PAIRED-END GENOMIC SIGNATURE TAGS: A METHOD FOR THE FUNCTIONAL ANALYSIS OF GENOMES AND EPIGENOMES

John J. Dunn, Sean R. McCorkle, Logan Everett and Carl W. Anderson

Biology Department
Brookhaven National Laboratory
Upton, NY 11973-5000

INTRODUCTION

Interactions between eukaryotic transcription factors and their cognate DNA binding sites form fundamental networks within cells that control critical steps during development and tissues-specific gene expression. These interactions also are important in regulating cellular responses to stresses, and their dysfunction contributes to numerous diseases. Therefore, determining the *in vivo* genome-wide binding distribution of transcription factors is an important step towards developing an understanding of the regulatory networks in a living cell as well as their changes in response to specific stimuli. Methods based on chromatin immunoprecipitation (ChIP) are beginning to provide an increasing detailed view of these dynamic events. This assay was originally developed to monitor histone modifications and then modified to detect binding of specific transcription factors to native chromatin (1-3). In this method, transcription factors are reversibly cross-linked to their binding sites using formaldehyde to freeze intracellular protein-DNA complexes, the DNA is sonicated to generate fragments with lengths

Genetic Engineering, Volume 28, Edited by J. K. Setlow

of ~500 to ~2,000 bp, and individual transcription factor-DNA complexes are immunoprecipitated using specific antibodies to the native protein or to a suitable epitope fused in-frame to the target protein's coding sequence. The DNA fragments enriched by ChIP can be identified by a variety of means such as cloning or amplification with gene specific primers, hybridization to microarrays containing subsets of the genomic sequences or by Serial Analysis of Gene Expression (SAGE)-type approaches (4) that extract short sequence identifier tags or Genomic Signature Tags (GSTs) from the ChIP DNA and then use this information to map the DNA back to the genome. In this article we will review the basic steps for generating GSTs, their application to analysis of ChIP data and will introduce several modifications of our original SACO *(for* **S***erial* **A***nalysis of* **C***hromatin* **O**ccupancy) method (5) that can simultaneously generate tags from both ends of the ChIP fragments and preserve their spatial relationship to each other. The sequences of each tag in combination uniquely identify the region of the genome from which the original SACO fragment was derived and encompass the sequence of the site to which the transcription factor was bound.

This same approach can also be used to obtain paired sequence tags from the ends of any DNA fragment. At the whole genome level any changes in the resulting paired-end profile can provide a sensitive method for distinguishing between closely-related genomes or genomes that have undergone deletions, insertions or other rearrangements that cause the appearance of new diTAG pairs. Detection and characterization of discrepancies between observed diTAG pairs from reference and test genomes can, in principle, detect structural variations with the same precision as afforded by paired-end sequencing of fosmid or bacterial-artificial chromosome libraries (6). Such changes are characteristic of many cancers as are changes in CpG methylation in CpG islands, which are clusters of CpG dinucleotides that are found in front of about half of human genes (7). Methylation of cytosine within these islands caused inhibition of downstream gene expression, and aberrant methylation is an important mechanism for gene activation or inactivation in cancer. In this article, we briefly review how paired-end diTAGs can be obtained from DNA fragments associated with methylated CpG islands.

WHAT ARE GENOMIC SIGNATURE TAGS?

Genomic Signature Tags (GSTs) are the products of a method we developed for identifying and qualitatively analyzing genomic DNAs (8). Two major principles underlie this method: first, short DNA sequences (18-21 bp) are sufficient to identify unique sites within a genome; second, concatenation of these short DNA sequences, as in SAGE (9), greatly increases sequence throughput. The original GST method begins with cutting the DNA sample with a type II restriction enzyme, also termed the fragmenting enzyme, to produce fragments with cohesive ends. After digestion with the first enzyme, the cohesive ends are biotinylated, and the sample is digested with *Nla* III, also called the anchoring enzyme, which cleaves leaving 4 base cohesive ends. Since *Nla* III has a 4 bp recognition sequence (CATG), it theoretically cleaves on average every 256 bp, and nearly every fragment in the original digest will be cleaved at least once to

produce two biotinylated end fragments which are recovered by binding to streptavidin-coated magnetic beads. The bound DNA fragments are then ligated with a linker cassette that creates partially overlapping *Mme* I (TCCRA**C**) and *Nla* III (**C**ATG) recognition sites; i.e., TCCRA**C**ATG with the C in bold being shared by both recognition sequences. *Mme* I is a type IIS restriction enzyme, with cut sites 20-21/18-19 bp past its recognition site. Cutting the linkered DNA with *Mme* I releases the linker and 17-18/15-16 bp immediately 3′ to the *Nla* III site. These CATG+17 or 18 bp sequences become the identifier tags which are PCR amplified and ligated together to form ≥500 bp long concatemers prior to cloning and DNA sequencing. Because each clone contains multiple tags, sequencing throughput increases accordingly.

