Submitted to

Genome Research

Revised 9/13/2002

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[##]

Biology Department, Brookhaven National Laboratory, Upton, NY 11973

Key words: Genome profiling, 21-bp tags

[*] To whom correspondence should be addressed:
Tel: +1-631-344-3012
Fax: +1-631-344-3407
Email: jdunn@bnl.gov

Division of Hematology, Department of Medicine, State University of New York at Stony Brook, Stony Brook, NY 11794-8151

[##] Department of Biology
Hofstra University
Hempstead, NY 11549

ABSTRACT

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 *Mme*I, 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 *Bam*HI as the fragmenting enzyme and *Nla*III 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.

INTRODUCTION

A variety of DNA-based fingerprinting techniques now exists 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 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 *Nla*III, 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 *Bsm*FI, a type IIS restriction endonuclease. Subsequent cleavage with *Bsm*FI releases short (13-14 bp) but positionally defined sequences, referred to as tags, which are eventually ligated to form "ditags", 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, i.e. from "uncharted transcriptomes" and in designing primers to obtain full-cDNAs from transcripts whose tags 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 *Mme*I which cleaves 20/18 bases past its non-palindromic (TCCRAC) recognition sequence (Boyd et al. 1986; Tucholski et al. 1995). *Mme*I has been used for development of long SAGE, which is an adaptation of the original SAGE approach that generates 21-bp tags from *Nla*III sites (Velculescu 2001; Saha et al. 2002). The long length of these tags (CATG + N_{17}) suggested to us that *Mme*I could be used to obtain unique tags directly from total microbial DNA owing to the number of *Mme*I 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, *Mme*I 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 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 *Mme*I recognition site is ligated to the non-biotinylated ends and *Mme*I 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 can also pinpoint areas of a genome that might have undergone changes which 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 *Nla*III, 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 *Nla*III to assure that *Nla*III digestion is complete. After the beads are washed, a duplex linker with *Nla*III cohesive termini is ligated to the bound DNA fragments. This linker generates a recognition site (TCCGAC) for the type IIS enzyme *Mme*I, the tagging enzyme, only when it is joined to *Nla*III cohesive ends. After washing to remove excess linkers, the beads are incubated with *Mme*I to release the linker and appended tags from the beads. Since the last C residue in the adaptor's *Mme*I recognition site partially overlaps the *Nla*III 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 *Mme*I. This adaptor was designed to add two consecutive T residues and a second *Nla*III site on the ends of the original *Mme*I generated fragments (TTCATG…). The ligation products are PCR-amplified using two linker-specific biotinylated primers, cleaved with *Nla*III and the two biotinylated end fragments removed by affinity capture on streptavidin-coated magnetic beads

(Powell,1998) leaving the 19 bp duplex GSTs with *Nla*III 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 *E. 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 where 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 Exo I, which preferentially hydrolyzes any remaining single-stranded primers (Hanke and Winke, 1994). Digestion with Exo I is also used to solubilize any free primers after the final amplification steps prior to digestion with *Nla*III to release the internal identifier tags from their flanking GST linker cassettes. Since 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 Bam*HI GST library

Shown in Figure 1 and Table 1 are the predicted numbers of tags which would be generated at each step of the procedure from *Y. pestis* DNA using either *Not*I or *Bam*HI 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 *Not*I, 699 sites for *Bam*HI and 16,572 sites for *Nla*III. Only one *Not*I fragment is predicted to lack an internal *Nla*III site but 36 of the smaller fragments generated by *Bam*HI should not be cleaved by *Nla*III.

The mean lengths of the resulting *Not*I-*Nla*III and *Bam*HI-*Nla*III fragments are 273 and 267 bp, respectively. The similarity in these mean fragment lengths reflects both the high density and nearly random distribution of *Nla*III sites in the *Y. pestis* genome. Only 11 of the *Not*I-*Nla*III and 90 of the *Bam*HI-*Nla*III fragments are predicted to be less than 21-bp long, all other fragments should generate full-length 21-bp tags. If only 21-bp tags are considered then the *Not*I –*Nla*III library should sample about 2.4 kb of the *Y. pestis* sequence while the *Bam*HI-*Nla*III library would sample about 10 times more DNA, about 26 kb.

One problem that is intrinsic to the method occurs when the *Mme*I recognition sequence (GTYGGA) is within 21-bp of the *Nla*III end. This sequence would direct cleavage back towards the *Nla*III end allowing *Mme*I to potentially cut within the attached *Mme*I linker which would interfere with subsequent PCR amplification. A GTYGGA sequence within the next 21-bp could potentially give rise to tags less than 21-bp long depending upon which site is first recognized by *Mme*I. Analysis of the *Y. pestis* sequence indicates that *Mme*I digestion would at most eliminate only 17 tags from a *Bam*HI library but none from the *Not*I-derived library. While all of the 21-bp *Not*I derived tags are unique, 47 of the *Bam*HI 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 *Y. pestis* GST library using *Bam*HI as the fragmenting enzyme since it was predicted to generate sufficient tags for meaningful data analysis. Sequence analysis of our initial library showed that *Mme*I 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 *Nla*III digestion step, also revealed the presence of a large fraction of tags which originated from *Nla*III sites that were not proximal to a *Bam*HI site. The presence of these tags in the library obviously was the result of incomplete *Nla*III digestion; therefore, we now routinely include a

second *Nla*III 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 Figure1. The cloned inserts in this library were typically several hundred bp to slightly less than one kb long.

