Genomic Signature Tags (GSTs): A System for Profiling Genomic DNA

John J. Dunn, Sean R. McCorkle, Laura A. Praissman, Geoffrey Hind, Daniel van der Lelie, Wadie F. Bahou, Dmitri V. Gnatenko, and Maureen K. Krause

Genomic signature tags (GSTs) are the products of a method we have developed for identifying and quantitatively analyzing genomic DNAs. The DNA is initially fragmented with a type II restriction enzyme. An oligonucleotide adaptor containing a recognition site for MmeI, a type IIS restriction enzyme, is then used to release 21-bp tags from fixed positions in the DNA relative to the sites recognized by the fragmenting enzyme. These tags are PCR-amplified, purified, concatenated, and then cloned and sequenced. The tag sequences and abundances are used to create a high-resolution GST sequence profile of the genomic DNA. GSTs are shown to be long enough for use as oligonucleotide primers to amplify adjacent segments of the DNA, which can then be sequenced to provide additional nucleotide information or used as probes to identify specific clones in metagenomic libraries. GST analysis of the 4.7-Mb Yersinia pestis EV766 genome using BamHI as the fragmenting enzyme and Ndal as the tagging enzyme validated the precision of our approach. The GST profile predicts that this strain has several changes relative to the archetype CO92 strain, including deletion of a 57-kb region of the chromosome known to be an unstable pathogenicity island.

A variety of DNA-based fingerprinting techniques now exist to characterize and compare whole genomes of prokaryotes and eukaryotes, either as independent organisms or as members of communities (Schloter et al. 2000; Kozdrój and van Elsas 2001; Torsvik and Øvreås 2002). These fingerprinting techniques—such as amplified fragment length polymorphism (AFLP), terminal restriction fragment length polymorphism (T-RFLP), denaturing gradient gel electrophoresis (DGGE), amplified rDNA restriction analysis, (ARDRA), and restriction landmark genome scanning (RLGS)—are generally based on some combination of restriction digestion of genomic DNA, PCR amplification, and gel electrophoretic separation. The DNA fingerprints are then visualized by means of autoradiography, phosphor-imaging, fluorescence, or other labeling methods. A drawback to these techniques is how to further analyze novel bands. Usually, individual fragments are extracted from the gels and the corresponding sequences determined by direct DNA sequencing; however, this approach is labor intensive and, in most cases, requires further PCR amplification or cloning of the eluted DNAs.

In this paper, we describe a new higher-throughput, direct sequence-based approach for characterizing prokaryotic or eukaryotic genomes by use of genomic signature tags (GSTs), which like AFLP-related methods does not rely on a priori knowledge of the genome (Vos et al. 1995). It is similar to long serial analysis of gene expression (long SAGE; Velculescu et al. 2001; Saha et al. 2002) in that it produces large numbers of positionally defined 21-bp tag sequences that can be used to examine intraspecific genomic variation and, if genome information is available, provide immediate species identity.

In the original SAGE procedure (Velculescu et al. 1995, 1997; Zhang et al. 1997; Yu et al. 1999), double-stranded cDNA is synthesized from poly (A) mRNA by priming first-strand cDNA synthesis with a biotinylated oligo (dT)$_{18}$ primer. The cDNA is then cut with a restriction endonuclease having a 4-bp recognition sequence (typically Ndal, recognition sequence CATG, which theoretically results in cleavage on average every 256 bp), and the 3′-terminal cDNA fragments are captured on streptavidin-coated magnetic beads. These fragments are ligated with two DNA cassettes, each containing a recognition sequence for BamHI, a type IIS restriction endonuclease. Subsequent cleavage with BamHI releases short (13 to 14 bp) but positionally defined sequences, referred to as tags, which are eventually ligated to form “ditaq,” concatenated into arrays, and cloned into a plasmid vector for DNA sequencing. The power of the method is that many SAGE tags can be read serially from each clone during the sequencing step which vastly increases throughput (Velculescu et al. 1995).

Since the SAGE technique was first reported, several groups have modified the original procedure in order to increase tag length (Ryo et al. 1998, 2000; Spinella 1999). These longer tags are particularly useful in characterizing expression patterns in the absence of complete genome sequence data, that is, from “uncharted transcriptomes,” and in designing
primers to obtain full-cDNAs from transcripts with tags that are not currently present in RefSeq or similar expression databases. One very useful type IIS enzyme for SAGE-based analysis that has only recently become commercially available is Mmel, which cleaves 20/18 bases past its nonpalindromic (TCCRAC) recognition sequence (Boyd et al. 1986; Tucholski et al. 1995). Mmel has been used for development of long SAGE, which is an adaptation of the original SAGE approach that generates 21-bp tags from NlaIII sites (Velculescu 2001; Saha et al. 2002). The long length of these tags (CATG + N17) suggested to us that Mmel could be used to obtain unique tags directly from total microbial DNA owing to the number of Mmel tag sequences, which theoretically exceeds 17 billion nucleotide combinations (Saha et al. 2002), by far surpassing the number of potential tags in most prokaryotic and many eukaryotic genomes. Consequently, Mmel tags should, in most cases, be able to uniquely identify their DNA source. This premise was confirmed by in silico analysis of ~60 complete microbial genomes in the National Center for Biotechnology Information (NCBI) database and several fungal genomes (data not shown).

The GST procedure we developed, like RLGS (Rouillard et al. 2001; Wimmer et al. 2002), involves the initial digestion of genomic DNA with two type II restriction enzymes. After the digestion with the first enzyme, the cut ends are biotinylated to allow their solid phase affinity capture after treatment with the second enzyme. A linker containing a Mmel recognition site is ligated to the nonbiotinylated ends, and Mmel digestion is then used to liberate 21-bp GST sequences from the untethered ends of the captured fragments. The released monomeric GSTs are PCR-amplified and randomly ligated on themselves prior to cloning. The resulting sequences are identified through database matches or used to create a new database that is specific for a particular DNA sample.

Using *Yersinia pestis* as a model system, we demonstrate that the basic GST procedure can not only identify the DNA source but also pinpoint areas of a genome that might have undergone changes that add or delete restriction sites. We further show that primers corresponding to GSTs can be used to directly convert tags into their corresponding longer genomic fragments, which are particularly useful for characterizing novel genomes or annotating known ones.

