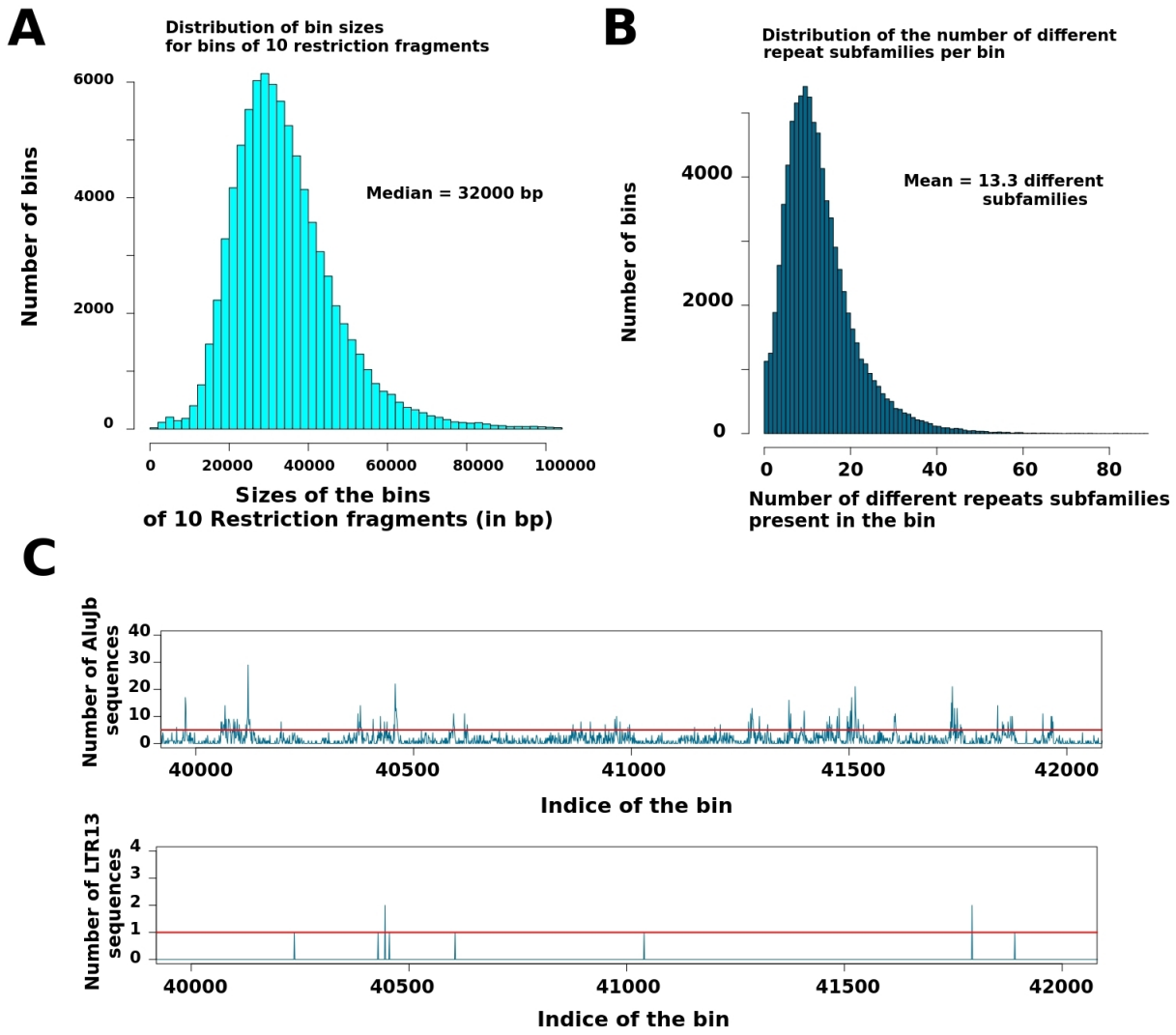
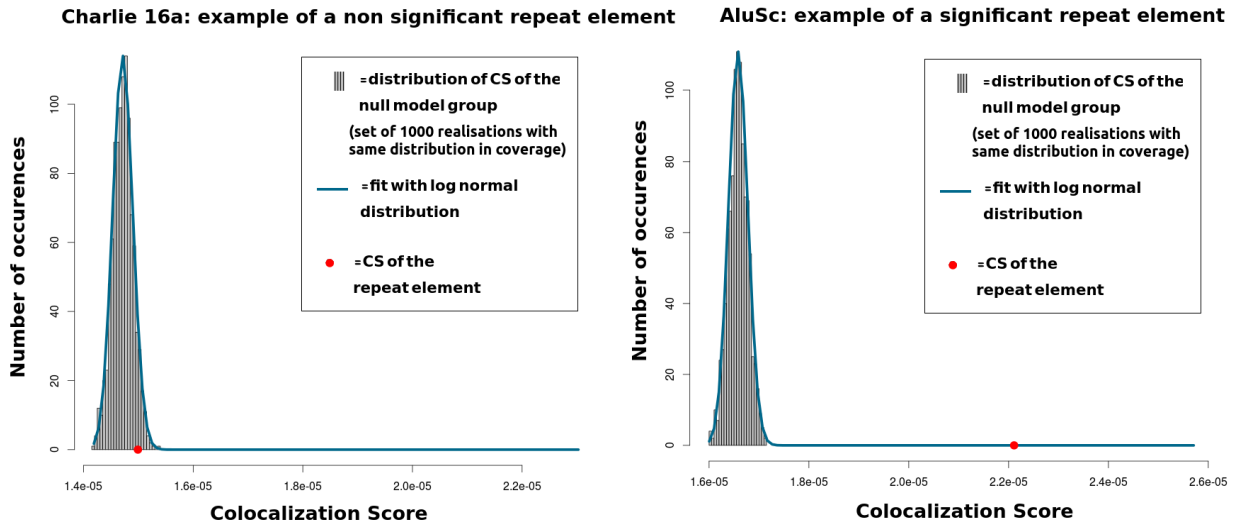
Supplementary Figures

