SUPPLEMENTAL INFORMATION

The epigenetic landscape of *Alu* repeats delineates the structural and functional genomic architecture of colon cancer cells

Mireia Jordà et al.

Contents

Supplemental Material

Tissues and cell lines

Three colorectal tumors and their paired non-adjacent normal colonic mucosa as well as the corresponding whole blood DNA were used in this analysis. Samples were obtained from the Hospital Germans Trias i Pujol following a protocol approved by the Local Ethics Committee (Ref. CEIC EO-11-134). Human colon cancer cell lines HCT116, SW480, CaCo-2, LoVo and HT29 were obtained from the American Type Culture Collection (ATCC, Manassas, VA), except DKO (Rhee et al. 2002), generously provided by B. Vogelstein. All cell lines were grown in DMEM:F12 medium plus 10% heat-inactivated FBS, 2mM L-glutamine and 1 mM pyruvate (Gibco, CA, USA) at 37 ºC in 5% CO_2.

Supplemental Methods

Next generation Sequencing of UnMethylated *Alu* (NSUMA)

The method is based in the AUMA technique (Rodriguez et al. 2008), although important modifications were introduced to expand significantly the genomic coverage and to introduce internal controls that allow normalization and control of some technical biases. One microgram of DNA was digested for 6 h at 25 °C with the methylation-sensitive restriction endonuclease SmaI (Roche Diagnostics GmH, Mannheim, Germany) leaving blunt ends (CCC/GGG), followed by a second digestion with the methylation-insensitive restriction enzyme MseI (T/TAA) (16h at 37 °C, Roche Diagnostics GmH, Mannheim, Germany) that leaves sticky ends. Adapters blunt-SmaI (ADPT-S1 GATAGTATGCCCGGGTGA plus the 5’ phosphorylated ADPT-S2 TCACCCGGGCATAC) and sticky-MseI (ADPT-M1 CTGAGGCTGGATCCCTG plus the 5’ phosphorylated ADPT-M2 TACAGGGATCCAGCCTCAG) were prepared by incubating the two oligonucleotides for 2 min at 65ºC and then cooling to room temperature for 30-60 min. Digested DNA and 2nmol of blunt and sticky adapters were ligated overnight at 16 °C using T4 DNA ligase (New England Biolabs, Beverly, MA). The product was purified using the Illustra GFX Purification kit (GE Healthcare, Buckinghamshire, UK) and eluted in 200 ul of bidistilled water. The ligation product consists of three types of molecules according to the flanking sites SmaI-SmaI, SmaI-MseI and MseI-MseI. Only the products containing SmaI adapters represent unmethylated fragments (Figure 1A). Next, a PCR (95 ºC 2 min; 95 ºC 30 sec, 60 ºC 1 min, 72ºC 1min for 30 cycles; 72 ºC 5min) was performed using two primers, one that anneals the MseI adapter (ADPT-M1A CTGAGGCTGGATCCCTGTAA) and another one homologous to the SmaI adapter plus TT at the

3’ end (ADPT-S1TT GATAGTATGCCCGGGTGAGGGTT) to enrich in *Alu* sequences (Figure 1A and Supplemental Fig. S1). The final product appeared as a smear when run in an agarose gel and most of the amplicons ranged from 50 bp to 1000 bp (data not shown). Primer sequences are reported in Supplemental Table S16.

Illumina sequencing

The NSUMA PCR product was sheared by sonication with a Bioruptor (Diagenode, Liege, Belgium) to a size of 100-300 bp. DNA fragments were blunt end repaired with T4 DNA polymerase and Klenow fragment (NEB, MA, USA) and purified with a QIAquick PCR purification kit (Qiagen, Venlo, Netherlands). Thereafter, 3’-adenylation was performed by incubation with dATP and the Klenow (3´→5´ exo-) fragment of DNA polymerase I (NEB). DNA was purified using MinElute spin columns (Qiagen) and ligated to double-stranded adapters (Illumina 1/2 adapters, Supplemental Table 15) using rapid T4 DNA ligase (NEB). The sample was purified again using a MinElute spin column and run on a 2% agarose gel, and fragments in the size range of interest, 150 bp plus 65 bp of adapters, were excised with a sterile single-use scalpel and recovered using the QIAquick gel extraction system (Qiagen). Then, adapter-ligated fragments were enriched, and adapters were extended, by selectively amplifying with an 18-cycle PCR reaction using Phusion DNA polymerase (Thermo Fisher Scientific, USA), and Illumina 3/4 amplification primers (Supplemental Table 15) (library size of approximately 150 plus 92 bp of adapters). Finally, the quality of libraries was confirmed on the Agilent Technologies 2100 Bioanalyzer and by cloning into a blunt TOPO vector. Six colonies were sequenced by conventional Sanger method to verify correct adapter ligation and sequence match. Libraries were quantified by TaqMan Universal PCR No AmpErase kit (Applied Biosystems, Foster City, CA, USA). DNA was loaded into a single read (SR) flow cell for cluster generation using a SR-cluster generation kit v4 (Illumina, CA, USA). During this process, DNA molecules were immobilized on the surface of the flow cell, amplified *in situ* to create same-sequence containing clusters, and following surface blocking and DNA denaturation, binding of sequencing primer was performed. The flow cell was then mounted on a Genome Analyzer II or a HiSeq2000 instrument for sequencing, and 35-50 sequencing cycles were carried out using v4 SBS kits. Each flow cell contained a PhiX control lane (loaded at a concentration of 4pM) that was used to monitor run quality.

NSUMA reads processing, alignment and counting

Pre-processing. Reads sequences were obtained from the Illumina instruments in fastaQ or qseq format and were pre-processed in different steps in order to remove misleading reads and improve the mapping accuracy. Bases at the 3’ end of the reads with low quality (PHRED score ≤ 20) were end clipped. The adapters sequence used in the PCR amplification were also removed using iterative searches of 5’ to 3’ fragments. Ambiguous bases (Ns) at both ends of the read were also removed. Finally, reads shorter than 10 bp or with more than 30% of ambiguous bases were filtered out from the analysis.

Mapping. Once trimmed and cleaned, reads were mapped to the human reference genome (build GRCh37/hg19) with Bowtie 0.12.7 using the following parameters: -S –p 8 –v 2 –phred64-quals –best –l 28 –k 2 (Langmead et al. 2009). Parameters description:

-v 2 to restrict the maximum number of mismatches to 2

-l 28 to seed with 28 bases on the high quality end of the read

-k 2 --best to report 2 valid alignments; as the --best flag is coupled to the -k, if more than one valid alignment exists, bowtie will report two of them belonging to the best alignment stratum.

Post-processing. Once mapped, the process reported three types of reads: unmapped, uniquely mapped and having two valid alignments. We further analyzed the latter to check whether they could be rescued and assigned to a unique location. This process was performed with a custom-made desambiguator using the following algorithm:

```
- for each read with two reported alignments, get the best and second hit chromosome name and stratum (i.e. the number of mismatches)
- if the strata differ
  - if one of the alignments points to a unknown or random chromosome (chrUn, chr_random; clone contigs not placed on a specific chromosome)
    - retrieve the other alignment and report the read as unique
  - else
    - report the alignment at the best stratum (i.e. with the lowest number of mismatches) as unique
- else
  - if one of them is located in a unknown or random chromosome (chrUn, chr_random)
    - retrieve the other alignment and report the read as unique
  - else
    - set the read as ambiguously mapped and therefore discard both alignments from further analysis.
```

The information was stored using the standard SAM/BAM format (Li et al. 2009).

Counting of reads in regions of interest (NSUMA universe) was performed using coverageBed of the bedtools suite 2.16.2 (Dale et al. 2011).

Definition of NSUMA universe and annotation of reads to NSUMA amplicons

The coordinates of canonical NSUMA amplicons were determined based on the distribution of the extended SmaI site (CCCGGGTT or AACCCGGG) and the MseI site (TTAA) along the human genome (GRCh37/hg19 assembly). Only amplicons with a length (distance between the two restriction sites) of 20-1,000 bp were considered (n=144,108). The canonical NSUMA universe was constituted by amplicons flanked by the SmaI restriction sequence and one MseI site (n=143,823) or by two SmaI restriction sites (n=285).

Virtual amplicons were labeled with the identity of the genomic element where the SmaI site was located. Amplicons containing two SmaI sites were considered twice for genome element annotation. RepeatMasker database (hg19 - Feb 2009 - RepeatMasker open-3.3.0 - Repeat Library 20120124) was used to annotate the repeat elements (Supplemental Table 1). The list and main features of the NSUMA amplicons are described in Supplemental Data S1.

A scheme summarizing the NSUMA data processing is shown in Supplemental Fig. S2. Processing of non canonical amplicons data is described in section: *“Analysis of chromosome imbalances based on NSUMA non-canonical amplicons”* in page 8 of this document.

NSUMA global output

NSUMA of 15 samples generated 232 million reads ≥35 bases long (Supplemental table 2). After quality control filtering, an average of 14 million reads per sample could be mapped to the human reference genome (NCBI build GRCh37/hg19). Half of the filtered and aligned reads were mapped to two or more loci (Supplemental Table 2) and discarded for further analyses due to its ambiguous origin. Most of ambiguous reads aligned to *Alu* repeats (data not shown). In average, 1 million reads per sample were mapped unambiguously to the NSUMA canonical amplicons (Supplemental Table 2) and were annotated and assigned to the corresponding SmaI site and the respective enclosing genomic element (see “Definition of NSUMA universe and annotation of reads to NSUMA amplicons” section). The number of reads mapping unambiguously inside every canonical amplicon was considered as a relative measure of unmethylation when two or more samples are compared.

Quantification and normalization

The competitive nature of the amplification preserves to some point the quantitativeness of the determination, allowing the comparison of the same amplicon among different samples. Nevertheless, different amplicons cannot be compared among them as multiple factors may affect the representativeness (i.e.: PCR efficiency, mappability, amplicon size, base composition, etc). In addition to technical factors that affect the overall performance of the technique, wide variations in the global levels and distribution of DNA methylation represent a handicap for the analysis of methylation using genome-scale screening approaches. To overcome some of these limitations we took advantage of the presence of multiple amplicons representing CpG islands to assess the overall performance and to normalize the reads. A total of 150 amplicons affecting CpG islands were selected based on their NSUMA representativeness (>10 reads in all the samples) and unmethylated state as determined by bisulfite sequencing (see Supplemental methods and Supplemental Fig. S6). The normalization factor was calculated based on the cumulated number of reads in the 150 reference amplicons, setting to 1 that of the sample with the lowest number of reads (Supplemental Table 3).

The inclusion of a large number of internal controls is instrumental to assess the efficiency and reproducibility of NSUMA enzymatic processes. Important differences in the representation of the selected controls among different samples are indicative of potential technical failures (e.g.: poor DNA quality, incomplete enzymatic digestion, etc.) and data should be discarded.

Assessment of potential biases in NSUMA representations

Base composition and amplicon length may have an effect on NSUMA representation (number of reads). Due to the comparative nature of the analysis, this effect should not affect the differential representation of an amplicon in different samples provided the bias is the same among the compared samples. This was verified by analysis of the distribution of reads in amplicons stratified by element type (*Alu* repeats, CpG islands, etc.) in relation to GC content and amplicon length for each sample (Supplemental Fig. S7-S8), as well as plotting the fold change ratio in biological and technical replicates (Supplemental Fig. S9). All these analyses showed very similar profiles indicating that potential biases had comparable effects on all experiments, and therefore were very unlikely to have an impact on the observed differences.

Representation and statistical analysis of differential methylation

We used DESeq package (Anders and Huber 2010) to find differentially methylated amplicons between samples. This package is based on a negative binomial distribution. Only canonical amplicons with at least one normalized read in at least one of the samples were included in the analysis. The normalization process within DESeq was not applied. We used the *pooled* method that uses all the samples from all conditions to estimate a single pooled empirical dispersion value. Differentially represented amplicons with an adjusted p-value < 0.05 and a fold change greater than 2 or less than -2 were considered as differentially methylated.

To compare NSUMA profiles among different samples, the *Alu* Differential Methylation Ratio (*Alu* DMR) was calculated as the log 2 of the ratio of the normalized reads +1 of the two samples: log 2 ((*nreads* Normal +1)/(*nreads* Tumor +1)).

To compare WGBS profiles, the *Alu* Differential Methylation (*Alu* DM) was calculated as the subtraction of normal-tumor mean beta values of the CpGs within each element. Only *Alu* repeats with a minimum of 3 informative CpGs were considered. The LINE DM was calculated in the same way considering LINE repeats with a minimum of 5 informative CpGs. The minimum number of informative CpGs was chosen arbitrarily to have groups of balanced size.

Nomenclature of genomic compartments based on differential methylation determined by NSUMA

To simplify the analysis of *Alu* features in relation to differential DNA methylation, *Alu* elements were classified and labeled based on their representation in NSUMA. Data are provided in Supplemental Table S4 and Supplemental Table S6. The nomenclature is as follows:

NSUMA virtual universe, *Alu* repeats represented by a canonical amplicon in NSUMA

Outside NSUMA (no NSUMA), *Alu* repeats not represented by a canonical amplicon in NSUMA

NSUMA informative, Alu repeats belonging to the NSUMA universe with at least 1 nreads in one or more samples

NSUMA(-), NSUMA informative *Alu* repeats with <5 nreads in all the samples analyzed.

NSUMA(+), NSUMA informative *Alu* repeats with >=5 nreads in at least one of the samples analyzed.

DKO(exc), NSUMA informative *Alu* repeats with ≥5 normalized reads in DKO but <5 nreads in the rest of samples.

UNM All, NSUMA *Alu* repeats unmethylated in all tissues analyzed (≥5 nreads).

Blood, NSUMA *Alu* repeats hypomethylated in blood as compared with normal colon mucosa (adjusted p-value <0.05 and |log 2 FC| >1).

NColon (Normal colon), NSUMA *Alu* repeats hypermethylated in blood as compared with normal colon mucosa (adjusted p-value <0.05 and |log 2 FC| >1).

Tumor, NSUMA *Alu* repeats hypermethylated in tumor as compared with normal colon mucosa (adjusted p-value <0.05 and |log 2 FC| >1).

Rest, *Alu* repeats with no statistically significant differences in pairwise comparisons

Identification of hypomethylated *Alu* and LINE regions (HMARs and HMLRs)

To detect large regions with prevalent hypomethylation or hypermethylation we applied the circular binary segmentation (CBS) algorithm implemented in the DNAcopy R package (Venkatraman and Olshen 2007). This algorithm is based in a maximum t-statistic and splits chromosomes into subsegments using a reference distribution estimated by permutation. We applied the *sundo* option to remove change-points detected due to local trends and setting SD=1 for NSUMA and SD=0.5 for WGBS data. To detect significant regions we set different thresholds depending on the samples we were comparing based on the distribution of the mean of the segments (Supplemental Figure S17).

Analysis of chromosome imbalances based on NSUMA non-canonical amplicons

Reads mapping outside NSUMA virtual universe correspond to amplicons flanked by two MseI sites and its representation is DNA methylation independent. Therefore they may be used to analyze copy number variations by comparing normal and tumor profiles. The NSUMA non-canonical reads (Supplemental Table 2) were normalized using total-sum scaling and the genome was divided in 100 kb windows. Copy number variations were detected using the DNAcopy R package as described above with the *sundo* option and setting SD=0.5. To detect significant regions we set different thresholds depending on the samples we were comparing based on the distribution of the mean of the segments.

Transcription Factor Binding sites

A weight matrix finding algorithm (MEME) and motif alignment was used to identify transcription factor (TF) binding sites either within the *Alu* sequence or in the 500 bp flanking regions.

Bisulfite sequencing

DNA was extracted from cell lines and tissues by phenol/chloroform method. Bisulfite reaction was carried out using the EZ DNA Methylation Kit (Zymo Research, Irvine, CA, USA)) according to the manufacturer's instructions. Following bisulfite conversion, DNA was amplified by PCR using primers listed in Supplemental Table 12. A minimum of 2 independent nested-PCRs were performed and pooled to ensure a representative methylation profile. Pooled PCR products were purified (JETquick PCR Product Purification Spin Kit, Genomed GmbH, Löhne, Germany) and when necessary cloned into the pGEMT-Easy Vector (Promega, Fitchburg, WI, USA) using the Rapid Ligation Buffer System (Promega). Direct sequencing of the pooled PCR or the individual clones (10-12 clones per sample) were performed using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). The degree of methylation was calculated by comparing the peak height of the cytosine residues with the peak of the thymine residues as described (Melki et al. 1999).

RNA-seq and transcriptional analysis of *Alu* expression.

Poly(A)$^+$ messenger RNA was obtained using the PureLink RNA mini kit (Ambion) from two independent cultures of the HCT116 cell line and processed separately. RNA was sequenced using standard Illumina protocols for paired-end RNA-seq (2x50). Sequencing quality was checked with qualimap v2.1.3 and no obvious outliers were detected. Paired-end reads were mapped with TopHat v2.1.0 (Kim et al. 2013) with the --b2-sensitive flag against the ENSEMBL GRCh37 69 human reference genome plus the ERCC92 RNA Spike-In Mix (Thermo Fischer Scientific). Accepted hits were marked for duplicates using picard v2.0.1(McKenna et al. 2010). A total of 177,681,451 and 164,386,234 aligned pair-reads were obtained from each replicate. To assign each uniquely mapped read to transcripts and repeats, we retrieved the annotations from ENSEMBL GRCh37 69 and the rmsk table from the hg19 assembly hosted at UCSC, respectively, as GTF/GFF files. Read counting was run on each GTF independently; we ignored sequencing duplicates, multimappers and ambiguous reads using featureCounts from subread v1.5.0-p1 (Liao et al. 2014). Results were summarized using BedTools (Quinlan 2014). A total of 4,663,115 and 3,557,588 reads mapped totally or partially on *Alu* sequences in replicates 1 and 2 respectively. Read counts were normalized using total-sum scaling upon overall alignment rates.

Annotation of expressed *Alu* elements

Based on the availability of RRBS DNA methylation data (Song et al. 2013; Varley et al. 2013), a total of 9,339 *Alu* elements with minimum of 3 informative CpGs (at least 3 reads per site) were selected for in depth analysis. Confirmatory analyses were performed using the 1.1 million *Alu* repeats.

Chromatin states of H1-hESC were used to annotate the *Alu* repeats. Genome segments were obtained by integrating ChIP-seq data for 8 chromatin marks, input data and the CTCF transcription factor, and two DNase-seq assays and a FAIRE-seq assay, using a multivariate Hidden Markov Model (HMM) that explicitly models the combinatorial patterns of observed modifications (Ernst and Kellis 2010; Ernst and Kellis 2012). . Chromatin segmentation was retrieved from the MySQL database at UCSC (hg19.wgEncodeAwgSegmentationChromhmmH1hesc table), which we further summarized into the recommended candidate annotations (Tss or TssF, Active Promoter; PromF, Promoter Flanking; PromP, Inactive Promoter; Enh or EnhF, Candidate Strong enhancer; EnhWF, EnhW, DNaseU, DNaseD or FaireW, Candidate Weak enhancer/DNase; CtrcfO or Ctcf, Distal CTCF/Candidate Insulator; Gen5', Elon, ElonW, Gen3', Pol2 or H4K20, Transcription associated;

Low, Low activity proximal to active states; ReprD, Repr or ReprW, Polycomb repressed; and Quies or Art, Heterochromatin/Repetitive/Copy Number Variation). Each *Alu* annotated with a single state. When overlapped to multiple segments, the state with highest priority was assigned from the following hierarchy: CTCF Candidate Insulator, Active Promoter, Inactive Promoter, Candidate Strong enhancer, Candidate Weak enhancer/DNase, Promoter Flanking, Transcription associated, Low activity proximal to active states, Polycomb repressed, Heterochromatin Repetitive/Copy Number Variation. Analyses were performed with the subset of *Alu* repeats with RRBS information (n=9,339) and all the *Alu* repeats (n=1,107,717).

Supplemental references

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**: 3423-3424.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**: 817-825.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215-216.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923-930.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Melki JR, Vincent PC, Clark SJ. 1999. Concurrent DNA hypermethylation of multiple genes in acute myeloid leukemia. *Cancer Res* **59**: 3730-3740.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**: 11 12 11-34.

Rhee I, Bachman KE, Park BH, Jair KW, Yen RW, Schuebel KE, Cui H, Feinberg AP, Lengauer C, Kinzler KW et al. 2002. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* **416**: 552-556.

Rodriguez J, Vives L, Jorda M, Morales C, Munoz M, Vendrell E, Peinado MA. 2008. Genome-wide tracking of unmethylated DNA *Alu* repeats in normal and cancer cells. *Nucleic Acids Res* **36**: 770-784.