The linker we used to biotinylate the *Bam*HI digest adds 12 bp to the ends of each fragment. In principle, the addition of this linker should allow *Mme*I to liberate 21-bp long tags even from the 90 *Bam*HI-*Nla*III fragments predicted from our in silico experiments to be less than 21-bp long. In these cases *Mme*I would have to cleave within the attached linker. Tags from these sites are easy to identify as they should contain a *Bam*HI 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 *Bam*HI fragment number 314, which would be followed by F314 (the next forward GST), R315, F315, etc.

GST analysis

A total of 5,432 GSTs were extracted from the sequenced arrays. The number of 21 and 22-bp long tags was approximately equal, 2,701 and 2,731 respectively. The vast majority, 5,268 (97%), exactly matched at 1,133 sites in the *Y. pestis* genome. This includes a total of 336 tags which 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 *Nla*III site next to a *Bam*HI fragmentation site, which indicates that the two-step *Nla*III digestion was virtually complete. Only 59 (1%) of the extracted tags exactly matched interior *Nla*III sites. These tags could result from over digestion with *Bam*HI or partial *Nla*III digestion; however, we suspect that several may have arisen because subtle changes in the genome introduced new *Bam*HI 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 *Nla*III sites in the *Y. pestis* DNA, is a highly improbable random event. A small number of tags (6) 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.

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 (8 times) in the *Y. pestis* chromosome. It was followed in order by the tag predicted to be the next most frequent, the one occurring 7 times. Only one tag should be present 5 times, one 4 times, three tags should each be found 3 times and seven tags should each occur twice. Two other redundant tags listed in Table 2 should not be recovered at all since each contains a *Bam*HI 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 4 times in the genome seems to be under represented in our database. This tag is associated with an IS100 element that is known to be a source for genetic variability in different *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 *Bam*HI 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 significantly elevated copy number in the strain used here, a prediction that was confirmed by inspection of agarose gel profiles of the total genomic DNA we used for this study (data not shown).

Such deviations in tag frequency or occurrence can also occur when sequence changes introduce or remove a fragmenting site or tagging site. Loss or gain of a single fragmenting site will at most affect the two GSTs flanking the site. Deletions or insertions on the other hand can simultaneously remove or add several tags. Analysis of our data for the absence of adjacent tags revealed several places where deletions must have occurred in the EV766 genome. The most striking example is our failure to recover any of the expected 25 consecutive tags from a segment beginning with F314 and ending with F327 (bp 2,172,627 through 2,254,447 if the 3' position of *Bam*HI site 327 is included). This region contains a 37 kb high-pathogenicity island encoding virulence genes involved in iron acquisition from the host via a siderophore called yersiniabactin (the *ybt* biosynthetic gene cluster) (Buchrieser et al. 1999). It is part of a larger, 100 kb region termed the *pgm* (pigmentation) locus. This locus can delete spontaneously, probably by homologous recombination between its two flanking IS100 elements (Fetherston et al. 1992). Such a deletion would eliminate tags F314-F327; therefore, we propose that strain EV766 lacks the entire *pgm* locus. Similar analysis also identifies a potential deletion of the region bounded by R194- R197 which normally harbors an IS1541 insertion element. Deletions or other changes may have eliminated tags F237-F238, another region associated with an IS100 element. Several other regions not associated with known IS elements that also seem to have been deleted or undergone DNA rearrangements that

eliminate consecutive tags are listed in Table 3. If these 44 tags are excluded, the number of unseen tags drops to 144.

A small fraction of our catalogued tags, totaling 164 (3%), appears to contain 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 mis-ligation between the *Mme*I 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 adaptor, 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 or 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 which 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 approximately 100 to 1000 bp away from their proximal *Bam*HI sites were selected and used for custom primer synthesis. Template *Y. pestis* DNA was digested with *Bam*HI and ligated with a linker cassette which introduced an identical priming site at both ends of each fragment. The DNA was then digested with *Nla*III to physically separate the linkered *Bam*HI ends. Aliquots were then subjected to ten 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 twenty-five 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

While the data we obtained show that we largely achieved our objectives, further analysis (Fig. 4) suggests that we are under sampling 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 *Sph*I site (GCATGC) tethered to a streptavidin bead by a short linker could be cut with *Sph*I but not by *Nla*III even though the linker contained a CATG sequence.

DISCUSSION

We have described a method for obtaining 21-22-bp Genomic Signature Tags 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 utilizes a 16-fold degenerate linker cassette to attach an oligonucleotide adapter to the unknown 3' overhangs of *Mme*I digested DNAs, thereby taking advantage of being able to use cohesive termini for high-efficiency linker addition. Addition of this linker provides not only an appended sequence for PCR amplification but it also attempts to reduce bias during amplification by flanking the monomeric GSTs on both sides with distinct, long linkers. Since the degenerate linker is in molar excess during ligation to the *Mme*I 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 which should suppress their amplification. Other nonstandard 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 21-22-bp sequence tags that can be used to identify the DNA source and, as shown, here 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 *Bam*HI-*Nla*III 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 *Bam*HI-*Nla*III GSTs, respectively. While they share 554 common tags which 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 *Not*I with an 8-base recognition sequence will cleave on average every 4^8 (65.5 kb) bases compared to every 4^6 (4 kb) bases for a restriction enzyme with a 6-base recognition sequence such as *Bam*HI. In practice this means that fragmenting the DNA with *Bam*HI will usually produce 10 xs 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 6 bases that can be used for fragmentation. Some of these enzymes, such as *Pac*I (recognition sequence TTAATTAA) cut only A+T rich DNAs while others cut mainly G+C rich DNAs but are not sensitive to CpG methylation. The ability to select fragmenting enzymes to suit a particular genome's characteristics (A+T or G+C rich, for example) is one strength of the GST method.

