Title: Comparative Genomics of Three Novel Lytic Jumbo Bacteriophages Infecting *Staphylococcus aureus*

Abby M. Korn^{a,b}, Andrew E. Hillhouse^{c,d}, Lichang Sun^{b,e}, Jason J. Gill^{a,b,*}

^a Department of Animal Science, Texas A&M University, Kleberg Center, Suite 133 2471, 474 Olsen Blvd, College Station, TX 77843, USA. abbykorn@tamu.edu

^b Center for Phage Technology, Texas A&M University, 300 Olsen Blvd, College Station, TX 77843, USA. abbykorn@tamu.edu

^c Department of Veterinary Pathobiology, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843

^d Texas A&M Institute for Genome Sciences and Society, Texas A&M University College Station, TX 77884

^e Key Laboratory of Control Technology and Standard for Agro-product Safety and Quality Ministry of Agriculture, Key Laboratory of Food Quality and Safety of Jiangsu Province-State Key Laboratory Breeding Base, Institute of Food Quality and Safety, Jiangsu Academy of Agricultural Sciences, Nanjing, China

*Corresponding author: jason.gill@tamu.edu

Running Title: Jumbo *S. aureus* phages

Key words: *Staphylococcus aureus*, bacteriophage, jumbo phage, MarsHill, Madawaska, Machias

Abstract

The majority of previously described *Staphylococcus aureus* bacteriophages belong to three major groups: P68-like *Podoviridae*, Twort-like or K-like *Myoviridae*, and a more diverse group of temperate *Siphoviridae*. Here we present three novel *S. aureus* “jumbo” phages: MarSHill, Madawaska, and Machias. These phages were isolated from swine production environments in the United States and represent a novel clade of *S. aureus Myoviridae* that is largely unrelated to other known *S. aureus* phages. The average genome size for these phages is ~269 kb with each genome encoding ~263 predicted protein-coding genes. Phage genome organization and content is most similar to known jumbo phages of *Bacillus*, including AR9 and vB_BpuM-BpSp. All three phages possess genes encoding complete viral and non-viral RNA polymerases, multiple homing endonucleases, and a retron-like reverse transcriptase. Like AR9, all of these phages are presumed to have uracil-substituted DNA which interferes with DNA sequencing. These phages are also able to transduce host plasmids, which is significant as these phages were found circulating in swine production environments and can also infect human *S. aureus* isolates.

Importance of work:

This study describes the comparative genomics of three novel *S. aureus* jumbo phages: MarSHill, Madawaska, and Machias. These three *S. aureus Myoviridae* represent a new class of *S. aureus* phage that have not been described previously. These phages have presumably hypermodified DNA which inhibits sequencing by several different common platforms. Therefore, not only are these phages an exciting new type of *S. aureus*

phage, they also represent potential genomic diversity that has been missed due to the limitations of standard sequencing techniques. The data and methods presented in this study could be useful for an audience far beyond those working in *S. aureus* phage biology. This work is original and has not been submitted for publication in any other journal.

Introduction

Staphylococcus aureus is an opportunistic pathogen of both humans and animals, and is a leading cause of bacteremia, skin, soft tissue and device-related infections in humans (1). The expansion and prevalence of methicillin-resistant *S. aureus* (MRSA) imposes a significant burden to the health care system (2). *S. aureus* infections, particularly MRSA infections, can be difficult and costly to treat, with one study reporting the median cost for treatment of a MRSA surgical site infection as \$92,363 (3).

Carriage of *S. aureus* in the general public in the continental US ranges from 26% to 32% (4). An estimated 1.3% of that *S. aureus* being MRSA (5). However, in individuals in the US that are swine farmers, production workers or veterinarians, carriage of multi-drug resistant *S. aureus* (MDRSA) is two to six times greater than individuals in the community, or those who are not exposed to swine (6, 7). As with humans, *S. aureus* is considered to be part of the normal bacterial flora of swine (8). Certain activities on swine farms such as pressure washing and tail docking generate particle sizes capable of depositing primarily in human upper airways but also the primary and secondary bronchi as well as terminal bronchi and alveoli (9). Livestock-

associated methicillin-resistant *S. aureus* (LA-MRSA) isolates have been found to have a half-life of five days in settled barn dust with an approximate die-off of 99.9% after 66-72 days (10). Therefore, a better understanding of the ecology of *S. aureus* in these environments and mitigation of MRSA in swine production facilities would be of benefit to farmers and workers for safety reasons, as MRSA isolates are able to persist and possibly spread throughout the environment and workers.

Bacteriophages (phages) are viruses that infect bacteria and are the most abundant organism on Earth with an estimated 10^{31} in the biosphere (11). However, despite their abundance and possible utility there are still many unknowns about the basic biology of most phages and their interactions with their hosts in the environment. Previous investigations of *S. aureus* phages have described three common classes of phages: small, virulent P68-like *Podoviridae* with genomes of ~18-20 kb, various temperate *Siphoviridae* with genomes of ~45 kb, and large, virulent Twort-like *Myoviridae* with genomes of ~130 kb (12, 13). There is a single report of a novel large *S. aureus* *Myoviridae* that appears to be distinct from the K-like *Myoviridae*; this phage was named S6 as reported by Uchiyama *et al.*, in 2014 (14). While S6 was estimated to have a 270 kb genome by pulse-field gel electrophoresis and contain DNA in which thymine was replaced by uracil, no genomic sequence for S6 has been reported. Phages that have genomes over 200 kb are classified as “jumbo” phages (15). Most jumbo phages have been isolated from Gram-negative hosts, with the only known jumbo phages for Gram-positive hosts infecting *Bacillus* spp. Jumbo phage phylogeny and taxonomy are complicated by their often distant relationships to each other and to other known phage types; jumbo phage genomes typically encode large numbers of

hypothetical proteins (15). This study describes three novel *S. aureus* jumbo phages isolated from swine environments: Marshill, Machias and Madawaska.

Methods

Culture and maintenance of bacteria and phages

S. aureus was routinely cultured on trypticase soy broth (Bacto TSB, Difco) or trypticase soy agar (TSA, TSB + 1.5% w/v Bacto agar, Difco) aerobically at 30 °C. Phages were cultured using the double-layer overlay method (16) with 4 ml of top agar (10 g/L Bacto Tryptone (Difco), 10 g/L NaCl, 0.5% w/v Bacto agar, (Difco)) supplemented with 5 mM each CaCl₂ and MgSO₄ over TSA bottom plates. Lawns were inoculated with 0.1 ml of a mid-log *S. aureus* bacterial culture grown to an OD₅₅₀ of ~0.5. Phage stocks were produced by the confluent plate lysate method (17) using the original phage isolation host and harvested with 4-5 ml of lambda diluent (100 mM NaCl, 25 mM Tris-HCl pH 7.4, 8 mM MgSO₄, 0.01% w/v gelatin). Harvested lysates were centrifuged at 10,000 x g, 10 min, 4 °C, sterilized by passage through a 0.2 µm syringe filter (Millipore, Burlington, MA) and stored in the dark at 4 °C. Phages that could not be cultured to high titers using the double-layer overlay method were propagated in liquid culture by inoculating TSB 1:100 with an overnight culture of the appropriate *S. aureus* host strain and incubating at 30 °C with shaking (180 rpm) until OD₅₅₀ of ~0.5 was reached. The culture was inoculated with phage at a multiplicity of infection (MOI) of 0.01 and incubated at 30 °C with shaking (180 rpm) overnight. The next day the lysate was harvested by centrifugation (10,000 x g, 10 min, 4 °C) and the supernatant was sterilized by passage through a 0.2 µm syringe filter (Millipore).