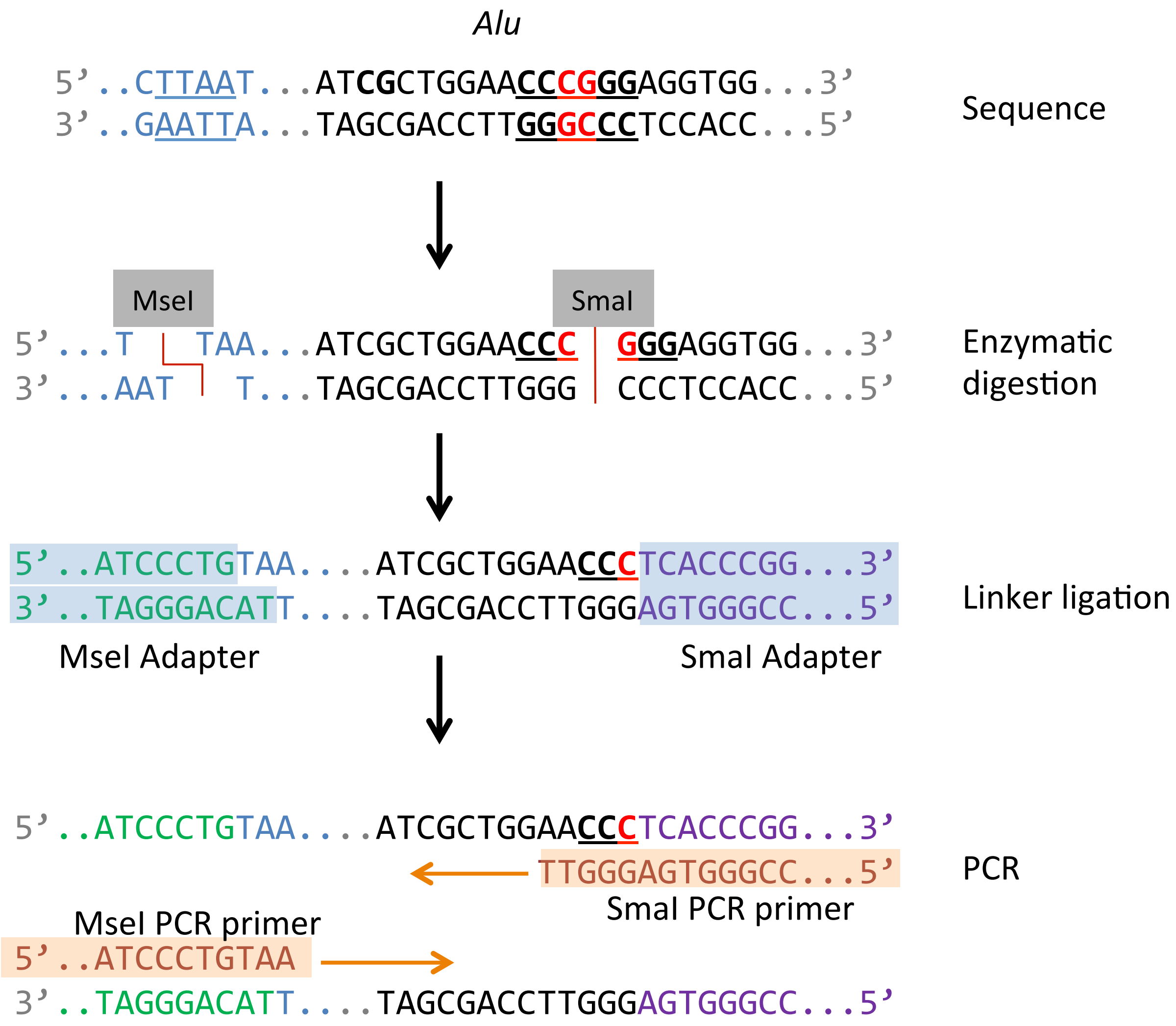
Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. 2013. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* **8**: e81148.

Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, Cross MK, Williams BA, Stamatoyannopoulos JA, Crawford GE et al. 2013. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* **23**: 555-567.

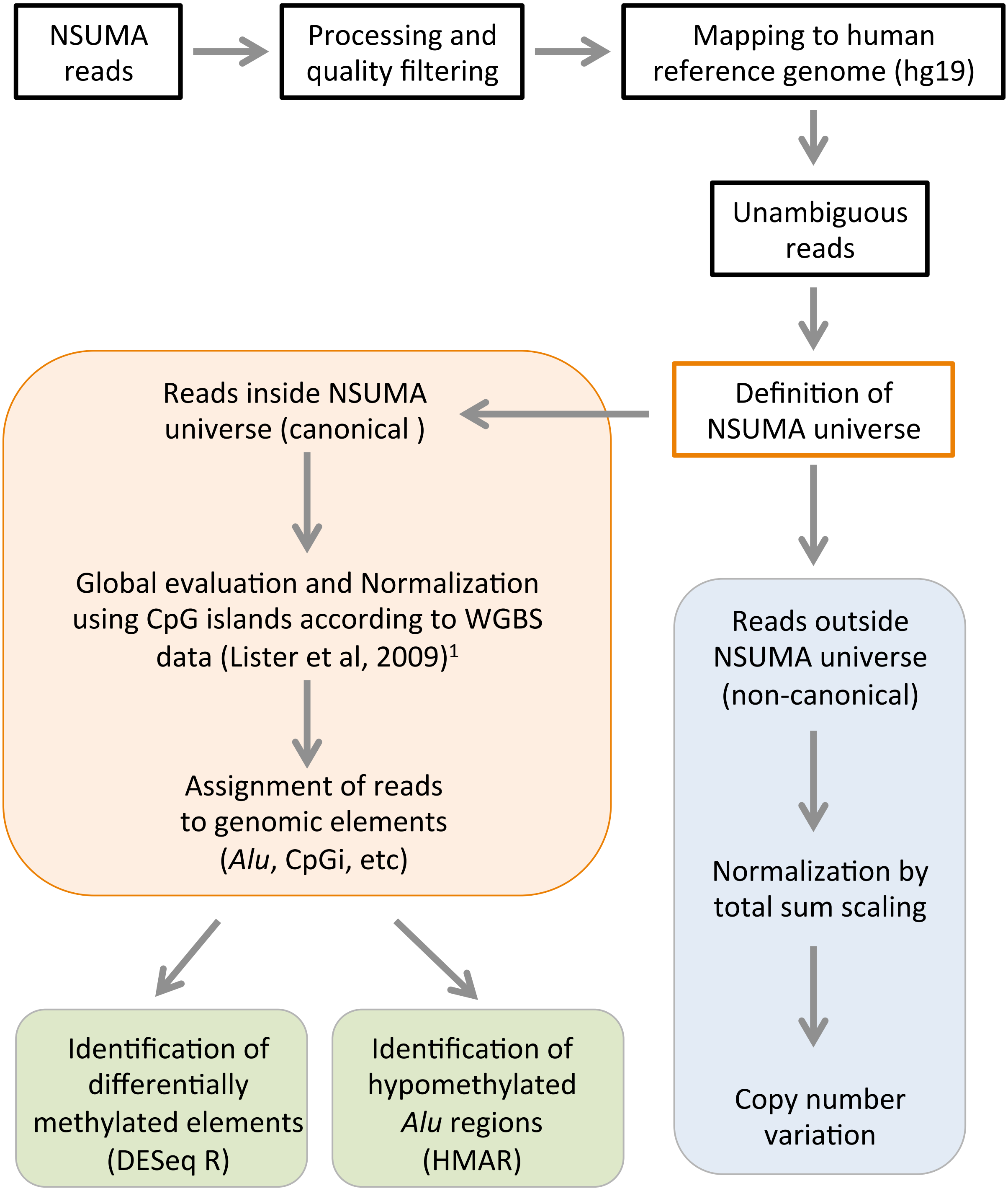
Venkatraman ES, Olshen AB. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**: 657-663.

List of abbreviations

- *Alu DMR*, *Alu* Differential Methylation Ratio, log2 (NSUMA *nreads* in sample X/ NSUMA *nreads* in sample Y)
- *Alu DM*, *Alu* Differential Methylation, WGBS methylation level (beta value) in sample X - WGBS methylation level (beta value) in sample Y
- *CBS*, Circular Binary Segmentation
- *CGH*, Comparative Genomic Hybridization
- *CNV*, Copy Number Variation
- *HB*, Hypomethylated Blocks
- *HMAR*, HypoMethylated *Alu* Region
- *HMLR*, HypoMethylated LINE Region
- *LINE DM*, LINE Differential Methylation, WGBS methylation level (beta value) in sample X - WGBS methylation level (beta value) in sample Y
- *LMA*, Low Methylated *Alu* (beta value <0.2)
- *nreads*, normalized count of NSUMA reads
- *NSUMA*, Next generation Sequencing of Unmethylated *Alu*
- *PCA*, Principal Component Analysis
- *ROC*, Receiver Operating Characteristic
- *RRBS*, Reduced Representation Bisulfite sequencing
- *SMAR*, Stable Methylated *Alu* Region
- *WGBS*, Whole Genome Bisulfite Sequencing

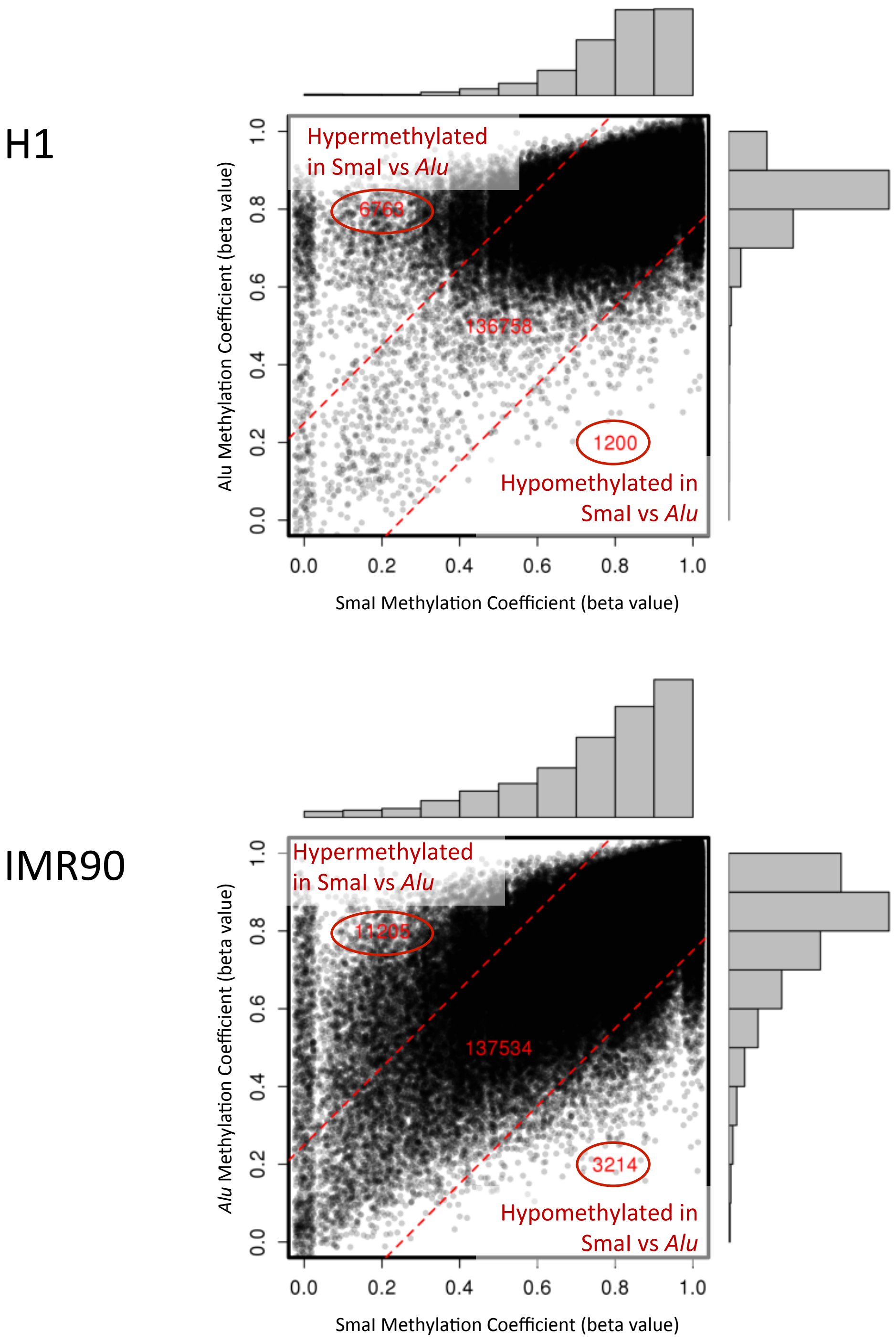

Alu
5'..CTTAAT...ATCGCTGGAACCCGGGAGGTGG...3'
3'..GAATTA...TAGCGACCTTGGGCCCTCCACC...5'
Sequence
MseI
SmaI
5'...T TAA...ATCGCTGGAACCC GGGAGGTGG...3'
3'...AAT T...TAGCGACCTTGGG CCCTCCACC...5'
Enzymatic digestion
5'..ATCCCTGTAA....ATCGCTGGAACCCTCACCCGG...3'
3'..TAGGGACATT....TAGCGACCTTGGGAGTGGGCC...5'
Linker ligation
MseI Adapter
SmaI Adapter
5'..ATCCCTGTAA....ATCGCTGGAACCCTCACCCGG...3'
TTGGGAGTGGGCC...5'
SmaI PCR primer
MseI PCR primer
5'..ATCCCTGTAA
3'..TAGGGACATT....TAGCGACCTTGGGAGTGGGCC...5'
PCR

Supplemental Fig. S1. Scheme of the NSUMA procedure. The partial sequence of an unmethylated *Alu* element containing a SmaI site and a MseI site outside the element is depicted. After digestion with both enzymes, the DNA fragments are ligated to specific adapters for each cut. Next, DNA fragments flanked by adapters are amplified by PCR using primers complementary to the adapters. To enrich for *Alu* sequences during the PCR, a SmaI primer (complementary to the SmaI adapter) extended at the 3' end with five additional nucleotides (5'-GGGTT-3') complementary to the consensus sequence of *Alu* repeats (5'-AACC-3'C) was used.

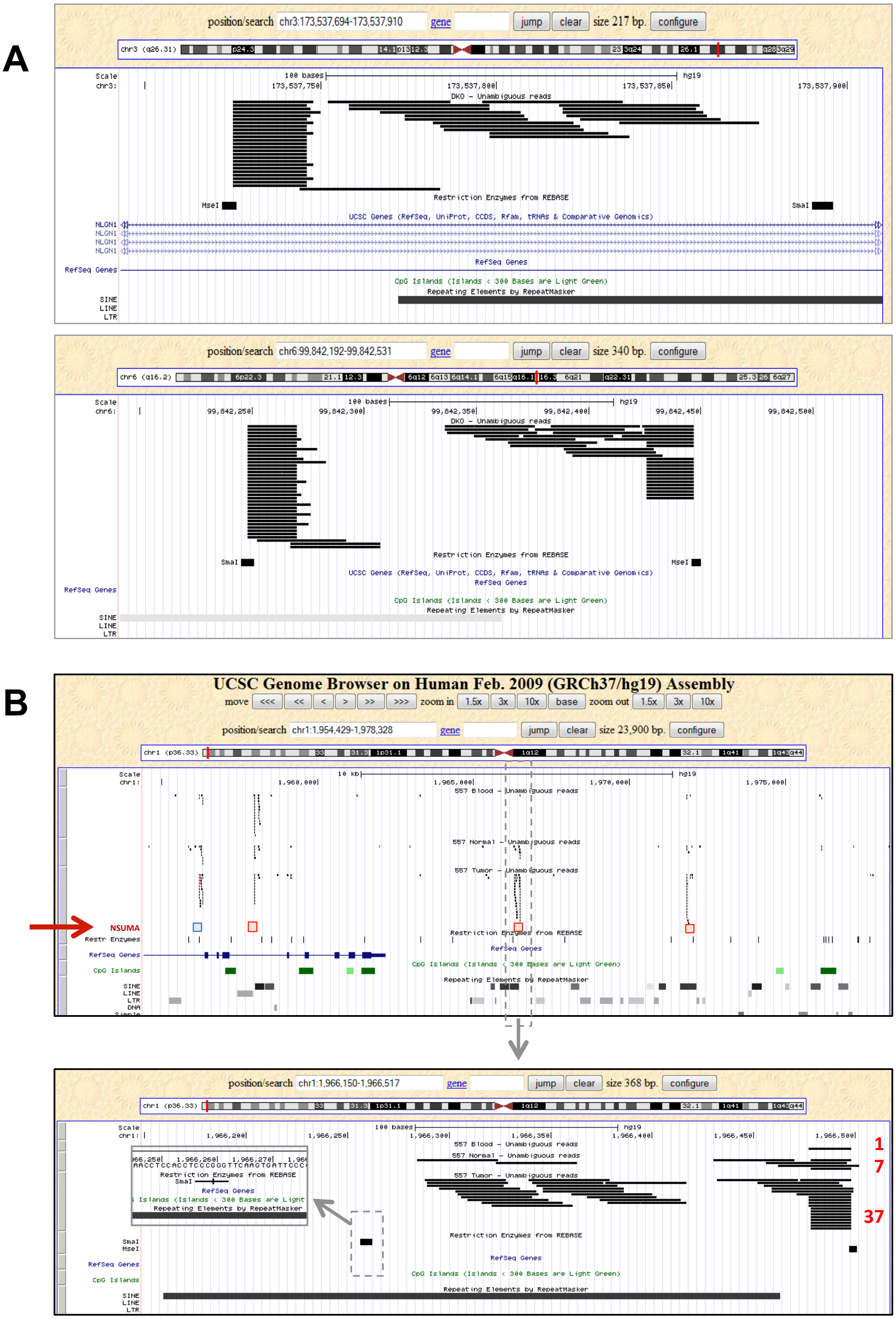

NSUMA reads
Processing and quality filtering
Mapping to human reference genome (hg19)
Unambiguous reads
Definition of NSUMA universe
Reads inside NSUMA universe (canonical)
Global evaluation and Normalization using CpG islands according to WGBS data (Lister et al, 2009)[1]
Assignment of reads to genomic elements (*Alu*, CpGi, etc)
Identification of differentially methylated elements (DESeq R)
Identification of hypomethylated *Alu* regions (HMAR)
Reads outside NSUMA universe (non-canonical)
Normalization by total sum scaling
Copy number variation

Supplemental Fig. S2. Scheme summarizing the NSUMA data processing. The NSUMA universe is composed by the virtual amplicons generated by in-silico NSUMA using the hg19 assembly of the human genome. The in-silico NSUMA included sequences ranging 20-1000 bp length flanked by a AACCCGGG and a TTAA (SmaI-MseI fragments) or by AACCCGGG and TTCCCGGG (SmaI-SmaI fragments).