**RESULTS**

**Overview of GST Methodology**

Figure 1 gives the general strategy for production of GSTs. The method depends on the ability of a type II restriction enzyme, termed the fragmenting enzyme, to cleave the starting DNA into a manageable number of fragments, all having the same complementary single-stranded extensions. The digest is then ligated with a molar excess of short biotinylated duplex complementary adaptors with only one cohesive end, to biotinylate both ends of all the fragments. The DNA is next digested with NlaIII, the anchoring enzyme, which cleaves, leaving four-base cohesive ends. Biotinylated end fragments are recovered by binding to streptavidin-coated magnetic beads and digested a second time with NlaIII to ensure that NlaIII digestion is complete. After the beads are washed, a duplex linker with NlaIII cohesive termini is ligated to the bound DNA fragments. This linker generates a recognition site (TCCGAC) for the type IIS enzyme Mmel, the tagging enzyme, only when it is joined to NlaIII cohesive ends. After washing to remove excess linkers, the beads are incubated with Mmel to release the linker and appended tags from the beads. Because the last C residue in the Mmel recognition site of the adaptor partially overlaps the NlaIII site of the bound DNA, the released fragments contain 21 bases of sequence information from the starting DNA. These products are recovered and ligated with an adaptor having a 16-fold degenerate 3’ overhang (Spinella et al. 1999) which renders it compatible with all possible two-base 3’ overhangs released by Mmel. This adaptor was designed to add two consecutive T residues and a second NlaIII site on the ends of the original Mmel-generated fragments (TTTCATG...). The ligation products are PCR-amplified using two linker-specific biotinylated primers and cleaved with NlaIII, and the two biotinylated end fragments

---

**Figure 1** Schematic for GST preparation. In this method, DNA is first fragmented with a rare cutter such as *Not*I or more frequent cutter such as *Bam*HI. Specific complementary biotinylated linkers are ligated to the free ends, and the DNA is then digested with NlaIII. All subsequent steps in the protocol are identical.
are removed by affinity capture on streptavidin-coated magnetic beads (Powell 1998), leaving the 19-bp duplex GSTs with NlaIII cohesive ends free in solution (Fig. 1). Each tag ends with two T/A base pairs donated by the degenerate linker, which help stabilize the identifier portion of the tag. They also act as a punctuation sequence to demarcate individual tags and aid in determining their polarity. The purified tag fragments are ligated together to form concatemers. Concatemers of sufficient minimal length are isolated by agarose gel electrophoresis and ligated into a pZero-based positive selection vector. The recombinant plasmids are electroporated into competent Escherichia coli cells to generate the GST library in preparation for DNA sequence analysis.

In developing the GST method, we reasoned that adaptor ligation would be more specific than enzymatically filling in the cohesive ends with biotinylated nucleotides. This might be especially important in cases in which obtaining nearly intact starting DNA is problematic. An additional benefit of adding a linker to the fragmented DNA is that it helps avert steric hindrance during the subsequent enzymatic reactions that are performed once the DNA is captured on magnetic beads.

Optimizing PCR Amplification

A critical step contributing to the robustness of the GST protocol is the amount of material produced during the first round of PCR amplification. Typically, when this reaction is analyzed by electrophoresis on a 10% polyacrylamide gel, a band with the expected mobility of the GSTs plus attached linker arms, 94 bp, is observed plus varying amounts of diffuse material with slower mobilities (Fig. 2, lane 2). The amount of this diffuse material in the reaction seemed to be proportional to the number of PCR amplification cycles; therefore, we reasoned that it most probably represents amplicon heteroduplexes, formed by preferential perfect annealing of the low-complexity linker arms but imperfect annealing of the internal tags at high product concentrations. As expected, the bulk of this material is sensitive to digestion with S1 nuclease (data not shown). To optimize amplicon recovery we introduced a linear amplification step to reduce heteroduplexes (LARHD), which uses one extra round of amplification to convert the bulk of the reaction products to double-stranded DNA (Fig. 2, lane 3). Several additional tests showed that the potential to form heteroduplexes could be avoided during additional rounds of PCR amplification of the LARHD products by doing repeated rounds of linear amplification with one GST linker-specific primer followed by one final amplification step after addition of the second linker-specific primer. Unwanted PCR primers that would be carried over from the LARHD step are eliminated by incubation with Exol, which preferentially hydrolyzes any remaining single-stranded primers (Hanke and Winnke 1994). Digestion with Exol is also used to solubilize any free primers after the final amplification steps prior to digestion with NlaIII to release the internal identifier tags from their flanking GST linker cassettes. Because the linker-specific primers used in amplification are biotinylated at their 5' end, streptavidin beads can be used to capture the liberated cassettes, thereby avoiding losses that would accompany gel purification of the 19-bp-long tags (Powell 1998).

Analysis of a Y. pestis BamHI GST Library

Shown in Figure 1 and Table 1 are the predicted numbers of tags that would be generated at each step of the procedure from Y. pestis DNA using either NotI or BamHI as the fragmenting enzyme. Using the 4.7-Mb Y. pestis CO92 complete genome (minus the pCD1 plasmid) as input (Parkhill et al. 2001), we determined in silico that there should be 64 cleavage sites for NotI, 699 sites for BamHI, and 16,572 sites for NlaIII. Only one NotI fragment is predicted to lack an internal NlaIII site, but 36 of the smaller fragments generated by BamHI should not be cleaved by NlaIII. The mean lengths of the resulting NotI-NlaIII and BamHI-NlaIII fragments are 273 and 267 bp, respectively. The similarity in these mean fragment lengths reflects both the high density and nearly random distribution of NlaIII sites in the Y. pestis genome. Only 11 of the NotI-NlaIII and 90 of the BamHI-NlaIII fragments are predicted to be <21 bp long, all other fragments should generate full-length 21-bp tags. If only 21-bp tags are considered, then the NotI-NlaIII library should sample ~2.4 kb of the Y. pestis sequence, whereas the BamHI-NlaIII library would sample ~10 times more DNA, ~26 kb.

One problem that is intrinsic to the method occurs when the MmeI recognition sequence (GTYGGA) is within 21 bp of the NlaIII end. This sequence would direct cleavage back towards the NlaIII end, allowing MmeI to potentially cut within the attached MmeI linker, which would interfere with subsequent PCR amplification. A GTYGGA sequence within the next 21 bp could potentially give rise to tags <21 bp long, depending on which site is first recognized by MmeI. Analysis of the Y. pestis sequence indicates that MmeI digestion would at most eliminate only 17 tags from a BamHI library but none from the NotI-derived library. Although all of the 21-bp NotI-derived tags are unique, 47 of the BamHI-derived 21-bp tags come from 14 repeated sequences and therefore occur two or more times within the database.

To validate the generality of this method, we prepared a

1758 Genome Research
Y. pestis GST library using BamHI as the fragmenting enzyme because it was predicted to generate sufficient tags for meaningful data analysis. Sequence analysis of our initial library showed that MmeI can liberate both 21- and 22-bp-long tags from the same location in the DNA. Analysis of this library, which was prepared using a single NlaIII digestion step, also revealed the presence of a large fraction of tags that originated from NlaIII sites that were not proximal to a BamHI site. The presence of these tags in the library obviously was the result of incomplete NlaIII digestion; therefore, we now routinely include a second NlaIII digestion step after the biotinylated fragments are captured on the magnetic beads. The data reported here are from a single library prepared following the steps outlined in Figure 1. The cloned inserts in this library were typically several hundred base pairs to slightly <1 kb long.

