2952–2961 *Nucleic Acids Research, 2005, Vol. 33, No. 9*
doi:10.1093/nar/gki582

CpG Island microarray probe sequences derived from a physical library are representative of CpG Islands annotated on the human genome

Lawrence E. Heisler[1], Dax Torti[1], Paul C. Boutros[2,3,8], John Watson[2,3], Charles Chan[1], Neil Winegarden[4], Mark Takahashi[4], Patrick Yau[4], Tim H.-M. Huang[5], Peggy J. Farnham[6], Igor Jurisica[3,7,8], James R. Woodgett[3,4,8], Rod Bremner[1,9], Linda Z. Penn[2,3] and Sandy D. Der[1,*]

[1]Department of Laboratory Medicine and Pathobiology, Program in Proteomics and Bioinformatics, University of Toronto, Toronto, ON M5S 1A8, Canada, [2]Division of Cancer Genomics and Proteomics, Ontario Cancer Institute, University Health Network, Toronto, ON M5G 2M9, Canada, [3]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 2M9, Canada, [4]University Health Network Microarray Centre, Toronto, ON M5G 2C4, Canada, [5]Human Cancer Genetics Program, Department of Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA, [6]Department of Medical Pharmacology and Toxicology, University of California-Davis, Davis, CA 95616, USA, [7]Department of Computer Science, University of Toronto, Toronto, ON M5S 2M9, Canada, [8]Division of Signaling Biology, Ontario Cancer Institute, Toronto, ON M5G 2M9, Canada and [9]Toronto Western Research Institute, Toronto, ON M5T 2S8, Canada

Received March 15, 2005; Revised and Accepted April 28, 2005

ABSTRACT

An effective tool for the global analysis of both DNA methylation status and protein–chromatin interactions is a microarray constructed with sequences containing regulatory elements. One type of array suited for this purpose takes advantage of the strong association between CpG Islands (CGIs) and gene regulatory regions. We have obtained 20 736 clones from a CGI Library and used these to construct CGI arrays. The utility of this library requires proper annotation and assessment of the clones, including CpG content, genomic origin and proximity to neighboring genes. Alignment of clone sequences to the human genome (UCSC hg17) identified 9595 distinct genomic loci; 64% were defined by a single clone while the remaining 36% were represented by multiple, redundant clones. Approximately 68% of the loci were located near a transcription start site. The distribution of these loci covered all 23 chromosomes, with 63% overlapping a bioinformatically identified CGI. The high representation of genomic CGI in this rich collection of clones supports the utilization of microarrays produced with this library for the study of global epigenetic mechanisms and protein–chromatin interactions. A browsable database is available on-line to facilitate exploration of the CGIs in this library and their association with annotated genes or promoter elements.

INTRODUCTION

An understanding of the epigenetic mechanisms and protein–DNA interactions that affect gene expression is an important component of functional genomics research. Chromatin-immunoprecipitation (ChIP) has been used to study these interactions by allowing the isolation of genomic fragments targeted by an antibody or bound by a specific protein *in vivo* (1–4). The presence of a specific chromosomal target in the immunoprecipitated fragments can be determined by amplifying the region of interest with appropriately designed primers. This approach is limiting in that it depends on prior knowledge of the expected gene targets, knowledge of the precise chromosomal location to which a protein binds near the gene, and requires separate PCR for each target. Global analysis of the DNA products from a ChIP assay is possible by hybridizing the fragments against a microarray of chromosomal regions, an approach that has been called ChIP-chip (5–7).

*To whom correspondence should be addressed. Tel: +1 416 978 8878; Fax: +1 416 978 5959; Email: sandy.der@utoronto.ca

Microarray analysis of ChIP samples requires the construction of arrays using probes against regulatory regions with which the protein might interact. The first ChIP-chip experiments were performed in yeast, using intergenic regions as probes on the microarray (6). ChIP-chip experiments in humans have been conducted using array sequences derived from regions directly upstream of the genes and into the first exon (8,9). This approach assumes that the majority of protein–DNA interactions that affect RNA transcription occur in regions near the transcription start site (TSS). A more comprehensive approach has been to use arrays with a set of probes that represent all regions within a given locus (10), chromosome (11) or genome (12,13).

An alternate approach to designing a microarray for ChIP-chip studies has been to use CpG Island (CGI) sequences. CGIs are genomic regions that are thought to have escaped the gradual loss of CpG dinucleotides that results from the transition of methylated cytosine to thymidine and is responsible for the CpG scarcity through most vertebrate genomes (14,15). It has been estimated that ~60% of all human genes are associated with a CGI usually in the 5′ end (16), and ~85% of CGIs have been determined to be within −500 to +1500 bp of a TSS. (17). Furthermore, a strong correlation has been noted between CGIs and clusters of transcription factor binding sites (18). This association of CGIs with promoter regions suggests that *trans*-acting factors bound to these sites prevent cytosine methylation and subsequent degradation (19). Supporting this idea is the observation that *de novo* methylation of CGIs can result in transcriptional repression and X-chromosome inactivation while demethylation with 5-azacytidine can remove gene repression (20).

The first large-scale effort to computationally identify CGIs was performed using GenBank sequences (14). With the completion of the human genome-sequencing project, bioinformatic approaches have been applied to identifying CGIs across the human genome (21) and CGI tracks are available on the major publicly accessible, genome browsers (NCBI, UCSC). The criteria and algorithms used to determine the bounds of the islands vary, but generally it is accepted that CGIs are 200 bp or greater in length with a G + C content >50% and a CpG percentage that exceeds 60% of that expected in random sequence (1/16) (14).

In contrast to bioinformatic identification of CGIs, a physical CGI library was constructed using a two-step cloning strategy involving isolation of GC-rich chromosomal fragments based on their lack of methylation *in vivo* and followed by reselecting fragments that could be methylated *in vitro* (22). After analysis of 113 clones in this library, it was concluded that 77% of the clones were derived from CpG-rich regions. The first array constructed from this clone library was utilized to determine the methylation status of CGIs in breast cancer cells (23). These arrays have since been used effectively for analysis of E2F (5) and c-myc targets (24) using a ChIP-chip approach. We have recently constructed microarrays containing probes derived from the clones isolated from this library. To better understand the nature of the CGI library and to facilitate analysis of CGI arrays, we have sequenced 20 736 clones and conducted a large-scale characterization that includes identification of the genomic origins of each clone as well as other potential hybridization targets in the genome.

