Title: Nanoarchaeota, its Sulfolobales host, and Nanoarchaeota virus distribution across Yellowstone National Park Hot Springs

Running Title: Nanoarchaea and *Sulfolobales* host distribution in YNP

Jacob H. Munson-McGee[1], Erin K. Field[2*], Mary Bateson[3]*, Colleen Rooney[4], Ramunas Stepanauskas[2], Mark J. Young[3,5]#.

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, Montana, 59717 USA, [2]Bigelow Laboratory for Ocean Sciences, East Boothbay ME 04544, [3]Thermal Biology Institute, Montana State University, Bozeman, Montana, 59717 USA, [4]Department of Chemistry and Biochemistry, Montana State University, Bozeman, Montana, 59717 USA, and [5]Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, Montana, 59717 USA

Correspondence:
Mark J. Young
Department of Plant Sciences and Plant Pathology
155 CBB
Montana State University
Bozeman MT, U.S.A
Phone: 1-406-994-5158
Email: myoung@montana.edu

24 *Present address: Mary Bateson, Bioscience Laboratory, Inc., Bozeman, MT, 59718 USA.

25 *Present address: Erin Field, University of Delaware, Newark, DE 19716.

Abstract

Nanoarchaeota are obligate symbionts with reduced genomes first described from marine thermal vent environments. Here, both community metagenomics and single-cell analysis revealed the presence of Nanoarchaeota in high temperature (~90˚C), acidic (pH~ 2.5-3.0) hot springs in Yellowstone National Park USA (YNP). Single-cell genome analysis of two cells resulted in two near identical genomes, with an estimated full-length of 650 kbp. Genome comparison showed that these two cells are more closely related to the recently proposed *Nanobsidianus stetteri* from a more neutral YNP hot spring than to the marine *Nanoarchaeum equitans*. Single-cell and CARD-FISH analysis of environmental hot spring samples identified the host of the YNP Nanoarchaeota as a Sulfolobales species known to inhabit the hot springs. Furthermore, we demonstrate that Nanoarchaeota are widespread in acidic to near neutral hot springs in YNP. An integrated viral sequence was also found within one Nanoarchaeota single-cell genome and further analysis of the purified viral fraction from environmental samples indicates that this is likely a virus replicating within the YNP Nanoarchaeota.

Key Words

Nanoarchaeota, CARD-FISH, Single-Cell Genomics, *Nanoarchaeum equitans*, *Nanobsidianus stetteri*, Sulfolobales, Yellowstone National Park, Archaeal Virus

Introduction

The Nanoarchaeota was first described in 2002 (1) as an obligate symbiont/parasite with a highly reduced genome lacking most biosynthetic capacity and found in association with its host, the marine hyperthermophilic crenarchaeon *Ignicoccus hospitalis*. The Nanoarchaeota provides the unique opportunity to study genome streamlining, as well as the transfer of

metabolites between Archaeal cells. In their original description based on 16S ribosomal sequence *Nanoarchaeum equitans* were classified as a deep-rooted branch of the Archaea forming a new phylum. Although, initially accepted, this original classification has been challenged and is still being debated. In part, this debate is based on phylogenetic reconstruction using concatenated protein sequences and analysis of the distribution of gene families among the major archaeal lineages, which suggests a close evolutionary relationship between *N. equitans* and the Thermococcales, a basal order of the Euryarchaeota phylum (2,3).

In the thirteen years since its discovery, there is evidence that Nanoarchaeota inhabit a diversity of environments beyond marine thermal vents. Using primers designed from the 16S rRNA gene of *N. equitans*, genetic evidence of Nanoarchaeota has been found to be widespread in terrestrial hot springs (4-5) and mesophilic hypersaline environments (5). Ribosomal sequences have also been detected from photic-zone water samples far removed from hydrothermal vents (6) suggesting that Nanoarchaeota are a widespread and diverse group of Archaea capable of inhabiting a broad spectrum of temperatures and geochemical environments. However, despite ribosomal evidence of Nanoarchaeota being widespread, the original *N. equitans* isolated with its host *I. hospitalis* remains the only cultured representative of this phylum. No *Ignicoccus* species have been identified in terrestrial hydrothermal samples, suggesting that terrestrial Nanoarchaeota use a different host and that Nanoarchaeota may associate with a wide diversity of host organisms.

The report of a second sequenced Nanoarchaeota, originally termed Nst1 and recently renamed as *Nanobsidianus stetteri*, from Yellowstone National Park (YNP) and provides more insight into Nanoarchaeota diversity (7). Single-cell analysis of *N. stetteri* resulted in a partial 593 kbp genome with an estimated complete genome size of 651 kbp. Analysis of the *N. stetteri*

Applied and Environmental Microbiology

genome provided additional insights into the metabolic capacity and phylogenetic position of the Nanoarchaeota. In contrast to *N. equitans*, the full-length genome of *N. stetteri* is estimated to be more than 20% larger (~651 versus 491 kbp), encodes a complete gluconeogenesis pathway, contains no CRISPR system, has components of RNase P, encodes for a smaller repertoire of split protein coding genes, contains an euryarchaeal type flagellum (archaellum), and has an inferred *Acidicryptum nanophilum* host.

We have investigated the presence of Nanoarchaeota species in high temperature acidic hot springs in YNP, and their relatedness to *N. stetteri* and *N. equitans.* We found *N. stetteri* is present in multiple YNP acidic hot springs, that *Nanobsidianus* in YNP are found in association with a *Sulfolobales* host, and that a virus is likely replicating within YNP *Nanobsidianus* species.

Materials and Methods

Sample collection

Two hot springs were initially selected for this survey (Fig. S1). The location, sampling date, pH, and temperature for each sample time are listed in Table S1. Hot spring water was collected from the Alice Springs, Crater Hills (CH09), and Nymph Lake 01 (NL01) sampling sites (Table S1). Upon returning to the lab, cells were collected by filtration of samples through 0.4 um filters (Isopore HTTP14250). Cells for CARD-FISH analysis were gently washed from filters and fixed in 1% paraformaldehyde for one hour at room temperature before being washed three times in phosphate buffered saline (PBS). After fixation, cells were stored in 50% EtOH/50% PBS at -20°C until needed for further processing.

Single cell genomics

One mL samples from NL01, collected on October 11, 2011 and September 20, 2012, were preserved with 5% glycerol and 1x TE buffer (final concentrations), immediately frozen in

Applied and Environmental Microbiology

liquid nitrogen and then stored at -80°C. The flow-cytometric separation of individual cells, cell lysis, whole genome amplification and the sequencing of 16S rRNA genes were performed at the Bigelow Laboratory Single Cell Genomics Center (scgc.bigelow.org), using previously described methods (8,9). Partial 16S rRNA sequences were determined for 99 single cells using universal Bacterial and Archaeal primers (Table S2). Based on the 16S rRNA results, 13 single cells representing the major archaeal species were selected for whole genome sequencing. Genomic library preparation and sequencing were performed at the Oregon State University's Center for Genome Research (cgrb.oregonstate.edu). SAG genomic DNA was sheared using S220 focused ultrasonicator (Covaris, Woburn, MA) and gel-fractionated for 450 bp fragments. Illumina sequencing libraries were prepared using TruSeq reagents and protocols (Illumina, San Diego, CA). 150x2-bp paired-end (PE) reads were sequenced using the Illumina HiSeq 2000 platform (Illumina). Twelve indexed SAG libraries were multiplexed, in equal proportions, into one lane of a flowcell. The obtained reads were quality-trimmed, digitally normalized and assembled in SPAdes v.2.2.1 Bankevich et al. 2012 (10), as described in Wilkins et al. (11). Potential contaminant contigs were removed after tetramer frequency evaluation and PCA analysis identified outliers (12) as well as BLAST comparisons to the NCBI nr database and between samples sequenced together as described in further detail in Field et al. 2015 (13). Genome size estimates were calculated using both arCOG (14) and CheckM (15) analysis.

Hot spring 16S rRNA phylogenetic analysis

Fifty-eight ~510 bp single cell 16S rRNA sequences from NL01 October 2011 and thirty-eight single cell 16S rRNA sequences from NL01 September 2012 were combined with eighteen 16S rRNA sequences from NL01 metagenomes (16,17) and twenty-one sequenced reference

Applied and Environmental Microbiology

genomes resulting in a total of 135 sequences. These sequences were aligned using MUSCLE (18) and a Bayesian analysis was performed using MrBayes (version 3.2.5) (19), with mixed nucleotide substation models and gamma shaped paramater. Posterior probability values were derived from ten million permutations while sampeling every 100,000 generations while using the default 25% burnin. The16S rRNA sequence from the bacterial species *Acidithiobacillus caldus* ATCC 51756 served as the outgroup.

Catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH) probe design

A total of eight 16S rRNA probes were designed, each of which detects one of the eight major 16S rRNA phylotypes in the NL01 hot spring (Table S2). An additional 16S rRNA probe was used to detect most Crenarchaeaota (20). Probes were synthesized by Biomers (Ulm, Germany) with a horseradish peroxidase (HRP) incorporated at their 5' end. Specificity of probes were validated by testing them in lab against near full length 16S rRNA clones generted for each major 16S rRNA present in NL10. Probes were confirmed to hybridize to only their inteneded target and not to the other 16S rRNA types presnet in NL10.

CARD-FISH

A modified CARD-FISH analysis (21-22) was used to probe Archaea-dominated acidic hot spring environmental samples. Fixed samples were placed in the wells of glass slides (PL-2026 Precision Lab Products) and air dried for 10 minutes at 46°C. Samples were subsequently dehydrated in 50%, 80%, and 100% ethanol for 3 minutes and dried at 46°C for five minutes. Wells were covered with permeabilization solution (50 mM glucose, 20 mM Tris pH 7.5, 10 mM EDTA, and 0. 2% Tween20) and placed on ice for one hour before being rinsed in 1x PBS and

air-dried at 46°C. Endogenous peroxidases were deactivated by a 10 minute incubation in 0.1 N HCl at room temperature, rinsed with 1x PBS and air-dried at 46°C. 16S rRNA probes were added to the hybridization buffer to a final concentration of 0.2 ng/ul with a varying amount of formamide (Sulfolobales 20%, *Nanobsidianus* 40%) to increase stringency of hybridization. Samples were allowed to hybridize at 46°C in Petri plates sealed with parafilm for three hours before being washed in washing buffer (22) at 48°C for 20 min. This was followed by 15 min of rinsing in 1x PBS at 37°C before air-drying at 46°C. All samples were overlaid with solution containing 1xPBS, 10% dextran sulfate, 0.1% blocking reagent (Roche, Germany cat. no. 11096176001), 2M NaCl, 0.0015% H_2O_2 and 0.33 ug/ul $Alexa_{488}$ or $Alexa_{594}$-labeled tyramides and allowed to incubate at 37°C for 30 min. Washes consisted of 5 min with 1x PBS at 46°C and 1 min with water, followed by air drying. For dual labeling, the protocol was repeated starting at the deactivation step using different 16S ribosomal probes and tyramide fluorophore. Following CARD amplification, slides were washed in 1x PBS and stained with DAPI 10ng/µl and fixed with VectaShield. Controls of equivalent *E. coli* rRNA probes (pB-02228), competiton with non-fluorphore labeled probes, and nuclease treatments were performed.

Samples were imaged on a Leica TBS SP8 confocal microscope fitted with a 63x oil immersion lens and images were collected sequentially. NIH ImageJ64 software was used to process the images.

Geographic distribution of *Nanobsidianus* within Yellowstone National Park

Seven conserved genes (Table S3) encoded within two single-cell genomes described here, AB_777_F03 (F03) and AB_777_O03 (O03), another single cell genome previously described, *N. stetteri*, *N. equitans*, and CH09 metagenomes (17), were concatenated and aligned using MUSCLE (18), and subjected to Bayesian analysis with MrBayes (19) as described above.

In addition, we recruited contigs from 23 publically available metagenomes (17) using the O03 contigs. A minimum overlap of 50 bp and 93% nucleic acid identity was used to recruit contigs onto the O03 genome.

***Nanobsidianus* virus identification**

The single-cell genomes were analyzed with the VirSorter app (https://de.iplantcollaborative.org/de/) to identify viral sequences in the single-cell genomes. After manual curation of the sequences, identified genes were subjected to BLASTx analysis to search for matches to viral proteins with a minimum 10e-5 score. Identified regions were additionally subjected to blast analysis against viral network clusters described in Bolduc et al. 2015 (10.1038/ismej.2015.28).

To gain further evidence for the presence of a *Nanobsidianus* associated virus, viral fractions were prepared by filtering NL01 hot spring water through a 0.45 um filter to remove cells. Viruses in the filtered water were concentrated by $FeCl_3$ flocculation (23). Virus particles were separated by buoyant cesium chloride density centrifugation gradients spanning 1.13-1.38 g/ml (116,000g for 20 hours). The total gradient was hand fractionated and DNA was extracted from individual fractions using the Invitrogen PureLink® Viral RNA/DNA Mini kit (Waltham, Massachusetts). Fractions were screened for the presence of the viral genome by PCR analysis using viral specific genome primers. PCR primers specific to *Nanobsidianus* and Sulfolobales were included as a control (Table S2). Total viral fractions from CH09 were also prepared by filtration to remove cells and $FeCl_3$ flocculation.

Results

Applied and Environmental Microbiology

Ninety-six 16S rRNA sequences derived from single cell sequencing and 18 16S rRNA sequences from cellular metagenomes both showed a simple archaeal microbial community present in NL01 (Fig. S2). This community is dominated by nine species based on their 16S rRNA gene. Based on metagenomic read abundance and single cell sequencing, eight archaeal species represent 97% of all cells and a single bacterial *Hydrogenobaculum* sp. comprises the remaining 3%. The eight archaeal clades are either related to *Acidianus hospitalis* (3%), and *Sulfolobus islandicus* (1%), or uncultured members of what likely constitute new crenarchaeal species, the recently proposed *Acidcryptum nanophilium* (46%), an *Acidolobus* sp. (6%), a *Vulcanisieta* sp. (5%), a *Nanobsidianus* sp. (16%) and two *Sulfolobus* sp. consisting of 11%, and 9% respectively (Fig .S2).

We selected representative single-cells from each of the eight archaeal clades for genome sequence analysis based on their 16S rRNA sequence (Table S4). We were successful in obtaining high quality sequence information for 13 single cells. The estimated range of genome completion for sequenced single cells ranged from 13-85%. The DNA sequence and assembly statistics are provided (Table S5).

Two of the thirteen single-cell genomes represented within NL01, designated O03 and F03, were classified as *Nanobsidianus* based on their 16S rRNA sequences. These two single amplified genomes (SAG) had high 16S rRNA gene homology (98.2%) to *N. stetteri* sequenced from a circumneutral hot spring within YNP and low homology to *N. equitans* (81.5%), which was isolated from a marine hydrothermal vent. A blastn analysis of four metagenomes from the same hot spring sampled over a seven-year period (min E score 1E-5) with the partial O03 16S rRNA gene revealed *Nanobsidianus* 16S rRNA gene sequences in all four metagenomes.

Applied and Environmental Microbiology

An examination of multiple YNP hot springs revealed that *Nanobsidianus* is broadly distributed. A BLASTn search of metagenomes from 23 hot springs (min E score 1E-5) with the O03 single-cell partial *Nanobsidianus* genome (described below) resulted in 7,024 contigs from 15 of the 23 hot springs (Table S6). Ten of the 15 hot springs also had *Nanobsidianus* 16S rRNA sequences (Table S7, Fig. S3). Based on read abundance, *Nanobsidianus* cells were most abundant in hot springs that had temperatures > 60˚C and pH < 4.5 (Fig. 1A). Outside of this temperature and pH range, *Nanobsidianus* sequences were rarely detected. Read recruitment analysis indicated the *Nanobsidianus* represent approximately 11% of the NL01 and 5% of the CH09 microbial community composition (Fig. 1B). However, one should be cautious in assigning quantitative community membership based on single cell genomics and metagenomic read recruitment analysis due to the inherent bias that each technique presents (24-28)

Assembly of the two *Nanobsidianus* genomes derived from two independent single cells produced two incomplete genomes of 499 kbp (F03) and 549 kbp (O03). However, upon further analysis, 50 kbp of the F03 genome was revealed to be from a Sulfolobales species (described below), resulting in a genome of 449 kbp, which was used in further analysis. The degree of genome completeness was estimated using conserved archaeal genes that are represented in the assembled genomic contigs (14). In the arCOG database as of March 2015, there were 95 genes conserved across all 168 sequenced Archaeal genomes including *N. equitans*, and 325 genes in all Crenarchaeota genomes. Completeness of Nanoarchaeota genomes was calculated using the 95 genes conserved in all archaeal genomes, while the genome completeness of all other SCGs was calculated using the 325 genes that are shared by all Crenarchaeota. F03 contained 74 of 95 genes, resulting in an estimated genome size of 576 kbp with 95% confidence limits of 530-644 kbp; O03 has 76 of 95 conserved genes and an estimated genome size of 686 kbp and 95%

confidence limits of 635-762 kbp (Table S5). The discrepancies in genome length between these two cells could reflect an estimation bias resulting from the uneven distribution of the conserved genes used in the analysis, a true difference in their genome size, and or the potential presence of sequences from the *Nanobsidianus* host (discussed below). Based on the calculated genome size of *N. stetteri* (650 kbp), F03 is 67% complete and O03 is 85% complete.

The partial genome sequences of *Nanobsidianus* from NL01 show a high degree of sequence identity to each other (Fig. 2, Fig. S4 (29)). Of the 472 ORFs that were identified in both F03 and O03, 372 (79%) had greater than 95% amino acid identity. 309 of 532 (58%) ORFs identified in both O03 and *N. stetteri* had greater than 95% amino acid identity. In contrast, none of the 313 ORFs identified in both O03 and *N. equitans* had 95% or greater amino acid identity. Mauve alignment showed more conserved regions between the three YNP *Nanobsidianus* genomes as compared to *N. equitans* (Fig. 2). Construction of a neighbor joining tree of 17 concatenated conserved genes in F03, O03, *N. stetteri*, *N. equitans* and *Nanobsidianus* contigs identified in CH09 metagenomes supports the close relationship between YNP *Nanobsidianus* species as compared to the marine *N. equitans* (Fig. 1C). The terrestrial *Nanobsidianus* sequences from YNP grouped together, while *N. equitans* formed its own distinct branch. Overall, these results support the conclusion that the F03 and O03 genomes are highly homologous and more closely related to *N. stetteri* than to *N. equitans*.

Additional analysis of specific genes supports the overall conclusion that F03 and O03 are closely related to *N. stetteri*, but there are some differences. In O03 we identified Rpp29 and Rpr2, two components of RNAse P. We also identified a truncated version (238 bp) of the RNase P RNA component in both of the genomes that is 98% similar to the one described by Podar et al. (7). The presence of these genes is similar to *N. stetteri,* where three protein components were

identified, but contrary to *N. equitans,* which remains the only identified cellular life that lacks this complex (30). We also searched for the twelve-split protein coding genes previously described in *N. equitans* and *N. stetteri* (7). All but the P-loop ATPase were found in at least one of the two partial NL01 *Nanobsidianus* genomes. All YNP single cells had identical split protein coding genes. Of these split genes, six are split in the same place as *N. equitans;* four genes that are split in *N. equitans* are not split in YNP single cells, and two are split in YNP isolates but not *N. equitans* (Table S8).

We searched the NL01 *Nanobsidianus* genomes for tRNA genes using tRNA scan (31). Between the two NL01 single cell genomes, all twenty standard amino acids have at least one tRNA represented. We did not find any non-standard tRNAs in either of these genomes. Within the two genomes, we found 36 total tRNA sequences (Table S9). In contrast, *N. stetteri* is missing a tRNA for Phe, most likely due to the incompleteness of the genome (7,32). Both F03 and O03 have four intron-containing tRNA genes, in contrast to *N. stetteri* which only has two intron containing tRNAs (Table S9). This discrepancy could again be explained by the incompleteness of the *N. stetteri* genome. Of the four intron-containing tRNA genes in F03 and O03, two, Ile (TAT) and Tyr (GTA) are also found in *N. stetteri* and *N. equitans.* One, Met (CAT) is not split in *N. stetteri* and has a longer intron (27 versus 65 bp) in *N. equitans* and Leu (TAA), is not split in *N. equitans* and not detected in *N. stetteri*. Additionally, like *N. equitans* and *N. stetteri*, F03 and O03 lacked central metabolic genes including nearly all the genes for lipid, cofactor, amino acid, and nucleotide synthesis indicating a symbiotic life style where the *Nanobsidianus* cells are incapable of replicating without their host (33). All the archaeal flagellum genes encoded in *N. stetteri* described by Podar et al. (7) are encoded in at least one of the NL01 *Nanobsidianus* genomes except for a FlaH homolog (the putative ATPase). The NL01

Applied and Environmental Microbiology

Nanobsidianus genomes also lacked any genes associated with the CRISPR/cas system like *N. stetteri* but in contrast to *N. equitans* in which crRNAs appear to be constituently expressed (34).

Analysis of the assembled F03 sequences revealed the presence of a Sulfolobales organism that may serve as a host for the nanoarchaeal cell. Five of the 25 assembled contigs from F03 (representing 50 kbp of DNA sequence or 9.9% of the total assembled sequences) did not map to any known Nanoarchaeota genomes. These five contigs also have a significant G/C skew (46% G/C) as compared to the 20 contigs that have homology to the O03 and *N. stetteri* genomes (24%). A Blast analysis of these five contigs against the eleven other SAGs reveals that each contig has the highest match to SAGs AB_777_J03, AB_777_K09, or AB_777_K20, here referred to as J03, K09, and K20 respectively, all members of the dominant Sulfolobales group, identified as *A. nanophilum* by Podar et al. (7). In addition, one of these contigs contained a CRISPR locus with a conserved leader and direct repeat sequences frequently found in *Sulfolobus* species and matched those found in the single-cell genomes of J03, K09, and K20. These contigs are likely present as a result of an intimate association between the *Nanobsidianus* and *Acidicryptum* cells, namely that of host-symbiont. One single-cell 16S rRNA PCR amplicon that was not selected for genome sequencing had 16S rRNA signatures from two organisms. One of the 16S rRNA sequences was *A. nanophilum* and 99.1% similar to J03 while the second is *Nanobsidianus* like and 99.1% similar to O03. Additionally *A. nanophilum* 16S rRNA signatures are present in all the YNP metagenomes where *Nanobsidianus* 16S rRNA sequences are located. All these lines of evidence suggest that the *Nanobsidianus* forms a host-symbiont relationship with *A. nanophilum* present in the NL01 hot springs.

To further investigate the nature of the host, CARD-FISH analysis was performed to confirm the genomic host identification of the *Nanobsidianus* present in the NL01 and CH09 hot

Applied and Environmental Microbiology

springs (Fig. 3). Briefly outlined, fluorescently labeled hybridization probe specific to the each of the eight major cell types were generated (see Materials and Methods). The probe specific to *Nanobsidianus* rRNA was labeled with $Alexa_{594}$ fluorophore, while the rRNA probes to the other seven individual cell types in NL01 were with labeled $Alexa_{488}$. The *Nanobsidianus* probe was used in combination with each individual cellular rRNA probes in a dual hybridization assay of cells collected from NL01 and CH09. CARD FISH analysis revealed that NL01 and CH09 *Nanobsidianus* is only found in association with an *Acidicryptum* species (Fig. 3A-D, G) and not in association with the other six archaeal species present in NL01 (Fig. 3E). This is a similar host species to the one previously described from another hot spring and the same phylotype identified by single-cell genomics analysis. CARD-FISH analysis demonstrated co-localization of greater than 95% of detected *Nanobsidianus* cells with *Acidicryptum* cells in both NL01 and CH09 hot springs. Further analysis showed the same co-localization in hot spring samples collected over the course of seventeen months (Fig 3F), indicating that the *Nanobsidianus* population is a persistent member of YNP hot spring enviornments. Controls of *Nanobsidianus* cells dual labeled with the other six potential host species in the hot springs showed no co-localization, indicating the host-symbiont relationship is limited to this *Acidicryptum* sp host.

Analysis of co-localization of *Acidicryptum* host and *Nanobsidianus* cells revealed fluctuations of both over time in both hot springs surveyed. In CH09 the *Acidicryptum sp* varied from 7-23% of total microbial community composition, 53 to 75% of *Acidicryptum* cells were associated with *Nanobsidianus*. In NL01 the proportion of the *Acidicryptum* cells ranged from 17-45% of the total microbial community. Regardless of their relative abundance, ~66% of these cells were associated with *Nanobsidianus* cells (Table S10). Overall, these results indicate that the *Acidicryptum* is the host for the *Nanobsidianus* in NL01 and CH09 hot springs.

Applied and Environmental Microbiology

322

323 ***Acidicryptum* host genome analysis**

324 The three *Acidicryptum* sp host single-cell genomes, J03, K09, and K20, showed high
325 sequence identity to *A. nanophilum*, the host of *N. stetteri* proposed by Podar and colleagues (7).
326 These three partial genomes, all from of the Sulfolobales order, range from 250 kpb to 850 kbp.
327 Of the 325 conserved genes in all Crenarchaeota genomes, J03 contains 260, K09 663, and K20
328 157; for an estimated genome size and genome completeness using the arCOG database of 1.070
329 Mbp (80.0%), 1.034 Mbp (20.3%), and 1.078 Mbp (48.3%) respectively (Table S5). Genome
330 size estimates were independently analalyzed with CheckM (15), which estimated an average
331 genome size of 1.149 Mbp for the three *N. stetteri* host cells. This estimate of approximately 1.1
332 Mbp is significantly smaller than the estimated size of the *A. nanophilum* genome (1.7 Mbp)
333 from Obsidian Pool previously described (7). It is possible that these methods are
334 underestimating the host genome size due to the inherent uncertainty using partial genome
335 sequences combined with the unevenness of MDA based genome amplification methods. It is
336 also 1.1 Mbp smaller than the closely related *S. acidocaldarius* genome (2.2 Mbp). *I. hospitalis,*
337 the host of *N. equitans*, has a genome size of 1.29 Mbp and has undergone similar genome
338 streamlining (7).

339 It has been proposed that Nanoarchaeota evolve faster then non-symbiotic Archaea, as a
340 consequence of its reduced genome size and content (3). We observe this trend with
341 *Nanobsidianus* cells from the different YNP hot springs showing more divergence from each
342 other as compared to the YNP *Acidicryptum*. Of ORFs identified in both J03 and *A. nanophilum*,
343 79% have greater than 95% amino acid identity (Fig. S5), while just 58% of ORFs in O03 and *N.*
344 *stetteri* have greater than 95% amino acid identity.

345 ***Nanobsidianus* virus detection**

346 Analysis of the *Nanobsidianus* O03 single cell genome with VirSorter identified one 10
347 kbp contig as having a virally enriched region. Annotation of this contig showed 10 ORFs, 5 of
348 which were annotated as hypothetical unknown proteins. Interestingly, two other ORFs were
349 annotated, one as a hypothetical protein related to a *Sulfolobus* monocaudavirus 1 (SMV1)
350 conserved archaeal viral protein, and the other a *N. equitans* 30S ribosomal protein, suggesting
351 that this contig contains a portion of an integrated viral genome. The archaeal virus-related ORF
352 is located at the end of the 10 kbp contig and has a GC content of 29.8%, which is significantly
353 different from the rest of the contig on which it is located (22.6%) and the GC content of all O03
354 contigs (23.4%). Further BLAST analysis of this contig to a collection of viral metagenomes
355 from NL01 resulted in 104 contigs matching to the conserved archaeal viral protein.
356 Furthermore, all of these 104 contigs mapped to a single viral network cluster of the viral
357 community structure previously reported (Bolduc et al., 2015). Assembly of this network cluster
358 generated a 3.0 kbp contig with the SMV1 conserved archaeal virus-like protein at one end and a
359 ~600 bp putative structural protein from a *Sulfolobus* rudivirus near the other end (Fig. S6) as
360 well as several other ORFs with no significant similarity. The GC content of the identified
361 conserved archaeal virus gene (29.8%) is nearly identical to the GC content of all the contigs that
362 make up viral network cluster (29.9%). The presence of this ORF in the viral fraction of NL01
363 hot spring water was confirmed by PCR analysis of virus-like particles (VLPs) isolated from
364 CsCl buoyant density gradients. This genome segment was successfully PCR amplified and
365 confirmed by DNA sequencing from the purified viral fraction on cesium chloride fractions with
366 a density of 1.36g/cm^3. The same genome segment was also successfully amplified from total
367 viral fractions from CH09 suggesting that the viral distribution is potentially similar to

Nanobsidianus in YNP. In order to show that cells were not contaminating the viral fraction, PCR with primers specific to both *Nanobsidianus* as well as *Acidicryptum* was performed on the same fractions but no sequence was PCR amplified (Fig. 4). Overall, these results indicate that there is likely a virus associated with *Nanobsidianus* present in YNP hot springs.

Discussion

We report here the sequencing of two *Nanobsidianus* single-cell genomes from high temperature acidic hot springs in YNP, and the detection of *Nanobsidianus* cells from high temperature acidic hot springs across YNP. Based on their high 16S rRNA similarity, >95% (35) to *N. stetteri,* and overall genome homology we propose that these cells are *N. stetteri,* and share the same *A. nanophilum* host as previously described (7). All three YNP *Nanobsidianus* genomes are closely related to each other and quite distinct from *N. equitans*. Even though members of the *Nanobsidianus* genus are widely distributed in YNP thermal features, including in Yellowstone Lake (6), they are most abundant in high temperature acidic hot springs found in YNP.

N. stetteri that were found in the Obsidian Pool, NL01 and CH09, share a common *A. nanophilum* host. In our analysis, approximately 66% of *A. nanophilum* cells are found in association with *N. stetteri* cells. The reason why this approximate ratio in environmental samples is maintained within different hot springs is unknown, but suggests that there is some control over these cell interactions. It remains to be determined if the physical interaction between the YNP *N. stetteri* and its host is similar to *N. equitans* and its *Ignicoccus hospitalis* host. Our estimates of only a 1.1 Mbp genome for the *A. nanophilum* host suggests an expanded mutualism with its *Nanobsidianus* partner in the NL01 hot spring system. However, the ability to find *A. nanophilum* lacking its *Nanobsidianus* partner indicates that *A. nanophilum* is capable of

independent replication, making it the smallest genome of a free-living organism. This speculation will need to be confirmed by complete genome sequencing and culturing of *A. nanophilum* from *NL01*. The fact that the marine *N. equitans* and terrestrial YNP *N. stetteri* are associated with different hosts suggests additional Nanoarchaeota-host interactions will be discovered in other environments and that there is a high degree of flexibility in the host-Nanoarchaeota partnership found in nature. It is clear that independent Nanoarchaeota-host partnerships have arisen over time, and it also suggests that the Nanoarchaeota-host association may be an old and essential feature of the archaeal communities.

Reno et al. (36) showed that microorganisms in the same geographic region are more closely related to each other than members of the same species isolated from geographically distinct areas. While that study was limited to major geographical areas, its results likely can be applied to the phylogenetic distribution of *N. stetteri* in YNP. Cells from the same hot spring cluster more closely together than cells from different hot springs, suggesting that *N. stetteri* in each hot spring may be independently adapting to local hot spring environments.

Similar to previous descriptions of *N. stetteri*, we did not observe a CRISPR/Cas system, which is in contrast to *N. equitans,* where the CRISPR/Cas system was detected (37). While it is possible that the three partial genomes of *N. stetteri* from YNP have all missed the CRISPR/Cas system, we find this unlikely. It is noteworthy that all Nanoarchaeota genomes contain split tRNA genes. The absence of the CRISPR/Cas system in some Nanoarchaeota that contain split tRNAs diminishes the argument that split tRNAs would be unlikely to arise in Nanoarchaeota due to the ability of the CRISPR/Cas system to eliminate foreign genetic elements (38). We also find it more likely that the marine lineage gained the CRISPR/cas system as opposed to the terrestrial lineage losing this system, as the YNP genomes are significantly larger. The likely

414 presence of a virus within *N. stetteri* diminishes the argument that the lack of viral pressure is
415 responsible for the loss of the CRISPR /Cas system and provides support for split tRNAs arising
416 as a mechanism to escape integration of mobile genetic elements in the anticodon loop (39).
417 Our results are consistent with those of Podar et al. (7), but additionally suggest a
418 widespread distribution of Nanoarchaeota in the thermal features of YNP. This expanded
419 collection of sequences provides a foundation for future studies on Nanoarchaeota genome
420 reduction. The current model of Nanoarchaeota presumes that symbiont and host have evolved
421 together to the point where *N. equitans* are only able to grow with *I. hospitalis* and not with any
422 closely related *Ignicoccus* sp. (33). It remains to be seen if the *Nanobsidianus* of YNP share the
423 host limitations of *N. equitans* or if *Nanobsidianus* from one hot spring are able to grow with a
424 host species isolated from a different hot spring.
425 The detection of a virus likely replicating within *Nanobsidianus* cells is the first report of
426 a virus associated with ultrasmall Archaea (Nanoarchaeota, Parvachaeota, and
427 Nanohaloarchaeota). The discovery of such a virus raises interesting questions about the possible
428 three-way interactions between the virus, *Nanobsidianus* and *Acidicryptum* cells. To our
429 knowledge, this is the smallest host genome supporting virus replication. This highly reduced
430 host genome, which contains all the tRNA genes but lacks most of the genes required for central
431 carbon metabolism, provides an opportunity to examine the minimum requirements for viral
432 replication.

433 Acknowledgements

434 This work was supported by the National Science Foundation grants DEB-4W4596 (to MJY)
435 and DEB-1441717 and DBI-1226726 (to RS). We thank Jennifer Wirth and Ross Hartman for
436 their critical reading of this text. We also thank Matthew Lavin for his advice regarding

phylogenetic analysis. The research was conducted in Yellowstone National Park under the conditions of permit numbers YELL-2011-SCI-5090, YELL-2012-SCI-5090, and YELL-2013-SCI-5090. We also thank the YNP research resource office and especially Stacey Gunther for their work facilitating sampling within YNP.

Conflict of Interest

The authors declare no conflict of interest.

Supplemental Information

Supplemental information accompanies the paper

446 # Reference

1. **Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO**. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature **417**:63–7.

2. **Petitjean C, Deschamps P, López-garcía P**. 2014. Rooting the Domain Archaea by Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota. Genome Biol. Evol. **7**:191–204.

3. **Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P**. 2005. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome Biol. **6**:R42.

4. **Hohn MJ, Hedlund BP, Huber H**. 2002. Detection of 16S rRNA sequences representing the novel phylum "Nanoarchaeota": indication for a wide distribution in high temperature biotopes. Syst. Appl. Microbiol. **25**:551–4.

5. **Casanueva A, Galada N, Baker GC, Grant WD, Heaphy S, Jones B, Yanhe M, Ventosa A, Blamey J, Cowan D a**. 2008. Nanoarchaeal 16S rRNA gene sequences are widely dispersed in hyperthermophilic and mesophilic halophilic environments. Extremophiles **12**:651–6.

6. **Clingenpeel S, Kan J, Macur RE, Woyke T, Lovalvo D, Varley J, Inskeep WP, Nealson K, McDermott TR**. 2013. Yellowstone lake Nanoarchaeota. Front. Microbiol. **4**:274.

7. **Podar M, Makarova KS, Graham DE, Wolf YI, Koonin E V, Reysenbach A-L**. 2013. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. Biol. Direct **8**:9.

8. **Stepanauskas R, Sieracki ME**. 2007. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. Proc. Natl. Acad. Sci. U. S. A. **104**:9052–9057.

9. **Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, Reinthaler T, Poulton NJ, Masland EDP, Gomez ML, Sieracki ME, DeLong EF, Herndl GJ, Stepanauskas R**. 2011. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science **333**:1296–300.

10. **Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev M a., Pevzner P a.** 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. **19**:455–477.

11. **Wilkins MJ, Kennedy DW, Castelle CJ, Field EK, Stepanauskas R, Fredrickson JK, Konopka AE**. 2014. Single-cell genomics reveals metabolic strategies for microbial growth and survival in an oligotrophic aquifer. Microbiol. (United Kingdom) **160**:362–372.

12. **Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, Saw JH, Senin P, Yang C, Chatterji S, Cheng J-F, Eisen J a, Sieracki ME, Stepanauskas R**. 2009. Assembling the marine metagenome, one cell at a time. PLoS One **4**:e5299.

13. **Field EK, Sczyrba A, Lyman AE, Harris CC, Woyke T, Stepanauskas R, Emerson D**. 2014. Genomic insights into the uncultivated marine Zetaproteobacteria at Loihi Seamount. ISME J. **9**:857–870.

14. **Wolf YI, Makarova KS, Yutin N, Koonin E V**. 2012. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol. Direct **7**:46.

15. **Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW**. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. **114**:gr.186072.114.

16. **Snyder JC, Bateson MM, Lavin M, Young MJ**. 2010. Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. Appl. Environ. Microbiol. **76**:7251–8.

17. **Inskeep WP, Jay ZJ, Tringe SG, Herrgård MJ, Rusch DB**. 2013. The YNP Metagenome Project: Environmental Parameters Responsible for Microbial Distribution in the Yellowstone Geothermal Ecosystem. Front. Microbiol. **4**:67.

18. **Edgar RC**. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113.

19. **Huelsenbeck JP, Ronquist F**. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754–755.

20. **Loy A, Maixner F, Wagner M, Horn M**. 2007. probeBase--an online resource for rRNA-targeted oligonucleotide probes: new features 2007. Nucleic Acids Res. **35**:D800–4.

21. **Brileya KA, Camilleri LB, Fields MW**. 2014. 3D-Fluorescence In Situ Hybridization of Intact, Anaerobic Biofi lm, p. 189–197. *In* Sun, L, Shou, W (eds.), Engineering and Analyzing Multicellular Systems. Springer New York, New York, NY.

22. **Pernthaler a, Pernthaler J**. 2004. Sensitive multi-color fluorescence in situ hybridization for the identification of environmental microorganisms. Mol. Microb. Ecol. Man. **3**:711–726.

23. **John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, Brum J, Polz MF, Boyle E a, Sullivan MB**. 2011. A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environ. Microbiol. Rep. **3**:195–202.

24. **Lasken RS, Stockwell TB**. 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol. **7**:19.

25. **Abulencia CB, Wyborski DL, Garcia J a, Podar M, Chen W, Chang SH, Hwai W, Watson D, Brodie EL, Hazen TC, Chang HW, Keller M**. 2006. Environmental Whole-Genome Amplification To Access Microbial Populations in Contaminated Sediments Environmental Whole-Genome Amplification To Access Microbial Populations in Contaminated Sediments. Appl. Environ. Microbiol. **72**:3291–3301.

26. **Yilmaz S, Allgaier M, Hugenholtz P**. 2010. Multiple Displacement amplification compromises quantative analysis of metagenomes. Nat. Methods **7**:943–944.

27. **Delmont TO, Robe P, Clark I, Simonet P, Vogel TM**. 2011. Metagenomic comparison of direct and indirect soil DNA extraction approaches. J. Microbiol. Methods **86**:397–400.

28. **Abbai NS, Govender A, Shaik R, Pillay B**. 2012. Pyrosequence analysis of unamplified and whole genome amplified DNA from hydrocarbon-contaminated groundwater. Mol. Biotechnol. **50**:39–48.

29. **Darling ACE, Mau B, Blattner FR, Perna NT**. 2004. Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements 1394–1403.

30. **Randau L, Schröder I, Söll D**. 2008. Life without RNase P. Nature **453**:120–3.

31. **Schattner P, Brooks AN, Lowe TM**. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. **33**:W686–9.

32. **Makarova KS, Sorokin A V, Novichkov PS, Wolf YI, Koonin E V**. 2007. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol. Direct **2**:33.

33. **Jahn U, Gallenberger M, Paper W, Junglas B, Eisenreich W, Stetter KO, Rachel R, Huber H**. 2008. Nanoarchaeum equitans and Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea. J. Bacteriol. **190**:1743–50.

34. **Randau L**. 2012. RNA processing in the minimal organism Nanoarchaeum equitans. Genome Biol. **13**:R63.

35 . **Thompson CC, Chimetto L, Edwards R a, Swings J, Stackebrandt E, Thompson FL**. 2013. Microbial genomic taxonomy. BMC Genomics **14**:913.

36. **Reno ML, Held NL, Christopher J, Burke P V, Whitaker RJ,** 2009. Biogeography of the Sulfolobus islandicus pan-genome. Proc. Natl. Acad. Sci. **106**:8605–8610.

37. **Lillestøl RK, Redder P, Garrett R a, Brügger K**. 2006. A putative viral defence mechanism in archaeal cells. Archaea **2**:59–72.

38. **Di Giulio M**. 2013. Is Nanoarchaeum equitans a paleokaryote? J. Biol. Res. **19**:83–88.

39. **She Q, Shen B, Chen L**. 2004. Archaeal integrases and mechanisms of gene capture. Biochem. Soc. Trans. **32**:222–226.

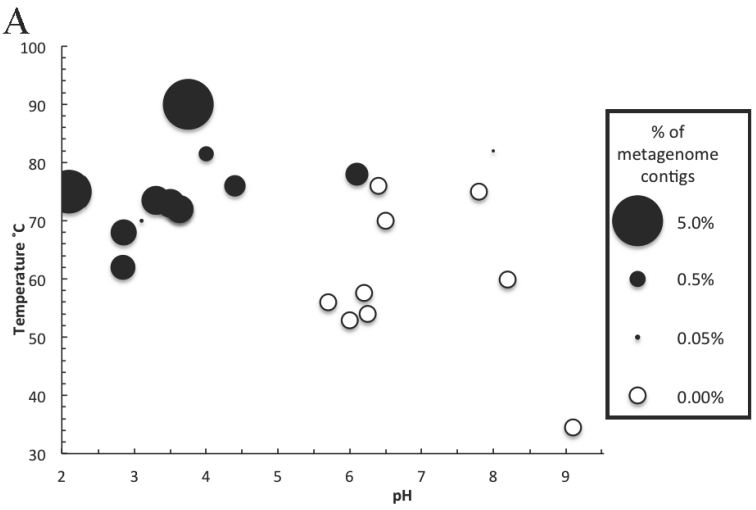
447 **Figure Legends**

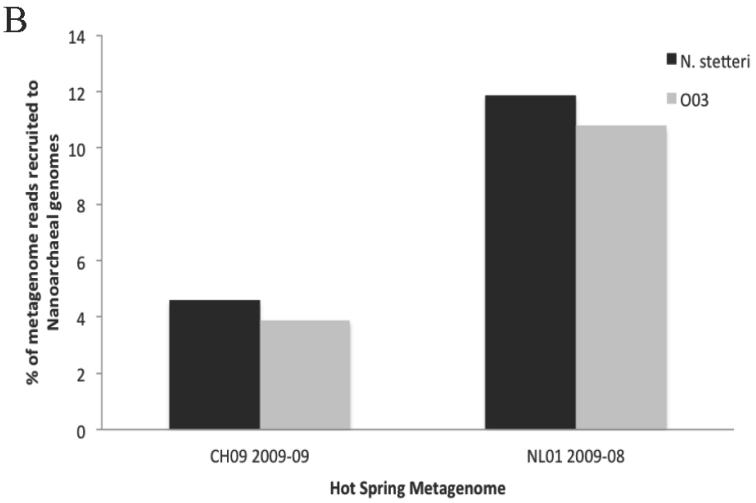
448 Fig. 1. Distribution of Nanoarchaea in 22 hot springs in Yellowstone National Park based on hot
449 spring pH and temperature. The size of the solid black circles indicates the percent of the
450 metagenome contigs that were recruited to the *N. stetteri* single cell genome O03, while open
451 circles indicate hot spring metagenomes with no nanoarchaeal sequences recruited (A).
452 Recruitment of metagenomic reads to *N. stetteri* Obsidian Pool (grey) and *N. stetteri* O03 (black)
453 partial genomes indicating relative abundance of *Nanobsidianus* in two YNP hot springs (B).
454 Transformed maximum likelihood phylogenetic tree of seven concatenated conserved
455 Nanoarchaeota genes in F03, O03, *N. stetteri*, *N. equitans* and CH09 cellular metagenomes.
456 Numbers indicate the posterior probability and *N. equitans* served as the outgroup (C).
457
458 Fig. 2. Mauve alignment of O03 concatenated partial length genome to F03 (A), *N. stetteri* (B)
459 and full length *N. equitans* (C) genomes. Blocks of the same color indicate regions of homology
460 between the genomes and blocks below the genome's center-line are inverted relative to the
461 reference sequence.
462
463 Fig. 3. CARD-FISH analysis of *N. stetteri*-host interaction from environmental samples. Total
464 cells from CH09 2012-01-stained blue with DAPI (A) cells imaged with 16S ribosomal probe to
465 *Acidicryptum* host cells (green, B). Cells imaged with 16S ribosomal probe to *Nanobsidianus*
466 (red, C). Merged image of A-C indicating co-localization of *Acidicryptum* and *N. stetteri* cells
467 (D). Control of a merged image of *Acidolobus* cells (green) and *Nanobsidianus* cells (red) from
468 the same hot spring and sampling date showing the lack of co-localization (E). Additional
469 merged images of *Nanobsidianus* cells (red) and *Acidicryptum* host cells (green) from a different
470 sampling date CH09 2013-05 (F) and hot spring NL01 2014-09 (G)

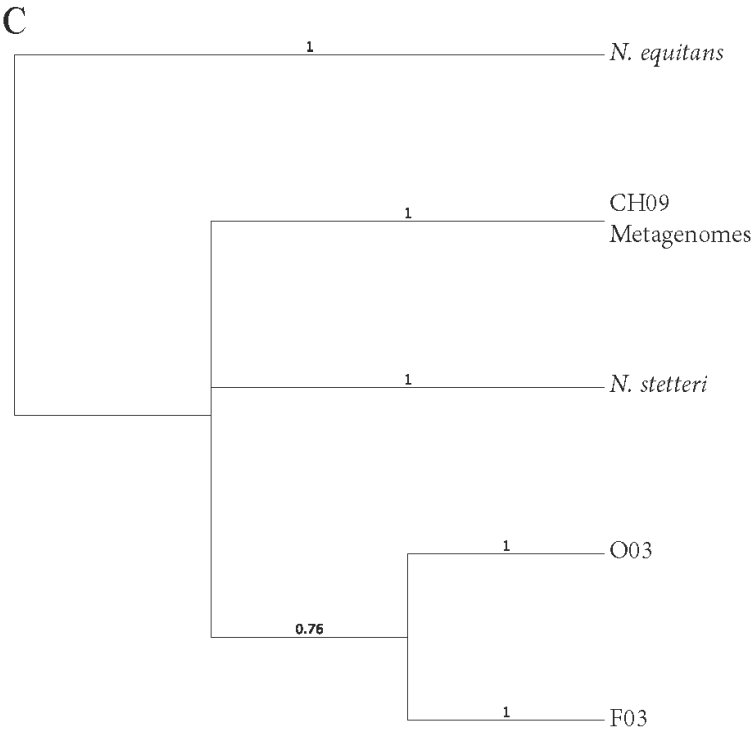
471

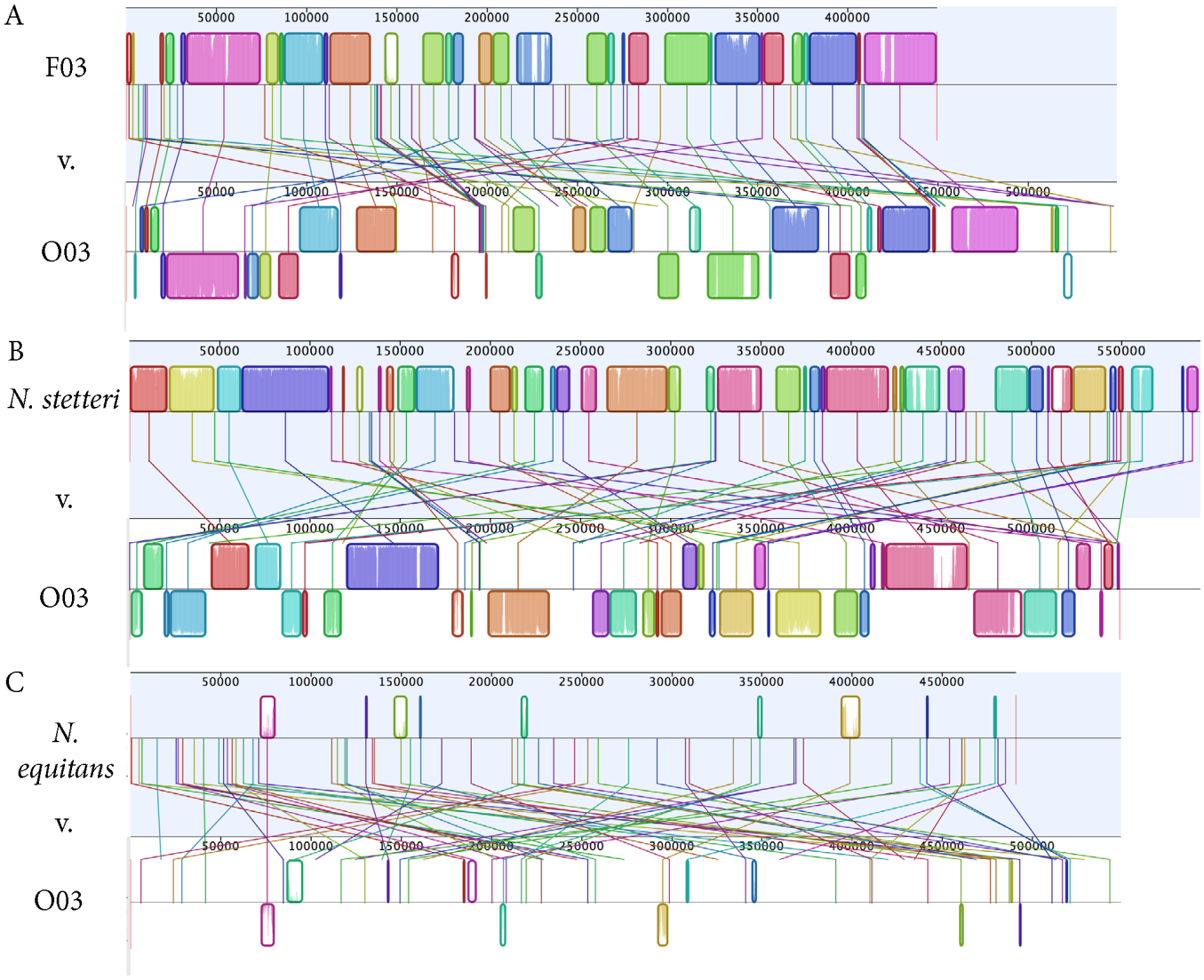
472 Fig. 4. PCR analysis of virus purified from NL01 environmental sample. Fractions of CsCl
473 buoyant density gradients were tested for a viral genome segment (A), and for contamination of
474 cellular sequences *Acidicryptum* (B) or *N. stetteri* (C) in the same viral fractions. Numbers
475 indicate the density of the CsCl gradient in g/cm^3, (+) indicates a positive control template, and
476 NTC is the no template control, white arrow indicates the presence of the expected viral product.

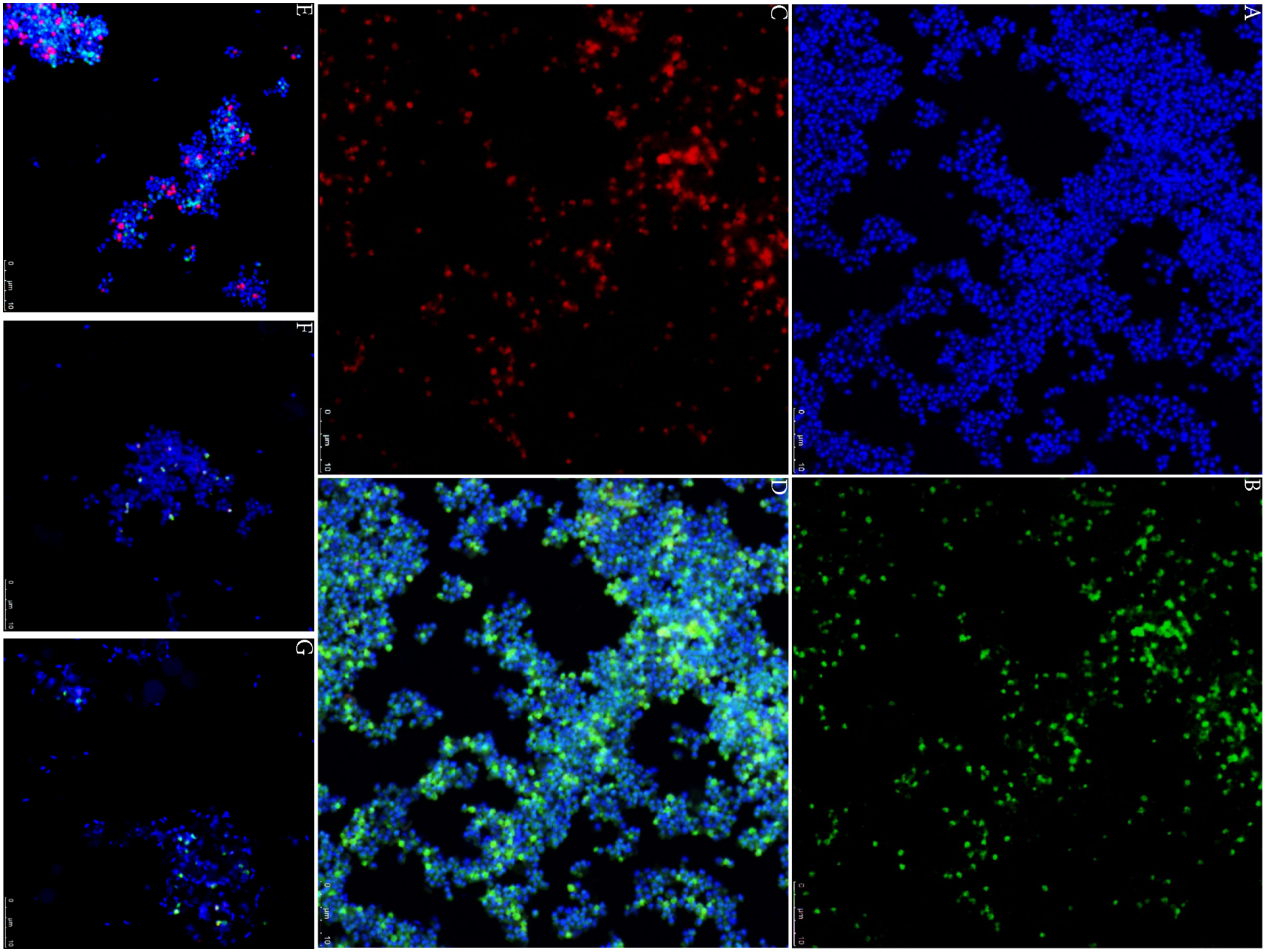
477

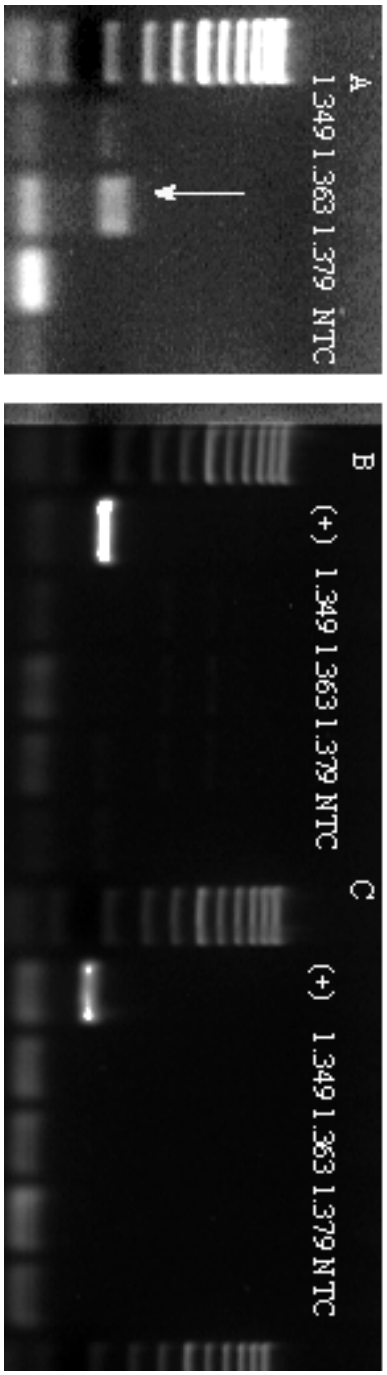









A
B
C
D
E
F
G


A
1.349 1.363 1.379 NTC
B
(+) 1.349 1.363 1.379 NTC
C
(+) 1.349 1.363 1.379 NTC