SERIAL ANALYSIS OF CHROMATIN OCCUPANCY (SACO)

In principle, *Mme* I derived tags can be used to identify the region of the genome from which any DNA or RNA (after conversion to cDNA) fragment is derived. As shown in Figure 1, *in silico* simulations of tag uniqueness *vs*. tag

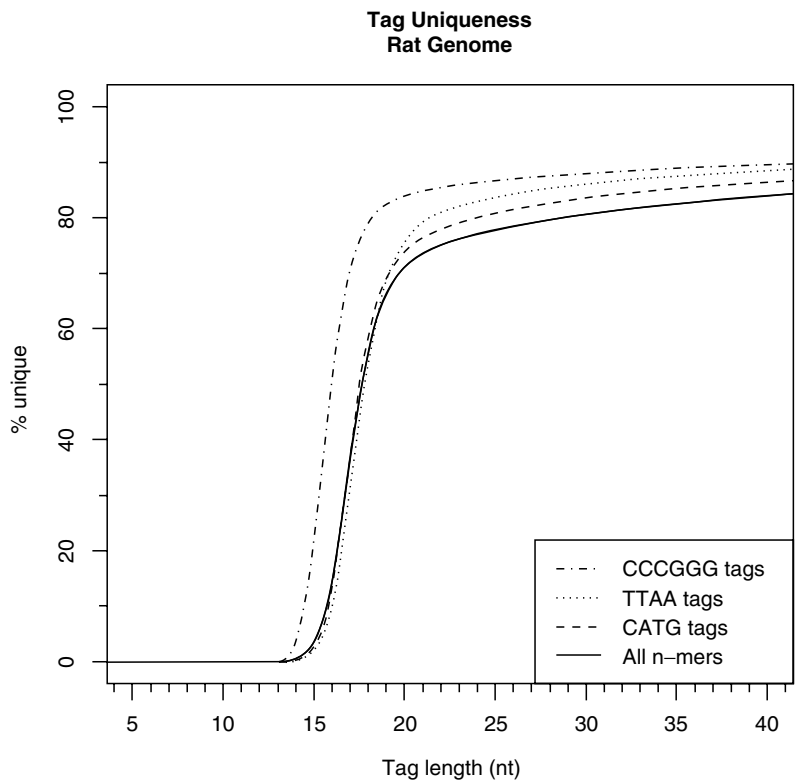

Figure 1. Plot of tag uniqueness *vs.* tag length for the entire rat genome. Plotted is tag length, which includes nucleotides at fragment ends specified by restriction sites: *Mse* I-TTAA; *Nla* III-CATG; *Sma* I CCCGGG or random ends (all N-mers) derived by sonication.

length for the rat genome show that uniqueness rapidly increases for lengths longer than ~12 bp and is limited only by the presence of highly repetitive regions in genomes. Similar profiles are obtained for the mouse and human genomes. With these as background, we reasoned that a ChIP-to-tag sequencing approach could be used to identify the genomic locations of ChIP-derived DNA fragments. To establish the effectiveness of the method, we set out to map globally the c-AMP response element binding protein (CREB) binding sites in the genome of rat PC12 cells (5). CREB was known to bind the cAMP-response element (CRE) (TGACGTCA) present in the promoters of many inducible genes (reviewed in 10). To increase the chances that CREB would be associated with CRE sequences, we first incubated the cells with forskolin to activate the enzyme adenylcyclase and increase the intracellular levels of cyclic AMP. The cells were then treated with formaldehyde, and, after randomly fragmenting the entire genome by sonication, the samples were subjected to ChIP using an anti-CREB antibody or, as a control, non-specific IgG. Real-time quantitative PCR showed that the CREB antibody provided an ~100-fold enrichment for c-fos (and other CREB targets) in the immunoprecipitates as compared to the IgG control. The ends of the CREB ChIP DNA were polished (protruding 3′ and 5′ ends were made flush by incubation with *E. coli* (Klenow fragment) and T4 DNA polymerases plus all four deoxynucleotide triphosphates and ligated to adapters for limited PCR amplification using biotinylated adapter-specific primers). The resulting DNA was digested with *Nla* III, and a modified Long-SAGE procedure was used to create concatemerized chains of randomly-associated 21 bp GSTs which were then cloned and sequenced. We termed this approach SACO; to demonstrate its utility, the sequences of ~75,000 tags from the PC12-derived library were determined. More than 40,000 CREB-SACO tags that mapped to unique loci in the rat genome were identified; 6,302 of these were identified two or more times. When these data were integrated with sequence annotation maps of the rat genome, forty percent of these loci were within 2 kb of the transcriptional start site of an annotated gene, and 72% were within 1 kb of a putative cAMP response element. In addition, CREB binding was confirmed for all loci supported by multiple tag hits (53 of 53 that were tested), and many of these loci were located upstream from genes not previously known to be regulated by CREB. These included genes for transcriptional regulators, chromatin modifying enzymes, coactivators, and co-repressors. A surprising result of the CREB SACO study was that CREB binding sites were commonly located in bi-directional promoters. Thus, the CRE that controls c-fos expression, for example, also regulates expression of a noncoding RNA transcribed in the opposite direction (5).

