Pseudo-pac site sequences used by phage P22 in generalized transduction of *Salmonella*

Authors: Jessie L. Maier[1], Craig Gin[2], Ben Callahan[2], Emma K. Sheriff[3], Breck A. Duerkop[3], Manuel Kleiner[1]

Affiliations:

[1]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC,

[2]Department of Population Health and Pathobiology, North Carolina State University, Raleigh, NC

[3]Department of Immunology and Microbiology, University of Colorado - Anschutz Medical Campus, School of Medicine, Aurora, CO

Correspondence:

Jessie Maier - jlmaier@ncsu.edu

Manuel Kleiner - manuel_kleiner@ncsu.edu

Abstract

Salmonella enterica Serovar Typhimurium (*Salmonella*) and its bacteriophage P22 are a model system for the study of horizontal gene transfer by generalized transduction. Typically, the P22 DNA packaging machinery initiates packaging when a short sequence of DNA, known as the pac site, is recognized on the P22 genome. However, sequences similar to the pac site in the host genome, called pseudo-pac sites, lead to erroneous packaging and subsequent generalized transduction of *Salmonella* DNA. While the general genomic locations of the *Salmonella* pseudo-pac sites are known, the sequences

themselves have not been determined. We used visualization of P22 sequencing reads mapped to host *Salmonella* genomes to define regions of generalized transduction initiation and the likely locations of pseudo-pac sites. We searched each genome region for the sequence with the highest similarity to the P22 pac site and aligned the resulting sequences. We built a regular expression (sequence match pattern) from the alignment and used it to search the genomes of two P22-susceptible *Salmonella* strains- LT2 and 14028S- for sequence matches. The final regular expression successfully identified pseudo-pac sites in both LT2 and 14028S that correspond with generalized transduction initiation sites in mapped read coverages. The pseudo-pac site sequences identified in this study can be used to predict locations of generalized transduction in other P22-susceptible hosts or to initiate generalized transduction at specific locations in P22-susceptible hosts with genetic engineering. Furthermore, the bioinformatics approach used to identify the *Salmonella* pseudo-pac sites in this study could be applied to other phage-host systems.

Importance

Bacteriophage P22 has been a genetic tool and a key model for the study of generalized transduction in *Salmonella* since the 1950s, yet certain components of the generalized transduction molecular mechanism remain unknown. Specifically, the locations and sequences of pseudo-pac sites, hypothesized to facilitate packaging of *Salmonella* DNA by P22, to date have not been determined. In this study, we identified the specific locations and sequences of the pseudo-pac sites frequently recognized by P22 in *Salmonella* genomes. The identification of highly efficient pseudo-pac sites in *Salmonella* provides fundamental insights into the sequence specificity necessary for P22 pac site recognition and opens the door to more targeted use of generalized transduction with P22.

Observation

Transduction, the transfer of DNA between bacterial cells by bacteriophages, can lead to horizontal gene transfer of entire operons of genetic material and can cause dramatic changes in bacterial phenotypes (1–4). Generalized transduction, one of several potential modes of transduction, was first discovered in 1952 in the bacteriophage P22 and *Salmonella enterica* Serovar Typhimurium LT2 (LT2), thus making these a model system for generalized transduction (5, 6). Generalized transduction was initially thought to be a random transfer of host DNA, but when the frequencies of transduced LT2 gene markers were quantified, it became clear that transduction frequencies differed widely across the genome (7–9). Similar transduction locations and frequencies were seen in two recent studies that used mapped P22 DNA sequencing reads to visualize transduction patterns in LT2. These studies demonstrate that the P22-facilitated generalized transduction of LT2 is non-random and consistent between methods and experiments (10, 11). The observed pattern consists of sharp increases in read coverage followed by sloping decreases of coverage across several regions of the LT2 genome (Fig. 1A).

P22 uses a headful packaging mechanism in which a short sequence of DNA, known as the pac site, is recognized prior to packaging initiation. After initiation, the P22 DNA packaging machinery packages several capsids in series using the same concatemer of DNA on which the pac site is located (12–14). Generalized transduction by P22 occurs when its small terminase, responsible for pac site recognition (15, 16), recognizes a sequence in the host genome that is similar to the phage's pac site (i.e pseudo-pac site), leading to initiation of packaging on the bacterial chromosome (9, 16–18). The locations of pseudo-pac sites along a bacterial chromosome lead to the non-random generalized transduction patterns observed between P22 and LT2. Despite P22's pac site sequence having been

previously described by Wu *et al.* (2002) (19), the exact pseudo-pac site sequences and locations remained unknown.

