Software Analysis of Genomic Sequence Tags for the *Yersinia pestis* Pilot Project

Sean R. McCorkle

April 2, 2002

GST Pilot Project

This document describes analysis of Genomic Sequence Tags (GSTs) for the *Yersinia pestis* (Yp) Pilot Project at Brookhaven National Laboratory. The purpose of the project was to examine the GST protocol's effectiveness in sampling the genome of an organism whose sequence is already known.

The sample, Yp strain EX766, was fragmented with the *Bam*HI restriction enzyme (recognition sequence GGATCC), tags were anchored with *Nla*III (CATG) and tags were cut with *Mme*I, which cuts $\sim$ 19 bases downstream from its recognition sequence TCCRAC. More details of the experimental aspects of DNA handling and GST extraction are described in Dunn, *et. al.*[1], and protocol details can be found in the Methods white paper, included in the `Preprints` section of this distribution.

Analysis of *Yersinia pestis* GSTs consisted of these steps:

1. construction of an indexed GST database from the published Yp genome
2. extraction of sample tags from concatemer sequences
3. sampled tags compared to database via exact matches
4. remaining tags compared to database via inexact matches

Locating GSTs

A GST is defined as the 3'-most *Nla*III site to a *Bam*HI end (in either direction). Every *Bam*HI fragment can give rise to 0 or 2 bona fide GSTs[2] Figure 1. shows GSTs relative to enzyme sites and explains some terminology, included 3' and 5' offsets and ranking in a fragment. GSTs have 3'rank = 1. We call 3'rank > 1 *Nla*III sites *internal*.

[1] submitted to Genome Research; a manuscript copy is also on this CD

[2] A *Bam*HI fragment with only one *Nla*III site will give rise to two tags, one in each direction

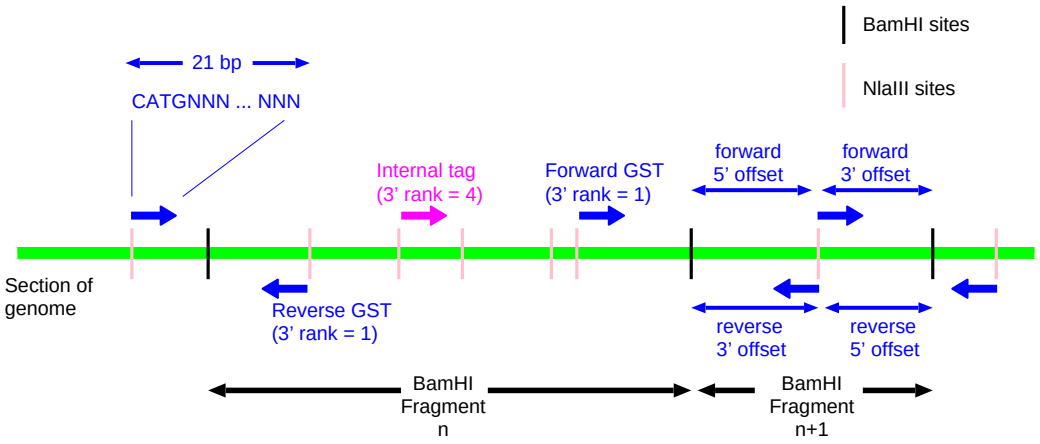


Figure 1. Examples of GSTs in a sequence; terminology

GSTs in the Yp Genome

The genome of *Yersinia pestis*, strain CO92, was sequenced by Parkhill, *et. al.*[3], who found it consisted of a circular 4.65 Mb chromosome, and three plasmids, sizes 96 kb, 70kb and 9.6kb. Because the EV766 strain is known to be lacking the 70kb plasmid, it was omitted from this analysis. The remaining three sequences were obtained from the National Center for Biotechnology Information (NCBI) at

`ftp://ncbi.nlm.nih.gov/genomes/Bacteria/Yersinia_pestis/`

and are referred to as sequence numbers 1, 2 and 3:

Sequence Id	RefSeq		Size	Component	*Bam*HI sites	*Nla*III sites
1	NC_003143	AL590842	4.65 Mb	chromosome	697	16180
2	NC_003134	AL117211.1	96 kb	plasmid pPMT1	1	360
3	NC_003132	AL109969.1	9.6 kb	plasmid pCCP1	1	32

36 *Bam*HI fragments in the chromosome contain no *Nla*III sites.

All sequences were taken to be circular. To handle the endpoint breaks of each sequence, the leading substring from the 5' end to the first (5'most) *Bam*HI site was removed and appended to the very 3' end of the sequence, before any further analysis. *Bam*HI fragment numbers were then assigned numerically, starting with 1, in each sequence.

