SUPPLEMENTAL FIGURES

Figure S1. Controls and biological replicates of nascent strands isolated from mouse ES cells

A. Schematic procedure of Nascent Strand (NS) isolation for mapping active origins. 0.5-2.5 Kb NS were isolated from genomic DNA by denaturation and sucrose gradient centrifugation, followed by λ-exonuclease (λ-Exo) digestion to remove contaminant DNA, before high-throughput sequencing.

B. We took particular care to define the most suitable background controls for our peak-calling analyses, because the choice of background crucially affects the peak positions. One control was sheared and heat-denatured genomic DNA (see, panel C). A second control was performed to exclude any signal due to a possible bias in the λ-Exo treatment by digesting with RNAse A the RNA primer in the 5' of the NS that protects the NS from λ-Exo digestion (RNAse A input). Specifically, DNA fractions containing NS isolated from exponentially growing cells were divided in two parts and one part was incubated with 50 μg/ml RNAse A at 37°C for 30 min after the sucrose gradient fractionation and before the purification steps. Then, NS purification was as described in panel A.

C. Alignment of three independent biological replicates and the combination result on mouse chromosome 11. The two controls used in this study are aligned below the replicates.

D. The NS-enriched zones (IZ - light green blocks) define origins; each IZ can contain one or more major peaks (dark green blocks) that identify initiation sites (IS) of replication. Only overlapping peaks and zones were retained for future analysis.

E. Comparison with array data of chromosome11 analyzed in Cayrou et al. 2011. Alignment of the NS, negative control raw profile from sequencing data (green) and microarray data (black) on parts of Chromosome 11. After computing overlap on all origins previously identified in Cayrou et al 2011, we found that 80.2% of IZ are covered by at least one microarray-origin on 84% of their length in mean.

Figure S2. Optimal λ-Exo digestion conditions are essential for proper digestion of the DNA template

In the first experiment, a plasmid containing the human c-MYC origin of DNA replication (pFRT-myc2.4) was used.

A. The BglII site used to linearize pFRT-myc2.4 is shown. The position and sequence of the G-quadruplexes (G4) found in the c-MYC origin are shown, as well as the size of the two expected fragments (G4F and G4R) if λ-Exo digestion was hindered by the G4.

B. Samples containing 700 ng of purified linearized plasmid were denatured (95°C for 5 min, then 5 min on ice) in standard λ-Exo Buffer (67 mM glycine-KOH (pH 9.4), 2.5 mM MgCl2, 0.01% (v/v) Triton X-100). Samples were incubated with 20 units of λ-Exo (the concentration used in Foulk et al, 2015) for 1h or 16h at 37°C, or with 100 units λ-Exo for 16h at 37°C (1st round of digestion in our protocol that calls for two or three rounds of λ-Exo digestion). Reactions were stopped by addition of loading buffer and samples were resolved on agarose gel and visualized using the Gel-Red staining dye. The size (in kb) of the 1Kb DNA ladder fragments is indicated. Note that limiting conditions lead to incomplete digestion of the substrate. On the other hand, G4-containing fragments resulting from a possible arrest of digestion are barely detectable. Most notable is the presence of high amount of incompletely digested substrate in the lower part of the gel that are still slightly detectable after overnight digestion with 20 units of λ-Exo (this lower part of the gel was not shown in Foulk et al, 2015). These DNA molecules are never observed (even after a single round of digestion) using the digestion conditions described in Cayrou et al. that, in addition, include two (or three) successive rounds of λ-Exo digestion.

C. Degradation of ssDNA containing the c-MYC G4 by λ-Exo. Substrates used in the λ-Exo digestion assay. A 1200 bp DNA molecule containing the G4 and the c-MYC replication origin was 3'labeled in a fill-in reaction using dCTP[a-32P] and the Klenow fragment. The position and sequence of the G4 motif is shown on both strands.

D. The 3'-labeled substrates were denatured and then incubated with increasing concentrations of λ-Exo at 37°C for 1 hour. Note that in our nascent strand purification protocol, two consecutive λ-Exo digestions (each at 37 °C overnight) are performed. The products were resolved on a denaturing polyacrylamide gel and revealed by autoradiography. At all concentrations, no detectable significant digestion intermediate was observed.

Figure S3. Random control of the origin classes and global view of chromosome 11

A. Clustering of origins based on read densities around randomly selected genomic fragments. The left panel displays the heatmap of read densities in 7kb regions on each side of the

random fragment center, by 100bp bins. Twelve groups are defined as in Figure 1. The right panel indicates the mean read density profiles per group.

B. Alignment of the four classes of origins along chromosome 11 above the raw NS signal.

C. Example of raw profile for each class.

Figure S4. Epigenetic marks and chromatin factors at origins

A. Precision-recall plot showing the mutual coverage by marks/factors of IS (left), IZ (right) and random sites. We defined recall (X axis, logarithmic scales) as the fraction of sites covered by marks/factors of a given type, and precision (Y axis, logarithmic scales) as the fraction of marks/factors covered by the sites.

B. Distribution of chromatin marks relative to initiation sites (IS, dark green) and initiation zones (IZ, light green). We considered as IS-specific the marks/factors the occurrence of which overlapped in IS and IZ (left panels). When the occurrence of a mark/factor was more than 2-fold higher in IZ than in IS, we considered that mark as IZ-specific (right panels).

C. Precision-recall plot showing the mutual coverage of marks/factors between IS and IZ. We defined recall (X axis, logarithmic scales) as the fraction of sites covered by marks/factors of a given type, and precision (Y axis, logarithmic scales) as the fraction of marks/factors covered by the sites.

D. Overlap between the chromatin marks described on the left and gene regions, including promoters (in red), 5'UTR (in pink), exons (in yellow) and introns (in brown) in all regions (left part of the plot) and in regions co-occupied by at least one IS (right part with green background).

Figure S5. Specificity of correlations

A-B. Hierarchical clustering of Pearson correlations between pairs of marks and/or chromatin factors for all random IS (A) or IZ (B). Marks and factors in the heatmap are organized according to the clustering described in Methods. Positive correlations are symbolized by a gradation of red that results from the localization of pairs of marks/factors at individual randomized regions. Negative correlations are symbolized by a gradation of blue. Four significant groups of chromatin mark/factor are highlighted on the right: Open 1 chromatin marks/factors group (mostly associated with transcription initiation; pale blue), Open 2 chromatin marks group (globally linked to decondensed chromatin; dark blue), Enhancer marks (grey) and Polycomb group mark/factors (brown).

C-D. Heatmap representing the Z-scores calculated by comparing the Pearson correlations between pairs of marks and/or chromatin factors for all IS (n=65,019 (**A**)) or IZ (n=35,512 (**B**)) among randomly distributed fragments (n=1000 x 65,019 and n=1000 x 35,512, respectively). Marks and factors in the heatmaps are ordered according to the clustering performed for the IS (see Fig. 3C). Positive z-scores are symbolized by a gradation of red and negative z-scores are symbolized by a gradation of blue.

E. Table showing the strongest z-scores found in IS. The highest correlations in all IS were obtained with DNAse1 hypersensitive sites (Dnase).

F. Table showing the strongest z-scores found in IZ. The highest correlations in all IZ were obtained with G4.

Figure S6. Distribution of RNA Pol II around origins

RNA Pol II ChIP-seq signals (GSE26658) from cells incubated (red line) or not (black line) with flavopiridol (Flavo) in ±1.5 Kb windows relative to the summit of all IS. The light green area shows the global profile of origins centered to the IS.

