ATAC-seq Pauline

Materials and Methods

Guidelines

We largely inspired by different guidelines papers[a,b].

Bioinformatic pretreatment

We used nfcore/atacseq[1] pipeline for trimming[2], alignment[3], peak calling[4], annotate peak[5], count reads[6] in consensus peaks and all for each step quality-control[7]. Pipeline generated data to make a differential accessibility analysis, PCA and clustering[8].

Bioinformatic analysis

Chromatine occupancy step

Tn5 tranposase used in ATAC-seq protocol working near to entry-exit sites of the nucleosomal DNA[c]. Objective is sequencing linker-DNA as represent of chromatine accessibility where the length is variable on genome and form is limited into 4 shapes (Fig.1). As chromatine is a suite of nucleosomes with a variable distance between each element, we separated reads mapped into two categories : inferred nucleosome-free fragments (NFR) and nucleosome-bound region (NBR). The first category contains mostly linker-DNA and short genomic regions. The second category contains differents larges genomics regions.
Linker-DNA length ranges between ~20–100 bp and varies among different species, tissues, and even fluctuates within a single cellular genome[d]. We used Fourier transformation analysis of Distribution from ATACgraph to identify the length for the first NBR (NBR+1 = mono-nucleosome) (Fig.2b). We indicates categories on fragment length distribution :

- NFR : nucleosomal free-region (23 to 100bp)
- inter 1 : region between NFR and NBR+1 (101 to 169 bp)
- NBR+1 : region flanked by mono-nucleosome (170 to 220 bp)
- inter 2 : region between NBR+1 and NBR+2 (221 to 339 bp)
- NBR+2 : region flanked by di-nucleosome (340 to 400 bp)

These informations are just descriptive. And there are not usage for the rest of analysis because I don’t have time now for integrate parameters and adapt

scripts to a custom analysis for each category. But in much publications, it's used to find enhancer for example.

Peak calling analysis

We focused on peak calling for NFR and inter 1 categories. First, we plotted a global view genome with a log10 peak calling score as scale. To facilitate reading, we filtered any peak inferior to log10(peak score) ≤ 3.1. (Fig.1d).
We created an union Granges object from broadpeak Setdb1$^{+/+}$ and Setdb$^{\Delta/\Delta}$. It contains for each sample,peak calling score, coverage on region overlapped, mean coverage and mean peak calling score for all sample.
From data, we plotted the distribution of ATAC-seq peak by sample (Fig. 1e).
We used merged peaks object to annotate peaks with *annotatePeak* from ChIPseeker R package on TxDb.Mmusculus.UCSC.mm10.knownGene references.
We calculated distances between each center of regions with nearest transposables elements positions from mm10 repeatmasker database, nearest CGI positions from TxDb.Mmusculus.UCSC.mm10.knownGene references. with *annotatr* R package and nearest TSSs positions with the same reference. We filtering the minimal distance between one transposable element for one subfamily to one peak.
We focused interest on CGI, TSS, DNA, ERV1, ERVK, ERVL (3 majors LTR subfamilies), LINE and SINE positions.
We plotted relative abundance for each sample to each interested positions for all peaks annotations based to biotype genomic content (Distal intergenic, Promoter ≤ 1kb to 3 kb, separetly into 3 subcategories), UTR regions (5' and 3'), exons, intron and dowstream region (≤ 300 bp). (Fig.2a).
We selected Promoter and Distal intergenic regions to evaluate conditions impacts (specific peaks for one condition) on transposable element. We compared peak width which contains transposable elements between conditions with a t-test (Fig.2b)

We plotted enrichment peaks over TSSs and CGI. We calculated normalized enrichment matrix from NFR and inter-1 groups over TSSs position and CGI regions positions with *normalizeToMatrix* from EnrichedHeatmap. To perform plot, we extracted data and calculated 2-means to merge enrichment profiles between TSSs and CGI signals. To allow a better comparison between conditions, we calculated a difference enrichment between conditions (Fig. 2c)

If I have time, I re-run Fig. 2c with correct name sample.s And I plotted also enchancer enrichment.
For 2 reasons : first, I used only peaks on NFR and inter 1 groups to maximise informations but In many papers, it's interesting to make on it for all groups separetly first or in combinaison, in second time. Second, these enrichment plot doesn't associated with annotation peaks. So, It's very interesting to merge informations to explain the signal enrichment, mostly in k-means groups.

Differential accessibility analysis

To evaluate Tn5 accessibility on samples, we performed differential analysis on count reads over open regions from peak calling. We build an union table between replicates and conditions. Then, we realised quality-control of samples and differential analysis with *DESeq2* (Fig. 3a). We associated deregulated open region information with annotation. (Fig. 3b).

Footprinting over Transposable element

To evaluate if open region contains a TE sequence, we performed enrichment peaks over differents families and subfamilies TE. We calculated coverage each normalized peak over TE. We plotted best candidates with the best coverage of all normalized enrichment peaks and we separated normalized enrichment peaks by K-means (Figure 4a, 4b and 4c).

Figure 1a : Different possibilities of linker-DNA. L is length (angström), β is torsion angle.
source : https://www.researchgate.net/figure/Linker-DNa-length-variations-change-spacing-and-orientation-between-nucleosomes_fig1_230879604
adapted to 10.4161/nucl.22168

