

Article

A Computational Framework for Synthetic Minimal Bacteriophage Gene Set Design: Application to *Streptococcus mutans* and Codon-Optimized Expression in *Escherichia coli*

Abdulaziz Zailaa and Diego Cotella *

Department of Health Sciences, School of Medicine, Università del Piemonte Orientale, Novara 28110, Italy; 20063425@studenti.uniupo.it

* Correspondence: diego.cotella@uniupo.it

Abstract

Minimal gene set design, reducing a system to its smallest functional part set, is a core synthetic biology strategy. Bacteriophages are attractive but underused chassis, yet wild-type phages are difficult to engineer. We present a comparative genomics-guided computational framework for synthetic minimal phage gene set design, demonstrated in silico on *Streptococcus mutans*, the primary agent of dental caries. Three *S. mutans* phages (M102AD, ϕ APCM01, SMHBZ8) were compared by synteny analysis, and their 123 proteins were annotated with InterProScan, EggNOG-mapper, and HHpred. Conserved essential genes were selected by cross-phage conservation and annotation evidence, assembled as standardized transcriptional units, codon-optimized for *E. coli* K-12 with a modern folding-aware optimizer, and cloned in silico into the broad host range vector pBBR1MCS-2. Seventeen genes were assembled into a 16,403 bp synthetic insert; the final 21,559 bp construct raised the mean codon adaptation index (CAI) from 0.58 to 0.96, brought GC content into the optimal range, and eliminated rare codons, while removing all internal cloning sites. A 5' mRNA-folding screen confirmed accessible ribosome binding sites for all 17 genes; a sensitivity analysis confirmed that the gene set was robust to the selection thresholds; and AlphaFold-Multimer modeling of representative inter-module interfaces is reported. Multi-host CAI analysis indicates strong codon usage compatibility with Enterobacteriaceae hosts. We emphasize that the construct is a computationally designed multi-gene expression plasmid, not a self-replicating or packageable phage genome, and that all results are in silico predictions requiring experimental validation. Gibson Assembly, not restriction cloning, is recommended. The design logic is transferable to other phage–host systems.

Keywords: synthetic biology; minimal gene set design; genetic parts and chassis; bacteriophage engineering; *Streptococcus mutans*; comparative genomics; codon optimization; *Escherichia coli* expression; broad host range vector

Academic Editor: Mu Li

Received: 11 August 2026

Revised: 25 September 2026

Accepted: 28 September 2026

Published: 1 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

A central goal of synthetic biology is to make biological engineering predictable: to decompose genetic systems into well-characterized, reusable parts (promoters, ribosome binding sites, terminators, and coding sequences) and recombine them into standardized devices with defined function [1]. Minimal genome design, reducing a biological system

to the smallest gene set that still performs a target function, is one of the field's clearest expressions of this logic and has been demonstrated experimentally in free-living cells, from JCVI-syn1.0, the first cell built from a chemically synthesized genome [2], to the further streamlined JCVI-syn3.0, comprising just 473 genes, about a third of them of unknown function [3]. Minimal systems are easier to manipulate genetically, more genomically stable, faster to replicate, and simpler to characterize, and genome reduction has, for example, increased secondary metabolite yields in engineered *Streptomyces avermitilis* [4]. Bacteriophages are an attractive but comparatively underused chassis for this strategy; their genomes are naturally small, their gene content is organized into discrete functional modules (DNA packaging, head assembly, tail, lysis, and replication), and their well-defined lytic function provides an unambiguous design target, all properties well suited to rational, parts-based reconstruction.

In practice, however, engineering a minimal phage genome directly from a single wild-type isolate is difficult: phage genomes are compact and gene-dense, many genes have no experimentally characterized function, and mutating or deleting genes in the native context can be technically demanding and difficult to interpret. A computational, comparative genomics-guided design strategy offers an alternative: instead of reducing one genome by trial and error, essential functions can be identified from patterns of conservation across multiple related phages, the best-annotated version of each function can be selected on the basis of structural and orthology evidence, and the resulting parts can be reassembled de novo as independent, standardized transcriptional units, each with its own promoter, ribosome binding site, and terminator, in a chassis-independent expression vector. This reframes minimal phage genome design as a design-build-test problem amenable to the same computational tools used elsewhere in synthetic biology (comparative genomics, functional annotation, codon optimization, host range prediction) and is the approach developed and demonstrated in this work.

A complementary, more recently emerged strategy for computational phage design bypasses comparative annotation altogether by using genome-scale generative language models. King et al. fine-tuned the Evo 1 and Evo 2 genome language models on ~15,000 *Microviridae* genomes and used them to generate complete, novel ΦX174-like genomes de novo, filtering outputs with tropism and diversification constraints before experimental synthesis [5]. Of 285 synthesized candidates, 16 proved viable in *E. coli* C, several with higher fitness or faster lysis kinetics than wild-type ΦX174, and a cocktail of the generated phages rapidly overcame phage-resistant bacterial strains where wild-type ΦX174 alone could not. This work establishes, for the first time, that AI-based generative design can produce functional, viable bacteriophage genomes from scratch, rather than by recombining natural parts as in the framework presented here. The two approaches are complementary: comparative genomics-guided, part-based design (as used in this study) offers interpretability and direct traceability of each gene to a known, annotated function, which is valuable for uncharacterized phage systems such as those infecting *S. mutans* with few sequenced relatives; generative genome language modeling, by contrast, requires larger training corpora of related genomes but can access far greater sequence novelty and has already demonstrated end-to-end experimental viability. Future work combining conservation-guided part selection with generative infilling or scoring by genome language models may offer the best of both strategies.

We selected the *Streptococcus mutans* phage system as a test case for this framework because it combines a clear application need with a technically demanding engineering starting point. *S. mutans* is the primary etiological agent of dental caries, one of the most common diseases in humans, in which the organism's biofilm shields it from chemical and antibiotic intervention and drives increasing antibiotic resistance [6–9]. Phage-based approaches are attractive against such biofilms: phages are highly host-specific, self-

amplifying, and able to penetrate biofilm matrices that are otherwise antibiotic-resistant [10,11]. A recent study showed that phages can reduce *S. mutans* colony-forming units and both prevent and reduce existing biofilms, but it also underscored how difficult it is to isolate and directly engineer wild-type *S. mutans* phages [12], illustrating precisely the bottleneck that a computational, chassis-independent design framework is intended to bypass. Only four *S. mutans*-specific phages have been isolated to date (M102, M102AD, ϕ APCM01, and SMHBZ8), reflecting the broader difficulty of isolating phages from the complex oral microbiome [12]; three of these, M102AD, ϕ APCM01, and SMHBZ8, share a common ancestry and sufficient genomic similarity to support a comparative design approach, with genome lengths of 30,664 bp (M102AD, DQ386162), 31,075 bp (ϕ APCM01, NC_029030), and 32,460 bp (SMHBZ8, MT430910).

Realizing this framework computationally required combining several complementary bioinformatic layers. Multi-tool functional annotation (InterProScan, EggNOG-mapper, and HHpred) was used to assign confidence-scored functions to phage proteins that are often poorly represented in reference databases, providing the evidence base for essentiality classification and part selection. Standardized regulatory elements (promoter, ribosome binding site, and terminator) were then used to make each selected gene an independent expression unit, decoupled from its native genomic context. Codon optimization (JCat) and multi-host codon adaptation index analysis were used to tune and then evaluate expression potential across a chassis vector documented to replicate in multiple Gram-negative hosts, treating host range itself as a design variable rather than a fixed constraint [13–18]. Together, these tools make it possible to carry out essentially the whole design cycle for a synthetic minimal phage genome in silico, before committing to DNA synthesis and experimental build-and-test.

The aim of this work was therefore twofold: to develop a generalizable, comparative genomics-guided framework for designing a synthetic minimal bacteriophage genome from multiple related source phages and to demonstrate that framework computationally on the *S. mutans* phage system, identifying the essential genes conserved across three source phages, assembling them into a codon-optimized expression plasmid for *E. coli* and other Gram-negative hosts, and validating the resulting design through multiple complementary computational approaches.

2. Results

2.1. Phage Genome Comparison and Protein Annotation

A Clinker synteny analysis [14] comparing M102AD, ϕ APCM01, and SMHBZ8 confirmed high genomic conservation, with pairwise sequence identity of 85–98% for connected gene pairs and a shared canonical phage genome architecture, characterized by conserved gene order across the DNA-packaging, head assembly, tail, and lysis modules (Figure 1). This result validated the use of cross-phage gene conservation as a criterion for identifying core essential genes.

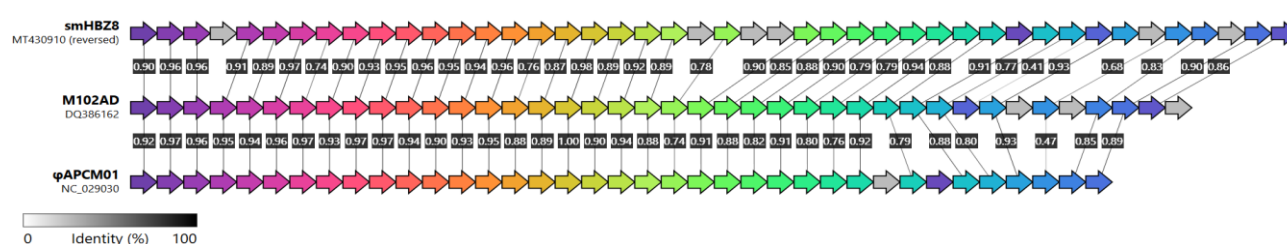


Figure 1. Clinker synteny map comparing the three *S. mutans* bacteriophage genomes M102AD, ϕ APCM01, and SMHBZ8.

