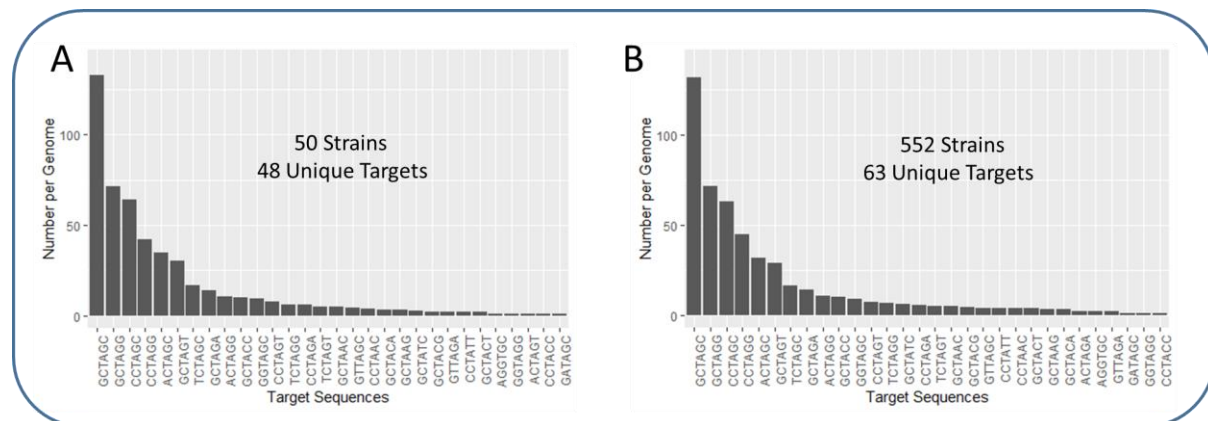
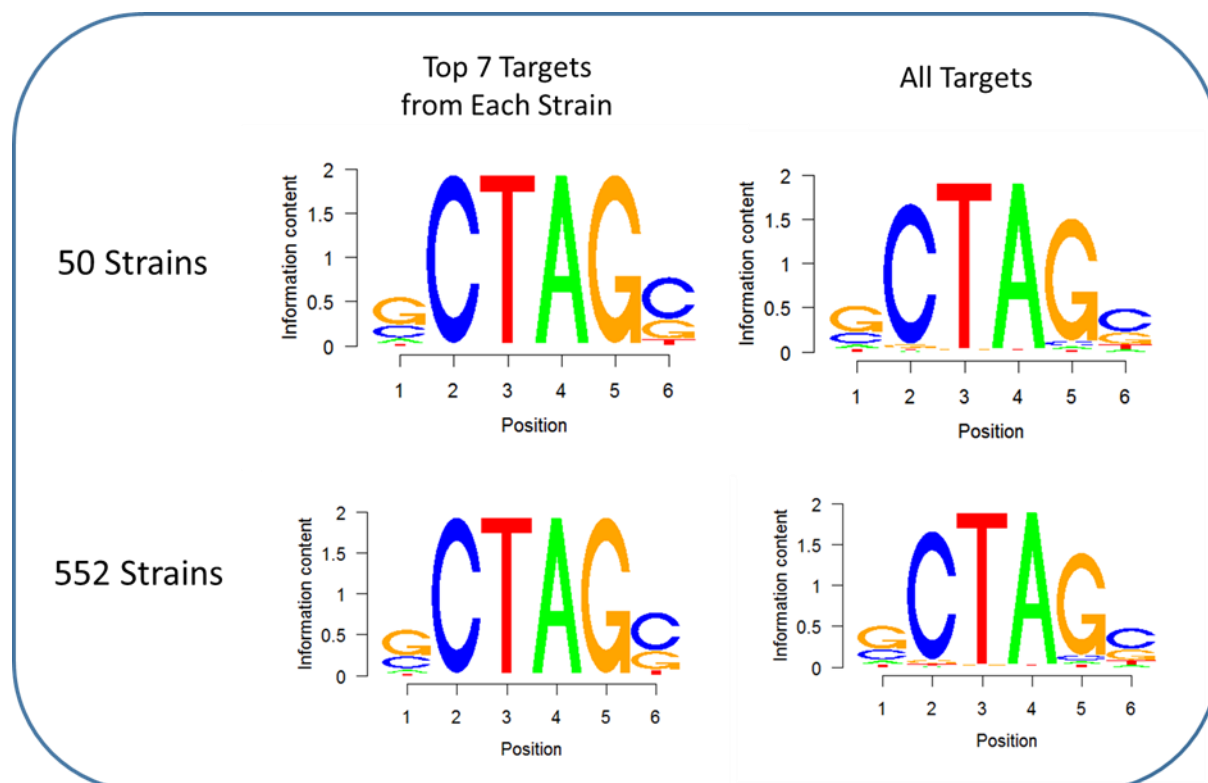


Supplementary File 1 – Figures and Tables



Supplementary Figure 1: Total unique target sequences for IS481 in 50 or 552 strains of *B. pertussis* and their average frequency within the genome. A) All target sequences with an average over 1 calculated from the closed genomes of 50 strains. Total unique target sequences found was 48. B) All target sequences with an average over 1 calculated from the closed genomes of 552 strains. Total unique target sequences found was 63.



Supplementary Figure 2: Consensus motifs of IS481 target sequences in *B. pertussis* created using 50 and 552 closed genomes. The top 7 target sequences for the IS481 sequence were merged from each strain and used to generate the consensus motif for 50 and 552 closed genomes as per analyses by Weigand et al in 2017. The all targets column shows consensus motifs generated when using all unique target sequences from the different sample sizes.

Supplementary Table 1: The number of tags from the simulated dataset suggested to be unique for each possible mismatch allowance (0 – 5 bp).

Maximum Mismatch Allowance to Known Tags (bp)	Tags Identified with First Part of the IS (Upstream Tags)	Tags Identified with Last Part of the IS (Downstream Tags)
0	2494	2497
1	2491	2494
2	2489	2493
3	2482	2488
4	2462	2457
5	2309	2286

Supplementary Table 2: The number of reads that were correctly and incorrectly identified from the simulated data as being unique. Giving total number of positives of the 5,000 reads in the controlled dataset.

Maximum Mismatch Allowance to Known Tags (bp)	Upstream Tags			Downstream Tags			Total Positive (/5,000)
	Fake Reads with First Part of <i>IS481</i> Correctly Identified (/2500)	Fake Reads Last Part of <i>IS481</i> Incorrectly Identified	Normal Reads Incorrectly Identified	Fake Reads Last Part of <i>IS481</i> Correctly Identified (/2500)	Fake Reads First Part of <i>IS481</i> Incorrectly Identified	Normal Reads Incorrectly Identified	
0	2500	3	0	2500	3	0	5006
1	2498	3	0	2498	2	0	5001
2	2496	3	0	2497	2	0	4998
3	2492	3	0	2492	2	0	4989
4	2471	3	0	2461	1	0	4936
5	2320	1	0	2291	1	0	4613

Supplementary Table 3: Breakdown of new insertion sequences and their locations found through investigation of raw reads compared to closed genomes per strain.

	UK36	UK38	UK76	Total
Unique Tags Identified by Method (Compared to full library of closed genomes)	29	13	31	73
Total New Insertions	13	8	9	30
Novel	12	7	9	28
New to Strain but Previously Known	1	1	0	2
Number of Insertions in Genes	4	5	5	14
Number of Insertions into Existing IS481	7	3	3	13
Number of Insertions in Pseudogenes	2	0	1	3