Wallenius, A. J., Dalcin Martins, P., Slomp, C. P. & Jetten, M. S. M. Anthropogenic and environmental constraints on the microbial methane cycle in coastal sediments. *Front. Microbiol.* **12**, 631621 (2021).
2. Thauer, R. K., Kaster, A.-K., Seedorf, H., Buckel, W. & Hedderich, R. Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat. Rev. Microbiol.* **6**, 579–591 (2008).
3. Hanson, R. S. & Hanson, T. E. Methanotrophic bacteria. *Microbiol. Rev.* **60**, 439–471 (1996).
4. Boetius, A. et al. A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407**, 623–626 (2000).
5. Hallam, S. J., Girguis, P. R., Preston, C. M., Richardson, P. M. & DeLong, E. F. Identification of methyl coenzyme M reductase A (mcrA) genes associated with methane-oxidizing archaea. *Appl. Environ. Microbiol.* **69**, 5483–5491 (2003).
6. Leu, A. O. et al. Lateral gene transfer drives metabolic fFlexibility in the anaerobic methane-oxidizing archaeal family Methanoperedenaceae. *mBio* **11**, e01325-20 (2020).
7. Ettwig, K. F. et al. Archaea catalyze iron-dependent anaerobic oxidation of methane. *Proc. Natl Acad. Sci. USA* **113**, 12792–12796 (2016).
8. Lee, S. et al. Methane-derived carbon flow through host-virus trophic networks in soil. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.16.423115 (2021).
9. Chen, L.-X. et al. Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat. Microbiol.* **5**, 1504–1515 (2020).
10. Ng, W. V. et al. Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? *Genome Res.* **8**, 1131–1141 (1998).
11. Ausiannikava, D. et al. Evolution of genome architecture in Archaea: spontaneous generation of a new chromosome in *Haloferax volcanii*. *Mol. Biol. Evol.* **35**, 1855–1868 (2018).
12. Wang, H., Peng, N., Shah, S. A., Huang, L. & She, Q. Archaeal extrachromosomal genetic elements. *Microbiol. Mol. Biol. Rev.* **79**, 117–152 (2015).
13. Lindell, D. et al. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl Acad. Sci. USA* **101**, 11013–11018 (2004).
14. Anantharaman, K. et al. Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**, 757–760 (2014).
15. Hug, L. A. et al. Aquifer environment selects for microbial species cohorts in sediment and groundwater. *ISME J.* **9**, 1846–1856 (2015).
16. Lawrence, J. G. & Ochman, H. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44**, 383–397 (1997).
17. Hua, Z.-S. et al. Insights into the ecological roles and evolution of methyl-coenzyme M reductase-containing hot spring Archaea. *Nat. Commun.* **10**, 4574 (2019).
18. DasSarma, S., Capes, M. & DasSarma, P. in *Microbial Megaplasmids* (ed. Schwartz, E.) 3–30 (Springer Berlin Heidelberg, 2009).
19. Al-Shayeb, B. et al. Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
20. Schoelmerich, M. C. et al. A widespread group of large plasmids in methanotrophic *Methanoperedens* archaea. Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.01.478723 (2022).

21. Hall, J. P. J., Botelho, J., Cazares, A. & Baltrus, D. A. What makes a megaplasmid? *Phil. Trans. R. Soc. B* **377**, 20200472 (2022).
22. Medema, M. H. et al. The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. *Genome Biol. Evol.* **2**, 212–224 (2010).
23. Wagenknecht, M. et al. Structural peculiarities of linear megaplasmid, pLMA1, from *Micrococcus luteus* interfere with pyrosequencing reads assembly. *Biotechnol. Lett.* **32**, 1853–1862 (2010).
24. Liu, Z. et al. Domain-centric dissection and classification of prokaryotic poly (3-hydroxyalkanoate) synthases. Preprint at *bioRxiv* https://doi.org/10.1101/693432 (2019).
25. Berger, W., Steiner, E., Grusch, M., Elbling, L. & Micksche, M. Vaults and the major vault protein: novel roles in signal pathway regulation and immunity. *Cell. Mol. Life Sci.* **66**, 43–61 (2009).
26. Cai, C. et al. A methanotrophic archaeon couples anaerobic oxidation of methane to Fe(III) reduction. *ISME J.* **12**, 1929–1939 (2018).
27. McGlynn, S. E., Chadwick, G. L., Kempes, C. P. & Orphan, V. J. Single cell activity reveals direct electron transfer in methanotrophic consortia. *Nature* **526**, 531–535 (2015).
28. Scheller, S., Yu, H., Chadwick, G. L., McGlynn, S. E. & Orphan, V. J. Artificial electron acceptors decouple archaeal methane oxidation from sulfate reduction. *Science* **351**, 703–707 (2016).
29. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**, 86–89 (2005).
30. Heider, J., Szaleniec, M., Sünwoldt, K. & Boll, M. Ethylbenzene dehydrogenase and related molybdenum enzymes involved in oxygen-independent alkyl chain hydroxylation. *J. Mol. Microbiol. Biotechnol.* **26**, 45–62 (2016).
31. Wang, Q. et al. Aerobic bacterial methane synthesis. *Proc. Natl Acad. Sci. USA* **118**, e2019229118 (2021).
32. Boyd, J. A. et al. Divergent methyl-coenzyme M reductase genes in a deep-subseafloor Archaeoglobi. *ISME J.* **13**, 1269–1279 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Sampling and creation of metagenomic datasets

We analysed sequences from sediments of an aquifer in Rifle, Colorado, USA, that were retrieved from cores from depths of 5 m and 6 m below the surface[15] in July 2011, and cell concentrates from pumped groundwater from the same aquifer collected at a time of elevated O_2 concentration in May 2013. Discharge from the Corona Mine, Napa County, California, USA, was sampled in December 2019. Shallow pore water was collected from the riverbed at the East River, Crested Butte, Colorado sampled in August 2016. Soil was sampled from depth intervals between 1 cm and 1 m from a permanently moist wetland located in Lake County, California. Wetland soils were sampled in late October and early November 2017, 2018 and 2019. DNA was extracted from each sample (DNeasy PowerSoil Pro) and submitted for Illumina sequencing (150-bp or 250-bp reads) at the QB3 facility, University of California, Berkeley. Reads were adapter and quality trimmed using BBduk[33] and sickle[34]. Filtered reads were assembled using IDBA-UD[35] and MEGAHIT, gene predictions were established using Prodigal[36] and USEARCH[37] was used for initial annotations[34,35,37,38]. Functional predictions and predictions of tRNAs followed previously reported methods[19].

Genome identification, binning and curation

Hundreds of kilobytes of de novo-assembled sequences were identified to be of interest as potential novel ECEs first based on their taxonomic profile. The taxonomic profiles were determined through a voting scheme in which the taxonomy is assigned at the species to domain level (Bacteria, Archaea, Eukaryotes and no domain) by comparison with a sequence database (protein annotations in the UniProt and ggKbase: https://ggkbase.berkeley.edu/) when the same taxonomic assignment received >50% votes. Assembled sequences selected for further analysis had no taxonomic profile, even at the domain level. The majority of contigs of interest had more genes with similarity to those of archaea of the genus *Methanoperedens* spp. than to any other genus (see Extended Data Fig. 4). The second feature of interest was dominance by hypothetical proteins yet absence of genes that would indicate identification as phage or viruses or plasmids.

These initially identified large fragments were manually curated to remove scaffolding gaps and local assembly errors, to extend and join contigs with the same profile, GC and coverage, and then to extend the near-complete sequences fully into their long terminal repeats. The last step required reassignment of reads mapped at one end and at double depth to both ends. The fully extended sequences had no unplaced reads extending outwards, despite genome-wide deep coverage. Given this, and the absence of any fragments that could potentially be part of a larger genome, it was concluded that sequences represented linear genomes.

In more detail, our curation method involved mapping of reads to the de novo fragments and extension within gaps and at termini using previously unplaced reads that we added based on overlap or by the relocation of misplaced reads (these could often be identified based on improper paired read distances and/or wrong orientation). Local assembly errors were sought by visualization of the reads mapped throughout the assembly and identified based on imperfect read support, or where a subset of reads was partly discrepant and discrepancies involved sequences that were shared by tandem direct repeats of the same region (that is, the tandem direct repeat regions were collapsed during assembly). De novo-assembled sequences often ended in tandem direct repeat regions because repeats fragment assemblies. To resolve local assembly errors, gaps were inserted and reads relocated to generate the sequence required to fill the gaps. This ensured comprehensive essentially perfect agreement between reads and the final consensus sequence. In some cases, the tandem direct repeat regions had greater than the expected depth of mapped reads and no reads spanned the flanking unique sequences. In these cases, the repeat number was approximated to achieve the expected read depth, but some arrays may be larger than shown. GC skew and cumulative GC skew were calculated using iRep[39] for the fully manually curated complete genomes, and the patterns were used to identify the origins and terminus of replication. The pattern of use of coding strands for genes (predicted in Bacterial Code 11) was compared with these origin and terminus predictions to resolve genome organization. The curated sequences were searched for perfect repeats of lengths of 50 or more nucleotides using Repeat Finder in Geneious. When repeat sequences overlapped, the unit of direct repeat was identified and the length of that repeat, number of repeats, location (within gene versus intergenic) and genome position were tabulated. Once the features characteristic of the ECEs of interest had been determined, we sought related elements. Sequences of interest were identified based on (1) credible partial alignment with the complete sequences, (2) no domain-level profile, (3) GC content of 30–35%, (4) regions with three or more direct tandem repeats scattered throughout the genome fragment, and (5) more best hits to *Methanoperedens* spp. proteins than to proteins from any other organisms. If scaffolds met criterion (1) they were immediately classified as targets. If they met most or all of the other criteria and had similar coverage values, they were binned together with other scaffolds from the same sample with these features. Often, ends of some of the contigs in the same bin overlapped perfectly and could be joined, increasing confidence in the bin quality. Genome sequences were aligned to each other using Mauve[40]. Where anomalously high (perfect) sequence identity suggestive of recent recombination was detected between Borgs, reads mapped to the region were visualized to verify that the assembly was correct (that is, not chimeric; also see information in the Extended Data).