Figure S7. The presence of specific sequences at origins is inversely proportional to the chromatin mark enrichment

Profile of origin-associated motifs. Distribution profiles of origin-associated motifs, obtained by scanning origin regions with the matrices described in Fig. 6. The y-axis indicates the number of motif occurrences within each bin (300 bp). We combined the distribution profiles in four different groups based on the presence of specific chromatin marks/factors in the origins.

Figure S8. A labile nucleosome at IS

A. The left panel displays the heatmap of NS read densities in 7 kb regions on each side of the peak summits represented in Figure 1. The two other heatmaps show the read density of the two MNase datasets on each side of the peak summits. The blue intensity is proportional to the read counts per 200 bp bins. This overview indicates stronger MNase digestion in the MNase 2 dataset.

B. ChIP-seq signals for the MNase1 digestion profile at ±1.5 kb around IS for each class. In each class, we noticed the presence of a nucleosome on the IS surrounded by an NDR. However, in Class 3 IS, the NDR was strongly oriented on the left compared to the three other classes.

C. Representative examples of four raw NS profiles (light green) aligned to the two MNase-seq dataset profiles (MNase1 in black and MNase 2 in grey). We observe strong differences between the two sets of MNase raw signals only within the origin site. The increase in the nucleosome depleted region (NDR), visible only in MNase 2, is correlated with the number of IS per origin and the IZ length. We noticed the presence of a single NDR per IS (red point) in the two MNase datasets, in vicinity of the IS, but not on it.

D. H3K64ac correlation with IS. Three levels of probe intensity were considered for the correlations between IS and the presence of this mark. We show that 86% of the IS that can align on the microarray (45% of all IS) are associated with positive probes above the 25th percentile.

E. Distribution of H3K64ac Chip-on-chip positive probes around IS. We notice a thin, well-positioned peak on the IS with the top 50% highest probes.

F. Distribution of the strongest H3K64ac Chip-on-chip probes around IS for each class of origins. The bottom panels show representative examples of NS-seq (green) and H3K64ac log2(ratio) (black) profiles. We noted the presence of two different types of profiles: sharp and strong peaks that perfectly coincide with the NS peak and very broad peaks surrounding the NS peaks.

G. Representative NS-seq profiles (green) and the associated MNase (black for MNase1 and grey for MNase2), H2AZ and H3K64ac patterns. H3K64ac, but not H2AZ, was detected on IS both in intergenic regions and in all gene regions.

Figure S9. Profile of histone H3 and G-rich occurrences around the IS considering the DNA strand on which the G-rich occurrences are located.

A. Profile of histone H3 and G-rich occurrences around the IS considering the DNA strand on which the G-rich occurrences are located. Nucleosome depletion is clearly related to the presence of G-rich occurrences, with the maximum point in G-rich occurrences corresponding to the strongest H3 depression.

Supplemental Table S1: List of the epigenetic features analyzed in this study.
The 43 histone modifications and chromatin factors used in this study are listed and their overlap with initiation sites, initiation zones, and each class of origins are indicated.

Supplemental Table S2: List of the relevant *k*-mers detected by our analyses, and the associated eight clusters further assembled to form degenerated DNA motifs.

SUPPLEMENTAL INFORMATION

Sequencing and read mapping

Sequenced reads were mapped against the mm9 mouse genome sequence (NCBI Build 37) with the CASAVA software (version 1.8 – Illumina). After mapping, unique mapping reads of the nascent strand (NS) replicates 1, 2 and 3 (59M, 43M, 10M reads, respectively) were combined (112M unique reads) for further analysis, and the genomic DNA (45M reads) and RNAse A (63M reads) samples were recovered.

Peak calling

A large variety of tools exist for peak calling (Pepke et al., 2009), many of which are specifically designed to treat ChIP-seq data types, such as short TFBS or larger histone marks. None of them was designed to treat NS sequencing data. Given the original nature of NS data, we tested several software tools: MACS (Zhang et al., 2008) combined with PeakSplitter (Salmon-Divon et al., 2010), SWEMBL (http://www.ebi.ac.uk/~swilder/SWEMBL/) and SICER (Zang et al., 2009). SICER was originally designed to identify regions ("broad peaks") enriched in histone modifications. SWEMBL identifies narrow peaks in read density landscapes. MACS is presented as a more generalist tool that can be used for both peak and region identification, depending on parameter tuning.

Based on the visual inspection of read densities and peak locations and on the significance of the discovered motifs, we discarded MACS, which tends to return large multi-peak regions rather than isolated peaks. The helper program PeakSplitter partly solves this problem by splitting the large MACS peaks in smaller peaks, but SWEMBL alone returned narrower peaks and their subsequent analysis revealed a more precise correspondence with peaks of motif occurrences. We thus retained SWEMBL for the identification of narrow peaks in NS reads. Based on this preliminary evaluation, we combined two peak-calling programs that show complementary properties (Fig. 1B): SWEMBL identifies well-delimited narrow peaks, whereas SICER returns wider regions enriched in reads, which often contain one or several SWEMBL peaks. We defined as *initiation sites* the narrow peaks (SWEMBL) overlapping a SICER broad region, and as *initiation zones* the broad regions (SICER) overlapping one or more SWEMBL narrow peaks. This combined analysis allows isolating consistent features, by discarding spurious peaks and regions showing some enrichment in reads, but without clear local summit.

Number of peaks returned by the respective approaches

Peak-callers	Region type	Number of peaks	Mean peak length (bp)	Total size (Mb)
SWEMBL (R=0.002)	Narrow peaks	70,945	165	11.74
SICER (FDR <= 0.01)	Broad peaks	78,688	3,023	237.88
SWEMBL + SICER	Initiation sites	65,019	170	11.08
SICER + SWEMBL	Initiation zones	35,512	3,938	139,84

Overlap with genomic features and data from other sources

To compute the overlap of origin sites with genomic annotations, we defined fixed-width intervals around peak summits (±500bp). Overlaps were computed using bedtools intersectBed (Quinlan et al., 2010). To draw profiles of CpG islands, promoters and genes along read density maps, we labeled each peak with a Boolean flag. The peak-wise Boolean values (overlap or not) were smoothed using a mean sliding window covering 500 peaks. We also tested alternative interval and sliding window size values (not shown). Genome localizations of initiation sites were computed using CEAS (http://liulab.dfci.harvard.edu/CEAS/) against the Refseq Gene annotation (Mouse genome version NCBI37/mm9 NCBI37).

To identify peaks overlapping with a CpG island (CGI), we considered 500bp intervals on each side of the peak summit. This fixed-width criterion avoids the bias of variable peak lengths when comparing CGI overlap rates between peak clusters. Negative controls were performed by selecting 65,019 random genomic summits and by applying the same procedure.