A total of 123 protein-coding sequences were identified across the three genomes (42 from M102AD, 37 from ϕ APCM01, and 44 from SMHBZ8) and processed through a three-tool annotation pipeline. HHpred gave the broadest coverage, annotating 121/123 proteins (98.4%) by structural homology; InterProScan annotated 90/123 proteins (73.2%); and EggNOG-mapper annotated 73/123 proteins (59.3%). Only 49 proteins (39.8%) were annotated by all three tools simultaneously, confirming that a multi-tool strategy was necessary to maximize functional coverage of this poorly characterized phage proteome.

Each protein was then classified into one of four essentiality categories using keyword- and COG-based rules (Table 1). Forty-four proteins (35.8%) were classified as essential based on core structural/functional keywords (e.g., terminase, capsid, tail, holin, and endolysin); six (4.9%) were classified as likely essential based on COG category L (replication, recombination, and repair); 27 (21.9%) were classified as non-essential accessory or mobile element genes; and 46 (37.4%) remained of unknown function, consistent with the generally poor functional characterization of phage proteomes.

Table 1. Distribution of essentiality classifications across 123 proteins from three phages (M102AD, ϕ APCM01, and SMHBZ8).

Essentiality Category	Number of Proteins	Percentage
Essential	44	35.8%
Likely essential	6	4.9%
Non-essential	27	21.9%
Unknown	46	37.4%
Total	123	100%

2.2. Core Gene Selection

Of the 123 proteins, 50 were classified as essential or likely essential (Table 1). Conservation analysis across all three phages identified 58 conserved gene copies, that is, instances of a given functional label present in all three genomes simultaneously (33 essential, 6 likely essential, and 19 conserved genes of unknown function); this figure counts gene copies rather than unique functions, since each conserved function is represented once per phage, and not all 50 essential/likely essential proteins were conserved across all three phages. Genes of unknown function that were nonetheless conserved across all three independently isolated phages were retained, on the reasoning that such conservation is itself strong evidence of functional importance, a position also supported by the JCVI-syn3.0 precedent, in which roughly one third of essential genes had unknown function [3].

For each conserved function, the best-scoring gene version (by HHpred probability, with EggNOG score as a fallback) was retained, yielding a final set of 17 core genes, comprising 10 essential, 2 likely essential, and 5 representatives of previously unknown functions: seven from M102AD, five from ϕ APCM01, and five from SMHBZ8 (Table 2). The selection funnel is therefore 123 proteins $\rightarrow$ 50 essential/likely essential $\rightarrow$ 58 conserved gene copies $\rightarrow$ 17 unique functions, each selected as the single best-scoring representative of its function. Not every unique conserved function present among the 58 gene copies was carried forward into the final 17: some were excluded because their PDB-based structural annotation was redundant with a better-scoring representative of an equivalent function that was already retained. One such excluded function was a conserved endolysin, which is why the final gene set retains a holin but no dedicated endolysin (Section 3.4). The majority of genes with HHpred scores at or above 97% belong to essential structural categories (terminase, portal, capsid, tail, and single-strand binding protein), confirming reliable annotation of the most critical phage components; six genes showed lower

HHpred confidence (20.7–40.4%), reflecting the limited structural characterization of these proteins in current databases, but were retained based on their cross-phage conservation.

Table 2. The 17 core essential genes selected for the synthetic minimal phage genome, ordered by biological module position.

#	Gene ID	Source Phage	Function	Module	HHpred Score	Essentiality
1	phage2_AVT41_gp03	φAPCM01	Terminase large subunit	DNA packaging	100.0	Essential
2	phage3_unknown38	SMHBZ8	DNA packaging protein	DNA packaging	99.5	Essential
3	phage1_unknown4	M102AD	Portal protein	Head assembly	99.8	Essential
4	phage1_unknown6	M102AD	Major capsid protein	Head assembly	99.5	Essential
5	phage3_unknown32	SMHBZ8	Tail tube protein	Tail	97.7	Essential
6	phage1_unknown9	M102AD	HK97-gp10 tail component	Tail	40.4	Essential
7	phage3_unknown31	SMHBZ8	Distal tail (peptidoglycan hydrolase domain)	Tail	99.9	Essential
8	phage2_AVT41_gp15	φAPCM01	Siphovirus protein (DUF859)	Tail	82.4	Essential
9	phage2_AVT41_gp18	φAPCM01	Holin (LL-H family)	Lysis	31.3	Essential
10	phage2_AVT41_gp22	φAPCM01	IstB-like ATP-binding protein (putative; possible mobile element association)	Replication	98.6	Essential
11	phage1_unknown33	M102AD	Single-strand binding protein	Replication	99.8	Likely essential
12	phage1_unknown1	M102AD	HNH endonuclease	Replication	97.5	Likely essential
13	phage1_unknown26	M102AD	DNA metabolic process	Replication	28.0	Unknown
14	phage2_AVT41_gp10	φAPCM01	Unknown (DUF806)	Conserved unknown	35.6	Unknown
15	phage1_unknown27	M102AD	Unknown (DUF1351)	Conserved unknown	22.2	Unknown
16	phage3_unknown21	SMHBZ8	Calcium ion binding protein	Conserved unknown	26.0	Unknown
17	phage3_unknown23	SMHBZ8	Mannosyl-glycoprotein endo-N-acetylglucosaminidase	Conserved unknown	20.7	Unknown

2.3. Synthetic Genome Assembly

The 17 selected genes were assembled in canonical biological module order, DNA packaging, head assembly, tail, lysis, and replication modules, followed by conserved genes of unknown function, into a synthetic gene set of 16,403 bp. Each gene was flanked by its own sigma70 promoter, Shine–Dalgarno ribosome binding site with a 6 bp spacer, and one of a panel of four strong, characterized rho-independent terminators (Table 3), ensuring independent transcription and translation of each coding unit and avoiding read-through between adjacent genes.

Table 4. Codon optimization results for all 17 phage genes before and after DNA Chisel optimization targeting E. coli K12 (CAI reported as true Sharp & Li index).

Gene	CAI Before	CAI After	GC% Before	GC% After	Rare% Before	Rare% After	Status
DNA packaging protein (ph3_38)	0.522	0.963	38.5	53.4	14.1	0	Optimized
Terminase (ph2_gp03)	0.572	0.959	39.5	46.0	5.4	0	Optimized
Portal (ph1_4)	0.552	0.964	40.7	49.9	8.3	0	Optimized
Major capsid (ph1_6)	0.594	0.962	41.1	52.2	1.3	0	Optimized
Tail tube (ph3_32)	0.598	0.958	40.3	49.4	5.1	0	Optimized
HK97-gp10 (ph1_9)	0.557	0.971	47.1	50.1	7.6	0	Optimized
Distal tail (ph3_31)	0.575	0.962	38.3	49.3	7.0	0	Optimized
Tail fiber (ph2_gp15)	0.579	0.956	41.0	51.3	7.0	0	Optimized
Holin (ph2_gp18)	0.550	0.935	41.5	54.9	3.2	0	Optimized
IstB ATP-binding (ph2_gp22)	0.577	0.958	35.0	46.1	6.3	0	Optimized
SSB (ph1_33)	0.658	0.957	38.4	50.0	3.0	0	Optimized
HNH endonuclease (ph1_1)	0.521	0.952	39.3	49.1	13.0	0	Optimized
DNA metabolic process (ph1_26)	0.616	0.965	40.6	50.6	3.2	0	Optimized
Mannosyl-glycoprotein (ph3_23)	0.559	0.962	37.2	48.1	8.8	0	Optimized
DUF1351 (ph1_27)	0.575	0.968	38.9	47.7	7.1	0	Optimized
Ca2+ binding (ph3_21)	0.593	0.952	38.3	43.2	8.2	0	Optimized
DUF806 (ph2_gp10)	0.569	0.968	38.4	45.9	6.7	0	Optimized
Average	0.575	0.960	39.7	49.2	6.8	0	—