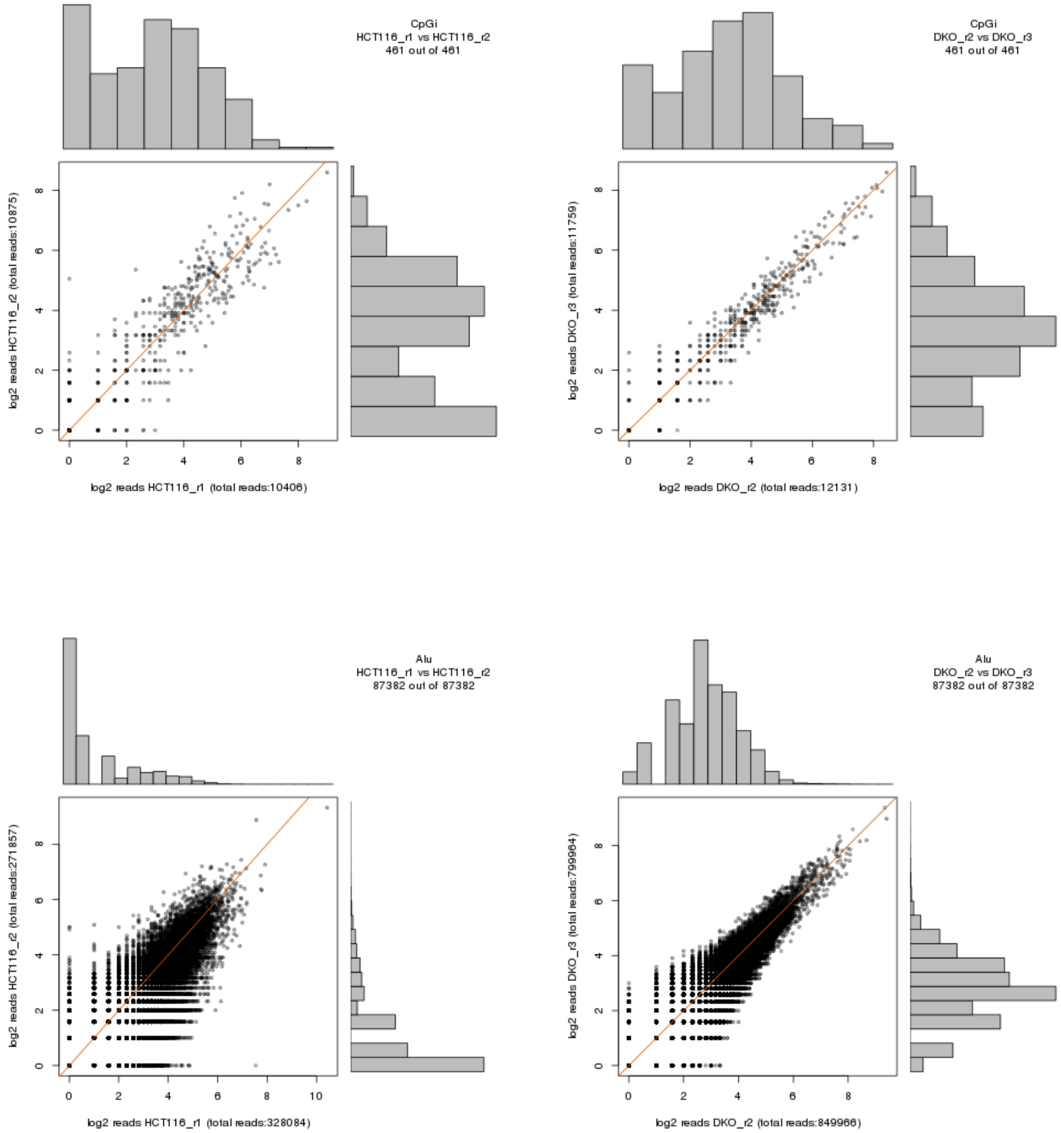
[1] Lister et al. Nature 2009;462:315-22



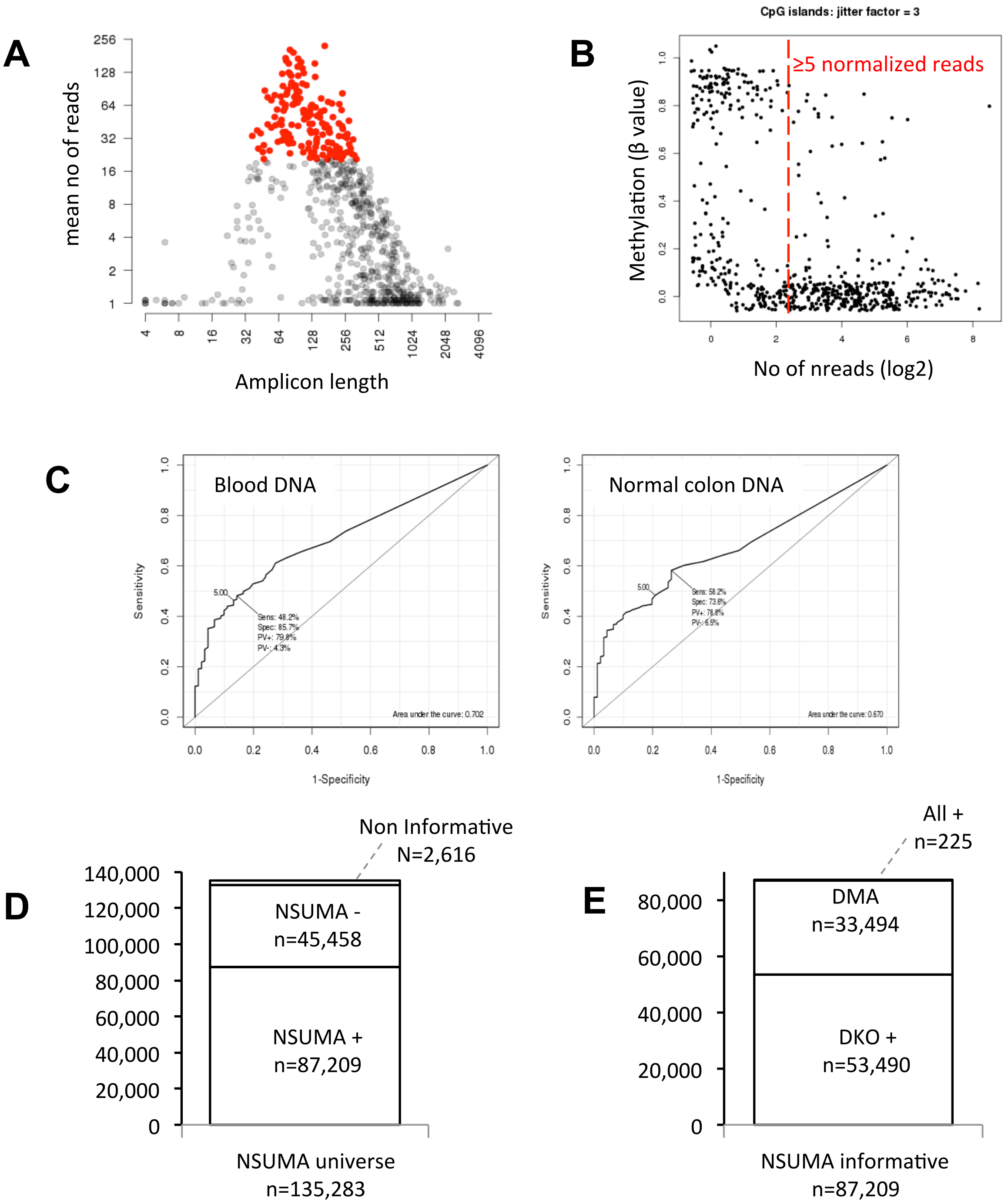
Supplemental Fig. S3. Correlation plots for the methylation coefficient between the SmaI and the corresponding *Alu* element for H1 and IMR90 samples as analyzed by whole genome bisulfite sequencing by Lister et al (Nature 2009;462:315-22). Dash lines delimit areas with differences >0.25 between the SmaI site and the corresponding *Alu* element. The numbers of points represented in each area of the graph and the distribution histograms of both axes are shown. Only *Alu* repeats with a SmaI site in the sequence and a minimum of 3 informative CpG sites in at least 3 reads each have been considered.


A
position/search chr3:173,537,694-173,537,910 gene jump clear size 217 bp. configure
chr3 (q26.31)
Scale chr3:
100 bases hg19
173,537,750| 173,537,800| 173,537,850| 173,537,900|
DKO - Unambiguous reads
Restriction Enzymes from REBASE
MseI SmaI
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
NLGN1 NLGN1 NLGN1 NLGN1
RefSeq Genes
CpG Islands (Islands < 300 Bases are Light Green)
Repeating Elements by RepeatMasker
SINE LINE LTR
position/search chr6:99,842,192-99,842,531 gene jump clear size 340 bp. configure
chr6 (q16.2)
Scale chr6:
100 bases hg19
99,842,250| 99,842,300| 99,842,350| 99,842,400| 99,842,450| 99,842,500|
DKO - Unambiguous reads
Restriction Enzymes from REBASE
SmaI MseI
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
RefSeq Genes
CpG Islands (Islands < 300 Bases are Light Green)
Repeating Elements by RepeatMasker
SINE LINE LTR
B
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr1:1,954,429-1,978,328 gene jump clear size 23,900 bp. configure
chr1 (p36.33)
Scale chr1:
10 kb hg19
1,960,000| 1,965,000| 1,970,000| 1,975,000|
557 Blood - Unambiguous reads
557 Normal - Unambiguous reads
557 Tumor - Unambiguous reads
NSUMA
Restr Enzymes
Restriction Enzymes from REBASE
RefSeq Genes
CpG Islands
CpG Islands (Islands < 300 Bases are Light Green)
Repeating Elements by RepeatMasker
SINE LINE LTR DNA Simple
position/search chr1:1,966,150-1,966,517 gene jump clear size 368 bp. configure
chr1 (p36.33)
Scale chr1:
100 bases hg19
1,966,200| 1,966,250| 1,966,300| 1,966,350| 1,966,400| 1,966,450| 1,966,500|
557 Blood - Unambiguous reads
557 Normal - Unambiguous reads
557 Tumor - Unambiguous reads
966,250| 1,966,260| 1,966,270| 1,966
AACCTCCACCTCCCGGGTTCAAGTGATTCCC
Restriction Enzymes from REBASE
SmaI
RefSeq Genes
Islands (Islands < 300 Bases are Light
Repeating Elements by RepeatMasker
SmaI MseI
RefSeq Genes
CpG Islands
Restriction Enzymes from REBASE
RefSeq Genes
CpG Islands (Islands < 300 Bases are Light Green)
Repeating Elements by RepeatMasker
SINE LINE LTR
1
7
37

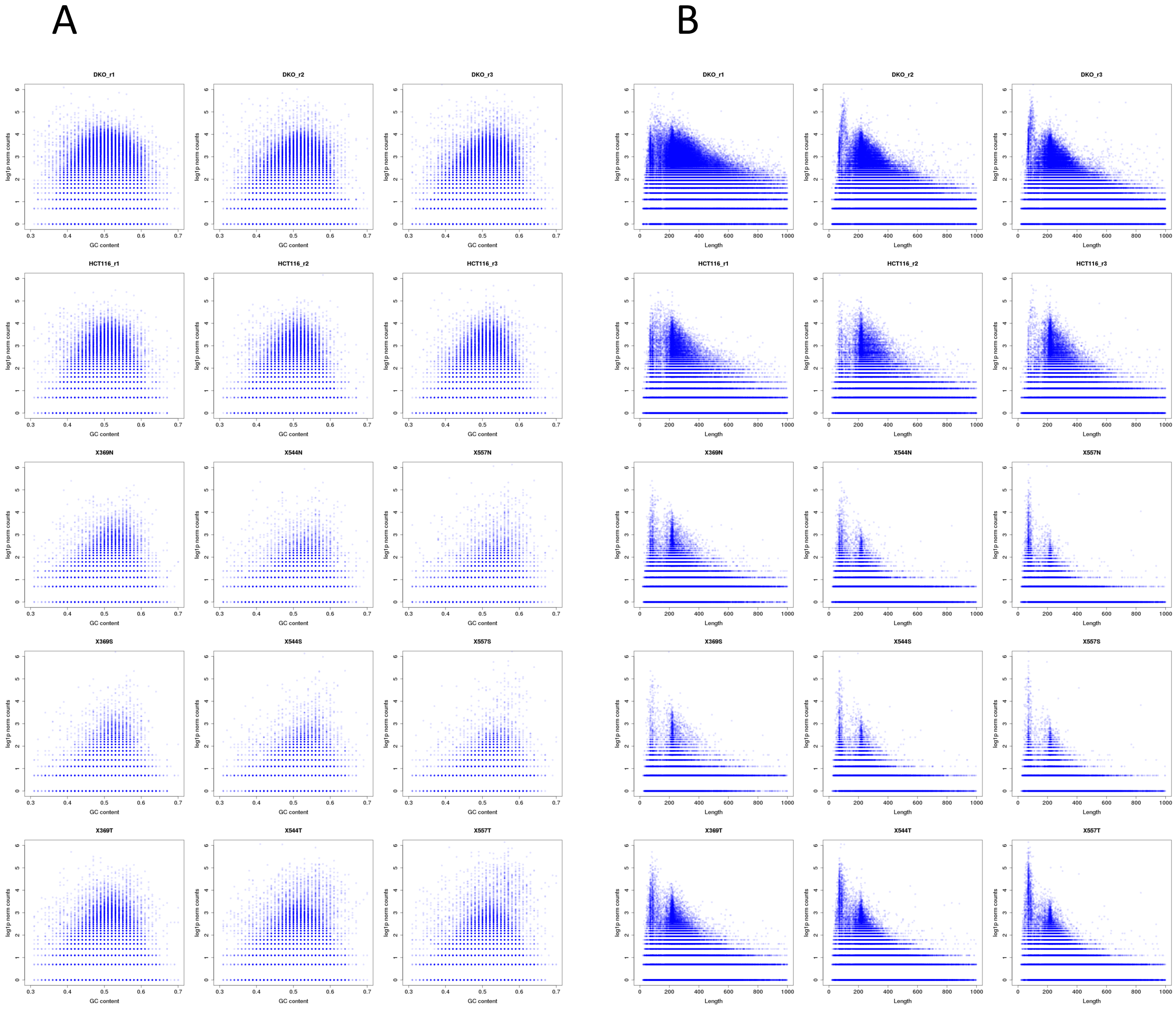
Supplemental Fig. S4. (**A**) Two examples of NSUMA display. The upper panel shows an amplicon representing an *Alu*Y located in an intron of the neuriligin (*NLGN1*) gene encoding for a neuronal cell surface protein. The predicted amplicon (flanked by the MseI and the SmaI restriction sites) is well covered by reads except for a small region near the SmaI that is ambiguous. The lower panel shows an amplicon representing an *Alu* located in the promoter region of the coenzyme Q3 homolog, methyltransferase (*COQ3*) gene which is mostly covered by reads along the unique sequence. (**B**) Illustrative example of reads distribution along a region containing four NSUMA amplicons represented. The tracks (from top to bottom) indicate: NSUMA reads counts in sample 557 blood, normal colon and tumor respectively; NSUMA canonical amplicons are shown as a blue box (amplicon without *Alu*) or a red (amplicon with *Alu*); restriction sites (SmaI and MseI); Refseq genes; CpG islands and repeats according to Repeat Masker. The lower panel B shows a zoom-in of the upper panel detailing one of the amplicons and the SmaI site in the *Alu* sequence (inset).



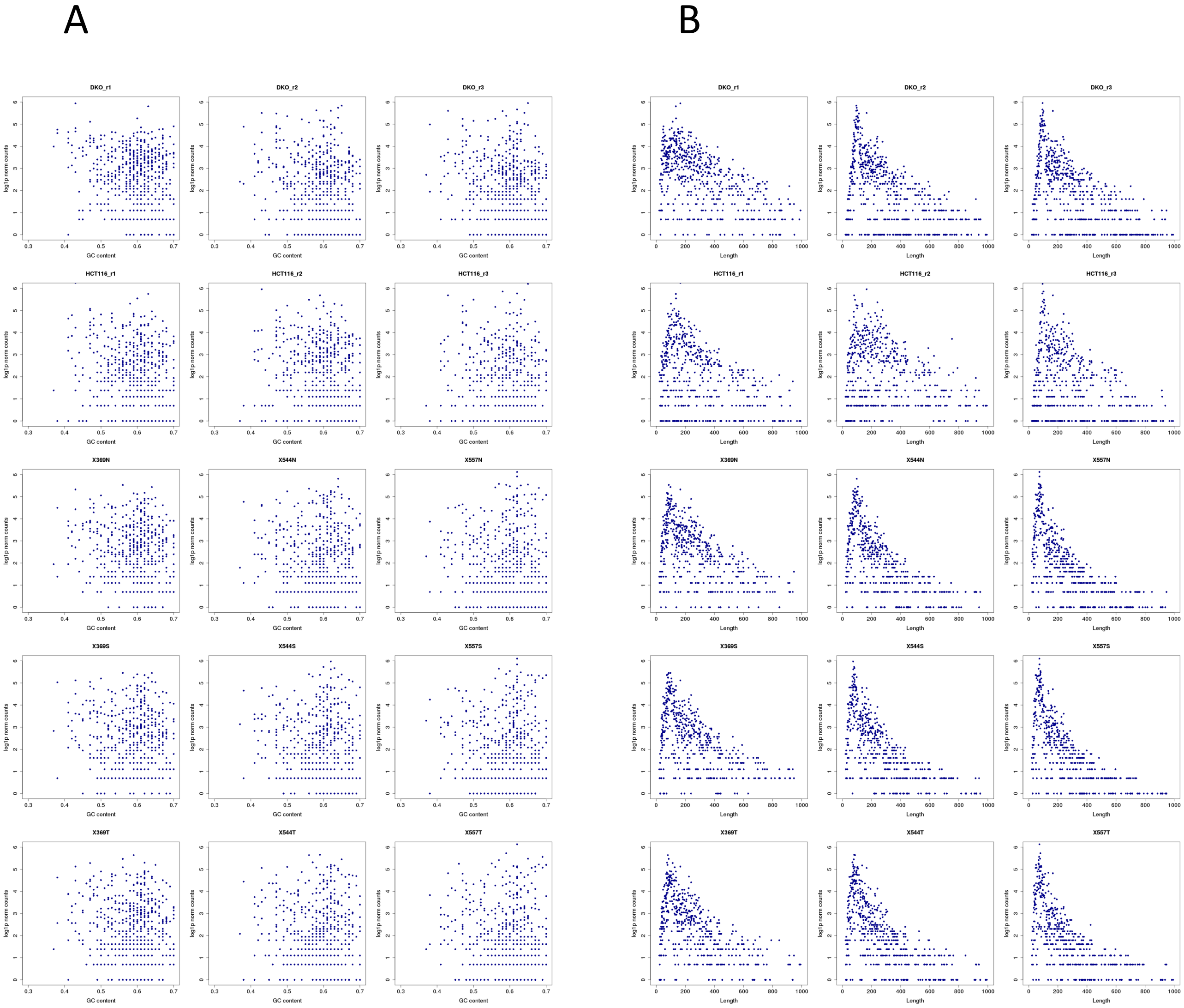
Supplemental Fig. S5. Scatterplot with marginal histograms of the comparison of two biological replicates of HCT116 and two technical replicates of DKO analyzed by NSUMA. Each dot corresponds to an amplicon and the axes show the log 2 of the number of reads in each replicate. CpG islands and *Alu* repeats associated amplicons are shown .



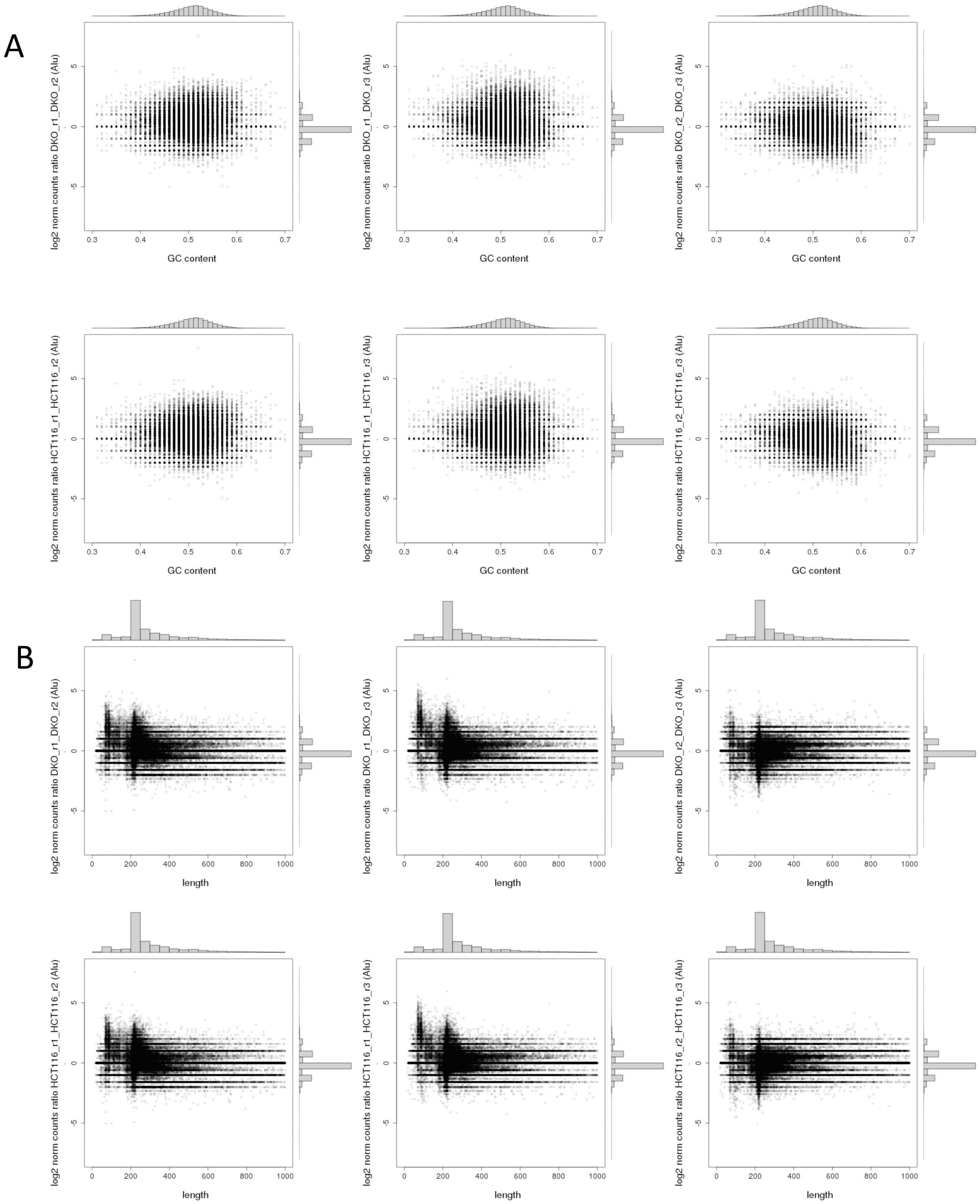
Supplemental Fig. S6. (**A**) Distribution of the number of reads ascribed to CpG islands according to the size of the NSUMA amplicons. Normalization among samples was performed using 150 amplicons generated in CpG islands with the highest mean and at least 5 normalized reads in all the samples (red dots). (**B**) Distribution of reads in 661 amplicons containing the SmaI site within a CpG island (X axis, log2 of the mean of all samples) against the methylation level of the respective CpG island in H1 embryonic stem cells as reported by Lister et al (Nature 2009;462:315-22). (**C**) ROC curve analysis of normalized reads determined that 5 or more reads were predictive of an unmethylated CpG island with 95% specificity and 50% sensitivity. (**D**) Distribution of *Alu* repeats according to the NSUMA representation in different samples. The virtual NSUMA universe comprises 135,283 *Alu* repeats, of which 87,209 were represented by at least 5 normalized reads in one or more samples (including DKO cells) and are therefore considered as partially or fully unmethylated, constituting the "NSUMA +" set. An additional 45,458 *Alu* repeats constitute the "NSUMA –" set and showed at least one read in at least one sample, but not reaching 5 reads in any of the analyzed samples, being considered as informative but methylated in all the samples. The remaining 2,616 *Alu* repeats were not represented in any sample and were considered non informative. (**E**) Within the NSUMA + set, 33,494 were detected in at least one sample and constitute differentially methylated *Alu* repeats (DMA) set. The rest of NSUMA + *Alu* repeats (n=53,490) were represented in DKO cells (DKO +) alone and are considered methylated in all the samples and unmethylated in DKO. Finally, 225 *Alu* repeats showed 5 or more normalized reads in all the samples and considered *"always unmethylated"*.



