Supplemental Materials
Chromatin signatures of the Drosophila replication program

Matthew L. Eaton[1], Joseph A. Prinz[1†] Heather K. MacAlpine[1†],
George Tretyakov[1], Peter V. Kharchenko[2] and David M. MacAlpine[1*]

[1] Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC 27710 USA.

[2] Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck St., Boston, Massachusetts 02115, USA.

[†]These authors contributed equally.

[*]To whom correspondence should be addressed; E-mail: david.macalpine@duke.edu

December 2, 2010

Supplemental Figures and Tables

Figure S1. Local chromatin signals around TSS proximal (black) and TSS distal (green) ORC binding sites. Each ORC binding site was classified as TSS proximal or TSS distal on the basis of 1 kb proximity to a transcription start site annotated in mb8 or flybase 5.12. The strength and relative location of the array probes in a 4 kb window around the ORC binding site were pooled and loess smoothed. The graphs represent the smoothed signal of PolII, GAF, ISWI and NURF in the cell lines for which we have data on those factors.

Figure S2. Distribution of RNA-seq expression scores for genes with ORC found within 1kb up- or down-stream of their transcription start site versus genes without ORC in the vicinity of the promoter. Expression values were taken from the cell-line specific modENCODE RNA-seq data Graveley et al. and were converted to log space by $log_2(fpkm + 1)$. Asterisks denote statistical significance by Wilcoxen unpaired test. In all cases, $p \leq 2 \times 10^{-16}$.

Figure S3. To test the possibility that active promoters with and without ORC may have different chromatin environments, we took the top quartile of genes by expression (3,662) in S2 cells and labeled them as having ORC overlapping the 1 kb 5' of their TSS (ORC; 1149) or not (No_ORC; 2513). We then displayed each factor's median enrichment as in figure 3. The most notable difference is the change in nucleosome dynamics which shows that bulk nucleosomes at ORC-occupied promoters are more depleted, enriched for H3.3 and experiencing rapid turnover. In addition, we see slight enrichment for NURF301 and ISWI. It is important to note that the scale of this figure is different from figure 2; since the top quartile of active promoters typically have stronger overall signals, the log_2 enrichment is from -3 to +3.

Figure S4. Motif distribution around the center of S2 ORC peaks. For the top 3 motifs returned by MEME, **(A)** (CT)n, **(B)** (CG)n, **(C)** (CA)n, the presence of each motif was binned relative to the center of an ORC peak within +/- 250 bp. For each motif, a background distribution of non-ORC bound regions is shown.

Figure S5. Enriched repetitive motif elements identified by MEME (CA)n,(CG)n and (CT)n had little predictive value in discriminating between ORC associated sites and random negative regions of the genome. Each ORC binding site was ranked by p-value for the presence of these motifs, and the ROC curve was generated by accepting ORC peaks with successively lower ranks.

Figure S6. SVM classification of ORC sites in Bg3 and Kc cells. **(A)** Using those features from the reduced feature set from figure 3 (in supplemental table S10) that have been profiled in Bg3 cells, an SVM was trained using each individual feature class (sequence, chromatin marks and binding proteins) and on the combination of the three. The area under the ROC curve characterizes the sensitivity and specificity of the classifier. **(B)** Each feature in BG3 cells was given an F-score and a t-statistic to gauge their relative contributions to the accuracy of the classifier. **(C)** As in A, but for Kc cells, for which fewer marks have been profiled. **(D)** Again, as in B, but for Kc cells.

Figure S7. Per-factor median signals in areas of differential ORC and early origins. Each factor was assayed under **(A)** ORC peaks and **(B)** early origin peaks from the union of the S2 and Bg3 ORC and early origin peak sets respectively. **(A)** ORC peaks that appear only in Bg3 cells were profiled using marks from Bg3 (first column) and S2 (second column), peaks that appear only in S2 cells were profiled using marks from S2 (third column) and Bg3 (fourth column), and peaks that appear in both cell lines were profiled using marks from BG3 (fifth column) and S2 (sixth column). **(B)** The same column pattern as in A is displayed, showing the enrichment for the factors within early origin peaks. In both panels, The green squares at the top act as a key to the identity of the column. All enrichments are shown from -2 to 2 median log_2 enrichment. Factor enrichments were collected as in figure 2.

Figure S8. A hierarchical clustering using the Ward metric of the correlation matrix of 36 chromatin factors at early origin meta-peaks in Bg3 cells. The red boxes indicate the clusters shown in figure 4A, while the numbers beneath the red boxes indicate the cluster label.

Figure S9. Pearson correlation of each chromatin factor cluster's strength with the signal from the early origin meta-peaks in Bg3. Pearson's r and the associated p-value are noted in the top right corner.

Figure S10. A table showing the features selected by feature selection in each of the three cell lines. Their F-score and t-statistic (as reported by libsvm) is also shown.

