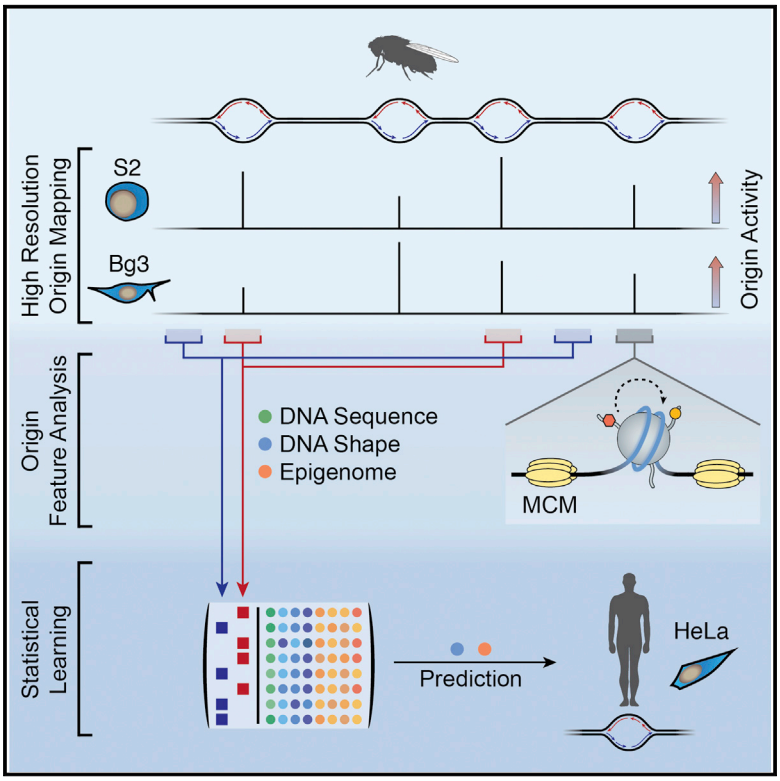
Resource

Cell Reports

High-Resolution Profiling of *Drosophila* Replication Start Sites Reveals a DNA Shape and Chromatin Signature of Metazoan Origins

Graphical Abstract



Authors

Federico Comoglio, Tommy Schlumpf, ..., Christian Beisel, Renato Paro

Correspondence

renato.paro@bsse.ethz.ch

In Brief

By mapping thousands of *Drosophila* replication origins at high resolution, Comoglio et al. identify DNA shape and specific chromatin features as predictive marks for active origins in the fly and human genomes. Differential origin activity across cell types mirrors cell-type-specific transcriptional programs.

Highlights

- *Drosophila* replication start sites are mapped in two cell types at high resolution
- Origin-proximal G-quadruplexes act as replication fork barriers in vivo
- DNA shape and chromatin configurations mark and predict metazoan origins
- Differential origin activity mirrors cell-type-specific transcriptional programs

Accession Numbers

GSE65692

Comoglio et al., 2015, Cell Reports *11*, 821–834
May 5, 2015
http://dx.doi.org/10.1016/j.celrep.2015.03.070

Cell Reports

Supplemental Information

High-Resolution Profiling of *Drosophila* Replication Start Sites Reveals a DNA Shape and Chromatin Signature of Metazoan Origins

Federico Comoglio, Tommy Schlumpf, Virginia Schmid, Remo Rohs, Christian Beisel, and Renato Paro

This Supplementary file accompanies the manuscript:

Comoglio F, Schlumpf T, Schmid V, Rohs R, Beisel C and Paro R.
High-resolution profiling of *Drosophila* replication start sites reveals a DNA shape and chromatin signature of metazoan origins

This file contains:

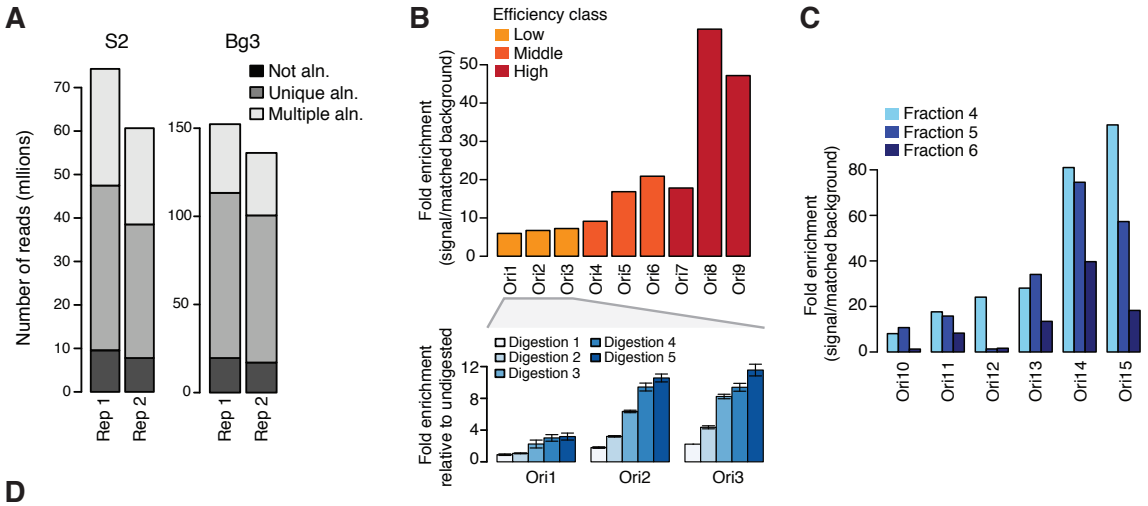
(1) Supplemental figures and legends (for Figures S1-S7, associated with Figures 1-7, respectively).

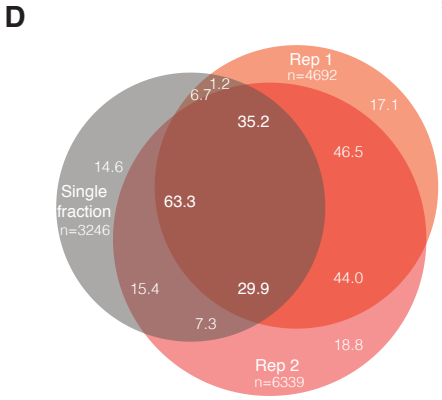
(2) Supplemental tables and legends (for Tables S1-S2, associated with Experimental Procedures and Figures 5-7, respectively).

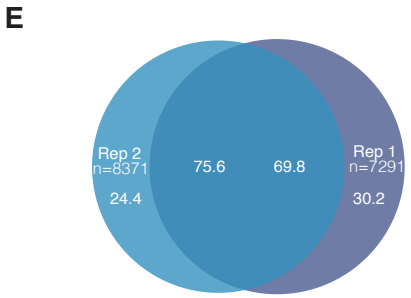
(3) Supplemental experimental procedures (details of experiments and computational analyses).