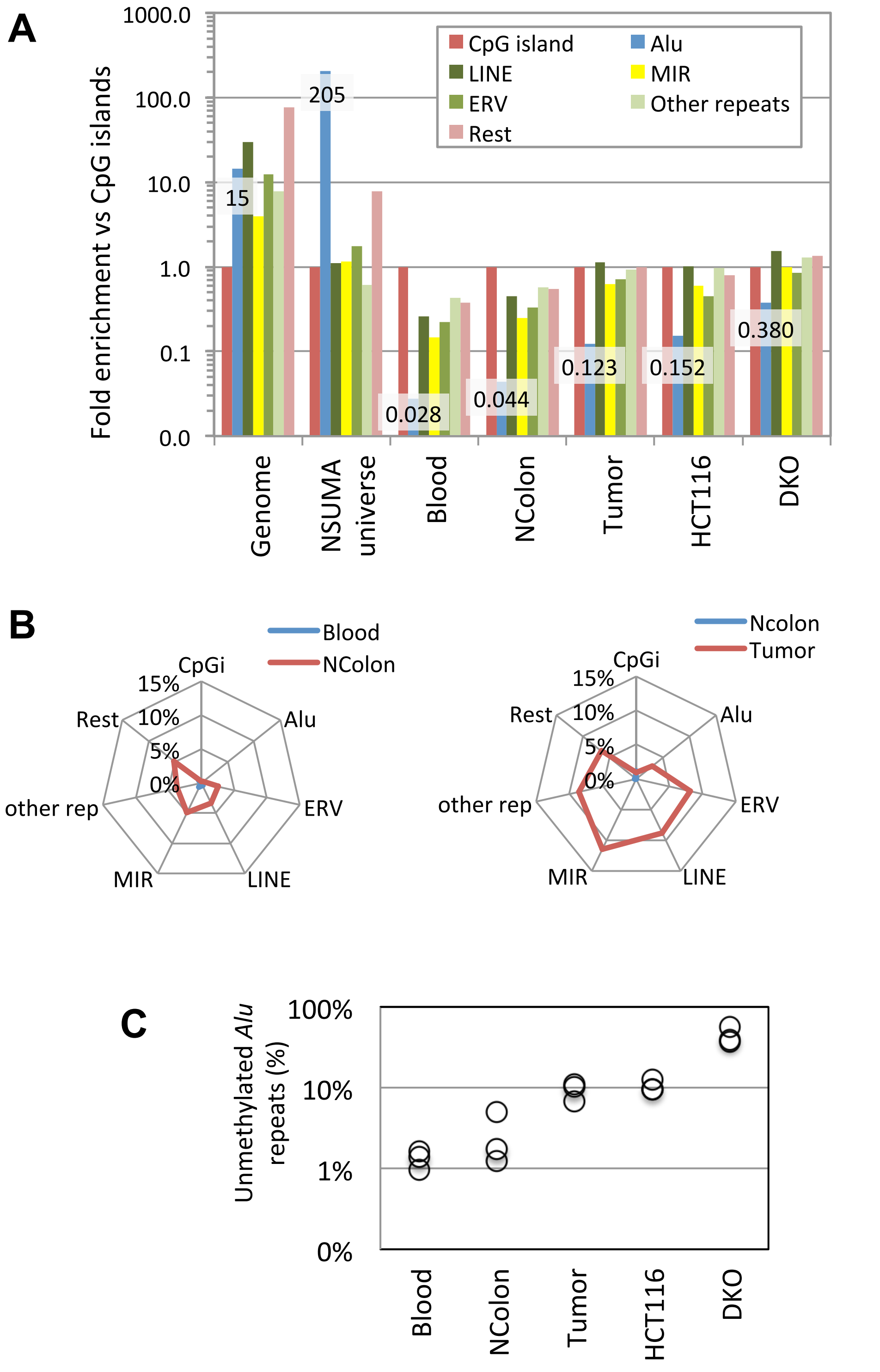
Supplemental Fig. S7. Distribution of NSUMA canonical amplicons annotated to *Alu* repeats according to the number of reads (Y axis) and the amplicon GC content (A) and amplicon size (B). All amplicons with at least 1 read in the given sample have been plotted.



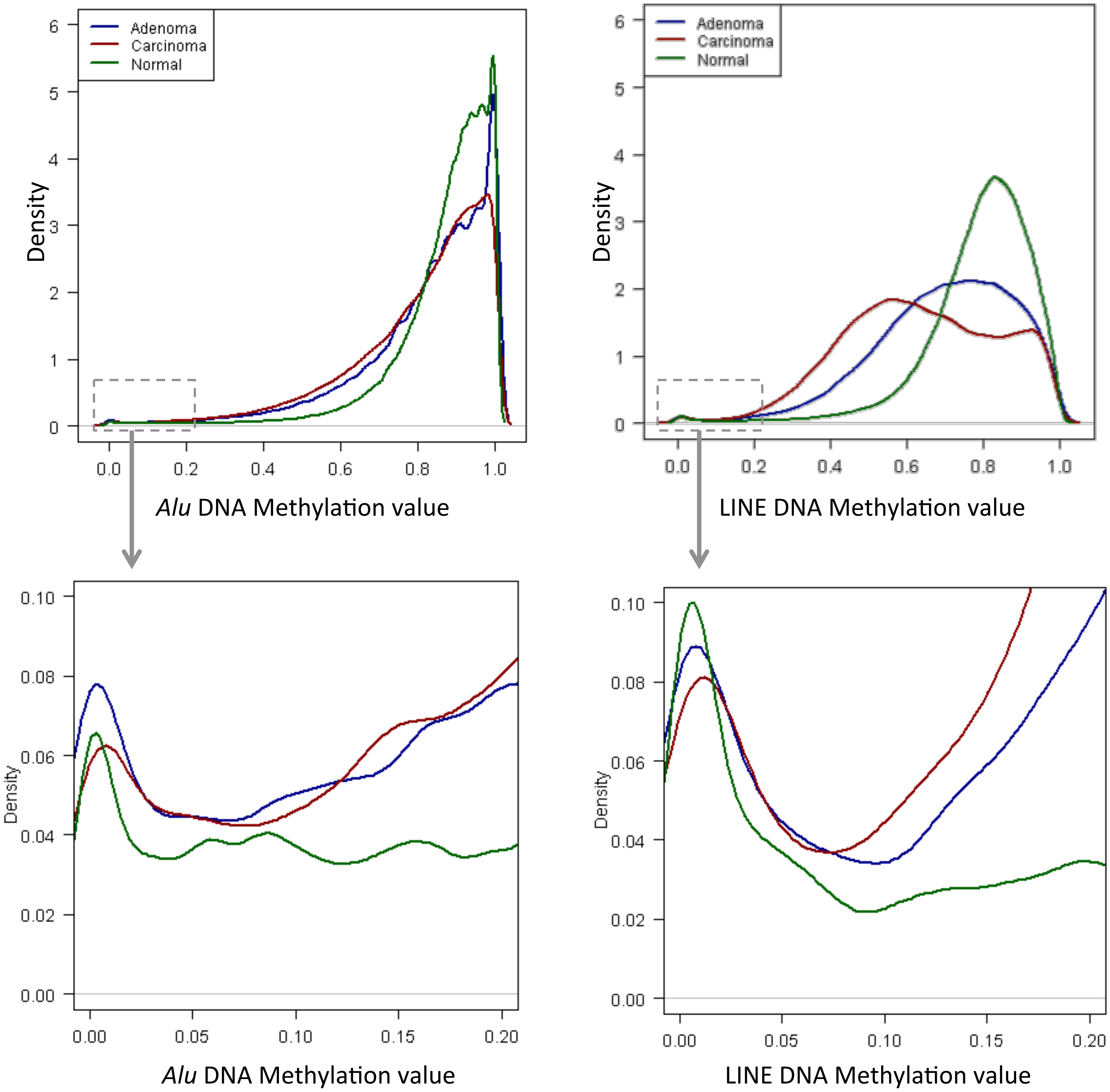
Supplemental Fig. S8. Distribution of NSUMA canonical amplicons annotated to CpG islands according to the number of reads (Y axis) and the amplicon GC content (A) and amplicon size (B). All amplicons with at least 1 read in the given sample have been plotted.



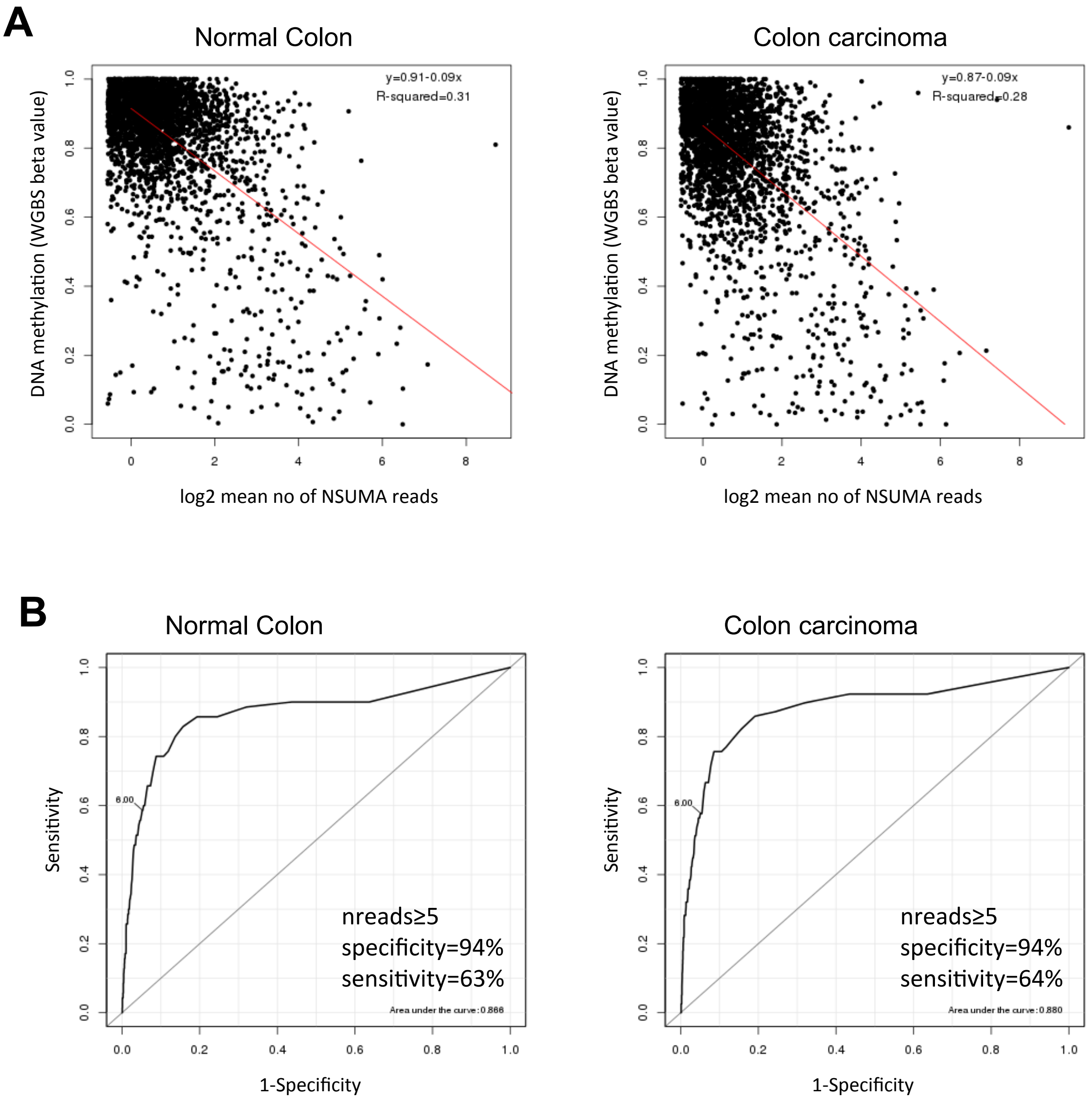
Supplemental Fig. S9. Assessment of potential biases due to GC content (A) and amplicon size (B) among biological (HCT116) and technical (DKO) replicates . All amplicons annotated to *Alu* repeats with at least 1 read in the given sample have been plotted.



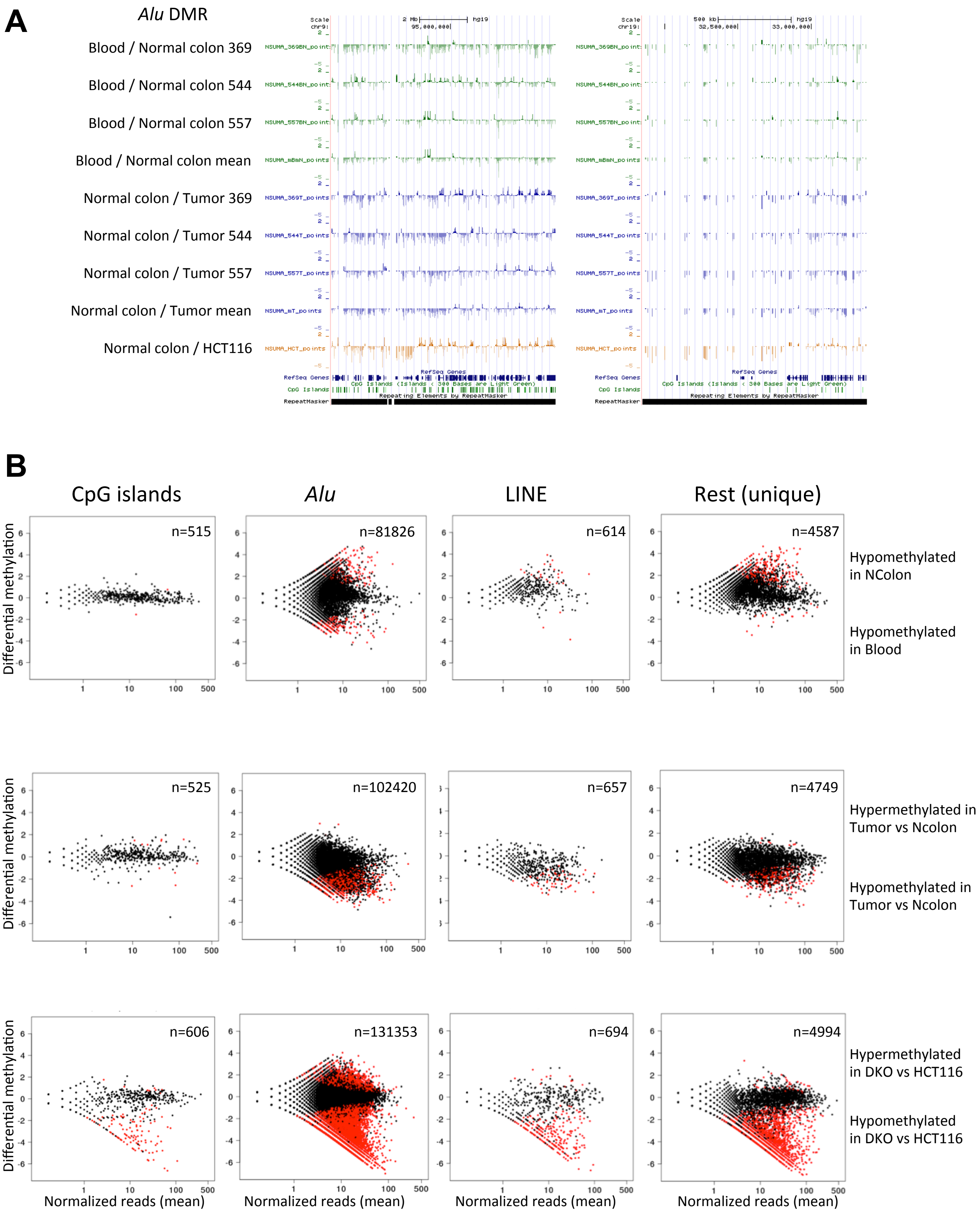
Supplemental Fig S10. (**A**) Relative abundance of different genetic elements in the human genome represented in reference to the number of CpG islands (fold enrichment). The relative enrichment has been calculated using the whole genome, the NSUMA virtual universe and in the analysis of different samples. Numbers on top of the blue bar indicate the fold enrichment of *Alu* repeats. The virtual NSUMA achieves a 205 fold enrichment of *Alu* potential representation as compared with the CpG islands. Nevertheless, this representativeness is dramatically reduced in actual experiments (and more especially in blood samples) as most *Alu* repeats are methylated. (**B**) Percentage of hypomethylated elements in the comparison between pair of tissues. (**C**) Proportion of unmethylated *Alu* repeats detected by NSUMA in three samples of each tissue (except the cell lines which correspond to replicates). Unmethylated *Alu* repeats are considered those with 5 or more *nreads* (see Supplementary fig. S6.)



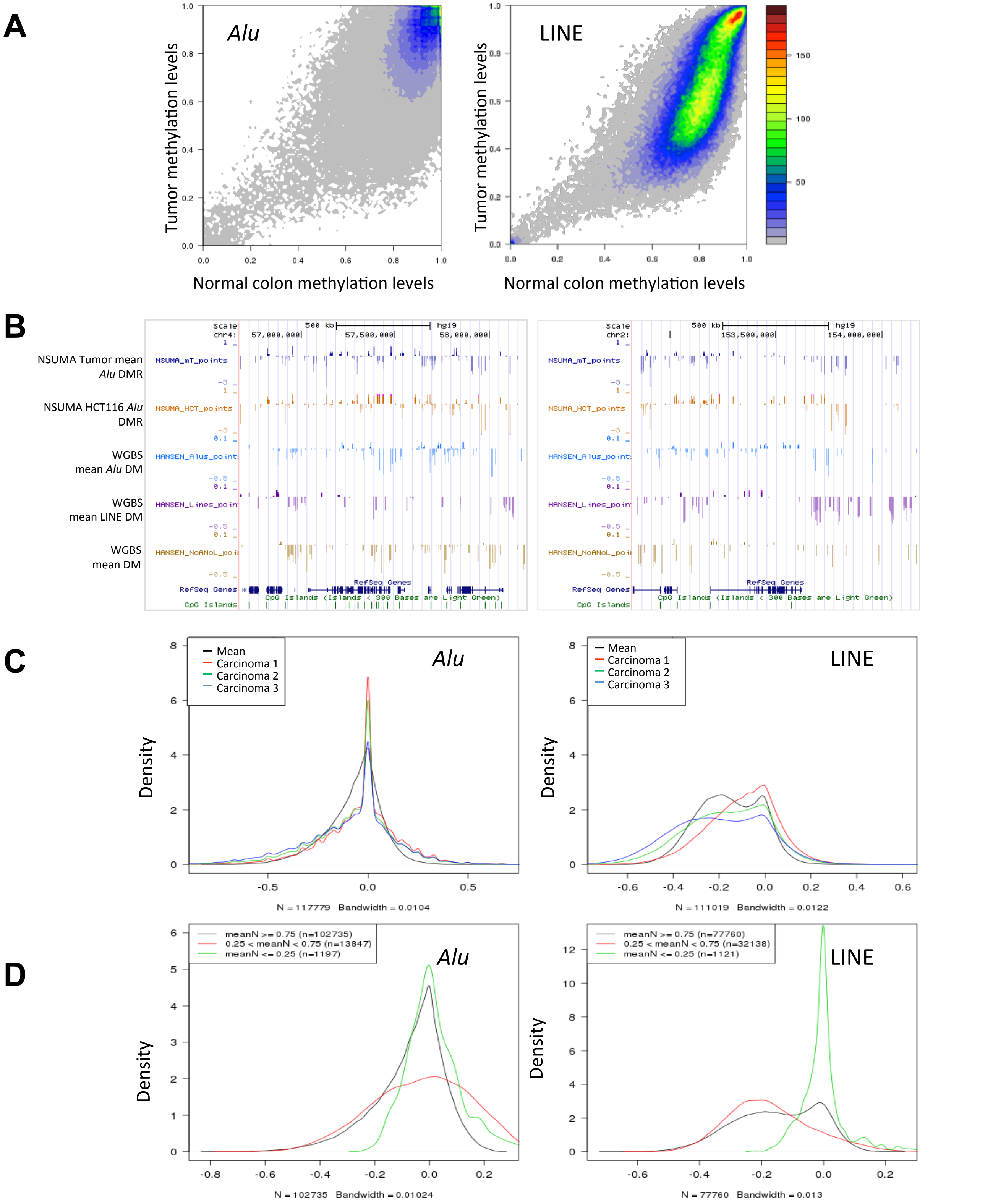
Supplemental Fig S11. Density plot of the mean beta values obtained by WGBS in *Alu* and LINE elements in Adenomas, Carcinomas and Normal colon samples. Only *Alu* repeats/LINEs with a minimum of 3/5 informative CpGs respectively in all samples were considered. The inset (elements with a DNA methylation value ranging 0 to 0.2) is shown enlarged below. Only *Alu* repeats with 3 or more informative CpGs (n=117,779 out of 574,343 with at least 1 informative CpG) were considered for further analyses.