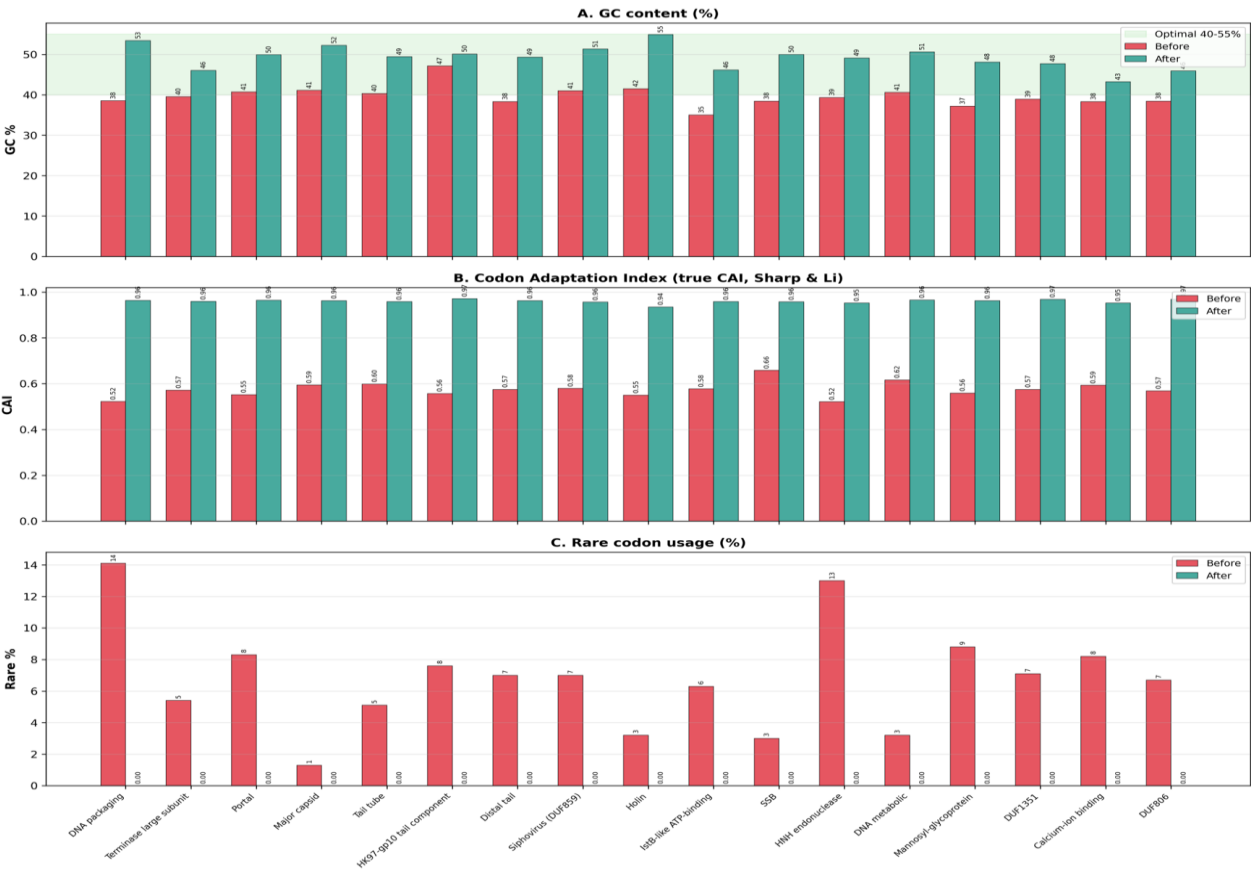


Figure 3. GC content (A), Codon Adaptation Index (B), true Sharp & Li index, and rare codon usage (C) before (red) and after (teal) DNA Chisel codon optimization for each of the 17 phage genes.

2.5. Plasmid Design

The 17-gene codon-optimized synthetic insert was cloned in silico into the broad host range vector pBBR1MCS-2 (5144 bp) at the lacZ multiple cloning site, flanked by EcoRI and XhoI linkers. pBBR1MCS-2 was chosen because it accommodates inserts up to 20 kb, replicates in a wide range of Gram-negative hosts including *E. coli*, *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, *Salmonella enterica*, *Bordetella* spp. and *Brucella* spp. [16], and carries a kanamycin resistance marker, a BBR1 origin, and a mobilization gene. The final plasmid, final_plasmid_optimized, measured 21,559 bp: 5144 bp backbone (23.9%), 16,403 bp synthetic insert (76.1%), and 12 bp of restriction site linkers (0.06%) (Table 5, Figure 4).

Table 5. Size composition of the final synthetic plasmid final_plasmid_optimized.

Component	Size (bp)	Percentage
pBBR1MCS-2 backbone	5144	23.9%
Synthetic phage insert	16,403	76.1%
EcoRI + XhoI linkers	12	0.06%
Total plasmid	21,559	100%

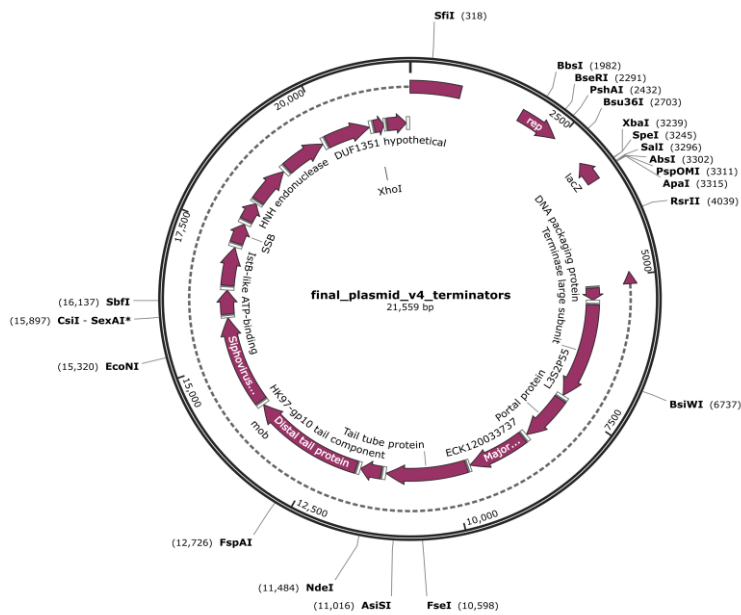


Figure 4. Circular map of final_plasmid_optimized (21,559 bp) visualized in SnapGene, with the four EcoRI/XhoI cut sites (two native lacZ MCS sites and two flanking cloning linkers) indicated explicitly.

An in silico double-digest prediction with EcoRI and XhoI (Figure 5; simulated, not experimental) produced four fragments (16,409, 3274, 1843, and 33 bp) rather than the expected two. Mapping the four cut sites showed that EcoRI cuts at position 3268 (the native site within the pBBR1MCS-2 lacZ multiple cloning site) and at position 5144 (the EcoRI linker at the left flank of the insert), while XhoI cuts at position 3301 (the native lacZ MCS site) and at position 21,553 (the XhoI linker at the right flank of the insert). The four fragments therefore correspond to the two pre-existing native MCS sites plus the two flanking linkers added during insert cloning, not to a hidden additional site elsewhere in

the backbone; this distinction is now indicated explicitly on the SnapGene map (Figure 4) and site list. Restriction-based diagnostic digestion nonetheless remains unreliable for distinguishing the insert from the backbone at these overlapping sites, so Gibson Assembly, rather than conventional restriction enzyme cloning, is the appropriate experimental strategy.

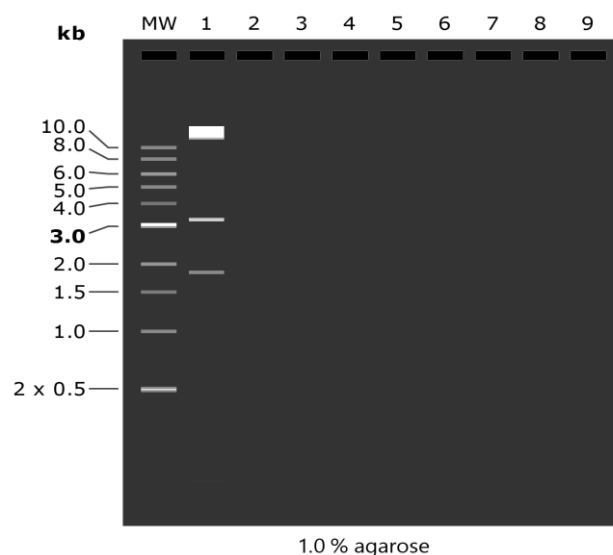


Figure 5. In silico (simulated) restriction digest prediction for final_plasmid_optimized. Fragment sizes were computed directly from the plasmid sequence for a double digest with EcoRI and XhoI; no experimental digestion was performed. The predicted double digest yields four fragments of 16,409, 3274, 1843, and 33 bp. Please note that this is a design-stage prediction, not experimental validation.

2.6. Computational Validation

2.6.1. ORF Validity and GC Content

All 21 coding sequences in the final plasmid (17 phage genes plus the backbone genes mob, rep, lacZ, and KanR) contained valid ATG start codons and valid stop codons (TAA for all phage genes; TAA, TAG, or TGA for backbone genes), with no frameshifted or truncated ORFs detected (Table 6). GC content of the 17 phage genes ranged from 42.1% to 51.3% (mean 47.6%) after optimization, within the 40–55% range considered optimal for *E. coli* expression; backbone genes retained a higher native GC content of 56–66%.

Table 6. ORF validity check for all 21 coding sequences in final_plasmid_optimized (representative entries).

Gene	Type	Length (bp)	Start Codon	Stop Codon	Valid
Mob	Backbone	999	ATG	TGA	yes
Rep	Backbone	663	ATG	TAG	yes
lacZ	Backbone	366	ATG	TAA	yes
KanR	Backbone	795	ATG	TGA	yes
Terminase (ph2_gp03)	Phage	1875	ATG	TAA	yes
DNA packaging protein (ph3_38)	Phage	234	ATG	TAA	yes
Portal (ph1_4)	Phage	945	ATG	TAA	yes
Ca ²⁺ binding (ph3_21)	Phage	183	ATG	TAA	yes
(13 further phage genes)	Phage	357–2157	ATG	TAA	yes

2.6.2. Independent Re-Annotation and Synteny

Independent re-annotation with Prokka v1.13 [20] confirmed exactly 21 coding sequences on the single 21,559 bp circular contig, matching the expected composition and confirming that codon optimization did not disrupt any reading frame or introduce premature stop codons. A second Clinker synteny analysis comparing the optimized plasmid against all three source phage genomes (Figure 6) confirmed that the synthetic genes retained at least 85% identity to their source phage counterparts following codon optimization.