For comparison and clustering, ChIP-seq data for histone marks and transcription factors from mouse ES cells were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). When available in the NCBI37/mm9 NCBI37 coordinates, we favored processed reads provided directly from the data sources. When the necessary data were only available in NCBI36/mm8 NCBI36 coordinates, they were remapped using the Bowtie short read aligner (Langmead et al., 2009) (Parameters: --end-to-end --sensitive -p 1 -t). Aligned reads were further manipulated with SAMtools (sort) (Li et al., 2009) and bedtools bamToBed (Quinlan and Hall, 2010). Read density profiles for

H3K4me3, H3K9ac, H3K27me3, SUZ12, H3K4me1, 5hmC, H3K27ac and RNA POL II were computed using mapped reads, as previously described. For average profile plots, reads were counted per 100bp bins ±7kb around IS, or per 50bp bins when looking at the ±1.5kb region around IS. On the other hand, the H3K64ac data presented in Figure 2C consisted on the analysis of chip-on-chip data (GEO: GSE35355). To plot histone peak profiles we retained chip-chip probes having a log2 (ChIP/Input) ratio greater than 0, and separated the probes in two groups (below or above the median value). The distribution of probes corresponding to these two groups (low, high) was then mapped around ±1.5kb of the IS summits. The chip-chip probes mapping around ±1.5kb of the IS summits were processed with BEDTools (Quilan A.R, Bioinformatics 2010). Figure 2C shows the profile of probes having the highest H3K64ac enrichment (high set of probes). The H3K64ac distribution around IS for both groups (low, high) is shown in Figure S4E and S4F for the four classes for the highest group.

Bed files defining mark/TFBS locations were taken from publications, Supplemental Materials or provided directly by authors (CpG, H3K4me3, ES_FMR, ES_LMR, ES_UMR, G4, H2A.Z, Ac_H2A.Z, SMARC4, POL2_global, POL2_Ser2P, POL2_Ser5P, TBP, RNF2, CHD4, SUZ12, mof, 5fC, 5hmC, p300, H3K27me3, H3K4me1, H3K09ac, H3K36me3, H3K27ac, H3K09me3, Dnase, H3K56ac, NANOG, SOX2, E2F1, CTCF, KLF4, ESET, H3K4me2, EZH2, SMC1, SMC3, MED1, WDR5, OCT4, bivalent_domain, Rest). Table S1 indicates the identifiers for all the marks used in this study.

We also investigated the potential regulatory role of the H3K4me3 mark in the establishment of replication origins. For this, we analyzed the average profile from our purified nascent strands (chip-on-chip, GEO: GSE68347) against the results obtained from a *Cfp1* knockout mouse ES cell line (GEO: GSE53490). In these cells, H3K4me3 is redistributed in the genome. Three genomic regions were identified by (Clouaire T et al. 2012 and 2014). They correspond to ectopic H3K4me3 regions, TSS with increased H3K4me3 and TSS with decreased H3K4me3. We retained regions with overlapping probes present in our chip-on-chip array. The ChIP-seq average profile analyses for RNA POL II and H3K4me3 were performed as previously described. The average profile analyses of nascent strands from wild type (WT) and *Cfp1*-/- cells were carried out using our previously published chip-on-chip data. We retained chip-chip probes with a log2 (ChIP/Input) ratio greater than 0, and separated these probes into tiers. The chip-chip probes mapping around ±3kb of the IS summits were processed with BEDTools. For the three regions identified by Clouaire et al,

Figure S6A shows the average profile for all probes as well as for the probes above 66% (tier3).

Timing regions in mouse ES cells were downloaded from the ReplicationDomain tool (Weddington et al., 2008). Coordinates were converted from NCBI36/mm8 to NCBI37/mm9 with the UCSC *liftover* tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver), using the default parameters (minimum ratio of bases that must remap: 0.7). Overlaps were computed using bedtools intersectBed. The random expectation was estimated by computing the mean overlap value for 1,000 random sets of 65,019 peaks.

For computing the average profiles, ChIP-seq and MNase data were processed as in (Fenouil et al., 2012). Briefly, raw (fastq) data files were aligned using bowtie and scores were piled up in WIG files and merged in 50bp bins after estimation of the average fragments size. For each set of annotations, all bin scores in the appropriate neighboring regions were extracted and linear interpolation was used to realign the bin scores to the exact annotations coordinates before computing the average score at each coordinate.

Ab initio motif discovery

Ab initio motif discovery was performed with the RSAT tool position-analysis (van Helden et al., 2000), but the program was improved by implementing new statistical models (Bernoulli, Markov) and by extending the post-processing (*k*-mer clustering and extraction of matrices from the *k*-mers). We tested three types of background models: uniform (flat) distribution, window-specific Bernoulli model (the frequency of each *k*-mer is estimated, for each window, as the product of the local frequencies of its residues) and window-specific Markov models. For the final results, we retained the Bernoulli model, which presents a good trade-off between specificity and sensitivity. The results obtained with the alternative models are available on the supporting web site. The significance of the difference between observed and expected distributions was assessed with the chi2 test, as described in the original publication about position-analysis (van Helden et al., 2000). The resulting significance score is defined as $sig = -log10(e\text{-}value) = -log10(p\text{-}value * n)$, where n is the number of *k*-mers considered. The top-ranking *k*-mers (e-value <= 0.1, rank <= 100) were clustered on the basis of their positional profile of occurrences relative to the peak centers, using the R function hclust (hierarchical clustering, complete agglomeration rule, correlation distance) to collect groups of words with similar occurrence profiles. The hierarchical tree of *k*-mers was cut at various levels (*k* from 6 to 12), and the relevant number of clusters was defined by visual inspection

of the positional profiles. *K*-mers belonging to the same cluster were assembled (*RSAT pattern-assembly*) to reveal longer and/or degenerated motifs and used as seeds to build position-specific scoring matrices (PSSM) that were used to scan peak sequences for the presence of motifs.

Sequences were scanned on each strand separately using *matrix-scan* (Turatsinze et al, 2008), with the following parameters: Markov background model of order 1 trained on the input sequences, p-value threshold set to 1e-4. Position profiles of the resulting matrices were used to scan peak regions and to estimate the position profiles of sequence coverage by each motif.

Supplementary references

Clouaire T, Webb S, Skene P, Illingworth R, Kerr A, Andrews R, Lee JH, Skalnik D, Bird A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes Dev. Aug 1;26(15):1714-28.

Clouaire T, Webb S, Bird A. (2014). Cfp1 is required for gene expression-dependent H3K4 trimethylation and H3K9 acetylation in embryonic stem cells. Genome Biol. Sep 4;15(9):451.

Foulk MS, Urban JM, Casella C, Gerbi SA. (2015). Characterizing and controlling intrinsic biases of Lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. Genome Research, Feb 18.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

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

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nature methods *6*, S22-32.

Quinlan, A. R., Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. *26*, 841-842.

Salmon-Divon, M., Dvinge, H., Tammoja, K., and Bertone, P. (2010). PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC bioinformatics *11*, 415.

Turatsinze, J.V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nature protocols *3*, 1578-1588.

van Helden, J., Olmo, M. & Perez-Ortin, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res 28(4), 1000-1010.

Weddington, N., Stuy, A., Hiratani, I., Ryba, T., Yokochi, T., and Gilbert, D.M. (2008). ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. BMC bioinformatics *9*, 530.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-seq data. Bioinformatics *25*, 1952-1958.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al*. (2008). Model-based analysis of ChIP-seq (MACS). Genome biology *9*, R137.