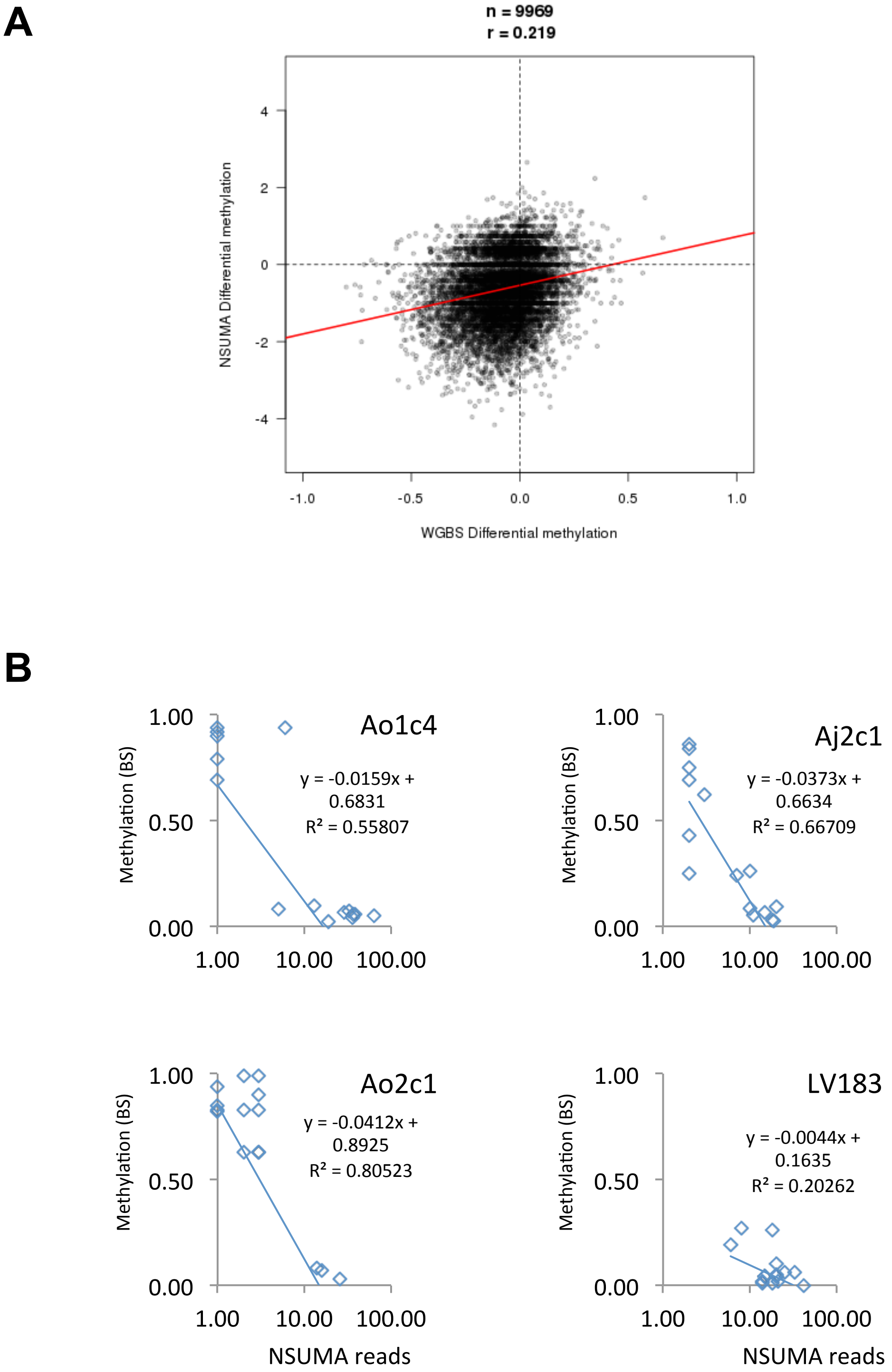
Supplemental Fig S12. **(A)** Comparison of DNA methylation levels determined by WGBS (beta values) with NSUMA (number of normalized reads) of *Alu* repeats repeats in normal and colon carcinomas. Only *Alu* repeats with at least 3 informative CpGs and a minimum of 5 NSUMA normalized reads in one or more samples were considered. **(B)** Receiver operating characteristic curves for the prediction of the methylation status of *Alu* repeats based on the number of normalized reads. The gold standard was considered the level of methylation detected by WGBS in Hansen *et al.* (*Nat Genet 2011, 43: 768-775)*. Only *Alu* repeats with at least 3 informative CpGs and a minimum of 5 normalized reads in one or more samples (NSUMA+ set) were considered.



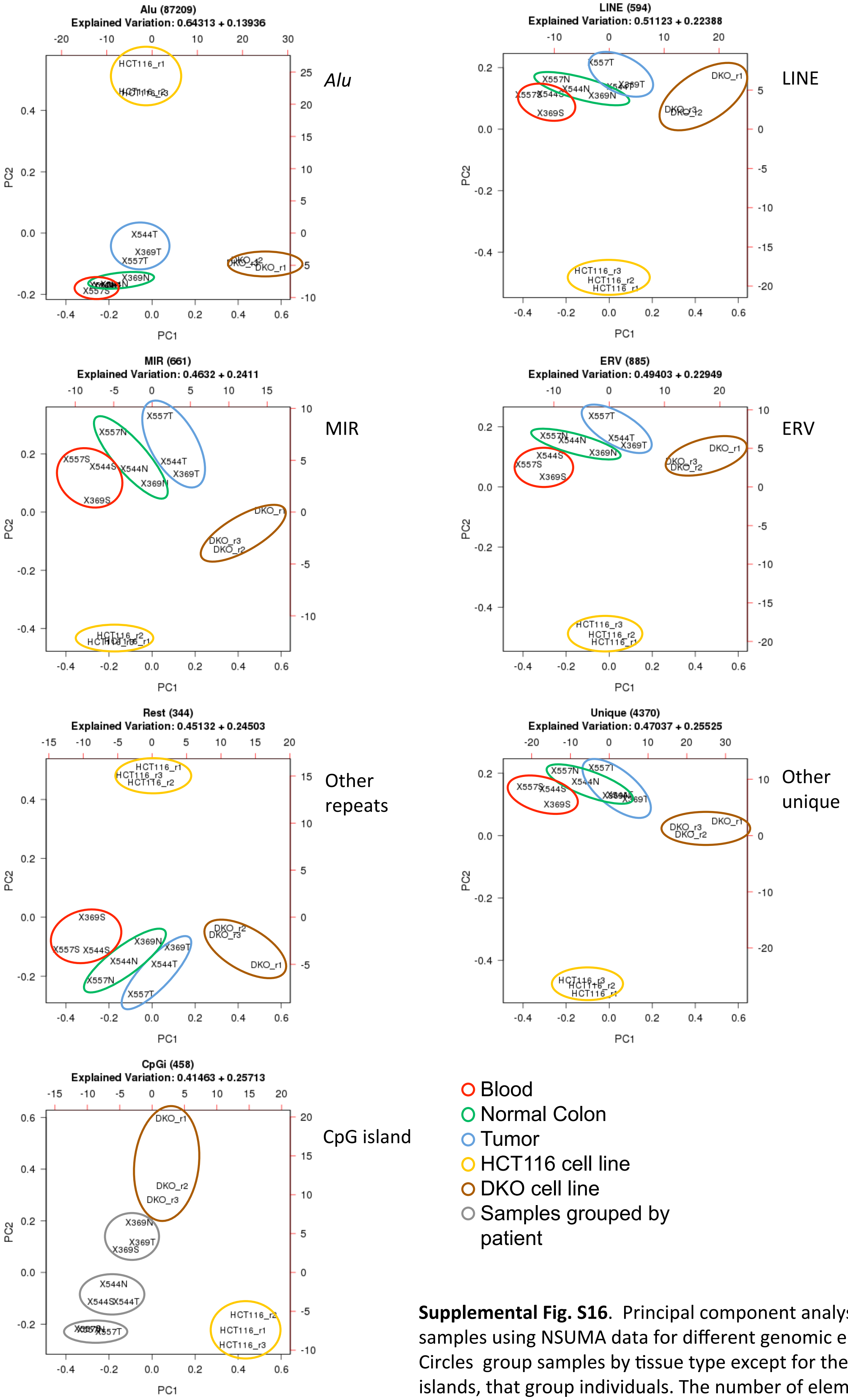
Supplemental Fig. S13. (A) UCSC Genome Browser representation of the *Alu* Differential Methylation Ratio (log 2 of the ratio of the number of NSUMA normalized reads between two samples) in sample comparisons as indicated. **(B)** MA plot of the differential DNA methylation of genomic elements analyzed by NSUMA when different samples are compared. Red spots indicate elements with statistically significant differences (adjusted p value <0.05). Summarizing data are illustrated in Fig. 1D. Detailed data are provided in Supplemental Data S2.



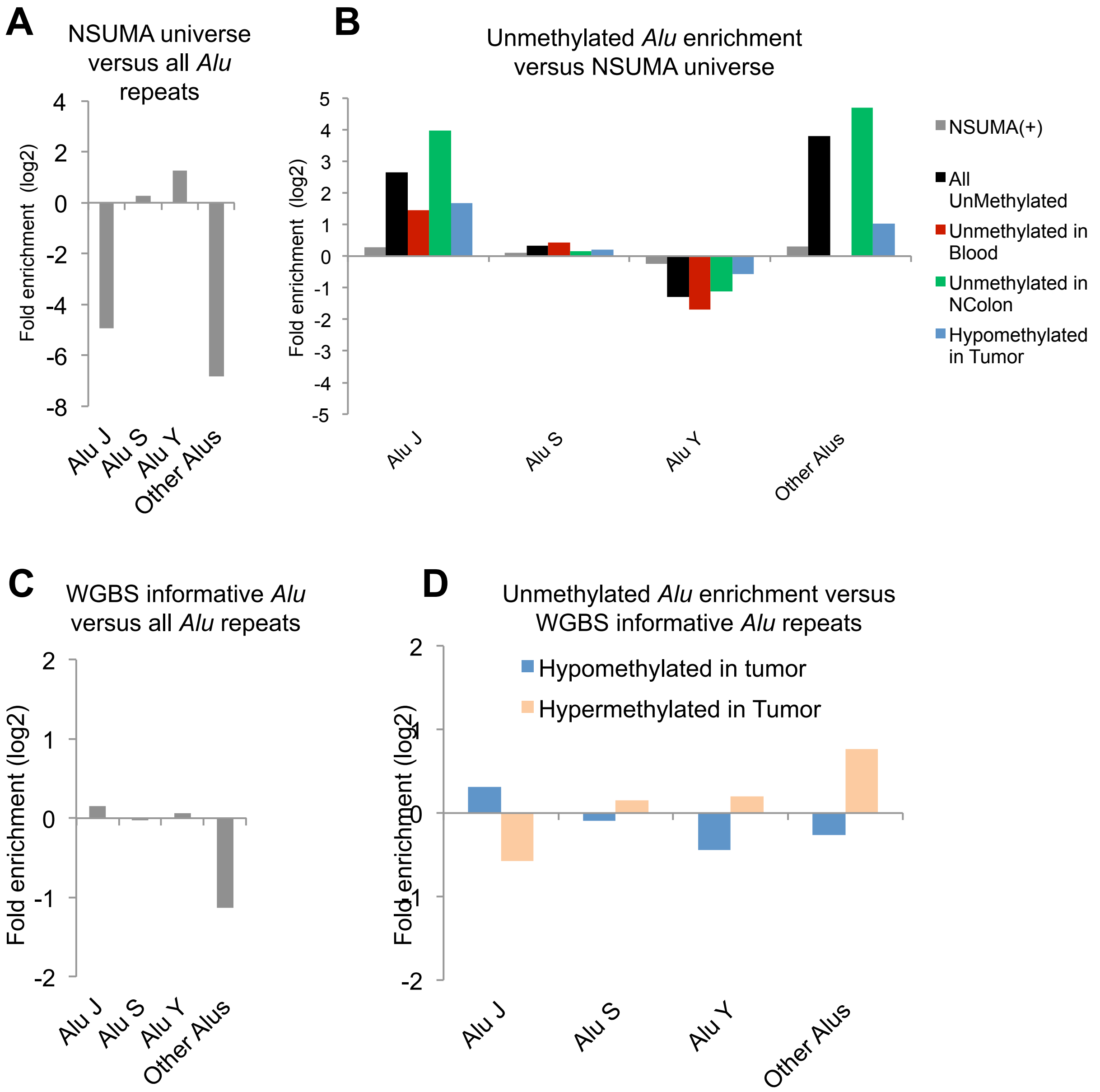
Supplemental Fig. S14. (**A**) Comparison of DNA methylation levels (beta value determined by WGBS) in *Alu* and LINE repeats between normal and colon carcinomas. (**B**) UCSC Genome Browser views of the mean differential methylation between colon tumors and the paired normal tissue as determined by NSUMA and WGBS in two regions. WGBS data were obtained from Hansen et al (*Nat Genet 2011, 43: 768-775).* (**C**) Tumor differential methylation (beta value normal-beta value tumor analyzed by WGBS) in *Alu* repeats (left) and LINEs (right) in three colon cancer patients and the average of the three. **(D)** Tumor differential methylation of *Alu* repeats (left) and Lines (right) stratified by the degree of methylation in the normal cells. The mean of three colon cancer patients is represented. High (black) and intermediate methylated repeat elements (red) in normal tissue tend to be hypomethylated in tumors, while unmethylated elements in normal tissue (green) show no variations in the tumor. Figure insets indicate the mean beta values intervals for high, intermediate and low methylation.



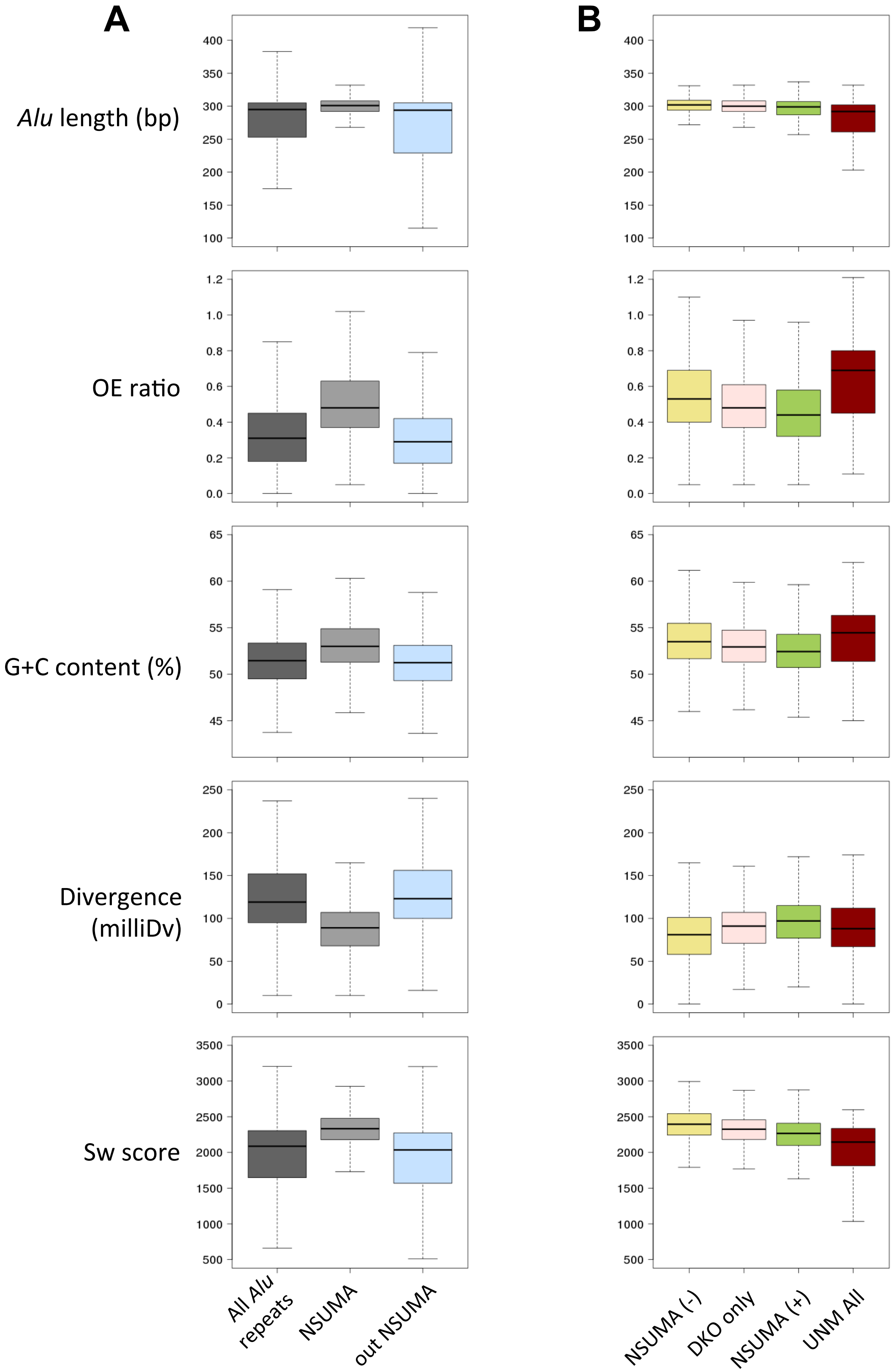
Supplemental Fig. S15. (**A**) Comparison of *Alu* differential methylation between three normal-colon tumor pairs analyzed by NSUMA (Y axis) and WGBS (X axis). NSUMA differential methylation ratio is represented as the log2 of the fold change of the normal normalized reads divided by the tumor normalized reads for each amplicon. Positive and negative values indicate hypermethylation and hypomethylation in the tumor, versus the normal, respectively. WGBS differential methylation is represented as the difference between beta values of the tumor and the normal tissues: positive values indicate hypermethylation, while negative values indicate hypomethylation. Only *Alu* repeats with at least 3 informative CpGs and a minimum of 3 normalized reads all the samples were considered. (**B**) Correlation between the number of reads in NSUMA amplicon and the methylation level measured by bisulfite sequencing.