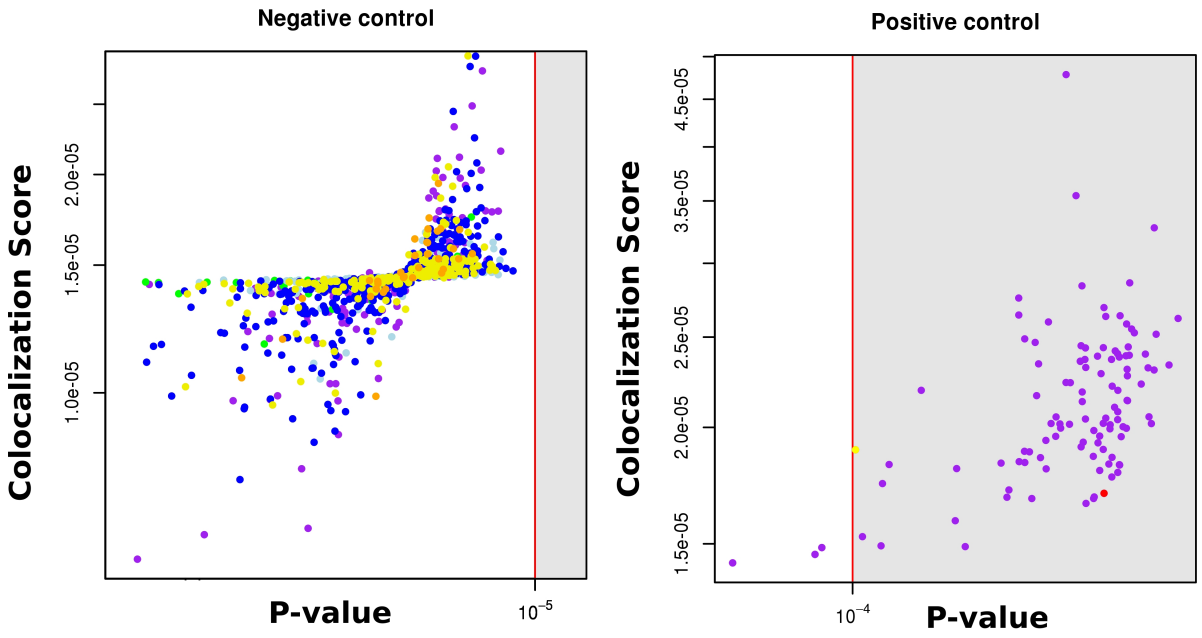


Supplementary Figure S1. A) Distribution of sizes of the bins used in the analysis. For the majority of libraries used in our study, each bin contains 10 consecutive HindIII fragments from a Hi-C experiment. The median size is ~32 kb for the human genome.
B) Distribution of the number of different subfamilies per bin.
C) Distribution along the genome (zoom) of the number of sequences per bin for 2 representative subfamilies: AluJb and LTR13. The vertical red line is the threshold above which bins are conserved for the computation of CS (which is respectively 5 and 1 for AluJb and LTR13).



Supplementary Figure S2. Illustration of the output of our pipeline to detect elements with a significant Co-localization Score.

For each repeat element, a random group of positions is generated with the same distribution in Hi-C coverage or GC content or the same chromosome distribution. 1000 realizations are generated which gives a distribution of the CS that can be expected by chance under the chosen null model (grey histogram). A log normal distribution is used to fit this distribution which allows to give a probability to observe a particular CS for the group of interest. Here, are presented the results for two repeat elements Charlie 16a and AluSc which are respectively present in 1607 bins and 1858 bins in the human genome (see Supplementary Tables S1 sheet B). The null model used here is the conservation in the coverage distribution. AluSc appears very significant das opposed to Charlie 16a. This statistical method is very similar to the one proposed by Witten et al. [1] and the use of null models is explained in details in [2].