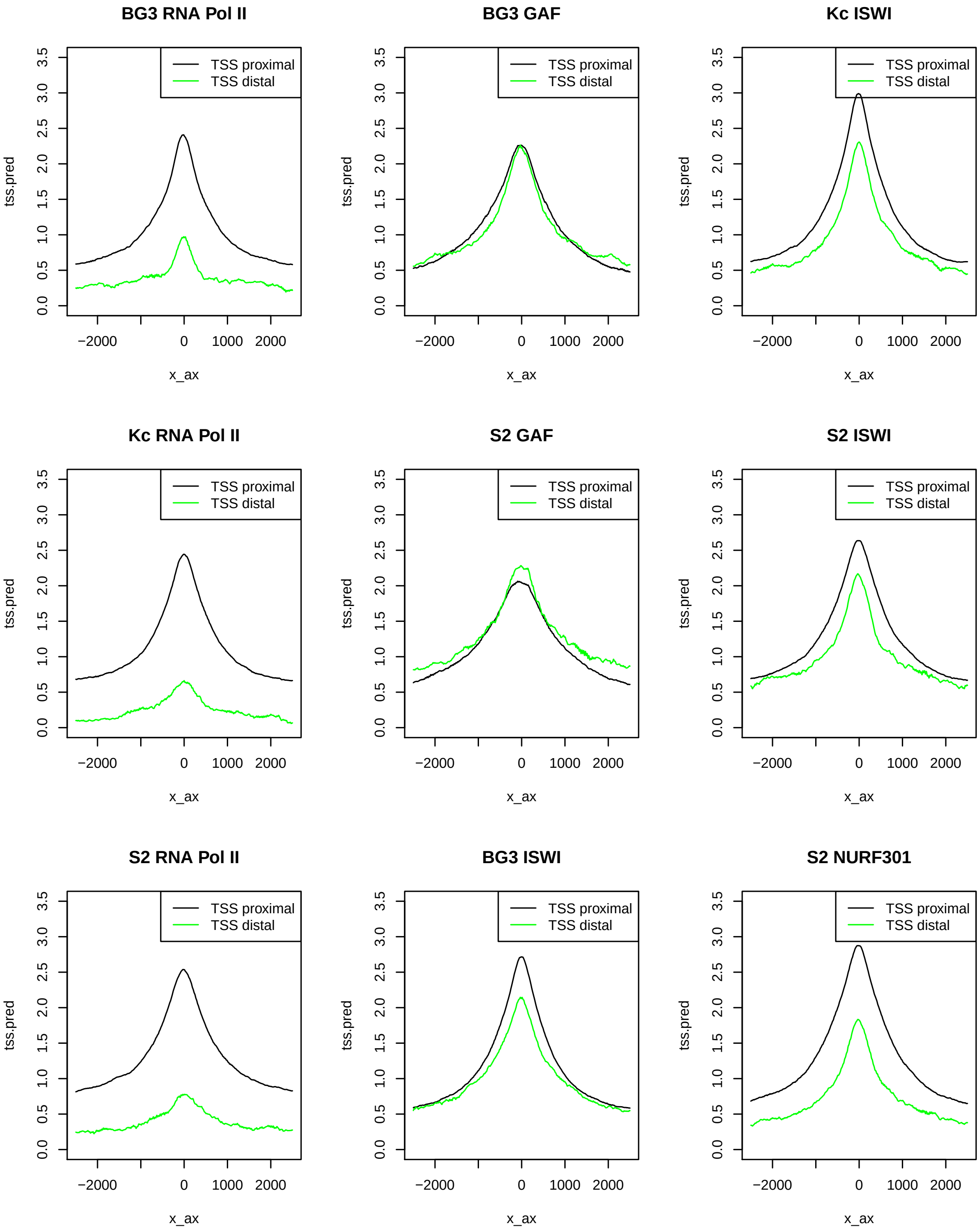

Figure S1: Local chromatin signals around TSS proximal (black) and TSS distal (green) ORC binding sites.

