Supplemental Material

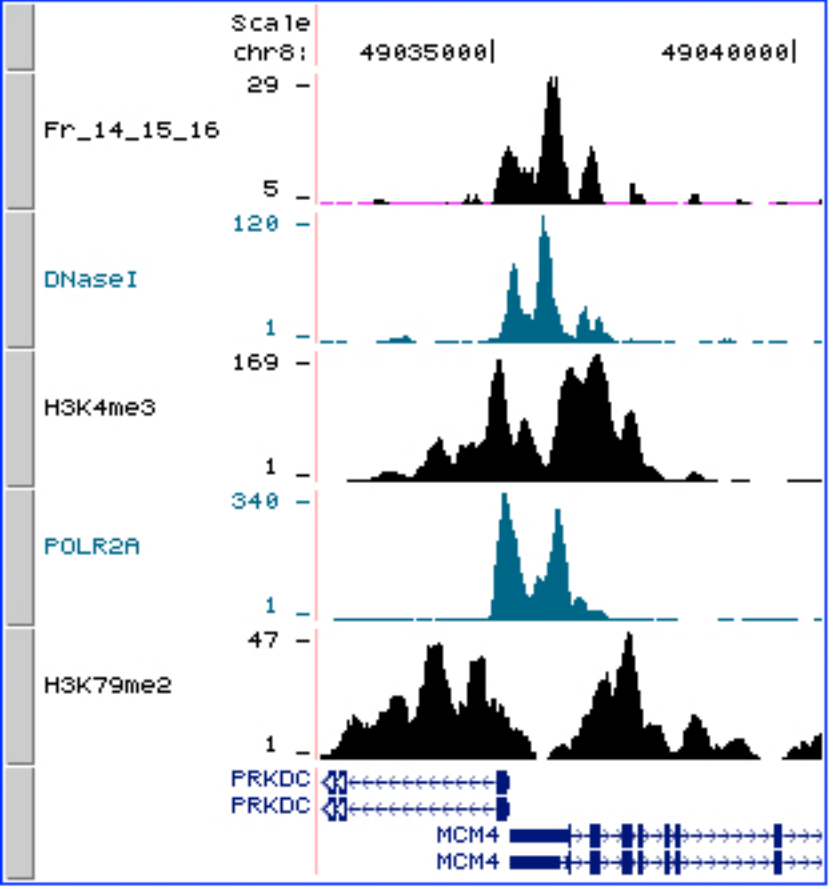
Section 1: Supplemental Figures (S1-S14) and Supplemental Tables (S1-S6)

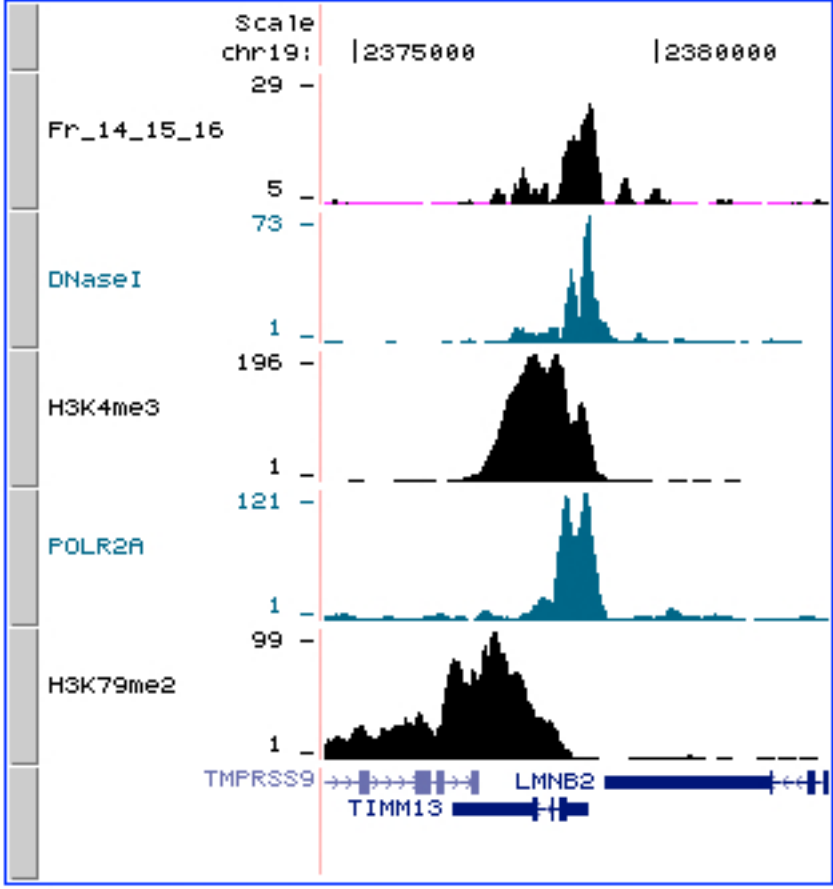
Section 2: Supplemental figure/table legends

Section 3: Supplemental Methods (Statistical analyses) and Supplemental References

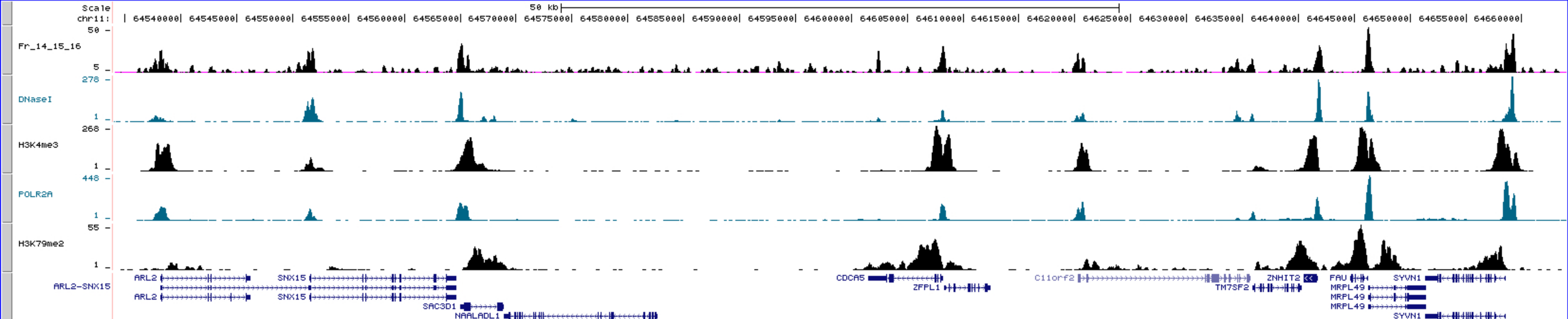
Section 1

Supplemental Figures (S1-S14) and Supplemental Tables (S1-S6)