Since publication of the SACO method, several papers have appeared that utilized similar approaches, attesting to the overall utility of tag-to-genome mapping of ChIP DNA fragments (11-14). In all of these procedures the tags, whether they are generated from an internal restriction site or directly from the 5′ and 3′ ends of the sonicated ChIP DNA fragments, are analyzed separately as independent bits of sequence data. When mapped correctly to the genome sequence, these tags locate within about 1 to 2 kb the site that was cross-linked *in vivo* to the immunoprecipitated protein. In practice finding these sites involves scanning the genome sequence in both directions from a tag's location for a

nearby binding motif. The distance scanned is usually set at around twice the upper limit of the size of the ChIP DNA since when tags are analyzed separately, it is not known where they originated in the fragment, i.e., were they close to an end or more towards the middle of the ChIP fragment. To overcome this limitation, a new cloning strategy was developed by Ng and co-workers (15) that covalently links the tag sequences from each end of a DNA fragment into a paired diTAG structure. This approach, which was originally developed for identifying simultaneously both ends of full-length cDNAs, can also be used to map ChIP fragments with high precision.

PAIRED-END GENOMIC SIGNATURE TAGS (PE-GST)

The first step in the procedure is cloning of the DNA fragments into a special vector, pBEST (Both End Signature Tags), which is based on the pSCANS vector developed at BNL (http://genome.bnl.gov/Vectors/pscans.php). This low-copy number vector, with an isopropyl-beta-D-thiogalactopyranoside (IPTG) inducible origin of replication, was modified for efficient cloning of single DNA fragments in a manner that places them immediately adjacent to oppositely oriented *Mme* I recognition sequences (Figure 2). These are the only *Mme* I sites in the vector. Two *Bbs* I sites were placed between the *Mme* I sites in opposite orientations such that when the vector is cut with *Bbs* I, the linearized vector DNA will have non-self-ligatable ends with 4 nt overhangs (5′-GTCG-3′). A synthetic

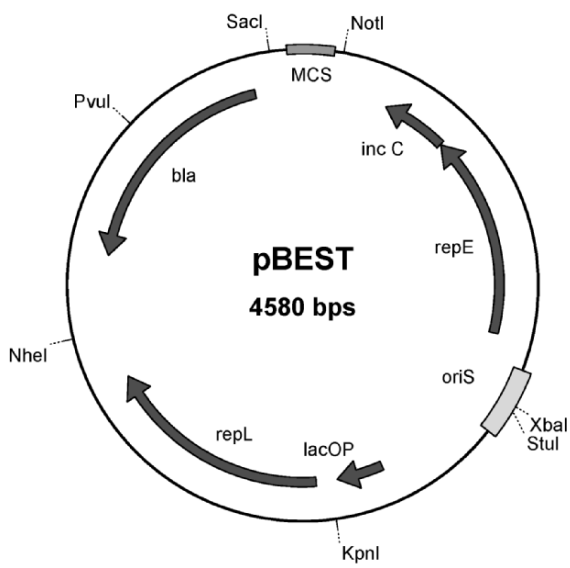

Figure 2a. Schematic diagram of pBEST paired-end vector. *ori*S, *rep*E and *inc* C are from the *E. coli* F factor, *lac*OP is the wild-type lac promoter, *rep*L is the lytic origin of replication from bacteriophage P1, and *bla* encodes β-lactamase activity (*amp*R). Several of the plasmid's unique restriction sites are indicated. MCS represents the cloning region, which is shown in greater detail in Figure 2b.

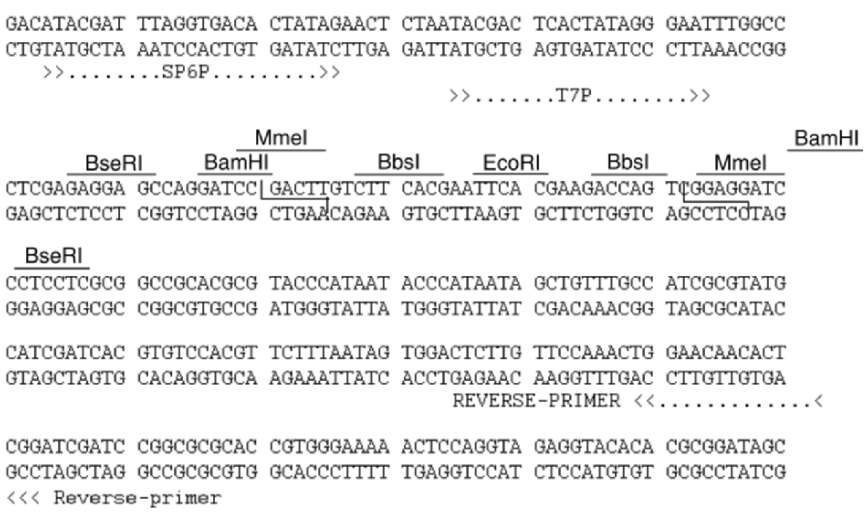

Figure 2b. MCS region of pBEST used for producing diTAGs. The locations of the relevant restriction enzyme recognition sites are indicated as are those for the primers used to PCR amplify the diTAG concatemers. Complete cutting with *Bbs* I generates a linearized vector with 5′-GTCG-3′ overhangs (indicated).

double-stranded DNA cassette is used to append simultaneously *Btg*Z I and *Bbs* I recognition sites to the ends of blunt-ended ChIP DNAs. The bottom strand of the cassette is 5′ phosphorylated (p) and its 3′ end is amino modified to prevent self-ligation higher than dimers.