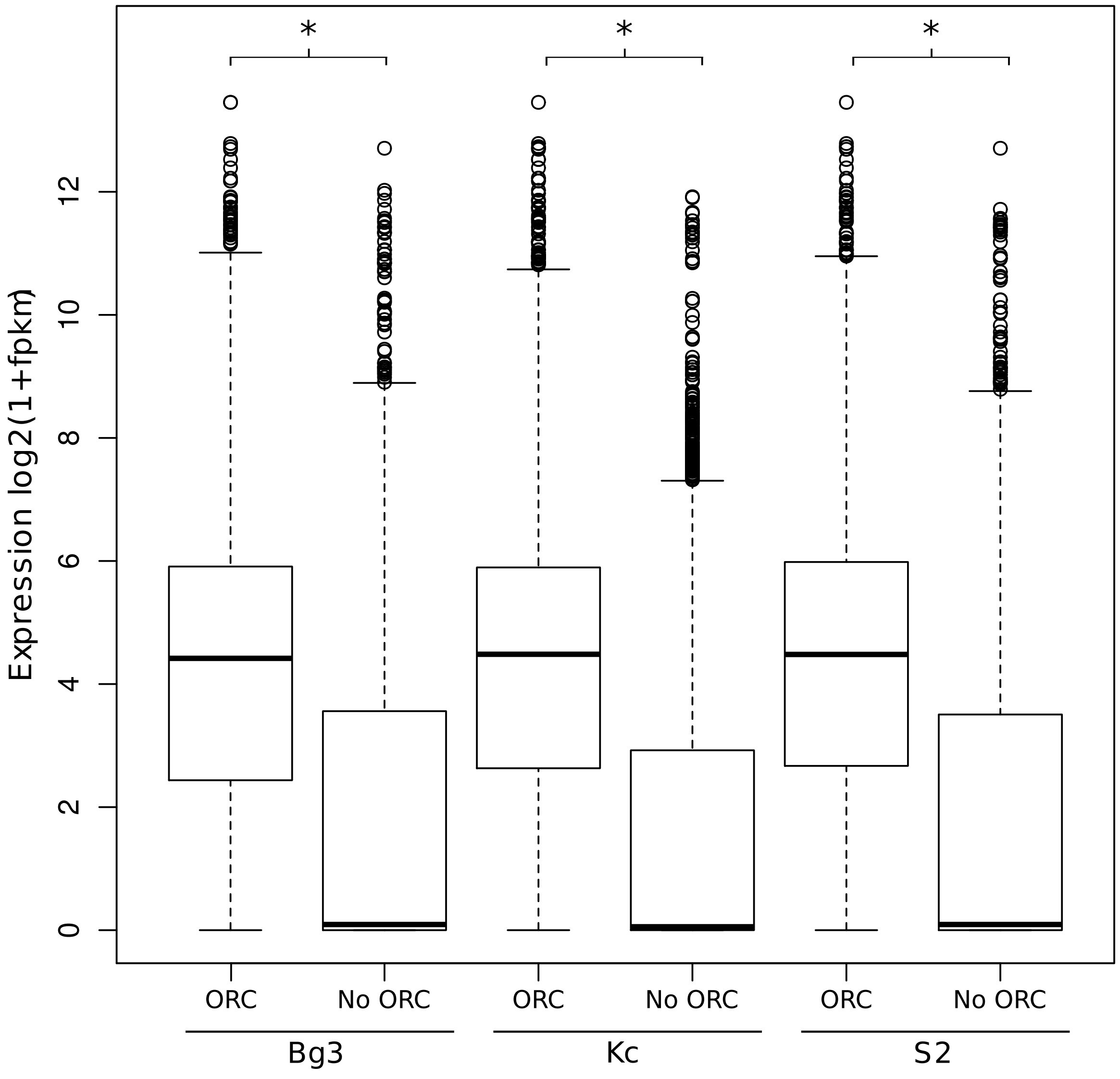

Figure S2: Expression of genes with and without ORC bound to their promoter.

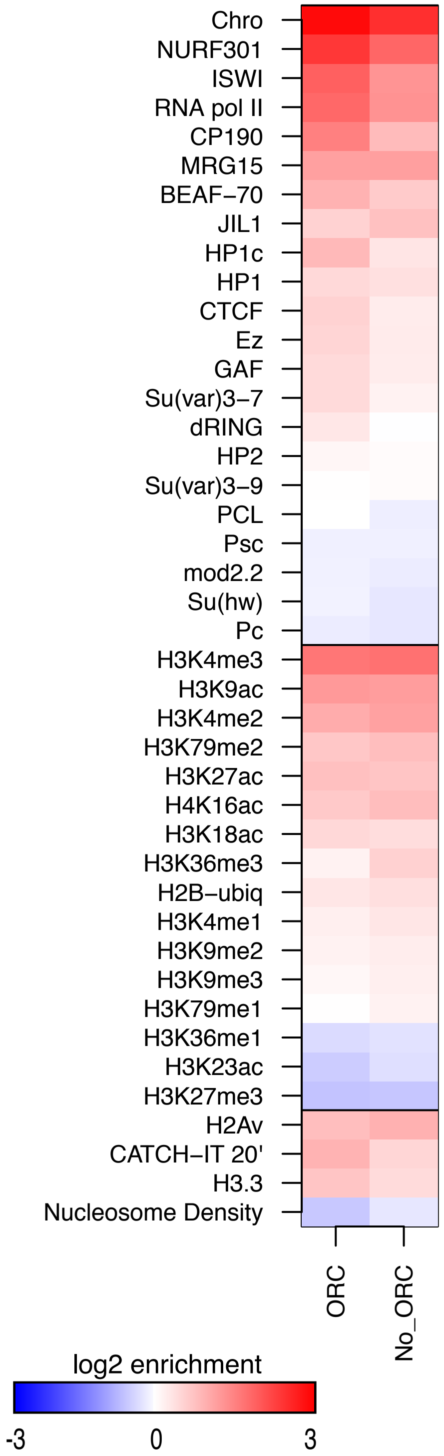

Figure S3: Median array signal for all factors at promoters with ORC, and those without.