Supplementary Figure S3. Negative and Positive controls of our pipeline for the detection of significant co-localizations of a group of elements

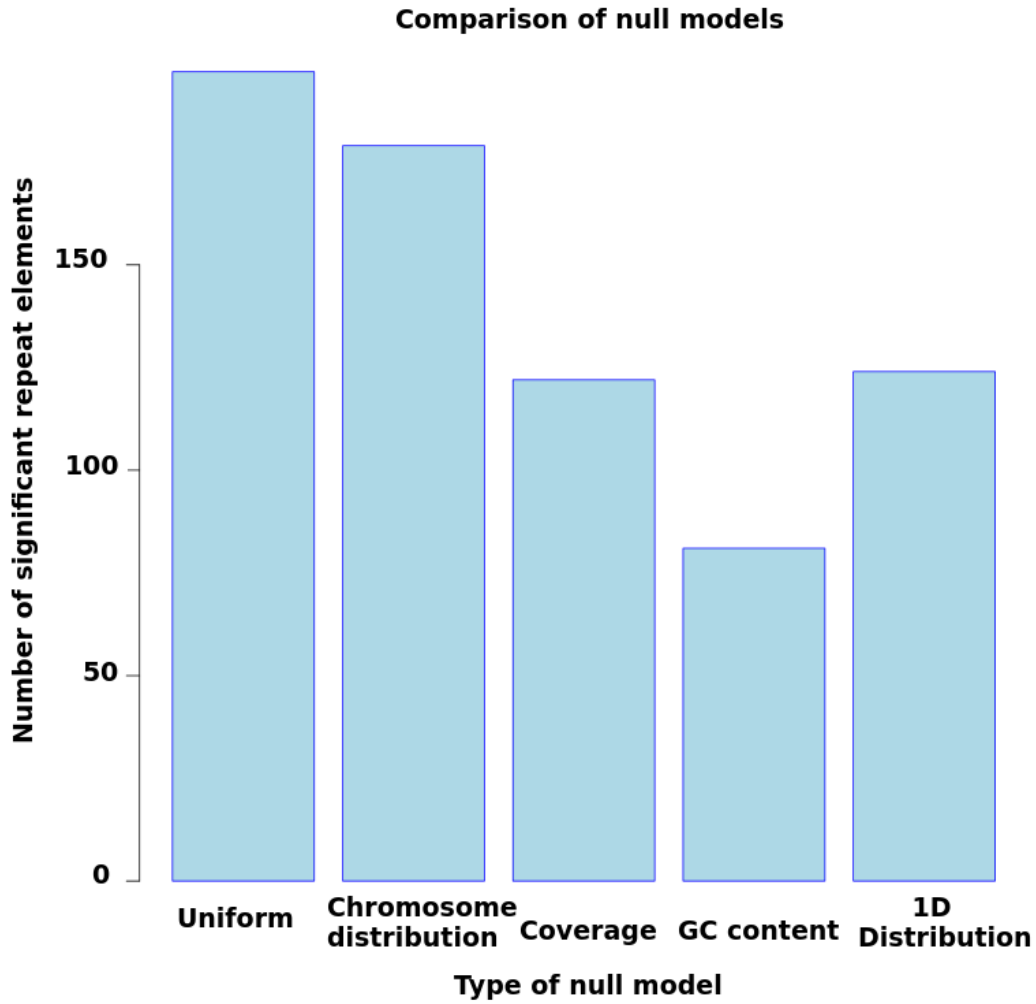
Left: Negative control We use the positions of all the repeat elements that we shifted with periodic conditions along the genome (value of the shift: 5000 bins i.e about 150 Mbp). None of the elements tested is significant confronted to the coverage null model. Colors correspond to repeat families as in Figure 2 of the main text.

Right: Positive control We use bins enriched in Transcription Factor Binding Sites, Histone modifications, tRNAs and NADs in the human embryonic stem cell.

Purple dots: Transcription Factors and Histone modifications. Yellow dot: tRNAs.