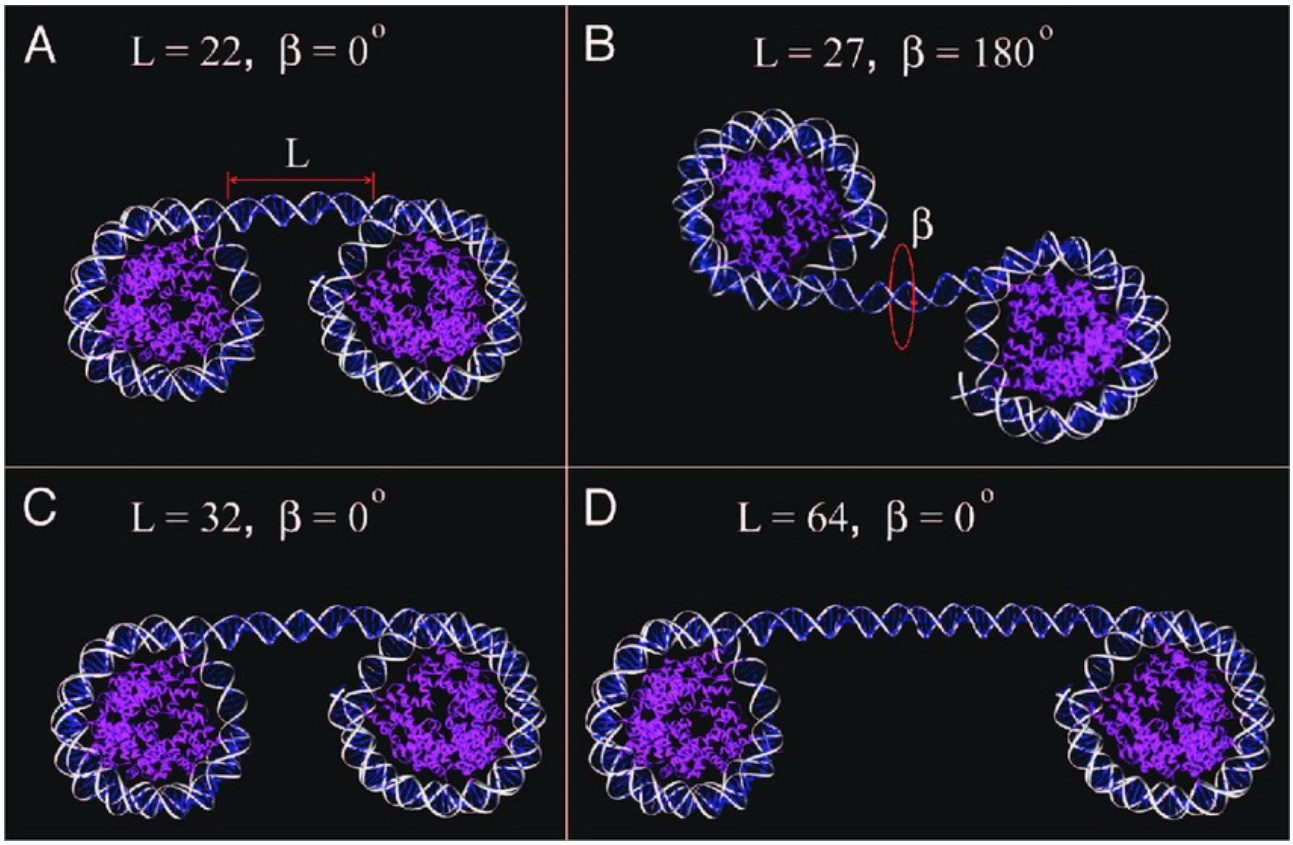


Figure 1b : Fourier Transformation Analysis of distribution

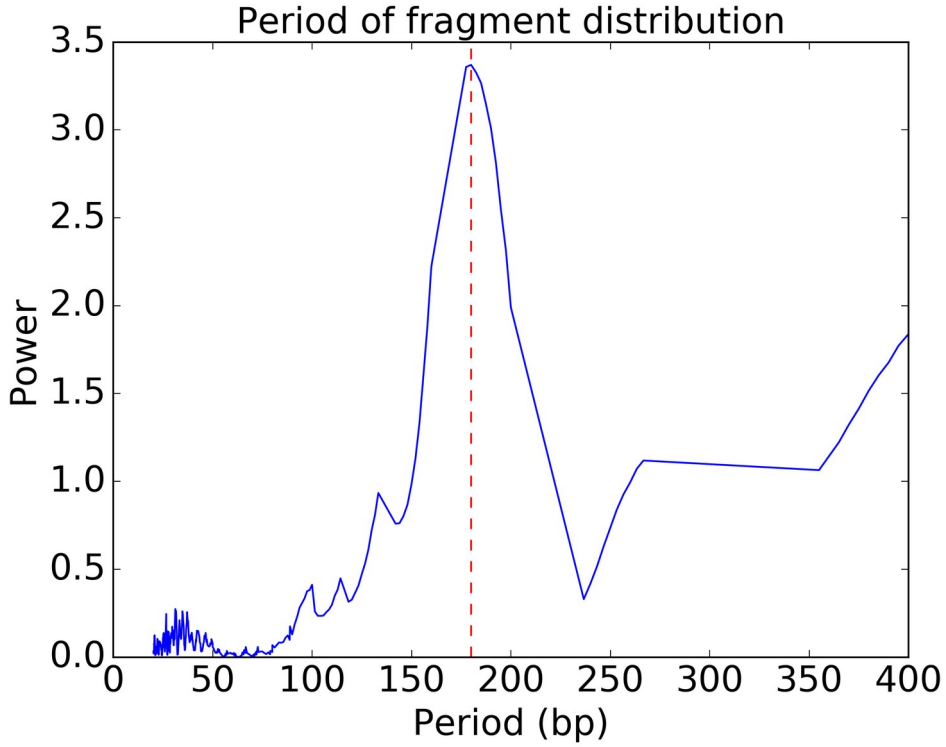


Figure 1c : Insert size distribution.

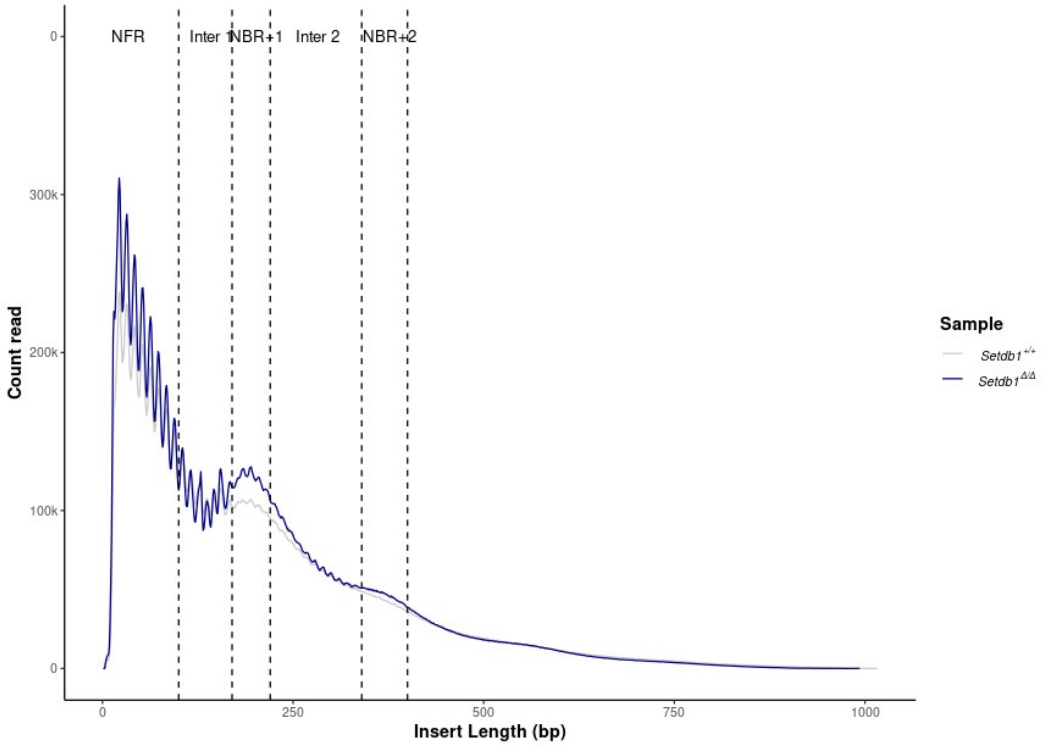


Figure 1d : Genomic view of high peak calling score.