Phage isolation

Phages were isolated from environmental swabs collected from 2015 - 2017 in swine production facilities in the United States. Five swabs per site were collected, targeting areas in production facilities with visible residue such as floor slats or water lines. Swabs were collected dry or were wetted in Stuart's Medium (BD BBL CultureSwab) for transport. Swab heads were aseptically clipped into a 50 ml conical tube (Falcon Corning) and eluted in sterile TSB with shaking for 2 hours at room temperature. Swabs were pooled by collection site. The swab eluates were centrifuged at 8,000 x g, 10 min, at 4 °C, and the supernatants passed through 0.2 µm syringe filters (Millipore). Samples were enriched for phage using a mixed enrichment approach (18, 19) by mixing 12 ml sample supernatant with 8 ml TSB, inoculating with 0.2 ml of a mixed *S. aureus* culture, and incubating overnight at 30 °C with aeration. Enrichment inocula were prepared by mixing equal volumes of overnight *S. aureus* cultures. The sample collected from the O.D. Butler, Jr. Animal Science Complex (College Station, TX) was enriched with a mixture of four *S. aureus* strains: USA300-0114, HT20020371, HT20020354 and CA-513 (**Table 1**). All other samples were enriched using two separate mixed enrichment panels: one panel consisted of a mixture of four human-associated *S. aureus* isolates and the other contained six swine-associated isolates (**Table 1**). Enrichment cultures were centrifuged (8,000 x g, 10 min, 4 °C) and the supernatants sterilized by passage through a 0.2 µm syringe filter (Millipore) and stored at 4 °C. Samples were screened for the presence of phage by spotting 10 µl aliquots onto soft agar lawns inoculated with each individual *S. aureus* strain used for

enrichment; samples were scored as positive for phage if they produced clearing zones or visible plaques on a lawn of any *S. aureus* host. Phages were isolated from phage-positive samples by dilution and plating to lawns of *S. aureus* followed by picking of well-isolated plaques. Each phage was subcultured three times to ensure clonality.

Phage DNA purification and sequencing

High titer ($>10^8$ PFU/mL) lysates of each phage were produced from plate lysates or liquid cultures as described above, and gDNA was extracted from 10-20 ml of phage lysate. Phage genomic DNA was extracted using the Wizard DNA Cleanup Kit (Promega) by a modified protocol as described previously (20, 21). Phage DNA was sheared to fragments of approximately 350 bp using a Diagenode Bioruptor® Pico (Diagenode Diagnostics). One μ g of DNA for each phage isolate was used for Illumina Truseq library preparation using the manufacturer's PCR-free library preparation protocol. Prepared libraries were quantified using KAPA qPCR library quantification kit (Roche) and diluted to 4 nM. Sequencing of phage DNA was performed by Illumina MiSeq V2 Nano 500 cycle chemistry. Library preparation and sequencing were performed at The Institute for Genomics and Society core laboratory at Texas A&M University.

FastQC (22), and SPAdes 3.5.0 (23) were used for read quality control and read assembly, respectively. Analysis of phage contigs by PhageTerm (24) indicated a circularly permuted genome. Marshill termini were determined by PCR using primers (forward 5'-AGCAGGTATTACAGGCCATTT-3') (reverse 5'-GAAGACGAAGTTAATGAGGCTAGA-3') facing off the contig ends to produce a ~1000

bp product. DNA polymerase Phusion U (Thermo Fisher) was used in this end closure PCR and the PCR product was sequenced by Sanger sequencing to close and correct the MarsHill contig. The Madawaska and Machias contigs were closed by reassembling the genomes with SPAdes (23) to produce contigs in which the original contig ends were internal to the genome with >40X coverage; this information was used to correct and close these contigs.

Restriction digests of phage DNA

To distinguish different phage types approximately 300 ng of extracted gDNA was digested with both DraI (5' TTAAA 3'), EcoRI-HF (5' GAATTC 3') (New England BioLabs Inc., Ipswich, MA). gDNA samples that showed poor banding patterns or could not be digested by the enzymes listed above were then digested with TaqαI (5' TCGA 3') and MspI (5'CCGG3'). For DraI, EcoRI-HF and MspI 300 ng of gDNA from each phage was incubated with each enzyme and CutSmart® Buffer at 37 °C overnight. For gDNA treated with TaqαI, 300 ng of gDNA was incubated with TaqαI and CutSmart® Buffer at 65 °C for two hours. After incubation, 4 µL of loading dye was then added and the total 24 µL for each sample was run on a 1% agarose gel at 90 V for 2.5 h.

Phage genome annotation

Genome annotation was conducted on the Center for Phage Technology Galaxy-Apollo genome annotation platform using the Galaxy pipelines described in (25). Briefly, genes were identified using Glimmer v3 (26) and MetaGeneAnnotator v1.0 (27), and tRNAs were identified using ARAGORN v2.36 (28). Protein function annotations

were assigned based on results from BLAST v2.9.0 against the nr and SwissProt databases (29, 30) InterProScan v5.33 (31), HHPred (32), and TMHMM v2.0 (33). Genomic DNA sequence was compared to that of other phages using BLASTn against the nt database, and using ProgressiveMauve v2.4 (34).

Transduction experiments

The ability of phage MarshHill to transduce host DNA into a recipient cell was determined using methods based on those described in Stanczak-Mrozek et al., 2015 (35). *S. aureus* strains Xen 36 (PerkinElmer; containing a plasmid-borne kanamycin resistance marker) and CA-347 (NRS648; a USA 600 MRSA strain with a chromosomal *mecA* marker) were used as donor *S. aureus* strains, and the swine-associated strain PD17 was used as the recipient. Phage MarshHill was propagated on donor strains in liquid culture as described above. PD17 was grown in 10 ml TSB supplemented with 5 mM MgCl₂ at 37 °C to ~5x10⁸ CFU/ml. MarshHill lysate propagated on a donor strain was added to the recipient culture at an MOI of 0.1 and incubated for 20 minutes at 30 °C. The culture was centrifuged at 13,000 x g for 2 minutes and the cell pellet was resuspended in 1 ml TSB with 50 mM sodium citrate. This culture was then incubated at 37 °C for 1 h, washed twice in TSB by pelleting and resuspension, and resuspended in 10 ml TSB. This culture was plated in 1 ml aliquots onto TSA plates supplemented with either 150 µg/ml kanamycin (for lysates propagated on Xen 36) or 5 µg/ml oxacillin (for lysates propagated on CA-347) and incubated overnight at 37 °C. Transduction efficiency was calculated as the number of transductants (CFU)/number of input phage particles (PFU), with a detection limit of 2x10⁻⁹.