MATERIALS AND METHODS

CGI library and sequence data

The CGI Library had been prepared and described (25). From this Library, 12 192 clones (12k Set) were obtained from the Wellcome Trust Sanger Institute (Cambridge, UK). Sequencing of these clones had been done previously at the Sanger Institute and this information is publicly available (http://www.sanger.ac.uk/HGP/cgi.shtml). A second set of 8544 clones (9k Set), derived from the same library but screened with human Cot1 DNA to remove clones with repetitive elements, was obtained from T. H. Huang (Ohio State University). Subsets of these clones had been used previously in the construction of several arrays (23,26), including the CGI Promoter arrays available through the UHN Microarray Centre in Toronto (www.microarray.ca). The complete set of 20 736 clones was replicated and sent for sequencing to the Genome Sciences Centre (Vancouver, BC).

Sequence alignment

BLAT (Blast-like Alignment Tool) (27) was obtained from UCSC Genome Bioinformatics (http://genome.ucsc.edu/FAQ/FAQblat). The May 2004 build (Hg17) of the human genome was obtained from the UCSC Genome Bioinformatics site (http://hgdownload.cse.ucsc.edu/downloads.html#human) (28) and formatted for local BLAT alignments. Annotated CGIs, Refseq and Known Gene positions from this build were obtained from UCSC and used for analysis of clone alignments.

End reads from each clone were aligned to the genomic sequence using BLAT ('-fastmap' option), masking out repetitive elements and low-complexity regions. Base-calls in the end read with a PHRED score <20 were also masked. Alignments to the genome for each clone were constructed by combining the BLAT alignment from each end read when within 5000 bp of each other. In cases where only a single read was available, clone alignments were constructed from a single BLAT alignment, but marked as incomplete. In cases where sequence reads for a single clone aligned independent of each other, alignment information was stored and this was noted. Once all clone sequences had been aligned, genomic loci were defined by combining overlapping alignments resulting from redundant clones in the collection. These loci were evaluated for chromosomal distribution, CpG content and proximity to TSS.

Two sets of loci were generated from the human genomic sequence for comparison with the CGI Library loci. To model the expected contents of the CGI Library, MseI fragments containing complete or partial annotated CGIs were identified. In addition, a set of random loci was created by selection of 5000 random positions distributed across all 23 chromosomes proportional to chromosome length, and extracting a 1000 bp downstream from each position.

CGI library sequence and alignment data has been stored in a MySQL database and made publicly available. Queryable web-based forms have been constructed (http://derlab.med.utoronto.ca/CpGIslands/) to facilitate analysis of this library, as well as retrieval of information associated with specific probe identifiers on the CGI Promoter arrays.

CGI array hybridization

Two samples of 100 ng of genomic DNA, prepared from 2ftgh fibrosarcoma cells, were spiked with 2 ng arabidopsis

DNA control and random-primed with a final 50 µl mixture of 1× Sequenase buffer, 0.12 mM amino-allyl-dUTP/dNTP [0.12 mM dATP/dGTP/dCTP, 0.048 mM dTTP (Invitrogen, 10297-018) and 0.072 mM aminoallyl-dUTP (Sigma, A0410)] and 1 µl (13 U) of Sequenase T7 DNA polymerase (USB, 70775Y/Z) for 4 h at 37°C. The resulting product was purified using the Cyscribe GFX purification kit (Amersham Biosciences, 27-9606-01) according to manufacturer's directions. Recovered DNA was reduced to a final volume of 8 µl, and incubated with DMSO reconstituted Alexa 647 or Alexa 555 dyes (Molecular Probes, A32755), in a final volume of 10 µl for 1 h at room temperature. The labeled DNA was again purified using the Cyscribe GFX purification kit. Both 100 ng samples were pooled and added to 85 µl of hybridization cocktail [0.5 µg/µl calf thymus DNA (Sigma, D8661) and 0.5 µg/µl yeast tRNA (Invitrogen, 15401-029) in DIG Easy Hyb solution (Roche, 1603558)] and hybridized to the array for 18 h. After washing (1× SCC and 0.1% SDS solution at 50°C), the arrays were scanned using the GenePix 4000A microarray scanner (Axon Instruments) at PMT voltages between 750 and 800 at 100% laser power.

Background subtraction was done using an algorithm that fits a convolution of exponential and normal distributions to foreground intensities (normexp) with an offset of 50 for low-intensity shrinkage. Data were normalized with a robust-spline algorithm (29). All algorithms were implemented in the limma package (v1.8.21) (30) of the BioConductor library (31) for the R statistical package.

RESULTS

Sequence information was obtained for two sets of CGI clones. The 12k set (12 192 clones) was a portion of the original clone selection that had been deposited at the Wellcome Trust Sanger Institute (22). A second 9k set (8544 clones) had been isolated from the same CGI Library by the Huang laboratory (32). Vector-trimmed sequences that were longer than 50 bp and had a PHRED score >20 were used for alignment to the genome. A summary of the available sequence and alignment to the human genome is presented in Table 1 for each set and for the combined set of 20 736 clones. The 9k set was generally of better quality, with 95% having acceptable sequence versus 78% of the 12k set (Table 1, CGI Library sequence information). Combined, 85% of the clone sequences were considered suitable for alignment to the human genomic sequence. The 9k set had a median sequence length of 499 bp with 2% of the reads <100 bp in length. In contrast, 11% of the 12k set reads were <100 bp in length, contributing to the shorter median sequence length of 307 bp. A second round of sequencing performed on 768 blinded clones confirmed the sequence data obtained in the first round (data not shown).

Of the 17 645 clones with sequence, 14 901 (84%) aligned to the human genome using BLAT (Table 1, Genomic alignment). The proportion of clones aligning was higher for the 9k set (92%) versus the 12k set (79%) as expected due to the longer reads in this set. Overall, 90% of the clones with sequence aligned at a single position in the genome. While many partial alignments were observed, the majority of sequences showed nearly full-length alignments to the genome.