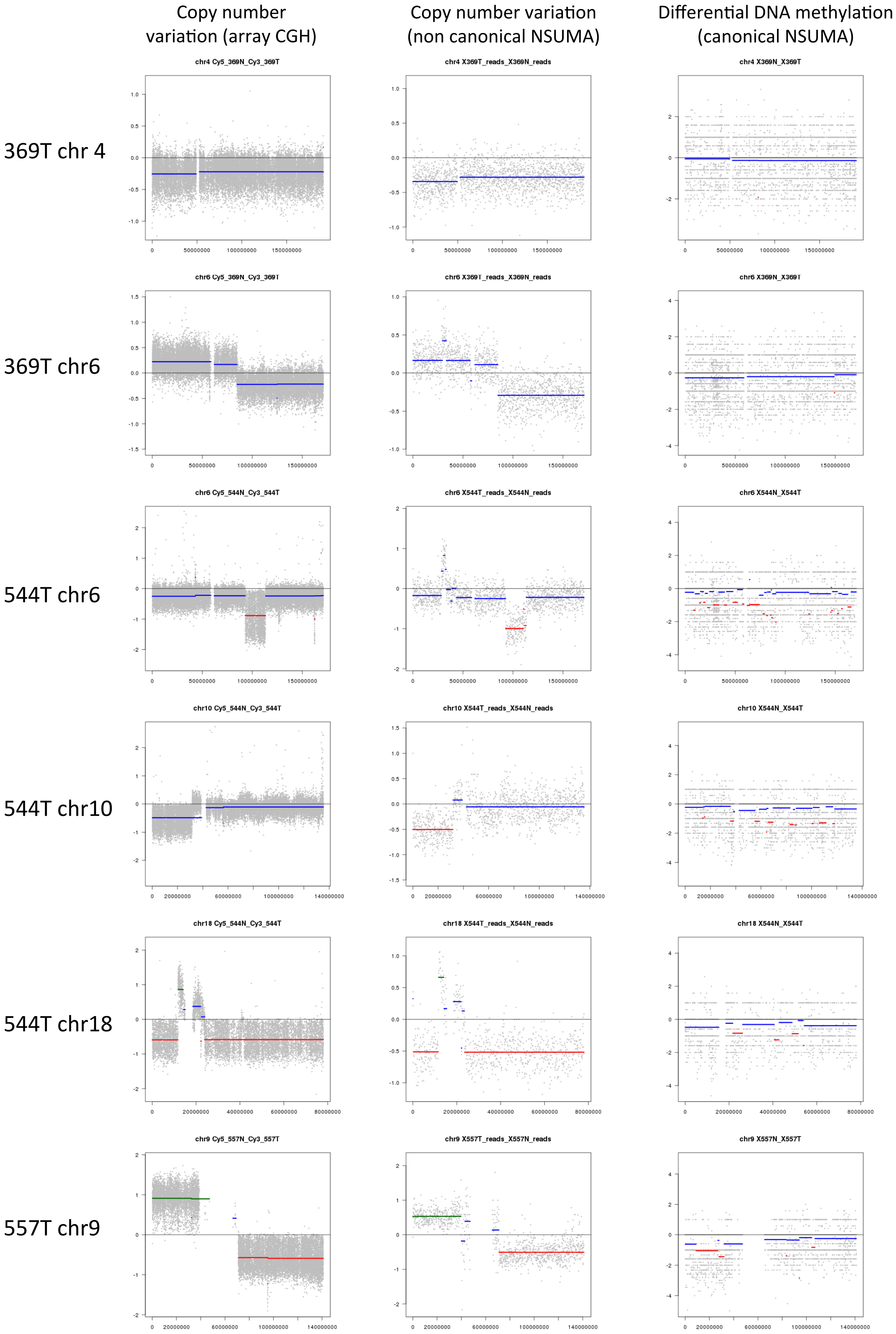
Supplemental Fig. S16. Principal component analysis (PCA) of samples using NSUMA data for different genomic elements. Circles group samples by tissue type except for the CpG islands, that group individuals. The number of elements considered in each analysis and the explained variation of the two principal components (component 1 + component 2) is indicated on top of each graph .



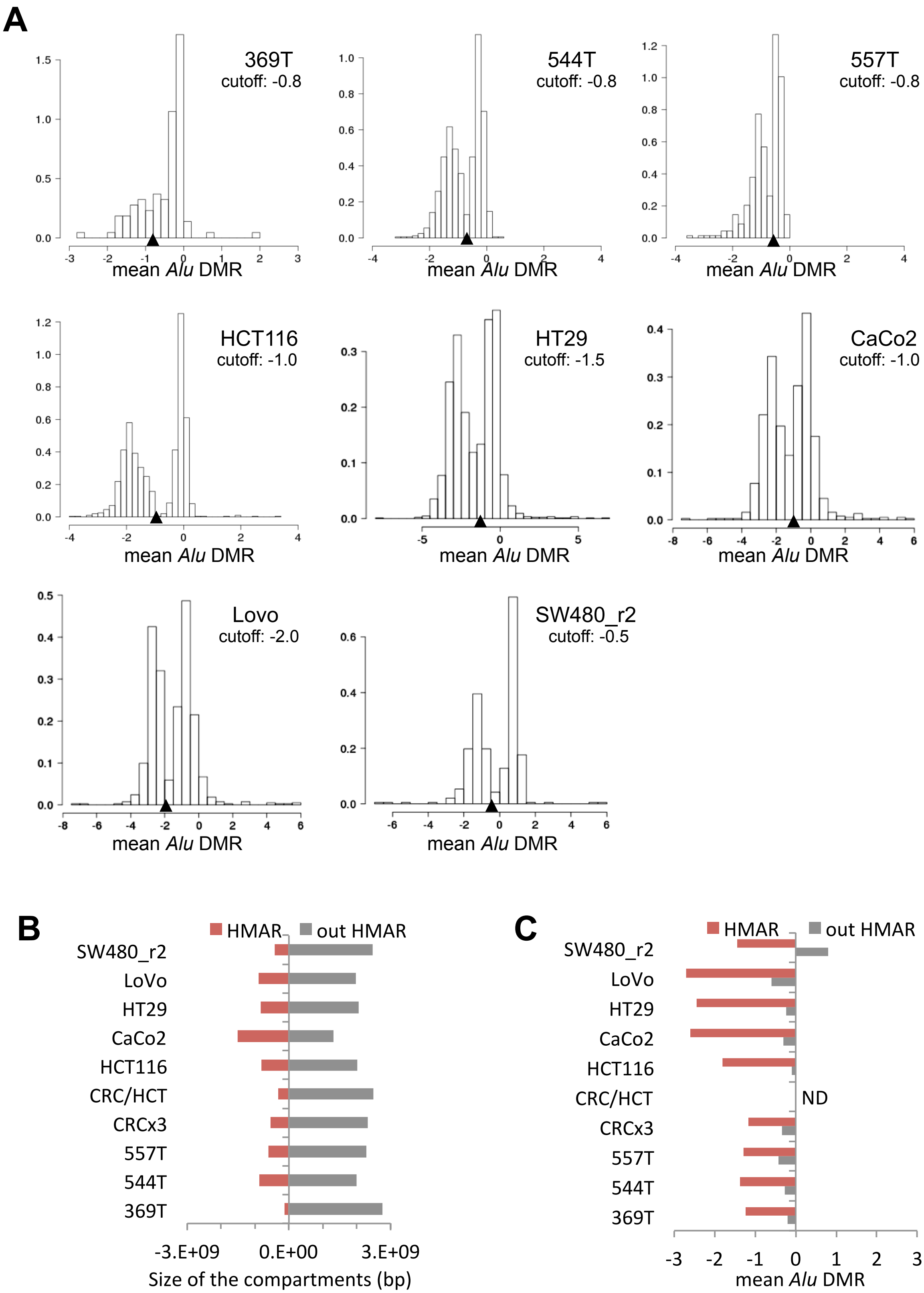
Supplemental Fig. S17. (**A**) *Alu* families bias in NSUMA representation. Y axis indicates the log 2 of the ratio of the subset of *Alu* repeats included in the NSUMA universe and all the *Alu* repeats in the human genome. The NSUMA universe is enriched in the *Alu* Y family (2.9x), with a balanced presence of *Alu* S repeats (1.1x) and an important depletion of the rest of families. This is, in fact, a bias towards *Alu* repeats with high CpG content because highly conserved *Alu* repeats, either young or old, are more likely to have a SmaI site and, therefore, be represented in NSUMA. Moreover, *Alu* repeats with high CpG content are the ones in which DNA methylation changes might have a stronger epigenetic impact. (**B**) Enrichment of *Alu* families in different experimental subsets in reference to the *Alu* repeats included in the NSUMA universe. NSUMA(+) represents the *Alu* repeats detected as unmethylated in at least one sample. Interestingly, *Alu* Y family unmethylation was depleted of all samples, but especially in normal tissues. (**C**) WGBS presented a balanced representation of *Alu* families as compared in regard to the whole human *Alu*ome. (**D**) WGBS analysis higher rates of hypomethylation in *Alu* J (8.6%) than *Alu* Y (5.1%) in tumors, consistent with NSUMA data.



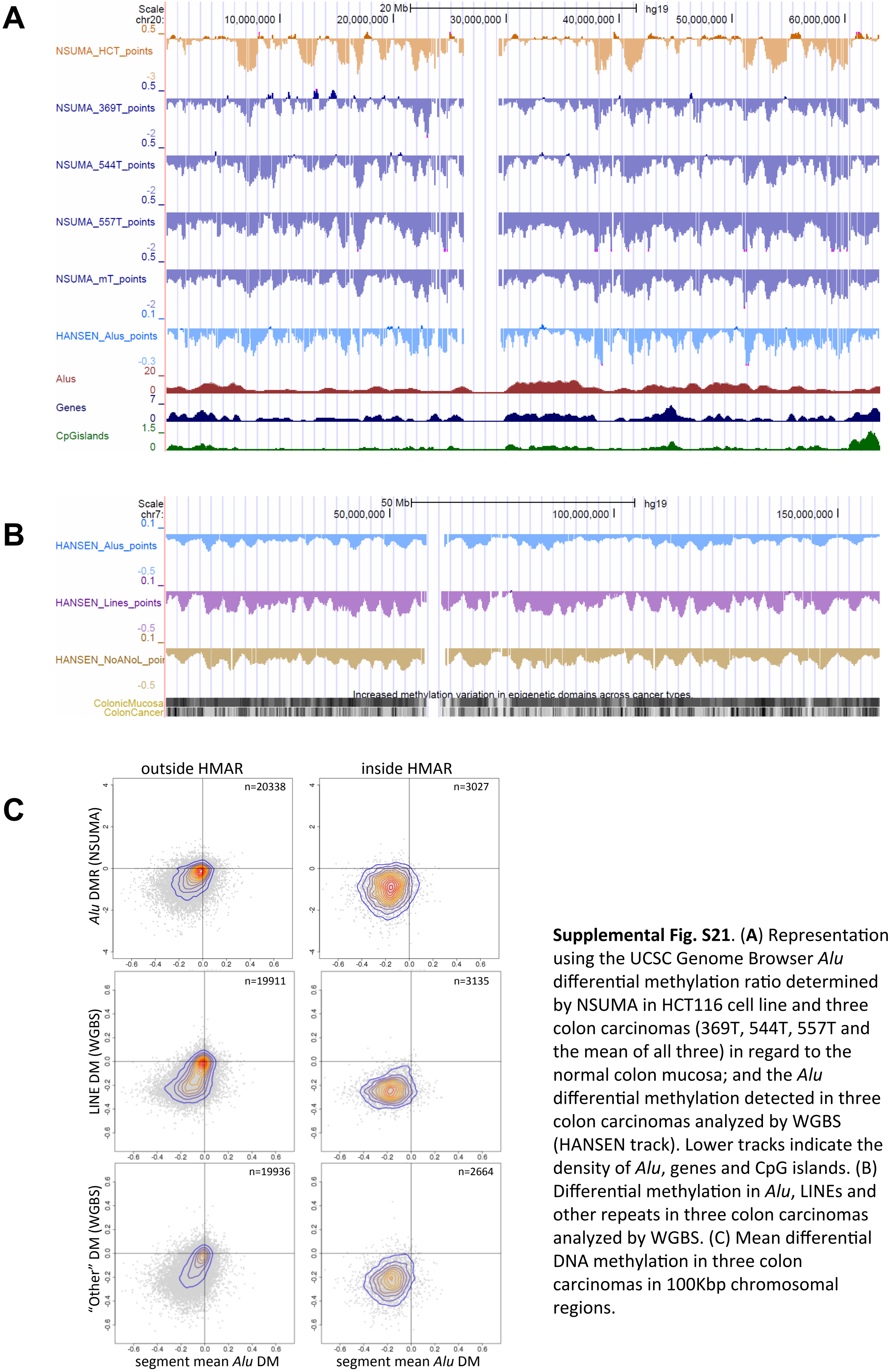
Supplemental Fig. S18. (**A**) Representativeness of *Alu* repeats included in the NSUMA virtual universe and the rest (out NSUMA) in regard to different genomic features. (**B**) Feature distribution of subsets of *Alu* repeats of the NSUMA universe according to their experimental representation: NSUMA (-), <5 *nreads* in all samples; DKO only, < 5 *nreads* in DKO but <5 *nreads* in the rest of samples, NSUMA (+), ≥5 *nreads* in at least one sample, UNM All, ≥ 5 *nreads* in all the samples.



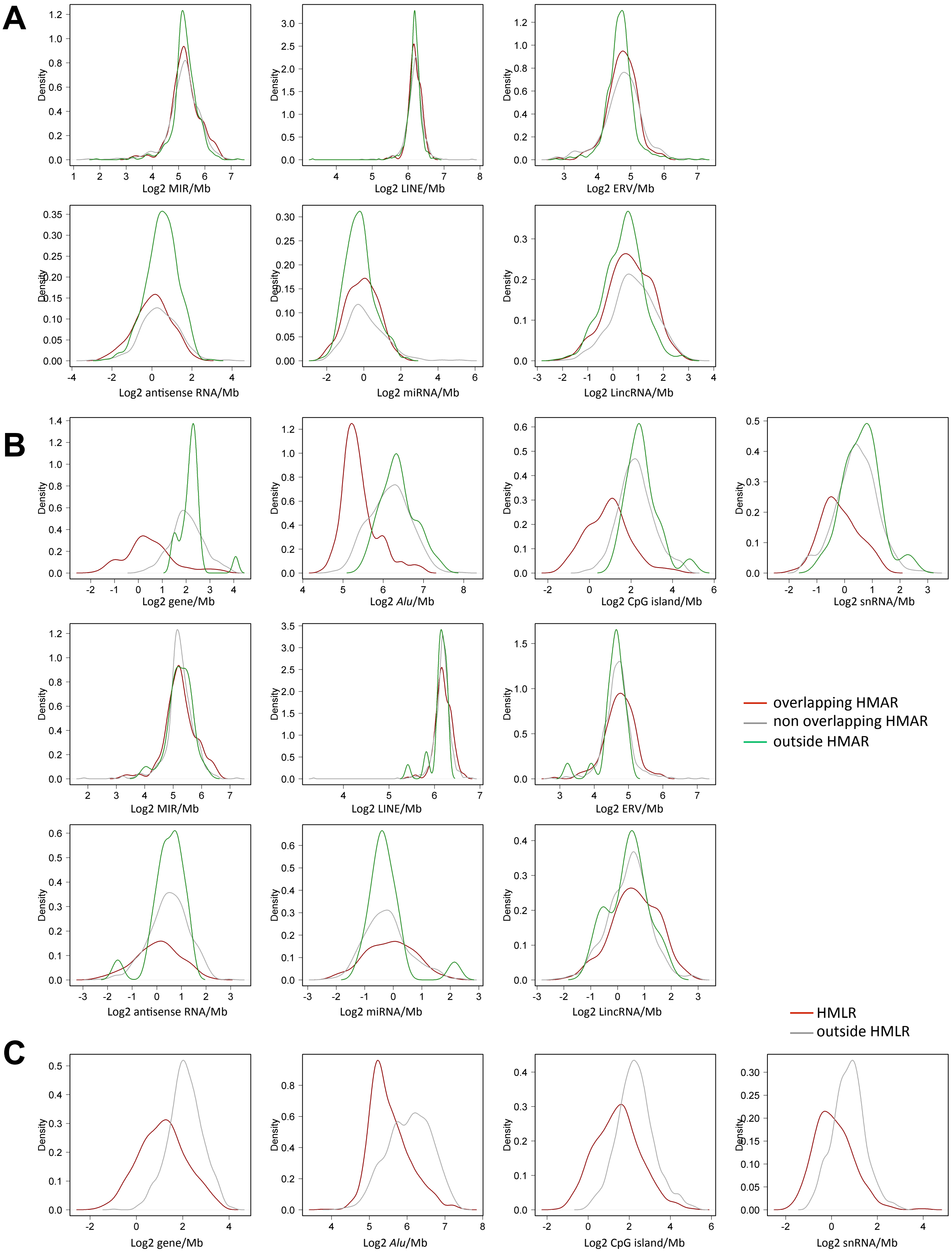
Supplemental Fig. S19. Chromosomal profiles of copy number variations determined by array CGH and counting of non canonical NSUMA reads are highly consistent. *Alu* Differential Methylation Ratios analyzed by NSUMA do not show to be affected by copy number variations. A circular binary segmentation algorithm was used to identify regions with different copy number of differential methylation, as appropriate. Y-axis displays log2 of the ratio, X-axis shows the chromosome position.



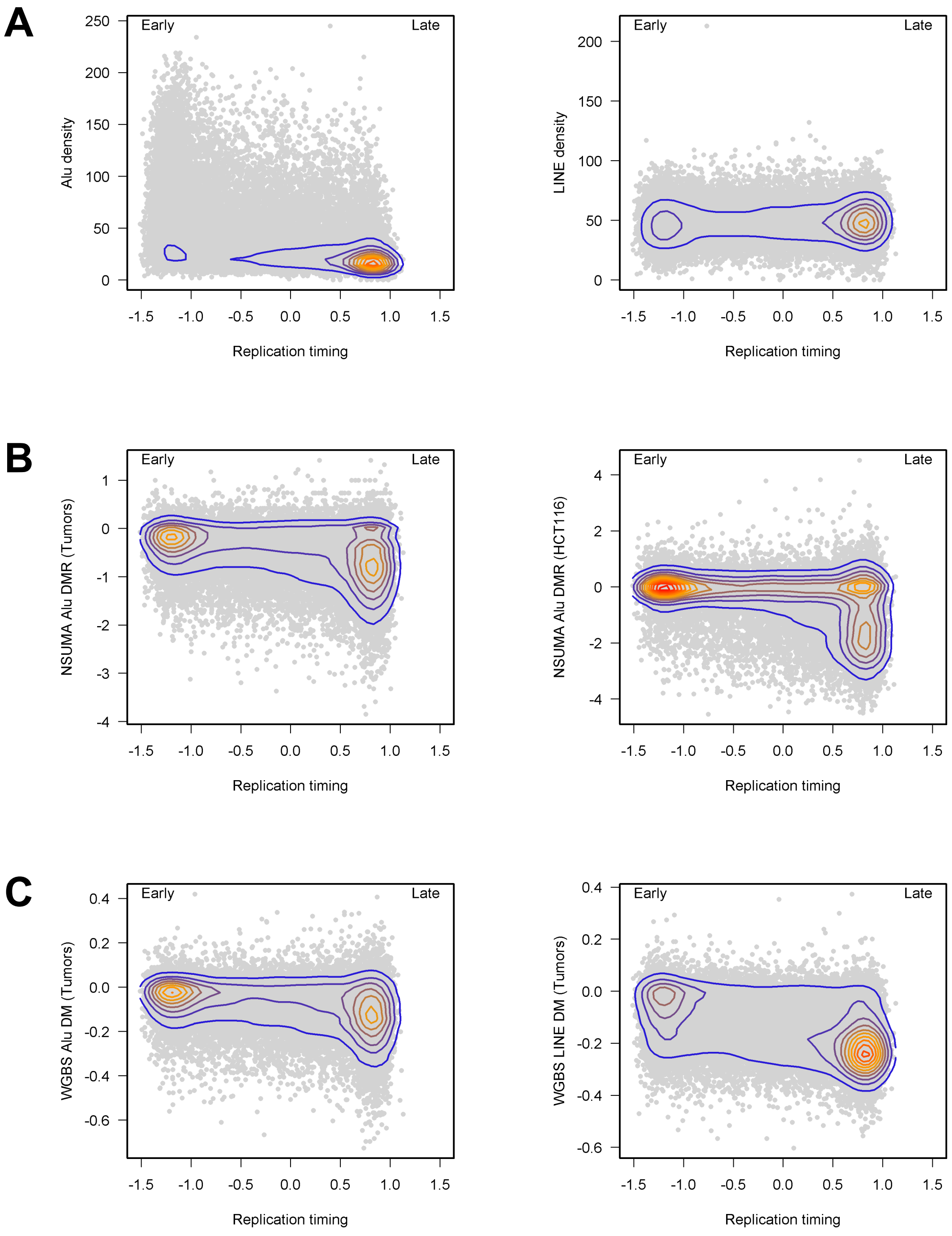
Supplemental Fig. S20. (**A**) Distribution of genomic segments according to the mean *Alu* Differential Methylation Ratio (*Alu* DMR). The arrowhead indicates the cutoff value used to define the two genomic compartments. Regions with mean *Alu* DMRs below the cutoff value were considered Hypomethylated *Alu* regions (HMAR). (**B**) Size of genomic compartments comprised in the HMAR and outside the HMAR in each of the samples analyzed. (**C**) Mean differential methylation ratio (log2 (normalized reads normal / normalized reads tumor) of the two genomic compartments. CRCx3 represents the combined analysis of three colon tumors against the normal tissue, CRC/HCT represents the CRC HMARs shared with the HCT116 cell line (see Supplemental Table 8 for details).