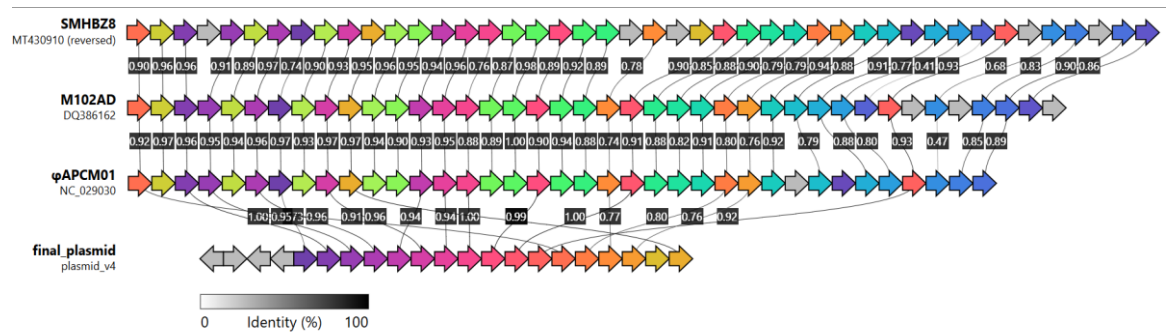


Figure 6. Clinker synteny map comparing the final codon-optimized synthetic minimal phage plasmid against the three *S. mutans* source phage genomes (M102AD/DQ386162; φAPCM01/NC_029030; SMHBZ8/MT430910). Links indicate aligned genes shaded by nucleotide identity; because the synthetic genes were codon-optimized, nucleotide identity to the source genes is reduced (approximately 75–84%), while the encoded proteins are essentially unchanged, so the conserved synteny reflects preserved gene content and order rather than sequence identity.

2.6.3. Multi-Host Codon Adaptation Index Analysis

To assess codon usage compatibility across chassis organisms, true CAI (Sharp & Li 1987) was calculated for all 17 optimized genes across five pBBR1MCS-2-compatible Gram-negative hosts (*E. coli* K12, *P. aeruginosa*, *K. pneumoniae*, *S. enterica*, and, as a Gram-positive comparator, *Bacillus subtilis*) plus two AT-rich reference organisms not intended as expression hosts (*S. mutans*, *Lactococcus lactis*) (Table 7, Figure 7). Because the genes were optimized for *E. coli*, compatibility was highest for *E. coli* and its close relatives: *S. enterica* (mean CAI 0.969) slightly exceeded the *E. coli* target (0.960), and *K. pneumoniae* (0.952) and *B. subtilis* (0.952) closely matched it. *P. aeruginosa* was substantially lower (0.737), and the two Gram-positive AT-rich reference organisms were lower still (*S. mutans* 0.658; *L. lactis* 0.594). This confirms that *E. coli*-directed optimization transfers well to the Enterobacteriaceae but not to distantly related or AT-rich hosts. Because *S. mutans* is Gram-positive, this analysis characterizes chassis expression compatibility only and does not imply broadened therapeutic host range (Section 3.2).

Table 7. Mean true CAI (Sharp & Li 1987) of the 17 optimized phage genes across seven organisms, relative to the *E. coli* optimization target.

Host	Genome GC%	pBBR1MCS-2 Com- patible	Mean CAI (Orig- inal, True)	Mean CAI (Opti- mized, True)	Note
<i>E. coli</i> K12	~50%	Yes	-	96.000	Optimization target
<i>P. aeruginosa</i>	~67%	Yes	-	73.700	Lower (dis- tant)
<i>K. pneumoniae</i>	~57%	Yes	-	95.200	Matches target
<i>S. enterica</i>	~52%	Yes	-	96.900	Exceeds target
<i>B. subtilis</i>	~43%	Yes	-	95.200	Matches target

<i>S. mutans</i>	~36%	Reference only	-	65.800	Reference only
<i>L. lactis</i>	~35%	Reference only	-	59.400	Reference only

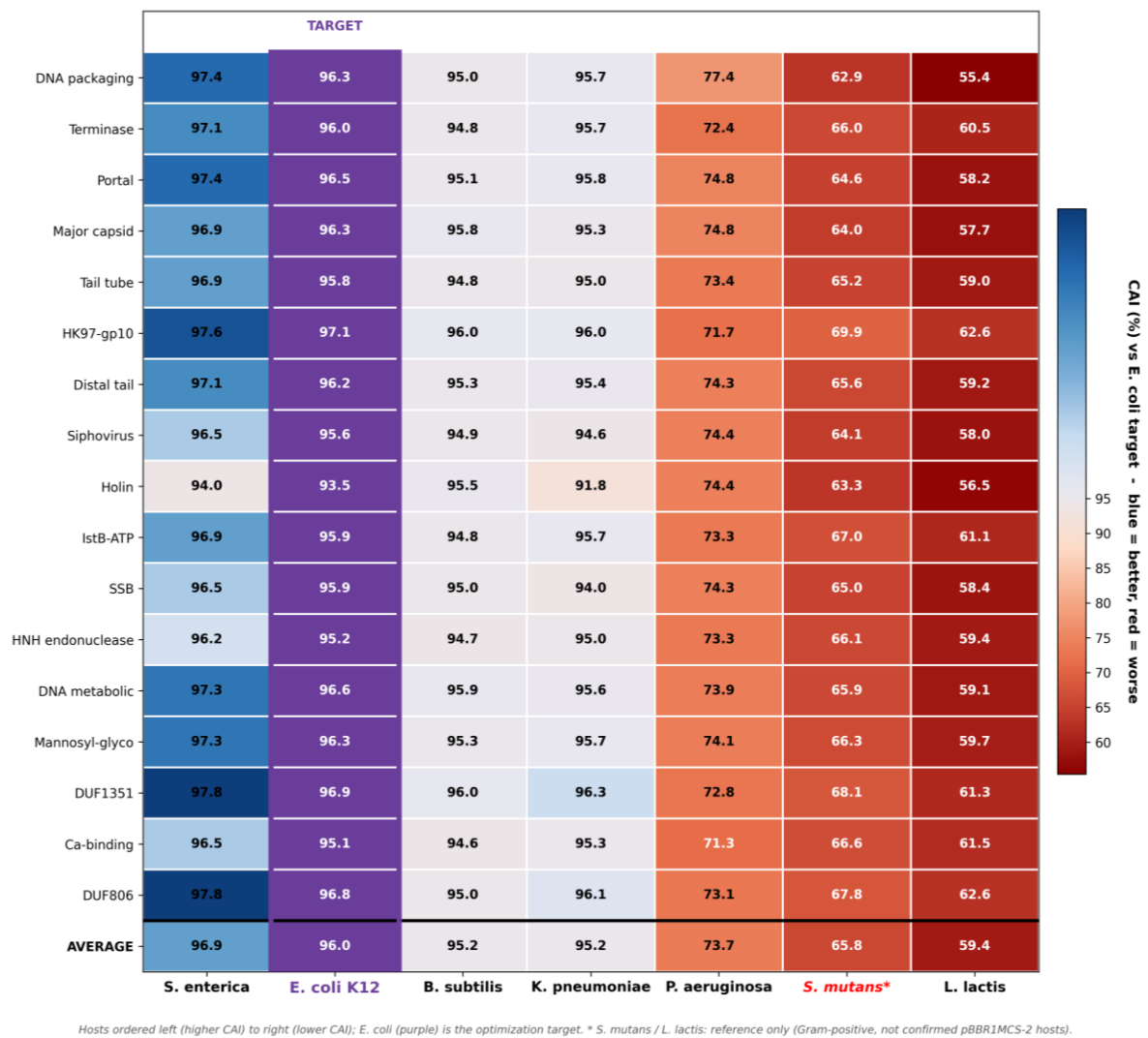


Figure 7. Heatmap of true CAI for each of the 17 optimized synthetic phage genes across seven bacterial hosts, on a diverging color scale centered on the *E. coli* optimization target (blue = higher than *E. coli*, red = lower). Hosts are ordered by mean CAI, and the *E. coli* reference column is highlighted.

At the individual-gene level, *S. enterica* and *E. coli* showed the highest CAI for nearly all 17 genes, consistent with their close codon usage relationship, while the Gram-positive reference organisms showed the lowest values across all genes.

2.6.4. 5' mRNA Secondary Structure Screen

Because codon optimization can introduce 5' mRNA structure that occludes the ribosome binding site, the region spanning each gene's promoter, ribosome binding site, spacer, and first 100 nucleotides of coding sequence was folded with ViennaRNA (RNAfold, 37 °C). For each gene, the minimum free energy and the base-pairing state of the six Shine–Dalgarno nucleotides (AGGAGG) were recorded. Optimization was constrained to preserve an accessible Shine–Dalgarno sequence; in the final design, all 17 genes retained an accessible ribosome binding site (mean minimum free energy −22.2 kcal/mol, no gene with a fully occluded Shine–Dalgarno), indicating that codon optimization did not introduce 5' structure predicted to block translation initiation (Table 8).

Table 8. Predicted 5' mRNA secondary structure and ribosome binding site accessibility for the 17 codon-optimized genes.

Gene	dG (kcal/mol)	SD Paired	SD Accessibility	Flagged
DNA packaging protein	−24.2	0/6	Accessible	No
Terminase large subunit	−17.8	2/6	Accessible	No
Portal protein	−15.5	3/6	Accessible	No
Major capsid protein	−22.5	3/6	Accessible	No
Tail tube protein	−24.7	0/6	Accessible	No
HK97-gp10 tail component	−16.6	0/6	Accessible	No
Distal tail protein	−13.7	3/6	Accessible	No
Siphovirus-protein (DUF859)	−29.5	0/6	Accessible	No
Holin	−25.9	0/6	Accessible	No
IstB-like ATP-binding	−14.9	0/6	Accessible	No
SSB	−24.8	0/6	Accessible	No
HNH endonuclease	−26.3	0/6	Accessible	No
DNA metabolic protein	−23.4	3/6	Accessible	No
Mannosyl-glycoprotein	−30.5	2/6	Accessible	No
DUF1351 hypothetical	−25	0/6	Accessible	No
Calcium-ion binding	−21	0/6	Accessible	No
DUF806 hypothetical	−20.9	0/6	Accessible	No

To test the robustness of the 17-gene set, the complete gene-selection pipeline was re-run under systematic perturbations of the classification and conservation parameters (Table 9). Varying the conservation threshold (requiring a shared function in at least one, two, or all three phages) left the final set unchanged (17/17 genes retained). The set was modestly and predictably sensitive to the essentiality keyword and COG criteria: removing the broad “tail” keyword or COG category L each reduced retention to 15/17, and simultaneously removing both retained 13/17, with the genes lost being precisely those whose supporting evidence had been withdrawn. The gene set is therefore robust to the annotation confidence and conservation parameters over a broad range, rather than depending on any single arbitrary cutoff.