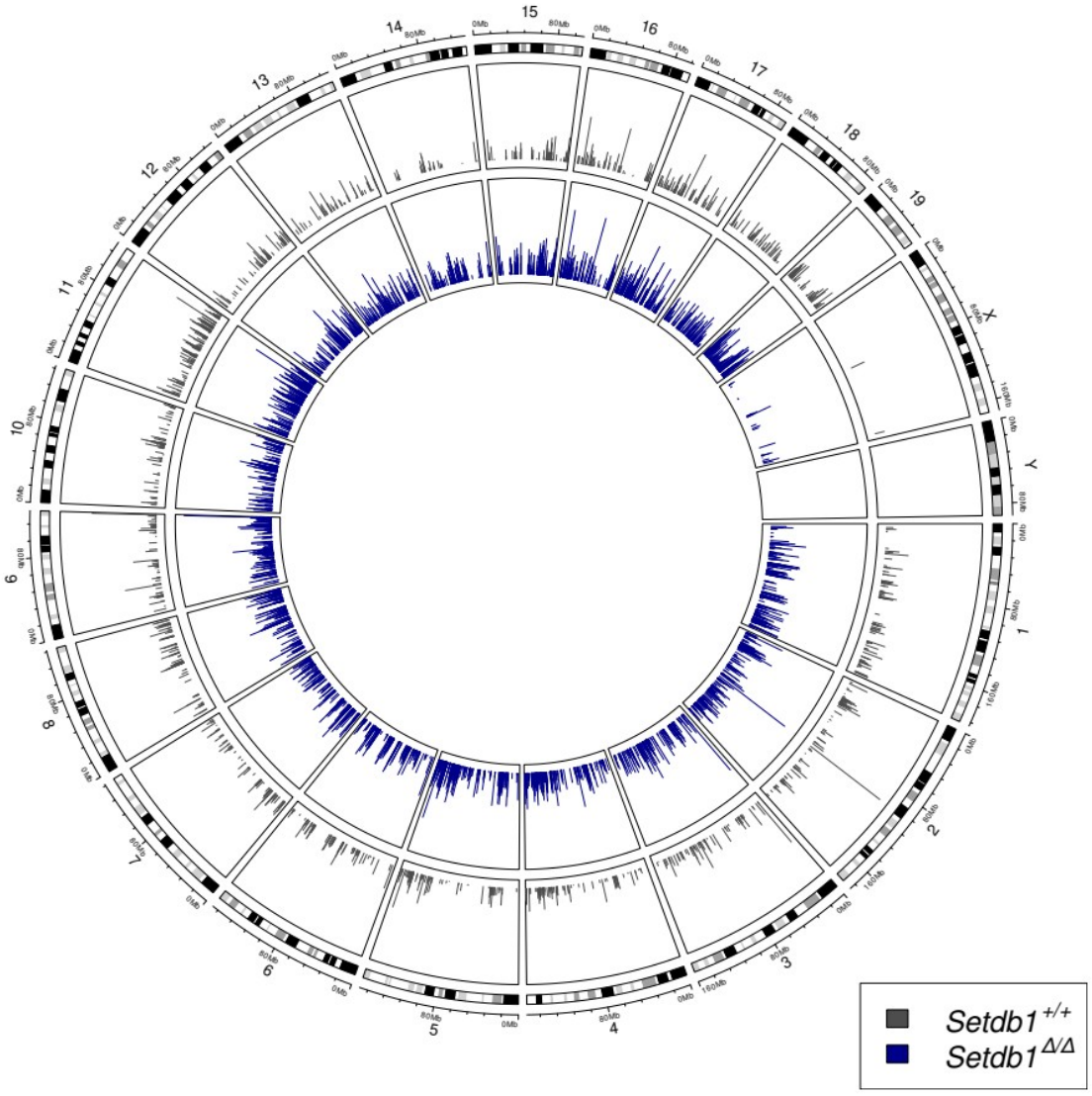


Figure 1e : Distribution of ATAC-seq peak by sample. We counted three situations : there are ATAC-seq peaks in same regions in all samples (Setdb1$^{+/+}$ (grey) and Setdb$^{\Delta/\Delta}$ (blue and blue stripe), only in Setdb1$^{+/+}$ samples and only in Setdb$^{\Delta/\Delta}$. ~76 % of ATAC-seq peaks shared in same regions between all samples. ~ 6 % of ATAC-seq peaks are only present in Setdb1$^{+/+}$. ~ 18 % of ATAC-seq peaks are only present in Setdb$^{\Delta/\Delta}$.

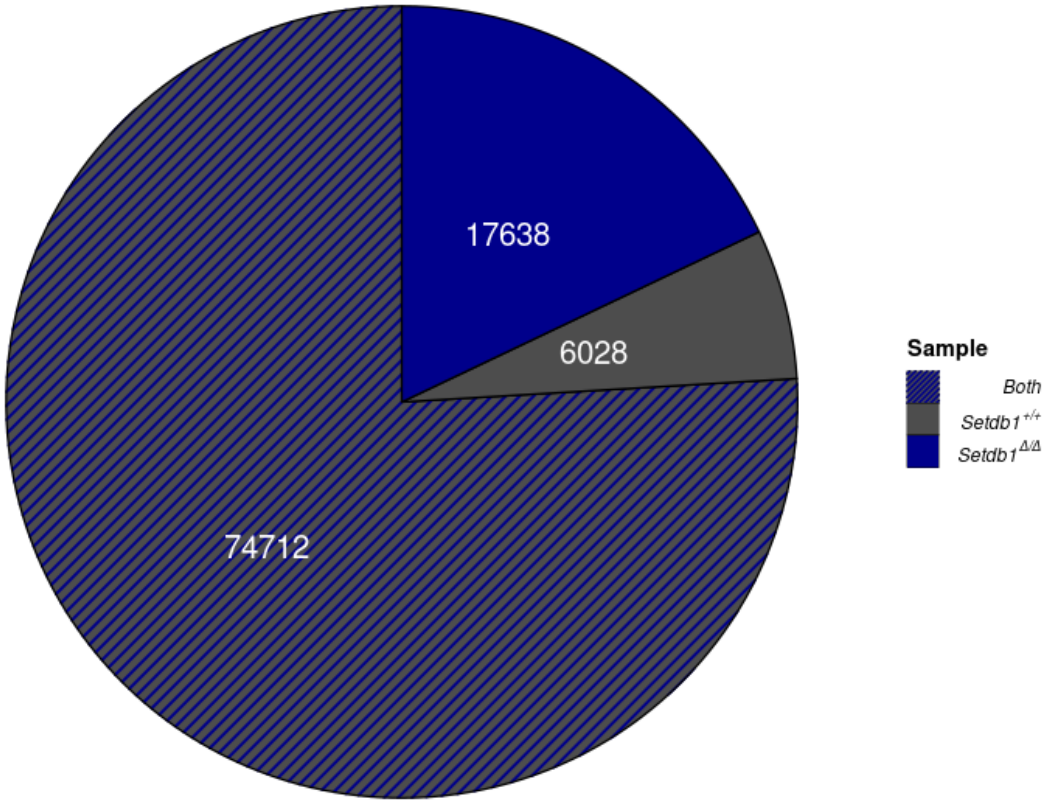


Figure 2a : Relative abundance for each sample to each interested positions for all peaks annotations based to biotype genomic content