Genome fragments were phylogenetically profiled to establish relatedness to sequences in public databases. Sequences were classified as having no detectable hit if the protein had no similar database sequence with an $E < 0.0001$.

Correlation analyses

Reads from each sample were aligned to each *Methanoperedens* and Borg genome. Alignments were performed using bbmap[41] using the following parameters: editfilter = 5, minid = 0.96, idfilter = 0.97, ambiguous = random. The number of reads aligning to each genome was then parsed into a matrix and the correlation between abundance patterns for *Methanoperedens* and Borg genomes was then calculated using Pearson correlation metric as implemented in scipy[42]. Correlation between a *Methanoperedens* genome and a Borg genome was deemed significant if the Pearson correlation between the two genomes was higher than 0.92. The code used for this analysis is available through Zenodo (https://doi.org/10.5281/zenodo.6887003).

CRISPR–Cas analysis

Borg and *Methanoperedens*-encoded CRISPR repeats and spacers were identified using CRISPRDetect[43]. The coding sequences from this study were searched against Cas gene sequences reported from previous studies[44] using hmmsearch with $E < 1 \times 10^{-5}$ to identify the full locus. Matches were checked using a combination of hmmscan and BLAST searches against the NCBI nr database and manually verified by identifying colocated CRISPR arrays and Cas genes. Spacers extracted from between repeats of the CRISPR locus were compared with sequence assemblies from the sites where Borgs were identified using BLASTN-short [45]. Matches with alignment length of more than 24 bp and 1 or less mismatch were retained and targets were classified as bacteria, phage or other. CRISPR arrays that had 1 or less mismatch, were further searched for more spacer matches in the target sequence by finding more hits with three or less mismatches.

Protein and gene content analysis

After the identification and curation of Borg genomes and accumulation of usearch annotations for coding sequences, functional annotations were further assigned by searching against PFAM r32, KEGG, pVOG. Transmembrane regions in proteins were predicted with TMHMM. All *Methanoperedens* genomes and genome assemblies, as well as 1,153 archaeal viruses and ECEs were downloaded from the NCBI RefSeq database. Open reading frames were predicted using Prodigal, and all proteins from Borg genomes and the reconstructed ECE database were clustered into protein families and compared across genomes as previously described[19]. In brief, the coding sequences were clustered into families using a two-step procedure; first an all-versus-all sequence search was performed using an *E* value cut-off of 1×10^{-3}, sensitivity of 7.5 and coverage of 0.5, and a sequence similarity network was built on the basis of the pairwise similarities and the greedy set cover algorithm to define protein subclusters. The resulting subclusters were grouped into protein families using a comparison of hidden Markov models. For subfamilies with probability scores of at least 95% and coverage at least 0.50, a similarity score (probability × coverage) was used as weight of the input network in the final clustering using the Markov clustering algorithm, with 2.0 as the inflation parameter. These clusters were defined as the protein families.

Functional annotation

Genes of interest were further verified and compared using the conserved domain search in NCBI and InterproScan[46] to identify conserved motifs within the amino acid sequence. MHCs were identified based on three or more CxxCH motifs within one gene. The cellular localization of proteins was predicted with Psort (v3.0.3) using archaea as the organism type. Proteins were compared using blastp and aligned using MAFFT[47] v.7.407 to visualize homologous regions and check conserved amino acid residues that constitute the active site or are required for cofactor and ligand binding.

Phylogenetic trees

For each gene, references were compiled by BLASTing the corresponding gene against the NCBI nr database, and their top 50 hits clustered by CD-HIT using a 90% similarity threshold[48]. The final set of genes was aligned using MAFFT v.7.407, and a phylogenetic tree was inferred using IQTREE v.1.6.6 using automatic model selection[49] and visualized using iTOL[50]. Synteny plots were generated using Mauve[51] and gene clusters through Adobe Illustrator and gggenes.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The Borg and *Methanoperedens* genomes and their proteins reported in this study are provided as Source Data (Supplementary Data), along with phylogenetic trees and alignments related to ribosomal protein analysis from Borgs and *Methanoperedens*. Genomes and reads can be accessed via PRJNA866293.

Code availability

The code used to perform the correlation analysis is available through Zenodo (https://doi.org/10.5281/zenodo.6887003). All other code is readily available at the cited sources.

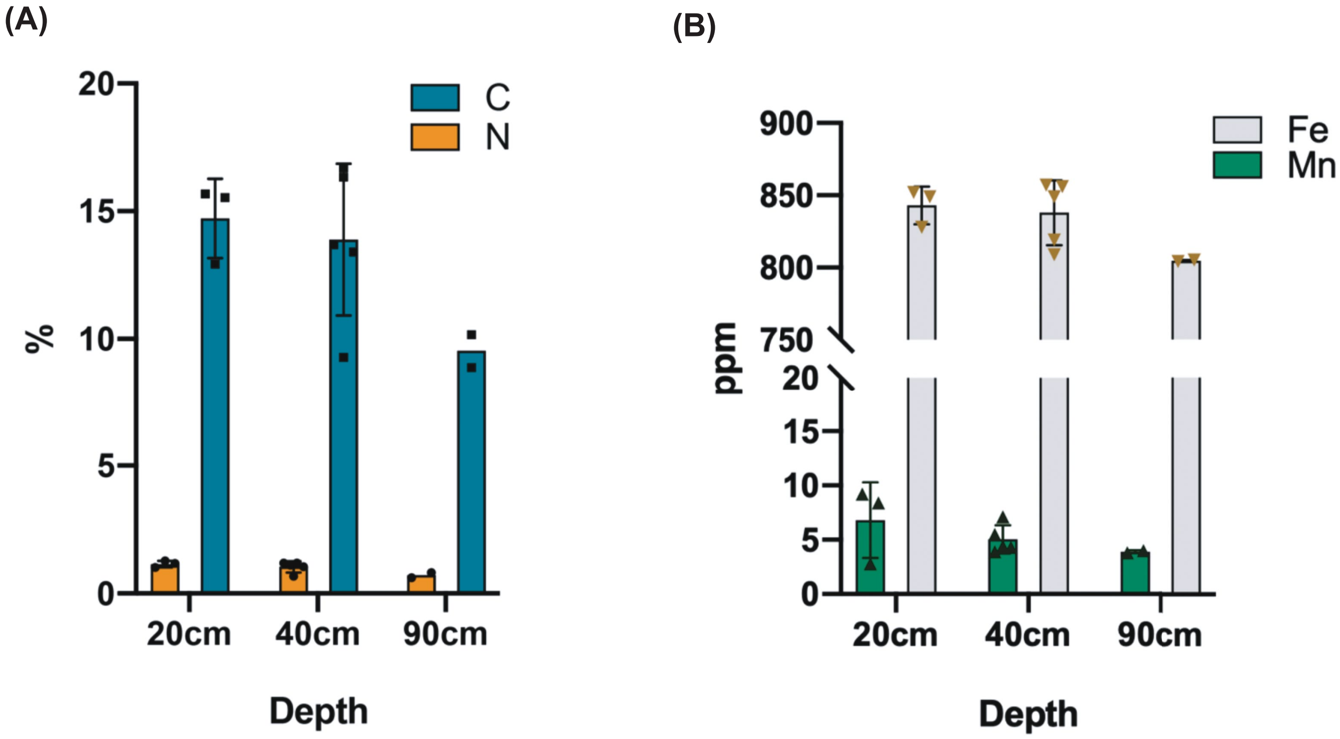
33. Bushnell, B. BBTools software package. http://sourceforge.net/projects/bbmap (Source Forge, 2014).
34. Joshi, N. & Fass, J. N. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. GitHub https://github.com/najoshi/sickle (2011).
35. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
36. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
37. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
38. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & 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 (2015).
39. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol.* **34**, 1256–1263 (2016).
40. Darling, A. E., Mau, B. & Perna, N. T.progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
41. Bushnell, B. BBMap: A fast, accurate, splice-aware aligner. *OSTI.gov* https://www.osti.gov/biblio/1241166 (2014).
42. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
43. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).
44. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
45. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
46. McWilliam, H. et al. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.* **41**, W597–W600 (2013).
47. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
48. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
49. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
50. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
51. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).