As a precautionary check of the GST protocols, as well as to alert us to possible strain differences (translocations) between EX766 and C092, all *Nla*III site + 17 tags were obtained from the genome; 31776 internal (3'rank $>$ 1) tags were recorded in addition to the the 1326 3'rank $=$ 1 tags. For all tags, the sequence and fragment numbers, directions, 5' and 3' offsets and ranks from both the

[3]Nature 2001 Oct 4;413(6855):523-7

fragment ends, were recorded in a database structure, indexed by a trie constructed from the tag sequences, without the leading CATG. The index allows rapid exact match searches.

1274 of the 3'rank 1 tags are unique within CATG+17 nt. The non-unique tags occur largely in repetitive elements, and are discussed in Dunn, *et. al.*.

Extraction of GSTs from Concatemers

Here is an example section of a concatemer sequence, highlighting CATG separators, and showing the forward (→), reverse (←), and ambiguous (⇄) polarity tags, whose directions are indicated by the orientation of the tail TT punctuation.

CATGAACTTATGATCGCGAAACCATGCCCTGGGCCTTCGATCATTTCATGTCCGCCAGCCCTAATCCTTCATG

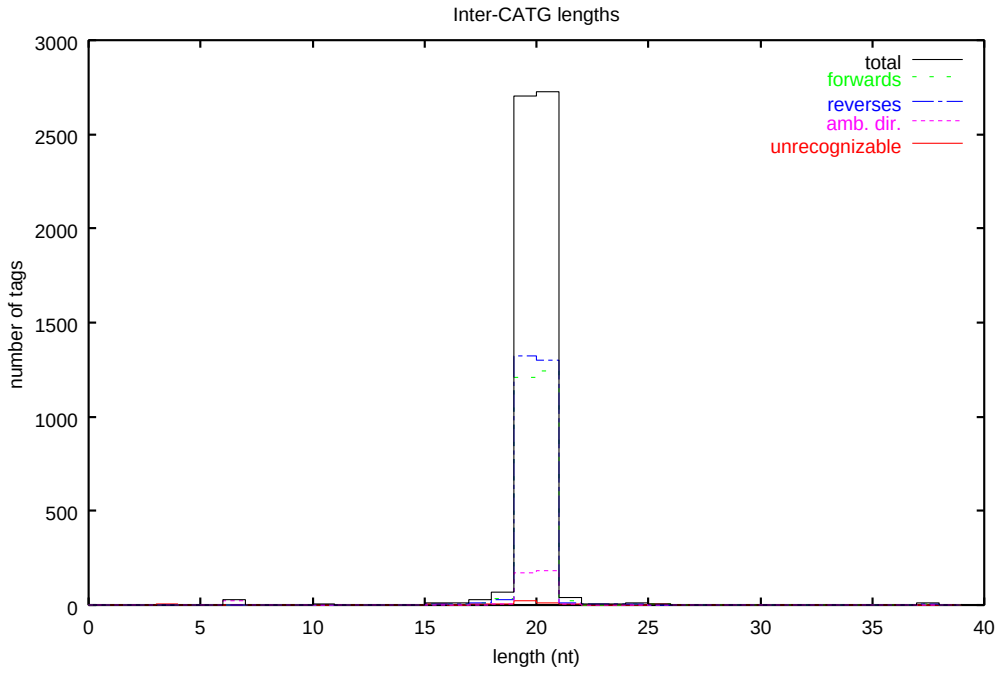
AAGTAAATGGCCGTTCACGTCATGAAATCAGCCCACAAGAGTTCATGTACGCGAACAAAATCAATTCATGTCG

TTAGTGGTTAATTACTTCATGAAGAACGCAGCAACCAGATCATG.....

The collected concatemer sequences were very clean; most of the inter-CATG sequences met our two criteria:

1. **Length** must be 19 or 20 nucleotides (nt) (note: this includes the TT punctuation). 95% of all candidates passed.

2. **Polarity punctuation** candidate sequence must either end with TT or begin with AA, or both. Only 33 sequences, 0.6% of those passing the length cut, were rejected because of missing punctuation.



The accepted 5432 GSTs were then divided into two groups which were analyzed separately:

1. **Unambiguous polarity**

   5077 tags were either clearly oriented forward (ending in TT, but no leading AA) or clearly reversed (beginning with AA, but no trailing TT)

   The reversed tags were flipped to forward, all the trailing TT's were removed, and the remaining sequences were then truncated to 17 nt to simplify handling, and were then compared with each other (exact matches), and found to represent 1170 unique GSTs. The binned tags can be found in file `unamb_tags.tbl` in this distribution. These were then searched for in the Yp genome trie. Tags for which no exact match was found in the Yp genome were saved for 2nd examination.

