Supplementary Material

Genome accessibility is widely preserved and locally modulated during mitosis

October 30, 2014

Chris C.-S. Hsiung*, Christapher S. Morrissey*, Maheshi Udugama, Christopher L. Frank, Cheryl A. Keller, Songjoon Baek, Belinda Giardine, Gregory E. Crawford, Myong-Hee Sung, Ross C. Hardison, and Gerd A. Blobel

Supplemental Results

Supplemental Methods

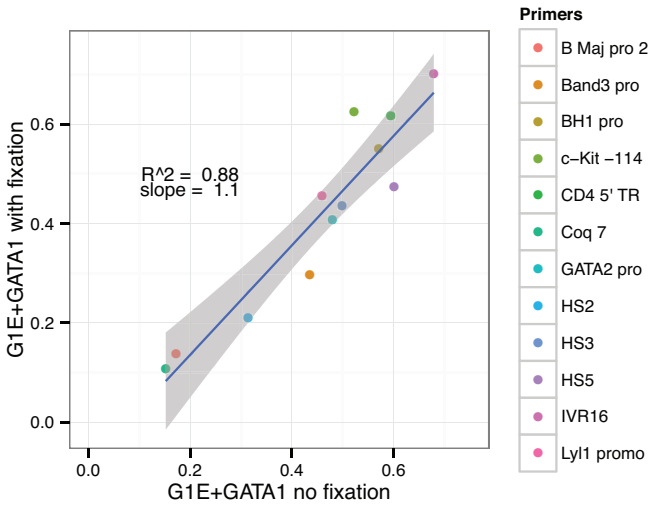

Figure S1 – Mild formaldehyde fixation does not appreciably alter DNase sensitivity. DNase-qPCR was performed on G1E+GATA1 cells with and without 0.1% formaldehyde fixation using the same set of primers. R-squared and slope of linear regression are shown, and the gray shaded area marks the 95% point-wise confidence interval.

G1E+GATA1 Asynchronous		
	REP1	**REP2**
REP2	0.84	
REP3	0.82	0.93
G1E+GATA1 Mitotic		
	REP1	**REP2**
REP2	0.79	
REP3	0.73	0.77
G1E Asynchronous		
	REP1	**REP2**
REP2	0.71	
REP3	0.82	0.87
G1E Mitotic		
	REP1	**REP2**
REP2	0.76	
REP3	0.74	0.90

Figure S2 – DNase-seq biological replicates show high degree of concordance. Pearson correlation coefficients between pairs of individual biological replicates are shown. The coefficients are calculated from DNase cut densities using reads from each individual replicate within a single set of regions defined by the union of all hotspots across all conditions (called from reads pooled from all replicates).