Figure S1

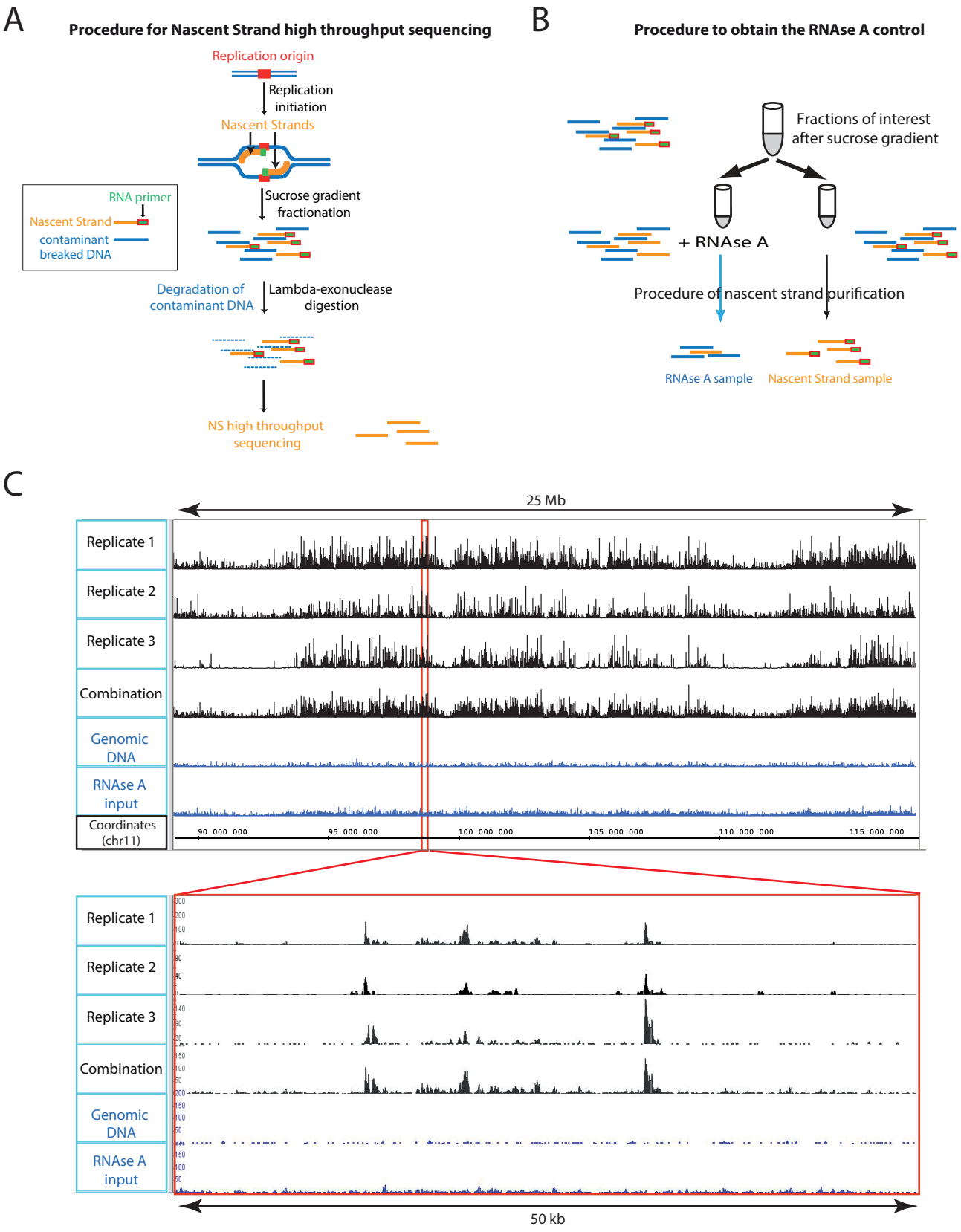

A
Procedure for Nascent Strand high throughput sequencing
Replication origin
Replication initiation
Nascent Strands
Sucrose gradient fractionation
RNA primer
Nascent Strand
contaminant breaked DNA
Degradation of contaminant DNA
Lambda-exonuclease digestion
NS high throughput sequencing
B
Procedure to obtain the RNAse A control
Fractions of interest after sucrose gradient
+ RNAse A
Procedure of nascent strand purification
RNAse A sample
Nascent Strand sample
C
25 Mb
Replicate 1
Replicate 2
Replicate 3
Combination
Genomic DNA
RNAse A input
Coordinates (chr11)
90 000 000
95 000 000
100 000 000
105 000 000
110 000 000
115 000 000
50 kb
D
Peak (Initiation Site=IS)
Zone (Initiation Zone=IZ)
Peak/zone
1
2
>2
18%
22%
60%
E
200 Kb
Nascent strands signal
ES.cont.3rep_logratio
negative Control

Figure S2

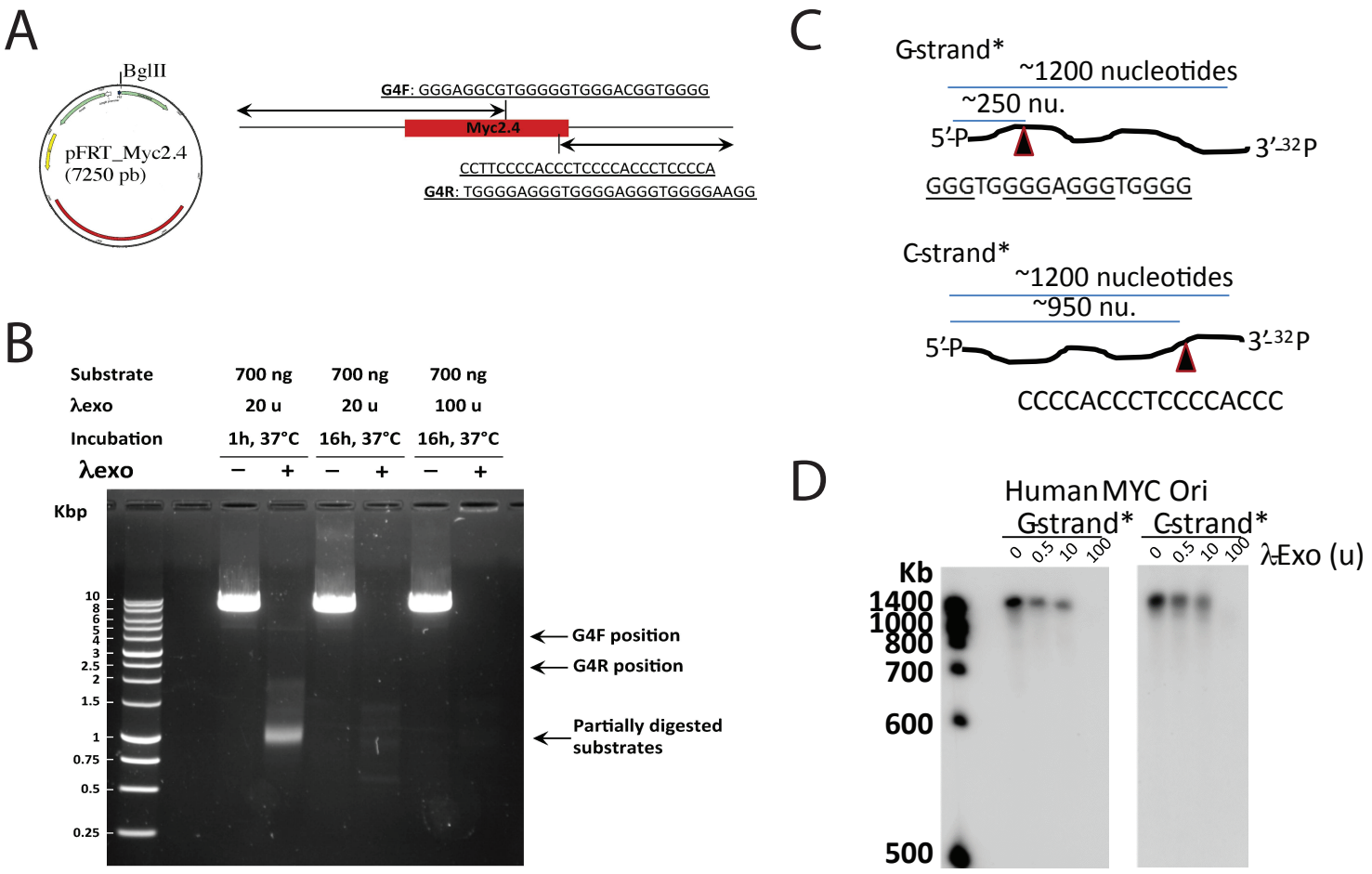

A
BglII
pFRT_Myc2.4
(7250 pb)
G4F: GGGAGGCGTGGGGGTGGGACGGTGGGG
Myc2.4
CCTTCCCCACCCTCCCCACCCTCCCCA
G4R: TGGGGAGGGTGGGGAGGGTGGGGAAGG
B
Substrate
700 ng
700 ng
700 ng
λexo
20 u
20 u
100 u
Incubation
1h, 37°C
16h, 37°C
16h, 37°C
λexo
– + – + – +
Kbp
10
8
6
5
4
3
2.5
2
1.5
1
0.75
0.5
0.25
G4F position
G4R position
Partially digested substrates
C
G-strand*
~1200 nucleotides
~250 nu.
5'-P
3'-32P
GGGTGGGGAGGGTGGGG
C-strand*
~1200 nucleotides
~950 nu.
5'-P
3'-32P
CCCCACCCTCCCCACCC
D
Human MYC Ori
G-strand*
C-strand*
0 0.5 10 100
0 0.5 10 100
λ-Exo (u)
Kb
1400
1000
800
700
600
500

Figure S3

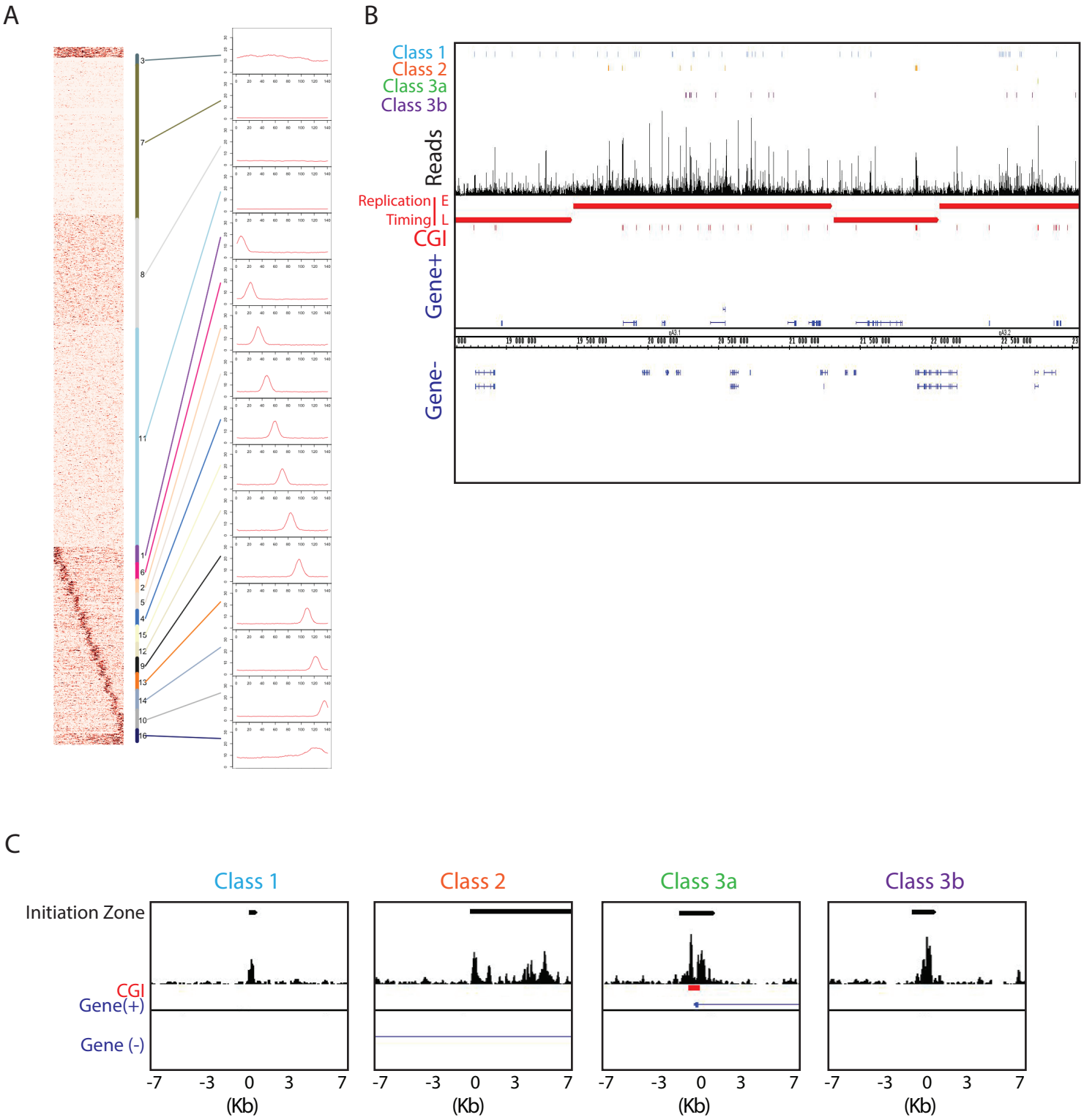

A
B
Class 1
Class 2
Class 3a
Class 3b
Reads
Replication Timing
E
L
CGI
Gene+
Gene-
C
Class 1
Class 2
Class 3a
Class 3b
Initiation Zone
CGI
Gene(+)
Gene (-)
-7
-3
0
3
7
(Kb)