Cassette #1

		*Btg*Z I	*Bbs* I	
5′	TCCGGTCTAC	TGAATTCCGA	ACGCGATGCT	GAAGACCACG AC
3′	Amino-AGGCCAGATG	ACTTAAGGCT	TGCGCTACGA	CTTCTGGTGC TGp

Similar cassettes with appropriate overhangs are used if dealing with fragments with cohesive ends. Cutting these cassettes with either *Btg*Z I or *Bbs* I generates 4 bp overhangs (5′-CGAC-3′) on the ends of the linkered DNA that are complementary to the overhangs of the *Bbs* I cut vector. After overnight ligation with excess linker, the ligation products are purified on a Qiagen Qiaquick PCR purification column, and the eluant is PCR amplified using 5′-biotin-TCCG-GTCTACTGAATTCCGAAC-3′ as primer. Ideally one should set up several different PCR reactions varying the amount of input template and PCR cycles. Amplified material should then be analyzed by agarose gel electrophoresis. The products should produce a smear that is similar in its size range to the DNA fragments in the sonicated ChIP starting material. The appropriate samples are phenol-chloroform extracted; then a portion is digested with *Btg*Z I and a similar portion with *Bbs* I to minimize loss of fragments with internal *Btg*Z I or *Bbs* I sites. After digestion the samples are combined, the cleaved linker cassettes are removed by gel electrophoresis or by binding to streptavidin beads, and the ChIP fragments are ligated into *Bbs* I cut pBEST to generate recombinant plasmids

with only single inserts. These are then electroporated into *E. coli* D1210, and the library is plated on ZYM5052 plus ampicillin (50 μg/ml) plates (16). Growth on ZYM5052 agar provides for solid-phase plasmid amplification without having to add IPTG to the plates, which should maintain library representation better than growth in liquid culture. The number of colonies required at this stage is determined by the estimated number of targets in the genome being investigated; we routinely target 1-10 X 10^5 cfu as a convenient benchmark. Cells can be plated at a density just below what is needed to provide for a confluent lawn. After overnight 37°C incubation, the resultant lawn of bacterial colonies is harvested by scraping into several ml of liquid medium and pelleted by centrifugation. Plasmid DNA preparation is performed e.g., by using a Qiagen Tip500 kit.

These clones now contain an *Mme* I site (TCCGAC) on each side of the DNA insert oriented so that digestion with *Mme* I cleaves 20-21 bp into the inserts from both their 5′ and 3′ ends. Consequently, despite the variable sizes of the original inserts, the vector-plus the two 20-21 bp tags on each end will be of a constant size (approx. 4,500 bp) that can be easily recognized upon agarose gel electrophoresis and can be purified from the unwanted internal ChIP-derived fragments that are produced during the *Mme* I digestion step. Approximately 5-10 μg of plasmid DNA is digested using *Mme* I as per the manufacturer's conditions (NEB), and the entire digestion reaction is then electrophoresed on a 0.7% low melt agarose gel. After staining, the vector plus tags band is excised, and the DNA is recovered. These molecules will eventually be ligated under dilute conditions to form circles that bring the tags at each end physically adjacent to each other as paired-end diTAGs. However, at this stage only 1 in 16 of the overhangs left following *Mme* I digestion are expected to be complementary to each other and able to form monomeric circles. Therefore, they either have to be removed by the 3′ exonuclease activity of T4 DNA polymerase prior to blunt end ligation or, alternatively, ligated with a special DNA adapter cassette with a 16-fold degenerate two-base 3′ overhang, which makes it compatible with all possible 3′ overhangs generated by *Mme* I digestion. Plasmid maps and detailed protocols are available on our web site (http://genome.bnl.gov/pBEST).

We initially used a blunt-ending approach to analyze DNA sequences associated with the product of the human p53 tumor suppressor gene (*TP53*). This 393 amino acid long polypeptide is known to function as a homotetrameric, sequence-specific transcription factor controlling cell cycle progression, DNA repair, and the induction of apoptosis and senescence in response to a variety of genotoxic and non-genotoxic stress signals (17-20). Genomic studies have shown that p53 induces or inhibits the expression of more than 1,500 human genes, but only a handful of p53 response elements (p53REs) have been characterized. The p53 tetramer binds a consensus DNA sequence, 5′-RRRCWWGYYY(N = 0-14) RRRCWWGYYY -3′, which consists of pairs of inverted repeats separated by 0 to 14 bp to create a 20 bp binding site (21-22). p53 also promotes the expression of some genes through elements that are of limited similarity to the consensus binding motif (e.g., *PIG3*, *PAC1*) (23-25); therefore, sequence pattern discovery algorithms alone cannot reliably predict where p53 will interact with its chromosomal targets nor does the presence of a consensus sequence itself determine whether the site will be occupied *in vivo* by p53. An added complication is that

the nuclear concentration of p53 increases one to two orders of magnitude, from a few hundred molecules per cell to perhaps a few thousand of tetramers per cell, in response to certain genotoxic and non-genotoxic stresses. Furthermore, post-translational protein modifications and the presence of other binding partners and their concentration all are thought to modulate p53's ability to transcriptionally activate or conversely repress target genes. Considerable effort will therefore be needed to map the global binding distribution of p53 in mammalian cells.

