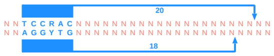
Restriction enzyme *Mme* I

Type IIs. A wonderful overhang cutter.

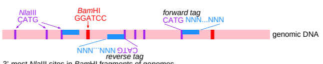


By ligating oligonucleotide cassettes containing the *Mme*I recognition sequence to DNA that has been cut with another restriction enzyme, called the **anchoring** enzyme (*Nla*III and *Mse*I are just two possible choices), tags of length roughly 20 bp can be cut off, collected, and sequenced in large quantities.

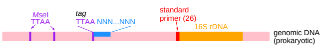
20 bp allows 4^{20}, or 1.2×10^{12} possible combinations. A single tag from a microbe is almost sufficient to identify family, or even species. By obtaining tags from specific locations relative to other genome features, even more information is added. Several tags collected from an organism can identify species, in many cases.

Here are four schemes for locating tags. Position reference features are shown in red, anchor enzyme sites are show in purple, tags are in blue.

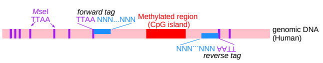
3'-most *Nla*III site in mRNA transcript (SAGE; Velculescu, p15, *Proceedings from SAGE 2001: Frontiers in Transcriptome Exploration,* San Diego, 2001)



3'-most *Nla*III sites in *Bam*HI fragments of genomes (Dunn, et. al, *Genome Research* 12, p1756, 2002)

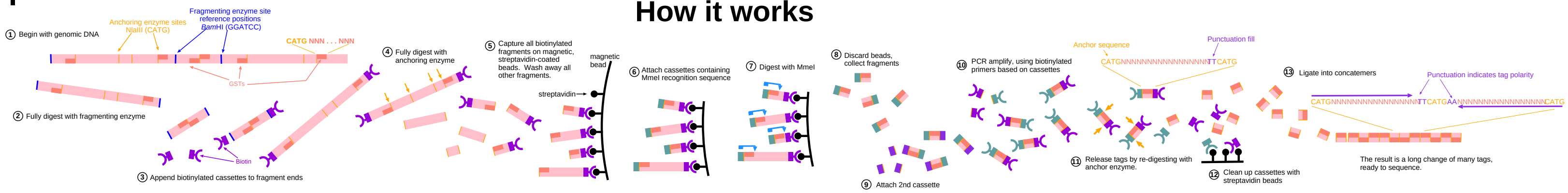


*Mse*I sites immediately upstream (3') from 16S rDNA standard primer (trials currently underway)



*Mse*I sites immediately upstream (3') from heavily methylated regions (CpG islands).

How it works



Profiling Genomes with Genomic Signature Tags (GSTs)

A powerful new technique that can

- **Rapidly identify microbes, even in complex communities**
- **Detect novel organisms**
- **Determine abundance levels**
- **Allow quick development of microarray probe sets**

Sean McCorkle
John Dunn
Daniel van der Lelie
http://genome.bnl.gov/GSTs

Brookhaven National Laboratory

Proof of principle: GSTs of a known genome

Genome Sequence Tags (GSTs) were collected from an avirulent strain of *Yersinia pestis* (EV766), for the purpose of comparing with the published genome sequence. The project is detailed in Dunn, et. al., *Genome Research* 12, p1756, 2002. GSTs were defined at 3'-most *Nla*III (CATG) anchor sites in *Bam*HI fragments.

Here's an example of a concatemer sequence that was collected:

punctuation — anchor enzyme sequence

```
CATGGTGGGCCGCGCCCCTATTTCATGAAATCGCGAATTCCAAAGCCATG
AATCGCAACATCATCATTGCCATGAAGGATGTCATTGATATCGGCATGAA
TGGTTAATCAAGATGCCCATGAAAATAAGGTCTGACGTCCCATGAAGGGG
CGTTGGCAGGATGGCATGAAGCCCCCAAGGCTTTTCGCCATGTTCCGTCC
CATATTGCACTTCATGCCAAAGAAGCTGGCTCTTTTCATGTTTTGCGGCA
GGCTAAATTCATGAAAGCAATACATCTGGTTCCATGAAGAACCGAACCTC
CAGATCATGAACAATCAGTACATCGCTGTCATGCCTCCCAGGAAGCCTGT
TTCATGAATCTGGCTGTCGCCCAATTCATGACTGTAACCAATTTTTCTTC
ATGGTCGCGCGCATACCTGCTTCATGAAGCCCACATCCTGTCGCTACATG
```

Concatemer sequences were split into tags and processed with relatively simple perl scripts, shown here in boxes. **split-tags** splits concatemer sequences at anchor sites, determines polarity from punctuation, and filters out tags which are too short or too long.