Figure S4

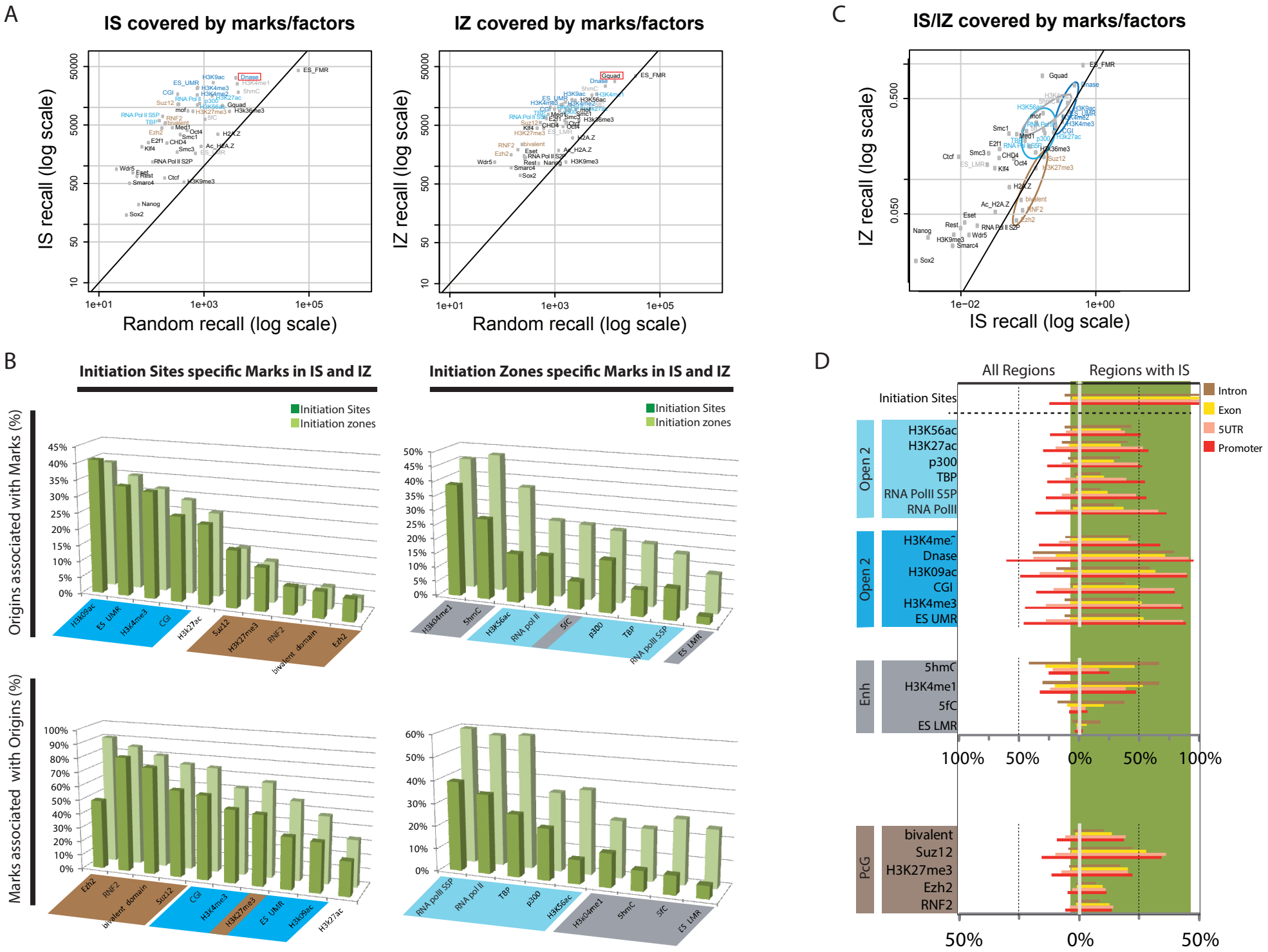
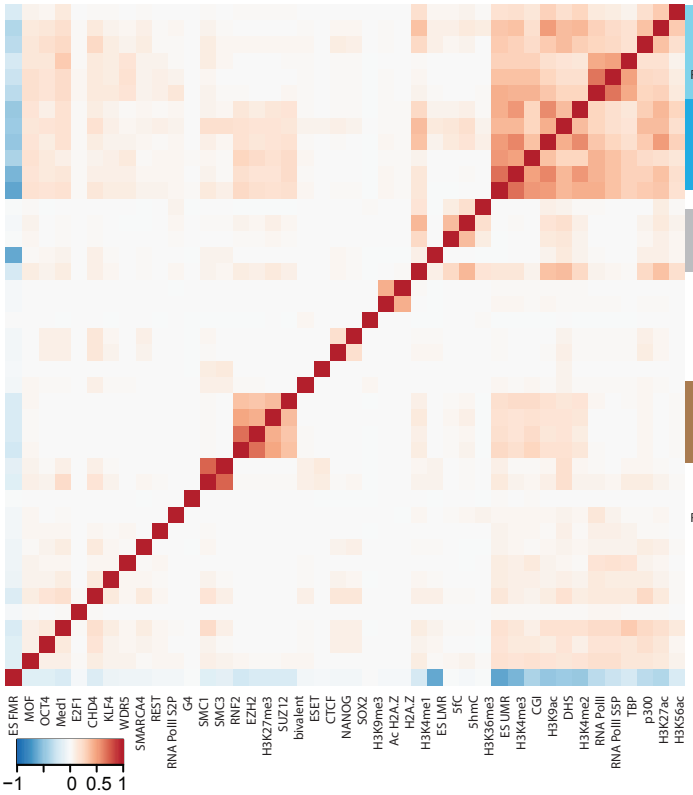
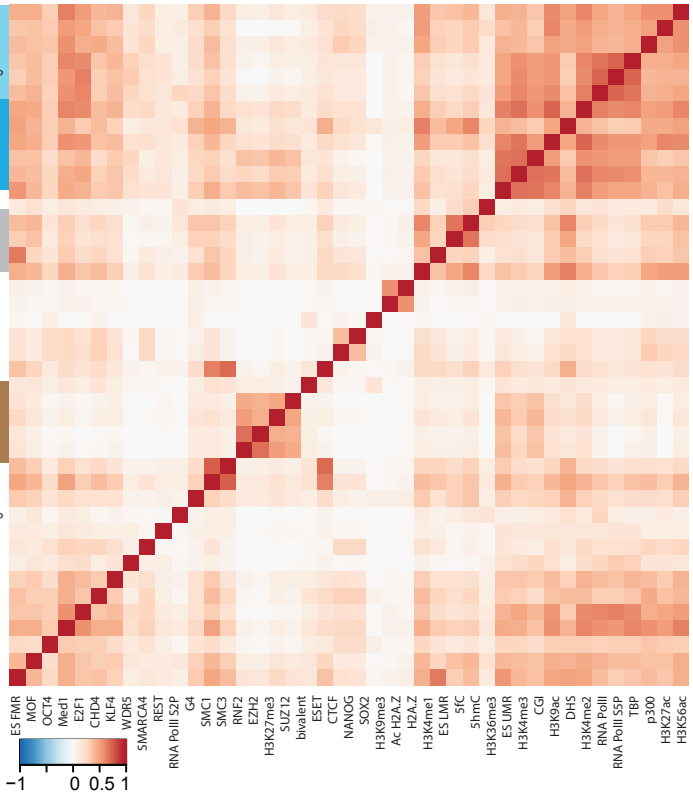

A
IS covered by marks/factors
IZ covered by marks/factors
IS recall (log scale)
IZ recall (log scale)
Random recall (log scale)
C
IS/IZ covered by marks/factors
IZ recall (log scale)
IS recall (log scale)
B
Initiation Sites specific Marks in IS and IZ
Initiation Zones specific Marks in IS and IZ
Initiation Sites
Initiation zones
Origins associated with Marks (%)
Marks associated with Origins (%)
D
All Regions
Regions with IS
Initiation Sites
Intron
Exon
5UTR
Promoter
Open 2
Open 2
Enh
PcG

Figure S5

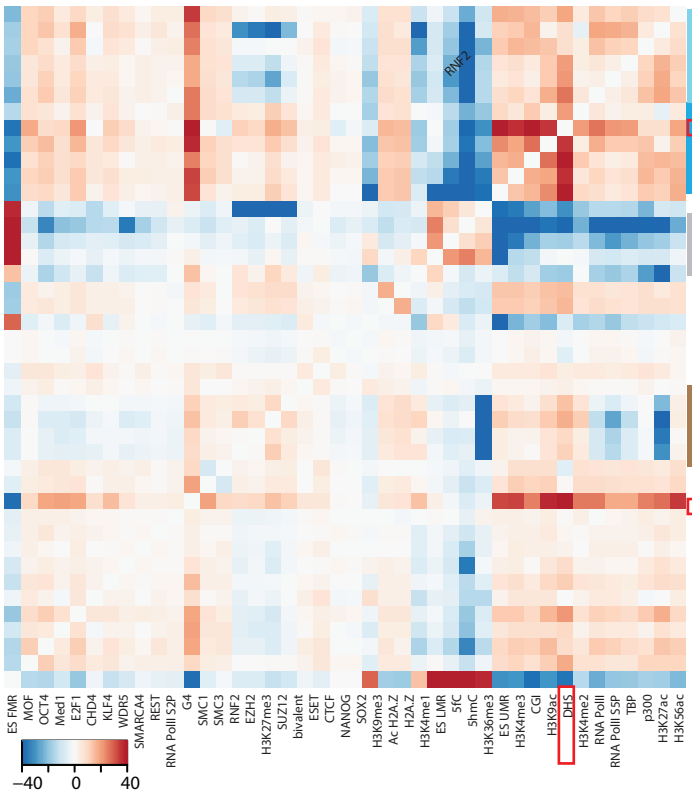
A Mean correlations betw. marks in random peaks