Transmission electron microscopy

Transmission electron micrographs of each phage were obtained by adhering the phage to a carbon film by the Valentine method (36), and staining with 2% uranyl acetate. Phage were viewed in a Jeol 1200EX TEM at 100 kV accelerating voltage.

Results and Discussion

Phage isolation and characterization

Phages MarshHill, Madawaska and Machias were isolated from *S. aureus*-enriched environmental samples collected from swine production facilities located in Texas and Minnesota (**Table 2**). MarshHill and Madawaska were isolated against a human clinical USA300 MRSA host, and Machias was isolated on an ST9 MSSA isolate of swine origin. Phage enrichment and culture was routinely conducted at 30 °C, and phages MarshHill, Madawaska and Machias were found to be unable to form plaques or propagate in liquid culture at 37 °C. Temperature sensitivity between 37 °C and 30 °C was also noted for the *S. aureus* jumbo phage S6 (14), and may be a limiting factor for recovery of this type of phage from the environment.

Transmission electron microscopy of MarshHill revealed a large phage with typical *Myoviridae* morphology (**Figure 1**). MarshHill was observed to have a mean head size of ~115 nm ($\pm$ 6 nm) and a tail length of ~233 nm ($\pm$ 4 nm). For comparison, the K-like *S. aureus* phage phi812 has a reported head diameter of 90 nm and a tail length of 240 nm (37). The measurements of the jumbo phage reported here are comparable to those of phage S6, with a reported head diameter of 118 nm and tail length of 237 nm

(14). Although the genome sequence of S6 is not available as of this writing, the large size and temperature-dependent growth characteristics reported for this phage makes it likely to be related to the jumbo phages reported here.

DNA sequencing and genomic characterization

Genomic DNA extracted from all three phages was not able to be digested by EcoRI-HF (5' GAATTC 3') or DraI (5' TTAAA 3'), but could be digested with MspI (5' CCGG 3') and Taq- α I (TaqI) (5' TCGA 3') (**Figure 2**). TaqI was used as it has been shown to cut heavily modified phage DNA such as that of SPO1, which contains hydroxymethyluracil in place of thymine in its DNA (38). These results are consistent with phage DNA modifications to adenine or thymine bases, as MspI, which lacks A or T bases in its recognition sequence, was also able to digest this DNA.

Obtaining the genomic DNA sequence of these phages was challenging. Multiple attempts at standard DNA sequencing approaches using library preparation by Illumina TruSeq Nano or Swift Accel-NGS 1S plus (Swift Biosciences) followed by MiSeq V2 Nano 500 cycle sequencing produced no detectable phage DNA sequence. Pacific BioSciences SMRTbell Express (Pacific Biosciences) failed to produce sequencable DNA libraries. Sequencing by MinION (Oxford Nanopore) produced sequence data that could not be interpreted, even though the DNA sequence of phage lambda could easily be recovered as the positive control for the sequencing run. Illumina library preparation by a PCR-free library protocol, which required significantly higher amounts of input DNA, produced DNA libraries sequencable by Illumina MiSeq. Modification or base substitution of phage DNA has been reported as an impediment to

sequencing for other phages, such as the 5-hydroxymethyluracil substitution for thymine found in the *Bacillus* phage CP-51 (39), or uracil substitution for thymine in the *Bacillus* jumbo phage AR9 (40). Phage AR9 was sequenced by Illumina following library amplification using a uracil-compatible DNA polymerase (40) and CP-51 was only sequencable by a combination of PacBio and Sanger sequencing (39). The previously reported *S. aureus* jumbo phage S6 was also determined to contain DNA with thymine completely replaced by uracil (14). The behavior of the DNA of the MarshHill-like phages reported here: resistance to digestion by restriction enzymes with A/T bases in the recognition site, sensitivity to digestion by TaqI, and inability to be amplified by PCR using polymerases other than the uracil-insensitive PhusionU, is consistent with such a thymine-uracil base substitution as reported for phages AR9 and S6 (14, 40).

The substitution of non-standard bases in phage DNA is likely an adaptation that provides broad protection of the phage chromosome from cleavage by host restriction systems (41). The glycosylated 5-hydroxymethylcytosine (glc-HMC) of coliphage T4 protects phage DNA from not only HMC-specific nucleases but also CRISPR-Cas9 (42). Substituting uracil for thymine is also a means of circumventing host defenses; *B. subtilis* phage SPO1 has hydroxymethyluracil in place of thymine, and replacement of this base with thymine renders the phage DNA more susceptible to restriction enzymes (43, 44). It is worth noting that the incompatibility of such hypermodified phage DNA with standard sequencing approaches likely impacts various metaviromic studies of natural phage communities, as this DNA will be rendered “invisible” to such approaches unless procedures are adapted to capture this sequence.

Genomic analysis of MarshHill-like phages, a clade of phages related to *Bacillus* jumbo phages

Phages MarshHill, Machias and Madawaska were sequenced by Illumina MiSeq to final coverages of 271-, 151- and 212-fold, respectively. These phages are “jumbo” phages with an average genome size of ~269 kb (**Table 2**). The phages’ G+C content of ~25% is lower than the median G+C content of 32.7% found in their *S. aureus* host. Comparison of phage DNA sequences by progressiveMauve (34) (**Figure 3**), found MarshHill and Madawaska were more closely related (92.4% identity), with the Machias DNA sequence more diverged (61.5% identity with MarshHill, and 61.8% identity with Madawaska). In addition to being ~8 kb longer than either MarshHill or Madawaska, Machias contains an inversion in its genome that corresponds to MarshHill genes 171 to 193. Differences between these genomes are primarily confined to loss or replacement of individual genes, typically in predicted homing endonucleases or hypothetical genes (**Figure 3**). A blastclust comparison of the 789 predicted proteins found across the three *S. aureus* jumbo phages (threshold of >50% identity aligned over >50% protein length) generated 311 protein clusters, with 197 of these conserved in all three phages. Sixty-two of these protein clusters were singletons found in only a single phage, with the majority (46/62, 74%) found only in Machias. The results of the protein cluster analysis are provided in **Supplementary Table S1**.

There is little detectable DNA sequence similarity between the MarshHill-like phages and any other organisms in the NCBI database (<5% alignable DNA sequence identity). Comparison of the predicted MarshHill, Madawaska and Machias proteomes to other phages by the CPT Galaxy comparative genomics workflow (25) shows that these

phages are most closely related to previously described *Bacillus jumbo* phages AR9 (NC_031039 (32)) PBS1 (NC_043027), and vB_BpuM-BpSp (KT895374.1 (45)) (**Table 3**). These phages are more distantly related to jumbo phages phiR-37 and MS32 infecting Gram-negative hosts *Yersinia enterocolitica* and *Vibrio mediterranei*, respectively. Phage Machias shares a greater number of proteins with these other jumbo phages, and also shares more proteins with the staphylococcal myophage Twort and other Twort-like phages (**Table 3**). Genes shared between the MarshHill-like and Twort-like phages are primarily those encoding homing endonucleases (14 of 21 shared proteins), ribonucleotide reductases (2 proteins) and a predicted amidase. The AR9-like jumbo phages are part of a larger clade of phiKZ-like jumbo phages that can be found infecting Gram-negative and Gram-positive hosts (38).

The genomes of phages MarshHill, Madawaska and Machias were annotated using the CPT Galaxy-Apollo toolset (25). Due to the overall similarity of the three phages, they will be discussed primarily using phage MarshHill as a reference. The phage genomes were closed and then analyzed by PhageTerm (24) to determine genomic termini, which indicated phage chromosomes with *pac*-like circular permutation. The phage genomes were reopened immediately upstream of the genes encoding predicted non-viral RNA polymerase β and β' subunits N-terminal domains (**Figure 4**). The MarshHill genome encodes 262 predicted proteins, of which 76 could be assigned functions. The genomes exhibit a general lack of clear modular organization, which is common in jumbo phages (15). Multiple structural components could be identified bioinformatically in the MarshHill genome, based on the presence of conserved domains or their similarity to proteins previously described in phages AR9 (40, 46) and

vB_BpuM-BpSp (45, 47), including the major capsid protein, predicted tail fibers, tail sheath, baseplate wedge, prohead scaffold and portal protein (**Table S1**). Two potential tape measure proteins were identified; both contain significant alpha-helical content as well as several predicted transmembrane domains which is indicative of tape measure proteins (48), but one contains a detectable peptidoglycan-degrading domains (PF18013, PF00877), and thus was annotated as a potential tail lysin. All three phages contain dCMP deaminase genes, but other genes associated with modifications to uracil such as dUMP hydroxymethylase or HMdUMP kinase were not identified in the genomes of the *S. aureus* jumbo phages (44). A GroEL-like chaperonin protein was identified in all three *S. aureus* jumbo phage genomes; a similar GroEL-like protein has also been identified in phage AR9 (40). The GroEL-like protein in AR9 has been shown to possess chaperone activity without requiring a co-chaperonin to function when purified and expressed in *E. coli* (49). In coliphage T4, a GroEL homolog (T4 gp31) is essential for folding of the major capsid protein and aids in the assembly of phage particles (50), and this jumbo phage chaperonin likely plays a similar role in folding key viral proteins.

MarsHill gp047 was annotated as the phage holin due to the presence of conserved domains associated with staphylococcal and streptococcal phages (pfam04531, TIGR01598), a conserved protein domain for the holin of lactococcal bacteriophage phi LC3 (51). However, two MarsHill proteins (gp153, gp203) possess characteristics of phage endolysins, thus both proteins have been annotated simply as N-acetylmuramoyl-L-alanine amidases. MarsHill gp203 has similarity to amidases in phages AR9 (40) and PBS1 (NC_043027.1). However, this similarity is to AR9 gp272, not

to AR9 gp194, which is noted to have a PlyB-like endolysin conserved domain (cd06523) and is not conserved in the MarshHill-like phages. The second MarshHill amidase, gp153, is more closely related to bacterial amidases and does not share a homolog in AR9. It is not clear if one or both of the identified amidases function as the endolysin in the MarshHill-like phages; these phages may have “captured” a second amidase gene with redundant endolysin activity.

A large class of phiKZ-like jumbo phages are known to encode their own DNA-directed RNA polymerases (40). These jumbo phage RNA polymerases differ from the bacterial RNA polymerase in that they lack α and ω subunits, with only β and β' subunits identified (52). These jumbo phages encode two complete RNA polymerase holoenzymes, one of which (the “viral” RNA polymerase) is packaged into the virion and presumably ejected into the host cell upon infection (53), and the other, “non-viral” RNA polymerase expressed during the phage lytic cycle (54). Identification and annotation of the jumbo phage RNA polymerase genes is complicated by their arrangement: each of the viral and non-viral β and β' RNA polymerase subunits are expressed from multiple loci, encoding subdomains of each of these subunits. The viral RNA polymerase is comprised of β and β' subunits, with the β subunit expressed from two genes encoding N- and C-terminal domains, and the β' subunit encoded by three genes encoding N-, mid- and recently-discovered C-terminal domains (54). The non-viral RNA polymerase is comprised of β and β' subunits, each encoded by two genes corresponding to N- and C-termini, and an additional, fifth protein subunit, which appears to be involved in phage-specific promoter recognition (55). The jumbo phage RNA polymerase subunit genes are also frequently disrupted by one or more introns (40). Genes encoding five

viral RNA polymerase subunits (β -N, β -C, β' -N, β' -mid, β' -C) and five non-viral RNA polymerase subunits (β -N, β -C, β' -N, β' -C, specificity determinant) were identified in the MarshHill-like jumbo phages based on the presence of conserved domains and similarity to homologs in related phages.

Comparison of the predicted MarshHill, Madawaska and Machias proteins with their homologs in other phages indicated that multiple genes are interrupted by one or more intron sequences. The presence of intron-disrupted genes is well-known in K-like *S. aureus* phages (56), in the *Bacillus* phages SPO1 and SP82 (57), and the *Bacillus* jumbo phage AR9 (40). As shown in **Table 4**, introns were detected in genes encoding multiple RNA polymerase subunits, the prohead scaffold/protease, DNA polymerase subunits, large terminase, and DNA gyrase B. A similar repertoire of genes is known to be disrupted by introns in AR9 (40). Based on protein sequence alignment to their intact homologs in AR9 and other related organisms, the number of exons in disrupted genes in MarshHill, Madawaska and Machias was predicted (**Table 4**). This approach allowed for the prediction of the mature peptides for all intron-disrupted genes except for the large terminase subunit (MarshHill genes 70 and 71), which could not be confidently reconstructed based on alignment to any intact protein homolog. Some intron-disrupted genes identified in AR9, such as the non-viral RNA polymerase β' N-terminal subunit, were found to be intact in MarshHill, while genes that are intact in AR9, such as the DNA polymerase I, were found to be disrupted in the MarshHill-like phages. The predicted number of exons and exon boundary positions was variable between the three *S. aureus* jumbo phages described here, which is consistent with the highly plastic nature of these regions (40).

Analysis of the protein sequence of the MarshHill DNA gyrase A (gp228) indicated that this gene is disrupted by an intein, based on the presence of intein-associated conserved domains (IPR006142, IPR003587, IPR003586). The intein boundaries in MarshHill gp228 were predicted based on alignment with its intein-less homolog in AR9 (AR9 gp044) and the presence of a conserved cysteine at the N-terminal boundary and the conserved residues His-Asn at the C-terminal boundary (58). Based on this analysis, the intein is predicted to be inserted between residues Y117 and T492 of gp228. The gp228 intein is also predicted to contain a homing endonuclease, based on the presence of a LAGLIDADG endonuclease conserved domain (IPR004860); such domains are a common feature of intein sequences (59). The AR9 homolog of the MarshHill DNA gyrase A appears to lack this intein but is instead disrupted by two intron sequences (40), indicating this gene is a common target for mobile DNA elements.

Previous analysis of AR9 transcription signals identified a AT-rich early promoter that is expressed by the viral RNA polymerase (46). DNA sequence 200 bp upstream of all protein-coding genes was extracted from the MarshHill genome and analyzed for sequence motifs using MEME (60). An AT-rich motif nearly identical to that observed in phage AR9, particularly the consensus motif 5' TATATTAT 3', was identified upstream of 29 genes in the MarshHill genome by MEME (**Figure 5**). A second conserved signal corresponding to the phage late promoter was not identified by this approach. Transcriptomic profiling of AR9 has indicated that this signal is somewhat less conserved with a consensus motif of AACAN₆TW (46), and thus is difficult to predict *de novo* without additional data from phage transcripts.

All three *S. aureus* jumbo phages encode a predicted RNA-dependent DNA polymerase (reverse transcriptase), exemplified by MarsHill gp206. This 501-residue protein contains a C-terminal retron-like reverse transcriptase domain (cd03487) and a ~200 residue N-terminal domain of unknown function. Reverse transcriptases have been previously identified in phages as part of diversity-generating retroelements (DGR's) which selectively mutagenize the phage tail fiber to extend phage host range (61, 62). The MarsHill gp206 reverse transcriptase does not appear to be a part of such a system, as the accessory variability determinant (Avd), TR, VR and IMH sequences typically associated with DGR's were not identified in MarsHill. The MarsHill reverse transcriptase is more closely related to proteins associated with bacterial retroelements, with homologs found in *Bacillus* (TQJ37342, $E=10^{-26}$) and *Marinimicrobium* (ROQ19867, $E=10^{-21}$). The MarsHill reverse transcriptase is quite diverged from other sequences present in NCBI, with the closest bacterial homologs sharing only ~30% sequence identity. Immediately upstream of MarsHill gp206 is a 1,200 bp non-coding region which is hypothesized to contain the retroelement-associated ncRNA (63). Bacterial retroelements have recently been implicated as a type of bacterial anti-phage defense mechanism which leads to abortive infection when an infecting phage inhibits RecBCD, however the MarsHill reverse transcriptase does not appear to be associated with a potential effector protein as described in anti-phage elements (63). If this element serves as an anti-phage system (e.g., to exclude co-infecting phages), it is possible that the gene encoding such an effector has become unlinked in the MarsHill genome. It is also possible that the gp206 reverse transcriptase serves some other function, or is simply present as a selfish genetic element (64).

There are no identified tRNAs in phage MarshHill, with Madawaska having one and Machias two pseudo-tRNA genes as identified by tRNAscan. The presence of tRNAs is positively associated with genome size in phages and is thought to be driven by divergence of phage codon usage from that of the host (65). In these phages, tRNA genes appear to be degenerating, with aberrant pseudo-tRNA genes appearing in two of the phages and these genes being undetectable in MarshHill, indicating a lack of selection pressure to maintain these functions (66).

Phage MarshHill is able to transduce host DNA

Phages AR9 of *Bacillus* and S6 of *S. aureus* have been previously described as generalized transducing phages, capable of aberrantly packaging host DNA and transferring it to a new cell (40) (14). The ability of MarshHill to transduce chromosomal and plasmid-borne antibiotic resistance marker genes was determined experimentally. MarshHill was able to transduce plasmid-borne kanamycin resistance from *S. aureus* strain Xen 36 to wild-type *S. aureus* isolate PD17 at an efficiency of 7×10^{-9} transductants/PFU, based on three replicate experiments. This transduction frequency is comparable to several different *pac*-type *S. aureus* *Siphoviridae*, which are able to transduce plasmids of varying sizes at a frequency of 10^{-5} to 10^{-11} transductants/ PFU (67). No transductants were recovered using lysates of the virulent *S. aureus* phage K or induced culture supernatants of Xen 36, indicating that any endogenous temperate phages residing in Xen 36 are not produced at high enough levels to account for the observed transduction. Phage K is not expected to be able to perform generalized transduction, as it has been reported to degrade the host chromosome shortly after

infection (68), an activity which is associated with lack of transducing capacity in other phages such as the coliphage T4 (69). Transduction experiments conducted with the donor strain CA-347, a USA600 MRSA strain, did not yield any detectable transductants expressing *mecA*-mediated oxacillin resistance in the recipient cultures using either phage or culture supernatants. This observed inability to transduce chromosomal markers is not conclusive as the measured transduction frequency for plasmid DNA was near the detection limit of the assay (2×10^{-9} transductants/PFU), and transduction of the *mecA* locus by temperate *Siphoviridae* phages has been reported ranging from 1×10^{-9} to 2×10^{-10} transductants/PFU (70). Phage S6 was reported to transduce plasmid pCU1 at an efficiency of 10^{-7} transductants per cell and a plasmid containing the *mecA* gene at an efficiency of 5.2×10^{-11} per cell into the restriction-deficient *S. aureus* strain RN4220. S6-mediated transduction of the pCU1 plasmid was also tested for several other staphylococci including *S. epidermidis*, *S. pseudintermedius*, *S. sciuri* and *S. felis*, with reported efficiencies of 10^{-7} to 10^{-10} per cell (14). The transduction efficiency of MarshHill may be expected to be lower than that observed for S6, as the recipient strain used to measure MarshHill transduction is a wild-type *S. aureus* strain recently isolated from the environment which presumably has a functional restriction system, unlike the restrictionless *S. aureus* strain RN4220.

Conclusion

Three novel *S. aureus* *Myoviridae* phages isolated from swine environments across the US presented unique challenges in both culturing and sequencing. Multiple commonly used approaches failed to produce usable sequence for these phages,

489 however using a PCR-free library preparation method produced libraries sequencable
 490 by Illumina. This study raises questions about the diversity of phages with modified DNA
 491 that are missed by metagenomic studies using traditional library preparations to identify
 492 phage sequences. Complete genomes were obtained for phages Marshill, Madawaska
 493 and Machias, with all phages having genomes well over 200 kb, classifying them as
 494 members of a new jumbo phage lineage in *S. aureus* (15). The *S. aureus* phages in
 495 this study were isolated from environmental samples at 30 °C instead of the traditional
 496 37 °C, in hopes of more closely mimicking ambient barn temperatures as well as human
 497 nasopharynx temperatures which average around 34 °C (71). Without this adjustment
 498 these phages would not have been isolated as they do not form observable plaques or
 499 propagate efficiently in liquid culture at 37 °C. These three phages are only distantly
 500 related to other known phages, with the most similar being *Bacillus* phage vB_BpuM-
 501 BpSp (45) and *Bacillus* phage AR9 (40). All three phages carry identifiable viral and
 502 non-viral RNA polymerase subunits, many of which are intron-disrupted. Intron and
 503 intein disruption of multiple phage genes was identified, including the viral and non-viral
 504 RNA polymerases, DNA polymerase, and DNA gyrase. Mature peptides were predicted
 505 for all intron-disrupted genes by alignment to homologs in AR9 and other organisms,
 506 except for the phages' TerL. Unlike phage AR9 all three phages contain an RNA-
 507 directed DNA polymerase, the function of which is unclear. Marshill was able to
 508 transduce plasmid DNA at an efficiency of 7×10^{-9} transductants/PFU which indicates
 509 that these phages, in addition to temperate *S. aureus* *Siphoviridae*, are also possible
 510 vectors for horizontal gene exchange in *S. aureus*. This is particularly relevant as these

phages were found to be circulating in swine production environments and are able to infect both swine and human *S. aureus* isolates.

Acknowledgements

This work was supported by the National Pork Board and Texas AgriLife Research. We would like to thank Dr. H. Morgan Scott and Dr. Andrew Hillhouse for their generosity and time in assisting in the sequencing of these phages. We would also like to thank Dr. Peter Davies for providing swine *S. aureus* isolates and for his help in locating phage sampling sites. We acknowledge the support of the Network on Antimicrobial Resistance in *Staphylococcus aureus* (NARSA) for provision of bacterial isolates.

Competing Interests

The authors declare no conflicts of interest.

References

1. Tong SY, Davis JS, Eichenberger E, Holland TL, Fowler VG, Jr. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clinical microbiology reviews*. 2015;28(3):603-61.
2. Valiquette L, Chakra CN, Laupland KB. Financial impact of health care-associated infections: When money talks. *The Canadian journal of infectious diseases & medical microbiology = Journal canadien des maladies infectieuses et de la microbiologie medicale*. 2014;25(2):71-4.
3. Engemann JJ, Carmeli Y, Cosgrove SE, Fowler VG, Bronstein MZ, Trivette SL, et al. Adverse clinical and economic outcomes attributable to methicillin resistance among patients with *Staphylococcus aureus* surgical site infection. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*. 2003;36(5):592-8.
4. Sivaraman K, Venkataraman N, Cole AM. *Staphylococcus aureus* Nasal Carriage and its Contributing Factors. *Future microbiology*. 2009;4:999-1008.

- 540 5. Salgado CD, Farr BM, Calfee DP. Community-acquired methicillin-resistant
541 *Staphylococcus aureus*: a meta-analysis of prevalence and risk factors. *Clinical*
542 *infectious diseases* : an official publication of the Infectious Diseases Society of
543 America. 2003;36(2):131-9.
- 544 6. Neyra RC, Frisancho JA, Rinsky JL, Resnick C, Carroll KC, Rule AM, et al.
545 Multidrug-resistant and methicillin-resistant *Staphylococcus aureus* (MRSA) in hog
546 slaughter and processing plant workers and their community in North Carolina (USA).
547 *Environmental health perspectives*. 2014;122(5):471-7.
- 548 7. Wardyn SE, Forshey BM, Farina SA, Kates AE, Nair R, Quick MK, et al. Swine
549 Farming Is a Risk Factor for Infection With and High Prevalence of Carriage of
550 Multidrug-Resistant *Staphylococcus aureus*. *Clinical Infectious Diseases*.
551 2015;61(1):59-66.
- 552 8. Frana TS. *Staphylococcosis*. In: Zimmerman JJ, Karriker LA, Ramirez A,
553 Stevenson GW, Schwartz KJ, editors. *Diseases of Swine* 10th ed. : Wiley; 2012. p. 834-
554 40
- 555 9. Madsen AM, Kurdi I, Feld L, Tendal K. Airborne MRSA and Total *Staphylococcus*
556 *aureus* as Associated With Particles of Different Sizes on Pig Farms. *Annals of work*
557 *exposures and health*. 2018;62(8):966-77.
- 558 10. Feld L, Bay H, Angen Ø, Larsen AR, Madsen AM. Survival of LA-MRSA in Dust
559 from Swine Farms. *Annals of work exposures and health*. 2018;62(2):147-56.
- 560 11. Comeau AM, Hatfull GF, Krisch HM, Lindell D, Mann NH, Prangishvili D.
561 Exploring the prokaryotic virosphere. *Res Microbiol*. 2008;159(5):306-13.
- 562 12. Lobocka M, Hejnowicz MS, Dabrowski K, Gozdek A, Kosakowski J, Witkowska
563 M, et al. Genomics of staphylococcal Twort-like phages--potential therapeutics of the
564 post-antibiotic era. *Advances in virus research*. 2012;83:143-216.
- 565 13. Xia G, Wolz C. Phages of *Staphylococcus aureus* and their impact on host
566 evolution. *Infection, Genetics and Evolution*. 2014;21:593-601.
- 567 14. Uchiyama J, Takemura-Uchiyama I, Sakaguchi Y, Gamoh K, Kato S, Daibata M,
568 et al. Intragenus generalized transduction in *Staphylococcus* spp. by a novel giant
569 phage. *The ISME journal*. 2014;8(9):1949-52.
- 570 15. Yuan Y, Gao M. Jumbo Bacteriophages: An Overview. *Frontiers in Microbiology*.
571 2017;8:403-.
- 572 16. Kropinski AM, Mazzocco A, Waddell TE, Lingohr E, Johnson RP. Enumeration of
573 bacteriophages by double agar overlay plaque assay. *Methods in molecular biology*
574 (Clifton, NJ). 2009;501:69-76.
- 575 17. Adams MH. *Bacteriophages*. New York: Interscience Publishers; 1959.
- 576 18. Gill JJ, Svircev AM, Smith R, Castle AJ. Bacteriophages of *Erwinia amylovora*.
577 *Appl Environ Microbiol*. 2003;69(4):2133-8.

19. Xie Y, Wahab L, Gill JJ. Development and Validation of a Microtiter Plate-Based Assay for Determination of Bacteriophage Host Range and Virulence. *Viruses*. 2018;10(4).
20. Summer EJ. Preparation of a phage DNA fragment library for whole genome shotgun sequencing. *Methods in molecular biology* (Clifton, NJ). 2009;502:27-46.
21. Gill JJ, Berry JD, Russell WK, Lessor L, Escobar-Garcia DA, Hernandez D, et al. The *Caulobacter crescentus* phage phiCbK: genomics of a canonical phage. *BMC Genomics*. 2012;13:542.
22. Andrew S. FastQC: a quality control tool for high throughput sequence data 2010 [Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
23. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-77.
24. Garneau JR, Depardieu F, Fortier L-C, Bikard D, Monot M. PhageTerm: a tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data. *Scientific Reports*. 2017;7(1):8292.
25. Ramsey J, Rasche H, Maughmer C, Criscione A, Mijalis E, Liu M, et al. Galaxy and Apollo as a biologist-friendly interface for high-quality cooperative phage genome annotation. *PLoS Comput Biol*. 2020;16(11):e1008214.
26. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. *Nucleic acids research*. 1999;27(23):4636-41.
27. Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res*. 2008;15(6):387-96.
28. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research*. 2004;32(1):11-6.
29. The UniProt C. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*. 2019;47(D1):D506-D15.
30. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10(1):421.
31. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236-40.
32. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of molecular biology*. 2018;430(15):2237-43.
33. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*. 2001;305(3):567-80.

34. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLOS ONE. 2010;5(6):e11147.
35. Stanczak-Mrozek KI, Manne A, Knight GM, Gould K, Witney AA, Lindsay JA. Within-host diversity of MRSA antimicrobial resistances. The Journal of antimicrobial chemotherapy. 2015;70(8):2191-8.
36. Valentine RC, Shapiro BM, Stadtman ER. Regulation of glutamine synthetase. XII. Electron microscopy of the enzyme from Escherichia coli. Biochemistry. 1968;7(6):2143-52.
37. Novacek J, Siborova M, Benesik M, Pantucek R, Doskar J, Plevka P. Structure and genome release of Twort-like Myoviridae phage with a double-layered baseplate. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(33):9351-6.
38. Huang LH, Farnet CM, Ehrlich KC, Ehrlich M. Digestion of highly modified bacteriophage DNA by restriction endonucleases. Nucleic acids research. 1982;10(5):1579-91.
39. Klumpp J, Schmuki M, Sozhamannan S, Beyer W, Fouts DE, Bernbach V, et al. The odd one out: Bacillus ACT bacteriophage CP-51 exhibits unusual properties compared to related Spounavirinae W.Ph. and Bastille. Virology. 2014;462-463(Supplement C):299-308.
40. Lavysh D, Sokolova M, Minakhin L, Yakunina M, Artamonova T, Kozyavkin S, et al. The genome of AR9, a giant transducing Bacillus phage encoding two multisubunit RNA polymerases. Virology. 2016;495:185-96.
41. Loenen WA, Raleigh EA. The other face of restriction: modification-dependent enzymes. Nucleic acids research. 2014;42(1):56-69.
42. Bryson AL, Hwang Y, Sherrill-Mix S, Wu GD, Lewis JD, Black L, et al. Covalent Modification of Bacteriophage T4 DNA Inhibits CRISPR-Cas9. MBio. 2015;6(3):e00648.
43. Hoet PP, Coene MM, Cocito CG. Replication cycle of Bacillus subtilis hydroxymethyluracil-containing phages. Annual review of microbiology. 1992;46:95-116.
44. Stewart CR, Casjens SR, Cresawn SG, Houtz JM, Smith AL, Ford ME, et al. The genome of Bacillus subtilis bacteriophage SPO1. Journal of molecular biology. 2009;388(1):48-70.
45. Yuan Y, Gao M. Characteristics and complete genome analysis of a novel jumbo phage infecting pathogenic Bacillus pumilus causing ginger rhizome rot disease. Arch Virol. 2016;161(12):3597-600.
46. Lavysh D, Sokolova M, Slashcheva M, Förstner KU, Severinov K. Transcription Profiling of Bacillus subtilis Cells Infected with AR9, a Giant Phage Encoding Two Multisubunit RNA Polymerases. mBio. 2017;8(1):e02041-16.
47. Yuan Y, Gao M. Proteomic Analysis of a Novel Bacillus Jumbo Phage Revealing Glycoside Hydrolase As Structural Component. Frontiers in Microbiology. 2016;7:745-.

48. Mahony J, Alqarni M, Stockdale S, Spinelli S, Feyereisen M, Cambillau C, et al. Functional and structural dissection of the tape measure protein of lactococcal phage TP901-1. *Scientific Reports*. 2016;6(1):36667.
49. Semenyuk PI, Moiseenko AV, Sokolova OS, Muronetz VI, Kurochkina LP. Structural and functional diversity of novel and known bacteriophage-encoded chaperonins. *International Journal of Biological Macromolecules*. 2020;157:544-52.
50. Linder CH, Carlson K, Albertioni F, Söderström J, Pålson C. A late exclusion of bacteriophage T4 can be suppressed by *Escherichia coli* GroEL or Rho. *Genetics*. 1994;137(3):613-25.
51. Birkeland NK. Cloning, molecular characterization, and expression of the genes encoding the lytic functions of lactococcal bacteriophage phi LC3: a dual lysis system of modular design. *Can J Microbiol*. 1994;40(8):658-65.
52. Sokolova ML, Misovets I, V Severinov K. Multisubunit RNA Polymerases of Jumbo Bacteriophages. *Viruses*. 2020;12(10):1064.
53. Wu W, Thomas JA, Cheng N, Black LW, Steven AC. Bubblegrams reveal the inner body of bacteriophage ϕ KZ. *Science*. 2012;335(6065):182.
54. Thomas JA, Benítez Quintana AD, Bosch MA, Coll De Peña A, Aguilera E, Coulibaly A, et al. Identification of Essential Genes in the Salmonella Phage SPN3US Reveals Novel Insights into Giant Phage Head Structure and Assembly. *Journal of virology*. 2016;90(22):10284-98.
55. Sokolova M, Borukhov S, Lavysh D, Artamonova T, Khodorkovskii M, Severinov K. A non-canonical multisubunit RNA polymerase encoded by the AR9 phage recognizes the template strand of its uracil-containing promoters. *Nucleic acids research*. 2017;45(10):gkx264.
56. O'Flaherty S, Coffey A, Edwards R, Meaney W, Fitzgerald GF, Ross RP. Genome of Staphylococcal Phage K: a New Lineage of Myoviridae Infecting Gram-Positive Bacteria with a Low G+C Content. *J Bacteriol*. 2004;186(9):2862-71.
57. Belfort M, Bonocora RP. Homing endonucleases: from genetic anomalies to programmable genomic clippers. *Methods in molecular biology (Clifton, NJ)*. 2014;1123:1-26.
58. Tori K, Dassa B, Johnson MA, Southworth MW, Brace LE, Ishino Y, et al. Splicing of the mycobacteriophage Bethlehem DnaB intein: identification of a new mechanistic class of inteins that contain an obligate block F nucleophile. *J Biol Chem*. 2010;285(4):2515-26.
59. Elleuche S, Pöggeler S. Inteins, valuable genetic elements in molecular biology and biotechnology. *Appl Microbiol Biotechnol*. 2010;87(2):479-89.
60. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*. 2009;37(Web Server issue):W202-8.