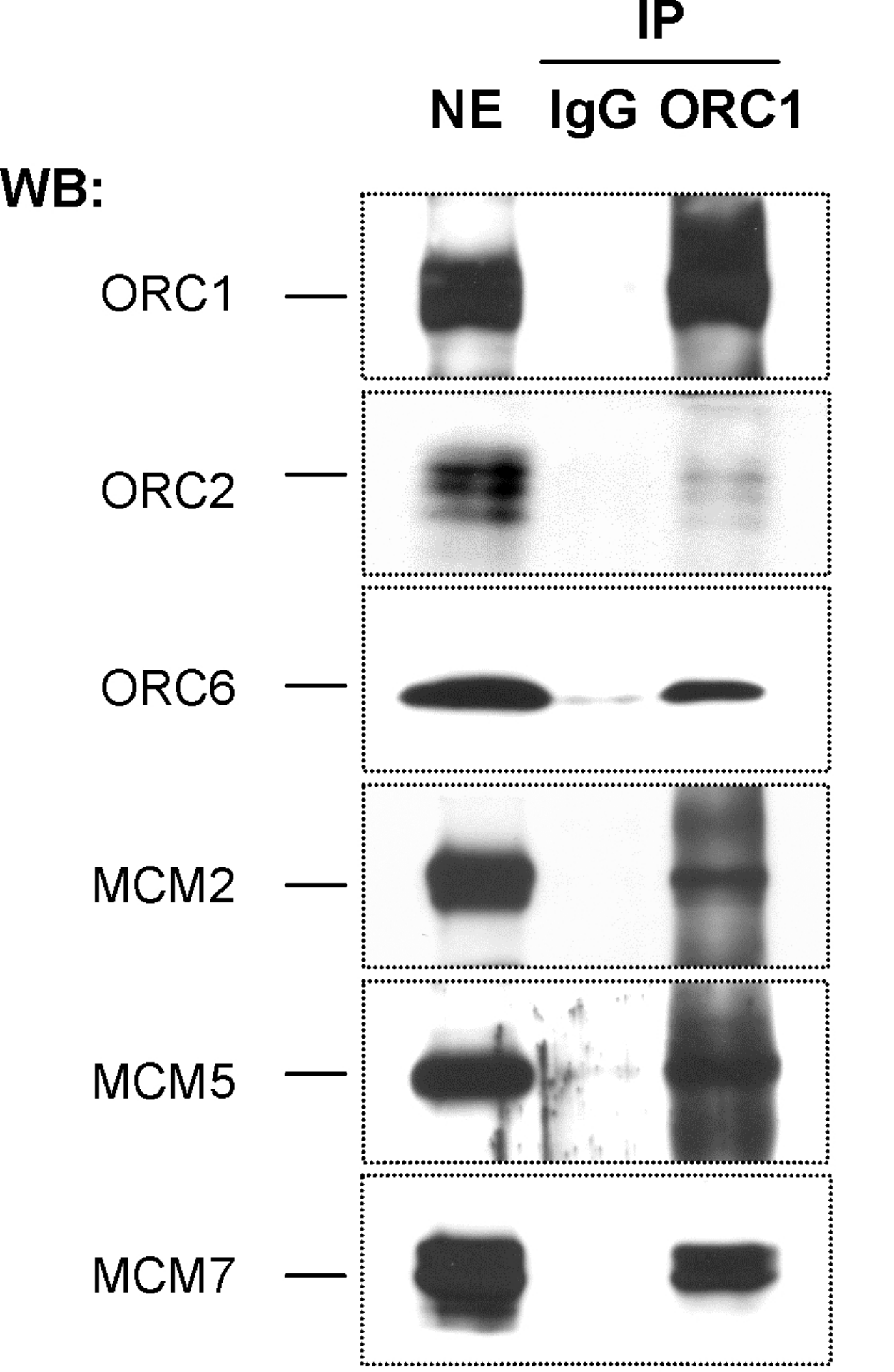




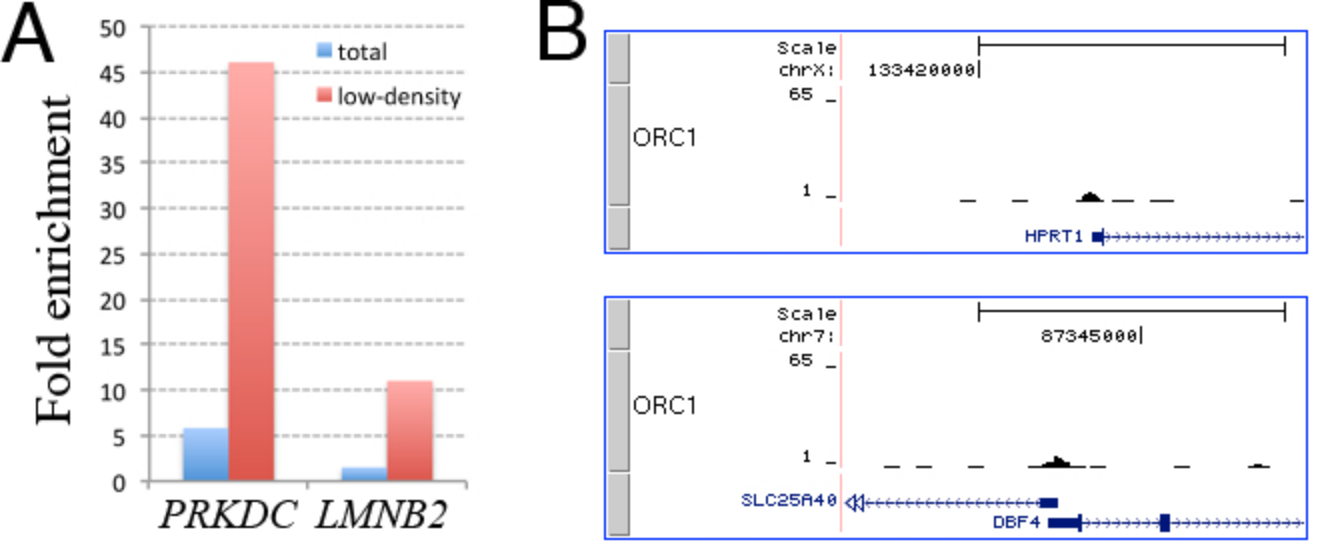
Supplemental Figure S1



Supplemental Figure S2

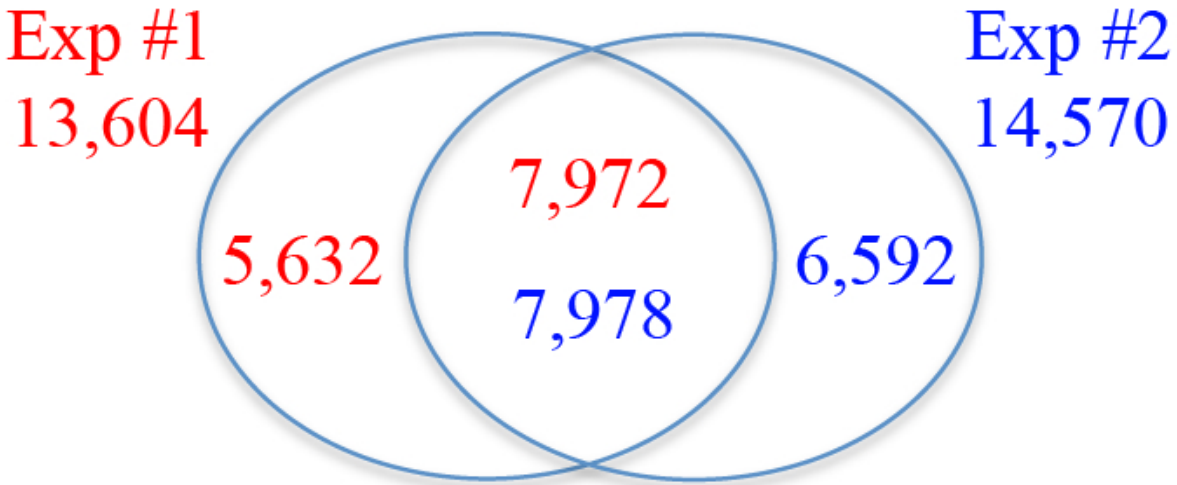


Supplemental Figure S3

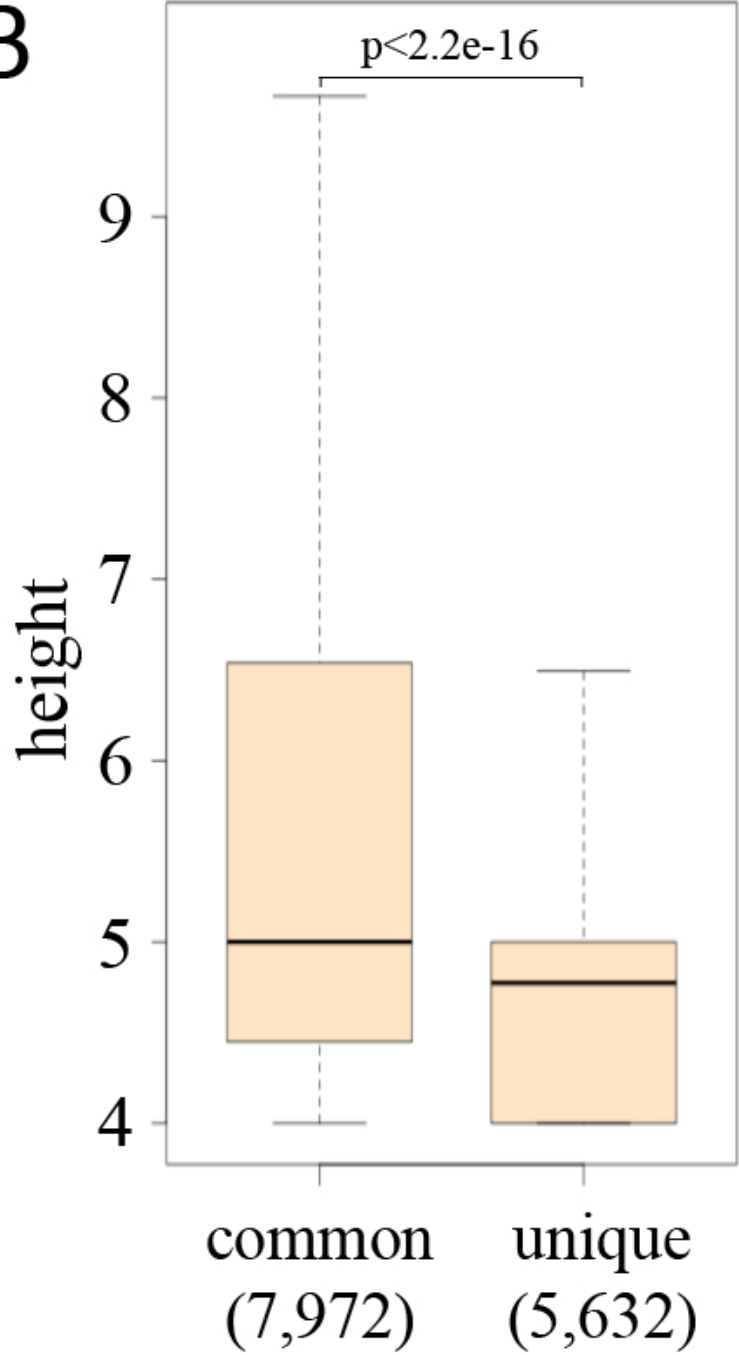


Supplemental Figure S4

A

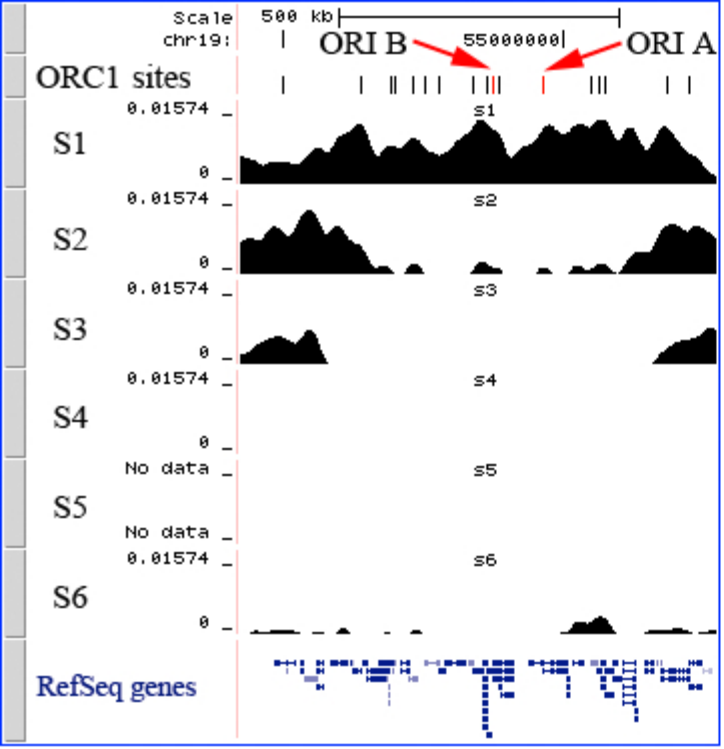


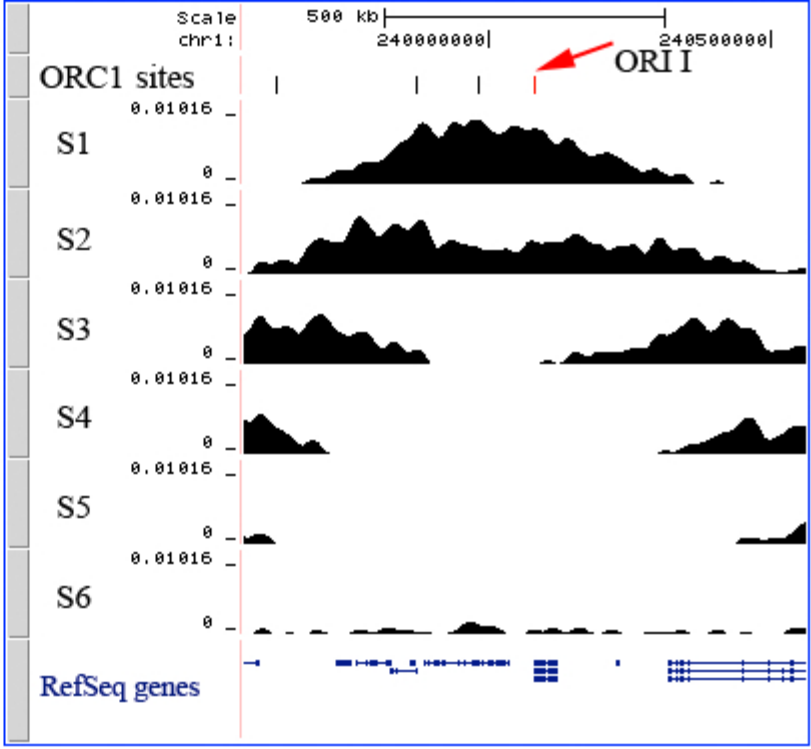
B

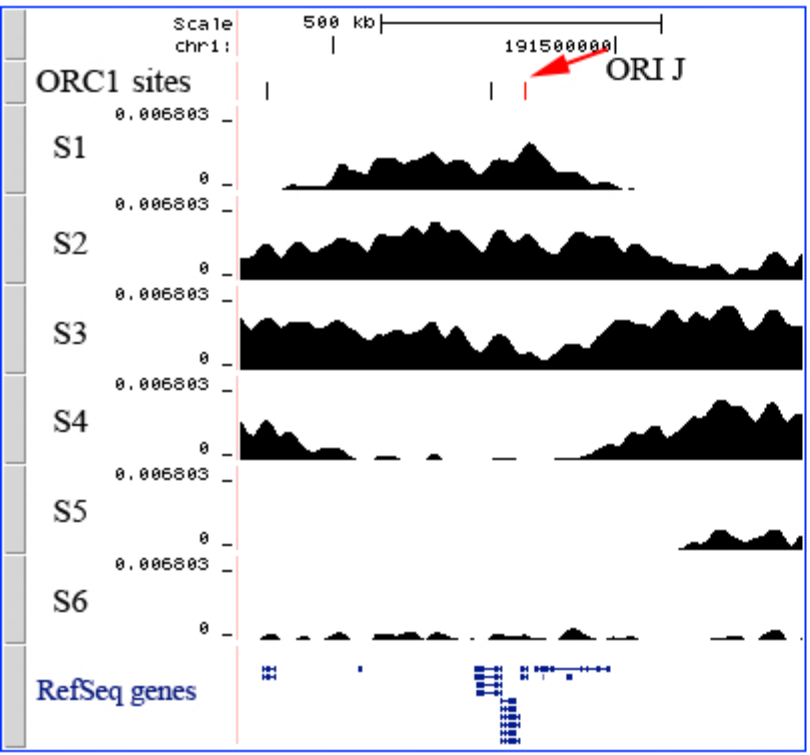


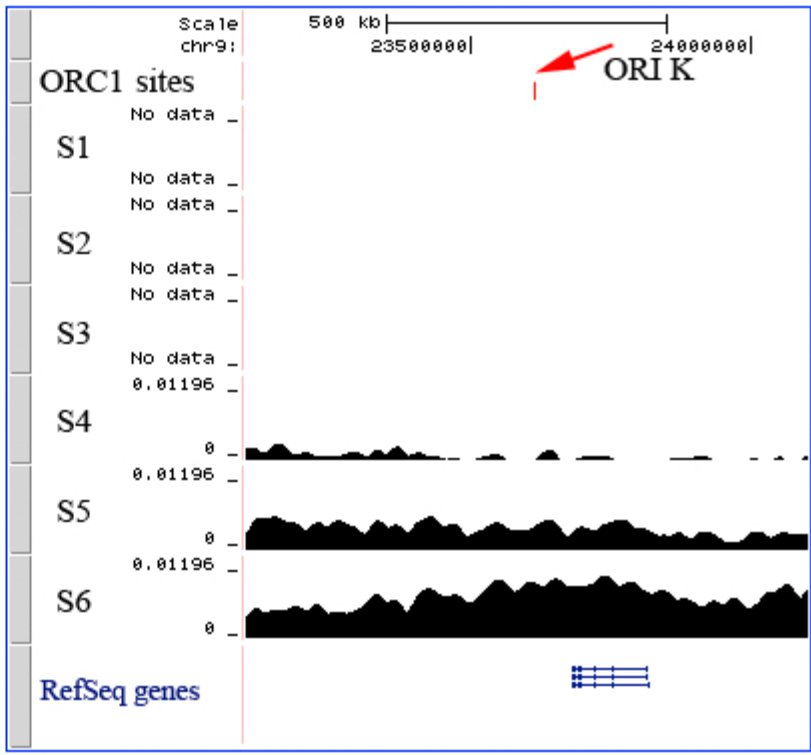
Supplemental Figure S5