For our studies we are treating human lung tumor A549 cells with adriamycin for 15 hr and then carrying out standard ChIP enrichment of the cross-linked DNA using D01 as the anti-p53 antibody. After the cross-links were reversed and the repaired DNA ends were ligated with the adapter shown above, limited PCR was used to amplify the fragments with cassette-specific primers, then the DNA was digested with *Bbs* I and cloned into pBEST. Purified plasmid DNA from this clone pool was digested with *Mme* I, and the protruding 3′ ends were removed by incubation with T4 DNA polymerase and deoxynucleotide triphosphates. After blunt-end ligation to form circles, the sample was electrophoresed on a low melt agarose gel and the monomeric circle band was recovered and electroporated into electrocompetent D1210 cells. The cells were plated on ZYM5052 agar plates, and plasmid DNA was prepared as above. These molecules now have the following paired-end diTAG structure:

```
          TAG (18-19bp)                          TAG (18-19bp)
Vector—CGACNNNNNNNNNNNNNNNNNN (N) (N) NNNNNNNNNNNNNNNNNNGTCG—Vector
  DNA--GCTGNNNNNNNNNNNNNNNNNN (N) (N) NNNNNNNNNNNNNNNNNNCAGC-DNA
```

Each tag is 18 (or 19) bp long which, in most cases, is sufficient to allow the site from which the fragments were derived to be uniquely positioned on the genomic map. In practice, since it can be hard to tell just from inspection where one tag ends and the next begins, tags of only 18 nucleotides are extracted.

INCREASING TAG LENGTH AT THE 3′ END

With the strategy described above, the two unique bp at the 3′ end of each tag are lost, which results in an inability to uniquely identify a tag's location in large genomes (Figure 1). One strategy for capturing these nucleotides in the tag is based on the approach used in the process called TALEST (tandem arrayed ligation of expressed sequence tags) developed by Spinella et al. (26) and modified by our laboratory for our original GST protocol (8). It employs ligation with a 16-fold degenerate oligonucleotide to capture all the sequence information in the *Mme* I′ site's 3′ extensions. To further simplify downstream processing of the data, we designed the linker to contain tandem copies of a *Bcl* I recognition site (5′-pTGATCACGTGATCANN-3′). After it is ligated to the *Mme* I 3′ overhangs, digestion with *Bcl* I leaves a single cohesive GATC overhang on the end of each tag, and these linear DNAs can now easily be ligated to form circles. Because *Bcl* I cutting is blocked by *dam* methylation, the enzyme will not cut in the tags or in the vector as the plasmid DNA was prepared from *E. coli* D1210, a *dam*$^+$ strain. After ligation and purification of the monomeric circles by agarose gel electrophoresis, the DNA is treated as before by being electroporated into

electrocompetent D1210 cells, and the library is plated on ZYM5052 agar plates. The resulting plasmid DNAs now have the following structure:

```
      TAG (20-21 bp)          TAG (20-21 bp)
Vector-CGACNNNNNNNNNNNNNNNNNNNN (N) TGATCA (N)
        NNNNNNNNNNNNNNNNNNNNGTCG-Vector
 DNA-GCTGNNNNNNNNNNNNNNNNNNNN (N) ACTAGT (N)
        NNNNNNNNNNNNNNNNNNNNCAGC-DNA
```

Each tag is 20 (or 21) bp long with the *Bcl* I recognition sequence serving as a clear punctuation mark to divide diTAGs into their respective left and right ends. It is also easy to tell if a tag is 20 or 21 bp long. In principle, several additional linkers based on the above *Bcl* I paradigm could be used provided the cognate methylase is available, e.g., *Bam*H I or *Eco*R I.

*BSE*R I DIGESTION TO RELEASE PAIRED-END DITAGS

Approximately 5-10 μg of plasmid DNA is prepared from plate scrapings and then digested using *Bse*R I as the manufacturer's conditions (NEB). Since each diTAG is flanked by a suitably positioned *Bse*R I recognition site, digestion releases one diTAG pair from every DNA circle. These are the only *Bse*R I sites in the vector, and the paired-end diTAGs can be easily purified from the linearized vector on a 1.5% low melt agarose gel and then concatemerized as described previously for Long-SAGE tags. After size fractionation, the concatemers are cloned back into pBEST cut with *Bse*R I and dephosphorylated to form the paired-end diTAG library. We routinely plate out this library on non-inducing agar plates, e.g., 2xYT, and then pick colonies into 96-well cultures using ZYM-5052 liquid autoinduction medium. Dilutions (1 to 10) of the overnight cultures are boiled for 10 min. to release DNA, which is then used as template in PCR reactions to amplify the concatemer inserts. After incubation with alkaline phosphatase and exonuclease I, the samples are sequenced using the same primers as were used for the PCR reactions. The concatemers have the following architecture if the degenerate *Bcl* I linker was used:

GTCGAC-Tag1-TGATCA-Tag1′-GTCGAC-Tag2-TGATCA-Tag2′-
GTCGAC-Tag3-TGATCA-Tag3′-GTCGAC-Tag4-TGATCA-Tag4′. . . .etc.

Each diTAG pair begins and ends with the sequence GTCGAC (a *Sal* I site), and in-between each set of paired-end tags is a single copy of the *Bcl* I recognition sequence (TGATCA), which makes parsing of the 20 or 21 bp tags straightforward.