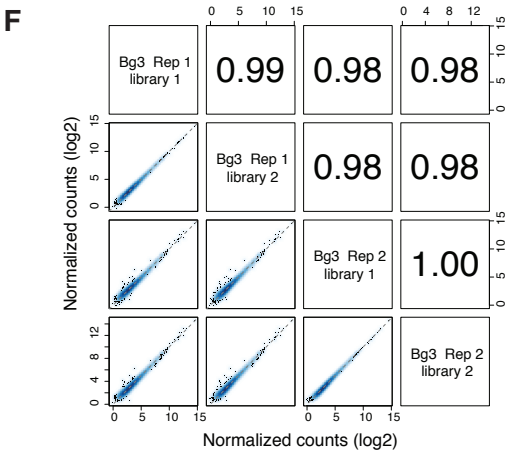
(4) Additional supplemental references

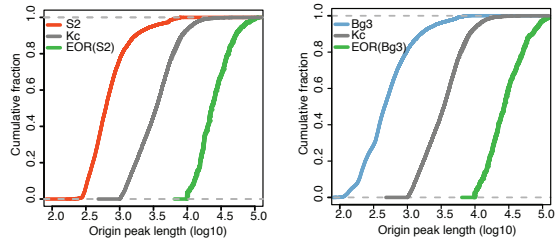
SUPPLEMENTAL FIGURES AND LEGENDS

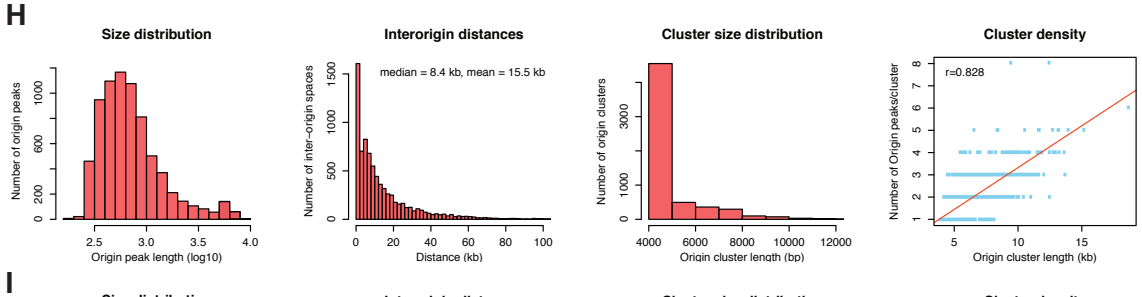












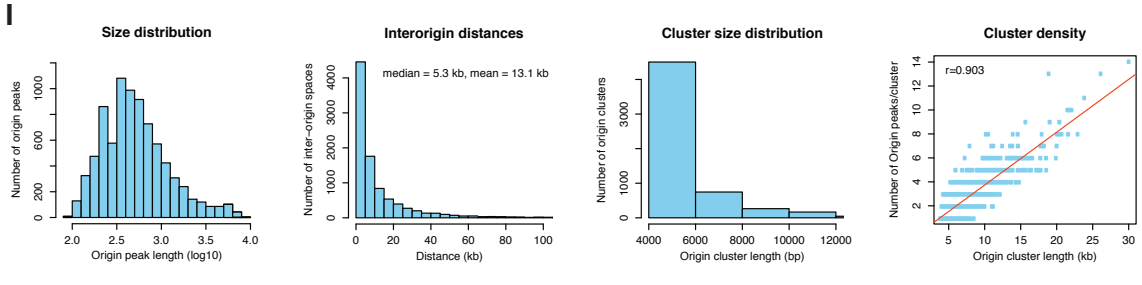


Figure S1. Consistency between SNS-Seq replicates, SNS-qPCR validation and qualitative analysis of origin peaks (related to Figure 1). (A) Short read alignment summary of SNS-Seq data from individual biological replicates (Rep1, Rep2) in S2 and Bg3 cells. (B) SNS-qPCR validation of SNS-Seq enrichments. Origins were partitioned in three efficiency classes based on quantiles of efficiency values estimated from S2 SNS-Seq data. The qPCR-based fold enrichments at three representative origins for each efficiency class are shown. Fold enrichments were computed with respect to a background genomic region located proximally to the origin peak. The inset shows the qPCR-based fold enrichments relative to undigested material after each round of Lexo digestion for the indicated origins from an independent biological replicate. Results are means $\pm$ s.d. of two measurements. See Table S1 for primer sequences. (C) Same as (B), where origin peaks were detected in single-fraction SNS-Seq experiments and qPCR carried out on Fractions 4-6. The six origin peaks are located in the 175 kb genomic region illustrated in Figure 1E and were independently detected in all fractions. See Table S1 for primer sequences. (D) Percentage overlap of origin peaks identified within each S2 SNS-Seq biological replicate (Rep1, Rep2). The overlap with origin peaks identified by all fractions from single fraction SNS-Seq experiments (Single fraction) is also shown for comparison (see also Figure S3). All detected peaks were considered for this analysis. (E) Percentage overlap of origin peaks identified within each Bg3 SNS-Seq biological replicate. All detected peaks were considered for this analysis. (F) Pairwise correlations of Bg3 SNS-Seq read counts within the union of origin peaks identified in individual biological replicates. Read counts were normalized to library sizes. (G) Empirical cumulative distribution functions of origin peak lengths for the indicated sets of replication origins. EOR, early origin regions. (H) Size distribution of S2 origin peaks, cluster size distribution, cluster density and interorigin distances. For the analysis of origin clusters, origin peaks were resized to 4 kb windows centered on RSSs. (I) Same as (H) for Bg3 origin peaks. Correlation values (r) are Pearson's correlation coefficients.

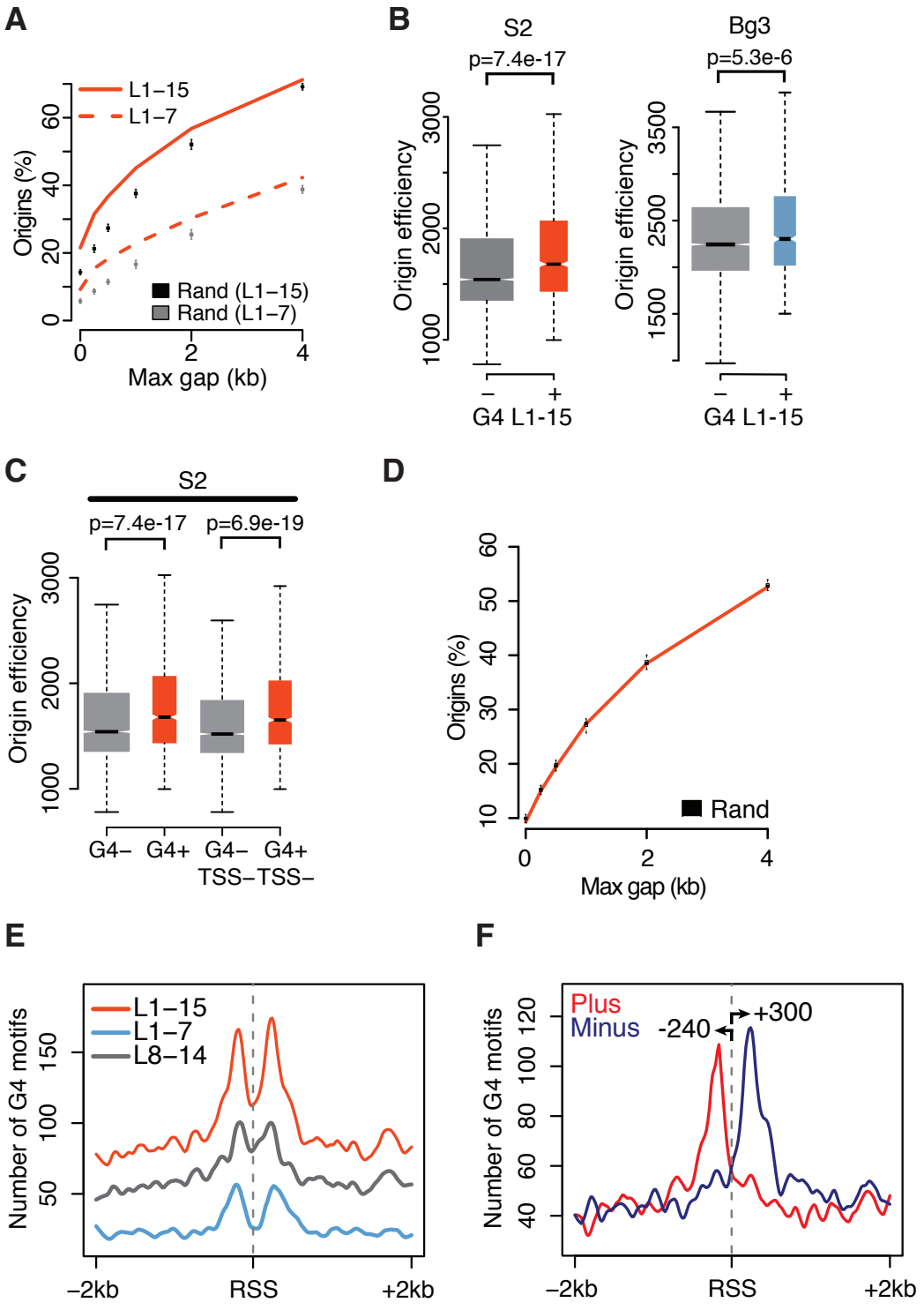


Figure S2. Additional properties of G4-associated origins (related to Figure 2). (A) Percentage overlap of S2 origin peaks with predicted G4 motifs and comparison of observed overlaps (lines) with random expectations (boxplots). (B) Efficiency of S2 and Bg3 origins associated (+) or not associated (-) with G4 L1-15 motifs. (C) Same as (B), but with origins further partitioned

by association with Transcription Start Sites (TSSs). (D) Percentage overlap of S2 origin peaks with TSSs and comparison of observed overlaps (line) with random expectations (boxplots). (E) Spatial distribution of G4 motifs within $\pm$2 kb of Bg3 RSSs. (F) Same as (E), for strand-specific annotation of G4 L1-15 motifs. Arrows indicate distances (bp) from the RSS. *p*-values are from Wilcoxon rank-sum test.