Supplemental Figure S6

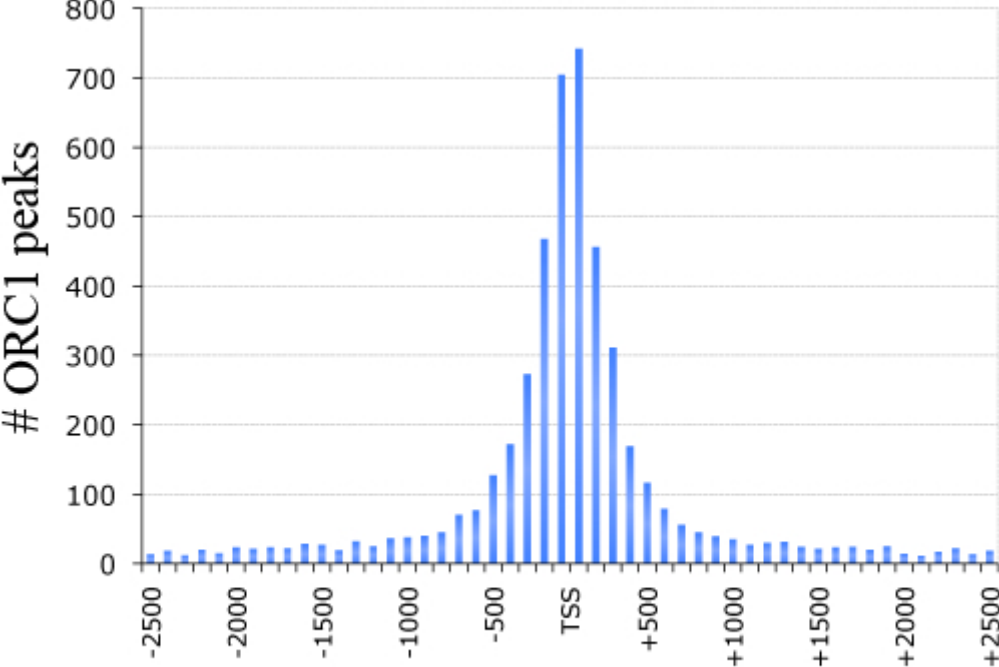




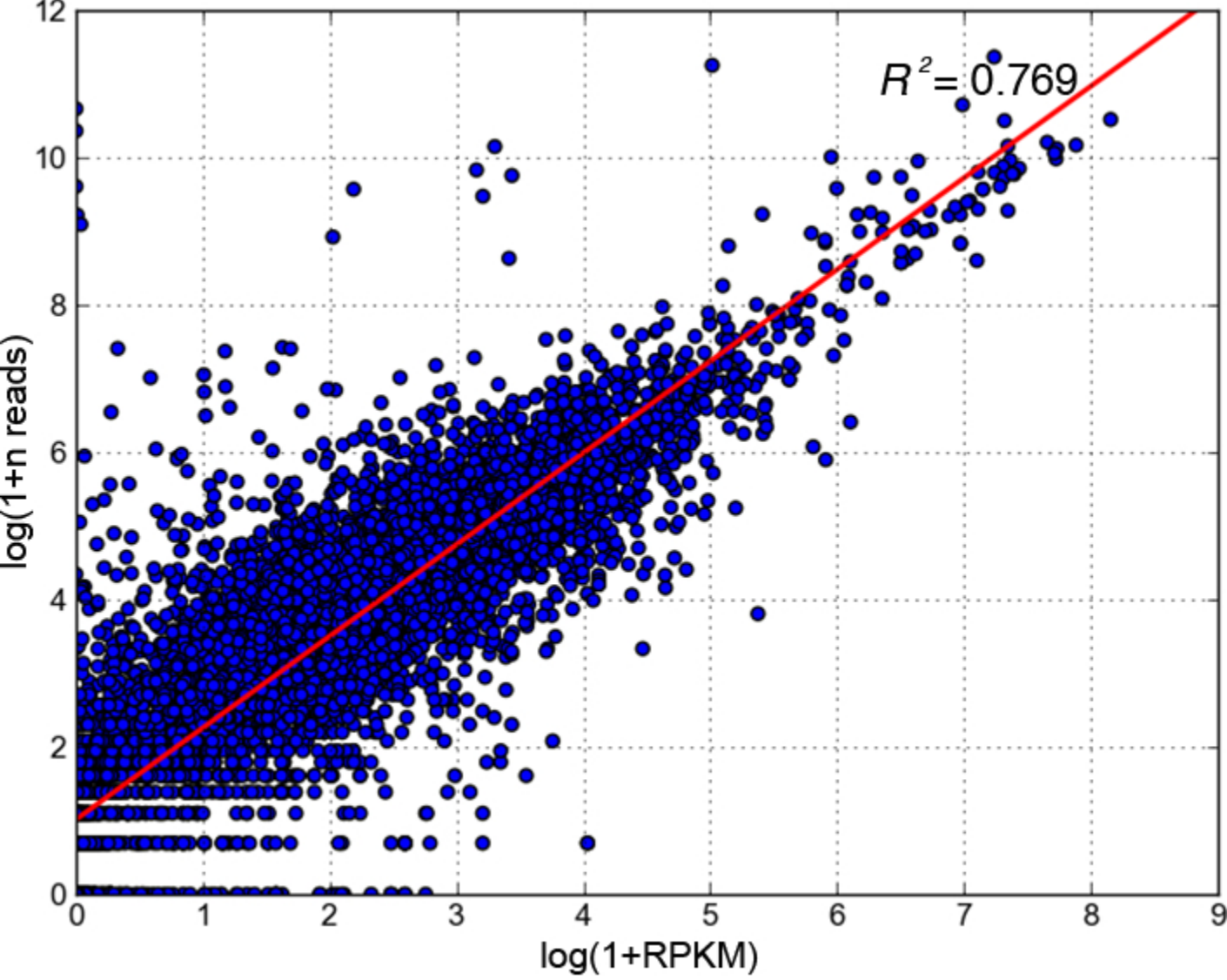




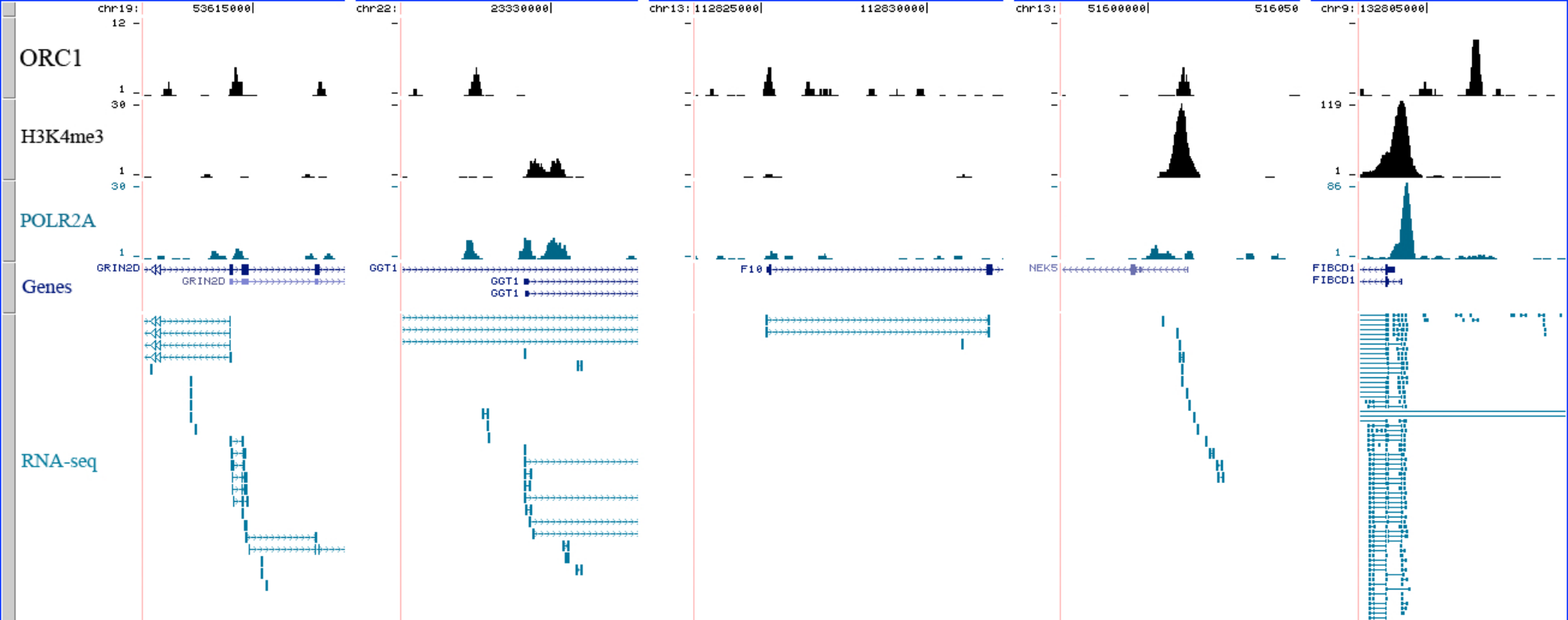
Supplemental Figure S7



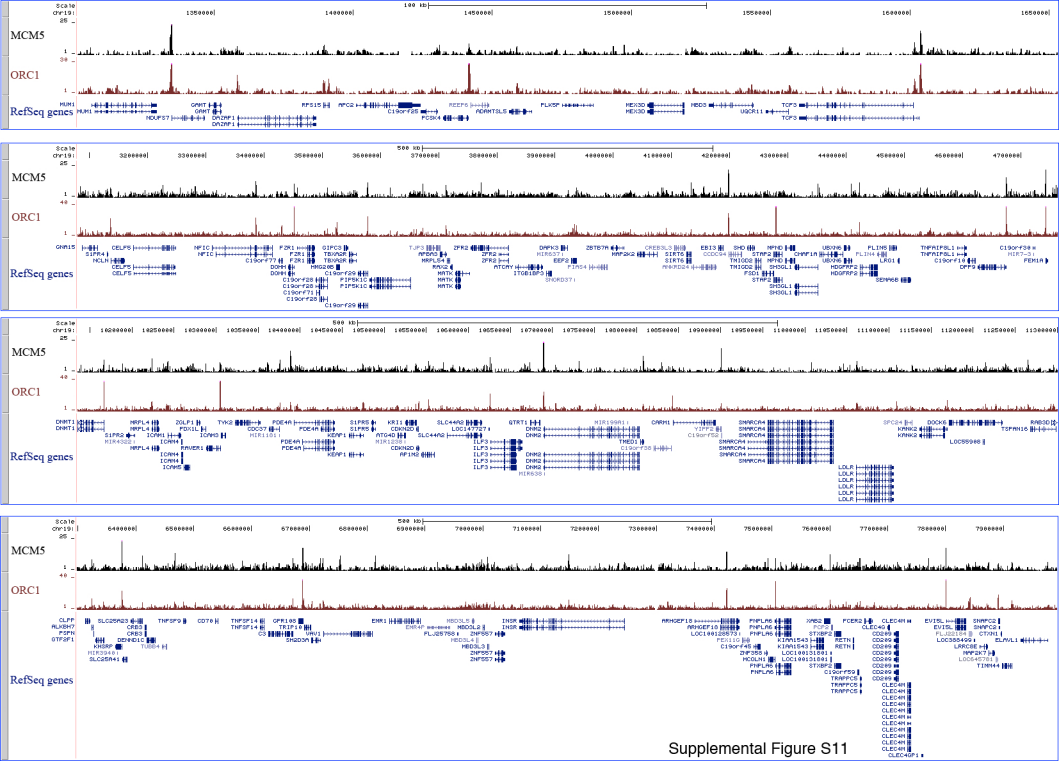
Supplemental Figure S8



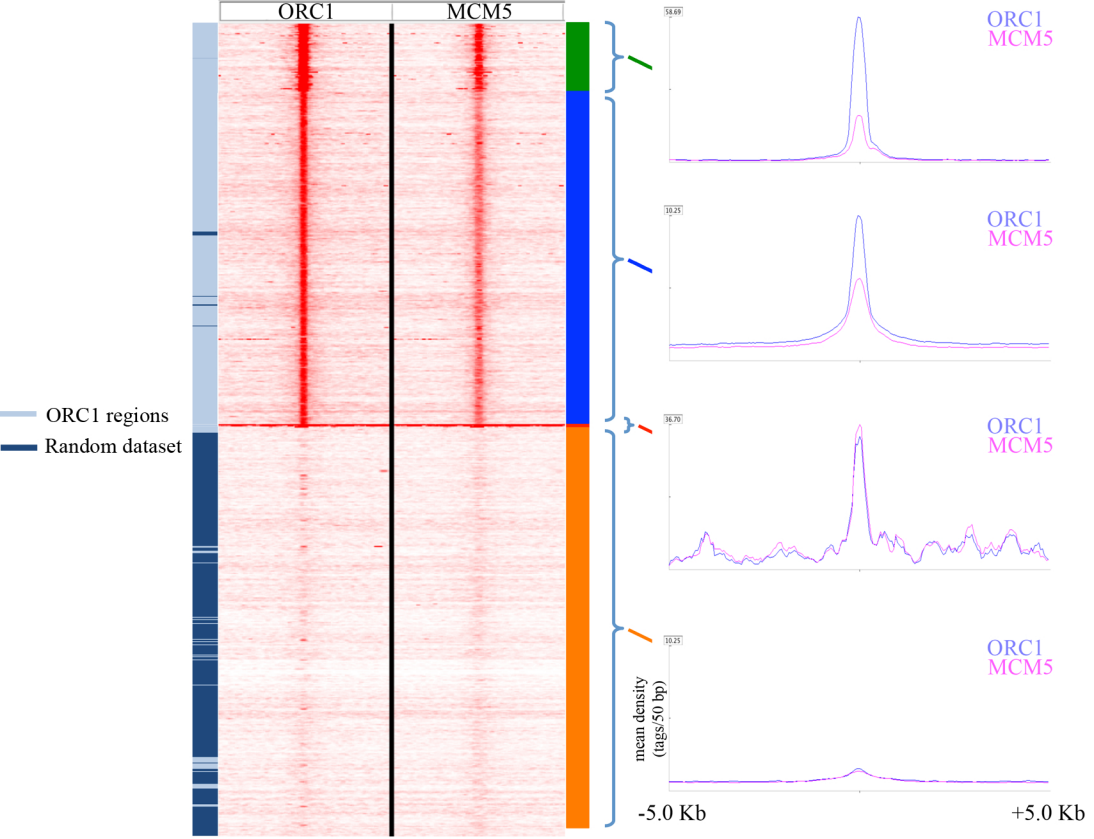
Supplemental Figure S9



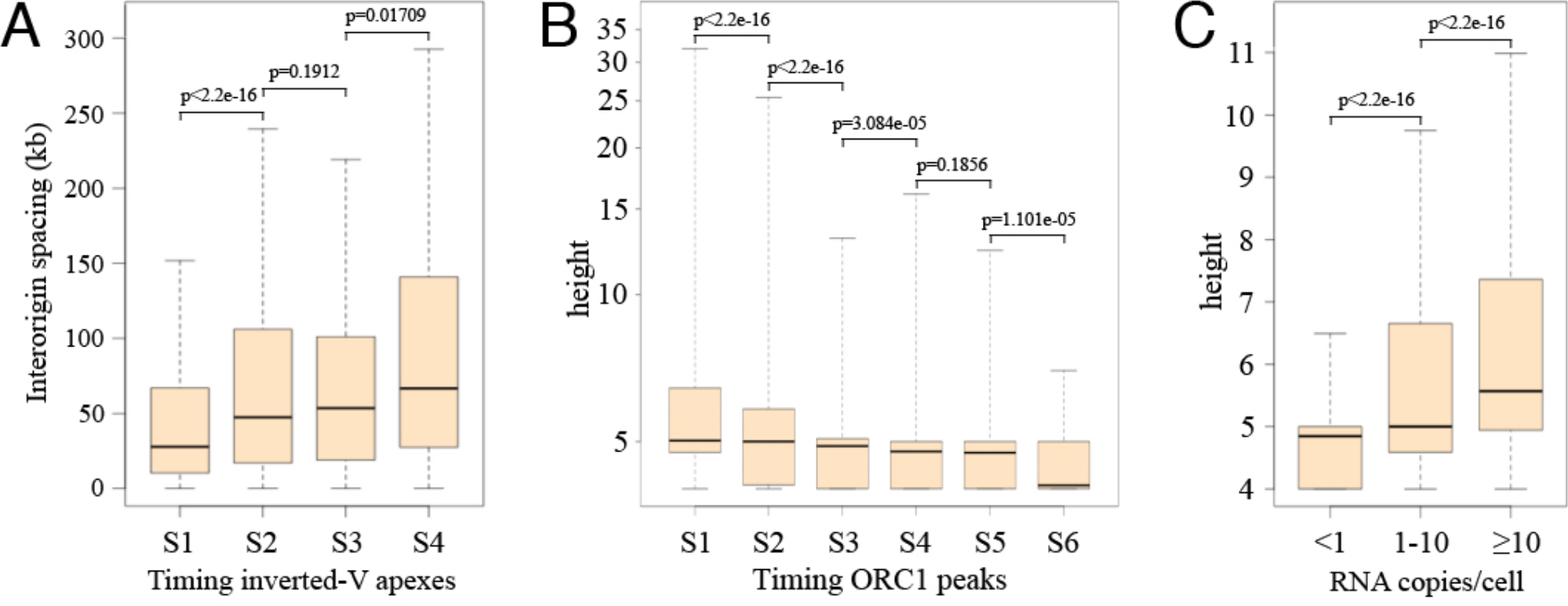
Supplemental Figure S10



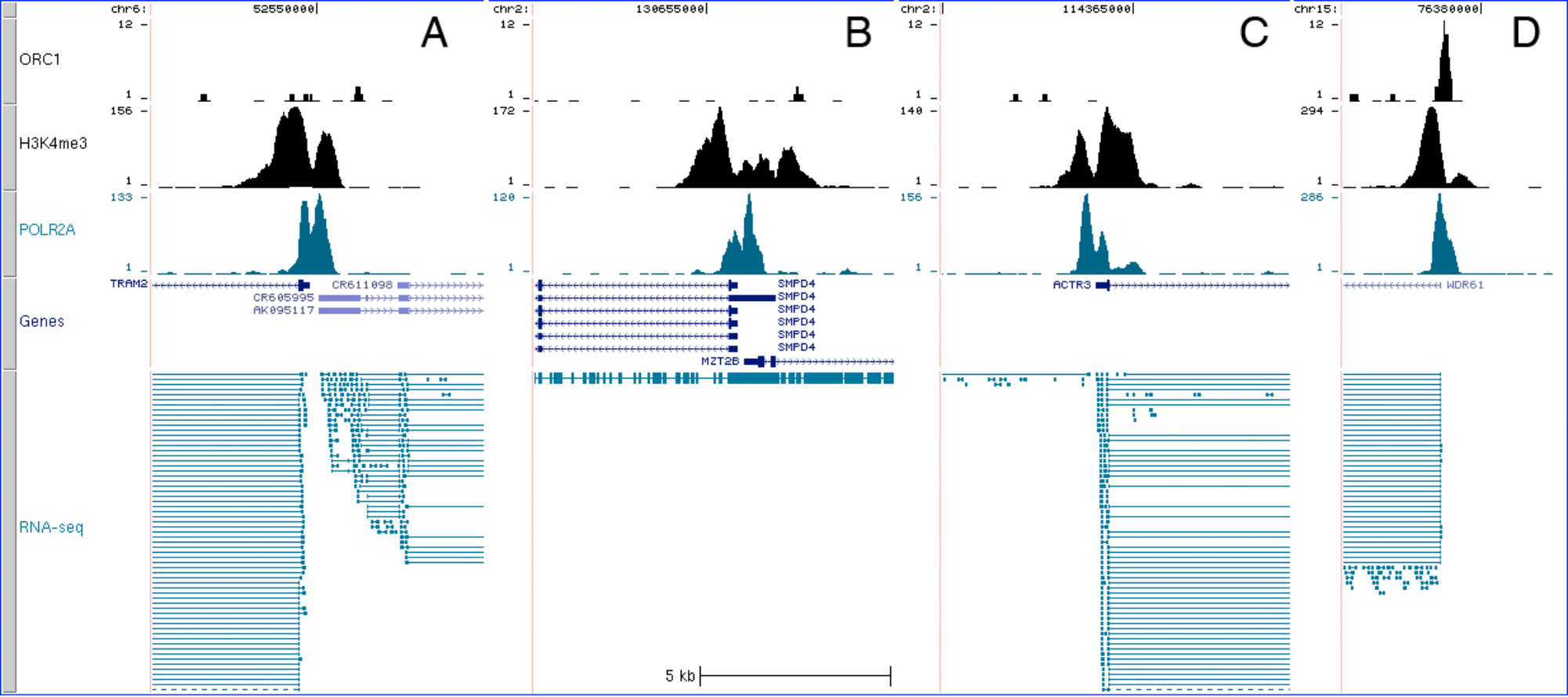
Supplemental Figure S11