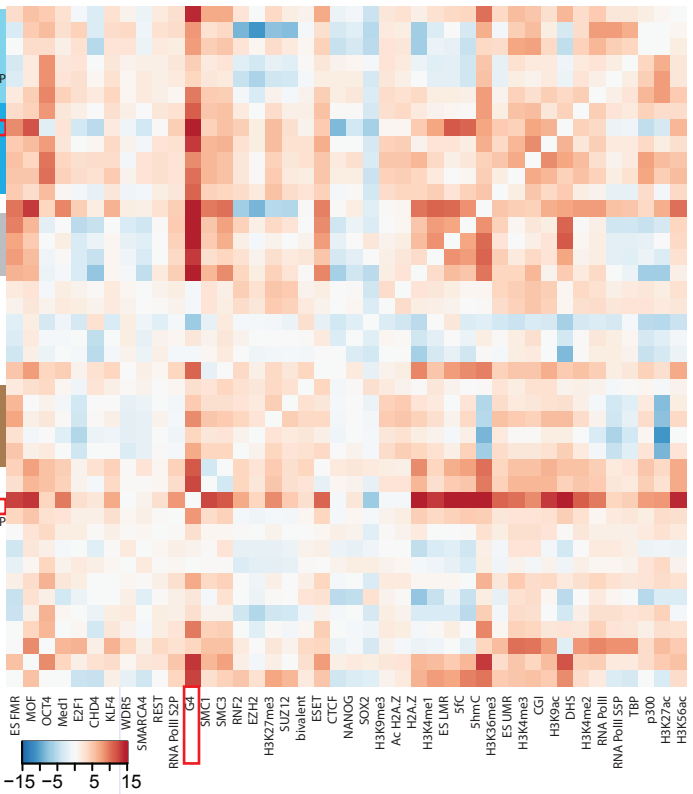
B Mean correlations betw. marks in random zones



C Correlation gain at IS (z−score)



D Correlation gain at IZ (z−score)



E

mark1	mark2	IS	rand mean	z (score)
Dnase	CpGii	0.579	0.264	34.0
Dnase	ES UMR	0.687	0.412	31.2
Dnase	H3K4me3	0.679	0.415	28.5
Dnase	H3K9ac	0.708	0.457	27.7
Dnase	RNA PolII	0.455	0.262	21.2
Dnase	SUZ12	0.433	0.271	17.5
Dnase	H3K27me3	0.308	0.174	14.3

F

mark1	mark2	IZ	rand mean	z (score)
G4	5hmC	0.609	0.360	17.6
G4	H3K4me1	0.564	0.333	17.4
G4	5fC	0.537	0.299	15.5
G4	ES LMR	0.382	0.197	13.2
G4	Dnase	0.652	0.427	17.1
G4	H3K9ac	0.485	0.284	13.2
G4	ES UMR	0.477	0.302	10.9
G4	H3K4me3	0.411	0.260	10.5
G4	CpGi	0.406	0.263	9.1
G4	SUZ12	0.373	0.213	9.5
G4	H3K27me3	0.303	0.173	9.0

Figure S6

Distribution of RNA pol II around IS

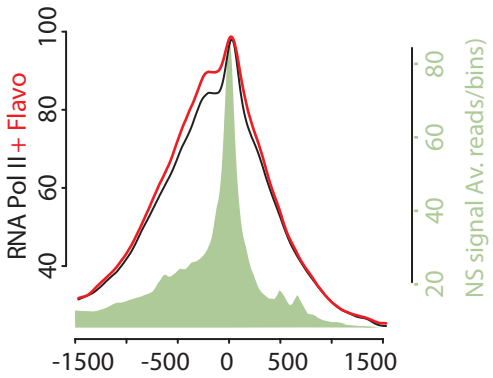


Figure S7

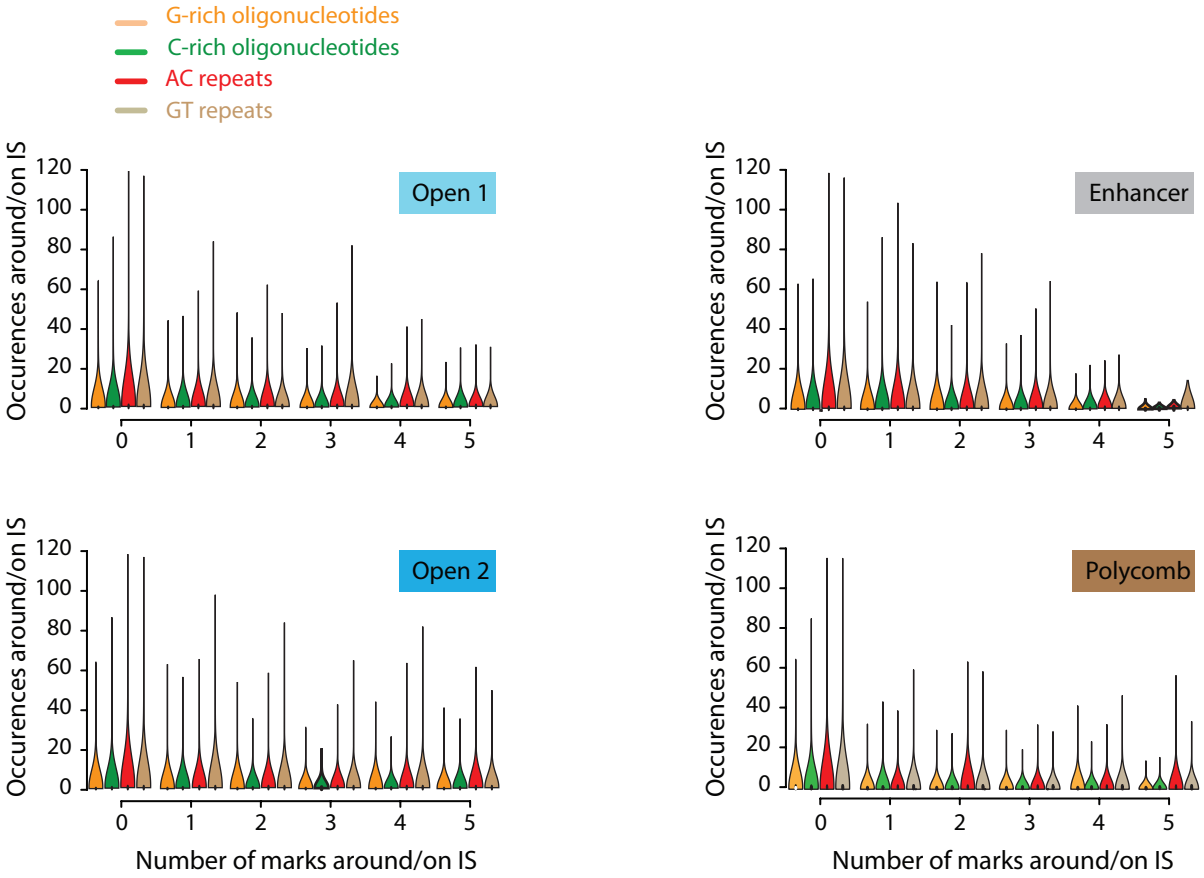

G-rich oligonucleotides
C-rich oligonucleotides
AC repeats
GT repeats
Open 1
Enhancer
Open 2
Polycomb
Occurences around/on IS
Number of marks around/on IS
0
20
40
60
80
100
120
0
1
2
3
4
5