Acknowledgements We thank G. Tyson, D. Nayak, N. Baliga, J. Cate, S. Diamond, R. Hatzenpichler and A. Murat Eren for helpful discussion; E. Smith, who proposed the name 'Borg'; L. Law and S. Lei for data management assistance; and Y. Amano for permission to mention *Methanoperedens* spp.-dominated metagenomic datasets in which we did not discover Borgs. This research was supported by an NSF Fellowship to B.A.-S., DFG Fellowship to M.C.S., the US Department of Energy, Office of Science, Office of Biological and Environmental Research under award number DE-AC02- 05CH11231, the Chan Zuckerberg Biohub and the Innovative Genome Institute, UC Berkeley.

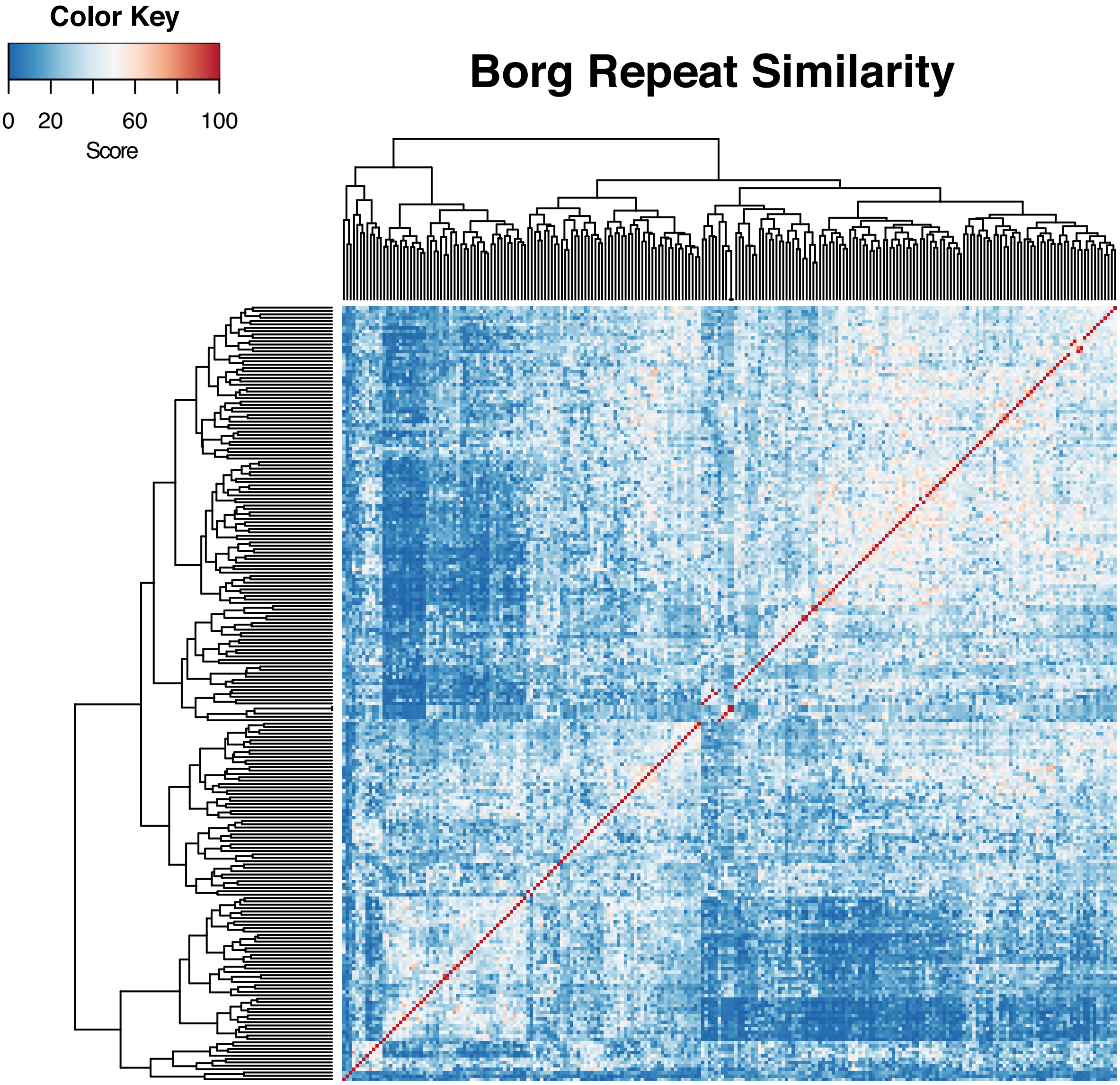
Author contributions The study was conceived by B.A.-S. and J.F.B. Metagenomic datasets were contributed by B.A.-S., R.S., L.E.V.A., A.C.-C., J.W.-R., M.J.W., S.M., K.H.W. and J.F.B. Genome binning was done by J.F.B., B.A.-S. and A.C.-C. Manual genome curation was conducted by J.F.B., with read mappings and CRISPR–Cas analysis from B.A.-S. Borg genome structure, taxonomic breakdowns, horizontal gene transfer and general feature analyses were conducted by B.A.-S. and J.F.B. J.W.-R., B.A.-S. and J.F.B. calculated relative abundances of Borgs and *Methanoperedens* spp. Phylogenetic analyses were conducted by B.A.-S., and repeat sequence comparisons across Borgs were done by B.A.-S. and J.F.B. General Borg and *Methanoperedens* spp. gene inventory and protein family analyses were done by B.A.-S. and J.F.B. Lilac Borg in-depth analysis was done by M.C.S. J.A.D. provided advisory support. B.A.-S., M.C.S. and J.F.B. wrote the manuscript, with input from all authors.

Competing interests J.F.B. is a co-founder of Metagenomi. J.A.D. is a cofounder of Caribou Biosciences, Editas Medicine, Scribe Therapeutics, Intellia Therapeutics and Mammoth Biosciences; is a scientific advisory board member of Vertex, Caribou Biosciences, Intellia Therapeutics, Scribe Therapeutics, Mammoth Biosciences, Algen Biotechnologies, Felix Biosciences, The Column Group and Inari Agriculture; is Chief Science Advisor to Sixth Street, a director at Johnson & Johnson, Altos and Tempus; and has research projects sponsored by Apple Tree Partners and Roche.

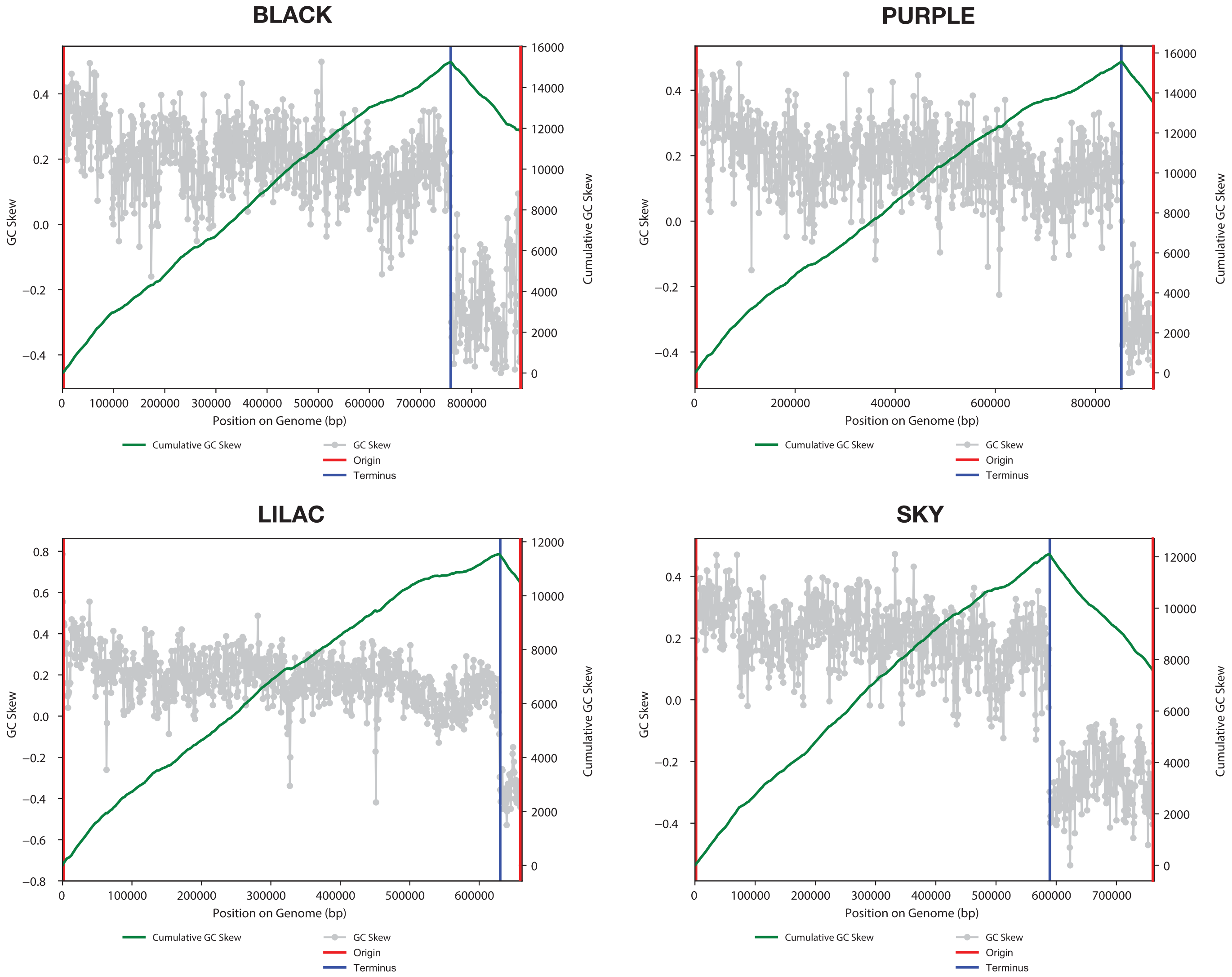
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05256-1.
Correspondence and requests for materials should be addressed to Jillian F. Banfield.
Peer review information *Nature* thanks Christian Rinke, Rudolf Thauer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