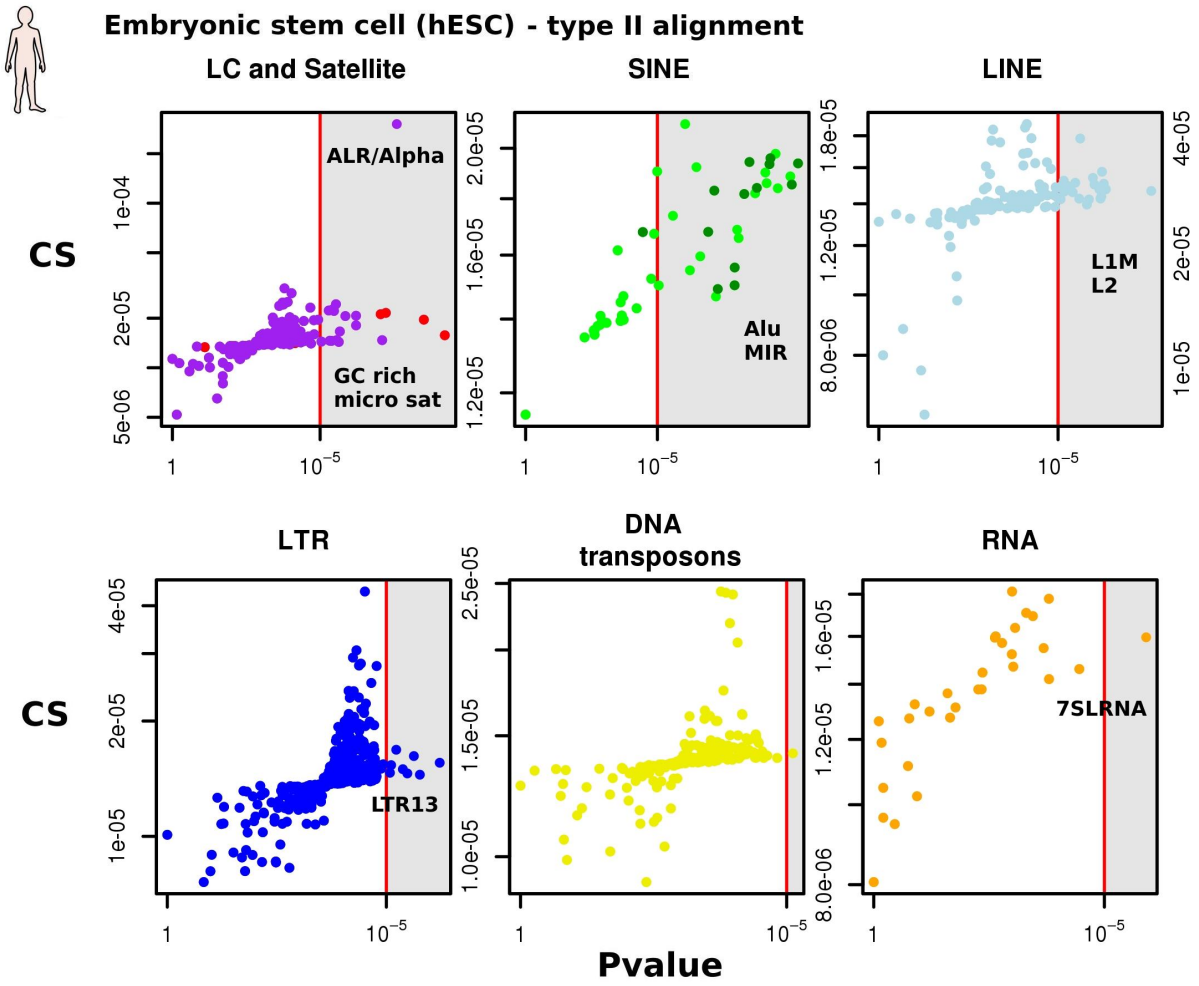
Red dot: NADs for Nucleolar Associated Domains (also known as Nuclear Organizing Regions [NORs]). 99 out of the 102 tested elements present significant CS when confronted to the coverage null model (see Table S1, sheet G), supporting the validity of the approach.

The most significant elements correspond to sets of bins enriched in opened chromatin, Pol2 and CTCF. Evidence of CTCF-mediated inter-chromosomal interactions has also been obtained using 4C (an advanced 3C technique) on the mouse *Igf2*/H19 locus [4-5] and as well with FISH techniques [6].