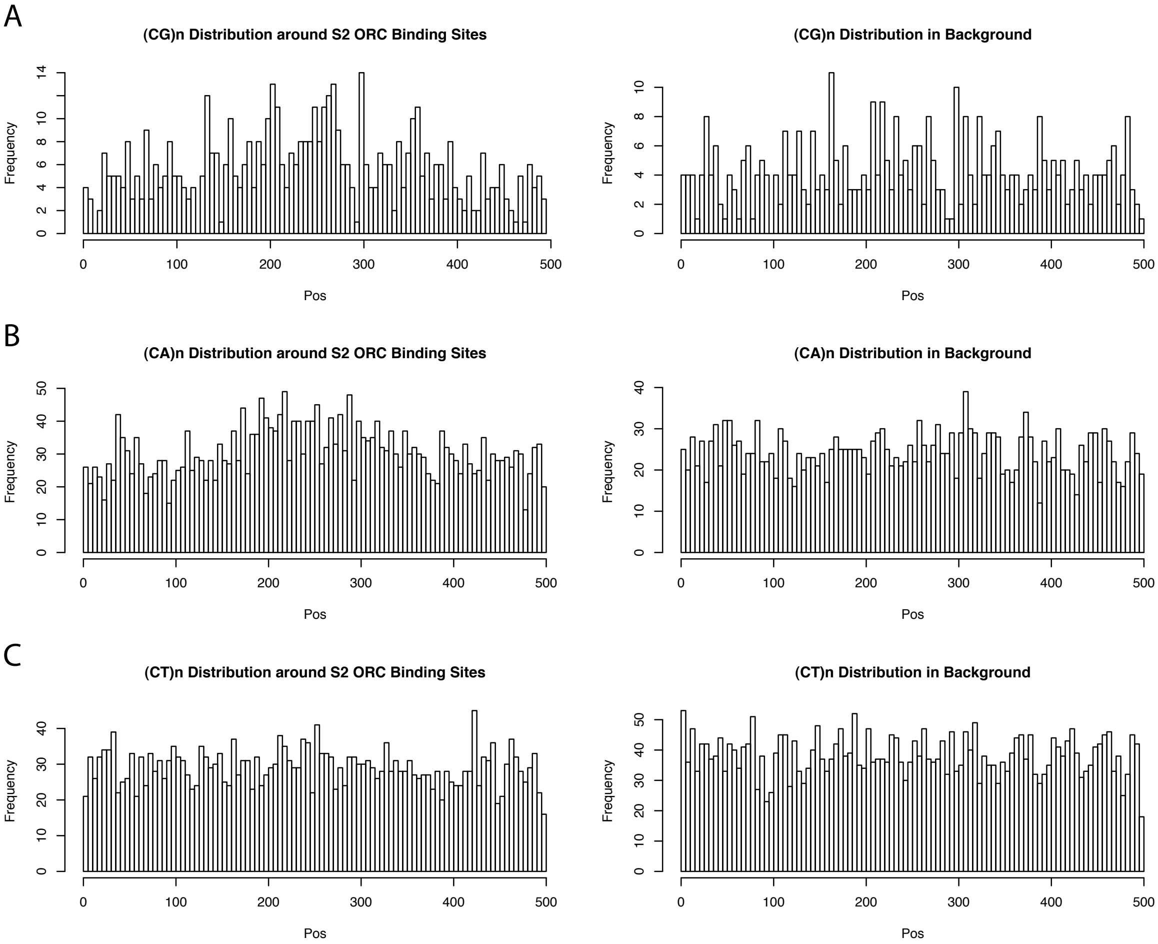

Figure S4: A histogram of the occurrences of high scoring motifs relative to the center of the ORC peak.