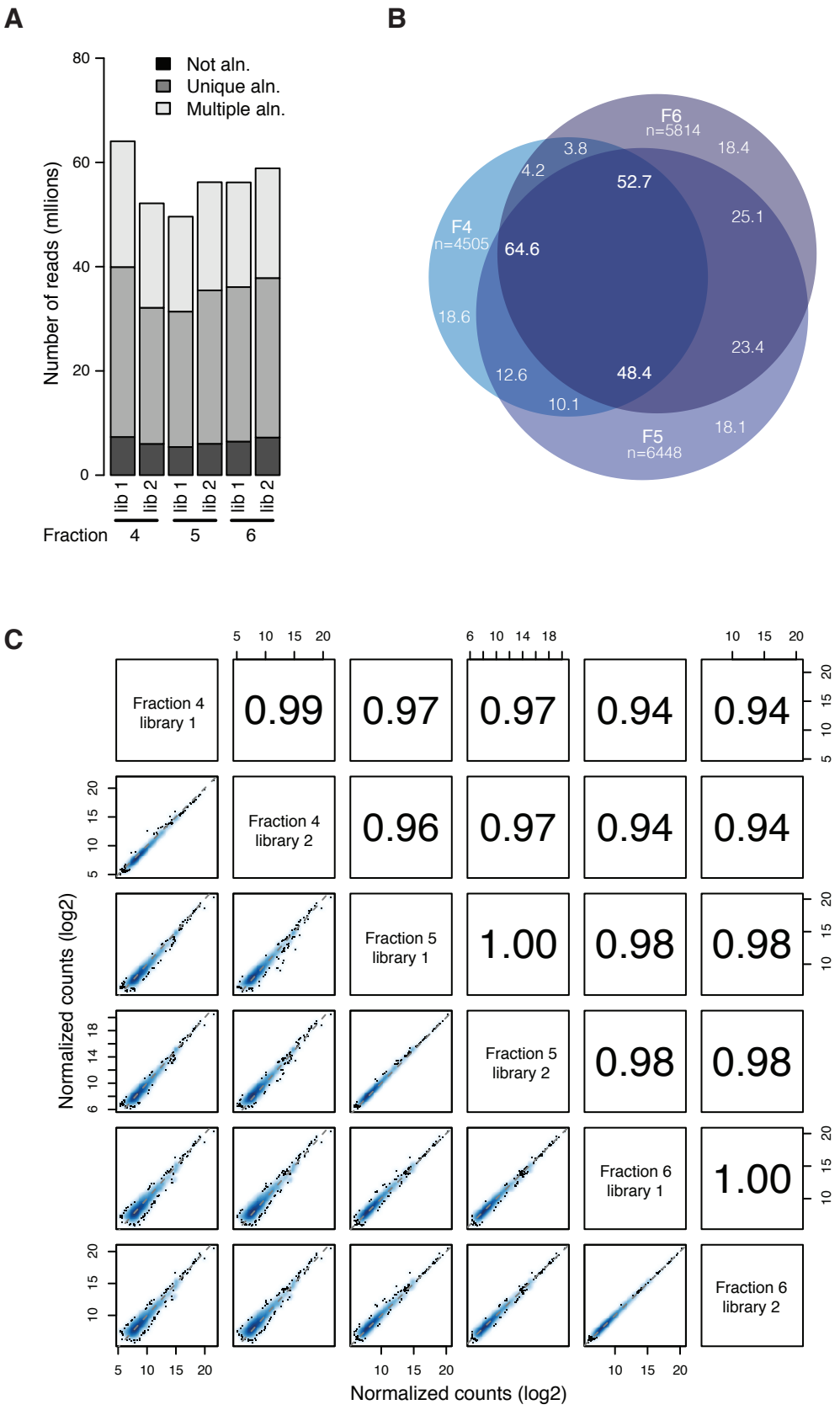

A
Not aln.
Unique aln.
Multiple aln.
Number of reads (mllions)
0
20
40
60
80
lib 1
lib 2
lib 1
lib 2
lib 1
lib 2
Fraction
4
5
6
B
F6
n=5814
F4
n=4505
F5
n=6448
3.8
18.4
4.2
52.7
25.1
64.6
18.6
23.4
12.6
48.4
10.1
18.1
C
Fraction 4 library 1
Fraction 4 library 2
Fraction 5 library 1
Fraction 5 library 2
Fraction 6 library 1
Fraction 6 library 2
0.99
0.97
0.97
0.94
0.94
0.96
0.97
0.94
0.94
1.00
0.98
0.98
0.98
1.00
Normalized counts (log2)
Normalized counts (log2)

Figure S3. Correspondence between fractions in single-fraction SNS-Seq experiments (related to Figure 3). (A) Short read alignment summary of single fraction SNS-Seq data from individual fractions and technical replicates (lib) in S2 cells. (B) Percentage overlap of origin peaks identified within each fraction (F4-F6) from single fraction SNS-Seq data. (C) Pairwise correlations of single fraction SNS-Seq read counts within the union of origin peaks identified in each fraction. Read counts were normalized to library sizes. Correlation values are Pearson's correlation coefficients.