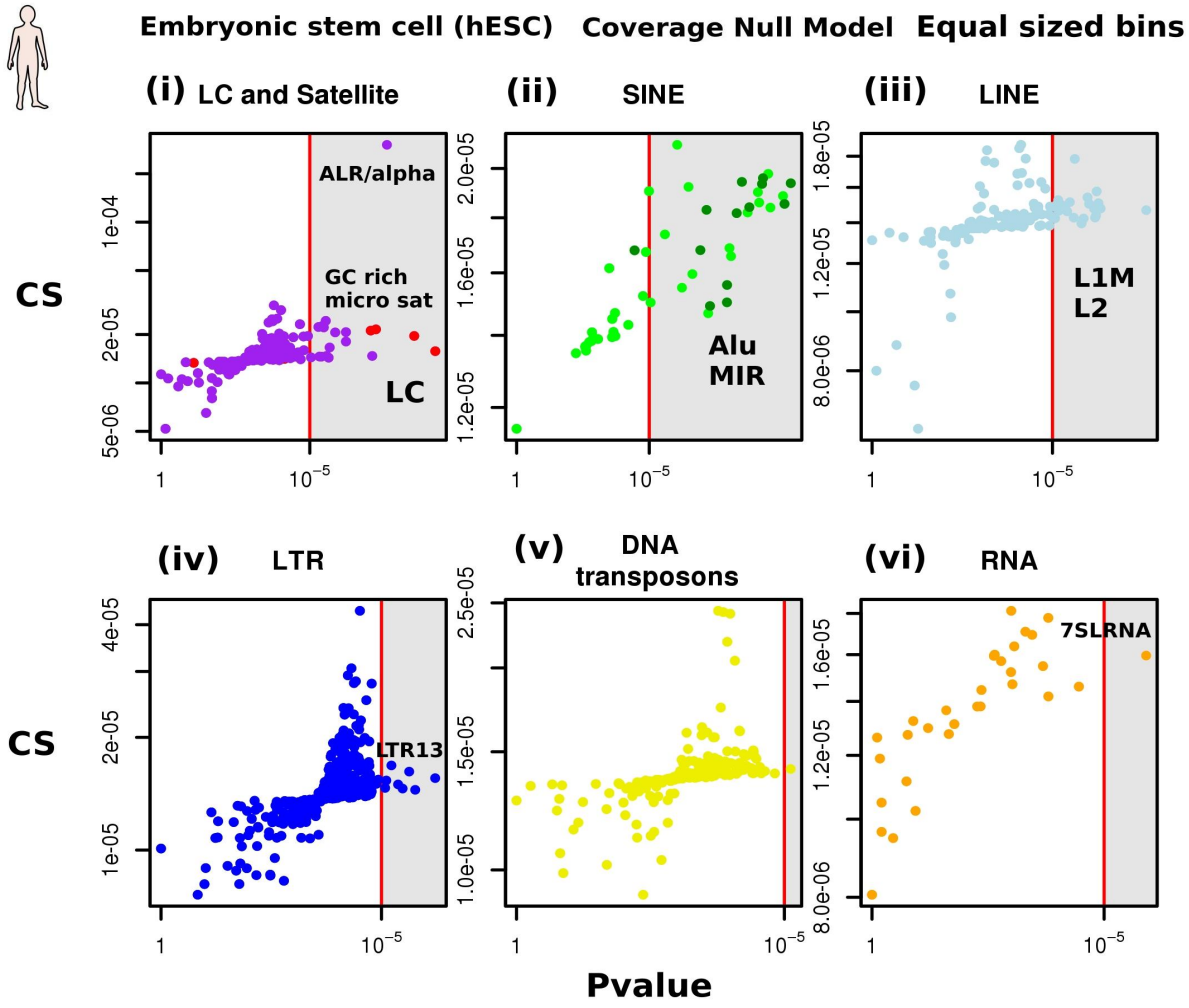
Supplementary Figure S4. Influence of the null model used in the number of significant repeat elements detected

The figure shows the number of significant repeat elements according to the null model used. The uniform null model assigns random bins in the genome without any constrain. The Chromosomes distribution null model conserve the distribution between the chromosomes when assigning the random bins. The coverage null model conserves the distribution in the reads coverage of the random bins. The GC content null model conserves the GC content distribution of the random bins. The 1D distribution null model conserves the distances between the random bins along the genome. The null model that conserves the GC content distribution of the group is the most stringent as it conserves the two compartments organization of the genome.



Supplementary Figure S5. Co-localization Scores and p-values of the different repeat elements for hESC with the *type II alignment*