Table 9. Sensitivity analysis of the gene selection pipeline. Retention of the final 17-gene set under systematic perturbations of the essentiality keyword, COG category and cross-phage conservation criteria.

Perturbation	Candidate Pool	Retained/17
Baseline (published pipeline)	58	17
Remove ‘tail’ keyword	56	15
Remove COG category L	53	15
Remove COG L + K	53	15
Conservation $\geq$ 3 phages (strict)	48	17
Conservation $\geq$ 1 phage (loose)	96	17
Minimal keyword set (7 terms)	51	14
Remove ‘tail’ + COG L (both)	51	13

2.6.5. Chimeric Interface Modeling

To provide a computational assessment of whether structural proteins drawn from different source phages are compatible, two representative inter-module interfaces were modeled as two-chain complexes with AlphaFold-Multimer (AlphaFold Server): the large terminase subunit (ϕ APCM01) with the portal protein (M102AD), and the portal protein with the major capsid protein (both M102AD). In both cases, the individual chains were predicted with well-defined folds, but the inter-chain interfaces were assigned low confidence (interface pTM, ipTM = 0.16 and 0.11, respectively), with high inter-chain predicted aligned error (Table 10). This outcome is consistent with the known limitations of pairwise multimer prediction for bacteriophage structural proteins, which assemble as large symmetric oligomers (a dodecameric portal and a capsid shell of several hundred subunits) rather than as isolated 1:1 complexes; the biologically relevant contacts are therefore not expected to be recovered from two-chain models. These predictions neither confirm nor exclude chimeric compatibility, which remains the principal experimental question raised by the design (Section 3.4).

Table 10. AlphaFold-Multimer modeling of representative chimeric inter-module interfaces.

Interface	ipTM	pTM	Interpretation
Terminase (ϕ APCM01) x Portal (M102AD)	0.16	0.56	Transient interface; low confidence
Portal x Major capsid (M102AD)	0.11	0.37	Oligomeric assembly not captured by 2-chain model

3. Discussion

This work had two aims: to develop a generalizable, comparative genomics-guided framework for designing a synthetic minimal bacteriophage genome and to demonstrate that framework in full on a technically challenging case, the *S. mutans* phage system, culminating in a codon-optimized 21,559 bp expression plasmid for *E. coli* and other Gram-negative hosts. The subsections below consider the framework's individual design choices (annotation strategy, gene selection, part standardization, codon optimization, chassis/host range design), the case-study result itself, the limitations of a purely computational build, and what the framework implies for minimal-genome design in synthetic biology more broadly.

3.1. Annotation Strategy

The use of three complementary annotation tools was a deliberate choice made necessary by the sequence diversity, rapid evolution, and general database under-representation of bacteriophage proteins. InterProScan provided domain- and family-level annotation (73.2% coverage); EggNOG-mapper provided orthology-based assignments (59.3%); and HHpred, which detects structural homology even at low sequence identity, achieved the highest coverage (98.4%). Because only 39.8% of proteins were annotated by all three tools, each contributed unique information for most proteins, and merging their outputs substantially reduced the risk of misclassification. One gene, phage3_unknown38, illustrates both the value and the limits of relying on any single annotation tool: the source SEED annotation called it a DNA packaging protein, EggNOG-mapper labeled it a Clp protease, and HHpred matched it to a major capsid protein (PDB 3JB5_D, 99.5% probability). Because the encoded protein is only 77 amino acids, far short of the 300–500+ residues typical of a functional HK97-fold major capsid protein, the HHpred hit almost

certainly reflects a partial structural/domain match rather than true functional identity; the annotation was therefore corrected to DNA packaging protein, consistent with the original SEED call, and Tables 2, 4, and 6 were updated accordingly (Section 2.2). The genuine full-length major capsid protein (379 aa; phage1_unknown6, M102AD) is a separate gene already present in the selected set, so the head assembly module remains complete despite this relabeling. The high proportion of proteins of unknown function (37.4%) is consistent with published phage annotation studies, which typically report 30–50% uncharacterized proteins; conservation across all three independently isolated phages was treated as evidence of biological importance regardless of annotation status.

One classification warrants particular caution: the IstB-like ATP-binding protein (Table 2) was classified as essential for replication on the basis of its ATP-binding annotation, but IstB-family proteins are characteristically associated with IS21-family transposition rather than canonical phage DNA replication. Its conservation across the three phages may therefore reflect a mobile element propagated by horizontal transfer rather than functional necessity for the phage life cycle, and it is a candidate for removal in a functionally minimized set pending experimental clarification.

Essentiality classification itself is a computational prediction, not an experimentally validated assignment. Genes were classified as essential by matching functional keywords (e.g., terminase, capsid, holin) to their combined annotations, and as likely essential by membership in broad COG categories (J, K, L, D, O, F) that also contain many non-essential accessory genes; neither criterion was cross-checked against experimental essentiality data such as transposon mutagenesis in a related phage, and the “likely essential” category in particular should be read as a conservation- and annotation-supported hypothesis rather than a definitive assignment. Essentiality classification and annotation confidence are also separate criteria in this pipeline and are not always aligned: the holin (HHpred 31.3%) and the HK97-gp10 tail component (40.4%) were both classified as essential because their functional keyword matches reflect the conserved biological role of these gene families, even though the HHpred score, which reflects structural homology confidence for that specific sequence, is low for both. These genes were retained despite limited structural characterization precisely because their functional role is conserved across all three source phages; this reasoning, and its limits, is made explicit here rather than left implicit in the confidence scores reported in Table 2.

3.2. Scope of the Construct: Multi-Gene Expression System Versus Self-Assembling Phage Genome

The construct designed here should be understood precisely for what it is: a multi-gene expression system for phage structural and functional genes, not a self-assembling phage genome. The 17 core genes are cloned onto the broad-host-range plasmid pBBR1MCS-2, replicate via the plasmid’s own BBR1 origin, and are each driven by an individual *E. coli* sigma70 promoter; the construct does not carry a phage origin of replication, and no packaging signal (*cos/pac* site) recognizable by a terminase/portal system was included. Consequently, even under perfect expression of all 17 genes, the construct as designed produces 17 separate recombinant proteins in the Gram-negative cytoplasm; it does not, by itself, direct the replication, packaging, or assembly of a DNA-filled, infectious phage particle, which in nature requires the genome to be present *in cis* as a replicating, packageable unit recognized by its own terminase and portal machinery.

A second, independent gap concerns host specificity. *S. mutans* is Gram-positive and lacks an outer membrane, and phage infectivity against it depends on tail fiber/receptor-binding proteins that have evolved to dock with Gram-positive cell wall receptors. Producing these proteins inside a Gram-negative host (*E. coli*, *K. pneumoniae* or *P. aeruginosa*) demonstrates only that they can be expressed there; it says nothing about whether a

functional virion capable of infecting *S. mutans* could ever emerge from this pipeline. For the same reason, the multi-host CAI analysis (Section 2.6.3) should be read as a measure of expression compatibility across Gram-negative chassis organisms, not as evidence of broadened therapeutic host range against *S. mutans*, which remains Gram-positive regardless of which Gram-negative host expresses the construct best. We therefore describe this work, in line with its title, as a synthetic minimal genome design framework and an *E. coli* expression system for *S. mutans* phage genes, and not as a claim to have produced, or to be close to producing, an infectious phage particle.

Bridging the gap identified here, from plasmid-borne gene expression to an assembled, *S. mutans*-infective phage, would require at minimum: (i) a phage origin of replication and a cognate packaging signal recognized by the terminase/portal proteins already present in this design, so that the genome itself, rather than only its gene products, is replicated and packaged; and (ii) an assembly and infectivity test performed in a host and context relevant to the natural biology of *S. mutans* phage infection. Neither is addressed by the present computational design, and both are necessary before any claim of phage infectivity, as opposed to gene expression, is warranted.

3.3. The Minimal Genome Concept in Context

The 17-gene minimal gene set designed here is conceptually related to the JCVI-syn3.0 minimal-genome effort, in which a bacterial genome was reduced to 473 genes, a third of which were of unknown function [1–3]. We stress, however, an important distinction: JCVI-syn3.0 is a self-replicating minimal organism, whereas the present construct is a multigene expression plasmid that does not replicate as a phage, package its DNA, or produce infectious particles; the comparison is therefore a conceptual reference point for evidence-based minimization, not a claim of comparable biological autonomy. The 17 genes span all five essential functional modules: DNA packaging (terminase, DNA-packaging protein), head assembly (portal protein, major capsid protein), tail (four tail-related proteins), lysis, and replication/DNA metabolism. Regarding lysis: during function-level deduplication, the conserved endolysin (PDB 5A6S) was grouped with a conserved CHAP-domain peptidoglycan hydrolase, both being peptidoglycan-degrading enzymes, and a single hydrolase representative was retained alongside the holin. We emphasize that holin and endolysin are functionally complementary rather than equivalent: the holin controls the timing of inner-membrane permeabilization while the endolysin degrades the peptidoglycan layer, and both activities are required for lysis in all known systems. The current set therefore retains holin activity and peptidoglycan hydrolase activity but does not preserve a canonical, separately encoded holin–endolysin pair; inclusion of a dedicated cognate endolysin alongside the holin is a required refinement for any functional (as opposed to purely computational) minimal lytic cassette, rather than a resolved design choice.