2. **Ambiguous polarity**

   355 tags (6.5%) began with AA *and* ended with TT. (They are listed in `amb_tags.tbl`). Each tag was first treated as though it were forward - the trailing TT was removed, the sequence truncated to 17nt, and then was searched for in the genome trie. If no match was found, the sequence was reversed and the process was repeated.

Sample Set to Genome Comparisons

We found that nearly all of the 5432 tags in the sample *exactly* matched GSTs from the Yp genome Many of the remaining tags were found to match by making Smith-Waterman comparisons with a threshold of less than 3 errors out of 17.

	unambiguous	ambiguous	total
Exact matches	4929 (90.7%)	339 (6.2%)	5268 (97.0%)
Smith-Waterman, 1 error	110 (2.0%)	6 (0.1%)	116 (2.1%)
Smith-Waterman, 2 error	13 (0.2%)	7 (0.1%)	20 (0.4%)
No match found	25 (0.5%)	3 (0.06%)	28 (0.5%)

Percentages reflect the fraction of the sample set of 5432 tags.

5404 (99.5%) of the tags were accounted for in the Yp genome. Of those, *most* (5346, or 99%) matched 3'-rank = 1 tags, but 58 (1%) matched internal *Nla*III sites (3'-rank > 1), and are listed in the file `internals.tbl` in this distribution. We are currently investigating whether any of these reflect translocational differences between the genomes of the sampled and published strains, or are indicative of biochemical systematics, such as a small background due to incomplete restriction digests.

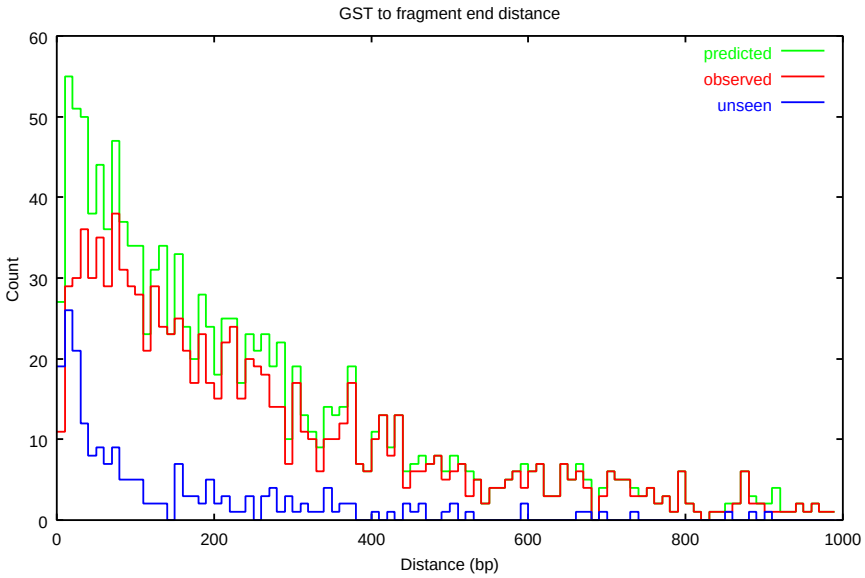
Exact matches of tags are detailed in `unamb_exact_matches.tbl` and `amb_exact_matches.tbl`. Smith-Waterman matches and alignments of unambiguous polarity tags are in `unamb_inexact_matches.txt`.

Possession of the complete Yp genome allows us to ask which 3'-rank 1 tags have been detected and which have not. 209 of the 1326 GSTs were not seen. A listing of number of tags sampled, viewed with respect to fragment number can be found in file `hits_by_fragment.tbl` in this distribution. (All 3'-rank 1 tags in the plasmids were detected)

A striking feature is the contiguous block of missing tags in *Bam*HI fragments 314–327[4] (genome positions 2172626–2254414), while the rest of the missing tags appear more randomly distributed. This most likely indicates a deletion in the sampled Yp strain with respect to the C092 genome.

[4] *Bam*HI fragment 320 contains no *Nla*III sites and thus does not appear in the list

`hits_by_fragment.tbl` also suggests that the other tend to be close to the *Bam*HI fragment ends. A histogram comparing tag-to-*Bam*HI-end distances for seen and missed tags shows more clearly that many tags within 30 nt of the fragment end are missed.



An interpretation and possible solution to this bias are offered in Dunn, *et. al.*.