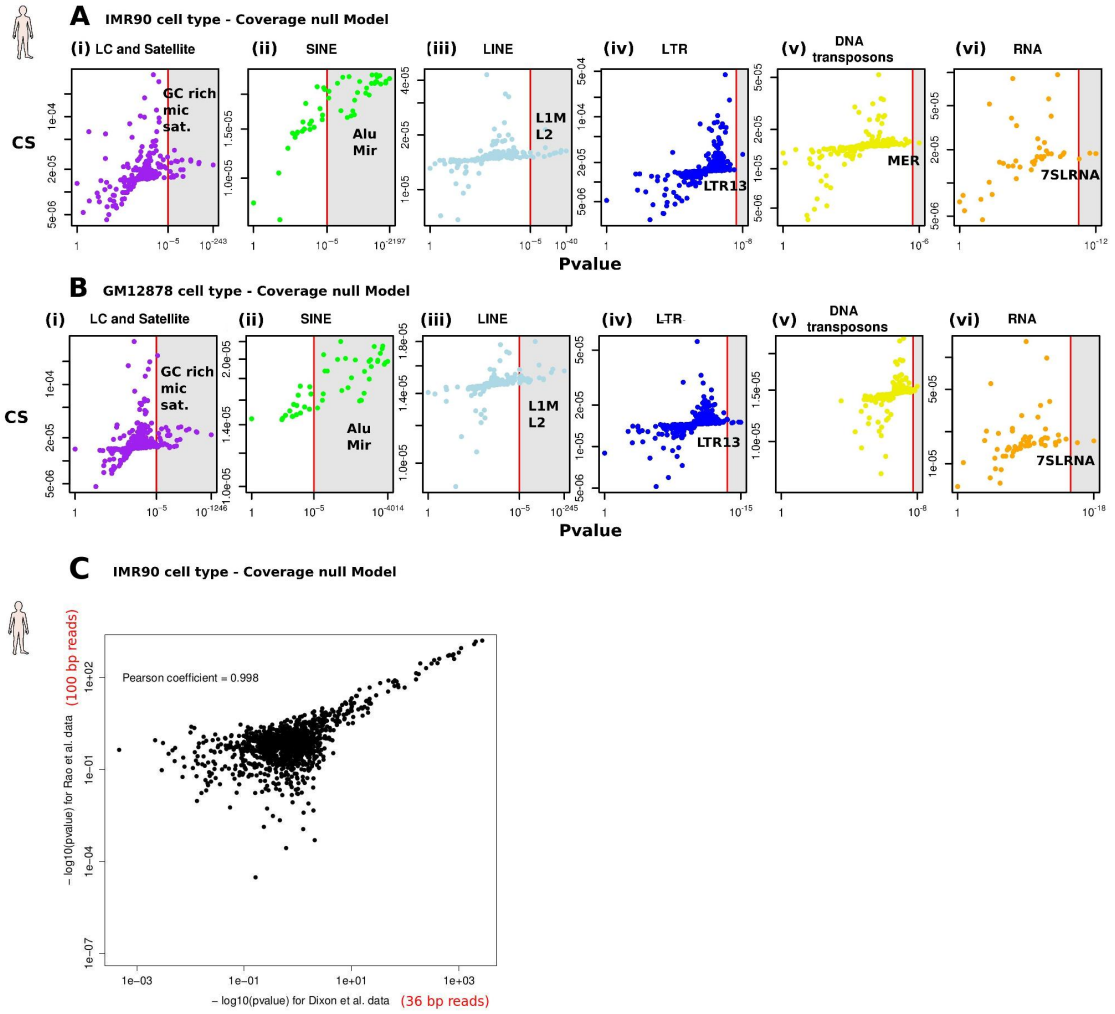
The CS is the average of Hi-C interactions between bins enriched with elements of the same repeat. The p-values correspond to the constant coverage null-model. The type II alignment is very stringent and keeps only reads that do not overlap any repetitive elements (referenced in repeat masker track of UCSC). This procedure only keeps around 23% of the initial reads for the human genome datasets.



Supplementary Figure S6. Co-localization Scores and p-values of the different repeat elements in hESC with bins of equal size

The CS is the average of Hi-C interactions between bins enriched with elements of the same repeat. The p-values correspond to a constant coverage null-model.

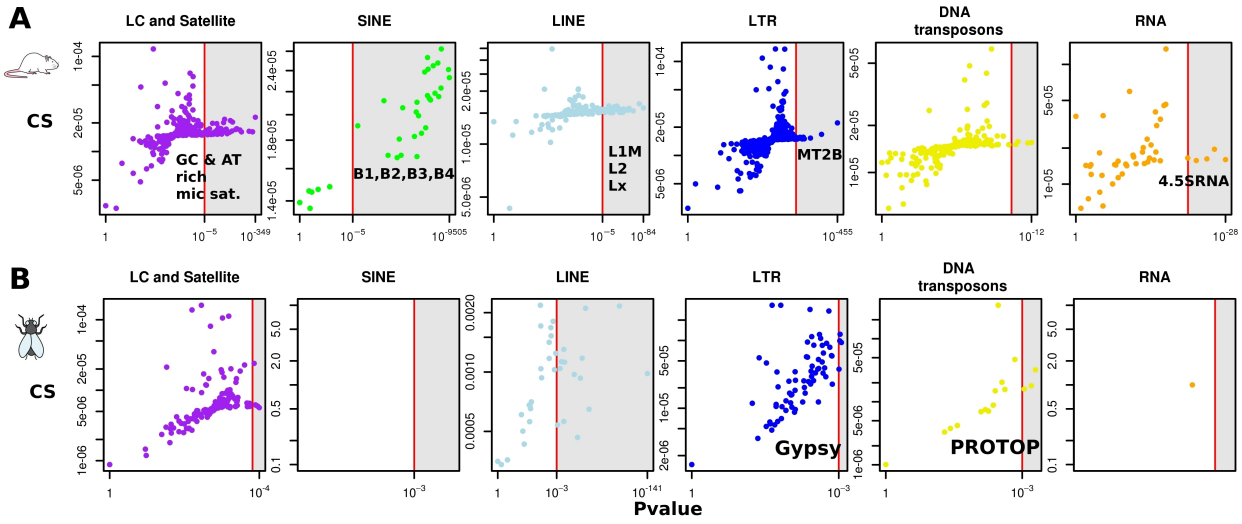
Here we used bins of equal size i.e 30 kb for the construction of the normalized matrices and the computation of CS.