In choosing a fragmentation enzyme we prefer to use ones that leave 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 non-laboratory sources where degradation may be a problem. In fact, for GST analysis the starting DNA does not have to be high molecular weight since as shown in Fig. 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.

Since GST analysis is a direct DNA sequencing approach for profiling DNA, perhaps the most exciting extension of the method would be for differential quantitative analyses of DNA in mixed microbial communities. In these communities the frequency of individual tags should approximate the frequency of cognate species abundance. By focusing on the differences in GST

abundance in different libraries, one could begin to identify subsets of tags which vary in abundance because of the community's response to environmental changes. Amplified segments adjacent to these tags could provide direct access to additional genetic information from the source DNA or could be used as probes 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 catalogue 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 (Virlon et al. 1999). The cDNA is then cleaved with *Nla*III leaving the 3' most portion of the cleaved cDNA with the cohesive overhang needed for ligation of the *Mme*I adaptor. All other steps then proceed as outlined in Fig. 1. We have implemented this method to obtain 21-22-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, submitted). 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 *Mme*I 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. While 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 since *Mme*I 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 concatemers 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. 2000). 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 Ca^{2+} 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, SUNY-SB. Ten micrograms was digested with 100 U of *Bam*HI (New England Biolabs, NEB, Tozer, MA), 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-HCl, pH 8.0, 0.1 mM EDTA-Na_3). A biotinylated GATC oligonucleotide adaptor cassette was created by mixing 3600 pmol each of two synthetic oligonucleotides (sense strand: CGA ACC CCT TCG; antisense strand: P-GAT CCG AAG GGG TTC GT-BIOTIN in 100 µl OFA buffer (10 mM Tris-acetate, pH 7.5, 10 mM Mg acetate, 50 mM K acetate, Amersham Bioscience, Piscataway, NJ) heating them to 95°C for 2 min and then allowing them to cool slowly to room temperature. An approximate 50-fold excess of biotinylated cassette (~600 pmol) relative to available *Bam*HI ends was ligated to the fragmented DNA in a total volume of 50 µl of 1X 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 *Nla*III and binding to magnetic beads

The fragmented DNA was next digested with 25 U of *Nla*III (NEB) in 100 µl *Nla*III digestion buffer (1X NEB buffer #4 supplemented with 1xBSA and 10 mM spermidine $(HCl)_3$ for

3 h at 37°C; *Nla*III digestion is stimulated 2 to 4-fold by addition of spermidine, J.J. Dunn, unpubl.). One hundred µl (1 mg) of streptavidin magnetic beads (Dynal Biotech. Inc., Lake Success, NY) were washed twice with 200 µl of 1X magnetic bead binding buffer (MBB: 10 mM Tris-HCl, pH 7.4, 1 mM $EDTANa_3$, 1 M NaCl and then resuspended in 100 µl of 2X MBB. The beads were then added to the *Nla*III digested DNA in a non-stick 1.5 ml microfuge tube (Ambion, Austin, TX). The beads and digest were mixed gently for 1 h at room temperature to bind biotinylated *Bam*HI-*Nla*III fragments.

Second digestion with *Nla*III and *Mme*I adaptor ligation

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

Digestion with *Mme*I

Beads were washed six times with 400 µl 1X MBB and then washed several times with 200 µl *Mme*I digestion buffer (100 mM HEPES, pH 8.0, 25 mM K acetate, pH 8.0, 50 mM Mg acetate, pH 8.0, 20 mM DTT, 4 mM S-adenosylomethionine-HCl). The beads were then resuspended in 100 µl 1X *Mme*I digestion buffer containing 8 U *Mme*I (Center of Technology Transfer, Gdansk, Poland) and incubated for 3 h at 37 °C with occasional mixing. The beads were collected and the supernatant containing the released tags was removed to a clean microfuge tube. The beads were washed with 100 µl TEsl and the wash combined with the first *Mme*I supernatant. The pooled *Mme*I digest is extracted with an equal volume of P/C and precipitated at -80°C for 1-2 h with 1 ml of ethanol after addition of 133 µl 7.5 M ammonium acetate and 2 µl Glyco blue(Ambion) as carrier. The resulting pellet was washed with cold 75% ethanol, dried *in vacuo* and resuspended in 29.5 µl TEsl plus 4 µl 10X 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: P-TTC ATG GCG GAG ACG TCC GCC ACT AGT GTC GCA ACT GAC TA-AmMC7; antisense strand: TAG TCA GTT GCG ACA CTA GTG GCG GAC GTC TCC GCC ATG AAN N). Thirty-five pmol of adaptor cassette (3.5 µl) was added to the resuspended tags and after 15 min at room temperature 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 consisting of an initial denaturation step at 95 °C for 2 min followed by 30 cycles of 95 °C for 30 s, 58 °C for 30 s and 72 °C for 30 s with a final extension step at 72 °C for 4 m using 5'-Biotin-GGA TTT GCT GGT CGA GTA CA and 5'-Biotin-TAG TCA GTT GCG