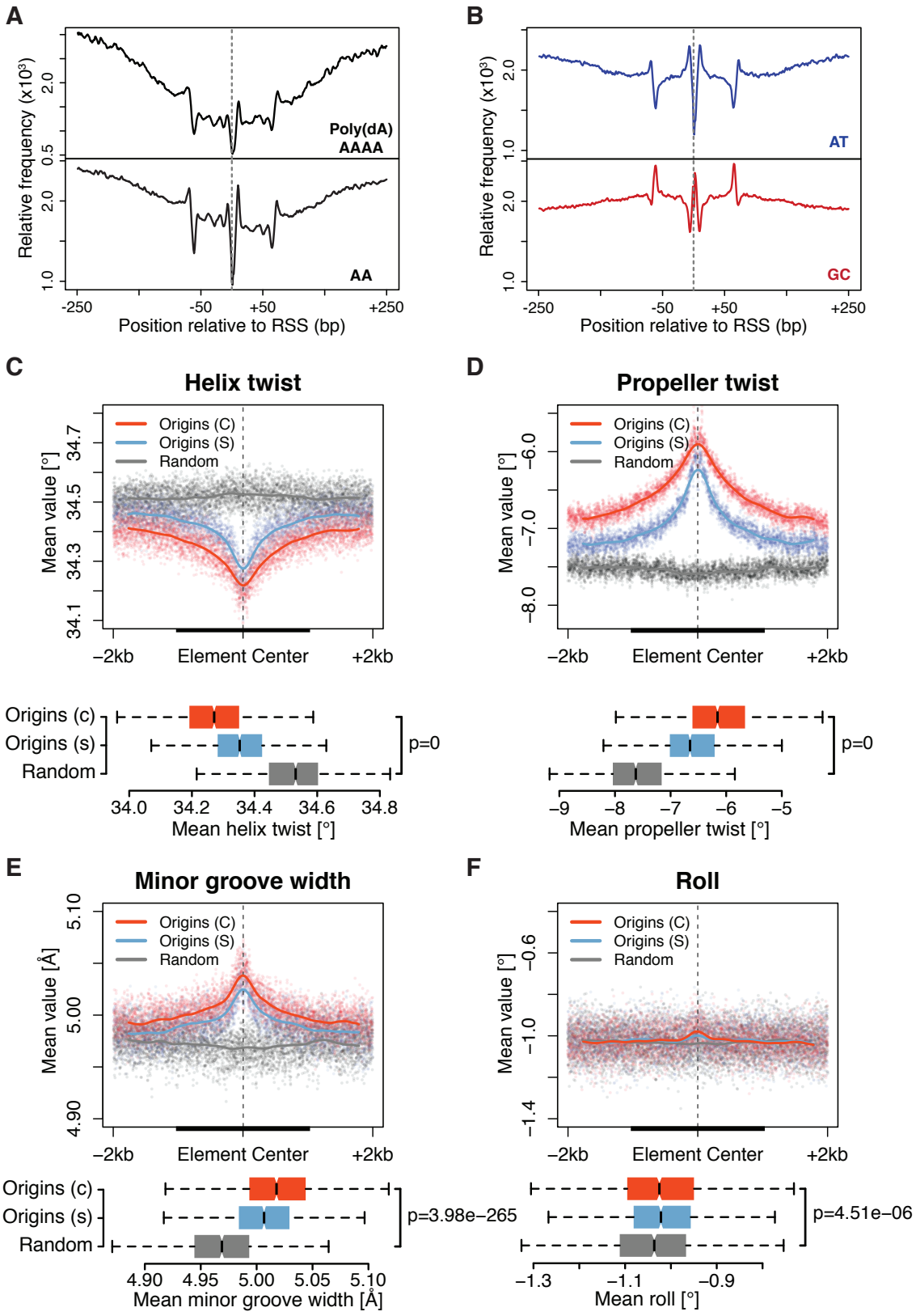

A
Relative frequency (x10³)
Poly(dA)
AAAA
AA
-250
-50
+50
+250
Position relative to RSS (bp)
B
Relative frequency (x10³)
AT
GC
-250
-50
+50
+250
Position relative to RSS (bp)
C
Helix twist
Mean value [°]
Origins (C)
Origins (S)
Random
−2kb
Element Center
+2kb
Origins (c)
Origins (s)
Random
p=0
Mean helix twist [°]
D
Propeller twist
Mean value [°]
Origins (C)
Origins (S)
Random
−2kb
Element Center
+2kb
p=0
Mean propeller twist [°]
E
Minor groove width
Mean value [Å]
Origins (C)
Origins (S)
Random
−2kb
Element Center
+2kb
Origins (c)
Origins (s)
Random
p=3.98e−265
Mean minor groove width [Å]
F
Roll
Mean value [°]
Origins (C)
Origins (S)
Random
−2kb
Element Center
+2kb
p=4.51e−06
Mean roll [°]

Figure S4. Nucleosome container and DNA shape features at human replication origins (related to Figure 4). (A) Relative frequency of AAAA polynucleotides and AA dinucleotides within $\pm$250 bp of HeLa RSSs. (B) Same as (A) for AT and GC dinucleotides. (C-F) Average of DNA shape features within $\pm$2 kb of RSSs for constitutive (C) and HeLa-specific (S) origins, and background regions. Solid lines are Loess fitted curves from single-nucleotide resolution shape predictions (dots). Boxplots of average feature values within 1 kb windows (thick black lines) are shown (bottom panels). *p*-values are from Wilcoxon rank-sum test.

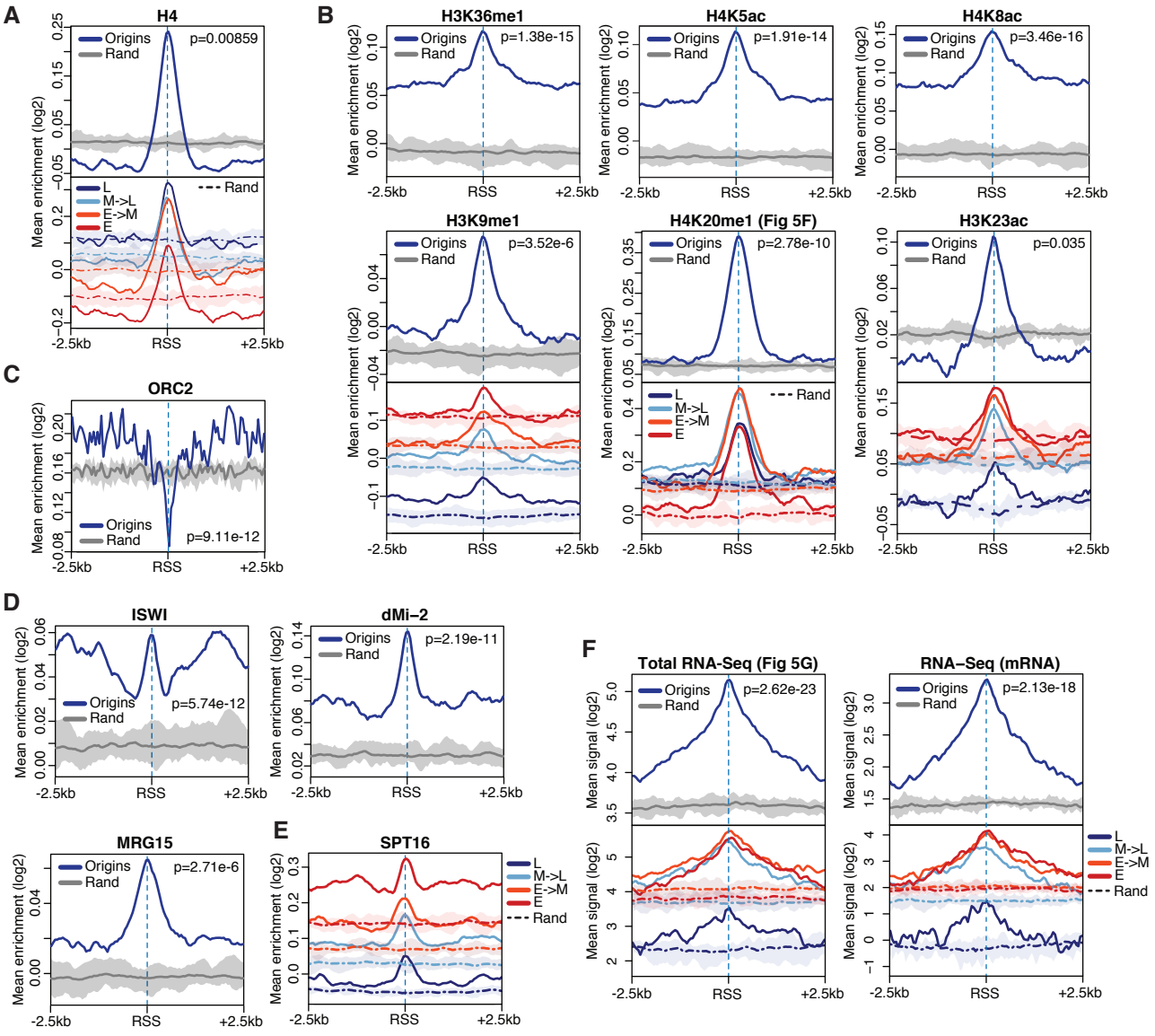


Figure S5. Absence of timing-specific origin signatures and spatial distribution of additional chromatin features at RSSs (related to Figure 5). (A) Average H4 enrichment (top) within ±2.5 kb of S2 RSSs and within ten sets of randomized genomic regions (Rand). The thick gray line traces average background values. Bottom panels show further partitioning of the signal above in four timing classes (L: late S-phase; M: mid; E: early) based on replication timing quartiles. (B-F) Same as (A) for the indicated features. Reference to Figure 5 panels (main text) that are reproduced here for sake of clarity is provided. *p*-values are from Wilcoxon rank-sum test.

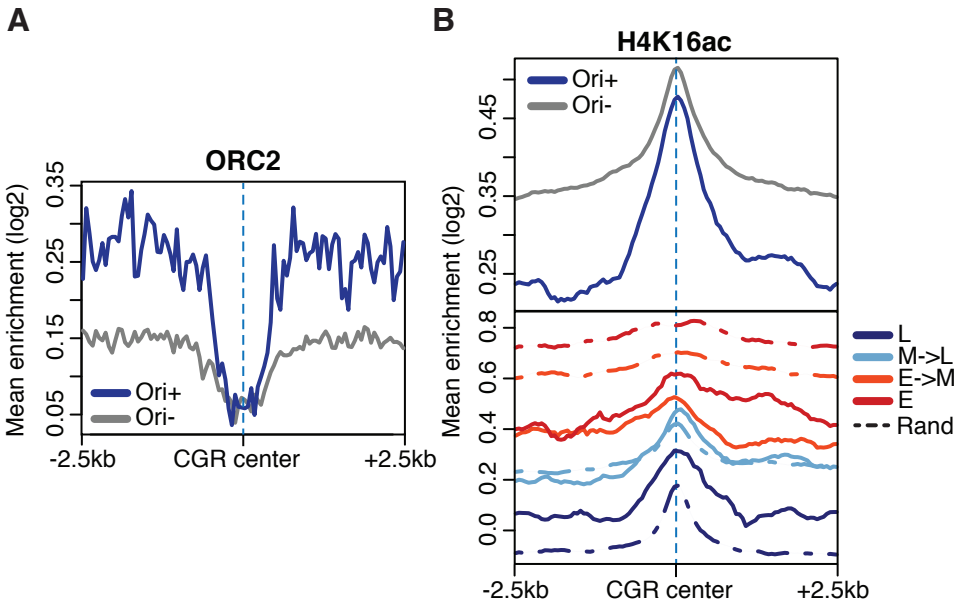


Figure S6. Spatial distribution of ORC2 and timing partitioning of the H4K16ac mark at CGRs (related to Figure 6). (A) Average ORC2 enrichment within ±2.5 kb of CGR midpoints for origin-CGRs (Ori+) and origin-negative CGRs (Ori-). (B) Average H4K16ac enrichment (top) within ±2.5 kb of S2 origin-CGR midpoints (Ori+) or of origin-negative CGRs (Ori-). Bottom panels show further partitioning of the signal above in four timing classes (L: late S-phase; M: mid; E: early) based on replication timing quartiles.

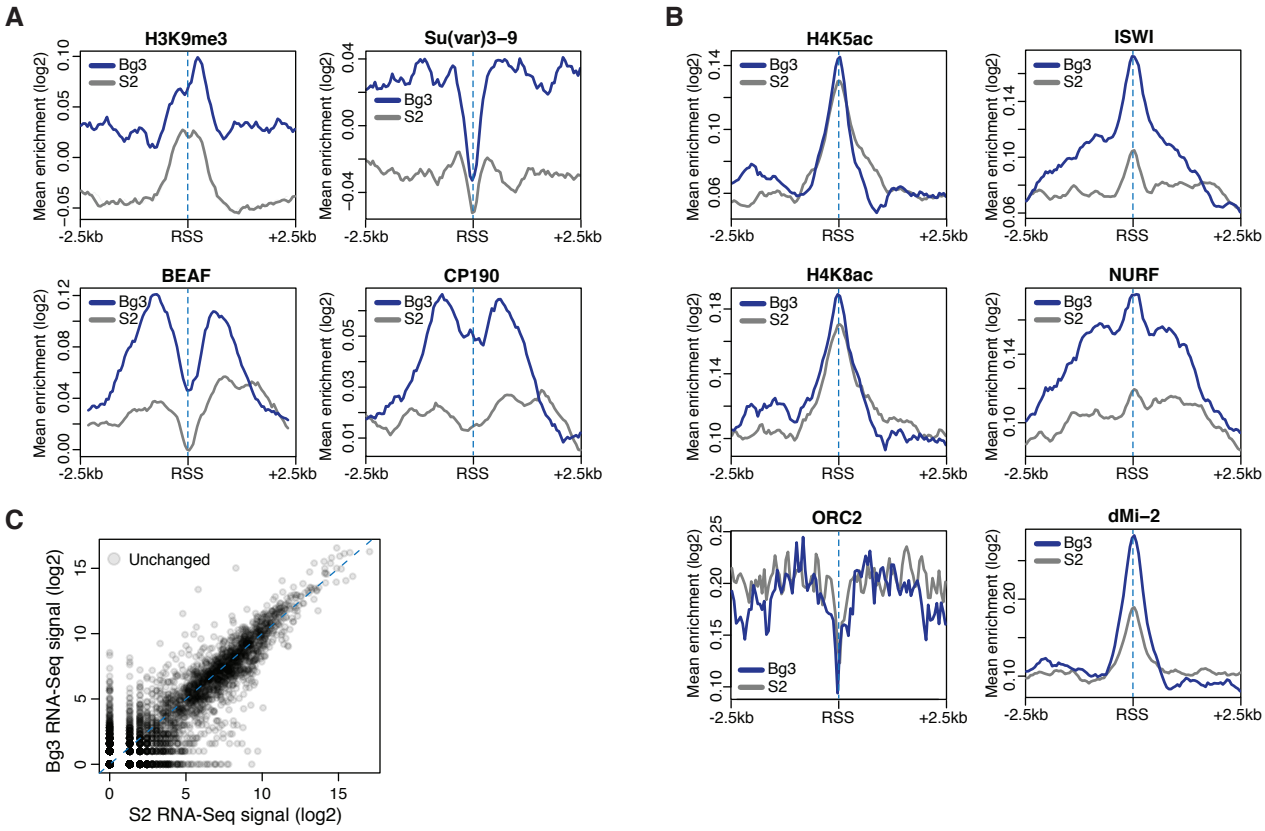


Figure S7. Spatial distribution of heterochromatin features, insulators and additional chromatin features at Bg3-preferred origins (related to Figure 7). (A) Average enrichments of heterochromatin marks and insulator proteins within ±2.5 kb of RSSs of origin peaks solely identified in Bg3 cells (blue) and comparison with all S2 origin peaks (gray). (B) Same as (A) for the indicated features. (C) Scatter plot of S2 and Bg3 RNA-Seq signals at origins that were not significantly differentially activated (unchanged) in S2 and Bg3 cells.

SUPPLEMENTAL TABLES AND LEGENDS

Name of the genomic region	Primer sequence	Tested on (P: pool; S: single fractions)
Ori 1	ACCACATTTTGACTTGGGTTTAC	P
	CCTTGGGATTTCGGGATCTGT	P
Background 1	TGAATCAACTGAACTGCGTTAAA	P
	CCATTTGGCTTCACATCCCATTT	P
Ori 2	CAGGGCCATAAACAGGGGAA	P
	TGACTGCCAAGAAGAGCCAG	P
Background 2	ACACAAACACAACTCACCTGG	P
	AACGTGCTTGGCTTTCTGTG	P
Ori 3	GCAGAGACATGATCGCCGAA	P
	TTGTTCATTGGGTGCATGTG	P
Background 3	ACGTACAACACAACCGACGA	P
	CAGGTTACAGGCCAAGCAAAA	P
Ori 4	CACAAAGTGCAGGTAGTCGC	P
	ATGCCCAGGAGATAGCCCTT	P
Background 4	AAGAGCAACAACAATGGCGG	P
	GGCCAGTCCCAGTGTATAACG	P
Ori 5	CATCGGGAGGAAGCCACTTT	P
	AAGAGATGATCGAGCCAGGC	P
Background 5	TTCCCCAGTCTCCACTCTCC	P
	CCAGCGTGTTTTCTGCGTTT	P
Ori 6	GCCGAGTCAAATGCCGAATG	P
	GTGTCCTATTGTTCGTTTGTTGA	P
Background 6	GGCGGACTGACTTAGGATCG	P
	TCACAGTACGCCCCAAAAGC	P
Ori 7	CAGTCATTGGCTGCACTCTT	P
	ACCCAAAATGCCTCGTCGAT	P
Background 7	GAAGTACGCAATGGAGCAGC	P
	AGACCTGGGATTTTGGAGCG	P
Ori 8	TCGTATGCCTGGAACGAACG	P
	GCAATGTGGTTATGGCTAGCA	P
Background 8	GCTCAGTAGTGCGTTCTTGC	P
	CGTTTCAAAGTCGCTCCATTCA	P
Ori 9	TCGTGTCGATGGAGTGGTG	P
	AACATGAGCTGGTCCGCTAC	P
Background 9	GGCATCGTCTCCATCAGTCA	P
	AGTCGGAAGTCTGCCAAGTT	P
Ori 10	CATTCCGCCGCTTGTTGTTC	S
	GTTGTAAAGGAAAGCGCCGA	S
Background 10	GGCTCATTCTATTGCCGTGG	S
	TTTTAACAGCCCAGCAGACG	S
Ori 11	AGCTCAACTACTCATGAACGCA	S
	TGTTTGGCAGGAACGGTTCA	S
Background 11	TCAAAGCAAATGTCGCCTGC	S
	CTGAAGTGGATCGCAGTGGT	S

Name of the genomic region	Primer sequence	Tested on (P: pool; S: single fractions)
Ori 12	TGGGTTACGTGGTTTGGCAT	S
(Background 11)	TGGCCAAATGAGCAGCCTTA	S
Ori 13	TGGCCGTGTTCTTAAGCGAT	S
(Background 11)	GTTATGGGGTCGGTGGATGG	S
Ori 14	GCTCACTCCGTTGCAAATCC	S
	CTCTGGCCGAAACATGTGTG	S
Background 12	GCGCATCAATTCGGTCAGTC	S
	CCGTAACCATACCCATCCCG	S
Ori 15	GATCGTTCATGGGCGAAACG	S
(Background 12)	AGTCTTCCAGCAGCGAATCC	S

Table S1. Primer used for the SNS-qPCR validation of origin peaks (related to Experimental Procedures). This table provides the sequence of the primers used to validate origin peaks from standard as well as single-fraction SNS-Seq experiments in S2 cells. Matched background regions are indicated.

Feature name	Symbol	Source	Identifier
absent, small, or homeotic discs, C-ter domain 1	Ash1C	GEO	GSE30820 (Kockmann et al., 2013)
absent, small, or homeotic discs, N-ter domain 1	Ash1N	GEO	GSE30820 (Kockmann et al., 2013)
Boundary element-associated factor of 32 kDa	BEAF	modENCODE	modENCODE 274
CG10630	CG10630	modENCODE	modENCODE 4176
Chromator	Chro	modENCODE	modENCODE 278
Centrosomal protein of 190 kDa	CP190	modENCODE	modENCODE 925
CCCTC-binding factor	CTCF	modENCODE	modENCODE 283
dMi-2	dMi-2	modENCODE	modENCODE 3676
Sex combs extra	dRING	modENCODE	modENCODE 928
Scm-related gene containing four mbt domains	dSFMBT	modENCODE	modENCODE 3751
Enhancer of Zeste	E(z)	modENCODE	modENCODE 2988
female sterile (1) homeotic L	FshL	GEO	GSE30820 (Kockmann et al., 2013)
female sterile (1) homeotic SL	FshSL	GEO	GSE30820 (Kockmann et al., 2013)
FK506-binding protein 2	fkbp		(Chen, Y. et al., in preparation)
Suppressor of variegation 205	HP1a	modENCODE	modENCODE 2668
Heterochromatin Protein 1b	HP1b	modENCODE	modENCODE 941
Heterochromatin Protein 1c	HP1c	modENCODE	modENCODE 943
rhino	HP1d	modENCODE	modENCODE 4187
Heterochromatin Protein 2	HP2	modENCODE	modENCODE 944
Heterochromatin Protein 4	HP4	modENCODE	modENCODE 3948
Heat shock protein 83	Hsp90	GEO	GSE31226 (Sawarkar et al., 2012)
Imitation SWI	ISWI	modENCODE	modENCODE 3032
Suppressor of variegation 3-1	JIL-1	modENCODE	modENCODE 945
Lysine (K)-specific demethylase 2	Kdm2	modENCODE	modENCODE 3033
Histone demethylase 4A	Kdm4A	modENCODE	modENCODE 3784
MBD-R2	MBD-R2	modENCODE	modENCODE 946
Minichromosome maintenance complex component 2-7	MCM	SRA	SRP006146
maleless	MLE	modENCODE	modENCODE 3788
modifier of mdg4	MOD2.2	modENCODE	modENCODE 2674
MRG15	MRG15	modENCODE	modENCODE 3047
male-specific lethal 1	MSL1	modENCODE	modENCODE 3293
Enhancer of bithorax	NURF	modENCODE	modENCODE 947
Origin Recognition Complex subunit 2	ORC2	SRA	SRP002091
Polycomb	Pc	GEO	GSE24521 (Enderle et al., 2011)
Polycomblike	PCL	modENCODE	modENCODE 3049
Polyhomeotic	Ph	GEO	GSE24521 (Enderle et al., 2011)
pleiohomeotic	Pho	modENCODE	modENCODE 3894
Painting of fourth	POF	modENCODE	modENCODE 3294
RNA polymerase II 215 kD subunit	PolII	modENCODE	modENCODE 329
PR-Set7	PR-Set7	modENCODE	modENCODE 3054
Posterior sex combs	Psc	GEO	GSE24521 (Enderle et al., 2011)

Feature name	Symbol	Source	Identifier
RPD3	RPD3	modENCODE	modENCODE 3057
dre4	SPT16	modENCODE	modENCODE 3058
Suppressor of Hairy wing	Su(Hw)	modENCODE	modENCODE 330
Suppressor of variegation 3-3	Su(var)3-3	modENCODE	modENCODE 3949
Suppressor of variegation 3-7	Su(var)3-7	modENCODE	modENCODE 2672
Suppressor of variegation 3-9	Su(var)3-9	modENCODE	modENCODE 2673
Topoisomerase II	TopoII	modENCODE	modENCODE 3945
Trithorax-like	Trl	modENCODE	modENCODE 285
Trithorax	Trx	GEO	GSE24521 (Enderle et al., 2011)
will die slowly	WDS	modENCODE	modENCODE 953
deformed wings	ZW5	modENCODE	modENCODE 3804
Histones	H1	modENCODE	modENCODE 3300
and	H2Av	modENCODE	modENCODE 2991
histone modifications	H2BK5ac	modENCODE	modENCODE 3283
	H2BUb	modENCODE	modENCODE 290
	H3	modENCODE	modENCODE 3301
	H3	GEO	GSE51598 (Bohla et al., 2014)
	H3K4me1	modENCODE	modENCODE 304
	H3K4me2	modENCODE	modENCODE 2655
	H3K4me3	modENCODE	modENCODE 305
	H3K9ac	modENCODE	modENCODE 309
	H3K9acS10P	modENCODE	modENCODE 2660
	H3K9me1	modENCODE	modENCODE 3770
	H3K9me2	modENCODE	modENCODE 3011
	H3K9me3	modENCODE	modENCODE 313
	H3K18ac	modENCODE	modENCODE 292
	H3K23ac	modENCODE	modENCODE 294
	H3K27ac	modENCODE	modENCODE 296
	H3K27me1	modENCODE	modENCODE 3943
	H3K27me2	modENCODE	modENCODE 3000
	H3K27me3	modENCODE	modENCODE 298
	H3K36me1	modENCODE	modENCODE 3170
	H3K36me3	modENCODE	modENCODE 303
	H3K79me1	modENCODE	modENCODE 2658
	H3K79me2	modENCODE	modENCODE 307
	H4	modENCODE	modENCODE 3304
	H4acTetra	modENCODE	modENCODE 201
	H4K5ac	modENCODE	modENCODE 321
	H4K8ac	modENCODE	modENCODE 322
	H4K16ac	modENCODE	modENCODE 320
	H4K20me1	modENCODE	modENCODE 3014
Others	DNase-Seq	GEO	GSE40739 (Arnold et al., 2013)
	MNase-Seq	GEO	GSE51599 (Bohla et al., 2014)
	RNA-Seq poly(A)+	GEO	GSE24521 (Enderle et al., 2011)
	Total RNA-Seq	GEO	GSE55872 (Ashwal-Fluss et al., 2014)
	RNA-Seq poly(A)+, Bg3	modENCODE	modENCODE 5954
	Replication timing, S2	GEO	GSE17280 (Eaton et al., 2011)
	Replication timing, Bg3	GEO	GSE17281 (Eaton et al., 2011)
	RNA-Seq poly(A)+, HeLa	ENCODE	(ENCODE Project Consortium, 2012)

Table S2. Detailed information on the genome-wide profiles included in the analysis (related to Figures 5-7). This table provides details, source and accession numbers of all *Drosophila* and human data included in the analysis.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

SNS Purification

S2 and Bg3 cells were seeded in 11 x 145 mm tissue culture plates (Greiner) at a density of 1.5 x 10^6 cells/ml 18-24 hours prior to SNS purification. Six sucrose gradients were prepared by carefully layering 1.6 ml each of 6 separate pre-cooled sucrose solutions (30%, 25%, 20%, 15%, 10% and 5% of sucrose w/v with RNase free Sucrose - Sigma 84097 in TEN300 - 10 mM Tris-HCl 7.9, 2 mM EDTA, 300 mM NaCl in RNase free water - Gibco 10977 hereinafter: water) in an ultracentrifuge tube (Beckman and Coulter 14x89 mm 331372) submerged in liquid nitrogen. Gradients were thawed overnight at 4°C prior to use, to generate linearity through limited mixing between layers. One additional gradient was prepared to assess linearity using an analog refractometer.

Exponentially growing S2 cells were washed twice with PBS (Gibco 10010023) and one 145 mm plate was covered in 5 mL of DNAzol (Invitrogen 10503-027). After cell lysis for 3 minutes at room temperature (RT), the DNAzol was transferred to the next plate and the process repeated for all plates. Proteinase K (Roche 03115844001) was added to the lysate/DNAzol mixture to a final concentration of 200 μg/ml and incubated at 37°C for 2 hours. Precipitation of the genomic DNA (gDNA) was then carried out following the DNAzol user manual. The dried pellet was dissolved in 1 ml of TEN20 (10 mM Tris-HCl 7.9, 2 mM EDTA, 20 mM NaCl, 0.1% SDS, 500 U RNasin - Promega N2511) at 70°C. DNA does not dissolve above concentrations of 300-400 ng/μl and when required, additional 500 μl of TEN20 were added to maximize solubilization. After dissolving, DNA was heated to 90°C for 15 minutes and snap chilled in ice-water. Five gradients were loaded at 4°C with a maximum of 100 μg (generally 30 μg) of gDNA each. One gradient was loaded with 5 μg of DNA ladder (Thermo Scientific - SM0314) diluted to the same volume as the gDNA aliquots. The gradients were then transferred to a pre-cooled SW-41 Ti rotor (Beckman-Coulter) and centrifuged at 4°C and 26700 rpm (121000 rcf) for 21 hours.

Fractions were collected by carefully pipetting 1 ml off the top of the gradient at 4°C. Quality check and selection of fractions for downstream processing was performed by taking 150 μl aliquots from each fraction of one gDNA gradient as well as from the ladder gradient. DNA was precipitated by adding 1/25 volume of 5 M NaCl and 2.5 volumes of 100% Ethanol, rotation at 4°C for 10 minutes and storage at -80°C for 30 minutes to overnight. The aliquots were subsequently spun at 21000 rcf and 4°C for 30 minutes, the supernatant removed and the pellet washed once with 500 μl cold 70% Ethanol in water. After air drying, DNA was resuspended in loading buffer and size separated on an agarose gel. Fractions containing DNA fragments of sizes between 500 and 2500 bases were identified, pooled and precipitated as described above. For single fraction experiments, matched fraction numbers were pooled across gradients. Air-dried pellets were resuspended in 85 μl of water with 100 U RNasin. 10 μl were kept for qPCR analysis whereas, to the 75 μl, 10 μl T4 PNK 10x buffer (Fermentas EK003) were added, the sample incubated at 90°C for 5 minutes and snap chilled in ice-water. 10 μl of 10 mM ATP (NEB P0756S) and 2 μl of T4 PNK (Fermentas EK003) were added and samples were incubated at 37°C for 1 hour, followed by enzyme inactivation at 75°C for 15 minutes. DNA was precipitated as described above except for addition of 1/10 volume of 5 M NaCl. The air-dried pellets were resuspended in a mix of 77 μl water, 2.5 μl - 100 U RNasin, 10 μl 10x Exonuclease Buffer, 0.5 μl 100x BSA (NEB B9000S) and 10 μl - 100 U lambda Exonuclease (Fermentas EN0562) and incubated overnight to 18 hours at 37°C. After addition of

100 μl water and 200 μl of PCI (Applichem A0889), the samples were mixed and centrifuged at 21000 rcf for 5 minutes. The upper phase was transferred and precipitated as specified above. T4 PNK phosphorylation, exonuclease digestion, phenol extraction and precipitation were repeated 4 to 5 times for each sample. Aliquots of digested and undigested DNA were run on a 1% agarose gel to assess digestion efficiency.

To prepare DNA for sequencing, pellets from exonuclease digestion were resuspended in 100 μl TE buffer (10 mM Tris-HCl 8.0, 1 mM EDTA) with 1 μl of RNase A (Roche, DNase free 11579681001). Digested DNA was extracted with Phenol-chloroform and precipitated as described above. Pellets were resuspended in a mix of 68 μl water, 8 μl 10x Reaction buffer and 4 μl second strand synthesis enzyme mix (NEB NEBNext E6111S) and incubated at 16°C for 2.5 hours. 0.7 Volumes of Agencourt AMPure XP beads (Beckman-Coulter A63880) were added, briefly vortexed and incubated for 5 minutes at RT. The beads were collected on a magnetic rack and washed twice for 30 seconds with 70% Ethanol. Following the second wash, beads were air-dried, resuspended in 40 μl of water and DNA was transferred to a new tube and subjected to library preparation. DNA libraries were prepared with the Nextera XT DNA Sample Preparation Kit (Illumina) following manufacturer's instructions. Each library was sequenced for 51 cycles in a single-end run on the HiSeq 2000 or 2500 system (Illumina) with SBS v3 and v4 chemistry, respectively.

Computational Methods

Annotations

Genomic coordinates of Kc origins (Cayrou et al., 2011) were retrieved from the DeOri database (Gao et al., 2012; http://tubic.tju.edu.cn/deori/). S2 and Bg3 early origin regions (Eaton et al., 2010) were retrieved from the modMine data warehouse (Contrino et al., 2012). These sites were mapped by BrdU immunoprecipitation from a synchronized cell population upon stalling replication forks by treatment with hydroxyurea. G-quadruplexes (G4) motif occurrences in the *Drosophila* genome were predicted with QuadParser (Huppert and Balasubramanian, 2005) with consensus sequence `G3+ N1-15 G3+ N1-15 G3+ N1-15 G3+`. Three classes of G4 motifs, L1-15, L1-7 and L8-14, were then defined based on the G4 loop length. Transcription Start Sites (TSSs) coordinates were extracted from Ensembl gene annotations. CG-rich regions in the *Drosophila* genome were predicted according to (Gardiner-Garden and Frommer, 1987). DNase I hypersensitive sites in S2 cells were obtained from (Arnold et al., 2013).

SNS-Seq data processing

S2 and Bg3 SNS-Seq reads were quality filtered and aligned to the dm3 *Drosophila* reference genome using Bowtie2 version 2.2.0 (Langmead et al., 2012) allowing one mismatch in seed alignment (`-N 1 -L 22`). The output was converted from SAM format to BAM using samtools version 0.1.19 (Li et al., 2009), and BAM files were sorted and indexed. Human SNS-Seq data from IMR-90, ESC h9 and HeLa cells (Besnard et al., 2012) were mapped to the hg19 human reference genome. As the polarity of leading strands at origins is expected to generate a bimodal enrichment pattern centered on the replication start site, origin peaks were called using MACS

version 1.4.1 (Zhang et al., 2008) with parameters `-g dm -m 5,30` for each replicate. Saturation of sequencing depth was evaluated using down-sampling and indicated that <40% of the sequenced reads sufficed to detect the entire S2 and Bg3 origin repertoire. Pairwise correlations between read counts normalized to library sizes were evaluated for each sequencing library in the union of origin peaks from individual biological replicates (see Figure S1D). The set of origin peaks for each cell type was then defined as the union of identified peaks across replicates. Only peaks supported by unique alignments were considered further. For single fraction SNS-Seq data, technical replicates for each fraction were pooled prior to peak calling. Constitutive origins were defined as the intersection of S2, Bg3 and Kc (Cayrou et al., 2011) origin peaks.
All subsequent analysis were performed using R version $\geq$ 3.1.0 (R Core Team, 2014; http://www.R-project.org/) and BioConductor (Gentleman et al., 2004; http://www.bioconductor.org) packages with custom scripts, which are available upon request. BigWig coverage tracks of individual replicates and pooled libraries were generated at single base pair resolution using wavClusteR 2.0 (Comoglio et al., 2015). For each origin peak, the coverage function from pooled libraries was used to infer the position of the replication start site (RSS), which was set to the peak summit. Read summarization within origin peaks was performed with the GenomicRanges package (Lawrence et al., 2013). For each origin, the weighted average of the read counts for each biological replicate normalized to the size of the peak expressed in kb was used as a proxy for its firing efficiency in a cell population. Weights were set to relative library sizes.
The replication timing of origins was computed as the average normalized smoothed M-value of tiling array probesets mapping within the origin peak region.

RNA-Seq, DNase-Seq and MNase-Seq data processing

RNA-Seq reads were aligned to the dm3 *Drosophila* reference genome using Tophat2 version 2.0.11 allowing one mismatch and reporting only one alignment for reads that map to multiple locations (`-N 1 -g 1`). MNase-Seq, DNA-Seq and input samples were dumped to fastq format using the NCBI Short Read Archive Toolkit version 2.3.1. Reads were aligned to the dm3 *Drosophila* reference genome using Bowtie2 and output files were converted from SAM format to BAM, sorted and indexed as described above.