Table 1. CGI Library sequence information, genomic alignment and genomic loci

	12k	9k	21k
CGI Library sequence information[a]			
Number of clones derived from CGI Library	12 192	8544	20 736
Clones with sequence	9552 (78%)	8093 (95%)	17 645 (85%)
Both end reads	7786 (63%)	7508 (88%)	15 294 (74%)
Single end reads	1766 (14%)	585 (7%)	2351 (11%)
Median read length (bp)	307	499	414
Genomic alignment[b]			
Clones with sequence	9552	8093	17 645
Number aligning	7495 (79%)	7446 (92%)	14 901 (84%)
Align once	6754 (71%)	6692 (83%)	13 410 (76%)
Multiple aligns	741 (8%)	754 (9%)	1495 (8%)
Non-aligning	2041 (21%)	646 (8%)	2751 (16%)
Genomic loci[c]			
Clones aligning	7495	7446	14 901
Loci	5411	4937	9595
Distinct loci (exclusive of the other clone set)	4658	4184	
Common loci (from both clone sets)	753		
Loci defined by 1 clone	3171	3008	6179
Loci defined by multiple redundant clones	2240	1929	3416

Statistics is shown for the 12k clone set, the 9k clone set and the combined 21k set.
[a]The number of clones sent for sequencing, the number with usable sequence.
[b]The number of clones with usable sequence that aligned to the Human Genome build. The number of clones aligning at a single versus multiple positions is also indicated.
[c]The number of loci generated from alignments in each clone set. The overlap between the two clone sets is indicated, along with the number of loci unique to each clone set. Also indicated are the number of loci generated by a single clone alignment and multiple clone alignments.

The average alignment length was 90% of the sequence length. The median sequence length for clones that did align was 480 bp versus only 217 bp for clones that did not align. Although it appears likely that read length affected the ability to align sequence, some sequences that did not align to genomic DNA were as long as 962 bp.

Alignments which overlapped each other on the genomic scaffold were combined to define 9595 distinct genomic loci (Table 1, Genomic loci). The 9k set generated 4937 loci while the 12k set generated 5411 loci. Interestingly, only 753 loci (8%) were shared, indicating that the two sets were very distinct from each other. The majority of loci (65%) were defined by a single clone (Figure 1, upper panel), although these accounted for only 30% of all clones which aligned (Figure 1, lower panel). Therefore, 70% of the clones showed some degree of redundancy within the combined sets. Approximately 39% had a low degree of redundancy (2–5 per locus) while 15% were highly redundant (11+ per locus). This last group consists of 2958 clones defining only 118 loci and includes a single locus defined by 582 clones.

Genomic loci ranged in size between 50 and 2589 bp, with a mean locus length of 511 bp. Many of the shorter loci result from partial sequence alignments and often these sequences align to a greater degree elsewhere in the genome. The percentage of each sequence contributing to the total alignment length is shown in Figure 2 (upper panel). Most alignments <200 bp tend to be due to partial alignments and this trend

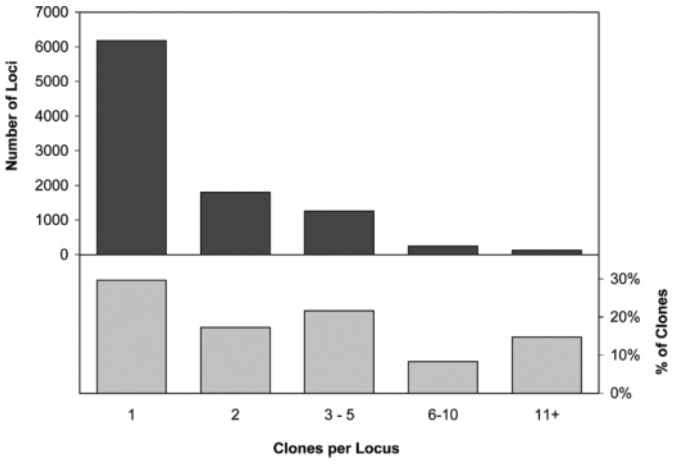

Figure 1. Clone composition of genomic loci. Each genomic locus is defined by one or more CGI library clone alignments. The majority of loci (6179/9552) are each defined by a single, non-redundant clone (upper panel, left-most bar). Fewer loci are represented by redundant clones with 1782 loci each defined by a pair of clones, 1240 by 3–5 clones, 235 by 6–10 clones and 118 loci each represented by 11 or more clones (range 11–582 clones). In the lower panel, the total number of clones represented in each group is indicated. The most redundant group consisting of 2958 clones defines only 118 loci.

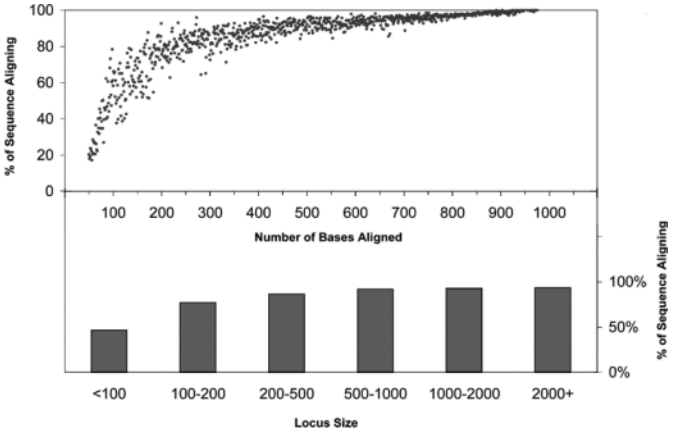

Figure 2. Percentage of sequence aligning. In the upper panel, the percentage of the sequence length aligning is plotted against the total number of bases aligned (alignment length). The shorter alignments generally result from partial sequence alignments and as the total length of the alignment increases, so does the % aligning. The lower panel shows the % sequence alignment for loci of various size ranges. Loci <100 bp in length are generated mostly from these partial alignments, while the longer loci (200 bp) are derived from nearly complete sequence alignments.

diminishes as the total aligned length increases. When the loci defined from these alignments are examined (Figure 2, lower panel), shorter loci are usually defined by partially aligning clones while loci 200 bp in size and greater generally result from nearly complete sequence alignments. Accordingly, we have not included loci <200 bp in length in the subsequent analysis, reducing the number from 9595 to 7184.

To evaluate the overall quality of the loci defined by the physical CGI Library, we calculated two metrics, the G + C content and CpG dinucleotide frequency, and compared this to the computationally annotated CGIs. To make the appropriate comparison, we modeled *in silico* the library construction undertaken by Cross *et al.* (22) by identifying MseI sites in the genomic sequence and generating fragments between MseI sites within or immediately flanking annotated CGIs (MseI-CGIs). Starting with the 27 801 computationally annotated CGIs from hg17, 42 450 MseI-CGIs were identified having an average length of 1106 bp (range 5–51, 939 bp). For additional comparison, a set of 5000 random loci, each 1000 bp in length and proportionally distributed across all 23 chromosomes, was also generated *in silico*. These three sets were then evaluated for the two above-mentioned metrics (Figure 3). Both criteria were met (CpG ratio > 0.6, G + C content > 0.5) for 64% of the physical loci. Although this was less than the 87% observed in the MseI-CGIs, this stood in stark contrast to the 2% observed in the random loci. Annotated CGI will, by definition, exceed both criteria but 13% of the MseI-CGIs do not, and this can be directly attributed to the non-CGI-containing flanking sequences imposed by the MseI positions. Given this, it is likely that the 36% of the physical CGIs not exceeding both criteria includes authentic CGI sequences. In contrast, 98% of the random loci fail to exceed both criteria.