Supplementary Figure S7. Co-localization Scores and p-values of the different repeat elements for two different human cell types: IMR90 and GM12878

(A) IMR90 cells (data from Dixon et al. Nature 2012 [7]).
(B) GM12878 cells (data from Khalor et al. Nature 2012 [8]).
The CS is the average of Hi-C interactions between bins enriched with elements of the same repeat. The p-value corresponds to the coverage null-model

(C) Comparison of p-values obtained with 2 different data sets on the IMR90 cell type : Dixon et al. [7] that contain reads of 36 bp and Rao et al. [10] that contain reads of 100bp. Both datasets give very similar results.



Supplementary Figure S8. Co-localization Scores and p-values of the different repeat elements for mESC and Drosophila with the *type II alignment*

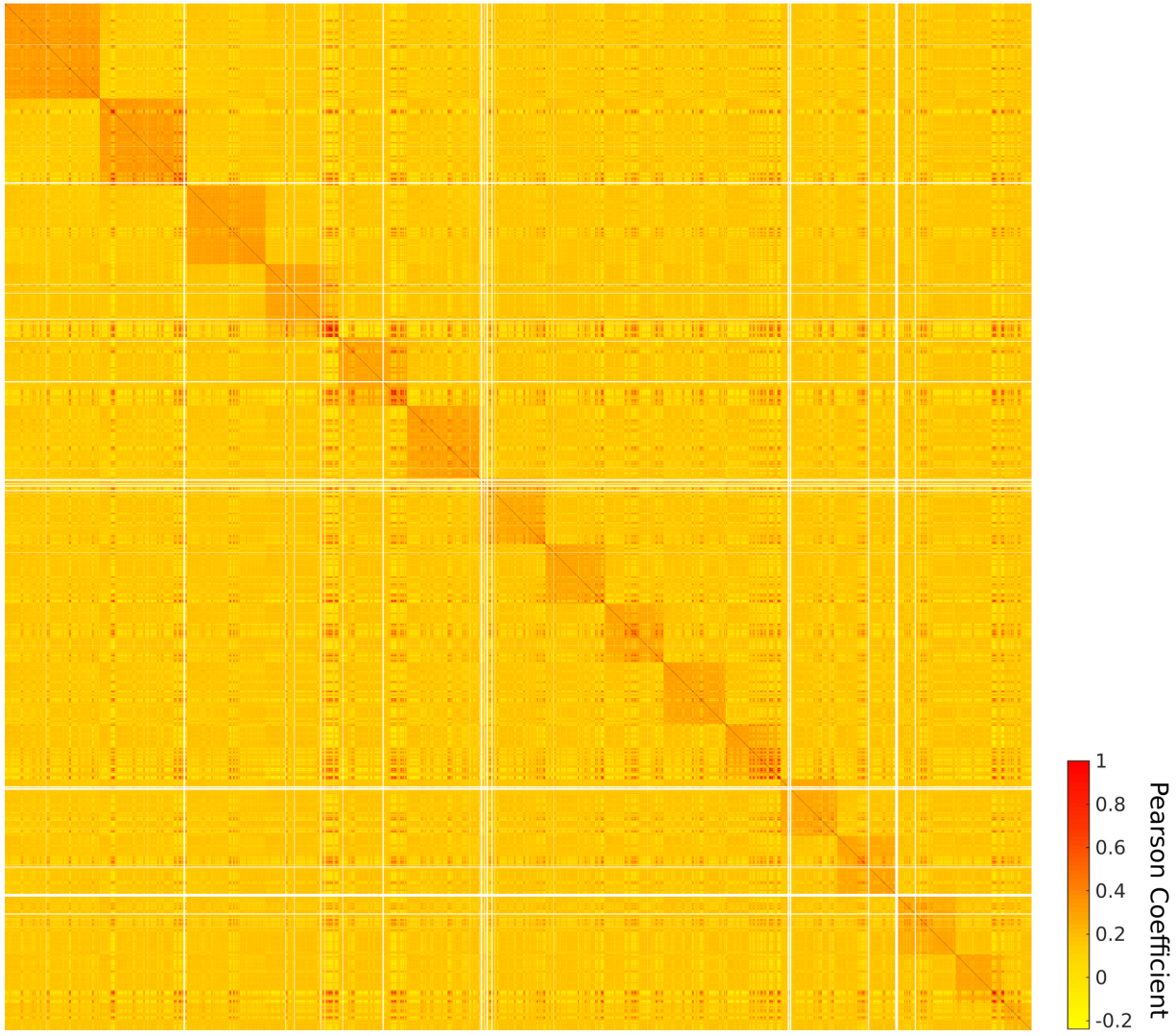
(A) mouse Embryonic Stem Cells (data from Dixon et al. Nature 2012 [7])