ACA CTA GTG GC as forward and reverse primers, respectively, each at a final concentration of 0.4 µM. Cycling was performed in 1X Promega buffer containing 2 mM Mg sulfate and 0.3 mM of each dNTP. Typically 0.8-1.0 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 (LARHD)

The PCR products were then subjected to one round of linear amplification to reduce heteroduplexes (LARHD) by diluting them to 1 ml with 800 µl 1X PCR buffer containing 4 µl Platinum Taq and an additional 400 pmol of each biotinylated primer. The reaction was then incubated at 95 °C for 2.5 m, 58 °C for 30s and 72 °C for 5 m. Unincorporated primers were digested by addition of 10 µl (200 U) of single-strand specific *E. coli* Exo I. 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), ***Nla*III digestion and concatemerization**.

Following 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 under 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 additional 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 Exo I as above. After P/C extraction and ethanol precipitation the amplified DNA is digested with 20 U of *Nla*III in 400 µl

at 37 °C for 4 h (after 2h 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 h and then centrifuged at 4 °C. The pellets are resuspended in 200 µl ice cold TEsl plus 25 mM NaCl, diluted with an equal volume of 2X MBB and added to 200 µl (2 mg) of streptavidin beads equilibrated with 1X MBB. After gentle mixing for 15 m at room temperature, the unbound fraction is transferred to a second 200 µl aliquot of beads to capture any remaining biotinylated DNA fragments. The unbound GST fraction is recovered and precipitated by addition of 2.5 volume of ethanol and Glyco bluecarrier and concatemerized with T4 DNA ligase (5 U/µl, Invitrogen) at 16 °C for 4 hr . The sample is subjected to electrophoresis on a 0.75% low melt agarose gel and products greater than 100 bp are recovered. These products are cloned into the *Sph*I-site of a pZero plasmid (Invitrogen) that was engineered to have a *Sph*I-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 ligated with the *Sph*I digested vector at a high DNA concentration, 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 electroporation of competent TOP10 cells (Invitrogen, Carlsbad, CA) are selected on 2X YT plates containing 50 µg/ml kanamycin. A schematic representation of the method is shown in Figure1 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 BioSystems 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 sequencers. Extracted data were ported to an Oracle® database and searched for valid tags using the GST software we developed. The software ensures that only unambiguous 21-22-bp tag sequences, see below, are extracted for further analysis (tags with Ns, lengths other than 21-22 bases or whose polarity 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 GGG TGC ACG ACC CGA (205R); [2,281,342] CAT GTG GCC GCC GCG CTT AA (384R); [2,894,318] CAT GAC TCT GCC ATA GCT TCG (1031R); [3,452,611] CAT GCA GGA CCG CGG ACA ATG (102F); and [4,145,945] CAT GCA GTG CCA TCC TCA CGG (230F). The values in brackets are the position of the *Nla*III tagging site in the *Y. pestis* chromosome. The values in parentheses are the distances between the respective *NlaI*II and *Bam*HI sites and the directionality of the PCR reaction. *Bam*HI digested *Y. pestis* DNA was ligated with a non-biotinylated GATC oligonucleotide adaptor created by mixing and annealing 3600 pmol each of two synthetic oligonucleotides (sense strand: CGT AAT ACG ACT CAC TAT AGG GA: antisense strand: GCA TTA TGC TGA GTT ATA TCC CTC TAG in 100 µl OFA as described above. The annealed GATC adaptor (40 pmol) was ligated to *Bam*HI fragmented DNA for 2 h at 16°C in a total volume of 50 µl of 1X ligase buffer (Takara) containing 350 U of T4 DNA ligase (Takara). Aliquots of the linkered DNA were

incubated at 94 °C for 2 min, followed by 10 rounds of linear amplification (94 °C for 20 sec, 55 °C for 30s and 68 °C for 4 min) 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 m at 68 °C and analyzed on 6% polyacrylamide gels. Extension with the sense strand primer should add an additional 23 bp to the *Bam*HI end of all the amplification products.

ACKNOWLEDGMENTS

We thank Willy Crockett, Kimberly Pellechi, Jutta Paparelli and Judi Romeo for assistance in DNA sequencing and Keith Thompson for statistical analysis. This project is supported by a Laboratory Directed Research and Development award (to JJD) and by the Offices of Biological and Environmental Research, and of Basic Energy Sciences (Division of Energy Biosciences) of the U.S. Department of Energy.

REFERENCES

Boyd, A. C., Charles, I. G., Keyte, J. W., and Brammar, W. J. 1986. Isolation and computer-aided characterization of *Mme*I, a type II restriction endonuclease from *Methylophilus methylotrophus*. Nucleic Acids Res. **14:** 5255-74.

Buchrieser, C., Rusniok, C., Frangeul, L., Couve, E., Billault, A., Kunst, F., Carniel, E., and Glaser, P. 1999. The 102-kilobase *pgm* locus of *Yersinia pestis*: sequence analysis and comparison of selected regions among different *Yersinia pestis* and *Yersinia pseudotuberculosis* strains. Infect. Immun. **67:** 4851-61.