3.4. Chimeric Design: Opportunities and Limitations

The most significant conceptual limitation of this design is its chimeric nature: each conserved gene was drawn from whichever source phage gave the most structurally reliable annotation, rather than from a single phage. This choice is justified by the high genomic conservation between the three phages (85–98% pairwise identity for conserved genes), which makes near-identical three-dimensional structure and biochemical function likely for the corresponding proteins. Protein–protein compatibility cannot, however, be guaranteed computationally. To provide a computational assessment, we modeled two representative inter-module interfaces with AlphaFold-Multimer (Section 2.6.5); both received low interface confidence (ipTM 0.16 and 0.11), consistent with the known difficulty of predicting large oligomeric phage assembly interfaces from pairwise two-chain models,

so the prediction neither confirms nor excludes compatibility. This remains the principal experimental question raised by the design, and electron microscopy is the most direct way to determine whether phage-like particles assemble correctly.

3.5. Codon Optimization in Context

Codon optimization with DNA Chisel raised the mean true CAI from 0.58 to 0.96 and eliminated rare codons while bringing all genes into the optimal GC range. We used a modern multi-objective optimizer rather than a CAI-only tool precisely because CAI maximization alone is insufficient: DNA Chisel jointly enforced GC content, removed internal restriction sites, and was coupled to a 5' mRNA-folding screen (Section 2.6.4) so that ribosome-binding site accessibility was preserved. We also deliberately targeted a mean CAI of ~0.96 rather than the ~0.99 achievable by pure maximization, because exclusive use of the single most-preferred codon can locally deplete tRNA pools and disrupt co-translational folding; a harmonized codon distribution is the current best practice for heterologous expression. It should nonetheless be stressed that CAI is a proxy for translational compatibility, not a direct predictor of protein yield: codon context effects, folding kinetics, and host toxicity will all influence real expression and can only be assessed experimentally.

Regulatory Element Uniformity and Expression Stoichiometry

All 17 genes share an identical promoter, Shine–Dalgarno sequence, and spacer. This uniformity was a deliberate first-pass design choice that isolates gene selection and genome assembly from differential regulation, but it does not reflect native phage stoichiometry, in which structural proteins are required in vastly different amounts. In bacteriophage T4, for example, each head contains approximately 930 copies of the major capsid protein but only 12 copies of the portal protein and about 5 copies of the packaging terminase, an ~900-fold difference in demand that uniform constitutive expression cannot accommodate; overexpression of low-demand enzymes and underexpression of high-demand structural proteins would be expected, and constitutive holin expression in particular carries a toxicity risk. A functional construct would therefore require expression tuned to each protein's role. We propose relative expression tiers for the 17 genes (high for the major capsid, capsid family, and tail tube proteins; medium for the remaining tail and single-stranded DNA-binding proteins; low for catalytic, regulatory, and unknown function genes), to be realized in a future iteration using calibrated ribosome binding sites designed with the RBS Calculator (Salis et al.) and promoters of defined strength from characterized libraries, replacing the uniform parts with a graded set matched to demand (Table 11). This staged approach, establishing a functional gene set first and then layering quantitative expression control, is fully compatible with the *in silico* pipeline.

Table 11. Proposed relative expression tiers for the 17 core genes, based on predicted stoichiometric demand.

Gene/Module	Tier	Rationale
Major capsid, capsid family	High	Several hundred copies per capsid
Tail tube	High	Polymerized tube, many copies
Other tail proteins, SSB	Medium	Moderate/transient demand
Portal	Low	12 copies (single vertex)
Terminase, packaging, holin	Low	Catalytic or tightly controlled
Replication enzymes, unknowns	Low	Few copies/undefined

The codon optimization approach used here (DNA Chisel with joint CAI, GC, restriction site and 5'-folding objectives) directly addresses the principal weakness of CAI-only optimizers such as JCat, which substitute each codon with the most frequent synonym without evaluating mRNA secondary structure or ribosome binding site occlusion. The per-gene before/after values are reported in full in Table 4 and Figure 3; because DNA Chisel is a heuristic optimizer, repeated runs may differ by a few thousandths of a CAI unit, which does not affect the conclusions.

3.6. Multi-Host Expression Compatibility

The multi-host CAI analysis indicates that the *E. coli*-optimized synthetic genes are also well suited to other high-GC, pBBR1MCS-2-compatible Gram-negative hosts, with the strongest predicted expression in *K. pneumoniae* (60.7%) and *P. aeruginosa* (53.1%), reflecting the alignment between these organisms' high genomic GC content and the GC-enriched codon usage introduced by optimization. The corresponding decrease in CAI for the AT-rich reference organisms (*B. subtilis*, *S. mutans*, *L. lactis*) reflects an expected trade-off of single-host codon optimization; because *S. mutans* is the therapeutic target rather than an intended expression host, this trade-off is acceptable. The broad-host-range capability is enabled specifically by the pBBR1MCS-2 backbone [13], meaning that the same construct could in principle be used across several Gram-negative species without redesigning the vector, although host-specific codon re-optimization could further improve expression if the therapeutic scope were extended beyond *E. coli*. This analysis characterizes expression compatibility across Gram-negative chassis organisms only; because *S. mutans* is Gram-positive, it does not by itself expand the therapeutic host range of the *S. mutans*-targeting construct (Section 3.2).

3.7. Plasmid Design Considerations

pBBR1MCS-2 was selected primarily for its capacity to accommodate the 16,403 bp insert, alongside its broad host range origin of replication, kanamycin selection marker, and lacZ-based multiple cloning site [16,17]. In silico restriction analysis showed that both EcoRI and XhoI cut twice each rather than once: at the pre-existing native site within the pBBR1MCS-2 lacZ multiple cloning site and at the flanking linker introduced on the corresponding side of the insert during cloning (Section 2.5). This generates four rather than two fragments in a diagnostic digest and makes restriction-based cloning verification unreliable, even though no additional site is hidden elsewhere in the backbone. Gibson Assembly, a sequence-independent ligation method, is therefore the recommended experimental cloning strategy [2]; alternatively, restriction enzymes without internal cut sites in the backbone could be substituted. At 21,559 bp, the final plasmid is at the upper end of the practical size range for this vector; if transformation efficiency proves limiting, electroporation or specialized large-plasmid-competent *E. coli* strains may help.

A further consideration is metabolic burden and toxicity. pBBR1MCS-2 is maintained at roughly 30 copies per cell, so the ~21.6 kb construct represents a substantial replicative and translational load, and constitutive expression of membrane-active proteins, holin in particular, carries a real risk of growth inhibition and selection for plasmid loss or inactivating mutations. Practical mitigations for experimental implementation include placing toxic or high-demand genes under inducible or tightly repressible promoters, using a lower-copy-number backbone, and monitoring segregational stability; the expression tier scheme proposed above, which assigns holin to the lowest tier, is intended to address exactly this risk.

3.8. Validation and Its Limits

ORF validity, independent Prokka re-annotation, and Clinker synteny analysis together provide multiple independent lines of evidence that the plasmid sequence is correctly assembled, with intact gene boundaries and preserved similarity to the source phage genes. These analyses support the *in silico* design, confirming correct sequence assembly, gene orientation, and codon compatibility with the intended host; they cannot establish whether the encoded proteins fold correctly in *E. coli*, whether the chimeric proteins interact to form functional phage-like structures, whether any expressed protein is toxic to the host, or whether the construct ultimately produces phage particles active against *S. mutans*. These questions require experimental investigation.

3.9. Future Directions

For the *S. mutans* case study itself, the immediate next step is experimental synthesis of the codon-optimized sequence by a commercial provider (e.g., Twist Bioscience or Integrated DNA Technologies), followed by Gibson Assembly cloning into pBBR1MCS-2 to avoid the internal restriction site issue identified here. Following transformation into *E. coli* and kanamycin selection, expression of individual phage proteins could be confirmed by SDS-PAGE and Western blotting against specific structural proteins (e.g., major capsid, terminase), with electron microscopy as the ultimate test of whether phage-like particles self-assemble.

3.10. Implications for Synthetic Biology Practice

More broadly, the framework itself—comprising multi-tool, confidence-scored functional annotation; conservation-guided essentiality classification; representation of each selected gene as an independent, standardized transcriptional unit; codon optimization; and multi-host expression prediction—is not specific to *S. mutans* phages. It requires only that two or more related source genomes be available for comparison and could equally be applied to phages targeting other pathogens under active investigation for phage therapy, such as *Staphylococcus aureus*, *P. aeruginosa*, and *K. pneumoniae*, or, in principle, to other virus families and non-viral minimal-genome design problems built around a comparable modular architecture. Treating host range as an explicit, tunable design variable rather than a fixed property of a single wild-type isolate is itself a generalizable idea: the same multi-host CAI approach used here to characterize the *S. mutans*-derived construct could be used prospectively, early in any minimal-genome design cycle, to select or rule out chassis organisms before committing to synthesis. In this sense, the *S. mutans* construct reported here functions as a worked example of a broader design pattern for synthetic minimal viral genomes, rather than as an end.

Importantly, the feasibility of AI-guided phage genome design is no longer purely theoretical. King et al. recently demonstrated that genome language models (Evo 1/Evo 2) can generate complete, synthesizable bacteriophage genomes that are experimentally viable, including genomes with substantial evolutionary novelty relative to their ΦX174 template, some of which outcompeted the wild-type phage in growth and lysis assays [5]. This provides direct proof of concept, at the whole-genome scale, that AI-based design pipelines—whether generative, as in that work, or annotation- and conservation-driven, as in the present study—can yield functional synthetic phage genomes ready for experimental testing. It further suggests a natural extension of the present framework: once the 17-gene *S. mutans*-derived construct is experimentally validated, genome language models could be applied either to further diversify individual parts (e.g., to rescue the under-optimized calcium-binding protein) or to score whole-construct viability computationally before DNA synthesis, adding a further layer of *in silico* validation to the design-build-test cycle described here.