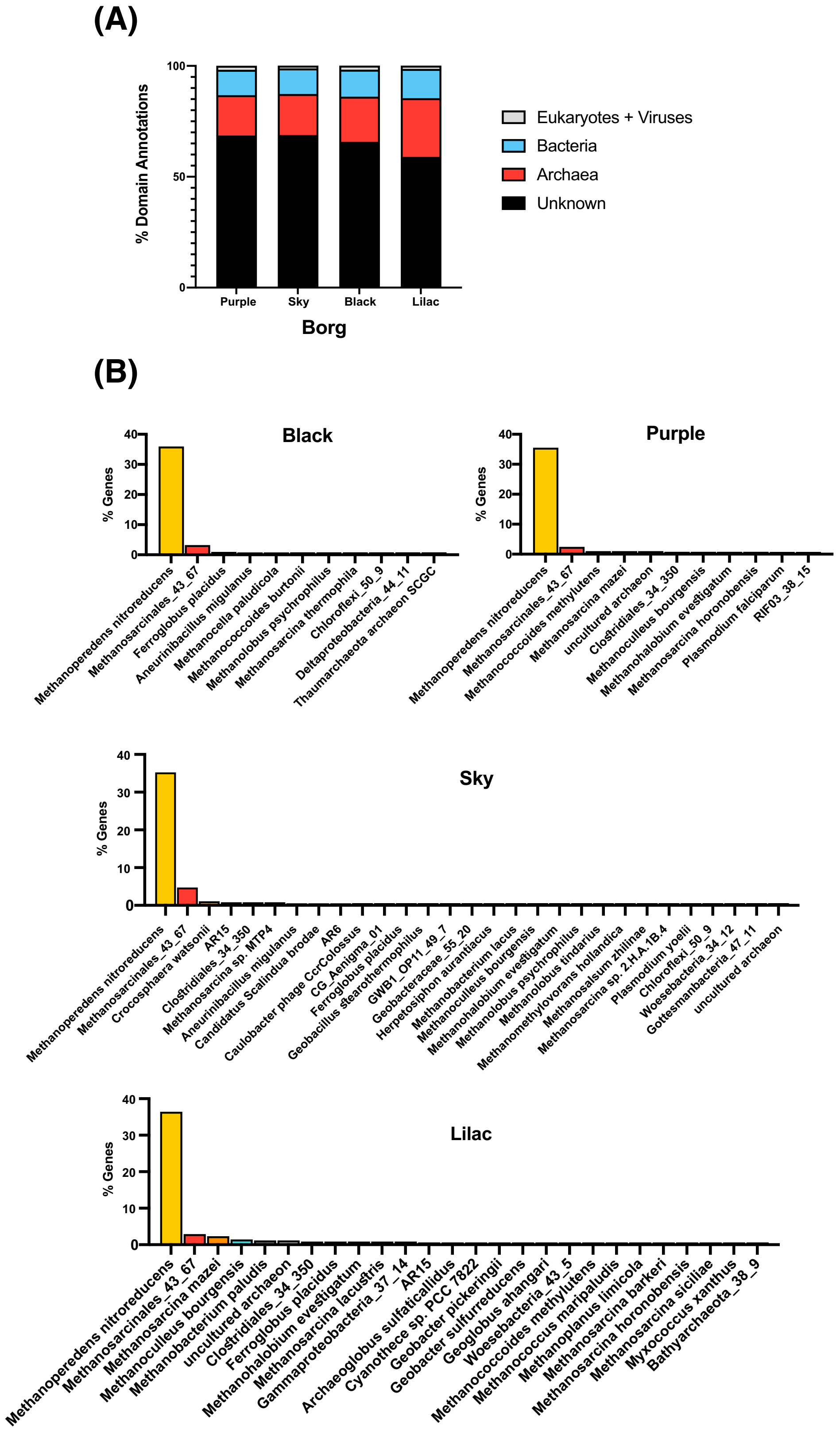
Extended Data Fig. 1 | Geochemical profiles of the permanently moist and organic-rich wetland soils. (A) The mean concentrations of total carbon, nitrogen as well as (B) iron and manganese in wetland soils at 20 cm (n = 3), 40 cm (n = 5), and 90 cm (n = 2) where n denotes the number of biological samples. Deeper soils, where these extrachromosomal elements are most abundant, are somewhat depleted in carbon, iron and manganese compared to shallow soils. Error bars denote standard deviation. 36 samples were collected and sequenced, with 1 to 10 independent samples collected from the same soil depth.



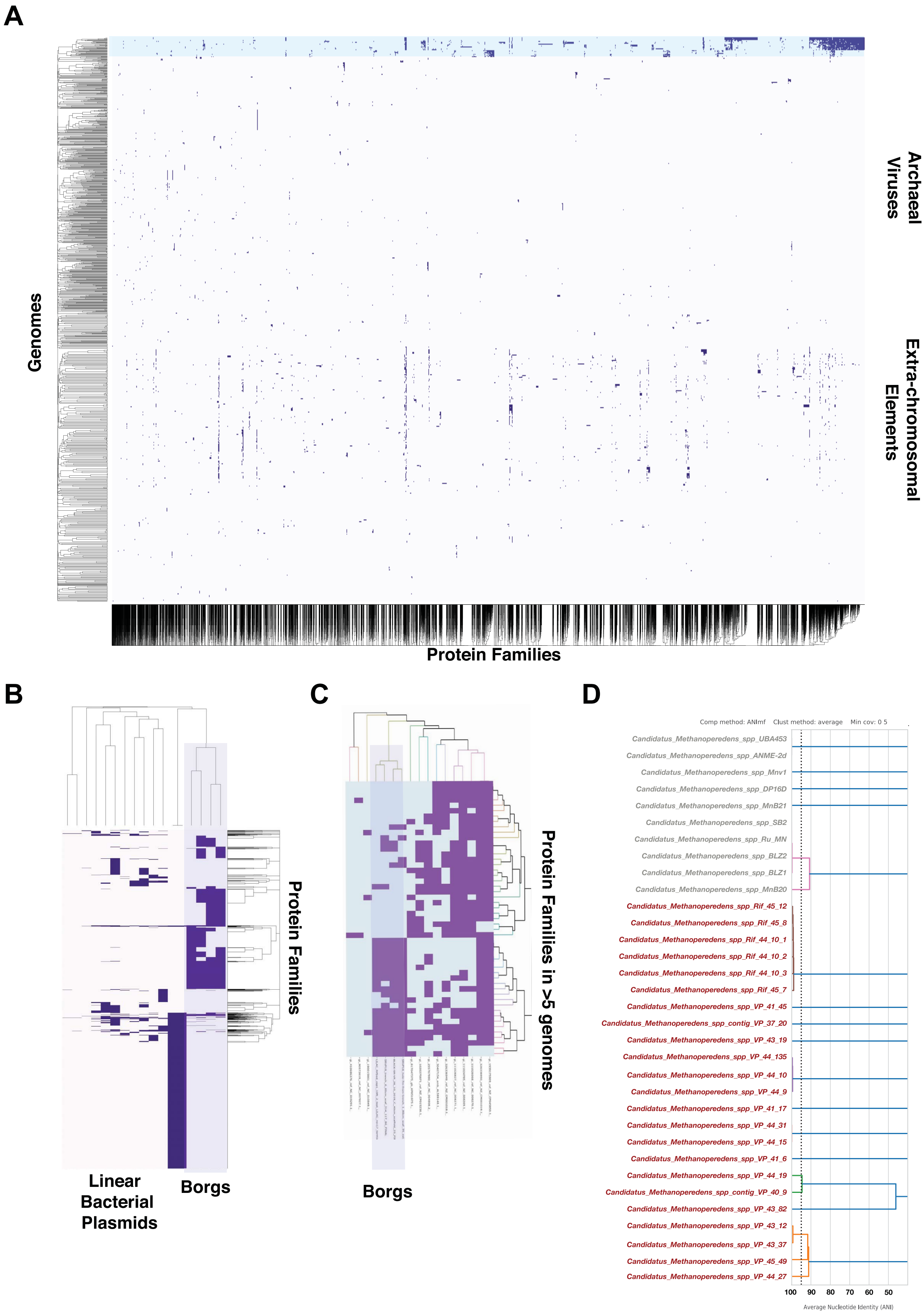
Extended Data Fig. 2 | Sets of three or more perfect tandem direct repeats (TDR) are a characteristic feature of the Borg genomes. Up to 54 instances occur in the four complete Borg genomes, with, on average, one repeat every 12 (Lilac) – 31 (Sky) kbp. These repeat regions fragment assemblies and cause local assembly errors, which we resolved by manual curation (Methods). Within the TDR regions of the four curated, complete genomes, the unit repeats occur up to 20 times and unit repeats are up to 54 bps in length (Supplementary Table 2). Between 54 and 64% of these perfect TDRs are encoded in intergenic regions, although part or all of the first repeat may occur within the C-terminus of a protein-coding gene. When the TDRs occur within proteins, the unit lengths are almost always divisible by 3, so they introduce perfect amino acid repeats. TDR sequences within a single Borg genome are almost always unique. Repeat sequence comparison from the four complete curated Borgs highlights the novelty of almost all TDR sequences (both within and across genomes).