Damak, S., and Bullock, D. W. 1993. A simple two-step method for efficient blunt-end ligation of DNA fragments. Biotechniques **15:** 448-50, 452.

Donson, J., Fang, Y., Espiritu-Santo, G., Xing, W., Salazar, A., Miyamoto, S., Armendarez, V., and Volkmuth, W. 2002. Comprehensive gene expression analysis by transcript profiling. Plant Mol. Biol. **48**:75-97.

Fetherston, J. D., Schuetze, P., and Perry, R. D. 1994. Loss of the pigmentation phenotype in *Yersinia pestis* is due to the spontaneous deletion of 102 kb of chromosomal DNA which is flanked by a repetitive element. Mol. Microbiol. **6:** 2693-704.

Hanke, M., and Wink, M. 1994. Direct DNA sequencing of PCR-amplified vector inserts following enzymatic degradation of primer and dNTPs. Biotechniques **17:** 858-60.

Kozdrój, J., and van Elsas, J.D. 2001. Structural diversity of microorganisms in chemically perturbed soil assessed by molecular and cytochemical approaches. J. Microbiol. Methods 43:197-212.

Kenzelmann, M, and Muhlemann, K. 1999. Substantially enhanced cloning efficiency of SAGE (Serial Analysis of Gene Expression) by adding a heating step to the original protocol. *Nucleic Acids Res.* **27**:917-8.

Motin, V. L., Georgescu, A. M., Elliott, J. M., Hu, P., Worsham, P. L., Ott, L. L., Slezak, T. R., Sokhansanj, B. A., Regala, W. M., Brubaker, R. R., and Garcia, E. 2002. Genetic variability of *Yersinia pestis* isolates as predicted by PCR-based IS100 genotyping and analysis of structural genes encoding glycerol-3-phosphate dehydrogenase glpD. J. Bacteriol. **184:** 1019-27.

Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B., Sebaihia, M., James, K. D., Churcher, C., Mungall, K. L., et al. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature **413:** 523-7.

Portnoy, D. A., and Falkow, S. 1981. Virulence-associated plasmids from *Yersinia enterocolitica* and *Yersinia pestis*. J. Bacteriol. **148:** 877-83.

Powell, J. 1998. Enhanced concatemer cloning- a modification to the SAGE Serial Analysis of Gene Expression technique. Nucleic Acids Res. **26:** 3445-6.

Qin, L., Prins, P., Jones, J.T., Popeijus, H., Smant, G., Bakker, J., and Helder, J. 2001. GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP. Nucleic Acids Res. **29**:1616-22 .

Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., Loiacono, K.A., Lynch, B.A., MacNeil, I.A., Minor, C., Tiong, C.L., Gilman, M., Osburne, M.S., Clardy, J., Handelsman, J., and Goodman, R.M. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. **66**:2541-7.

Rouillard, J. M., Erson, A. E., Kuick, R., Asakawa, J., Wimmer, K., Muleris, M., Petty, E. M., and Hanash, S. 2001. Virtual genome scan: a tool for restriction landmark-based scanning of the human genome. Genome Res. **11:** 1453-9.

Ryo, A., Kondoh, N., Wakatsuki, T., Hada, A., Yamamoto, N., and Yamamoto, M. 1998. A method for analyzing the qualitative and quantitative aspects of gene expression: A transcriptional profile revealed for HeLa cells. Nucleic Acids Res. **26**, 2586-92.

Ryo, A., Kondoh, N., Wakatsuki, T., Hada, A., Yamamoto, N., and Yamamoto, M. 2000. A modified serial analysis of gene expression that generates longer sequence tags by nonpalindromic cohesive linker ligation. Anal. Biochem. **277:** 160-2.

Schloter, M., Lebuhn, M., Heulin, T., and Hartmann, A. 2000. Ecology and evolution of bacterial microdiversity. FEMS Microbiol. Rev. **24**:647-600.

Saha. S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, KW., and Velculescu, V.E. 2002. Using the transcriptome to annotate the genome. Nat. Biotechnol. **20**:508-12

Spinella, D. G., Bernardino, A. K., Redding, A. C., Koutz, P., Wei, Y., Pratt, E. K., Myers, K. K., Chappell, G., Gerken, S., and McConnell, S. J. 1999. Tandem arrayed ligation of expressed sequence tags TALEST: a new method for generating global gene expression profiles. Nucleic Acids Res. **27:** e22.

Torsvik, V., and Øvreås, L. 2002. Microbial diversity and function in soil: from genes to ecosystems. Curr. Opin. Microbiol. 5:240-5.

Tucholski, J., Skowron, P. M., and Podhajska, A. J. 1995. *Mme*I, a class-IIS restriction endonuclease: purification and characterization. Gene **157:** 87-92.

Velculescu, V. E. 2001. Using SAGE to explore the genome. In proceedings from *SAGE 2001: Frontiers in transcriptome exploration*. San Diego CA. p. 15.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. 1995. Serial analysis of gene expression. Science **270:** 484-7.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B., and Kinzler, K. W. 1997. Characterization of the yeast transcriptome. Cell **88:** 243-51.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frifiters, A., Pot, J., Peleman, J., Kuiper, M., and Xabeau, M. 1995. AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res. 23: 4407-14.