PAIRED-END PROFILING OF THE METHYLOME

Because the degenerate linker strategy maximizes the information content at the 3′ ends of the tags, it has become the core strategy for our ongoing analysis of p53 binding sites, and it also is being adapted for global analysis of alterations in the human genome involving 5′ methylation of cytosine in CpG dinucleotides. These alterations are regarded as epigenetic as they control gene

expression in cells and during development but do not change the DNA sequence. Seventy percent of all cytosines in CpG dinucleotides in the human genome are methylated and prone to deamination, resulting in a cytosine to thymine transition, CpG to TpG or CpA on the complementary DNA strand (27-28). This process is believed to have led to an overall reduction in the frequency of guanine and cytosine in the human genome to about 40% of all nucleotides and a further reduction in the frequency of CpG dinucleotides to about a quarter of their expected frequency (29). The exception to CpG under representation in the genome is within CpG islands, which were originally called HTFs, for *Hpa* II tiny fragments that remained uncut after digestion with the 5 mC sensitive restriction enzyme *Hpa* II (CCGG) (29). CpG islands were later formally defined as sequences >200 bp in length with a GC content >0.5, and a CpGobs/CpGexp (observed to expected ratio based on GC content) >0.6 (29-30). However, more recent studies have shown that CpG islands located near transcription start sites are usually longer than 500 bp while those less than 500 bp tend to be associated with repetitive elements (31-32).

Determining the global pattern of DNA methylation, or the methylome (33), and its variation in cells is an area of considerable interest because of its potential use as an early diagnostic biomarker for cancer (34-35). Tumor cells exhibit hypomethylation of their genomes, but the promoters of certain tumor suppressor genes (e.g., $p16^{ARF}$) frequently are silenced in tumor cells through hypermethylation (reviewed in 36). Accordingly, numerous approaches are being developed to identify methylation-silenced or demethylation activated genes. In one approach we are taking, total genomic DNA is digested to completion with *Mse* I (T/TAA), whose recognition site is found rarely within CpG islands but occurs about once every 140 bp in bulk DNA. DNA fragments with methylated cytosines in the digest are then separated from the remainder of the genomic fragments by affinity chromatography (37-39). The methyl CpG fragments can be ligated with

Cassette #2

```
                                        BtgZ I       Bbs I
5′        TCCGGTCTAC TGAATTCCGA  ACGCGATGCT  GAAGACCACG  AC
3′ Amino-AGGCCAGATG ACTTAAGGCT  TGCGCTACGA  CTTCTGGTGC  TGATp
```

and then digested with *Btg*Z I and *Bbs* I as in the ChIP protocol. After cloning and *Mme* I digestion, the resulting paired-end diTAGs have the following structure

```
     TAG (20-21bp)                      TAG (20-21bp)
Vector-CGACTAANNNNNNNNNNNNNNNNNN (N) TGATCA (N)
        NNNNNNNNNNNNNNNNNNTTAGTCG-Vector
 DNA-GCTGATTNNNNNNNNNNNNNNNNNN (N) ACTAGT (N)
          NNNNNNNNNNNNNNNNNNAATCAGC-DNA
```

with the nucleotides in bold coming from the *Mse* I recognition sequence. In this case, as shown in Figure 1, tag length is critical since the first 3 bases are already fixed by the remainder of the *Mse* I recognition sequence. Decreasing tag length by trimming off the 3′ extensions after *Mme* I cutting would inflict a sizable

penalty on the chances of the tags being unique. Another example of tag size counting is shown in Figure 3, which illustrates the basic principles of the Methylated CpG island Amplification (MCA) protocol developed by Issa and co-workers (40) and how it can be modified to provide paired-end diTAGs. In this case the DNA is first digested with *Sma* I, which only cleaves leaving blunt ends provided the central CpG dinucleotide in its recognition sequence (CCC/GGG) is unmethylated. These methylated sites, however, can be cleaved with *Xma* I (C/CCGG G) to leave a 4 base overhang. Ligation of the overhang with the DNA adapter cassette #3 shown below followed by cleavage with *Btg*Z I or *Bbs* I places 5′ CGAC 3′ overhangs on the ends of what were methylated CCCGGG sequences in the genome.

Cassette #3

```
                                    BtgZ I        Bbs I
5′ TCCGGTCTAC        TGAATTCCGA  ACGCGATGCT  GAAGACCACG  A
3′ Amino-AGGCCAGATG ACTTAAGGCT  TGCGCTACGA  CTTCTGGTGC  TGGCCp
```

About 70-80% of CpG islands contain at least two closely spaced (≤1 kb) *Sma* I sites. If they are consecutively methylated they can be used for cloning the intervening CpG-rich segments since after *Btg*Z I and/or *Bbs* I digestion they will have the CGAC overhangs needed for ligation into the *Bbs* I-digested pBEST vector. During cloning the two *Mme* I recognition sequences flanking the inserts are recreated, and the 3˚ C in the overhang now becomes the last residue in the *Mme* I recognition site. Therefore, cutting with *Mme* I will generate tags that are CGGG plus 16 or 17 nt, which maximizes their information content for determining where these fragments map in the genome (see Figure 1).