Extended Data Fig. 3 | All genomes have two replichores of unequal lengths. GC skew (grey plots) and cumulative GC skew (green lines) across the four complete Borg genomes, all of which end in long inverted terminal repeats (1.4–2.7 kbp in length). The cumulative GC skew plots indicate replication is initiated in these terminal repeats (red lines). Blue lines mark the predicted replication termini. The red and blue lines define two replichores of unequal length that correspond almost completely to distinct coding strands (almost all genes on the +ve strand of the large replichore and on the -ve strand of the small replichore).

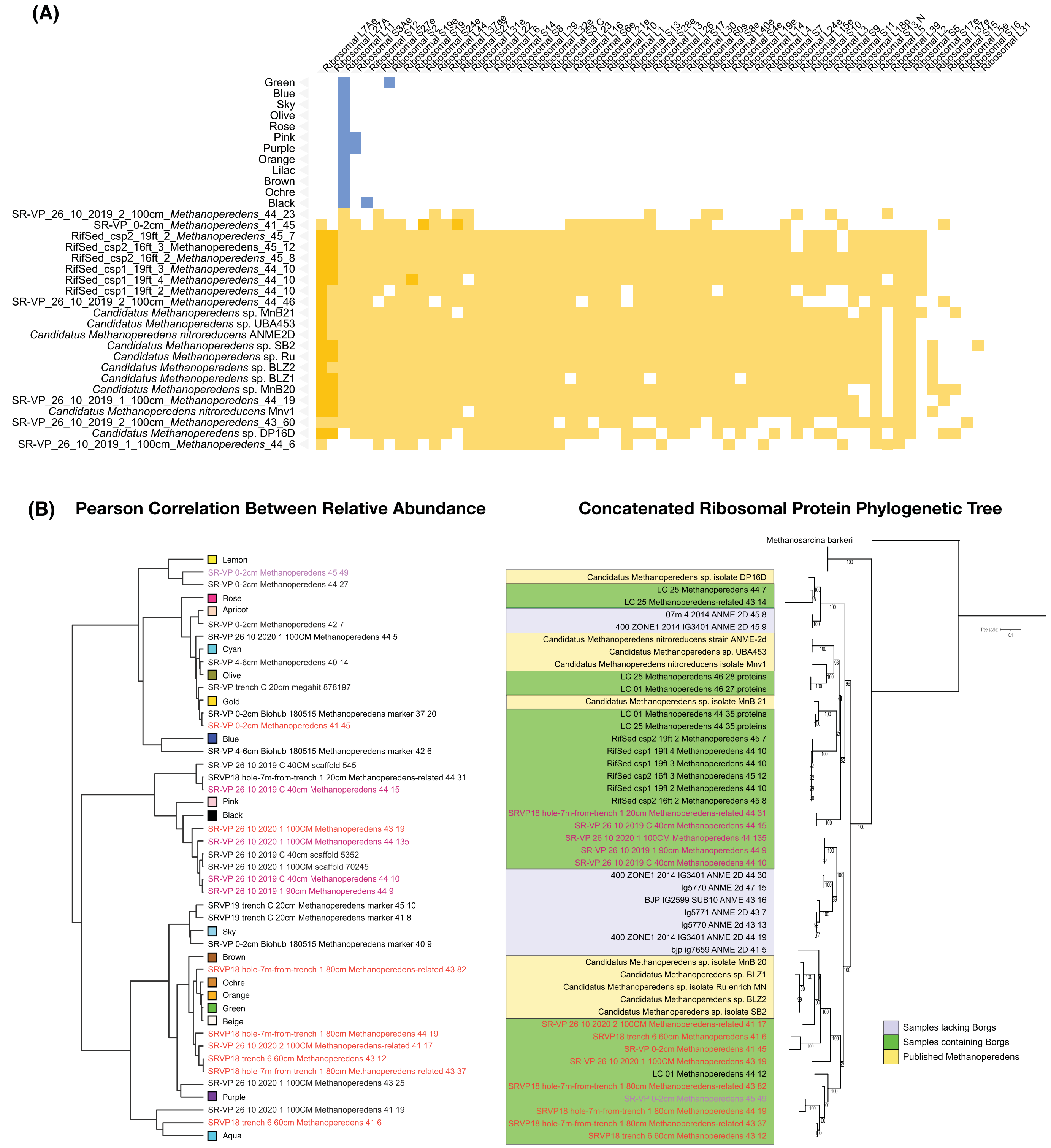


Extended Data Fig. 4 | Taxonomic profiles of the four complete Borg genomes. A. In all cases, the majority of proteins have no similarity to proteins in the reference database ("Unknown"; e-value of > 0.0001). For the cases where a protein has an identifiable hit (blue and red bars in A), the plots in **B**. show the taxonomy of the organisms in which those hits were identified. Only cases where the same organism accounted for hits for > 0.5% of genes are shown. The results clearly indicate that the vast majority of cases where proteins have identifiable matches involve matches to proteins of *Methanoperedens* spp. (gold bars).

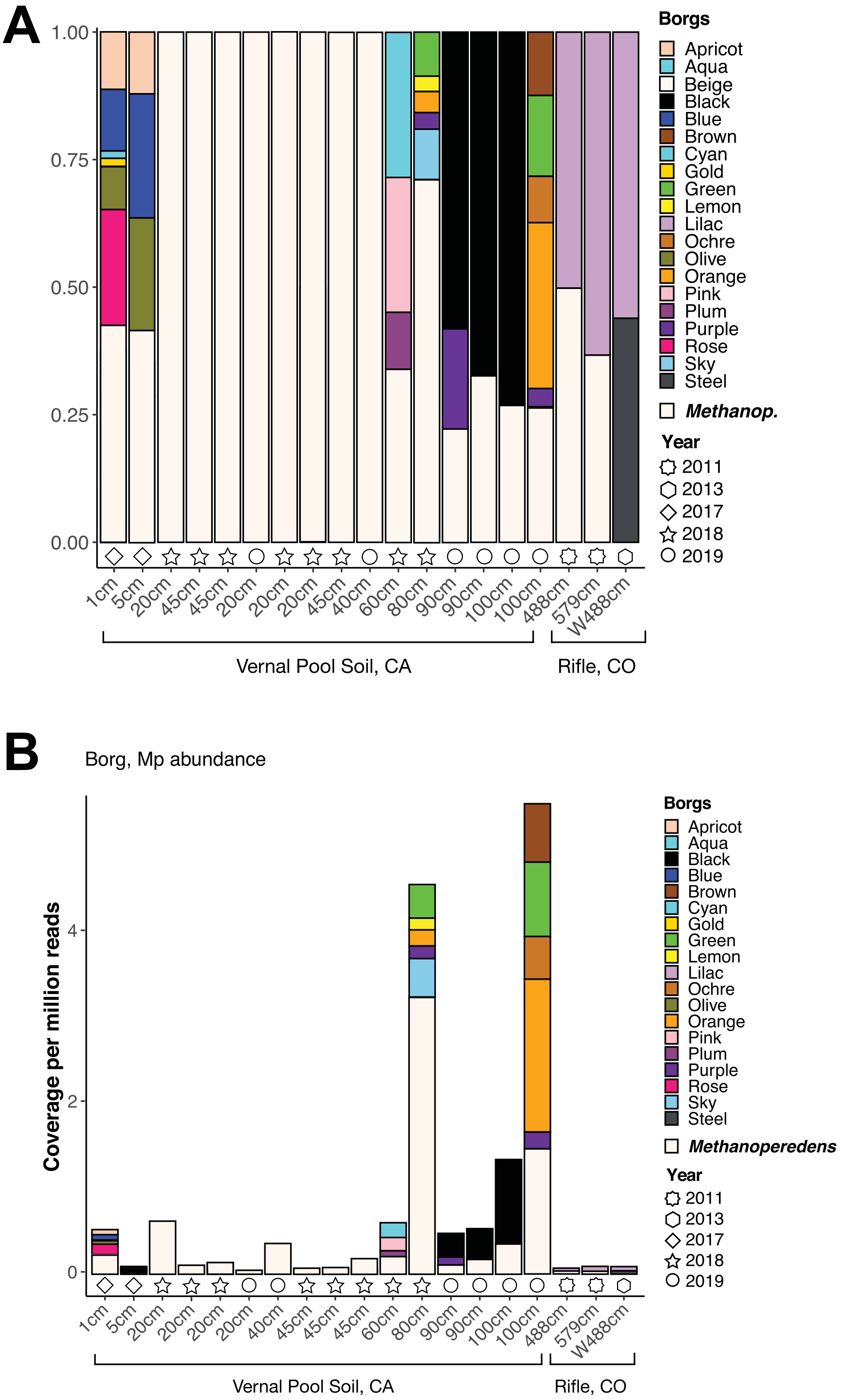


Extended Data Fig. 5 | See next page for caption.