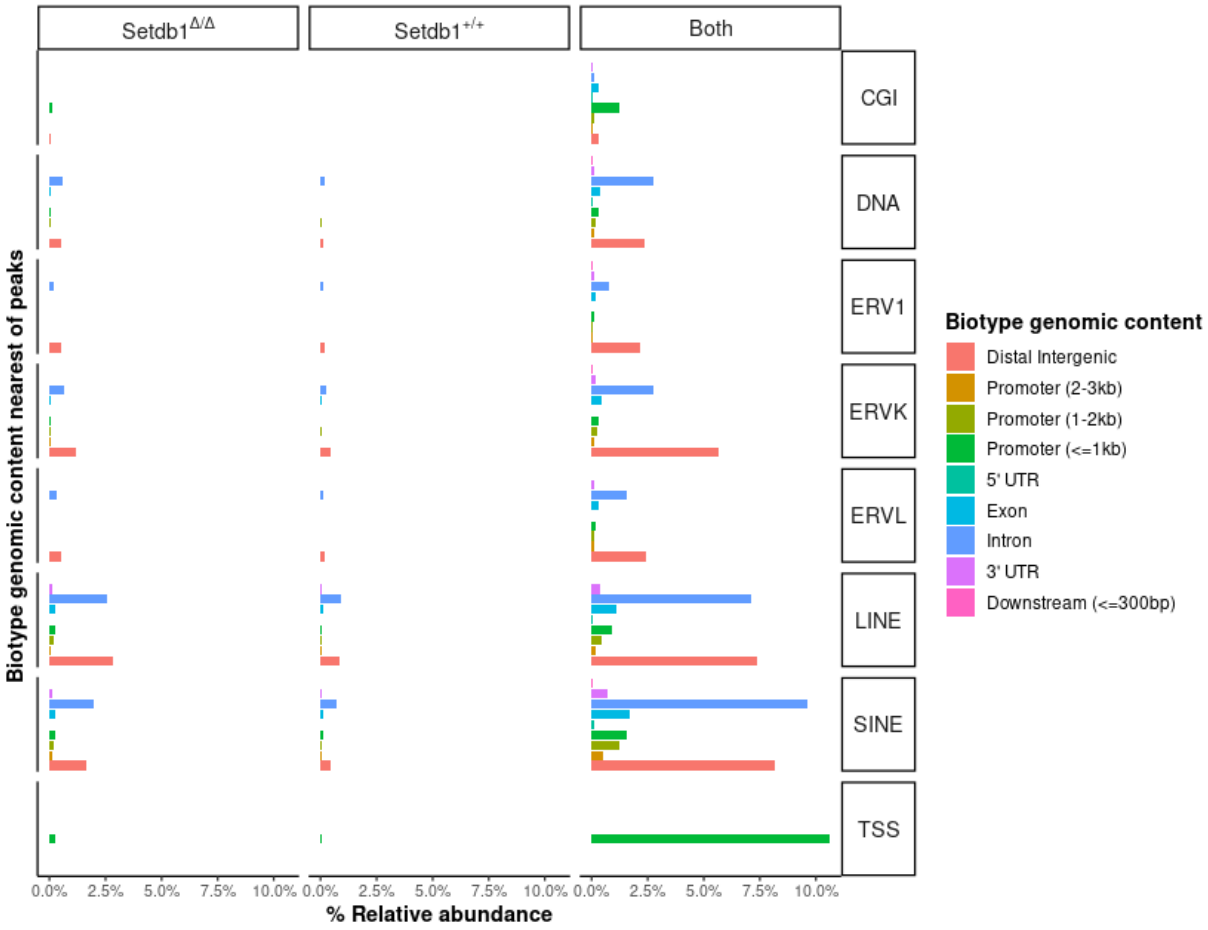


Figure 2b : Width comparisons containing nearest TE between specific regions annotated Setdb1$^{+/+}$ and Setdb$^{\Delta/\Delta}$. We selected only Promoters and Distal intergenic regions and we plotted widths for each peak by nearest TE. Then, we did a t-test between conditions.

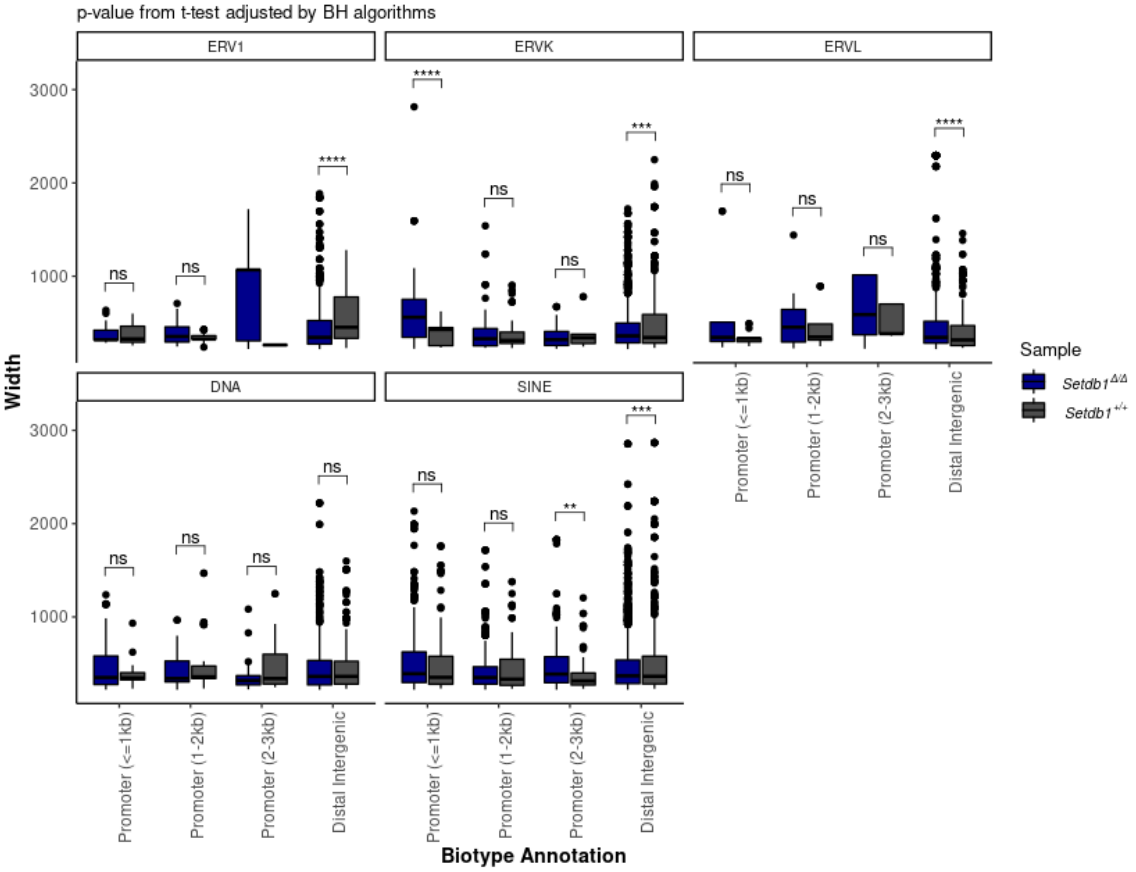


Figure 2c : Enrichment of TSS and CpG islands (CGI). Clusters 1 and 2 contains high value upstream TSS and around CGI regions. Clusters 3 and 4 contains low value upstream TSS and around CGI regions. TSS and CGI differences (third and fifth heatmap) are difference between Setdb1$^{+/+}$ and Setdb1$^{\Delta/\Delta}$ conditions.