We next determined the chromosomal distribution of the physical CGI loci and compared this to the distribution of MseI-CGIs (Figure 4). In total, 63% of the physical CGI loci show overlap with an MseI-CGI. The proportional representation of both sets relative to chromosome size were similar (Figure 4, right-hand side), suggesting that the clones isolated from the library are generally representative of annotated CGI. A similar over-representation (chromosomes 16,17 and 19) or under-representation (chromosomes 13, X, Y) of CGI on particular chromosomes was also observed in both sets.

Given that CGIs have been demonstrated to have a close relationship with the 5′ upstream region of genes, we next examined the distance of each CGI Library locus from the nearest TSS. Distance from a TSS to the annotated CGIs as well as to the random loci was also calculated and these results are summarized in Figure 5. For the CGI-derived MseI fragments, 47% are in the proximal promoter region (+200 bp to −1 kb), 41% of which directly overlap a TSS. An additional 12% are found in more distal promoter regions (+1 kb to −10 kb) and 14% are found further into the gene sequence, >1000 bp downstream from the TSS. Less than 10% of the annotated CGIs are found in regions far upstream of a TSS (−100k) in contrast to the randomly generated loci which are predominantly in these regions.

The CGI Library loci that overlapped a CGI-derived MseI fragment showed a similar preferential localization around the TSS and in promoter regions. The CGI Library loci that were not associated with the annotated CGI also showed a stronger association with TSSs than observed with the random loci. Although few were directly at a TSS, 25% were within the distal promoter region, compared to only 10% of the random loci. Furthermore, 43% of the random loci were >100 kb away from a TSS compared to only 25% of this subset of the CGI library loci.

To verify the overall ability to align these sequences to the human genome, a 12k microarray constructed from the CGI clones was hybridized with DNA fragments generated by sonication of isolated human genomic DNA. Fluorescence measurements from non-human probes on this array were

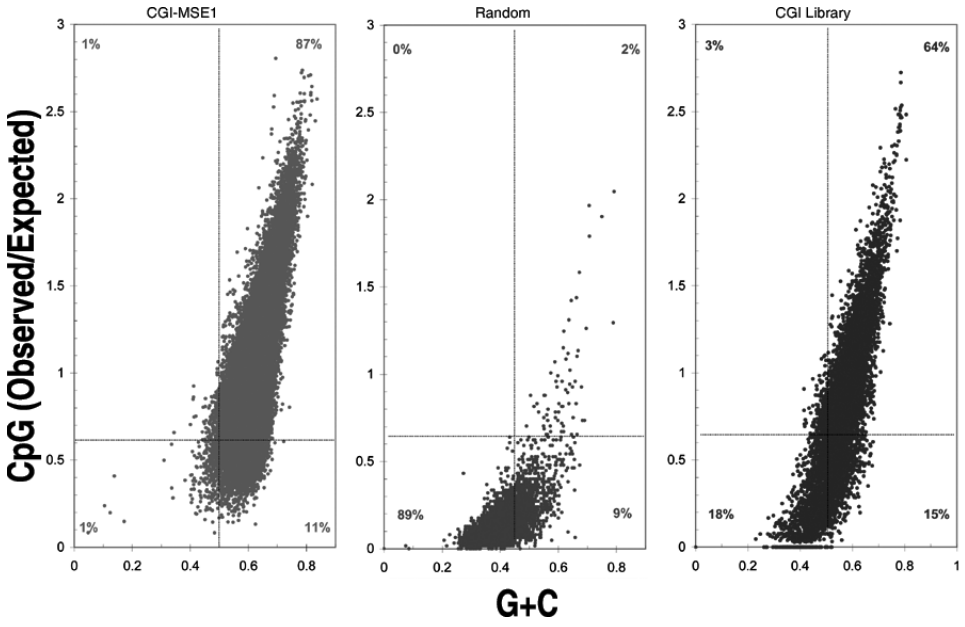

Figure 3. Evaluation of GC and CpG dinucleotide content. CGI Library Loci (right panel) were evaluated for G + C content and CpG dinucleotide content (expressed as a ratio of the expected frequency of 1/16). For comparison, MseI fragments containing annotated CGIs (left panel) and random loci (center panel) were also evaluated. Dotted lines indicate the values frequently used for assessment of CGIs (G + C > 0.5, CpG observed/expected > 0.6). The percentage of Loci in each quadrant is indicated.

