Supplementary Materials

1. Microarray response

A control experiment defined the longest distance from a probe at which endonuclease cuts can be measured using the microarray method. DNA was cleaved with PmeI, and processed by isothermal whole genome amplification followed by microarray analysis using uncleaved, amplified DNA as a reference channel. All probes containing a single predicted PmeI cut, and not bounded by another cut within a distance of +- 40 kb were plotted at x=0 in an xy plot. Other probes proximal to the cut site (upstream as well as downstream) were plotted according to their position in the x axis. The ratio of the two microarray channels (cleaved and uncleaved DNA) was plotted in the y axis (**Supplementary Figure S1**). The deflection of the y axis in the xy plot indicates that a single endonuclease cut produces large changes in the ratio (y) within a window of +/- 3.0 kb, with the most pronounced deflection of the ratio occurring within a window of +/- 1.2 kb.

2. Validation of microarray-based observations using bisulfite DNA sequencing

We selected a total of 74 probe loci that showed DNA methylation changes in tumors, and examined the DNA methylation status using bisulfite DNA sequencing across a total of 12 experiments, for a total of 207 probe validation data points. DNA sequencing was performed using two different experimental approaches. In the first approach, bisulfite-treated DNA was used to amplify by PCR the genomic regions of interest, and the PCR amplicons were cloned. Individual clones were processed for Sanger sequencing in both strand orientations. In the second approach, bisulfite-treated DNA was used to amplify by PCR the genomic regions of interest, and the PCR amplicons were then transcribed to generate complementary DNA using reagents provided by Sequenom, Inc. as part of their EpiTYPER kit. The RNA was then cleaved with ribonuclease A, and subjected to mass spectrometry analysis. Using software provided by Sequenom, the mass spectrograms were processed to generate a fractional value of DNA methylation between 0.0 and 1.0. When multiple probes associated with a single CpG island were averaged, the concordance of the microarray calls and the bisulfite sequencing results was 87.6% (See also **Figure 3**)

3. Plotting the data

3.1. Per-experiment plots

Given that each probe in the microarray is annotated with its association to the proximal genomic elements (repetitive element category, gene, miRNA) for every experiment in the library a query is issued to retrieve from our lab's database a subset of probes in the vicinity of a specific element. A set of values from these probes is then averaged per experiment and plotted accordingly. (Note, instead of average, any other function can be applied here) This is repeated for every experiment and for every category requested. **Supplementary Figure S4A** is an example of this plot for 4 categories of genomic compartments. Contrast with **Supplementary figure S4B**, which shows the same information in per-category view.

A standard boxplot implementation included in R programming language was embedded in a custom script to generate these plots.

3.2. Per-category plots

An alternative to the plots described in supplementary section 3.1 are the per-category plots, devised to simplify the presentation of information especially when many categories of repetitive elements are to be plotted. For these plots, once an average of a given category of probes is calculated for all experiments, a box-and-whisker plot is then generated to summarize these values for experiment subsets: normals (top), non-tumor adjacents, tumor and sperm (bottom). **Supplementary Figure S4B** is an example of this plot for 4 categories of genomic compartments. Contrast with **Supplementary Figure S4A**, which shows the same information in per-experiment mode.
A standard boxplot implementation included in R programming language was embedded in a custom script to generate these plots.

3.3. On Order of experiments

The experiments are always grouped (top to bottom) into normal, non-tumor adjacent, tumor and sperm-replicate classes. The order of experiments across all plots (unless stated otherwise) is kept constant. The order has been established based on the difference of most informative category of L1P (**Supplementary Table 3**) versus the most stable across all experiments categories of repetitive elements: AluSq and DNA transposons. **Supplementary Figure S3A and S3B** depicts the element values used.

3.4. Shannon Information Value

The order of categories in the legend of per-experiment plot and the category-list of per-category plots is not accidental. The categories are ordered based on the extent of their variation using Shannon information content metric[1] . Only Normal and Tumor experiments were used to establish the order of categories. Specifically, once the per-category values (average methylation for **Figure 4A and 4B**) across normal (10 values per category) and tumor (33 values per category) experiments are calculated, a Shannon information measure function is applied to the distribution of 43 values.

The Shannon Information measure is a foundation of modern Information theory and was devised to estimate the minimum number of bits needed to encode sentence or a string of characters of text, if one wanted to transmit such string digitally. The information measure takes into consideration the frequency of the symbols. As a result, a string made up of the same symbol would require a very simple encoding using one bit of information, whereas a string made up of all the letters in the alphabet would need considerably more bits to represent all the letters unambiguously.

Analogously, the 43 values can be considered as the individual letters of Shannon's string. Shannon's entropy measures how dissimilar the 43 values are from each other. The more dissimilar, the more information is in the set.

[1] http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

The categories are listed from lowest information content (top) to the highest information content (bottom). The most informative categories are highlighted using colors. A custom R script was used to generate the plots and calculate the information content.