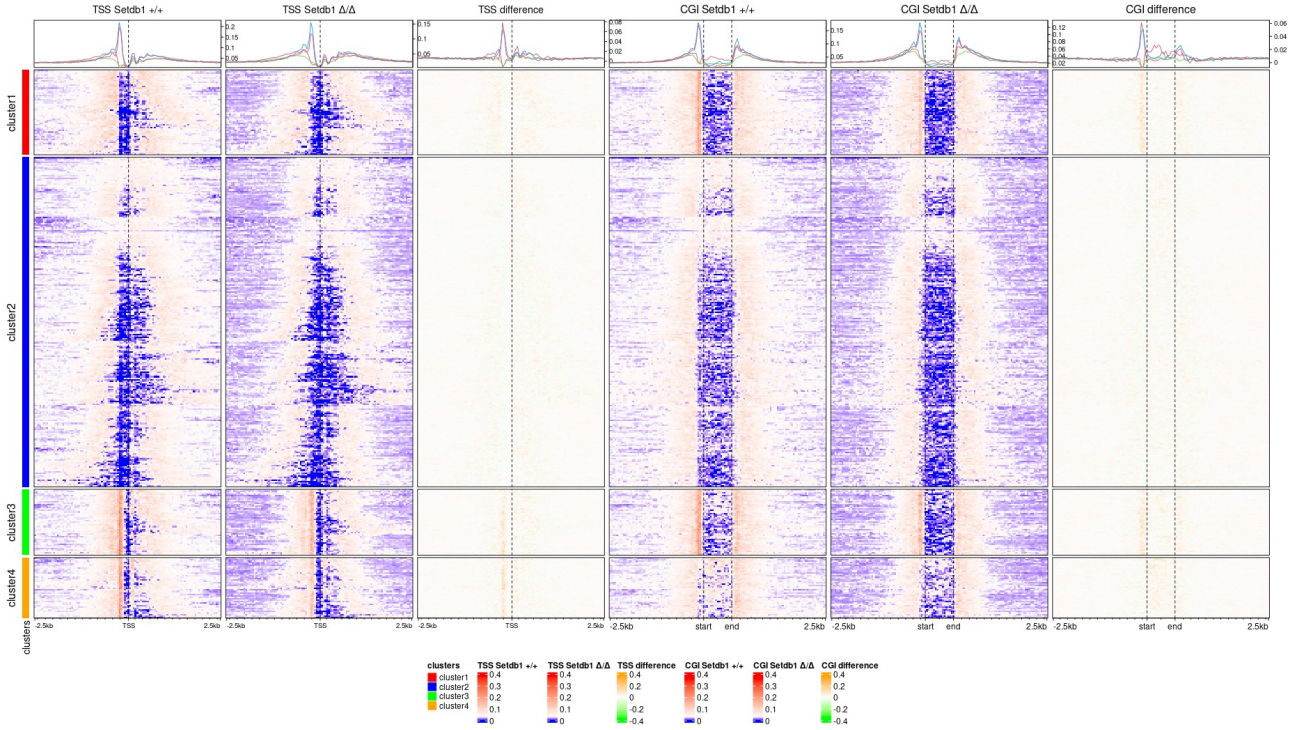


Figure 3 : Differential accessibility analysis. A : unsupervised clustering on replicates and conditions. B : PCA and UMAP representation of replicates and samples. C : Volcanoplot Setdb1$^{\Delta/\Delta}$ vs Setdb1$^{+/+}$. D : MAplot Setdb1$^{\Delta/\Delta}$ vs Setdb1$^{+/+}$. E : Distribution of deregulated open regions.

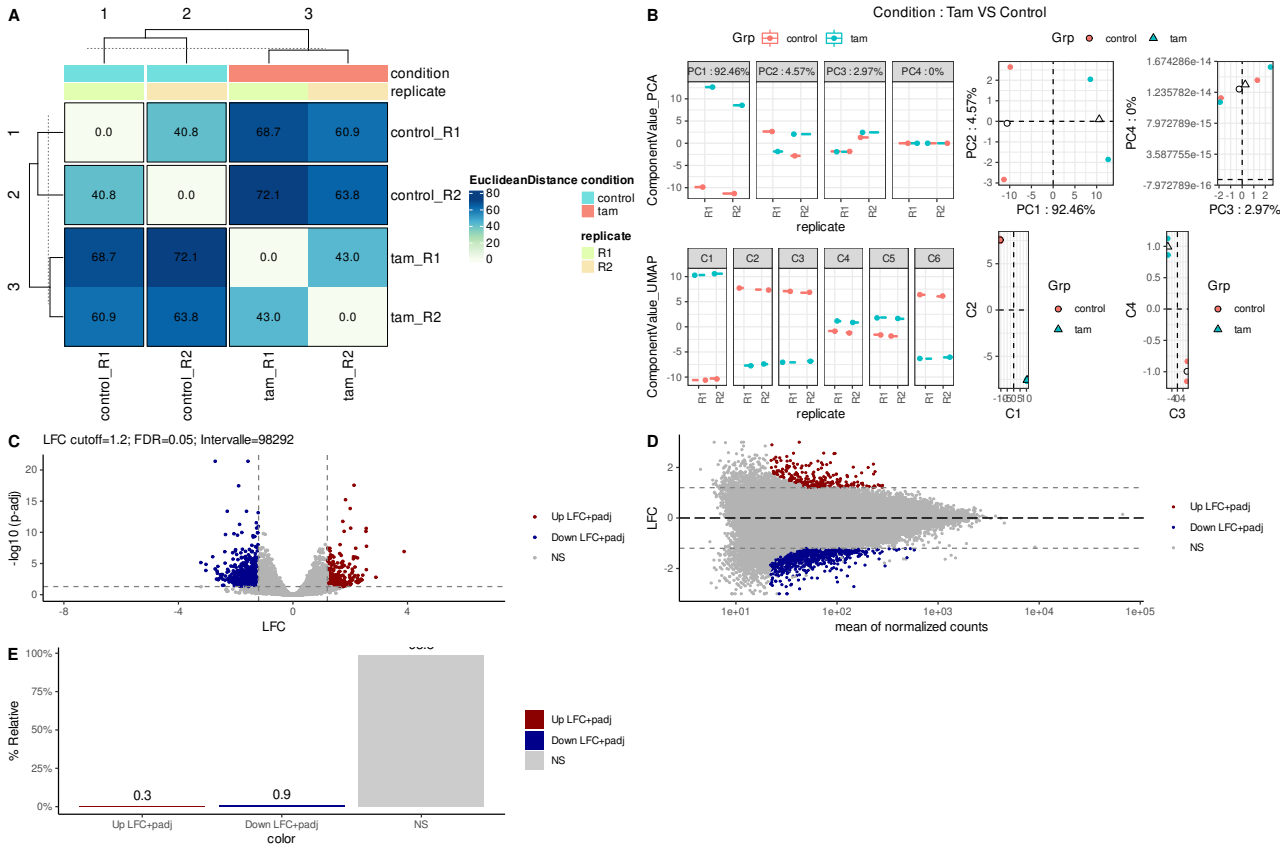


Figure 3b : Z-score normalized count on open region Setdb1$^{\Delta/\Delta}$ vs Setdb1$^{+/+}$ associated with deregulated information (Log2FC), if open-region is shared by all samples (Both) or only one (Setdb1$^{\Delta/\Delta}$, Setdb1$^{+/+}$), nearest TE of open-region and the distance between center of open-region and nearest TSS.

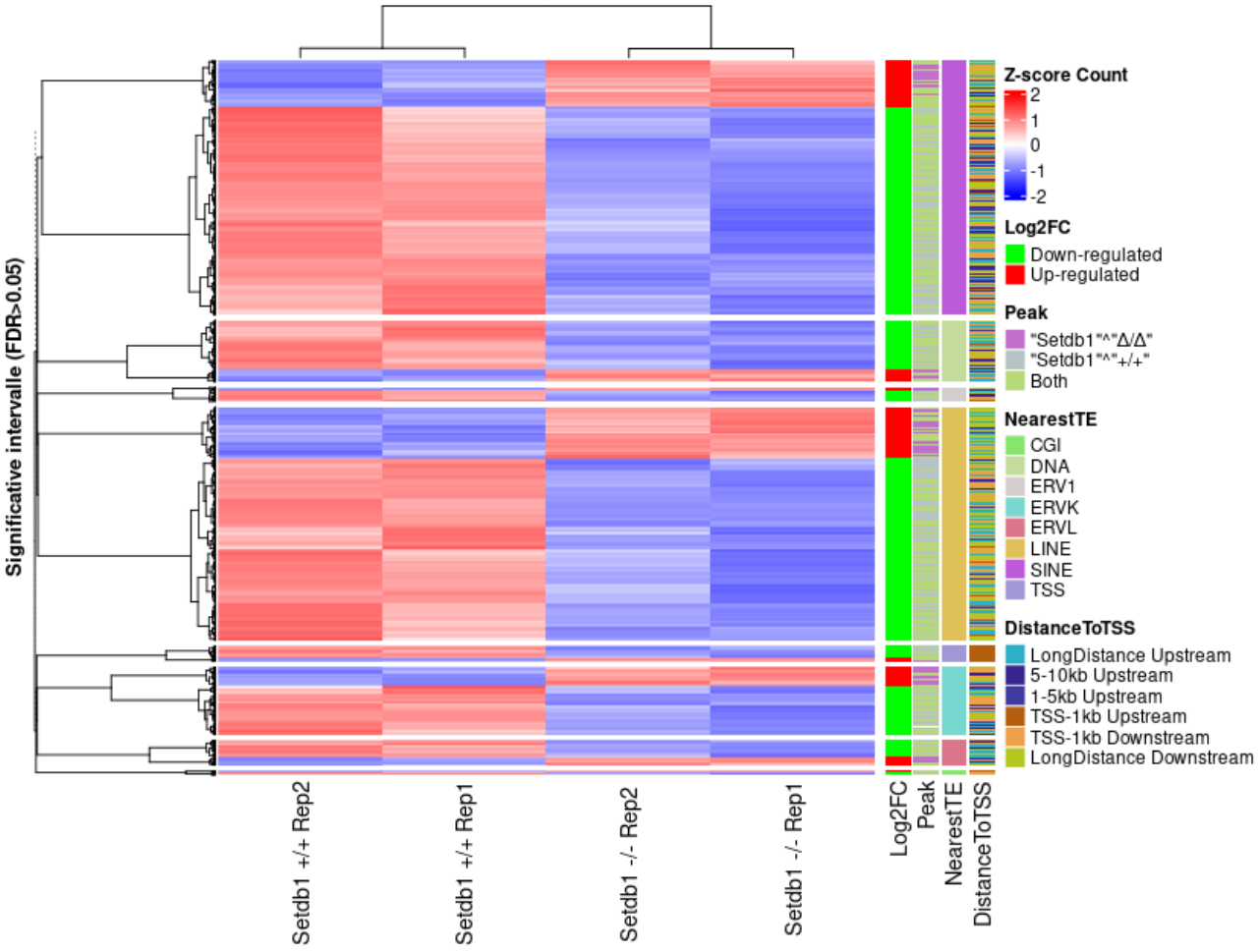
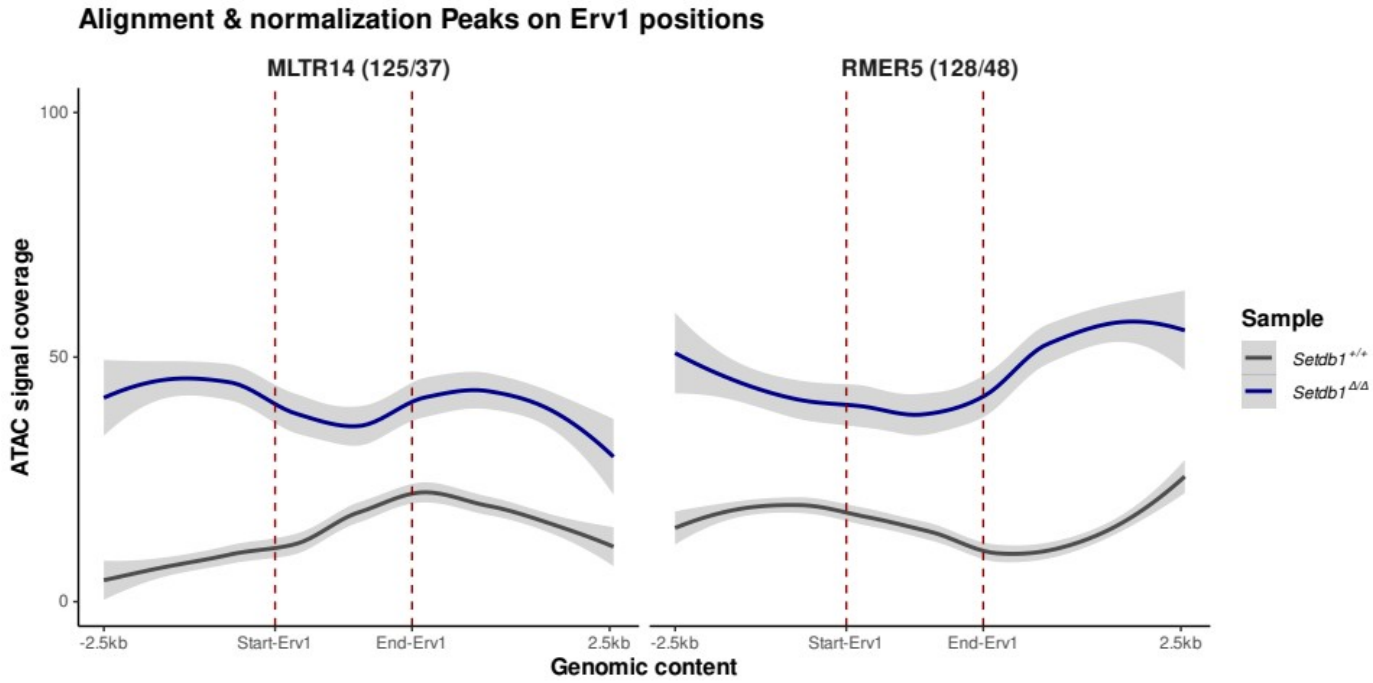


Figure 4a : Enrichment of peaks over ERV1. Top : best ERV1 class with best coverage. Bottom : Clustering of peak enrichment over ERV1 by K-means.



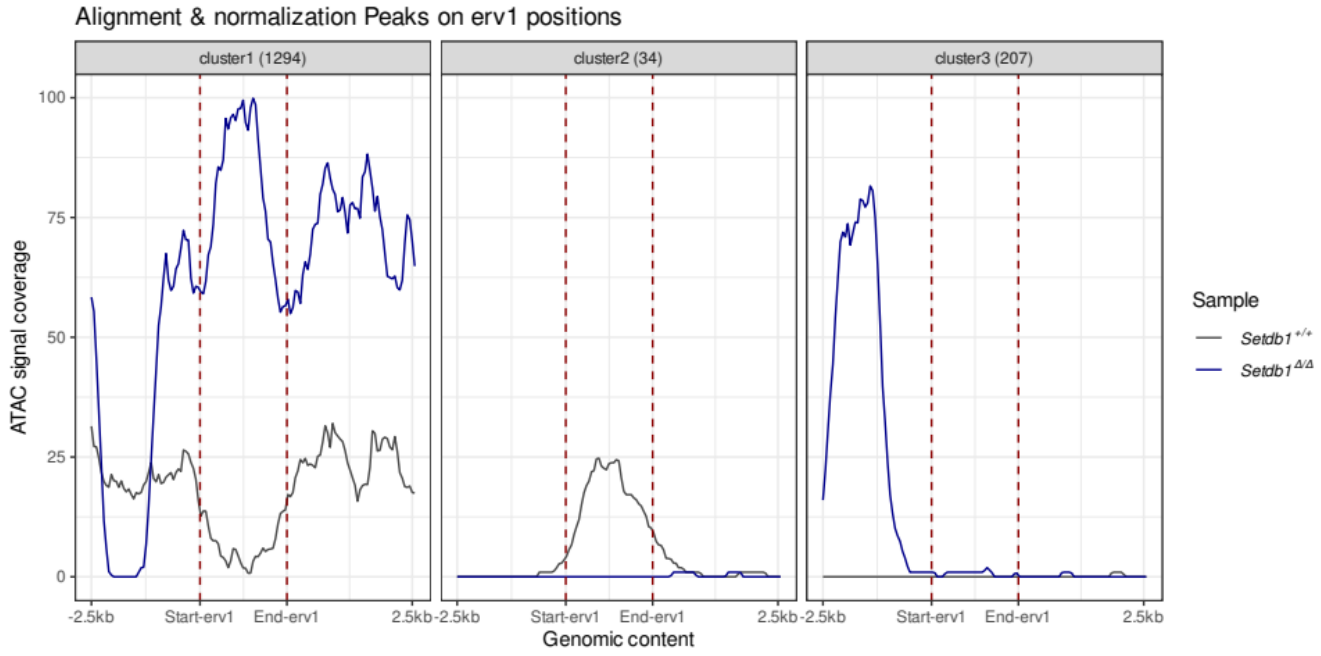
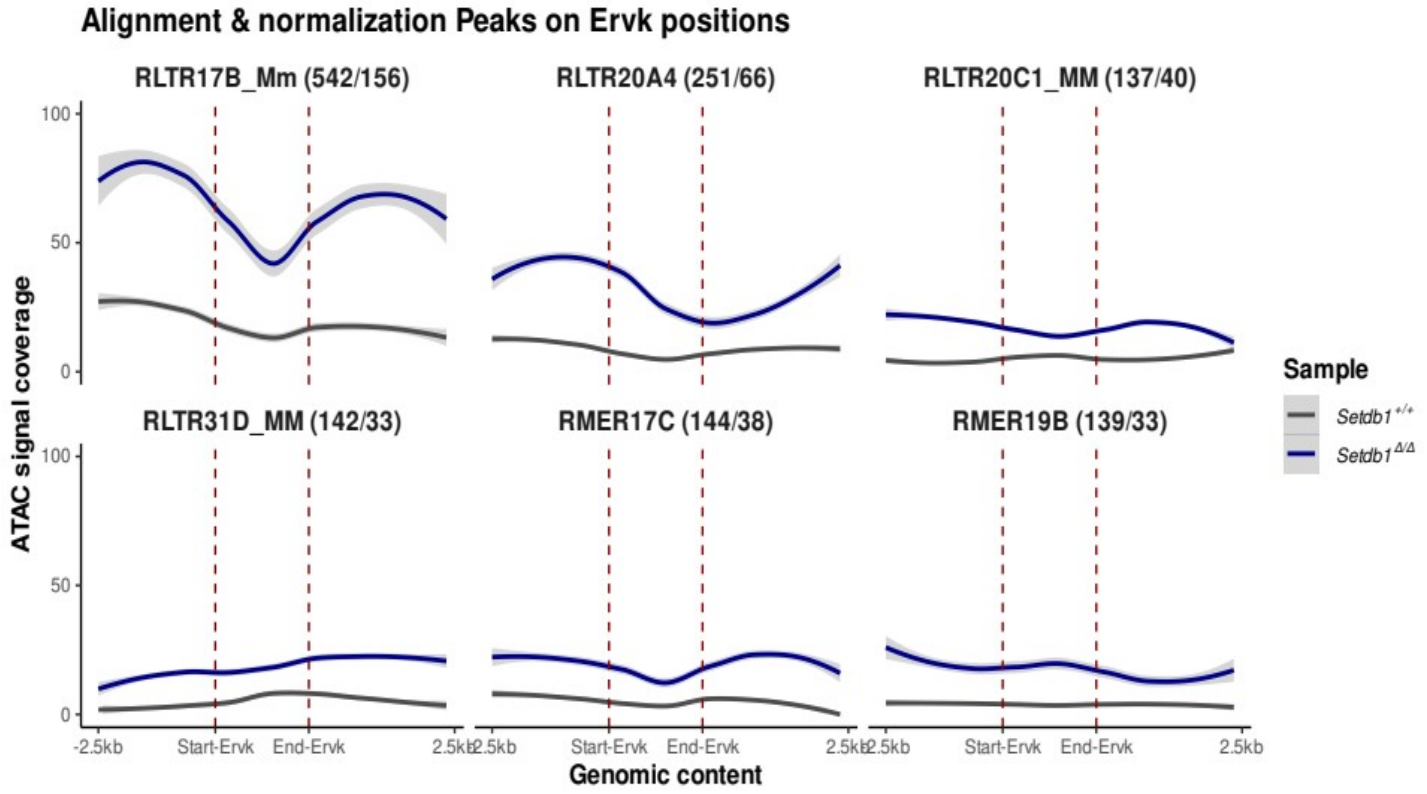


Figure 4b : Enrichment of peaks over ERV1. Top : best ERVK class with best coverage. Bottom : Clustering of peak enrichment over ERVK by K-means.



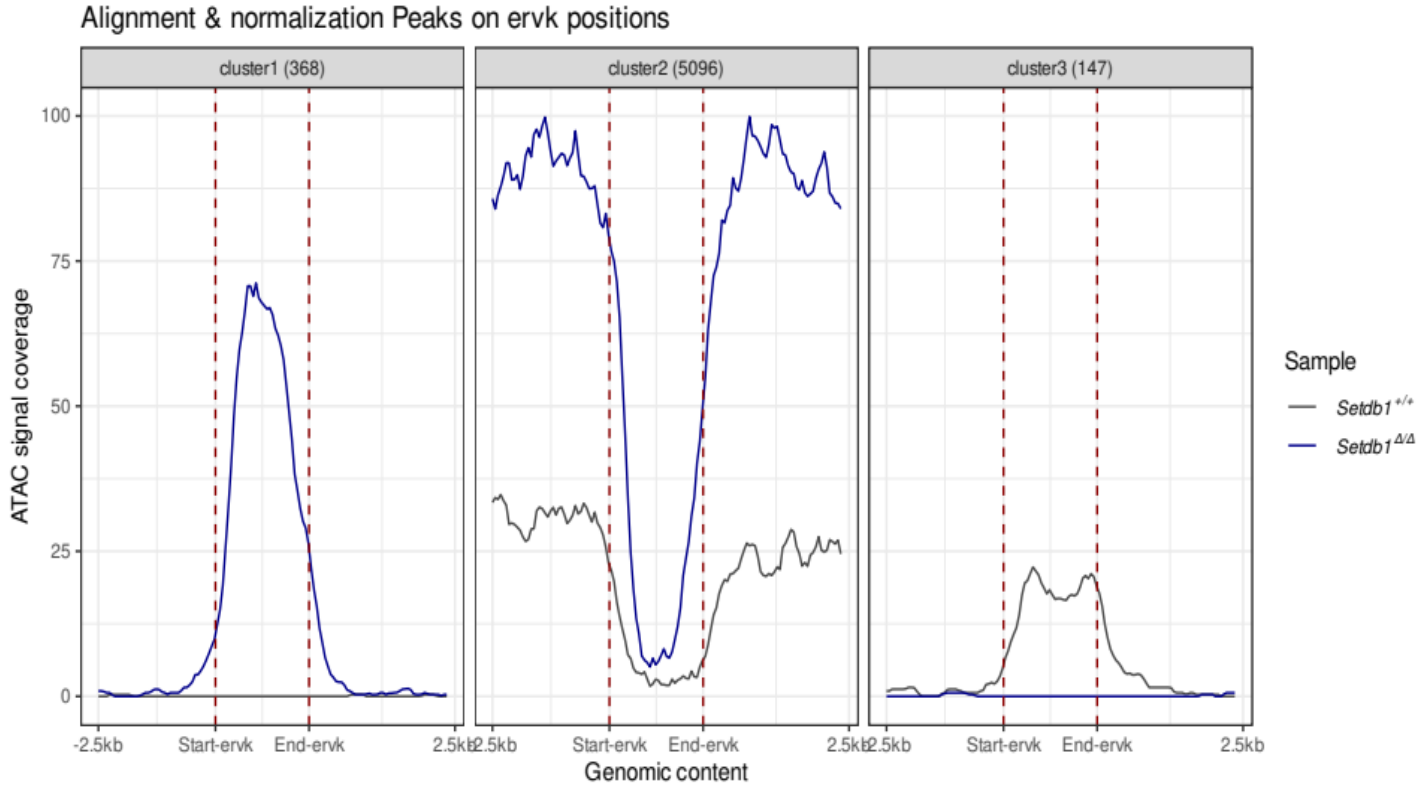
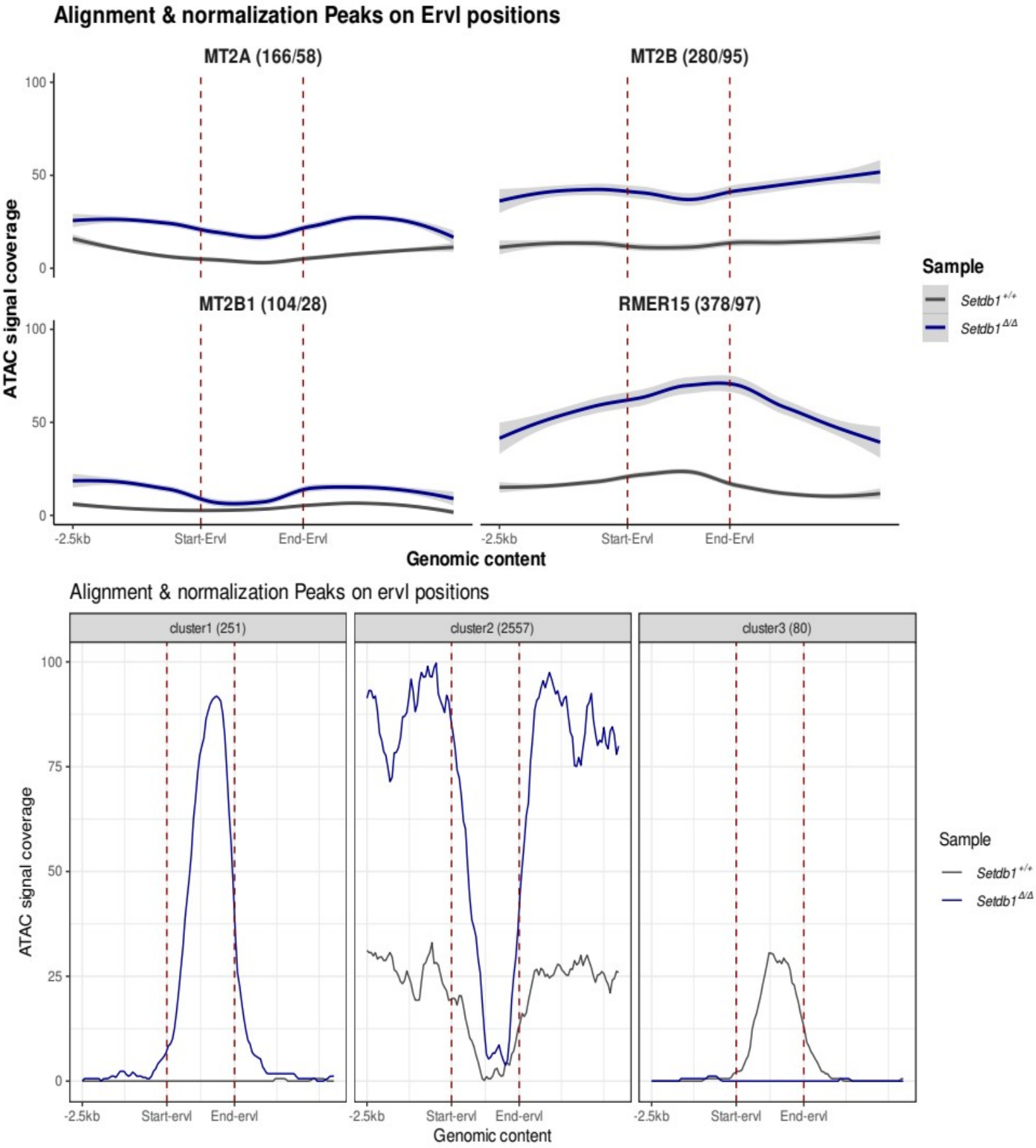


Figure 4c : Enrichment of peaks over ERV1. Top : best ERVL class with best coverage. Bottom : Clustering of peak enrichment over ERVL by K-means.



a: https://www.nature.com/articles/s41467-021-21583-9
b: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1929-3
c: https://royalsocietypublishing.org/doi/10.1098/rsob.190116
d: 10.1139/O10-139

[1]: https://nf-co.re/atacseq. DOI : 10.1038/s41587-020-0439-x
[2]: Trim Galore ! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
[3]: BWA http://bio-bwa.sourceforge.net/ DOI : 10.1093/bioinformatics/btp324
[4]:

- picard
- SAMtools, BEDTools, BAMTools
- Pysam
- Preseq
- bedGraphToBigWig
- MACS2
- deepTools

[5]: HOMER
[6]: featureCounts
[7]:ataqv,MultiQC
[8]: R, DESeq2