Supplemental Figure S12



Supplemental Figure S13



Supplemental Figure S14

Processing steps	HelaS3_INPUT_Exp #1	HelaS3_INPUT_Exp #2	HelaS3_ORC1_Exp #1	HelaS3_ORC1_Exp #2
raw reads	11,723,152	11,662,848	10,367,284	10,583,902
aligned reads with duplicates	9,365,313	9,028,301	7,617,047	7,982,089
aligned reads w/o duplicates	8,756,061	8,359,592	2,638,288	2,735,996

Supplemental Table S1

Timing		Inverted-V apexes				ORIs			
		Total	%	with ORIs	%	Total	%	within inverted-V apexes	%
Early	S1	568	25.8%	546	96.1%	4097	30.1%	3541	86.4%
	S2	743	33.7%	655	88.2%	4433	32.6%	3245	73.2%
Middle	S3	360	16.3%	269	74.7%	2068	15.2%	1035	50.0%
	S4	304	13.8%	164	53.9%	1365	10.0%	466	34.1%
Late	S5	229	10.4%	92	40.2%	918	6.7%	267	29.1%
	S6	-	-	-	-	721	5.3%	-	-
Total		2204	100.0%	1726	78.3%	13602	100%	8554	66.4%

Supplemental Table S2

Genomic position	All		RNA-seq$^+$						RNA-seq$^-$					
			All		TSS-seq$^+$		TSS-seq$^-$		All		TSS-seq$^+$		TSS-seq$^-$	
Proximal promoters	4794	35.2%	4560	95.1%	4377	96.0%	183	4.0%	234	4.9%	138	59.0%	96	41.0%
Other positions	8810	64.8%	5248	59.6%	3247	61.9%	2001	38.1%	3562	40.4%	1031	28.9%	2531	71.1%
Intragenic	4272	31.4%	3198	74.9%	2302	72.0%	896	28.0%	1074	25.1%	475	44.2%	599	55.8%
Intergenic	4538	33.4%	2050	45.2%	945	46.1%	1105	53.9%	2488	54.8%	556	22.3%	1932	77.7%
Total	13604	100.0%	9808	72.1%	7624	77.7%	2184	22.3%	3796	27.9%	1169	30.8%	2627	69.2%