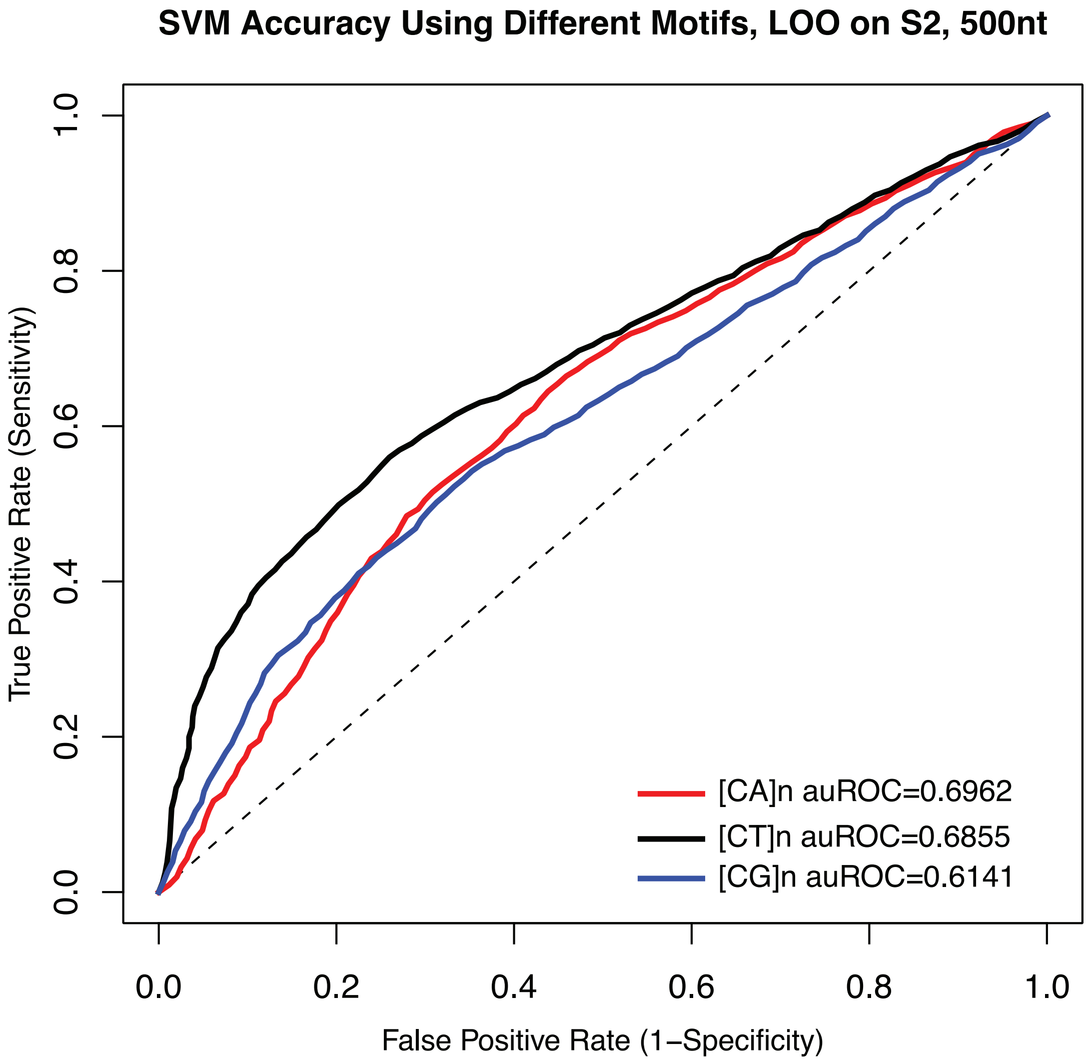

Figure S5: ROC curve for motif classification of ORC binding sites.