The linker we used to biotinylate the BamHI digest adds 12 bp to the ends of each fragment. In principle, the addition of this linker should allow MmeI to liberate 21-bp-long tags even from the 90 BamHI-NlaIII fragments predicted from our in silico experiments to be <21 bp long. In these cases, MmeI would have to cleave within the attached linker. Tags from these sites are easy to identify, as they should contain a BamHI recognition sequence near their 3’ ends. To simplify discussion, we number the fragments according to their order along the DNA and use R (reverse) and F (forward) to indicate the relative location of the GST within the fragment; thus, R314 indicates the reverse GST from BamHI fragment number 314, which would be followed by F314 (the next forward GST), R315, F315, etc.

### GST Analysis

A total of 5432 GSTs were extracted from the sequenced arrays. The number of 21- and 22-bp-long tags was approximately equal, 2701 and 2731, respectively. The vast majority, 5268 (97%), exactly matched at 1133 sites in the Y. pestis genome. This includes a total of 336 tags that were uniquely matched at 88 correct tagging sites, even though their initial polarities were ambiguous. Most of these unique matches could be assigned to the first NlaIII site next to a BamHI fragmentation site, which indicates that the two-step NlaIII digestion was virtually complete. Only 59 (1%) of the extracted tags exactly matched interior NlaIII sites. These tags could result from over digestion with BamHI or partial NlaIII digestion; however, we suspect that several may have arisen because subtle changes in the genome introduced new BamHI sites. This seems to be the case for fragments 90 and 459, which each gave rise to two internal tags. Two other internal tags occurred twice, which, because of the large number of total NlaIII sites in the Y. pestis DNA, is a highly improbable random event. A small number of tags (six) that passed all our editing criteria have no obvious close match to the Y. pestis genome or any other sequence in GenBank. These might originate from sequences that are unique to the EV766 genome or represent spurious tags generated during library construction, amplification, and cloning. Of the total predicted potential tagging sites, 209 were still unseen. We believe that many, but not all, of these unseen sites would be matched if the sample size were increased (see below). A detailed analysis of the data is available at [http://genome.bnl.gov/GSTs](http://genome.bnl.gov/GSTs).

To a first approximation, cloning and sequencing of GSTs should be random processes and on average, the relative frequency of occurrence of a particular GST in a library should reflect its frequency in the DNA sample. Therefore, tags from highly repetitive regions of the chromosome or from higher-copy-number plasmids should be more numerous than tags from unique regions. This prediction seems to hold true for our GST library. As shown in Table 2, the most numerous tag we encountered is the one predicted to occur most frequently (eight times) in the Y. pestis chromosome. It was followed in order by the tag predicted to be the next most frequent, the one occurring seven times. Only one tag should be present five times; one should be present four times; three tags should each be found three times; and seven tags should each occur twice. Two other redundant tags listed in Table 2 should not be recovered at all because each contains a BamHI fragmentation site very close to its 5’ end. The actual observed frequency of the multiple tags is highly correlated (r = 0.88) with the predicted frequency. However, one tag that is predicted to be present four times in the genome seems to be underrepresented in our database. This tag is associated with an IS100 element that is known to be a source for genetic variability in Y. pestis isolates (Motin et al. 2002), which may in part explain our results. The two plasmids, pMT1 and pPCP1, thought to be present in the EV766 genome, each contain a single BamHI site, and each should have contributed two unique tags to our library. All four tags were catalogued at about the same frequency as single-copy chromosomal tags. This would suggest that neither of these plasmids had a sig-

<table>
<thead>
<tr>
<th>Tags of length ≥21</th>
<th>start*</th>
<th>after MmeI digestion</th>
<th>start</th>
<th>after MmeI digestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted tags</td>
<td>115 (7)</td>
<td>115 (7)</td>
<td>1236 (96)</td>
<td>1214 (93)</td>
</tr>
<tr>
<td>Unique tags</td>
<td>115 (7)</td>
<td>115 (7)</td>
<td>1203 (94)</td>
<td>1181 (91)</td>
</tr>
<tr>
<td>Single tags</td>
<td>115 (7)</td>
<td>115 (7)</td>
<td>1189 (92)</td>
<td>1167 (89)</td>
</tr>
<tr>
<td>Multiple tags</td>
<td>0 (0)</td>
<td>14 (2)</td>
<td>14 (2)</td>
<td></td>
</tr>
<tr>
<td>Tags of length &lt;20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted tags</td>
<td>7 (0)</td>
<td>7 (0)</td>
<td>89 (12)</td>
<td>89 (12)</td>
</tr>
<tr>
<td>Unique tags</td>
<td>7 (0)</td>
<td>7 (0)</td>
<td>86 (12)</td>
<td>86 (12)</td>
</tr>
<tr>
<td>Single tags</td>
<td>7 (0)</td>
<td>7 (0)</td>
<td>84 (12)</td>
<td>84 (12)</td>
</tr>
<tr>
<td>Multiple tags</td>
<td>0 (0)</td>
<td>2 (0)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Zero length tags</td>
<td>4 (0)</td>
<td>4 (0)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>126 (7)</td>
<td>126 (7)</td>
<td>1326 (108)</td>
<td>1303 (105)</td>
</tr>
</tbody>
</table>

*aValues in parenthesis are the numbers of tags with ambiguous directions, i.e., they begin with the sequence CATGAA.

*bZero length tags occur when the fragmenting site is immediately adjacent to a NlaIII site.
Table 2. Correspondence Between Predicted and Actual Identifier Tag Frequencies

<table>
<thead>
<tr>
<th>Identifier tag sequence</th>
<th>Frequency</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATCTGAGGTTTCGTTTC</td>
<td>8</td>
<td>65</td>
</tr>
<tr>
<td>GTCATCTTCTCTGGAAAC</td>
<td>7</td>
<td>45</td>
</tr>
<tr>
<td>GAATATTATCCGCCTTTCT</td>
<td>5</td>
<td>34</td>
</tr>
<tr>
<td>CCGTCGCTACGGCAGAC</td>
<td>3</td>
<td>34</td>
</tr>
<tr>
<td>GTCTAGCGGCCCATTTT</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>CCAAGTACGCGCACGACC</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>TAGAATCGAACCGACACC</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>TCCTTCAAAATCAAGGGA</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>CGATAAACCGGACATG</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>AATCCTACTCACTACACG</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>CTTCTGCTGGTTAGTCAACG</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>CCCAACCCCTGTGCACCC</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>AACCCCTATACATCAG</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>TGCGTTCCCAAGAGCTT</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>TGGAACCGAAGGGGTTT</td>
<td>3</td>
<td>unseen-contains BamHI site</td>
</tr>
<tr>
<td>GGGATCCGAAAGGGGTTT</td>
<td>2</td>
<td>unseen-contains BamHI site</td>
</tr>
</tbody>
</table>

Complete lists of GSTs, in both order of abundance and position in the Yersinia pestis genome, are available via the internet at http://genome.bnl.gov/GSTs. “<CATG omitted.”