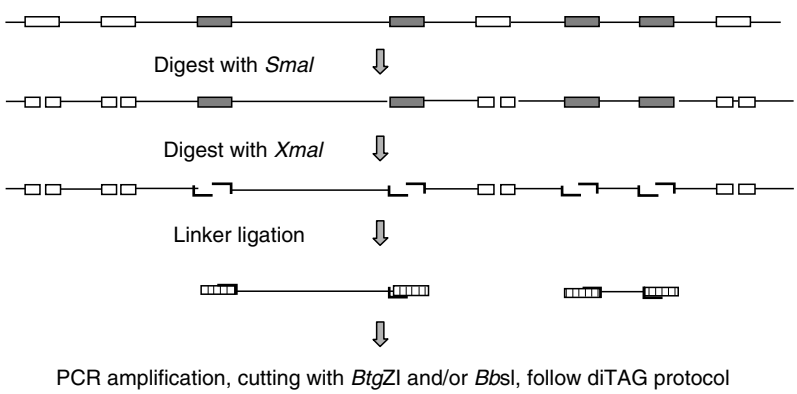

Figure 3. Schematic diagram of *Smn* I/*Xma* I double-digestion protocol. Genomic DNA is represented by a solid line with eight *Sma* I CCCGGG recognition sequences; four sites are non-methylated (open boxes), and four are methylated (filled boxes). The non-methylated sites are cut in a first digestion with methyl-sensitive *Sma* I leaving blunt ends. A second digestion is performed using the methyl insensitive *Sma* I isoschizomer *Xma* I, which leaves CCGG overhangs. DNA adapters with appropriately positioned *Btg*Z I and *Bbs* I sites are ligated to the overhangs, and DNA fragments with an adapter at each end are PCR amplified using primers complementary to the adapters.

SUMMARY

Because paired-end genomic signature tags are sequenced-based, they have the potential to become an alternate tool to tiled microarray hybridization as a method for genome–wide localization of transcription factors and other sequence-specific DNA binding proteins. As outlined here the method also can be used for global analysis of DNA methylation. One advantage of this approach is the ability to easily switch between different genome types without having to fabricate a new microarray for each and every DNA type. However, the method does have some disadvantages. Among the most rate-limiting steps of our PE-GST protocol are the need to concatemerize the diTAGs, size fractionate them and then clone them prior to sequencing. This is usually followed by additional steps to amplify and size select for long (≥500) concatemer inserts prior to sequencing. These time-consuming steps are important for standard DNA sequencing as they increase efficiency ~20-30-fold since each amplified concatemer can now provide information on multiple tags; the limitation on data acquisition is read length during sequencing. However, the development of new sequencing methods such as Life Sciences' 454 new nanotechnology-based sequencing instrument (41) could increase tag sequencing efficiency by several orders of magnitude (≥100,000 diTAG reads/run), which is sufficient to provide in-depth global analysis of all ChIP PE-GSTs in a single run. This is because the lengths of our paired-end diTAGs (~60 bp) fall well within the region of high accuracy for read lengths on this instrument. In principle, sequence analysis of

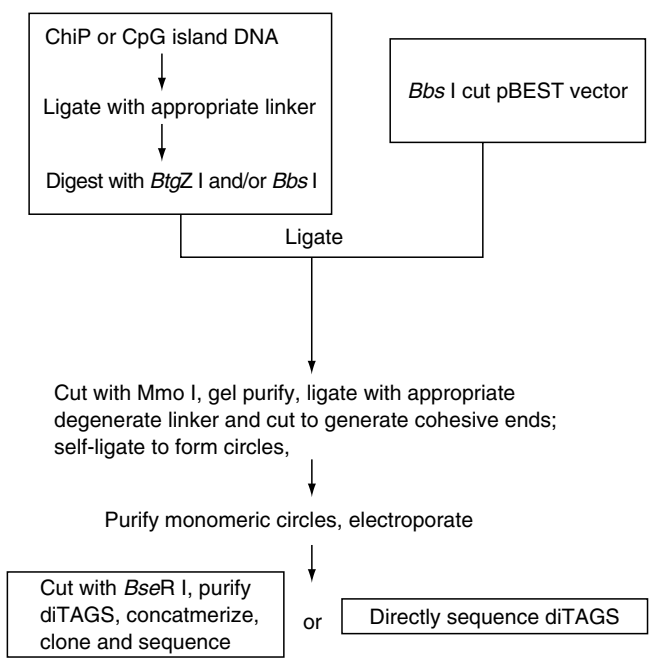

Figure 4. Schematic diagram of the diTAG protocol.

diTAGs could begin as soon as they are generated, thereby completely bypassing the need for the concatemerization, sizing, downstream cloning steps and sequencing template purification. In addition, our protocol places any one of several unique four-base long nucleotide sequences, such as GATC, between each and every diTAG pair, which could be used to help the instrument's software keep base register and also provide a well-located peak height indicator in the middle of every sequence run. This additional feature could permit multiplexing of the data by simultaneous sequencing of several pooled libraries if each used a different linker sequence during diTAG formation (Figure 4).

ACKNOWLEDGMENTS

Support from National Institutes of Health Grant AI056480, the Low Dose Radiation Research Program of the Office of Biological and Environmental Research (BER), U.S. Department of Energy and the BNL Laboratory Directed Research and Development Program is gratefully acknowledged.