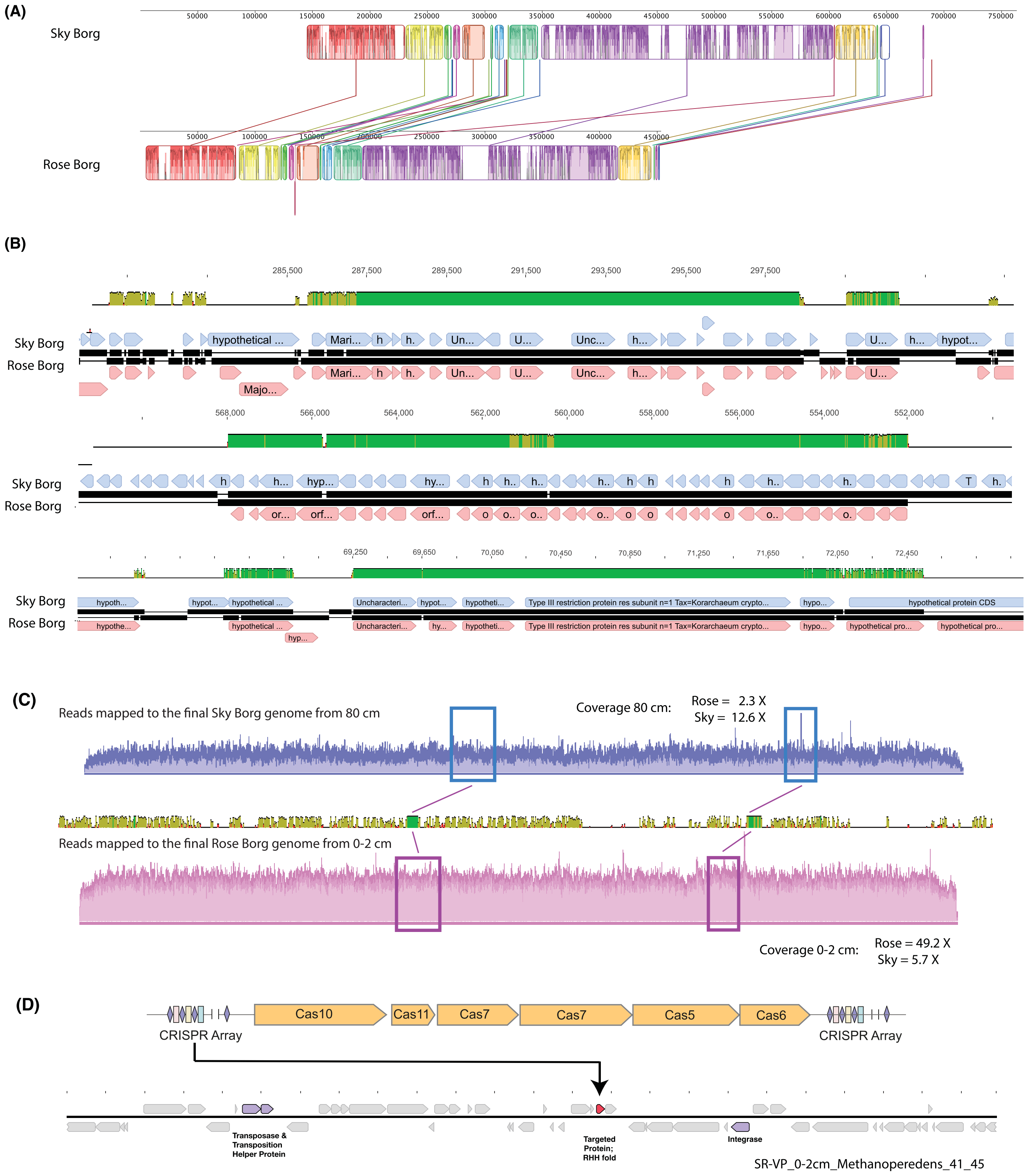
Extended Data Fig. 5 | The clustering based on protein family content demonstrates that the *Methanoperedens*, Borgs, archaeal viruses and plasmids/minichromosomes are distinct from each other. (**A**) Colored blocks indicate presence of each protein family in the corresponding genome. The blue highlight at the top indicates the *Methanoperedens* spp. (top) and Borg (bottom) protein family profiles. For details see Fig. 2d. We note that archaeal plasmids are highly undersampled. If Borgs are ultimately classified as plasmids, they dramatically expand the known characteristics (e.g., size, linear genomes) and diversity of archaeal plasmids. (**B**) Borg protein inventories (purple highlight) compared to giant linear bacterial plasmids. (**C**) Protein families occurring in more than 5 genomes of Borgs and giant linear bacterial plasmids. Few protein families are shared between Borgs and linear plasmids in bacteria beyond methyltransferases, histidine kinases, and other enzymes unrelated to replication. (**D**) Average Nucleotide Identity of different *Methanoperedens* species that coexist with Borgs (red) and previously reported genomes (gray) and the 95% species threshold shown with a dashed line.



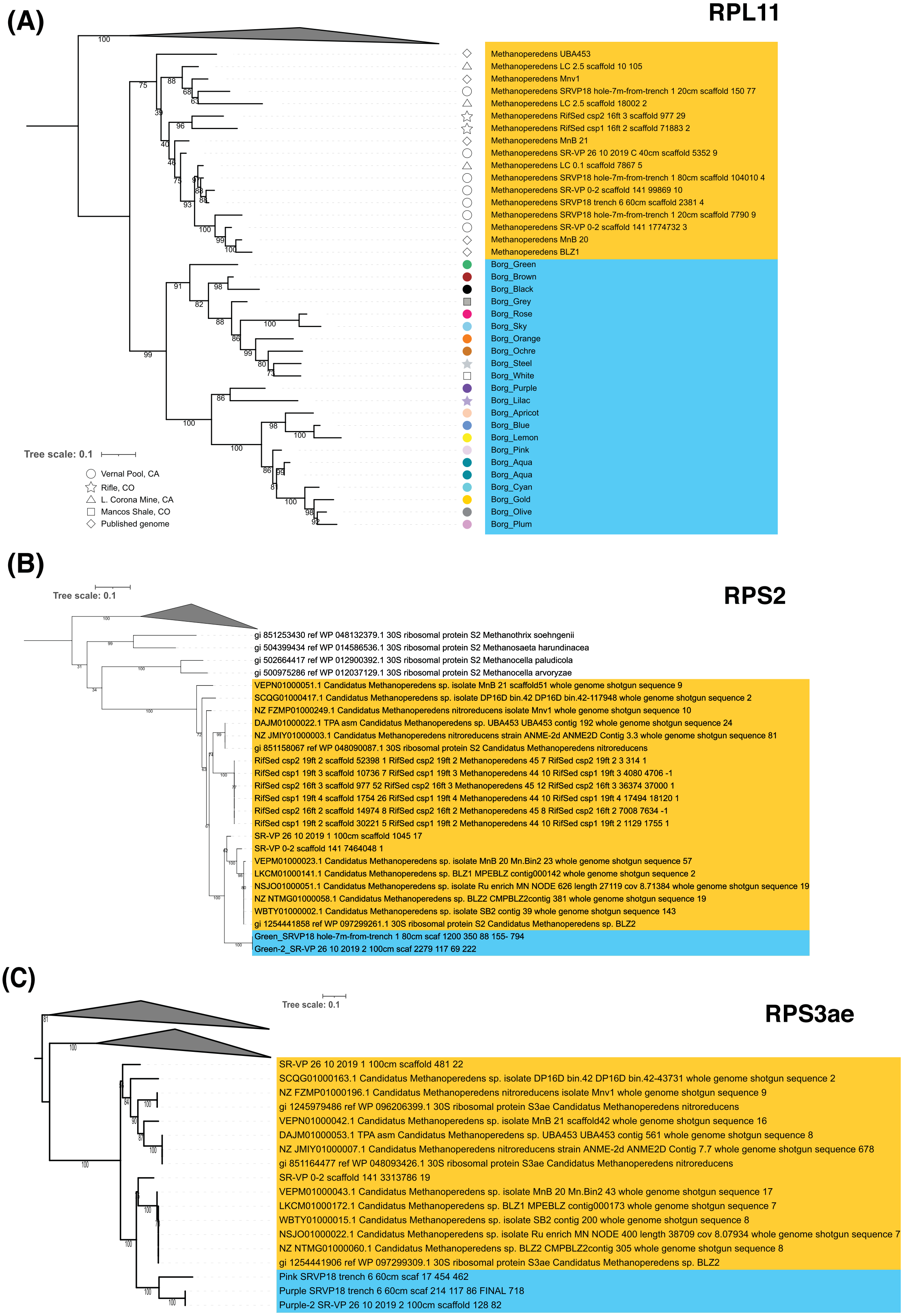
Extended Data Fig. 6 | Ribosomal protein analyses and phylogenies. (**A**) The array of single-copy archaeal ribosomal genes (columns) vs. Borg (blue) and *Methanoperedens* spp. (gold) genomes illustrating that although Borgs often have rpL11 and occasionally, other ribosomal proteins, they do not have the gene inventory needed to construct ribosomes. (**B**) **Left**; Dendrogram of hierarchical clustering of all-vs-all Pearson correlation values between all Borgs and *Methanoperedens* spp. from the wetland. **Right**; Maximum Likelihood Phylogeny of concatenated ribosomal proteins from *Methanoperedens* species that do and do not coexist with Borgs and previously reported genomes. We found no data indicating the presence of Borgs in samples containing previously reported *Methanoperedens* genomes. We searched for Borgs in the samples highlighted in blue using the same methods used to detect Borgs in this study and concluded that they do not contain Borgs. A subset of the Borg-free samples contain *Methanoperedens* spp. at very high abundance levels.