Identifying pseudo-pac site candidate sequences

We used previously published Illumina sequencing reads from ultra-purified P22 propagated on LT2 (10) and mapped them back to the LT2 genome to identify the locations where pseudo-pac site facilitated packaging of the LT2 genome occurred. The regions where packaging is initiated are characterized by a sudden, sharp increase in read coverage (Fig. 1A). We visually identified eight sites that matched this profile and extracted 120 base pair (bp) regions of the LT2 genome surrounding these sites (Fig. 1A). We chose 120 bp because the P22 packaging machinery makes its packaging initiation cuts in a 120 bp region surrounding its pac site (15). We searched each of the eight 120 bp regions for the sequence that best matched the 12 bp P22 consensus pac site sequence- 5' AAGATTTATCTG 3'- identified in Casjens *et al.* (1987) (20) and further characterized by Wu *et al*. (2002) (19) using P22 mutants. For generalized transduction events whose read coverage patterns sloped left to right across the LT2 genome (sites 3, 4, 5, 7 and 8), we searched the forward strand of the genome and for events that sloped right to left (sites 1, 2 and 6), we searched the reverse strand (Suppl. Text). The eight sequences that best matched the P22 pac site in each of the 120 bp regions (Fig. 1A) are henceforth referred to as pseudo-pac site candidates. We performed a multiple sequence alignment (MSA) with ClustalW (21) of both the P22 pac-site and the pseudo-pac site candidates including the respective neighboring genome regions. We discovered that sequence conservation between the candidates extended beyond the 12 bp pseudo-pac site consensus region (Fig. 1B) and adjusted the consensus region to include all strongly conserved regions of the MSA accordingly.

Confirming accuracy of the pseudo-pac site consensus sequence

To determine if the consensus sequence obtained using the MSA was accurate, we used it to 'scan' the genome of *Salmonella enterica* Serovar Typhimurium 14028S (14028S). Strain 14028S is susceptible to P22 infection and based on read mapping, we determined that 14028S shares the same generalized transduction sites and associated candidate pseudo-pac site sequences as LT2. However, when infected with P22, we observed one additional generalized transduction site in the 14028S read coverages. The additional pattern was located on the reverse strand around 1.3 Mbp and was seemingly not present in LT2 when infected with P22. (Fig. 2A). We hypothesized that if our pseudo-pac site consensus sequence was correct, it could be used to identify the pseudo-pac site present at the additional generalized transduction site in 14028S. We built a regular expression (sequence match pattern) - 5' AAG[AG][TC][AT][AT][ATC][TC][TC]T[GT][ACG][ACG][ACG]TC 3'- that represents the bases observed for each position of the MSA. This regular expression specifies that any matching sequence will have an AAG in the first three positions, the fourth position must be either A or G, the fifth position must be either T or C and so on. We used the regular expression to search both the forward and reverse strands of LT2 and 14028S (Suppl. Text). Five new sequences were identified in both LT2 and 14028S using the regular expression, two of which, located on the forward strand around 2.5 and 2.8 Mbp, were associated with right to left sloping read coverages immediately following the match location. There was a sharp jump in read coverage for the forward strand match at 2.8 Mbp that was more obvious in 14028S than LT2 (Fig. 2B). These sequence matches likely represent two additional pseudo-pac sites in the LT2 and 14028S genomes (sites 9 and 10, respectively, in Fig. 2C) which were not identified in our initial visual screen as they are not as prominent in LT2. Despite the newly identified sites, the additional generalized transduction site present in 14028S was not identified by the regular expression which indicated an error in the consensus sequence. After testing various changes to the regular expression, we

ultimately found that changing the conserved G in position three to G **or** C enabled the identification of the additional pseudo-pac site in 14028S (Fig. 2C) with minimal false positives. While we only tested a relatively small number of all the possible changes to the regular expression, we found that changes to other conserved positions, like the As in the first two positions or the C in the last position, caused large increases in false positive matches (Fig. S1). Eight out of the eighteen and nineteen matches in LT2 and 14028S, respectively, to the final regular expression do not appear to be associated with large jumps and/or sloping read coverages. This could be due to the corresponding generalized transduction patterns being covered by more prominent patterns at these positions or by secondary DNA structures preventing the P22 packaging machinery from binding.

Conclusions

Based on the evidence presented, we are confident that the ten pseudo-pac site sequences identified in LT2 (Fig. 2D) are the exact sequences that P22 routinely recognizes for generalized transduction. We are also confident that our final regular expression pseudo-pac site consensus sequence- 5’ AA[GC][AG][TC][AT][AT][ATC][TC][TC]T[GT][ACG][ACG] [ACG]TC 3’- can identify highly efficient pseudo-pac sites in other P22-susceptible *Salmonella* strains, like the eleven sites identified in 14028S (Fig. 2D). Our results could be further validated *in vitro* by genetically engineering the pseudo-pac site sequences identified in this study into a P22-susceptible host bacteria, infecting the host with P22, and sequencing the purified P22 to determine if generalized transduction was induced at the location of the inserted pseudo-pac site sequence. We hope that the methods used to identify the P22 pseudo-pac sites in LT2 and 14028S can be adapted by others to identify pseudo-pac sites used for generalized transduction in diverse phage-host systems.

Acknowledgements:

This work was supported by funding from the NC State University Data Science Academy and by the National Institute of General Medical Sciences and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R35GM138362 (MK) and R01AI141479 (BAD). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data availability:

P22 sequencing reads used for LT2 read mapping were generated in Kleiner *et al.* (2020) and are available at ENA (Study: PRJEB6941, Sample: SAMEA2690949). P22 sequencing reads used for 14028S read mapping are available at ENA (Study: PRJEB72417, Sample: SAMEA115180785). The reference genomes used for LT2 and 14028S read mapping are from NCBI RefSeq NC_003197.2 and NC_016856.1, respectively.

References:

1. Gozzi K, Tran NT, Modell JW, Le TBK, Laub MT. 2022. Prophage-like gene transfer agents promote Caulobacter crescentus survival and DNA repair during stationary phase. PLOS Biol 20:e3001790.

2. Penadés JR, Chen J, Quiles-Puchalt N, Carpena N, Novick RP. 2015. Bacteriophage-mediated spread of bacterial virulence genes. Curr Opin Microbiol 23:171–178.

3. Haaber J, Leisner JJ, Cohn MT, Catalan-Moreno A, Nielsen JB, Westh H, Penadés JR, Ingmer H. 2016. Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. 1. Nat Commun 7:13333.

4. Fillol-Salom A, Martínez-Rubio R, Abdulrahman RF, Chen J, Davies R, Penadés JR. 2018. Phage-inducible chromosomal islands are ubiquitous within the bacterial universe. 9. ISME J 12:2114–2128.

5. Zinder ND, Lederberg J. 1952. Genetic Exchange in Salmonella. J Bacteriol 64:679–699.

6. Zinder ND. 1955. Bacterial transduction. J Cell Comp Physiol 45:23–49.

7. Ozeki H. 1959. Chromosome Fragments Participating in Transduction in Salmonella Typhimurium. Genetics 44:457–470.

8. Schmieger H, Backhaus H. Altered Cotransduction Frequencies Exhibited by I-IT-Mutants of Salmonella-PhageP22.

9. Schmieger H. 1982. Packaging signals for phage P22 on the chromosome of Salmonella typhimurium. Mol Gen Genet MGG 187:516–518.

10. Kleiner M, Bushnell B, Sanderson KE, Hooper LV, Duerkop BA. 2020. Transductomics: sequencing-based detection and analysis of transduced DNA in pure cultures and microbial communities. Microbiome 8:158.

11. Fillol-Salom A, Bacigalupe R, Humphrey S, Chiang YN, Chen J, Penadés JR. 2021. Lateral transduction is inherent to the life cycle of the archetypical Salmonella phage P22. 1. Nat Commun 12:6510.

12. Jackson EN, Jackson DA, Deans RJ. 1978. EcoRI analysis of bacteriophage P22 DNA packaging. J Mol Biol 118:365–388.

13. Tye B-K, Botstein D. 1974. P22 morphogenesis II: Mechanism of DNA encapsulation. J Supramol Struct 2:225–238.

14. Casjens S, Huang WM. 1982. Initiation of sequential packaging of bacteriophage P22 DNA. J Mol Biol 157:287–298.

15. Casjens S, Sampson L, Randall S, Eppler K, Wu H, Petri JB, Schmieger H. 1992. Molecular genetic analysis of bacteriophage P22 gene 3 product, a protein involved in the initiation of headful DNA packaging. J Mol Biol 227:1086–1099.

16. Raj AS, Raj AY, Schmieger H. 1974. Phage genes involved in the formation of generalized transducing particles in Salmonella-phage P22. Mol Gen Genet MGG 135:175–184.

17. Thierauf A, Perez G, Maloy and S. 2009. Generalized Transduction, p. 267–286. *In* Clokie, MRJ, Kropinski, AM (eds.), Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions. Humana Press, Totowa, NJ.

18. Chelala CA, Margolin P. 1976. Evidence that HT mutant strains of bacteriophage P22 retain an altered form of substrate specificity in the formation of transducing particles in Salmonella typhimurium. Genet Res 27:315–322.

19. Wu H, Sampson L, Parr R, Casjens S. 2002. The DNA site utilized by bacteriophage P22 for initiation of DNA packaging. Mol Microbiol 45:1631–1646.

20. Casjens S, Huang WM, Hayden M, Parr R. 1987. Initiation of bacteriophage P22 DNA packaging series: Analysis of a mutant that alters the DNA target specificity of the packaging apparatus. J Mol Biol 194:411–422.

21. Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

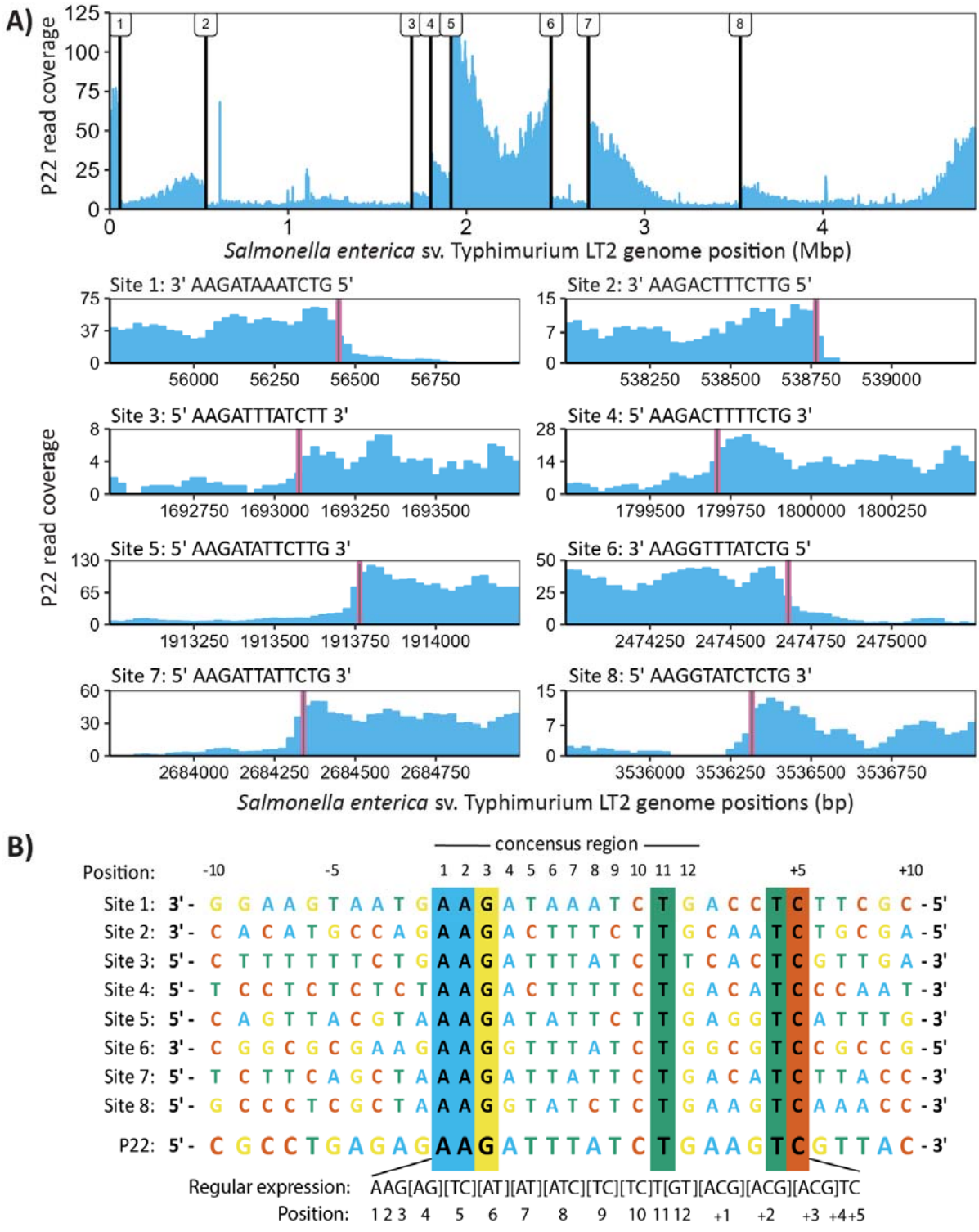
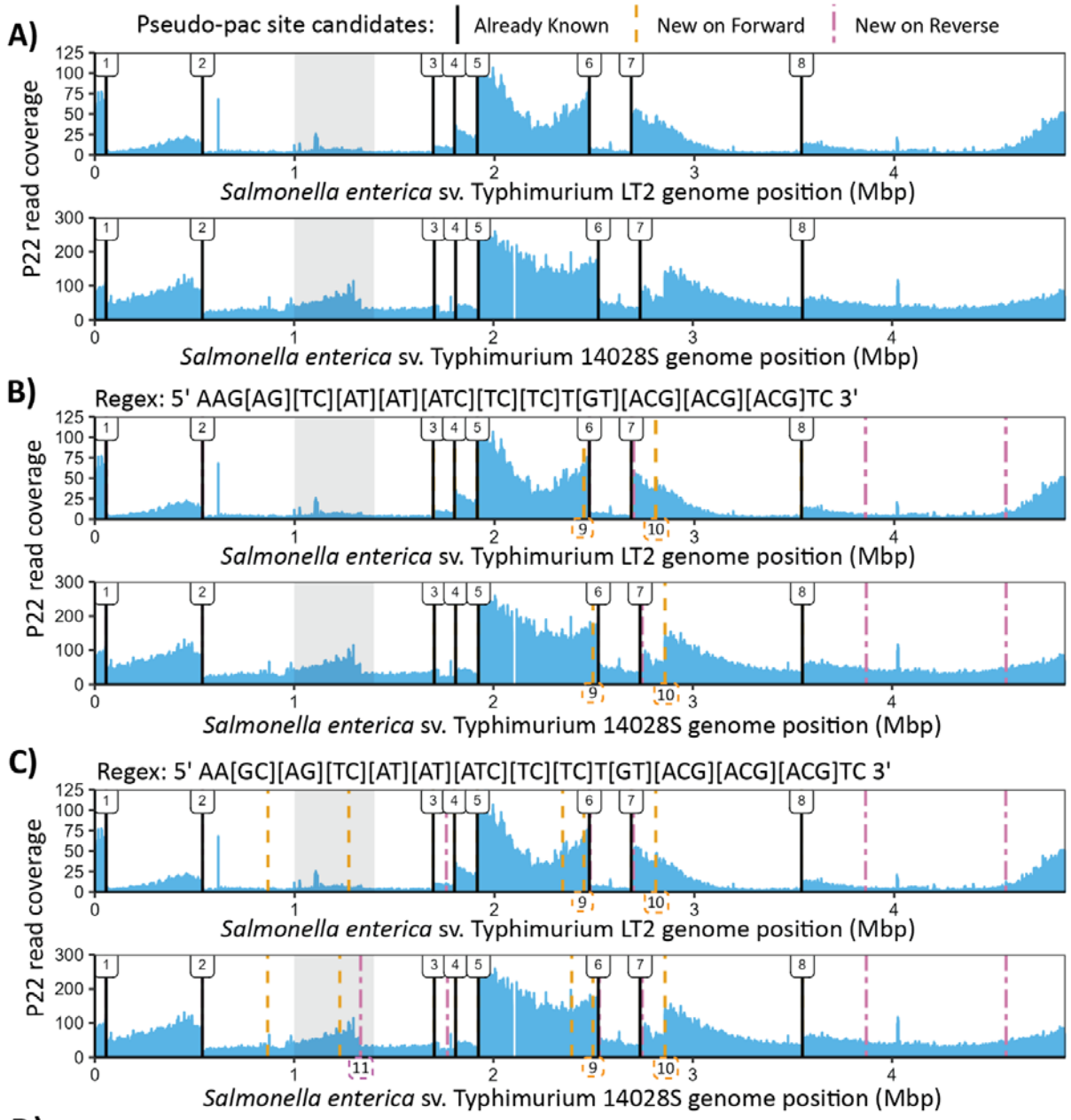


Figure 1: *Salmonella* pseudo-pac site candidate sequences identified using P22 reads mapped to the *Salmonella* genome

A) Coverage plot of the *Salmonella enterica* Serovar Typhimurium LT2 (LT2) genome with sequencing reads from purified P22. Black vertical lines indicate the eight initiation sites for generalized transduction and the locations of the associated pseudo-pac sites along the LT2 genome. The exact locations of the pseudo-pac site sequences identified in this study are highlighted in pink on subsets of the LT2 read coverage plot associated with each site. The sequence of the pseudo-pac site candidate associated with each site is displayed above its respective read coverage plot. B) A multiple sequence alignment (MSA) generated with ClustalW of the eight pseudo-pac site candidate sequences, the P22 pac site and the respective neighboring genome sequences. The location of the 12 bp pac site consensus region, identified in Casjens *et al*. (1987) and further characterized by Wu *et al.* (2002), is defined above the MSA. The regular expression pattern built from the pseudo-pac site candidate sequences is displayed below the MSA.



D)

Site	Pseudo-pac site seuqence	LT2 genome position (bp)	14028S genome position (bp)	Found in Strain:
1	3' AAGATAAATCTGACCTC 5'	56,438 - 56,454	56,438 - 56,454	LT2, 14028S
2	3' AAGACTTTCTTGCAATC 5'	538,755 - 538,771	539 446 - 539,462	LT2, 14028S
3	5' AAGATTTATCTTCACTC 3'	1,693,070 - 1,693,086	1,703,053 - 1,703,069	LT2, 14028S
4	5' AAGACTTTTCTGACATC 3'	1,799,703 - 1,799,719	1,810,398 - 1,810,414	LT2, 14028S
5	5' AAGATATTCTTGAGGTC 3'	1,913,758 - 1,913,774	1,924,453 - 1,924,469	LT2, 14028S
6	3' AAGGTTTATCTGGCGTC 5'	2,474,669 - 2,474,685	2,526,222 - 2,526,238	LT2, 14028S
7	5' AAGATTATTCTGACATC 3'	2,684,334 - 2,684,350	2,735,886 - 2,735,902	LT2, 14028S
8	5' AAGGTATCTCTGAAGTC 3'	3,536,311 - 3,536,327	3,550,203 - 3,550,219	LT2, 14028S
9	5' AAGGTATTTCTTCCATC 3'	2,447,807 - 2,447,823	2,499,360 - 2,499,376	LT2, 14028S
10	5' AAGGTTTCTTTGGCGTC 3'	2,807,035 - 2,807,051	2,860,949 - 2,860,965	LT2, 14028S
11	3' AACGCATTTCTGACGTC 5'	NA	1,333,769 - 1,333,785	14028S

Figure 2: *Salmonella* pseudo-pac site sequences identified using regular expression searches

A-C) Coverage plots of the *Salmonella enterica* Serovar Typhimurium LT2 (LT2) and *Salmonella enterica* Serovar Typhimurium 14028S (14028S) genomes with sequencing reads from purified P22. The additional generalized transduction site present in 14028S but not LT2 is shaded in grey. The regular expression (Regex) patterns used to search the *Salmonella* genomes for additional pseudo-pac sites are displayed above their associated plots. Black vertical lines indicate the locations of the pseudo-pac site sequences that were previously identified in this study. Orange and pink dashed lines indicate the locations of regex matches on the forward and reverse Salmonella genome strands, respectively. The additional pseudo-pac sites identified with the regex searches, sites 9-11, are indicated in dashed boxes below their respective pattern match locations. D) A summary table with information about each pseudo-pac site identified in this study numbered in order of detection. Sites 1-8 were initially visually identified using P22 read coverages mapped to the *Salmonella* genome. Sites 9-11 were identified using the regex searches and were confirmed visually with read coverages.