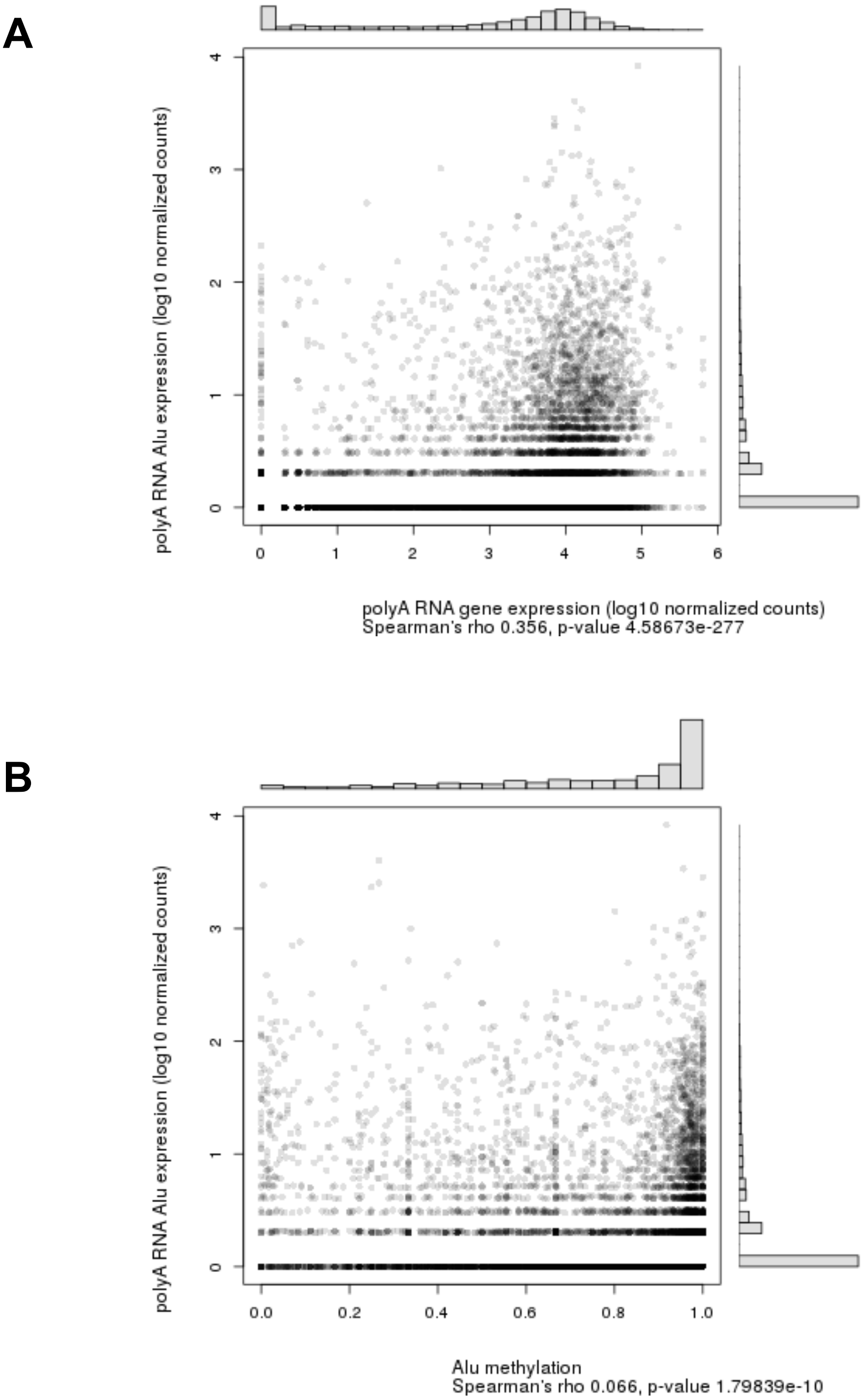
Supplemental Fig. S21. (**A**) Representation using the UCSC Genome Browser *Alu* differential methylation ratio determined by NSUMA in HCT116 cell line and three colon carcinomas (369T, 544T, 557T and the mean of all three) in regard to the normal colon mucosa; and the *Alu* differential methylation detected in three colon carcinomas analyzed by WGBS (HANSEN track). Lower tracks indicate the density of *Alu*, genes and CpG islands. (B) Differential methylation in *Alu*, LINEs and other repeats in three colon carcinomas analyzed by WGBS. (C) Mean differential DNA methylation in three colon carcinomas in 100Kbp chromosomal regions.



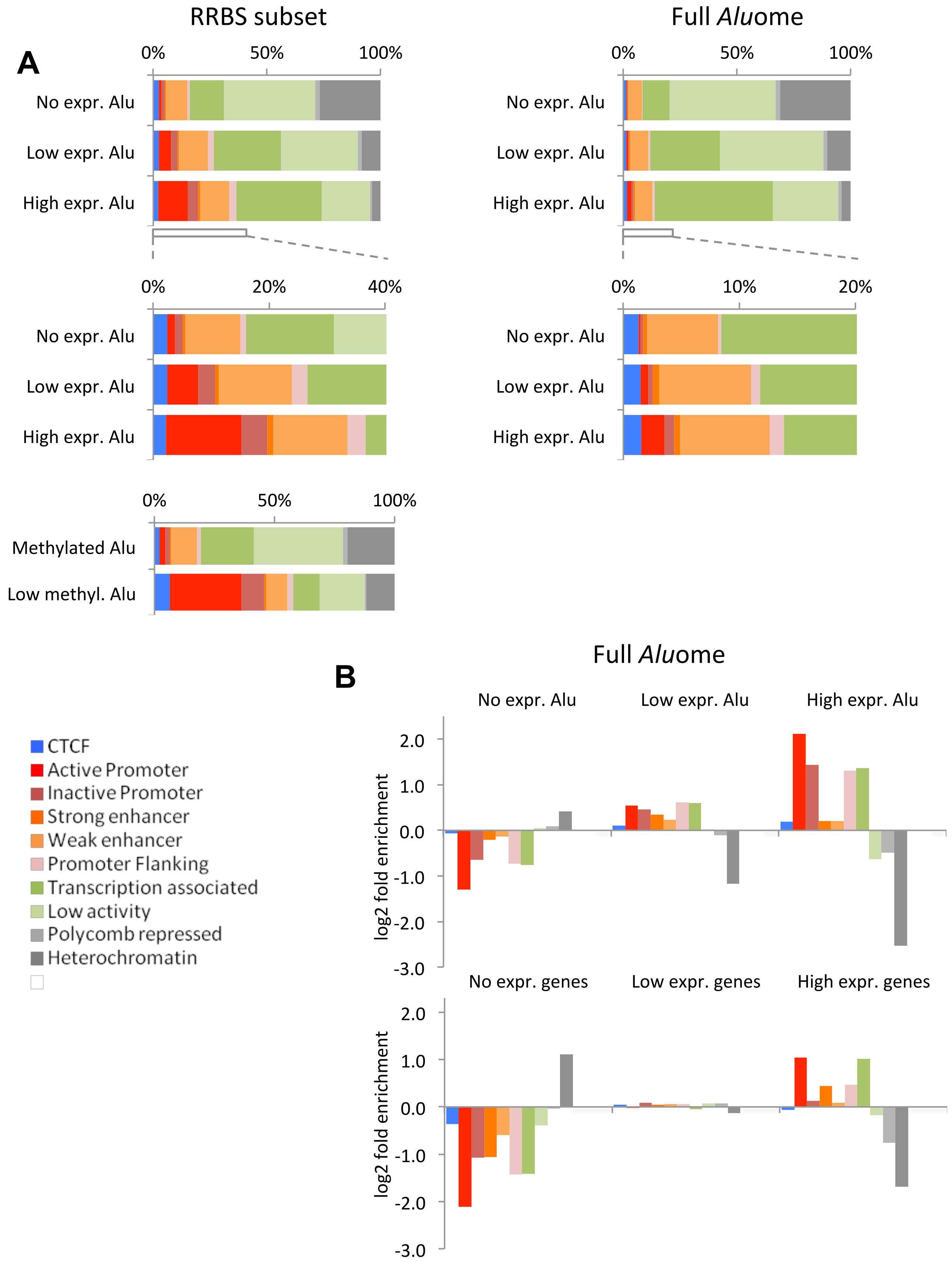
Supplemental Fig S22. (**A**) Density of genomic elements in genomic segments according to the *Alu* hypomethylation profiles determined by NSUMA. Overlapping HMAR regions (red line) result from the shared HMARs of three colon tumors and the HCT116 cell line (Supplemental Table 8). Non overlapping HMARs (grey line) correspond to hypomethylated regions in tumors or HCT116 but not in both. The rest of the genome (outside HMARs, green line) is neither hypomethylated in tumors nor HCT116 cells. The distribution of other genomic elements is shown in Fig. 5A. (**B**) Density of genomic elements in genomic segments according to the *Alu* hypomethylation profiles determined by WGBS following the same nomenclature. (**C**) Density of genomic elements in hypomethylated LINE regions (HMLR, red line) and the rest of genomic segments (gray line) as determined by WGBS.



Supplemental Fig S23. (**A**) Distribution of *Alu* and LINE repeats according to the its DNA replication timing in IMR90 cells. (**B**) Distribution of *Alu* repeats according to the its differential methylation ratio measured by NSUMA in three colon tumors and the HCT116 cell line in relation to IMR90 replication timing. (**C**) Distribution of *Alu* and LINE repeats according to the its differential methylation measured by WGBS in three colon tumors in relation to IMR90 replication timing.



Supplemental Fig S24. (**A**) Correlation between *Alu* expression and that of the nearest gene. Expression data were obtained from two replicates of poly(A)$^+$ RNA-seq of the HCT116 cell line (see Supplemental Material and Methods). (**B**) Correlation between *Alu* expression and DNA methylation levels. DNA methylation information was obtained by RRBS (see Supplemental Material and Methods).



Supplemental Fig S25. (A) Distribution of chromatin states in *Alu* elements included in the RRBS subset (n=9,339) and the full *Alu*ome (n=1,107,717) according to the expression levels and the *Alu* itself (top) and the associated gene (bottom), as well as the DNA methylation status in HCT116 cells. (**B**) Enrichment of chromatin states in *Alu* elements according to the expression levels and the *Alu* itself (top) and the associated gene (bottom) in HCT116 cells. *Alu* repeats with high expression were enriched in promoter associated histone marks. See supplemental Material and Methods for details.

The epigenetic landscape of *Alu* repeats delineates the structural and functional genomic architecture of colon cancer cells

Mireia Jordà et al.

SUPPLEMENTAL TABLES

Supplemental Table S1. NSUMA universe in the human genome (hg19)

Genomic Element	No in the human genome	Average size (bp)	Cumulated size (bp)	% of the genome	no of elements with SmaI site[1]	no of elements represented in NSUMA[2]	no of amplicons in NSUMA[3]
CpG island	27,718	761	21,842,742	0.70%	16,792	1,547	661
***Alu* (total)**	1,194,734	261	311,740,887	10.10%	187,697	155,246	135,282
Alu J	312,138	236	73,692,917	2.38%	13,161	1,119	940
Alu S	686,962	279	191,473,741	6.18%	116,814	100,510	87,191
Alu Y	143,178	279	39,909,490	1.29%	56,919	53,567	47,109
Other *Alu*	52,456	127	6,664,739	0.21%	803	50	42
LINE	1,418,218	440	623,661,160	20.15%	10,917	913	726
MIR	595,094	143	84,971,757	2.74%	4,075	926	761
ERV	694,835	379	263,304,945	8.50%	12,550	1,382	1,147
Other repeats	1,395,249	132	183,718,239	5.93%	11,051	567	400
Rest	-	-	1,606,437,682	51.89%	-	7,251	5,131
TOTAL	-	-	3,095,677,412	100.00%	-	167,832	144,108

[1] Genomic elements containing at least one SmaI site (CCCGGG) inside the sequence

[2] NSUMA includes canonical amplicons (20-1000 bp) flanked by SmaITT-AASmaI and SmaITT-MseI based on hg19 reference genome and with at least 1 read in one sample.

[3] The number of different SmaI sites analyzed is 143,897: 143,686 are represented by a single amplicon and 211 are represented by two amplicons (this makes a total of 144,108 amplicons). For the sake of simplicity, SmaI sites represented by two amplicons (n=211) are considered as two independent points in calculations.

Supplemental Table S2. Distribution of reads from NSUMA products in different samples

Experiment Code	Patient ID	Tissue	no of reads	Filtered-in [1]	Mapped	Unambiguous mapping	Ambiguous mapping	Non-canonical NSUMA [2]	canonical NSUMA [3]
ZUMA-0004	369	Blood	12,385,295	10,815,329	9,919,643	6,119,917	3,799,726	5,663,186	456,731
ZUMA-0008	544	Blood	14,724,716	13,407,853	12,726,308	8,204,197	4,522,111	7,397,721	806,476
ZUMA-0011	557	Blood	10,482,334	9,669,007	9,157,026	5,954,544	3,202,482	5,387,073	567,471
ZUMA-0003	369	Normal Colon	13,716,233	12,169,837	11,226,521	6,765,306	4,461,215	6,076,524	688,782
ZUMA-0006	544	Normal Colon	10,427,855	9,632,860	9,097,633	5,835,708	3,261,925	5,232,545	603,163
ZUMA-0009	557	Normal Colon	11,358,188	10,429,612	9,950,983	6,237,246	3,713,737	5,520,424	716,822
ZUMA-0002	369	Tumor	14,415,492	12,681,166	11,743,274	6,638,440	5,104,834	5,675,790	962,650
ZUMA-0007	544	Tumor	12,275,754	11,341,042	10,800,070	6,086,788	4,713,282	5,005,558	1,081,230
ZUMA-0010	557	Tumor	13,420,745	12,263,314	11,766,494	6,353,998	5,412,496	4,963,215	1,390,783
ZUMA-0001	HCT116	HCT116_r1	17,675,783	15,497,388	13,934,300	6,057,981	7,876,319	4,947,396	1,110,585
ZUMA-0012	HCT116	HCT116_r2	13,432,728	12,007,136	11,445,200	3,844,961	7,600,239	3,395,206	449,755
ZUMA-0014	HCT116	HCT116_r3	22,859,452	20,255,385	19,472,733	7,672,832	11,799,901	6,716,951	955,881
ZUMA-0013	DKO	DKO_r1	14,624,565	13,475,393	12,489,863	7,296,515	5,193,348	5,729,799	1,566,716
ZUMA-0015	DKO	DKO_r2	25,140,031	22,603,228	22,020,674	7,677,290	14,343,384	5,847,393	1,829,897
ZUMA-0016	DKO	DKO_r3	24,954,452	22,453,548	21,913,316	7,332,709	14,580,607	5,553,201	1,779,508
Mean	-	-	15,459,575	13,913,473	13,177,603	6,538,562	6,639,040	5,540,799	997,763

[1] Reads available for study after trimming-out primer sequences and passing quality and length criteria.
[2] Unambiguous reads mapped outside NSUMA universe (NSUMA non canonical amplicons).
[3] Reads mapped in NSUMA canonical amplicons (flanked by SmaI-TT/AA-SmaI or SmaI-TT/MseI; length 20-1,000 bp).

Supplemental Table S3. Distribution of NSUMA normalized reads (*nreads*) by genomic element class[1]

Experiment Code	Patient ID	Tissue	filtered reads	normalization factor	nreads	CpG islands	*Alu*	ERV	LINE	MIR	Other repeats	Rest
ZUMA-0004	369	Blood	456,731	0.32	141,093	12,208	74,714	3,761	2,126	4,705	3,484	40,095
ZUMA-0008	554	Blood	806,476	0.17	133,181	11,693	71,247	3,782	2,292	4,900	3,235	36,032
ZUMA-0011	557	Blood	567,471	0.18	95,710	11,669	43,695	2,690	1,564	4,114	2,745	29,233
		mean Blood	610,226		123,328	11,857	63,219	3,411	1,994	4,573	3,155	35,120
ZUMA-0003	369	Normal Colon	688,782	0.43	284,037	12,517	182,085	7,777	4,752	7,844	4,878	64,184
ZUMA-0006	554	Normal Colon	603,163	0.29	171,520	11,552	89,832	6,009	3,195	6,530	4,181	50,221
ZUMA-0009	557	Normal Colon	716,822	0.20	134,803	11,532	63,674	4,763	2,745	6,226	3,575	42,288
		mean Colon	669,589		196,787	11,867	111,864	6,183	3,564	6,867	4,211	52,231
ZUMA-0002	369	Tumor	962,650	0.48	443,707	11,814	298,763	14,954	8,799	13,378	6,698	89,301
ZUMA-0007	554	Tumor	1,081,230	0.41	434,524	10,925	292,436	14,590	8,007	13,593	6,750	88,223
ZUMA-0010	557	Tumor	1,390,783	0.27	375,162	10,862	243,641	12,642	7,551	13,892	5,719	80,855
		mean Tumor	1,144,888		417,798	11,200	278,280	14,062	8,119	13,621	6,389	86,126

Experiment Code	Cell lines	filtered reads	normalization factor	nreads	CpG islands	*Alu*	ERV	LINE	MIR	Other repeats	Rest
ZUMA-0001	HCT116_r1	1,110,585	0.43	464,859	10,487	344,164	13,108	8,591	9,700	6,809	72,000
ZUMA-0012	HCT116_r2	449,755	0.85	389,471	10,987	288,028	10,418	7,018	7,492	5,436	60,092
ZUMA-0014	HCT116_r3	955,881	0.41	381,889	9,586	290,623	9,865	5,635	6,272	5,440	54,468
	mean HCT	838,740		412,073	10,353	307,605	11,130	7,081	7,821	5,895	62,187
ZUMA-0013	DKO_r1	1,566,716	1.00	1,566,71	14,930	1,294,930	27,842	21,017	26,516	12,179	169,30
ZUMA-0015	DKO_r2	1,829,897	0.61	1,115,91	12,184	918,327	19,796	12,638	17,159	9,378	126,42
ZUMA-0016	DKO_r3	1,779,508	0.59	1,053,42	11,802	866,504	19,037	11,973	15,800	8,885	119,42
	mean DKO	1,725,374		1,245,35	12,972	1,026,587	22,225	15,209	19,825	10,147	138,38
ZUMA-0042	SW480-r2	8,746,432	0.49	4,263,90	258,804	1,988,534	199,086	112,93	256,82	144,887	1,302,8
ZUMA-0038	CaCo2-wt	8,841,452	0.83	7,335,51	268,092	4,457,718	286,152	172,88	320,47	167,296	1,662,9
ZUMA-0045	LoVo	10,829,511	0.87	9,423,48	263,648	5,266,032	471,338	246,94	485,82	218,446	2,471,2
ZUMA-0044	HT29	8,215,098	1.00	8,215,09	265,197	4,384,443	413,958	207,92	470,27	182,459	2,290,8

[1] The assigned genomic element corresponds to the SmaI position of the amplicon in which the read maps.

Supplemental Table S4. Genomic elements represented in NSUMA amplicons

Genomic Element	No of elements in the genome (hg19)	NSUMA universe (canonical amplicons)	NSUMA (+) [1]	Positive in Colon (Normal + Tumors) [2]	Positive in normal tissue (Colon+Blood) [3]
CpG islands	28,691	661	458	368	363
Alu	1,194,734	135,282	87,209	23,038	8,086
Other repeats[4]	4,103,396	3,034	2,484	1,745	1,239
Rest	-	5,131	4,370	3,243	2,653
Total	-	144,108	94,521	28,394	12,341

[1] Amplicons with ≥ 5 nreads in at least one sample (including DKO cells) constitute the collection.
[2] Amplicons with ≥ 5 nreads in at least one sample among all colon tissues (normal and cancer) from cancer patients.
[3] Amplicons with ≥ 5 nreads in at least one sample among all normal tissues from cancer patients.
[4] Includes LINEs, MIR, ERV and other repeats from REPBASE (http://www.girinst.org/repbase/).

Supplemental Table S5. Differential NSUMA representation between different tissues [1]

	Blood vs Colon				Normal Colon vs Tumor		
	Informative[2]	Unmethylated in Blood	Unmethylated in NColon		Informative	Hypermethylated in Tumor	Hypomethylated in Tumor
CpG islands	515	1	0		525	6	4
Alu	81,826	206	243		102,420	4	3,118
ERV	893	2	23		988	0	78
LINE	614	3	21		657	0	57
MIR	689	5	33		715	0	80
Other repeats	356	0	13		373	1	31
Rest	4,587	16	232		4,749	6	299

	Normal Colon vs HCT116				HCT116 vs DKO		
	Informative[2]	Hypermethylated in HCT116	Hypomethylated in HCT116		Informative	Hypermethylated in DKO [3]	Hypomethylated in DKO
CpG islands	553	81	53		606	2	105
Alu	104,787	1,109	11,143		131,353	1,120	62,883
ERV	999	83	107		1,109	6	372
LINE	659	50	82		694	10	245
MIR	718	103	54		736	2	288
Other repeats	382	46	39		389	3	128
Rest	4,806	672	491		4,994	18	2,039

[1] Columns indicate the number of hypomethylated elements in the named tissue versus the counterpart (adjusted p-value <0.05 and |log 2 FC| >1).

[2] Amplicons with at least 1 normalized read in at least one of the samples included in the comparison.