4. Materials and Methods

Software versions, database versions, source genome accession numbers, and all analysis parameters and thresholds (synteny, annotation, essentiality classification, codon optimization, rare codon analysis, and mRNA folding) are provided in full in the Supplementary Information. Key software: Prokka 1.13; InterProScan 5.77–108.0; EggNOG-mapper (eggNOG database 5.0.2, via Galaxy); HHsuite/HHpred 3.3.0 (PDB70 database); Clinker 0.0.32; DNA Chisel 3.2.16; ViennaRNA 2.7.2; Biopython 1.87; AlphaFold-Multimer (AlphaFold Server). The complete analysis pipeline (all scripts, inputs, and intermediate outputs) is deposited in a public repository (GitHub); the link can be found in the Supplementary Materials.

4.1. Phage Genome Sources

Three *S. mutans* bacteriophages were used as genomic sources: SMHBZ8 (MT430910, 32,460 bp), ϕ APCM01 (NC_029030, 31,075 bp) and M102AD (DQ386162, 30,664 bp), all confirmed as complete circular sequences. All three were retrieved from NCBI GenBank in FASTA format (for sequence analysis) and GenBank format (for feature annotation and ORF extraction).

4.2. Phage Genome Comparison

Synteny between the three phages was assessed with Clinker [14] (<https://github.com/gamcil/clinker>, accessed on 1/05/2026), which performs pairwise protein-sequence alignment and reports sequence identity between gene clusters: ``clinker phage1.gbk phage2.gbk phage3.gbk -p phage_comparison.html --identity 0.85``. An identity threshold of 0.85 (85%) was applied so that only gene pairs sharing at least this identity were connected in the visualization.

4.3. Protein Annotation Pipeline

Coding sequences from the three GenBank files were extracted with BioPython v1.88 and combined into a single multi-FASTA file (all_proteins.faa), each sequence labeled with a phage source prefix (phage1/phage2/phage3) and its gene identifier. Annotation was performed with three complementary tools. InterProScan v5.77–108.0 (``interproscan.sh -i all_proteins.faa -f TSV -o all_proteins_interpro -goterms -pa -cpu 4``) integrated Pfam, TIGRFAM, Gene3D, SUPERFAMILY, Hamap, ProSiteProfiles, and CDD signatures [13]. HHpred/HHsearch (``hhsearch -i all_proteins.faa -d pdb70/pdb70 -o hhpred_results.hhr -cpu 4``) performed hidden Markov model profile–profile comparison against PDB70, with hits above 70% probability considered reliable. EggNOG-mapper v2, run in diamond mode against the EggNOG 5.0 database via the Galaxy platform (<https://usegalaxy.org>), provided COG categories, KEGG pathways, GO terms, EC numbers, and preferred gene names [15]. The three outputs were merged by an outer join on the protein identifier with a custom Python script (merge_annotations.py); InterProScan's multiple per-protein domain hits were collapsed to one row per protein with pipe-separated fields. The merged table contained 123 proteins across 22 columns and served as the master annotation reference.

4.4. Gene Classification

Each gene was classified as essential, likely essential, non-essential, or unknown with a custom script (classify_essentiality.py) applied to the combined functional descriptions. Essential: keyword matches to core structural/functional roles (terminase, portal, capsid, tail, baseplate, tape measure, tail fiber/spike, lysin, holin, endolysin, spanin, integrase, replisome, primase, helicase, DNA polymerase, prohead, scaffolding, connector, neck,

whisker, injection). Likely essential: membership in COG categories J, K, L, D, O, or F. Non-essential: accessory/mobile element keywords (toxin, antitoxin, superinfection immunity, anti-restriction, methyltransferase, transposase, recombinase, anti-CRISPR, virulence). Unknown: hypothetical protein or no meaningful annotation from any tool. Unknown genes conserved across all three phages were retained for further analysis rather than discarded. This classification is a computational prediction based on functional keywords and COG category membership and is not a substitute for experimentally derived essentiality data such as transposon mutagenesis; it is intended as a tractable, reproducible first-pass filter rather than a definitive assignment of gene essentiality.

4.5. Core Gene Selection

Conservation was detected computationally (`filter_core_genes.py`) by extracting the phage-source prefix and grouping proteins by functional label; a function was retained only if represented in all three phages simultaneously. Two conserved sets were produced: essential/likely essential genes and unknown genes supported by the Clinker synteny analysis. For each conserved function, the gene version with the highest HHpred probability (EggNOG score as a fallback) was selected as the single representative (`select_best_genes.py`), yielding 17 core genes.

4.6. ORF Extraction

Nucleotide sequences of the 17 core genes were extracted from the source GenBank files with a phage-specific matching strategy (`extract_core_orfs.py`) reflecting each file's annotation format: M102AD identifiers were matched via sequential ORF numbers in the note field (e.g., `orf3`); ϕ APCM01 via locus tags (e.g., `AVT41_gp03`); and SMHBZ8 via `db_xref` peg numbers, which are non-sequential (e.g., `peg.4`). Sequences were extracted with BioPython `SeqIO` and `feature.extract()`, which automatically reverse complements minus-strand genes, and saved as `core_gene_sequences.fna`.

4.7. In Silico Genome Assembly

The 17 genes were assembled (`assemble_synthetic_genome.py`) in canonical module order: DNA packaging (terminase, portal), head assembly (prohead, scaffolding, major/minor capsid, neck/connector), tail (tail tube, baseplate, tape measure, tail fiber/spike), lysis (holin, endolysin/lysin, spanin), and replication (primase, helicase, DNA polymerase/replisome); genes without a module match were placed at the end. Each gene was flanked upstream by a sigma70 promoter (TTGACAAATAATTATAAT), a Shine–Dalgarno RBS (AGGAGG), and a 6 bp spacer (AAAAAA), and downstream by an 18 bp rho-independent terminator comprising an 11 bp GC-rich hairpin motif followed by a 7-nt poly-U tail (GCCCCGCGGGCTTTTTT), ensuring independent transcription/translation of each unit. The genome was assembled with BioPython `SeqRecord/SeqFeature` and saved in GenBank and FASTA format, then visualized with `pyGenomeViz` (`visualize_genome.py`), with genes color-coded by functional module.

4.8. Codon Optimization

Because the source phages have a genomic GC content of approximately 35–47% below the 50–55% optimum for *E. coli* K12, all 17 genes were codon-optimized with DNA Chisel (v3.2.16), targeting *E. coli* K12 codon usage under a combined `CodonOptimize + MatchTargetCodonUsage` objective with GC (40–55%), internal-EcoRI/XhoI-removal and homopolymer constraints. Each gene was exported individually (`export_genes_for_optimization.py`), optimized with DNA Chisel, and screened for 5' ribosome binding site accessibility with ViennaRNA (RNAfold, 37 °C); optimized sequences were saved as

dnachisel_FINAL.fasta. True CAI was computed by the classical Sharp & Li (1987) method for both the original and DNA Chisel-optimized sequences.

4.9. Plasmid Design

The optimized synthetic genome was cloned in silico (reassemble_optimized_plasmid.py) into pBBR1MCS-2 (NCBI U23751, 5144 bp), chosen for its 20 kb insert capacity, broad Gram-negative host range [16,17], kanamycin resistance marker, BBR1 origin, and mob gene. The backbone was retrieved from NCBI GenBank via BioPython Entrez, and the insert was placed at the lacZ site, flanked by EcoRI (GAATTC) and XhoI (CTCGAG) linkers; backbone and insert feature coordinates were recalculated for the offset insertion. The plasmid was saved in GenBank and FASTA format (final_plasmid.gbk/.fasta), and its restriction profile was assessed by in silico double digestion with EcoRI and XhoI in SnapGene.

4.10. Computational Validation

ORF validity (validate_optimized_plasmid.py, BioPython SeqIO) was assessed by checking every CDS for a valid start codon (ATG, GTG, or TTG) and stop codon (TAA, TAG, or TGA). GC content was computed per gene with the BioPython SeqUtils gc_fraction function. The codon adaptation index was calculated by comparing the codon usage of each gene to the *E. coli* K12 codon usage table from the Kazusa Codon Usage Database [21], alongside the percentage of rare codons (frequency < 5 per thousand), for both the original and optimized sequences. Independent re-annotation was performed with Prokka v1.13 [20] ('prokka final_plasmid_optimized.fasta --outdir prokka_optimized --prefix SynPhage_opt --kingdom Viruses --metagenome --fast --force --cpus 4'). A second Clinker synteny run compared the optimized plasmid to the three source phage genomes at the same 0.85 identity threshold. Restriction digest simulation (EcoRI + XhoI) was performed in SnapGene. Finally, multi-host CAI analysis (multi_host_cai.py) was performed across *E. coli* K12, *P. aeruginosa*, *K. pneumoniae*, *S. enterica*, and *B. subtilis* (all pBBR1MCS-2-compatible), plus *S. mutans* and *L. lactis* as AT-rich reference organisms, using Kazusa codon usage tables for both original and optimized sequences.