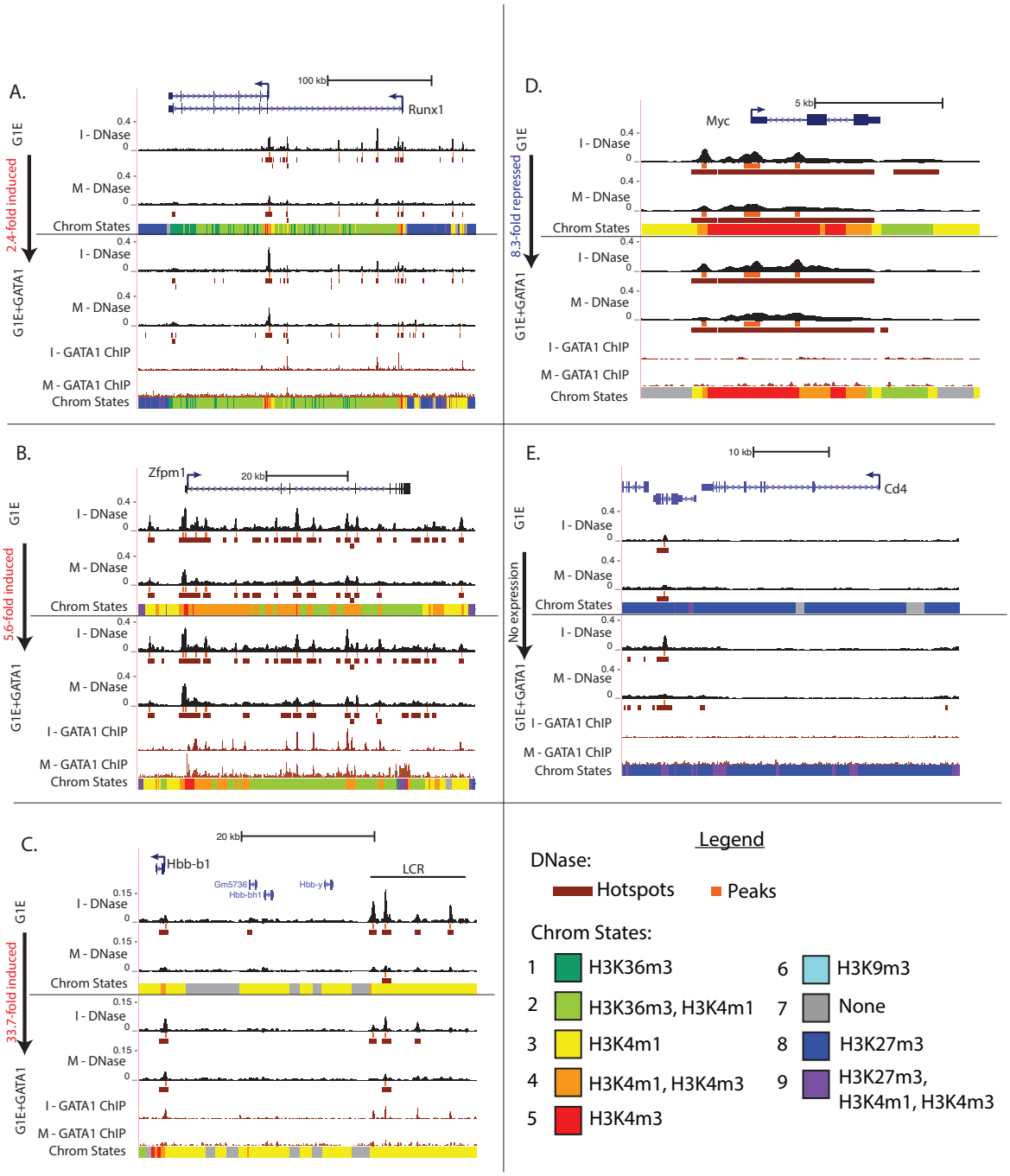

Figure S3 – Integrative browser track views at additional loci. Legend indicates the predominant histone modifications associated with each chromatin state. **a)** *Runx1*: Significant retention of accessibility during mitosis at the alternative promoter on the left, and less so at the other promoter. Distal DNase peaks generally diminished during mitosis. **b)** *Zfpm1*: note the lack of change in mitotic G1E and G1E+GATA1 DNase profiles despite multiple GATA1 mitotic binding sites. Also, the promoter accessibility relatively unchanged despite 5.6-fold induction in mRNA in G1E+GATA1, compared to G1E. **c)** *Hbb-b1* (beta-globin): note the dramatic diminishment of DNase sensitivity at the LCR during mitosis. **d)** *Myc*: Example of a gene with broad moderate DNase sensitivity captured by a large hotspot. Note strong repression at mRNA level from G1E to G1E+GATA1, but DNase sensitivity remains unchanged. **e)** *Cd4*: a non-expressed gene marked by heterochromatin-associated histone modifications and no DNase sensitivity.

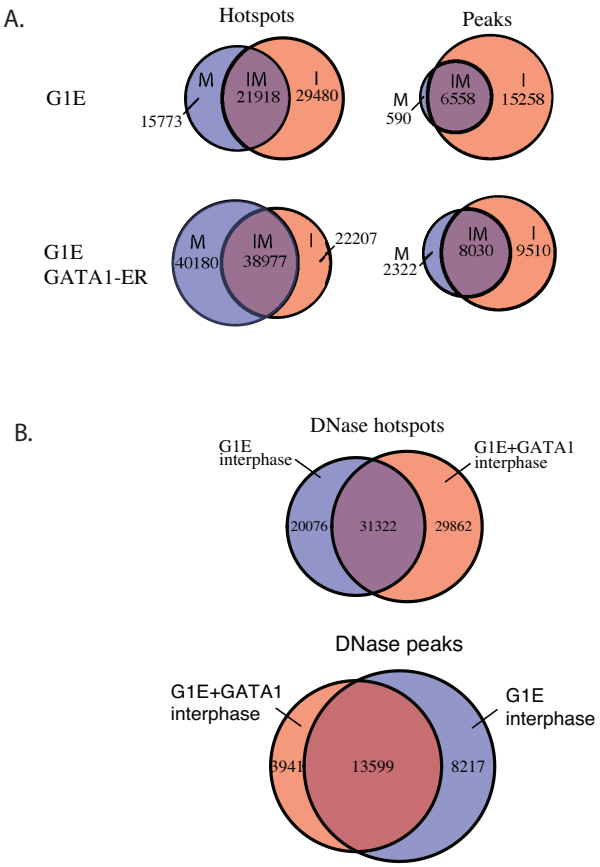

Figure S4 – Venn diagram summaries of hotspots and peaks called in each experimental condition. a) Overlap between interphase and mitosis hotspots and peaks in each experimental condition (G1E and G1E+GATA1). Quantitative assessments of DNase cut densities for parts of the venn diagrams are shown in Fig. S5. b) Overlap of hotspots and peaks between G1E and G1E+GATA1 in interphase.

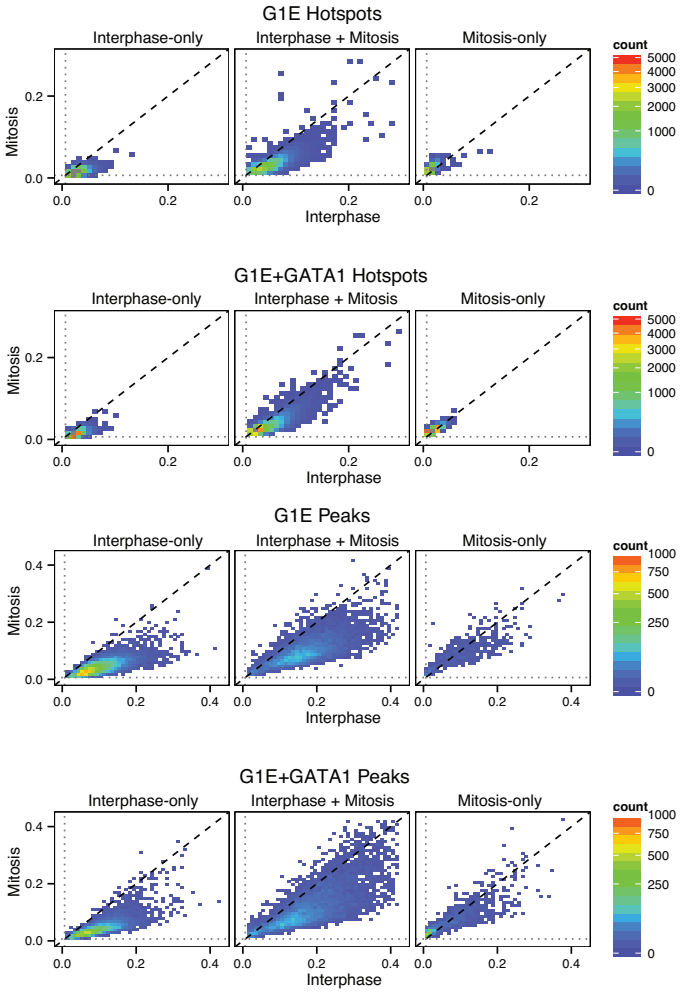

Figure S5 – DNase cut densities of hotspots called in mitosis-and-interphase, mitosis-only, and interphase-only. Note that "mitosis-only" and "interphase-only" hotspots can have comparable signal in both interphase and mitosis, despite their meeting thresholds in only one condition, and in such cases are generally lower in signal. For most analyses described in the main text, we interrogated a single set of regions consisting of the union of all hotspots called in at least one of the four experimental conditions, then assessed their changes DNase cut densities across conditions.