Table 3. Potential Deletions in the Yersinia pestis EV766 Genome

<table>
<thead>
<tr>
<th>Start-end</th>
<th>Position bp</th>
<th>IS element</th>
<th>No. of tags affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>F344–F327</td>
<td>2,172,627–2,254,447</td>
<td>yes IS100</td>
<td>25</td>
</tr>
<tr>
<td>R194–R197</td>
<td>1,307,243–1,316,087</td>
<td>yes IS154</td>
<td>7</td>
</tr>
<tr>
<td>F227–F228</td>
<td>1,554,643–1,556,368</td>
<td>no</td>
<td>3</td>
</tr>
<tr>
<td>F327–F328</td>
<td>1,618,033–1,652,133</td>
<td>yes IS100</td>
<td>3</td>
</tr>
<tr>
<td>F381–F382</td>
<td>2,662,263–2,685,036</td>
<td>no</td>
<td>3</td>
</tr>
<tr>
<td>F453–F454</td>
<td>3,069,009–3,122,226</td>
<td>no</td>
<td>Total 44</td>
</tr>
</tbody>
</table>

apparently contains point mutations. Inspection of the relevant single-pass sequencing chromatograms indicates that the original base calls were accurate. In nearly every case, the corresponding correct GST could be found in the data set. Presumably these differences represent errors introduced during library preparation rather than true polymorphisms in the DNA sample. The distribution of mismatches within the tags was not totally random; discrepancies were somewhat more frequent within the last two bases at the 3’ end of the tag. This most likely reflects mislabeling between the MinI overhangs and the 16-fold degenerate cassette during this step in the GST protocol. Increased fidelity should be possible by using a lower concentration of the degenerate primer, shorter incubation times, or higher temperature during the ligation step. One empirical way to eliminate most of these errors is to omit tags encountered only once from further analysis, as is typically done to help eliminate sequencing and other errors from SAGE libraries. This type of filtering would eliminate all but 23 of the imperfectly matched tags from further consideration.

Generation of Longer Sequences From GSTs

The sequence complexity and length of a GST, 21 to 22 bp, should in most cases be sufficient to enable its use directly as a primer to amplify the stretch of DNA between the tagging site and the proximal site for the fragmenting enzyme. This is especially important as a GST library readily generates large numbers of tags that can then be converted into longer genomic DNA fragments for more detailed analysis of the source DNA or for further characterization of novel genomes. To test this concept a group of five tags predicted to begin –100 to 1000 bp away from their proximal BamHI sites were selected and used for custom primer synthesis. Template Y. pestis DNA was digested with BamHI and ligated with a linker cassette that introduced an identical priming site at both ends of each fragment. The DNA was then digested with NlaIII to physically separate the linker BamHI ends. Aliquots were then subjected to 10 rounds of linear PCR amplification using just the GST-specific primer to increase the amount of complementary single-stranded targets in the sample. This step was then followed by 25 PCR cycles with both primers. As shown in Figure 3, each reaction generated a distinct band of the expected length. Direct sequencing of these five bands unequivocally confirmed their correct location in the Y. pestis genome.

Potential Enhancements

Although the data we obtained show that we largely achieved our objectives, further analysis (Fig. 4) suggests that we are
undersampling tags that lie a short distance from the fragmenting site. This deficiency can be easily addressed by increasing the length of the biotinylated cassette used to attach the DNA to the streptavidin beads. In this context it is worth noting that Wang and Rowley (1998) observed that a SphI site (GCATGC) tethered to a streptavidin bead by a short linker could be cut with SphI but not with NlaIII, even though the linker contained a CATG sequence.

**DISCUSSION**

We have described a method for obtaining 21- to 22-bp GSTs from predetermined positions in genomic DNAs. In principle, the method can provide limited representation of all the DNA molecules in a sample without prior knowledge of the DNA sequence. The approach can be fine-tuned by the user to provide different degrees of coverage and discriminatory power, depending on the choice of fragmenting enzyme. The method is similar to the TALEST modification (Spinella et al. 1999) of the original SAGE protocol in that it uses a 16-fold degenerate linker cassette to attach an oligonucleotide adapter to the unknown 3' overhangs of MmeI-digested DNAs, thereby taking advantage of being able to use cohesive termini for high-efficiency linker addition. Addition of this linker not only provides an appended sequence for PCR amplification but also attempts to reduce bias during amplification by flanking the monomeric GSTs on both sides with distinct long linkers. Because the degenerate linker is in molar excess during ligation to the MmeI-generated ends, few tags should self-ligate and be sandwiched by the same GST linker. GST panhandle structures, which would result in low amplification efficiency, are thereby avoided. In contrast, excess degenerate linker, which should dimerize during ligation, is expected to form panhandles that should suppress their amplification. Other non-standard steps in our GST amplification strategy include two separate rounds of linear amplification to generate sufficient material for library construction while at the same time reducing product heteroduplexes.

The results of this study show that the GST technique provides a route to obtaining numerous 21- to 22-bp sequence tags that can be used to identify the DNA source, and as shown, the presence or absence of particular tags can provide some indication of the genetic variability between two closely related strains. The length of the tags allows direct determination of the source DNA if the sequence is available. An in silico comparison of all the BamHI-NlaIII GSTs that would be generated from a mixture of the 60 complete microbial genomes in the NCBI database demonstrated that these different bacterial strains share few GSTs in common. Table 4 contains a list of the top 30 shared tags. The worst case scenario is the occurrence of a single tag that was found three times in E. coli and once in Y. pestis. No GST was shared by three strains, although this might change as more closely related organisms are sequenced. Even between closely related strains, the frequency of unique unshared identifiers is more than adequate to allow strain differentiation. A comparison between the 4.6-Mb E. coli K12 and 5.5-Mb O157H7 genomes predicts that they would generate 863 and 1018 unique BamHI-NlaIII GSTs, respectively. Although they share 554 common tags that would classify the DNA as being E. coli, the K12 genome has 309 unique GSTs and the O157H7 genome has 464 that might be used to accurately differentiate between them.