Extended Data Fig. 7 | Abundance and distribution of Borgs and *Methanoperedens spp.* in the wetland soil and Rifle aquifer. A. Relative abundances of *Methanoperedens spp.* and Borgs in samples collected over time and arrayed by sample collection depth from the wetland soils, sediments and groundwater. The absolute abundances of Borgs are far greater in the deeper compared to shallower soils B. Although some Borgs can substantially exceed all the combined abundance of *Methanoperedens spp.*, no Borgs were detected in some *Methanoperedens*-bearing samples. "W" indicates that the sample was pumped groundwater.

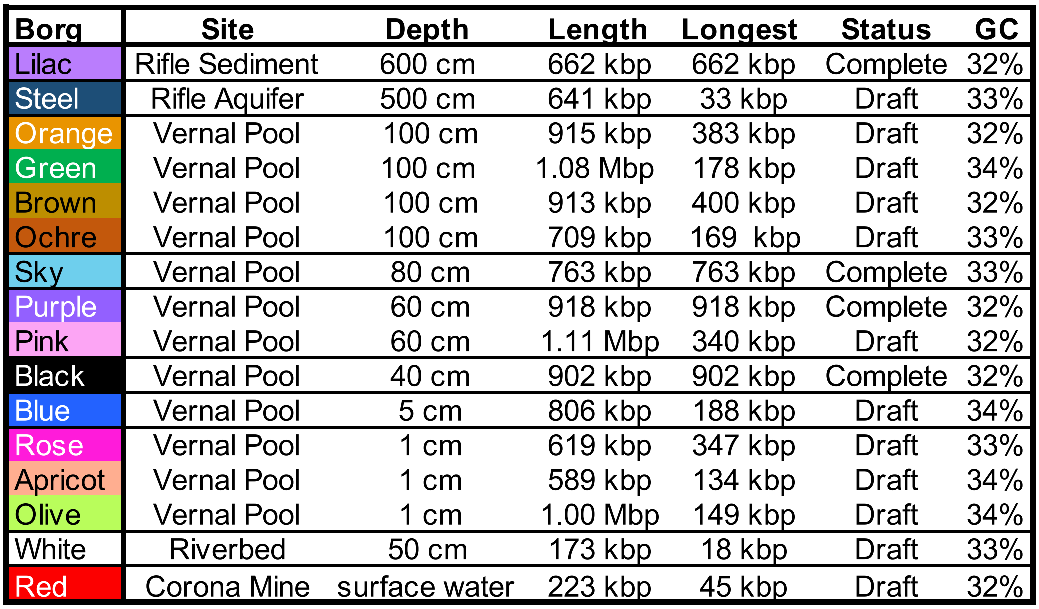


Extended Data Fig. 8 | Genome comparisons and CRISPR-Cas interactions. (A) Genome-to-genome comparisons provide evidence for recombination between two of the mostly closely related Borgs, Sky and Rose. These Borgs share only moderate overall genomic nucleic acid identity although, as is the case for other Borgs (Fig. 1a), have blocks of partially alignable sequence throughout their genomes. Notable, and indicating recent homologous recombination, are 100% identical regions of up to ~11 kbp in length (B). Although not fully manually curated to completion, the relevant Rose Borg genome regions were carefully checked by inspection of the mapped reads to rule out chimeric assembly that could otherwise explain perfect identity with the Sky Borg sequence (Sky is one of the four curated complete genomes). (C) Read coverages over the Rose and Sky genomes are consistent throughout, with the regions in B noted with green boxes. (D) Diagram illustrating the organization of the Type III-A CRISPR-Cas system variant (lacking acquisition machinery and Csm6) in the Orange Borg. One spacer from the CRISPR array targets a small protein with a ribbon-helix-helix motif, a common transcriptional regulator in archaeal mobile elements, in a mobile region of a *Methanoperedens* genome bin from the same wetland site.


(A)
RPL11
Methanoperedens UBA453
Methanoperedens LC 2.5 scaffold 10 105
Methanoperedens Mnv1
Methanoperedens SRVP18 hole-7m-from-trench 1 20cm scaffold 150 77
Methanoperedens LC 2.5 scaffold 18002 2
Methanoperedens RifSed csp2 16ft 3 scaffold 977 29
Methanoperedens RifSed csp1 16ft 2 scaffold 71883 2
Methanoperedens MnB 21
Methanoperedens SR-VP 26 10 2019 C 40cm scaffold 5352 9
Methanoperedens LC 0.1 scaffold 7867 5
Methanoperedens SRVP18 hole-7m-from-trench 1 80cm scaffold 104010 4
Methanoperedens SR-VP 0-2 scaffold 141 99869 10
Methanoperedens SRVP18 trench 6 60cm scaffold 2381 4
Methanoperedens SRVP18 hole-7m-from-trench 1 20cm scaffold 7790 9
Methanoperedens SR-VP 0-2 scaffold 141 1774732 3
Methanoperedens MnB 20
Methanoperedens BLZ1
Borg_Green
Borg_Brown
Borg_Black
Borg_Grey
Borg_Rose
Borg_Sky
Borg_Orange
Borg_Ochre
Borg_Steel
Borg_White
Borg_Purple
Borg_Lilac
Borg_Apricot
Borg_Blue
Borg_Lemon
Borg_Pink
Borg_Aqua
Borg_Aqua
Borg_Cyan
Borg_Gold
Borg_Olive
Borg_Plum
Tree scale: 0.1
○ Vernal Pool, CA
☆ Rifle, CO
△ L. Corona Mine, CA
□ Mancos Shale, CO
◇ Published genome
(B)
Tree scale: 0.1
RPS2
gi 851253430 ref WP 048132379.1 30S ribosomal protein S2 Methanothrix soehngenii
gi 504399434 ref WP 014586536.1 30S ribosomal protein S2 Methanosaeta harundinacea
gi 502664417 ref WP 012900392.1 30S ribosomal protein S2 Methanocella paludicola
gi 500975286 ref WP 012037129.1 30S ribosomal protein S2 Methanocella arvoryzae
VEPN01000051.1 Candidatus Methanoperedens sp. isolate MnB 21 scaffold51 whole genome shotgun sequence 9
SCQG01000417.1 Candidatus Methanoperedens sp. isolate DP16D bin.42 DP16D bin.42-117948 whole genome shotgun sequence 2
NZ FZMP01000249.1 Candidatus Methanoperedens nitroreducens isolate Mnv1 whole genome shotgun sequence 10
DAJM01000022.1 TPA asm Candidatus Methanoperedens sp. UBA453 UBA453 contig 192 whole genome shotgun sequence 24
NZ JMIY01000003.1 Candidatus Methanoperedens nitroreducens strain ANME-2d ANME2D Contig 3.3 whole genome shotgun sequence 81
gi 851158067 ref WP 048090087.1 30S ribosomal protein S2 Candidatus Methanoperedens nitroreducens
RifSed csp2 19ft 2 scaffold 52398 1 RifSed csp2 19ft 2 Methanoperedens 45 7 RifSed csp2 19ft 2 3 314 1
RifSed csp1 19ft 3 scaffold 10736 7 RifSed csp1 19ft 3 Methanoperedens 44 10 RifSed csp1 19ft 3 4080 4706 -1
RifSed csp2 16ft 3 scaffold 977 52 RifSed csp2 16ft 3 Methanoperedens 45 12 RifSed csp2 16ft 3 36374 37000 1
RifSed csp1 19ft 4 scaffold 1754 26 RifSed csp1 19ft 4 Methanoperedens 44 10 RifSed csp1 19ft 4 17494 18120 1
RifSed csp2 16ft 2 scaffold 14974 8 RifSed csp2 16ft 2 Methanoperedens 45 8 RifSed csp2 16ft 2 7008 7634 -1
RifSed csp1 19ft 2 scaffold 30221 5 RifSed csp1 19ft 2 Methanoperedens 44 10 RifSed csp1 19ft 2 1129 1755 1
SR-VP 26 10 2019 1 100cm scaffold 1045 17
SR-VP 0-2 scaffold 141 7464048 1
VEPM01000023.1 Candidatus Methanoperedens sp. isolate MnB 20 Mn.Bin2 23 whole genome shotgun sequence 57
LKCM01000141.1 Candidatus Methanoperedens sp. BLZ1 MPEBLZ contig000142 whole genome shotgun sequence 2
NSJO01000051.1 Candidatus Methanoperedens sp. isolate Ru enrich MN NODE 626 length 27119 cov 8.71384 whole genome shotgun sequence 19
NZ NTMG01000058.1 Candidatus Methanoperedens sp. BLZ2 CMPBLZ2contig 381 whole genome shotgun sequence 19
WBTY01000002.1 Candidatus Methanoperedens sp. isolate SB2 contig 39 whole genome shotgun sequence 143
gi 1254441858 ref WP 097299261.1 30S ribosomal protein S2 Candidatus Methanoperedens sp. BLZ2
Green_SRVP18 hole-7m-from-trench 1 80cm scaf 1200 350 88 155- 794
Green-2_SR-VP 26 10 2019 2 100cm scaf 2279 117 69 222
(C)
Tree scale: 0.1
RPS3ae
SR-VP 26 10 2019 1 100cm scaffold 481 22
SCQG01000163.1 Candidatus Methanoperedens sp. isolate DP16D bin.42 DP16D bin.42-43731 whole genome shotgun sequence 2
NZ FZMP01000196.1 Candidatus Methanoperedens nitroreducens isolate Mnv1 whole genome shotgun sequence 9
gi 1245979486 ref WP 096206399.1 30S ribosomal protein S3ae Candidatus Methanoperedens nitroreducens
VEPN01000042.1 Candidatus Methanoperedens sp. isolate MnB 21 scaffold42 whole genome shotgun sequence 16
DAJM01000053.1 TPA asm Candidatus Methanoperedens sp. UBA453 UBA453 contig 561 whole genome shotgun sequence 8
NZ JMIY01000007.1 Candidatus Methanoperedens nitroreducens strain ANME-2d ANME2D Contig 7.7 whole genome shotgun sequence 678
gi 851164477 ref WP 048093426.1 30S ribosomal protein S3ae Candidatus Methanoperedens nitroreducens
SR-VP 0-2 scaffold 141 3313786 19
VEPM01000043.1 Candidatus Methanoperedens sp. isolate MnB 20 Mn.Bin2 43 whole genome shotgun sequence 17
LKCM01000172.1 Candidatus Methanoperedens sp. BLZ1 MPEBLZ contig000173 whole genome shotgun sequence 7
WBTY01000015.1 Candidatus Methanoperedens sp. isolate SB2 contig 200 whole genome shotgun sequence 8
NSJO01000022.1 Candidatus Methanoperedens sp. isolate Ru enrich MN NODE 400 length 38709 cov 8.07934 whole genome shotgun sequence 7
NZ NTMG01000060.1 Candidatus Methanoperedens sp. BLZ2 CMPBLZ2contig 305 whole genome shotgun sequence 8
gi 1254441906 ref WP 097299309.1 30S ribosomal protein S3ae Candidatus Methanoperedens sp. BLZ2
Pink SRVP18 trench 6 60cm scaf 17 454 462
Purple SRVP18 trench 6 60cm scaf 214 117 86 FINAL 718
Purple-2 SR-VP 26 10 2019 2 100cm scaffold 128 82