Virlon, B., Cheval, L., Buhler, J. M., Billon, E., Doucet, A., and Elalouf, J. M. 1999. Serial microanalysis of renal transcriptomes. Proc. Natl. Acad. Sci. **96:** 15286-91.

Wang, S.M, and Rowley, J.D. 1998. A strategy for genome-wide analysis: Integrated procedure for gene identification. Proc. Natl. Acad. Sci. **95:**11909-94.

Welle, S., Bhatt, K., and Thornton, C.A. 1999. Inventory of high-abundance mRNAs in skeletal muscle of normal men. Genome Res. 9:506-13.

Wimmer, K., Zhu Xx, X. X., Rouillard, J. M., Ambros, P. F., Lamb, B. J., Kuick, R., Eckart, M., Weinhausl, A., Fonatsch, C., and Hanash, S. M. 2002. Combined restriction landmark genomic scanning and virtual genome scans identify a novel human homeobox gene, ALX3, that is hypermethylated in neuroblastoma. Genes Chromosomes Cancer **33:** 285-94.

Yu, J., Zhang, L., Hwang, P. M., Rago, C., Kinzler, K. W., and Vogelstein, B. 1999. Identification and classification of p53-regulated genes. Proc. Natl. Acad. Sci. **96:** 14517-22.

Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. 1997. Gene expression profiles in normal and cancer cells. Science **276:** 1268-72.

Figure Legends:

Fig. 1. Schematic for GST preparation. In this method DNA is first fragmented with a rare cutter such as *Not*I or a more frequent cutter such as *Bam*HI. Specific complementary biotinylated linkers are ligated to the free ends and the DNA is then digested with *Nla*III. All subsequent steps in the protocol are identical.

Fig. 2. (A) PCR amplification of GSTs. Samples were electrophoresed on a 10% polyacrylamide gel to show the size distributions of the products after first round amplification. Lane 1 contains a 100 bp ladder; lanes 2-7 contain GST samples after: PCR amplification, LARHD, LARHD+ExoI digestion, LARHD2, LARHD2+ExoI digestion, and *Nla*III digestion, respectively. (B) Concatemer separation. Ligated tags (lane 2) were heated to dissociate aggregates (Kenzelmann and Muhlemann 1999) and then sized by agarose gel electrophoresis.

Fig. 3. Specific amplification of end sequences corresponding to a specific GST in the *Y. 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 1,031 +23= 1,054 bp, lanes 2-6, respectively. Lane 1 contains a 100 bp ladder.

Fig. 4. Length Distribution of *Y. pestis Bam*HI-*Nla*III 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.

Table 1 Predicted GSTs for *Y. pestis* EV766

	*Not*I fragmentation [64 fragments]		*Bam*HI fragmentation [699 fragments]	
Tags of length ≥21	start [a]	after *Mme*I digestion	start	after *Mme*I digestion
predicted tags	115 (7)	115 (7)	1236 (96)	1214 (93)
unique tags	115 (7)	115 (7)	1203 (94)	1181 (91)
single tags	115 (7)	115 (7)	1189 (92)	1167 (89)
multiple tags	0	0	14 (2)	14 (2)
Tags of length ≤20				
predicted tags	7 (0)	7 (0)	89 (12)	89 (12)
unique tags	7 (0)	7 (0)	86 (12)	86 (12)
single tags	7 (0)	7 (0)	84 (12)	84 (12)
multiple tags	0	0	2 (0)	0
zero length tags[b]	4	4		1
SUM	126 (7)	126 (7)	1326 (108)	1303 (105)

(a) Values in parenthesis are the numbers of tags with ambiguous directions, *i.e.*, they begin with the sequence CATGAA.

(b) Zero length tags occur when the fragmenting site is immediately adjacent to a *Nla*III site

Table 2. Correspondence Between Predicted and Actual IDentifier Tag Frequencies

IDentifier tag sequence[(a)]	FREQUENCY predicted	actual	
ATCTGGAGGTTCGGTTC	8	65	
CGTCATCTCGCTGAACG	7	45	
GATGTATTTACGGCGTC	5	34	
CCCTGCGGTACGGGAGC	3	34	
GCTGCATTGGCACCGTT	2	23	
CCAGCATCAGCCAGCGC	2	22	
TAGGCTCGAGCCGCGCC	3	20	
TCGTTCAAATCAAAGGA	4	13	
CTGATAAACCGGGATCG	2	13	
AATCCTCACCTAACCGA	2	12	
CTTTCGTTGGTTAGCGA	3	11	
CCCCAGCCCTGGCCCGC	2	11	
AACCGCGTATCAATCAG	2	11	
TGCGTTTTCAGGACGGT	2	9	
TT**GGATCC**GAAGGGGTT	3		unseen-contains ***Bam*HI** site
G**GGATCC**GAAGGGGTTC	2		unseen-contains ***Bam*HI** site

Complete lists of GSTs, in both order of abundance and position in the *Y. pestis* genome, are available via the internet at http://genome.bnl.gov/GSTs.