Assuming a 50% G+C content, an enzyme such as NotI with an eight-base recognition sequence will cleave on average every 48 (65.5 kb) bases compared to every 46 (4 kb) bases.

![Figure 3](https://example.com/figure3.png)

**Figure 3** Specific amplification of end sequences corresponding to a specific GST in the *Yersinia pestis* genome. In each PCR, a specific GST sequence was used as a primer along with a primer complementary to the common GATC adaptor ligated to the fragmentation ends. The sizes of the resulting amplicons are 102 + 23 = 125; 205 + 23 = 228; 230 + 23 = 253; 384 + 23 = 407; and 1031 + 23 = 1054 bp (lanes 2–6, respectively). Lane 1 contains a 100-bp ladder.

![Figure 4](https://example.com/figure4.png)

**Figure 4** Length distribution of *Yersinia pestis* BamHI-NlaIII fragments. The number of GSTs is plotted on the Y-axis. Fragment lengths are plotted on the X-axis. Shown in green are the fragment lengths associated with the predicted GSTs; in blue, for the observed GSTs; and in red, for the unseen GSTs.
for a restriction enzyme with a six-base recognition sequence, such as **Bam**HI. In practice, this means that fragmenting the DNA with **Bam**HI will usually produce 10 times more GST tags from a genome than would fragmentation with **Not**I. Other factors that influence the average fragment size generated by the fragmenting enzyme are G+C content, dinucleotide frequency, and sensitivity to methylation. CpG methylation completely blocks cleavage by **Not**I, and such sites would be missed if only **Not**I were used for fragmentation. Fortunately, there are at least 10 other commercially available enzymes with specificities greater than six bases that can be used for this purpose. Some of these enzymes, such as **Pac**I (recognition sequence TTAATTAA), cut only A+T-rich DNAs, whereas others cut mainly G+C-rich DNAs but are not sensitive to CpG methylation. The ability to select fragmenting enzymes to suit the characteristics of a particular genome (e.g., A+T-rich or G+C-rich) is one strength of the GST method.

Choosing a fragmenting enzyme, we prefer to use one that leaves cohesive ends for ligation with appropriate biotinylated linker cassettes. We believe that cohesive end-mediated ligation with a biotinylated linker cassette is an important discriminatory GST tool, as it alleviates the problem of having to enzymatically biotinylate only the ends of the DNA that were generated by enzymatic cleavage, which in practice can be very difficult when dealing with DNA isolated from nonlaboratory sources in which degradation may be a problem. In fact, for GST analysis the starting DNA does not have to be high molecular weight because, as shown in Figure 4, even a relatively small fragment containing a site for the fragmenting enzyme should carry a nearby site for the **Nla**III tagging enzyme.

The only mathematical assumption behind the GST method is that the probability of observing specific GSTs should closely follow the Poisson distribution; therefore, the probability of observing a given tag with 1/N abundance while sequencing N tags is 0.63. Tags with abundance larger than 1/N should be sampled more frequently, provided that the PCR amplification and subsequent cloning steps used to obtain the library are not biased, which would compromise the quantitative aspects of the method. In developing the GST method, several steps were critically evaluated to help ensure that the frequency of tags in our library reflected the predicted frequency of tags in the *Y. pestis* genomic DNA. The frequency distribution of the tags in our *Y. pestis* database appears to be quite flat, and as might be expected, many of the most abundant GSTs were derived from repetitive sequences.

---

**Table 4. Shared GSTs Between Two Different Bacteria**

<table>
<thead>
<tr>
<th>GST sequence&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Organisms</th>
<th>Total</th>
<th>Organism (count)</th>
<th>Organism (count)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCCGCCCTTAACCGCCGCA</td>
<td>2</td>
<td>4</td>
<td>Escherichia coli (3)</td>
<td>Yersinia pestis (1)</td>
</tr>
<tr>
<td>GCCGCCCTGATCTCCGGG</td>
<td>2</td>
<td>3</td>
<td>Mycobacterium leprae (1)</td>
<td>Mycobacterium tuberculosis (2)</td>
</tr>
<tr>
<td>GCAGCGTATTGGTAC</td>
<td>2</td>
<td>3</td>
<td>Mycobacterium leprae (1)</td>
<td>Mycobacterium tuberculosis (2)</td>
</tr>
<tr>
<td>CCGCCCCGGAAAAATCAAC</td>
<td>2</td>
<td>3</td>
<td>Mycobacterium leprae (1)</td>
<td>Mycobacterium tuberculosis (2)</td>
</tr>
<tr>
<td>GAAGCTGGCGCACCAGGCA</td>
<td>2</td>
<td>3</td>
<td>Mycobacterium leprae (1)</td>
<td>Mycobacterium tuberculosis (2)</td>
</tr>
<tr>
<td>CCTGCGCTTACACGAC</td>
<td>2</td>
<td>3</td>
<td>Pyrococcus abyssi (1)</td>
<td>Pyrococcus horikoshii (2)</td>
</tr>
<tr>
<td>CTCGCCCTTTGTACG</td>
<td>2</td>
<td>3</td>
<td>Mycobacterium leprae (1)</td>
<td>Mycobacterium tuberculosis (2)</td>
</tr>
<tr>
<td>TGGAATGCTCTTGGATAC</td>
<td>2</td>
<td>3</td>
<td>Mycobacterium leprae (1)</td>
<td>Mycobacterium tuberculosis (2)</td>
</tr>
<tr>
<td>GCCTCCCGTACCACGCA</td>
<td>2</td>
<td>3</td>
<td>Mycobacterium leprae (1)</td>
<td>Mycobacterium tuberculosis (2)</td>
</tr>
<tr>
<td>GAAGCTGGATTTGTCG</td>
<td>2</td>
<td>3</td>
<td>Mycobacterium leprae (1)</td>
<td>Mycobacterium tuberculosis (2)</td>
</tr>
<tr>
<td>AACGCCTCGCCTCCACTAT</td>
<td>2</td>
<td>2</td>
<td>Mesorhizobium loti (1)</td>
<td>Sinorhizobium meliloti (1)</td>
</tr>
<tr>
<td>GTGACTTCTTCTGAAATGT</td>
<td>2</td>
<td>2</td>
<td>Chlamydia muridarum (1)</td>
<td>Chlamydia trachomatis (1)</td>
</tr>
<tr>
<td>TGCACCCGAATCTCGGAT</td>
<td>2</td>
<td>2</td>
<td>Mesorhizobium loti (1)</td>
<td>Sinorhizobium meliloti (1)</td>
</tr>
<tr>
<td>ACCACCTTCTCTCTCTCTA</td>
<td>2</td>
<td>2</td>
<td>Thermoplasma acidophilum (1)</td>
<td>Thermoplasma volcanium (1)</td>
</tr>
<tr>
<td>TGGAGCAGACCCTTGCG</td>
<td>2</td>
<td>2</td>
<td>Agrobacterium tumefaciens (1)</td>
<td>Sinorhizobium meliloti (1)</td>
</tr>
<tr>
<td>ACCCGAGGATGGGCCG</td>
<td>2</td>
<td>2</td>
<td>Mesorhizobium loti (1)</td>
<td>Sinorhizobium meliloti (1)</td>
</tr>
<tr>
<td>ACCAGAAGATCTATGCTCC</td>
<td>2</td>
<td>2</td>
<td>Thermoplasma acidophilum (1)</td>
<td>Thermoplasma volcanium (1)</td>
</tr>
<tr>
<td>AATTTAAAATATGAC</td>
<td>2</td>
<td>2</td>
<td>Chlamydia muridarum (1)</td>
<td>Chlamydia trachomatis (1)</td>
</tr>
<tr>
<td>CGGACTCTGGCTGCTT</td>
<td>2</td>
<td>2</td>
<td>Mesorhizobium loti (1)</td>
<td>Sinorhizobium meliloti (1)</td>
</tr>
<tr>
<td>TTCTGAAAGTGGAAC</td>
<td>2</td>
<td>2</td>
<td>Pyrococcus abyssi (1)</td>
<td>Pyrococcus horikoshii (1)</td>
</tr>
<tr>
<td>CCCTCACTATGAAACC</td>
<td>2</td>
<td>2</td>
<td>Chlamydia muridarum (1)</td>
<td>Chlamydia trachomatis (1)</td>
</tr>
<tr>
<td>AACGATTTGAAATGGA</td>
<td>2</td>
<td>2</td>
<td>Thermoplasma genitalium (1)</td>
<td>Mycoplasma pneumoniae (1)</td>
</tr>
<tr>
<td>ATGCCAGATGCGATTCT</td>
<td>2</td>
<td>2</td>
<td>Agrobacterium tumefaciens (1)</td>
<td>Sinorhizobium meliloti (1)</td>
</tr>
<tr>
<td>CAGTGGCTGGATTTACC</td>
<td>2</td>
<td>2</td>
<td>Pyrococcus abyssi (1)</td>
<td>Pyrococcus horikoshii (1)</td>
</tr>
<tr>
<td>AAGGTCCTTCAATGTGCCG</td>
<td>2</td>
<td>2</td>
<td>Thermoplasma acidophilum (1)</td>
<td>Thermoplasma volcanium (1)</td>
</tr>
<tr>
<td>CGGGGAAATGCGAATTC</td>
<td>2</td>
<td>2</td>
<td>Chlamydia muridarum (1)</td>
<td>Chlamydia trachomatis (1)</td>
</tr>
<tr>
<td>CAGCGACCACTCGAGCCG</td>
<td>2</td>
<td>2</td>
<td>Mesorhizobium loti (1)</td>
<td>Sinorhizobium meliloti (1)</td>
</tr>
<tr>
<td>AAATTTGACCTTATTTTTA</td>
<td>2</td>
<td>2</td>
<td>Chlamydia muridarum (1)</td>
<td>Chlamydia trachomatis (1)</td>
</tr>
<tr>
<td>CAGCGACCTTATTTTTA</td>
<td>2</td>
<td>2</td>
<td>Clostridium acetobutylicum (1)</td>
<td>Pyrococcus horikoshii (1)</td>
</tr>
</tbody>
</table>