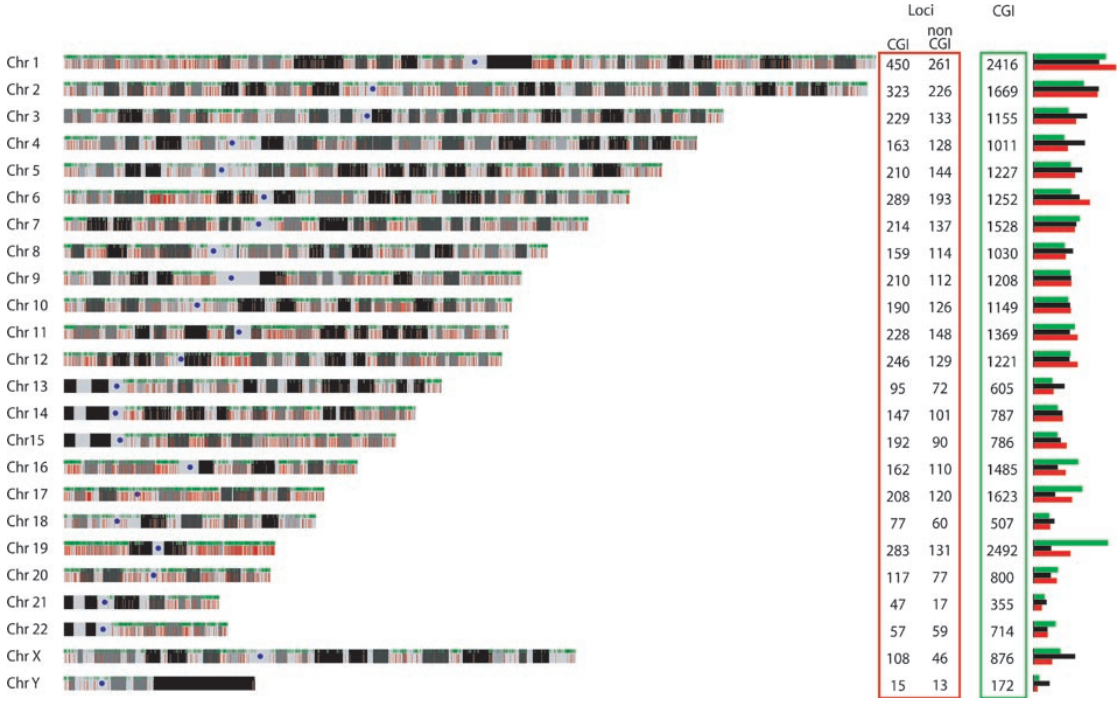

Figure 4. Distribution of CGIs across the human genome. A schematic diagram of the 23 human chromosomes is shown with Giemsa staining patterns in grayscale. Annotated CGIs are indicated in green (top of schematic diagram) and the number identified on each chromosome is indicated to the right (CGI). The position of each mapped locus is indicated in red (bottom of schematic diagram). The number identified on each chromosome is indicated to the right (Loci). The first column is the number of loci with a position that overlaps an Annotated CGI/MseI. The second column is the number of loci that do not overlap an annotated CGI/MseI. To the far right is indicated the proportional representation of the number of loci or annotated CGIs relative to chromosome length. Loci were also identified on mitochondrial DNA (14 loci) as well as on undesignated chromosome sequence collected into UCSC random sequence files (94 loci).

used as a measure of non-specific binding since the corresponding spike-in cDNA was not included in the hybridization mixture (Figure 6A). Signal from only 396 of the 12 196 CGI probes were within 2 SD of the mean signal intensity of the non-human DNA probes, suggesting that genomic DNA bound specifically to >97% of the CGI probes.

The array was hybridized with two independently labeled aliquots (100 ng each) of sonicated genomic DNA. An MA plot

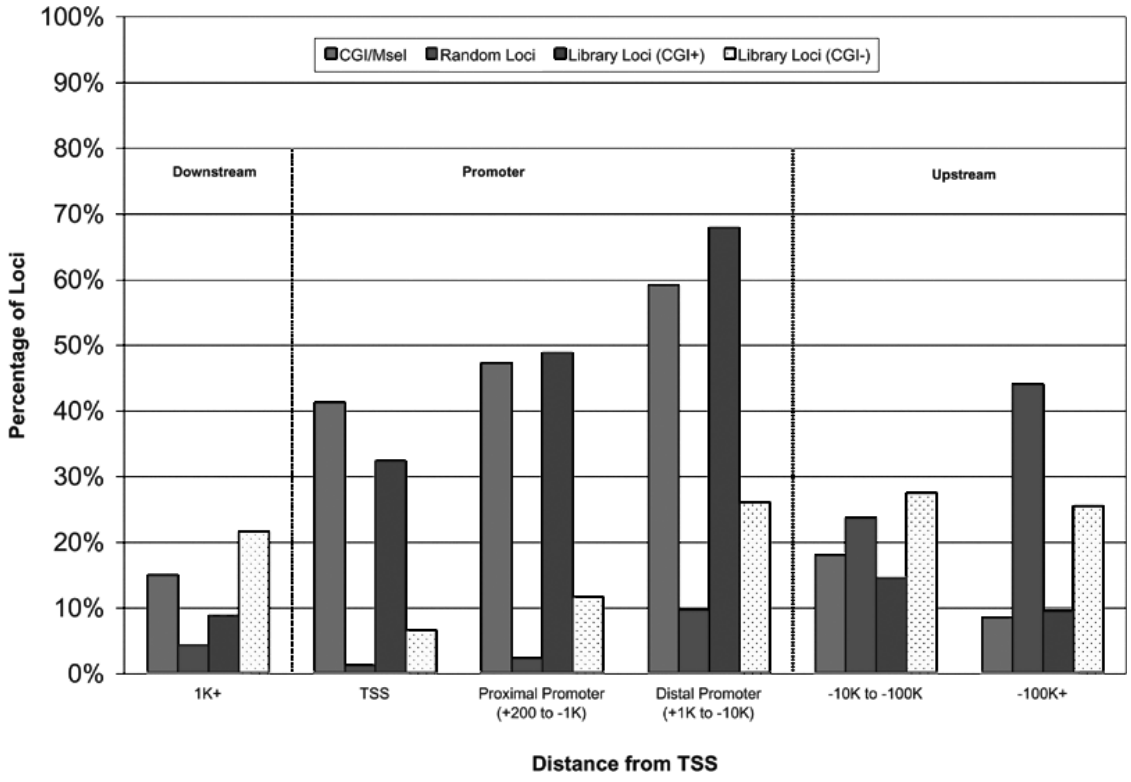

Figure 5. Position of Loci relative to gene TSSs. The distance to the nearest annotated TSS is shown for three sets of loci: the annotated CGI in the current build of the human genome (Hg17, May 2004); the random loci; and the loci derived from the CGI Library. The last set has been subdivided into loci which overlap an annotated CGI (solid) and those that do not (speckled). The percentage of loci in each set at various positions relative to the TSS is shown. (i) Promoter regions. Percentage of loci overlapping an annotated TSS, in the proximal promoter region (+200 to −1000 bp) and in the distal promoter region (+1000 to −10000 bp). (ii) Downstream (1000 bp or greater within a gene). (iii) Upstream between 10 and 100 kb upstream of a TSS or >100 kb upstream of a TSS.