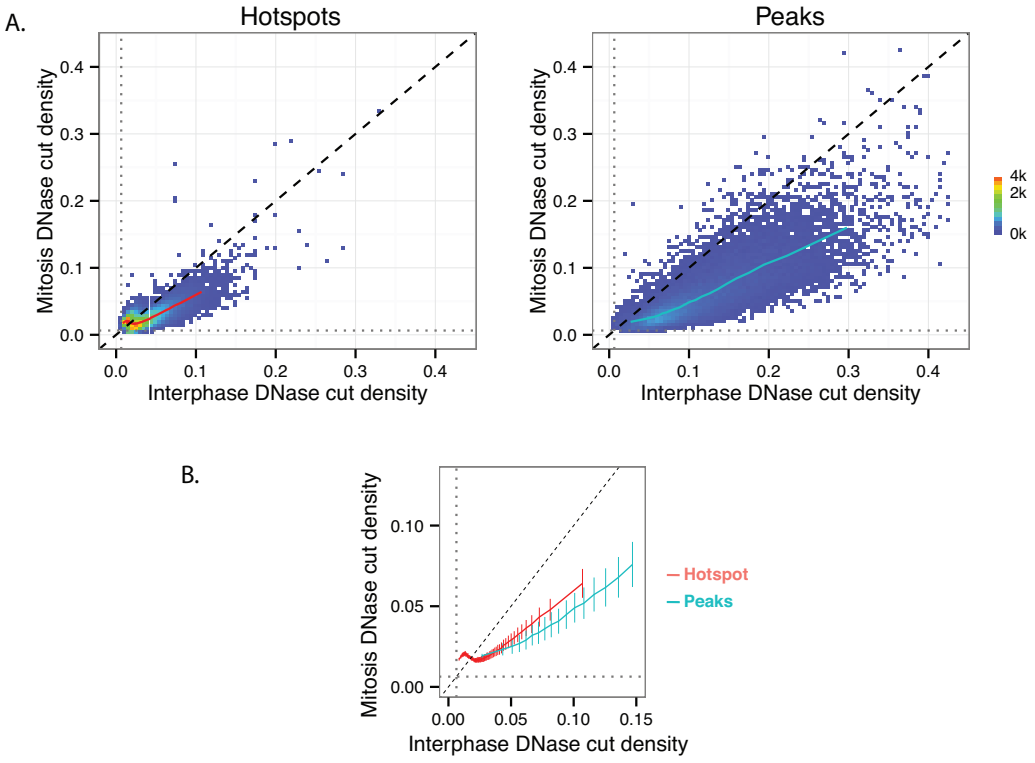

Figure S6 – Related to Fig. 3: Mitosis and interphase dynamics of hotspots and peaks for G1E. The same analysis of the mitosis and interphase DNase cut densities of hotspots versus peaks as in Fig. 3A and Fig. 3B, which show the results for G1E+GATA1, are presented for G1E.

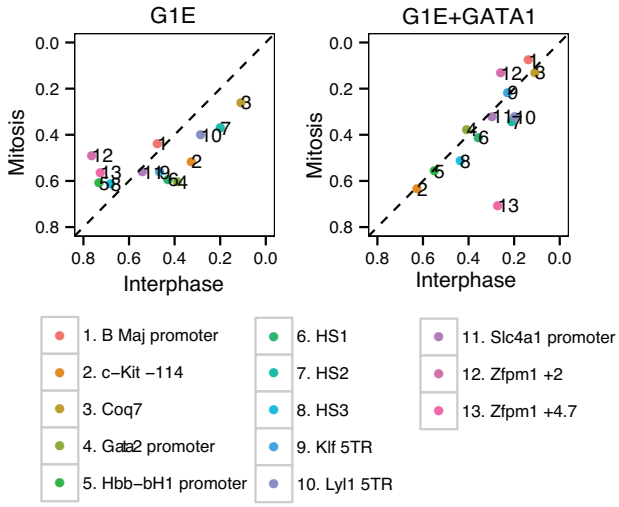

Figure S7 – DNase-qPCR of individual sites show significant preservation of chromatin accessibility during mitosis. Data from our previous study (Kadauke et al., 2012) are re-plotted as a scatterplot of mitotic versus interphase qPCR quantities normalized by input DNA. Primer sequences can be found in Kadauke et al. (2012).

State	H3K36m3	H3K4m1	H3K4m3	H3K27m3	H3K9m3	Predominant feature
1	0.812	0.033	0.003	0.004	0.019	H3K36m3
2	0.942	0.882	0.046	0.009	0.016	H3k36m3, H3K4m1
3	0.025	0.698	0.009	0.014	0.006	H3K4m1
4	0.280	0.998	0.967	0.019	0.017	H3K4m1, H3K4m3
5	0.059	0.067	0.983	0.006	0.027	H3K4m3
6	0.005	0.003	0.018	0.009	0.564	H3K9m3
7	0.001	0.002	0.001	0.007	0.011	None
8	0.001	0.013	0.005	0.543	0.018	H3K27m3
9	0.046	0.808	0.231	0.883	0.095	H3K27m3, H3K4m1, H3K4m3

Figure S8 – Chromatin states defined by ChromHMM. Emission probabilities for histone modifications are shown for the states derived from ChromHMM. States 2, 3 and 4 are used as part of distal CRM definition, and State 5 is used as part of promoter definition (detailed in Supplemental Methods).