(B) Drosophila kc167 cells (data from Hou et al. Cell 2012 [9])

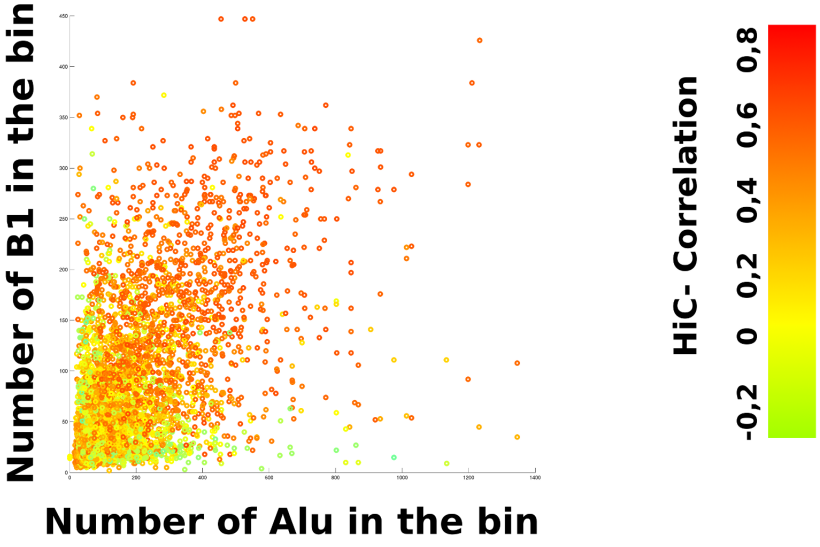
The CS is the average of Hi-C interactions between bins enriched with elements of the same repeat. The used bin is 10 HindIII restriction fragments for the mouse set and 1 HindIII restriction fragment for the drosophila data set.

The p-value corresponds to a coverage null-model.
The most significant families of each class of repeat elements have been annotated.

A


1
0.8
0.6
0.4
0.2
0
-0.2
Pearson Coefficient

B


Number of B1 in the bin
Number of Alu in the bin
HiC- Correlation
-0,2
0
0,2
0,4
0,6
0,8

Supplementary Figure S9. Supplemental information of Supplementary Figure 5 in the main text

(A) Correlation matrix of the contact map obtained for mESC cells. (B) Scatter plot representing the number of Alu in human (300 kb bins) versus the number of B1 SINE in mouse syntenic bins of the genome. The color of each point in the scatter plot corresponds to the correlation between human and mouse contact maps of Fig 5. The scale is the same as the one provided on the colors stripe in between the two contacts maps.

References

1. Witten,D.M. and Noble,W.S. (2012) On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucl. Acids Res.*, **40**, 3849–3855.

2. Paulsen J, Lien TG, Sandve GK, Holden L, Borgan O, Glad IK, Hovig E. (2013) Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucl. Acids Res.,* **41**, 5164-5174.

3. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Meth*, **9**, 999–1003.

4.Kurukuti S1, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, Lobanenkov V, Reik W, Ohlsson R. (2014) CTCF binding at the *H19* imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to *Igf2. Proc Natl Acad Sci U S A. 2006 Jul 11;* ***103****(28): 10684–10689.*

5. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.

Nat Genet. 2006 Nov;**38**(11):1341-7.

6. Jian Qun Ling, Tao Li, Ji Fan Hu, Thanh H. Vu, Hui Ling Chen, Xin Wen Qiu, Athena M. Cherry2, (2006) CTCF Mediates Interchromosomal Colocalization Between Igf2/H19 and Wsb1/Nf1 *Science* 14 April 2006: Vol. 312 no. **5771** pp. 269-272

7. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

8. Kalhor,R., Tjong,H., Jayathilaka,N., Alber,F. and Chen,L. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotech*, **30**, 90–98.

9. Hou,C., Li,L., Qin,Z. and Corces,V. (2012) Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell*, **48**, 471–484.

10. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I3, Omer AD, Lander ES6, Aiden EL. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping (2014) Cell. 2014 Dec 18;**159**(7):1665-80.

Code availability

All the codes used for the analysis (in C and R) and instructions to reproduce the main figures are available here:

https://github.com/axelcournac/Repeats_elements