61. Dai W, Hodes A, Hui WH, Gingery M, Miller JF, Zhou ZH. Three-dimensional structure of tropism-switching Bordetella bacteriophage. *Proceedings of the National Academy of Sciences*. 2010;107(9):4347.
62. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, et al. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science*. 2002;295(5562):2091-4.
63. Millman A, Bernheim A, Stokar-Avihail A, Fedorenko T, Voichek M, Leavitt A, et al. Bacterial Retrons Function In Anti-Phage Defense. *Cell*. 2020.
64. Wang C, Villion M, Semper C, Coros C, Moineau S, Zimmerly S. A reverse transcriptase-related protein mediates phage resistance and polymerizes untemplated DNA in vitro. *Nucleic acids research*. 2011;39(17):7620-9.
65. Bailly-Bechet M, Vergassola M, Rocha E. Causes for the intriguing presence of tRNAs in phages. *Genome research*. 2007;17(10):1486-95.
66. Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. *Trends in genetics : TIG*. 2001;17(10):589-96.
67. Mašláňová I, Stříbná S, Doškař J, Pantůček R. Efficient plasmid transduction to Staphylococcus aureus strains insensitive to the lytic action of transducing phage. *FEMS microbiology letters*. 2016;363(19).
68. Rees PJ, Fry BA. The Morphology of Staphylococcal Bacteriophage K and DNA Metabolism in Infected Staphylococcus aureus. *Journal of General Virology*. 1981;53(2):293-307.
69. Fineran PC, Petty NK, Salmond GPC. Transduction: Host DNA Transfer by Bacteriophages. In: Schaechter M, editor. *Encyclopedia of Microbiology (Third Edition)*. Oxford: Academic Press; 2009. p. 666-79.
70. Scharn CR, Tenover FC, Goering RV. Transduction of staphylococcal cassette chromosome mec elements between strains of Staphylococcus aureus. *Antimicrob Agents Chemother*. 2013;57(11):5233-8.
71. Keck T, Leiacker R, Riechelmann H, Rettinger G. Temperature Profile in the Nasal Cavity. *The Laryngoscope*. 2000;110(4):651-4.

Tables and Figures

Table 1. *S. aureus* strains used in this study.

Strain No.	Strain Name	ST	CC or PFGE	SCCmec	Geographic Origin	Isolation
NRS384	USA300-0114	8	USA300	IV	Mississippi, USA	Human
NRS255	HT20020371	80	80	IV	France	Human
NRS253	HT20020354	398	398	NONE	France	Human
NRS653	CA-513	5	USA800	IV	California, USA	Human
N/A	PD6	9	N/A	NONE	Illinois, USA	Swine
N/A	PD10	398	N/A	YES	Iowa, USA	Swine
N/A	PD17	398	N/A	NONE	Texas, USA	Swine
N/A	PD18	9	N/A	NONE	North Carolina, USA	Swine
N/A	PD19	5	N/A	NONE	North Carolina, USA	Swine
N/A	PD32	9	N/A	NONE	Alabama, USA	Swine

Table 2. Summary statistics for three *S. aureus* jumbo myophage genomes.

Phage Name	<i>S. aureus</i> host	Isolation site	Genome Length (bp)	Protein-coding genes	Disrupted genes	Accession no.
MarsHill	USA300-0114	Texas swine barn	266,637	262	9	MW248466
Madawaska	USA300-0114	Texas swine barn	265,446	264	10	MW349129
Machias	PD32	Minnesota swine barn	274,478	263	13	MW349128

Table 3. Predicted proteins shared between the MarsHill-like phages and other closely related phages, as determined by BLASTp ($E < 10^{-5}$). Top row lists the related phage name, NCBI accession and phage host.

	AR9 NC_031039 <i>B. subtilis</i>	PBS1 NC_043027 <i>B. subtilis</i>	BpuM-BpSp KT895374 <i>B. pumilus</i>	phiR-37 NC_016163 <i>Y. enterocolitica</i>	MS32 MK308677 <i>V. mediterranei</i>	Twort NC_007021 <i>S. aureus</i>
Mars Hill	77	76	75	60	20	13
Madawaska	76	75	75	59	19	12
Machias	81	81	80	62	19	21

Table 4. Predicted intron content in the genomes of three *S. aureus* jumbo phages, in comparison with introns identified in the related Bacillus jumbo phage AR9. Only genes disrupted in at least one of the four phages are shown. Intron content in the *S. aureus* phages was predicted based on protein sequence alignment with intact exons or experimentally determined mature mRNA coding sequences found in AR9 or other related organisms. The DNA gyrase A protein was found to be encoded by a single exon in the MarshHill-like phages but was disrupted by an intein.

Gene product	Phage genome							
	MarshHill		Madawaska		Machias		AR9	
	Gene	Predicted exons	Gene	Predicted exons	Gene	Predicted exons	Gene	Exons
nvRNAP β' N	002	1	002	1	002	1	g270	2
scaffold/prohead protease	063	2	064	3	064	3	g114	2
TerL	070, 071	≥ 2	069, 071	≥ 2	070, 073	≥ 2	g121	5
NrdE	077	1	077	1	077	3	g223	1
virion protein/ hypothetical protein	085	1	085	1	090	2	g205	3
DNA pol catalytic subunit	108	2	108	2	110	3	g152	1
vRNAP β' N	113	3	113	2	117	3	g145	3
DNA pol exonuclease subunit	138	1	137	2	137	3	g132	1
vRNAP β C2	161	2	160	2	170	4	g189	6
vRNAP β' C	162	1	161	1	171	2	g264	1
DNA gyrase B	216	2	218	2	225	2	g024	1
DNA gyrase A	228	1 (intein)	231	1 (intein)	238	1 (intein)	g044	3
vRNAP β C1	230	2	233	3	239	3	g078	3
nvRNAP β C	248	4	254	3	266	6	g089	1

Figure 1. Transmission electron micrograph of the jumbo *S. aureus* myophage MarsHill.

Figure 2. Restriction digests of MarshHill genomic DNA. DNA was not digestible with DraI but was cut with the enzymes MspI and TaqI. Marker is a 1 kb ladder (NEB).

Figure 3. Comparison of three *S. aureus* jumbo phage genomes by progressiveMauve. The annotated genome map for phage MarshHill is added to scale at the top of the figure. The genomes are generally syntenic, with the exception of an inverted region in phage Machias. DNA sequence discontinuities are primarily located at homing endonuclease genes (light blue) or hypothetical genes (grey).

Figure 4. Annotated genome map of the jumbo *S. aureus* phage MarsHill. Gene transcribed from the top DNA strand are drawn above the black line, and genes transcribed from the minus strand are drawn below. Genes are color coded based on their predicted functions as shown in the legend, and selected genes are annotated with their predicted functions. Open reading frames that are predicted to be part of intron-disrupted genes are joined by heavy black lines.

Figure 5. Sequence logo of the predicted MarshHill early promoter sequence recognized by the phage viral RNA polymerase.

Supplementary Table S1. Related proteins in three *S. aureus* jumbo phages as determined by blastclust with a cutoff of 50% identity over 50% of the protein length. Proteins on the same line were determined to be in the same cluster.

Provided in attached Excel file.