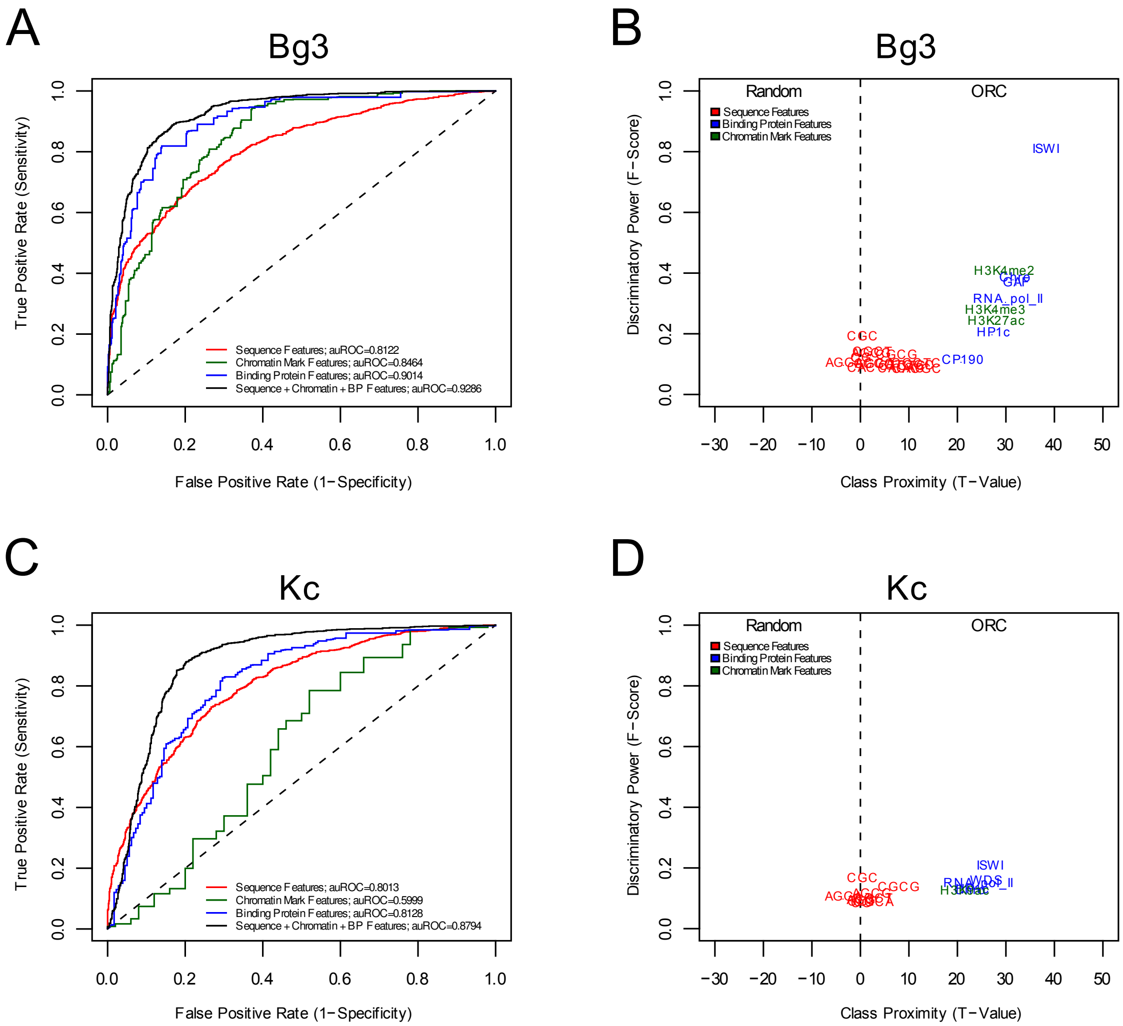

Figure S6: SVM classification of ORC sites in Bg3 and Kc.

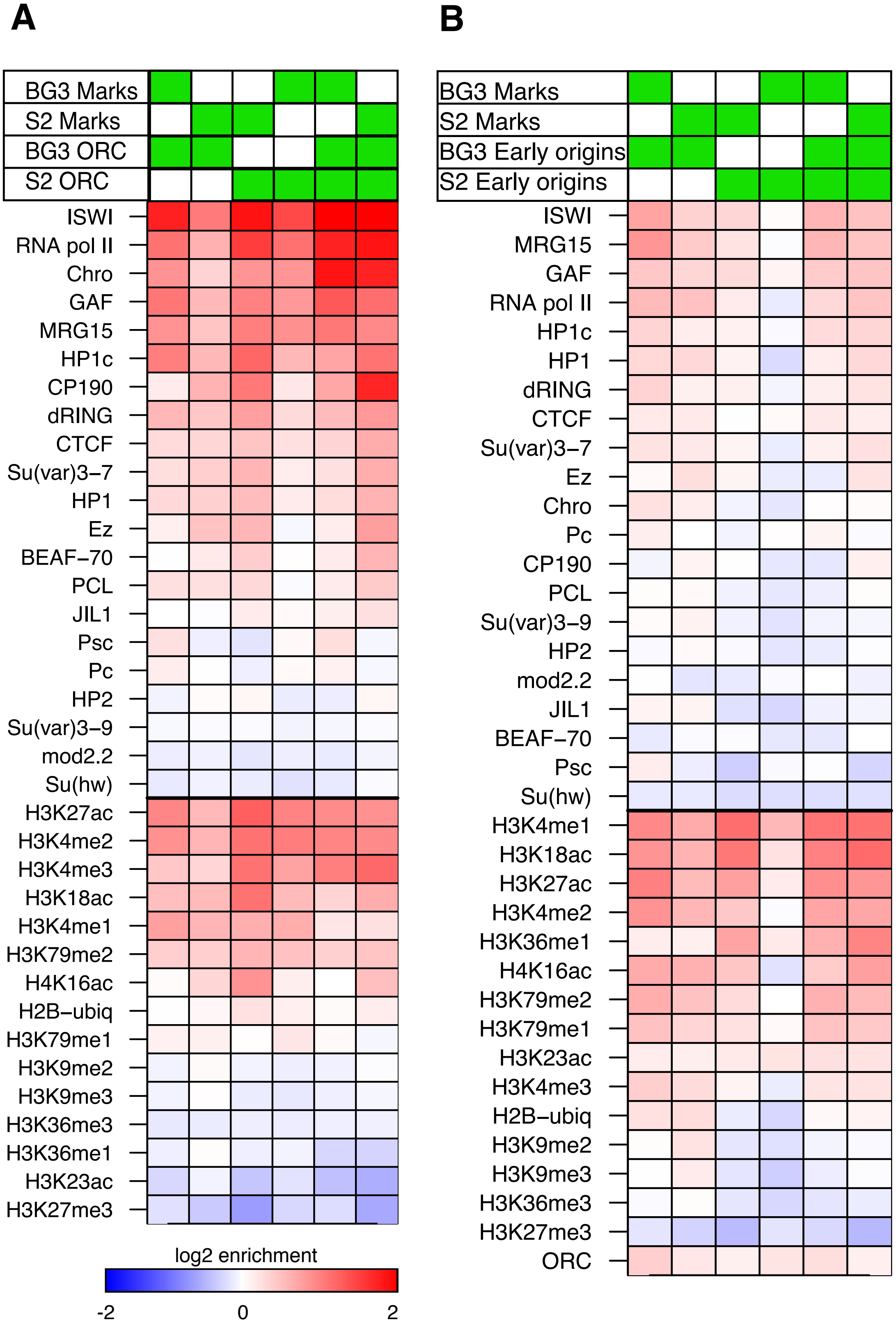

Figure S7: Per-factor median signals in areas of differential ORC (A) and early origin (B) activity.