(a) CATG omitted

Table 3. Potential Deletions in the *Y. pestis* EV766 genome

Start-End	Position bp	IS Element	No. of tags affected
F314-F327	2,172,627 - 2,254,447	yes IS100	25
R194-R197	1,307,243 - 1,316,087	yes IS1541	7
F227-F228	1,554,643 - 1,556,368	no	3
F237-F238	1,618,033 - 1,652,133	yes IS100	3
F381-F382	2,662,263 - 2,685,036	no	3
F453-F454	3,069,009 - 3,122,266	no	3
		Total	44

Table 4. Shared GSTs Between Two Different Bacteria [(a)]

GST sequence[(b)]	organisms	total	organism (count)	organism (count)
GCCGCTTAACCGCCGCA	2	4	*Escherichia coli (3)*	*Yersinia pestis (1)*
GATCGCCGATCGTCCCG	2	3	*Mycobacterium leprae (1)*	*Mycobacterium tuberculosis (2)*
GCAACGATATTGGTGAC	2	3	*Mycobacterium leprae (1)*	*Mycobacterium tuberculosis (2)*
CCGCCCCGGAAATCACC	2	3	*Mycobacterium leprae (1)*	*Mycobacterium tuberculosis (2)*
GACCTGTCCACCGGCAA	2	3	*Mycobacterium leprae (1)*	*Mycobacterium tuberculosis (2)*
GGCTGTGGGTGGCGTTC	2	3	*Mycobacterium leprae (1)*	*Mycobacterium tuberculosis (2)*
CTTGGCCGCTACACCAC	2	3	*Pyrococcus abyssi (1)*	*Pyrococcus horikoshii (2)*
CTCCGCCGCTTGTGCGG	2	3	*Mycobacterium leprae (1)*	*Mycobacterium tuberculosis (2)*
GTGGATGCCTTGGCATC	2	3	*Mycobacterium leprae (1)*	*Mycobacterium tuberculosis (2)*
GCGACCCAGGAACAGCA	2	3	*Mycobacterium leprae (1)*	*Mycobacterium tuberculosis (2)*
GGAGTCGATGTTATCGG	2	3	*Mycobacterium leprae (1)*	*Mycobacterium tuberculosis (2)*
AAGCCGGTCGCCATCAT	2	2	*Mesorhizobium loti (1)*	*Sinorhizobiumme liloti (1)*
GTGACTTCTGCGGATGT	2	2	*Chlamydia muridarum (1)*	*Chlamydia trachomatis (1)*
TGCACCGGAATGCGGAT	2	2	*Mesorhizobium loti (1)*	*Sinorhizobiumme liloti (1)*
CACCACCTCTCCTTCTA	2	2	*Thermoplasma acidophilum (1)*	*Thermoplasma volcanium (1)*
TCGGACAGAACCTTGCG	2	2	*Agrobacterium tumefaciens (1)*	*Sinorhizobium meliloti (1)*
ACGCCGAAGGTGATGGC	2	2	*Mesorhizobiumloti (1)*	*Sinorhizobiumme liloti (1)*
AACGAAGATCAATTTCC	2	2	*Chlamydia muridarum (1)*	*Chlamydia trachomatis(1)*
AATTAGAAAATTATGAC	2	2	*Haemophilus influenzae (1)*	*Pasteurella multocida (1)*
CGGACTTCGGTCGGCTT	2	2	*Mesorhizobiumloti (1)*	*Sinorhizobium meliloti (1)*
CTCTCAACGTAGGGAAC	2	2	*Pyrococcusabyssi (1)*	*Pyrococcushorikoshii (1)*
CCCATCACTATCAAGCC	2	2	*Chlamydiamuridarum (1)*	*Chlamydia trachomatis (1)*
AGCAGGTTGAAGGTTGA	2	2	*Mycoplasmagenitalium (1)*	*Mycoplasma pneumoniae (1)*
ATGCGCAAGTGCCATCT	2	2	*Agrobacterium tumefaciens (1)*	*Sinorhizobium meliloti (1)*
CAGGTCGGCATTTAACC	2	2	*Pyrococcus abyssi (1)*	*Pyrococcus horikoshii (1)*
AAGGTTCAACGTGGGTC	2	2	*Thermoplasma acidophilum (1)*	*Thermoplasma volcanium(1)*
CGGGGAAACGTAGTAGC	2	2	*Chlamydia muridarum (1)*	*Chlamydia trachomatis(1)*
CACAAGATCCAGGACCG	2	2	*Mesorhizobium loti (1)*	*Sinorhizobium meliloti (1)*
AGCTAACCCCATTTTGT	2	2	*Chlamydiam uridarum (1)*	*Chlamydia trachomatis (1)*
CAGCACTCCATATTTTA	2	2	*Clostridium acetobutylicum (1)*	*Pyrococcus horikoshii (1)*

*(a)*GSTs within 25 bp of the BamHI fragmentation site were omitted; *(b)* CATG omitted