Figure S8

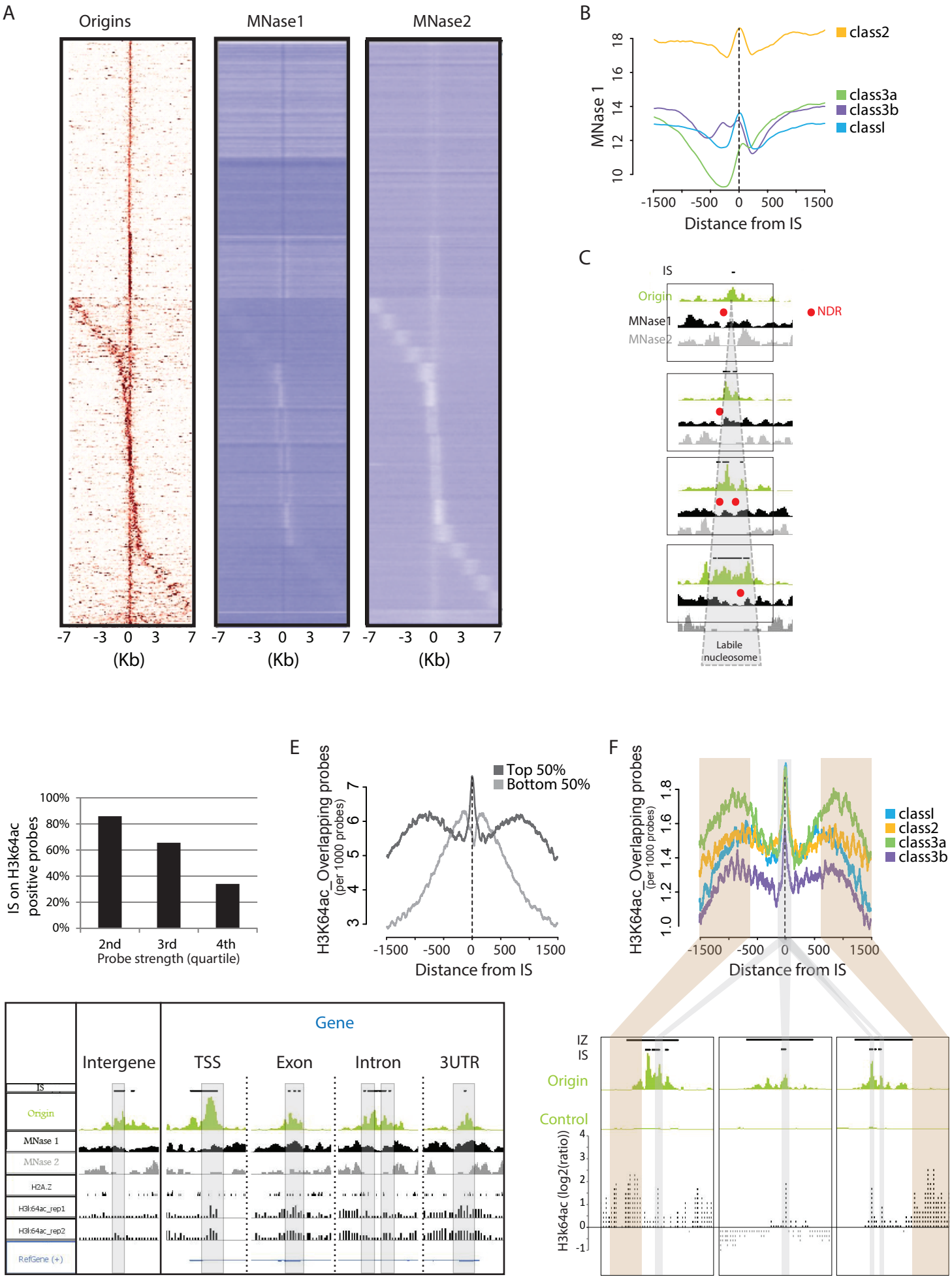
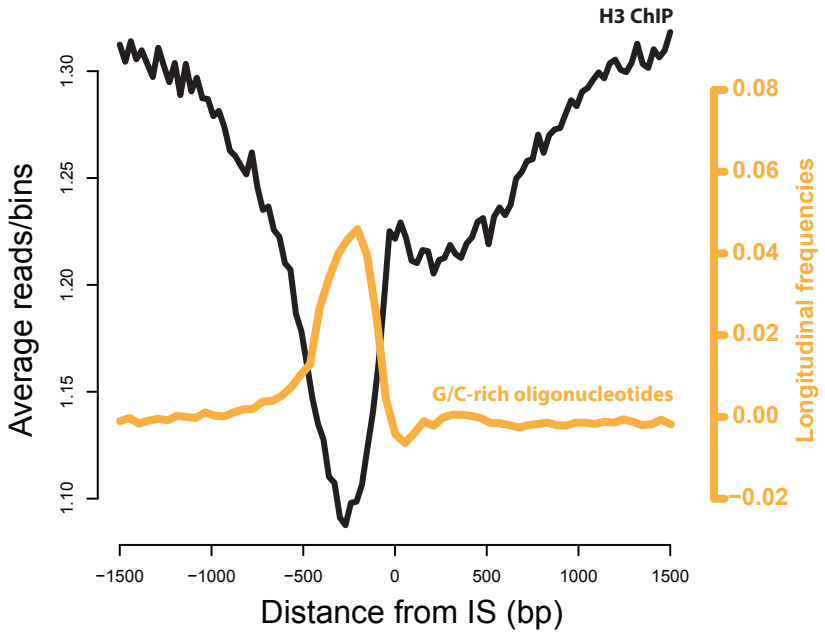

A
Origins
MNase1
MNase2
-7 -3 0 3 7
(Kb)
B
MNase 1
class2
class3a
class3b
classI
-1500 -500 0 500 1500
Distance from IS
C
IS
Origin
MNase1
MNase2
NDR
Labile nucleosome
D
IS on H3k64ac positive probes
100%
80%
60%
40%
20%
0%
2nd
3rd
4th
Probe strength (quartile)
E
H3K64ac_Overlapping probes (per 1000 probes)
Top 50%
Bottom 50%
Distance from IS
F
H3K64ac_Overlapping probes (per 1000 probes)
classI
class2
class3a
class3b
Distance from IS
G
Gene
Intergene
TSS
Exon
Intron
3UTR
IS
Origin
MNase 1
MNase 2
H2A.Z
H3k64ac_rep1
H3k64ac_rep2
RefGene (+)
IZ
IS
Origin
Control
H3k64ac (log2(ratio))

Figure S9

A


H3 ChIP
G/C-rich oligonucleotides
Average reads/bins
Longitudinal frequencies
1.10
1.15
1.20
1.25
1.30
0.08
0.06
0.04
0.02
0.00
−0.02
−1500
−1000
−500
0
500
1000
1500
Distance from IS (bp)