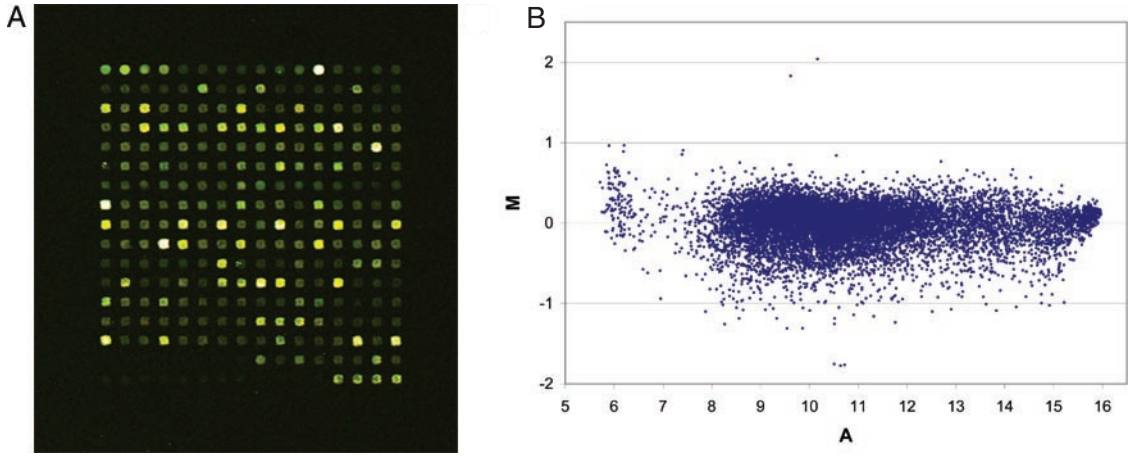

Figure 6. 12k CGI microarray. (**A**) A representative block from the CGI array after hybridization with 100 ng sonicated human genomic DNA. The faint spots in the lower left are the non-human DNA control probes. The corresponding spike-in controls were not added to the hybridization mixture. Arabidopsis controls were added to the hybridization and the Arabidopsis probes are in the lower right. (**B**) MA plot of differential expression (M) versus mean signal intensity (A) in $\log_2$-space for all CGI probes on the array. Only 32 of 12 196 probes display >2-fold differential expression ($|M| > 1$).

of background-corrected and normalized $\log_2$ signal versus $\log_2$ differential expression is shown in Figure 6B. Signal consistency is very strong between the two channels with >99.5% of CGI probes showing <2-fold differential expression ($|M| < 1$) and signal from the two channels showing a correlation (Pearson) of 0.99. The mean $\log_2$ signal intensity for all probes was 10.9 ± 1.9; in total 85% of CGI probes had signals within two orders of magnitude range.

To facilitate examination of the CGI Library clone alignments to the genome, a CGI Library Browser has been constructed and is available at http://derlab.med.utoronto.ca/CpGIslands/ (Figure 7) Clone identifiers corresponding to

Figure 7. CGI library browser. An online CGI Library Browser (http://derlab.med.utoronto.ca/CpGIslands/) has been constructed to allow users of the CGI Promoter Microarrays as well as users interested in the CGI Library to explore alignments of the clones to the human genome. For each clone, alignments to the genome are listed and displayed diagrammatically, including the relative positions of annotated CGI and nearby genes.

spots on the 12k CGI Microarray can be entered to obtain a graphical view of the genomic alignment, as well as the relative positions of upstream and downstream genes. Other features of the browser include the ability to view other clones aligning to the same locus and to identify clones aligning near a specified gene. Links have been created to map directly to the UCSC genome browser (33) to allow exploration of other genomic features near the locus of interest.

DISCUSSION

Arrays constructed from a CGI library (22) have been applied successfully to study DNA methylation status and to identify genomic fragments immunoprecipitated with antibodies against several target proteins (5,24,32,34–36). The rationale for using a CGI array to analyze ChIP DNA is based on the association between CGIs and gene TSSs (16,17) as well as the hypothesis that CpG conservation in these regions results from protein–chromatin interactions which prevent methylation and subsequent CpG degeneration (19). The CGI library from which the probes for the array were selected was created by isolating genomic fragments that have a high G + C content, are rich in CpG dinucleotides, but are poorly methylated (22). The clones used as probes on earlier arrays were selected randomly from this library and therefore their sequence and identity was unknown. In contrast to arrays constructed from defined genomic fragments, it had been necessary to identify probes of interest post-hybridization by sequencing the clone, aligning to known sequences, then identifying associated genes. This study provides a comprehensive analysis of a set of 20 736 clones isolated from the CGI library validating the approach taken by Cross *et al*. (22) in construction of the library as well as the use of this library for construction of a microarray suitable for ChIP-chip analysis.

Initial attempts to align clone sequences to genomic DNA utilized the sequence information which was publicly available for the 12k set (http://www.sanger.ac.uk/HGP/cgi.shtml). While sequence information was available for only 60% of the clones in this data set, nearly 85% of the clones had usable sequence following resequencing in this study. Furthermore, the average read length in the original data set was 215 bp compared to 414 bp in our sequences. The higher quality of sequence data probably contributed to a higher percentage of successful alignments. More importantly though, since discrepancies were observed between the two data sets (data not shown) and our sequence data is known to directly correspond to the clones used for array construction, this ensures an accurate annotation of the probes on the microarray.

Despite resequencing, ~15% of the clones did not have adequate sequence information to generate genomic alignments. This included clones in which no sequence information was obtained as well as clones with <50 bp of quality sequence (PHRED score >20), which would generate short, less informative alignments. This was not unexpected and may be due to lack of an insert in the clone, or to the presence of multiple clones in some wells, preventing proper sequencing of the inserts.

Two sets of clones were sequenced and analyzed. The first set of 12 192 clones (12k set) was obtained directly from the Sanger Institute. The second subset of 8544 clones (9k set) had been previously isolated by Huang *et al*. (23) from the same library. The clones from this set have been used in the construction of various CGI arrays (5,24,32). The quality of the sequence information in the 9k set was superior to the 12k set. Fewer clones in the 9k set returned no sequence, and the read lengths were generally longer. This is probably due to the fact that this set had been prescreened with human Cot-1 DNA to remove sequences with a high degree of repeat elements. Consequently, a higher percentage of clones in the 9k set aligned to genomic sequence than in the 12k set. CGI arrays constructed from the 12k set have been available through the UHN Microarray Centre (http://www.microarray.ca/) since 2003. Arrays constructed from the full set of 20 736 clones are now being produced.

We have used the BLAT software (27) to align sequences to genomic DNA, both for its speed and because it is well-suited for identifying highly similar sequences. Some post-alignment processing was necessary to separate out partial alignments on a single chromosome that had been combined by BLAT's stitching routine. Using BLAT, it was possible to align 84% of all sequences. The 16% that did not align can be partially accounted for by their shorter sequence lengths although there did exist long sequences that did not align. In addition, sequences that did not have long contiguous stretches of quality sequence may have not aligned due to masking of the poor quality bases. Use of other alignment algorithms may be more suitable for these sequences, but the results would include similar or homologous regions that would not necessarily represent either the genomic origin of a clone, nor an area that would hybridize against the clone on an array.