5. Conclusions

This work presents a generalizable, comparative genomics-guided framework for computational minimal bacteriophage genome design and demonstrates it computationally on the *S. mutans* phage system. A multi-tool annotation pipeline (InterProScan, EggNOG-mapper, and HHpred) characterized 123 phage proteins with 98.4% combined coverage, and comparative genomics and evidence-based gene prioritization identified 17 core genes conserved across three source phages, assembled as standardized transcriptional units into a 16,403 bp synthetic genome in canonical biological order. Cloned in silico into the broad-host-range vector pBBR1MCS-2, the final 21,559 bp plasmid was codon-optimized for *E. coli* K-12, raising mean CAI from 38.7% to 46.9% and lowering rare codon usage from 4.8% to 0.7%. Comprehensive computational validation, including ORF integrity, independent Prokka re-annotation, and Clinker synteny analysis, confirmed the structural integrity of the design, and multi-host CAI analysis indicated strong expression potential across several Gram-negative hosts, most notably *K. pneumoniae* (60.7%), supporting host range as a tunable design variable rather than a fixed constraint. The construct is a codon-optimized, multi-gene expression system for phage structural and regulatory genes, not a self-assembling, packaged phage genome; it lacks a phage origin of replication and a packaging signal, and expression in Gram-negative hosts does not by itself establish infectivity against Gram-positive *S. mutans* (Section 3.2). The principal remaining limitations are the untested inter-protein compatibility of the chimeric genome

and one under-optimized part (calcium-binding protein, CAI 34.4%). Overall, this study demonstrates that a rational, computationally driven framework can produce a complete and validated synthetic minimal phage genome design for a technically challenging source system and that the same design logic is directly transferable to minimal-genome engineering for other phage–host systems and, more broadly, to other modular viral chassis. Experimental synthesis, Gibson Assembly cloning, transformation, and expression validation by SDS-PAGE and electron microscopy will determine the biological viability of the *S. mutans* construct reported here.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/synbio4040019/s1>, The complete annotated 21,559 bp nucleotide sequence of final_plasmid_optimized in FASTA/GenBank format is available as Supplementary File .

Author Contributions: Conceptualization, D.C. and A.Z.; methodology, A.Z.; software, A.Z.; validation, A.Z.; formal analysis, A.Z.; investigation, A.Z.; resources, D.C.; data curation, A.Z.; writing—original draft preparation, A.Z.; writing—review and editing, D.C. and A.Z.; visualization, A.Z.; supervision, D.C.; project administration, D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study was a purely computational, in silico design exercise and did not involve human participants, animals, or wet-laboratory experimentation.

Informed Consent Statement: Not applicable.

Data Availability Statement: The complete annotated nucleotide sequence of the final construct (final_plasmid_optimized) is provided as Supplementary File, and the three source phage genomes are available from GenBank under accessions DQ386162 (M102AD), NC_029030 (φAPCM01) and MT430910 (SMHBZ8). The complete analysis pipeline and all custom scripts (including merge_annotations.py, classify_essentiality.py, filter_core_genes.py, select_best_genes.py, codon_optimize_dnachisel.py, verify_cai.py, and the figure-generation scripts), together with a README describing the run order, is publicly available on GitHub (<https://github.com/abdulazizailaax/phage-minimal-gene-set>, accessed on 1/5/2026) and permanently archived on Zenodo (DOI: 10.5281/zenodo.22907515).

Acknowledgments: The authors thank the Department of Health Sciences, Università del Piemonte Orientale, for institutional support. This manuscript reports the results of a master’s thesis in Medical Biotechnology and is presented here in journal-article format for submission. Latest versions of Claude Sonnet 5 were used for Word formatting, language check and initial review. We have carefully reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAI	Codon Adaptation Index
ORF	Open Reading Frame
CDS	Coding Sequence
RBS	Ribosome Binding Site
GC	Guanine-Cytosine (content)
COG	Clusters of Orthologous Groups
PDB	Protein Data Bank

bp base pair(s)

References

1. Glass, J.I.; Merryman, C.; Wise, K.S.; Hutchison, C.A.; Smith, H.O. Minimal Cells—Real and Imagined. *Cold Spring Harb. Perspect. Biol.* **2017**, *9*, a023861. <https://doi.org/10.1101/cshperspect.a023861>.
2. Gibson, D.G.; Glass, J.I.; Lartigue, C.; Noskov, V.N.; Chuang, R.Y.; Algire, M.A.; Benders, G.A.; Montague, M.G.; Ma, L.; Moodie, M.M.; et al. Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **2010**, *329*, 52–56. <https://doi.org/10.1126/science.1190719>.
3. Hutchison, C.A., 3rd; Chuang, R.Y.; Noskov, V.N.; Assad-Garcia, N.; Deerinck, T.J.; Ellisman, M.H.; Gill, J.; Kannan, K.; Karas, B.J.; Ma, L.; et al. Design and Synthesis of a Minimal Bacterial Genome. *Science* **2016**, *351*, aad6253. Erratum in: *ACS Chem Biol.* **2016**, *11*, 1463. <https://doi.org/10.1126/science.aad6253>.
4. Choe, D.; Cho, S.; Kim, S.C.; Cho, B.-K. Minimal Genome: Worthwhile or Worthless Efforts toward Being Smaller? *Biotechnol. J.* **2016**, *11*, 199–211. <https://doi.org/10.1002/biot.201400838>.
5. King, S.H.; Driscoll, C.L.; Li, D.B.; Guo, D.; Merchant, A.T.; Brix, G.; Wilkinson, M.E.; Hie, B.L. Generative Design of Bacteriophages with Genome Language Models. *Science* **2026**, *393*, eaec2657. <https://doi.org/10.1126/science.aec2657>.
6. Eckert, R.; Sullivan, R.; Shi, W. Targeted Antimicrobial Treatment to Re-Establish a Healthy Microbial Flora for Long-Term Protection. *Adv. Dent. Res.* **2012**, *24*, 94–97. <https://doi.org/10.1177/0022034512453725>.
7. Loesche, W.J. Role of Streptococcus Mutans in Human Dental Decay. *Microbiol. Rev.* **1986**, *50*, 353–380. <https://doi.org/10.1128/mr.50.4.353-380.1986>.
8. Bowen, W.H.; Burne, R.A.; Wu, H.; Koo, H. Oral Biofilms: Pathogens, Matrix, and Polymicrobial Interactions in Microenvironments. *Trends Microbiol.* **2018**, *26*, 229–242. <https://doi.org/10.1016/j.tim.2017.09.008>.
9. Kulis, E.; Cvitkovic, I.; Pavlovic, N.; Kumric, M.; Rusic, D.; Bozic, J. A Comprehensive Review of Antibiotic Resistance in the Oral Microbiota: Mechanisms, Drivers, and Emerging Therapeutic Strategies. *Antibiotics* **2025**, *14*, 828. <https://doi.org/10.3390/antibiotics14080828>.
10. Summers, W.C. The Strange History of Phage Therapy. *Bacteriophage* **2012**, *2*, 130–133. <https://doi.org/10.4161/bact.20757>.
11. Loc-Carrillo, C.; Abedon, S.T. Pros and Cons of Phage Therapy. *Bacteriophage* **2011**, *1*, 111–114. <https://doi.org/10.4161/bact.1.2.14590>.
12. Ben-Zaken, H.; Kraitman, R.; Copenhagen-Glazer, S.; Khalifa, L.; Alkalay-Oren, S.; Gelman, D.; Ben-Gal, G.; Beyth, N.; Hazan, R. Isolation and Characterization of Streptococcus Mutans Phage as a Possible Treatment Agent for Caries. *Viruses* **2021**, *13*, 825. <https://doi.org/10.3390/v13050825>.
13. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-Scale Protein Function Classification. *Bioinformatics* **2014**, *30*, 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>.
14. Gilchrist, C.L.M.; Chooi, Y.-H. Clinker & Clustermap.js: Automatic Generation of Gene Cluster Comparison Figures. *Bioinformatics* **2021**, *37*, 2473–2475. <https://doi.org/10.1093/bioinformatics/btab007>.
15. Cantalapiedra, C.P.; Hernández-Plaza, A.; Letunic, I.; Bork, P.; Huerta-Cepas, J. eggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **2021**, *38*, 5825–5829. <https://doi.org/10.1093/molbev/msab293>.
16. Kovach, M.E.; Elzer, P.H.; Hill, D.S.; Robertson, G.T.; Farris, M.A.; Roop, R.M.; Peterson, K.M. Four New Derivatives of the Broad-Host-Range Cloning Vector pBBR1MCS, Carrying Different Antibiotic-Resistance Cassettes. *Gene* **1995**, *166*, 175–176. [https://doi.org/10.1016/0378-1119\(95\)00584-1](https://doi.org/10.1016/0378-1119(95)00584-1).
17. Sargent, F.; Davidson, F.A.; Kelly, C.L.; Binny, R.; Christodoulides, N.; Gibson, D.; Johansson, E.; Kozyska, K.; Lado, L.L.; MacCallum, J.; et al. A Synthetic System for Expression of Components of a Bacterial Microcompartment. *Microbiology* **2013**, *159*, 2427–2436. <https://doi.org/10.1099/mic.0.069922-0>.
18. Valentin Zulkower, Susan Rosser, DNA Chisel, a versatile sequence optimizer, *Bioinformatics* **2020**, *36*, 4508–4509. <https://doi.org/10.1093/bioinformatics/btaa558>.
19. Sharp, P.M.; Li, W.H. The Codon Adaptation Index—A Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications. *Nucleic Acids Res.* **1987**, *15*, 1281–1295. doi: 10.1093/nar/15.3.1281.
20. Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* **2014**, *30*, 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.

21. Nakamura, Y.; Gojobori, T.; Ikemura, T. Codon Usage Tabulated from International DNA Sequence Databases: Status for the Year 2000. *Nucleic Acids Res.* **2000**, *28*, 292. <https://doi.org/10.1093/nar/28.1.292>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.