Supplemental Table S3

Genomic position	RNA-seq+						
	Transcription levels	All		TSS-seq+		TSS-seq-	
Proximal Promoter	≥5	2476	54.3%	2436	55.7%	40	21.9%
	<5	2084	45.7%	1941	44.3%	143	78.1%
	Total	4560	100.0%	4377	100.0%	183	100.0%
Other positions	≥5	1049	20.0%	878	27.0%	171	8.5%
	<5	4199	80.0%	2369	73.0%	1830	91.5%
	Total	5248	100.0%	3247	100.0%	2001	100.0%
Intragenic	>5	932	29.1%	793	34.4%	139	15.5%
	<5	2266	70.9%	1509	65.6%	757	84.5%
	Total	3198	100.0%	2302	100.0%	896	100.0%
Intergenic	>5	117	5.7%	85	9.0%	32	2.9%
	<5	1933	94.3%	860	91.0%	1073	97.1%
	Total	2050	100.0%	945	100.0%	1105	100.0%
All peaks		9808		7624		2184	

Supplemental Table S4

Available datasets in HeLa cells		Exp. Conditions		Intersections (%)			
Studies	Nr of ORIs/*ORI-containing regions*	Cell type	Method	Dellino et al.	Cadoret et al.	Karnani et al.	Mesner et al.
Dellino et al. (whole genome)	229	HeLa S3	α-ORC1 ChIPseq	-	29.3	10.9	46.7
Cadoret et al. (ENCODE)	282	HeLa S3	SNS (λ digestion)	23.0	-	13.8	35.0
Karnani et al. (ENCODE)	320	HeLa	SNS (λ digestion)	7.8	12.8	-	26.6
Mesner et al. (ENCODE)	*657*	HeLa	Bubble trapping	20.2	14.0	10.7	-

Supplemental Table S5

	S1	S2	S3	S4	S5	S6	TOT
Timing of 50 kb windows within ENCODE regions	123	222	85	52	71	90	643
Timing of ORI Cadoret et al. (ENCODE only)	95	117	34	16	14	6	282
Timing of ORI-containing regions Mesner et al. (ENCODE only)	297	233	51	33	16	26	656
Timing of ORC1 peaks within ENCODE regions	97	87	26	9	5	5	229
ORC1 peaks common to Mesner et al.	59	38	6	3	0	1	107
% of ORC1 peaks common to Mesner et al.	60.8%	43.7%	23.1%	33.3%	0.0%	20.0%	46.7%
ORC1 peaks common to Cadoret et al.	39	24	3	1	0	0	67
% of ORC1 peaks common to Cadoret et al.	40.2%	27.6%	11.5%	11.1%	0.0%	0.0%	29.3%

Supplemental Table S6

Section 2
Supplemental Figure/Table Legends

Figure S1. Chromatin features of known ORIs. *Top track:* sequencing profile of DNA fragments from high-density fractions at *LMNB2* and *PRKDC* ORIs. *Lower tracks*: i) ChIP-seq profile of K79me2, K4me3 and POLR2A; ii) sequencing profile of DNA fragments from DNase I digestion (DNase hypersensitive - DH - sites); iii) RefSeq genes, as shown in the UCSC Genome Browser.

Figure S2. Chromatin features of genomic regions enriched in high-density fractions. Visualization of (bottom to top): RefSeq genes; ChIP-seq profiles of K79me2, K4me3 and POLR2A; DH sites and sequencing profile of DNA fragments from high-density fractions, within 100 kb from human chr. 11, as shown in the UCSC Genome Browser.

Figure S3. Western blot showing proteins co-immunoprecipitated by the anti-ORC1 antibodies used in ChIP-seq assays. The anti-ORC1 (and rabbit IgG) antibodies were cross-linked to Protein G sepharose beads prior to standard immune-precipitation assays on HeLa nuclear extracts. The immunoprecipitated proteins were resolved by SDS-PAGE prior to visualization of ORC1 (positive control) and other Pre-RC subunits by immunoblotting with the indicated antibodies (the same as in Fig. 1C, plus anti-ORC6: upstate 05-938).

Figure S4. ORC1 binding to known ORIs. (*A*) QChIP data of ORC1 binding to the *PRKDC* and *LMNB2* ORIs, using unfractionated (blue) or low-density (red) chromatin. Fold enrichment of ORC1 binding as compared to the corresponding flanking regions (*LMNB2* -1 kb and *PRKDC* +5 kb) is shown. (*B*) ChIP-seq profile of ORC1 at the *HPRT1* and *DBF4* ORIs (Y-axis value range is as in Fig. 1F). Scale bars indicate 5 kb.

Figure S5. Overlap between two independent anti-ORC1 ChIP-seq experiments. (*A*) Venn diagram showing the overlap of the ORC1 peaks identified in the two replicates. Experiment #1 was analysed and discussed in details. (*B*) Boxplot representation of amplitude (height) of common and unique ORC1 peaks identified in

the two replicates.

Figure S6. ChIP-seq profile of ORC1 peaks with small amplitude. I, J and K ORC1 peaks (same scale as in Figures 1F and 2A) and position of ORI probes used for NSA assay (Figure 2D), as shown in the UCSC Genome Browser. Distances from ORI are in kb. Scale bar indicates 2 kb.

Figure S7. Replication timing of regions containing validated ORIs. Replication patterns of genomic regions containing the ORC1 peaks validated by NSA (red vertical lines): A and B ORIs (mapping within visually identified very early-replicating *inverted-V apexes*); ORI I (early); ORI J (early-middle); ORI K (late). RefSeq genes are shown.

Figure S8. Distribution of ORC1 sites mapping within proximal promoters of RefSeq/UCSC genes (distance from TSS is in bp).

Figure S9. Scatter plot. The correlation between RPKM values and the highest number of overlapping RNA tags measured at the 7,011 ORC1-sites for which RPKM values were available in the UCSC Genome Browser, is shown (see Methods).

Figure S10. Chromatin features of POLR2A^{neg} ORIs. ChIP-seq profiles of ORC1, H3K4me3 and POLR2A at newly identified POLR2A^{neg} ORIs are shown in the UCSC Genome Browser. Y-axis value ranges, where not indicated, are: 1-12 (ORC1); 1-30 (POLR2A and H3K4me3). RefSeq genes and RNA-seq data are as of March 2011.

Figure S11. Genomic distribution of ORC1 and MCM5 sequence tags (screenshots). Visualization of (bottom to top): ChIP-seq profiles of MCM5 and ORC1 of U937 cells (and RefSeq genes) within representative regions of human chr. 19 (350 kb-1.8 Mb wide), as shown in the UCSC Genome Browser.

Figure S12. Genomic distribution of ORC1 and MCM5 sequence tags (heat map). Heat map representation of the signal densities of ORC1 and MCM5 observed within 7,608 regions, half of which contained high-confidence (p<0.01 and h≥10)

ORC1 binding sites (±5 kb from the ORC1 peak summit) and half were randomly chosen around ORC1-free TSSs – containing regions (±5 kb from TSS; "random dataset"). The density map was subjected to clustering in order to identify groups of loci sharing similar ORC1 and MCM5 genomic profiles. The two datasets segregate in two main groups: cluster "1+2+3" (green, blue and red boxes), which contains nearly all (>95%) the identified ORC1 binding sites (showing identical distribution of ORC1 and MCM5 sequenced tags), and cluster "4" (orange box), containing ORC1-free TSSs, in which both ORC1 and MCM5 are virtually absent.

Figure S13. Inter-origin spacing within inverted-V apexes showing different replication timing and amplitude of ORC1 peaks showing different replication timing or transcription levels. Boxplot representations of: (*A*) inter-origin spacing measured within inverted-V apexes showing different replication timing, as indicated; (*B*) amplitude (height) at ORC1 peaks replicating in the S1 to S6 windows; (*C*) amplitude (height) of ORC1 peaks showing low (<1), moderate (1-10) or high (≥10 RNA copies/cell) transcription levels.