[3] This figure is likely to represent false positives as the extensive demethylation of DKO cells results in a very high number of amplicons as compared with the rest of samples. This high number of amplicons is likely to affect the deepness of the coverage and diminish the representativeness of fully unmethylated elements and hence the normalization method. In addition to the detection of false hypermethylations, the expected bias may also result in an underestimation of hypomethylations in DKO cells.

Supplemental Table S6. Representativeness of *Alu* families in NSUMA, WGBS analyses

	Human genome	NSUMA universe	NSUMA informative[1]	NSUMA(+) [2]	DKO (exc) [3]	Unmethylated in All[4]	Unmethylated in Blood[5]	Unmethylated in Colon[6]
***Alu* J**	312,138	940	933	746	286	10	4	27
***Alu* S**	686,962	87,191	85,787	61,100	37,125	182	180	175
***Alu* Y**	143,178	47,109	45,905	25,329	16,068	32	22	39
Other *Alu*	52,456	42	42	34	11	1	0	2
TOTAL	1,194,734	135,282	132,667	87,209	53,490	225	206	243

	Human genome	WGBS informative[7]	Tumor hypomethylated[8]	Tumor hypermethylated[9]
***Alu* J**	312,138	34,168	2945	120
***Alu* S**	686,962	66,533	4318	386
***Alu* Y**	143,178	14,720	748	88
Other *Alu*	52,456	2,358	136	21
TOTAL	1,194,734	117,779	8,147	615

[1] *Alu* repeats belonging to the NSUMA universe with at least one normalized read in one or more samples. A total of 2,615 *Alu* included in the NSUMA universe produced no reads in any of the experiments.
[2] NSUMA informative *Alu* repeats with >=5 nreads in at least one of the samples analyzed.
[3] NSUMA informative *Alu* repeats with ≥5 nreads in DKO but <5 nreads in the rest of samples.
[4] *Alu* repeats unmethylated in all tissues analyzed (≥5 nreads).
[5] *Alu* repeats hypomethylated in blood as compared with normal colon mucosa (adjusted p-value <0.05 and |log 2 FC| >1).
[6] *Alu* repeats hypermethylated in blood as compared with normal colon mucosa (adjusted p-value <0.05 and |log 2 FC| >1).
[7] *Alu* repeats informative in the WGBS analysis (at least 3 informative CpG sites per *Alu* in all the samples).
[8] Mean methylation difference (Beta value) ≤ -0.2 and p<0.05
[9] Mean methylation difference (Beta value) ≥ 0.2 and p<0.05

Supplemental Table S7. Tukey's pairwise comparisons[1] adjusted p-values of CpG content in *Alu* repeats and the flanking sequences.

Alu	LOW 500 bp [2]			*Alu*			HIGH 500 bp [3]		
Compartment	**DKO (exc) [5]**	**NSUMA (+) [6]**	**UNM All [7]**	**DKO (exc)**	**NSUMA (+)**	**UNM All**	**DKO (exc)**	**NSUMA (+)**	**UNM All**
NSUMA(-) [4]	6.85E-14	<0.00E-19	<0.00E-19	<0.00E-19	<0.00E-19	1.62E-01	5.15E-14	3.61E-14	<0.00E-19
DKO(exc)		<0.00E-19	<0.00E-19		<0.00E-19	1.01E-09		<0.00E-19	<0.00E-19
NSUMA(+)			<0.00E-19			3.73E-14			<0.00E-19
ANOVA	<0.00E-99			<0.00E-99			<0.00E-99		

Alu	LOW 500 bp [2]		*Alu*		HIGH 500 bp [3]	
Compartment	**NCOLON [9]**	**REST [10]**	**NCOLON**	**REST**	**NCOLON**	**REST**
BLOOD [8]	5.23E-02	1.57E-01	1.49E-03	1.55E-05	5.12E-01	2.77E-01
NCOLON		1.73E-07		3.31E-14		2.91E-03
ANOVA	7.67E-08		3.59E-27		1.42E-03	

Alu	LOW 500 bp [2]		*Alu*		HIGH 500 bp [3]	
Compartment	**TUMOR [12]**	**REST [13]**	**TUMOR**	**REST**	**TUMOR**	**REST**
NCOLON [11]	9.36E-01	9.63E-01	8.97E-01	3.53E-01	7.48E-01	9.94E-01
TUMOR		<0.00E-19		<0.00E-19		<0.00E-19
ANOVA	1.09E-61		<0.00E-99		3.37E-64	

[1] Overall ANOVA was performed for each comparison followed by Tukey's pairwise comparisons. Detailed data are shown in Supplemental Data S2 and summarized in Supplemental Tables S4 and S6.

[2] *Alu* flanking region (500 bp) with the lower CpG content

[3] *Alu* flanking region (500 bp) with the higher CpG content

[4] NSUMA informative *Alu* repeats with <5 nreads in all the samples analyzed.

[5] NSUMA informative *Alu* repeats with ≥5 nreads in DKO and <5 nreads in the rest of samples.

[6] NSUMA informative *Alu* repeats with ≥5 nreads in at least one of the samples analyzed.

[7] *Alu* repeats unmethylated in all tissues analyzed (≥5 nreads).

[8] *Alu* repeats hypomethylated in blood as compared with normal colon mucosa (adjusted p-value <0.05 and |log 2 FC| >1).

[9] *Alu* repeats hypomethylated in normal colon mucosa as compared with blood (adjusted p-value <0.05 and |log 2 FC| >1).

[10] *Alu* repeats informative in blood and normal colon mucosa with no differential NSUMA representation. Full dataset is provided as Supplemental Data S2.

[11] *Alu* repeats hypomethylated in normal colon mucosa as compared with tumor (adjusted p-value <0.05 and |log 2 FC| >1).

[12] *Alu* repeats hypomethylated in tumor as compared with normal colon mucosa (adjusted p-value <0.05 and |log 2 FC| >1).

[13] *Alu* repeats informative in tumor and normal colon mucosa with no differential NSUMA representation.

Supplemental Table S8. Characteristics of hypomethylated *Alu* regions detected by NSUMA

	Tumor vs NColon			HCT116 vs NColon	
	HMAR^{Tumor} (FC<-0.8)	Outside HMAR		HMAR^{HCT116} (FC<-1.0)	Outside HMAR
no of regions	248	356		498	493
size total (bp)	544,343,109	2,336,572,019		808,090,459	2,018,137,398
mean	2,194,932	6,563,405		1,622,672	4,093,585
SD	3,035,166	9,601,180		1,900,437	5,015,329
min/max	10,180 / 29,245,454	62,318 / 8,3043,923		630 / 18,460,277	70,629 / 37,146,804
***Alu* mean DMR**[1]	-1.161	-0.345		-1.806	-0.091
No of *Alu* repeats[2]	133,063	1,034,571		179,462	982,618
***Alu*/Mb (enrichment)**	244 (0.603)	443 (1.092)		222 (0.540)	487 (1.184)
NSUMA informative[3]	15,270	82,257		21,361	78,072
No of HM *Alu* repeats[4]	1,494	1,553		8,046	3,611
% of HM *Alu* repeats	9.8	1.9		37.7	4.6
HM *Alu* enrich.[5]	3.132	0.604		3.213	0.395

	Overlapping HMAR in Tumor/HCT116		
	Overlapping HMAR	Non overlapping HMAR	Outside HMAR
no of regions	203	706	602
size total (bp)	324,917,124	687,960,885	1,797,040,110
mean size	1,600,577	974,449	2,985,116
SD	1,873,534	1,407,728	3,539,949
min/max	11,356 / 14,031463	0/18,460,277	0 / 26,286,752
No of *Alu* repeats[2]	71,776	165,032	919,838
***Alu*/Mb (enrichment)**	221 (0.537)	240 (0.583)	512 (1.243)
NSUMA informative [3]	8,767	17,330	70,706

[1] *Alu* Differential methylation ratio (*Alu* DMR) was calculated as the mean of: log2 ((Normal Colon nreads+1) /(Tumor nreads+1))

[2] Number of *Alu* repeats as annotated in the hg19 human genome map

[3] No of informative *Alu* repeats in the NSUMA universe

[4] No of hypomethylated (HM) *Alu* repeats as compared with NColon (adjusted p-value <0.05)

[5] Hypomethylated *Alu* fold enrichment respect the NSUMA informative universe.

Supplemental Table S9. Characteristics of hypomethylated *Alu* regions detected by WGBS

	Carcinoma vs Normal Colon (WGBS data) [1]		
	HMAR (FC<-0.15)	**Out HMAR**	**SMAR [2]**
no of segments	134	274	22
size total (bp)	328,339,685	2,526,708,463	99,839,644
mean size (bp)	2,450,296	9,221,564	4,538,166
SD size	2,573,685	11,761,469	3,582,394
min/max size	117 / 12,343,818	44,225 / 92,967,349	440,403 / 12,485,355
***Alu* DM mean [3]**	-0.216	-0.063	-0.025
No of *Alu* repeats [4]	74,260	1,084,683	56,243
***Alu*/Mb (enrichment)**	226 (0.557)	429 (1.058)	563 (1.387)

[1] Hypomethylated *Alu* Regions (HMAR) identified using Whole Genome Bisulfite Sequencing (WGBS) data from Hansen et al. (*Nat Genet* **43**: 768-775, 2011). *Alu* methylation beta value differences between 3 normal colon and 3 colon cancers were calculated. Only *Alu* repeats with ≥3 informative CpGs in all the samples (n= 117,779) have been considered.

[2] Stable methylation *Alu* regions (SMAR) are defined as regions with *Alu* DM > -0.05 in all three tumors. SMAR are a subset within the No HMAR compartment.

[3] Mean of the differential methylation (Normal colon beta value – Tumor beta value) of the *Alu* elements located in the considered chromosomal segments.

[4] Number of *Alu* repeats within each compartment as annotated in the human genome assembly (hg19).

Supplemental Table S10. Overlap between HMARs and Hypomethylated Blocks (HB) [1]

	Hypomethylated blocks (HB)	Non-Hypomethylated blocks
no of regions	14,461	19,791
size total (bp)	1,897,786,874	1,070,114,715
mean size (bp)	131,235	54,071
No of *Alu* repeats[2]	614,390	542,162
***Alu*/Mb**	324 (0.83)	507 (1.30)
NSUMA *Alu* repeats [3]	59,602 (0.97)	36,866 (1.06)
No of HM *Alu* repeats[4]	2,795 (1.45)	212 (0.2)
% HM *Alu* repeats [5]	4.69	0.57

	HMAR (NSUMA)	No HMAR (NSUMA)
no of regions	599	13861
size total (bp)	313,609,185 (97% of HMAR)	1,584,073,455
mean size (bp)	523,555	114,283
No of *Alu* repeats[2]	69,036	545,336
***Alu*/Mb**	220 (0.68)	344 (1.06)
NSUMA *Alu* repeats [3]	8,535 (0.87)	51,065 (1.03)
No of HM *Alu* repeats[4]	962 (2.08)	1,833 (0.79)
% HM *Alu* repeats [5]	11.27	3.59

[1] Hypomethylated blocks (HB) as reported in by Hansen et al. (*Nat Genet* **43**: 768-775, 2011). HMARs correspond to the comparison Normal Colon-Tumor/HCT116 described in Supplemental Table S8.

[2] Number of *Alu* repeats as annotated in the hg19 human genome map. Enrichment is calculated in regard to the considered genome fraction.

[3] No of *Alu* repeats in the NSUMA informative universe. Enrichment is calculated in regard to the no of *Alu* repeats in the considered genome fraction.

[4] No of NSUMA informative *Alu* repeats that are hypomethylated in the Tumor as compared with Normal Colon (adjusted p-value<0.05). Enrichment is calculated in regard to the no of NSUMA informative *Alu* repeats in the considered genome fraction.

[5] Percentage of NSUMA informative *Alu* repeats that are hypomethylated in the Tumor as compared with Normal Colon (adjusted p-value<0.05). Enrichment is calculated in regard to the number of NSUMA informative *Alu* repeats in the considered genome fraction.

Supplemental Table S11. Distribution of HMARs by NSUMA and WGBS

	HMAR (NSUMA/WGBS) [1]			
	+/+	+/-	-/+	-/-
no of regions	54	189	101	681
size total (bp)	109,839,939	211,668,229	67,476,269	1,697,760,743
mean size (bp)	2,034,073	1,119,938	668,082	2,493,041
SD size	1,984,650	1,512,955	1,134,875	2,909,835
min/max size	124,362 / 7,412,169	210 / 14,031,463	117 / 6,359,758	2,149 / 21,021,304
No of *Alu* repeats [2]	24,166	47,029	17,647	890,179
***Alu*/Mb**	220 (0.542)	222 (0.547)	262 (0.644)	524 (1.292)

[1] Overlapping HMAR identified by NSUMA and WGBS in colon tumors and the colon cancer cell line HCT116 (analyzed by NSUMA).

[2] Number of *Alu* elements in the region.

Supplemental Table S12. Analysis of differential DNA methylation in *Alu* repeats, performance comparison

	***Alu* targeted bisulfite sequencing (Xie et al, 2009)**	**Methyl-MAPS (Edwards et al, 2010)**	**WGBS (Hansen et al, 2011)**	**HT-TREBS [a] (Bakshi et al, 2016)**	**NSUMA (Jordà et al)**
Sequencing platform	454 Roche	SOLiD	SOLiD	IonTorrent	Illumina
Amount of DNA	1µg	10-15 µg	5 µg	1 µg	1 µg
No of reads/sample	$3.7 \cdot 10^5$	$2 \cdot 10^7$	$2 \cdot 10^9$	$3.9 \cdot 10^6$	$1\text{-}2 \cdot 10^7$
No of informative *Alu* repeats	31,178	Not reported	117,779 [b]	5,238	135,282
Methylation measure	Mean beta value	Differential representation	Mean beta value	Mean beta value	Differential representation

[a] Only *Alu* Ya5 and Yb8 were targeted and included in the study

[b] *Alu* repeats with at least 3 informative CpGs, this is covered by at least 3 reads each. Thirty six % of CpGs in *Alu* repeats were mappable and 18% were covered in this study (Hansen et al, Supplemental Table 4).

References

Bakshi A, Herke S, Batzer MA, Kim J. 2016. DNA methylation variation of human-specific *Alu* repeats. *Epigenetics* **11**: 163-173.

Edwards JR, O'Donnell AH, Rollins RA, Peckham HE, Lee C, Milekic MH, Chanrion B, Fu Y, Su T, Hibshoosh H et al. 2010. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res* **20**: 972-980.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D et al. 2011. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**: 768-775 .

Xie H, Wang M, Bonaldo Mde F, Smith C, Rajaram V, Goldman S, Tomita T, Soares MB. 2009. High-throughput sequence-based epigenomic analysis of *Alu* repeats in human cerebellum. *Nucleic Acids Res* **37**: 4331-4340.

Supplemental Table S13. Expression of *Alu* elements in relation to transcriptomic and epigenetic features (RRBS subset [a]).

	No of *Alu* repeats	*Alu* expression (mean ± SD)	Gene expression (mean ± SD)	DNA methylation (mean ± SD)	No of LMA [b]
RRBS subset	9,339	10.7 ± 125.4	12,477 ± 30,554	0.77 ± 0.26	417
Non expressed *Alu* repeats (0 read)	5,763	0	8,356 ± 22,497	0.76 ± 0.25	192
Expressed *Alu* repeats (>=1 read)	3,576	27.9 ± 201.5	19,865 ± 19,119	0.77 ± 0.28	225
High expressed *Alu* repeats (>=10 read)	936	98.5 ± 385.4	24,003 ± 50,935	0.74 ± 0.32	96

[a] Subset of the *Alu* repeats considered in the analyses by RNA-seq and NSUMA and with minimal information on DNA methylation by Reduced Representation Bisulfite Sequencing (see Supplemental Material and Methods).

[b] Number of Low Methylated *Alu* repeats, mean methylation of all informative CpGs <0.2.

Supplemental Table S14. Expression of *Alu* elements in relation to transcriptomic and epigenetic features and HMAR location (RRBS subset [a]).

	No of *Alu* repeats	*Alu* expression (mean ± SD)	Gene expression (mean ± SD)	DNA methylation (mean ± SD)	No of LMA [b]
HMAR	1,574	0.7 ± 7.2	2,430 ± 8,936	0.67 ± 0.25	89
Out HMAR	7,765	12.7 ± 137.4	14,514 ±32,894	0.79 ± 0.26	328

Non expressed *Alu* repeats (0 read)	No of *Alu* repeats	*Alu* expression (mean ± SD)	Gene expression (mean ± SD)	DNA methylation (mean ± SD)	No of LMA
HMAR	1,422	0	2,443 ± 9,235	0.67 ± 0.25	77
Out HMAR	4,341	0	10,293 ± 25,075	0.79 ± 0.24	115

Expressed *Alu* repeats (>=1 read)	No of *Alu* repeats	*Alu* expression (mean ± SD)	Gene expression (mean ± SD)	DNA methylation (mean ± SD)	No of LMA
HMAR	152	7.6 ± 22.0	2,310 ± 5,398	0.67 ± 0.27	12
Out HMAR	3,424	28.8 ± 205.9	19,865 ± 40,073	0.78 ± 0.28	213

High expressed *Alu* repeats (>=10 read)	No of *Alu* repeats	*Alu* expression (mean ± SD)	Gene expression (mean ± SD)	DNA methylation (mean ± SD)	No of LMA
HMAR	23	41.0 ± 44.0	2,402 ± 5,628	0.73 ± 0.30	2
Out HMAR	913	99.9 ± 390.0	24,547 ± 51,449	0.74 ± 0.32	94

[a] Subset of the *Alu* repeats considered in the analyses by RNA-seq and NSUMA and with minimal information on DNA methylation by Reduced Representation Bisulfite Sequencing (see Supplemental Material and Methods).

[b] Number of Low Methylated *Alu* repeats, mean methylation of all informative CpGs <0.2.

Supplemental Table S15. Expression of *Alu* elements in relation to transcriptomic and epigenetic features and HMAR location (full *Alu*ome).

All *Alu* repeats (n=1,107,717)	**No of *Alu* repeats**	***Alu* expression (mean ± SD)**	**Gene expression (mean ± SD)**
HMAR	170,812	0.8 ± 14.6	2,408 ± 9,687
Out HMAR	936,905	9.0 ± 103.0	13,163 ± 26,979

Expressed *Alu* repeats [a] (n=387,897)	**No of *Alu* repeats**	***Alu* expression (mean ± SD**	**Gene expression (mean ± SD)**
HMAR	13,261	9.8 ± 51.6	4,158 ± 13,005
Out HMAR	374,636	22.6 ± 161.9	17,782 ± 30,396

[a] *Alu* repeats with at least 1 overlapping normalized read.

Supplemental Table S16. Primers and adapters sequence

Primer Id	Use	Sequence 5’-3’
ADPT-S1	NSUMA blunt adapter	GATAGTATGCCCGGGTGA
ADPT-S2	NSUMA blunt adapter	5’P-TCACCCGGGCATAC
ADPT-M1	NSUMA sticky adapter	CTGAGGCTGGATCCCTG
ADPT-M2	NSUMA sticky adapter	5’P-TACAGGGATCCAGCCTCAG
ADPT-M1A	NSUMA PCR primer	CTGAGGCTGGATCCCTGTAA
ADPT-S1TT	NSUMA PCR primer	GATAGTATGCCCGGGTGAGGGTT
Illumina 1/2	Library adapters	5’P-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Illumina 3/4	Library amplification	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT
Ao2c1 F1	Bisulfite PCR primer	ATATTTAGAGAATTAAATGGTTT
Ao2c1 R1	Bisulfite PCR primer	CCATTTACCATACCCTATAA
Ao2c1 F2	Bisulfite PCR primer	TGGATTTTGGAGAAGTAGATTT
Ao2c1 R2	Bisulfite PCR primer	ACCTTCAAAACCCCATTCAA
Ao1c4 F1	Bisulfite PCR primer	GTTTAAAATTTTTTGTTTGGGA
Ao1c4 R1	Bisulfite PCR primer	ACCTATAACCTCTAAAACCA
Ao1c4 F2	Bisulfite PCR primer	GATATATGTTTATTTATTTATGTTT
Ao1c4 R2	Bisulfite PCR primer	ATTTTTTTCTTTTTCTATTTTTTTTAA
Aj2c1 F1	Bisulfite PCR primer	TAGATTTTAATGGGTAAAAGTA
Aj2c1 R1	Bisulfite PCR primer	CAATAACTTTACAATATACTAC
Aj2c1 F2	Bisulfite PCR primer	TGTGGTTGTATTAGTTAAGG
Aj2c1 R2	Bisulfite PCR primer	ACCAACATTTAAAAATACCTAA
LV183 F1	Bisulfite PCR primer	GTAGATTAAGTAGAAAGAGG
LV183 R1	Bisulfite PCR primer	ACCAACTAAATAAACAACATAA
LV183 F2	Bisulfite PCR primer	GGAGGTATTTGTTTTATTTTTTATGATG
LV183 R2	Bisulfite PCR primer	ACACACACTAACCAAAAATAAT