REFERENCES

1 Orlando, V. (2000) Trends Biochem. Sci. 25, 99-104.
2 Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P. and Young, R.A. (2000) Science 290, 2306-2309.
3 Wells, J., Graveel, C.R., Bartley, S.M., Madore, S.J. and Farnham, P.J. (2002) Proc. Nat. Acad. Sci. U.S.A. 99, 3890-3895.
4 Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002) Nat. Biotechnol. 20, 508-512.
5 Impey, S., McCorkle, S.R., Cha-Molstad, H., Dwyer, J.M., Yochum, G.S., Boss, J.M., McWeeney, S., Dunn, J.J., Mandel, G. and Goodman, R.H. (2004) Cell 119, 1041-1054.
6 Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M.V. and Eichler, E.E. (2005) Nat. Genet. 37, 727-732.
7 Bird, A.P. (1986) Nature 321, 209-213.
8 Dunn, J.J., McCorkle, S.R., Praissman, L.A., Hind, G., Van Der Lelie, D., Bahou, W.F., Gnatenko, D.V. and Krause, M.K. (2002) Genome Res. 12, 1756-1765.
9 Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Science 270, 484-487.
10 Mayr, B. and Montminy, M. (2001) Nat. Rev. Mol. Cell Biol. 2, 599-609.
11 Chen, J. and Sadowski, I. (2005) Proc. Nat. Acad. Sci. U.S.A. 102, 4813-4818.
12 Kim, J., Bhinge, A.A., Morgan, X.C. and Iyer, V.R. (2005) Nat. Methods 2, 47-53.
13 Labhart, P., Karmakar, S., Salicru, E.M., Egan, B.S., Alexiadis, V., O'Malley, B.W. and Smith, C.L. (2005) Proc. Nat. Acad. Sci. U.S.A. 102, 1339-1344.

14 Roh, T.Y., Cuddapah, S. and Zhao, K. (2005) Genes Dev. 19, 542-552.
15 Ng, P., Wei, C.L., Sung, W.K., Chiu, K.P., Lipovich, L., Ang, C.C., Gupta, S., Shahab, A., Ridwan, A., Wong, C.H., Liu, E.T. and Ruan, Y. (2005) Nat. Methods 2, 105-111.
16 Studier, F.W. (2005) Protein Expr. Purif. 41, 207-234.
17 Wahl, G.M. and Carr, A.M. (2001) Nat. Cell Biol. 3, E277-286.
18 Vousden, K.H. and Lu, X. (2002) Nat. Rev. Cancer 2, 594-604.
19 Yang, A., Kaghad, M., Caput, D. and McKeon, F. (2002) Trends Genet. 18, 90-95.
20 Oren, M. (2003) Cell Death Differ. 10, 431-442.
21 el-Deiry, W.S., Kern, S.E., Pietenpol, J.A., Kinzler, K.W. and Vogelstein, B. (1992) Nat. Genet. 1, 45-49.
22 Hoh, J., Jin, S., Parrado, T., Edington, J., Levine, A.J. and Ott, J. (2002) Proc. Nat. Acad. Sci. U.S.A. 99, 8467-8472.
23 Contente, A., Dittmer, A., Koch, M.C., Roth, J. and Dobbelstein, M. (2002) Nat. Genet. 30, 315-320.
24 Kim, E. and Deppert, W. (2003) Biochem. Cell Biol. 81, 141-150.
25 Yin, Y., Liu, Y.X., Jin, Y.J., Hall, E.J. and Barrett, J.C. (2003) Nature 422, 527-531.
26 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) Nucl. Acids Res. 27, e22.
27 Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Nature 274, 775-780.
28 Bird, A.P. and Taggart, M.H. (1980) Nucl. Acids Res. 8, 1485-1497.
29 Bird, A. (2002) Genes Dev. 16, 6-21.
30 Gardiner-Garden, M. and Frommer, M. (1987) J. Mol. Biol. 196, 261-282.
31 Ponger, L., Duret, L. and Mouchiroud, D. (2001) Genome Res. 11, 1854-1860.
32 Ponger, L. and Mouchiroud, D. (2002) Bioinformatics 18, 631-633.
33 Feinberg, A.P. (2001) Nat. Genet. 27, 9-10.
34 Baylin, S.B., Herman, J.G., Graff, J.R., Vertino, P.M. and Issa, J.P. (1998) Adv. Cancer Res. 72, 141-196.
35 Marsit, C.J., Kim, D.H., Liu, M., Hinds, P.W., Wiencke, J.K., Nelson, H.H. and Kelsey, K.T. (2005) Int. J. Cancer 114, 219-223.
36 Feinberg, A.P., Ohlsson, R. and Henikoff, S. (2006) Nat. Rev. Genet. 7, 21-33.
37 Cross, S.H., Charlton, J.A., Nan, X. and Bird, A.P. (1994) Nat. Genet. 6, 236-244.
38 Shiraishi, M., Chuu, Y.H. and Sekiya, T. (1999) Proc. Nat. Acad. Sci. U.S.A. 96, 2913-2918.
39 Rauch, T. and Pfeifer, G.P. (2005) Lab. Invest. 85, 1172-1180.
40 Toyota, M. and Issa, J.P. (2002) Methods Mol. Biol. 200, 101-110.
41 Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P.,

Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) Nature 437, 376-380.