Clone alignments were constructed after aligning each read separately, allowing the genomic sequence to act as a scaffold to determine the positioning of the two reads with respect to each other. In clones with non-overlapping ends, the gap was therefore defined by the genomic scaffold. There were also instances of alignments of one read in the absence of a nearby alignment of the other read. Although these alignments may not represent the genomic origin of a particular clone, they do indicate a potential hybridization target in genomic DNA against the clone when acting as a probe on a microarray. Such clones may have arisen by ligation of two or more distinct fragments into the same plasmid. Evidence suggesting this is observed in 2% of the clones (213 clones), where the end reads align to distinct genomic positions, often on different chromosome. The absence of the other read's contribution to the alignment is noted in the annotation available through the web browser.

The genomic scaffold was also used to identify redundancy in the clones derived from the CGI Library. CGI Library Loci were generated by combining overlapping alignments and identifying all clones contributing to each locus. There is the possibility also that a locus might capture contiguous clones in cases where incompletely digested fragments were captured, but in the absence of this, contiguous clones will be represented by adjacent loci. In this way, 9595 loci were defined, slightly less than one-half of the total number of clones available and slightly more than one-half of the clones with sequence. This number is reduced to 7184 when eliminating shorter loci resulting from partial sequence alignments. In this way, it was determined that only ~30% of the clones were unique, the other 70% showing some degree of redundancy.

Interestingly, the 12k and 9k sets show only a small degree of overlap which was not expected given the degree of redundancy in each set. Of the 9595 loci, only 753 were present in both sets. This finding suggests there may be value in continuing to isolate clones from the library. Approximately 60% of the loci that we have identified are within the MseI restriction fragments associated with the annotated CGI. This number corresponds well with the observation that 64% of the loci evaluate as a CGI based on G + C content and CpG dinucleotide frequency. CGIs by definition are regions of CpG conservation due to the absence of methylation and will therefore have high G + C content and a CpG frequency that approaches the expected random frequency of 1/16. Identification of CGI in promoter regions or in genomic sequence exploits algorithms that evaluate when these parameters exceed arbitrary values designed to minimize signal-to-noise. Within the 40% of the loci that do not correspond to the annotated CGI-associated MseI restriction fragments, there may exist CGIs not identified by these algorithms. Alternately, some of these loci may represent noise in the original library due to capture of fragments not expected from the selection procedure.

The documented association of CGIs with regions upstream of genes is based on bioinformatic analysis of the genome (14,17,21) and is dependent on the analysis of the nucleotide sequence alone. We have approached this from a different direction, mapping the CGI library clones that were selected based on sequence characteristics and methylation status, to the genomic sequence. Approximately 60% of the CGI Library loci in the subset that overlap annotated CpG sites are in the distal promoter region or closer to a TSS. This is nearly identical to the distribution of the annotated CGIs, supporting the idea that CGIs are associated with TSSs.

In addition to collecting CGIs near the TSS, the CGI library also contains fragments representing CGIs that are distant from a gene. Potentially these islands may represent regions to which transcriptional enhancers and repressors may bind outside the proximal or distal promoter regions. Alternately, they may be a part of promoter regions in proximity to an unknown TSS and may ultimately play a role in identifying new genes (37).

Nearly 25% of the loci in the subset that does not overlap annotated CpG sites are within the distal promoter region or closer. This is a significant enrichment relative to random loci in this region (~10%). Furthermore, while >40% of random loci are 100 kb or further from a TSS, only ~25% of this second subset is found here. Despite not being associated with annotated CGIs, this suggests that these loci are not simply due to non-specific collection of random fragments into the library. Possibly, the CpG content is too low to characterize these regions as CGIs by the established criteria, but still high enough to have been effectively methylated and isolated during the procedure used to create the CGI library.

There have been a variety of approaches taken towards the design of microarrays suitable for analysis of ChIP DNA. Proximal promoter arrays constructed using regions directly upstream of known TSSs (8,9) can identify binding to transcriptional regulatory elements which exist close to the gene. It has been well established though that regulatory factors can bind at more distal regions, and even within introns, exons or downstream of a gene (10,11,38). More comprehensive tiling arrays have been constructed targeting intergenic regions (6), distinct loci (10,11) and the full genome (12,13). The CGI array approach is based primarily on the association of CGIs with regulatory activity (16,17), but in contrast to the proximal promoter arrays the location of CGIs relative to the TSS is not limited to 1000 bp upstream. While many of the loci identified in this study are still within 10 kb of a TSS, they also occur at far more distant regions. Although full genome tiling arrays obviously represent the most comprehensive tool for global methylation or protein–chromatin interaction studies, technical issues such as costs, multiple array platforms and data analysis software, represent aspects which currently preclude widespread use of these arrays for a typical molecular biology laboratory. However, it is likely that all these array types

represent complementary approaches to studying global epigenetic mechanisms and transcription factor binding sites.

The primary focus of this study is to annotate clones isolated from the CGI library and the CGI arrays which use these clones as probes. While more extensive studies are in progress to further evaluate these arrays, we conducted a straightforward experiment to examine probe hybridization to the 12k CGI array. At least 97% of the probes bound sonicated genomic DNA. Furthermore, the degree of binding observed after hybridization with equal amounts of genomic DNA in both channels was highly consistent, with only 32 of 12 192 CGI probes showing differential hybridization. The signal obtained from each probe varied over ~2 orders of magnitude, probably due to differences in hybridization strength, dye incorporation and hybridization specificity between different DNA fragments.

The data produced in this analysis is available online through the CpG Island Library browser (http://derlab.med.utoronto.ca/CpGIslands/). This has been designed to allow detailed examination of individual probes on the CGI array as well as providing a means to summarize information for a list of probes. The alignments detailed in this browser include complete and near-complete alignments of clone sequences to genomic loci, as well as partial alignments that may require consideration when analyzing hybridization results. This interactive tool will facilitate analysis of ChIP-chip experiments by allowing faster identification of targets. In addition, redundant probes on the array have been identified and probes that align near a gene of interest as well as the spatial relationship of the alignment to the gene can be easily determined.

The CGI library is expected to contain only a subset of all genomic CGI, those that are unmethylated in whole blood genomic DNA from which the library was constructed (22). The analysis presented here will contribute to the production of next generation CGI arrays by allowing removal of redundant probes, enrichment with unique clones, and the inclusion of probes designed against other regulatory regions which complement this CGI approach. A clearer understanding of the nature of the CGI probes will also facilitate the design of coordinated expression arrays that can be used in conjunction with ChIP-chip studies with CGI arrays to study and define transcriptional networks. Furthermore, the development of high quality CGI arrays will contribute to more accurate analysis of global methylation status and the relationship of epigenetic mechanisms to these networks.