Three possible polarity punctuation conditions must be considered:

GTGGGCCGCGCCCCTATTT
Forward. tag is: CATG + GTGGGCCGCGCCCCTAT

AATGGTTAATCAAGATGCC
Reverse. tag is: CATG + GGCATCTTGATTAACCA

AACAGAGGATCGCCAGTATT
Ambiguous (Tag begins with AA in genome) could be
CATG + AACAGAGGATCGCCAGTA
or
CATG + AATACTGGCGATCCTCTG

split-tags

The output from **split-tags** is a tabular stream of tag sequences, directions and positions. The common CATG prefix in all tags is not shown. Tag lengths are evenly split between CATG +17 and +18; the variation is a product of the *Mme*I enzyme

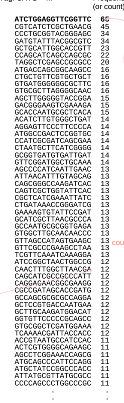
.	.	.
.	.	.
.	.	.
GTGGGCCGCGCCCCTAT	f	4
GCTTTGGAATTCGCGAT	r	27
GCAATGATGATGTTGCGA	r	50
CCGATATCAATGACATCC	r	74
GGCATCTTGATTAACCA	r	98
GGACGTCAGACCTTATT	r	121
CCATCCTGCCAACGCCCC	r	144
GCGAAAAGCCTTGGGGGC	r	168
TTCCGTCCCATATTGCAC	f	192
CCAAAGAAGCTGGCTCTT	f	216
TTTTGCGGCAGGCTAAA	f	240
GAACCAGATGTATTGCT	r	263
ATCTGGAGGTTCGGTTC	r	286
ACAGCGATGTACTGATTG	r	309
CCTCCCAGGAAGCCTGT	f	333
AATCTGGCTGTCGCCCAATT	**a**	**356**
ACTGTAACCAATTTTTC	f	380
GTCGCGCGCATACCTGC	f	403
TAGCGACAGGATGTGGGC	r	426
.	.	.
.	.	.
.	.	.

Ambiguous tags are filtered out for rejection or seperate processing

```
clust-tags

#!/usr/bin/perl


while ( <> )
   {
    /^([A-Z]{$ext,})\s+([abfr])\s/ || die "bad line $_";
    ($t,$dir) = ($1,$2);
    next unless ( $dir eq 'f' || $dir eq 'r' );
    $tag = substr( uc( $t ), 0, $ext );
    $count{$tag}++;
   }

foreach $t ( sort by_count keys( %count ) )
   {
    printf $format, $t, $count{$t};
   }

sub by_count
   {
    if ( $count{$a} == $count{$b} )
       { return( $a cmp $b ); }
    else
       { return( $count{$b} <=> $count{$a} ); }
   }
```

clust-tags truncates tags to CATG + 17 and tabulates tag frequencies. The output is a simple table of tag and frequency (or count), sorted by decending frequencies

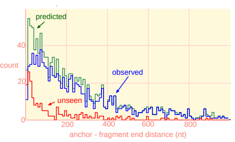
Tag: CATG + ... Frequency (or count)



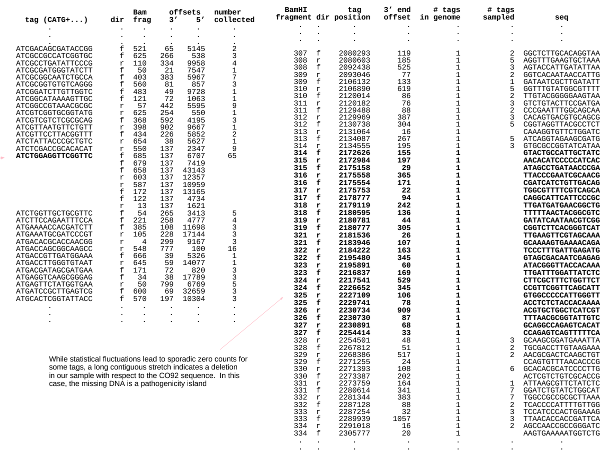
5432 GSTs were collected in total, from concatemers. These were then compared with a small database of 1326 GSTs that was generated from the *Y. pestis* genome sequence (CO92). 97% exactly matched 1133 GST sites in the genome. The remainder were almost all matched imperfectly (sequencing errors) or matched other NlaIII sites.