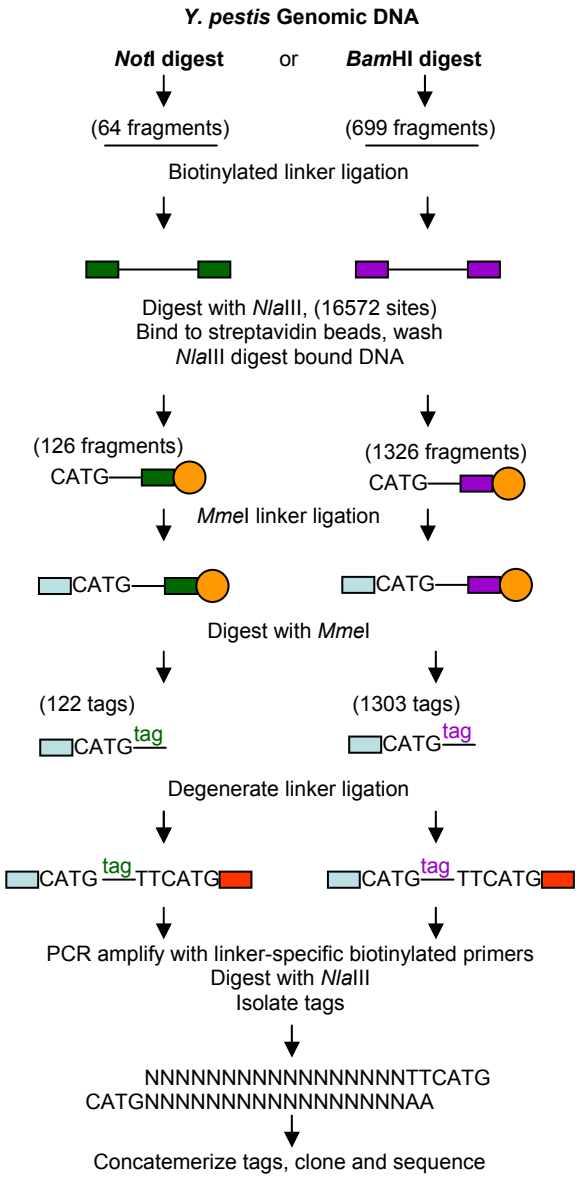
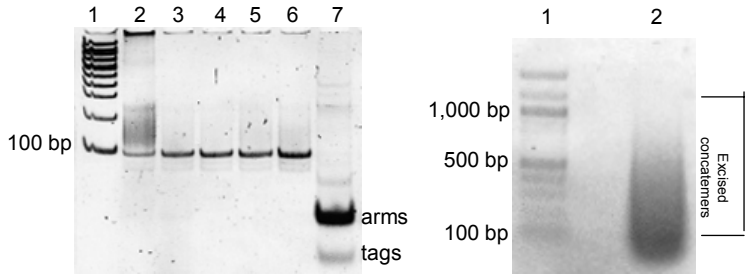


Figure 1

Dunn et al.



2A 2B

Figure 2A and 2B

Dunn et al.

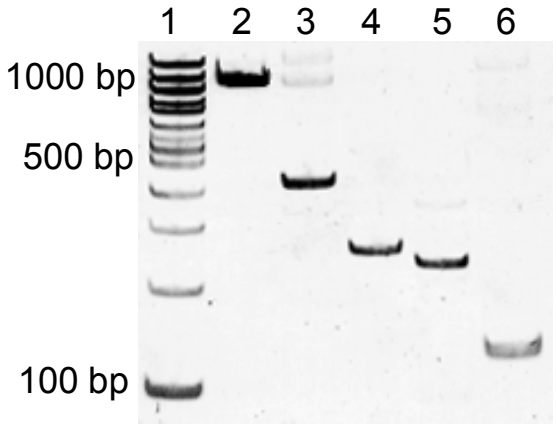


Figure 3

Dunn et al.

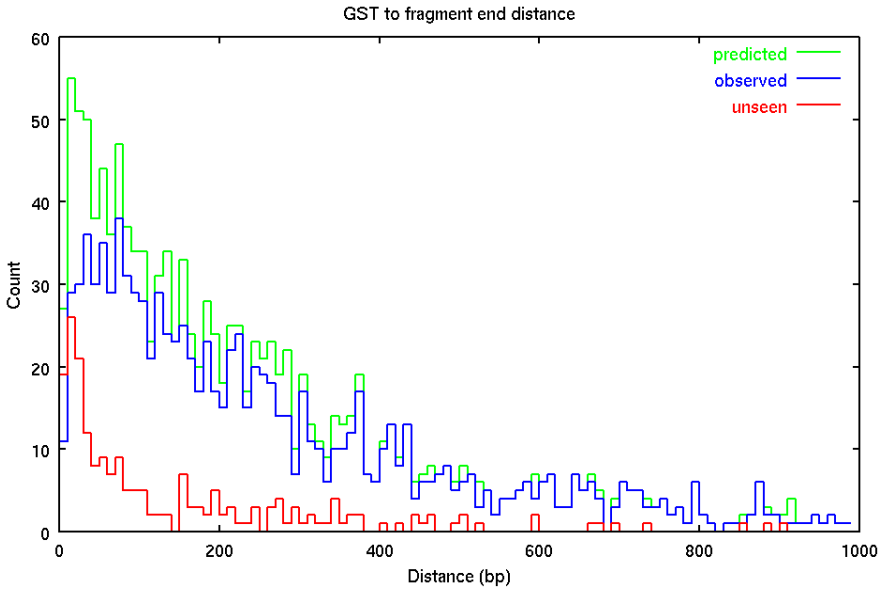

GST to fragment end distance
predicted
observed
unseen
Count
Distance (bp)
0
10
20
30
40
50
60
0
200
400
600
800
1000

Figure 4