<sup>a</sup>GSTs within 25 bp of the **Bam**HI fragmentation site were omitted.

<sup>b</sup>CATG omitted.
to isolate overlapping cloned DNAs in metagenome libraries (Rondon et al. 2000). Differential tag information could be used in conjunction with traditional culture techniques to help complete the catalog of species present in a sample. We are actively pursuing a pilot study to demonstrate this application. Application of GSTs to analyze the complexity of microbial communities may necessitate the use of two or more fragmentation enzymes to ensure adequate depth and resolving power of the GST coverage.

Only minor changes in the GST protocol are needed to use the method for modified long SAGE analysis of poly(A)⁺ eukaryotic mRNAs. In this case, double-stranded cDNA is synthesized from the mRNA by means of a biotinylated oligo (dT) primer anchored to streptavidin beads (Virion et al. 1999). The cDNA is then cleaved with NlaIII, leaving the 3′-most portion of the cleaved cDNA with the cohesive overhang needed for ligation of the Mmel adaptor. All other steps then proceed as outlined in Figure 1. We have implemented this method to obtain 21- to 22-bp SAGE tags to profile gene expression in human platelets (D.V. Gnatenko, J.J. Dunn, S.R. McCorkle, D. Weissmann, P.L. Perrotta, and W.F. Bahou, in prep.). Likewise, the long SAGE protocol (Saha et al. 2002) could easily be modified to obtain GSTs by starting with biotinylated genomic DNA fragments rather than poly (A)⁺-derived cDNA. Approximately the same amount of time, about a week, would be needed to generate a GST library using either method. The major difference between the two procedures occurs during the formation and subsequent amplification of the resulting tags. In the long SAGE protocol, self-ligation of Mmel overhangs is used to form ditags (Saha et al. 2002), whereas in the GST method, an excess of a 16-fold degenerate cassette (Spinella et al. 1999) is used to add the oligonucleotide adapter needed for PCR amplification. Although ligation of the degenerate cassette and subsequent PCR of the monotags might be more efficient under some conditions, the orientation of the cloned monotags can only be independently determined by the position of the extra nonpalindromic bases that are added during ligation with the degenerate cassette. These added bases also define the exact length of the tag because Mmel can cleave, as shown here, 20 or 21 bases past its recognition sequence with nearly equal probability. In long SAGE, the orientations of the individual tags in the concatamers are unambiguous; however, because of the variability in tag length, some caution is needed in determining where one tag ends and another begins in each ditag. An additional but more subtle difference between the two methods is that during formation of ditags, the most redundant SAGE tags can ligate to one another to form the same ditag more than once. This can cause preferential PCR amplification of certain ditags in cDNA-based SAGE libraries. These replicate ditags, which arise mainly from the most abundant mRNA species, are usually excluded from the tag database, which may cause underestimation of the actual frequency for some abundant mRNA species in highly specialized tissues, as was recently demonstrated in a SAGE study of human skeletal muscle (Welle et al. 1999). However, it remains to be seen whether a combined GST long SAGE-based approach that relies on amplification of individual monotags is in reality less prone to underestimation of mRNA abundance.

In summary, the basic GST procedure described here provides a means for genome-wide fingerprinting of chromosomal and episomal DNAs and, by extension, for profiling DNA genomes in natural populations. Like SAGE, it can be performed with equipment available in most molecular biology laboratories. The GST technique can be used, with minor modifications, for long SAGE analysis of eukaryotic mRNAs and might, like AFLP of cDNA (Qin et al. 2001; Donson et al. 2002), be adaptable for profiling gene expression in prokaryotes.