ACKNOWLEDGEMENTS

This project has been funded in part with Federal funds from the National Cancer Institute and National Institutes of Health, under Contract No. N01-C0-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the US Government. This study was funded by grants from CIHR and Genome Canada to S.D.D. Computational analysis was in part supported by NSERC and IRIS grants to I.J. Sequencing was carried out by the British Columbia Genome Sciences Centre, Suite 100, 570 West 7th Ave, Vancouver, BC V5Z 4S6, Canada. Funding to pay the Open Access publication charges for this article was provided by Genome Canada.

Conflict of interest statement. None declared.

REFERENCES

1. Oberley,M.J., Tsao,J., Yau,P. and Farnham,P.J. (2004) High-throughput screening of chromatin immunoprecipitates using CpG-island microarrays. *Methods Enzymol*., **376**, 315–334.
2. Kuras,L. (2004) Characterization of protein–DNA association *in vivo* by chromatin immunoprecipitation. *Methods Mol. Biol*., **284**, 147–162.
3. Im,H., Grass,J.A., Johnson,K.D., Boyer,M.E., Wu,J. and Bresnick,E.H. (2004) Measurement of protein–DNA interactions *in vivo* by chromatin immunoprecipitation. *Methods Mol. Biol*., **284**, 129–146.
4. Yan,P.S., Wei,S.H. and Huang,T.H. (2002) Differential methylation hybridization using CpG island arrays. *Methods Mol. Biol*., **200**, 87–100.
5. Weinmann,A.S., Yan,P.S., Oberley,M.J., Huang,T.H. and Farnham,P.J. (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev*., **16**, 235–244.
6. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
7. Ren,B., Cam,H., Takahashi,Y., Volkert,T., Terragni,J., Young,R.A. and Dynlacht,B.D. (2002) E2F integrates cell cycle progression with DNA repair, replication and G(2)/M checkpoints. *Genes Dev*., **16**, 245–256.
8. Li,Z., Van Calcar,S., Qu,C., Cavenee,W.K., Zhang,M.Q. and Ren,B. (2003) A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl Acad. Sci. USA*, **100**, 8164–8169.
9. Odom,D.T., Zizlsperger,N., Gordon,D.B., Bell,G.W., Rinaldi,N.J., Murray,H.L., Volkert,T.L., Schreiber,J., Rolfe,P.A., Gifford,D.K. *et al*. (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.
10. Horak,C.E., Mahajan,M.C., Luscombe,N.M., Gerstein,M., Weissman,S.M. and Snyder,M. (2002) GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc. Natl Acad. Sci. USA*, **99**, 2924–2929.
11. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J. *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
12. Ishkanian,A.S., Malloff,C.A., Watson,S.K., DeLeeuw,R.J., Chi,B., Coe,B.P., Snijders,A., Albertson,D.G., Pinkel,D., Marra,M.A. *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genet*., **36**, 299–303.
13. Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M., Weissman,S. *et al*. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
14. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol*., **196**, 261–282.
15. Sved,J. and Bird,A. (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl Acad. Sci. USA*, **87**, 4692–4696.
16. Antequera,F. and Bird,A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA*, **90**, 11995–11999.
17. Ioshikhes,I.P. and Zhang,M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nature Genet*., **26**, 61–63.
18. Murakami,K., Kojima,T. and Sakaki,Y. (2004) Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression. *BMC Genomics*, **5**, 16.
19. Antequera,F. (2003) Structure, function and evolution of CpG island promoters. *Cell. Mol. Life Sci*., **60**, 1647–1658.
20. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev*., **16**, 6–21.
21. Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 3740–3745.

22. Cross,S.H., Charlton,J.A., Nan,X. and Bird,A.P. (1994) Purification of CpG islands using a methylated DNA binding column. *Nature Genet*., **6**, 236–244.
23. Huang,T.H., Perry,M.R. and Laux,D.E. (1999) Methylation profiling of CpG islands in human breast cancer cells. *Hum. Mol. Genet*., **8**, 459–470.
24. Mao,D.Y., Watson,J.D., Yan,P.S., Barsyte-Lovejoy,D., Khosravi,F., Wong,W.W., Farnham,P.J., Huang,T.H. and Penn,L.Z. (2003) Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr. Biol*., **13**, 882–886.
25. Cross,S.H., Clark,V.H. and Bird,A.P. (1999) Isolation of CpG islands from large genomic clones. *Nucleic Acids Res*., **27**, 2099–2107.
26. Weinmann,A.S., Bartley,S.M., Zhang,T., Zhang,M.Q. and Farnham,P.J. (2001) Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Mol. Cell. Biol*., **21**, 6820–6832.
27. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
28. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.
29. Workman,C., Jensen,L.J., Jarmer,H., Berka,R., Gautier,L., Nielser,H.B., Saxild,H.H., Nielsen,C., Brunak,S. and Knudsen,S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol*., **3**, research0048.
30. Smyth,G. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*., **3**, 1–26.
31. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.
32. Yan,P.S., Chen,C.M., Shi,H., Rahmatpanah,F., Wei,S.H., Caldwell,C.W. and Huang,T.H. (2001) Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Res*., **61**, 8375–8380.
33. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.
34. Yan,P.S., Efferth,T., Chen,H.L., Lin,J., Rodel,F., Fuzesi,L. and Huang,T.H. (2002) Use of CpG island microarrays to identify colorectal tumors with a high degree of concurrent methylation. *Methods*, **27**, 162–169.
35. Yan,P.S., Shi,H., Rahmatpanah,F., Hsiau,T.H., Hsiau,A.H., Leu,Y.W., Liu,J.C. and Huang,T.H. (2003) Differential distribution of DNA methylation within the RASSF1A CpG island in breast cancer. *Cancer Res*., **63**, 6178–6186.
36. Testa,A., Donati,G., Yan,P., Romani,F., Huang,T.H., Vigano,M.A. and Mantovani,R. (2005) ChIP on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters. *J Biol Chem*., **280**, 13606–13615.
37. Cross,S.H., Clark,V.H., Simmen,M.W., Bickmore,W.A., Maroon,H., Langford,C.F., Carter,N.P. and Bird,A.P. (2000) CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. *Mamm. Genome*, **11**, 373–383.
38. Taverner,N.V., Smith,J.C. and Wardle,F.C. (2004) Identifying transcriptional targets. *Genome Biol*., **5**, 210.