This tag has a high frequency because it occurs in a repeat element in the genome.

209 tags present in the database were not observedin concatemers. While this is partly due to the non-oversampled statistics, a histogram of anchor-fragment end distances for observed and unseen tags indicates a distinct bias towards the short end. A longer biotinylated cassette should rectify this problem.



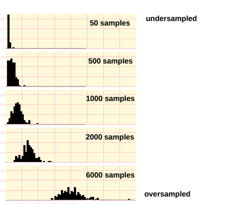
Nearly all tags matched the Genome or could be otherwise accounted for.



While statistical fluctuations lead to sporadic zero counts for some tags, a long contiguous stretch indicates a deletion in our sample with respect to the CO92 sequence. In this case, the missing DNA is a pathogenicity island

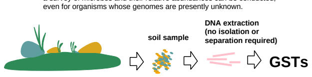
GST sampling is democractic. Tag frequencies follow a binomial distribution. For low sample numbers, the distribution is approximately Poisson. As the number of samples grows, the distribution becomes normal (Gaussian).

Shown here is a series of Monte Carlo samplings of *Y. pestis Bam*HI fragmement defined *Nla*III GSTs, taken from the genome.



Surveying microbial communities and discovering new organisms

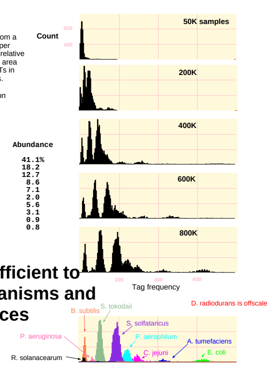
By oversampling GSTs from a community (even a complex one), a survey of microbes and their relative abundances can be conducted, even for organisms whose genomes are presently unknown.



Tag frequency histograms of oversampled GSTs from a mixture of genomes yield peaks of tags, one peak per organism. The means of the peaks are simply the relative abundances of the organisms in the mix. The total area under each peak is a product of the number of GSTs in the organisms genome and the number of samples.

This series of histograms for Monte Carlo simulation of drawing GSTs (SpeI fragment, NlaIII anchor) demonstrates this effect. The organisms and abundances used are:

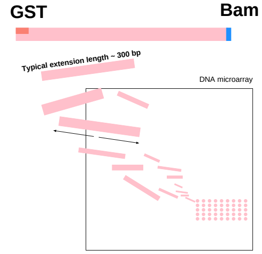
Organism	# GSTs	Abundance
Deinococcus radiodurans	42	41.1%
Escherichia coli K12	140	18.2
Agrobacterium tumefaciens C58	96	12.7
Campylobacter jejuni	262	8.6
Pyrobaculum aerophilum	452	7.1
Pseudomonas aeruginosa	74	2.0
Sulfolobus solfataricus	1518	5.6
Sulfolobus tokodaii	1156	3.1
Bacillus subtilis	308	0.9
Ralstonia solanacearum	72	0.8



800K samples is sufficient to isolate 9 out 10 organisms and determine abundances

Making libraries for new microarray probes

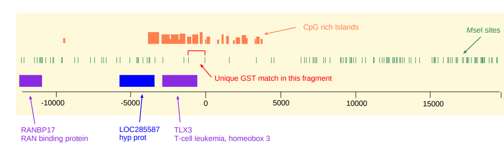
Armed with groups of tags identified with organisms based on peak clustering, the GST tag can serve as a PCR priming site, along with cassettes ligated onto the fragment enzyme site. Depending on enyme choices, these PCR products can fall comfortablng in the 50 to 500 bp size range. This provides a quick entry point for detailed study of novel organisms, and also rapidly generates large amounts of sequence for making new microarrays to test for the new organism(s).



Finding heavily methylated regions in Human DNA

Methylation-regulation irregularities may play a role in cancer development. By running *Mse*I fragments from cancerous cell DNA through a methylation column and then collecting GSTs from the fragment ends, its is possible to locate the fragments to unique positions in chromosome sequences.

There are roughly 40 million *Mse*I sites in the Human Genome. A test protocol conducted on pancreatic cancer cell DNA has yielded over a hundred GSTS, the bulk matching unique sites in the Human Genome. This diagram shows the neighborhood of one such match in chromsome 5. The fragment matched contains a very CpG rich island, believe to play an important role in regulation in cancer cells.



This poster prepared using TGIF and gnuplot.