**METHODS**

**DNA Fragmentation and Biotinylated Adaptor Ligation**

DNA from avirulent *Y. pestis* EV766, a Ca²⁺-independent strain cured of the 70.5-kb pCD1 plasmid but retaining the pPCP1 9.5-kb and 100-kb pMT1 plasmids (Portnoy and Falkow 1981), was kindly provided by James Bliska (State University of New York at Stony Brook). Ten micrograms was digested with 100 U of BsmHI (New England Biolabs [NEB]), extracted with an equal volume of phenol/chloroform (P/C), and precipitated with ethanol. After centrifugation, the pellet was resuspended in 34 µL TEsl (10 mM Tris-acetate at pH 8.0, 0.1 mM EDTA-Na₂). A biotinylated GATC oligonucleotide adaptor cassette was created by mixing 3600 pmole each of two synthetic oligonucleotides (sense strand, CGA ACC CCT TTCG; antisense strand, P-GAT CCG AAG GGG TTC GT-BIOTIN) in 100 µL OFA buffer (10 mM Tris-acetate at pH 7.5, 10 mM Mg-acetate, 50 mM K acetate; Amersham Bioscience), heating them to 95°C for 2 min, and then allowing them to cool slowly to room temperature. An ~50-fold excess of biotinylated cassette (~600 pmole) relative to available BsmHI ends was ligated to the fragmented DNA in a total volume of 50 µL of 1× ligase buffer (Takara) containing 350 U of T4 DNA ligase (Takara). The reaction was incubated overnight at 16°C, followed by extraction with an equal volume of P/C. The sample was precipitated with ethanol, centrifuged, and resuspended in 83 µL TEsl.

**First Digestion With NlaIII and Binding to Magnetic Beads**

The fragmented DNA was next digested with 25 U of NlaIII (NEB) in 100 µL NlaIII digestion buffer (1× NEB buffer 4 supplemented with 1× BSA and 10 mM spermidine [HCl₃], for 3 h at 37°C; NlaIII digestion is stimulated two- to fourfold by addition of spermidine; J.J. Dunn, unpubl.). One hundred microliters (1 mg) of streptavidin magnetic beads (Dynal Biotech Inc.) were washed twice with 200 µL of 1× magnetic bead binding buffer (MBB; 10 mM Tris-HCl at pH 7.4, 1 mM EDTA-Na₂, 1 M NaCl) and then resuspended in 100 µL of 2× MBB. The beads were then added to the NlaIII-digested DNA in a nonstick 1.5-mL microfuge tube (Ambion). The beads and digest were mixed gently for 1 h at room temperature to bind biotinylated BsmHI-NlaIII fragments.

**Second Digestion With NlaIII and Mmel Adaptor Ligation**

A second incubation with NlaIII was performed on the bound fragments by resuspending the beads in 200 µL NlaIII digestion buffer containing 25 U of enzyme and incubating for 2 h at 37°C, after which an additional 25 U of enzyme was added and incubation continued for 2 h. The beads were washed three times with 200 µL TEsl, to remove nonbound DNA fragments, and one time with 200 µL 1× T4 ligase buffer. A Mmel oligonucleotide adaptor was created by mixing and annealing as described above; 1000 pmole each of two synthetic oligonucleotides (sense strand, TTT GGA TTT GCT GGT CGA GTA CAA CTA GCC TTC TTA ATC CGA CAT G; antisense strand, P-TCG GAT TAA GCC TAG TCG TAC TCG ACC AGC AAA TCC-AmMC7) in 100 µL 1× OFA. The annealed Mmel adaptor cassette (40 pmole) was ligated to the fragmented solid-
phase DNA for 2 h at 16°C in a total volume of 50 µL of 1×
ligase buffer (Takara) containing 350 U of T4 DNA ligase
(Takara).

Digestion With Mme1

Beads were washed six times with 400 µL 1× MBB and then
washed several times with 200 µL Mme1 digestion buffer (100
mM HEPEs at pH 8.0, 25 mM K acetate at pH 8.0, 50 mM Mg
acetate at pH 8.0, 20 mM DTT, 4 mM S-adenosylmethionine-
HCl). The beads were then resuspended in 100 µL 1× Mme1
digestion buffer containing 8 U Mme1 (Center of Technology
Transfer) and incubated for 3 h at 37°C with occasional mix-
ing. The beads were collected, and the supernatant containing
the released tags was a new microfuge tube. The beads were washed with 100 µL TEsL, and the wash was com-
bined with the first Mme1 supernatant. The pooled Mme1 di-
gest is extracted with an equal volume of P/C and precipitated
at −80°C for 1 to 2 h with 1 mL of ethanol after addition of
133 µL 7.5 M ammonium acetate and 2 µL glyco blue (Am-
bia) as carrier. The resulting pellet was washed with cold
75% ethanol, dried in vacuo, and resuspended in 29.5 µL TEsl
plus 4 µL 10× T4 DNA ligase buffer.

Second Cassette Ligation and Initial
PCR Amplification

A second, 16-fold degenerate adaptor cassette was prepared by
annealing two synthetic oligonucleotides as described above
(sense strand, 5′-TTC ATG GCC GAG TCC GCC ACT AGT
gTCA GGT GCA ACT TA-AmnMcT; antisense strand, TAG
tCA GGT TCG ACA CTA GTG GCC GAC GTC TCC GCC AFG
AAC N). Thirty-five picomoles of adaptor cassette (3.5 µL) was
added to the resuspended tags, and after 15 min at room tem-
perature, 3 µL of ligase (1000 U-Takara) was added and the
reaction incubated overnight at 16°C (Ryo et al. 2000). The
ligation products were subjected to PCR amplification, con-
sisting of an initial denaturation step for 2 min at 95°C, fol-
lowed by 30 cycles of 30 sec at 95°C, 30 sec at 58°C, and for
30 sec at 72°C, with a final extension step for 4 min at
72°C using 5′-Biotin-GGA TTT GCT GGT CGA GTA CA and
5′-Biotin-TAG TCA GGT TCG ACA CTA GTG GC as forward
and reverse primers, respectively, each at a final concentra-
tion of 0.4 µM. Cycling was performed in 1× Promega buffer
containing 2 mM Mg sulfate and 0.3 mM of each dNTP. Typi-
cally 0.8 to 1.0 µL of ligation product was amplified in a 200
µL reaction containing 0.8 µL Platinum Taq DNA polymerase
mixture (Invitrogen).