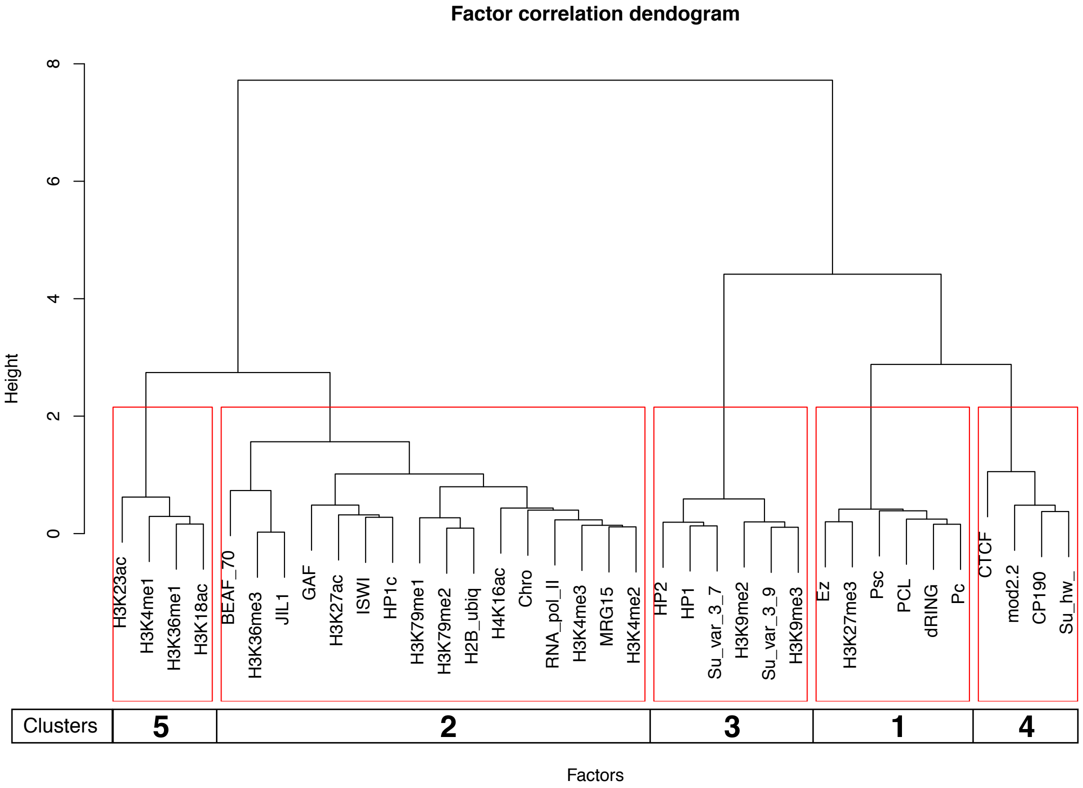

Figure S8: Hierarchical clustering of chromatin factors at early origin meta-peaks.