Extended Data Fig. 9 | The Borg ribosomal sequences form monophyletic groups that cluster adjacent to those from *Methanoperedens* spp. Phylogenetic tree constructed using the protein sequences for (A) ribosomal protein L11 (rpL11), (B) Ribosomal protein S2 (C) Ribosomal protein 3ae.

Extended Data Table 1 | Manually curated complete and draft genomes for the best sampled Borgs



Borg	Site	Depth	Length	Longest	Status	GC
Lilac	Rifle Sediment	600 cm	662 kbp	662 kbp	Complete	32%
Steel	Rifle Aquifer	500 cm	641 kbp	33 kbp	Draft	33%
Orange	Vernal Pool	100 cm	915 kbp	383 kbp	Draft	32%
Green	Vernal Pool	100 cm	1.08 Mbp	178 kbp	Draft	34%
Brown	Vernal Pool	100 cm	913 kbp	400 kbp	Draft	32%
Ochre	Vernal Pool	100 cm	709 kbp	169 kbp	Draft	33%
Sky	Vernal Pool	80 cm	763 kbp	763 kbp	Complete	33%
Purple	Vernal Pool	60 cm	918 kbp	918 kbp	Complete	32%
Pink	Vernal Pool	60 cm	1.11 Mbp	340 kbp	Draft	32%
Black	Vernal Pool	40 cm	902 kbp	902 kbp	Complete	32%
Blue	Vernal Pool	5 cm	806 kbp	188 kbp	Draft	34%
Rose	Vernal Pool	1 cm	619 kbp	347 kbp	Draft	33%
Apricot	Vernal Pool	1 cm	589 kbp	134 kbp	Draft	34%
Olive	Vernal Pool	1 cm	1.00 Mbp	149 kbp	Draft	34%
White	Riverbed	50 cm	173 kbp	18 kbp	Draft	33%
Red	Corona Mine	surface water	223 kbp	45 kbp	Draft	32%

Length is the genome length. Longest is the size of the largest genome fragment. Status indicates degree of genome completeness: complete genomes have been corrected and fully verified throughout. GC is the genome-wide average GC content. For details for these and less abundant examples, see Supplementary Table 1.

nature portfolio

Corresponding author(s): Jillian Banfield

Last updated by author(s): Mar 3, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ ☐ A description of all covariates tested

☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Geneious v9.1.8 (Licensed, paid version used in this study, free versions available) BBmap v37.5 IDBA_UD v1.1.1 Bowtie2 v2.3.4.1 MEGAHIT v1.1.3
Data analysis	Prodigal v2.6.3 tRNAscan-SE v2.0 MMseqs2 Version: 9f493f538d28b1412a2d124614e9d6ee27a55f45 HHsuite v3.0.3 SignalP v4.1 DAMA v1.0 PSORT v3.0 TMHMM v2.0 MAFFT v7.407 RAxML v8.0.26 IQTREE v1.6.6 HMMER v3.1b2 MinCED v0.2.0 CD-HIT v4.8.1 iTOL v5 Mauve v2.4.0 InterProScan

iRep
CRISPRDetect v2.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All genomes are provided in the data availability statement and sequencing reads can be accessed via NCBI accessions PRJNA728365, PRJNA268031, and PRJNA441604. Sequence databases used include UniProt, ggKbase, PFAM r32, KEGG, pVOG.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was chosen to provide a high breadth of ecosystem coverage for the recovery of Borg genomes. Accordingly, there were no statistical methods to determine sample sizes. Data were compiled from multiple sources where Borg genomes could be found. Samples from each sampling site is listed in Table 1 and Table S1
Data exclusions	No databases or datasets were excluded during our survey.
Replication	Borg genomes were recovered from 36 independent soil samples collected in 2017, 2018, 2019, and 2020 from the same wetland field site in Lake County, CA. Up to 10 samples were collected from the same soil depth interval. DNA extracted from all samples was sequenced separately and the sequence data assembled and the datasets analyzed individually. Genomes were recovered from single samples, but in some cases the same genome was sampled independently in more than 1 sample. Sequencing reads from all samples were mapped to the most complete version of each genome to test for the presence of each genotype in each sample and to quantify the pattern of abundance of each genotype across samples. The only use of statistics was for correlation analysis that used the pattern of abundance data. The correlation analysis used a two-sided Pearson correlation test to generate a correlation metric. Borg genomes were also recovered from samples collected in 2011 and 2013 from an aquifer in Rifle, Colorado. Sediment samples were taken from cores from depths of 5 and 6 m below the surface. Four replicates were collected at each depth and the genomic datasets from them were processed individually. The same Borg genotype was recovered in the 5 and 6 m depth samples and from co-assemblies. The other sites from where Borg genomes were sampled (groundwater from the Rifle site and from below the riverbed at the East River, Crested Butte, Colorado) were sampled only once. The metagenomic datasets from these samples were assembled and analyzed independently. Host identification was verified by a combination of CRISPR targeting (sequence identity between the CRISPR spacer and Borg genome), phylogenetic analysis of ribosomal proteins, and phylum-level taxonomic profiles. Annotations were verified across multiple databases.
Randomization	After samples were collected from the natural environment they were homogenized to reduce the effects of small-scale heterogeneity, and DNA was extracted from the homogenized material. Other forms of randomization are not applicable to this study because the research relied on DNA sequences that were used to reconstruct genomes and not laboratory experiments.
Blinding	Blinding was not performed because it was not applicable to this study. We analyzed samples collected from the environment and thus the conclusions are not dependent on trial outcomes.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging