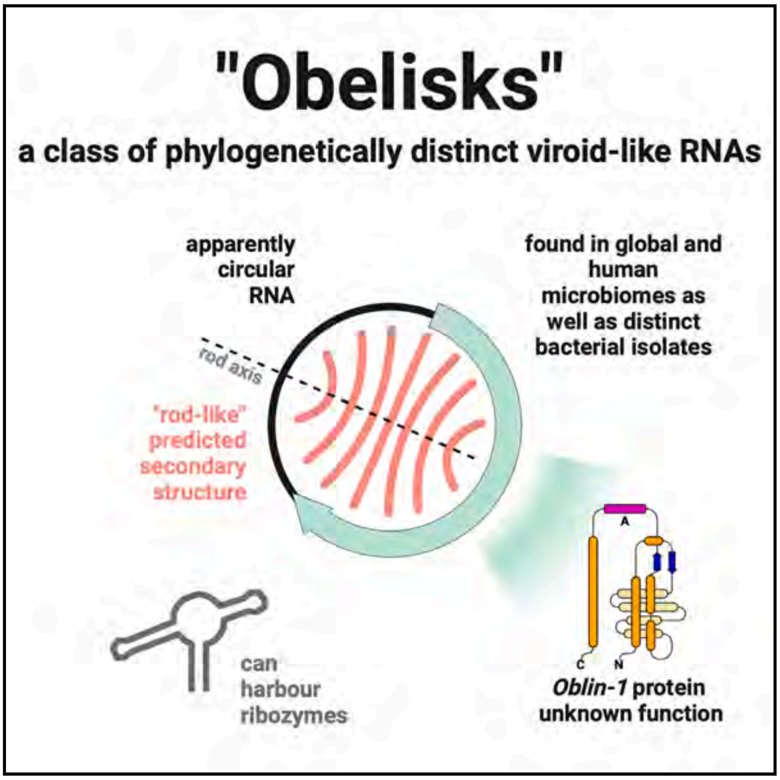
Cell

Article

Viroid-like colonists of human microbiomes

Graphical abstract


"Obelisks"
a class of phylogenetically distinct viroid-like RNAs
apparently circular RNA
found in global and human microbiomes as well as distinct bacterial isolates
rod axis
"rod-like" predicted secondary structure
A
C
N
can harbour ribozymes
Oblin-1 protein unknown function

Authors

Ivan N. Zheludev, Robert C. Edgar, Maria Jose Lopez-Galiano, Marcos de la Peña, Artem Babaian, Ami S. Bhatt, Andrew Z. Fire

Correspondence

zheludev@stanford.edu (I.N.Z.), afire@stanford.edu (A.Z.F.)

In brief

Obelisks are widespread RNA-based agents found in diverse environmental and human-associated microbiomes. These apparently unbranched and circular RNAs represent a previously uncharacterized class of genetic elements.

Highlights

- Obelisks are a phylogenetically distinct group of microbiome-associated, viroid-like RNAs
- Found globally in diverse niches, obelisks also occur in human stool and oral microbiomes
- The human oral bacterium *Streptococcus sanguinis* SK36 harbors a distinct “obelisk-*S.s*”
- Under replete growth conditions, obelisk-*S.s* appears to be disposable for SK36 growth

Zheludev et al., 2024, Cell *187*, 6521–6536
November 14, 2024
https://doi.org/10.1016/j.cell.2024.09.033

CellPress

Article

Viroid-like colonists of human microbiomes

Ivan N. Zheludev,[1,9,*] Robert C. Edgar,[2] Maria Jose Lopez-Galiano,[3] Marcos de la Peña,[3] Artem Babaian,[4,5]
Ami S. Bhatt,[6,7] and Andrew Z. Fire[6,8,*]
[1]Stanford University, Department of Biochemistry, Stanford, CA, USA
[2]Independent researcher, Corte Madera, CA, USA
[3]Instituto de Biología Molecular y Celular de Plantas, Universidad Politécnica de Valencia-CSIC, Valencia, Spain
[4]University of Toronto, Department of Molecular Genetics, Toronto, ON, Canada
[5]University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, Toronto, ON, Canada
[6]Stanford University, Department of Genetics, Stanford, CA, USA
[7]Stanford University, Department of Medicine, Division of Hematology, Stanford, CA, USA
[8]Stanford University, Department of Pathology, Stanford, CA, USA
[9]Lead contact
*Correspondence: zheludev@stanford.edu (I.N.Z.), afire@stanford.edu (A.Z.F.)
https://doi.org/10.1016/j.cell.2024.09.033

SUMMARY

Here, we describe "obelisks," a class of heritable RNA elements sharing several properties: (1) apparently circular RNA ~1 kb genome assemblies, (2) predicted rod-like genome-wide secondary structures, and (3) open reading frames encoding a novel "Oblin" protein superfamily. A subset of obelisks includes a variant hammerhead self-cleaving ribozyme. Obelisks form their own phylogenetic group without detectable similarity to known biological agents. Surveying globally, we identified 29,959 distinct obelisks (clustered at 90% sequence identity) from diverse ecological niches. Obelisks are prevalent in human microbiomes, with detection in ~7% (29/440) and ~50% (17/32) of queried stool and oral metatranscriptomes, respectively. We establish *Streptococcus sanguinis* as a cellular host of a specific obelisk and find that this obelisk's maintenance is not essential for bacterial growth. Our observations identify obelisks as a class of diverse RNAs of yet-to-be-determined impact that have colonized and gone unnoticed in human and global microbiomes.

INTRODUCTION

RNA viruses (*Riboviria*) are in part defined by their encoding of their own replicative polymerases, a feature that can be leveraged for homology-based viral discovery.[1–5] By contrast, viroids[6,7] and hepatitis delta-like viral (HDV) "satellites"[8] (Figure S1) co-opt eukaryotic host RNA polymerases for their replication, resulting in some of biology's smallest known genomes (viroids: ~350 nt; delta: ~1.7 kb). These streamlined genomes define the working limits of biological information transfer,[9,10] and their simplicity raises the question of why, compared with *Riboviria*, there are so few known examples of viroids and similar agents. Recently, inquiries based on protein similarity have uncovered new delta-like agents.[2,11] Likewise, viroids, which lack any protein-coding capacity, are beginning to be surveyed at a larger scale based in part on circular genome maps and the presence of ribozyme-like features. These searches have led to an expanded family of known viroid-like RNAs and a revision of earlier models that their distribution is limited to plants.[12–14] Thus, these studies have already shifted virological paradigms, leaving open the possibility that an even broader category of viroid-like elements is present in living systems, which might have been overlooked due to a lack of detectable similarity to known viroids and HDV family members.

The human gut microbiome (hGMB) is experimentally attractive for the discovery of novel genetic agents. Indeed, "metagenomic" and metatranscriptomic[15] profiling of the hGMB has yielded new insights into prokaryotic, viral,[16–18] and plasmid[19] ecology. To this end, we developed a reference-free bioinformatic approach ("viroid nominator" [VNom]) to identify novel viroid-like elements. We initially applied VNom to published Integrative Human Microbiome Project (iHMP) data,[20] resulting in the identification of a new class of hGMB-colonizing RNA agents, which we term obelisks. Obelisks form a distinct phylogenetic group restricted to RNA datasets and lack any evident homology to characterized genomes or viromes. Obelisk RNA reads assemble into ~1,000 nt circles, which are predicted to fold into rod-like RNA secondary structures and code for at least one member of an apparently novel Oblin protein superfamily. We further found that a subset of obelisks harbors obelisk-specific hammerhead ribozyme motifs. While querying 5.4 million public sequencing datasets, we identified 29,959 distinct obelisks (90% identity threshold) present across ~220,000 datasets representing diverse ecosystems beyond the hGMB. Among the datasets with clear obelisk representatives, we identified a definitive obelisk-host pair, with *Streptococcus sanguinis* (*S. sanguinis*) acting as a replicative host. We show that under replete laboratory growth conditions, this obelisk is able to

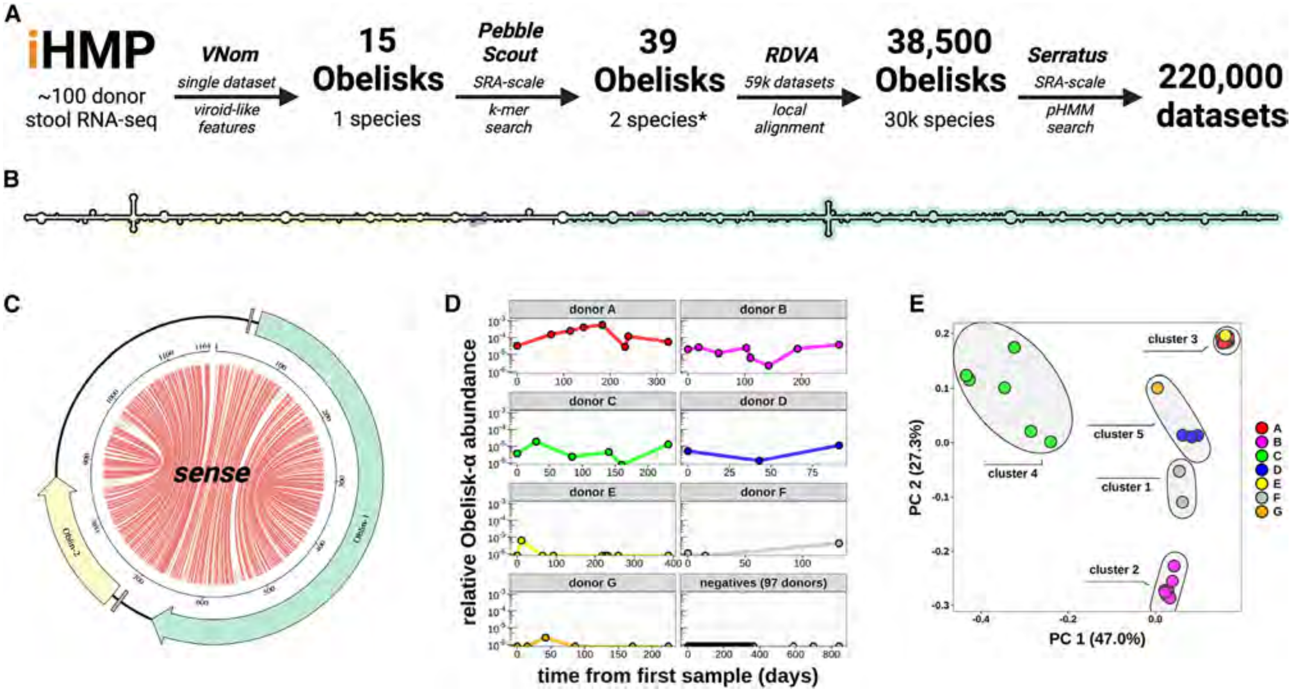


Figure 1. Obelisk-alpha has a predicted extensive secondary structure and appears to colonize and speciate within the human gut

(A) Overview of the iterative approach taken in obelisk discovery, (see STAR Methods).

(B and C) (B) Schematic of the predicted *sense* consensus secondary structure derived from all non-redundant, 1,164 nt obelisk-αs found using SRA-scale k-mer matching (PebbleScout). Predicted open reading frames (ORFs) 1 and 2 (green/yellow), and Shine-Dalgarno sequences (purple) shown and (C) jupiter plot of obelisk-α colored as in (B), chords illustrate predicted base pairs (base pair probabilities gray, 0.1, to red, 1.0).

(D) Obelisk-α relative read abundance for six donors (A–G); sequence data from in Lloyd-Price et al.[20] and time in days from first sample.

(E) Principal component analysis of sequence variation seen in obelisk-α reads in Lloyd-Price et al.[20] (the initial iHMP dataset), grouped by k-means clustering with 5 centers, colored as in (D).

See also Figure S1.

persist in *S. sanguinis* but that it is non-essential to bacterial fitness under these conditions. Lastly, we surveyed obelisks in 5 published human oral and gut microbiome studies from 472 donors, finding an estimated ~9.7% donor prevalence within these datasets, with an apparent anatomy-specific obelisk distribution.

RESULTS

A previously unnoticed, human microbiome-associated, viroid-like RNA

Viroids and delta viruses are in part typified by their single-stranded, circular genomes, both of which are molecular features that can be detected in strand-specific RNA sequencing (RNA-seq). To search for such features in microbiome RNA-seq datasets, we created a bioinformatic tool, VNom (see VNom and Figure S1D), and applied it to microbiome RNA-seq datasets (see initial obelisk identification). In particular, we chose an iHMP human stool dataset[20] for its strand-specific RNA-seq, its longitudinal nature (regular sampling over ~1 year), and its cohort size (104 donors), all qualities well suited for identifying persistent hGMB colonists.

We next filtered VNom-nominated RNAs to retain contigs with no evident homology to the NCBI BLAST (nt or nr) databases[21] (see initial obelisk identification). One class of 15 related (<2% sequence variation; Table S2) 1,164 nt RNAs stood out with their extended predicted secondary structure reminiscent of HDV and *Pospiviroidae* (Figures 1B, S1A, S1B, and S2B). Owing to a strong predicted rod-like secondary structure, we term this group of RNAs obelisk-alpha (obelisk-α, "*Obelisk_000001*" in Table S2). At 1,164 nt in length, the rod-like secondary structure was striking because typical mRNA sequences are not predicted to readily fold in this manner (as evidenced by the efforts required to maximize the degree of "rod-ness" in mRNA vaccines[22]). Based on open reading frame (ORF) predictions, obelisk-α has the capacity to code for two proteins (202 and 53 amino acids [aa]). Both ORFs lack evident nucleotide or protein sequence homology when querying a number of reference databases (NCBI nt, nr, or CDD,[23] Pfam[24]). Tertiary structure protein alignment yielded similar negative results (see protein tertiary structure prediction). Therefore, we chose new names, terming these two proteins "Oblin-1" and "Oblin-2," respectively. We specifically note that despite some similar characteristics between obelisk-α and HDV (apparently circular, predicted highly structured RNA genome and the ability to code for at least one ~200-aa ORF; Figure S1A), there is no evident sequence homology at the RNA level or protein level or structural homology at the protein level between obelisks and HDV. In further contrast to HDV, whose large hepatitis delta antigen (L-HDAg) occurs on one strand of the extended HDV predicted secondary structure (Figure S1A), the obelisk-α Oblin-1-encoding region is largely self-complementary within the ORF, forming a ∼300-bp hairpin making up half of the predicted obelisk-α RNA secondary structure (Figures 1B and 1C). Obelisk-α sequences were found to occur in 7 of the 104 iHMP donors (Table S1), with donors A–C exhibiting consistent prevalence for over 200 days (Figure 1D; note: positive donors are renamed for brevity, with original donor alias equivalences in Table S1). Further, obelisk-α sequences were found to largely cluster together based on donor identity, when grouped by sequence variation (Figure 1E). We noted some co-clustering of sequences between donors (A and E in cluster 3 and D and G in cluster 5); this co-clustering could be explained by either transient prevalence or by library cross contamination, as each minor member of such clusters was both low prevalence (few positive time points) and low abundance (low counts in positive time points) (see Table S1). Regardless of the source of the relatively rare cross-sample reads, obelisk-α appears to persist within human donors, with each donor appearing to harbor their own distinct "strain." Lastly, in companion DNA-seq data from this project, no detectable obelisk reads are found (Table S1). Taken together, these findings are consistent with obelisk-α representing an as yet uncharacterized RNA element with viroid-like features, which occurs in human stool and further comprises subspecies that persist in individual donors over time.

Public data are replete with obelisk-like elements

Using obelisk-α as a starting point, 21 additional full-length examples of obelisk-α (<4% sequence variation; Table S2) were found in 7 datasets using a k-mer search (PebbleScout[25]) of ∼3.2 million metagenomic annotated sequence read archive (SRA) datasets. All seven datasets were human-derived metatranscriptome (metagenomic RNA) BioProjects (Table S1; see obelisk homolog detection in additional public data); zero sequences were found in metagenomic DNA samples. The repeated finding of obelisk-α in disparate BioProjects supported the notion that obelisk-α is a *bona fide* biological entity. Based on the prevalence of obelisk-α in these human microbiome transcriptome datasets (Table S1), we investigated the possibility that additional obelisks might be present in such data (as identified by both VNom and Oblin-1 protein similarity). This search ultimately led to the discovery of obelisk-beta (obelisk-β, "*Obelisk_000002*" in Table S2), a 1,182-nt, likely hGMB-resident, obelisk-like RNA with similar characteristics to obelisk-α (i.e., circular assembly map, rod-like predicted secondary structure, and absence in paired DNA sequence) and a low but evident protein sequence similarity to Oblin-1 (∼38% protein similarity and pairwise mean BLASTp E value: 5.2×10^{-14}). Thus, both obelisks appear to be Oblin-1-encoding elements. Analysis of the Oblin-2 homology at this stage was limited by the short size of the proteins—nonetheless, both obelisk-α and obelisk-β encode second proteins of ∼50 aa rich in helix-forming residues (Figures 2C and 2D, S2A–S2C, and S4C and S4D). Next, utilizing the uniqueness of the obelisk-α/β Oblins-1 and -2 as obelisk-specific hallmark sequences, we searched over 12 trillion contigs in the RNA deep virome assemblage (RDVA), a database of assembled metatranscriptomes[13,26] (see obelisk homolog detection in additional public data), yielding over 38,500 Oblin-encoding RNA assemblies. Following this search, the smaller obelisk-α/β proteins were determined to be likely Oblin-2 homologs (∼31% protein similarity and mean BLASTp E value against the Oblin-2 consensus sequence: 2.5×10^{-6}). Ultimately, by insisting on evidence of apparent circularity, we created a "stringent" subset of 7,202 clustered obelisks (1,744 clusters at 80% nucleotide identity) as a conservative database for future studies (Table S2). Building from these RDVA hits, we then queried ∼5.4 million SRA datasets for distant Oblin-1 and -2 homology, using Serratus[2] (applying an inclusion threshold from earlier Serratus projects, see serratus), yielding over 220,000 putatively obelisk-positive datasets. From these datasets, we followed up on the 4,505 datasets with confident Oblin-1 hits (see serratus). These searches suggest that obelisk-like elements are found globally (Figure 3C), and they represent a distinct, diverse group of apparently phylogenetically related RNA-based elements (Figure 3A).

An oral commensal bacterium, *S. sanguinis*, serves as one obelisk host

The task of identifying specific host-agent pairings from metagenomic data presented a number of challenges. Most samples with obelisk homologs that were retrieved from the various searches were from metatranscriptomic samples derived from complex mixtures such as highly biodiverse microbiome and wastewater samples (Figure 3B). Thus, the potential host(s) of obelisk elements were not immediately clear. While correlation- and co-occurrence-based methods for inferring potential hosts are possible,[3,4,16] concerns about their statistical validity and interpretability[27–29] motivated a more direct strategy for obelisk-host identification. Consequently, we combed the serratus results for obelisk-like elements found in limited-complexity samples, such as defined monoculture and/or co-cultures. This search yielded a set of independent sequencing datasets from *S. sanguinis* (strain SK36), a commensal bacterium of the healthy human oral microbiome.[30] Several RNA-seq datasets (Table S1) from *S. sanguinis* strain SK36 contained an Oblin-1 coding obelisk-like sequence (see *streptococcus sanguinis* bioinformatics). These datasets evidenced a well-defined RNA element, which we refer to as "obelisk-*S.s*" ("*Obelisk_000003*" in Table S2). This RNA has the hallmark features of an obelisk: a characteristic length (1,137 nt), circular assembly with an obelisk-shaped predicted RNA secondary structure; genome similarity to obelisks-α and -β (41% and 35% nucleotide identity, respectively); and an Oblin-1 homolog (α and β: 33% protein similarity and mean pairwise E values of 5.2×10^{-5} and 4.5×10^{-7}, respectively). Unlike the other two obelisks, however, it lacks a predicted Oblin-2 homolog (Figures S2A/S2D). Overall, based on sequence homology, the predicted genomic secondary structure, the Oblin-1 tertiary structure, and the obelisk-characteristic Oblin-1 self-complementarity (Figure S2D), this RNA element is a *bona fide* obelisk. Further, the robust co-occurrence of *S. sanguinis* SK36 with obelisk RNA-seq reads (Table S1) positions *S. sanguinis* SK36 as a model system for future obelisk characterization.

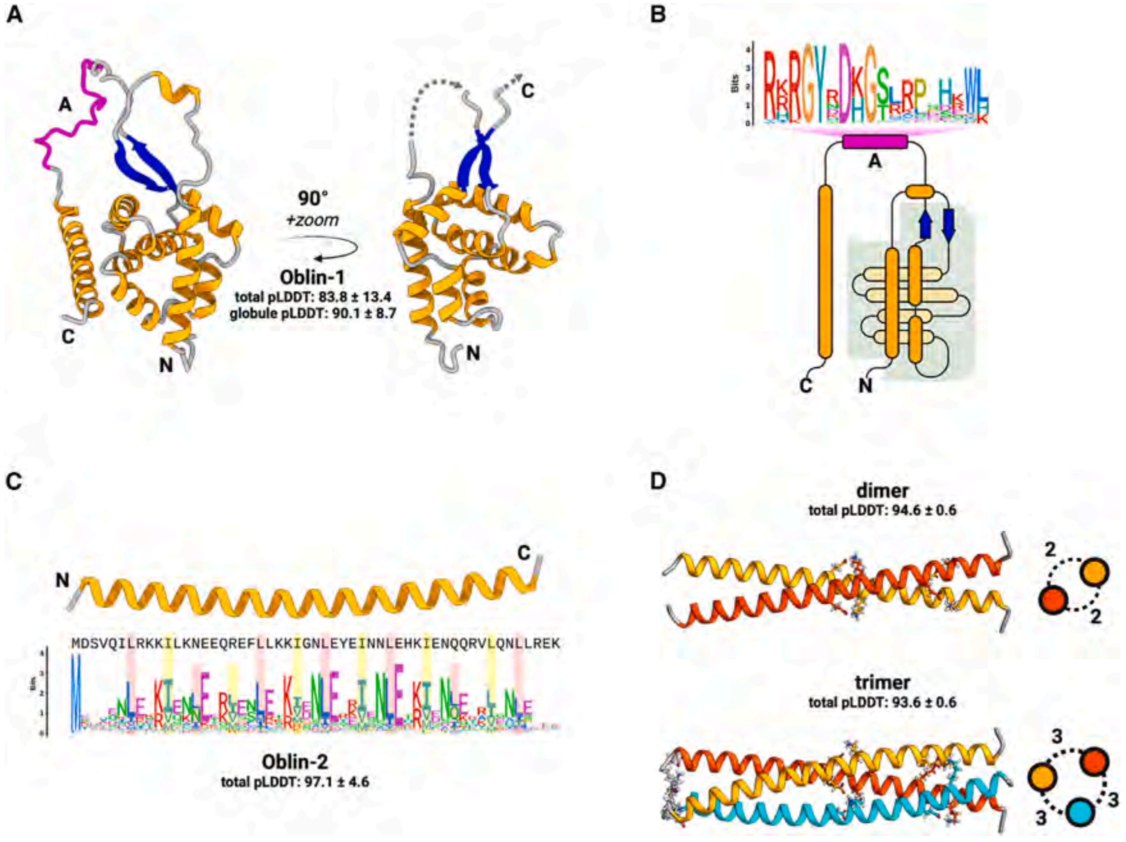


Figure 2. Obelisks encode putatively well-folded proteins

(A) Obelisk open reading frame 1 (Oblin-1) is predicted (total mean-pLDDT ± SD = 83.8 ± 13.4, see STAR Methods) to fold into a stereotyped N-terminal globule formed of a three alpha helix (orange) bundle partially wrapping around an orthogonal four helix bundle, capped with a beta sheet clasp (blue, globule mean-pLDDT = 90.1 ± 8.7), joined by an intervening region harboring the conserved *domain-A* (magenta) with no predicted tertiary structure, to an arbitrarily placed C-terminal alpha helix. Globule emphasized on the right.

(B) A to-scale (secondary structure) topological representation of Oblin-1 with the globule shaded in gray, and the *domain-A* emphasized with this bit-score sequence logo (see STAR Methods).

(C) Obelisk Oblin-2 is confidently predicted (mean-pLDDT = 97.1 ± 4.6) to fold into an alpha helix which appears to be a leucine zipper. Sequence logo of an i+7 leucine spacing emphasized in red, with hydrophobic "d" position residues emphasized in yellow (expanded in Figure S4C).

(D) Homo-multimer predictions of obelisk-*alpha* Oblin-2. Top: dimer (mean-pLDDT = 94.6 ± 0.6); bottom: trimer (mean-pLDDT = 93.6 ± 0.6). Side-on representations of homomultimers shown with numbers of inter-helix salt bridges (see Figure S4D).

See also Figure S4.

Derivatives of *S. sanguinis* SK36 that carry and lack obelisk-*S.s* sequences

We next sought to characterize the apparent *S. sanguinis* SK36 (hereafter "SK36")-obelisk-*S.s* in microbial monoculture, asking the following: (1) can SK36 monocultures retain obelisk-*S.s*, (2) does obelisk-*S.s* have any detectable DNA counterparts, (3) are the short read assemblies supported by long read sequencing, (4) are both strands of the obelisk represented in RNA-seq data, and (5) are there molecular or gross physiological consequences of harboring obelisk-*S.s* in standard (replete) laboratory conditions for bacterial growth?

To initially detect obelisk-*S.s* RNA, we used a reverse-transcriptase PCR assay (RT-PCR; see STAR Methods) followed by gel electrophoresis. RNA from double-colony purified SK36 substrains (Figure S3B) produced a positive signal for obelisk-*S.s* in such assays, while DNA did not. Being an "RNA-only" element, traditional genetic knockout approaches to generating a null substrain were not appropriate; instead, RT-PCR screening of multiple colonies not only revealed a general retention of the obelisk-*S.s* signal but also yielded an SK36 strain that appears to have serendipitously lost its obelisk-*S.s*, suggesting that the growth conditions used (see STAR Methods) are not fully selective for obelisk-*S.s* maintenance. This apparent spontaneous null "obelisk-negative-1" (ObN1) strain was paired with an arbitrary, similarly passaged "obelisk-positive-1" (ObP1) strain for further comparative analysis. We note that such apparent spontaneous loss of obelisk-*S.s* has occurred unnoticed in previous studies of the SK36 transcriptome (Table S1).

Total RNAs extracted from ObN1 and ObP1 were resolved by gel electrophoresis (Figure S3A) and stained with ethidium bromide. In addition to the strong ribosomal bands present in both,

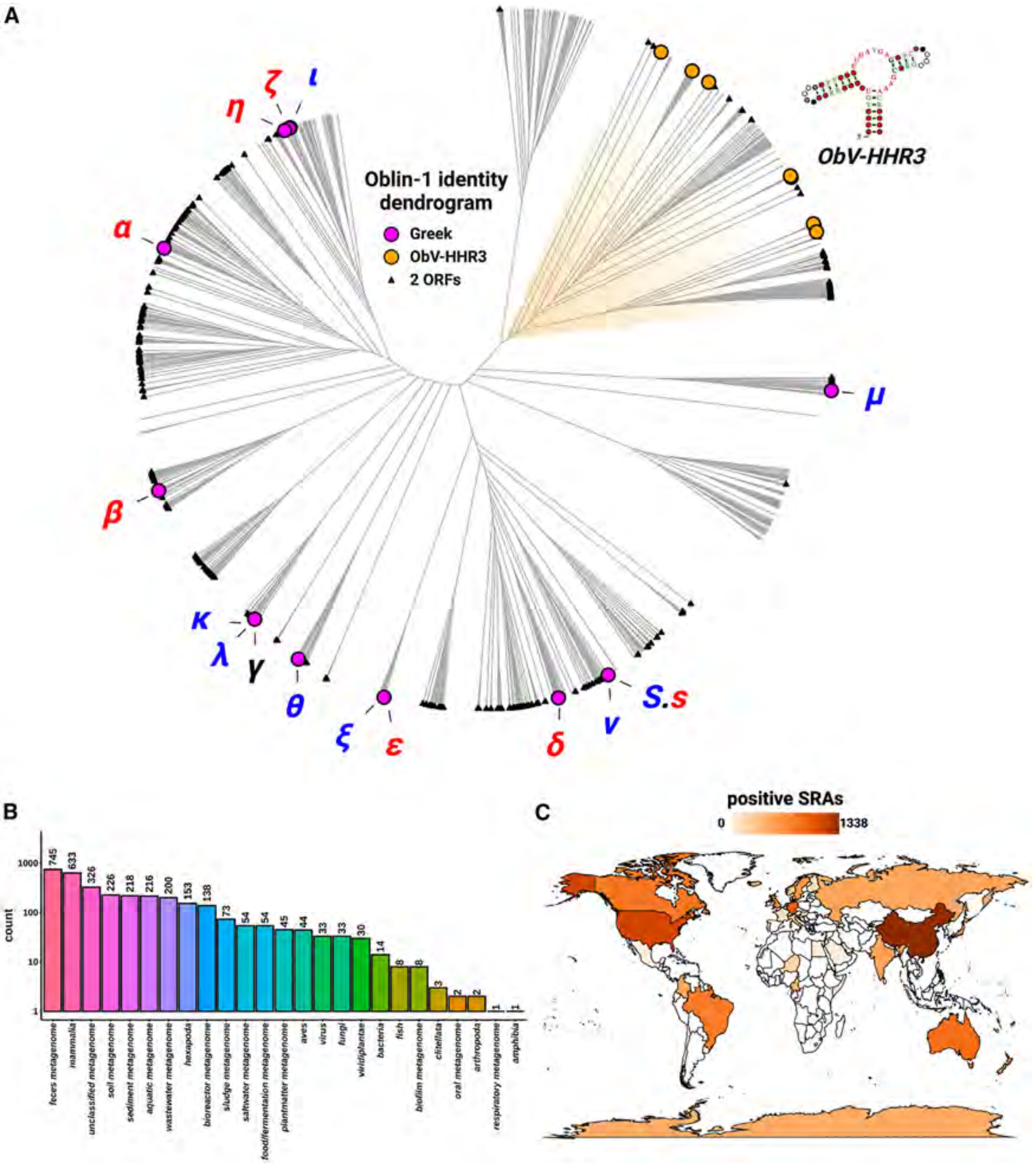


Figure 3. Obelisks form a globally distributed phylogenetic group

(A) A pairwise distance, neighbor joining, midpoint-rooted, dendrogram (branch lengths ignored, see STAR Methods) constructed from a non-redundant set of 1641 RDVA *Oblin-1* sequences, with obelisk-variant self-cleaving hammerhead type-III ribozymes illustrated as orange circles on leaves, and obelisks possessing exactly two predicted ORFs indicated with black triangles. Leaves that correspond to sequences from Figure 6 are illustrated with magenta circles and are colored by their microbiome of discovery (red = gastric, blue = oral, black = unknown).

(B) Counts of filtered SRA datasets from serratus and RDVA sorted by their host metadata (see STAR Methods).

(C) Datasets from (B) arranged by sample geolocation (where known) illustrated on a world map (darker orange = more SRA datasets). We note that SRA counts are not expected to correlate with true geo-/ecological prevalence, but are still indicative of global presence.

See also Figure S5.

we observed conspicuous ObP1-specific "extra material," with a band at approximately the size range expected for obelisk-*S.s*. The presence of an apparent ObP1-specific band suggests that obelisk-*S.s* may comprise an appreciable fraction of total ObP1 RNA. Strikingly, no noticeable differences in liquid aerobic culture growth (brain heart infusion broth, at 37°C; see STAR Methods) was observed between ObN1 and ObP1, either in lag time, doubling time, or final density (Figures 4A–4C).

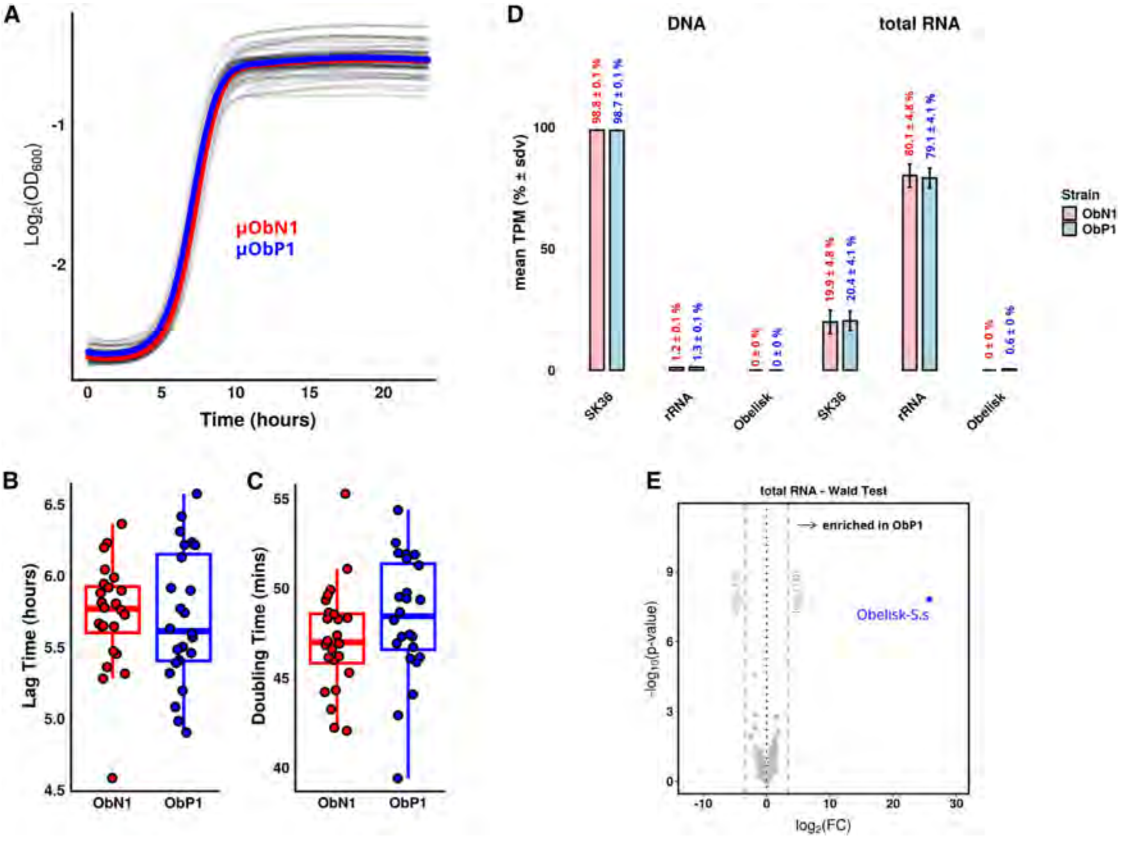


Figure 4. Obelisk-*S.s* is dispensable for SK36 growth in replete conditions

(A–C) (A) *Streptococcus sanguinis* SK36 substrains positive (ObP1) and negative (ObN1) for obelisk-*S.s* do not appear to grow discernibly differently in replete aerobic liquid culture (octuplet cultures of triplicate isolated substrains per ObP1/ObN1, brain heart infusion broth, 37°C, see STAR Methods). Likewise, computed growth characteristics do not show discernible effects from loss of obelisk-*S.s* either in lag time (B, mean ± SD, 5.7 ± 0.37 h for ObN1, and 5.7 ± 0.47 h for ObP1), or in doubling time (C, mean ± SD, 47.2 ± 2.9 min for ObN1, and 48.4 ± 3.4 min for ObP1).

(D) Short read sequencing (see STAR Methods) of triplicate cultures of ObN1 and ObP1 (red and blue, respectively) indicate that obelisk-*S.s* is exclusively RNA (see also Figure S3B), with the RNA and accountings for 0.6% ± 0.04 % of the total ObP1 transcriptome.

(E) Differential expression analysis indicates that under these growth conditions and statistical methods (see STAR Methods) that no transcripts other than obelisk-*S.s* were significantly differentially expressed between ObP1 and ObN1 (blue = q-value ≤ 0.05). Links to analysis results for rRNA-depleted and RNaseR-treated data are available in the key resources table.

See also Figure S3.

Short read sequencing (see STAR Methods) of RNA and DNA populations from 12-h growths provided further information about obelisk-*S.s*, notably supporting an RNA-only nature of obelisk-*S.s* (Figure 4D), as no reads were observed in DNA. Supporting the loss of the obelisk in the ObN1 substrain, RNA from this strain was likewise devoid of obelisk-*S.s* reads. By contrast, a remarkable abundance of obelisk-*S.s* reads was evident in the sequencing data of ObP1 reads; obelisk-*S.s* accounted for 0.6% ± 0.04% of total RNA reads in ObP1, compared with 79.1% ± 4.1% from ribosomal RNA (rRNA), and 20.4% ± 4.1% other SK36-mapping reads. The ObP1 obelisk-*S.s* fraction further increased when total RNA was either ribosomally depleted (3.5% ± 0.9% obelisk-*S.s* and 0.1% ± 0.1% rRNA) or treated with a 3′-→5′ specific RNA exoribonuclease, RNaseR (6.6% ± 1.1% obelisk-*S.s* and 9.8% ± 1.8% rRNA; see STAR Methods and the key resources table). Together with the ObP1-correlating "extra band" in the total RNA gel, these data suggest that obelisk-*S.s* comprises a non-negligible fraction of carrier SK36 strains' transcriptomes; strikingly however, but in-line with the growth curve data, an analysis of chromosomal gene expression in triplicate RNA-seq data from ObP1 and ObN1 shows a remarkably similar pattern with few, if any, outliers (see STAR Methods, Figure 4E). These data argue against an essential role for the obelisk under these growth conditions.

Analyzing the DNA-seq data from ObP1 and ObN1 revealed no large-scale changes, with six well-supported SNP variants, three in each substrain (see STAR Methods). There was no evident difference between the two substrains in variant frequency. While these SNPs could reflect adaptive changes related to the presence or absence of obelisk-*S.s*, an alternative possibility is that these represent background mutations that are present in the original *S. sanguinis* population from

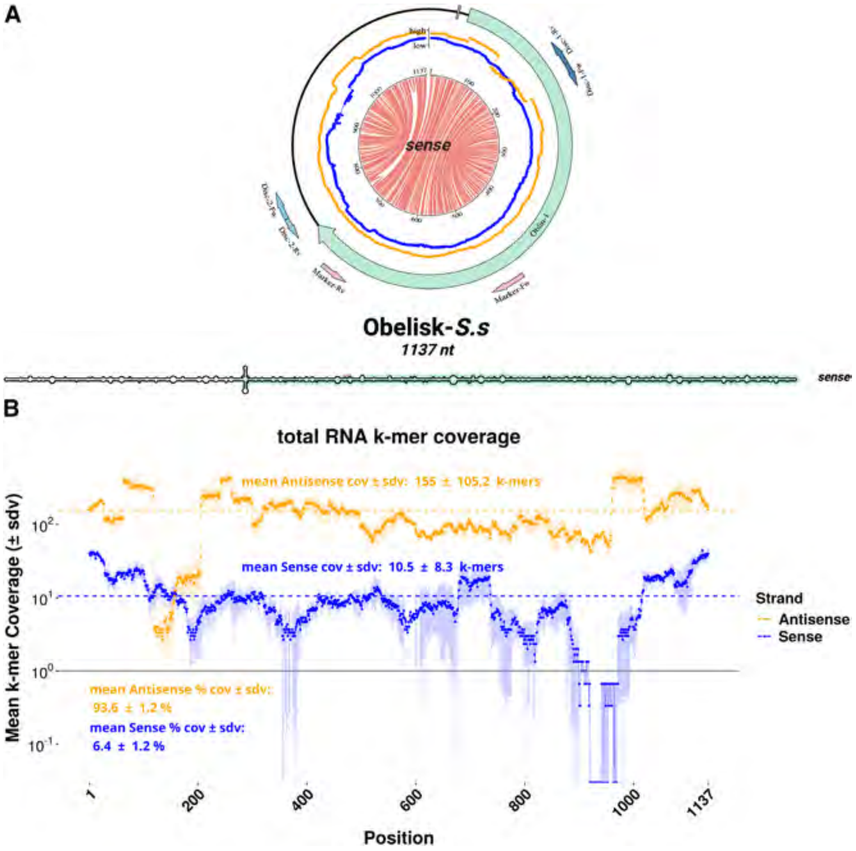


Figure 5. *Streptococcus sanguinis* SK36 harbors obelisk-*S.s*

(A) Jupiter plot the *Streptococcus sanguinis* SK36 obelisk-*S.s* ("*Obelisk_000003*" in Table S2) illustrated as in Figure 1C, chords illustrate predicted base pairs (base pair probabilities gray, 0.1, to red, 1.0) with the addition of annotations for primer sites used in characterization (outer track, provided in the key resources table) and of coverage plots of data shown in (B) (inner track).

(B) Distribution of total RNA sequence k-mers matching the obelisk-*S.s* in the ObP1 RNA sequence data (see STAR Methods): under these experimental and analytical conditions, obelisk-*S.s* appears to be predominantly antisense (relative to Oblin-1), with 93.6% ± 1.2 % k-mers mapping to the antisense strand. Links to analysis results for rRNA-depleted and RNaseR-treated data are available in the key resources table.

which the colonies were isolated or that occur during strain passage.

To derive a full-length obelisk-*S.s* RNA genome from our monoculture materials, we assembled an additional reference from longer read sequencing (nanopore consensus sequences from a set of PCR reactions). This assembly yielded an 1,137-base circular contig identical to *Obelisk_000003* (Table S2), with the assembly derived from two differently, circularly permuted "divergent RT-PCR" reactions (see STAR Methods, Figure S3C). The circular assembly is consistent with obelisk-*S.s* existing as a multi-repeat concatemer and/or as a circular RNA (features used by VNom for obelisk discovery). The combined PCR results and the apparent RNaseR resistance of obelisk-*S.s* are consistent with a circular obelisk-*S.s* genome topology, although we note that both methods could have been confounded by concatemeric sequences or by high degrees of RNA secondary structure. Over 99.9% of mapped total RNA bases against this consensus contig retained the reference sequence (see STAR Methods), indicating a relative stability of the obelisk-*S.s* genome under these experimental conditions.

All short read sequencing datasets were assayed for any detectable RNA-dependent RNA polymerase (RdRP) homologous reads (see STAR Methods); no such reads were found in either the DNA or RNA data, suggesting that obelisk-*S.s* might co-opt cellular replicative mechanisms, perhaps using a strategy similar to delta-like and viroid-like agents.

We see representation of both strands and all regions in the RNA-seq data (see STAR Methods; Figure 5 and linked in the key resources table). The antisense strand (non-coding) sequences are remarkably not only present but also comprises the vast majority of observed reads (evident with total RNA, rRNA-depleted, and RNaseR-treated protocols), with antisense reads comprising 93.6% ± 1.2% of total RNA, 98% ± 0.3% of rRNA-depleted, and 98.6% ± 0.3% of RNaseR-treated mapped reads (Figure 5B and linked in the key resources table); similarly, such an antisense bias has also been observed in HDV.[31]

Structural prediction indicates a novel globular domain characteristic of Oblin-1 proteins

Due to the lack of obvious Oblin-1 and -2 protein sequence homology in existing, non-obelisk databases, we performed protein tertiary structure predictions in an attempt to identify both shared predicted structural elements and homology through tertiary structure similarity searches. Owing to Oblin-1 and -2's previously unrecorded nature and apparent monophyly, we avoided automated multiple sequence alignment construction during conventional tertiary structure prediction using ColabFold (an implementation of AlphaFold2)[32,33] and instead opted for custom RDVA alignments (see protein tertiary structure prediction). This yielded a folding prediction of Oblin-1 (mean per-residue confidence estimate, μ-pLDDT ± standard deviation, of 83.8 ± 13.4, where 70–90 pLDDT values are "*a generally good backbone prediction*"[34] and higher is better) with a more confidently predicted N-terminal "globule" (μ-pLDDT of 90.1 ± 8.7, Figure 2A). ColabFold was not able to confidently place the two flanking backbones between the first and last predicted α helices and the rest of the Oblin-1 sequence (Figure S4A), and owing to heterogeneity in the last α helix's placement across predictions (see data and code availability), the globule was further focused on. The globule was predicted to form a consistent fold

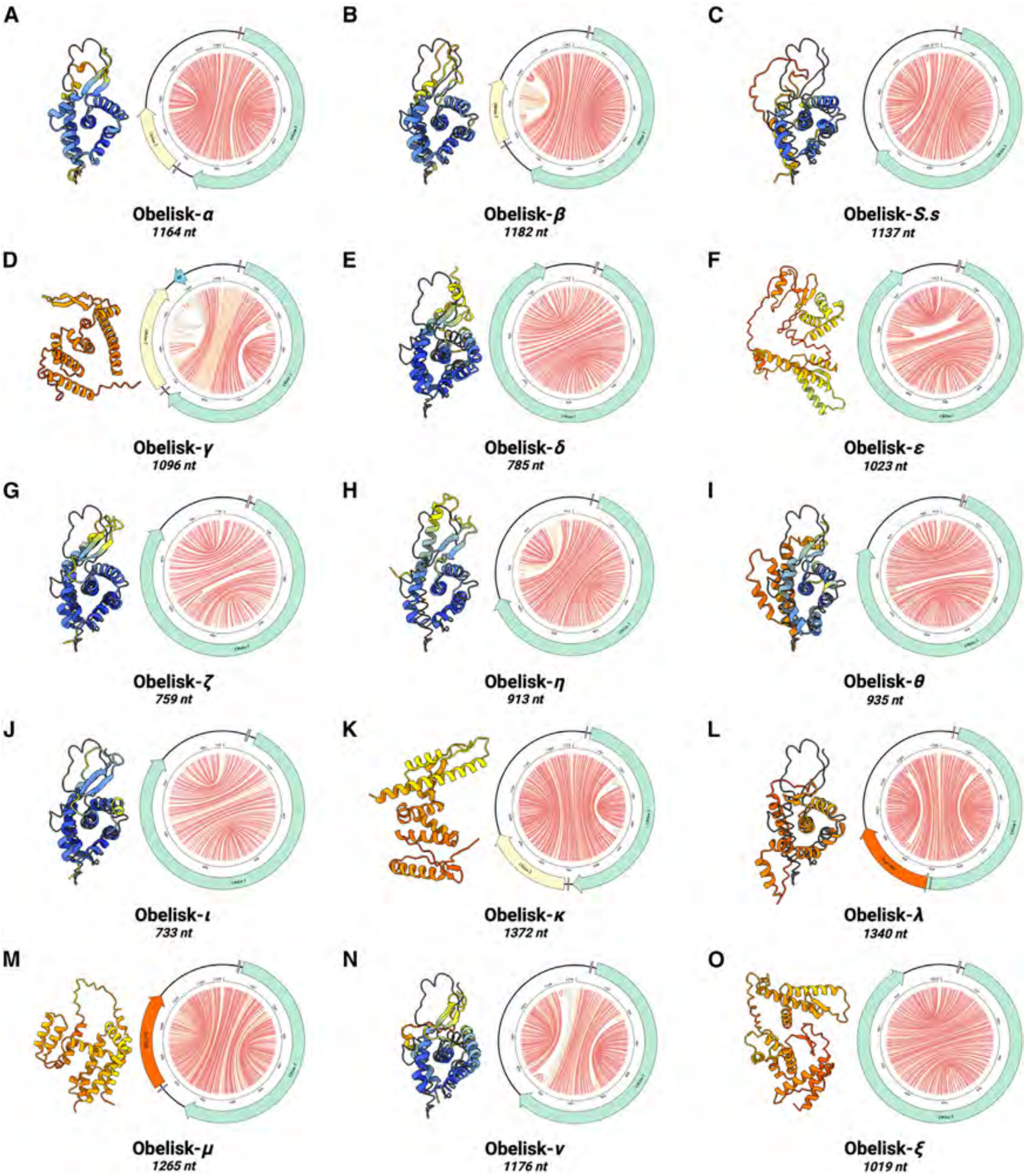


Figure 6. Obelisks form a self-consistent set

Predicted obelisk secondary structures depicted as jupiter plots where chords represent predicted base pairs (colored by base pair probability from 0, gray, to 1, red, see STAR Methods) with predicted ORFs (preceded by predicted Shine-Dalgarno sequences, purple) depicted: Oblin-1 (green), Oblin-2 (yellow, based on BLASTp hits against the Oblin-2 consensus), and "2ndORF" (orange). Obelisk-ɣ's suggested CRISPR spacer match illustrated in light blue. ColabFold predictions of Oblin-1 tertiary globule structures built with *ad hoc* multiple sequence alignment (MSA) construction (colored cartoons) superimposed over the RDVA-derived MSA prediction for obelisk-α where possible (black line, Figure 2A, see STAR Methods). Prediction confidence (pLDDT) shown as cartoon coloring as in Figure S2. Greek letter key: α: alpha, β: beta, ɣ: gamma, δ: delta, ε: epsilon, ζ: zeta, η: eta, θ: theta, ι: iota, κ: kappa, λ: lambda, μ: mu, ν: nu, and ξ: xi.
See also Figure S6.

(Figure 6): a three α-helical bundle (two smaller α helices coaxially aligned along the larger α helix) partially wrapping over a semi-orthogonal four α-helical bundle—all bookended with a two-strand β sheet "clasp" (Figure 2B). Interestingly, no confident fold was predicted for the largest conserved region in Oblin-1 (Figure S4A), termed *domain-A*, (Figures 2A and 2B—magenta). Suggestive of an anion binding function, this 18-aa stretch is enriched for positively charged residues (arginine, histidine, and lysine) with the obelisk-ɑ *domain-A* containing five arginines, three histidines, and a lysine residue (50% of *domain-A*;

Figure 2B; see protein homology bioinformatics). Additionally, a "GYxDxG" motif appears prominently in *domain-A*. If Oblin-1 represents a new class of RNA-binding protein, ColabFold may miss the fold of *domain-A* due to the absence of its client RNA ligand, absence of other key interactors, posttranslational modification, or lack of examples in previous databases.

Oblin-2 modeling suggests a leucine zipper α helix

Oblin-2 modeling with ColabFold resulted in a high-confidence prediction (μ-pLDDT of 97.1 ± 4.6) that this protein forms a solitary α helix (Figure 2C). In the RDVA consensus, the Oblin-2 α helix consists of a leucine zipper motif (see protein conservation and phylogenetics), with the characteristic "i+7" spacing of leucines at the "a" position; another hydrophobic residue (leucine or isoleucine) with i+7 spacing at the "d" position; and complementary charged residues (glutamic acid and lysine or arginine) at the "e" and "g" positions, respectively[35] (Figure S4C). Based on μ-pLDDT, ColabFold predicts that Oblin-2 might be able to homo-multimerize as a dimer (μ-pLDDT of 94.6 ± 0.6) or a trimer (μ-pLDDT of 93.6 ± 0.6) with a coiled-coil forming with two or three inter-helical salt bridges per helical pair, respectively (Figures 2D and S4D). Although conceivable, a higher order Oblin-2 homotetramer is less well supported by ColabFold (μ-pLDDT of 65.3 ± 7.9, Figure S4D). Leucine zippers typically act as multimerization motifs that bring together other protein domains such as the DNA-binding basic leucine zipper domain (bZIP).[36] Oblin-2 does not appear to include any other sequence motifs (e.g., a non-zipper poly-basic patch similar to bZIP proteins), suggesting potential function as a homo-multimer or as a binding partner to other host leucine zippers. Oblin-2 does not appear to be the only secondary ORF present in obelisks, with 788 of 1,744 stringent obelisks harboring at least two ORFs (Table S1), suggesting a range of potential accessory obelisk functions.

A subtype of obelisks bears ribozyme signatures of a viroid-like replication mechanism

Viroids of the family *Avsunviroidae* and HDV code for self-cleaving ribozymes used in their respective replicative cycles[7,8] (Figures S1A and S1C), and previous bioinformatic studies have found self-cleaving ribozymes in candidate viroid-like genomes.[12–14] Upon querying for hammerhead type-III ribozyme-coding obelisks, we identified 23 initial hits and noticed that these ribozymes slightly differed from the reference covariance model (Rfam: RF00008). Therefore, we constructed an "obelisk-variant hammerhead type-III" ribozyme (ObV-HHR3) covariance model (Figure S5B, see RNA homology bioinformatics), yielding 339 total obelisks containing HHRs in the RDVA set with stringent similarity (35 clustered at 80% identity in Table S2, "ObV-HHR3" column). These "HHR-obelisks" are similarly rod-shaped, ∼1 kb in length, and code for diverged Oblin-1 proteins (20.6% identity and 31.7% similarity to the obelisk-ɑ Oblin-1) that are similarly largely self-complementary (Figure S5A), do not code for Oblin-2, but do include an unrelated "smaller ORF." Additionally, some obelisks appear to include a bidirectional pair of ObV-HHR3 ribozymes (Figure S5A), a feature used by *Avsunviroidae*, HDV, and ambiviruses for their rolling-circle replicative cycles. For the subset of ObV-HHR3 ribozyme-containing obelisks, ColabFold predicts a globule fold (total μ-pLDDT of 76.8 ± 20.1, and globule μ-pLDDT of 88.3 ± 8.6; Figures S5C/S5E), that is similar to the non-HHR Oblin-1 model but with additional specific tertiary structure features. Namely, the β sheet clasp region is expanded by an extra sheet as well as some small α helices, and the C-terminal α helix is predicted to be shorter (Figure S5D). Additionally, the *domain-A* region appears to be diverged in the ObV-HHR3 class, yet still exhibits the positive residue skew as well as the GYxDxG protein motif also found in non-HHR-obelisks (Figure S5D). These subset-specific features, and the correlation with HHR co-occurrence, suggest that at least HRR-obelisks may replicate via a viroid-like mechanism, with Oblin-1 and/or Oblin-2 as potential co-factors.

An Oblin-1 dendrogram provides evidence for in-family evolution and places ribozyme-bearing obelisks in a distinct clade

Following the RDVA search, an initial obelisk dendrogram spanning diverse sampling sites (Figure 3B) from around the globe (Figure 3C) was constructed using full-length Oblin-1 sequences (see protein homology bioinformatics and protein conservation and phylogenetics; Figure 3A). This dendrogram was sufficient to partially explain the distribution of ObV-HHR3-bearing obelisks, which segregate into one clade (Figure 3A, orange circles and orange shading), suggesting the possibility of an evolutionary link between obelisk genome processing and Oblin-1. Additionally, this dendrogram indicates that the human microbiome-associated obelisks (Figure 3A, magenta circles and Greek letters) are widely distributed, implying a complex intersection between human and obelisk biology. Lastly, the occurrence of exactly two predicted ORFs (such as Oblin-1 and Oblin-2) appears to be biased toward the non-ObV-HHR3-bearing obelisks' clades (Figure 3A, black triangles); however, a clear trend cannot be determined.

Absence of captured obelisk matches among available CRISPR spacer datasets

Searches through CRISPR spacer databases offer an opportunity to deduce past associations between specific mobile genetic elements and potential cellular prokaryotic hosts.[4,14] We applied a conservative k-mer matching approach (see obelisk spacer analysis) to gauge the extent to which obelisks appear to be sampled by the CRISPR spacer arrays, using a dataset of 29,857,318 spacers predicted by the Joint Genome Institute's (JGI's) IMG/M database.[37] Ultimately, no convincing candidates for obelisk capture by CRISPR loci appeared in this analysis (one apparent match was observed, but further characterization of the assembly containing this obelisk-"gamma," see the key resources table; *Obelisk_000004* in Table S2 indicated that this region of the assembly was likely spurious [see obelisk spacer analysis]).

Obelisks are prevalent in tested human microbiomes

Next, we sought to roughly estimate the prevalence of obelisks in human gut and oral microbiomes by searching five datasets (three gut and two oral; Table S1) spanning 472 human donors primarily from North America (due to representational bias on the SRA). Of these, 25 donors (5.3%) were identified as positive for obelisk-ɑ, obelisk-β, or obelisk-*S.s*; and a further 21 donors

(4.4%) appeared to be positive for novel obelisks (Figure S6), for a total of 9.7% obelisk positivity (see surveying for obelisks in human data). Upon separating by microbiome source, 6.6% (29 donors) of gut microbiome, and 53% (17 donors) of oral microbiome samples contained obelisks. These data therefore implicate the oral microbiome as a reservoir of obelisks with more than half of the donors positive for such elements, although this could also be explained by an idiosyncrasy of the major oral dataset (Belstrøm and Constancias et al.[38]) that contributes to this count. Ultimately, 11 new, distinct, full-length obelisks were identified upon examining the obelisk-positive donors without obelisk-ɑ, obelisk-β, or obelisk-*S.s* homology—which we name "delta" through "xi" (see surveying for obelisks in human data; Figure 6; *Obelisk_000005* through *Obelisk_000015* in Table S2). Obelisks "alpha," "beta," "epsilon (ε)," "zeta," and "eta" were restricted to gut microbiome samples (obelisk-ε was found in one oral sample), whereas obelisks "*S. sanguinis*," and "theta" through "xi" were primarily orally restricted (obelisk-*S.s* was found in one stool sample)—indicating an anatomical specificity of obelisks despite the oral-gastric connection. We note that these studies used different library preparation strategies (Table S1) and show varying obelisk sensitivity as a function of read depth (Figure S6, scale bars), consistent with the technical expectation that not all metatranscriptomic sequencing workflows would be equally good at detecting obelisks. This raises the question of a potential technological blind spot to these (and similar) elements with some protocols. In any case, the observed values certainly represent a lower bound, and these data point to obelisks being a non-negligible member of the tested adult oral and gut microbiomes. By their public nature, these datasets lack complete donor medical metadata; this lack and the relatively small sample size leave the investigation of correlations between obelisk prevalence (and abundance) and the health of human hosts for future studies.

DISCUSSION

The RNA viroid/sub-viral component of the biosphere is beginning to be estimated,[12–14] but sequence-matching-based strategies, although potent for RNA viral discovery,[1–5] are blind to previously unnoticed classes of agents. Here, we applied a generic molecular-feature-focused search strategy (VNom) to identify viroid-like RNAs in public RNA-seq datasets. We ultimately focused on a large monophyletic group of viroid-like elements that we term "obelisks." A single clear obelisk-host pairing (*S. Sanguinis* SK36-obelisk-*S.s*) indicates that obelisks can be a component of bacterial cells; while we do not know the hosts of other obelisks, it is reasonable to assume that at least a fraction may be present in bacteria.

Obelisk genomes are predicted to fold into conspicuous rod-like secondary structures, with a largely self-complementary and conserved Oblin-1 ORF that accounts for at least half of the circular sequence assembly. Oblin-1 itself is predicted to fold into a stereotyped globule tertiary structure (Figure 6) with its most conserved motif, *domain-A*, lacking a confident tertiary structure prediction (Figure 2A). Furthermore, the presence of a subset of hammerhead ribozyme-bearing obelisks with distinct Oblin-1 features (Figure S5) and an interplay with Oblin-1 evolution (Figure 3A) suggests an Oblin-ribozyme functional relationship, perhaps in viroid-like rolling-circle or rolling-hairpin[39] replication. We note that conservative ribozyme detection thresholds were used in this work, leaving open the possibility that a larger diversity of ribozymes could be present in the obelisks, including potentially novel self-cleaving ribozymes. Thus, the exact interplay between obelisk genome processing (via ribozymes) and Oblins-1, -2, and others is currently unknown.

Obelisks appear to be globally distributed (Figure 3C) and are a constituent member of the human oral and gut microbiomes, occurring in ∼10% of human donors in five assayed human metatranscriptomic studies (Table S1; Figure S6). Interestingly, we note one oral microbiome study showing a ∼50% obelisk prevalence (Figure S6D). We also note that observed obelisk prevalence is likely to be quite dependent on the population in question, sampling scheme, type and depth of sequencing, and other features. Lastly, a specific obelisk strain, obelisk-ɑ, appears to persist and speciate within microbiomes of human donors (Figures 1D and 1E). The prevalence and apparent novelty of these elements implies more is yet to be learned about their interplay with microbial and human life.

Constructing a full obelisk dendrogram with explanatory power proved difficult (Figure 3A). This is likely due to several factors including the fact that obelisks appear to be under selection for a highly base-paired genomic coding region that must also code for stereotyped protein fold (Figure 6). Classical phylogenetic tools cannot account for evolutionary signals from non-position-independent RNA secondary structure constraints,[40] consistent with the complexities in estimating trees from such families.[41] Further, recent advances in protein tertiary structure prediction may now allow for protein structure-based phylogenetic reconstruction that may be tolerant of greater sequence divergence.[42,43] Therefore, definitive phylogenetic work on obelisks might benefit from future tools that incorporate both evolutionary signals from RNA secondary structure conservation and from structural alignment of predicted Oblin-1 globule tertiary structures, as well as employing maximum-likelihood methods. Lastly, the serratus approach taken for large-scale obelisk discovery was run using homology models built from sequences initially homologous to obelisk-ɑ (see protein homology bioinformatics) and thresholds derived from RNA viral discovery campaigns;[2] so while a mammalian sample-origin bias is seen (Figure 3B), this could be explained by an auto-correlation based on the mammalian origin of obelisk-ɑ, potentially confounded by the choice of RNA viral discovery threshold. Due to this aforementioned bias, as well as a lack of a systematized method for discovery, it should be noted that the breadth of obelisk diversity reported in this study could be an underestimate. Further, while we focused on obelisks, their prevalence and diversity suggest that similar, unrelated viroid-like RNAs are likely widespread and waiting to be discovered in public sequencing data.

The observation that distinct subsets of obelisks appear to occur in human oral versus gastric sites, an anatomic specificity that mirrors the site specificity of human microbiomes[44] (Figure S6), supports the notion that obelisks might include colonists of said human microbiomes. Building on this, donor-specific factors such as diet or lifestyles therefore likely play a role in obelisk (re-)colonization and retention. Further, given that *S. sanguinis* is not only a commensal of the healthy human oral microbiome[45] but also a causative agent of bacterial endocarditis,[46] a study of the implied *S. sanguinis*-obelisk-*S.s* relationship might begin to reveal the relevance of obelisks to the natural oral niche and potentially to human health, as well as offer a tractable model system to study obelisk molecular biology. As an initial laboratory characterization, we used this *S. sanguinis* SK36-obelisk-*S.s* system for *in vitro* culture and RNA-seq. The produced data support a model in which the obelisk-*S.s* is an RNA-only element, with representation of both obelisk-*S.s* strands. We saw no evidence of an RdRP for replication and no evidence for an essential role in *S. sanguinis* SK36 host fitness (under replete growth conditions; Figures 4A–4C/4E).

From clustering analysis, obelisks appear to be a broad family, with obelisk-S.s clustering with some but distinctly from other human-associated obelisks (Figure 3A), Thus, we might expect both shared and unique aspects to the functions, host interactions, and distributions of each obelisk member and group. However, with 15 exemplar obelisk sequences (Figure 6), an "obelisk blueprint" arises: with an ~730- to 1,340-nt apparently circular RNA; an extended rod-like, largely symmetrical predicted RNA secondary structure (Figure 6, "jupiter" plots); an Oblin-1 homolog whose RNA sequence is largely self-complementary (which ColabFold predicts occupies a "globule-like" tertiary structure in 9 of 15 examples; Figure 6, tertiary structures); and an occasionally present second, smaller protein (e.g., Oblin-2). The observation of multiple, diverse (see STAR Methods) "solutions" to the constrained problem of coding for Oblin-1, while also coding for a highly base-paired RNA genome, indicates that obelisks are *bona fide* biological agents with a shared evolutionary origin.

Many questions arise about the obelisks. Does their transmission involve a separate, more complex, infectious agent (like HDV)? Do they primarily spread via virus-like particles or cytoplasmically, like viroids? Are obelisks plasmid-like in that they can coexist and, in some cases, contribute to host adaptability and fitness? Like viroids and HDV, do obelisks replicate via rolling-circle replication using a co-opted host RNA polymerase? What roles do the apparently circular obelisk genome topology and the evidently conserved obelisk genomic secondary structure play in the obelisk life cycle? Is Oblin-1 an RNA-binding protein, and how does *domain-A* factor into its function? Does Oblin-2 act as a competitive inhibitor of host leucine zippers, as a multimerizing element, and/or can it interact with Oblin-1? How do obelisks that lack Oblin-2 complement its function(s)? What role do the obelisk-specific self-cleaving ribozymes play, and how do they interact with the Oblin proteins? How do obelisks affect their host, and are they largely a deleterious or beneficial element to harbor? And what impact, if any, does harboring an obelisk have on "meta"-host physiology, and is obelisk positivity predictive of human health states?

Lastly, obelisks do not closely resemble any existing mobile genetic elements, raising the question of their appropriate designation. Throughout this work, obelisks have been referred to as viroid-like, drawing comparisons to viroids and HDV. However, viroids are in part defined by their non-coding nature,[6,7] and HDV-like elements are defined by homology to the large L-HDAg (and in the case of HDV, human tropism and a satellite relationship to hepatitis B virus).[8] By virtue of their predicted coding capacity, which does not resemble L-HDAg, obelisks are then neither strictly viroids nor delta-like elements. The predicted self-complementarity of the Oblin-1 further deviates from L-HDAg, likely imposing a set of unique evolutionary constraints (protein tertiary structure in addition to RNA secondary structure), which is not experienced by viroids and HDV-like elements. We therefore propose these proteins be referred to as Oblins. Viruses are already ill-defined, with sub-viral agents (such as viroids and HDV) being defined within the then more nebulous "perivirosphere"[47], but part of "sub-virality" is the implication of virus-like behavior, either in transmission (e.g., via virions), in host impact (e.g., a pathology), or in replication (e.g., a co-opting viral replication machinery). Currently, it is not possible to assign transmission mode, host impact, or replication mode of obelisks, suggesting that these elements might not even be "viral" in nature and might more closely resemble "RNA plasmids." As such, we propose that the term "obelisk" be used to refer to these RNA-only agents as they are distinct from other sub-viral satellites,[48] viroids, and HDV.

Limitations of the study

(1) Despite the unambiguous identification of one obelisk-bacterium pair (*S. sanguinis* SK36 and obelisk-*S.s*), we do not know if all obelisks reside in bacteria—and indeed, there may be numerous biological niches (bacterial or otherwise) that could harbor obelisks.

(2) Despite not finding any evidence of obelisk-CRISPR spacer interactions, we do not know if obelisks are ever surveilled by CRISPR (or other phage defense) systems.

(3) Despite the large family of obelisks identified in this exploration of metagenomic sequencing data, we do not know if this family encompasses all previously undefined viroid-like elements, and indeed we expect that there could be many more such families, either distantly related or entirely unrelated.

(4) Similarly, despite being able to detect obelisks over a wide range of experimental designs, we do not know how conventional RNA-seq approaches are biased for or against obelisks and related elements.

(5) Despite inferring phylogenetic estimations for the large obelisk family, we do not confidently know how obelisks are interrelated. This is in part due to how obelisk genome topology violates the site independence assumption of typical phylogenetic tools. Additionally, while clustering of obelisks is seen, the method used for dendrogram construction does not provide measurements of cluster support, meaning that the true topology of the obelisk phylogenetic tree could be different from the dendrogram presented here.

(6) Despite the consistent computational predictions of a rod-shaped RNA secondary structure, we do not know how obelisk genomes may actually fold and how these folds may vary over the obelisk life cycle.

(7) Likewise, despite the consistent computational prediction of an obelisk-specific Oblin-1 globule fold, we do not know how any of the Oblin proteins may fold in reality.

(8) Despite noting that obelisks can harbor secondary, smaller ORFs (e.g., Oblin-2), the relationships of these ORFs to Oblin-1 and obelisk phylogenetics as a whole remain unclear. We expect that future work into obelisk-optimized ORF prediction tools will begin to address these questions.

(9) Despite their consistent circular assemblies, we do not know if obelisks ever exist as covalently closed circles; indeed, we might expect that other topological forms may be present and participate in obelisk replication.

(10) Despite the low apparent DNA divergence between ObN1 and ObP1, the precise nature of the interplay between the SK36 host genome and obelisk-*S.s* as it relates to SK36 fitness has yet to be determined.

(11) Despite the lack of observed phenotype from obelisk-*S.s* loss during growth in rich media in the lab, we do not know if obelisks are dispensable to their hosts and "meta"--hosts in their native environments, and indeed, we would expect such roles/consequences to emerge as these elements are further studied.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ivan N. Zheludev (zheludev@stanford.edu).

Materials availability

The parent SK36 strain is commercially available (see key resources table). Please contact the lead contact for the ObP1 and ObN1 derivatives described herein.

Data and code availability

- Raw, DNA, and RNA short read sequencing data have been deposited at the Sequence Read Archive (SRA). See the key resources table for accession number.
- Raw, DNA amplicon long read sequencing data are publicly available at the Stanford Digital Repository. See the key resources table for accession DOI.
- Raw, A_{600} kinetic growth curve data are publicly available at the Stanford Digital Repository. See the key resources table for accession DOI.
- This paper analyzes existing, publicly available SRA data, the details of which are publicly available at the Stanford Digital Repository. See the key resources table for accession DOI.
- All data, analytical code, and tabular summaries are available at the Stanford Digital Repository. See the key resources table for accession number.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work is dedicated to the memory of mentor and friend, Paul Berg.

I.N.Z. would like to thank Dr. K.L. Artiles, Dr. M.J. McCoy, Dr. M. Shoura, S.U. Enam, E. Greenwald, S.M. Bowden, Dr. T. Kitten, Dr. D. Dodd, Dr. R. Maddamsetti, Dr. B. Navarro, Dr. J.F. Bazan, as well as the whole Fire laboratory for their fruitful discussion, advice, and reagents.

This work was funded by the Stanford Graduate Fellowship (I.N.Z.), University of Valencia Margarita Salas Fellowship MS21-067 (M.J.L.-G.), Generalitat Valenciana Grant PROMETEO CIPROM/2022/21 (M.d.I.P.), Ministry of Economics and Competitivity of Spain-FEDER grant PID2020-116008GB-I00 (M.d.I.P.), and Canadian Institutes of Health Project grant PTJ-496709 (A.B.). Computing resources were provided by the University of Toronto Cloud Research Lab at The Donnelly, powered by AWS (A.B.). It was also funded by the NIH grant R01AI148623 (National Institute of Allergy and Infectious Diseases, NIAID) (A.S.B.), NIH grant R01AI143757 (NIAID) (A.S.B.), Convergence Grant 3.1416 (Stand up 2 Cancer) (A.S.B.), Paul Allen Distinguished Investigator Award (A.S.B.), and NIH grant R35GM130366 (National Institute of General Medical Sciences) (A.Z.F.).

We are grateful to the greater biology community for open data sharing. Some figures were arranged in BioRender.

AUTHOR CONTRIBUTIONS

Conceptualization, I.N.Z. and A.Z.F.; methodology, I.N.Z. and A.Z.F.; software, I.N.Z., R.C.E., and A.Z.F.; formal analysis, I.N.Z., M.J.L.-G., M.d.I.P., and A.Z.F.; investigation, I.N.Z.; resources, A.B., A.S.B., and A.Z.F.; data curation, I.N.Z.; writing—original draft, I.N.Z. and A.Z.F.; writing—review & editing, I.N.Z., R.C.E., M.J.L.-G., M.d.I.P., A.B., A.S.B., and A.Z.F.; visualization, I.N.Z.; supervision, A.Z.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENT MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Microbial strains
- METHOD DETAILS
  - ○ VNom
  - ○ Initial Obelisk identification
  - ○ Taxonomic classification
  - ○ Obelisk homologue detection in other public data
  - ○ Serratus
  - ○ Protein homology bioinformatics
  - ○ Protein tertiary structure prediction
  - ○ Protein conservation and phylogenetics
  - ○ ScanRabbit
  - ○ Obelisk spacer analysis
  - ○ Identity and similarity measurements
  - ○ RNA homology bioinformatics
  - ○ *Streptococcus sanguinis* bioinformatics
  - ○ Surveying for Obelisks in human data
  - ○ *S. sanguinis* SK36 culturing and RT-PCR screening
  - ○ Assays on SK36 ObP1 and ObN1 nucleic acids
  - ○ SK36 ObN1 and ObP1 sequencing analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.09.033.

Received: January 27, 2024
Revised: July 3, 2024
Accepted: September 18, 2024
Published: October 22, 2024

REFERENCES

1. Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., et al. (2016). Redefining the invertebrate RNA virosphere. Nature *540*, 539–543. https://doi.org/10.1038/nature20167.

2. Edgar, R.C., Taylor, B., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., et al. (2022). Petabase-scale sequence alignment catalyses viral discovery. Nature *602*, 142–147. https://doi.org/10.1038/s41586-021-04332-2.

3. Zayed, A.A., Wainaina, J.M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M., Tian, F., Pratama, A.A., Bolduc, B., Zablocki, O., et al. (2022). Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. Science *376*, 156–162. https://doi.org/10.1126/science.abm5847.

4. Neri, U., Wolf, Y.I., Roux, S., Camargo, A.P., Lee, B., Kazlauskas, D., Chen, I.M., Ivanova, N., Zeigler Allen, L., Paez-Espino, D., et al. (2022). Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell *185*, 4023–4037.e18. https://doi.org/10.1016/j.cell.2022.08.023.

5. Olendraite, I., Brown, K., and Firth, A.E. (2023). Identification of RNA Virus–Derived RdRp Sequences in Publicly Available Transcriptomic Data Sets. Mol. Biol. Evol. *40*, msad060.

6. Di Serio, F., Owens, R.A., Li, S.-F., Matoušek, J., Pallás, V., Randles, J.W., Sano, T., Verhoeven, J.T.J., Vidalakis, G., Flores, R., et al. (2021). ICTV Virus Taxonomy Profile: Pospiviroidae. J. Gen. Virol. *102*, 001543. https://doi.org/10.1099/jgv.0.001543.

7. Di Serio, F., Li, S.-F., Matoušek, J., Owens, R.A., Pallás, V., Randles, J.W., Sano, T., Verhoeven, J.T.J., Vidalakis, G., Flores, R., et al. (2018). ICTV Virus Taxonomy Profile: Avsunviroidae. J. Gen. Virol. *99*, 611–612. https://doi.org/10.1099/jgv.0.001045.

8. Magnius, L., Taylor, J., Mason, W.S., Sureau, C., Dény, P., and Norder, H.; Ictv Report Consortium (2018). ICTV Virus Taxonomy Profile: Deltavirus. J. Gen. Virol. *99*, 1565–1566. https://doi.org/10.1099/jgv.0.001150.

9. Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften *58*, 465–523. https://doi.org/10.1007/BF00623322.

10. Gago, S., Elena, S.F., Flores, R., and Sanjuán, R. (2009). Extremely High Mutation Rate of a Hammerhead Viroid. Science *323*, 1308. https://doi.org/10.1126/science.1169202.

11. Bergner, L.M., Orton, R.J., Broos, A., Tello, C., Becker, D.J., Carrera, J.E., Patel, A.H., Biek, R., and Streicker, D.G. (2021). Diversification of mammalian deltaviruses by host shifting. Proc. Natl. Acad. Sci. USA *118*, e2019907118. https://doi.org/10.1073/pnas.2019907118.

12. Weinberg, C.E., Olzog, V.J., Eckert, I., and Weinberg, Z. (2021). Identification of over 200-fold more hairpin ribozymes than previously known in diverse circular RNAs. Nucleic Acids Res. *49*, 6375–6388. https://doi.org/10.1093/nar/gkab454.

13. Forgia, M., Navarro, B., Daghino, S., Cervera, A., Gisel, A., Perotto, S., Aghayeva, D.N., Akinyuwa, M.F., Gobbi, E., Zheludev, I.N., et al. (2023). Hybrids of RNA viruses and viroid-like elements replicate in fungi. Nat. Commun. *14*, 2591. https://doi.org/10.1038/s41467-023-38301-2.

14. Lee, B.D., Neri, U., Roux, S., Wolf, Y.I., Camargo, A.P., Krupovic, M., RNA Virus Discovery Consortium, Simmonds, P., Kyrpides, N., Gophna, U., et al. (2023). Mining metatranscriptomes reveals a vast world of viroid-like circular RNAs. Cell *186*, 646–661.e4. https://doi.org/10.1016/j.cell.2022.12.039.

15. Tisza, M.J., and Buck, C.B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. Proc. Natl. Acad. Sci. USA *118*, e2023202118. https://doi.org/10.1073/pnas.2023202118.

16. Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. *5*, 4498.

17. Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D., and Lawley, T.D. (2021). Massive expansion of human gut bacteriophage diversity. Cell *184*, 1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.

18. Dahlman, S., Avellaneda-Franco, L., Kett, C., Subedi, D., Young, R.B., Gould, J.A., Rutten, E.L., Gulliver, E.L., Turkington, C.J.R., Nezam-Abadi, N., et al. (2023). Temperate gut phages are prevalent, diverse, and predominantly inactive. Preprint at bioRxiv. https://doi.org/10.1101/2023.08.17.553642.

19. Fogarty, E.C., Schechter, M.S., Lolans, K., Sheahan, M.L., Veseli, I., Moore, R.M., Kiefl, E., Moody, T., Rice, P.A., Yu, M.K., et al. (2024). A cryptic plasmid is among the most numerous genetic elements in the human gut. Cell *187*, 1206–1222.e16. https://doi.org/10.1016/j.cell.2024.01.039.

20. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature *569*, 655–662. https://doi.org/10.1038/s41586-019-1237-9.

21. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Res. *50*, D20–D26. https://doi.org/10.1093/nar/gkab1112.

22. Zhang, H., Zhang, L., Lin, A., Xu, C., Li, Z., Liu, K., Liu, B., Ma, X., Zhao, F., Jiang, H., et al. (2023). Algorithm for optimized mRNA design improves stability and immunogenicity. Nature *621*, 396–403. https://doi.org/10.1038/s41586-023-06127-z.

23. Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. *48*, D265–D268. https://doi.org/10.1093/nar/gkz991.

24. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

25. Shiryev, S.A., and Agarwala, R. (2024). Indexing and searching petabase-scale nucleotide resources. Nat. Methods *21*, 994–1002. https://doi.org/10.1038/s41592-024-02280-z.

26. Lin, V., Ravichandran, G., Ha, K., Kinoshita, B.P., and Babaian, A. (2022). RNA Deep Virome Assemblage. GitHub. https://github.com/ababaian/serratus/wiki/RNA-Deep-Virome-Assemblage.

27. Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. Front. Microbiol. *8*, 2224. https://doi.org/10.3389/fmicb.2017.02224.

28. Coenen, A.R., and Weitz, J.S. (2018). Limitations of Correlation-Based Inference in Complex Virus-Microbe Communities. mSystems *3*, e00084–18. https://doi.org/10.1128/mSystems.00084-18.

29. Hirano, H., and Takemoto, K. (2019). Difficulty in inferring microbial community structure based on co-occurrence network approaches. BMC Bioinformatics *20*, 329. https://doi.org/10.1186/s12859-019-2915-1.

30. Caufield, P.W., Dasanayake, A.P., Li, Y., Pan, Y., Hsu, J., and Hardin, J.M. (2000). Natural history of Streptococcus sanguinis in the oral cavity of infants: evidence for a discrete window of infectivity. Infect. Immun. *68*, 4018–4023. https://doi.org/10.1128/IAI.68.7.4018-4023.2000.

31. Chen, P.J., Kalpana, G., Goldberg, J., Mason, W., Werner, B., Gerin, J., and Taylor, J. (1986). Structure and replication of the genome of the hepatitis delta virus. Proc. Natl. Acad. Sci. USA *83*, 8774–8778. https://doi.org/10.1073/pnas.83.22.8774.

32. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

33. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682. https://doi.org/10.1038/s41592-022-01488-1.

34. DeepMind; Embl-Ebi (2022). AlphaFold Protein Structure Database: Frequently Asked Questions. https://alphafold.ebi.ac.uk/faq.

35. O'Shea, E.K., Lumb, K.J., and Kim, P.S. (1993). Peptide "Velcro": design of a heterodimeric coiled coil. Curr. Biol. *3*, 658–667. https://doi.org/10.1016/0960-9822(93)90063-t.

36. Sinden, R.R. (1994). Chapter 8 - DNA–Protein Interactions. In DNA Structure and Function, R.R. Sinden, ed. (Academic Press), pp. 287–325. https://doi.org/10.1016/B978-0-08-057173-7.50013-4.

37. Chen, I.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S., Varghese, N., Seshadri, R., et al. (2021). The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Res. *49*, D751–D763. https://doi.org/10.1093/nar/gkaa939.

38. Belstrøm, D., Constancias, F., Drautz-Moses, D.I., Schuster, S.C., Veleba, M., Mahé, F., and Givskov, M. (2021). Periodontitis associates with species-specific gene expression of the oral microbiota. npj Biofilms Microbiomes *7*, 76. https://doi.org/10.1038/s41522-021-00247-y.

39. Tattersall, P., and Ward, D.C. (1976). Rolling hairpin model for replication of parvovirus and linear chromosomal DNA. Nature *263*, 106–109. https://doi.org/10.1038/263106a0.

40. Pedersen, J.S., Forsberg, R., Meyer, I.M., and Hein, J. (2004). An evolutionary model for protein-coding regions with conserved RNA structure. Mol. Biol. Evol. *21*, 1913–1922. https://doi.org/10.1093/molbev/msh199.

41. Patiño-Galindo, J.Á., González-Candelas, F., and Pybus, O.G. (2018). The Effect of RNA Substitution Models on Viroid and RNA Virus Phylogenies. Genome Biol. Evol. *10*, 657–666. https://doi.org/10.1093/gbe/evx273.

42. Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., and Dessimoz, C. (2023). Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. Preprint at bioRxiv. https://doi.org/10.1101/2023.09.19.558401.

43. Puente-Lelievre, C., Malik, A.J., Douglas, J., Ascher, D., Baker, M., Allison, J., Poole, A.M., Lundin, D., Fullmer, M., Bouckaert, R., et al. (2023). Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. Preprint at bioRxiv. https://doi.org/10.1101/2023.12.12.571181.

44. Kennedy, M.S., and Chang, E.B. (2020). The microbiome: composition and locations. Prog. Mol. Biol. Transl. Sci. *176*, 1–42. https://doi.org/10.1016/bs.pmbts.2020.08.013.

45. Xu, P., Alves, J.M., Kitten, T., Brown, A., Chen, Z., Ozaki, L.S., Manque, P., Ge, X., Serrano, M.G., Puiu, D., et al. (2007). Genome of the Opportunistic Pathogen Streptococcus sanguinis. J. Bacteriol. *189*, 3166–3175. https://doi.org/10.1128/JB.01808-06.

46. Mylonakis, E., and Calderwood, S.B. (2001). Infective Endocarditis in Adults. N. Engl. J. Med. *345*, 1318–1330. https://doi.org/10.1056/NEJMra010082.

47. Koonin, E.V., Dolja, V.V., Krupovic, M., and Kuhn, J.H. (2021). Viruses Defined by the Position of the Virosphere within the Replicator Space. Microbiol. Mol. Biol. Rev. *85*, e0019320. https://doi.org/10.1128/MMBR.00193-20.

48. Symons, R.H. (1991). The intriguing viroids and virusoids: what is their information content and how did they evolve? Mol. Plant Microbe Interact. *4*, 111–121. https://doi.org/10.1094/mpmi-4-111.

49. Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A.D. (2019). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience *8*, giz100. https://doi.org/10.1093/gigascience/giz100.

50. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461.

51. Ayad, L.A.K., and Pissis, S.P. (2017). MARS: improving multiple circular sequence alignment using refined sequences. BMC Genomics *18*, 86. https://doi.org/10.1186/s12864-016-3477-5.

52. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

53. Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26. https://doi.org/10.1186/1748-7188-6-26.

54. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). Blast+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

55. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935. https://doi.org/10.1093/bioinformatics/btt509.

56. Rivas, E. (2020). RNA structure prediction using positive and negative evolutionary information. PLoS Comput. Biol. *16*, e1008387. https://doi.org/10.1371/journal.pcbi.1008387.

57. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

58. Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L. (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. *3*, e104. https://doi.org/10.7717/peerj-cs.104.

59. Vasimuddin, M., Misra, S., Li, H., and Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 314–324. https://doi.org/10.1109/IPDPS.2019.00041.

60. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. Preprint at arXiv. https://doi.org/10.48550/arXiv.1207.3907.

61. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

62. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328. https://doi.org/10.1093/bioinformatics/bts606.

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

64. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

65. Edgar, R.C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat. Commun. *13*, 6968. https://doi.org/10.1038/s41467-022-34630-w.

66. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

67. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

68. Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175. https://doi.org/10.1038/nmeth.1818.

69. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. *10*, 845–858. https://doi.org/10.1038/nprot.2015.053.

70. Holm, L. (2022). Dali server: structural unification of protein families. Nucleic Acids Res. *50*, W210–W215. https://doi.org/10.1093/nar/gkac387.

71. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. *42*, 243–246. https://doi.org/10.1038/s41587-023-01773-0.

72. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., and Hochreiter, S. (2015). msa: an R package for multiple sequence alignment. Bioinformatics *31*, 3997–3999. https://doi.org/10.1093/bioinformatics/btv494.

73. Pagès, H., Aboyoun, P., Gentleman, R., DebRoy, S., Carey, V., Delhomme, N., Ernst, F., Lakshman, A., O'Neill, K., Obenchain, V., et al. (2023). Biostrings: Efficient manipulation of biological strings. Bioconductor version: Release (3.17). https://doi.org/10.18129/B9.bioc.Biostrings.

74. Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics *33*, 3645–3647. https://doi.org/10.1093/bioinformatics/btx469.

75. Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Mol. Biol. Evol. *32*, 2798–2800. https://doi.org/10.1093/molbev/msv150.

76. Tumescheit, C., Firth, A.E., and Brown, K. (2022). CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. PeerJ *10*, e12983. https://doi.org/10.7717/peerj.12983.

77. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

78. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589. https://doi.org/10.1038/nmeth.4285.

79. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. *35*, 518–522. https://doi.org/10.1093/molbev/msx281.

80. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296. https://doi.org/10.1093/nar/gkab301.

81. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

82. Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. BioTechniques *28*, 1102–1104. https://doi.org/10.2144/00286ir01.

83. Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., and Stadler, P.F. (2008). RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics *9*, 474. https://doi.org/10.1186/1471-2105-9-474.

84. Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812. https://doi.org/10.1093/bioinformatics/btu393.

85. Weinberg, Z., and Breaker, R.R. (2011). R2R - software to speed the depiction of aesthetic consensus RNA secondary structures. BMC Bioinformatics *12*, 3. https://doi.org/10.1186/1471-2105-12-3.

86. Blazanin, M. (2024). gcplyr: an R package for microbial growth curve data analysis. Bioinformatics *25*, 23. https://doi.org/10.1186/s12859-024-05817-3.

87. Deatherage, D.E., and Barrick, J.E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol. Biol. *1151*, 165–188. https://doi.org/10.1007/978-1-4939-0554-6_12.

88. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.

89. Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods *14*, 687–690. https://doi.org/10.1038/nmeth.4324.

90. Chaung, K., Baharav, T.Z., Henderson, G., Zheludev, I.N., Wang, P.L., and Salzman, J. (2023). SPLASH: A statistical, reference-free genomic algorithm unifies biological discovery. Cell *186*, 5440–5456.e26. https://doi.org/10.1016/j.cell.2023.10.028.

91. Kokot, M., Dehghannasiri, R., Baharav, T., Salzman, J., and Deorowicz, S. (2024). Scalable and unsupervised discovery from raw sequencing reads using SPLASH2. Nat Biotechnol. https://doi.org/10.1038/s41587-024-02381-2.

92. Qin, Y., Xu, T., Lin, W., Jia, Q., He, Q., Liu, K., Du, J., Chen, L., Yang, X., Du, F., et al. (2020). Reference-free and de novo Identification of Circular RNAs. Preprint at bioRxiv. https://doi.org/10.1101/2020.04.21.050617.

93. Leinonen, R., Sugawara, H., and Shumway, M.; International; Nucleotide; Sequence; Database Collaboration (2011). The Sequence Read Archive. Nucleic Acids Res. *39*, D19–D21. https://doi.org/10.1093/nar/gkq1019.

94. Pinto, Y., Chakraborty, M., Jain, N., and Bhatt, A.S. (2024). Phage-inclusive profiling of human gut microbiomes with Phanta. Nat. Biotechnol. *42*, 651–662. https://doi.org/10.1038/s41587-023-01799-4.

95. adel922 (2019). Working with VCF files and Trees. https://rpubs.com/adel922/560260.

96. Abu-Ali, G.S., Mehta, R.S., Lloyd-Price, J., Mallick, H., Branck, T., Ivey, K.L., Drew, D.A., DuLong, C., Rimm, E., Izard, J., et al. (2018). Metatranscriptome of human faecal microbial communities in a cohort of adult men. Nat. Microbiol. *3*, 356–366. https://doi.org/10.1038/s41564-017-0084-4.

97. Petersen, L.M., Bautista, E.J., Nguyen, H., Hanson, B.M., Chen, L., Lek, S.H., Sodergren, E., and Weinstock, G.M. (2017). Community characteristics of the gut microbiomes of competitive cyclists. Microbiome *5*, 98. https://doi.org/10.1186/s40168-017-0320-4.

98. Zhang, Y., Brady, A., Jones, C., Song, Y., Darton, T.C., Jones, C., Blohmke, C.J., Pollard, A.J., Magder, L.S., Fasano, A., et al. (2018). Compositional and Functional Differences in the Human Gut Microbiome Correlate with Clinical Outcome following Infection with Wild-Type Salmonella enterica Serovar Typhi. mBio *9*, e00686–18. https://doi.org/10.1128/mBio.00686-18.

99. Richter, T.K.S., Michalski, J.M., Zanetti, L., Tennant, S.M., Chen, W.H., and Rasko, D.A. (2018). Responses of the Human Gut Escherichia coli Population to Pathogen and Antibiotic Disturbances. mSystems *3*, e00047–18. https://doi.org/10.1128/mSystems.00047-18.

100. Peters, B.A., Wilson, M., Moran, U., Pavlick, A., Izsak, A., Wechter, T., Weber, J.S., Osman, I., and Ahn, J. (2019). Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. Genome Med. *11*, 61. https://doi.org/10.1186/s13073-019-0672-4.

101. Sinha, S.R., Haileselassie, Y., Nguyen, L.P., Tropini, C., Wang, M., Becker, L.S., Sim, D., Jarr, K., Spear, E.T., Singh, G., et al. (2020). Dysbiosis-Induced Secondary Bile Acid Deficiency Promotes Intestinal Inflammation. Cell Host Microbe *27*, 659–670.e5. https://doi.org/10.1016/j.chom.2020.01.021.

102. Campbell, S.J., Ashley, W., Gil-Fernandez, M., Newsome, T.M., Giallonardo, F.D., Ortiz-Baez, A.S., Mahar, J.E., Towerton, A.L., Gillings, M., Holmes, E.C., et al. (2020). Red fox viromes across an urban-rural gradient. Preprint at bioRxiv. https://doi.org/10.1101/2020.06.15.153858.

103. Maghini, D.G., Dvorak, M., Dahlen, A., Roos, M., Kuersten, S., and Bhatt, A.S. (2024). Quantifying bias introduced by sample collection in relative and absolute microbiome measurements. Nat. Biotechnol. *42*, 328–338. https://doi.org/10.1038/s41587-023-01754-3.

104. Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C.L.M., Wein, T., Varadi, M., Velankar, S., Beltrao, P., and Steinegger, M. (2023). Clustering predicted structures at the scale of the known protein universe. Nature *622*, 637–645. https://doi.org/10.1038/s41586-023-06510-w.

105. Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA *89*, 10915–10919. https://doi.org/10.1073/pnas.89.22.10915.

106. PyPy Team (2019). PyPy. https://pypy.org/.

107. Avinery, R., Kornreich, M., and Beck, R. (2019). Universal and Accessible Entropy Estimation Using a Compression Algorithm. Phys. Rev. Lett. *123*, 178102. https://doi.org/10.1103/PhysRevLett.123.178102.

108. Katz, P. (1989). ZIP (PKWare).

109. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.

110. Edgar, R.C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics *8*, 18. https://doi.org/10.1186/1471-2105-8-18.

111. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics *8*, 209. https://doi.org/10.1186/1471-2105-8-209.

112. Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. *49*, D192–D200. https://doi.org/10.1093/nar/gkaa1047.

113. Jacobs, J.P., Lagishetty, V., Hauer, M.C., Labus, J.S., Dong, T.S., Toma, R., Vuyisich, M., Naliboff, B.D., Lackner, J.M., Gupta, A., et al. (2023). Multi-omics profiles of the intestinal microbiome in irritable bowel syndrome and its bowel habit subtypes. Microbiome *11*, 5. https://doi.org/10.1186/s40168-022-01450-5.

114. Tong, F., Tang, G., and Wang, X. (2023). Characteristics of Human and Microbiome RNA Profiles in Saliva. RNA Biol. *20*, 398–408. https://doi.org/10.1080/15476286.2023.2229596.

115. Song, F., Kuehl, J.V., Chandran, A., and Arkin, A.P. (2021). A Simple, Cost-Effective, and Automation-Friendly Direct PCR Approach for Bacterial Community Analysis. mSystems *6*, e0022421. https://doi.org/10.1128/mSystems.00224-21.

116. Stead, M.B., Agrawal, A., Bowden, K.E., Nasir, R., Mohanty, B.K., Meagher, R.B., and Kushner, S.R. (2012). RNAsnap™: a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. Nucleic Acids Res. *40*, e156. https://doi.org/10.1093/nar/gks680.

117. Aranda, P.S., LaJoie, D.M., and Jorcyk, C.L. (2012). Bleach gel: a simple agarose gel for analyzing RNA quality. Electrophoresis *33*, 366–369. https://doi.org/10.1002/elps.201100335.

118. Nakamura, T., Yamada, K.D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics *34*, 2490–2492. https://doi.org/10.1093/bioinformatics/bty121.

119. Sequence correction provided by ONT Research. GitHub. https://github.com/nanoporetech/medaka.

120. Wang, W., Artiles, K.L., Machida, S., Benkirane, M., Jain, N., and Fire, A.Z. (2023). Combined direct/indirect detection allows identification of DNA termini in diverse sequencing datasets and supports a multiple-initiation-site model for HIV plus-strand synthesis. Preprint at bioRxiv. https://doi.org/10.1101/2023.06.12.544617.

121. Saldanha, J.A., Thomas, H.C., and Monjardino, J.P. (1990). Cloning and sequencing of RNA of hepatitis delta virus isolated from human serum. J. Gen. Virol. *71*, 1603–1606. https://doi.org/10.1099/0022-1317-71-7-1603.

122. Gross, H.J., Domdey, H., Lossow, C., Jank, P., Raba, M., Alberty, H., and Sänger, H.L. (1978). Nucleotide sequence and secondary structure of potato spindle tuber viroid. Nature *273*, 203–208. https://doi.org/10.1038/273203a0.

123. Bussière, F., Ouellet, J., Côté, F., Lévesque, D., and Perreault, J.P. (2000). Mapping in Solution Shows the Peach Latent Mosaic Viroid To Possess a New Pseudoknot in a Complex, Branched Secondary Structure. J. Virol. *74*, 2647–2654. https://doi.org/10.1128/JVI.74.6.2647-2654.2000.

124. Johnson, A.D. (2010). An extended IUPAC nomenclature code for polymorphic nucleic acids. Bioinformatics *26*, 1386–1389. https://doi.org/10.1093/bioinformatics/btq098.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Streptococcus sanguinis SK36	American Type Culture Collection (ATCC)	BAA-1455
Chemicals, peptides, and recombinant proteins		
Brain Heart Infusion broth	Millipore®	53286
Glycerol	Sigma-Aldrich®	G5516
Igepal-CA630	Sigma-Aldrich®	I8896
UltraPure™ Agarose	Invitrogen™	16500500
Tris-Acetate-EDTA	BioRad	1610743
Ethidium Bromide	Sigma-Aldrich®	E1510
Orange loading dye	New England Biolabs	B7022
Purple loading dye	New England Biolabs	B7024
100 bp marker DNA ladder	New England Biolabs	N0551
1 Kb Plus DNA ladder	Invitrogen™	10488085
UltraPure™ Distilled Water	Invitrogen™	10977015
RNAprotect™ bacteria reagent	Qiagen®	76506
Formamide	Millipore®	344206
β-Mercaptoethanol	Sigma-Aldrich®	63689
Sodium Dodecyl Sulphate	Invitrogen™	AM9820
Ethylenediaminetetraacetic acid	Invitrogen™	AM9260
TURBO™ DNase	Invitrogen™	AM2238
pH 8.0 Tris-EDTA buffer	Sigma-Aldrich®	93283
T4 lysozyme	New England Biolabs	P8115
Lytic Enzyme Solution	Qiagen®	158928
RNaseR	Biosearch Technologies	RNR07250
Critical commercial assays		
LunaScript® RT SuperMix	New England Biolabs	E3010
NEBNext® Ultra™ II Q5® Master Mix	New England Biolabs	M0544
RNA Clean & Concentrator 5	Zymo Research	R1013
Monarch® Genomic DNA Purification Kit	New England Biolabs	T3010
Zymoclean™ Agarose DNA gel extraction kit	Zymo Research	D4001
NEBNext®, rRNA bacteria depletion kit	New England Biolabs	E7860
Qubit™ RNA high sensitivity kit	Invitrogen™	Q32852
Qubit™ dsDNA high sensitivity kit	Invitrogen™	Q32851
SMART-Seq Total RNA Mid Input	TaKaRa Bio	635049
NucleoMag™ NGS SPRI beads	MACHEREY-NAGEL	744970
Unique Dual Index Kit 96U Set B	TaKaRa Bio	634457
Nextera XT DNA library preparation kit	Illumina	FC1311024
DNA/RNA UD Indexes Set C	Illumina	20091648
Deposited data		
RNA and DNA sequencing of *S. sanguinis* SK36 harbouring and lacking Obelisk-*S.s*	this paper	BioProject: PRJNA1129866
data, supplementary data, code, and metadata	this paper	https://purl.stanford.edu/wb363nt3637
publically available sequencing data	Sequence Read Archive	see supplementary data above

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Oligonucleotides		
GCTAGAAATAGAAAGGTACCTTTACAGTAAAAAGATGTATCC	Integrated DNA Technologies	Obelisk-S.s-Marker-Fw
CGTTTTTCAGAGTAACCATGATATAGTTCGAACGG	Integrated DNA Technologies	Obelisk-S.s-Marker-Rv
GCTGTTTAGGCTGTGGTCTTCC	Integrated DNA Technologies	SK36-Fw
TCGCAGGCTAACCATTCATGCG	Integrated DNA Technologies	SK36-Rv
GGAACGATCTATCCTCTGAATAAATCACG	Integrated DNA Technologies	Disc-1-Fw
TTTGTATCCAAACTCGTAAGGAATTCCATCC	Integrated DNA Technologies	Disc-1-Rv
TCGAACTTCTTCTTTCAAGAATTTCCTAATTGG	Integrated DNA Technologies	Disc-2-Fw
CCTTAAGTTCTTAGGCTTTCCGTTGCC	Integrated DNA Technologies	Disc-2-Rv
Software and algorithms		
VNom	this paper	N/A
rnaSPAdes	Bushmanova et al.[49]	N/A
circUCLUST	https://github.com/rcedgar/circuclust	N/A
USEARCH	Edgar[50]	N/A
MARS	Ayad and Pissis[51]	N/A
fasterq-dump	https://github.com/ncbi/sra-tools	N/A
fastp	Chen et al.[52]	N/A
RNAfold	Lorenz et al.[53]	N/A
blast+	Camacho et al.[54]	N/A
Infernal	Nawrocki and Eddy[55]	N/A
CaCoFold	Rivas[56]	N/A
Kraken2	Wood et al.[57]	N/A
KrakenGrafter	https://github.com/Zheludev/FireTools/	N/A
Bracken	Lu et al.[58]	N/A
Bracken2OTU	https://github.com/Zheludev/FireTools/	N/A
bwa-mem2	Vasimuddin et al.[59]	N/A
picard	http://broadinstitute.github.io/picard/	N/A
freebayes	Garrison and Marth[60]	N/A
SAMtools	Li et al.[61]	N/A
bamaddrg	https://github.com/ekg/bamaddrg	N/A
SNPRelate	Zheng et al.[62]	N/A
R	https://www.R-project.org/	N/A
PebbleScout	Shiryev and Agarwala[25]	N/A
diamond	Buchfink et al.[63]	N/A
prodigal	Hyatt et al.[64]	N/A
Serratus	Edgar et al.[2]	N/A
Muscle5	Edgar[65]	N/A
HMMer	Eddy[66]	N/A
MSACleaner	https://github.com/Zheludev/FireTools/	N/A
FASTACleanUp	https://github.com/Zheludev/FireTools/	N/A
msaconverter	https://github.com/linzhi2013/msaconverter	N/A
seqkit	Shen et al.[67]	N/A
ColabFold	Mirdita et al.[33]	N/A
HHblits	Remmert et al.[68]	N/A
Phyre2	Kelley et al.[69]	N/A
Dali	Holm[70]	N/A
FoldSeek	van Kempen et al.[71]	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MSA	Bodenhofer et al.[72]	N/A
Biostrings	Pagès et al.[73]	N/A
ggseqlogo	Wagih[74]	N/A
pseqid	https://github.com/amaurypm/pseqsid/	N/A
factoextra	https://github.com/kassambara/factoextra/	N/A
FastME2	Lefort et al.[75]	N/A
phangorn	https://github.com/KlausVigo/phangorn/	N/A
ggtree	https://github.com/YuLab-SMU/ggtree/	N/A
CIAlign	Tumescheit et al.[76]	N/A
iqtree	Minh et al.[77]	N/A
ModelFinder	Kalyaanamoorthy et al.[78]	N/A
UFBoot2	Hoang et al.[79]	N/A
iTOL	Letunic and Bork[80]	N/A
ScanRabbit	https://github.com/FireLabSoftware/ScanRabbit/	N/A
KmerCatcher	https://github.com/Zheludev/FireTools/	N/A
EMBOSS	Rice et al.[81]	N/A
Ident and Sim	Stothard[82]	N/A
RNAalifold	Bernhart et al.[83]	N/A
circlize	Gu et al.[84]	N/A
R2R	Weinberg and Breaker[85]	N/A
GCplyr	Blazanin[86]	N/A
breseq	Deatherage and Barrick[87]	N/A
kallisto	Bray et al.[88]	N/A
sleuth	Pimentel et al.[89]	N/A
PolyBench	https://github.com/FireLabSoftware/PolyBench/	N/A
SPLASH2	Chaung et al.[90] and Kokot et al.[91]	N/A
VariantRabbit	https://github.com/FireLabSoftware/VariantRabbit/	N/A

EXPERIMENT MODEL AND STUDY PARTICIPANT DETAILS

Microbial strains

Streptococcus sanguinis SK36 was obtained from the American Type Culture Collection (ATCC, BAA-1455) and was cultured in autoclaved Brain Heart Infusion (BHI) broth at 37 °C; or on 1 % agar - BHI plates; both in an ambient, oxic atmosphere. Derivative substrains ObP1 and ObN1 were identified as described in the experimental methods below and always grown in triplicate (from 3 individual, derived colonies per substrain) along with triplicate mock-inoculated media controls. Strain verification was performed *post hoc* on the short read sequencing data.

METHOD DETAILS

VNom

VNom (pronounced *venom*, short for ''Viroid Nominator'') was written to sequentially filter, in a homology-independent manner, for contigs with molecular features consistent with viroid-like biology from *de novo* assembled stranded RNA-seq data, namely: apparent circularity, and the co-occurrence of both positive- and negative-sense strands within a given sample (Figure S1D). As an input, VNom can take in any De Bruijn graph assembled contigs from stranded RNA-seq data; however, VNom is optimised to work on the output from rnaSPAdes. Initially, apparent circularity is inferred by identifying perfect k-mer repeats between the start and end of a contig: a previously exploited[2,12,92] sequence feature produced from circular De Bruijn graphs which are in turn produced from repetitive or circular transcripts during assembly. These apparently circular contigs are further de-concatenated into apparent unit-length, monomeric sub-sequences if a regular repetition of the identified k-mer is found, as is analogously done in Lee et al.[14] The resulting apparently circular contigs are then clustered with circUCLUST and clusters containing at least one apparent sense and one antisense contig are kept (as inferred by k-mer counting). Any previously filtered out contigs that produce strong global alignments (usearch -usearch_global) to these resulting sense-antisense clusters are then re-introduced where any clusters with now mutual contigs are merged. Local alignment (usearch -usearch_local) is then used to resolve and annotate any new multi-unit-length contigs into monomeric sub-sequences, and any sub-unit-length sequences into fragments. Finally, the resulting clusters are all "phased" to the same circular permutation using the multiple sequence aligner MARS. VNom is freely available at github.com/Zheludev/VNom.

Initial Obelisk identification

Stranded RNA-seq data were fetched from the SRA[93] using fasterq-dump, adapter and quality filtered using fastp (–average_qual=30 –n_base_limit=0 –cut_front –cut_tail), and *de novo* assembled with rnaSPAdes (default settings). Viroid-like sequences were identified using VNom (-max 2000 -CF_k 10 -CF_simple 0 -CF_tandem 1 -USG_vs_all 1).

Obelisk RNA was initially identified in a longitudinal dataset of human stool stranded metatranscriptomics from the Integrative Human Microbiome Project (iHMP).[20] All paired-end RNA-seq datasets were downloaded (104 donors), trimmed, and assembled as described. Contigs were then grouped by donor ID and passed through VNom. The 2306 resulting VNom-nominated sense contigs were then queried manually for apparent lack of nucleotide, or protein-coding homology to the NCBI nt/nr (see later in this paragraph). Amongst these, we chose a sequence with striking predicted RNA secondary structure (high degree of basepairing, by eye, RNAfold -p -d2 –noLP –circ). Obelisk RNAs were also manifest when VNom nominated contigs were passed through the following pipeline: the sense contigs were queried against a custom database (see the key resources table) of self-cleaving ribozymes (CMscan, default settings, keeping any, including likely spurious, hits), these resulting 196 contigs were then assayed against the NCBI nt database (*11 Oct 2021*, blastn, default settings), and contigs that yielded no hits, or whose best (by E-value) hits aligned to less than 40 % of contig's length were kept. These resulting 20 contigs were then queried against the NCBI nr database (*8 Nov 2021*, blastx, default settings), similarly keeping sub-40 % alignment length best hits, yielding 11 contigs, of which 5 had a unit length of 1164 nt (one contig was 1166nt) - suggesting a common class of RNA. These were later defined as the Obelisk RNAs. Similarly, blastn/p filtering the 2306 sense contigs but without the CMscan step yielded 107 contigs, 8 of which were over 1000 nt in length, comprising the 6 Obelisk RNAs. Lastly, running blastn on all the iHMP contigs against the 6 Obelisk RNAs resulted in a final total of 15 unique Obelisk RNA sequences.

Taxonomic classification

Taxa from length-filtered reads (fastp, as above with –length_required 75) were classified using Kraken2 (default settings) against the Phanta[94] database, modified with non-redundant Obelisk-α/β sequences using KrakenGrafter, followed by Bayesian re-estimation using Bracken (-r 75), lastly taxon counts were combined using Bracken2OTU, summing any samples that came from the same donor on the same day (indicative of split sequencing lanes).

Obelisk-α positive length-filtered read datasets, were assessed for sequence diversity relative to a fixed, arbitrarily chosen Obelisk-α reference. Namely, single nucleotide polymorphisms (SNPs) and small structural variants were measured by aligning reads (bwa-mem2, default settings) to the reference, followed by deduplication (picard, MarkDuplicates), and detection freebayes (–ploidy 1 –pooled-discrete –pooled-continuous). SAMtools and bamaddrg were used throughout. Principal component analysis (PCA) on the resulting vcf file was computed using SNPRelate (snpgdsPCA), as described in [95], clusters were identified by kmeans (centers = 5).

Obelisk homologue detection in other public data

Close Obelisk-α homologues were identified in the Short Read Archive (SRA) using PebbleScout ("Metagenomic" database, default settings), a recently released tool that efficiently queries ~3.2 million (*mid 2022*) raw sequencing data for exact, fixed 42 k-mer matches. 9 metatranscriptome BioProjects (comprising 34 short read datasets) were identified (PBSscore > 65) with close (~1 % nucleotide divergence) matches to Obelisk-α, of which 3 were part of iHMP or its predecessor,[96] 5 were from other human stool studies,[97–101] and 1 was from a fox gut autopsy.[102] Using the VNom pipeline (see above), 21 datasets (from 7 BioProjects) yielded full length Obelisk-α sequences, all from human hGMB studies (Table S1).

Finding Obelisk-α homologues in studies separate from the iHMP lent support to these RNA elements being legitimate biological entities. Further, one Obelisk-α homologue was found in a study from our own institution,[101] suggesting that Obelisk-like RNAs could be locally present. Emboldened by this, we solicited hGMB stranded RNA-seq data from the local academic community and identified closely related Obelisk-α homologues in a dataset that at the time had not been uploaded onto the SRA (now available at PRJNA940499: donors D01 - both Obelisks -α and -β; and D10 - just Obelisk-α).[103] Further, within this dataset we identified a diverged Obelisk-like sequence with similar: length (1182 nt), lack of apparent homology to reference databases, predicted obelisk-like secondary structure, and two ORFs but with low homology to Obelisk-α. In comparison to Obelisk-α, this new "Obelisk-β" had a 41.30 % nucleotide sequence identity, and 23.42/38.29 % and 18.75/31.25 % on the amino acid level identities/similarities for ORFs 1 and 2, respectively (see below, Figures S2A/S2C).

Owing to their apparent sequence novelty, the Obelisk-α/β Oblin-1 and -2 protein sequences were next used as hallmark sequences specific to Obelisk-like RNAs - analogous to the use of RNA-dependant RNA polymerase (RdRP) hallmark sequences in RNA viral discovery.[1–5] To identify divergent Obelisk-like elements, we searched the RNA Deep Virome Assemblage (RDVA, v0.2),[13,26] a collection of 58,557 assemblies of ~12.5 trillion contigs, with diamond (–very-sensitive) using Obelisk-α/β Oblin -1 and -2 protein sequences deduplicated at 90 % sequence identity (UCLUST, default settings) as queries. This resulted in 38,545 sub-5000 nt hits which when de-replicated, circularly clustered (circUCLUST) into 29,859 and 19,808 clusters at 90 % and 75 % nucleotide sequence identity, respectively (see the key resources table). A conservative database of 7,202 Obelisks was built by keeping assemblies with a CircleFinder (VNom defaults) implied circularity, with each genome "phased" to 50 nt from the start codon of its largest predicted ORF (prodigal, -p meta). This database was clustered (circUCLUST) into 1,744 80 % identity clusters which were then sub-clustered at 95 % identity (Table S2). The assemblies were then named based on these nested clusterings. A naming convention is proposed with the following pattern *"Obelisk_X_Y_Z"* where "X" refers to the 80 % cluster ordinate, "Y" to the 95 % cluster ordinate, and "Z" as a unique identifier within the 95 % cluster. The first 15 80 % ordinates are defined as the Obelisks depicted in Figure 6, the next 10 80 % ordinates are defined as the remaining letters in the Greek alphabet (*omicron* through *omega*). As such, the centroid Obelisk-α sequence that is also the centroid of the first 95 % sub-type is defined as "*Obelisk_000001_000001_000001*".

Serratus

Extending from the RDVA search, a larger breadth of public datasets (5,470,176 runs) was next assessed for diverged Obelisk-like sequence presence. Profile hidden Markov models (pHMMs) of ORFs 1 and 2 were derived from the RDVA hits (see below) and used as queries in the Serratus architecture,[2] an optimised, cloud-based pipeline for efficiently identifying sequencing reads that align to pHMMs. By looking for pHMM matches, Serratus is able to find more distantly related Obelisk-like sequences where k-mer match searches (*e.g.* PebbleScout) would fail, but at a considerable computational expense. Datasets were defined as a Serratus hit if at least one read aligned (E-value $<1x10^{-4}$) to either Oblin-1 or Oblin-2. Of the resulting 949,810 non-redundant SRA hits, 215,398 datasets were selected by filtering with a virus-presence score (≥25, explained in github.com/ababaian/serratus/wiki/.summary-Reports) which attempts to predict ORF *de novo* assembly success, ultimately yielding 1,499 datasets containing both Oblin-1 and Oblin-2, 3,006 containing only Oblin-1, and 213,891 containing only Oblin-2. Per hit SRA, high confidence ORF mapping reads were then *de novo* assembled using rnaSPAdes (default settings) yielding Obelisk "micro-assemblies". This Serratus run was conducted along with other pHMM queries, meaning that *de novo* assembly happened in aggregate with all other hits, as such, diamond (–very-sensitive) was used to extract Oblin-1/-2 micro-assembly protein sequences.

Protein homology bioinformatics

To probe the deep sequence diversity of Oblins 1 and 2, corresponding single domain profile hidden Markov models (pHMMs) were individually constructed from the RDVA hits using an iterative approach: A multiple sequence alignment (MSA) from the initial PebbleScout set was computed using Muscle5 (default settings), from which an initial pHMM was computed using HMMbuild (default settings). Each genome in the RDVA non-redundant 90 % sequence identity cluster centroid set was doubled in length using SeqDoubler[55] and ORFs were predicted using Prodigal (-p meta). Note, Prodigal attempts to predict Shine-Dalgarno sequences for the ORFs it identifies, in this study, we have included annotations of these predicted motifs (Figure 6, where made), however, we chose not to interpret the presence (or absence) of these motifs as no estimate of reliability is given. ORFs with predicted N- or C- terminal truncation were omitted and a non-redundant set was kept (usearch -fastx_uniques). This ORF database was queried against (HMMsearch, default settings) the initial pHMM and hits with global E-values lower than $1x10^{-15}$ for Oblin-1 or $1x10^{-8}$ for Oblin-2 were kept. HMMalign (–trim) and MSACleaner (-ref from the PebbleScout set and -fxn 0.01) were used recursively (until no new sequences were omitted) to filter the constituent MSA sequences to omit sequences that contributed large indels relative to the initial pHMM. A new pHMM was computed and the HMMsearch (on the remaining ORFs), HMMalign (without –trim), and MSACleaner steps were repeated once. This resulting MSA was filtered by sequence length FASTACleanUp (-lower 150 for Oblin-1, -lower 40 for Oblin-2) and a final pHMM was computed. Msaconverter was used throughout. There were no overlapping sequences between the resulting Oblin-1 and -2 pHMMs.

A contiguous alignment block of 18 amino acids was noticed in the resulting Oblin-1 pHMM (Obelisk-α: 152-RRRGYKDHGSRRFPHEVH-169) and was selected as a marker sequence, terming it *domain-A*. Because the Serratus Oblin-1 micro-assemblies may include some that are not full-length (*wrt* Oblin-1), further aggregation from the Serratus data utilised a search for similarity to *domain-A*. To incorporate the Serratus results, an initial 503 sequence *domain-A* alignment was extracted from the RDVA pHMM (and later used with K-mer Rabbit, below) and a new pHMM was constructed (HMMbuild, default settings). A length sorted (seqkit sort -l -r), non-redundant (usearch -fastx_uniques) set of Serratus Oblin-1 micro-assemblies was then iteratively queried with an ever-rebuilt *domain-A* pHMM: keeping HMMsearch (default settings) hits with E-values lower than $1x10^{-4}$, intermediate MSAs were re-built (HMMalign –trim) relative to the previous iteration and sequences with at least 8 amino acids (seqkit seq -g -m 8) were kept, next, the resulting sequences were re-aligned to the current pHMM and a new pHMM was built, lastly, all $<1x10^{-4}$ E-value hits were omitted and a new iteration was started. A finalised Serratus-inclusive *domain-A* pHMM was constructed with 30,686 sequences after 12 cycles. This process was repeated for two other less well-conserved domains, *domain-B* (Obelisk-α: 96-CLTSKSGMLNFLEDTTLY-113), and *domain-C* (Obelisk-α: 53-RSKKDLLALAIISWWLEE-70), with 5076 and 5103 resulting sequences, respectively. *Domains -B/-C* were not studied further in this work.

Protein tertiary structure prediction

For initial, monomeric tertiary structure prediction, RDVA pHMM MSAs were re-aligned (Muscle5, default settings) relative to ORFs-1/2 from Obelisk-ɑ and used with ColabFold (v1.5.2-patch) implementation of AlphaFold2 (default settings, no amber, no dropout). The HHblits suite was used to convert between fasta and a3m MSA formats. Tertiary structure homology was assessed using the Phyre2 (default settings), Dali (PDB Search), FoldSeek (all databases, 3Di/AA and TM-align scoring), and the Clustered AlphaFold Database[104] webservers (see data and code availability). For all other tertiary structure predictions, ColabFold was used with mmseqs2 uniref env for MSA generation. For 9 in 15 predictions, including Obelisks -ɑ, -β, and -*S.s*, this yielded qualitatively similar "globule" predictions (Figure 6 - tertiary fold predictions). An equivalent 73 sequence MSA was constructed for Oblin-1 homologues from ribozyme-baring Obelisks (see RNA homology bioinformatics) by first filtering any Prodigal-predicted proteins for length (seqkit seq -m 200 -M 250), aligning the resulting sequences (Muscle5), and manually removing any sequences that appeared to disrupt the MSA. ColabFold v1.5.3 was used for ribozyme-baring Oblin-1 protein tertiary fold predictions and Obelisk-nu.

Protein conservation and phylogenetics

Oblin-1/-2 conservation analysis was conducted on Obelisk-ɑ-relative a3m alignments against the BLOSUM62 substitution matrix[105] using msaConservationScore (gapVsGap = 0) and the Biostrings package. The Oblin-2 sequence logo was constructed using ggseqlogo, and a consensus sequence was generated with msaConsensusSequence (upperlower, thresh = 20,0).

Owing to the micro-assembly used in the Serratus search, phylogenetic analysis was limited to stringently filtered, representative (80 % sequence identity clustering), full-length Obelisks genomes (Table S2). Per representative "centroid" Obelisk, each largest Prodigal-predicted, full-length ORF amino-acid sequence was iteratively pair-wise aligned (Muscle5) against each one-another, and for each alignment, the identity, similarity, and normalised similarity score (NSS) were calculated using pseqsid (default settings). To enrich for *bona-fide* Oblin-1 sequences, comparisons were limited to ORFs between 180 and 320 amino acids (inclusive) resulting in 1651 remaining sequences. Ignoring self-self comparisons, the minimum and maximum observed values were: identity 9.22 / 99.53 %; similarity 15.27 / 99.53 %; and NSS -0.12 / 0.99. For each type of measurement, a Euclidean distance matrix was computed using get_dist() from which dendrograms were computed using FastMe2 (BioNJ tree building, with best of Nearest Neighbour Interchanges and Subtree Pruning and Regrafting topology optimisation). Of the resulting dendrograms, the identity tree had the highest explained variance (0.675, vs 0.549 for similarity, and 0.669 for NSS) and so was subsequently used. A midpoint-rooted (Phangorn), equal-daylight tree (ggtree) was then annotated with ribozyme-baring status (see RNA homology bioinformatics), if the Obelisk's full length ORF count was exactly 2 (overall, 788 of 1744 of the "stringent" Obelisks harboured at least 2 ORFs), and "Greek" identity if available (see Figures 6 and S6, and Table S1).

To summarise over the RDVA and Serratus search results, a *domain-A* specific multiple sequence alignment (MSA) was also constructed. First Oblin-1 homologues from ribozyme-baring Obelisks (see RNA homology bioinformatics) were queried (HMMsearch –max, E-value $\leq$ 1x10^{-8}) against the initial Oblin-1 pHMM, yielding only sequences homologous to *domain-A*. These sequences were re-aligned (Muscle5) and an initial ribozyme-associated *domain-A* pHMM was built. This ribozyme-associated pHMM was then iteratively built upon with successive rounds of similarity searches (HMMsearch –max, E-value $\leq$ 1x10^{-8}) against the RDVA's ribozyme-baring Obelisk's predicted proteins followed by re-alignment with Muscle5. Once no new sequences were found, the cycle was continued at an E-value threshold of 1x10^{-5}. This resulting ribozyme-associated MSA was then re-aligned to the initial Oblin-1 MSA (HMMalign, default settings) and the alignment column corresponding to *domain-A* was manually excised, and re-aligned (Muscle5). The entirety of the full-length predicted proteins from the RDVA were then similarly iteratively queried but at a E-value threshold of 1x10^{-4}, and without an intermediate Muscle5 step. The converged alignment was then re-aligned with Muscle5 (Super5) and similarly iteratively queried against the Serratus micro-assemblies, keeping the best hit per micro-assembly until convergence. The resulting 46,884 total *domain-A* sequences were finally re-aligned with Muscle5 (Super5). This MSA was then deduplicated, and optimised using CIAlign[88] to remove insertions (minimum size 1, minimum 0.05 %), to crop divergent sequences (minimum identity proportion 0.01, minimum non-gap proportion 0.5, buffer size 4), and to remove any resulting sequences shorter than or equal to 16 aa. A final round of deduplication yielded a 3265 non-redundant sequence *domain-A* no-gap alignment of 17 aa (Table S1).

A maximum likelihood phylogenetic tree was then constructed from this 17 aa alignment using iqtree. The LG+G4 substitution model (testnewonly) was selected (ModelFinder) based on a consensus between the Akaike and Bayesian Information Criteria. Tree construction was run with 33,000 UFBoot bootstraps, Nearest Neighbour Interchange optimization, and 33,000 SH-like approximate likelihood ratio tests (-B 33000 -bnni -alrt 33000). The resulting tree was plotted using iTOL and is available in the supplementary data (see the key resources table).

ScanRabbit

For rapidly searching smaller, locally-held datasets for novel Obelisk homologues, we developed a second tool, ScanRabbit, which focuses on a short segment of any multiple sequence alignment. ScanRabbit was run using the position-specific-scoring matrix (PSSM) based on the multiple sequence alignment used to build the Oblin-1 profile hidden Markov model (see above) from the RDVA hits corresponding to *Domain-A*. ScanRabbit accelerates searches on local hardware through direct bitwise conversion of the PSSM to a local bitwise scoring that can be applied to the raw binary representation of RNA-seq reads, and a just-in-time compiler PyPy.[106] ScanRabbit is available on GitHub at github.com/FireLabSoftware/ScanRabbit.

Obelisk spacer analysis

The presence of Obelisks in known prokaryotic CRISPR spacer arrays was assessed using a conservative k-mer matching approach. Namely, the RDVA Obelisk dataset was queried against predicted CRISPR spacers in the Joint Genome Institute (JGI) IMG/M spacer database (*May 2023*). To estimate a lower length bound on matching noise, a parallel analysis was conducted on "reversed" (*not* reverse complemented) Obelisk sequences. Initially, RDVA Obelisk sequences were searched against the IMG/M spacer database[37] using blastn (default settings), only keeping perfect matches with no gaps or mismatches (k-mers) - the longest k-mer match between a given spacer/Obelisk pairing was kept. Next, all kept spacers containing any 12-mer match to common Illumina sequencing adaptors were omitted using KmerCatcher (default settings). For each remaining spacer, the information content was estimated[107] by comparing how efficiently the compression algorithm zip (-9)[108] could "deflate" a given spacer - a larger length normalised deflation indicates a less complex spacer sequence that is less likely to be unambiguously mapped to a specific (Obelisk) sequence. The repetitive content of each spacer was also assessed using etandem (-minrepeat 4, -maxrepeat 15, -threshold 2). Spacers with a length normalised deflation less than 1.0 percent per nucleotide were kept (137,667 forward, 118,411 reverse), these spacers also qualitatively had a low etandem score though this metric was not used for filtration (see the key resources table). Next, only the 23 forward spacers longer than the maximum length of the reverse spacers (25 nt) were kept as any mappings below this threshold would be indistinguishable from noise (reverse-mapping, see the key resources table). Lastly, the corresponding Obelisks mapping to these spacers were minimum length filtered to 1000 nt (seqkit seq -m 1000), resulting in two contigs. Only one of these contigs gave blastn (default settings, NCBI webserver, *August 2023*) a largely ($\sim$95 %) unknown sequence with a singular $\sim$45 nt sequence mostly showing up in high G+C Gram-positive bacteria and cyanobacteria (consistent with a CRISPR spacer array, see data and code availability). This largely unknown 1096 nt contig was found to encode (prodigal -p meta) homologues of both Oblin-1 and Oblin-2 (HMMsearch, default settings, against the RDVA pHMMs), and is predicted to fold (see below) into an obelisk-like RNA secondary structure (see the key resources table) - features consistent with being an Obelisk which we term Obelisk-"gamma" (Obelisk-ɣ). Two spacers were found to map to Obelisk-ɣ, both from the same *Bombella mellum* genome (RefSeq GCF_014048465.1)[109] - these spacers (which differ by one extra nucleotide) were found at the same putative CRISPR locus but predicted in the IMG/M database with two different tools (PILER-CR and CRT),[110,111] as such, this is likely one spacer. Obelisk-ɣ's predicted secondary structure is not as "rod-like" as other Obelisks (Figure 6 - "jupiter" plots), with the spacer mapping to the "frayed" end; additionally, the spacer mapping position coincides with the locus identified by blastn; and lastly, CircleFinder (VNom default settings) did not identify a start-end k-mer repeat indicative of a circular genome. The Obelisk-ɣ Oblin-1 was also not predicted (see above) to fold into the characteristic "globule" fold (Figure 6 - tertiary fold predictions), though the discriminatory power of this is unclear and so ignored. These features suggest that the Obelisk-ɣ genome might be mis-assembled, with the putative spacer mapping sequence arising from a chimeric assembly. As such, this conservative approach to CRISPR spacer mapping was not able to unambiguously identify any Obelisk relationships to CRISPR spacer arrays as we currently recognise them.

Identity and similarity measurements

Unless otherwise stated, all nucleotide identity, and protein identity and similarity measurements were computed by first building a pairwise alignments Muscle5 (default settings) of "phased" genomes (as below) followed by calculation with Ident and Sim (default settings).

RNA homology bioinformatics

Figures 1B, 5A, 6, S2B, and S5A

RNA secondary structures were predicted using RNAalifold (-p, -r, -d2, –noLP, –circ) on the non-redundant (usearch -fastx_uniques), 1164 nt long, PebbleScout set of the above "phased" Obelisk-ɑ sequences, split by genome polarity, using a Muscle5 (default settings) derived MSA. Figures S1A, S2C and S2D secondary structures were predicted on singular genomes using RNAfold (-p -r -d2 –noLP –circ). RNA secondary structures were illustrated using circlize for "jupiter" plots, and R2R for "skeleton" diagrams. Conserved RNA element (*e.g.* ribozymes) coordinates in Figure S1 were identified using CMscan (–rfam –cut_ga) against the Rfam 14.6 database.[112]

23 Obelisk-encoded hammerhead type-III ribozyme homologous sequences were initially identified (CMsearch) using the RF00008 reference covariance model against the 90 % identity-clustered (circUCLUST), sequence-doubled (SeqDoubler) RDVA dataset, using stringent cutoffs for confident (E-value $\leq$ 1x10^{-5}), full-length (–notrunc) hits, keeping only the best hit per Obelisk genome. An Obelisk-specific, "Obelisk-variant hammerhead type-III" (ObV-HHR3) covariance model (CM) was constructed using an iterative approach: an initial CM was constructed using the 23 hit sequences by aligning them against RF00008 (CMalign, default settings), optimising the alignment using CaCoFold (R-scape: -s, –cacofold, –rna), and finally building (CMbuild, default settings), and calibrating (CMcalibrate, default settings) the CM. Using this initial CM as a starting point, the sequence-doubled RDVA dataset was iteratively passed through the CMsearch, CMalign, CaCoFold, CMbuild, and CMcalibrate pipeline, each time only keeping the best, non-truncated, E-value $\leq$ 1x10^{-5} hits (one hit per Obelisk genome) and additively appending them to the CM, subtracting the hits from the RDVA set as they were found, until no new hits were found. Ultimately, a 178 sequence ObV-HHR3 was constructed with 15 significantly covarying positions identified (Figure S5B). When re-querying (CMsearch, –no-trunc) the full RDVA dataset with this finalised CM at an E-value $\leq$ 1x10^{-4}, 339 Obelisk genomes were identified. The ObV-HHR3 column in Table S2 was annotated with CMsearch, –no-trunc, $\leq$ 1x10^{-5} on sequence-doubled genomes.

Streptococcus sanguinis bioinformatics

In an attempt to identify Obelisk-like elements that had been serendipitously sequenced in isolation with their putative cellular host(s), Oblin-1 positive filtered Serratus hits were screened for potentially low biodiversity experimental designs such as defined co-culture, single-cell RNA-seq, and isolate culture. As such, isolate RNA-seq experiments of *Streptococcus sanguinis* (strain SK36, a commensal of the human oral microbiome) stood out (Table S1). Upon further investigation (using CircleFinder from VNom), a 1137 nt, obelisk-shaped RNA coding only for Oblin-1 was identified. This so-called "Obelisk-*S.s*" exhibited 40.65 % and 35.47 % nucleotide sequence identity with Obelisk-α and Obelisk-β, respectively, and 19.92/33.47 % and 21.05/32.71 % Oblin-1 amino acid identity and similarity to Obelisk-α and Obelisk-β, respectively (see above, Figures S2A/S2D). Additionally, Obelisk-*S.s* was further found in human oral microbiome samples (Table S1, Figure S6), and by comparing isolate cultures from different growth media, *S. sanguinis* was determined to be the likely cellular host as opposed to Obelisk-*S.s* being a contamination from complex media.

Surveying for Obelisks in human data

The prevalence of Obelisks in five human microbiome datasets (three gastric, hGMB, and two oral, hOMB, Table S1) was (re-)evaluated after both Obelisks -α, -β, and -*S.s* were identified, and the RDVA pHMMs were constructed. For human gut metatranscriptome data, the 104 iHMP donors,[20] and the 10 "ZF" donors from the dataset where Obelisk-β was found[103] were reanalysed; additionally, 326 new donor samples from an irritable bowel syndrome study[113] were queried, for a total of 440 hGMB donors analysed. For human oral metatranscriptome data, 22 (50/50 healthy/case) donors from a Dutch cohort studying periodontitis,[38] and 10 healthy donors from an oral extracellular vesicles study[114] were queried for a total of 32 hOMB donors analysed. We note that the Serratus search identified additional human-associated metatranscriptome data, and that these five datasets were chosen for detailed analysis primarily on the niche of sampling (gastric / oral) and their sample size. To identify more diverged Obelisk elements, a pHMM mapping approach was taken - similarly to Serratus. Namely, each dataset's trimmed read-1 reads (as before) were translated in all six frames (seqkit -f 6 -F) and assessed for Oblin-1 homology using HMMsearch (default settings) against the RDVA Oblin-1 pHMM. Donors with greater than or equal to 10 translated reads (averaging over per-donor replicates, time points, or sampling locations if present) mapping with an E-value less than or equal to $1\text{x}10^{-5}$ were counted as true Oblin-1 hits. Additionally, these trimmed reads were assessed for Obelisk -α, -β, and -*S.s* presence using a modified Phanta Kraken2 and Bracken database constructed as before incorporating all non-redundant Obelisk -α, -β, and -*S.s* sequences (only the previous Obelisk-α positive iHMP datasets were re-assessed in this way). Across these five datasets, 21 donors were identified as positive for Obelisk homologues (>10 HMMsearch hits) but negative for Obelisks -α, -β, or -*S.s* (<10 Kraken2 hits), additionally, 25 donors were identified as positive for Obelisks -α, -β, or -*S.s* (>10 Kraken2 hits, Figure S6). The presence of pHMM-mapping reads in the absence of k-mer reads suggested the existence of new Obelisks, as such, these 21 donors' datasets were assessed for new Obelisks. Briefly, these donor's trimmed reads were assembled as before, keeping any contigs with Oblin-1 homology (HMMsearch, –max, E-value $\leq 1\text{x}10^{-5}$), and then selecting for apparently circular contigs with CircleFinder (default VNom settings). These selected contigs were next assessed for Oblin-2 coding capacity (prodigal -p meta, followed by HMMsearch and blastn versus the Oblin-2 RDVA pHMM, and Obelisk-α Oblin-2 sequence and consensus, respectively E-value $\leq 1\text{x}10^{-4}$), and obelisk-like secondary structure as before. Clustering all resulting and previously identified contigs (circUCLUST -id 0.8), 11 new full-length Obelisks were identified, which we named "delta" through "xi" ("*Obelisk_000005*" to "*Obelisk_000015*" in Table S2; Figure 6). "Delta," "epsilon," "zeta," and "eta" were found in the hGMB datasets and all remaining Obelisks were found in the Dutch hOMB dataset - indicating a human sampling site specificity to Obelisk species. Of these 11 Obelisks, eight apparently only code for an Oblin-1 homologue, Obelisk-"kappa" codes for an Oblin-2 homologue, and Obelisks -"lambda" and -"mu" code for a second ORF similar in size to Oblin-2 but with no obvious homology (which we term the "2ndORF" as more study is needed to determine if this is actually a *bona fide* new ORF). Four of these new Obelisks' ("epsilon," "kappa," "mu," and "xi") Oblin-1 sequences were not predicted to fold (as before) into the otherwise characteristic "globule" tertiary structure (Figure 6 - tertiary fold predictions). These new Obelisks span between 733 nt (Obelisk- "iota") to 1372 nt (Obelisk-"kappa"). Considering these new Obelisk sequences, as well as donors which did not yield full-length Obelisk candidates, Obelisks appear to occur in 9.5 % of the human donors assayed (6.6 % of hGMB samples, and 53 % of hOMB samples) and describe a wider breadth of characteristics that Obelisks seem to be able to possess (length and coding capacity).

S. sanguinis SK36 culturing and RT-PCR screening

To follow up on the apparent Obelisk-*S.s* - *Streptococcus sanguinis* SK36 (hereafter "SK36") association, the SK36 strain (American Type Culture Collection, ATCC, BAA-1455) was grown at 37 °C, in an ambient, oxic atmosphere, in autoclaved, Brain Heart Infusion broth (BHI, Millipore® NutriSelect®, 53286) or on 1 % agar (Millipore®, 01916) - BHI in 100 mm, polystyrene bacteriological petri dishes (VWR®, 25384302). Liquid cultures were grown in 5 mL volumes, in 14 mL round-bottom, capped but not sealed polypropylene test tubes (BD Falcon®, 352059) with constant orbital shaking. The initial ampule from ATCC was resuspended in BHI, and grown for ~48 hours, after which, multiple 1.5 mL, 30 % glycerol stocks were made by diluting the turbid culture 1:1 in 60 % glycerol (Sigma-Aldrich®, G5516, made in SynergyUV®, EMD Millipore water and 0.2 μm filtered, Thermo Scientific™, Nalgene™, PES, 5650020) before freezing at -80 °C in CryoVials (Thermo Scientific™, Nalgene™, 1167649). Following, a new ~48 hour BHI culture was inoculated from a glycerol stock (Fisherbrand™ disposable, 1 μL inoculating loops, 22363595) after which a 10 μL suspension was used for plating. After ~72 hours, single colonies were picked into BHI liquid cultures and grown for a further 48 hours before being pelleted and resuspended in 500 μL 30 % glycerol in 0.5 x BHI for stocks.

These six clonal-origin stocks were then screened for Obelisk-*S.s* using a direct, duplex, endpoint RT-PCR assay based on an established method[115]: 200 μL aliquots of overnight, turbid BHI cultures were pelleted and resuspended in 100 μL, 0.1 % Igepal-CA630 (Sigma-Aldrich®, I8896), heat treated at 98 °C for 5 minutes before reverse transcription of 6 μL of the resulting supernatant in 10 μL final volume reactions (New England Biolabs, NEB, LunaScript® RT SuperMix, E3010) with 1.5 μM final concentrations of both Obelisk-*S.s* Marker-Fw, and SK36-Rv primers added in (note, SK36-Fw is likely the more appropriate primer to use for RT-PCR, however this does not appear to be an issue, likely due to the random hexamer and poly-T primers supplied in the LunaScript® RT SuperMix as well as the SK36 genomic DNA in the reaction). Reverse transcription reactions were performed at 25 °C for 2 minutes, 55 °C for 10 minutes, 65 °C for 10 minutes, and 95 °C for 1 minute followed by direct dilution into 25 μL final volume PCR reactions (NEB, NEBNext® Ultra™ II Q5® Master Mix, M0544) with 250 nM final concentration paired Obelisk-*S.s* Marker-Fw/Rv and SK36-Fw/Rv primers added in (Integrated DNA Technologies, IDT, see the key resources table). Note in order to ensure RT-PCR (and sequencing) interpretability, no Obelisk-S.s (or other Obelisk) RNAs were ever synthetically produced in the laboratories used for this study. PCR cycling was performed at 98 °C for 30 seconds, followed by 35 cycles of 98 °C for 10 seconds, 68 °C for 20 seconds, and 72 °C for 20 seconds; ending with 72 °C for 2 minutes. Amplicons were then analysed by 2 % agarose (Invitrogen™, UltraPure™ agarose, 16500500) gel electrophoresis in 1X Tris-Acetate-EDTA (TAE, BioRad, 1610743) - 0.2 μg/mL Ethidium Bromide (EtBr, Sigma-Aldrich®, E1510) buffer in 1x Orange loading dye (NEB, B7022), run at 7 V/cm along with a 100 bp marker DNA ladder (NEB, N0551). The Obelisk-*S.s* and SK36 Marker-Fw/Rv amplicons were expected to run at 249 bp and 301 bp, respectively. For all RT-PCR assays, negative controls (mock inoculated BHI) and no template controls (NTCs, RT and PCR of Invitrogen™, UltraPure™ DNase/RNase-Free Distilled Water, ddH_2O, 10977015) were performed. From the initial six clonal SK36 isolates, one stock was serendipitously found to be consistently negative for the Obelisk-*S.s* marker amplicon, this stock was designated "Obelisk Negative 1" (ObN1), and an arbitrary, amplicon-positive, second stock was chosen to be "Obelisk Positive 1" (ObP1). Both ObN1 and ObP1 glycerol stocks were then streaked out and grown for ~72 hours, after which three single colonies (designated A, B, and C) for each were picked into 10 μL of fresh BHI, of which 5 μL was used to inoculate 24 hour BHI liquid cultures followed by glycerol stocking, and the remaining 5 μL were added to 1 μL of 0.6 % Igepal, for validation duplex RT-PCR as described above using the entire 6 μL volume and 45 cycles of PCR. For the remainder of the study, ObN1 was used as a *spontaneous loss* (emphasised in distinction to the term *knockout*) comparator to the apparently isogenic ObP1 (see below) for investigating the impact of Obelisk-*S.s* positivity.

ObN1 A, B, and C, and ObP1 A, B, and C growth characteristics were assessed by plate-based (Corning®, Costar®, 3361) OD_{600} growth curve assays. Clonal stocks were out-grown in BHI liquid cultures from glycerol stocks for 24 hours, before dilution into 5 mL fresh BHI to a final starting OD_{600} of 0.002 (measured on an IMPLEN, OD_{600} DiluPhotometer™ spectrophotometer with half-width polystyrene cuvettes, Brand GmbH, 759015) and pipetting 200 μL volumes into 8 wells per clone, and a matching 8 wells of mock-inoculated BHI. We note that this experimental set-up matches that of the short read sequencing experiments (see below) and as such we do not anticipate abrupt Obelisk-*S.s* loss from ObP1 over this time course. Each column of inoculated wells was spaced apart with intervening mock-inoculated wells. The filled plate was sealed with a breathable membrane (Sigma-Aldrich®, Breath-Easy®, Z380059) and placed in a 37 °C, heated plate reader (Agilent BioTek Synergy H1), without its lid for 24 hours of growth with constant orbital shaking and maximally frequent A_{600} kinetic measurement. Raw A_{600} measurements were then processed using GCplyr to extract growth rates and lag times for each inoculated well (derivative window size = 25).

Assays on SK36 ObP1 and ObN1 nucleic acids

To characterise the nucleic acids of SK36 ObN1 and ObP1, total RNA and genomic DNA were prepared from liquid cultures, followed by nuclease and column-based purification, and then molecular tests with RT-PCR, PCR, DNA sequencing, and RNA sequencing. Mock-inoculated BHI-only controls were conducted in triplicate throughout. The six clonal stocks (ObN1 and ObP1 A, B, and C) were out-grown in BHI liquid cultures from glycerol stocks for 24 hours, before inoculating 5 mL fresh BHI to a final starting OD_{600} of 0.002. At 12 hours (note, under these growth conditions, going beyond 12 hours would result in substantially lower RNA integrity), the cultures were centrifuged at 4000 rcf, at 4 °C, for 5 minutes before removing the BHI and resuspending the pellet in 1 mL of RNA preservative (BRP, Qiagen®, RNAprotect™ bacteria reagent, 76506). The pellets were then incubated in the BRP at room temperature for 5 minutes, before being split into thirds and frozen at -80 °C.

Total RNA was extracted from one-third BRP aliquots using an adaptation of the RNAsnap™ protocol.[116] BRP-stored samples were thawed at room temperature, before centrifugation (16,000 rcf, room temperature, 1 minute) and aspiration of BRP. The pellets were then resuspended in 100 μL RNA extraction solution (RES), composed of: 95 % formamide (Millipore®, 344206), 1 % β-mercaptoethanol (Sigma-Aldrich®, BioUltra, 63689), 0.025 % sodium dodecyl sulphate (SDS, Invitrogen™, AM9820), and 18 mM Ethylenediaminetetraacetic acid (EDTA, Invitrogen™, AM9260). The RES resuspensions were then vortexed on high for 10 minutes at 4 °C, followed by 7 minutes at 60 °C on an Eppendorf ThermoMixer® at 800 RPM. The still warm suspensions were then centrifuged (16,000 rcf, room temperature, 5 minutes) and the supernatants were column purified (Zymo Research, RNA Clean & Concentrator 5, RCC-5, R1013) without an on-column DNaseI treatment, and elution in 20 μL ddH_2O. These eluates were then treated with 1 μL TURBO™ DNase (Invitrogen™, AM2238), in 1 x reaction buffer for 30 minutes at 37 °C. After DNase treatment, the samples were column purified once more (RCC-5) with no on-column DNaseI treatment and elution in 10 μL ddH_2O. The resulting DNA-free total RNA was then quantified on a Qubit™ 2.0 spectrophotometer with the RNA high sensitivity kit (Invitrogen™, Q32852), and high RNA integrity was confirmed (Figure S3A) by 1 % agarose, 1 % bleach (The Clorox Company), TAE-EtBr gel electrophoresis per the "Bleach Gel" protocol[117] with ~500 ng loads (and equal volume media control loads).

Genomic DNA (gDNA) was extracted from another set of one-third BRP aliquots using a combined enzyme and detergent lysis, RNA digestion, and column purification (NEB, Monarch®, genomic DNA purification kit, T3010). As before, cells were centrifuged out and BRP was removed. The resulting cell pellets were then resuspended and washed in 900 μL of ice chilled pH 8.0 Tris-EDTA buffer (TE, Sigma-Aldrich®, BioUltra, 93283) before centrifugation, aspiration of the wash TE, and resuspension in 60 μL of fresh pH 8.0 TE. Enzymatic cell lysis was then performed with both 10 μL of T4 lysozyme (NEB, NEBExpress®, P8115) and 10 μL of Lytic Enzyme Solution (Qiagen®, 158928) at 15 minutes at room temperature. Next, 10 μL of proteinase K and 100 μL of Monarch Tissue Lysis buffer (NEB, T3010) was added and incubated at 56 °C for 30 minutes at max ThermoMixer speed. Finally, 10 μL of RNaseA (NEB, T3010) was added and the 56 °C incubation continued for 40 further minutes. Column purification, eluting in 35 μL of 60 °C elution buffer, was then used to purify gDNA per manufacturer specifications (NEB, T3010). The resulting RNA-free gDNA was then quantified on a Qubit™ 2.0 spectrophotometer with the double stranded DNA high sensitivity kit (Invitrogen™, Q32851).

To test for the existence of an Obelisk-*S.s* DNA component, 40 x cycle, duplex RT-PCR and PCR were conducted as before (see above) priming for both the Obelisk-*S.s* and SK36 marker amplicons. 2 ng of gDNA and 2 ng of total RNA (and equal volume media-only and NTC controls) were used as input, with the PCR-only samples being added to the RT-minus SuperMix (NEB, E3010). No igepal was used, and samples were diluted in ddH_2O. The amplicons were then assayed by 4 % agarose, TAE-EtBr gel electrophoresis. SK36 marker amplicons were seen for all cultures (and a faint band for one of the gDNA media controls) and both RT-PCR and PCR-only conditions, whereas, Obelisk-*S.s* amplicons were strictly only seen in the ObP1 RT-PCR samples (Figure S3B).

To obtain full-length Obelisk-*S.s* sequences, each total RNA sample was subjected to two different divergent RT-PCR amplifications (see Figure 5A for primer layout), similar to the above marker RT-PCR. 2 ng total RNA inputs were used with Disc-1-Fw and Disc-2-Fw reverse transcription primers, followed by PCR with their corresponding pairs, Disc-1-Rv, and Disc-2-Rv, respectively (see the key resources table). PCR amplification was performed with an adjusted protocol: 98 °C for 30 seconds, followed by 30 x cycles of: 98 °C for 10 seconds, 67 °C for 20 seconds, 72 °C for 30 seconds; followed by 72 °C for 2 minutes. The PCR products were then assessed by 0.75 % agarose TAE-EtBr gel electrophoresis (Figure S3C). Only the ObP1 samples produced any visible amplicons and these amplicons corresponded to the expected 1137 bp length. A second preparative 0.75 % agarose gel electrophoresis was used to purify the amplicons, extracting the DNA by column purification (Zymo Research, Zymoclean™, D4001), and quantifying the DNA on a Qubit™ 2.0 spectrophotometer with the double stranded DNA high sensitivity kit. The purified amplicons were then sent for full-length, tagmentation-free Oxford Nanopore sequencing (PlasmidSaurus Inc., Premium PCR Sequencing, v14 library chemistry, R10.4.1 flow cell). The consensus amplicon sequences (produced by PlasmidSaurus Inc. using MAFFT[118] and Medaka[119]), were pooled between clones and Disc-pairs, re-oriented to match the sense orientation, and trimmed on either end by 35 nt (seqkit) to remove any influence from the Disc primers. These re-oriented, trimmed sequences were then circularly aligned (MARS and Muscle5), and a global consensus sequence was computed (EMBOSS_cons, default settings). The resulting consensus sequence was a perfect match to Obelisk_000003_000001_000001 in Table S2 despite this assembly process not using this reference sequence. This resulting Obelisk-*S.s* reference sequence was used for all further bioinformatic analysis.

To potentially improve Obelisk-*S.s* coverage, two different enrichment strategies were tried: RNaseH-based ribosomal depletion, and RNaseR-based circular and structured RNA enrichment. SK36 ribosomes were depleted from the total RNA (and volume-matched media-only controls) using 1 μg inputs and following the manufacturer's instructions (NEB, NEBNext®, rRNA bacteria depletion kit, E7860). Alternatively, 5 μg total RNA inputs (and volume-matched media-only controls) were treated with 5 U/μg RNaseR (Biosearch Technologies, RNR07250) in 1 x reaction buffer in 10 μL reaction volumes at 37 °C for 30 minutes, followed by column purification (RCC-5, no on-column DNaseI). The resulting RNA fractions were quantified by Qubit™ 2.0 spectrophotometry with the RNA high sensitivity kit.

All RNA samples (and their corresponding, volume-matched media-only controls) were prepared for RNA sequencing using a strand-specific, random hexamer primed, template-switching -based library preparation method (TaKaRa Bio, SMART-Seq Total RNA Mid Input, 635049, with TaKaRa Bio, NucleoMag™ NGS SPRI beads, 744970). All reactions were conducted at half-volumes relative to the manufacturer's instructions and 25 ng inputs were used throughout. The optional second cDNA cleanup step was performed in all cases. Each type of RNA library was prepared on separate days. The total RNA samples were fragmented for 5 minutes, while the ribosome-depleted and RNaseR samples were fragmented for 3 minutes. All samples were amplified with 12 x cycles of PCR, using pre-mixed unique dual indexing primers (TaKaRa Bio, Unique Dual Index Kit 96U Set B, 634457). The resulting DNA libraries were quantified by Qubit™ 2.0 spectrophotometry with the double stranded DNA high sensitivity kit.

The genomic DNA samples (and their corresponding, volume-matched media-only controls) were prepared for DNA sequencing using a modified tagmentation-based protocol (Illumina, Nextera XT DNA library preparation kit, FC1311024). All reactions were conducted at third-volumes relative to the manufacturer's instructions and 1 ng inputs were used throughout. The tagmentation step was performed at 37 °C for 5 minutes, and 12 x cycles of indexing PCR were performed using pre-mixed unique dual indexing primers (Illumina, DNA/RNA UD Indexes Set C, 20091648). The resulting DNA libraries were size selected by 1 % agarose, TAE-EtBR gel electrophoresis, cutting at the 300 to 650 bp range. DNA was extracted from the gel and quantified as before.

The resulting 36 libraries (9 x per condition, 4 x conditions) were then pooled in equimolar ratios (assuming a 300 bp average insert size for the RNA samples, and a 450 bp insert size for the DNA samples, using the media-only libraries as diluent) and loaded at 12 pM on a MiSeq v3 300 cycle kit.

SK36 ObN1 and ObP1 sequencing analysis

The resulting short read sequencing data were then used to: (a) test the presence of Obelisk-*S.s* in ObN1 and ObP1; (b) gauge the molecular impact of Obelisk-*S.s* presence on SK36 RNA expression profiles; and (c) gauge the completeness of coverage on both strands for Obelisk-*S.s* - all in the context of 12 hours of aerobic growth in BHI. All short read sequencing data were first adapter- and quality- trimmed as before (fastp, no length limit). First, to account for any genetic changes accumulated during the passaging and screening for the ObN1 and ObP1 strains, the per-strain DNA-seq reads were mapped to the established SK36 reference genome[45] (GenBank: CP000387.1) and clonal mutations were marked (breseq, default settings). Then, all shared mutations over the six strains were extracted and applied to the reference genome to yield a new bespoke reference genome (gdtools intersect, apply, and annotate). This "SK36-Ob" reference genome was used for all further analysis. By manually inspecting these breseq-identified variants by read pileup inspection, with VariantRabbit (default settings), and with SPLASH2 (default settings, corrected p-value < 0.05, effect size >= 0.95); six confident single nucleotide polymorphisms (SNPs) were identified between ObP1 and ObN1, with 3 SNPs per substrain (provided in Table S1 and the key resources table).

Obelisk-*S.s* presence and strand-specific coverage was assessed by k-mer pseudoalignment to both the SK36-Ob transcriptome, and both the Obelisk-*S.s* sense and antisense strands from the Nanopore-derived reference sequence. Prior to building the k-mer index (kallisto index, default settings), in order to account for the apparent circular nature of Obelisk-*S.s*, the reference sequences were each extended by 30 nt. K-mer pseudoalignment was performed while enforcing forward-stranded mapping (not enforced for DNA-seq data), with 1000 bootstraps (kallisto quant, –fr-stranded, –bootstrap-samples=1000). Obelisk-*S.s* mapping reads were only seen in the ObP1 RNA-seq datasets. Differential expression analysis was performed on the effective counts (grouping over sense and antisense Obelisk-*S.s* counts) testing the ObN1 and ObP1 biological replicates against each other using both a likelihood ratio test and a Wald test (sleuth), a false discovery rate -adjusted p-value (q-value) threshold of 0.05 was used to score significant differential expression, and mean $\log_2$ fold change was computed with a $+10^{-4}$ offset on transcript per million counts to account for the zero-count representation of Obelisk-*S.s* in the ObN1 samples. Strand bias and position-wise Obelisk-*S.s* coverage was assessed by k-mer pseudoalignment using a k-mer mapping tool that is tolerant of SNPs (PolyBench)[120] while accounting for the apparent circularity of the Obelisk-*S.s* reference (Circular=True), and filtering for high-confidence stranded reads by filtering for expected template-switching sequence-end artefacts (SmarterStrandedFilter=True). To gauge the overall stranded-ness of the RNA-seq data, k-mer pseudoalignment was also performed against a 16S rRNA locus (SSA_2400, Circular=False).

The presence of RNA-dependent RNA polymerases (RdRPs) in all short read sequencing data was assessed by mapping translated reads against RdRP pHMMs (as above, Pfam models RdRP_1 through RdRP_5). No mapping translated reads (at any E-value threshold) were found. Lastly, Obelisk-*S.s* sequence polymorphisms in the RNA-seq data were measured using Breseq and PolyBench and are provided in the key resources table. >99.9% of total RNA bases (with PolyBench) retain the consensus assembled sequence.

Supplemental figures

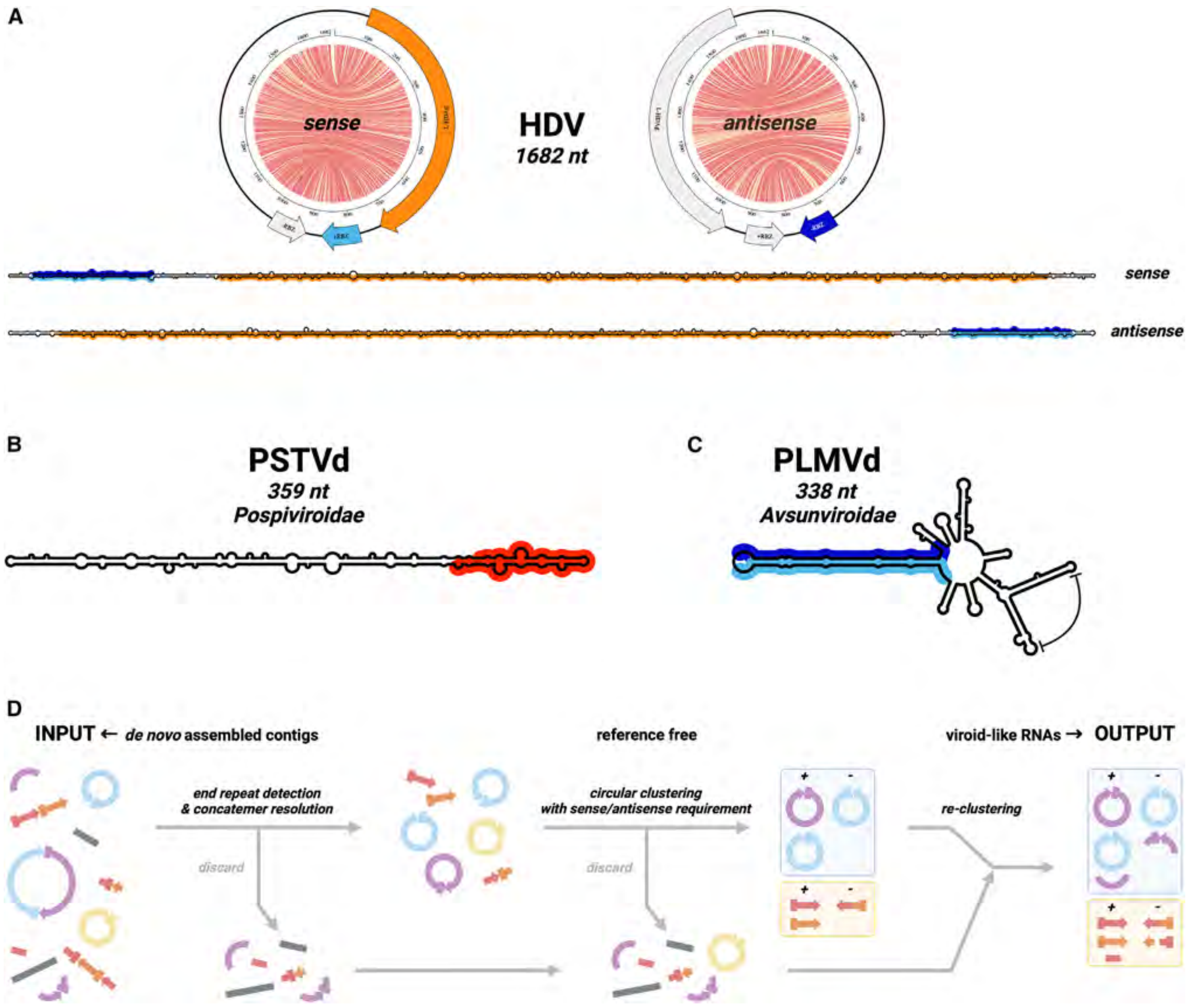


Figure S1. Background on viroid and HDV families and VNom, related to method details

(A) The hepatitis delta virus (HDV) genome (NC_001653.2)[121] is predicted to fold into a rod-shaped RNA secondary structure in both sense, and antisense—depicted here as both jupiter plots where chords represent predicted base pairs (colored by base pair probability from 0, gray, to 1, red) with features grayed out in antisense, and "skeleton" diagrams. Large hepatitis delta antigen (L-HDAg, orange), and hepatitis delta ribozymes (RBZ, Rfam: RF00094, antisense: dark blue, sense: light blue) indicated.

(B) Potato spindle tuber viroid (PSTVd) of the family *Pospiviroidae* folds[122] into a rod-like RNA secondary structure similar to HDV but encodes no ORFs, though does possess a conserved Pospiviroid RY motif (Rfam: RF00362, red).

(C) Peach latent mosaic viroid (PLMVd) folds[123] into a highly base paired, but "branched" RNA secondary structure as is characteristic of the *Avsunviroidae* family. Type-III hammerhead ribozymes (Rfam: RF00008, antisense: dark blue, sense: light blue) and "P8" pseudoknot (curved flat-headed arrow) illustrated.

(D) VNom (short for "viroid nominator," pronounced *venom*) attempts to enrich for RNAs that are apparently circular and are present in the dataset in both polarities (a hallmark of RNA replication). To do this, VNom takes in *de novo* De Bruijn graph assembled contigs (from stranded RNA-seq data) and filters for potentially circular contigs by looking for perfect k-mer matches at the ends of each contig. Further, VNom also attempts to resolve concatemeric contigs by looking for regular repetition of such identified k-mers. These potentially circular contigs are then clustered based on sequence similarity using a circularly-permuting clustering algorithm. These resulting clusters are then kept if at least one contig of each polarity is identified by k-mer counting. Finally, these filtered clusters are compared against all of the previously discarded contigs to identify any remaining cluster members. While these filters should enrich for viroid-like RNAs, highly repetitive sequences also satisfy these requirements and so are often also enriched. VNom was found to work adequately well on deeply sequenced viroid-positive plant RNA-seq datasets (e.g., SRR11060618, SRR11060619, SRR11060620, and SRR16133646), especially when assemblies from the same bioProject were grouped together.

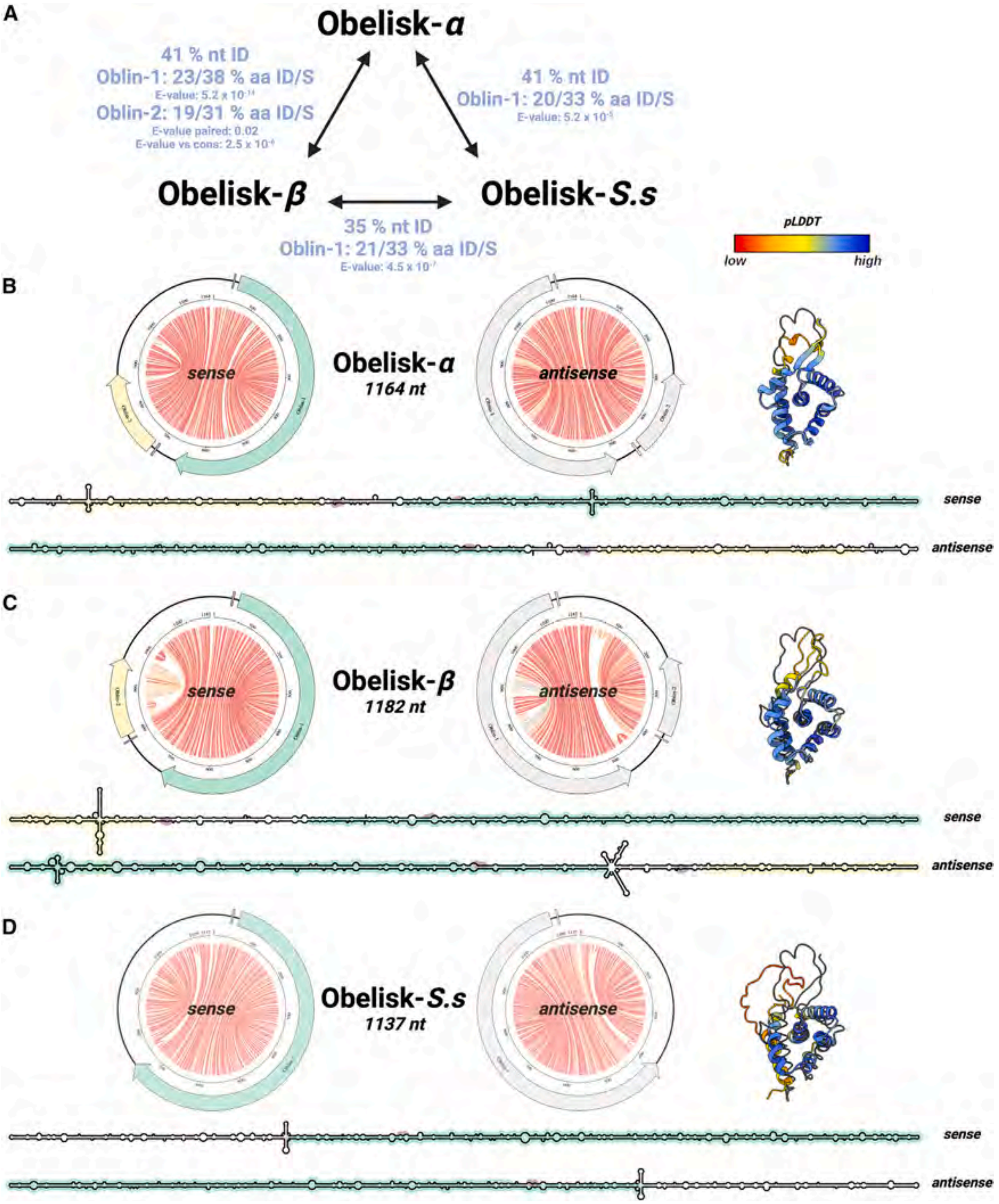
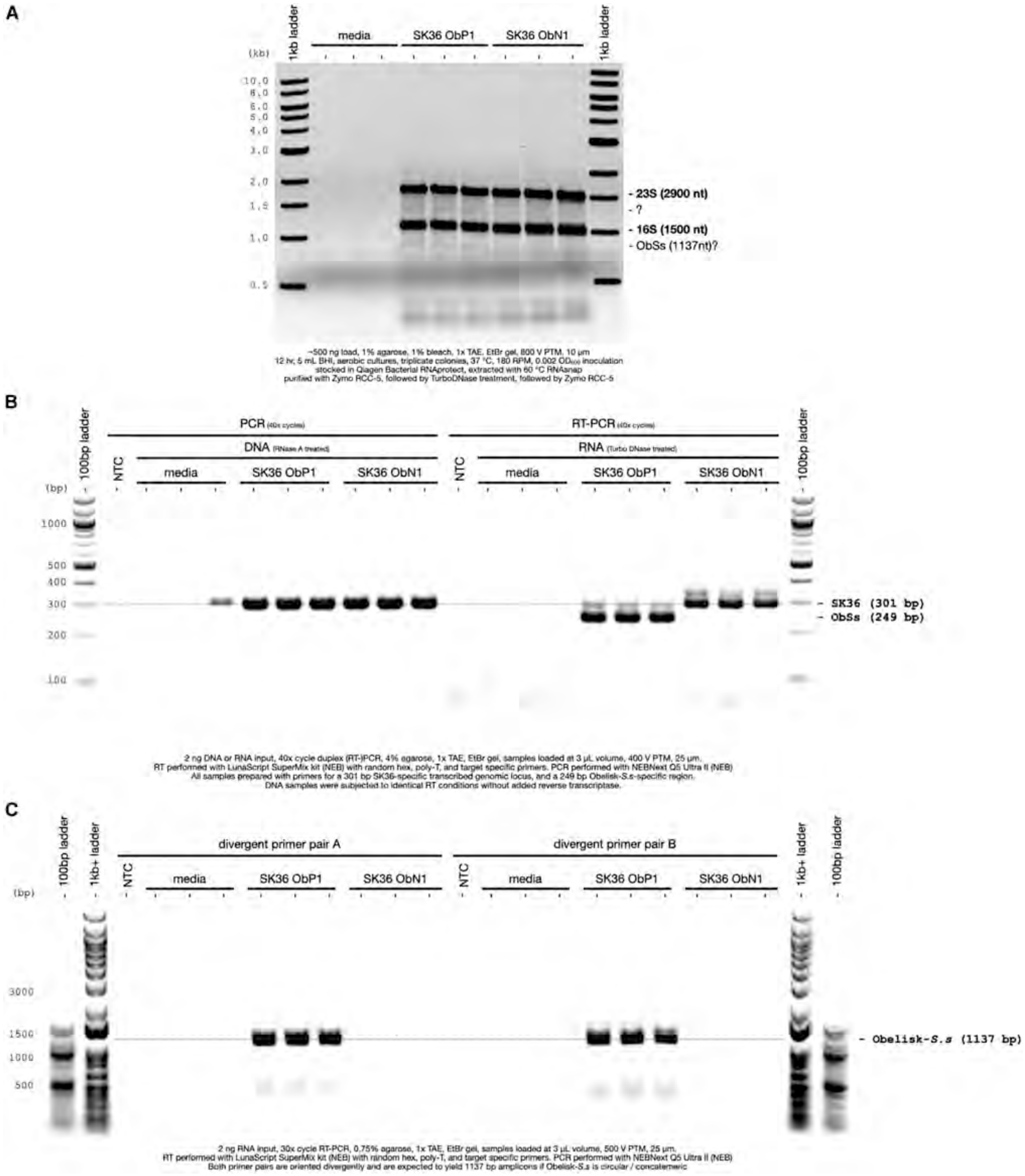


Figure S2. Obelisks-alpha, -beta, and -*S. sanguinis* appear to belong to the same, diverse family, related to Figure 1

(A) Nucleotide (nt) and amino acid (aa) -level pairwise sequence identities (ID) and similarities (S) between obelisks-α, -β, and -*S.s*. For Oblin protein sequences, mean pairwise BLASTp E values are shown. Note, for Oblin-2 the pairwise BLASTp E value relative to the Oblin-2 consensus (see STAR Methods) is also shown, indicating a distant, but evident homology between the α and β Oblin-2s.

(B–D) These obelisks are similar in lengths; 1,164, 1,182, and 1,137 nt, respectively, and share globally similar obelisk-like predicted RNA secondary structures in both their sense and antisense—depicted here as both jupiter plots where chords represent predicted base pairs (colored by base pair probability from 0, gray, to 1, red) with features grayed out in antisense, and skeleton diagrams. Likewise, the genomic synteny of predicted ORFs (preceded by predicted Shine-Dalgarno sequences, purple) appear to be shared, with Oblin-1 (green) consistently being present on one half of the predicted RNA secondary structure, and Oblin-2 (yellow), when present, following shortly after Oblin-1. ColabFold predictions of Oblin-1 tertiary globule structures built with *ad hoc* MSA construction (colored cartoons) superimposed over the RDVA-derived MSA prediction for obelisk-α (black line, Figure 2A, see STAR Methods) indicating a conserved tertiary structure. Prediction confidence (pLDDT) shown as a color bar (low confidence: 0, red; high confidence: 100, blue).



~500 ng load, 1% agarose, 1% bleach, 1x TAE, EtBr gel, 800 V PTM, 10 µm
12 hr, 5 mL BHI, aerobic cultures, triplicate colonies, 37 °C, 180 RPM, 0.002 OD_{600} inoculation
stocked in Qiagen Bacterial RNAprotect, extracted with 60 °C RNAsnap
purified with Zymo RCC-5, followed by TurboDNase treatment, followed by Zymo RCC-5

2 ng DNA or RNA input, 40x cycle duplex (RT-)PCR, 4% agarose, 1x TAE, EtBr gel, samples loaded at 3 µL volume, 400 V PTM, 25 µm.
RT performed with LunaScript SuperMix kit (NEB) with random hex, poly-T, and target specific primers. PCR performed with NEBNext Q5 Ultra II (NEB)
All samples prepared with primers for a 301 bp SK36-specific transcribed genomic locus, and a 249 bp Obelisk-*S.s*-specific region.
DNA samples were subjected to identical RT conditions without added reverse transcriptase.

2 ng RNA input, 30x cycle RT-PCR, 0.75% agarose, 1x TAE, EtBr gel, samples loaded at 3 µL volume, 500 V PTM, 25 µm.
RT performed with LunaScript SuperMix kit (NEB) with random hex, poly-T, and target specific primers. PCR performed with NEBNext Q5 Ultra II (NEB)
Both primer pairs are oriented divergently and are expected to yield 1137 bp amplicons if Obelisk-*S.s* is circular / concatemeric

Figure S3. Obelisk-*S.s* correlates with an extra band in total RNA and does not have a detectable DNA component, related to Figure 4

(A) "Bleach gel"[117] electrophoretogram of triplicate total RNA extractions from cultures (see STAR Methods) of obelisk-positive 1 (ObP1), a strain of *Streptococcus sanguinis* SK36 that harbors obelisk-*S.s* ("*Obelisk_000003*" in Table S2), and obelisk-negative 1 (ObN1), indicating a ObP1-specific extra band running beneath the apparent 16S band.

(legend continued on next page)

(B) Agarose gel of amplicons from PCR and RT-qPCR probing for markers for a transcribed SK36 genomic fragment (SSA_0213 gene; 301 bp expected fragment size), and obelisk-*S.s* (249 bp expected fragment size). DNA samples show amplification from RNaseA-treated DNA from ObP1 and ObN1, while RNA samples show amplification from TURBO-DNase-treated SK36 total RNA (see STAR Methods, primers provided in the key resources table, and illustrated on Figure 5A).
(C) Agarose gel of amplicons from two different sets of "divergent" RT-PCR experiments targeting the full length of obelisk-*S.s* which are expected to be produced if obelisk-*S.s* is either circular and/or concatemeric in nature. All raw gel images are linked in the key resources table.

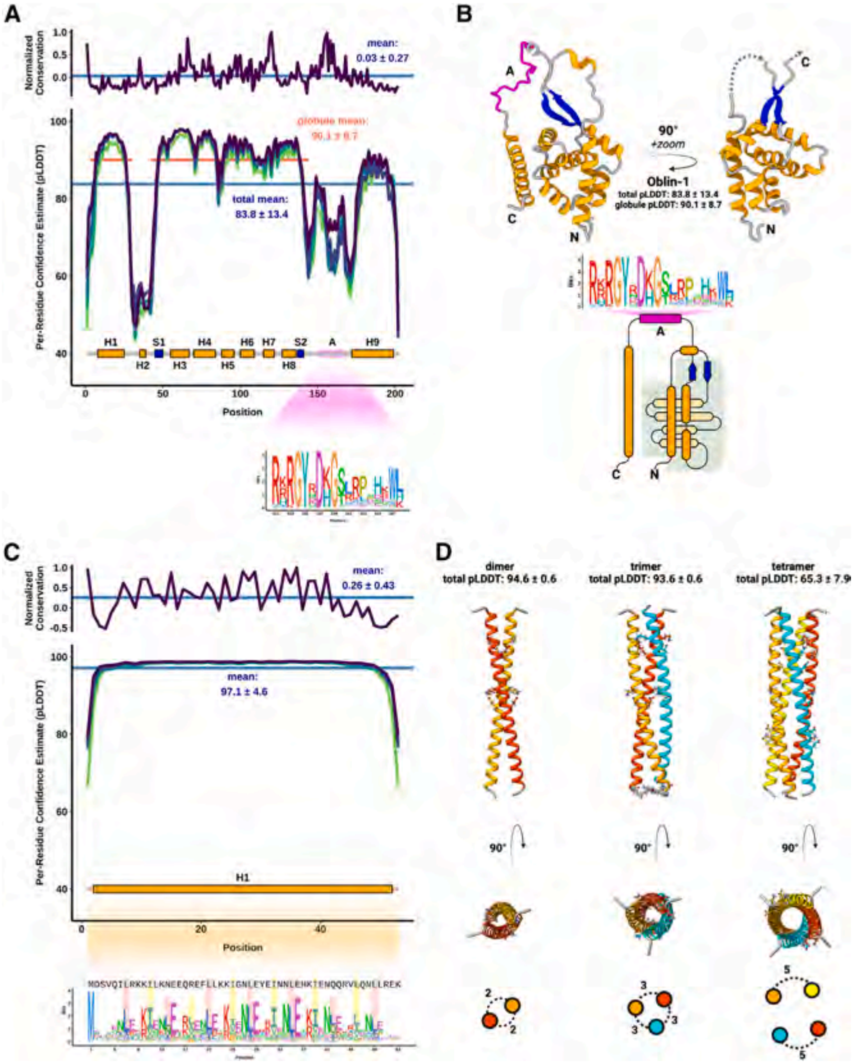


Figure S4. Oblins are diverse and generate robust protein fold predictions, related to Figure 2

(A) Normalized conservation (top, above zero = more conserved, see STAR Methods) of obelisk open reading frame 1 (Oblin-1) relative to obelisk-ɑ indicates that Oblin-1 is largely poorly conserved (mean per-residue confidence estimate, μ-pLDDT ± standard deviation of 0.0 ± 0.3) but has three regions of conservation, around the C termini of alpha helices 3 and 7, and *domain-A* (see sequence logo callout, bottom). Oblin-1 tertiary structure prediction per-residue confidence estimate (bottom, see STAR Methods) suggests a medium confidence total fold (μ-pLDDT: 83.8 ± 13.4), and a high confidence N-terminal globule (μ-pLDDT: 90.1 ± 8.7) that is consistently predicted over the top five models (green lines). *domain-A* is consistently predicted without a confident tertiary structure.

(B) Top: tertiary structure representation of the predicted obelisk-ɑ globule fold. Bottom: a to-scale (secondary structure) topological representation of Oblin-1 with the globule shaded in gray, and the *domain-A* emphasized with this bit-score sequence logo (see STAR Methods).

(C) Obelisk Oblin-2 has a higher mean normalized conservation (top, 0.26 ± 0.43), and is confidently predicted to form an alpha helix (μ-pLDDT: 97.1 ± 4.6). The Oblin-2 sequence logo (callout, bottom) shows leucine zipper features with i+7 leucine spacing emphasized in red, with hydrophobic “d” position residues emphasized in yellow (obelisk-ɑ Oblin-2 sequence shown for reference). Obelisk-ɑ alpha helices (orange boxes, “H” labels), and beta sheets (blue boxes, “S” labels) illustrated for clarity.

(D) Tertiary structure predictions of obelisk-*alpha* open reading frame 2 (Oblin-2) homo-multimers: left: dimer (mean pLDDT ± standard deviation: 94.6 ± 0.6), middle: trimer (mean pLDDT: 93.6 ± 0.6), and right: tetramer (mean pLDDT: 65.3 ± 7.9). Residues involved in inter-helix salt bridges emphasized, and salt bridge counts illustrated on bottom.

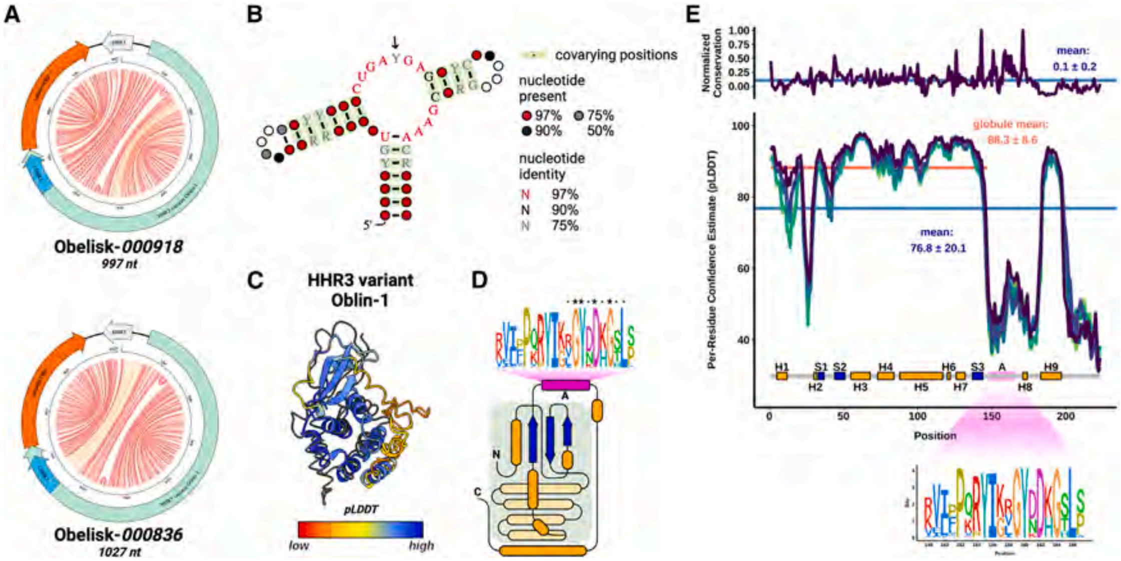


Figure S5. Ribozyme-bearing obelisks encode a diverged Oblin-1, related to Figure 3

(A) Two "obelisk-variant hammerhead type-III" (ObV-HHR3) -positive obelisk genomes from Table S2, illustrated as jupiter plots where chords represent predicted base pairs (colored by base pair probability from 0, gray, to 1, red), Oblin-1 homologs illustrated in green, smaller, non-Oblin-2 ORFs in orange, and sense ObV-HHR3 in blue (with antisense ObV-HHR3 in gray). Note the conspicuous placement of ObV-HHR3 relative to Oblin-1 and the smaller ORF.

(B) The RDVA-derived, stringently-thresholded ObV-HHR3 covariance model summarized as a secondary structure with base pair-forming, significantly covarying positions indicated with a green highlight. IUPAC "ambiguity codes"[124] used to represent RNA diversity: Y = U or C, R = A or G.

(C) ColabFold prediction of the "HHR-variant" Oblin-1 tertiary ("*Obelisk_000918*" as the reference sequence) structure built with a custom MSA construction (colored cartoons) superimposed over the RDVA-derived MSA prediction for obelisk-α where possible (black line, Figure 2A, see STAR Methods). Prediction confidence (pLDDT) shown as cartoon coloring as in Figure S2.

(D) A to-scale (secondary structure) topological representation of HHR-variant Oblin-1 with the globule shaded in gray (as in Figure 2B), and the *domain-A* emphasized with this bit-score sequence logo (see STAR Methods). Conserved GYxDxG motif emphasized.

(E) Normalized conservation (top, above zero = more conserved, see STAR Methods) of "obelisk-variant hammerhead type-III" (ObV-HHR3) "HHR3-variant" Oblin-1 indicates that, similarly to the non-HHR3 Oblin-1 (Figure S4), the HHR3-variant Oblin-1 is largely poorly conserved (mean normalized conservation ± standard deviation: 0.1 ± 0.2) but retains a conserved *domain-A* (see sequence logo callout, bottom). HHR3-variant Oblin-1 tertiary structure prediction per-residue confidence estimate (bottom, see STAR Methods) suggests a medium confidence total fold (mean per-residue confidence estimate, μ-pLDDT ± standard deviation of 76.8 ± 20.1), and a higher confidence N-terminal globule (μ-pLDDT: 88.3 ± 8.6) that is consistently predicted over the top five models (green lines). *domain-A* is consistently predicted without a confident tertiary structure.

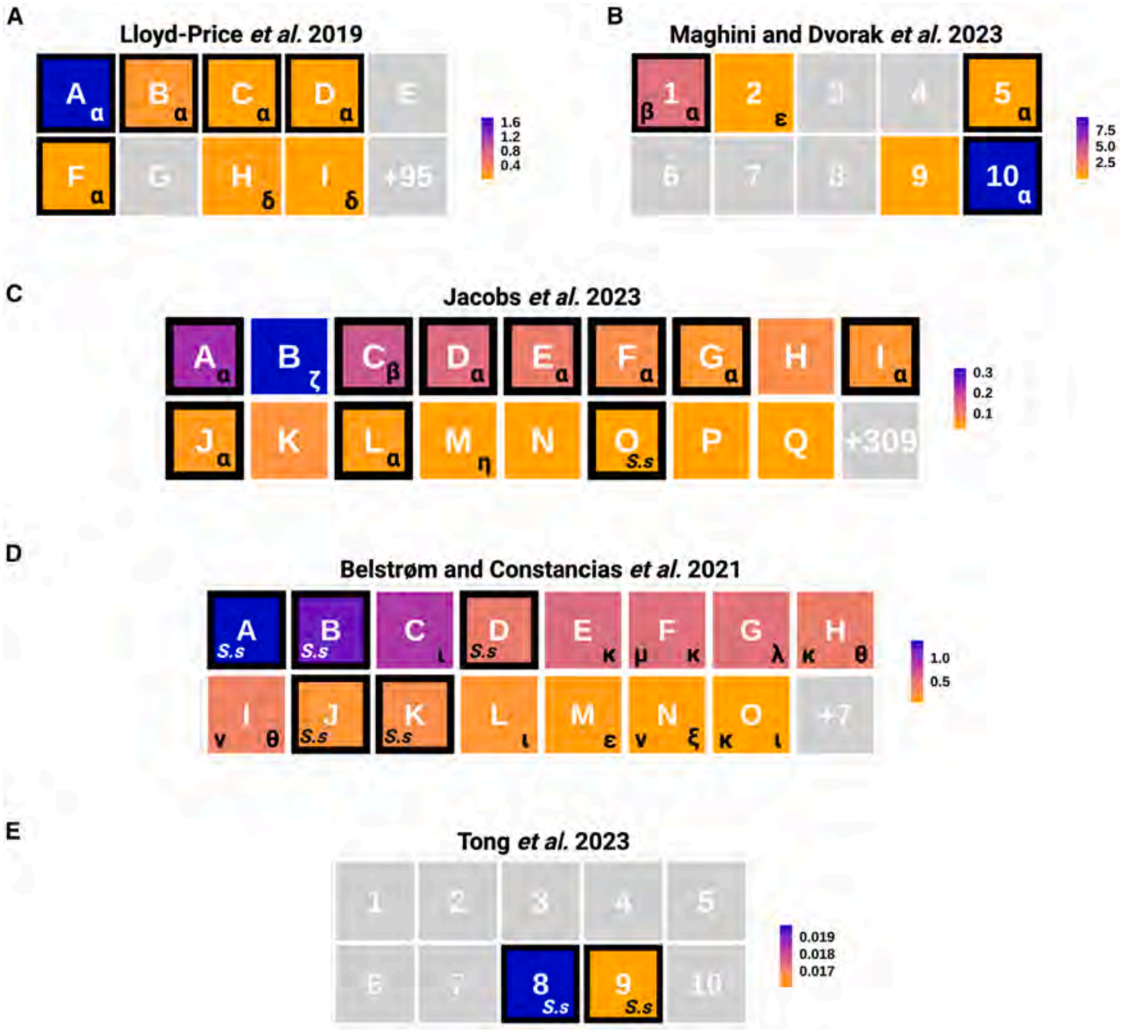


Figure S6. Human gut and oral microbiomes harbor diverse obelisks, related to Figure 6

Heatmaps of obelisk-positive donors (>10 reads, averaged over donor if multiple samples) as inferred by k-mer and Oblin-1 pHMM matching (see STAR Methods and Table S1, donors with complex internal nomenclature were renamed for clarity see Table S1). Samples emphasized with black boxes were k-mer positive (but not exclusively). Lowercase Greek lettering indicate which obelisks were found in a given donor as inferred by either k-mer counting (black boxes—k-mer profiling obelisks -α, -β, and -*S.s*), or by *post hoc* classification of newly assembled and independently clustering obelisks (see STAR Methods). Human gut microbiome samples: (A) Lloyd-Price et al.[20], (B) Maghini and Dvorak et al.[103], and (C) Jacobs et al.[113] Human oral microbiome samples: (D) Belstrøm and Constancias et al.[38], and (E) Tong et al.[114]. Color scales indicate obelisk read counts relative to total donor reads $\times 10^{-4}$. Greek letter key: α: alpha, β: beta, δ: delta, ε: epsilon, ζ: zeta, η: eta, θ: theta, ι: iota, κ: kappa, λ: lambda, μ: mu, ν: nu, and ξ: xi. obelisks diagrammed in Figure 6.