Linear Amplification to Reduce Heteroduplexes

The PCR products were then subjected to one round of
LARHD by diluting them to 1 mL with 800 µL 1× PCR buffer
containing 4 µL Platinum Taq and an additional 400 pmole of
each biotinylated primer. The reaction was then incubated for
2.5 min at 95°C, for 30 sec at 58°C, and for 5 min at 72°C.
Unincorporated primers were digested by addition of 10 µL
(200 U) of single strand–specific E. coli Exol. After 1 h at 37°C,
the sample was P/C extracted and precipitated by addition of
2.5 volumes of ethanol in the presence of 0.3 M Na acetate
(pH 6.0).

Second Linear Amplification (LARHD2), NlaII
Digestion, and Concatemerization

After centrifugation, the pellet was washed in 70 % ethanol,
dried, and then dissolved in 200 µL TEsL. A portion (20%) was
subjected to 25 additional rounds of linear amplification un-
der the above LARHD conditions, except only the forward
primer was added. This was then followed by one round of
amplification after addition of the reverse primer and addi-
tional DNA polymerase to convert the linear amplification
products to double-stranded DNA. Typically, 1 mL of sample
is amplified, and any unincorporated primers are hydrolyzed
by incubation with Exol as above. After P/C extraction and
ethanol precipitation, the amplified DNA is digested with 20
U of NlaII in 400 µL for 4 h at 37°C (after 2 h, the completion
of digestion is checked by electrophoresis of a small aliquot
on a 10% polyacrylamide gel). The digest is then extracted on
ice with chilled P/C to prevent denaturation of the smaller
duplexes and ethanol-precipitated from Na acetate in the
presence of glyco blue carrier. The sample is chilled for several
hours and then centrifuged at 4°C. The pellets are resus-
pended in 200 µL ice-cold TEsL plus 25 mM NaCl, diluted with
an equal volume of 2× MBB, and added to 200 µL (2 mg) of
streptavidin beads equilibrated with 1× MBB. After gentle
mixing for 15 min at room temperature, the unbound frac-
tion is transferred to a second 200 µL aliquot of beads to
capture any remaining biotinylated DNA fragments. The un-
bound GST fraction is recovered, precipitated by addition of
2.5 volume of ethanol and glyco blue carrier, and concate-
merized with T4 DNA ligase (5 U/µL; Invitrogen) for 4 h at
16°C. The sample is subjected to electrophoresis on a 0.75%
low-melt agarose gel, and products >100 bp are recovered.

These products are cloned into the Spli-site of a pZero plas-
maidivector (Invitrogen) that was engineered to have a Spli-minus
KanR gene (J. J. Dunn, unpubl.). To increase the efficiency of
cloning longer inserts, we used a two step ligation strategy
(Damak and Bullock 1993). Initially, an excess of GSTs is li-
gated with the Spli-digested vector at a high DNA concentra-
tion, a condition which promotes further concatemerization
of the tags onto ends of the linearized vector. The reaction is
then diluted and incubated overnight under conditions that
favor circularization. The resulting clones typically contained
20 to 40 GSTs.

Recombinant clones obtained after electrophoration of
competent TOP10 cells (Invitrogen) are selected on 2× YT
plates containing 50 µg/mL kanamycin. A schematic repre-
sentation of the method is shown in Figure 1, and a complete
description of all steps is available at the Web site (http://
genome.bnl.gov/GSTS).

DNA Sequencing

Plasmid DNA for sequencing was prepared using Edge BioSys-
tems reagents and protocols in 96-well plates. Templates were
cycle sequenced using ABI Prism BigDye terminator chemistry
from the M13 forward primer and analyzed on ABI 377 se-
querencers. Extracted data were ported to an Oracle database
and searched for valid tags using the GST software we de-
veloped. The software ensures that only unambiguous 21- to
22-bp tag sequences (see below) are extracted for further
analysis (tags with Ns, lengths other than 21 to 22 bases or
with polarity that is unambiguous) are extracted to separate
files for manual editing or further examination.

Ligation-Mediated PCR

Five Y. pestis–specific primers were synthesized: [535,384]
CAT GCA GGA TGG TGC ACG ACC CGA (205R); [2,281,342]
CAT GTG GCC GCC GGG CTT AA (384R); [2,894,318] CAT GAC
TCT GCC ATG TCT TG (1031R); [3,452,611] CAT GCA GGA
CCG ACA ATG (102F); and [4,145,945] CAT GCA GTG
CCA TCC TCA CGG (230F). The values in brackets are the
position of the NlaII tagging site in the Y. pestis chromosome.
The values in parentheses are the distances between the re-
spective NlaII and BamHI sites and the directionality of the
PCR reaction. BamHI-digested Y. pestis DNA was ligated with
a nonbiotinylated GATC oligonucleotide adaptor created by
mixing and annealing 3600 pmole each of two synthetic oli-
gonucleotides (sense strand, CGT AAT ACG ACT CAC TAT
AGG G; antisense strand, GCA TTA TGG TGA ATT ATC
GTC TAG) in 100 µL OFA as described above. The annealed
GATC adaptor (40 pmole) was ligated to BamHI-fragmented
DNA for 2 h at 16°C in a total volume of 50 μl of 1× ligase buffer (Takara) containing 350 U of T4 DNA ligase (Takara). All aliquots of the mix of amplified DNA were incubated for 2 min at 94°C, followed by 10 rounds of linear amplification (20 sec at 94°C, 30 sec at 55°C, and 4 min at 68°C) with the above Y. pestis-specific primers. This was followed by 25 additional rounds of amplification under the same conditions after addition of the common GATC-specific primer, the GATC sense strand. Products were extended for 10 min at 68°C and analyzed on 6% polyacrylamide gels. Extension with the sense-strand primer should add an additional 23 bp to the BamHI end of all the amplification products.

ACKNOWLEDGMENTS

We thank Willy Crockett, Kimberly Pellechi, Jutta Paparelli, and Judy Romeo for assistance in DNA sequencing, as well as Keith Thompson for statistical analysis. This project is supported by a Laboratory Directed Research and Development award (to J.D.) and by the Office of Biological and Environmental Research, and of Basic Energy Sciences (Division of Energy Biocatalysis) of the U.S. Department of Energy. The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


Received March 28, 2002; accepted in revised form September 11, 2002.

WEB SITE REFERENCE

http://genome.bnl.gov/GSTs; a detailed analysis of the data.
Genomic Signature Tags (GSTs): A System for Profiling Genomic DNA

John J. Dunn, Sean R. McCorkle, Laura A. Praissman, et al.

*Genome Res.* 2002 12: 1756-1765
Access the most recent version at doi:10.1101/gr.306102

References
This article cites 31 articles, 11 of which can be accessed free at:
http://genome.cshlp.org/content/12/11/1756.full.html#ref-list-1

License

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Cold Spring Harbor Laboratory Press