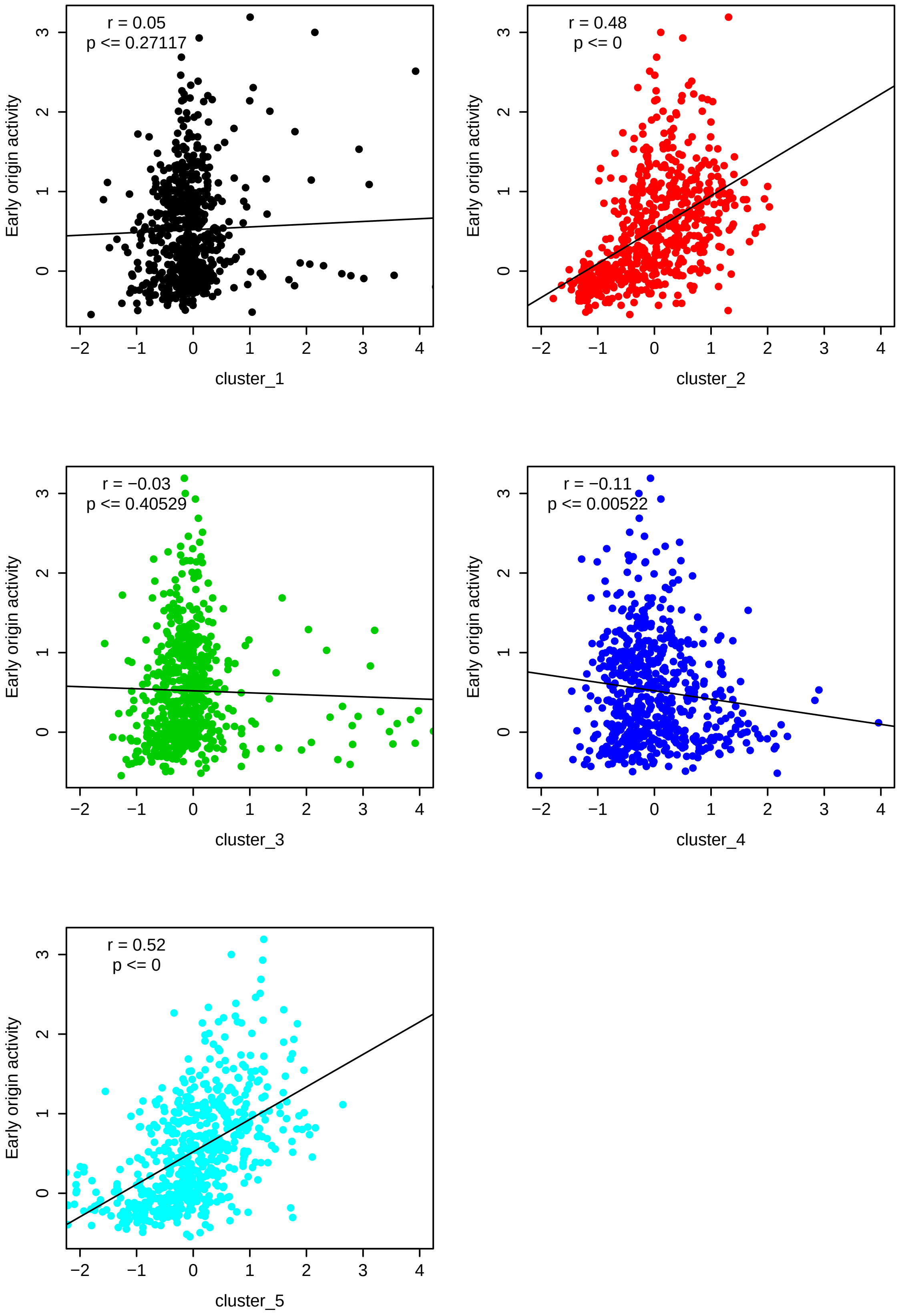

Figure S9: Correlations between chromatin factor cluster signal with early origin meta-peak strength in Bg3 cells.

	S2			Bg3			Kc		
	Feature	T-Score	F-Score	Feature	T-Score	F-Score	Feature	T-Score	F-Score
Sequence	CG	−0.14	0.11	CG	−0.08	0.11	CG	0.01	0.08
	AGG	−0.17	0.10	CAC	−0.03	0.08	AGG	0.04	0.09
	CCT	−0.18	0.10	CGC	0.10	0.18	CCT	0.00	0.08
	CGC	−0.04	0.20	ACGC	2.63	0.09	CGC	0.12	0.16
	ACGC	2.10	0.09	AGCG	2.03	0.13	AGCG	2.13	0.11
	AGCG	2.16	0.14	AGGA	−3.44	0.10	AGGA	−3.57	0.10
	AGGA	−3.98	0.12	CGCA	2.44	0.09	CGCA	2.32	0.08
	CGCA	2.27	0.10	CGCG	7.03	0.12	CGCG	7.48	0.13
	CGCG	7.15	0.15	CGCT	2.05	0.13	CGCT	2.17	0.09
	CGCT	2.10	0.13	CACAC	8.02	0.08			
	AGCGA	8.04	0.08	CACGC	11.14	0.08			
	AGCGC	11.89	0.08	CGCTC	11.01	0.09			
	AGGAC	−13.68	0.08	CTCGC	9.59	0.09			
	CACAC	6.85	0.08	CTCTC	8.19	0.09			
	CGCGC	14.69	0.08						
	CGCTC	10.12	0.08						
	CTCGC	10.71	0.10						
	CTCTC	7.31	0.08						
Binding Proteins	Chro	30.10	0.41	Chro	31.44	0.38	Chro	22.19	0.12
	CP190	32.50	0.41	CP190	20.77	0.11	HP1c	22.82	0.12
	dRING	20.50	0.10	GAF	32.06	0.36	ISWI	26.78	0.20
	GAF	30.85	0.36	HP1c	27.16	0.20	RNA_pol_II	23.79	0.14
	HP1c	29.81	0.29	ISWI	38.29	0.80	WDS	25.65	0.15
	ISWI	37.26	0.90	RNA_pol_II	29.89	0.30			
	NURF301	36.97	0.99						
	RNA_pol_II	30.87	0.55						
	WDS	36.48	0.78						
Chromatin Marks	H3K18ac	23.97	0.20	H3K27ac	27.47	0.23	H3K9ac	21.06	0.12
	H3K27ac	28.92	0.34	H3K4me2	29.21	0.40			
	H3K4me2	30.08	0.54	H3K4me3	27.35	0.27			
	H3K4me3	29.84	0.40						
	H3K79me2	16.82	0.09						
	H3K9ac	30.04	0.44						
	H4K16ac	21.74	0.17						

Figure S10: A list of the top SVM features by F-score, used in the SVM analysis of figure 4.

References

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J., Yang, L., Artieri, C., van Baren, M. J., Booth, B. W., Brown, J. B., *et al.*, 2010. The developmental transcriptome of drosophila melanogaster. *Nature*, **Submitted**.