Figure S14. ORC1^{neg} TSSs of active genes. Representative ChIP-seq profiles of ORC1, POLR2A and H3K4me3 at four active TSSs without (*A*-*C*) or with (*D*) ORC1 binding. The RNA-seq compacted track in (*B*) indicates a very highly expressed gene (as viewed in the UCSC browser). The active TSS bound by ORC1 in (*D*) is the same as in Figure 4B. RefSeq genes and RNA-seq data are as of March 2011.

Table S1. Numbers of DNA sequencing reads. Numbers of total reads (raw reads) and of aligned reads prior (with duplicates) and after (without duplicates) filtering out duplicate reads (potential PCR artifacts), in two anti-ORC1 ChIP-seq experiments in HeLa cells.

Table S2. Replication timing of ORC1 peaks and inverted-V apexes. Columns 3-6 (inverted-V apexes): total number and percentage of *inverted-V apexes*, and number and percentage of *inverted-V apexes* containing at least one ORC1-site, for each S-phase window (S1 to S5). Columns 7-10 (ORIs): total number and percentage of ORC1 sites (ORIs, S1 to S6), and number and percentage of ORC1 sites within the *inverted-V apexes*, for each S-phase window (S1 to S5). Since S6 *inverted-V apexes*

cannot be identified, the 721 S6 ORIs were excluded from the analysis of ORIs mapping within all the *inverted-V apexes*.

Table S3. Association of ORC1 sites with functional TSSs. Number and percentage of ORC1 peaks mapping within or outside proximal promoters associated, or not, with RNA-seq tags and TSSs from the TSS-seq dataset (Yamashita et al. 2011).

Table S4. Transcription levels measured at RNA-seq$^+$ ORC1-sites showing different genomic locations. Number and percentage of RNA-seq$^+$ ORC1-peaks (≥5 or <5 RNA copies/cell, as indicated) mapping within or outside proximal promoters and associated, or not, with TSSs from the TSS-seq dataset (Yamashita et al. 2011).

Table S5. Overlap between available ORI datasets in HeLa cells. Percentage of overlap (within 2.5 kb from the ORC1 and the nascent-strand peaks) between datasets of ORIs identified in different genome-wide studies of HeLa cells, using different methods. In blue: overlap between previously published datasets obtained with the same approach; in red: overlap between ORC1 peaks identified in the ENCODE regions and the previously published datasets of ORIs (Cadoret et al. 2008; Karnani et al. 2010) or ORI-containing regions (Mesner et al. 2011).

Table S6. Replication timing of common ORIs. Number, percentage and replication timing of the common ORIs identified in this study, in Cadoret et al. 2008 and in Mesner et al. 2011, within the ENCODE regions. Replication timing (S1-S6) in HeLa cells of the non-overlapping 50 kb windows contained within the ENCODE regions is also shown.

Section 3

Supplemental Methods (Statistical analyses) and Supplemental References

1) **Significance of the association between ORC1 peaks and POLR2A, H3K4me3 or DH sites.** Datasets of RefSeq genes, DNase Hypersensitive sites, RNA Polymerase II and TSS-seq were obtained from (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=214617233&c=chr8&g=refGene), ENCODE Univ. Washington DNaseI Hypersensitivity by Digital DNaseI at UCSC (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=214617233&c=chr8&g=wgEncodeUwDnaseSeq), (http://genome.ucsc.edu/cgi-bin/hgFileSearch?hgsid=222594973&db=hg18&hgt_tsDelRow=&hgt_tsAddRow=&tsName=&tsDescr=&tsGroup=Any&fsFileType=fastq&hgt_mdbVar1=cell&hgt_mdbVal1=HeLa-S3&hgt_mdbVar2=lab&hgt_mdbVal2=Yale&hgt_mdbVar3=antibody&hgt_mdbVal3=Pol2&hgfs_Search=search) and Solexa Tag Data at dbtss V7.0 of (http://dbtss.hgc.jp; Yamashita et al. 2011), respectively. TSS-seq Solexa tag mapping data identifying TSS clusters (TSCs) of 12 different human samples were pooled, sorted and made not redundant keeping the highest scoring record using a custom script. For all the statistical analyses testing the enrichment of specific features associated to the ORC1 sites, namely the intersection of the ORC1 peaks with POLR2A, H3K4me3 or DH sites (Table 1), we simulated 3,000 random permutations of the ORC1 sites. A one-sided Wilcoxon signed-rank test (alternative hypothesis: ‘‘greater’’) was used to compare the mean of the random distribution with the actual observed value of overlapping events.

2) **Boxplots**. Statistical significance of differences among distributions shown in boxplots (two-sample test) was assessed by the non-parametric Wilcoxon signed-rank test, due to the absence of normality verified with Kolmogorov–Smirnov goodness-of-fit test. 2x75 nt raw signal file alignment of the RNA-seq experiment from HeLaS3 cells produced as part of the ENCODE Project were downloaded from UCSC site *(http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=213836427&c=chr5&g=wgEncodeCaltechRnaSeq)*; using a

custom script, we combined the information of the original files and the intervals of the ORC1 peaks to compute, for each ORC1-site, the value corresponding to the highest number of overlapping RNAseq tags within ±2.5 kb from the peak summit.

3) **Heat maps** were generated using seqMINER (Version 1.3; Ye et al. 2011).

4) **Evaluation of expression levels at the ORC1 sites**. The RPKM values of the Caltech experiment used in the correlation analysis were obtained from the "encodeDCC/wgEncodeCaltechRnaSeq/" repository at the UCSC site (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeCaltechRnaSeq/). For 7,011 ORC1 peaks we could match the number of underlying reads and the RPKM values obtained from Caltech. We performed linear regression in loglog space between RPKM and the corresponding number of overlapping reads; in order to consider also RPKM values=0, log(1+x) was used instead of log(x), while loglog space was adopted to better discriminate low values. We used the regression line to interpolate the number of overlapping reads (n=5.53) corresponding to RPKM=1 (R^2 =0.769) (Figure S6). RPKM=2 corresponds to approximately 1 RNA copy/cell (Mortazavi et al. 2008).

5) **Significance of the occurrence of ORC1 peaks within *inverted-V apexes***. We simulated peak occurrences by randomizing their positions. Each 50 kb window of the genome was assigned an S label (S1 to S6) according the its s50 values (S1: $0<s50\leq0.167$, etc., see main text). Each ORC1 peak was then assigned the same label as the 50 kb window containing its middle point. Each of the six groups of ORC1 peaks (S1 to S6) was randomized 1,000 times within regions with the same s50 label (i.e. S1 peaks were randomized only within S1 windows) using shuffleBed (Quinlan and Hall 2010). The 6 groups of random peaks were first merged in order to obtain 1,000 lists of random peaks with the same replication status as the 13,604 ORC1 peaks and then intersected with the 2,317 s50 *inverted-V apexes*. The Normal behavior was assessed (μ=8,409.865, σ=47.662) and finally compared to the number of observed interactions (8,641).

6) **Significance of the occurrence of *inverted-V apexes* containing ORC1-sites**. We calculated the probability that the number of *inverted-V apexes* with ≥1 peak would be found by chance. To this end, we produced 3,000 randomizations of two groups of inverted-V apexes ("all the *inverted-V apexes*" and "S1-*inverted-V apexes*") using shuffleBed utility (Quinlan and Hall 2010). In order to produce the best simulation for the "S1-*inverted-V apexes*", we shuffled only S1- to S3-replicating 50 kb windows.

Simulation of "all the *inverted-V apexes*" was sampled using the whole genome. We assessed significance of the enrichment by comparing distributions of simulated results with the observed number of *inverted-V apexes* containing at least one ORC1-site and applying a Wilcoxon signed-rank test using R (R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/).

Supplemental References

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5: 621-628.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6): 841-842.

Yamashita R, Sathira NP, Kanai A, Tanimoto K, Arauchi T, Tanaka Y, Hashimoto S, Sugano S, Nakai K, Suzuki Y. 2011. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome research* 21(5): 775-789.

Ye T, Krebs AR, Choukrallah MA, Keime C, Plewniak F, Davidson I, Tora L. 2011. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic acids research* 39(6): e35.