Background regions and randomization procedure

Matched background regions were used to i) estimate SNS-Seq noise levels; ii) determine whether observed pairwise overlaps between replication origins across cell types or association of origin peaks with distinct genomic features (e.g. DHSs, G4, CG-rich regions) occurred significantly more frequently than random expectation; iii) computing control metaprofiles of chromatin features. Given an origin set, a matched background set was obtained by randomly sampling an equal number of genomic regions of the same length as the origin peaks from the mappable regions of the *Drosophila* genome. To account for the observed genomic distribution of replication origins, the number and sizes of origin peaks for each chromosome were matched by sampling an identical number of background regions exhibiting the same length as the origin peaks therein.
For analysis of pairwise overlaps across cell types or association of origin peaks with genomic features, 1000 background sets were generated. The number of intervals overlapping by at least one base pair, or separated by no more than 250, 500, 1000, 2000 and 4000 base pairs was com-

puted. Statistical significance of the overlap between two features was determined by computing the empirical *p*-value from the sampling distribution of the background set overlaps. To compute control metaprofiles of chromatin features, 10 background sets were used.
The replication timing of background regions was computed as described above.

Origin DNA sequence content analysis

DNA sequences of S2 and HeLa origins were extracted from the dm3 *Drosophila* reference genome and the hg19 human reference genome, respectively. The relative frequency of AAAA polynucleotides and AA, AT and GC dinucleotides was computed at single-nucleotide resolution in 500 bp windows centered on RSSs.

DNA shape features

DNA sequences of *Drosophila* constitutive origins ($n = 1286$) were extracted from the dm3 reference genome in 2 kb windows centered on the RSSs. DNA sequences from an equally sized random sample of background regions, and from a random sample of 500 TSSs were also retrieved. For human origins, the DNA sequences of a random sample ($n = 2000$) of constitutive origins, HeLa-specific origins (i.e. origin peaks solely identified in HeLa SNS-Seq data) and background control regions were extracted from the hg19 human reference genome in 4 kb windows centered on the RSSs and on the range midpoint, respectively.
DNA shape features were derived from the high-throughput approach DNAshape, which is based on all-atom Monte-Carlo predictions (Zhou et al., 2013; http://rohslab.cmb.usc.edu/DNAshape/). The four DNA shape features used in this study were helix twist, propeller twist, minor groove width and roll.

Lasso logistic regression

Logistic regression models are a class of probabilistic classifiers widely applied to binary classification problems in several scientific areas. Lasso logistic regression is an L1-regularized classifier that penalizes model complexity through an L1 norm penalty, thereby generating a sparse model representation, performing intrinsic feature selection and contrasting overfitting in high-dimensional feature spaces (Tibshirani, 1996). The model parameters are generally first estimated from a training set of labeled data points. The ensuing model is then used to classify new observations. Lasso logistic regression models were used to discriminate CG-rich regions coinciding with sites of replication initiation (origin-CGRs) from origin-negative CGRs (hereinafter CGR-classifier) and to discriminate *Drosophila* and human active replication origins from the rest of the *Drosophila* and human genomes, respectively (hereinafter origin-classifier).
The following sets of features were scored in a 500 bp window centered on each CGR midpoint (CGR-classifier) or on each RSS (origin-classifier):

1. DNA sequence content (k-mers, $k \leq 4$).
2. Enrichment of chromatin binding proteins, histone modifications and DNA-Seq, and MNase-Seq, RNA-Seq signals. These values were computed essentially as described (Comoglio and Paro, 2014). See Table S2 for detailed information on the datasets.

3. Mean DNA shape feature values (origin classifier only). Helix twist, propeller twist, minor groove width and roll, were considered.

These sets of features were used alone or in combination for the *Drosophila* origin-classifier. In contrast, helix twist, propeller twist and RNA-Seq were used for the human origin-classifier. For statistical learning, a balanced set composed of an equal number of observations for each class was created. For the CGR-classifier, negative instances were randomly selected among the remaining origin-negative CGRs. For the *Drosophila* origin-classifier, constitutive origins were used as positive instances and matched background regions (see above) were used as negative instances. The observations were randomly partitioned in a training set (80%) and in a test set (20%) and the Lasso logistic regression models were trained on the training set with ten-fold cross validation using the glmnet implementation (Friedman et al., 2010). The value of the regularization parameter minimizing the cross-validated misclassification error was used to predict the class labels of the genomic regions in the test set. Model performances were evaluated by computing the area under receiver operating characteristic curves (AUC) using the ROCR package (Sing et al., 2005). To estimate feature importance and identify the most predictive set of features, we performed stability analysis of model coefficients by computing feature selection probabilities (i.e. normalized frequencies of non-zero coefficients) using the bootstrap-Lasso algorithm as previously described (Sakoparnig et al. 2012; Comoglio and Paro 2014). Features were ranked by selection probabilities and minimal models were then trained using predictors selected with a probability ≥ 0.8.

Differential origin activity analysis

Differential origin activity analysis was carried out using the DESeq2 package (Love et al., 2014). Read summarization was performed for each biological replicate in the union of S2 and Bg3 origin peaks. An adjusted p-value cutoff of 10^{-5} was used to call significantly differentially activated origins.

Supplemental References

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339, 1074-1077.

Ashwal-Fluss, R., Meyer, M., Pamudurti, N.R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N., and Kadener, S. (2014). circRNA biogenesis competes with pre-mRNA splicing. Mol Cell 56, 55-66.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.M., and Lemaitre, J.M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat. Struct. Mol. Biol. 19, 837-844.

Bohla, D., Herold, M., Panzer, I., Buxa, M.K., Ali, T., Demmers, J., Krger, M., Scharfe, M., Jarek, M., Bartkuhn, M., and Renkawitz, R. (2014). A functional insulator screen identifies NURF and dREAM components to be required for enhancer-blocking. PLoS One 9, e107765.

Cayrou, C., Coulombe, P., Vigneron, A., Stanojcic, S., Ganier, O., Peiffer, I., Rivals, E., Puy, A., Laurent-Chabalier, S., Desprat, R., and Mchali, M. (2011). Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res. 21, 1438-1449.

Comoglio, F., and Paro, R. (2014). Combinatorial modeling of chromatin features quantitatively predicts DNA replication timing in Drosophila. PLoS Comput. Biol. 10, e1003419.

Comoglio, F., Sievers, C. and Paro, R. (2015). Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. BMC Bioinformatics 16, 32.

Contrino, S., Smith, R.N., Butano, D., Carr, A., Hu, F., Lyne, R., Rutherford, K., Kalderimis, A., Sullivan, J., Carbon, S., Kephart, E.T., Lloyd, P., Stinson, E.O., Washington, N.L., Perry, M.D., Ruzanov, P., Zha, Z., Lewis, S.E., Stein, L.D., Micklem, G. (2012). modMine: flexible access to modENCODE data. Nucleic Acids Res. 40 D1082-1088.

Eaton, M.L., Prinz, J.A., MacAlpine, H.K., Tretyakov, G., Kharchenko, P.V., and MacAlpine, D.M. (2011). Chromatin signatures of the Drosophila replication program. Genome Res. 21, 164-174.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.

Enderle, D., Beisel, C., Stadler, M.B., Gerstung, M. Athri, P., and Paro, R. (2011). Polycomb preferentially targets stalled promoters of coding and noncoding transcripts. Genome Res. 21, 216-226.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33, 1-22.

Gao, F., Luo, H., and Zhang, C.T. (2012). DeOri: a database of eukaryotic DNA replication origins. Bioinformatics 28, 1551-1552.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261-282.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li,

C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

Huppert, J.L., and Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 33, 2908-2916.

Kockmann, T., Gerstung, M., Schlumpf, T., Xhinzhou, Z., Hess, D., Beerenwinkel, N., Beisel, C., and Paro, R. (2013). The BET protein FSH functionally interacts with ASH1 to orchestrate global gene activity in Drosophila. Genome Biol. 14, R18.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357-359.

Lawrence, M., Huber, W., Pags, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. 9, e1003118.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25, 2078-2079.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sakoparnig, T., Kockmann, T., Paro, R., Beisel, C., and Beerenwinkel, N. (2012). Binding profiles of chromatin-modifying proteins are predictive for transcriptional activity and promoter-proximal pausing. J. Comput. Biol. 19, 126-138.

Sawarkar, R., Sievers, C., and Paro, R. (2012). Hsp90 globally targets paused RNA polymerase to regulate gene expression in response to environmental stimuli. Cell 149, 807-818.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940-3941.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 58, 267-288.

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

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. 41, W56-62.