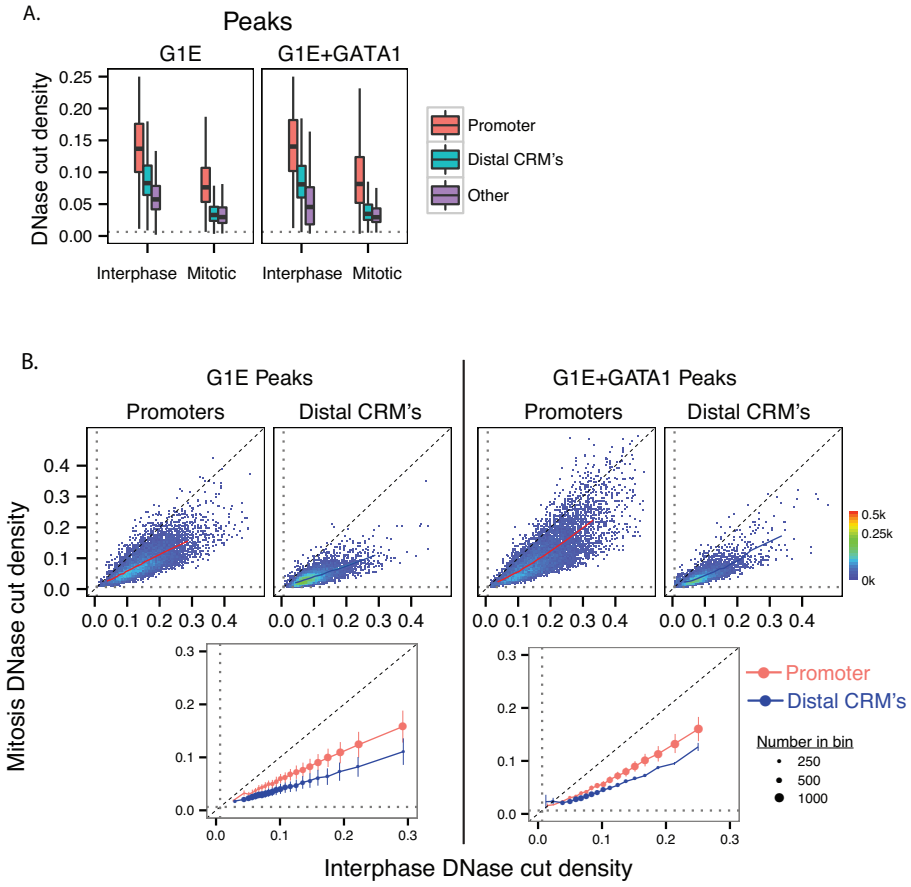

Figure S9 – Related to Fig. 4: Promoter peaks preserve mitotic accessibility more than distal CRM peaks. The plots contrasting the mitosis and interphase DNase cut densities for promoters and distal CRM peaks, using the same conventions as Fig. 4, which showed the same information for hotspots.

A.

G1E+GATA1 >0.85 mitosis/interphase

Description	P–value	Fold Enrichment
signaling receptor activity	4.8e–21	1.6
transmembrane signaling receptor activity	3.7e–19	1.7
sequence–specific DNA binding	1.1e–17	1.5
G–protein coupled receptor activity	1.7e–12	1.7
transcription regulatory region sequence–specific DNA binding	2.0e–06	1.9
olfactory receptor activity	1.3e–04	2.4
polysaccharide binding	1.3e–03	1.7
cation channel activity	1.7e–03	1.5
glycosaminoglycan binding	3.9e–03	1.7
RNA polymerase II regulatory region DNA binding	1.1e–02	1.9

B.

G1E+GATA1 ≥10 cell/tissue type

Description	P–value	Fold Enrichment
DNA binding	1.2e–30	1.8
sequence–specific DNA binding transcription factor activity	1.7e–27	2.2
nucleic acid binding transcription factor activity	2.4e–27	2.2
sequence–specific DNA binding	4.0e–27	2.4
sequence–specific DNA binding RNA polymerase II transcription factor activity	1.0e–10	3.0
transcription regulatory region DNA binding	2.2e–10	2.3
regulatory region DNA binding	9.7e–10	2.2
transcription factor binding	1.7e–09	2.2
transcription regulatory region sequence–specific DNA binding	2.0e–09	3.4
chromatin binding	1.2e–07	2.2

C.

G1E+GATA1 >0.85 mitosis/interphase and ≥10 cell/tissue type

Description	P–value	Fold Enrichment
DNA binding	2.9e–28	2.1
sequence–specific DNA binding	1.0e–26	3.0
sequence–specific DNA binding transcription factor activity	2.4e–26	2.7
nucleic acid binding transcription factor activity	3.1e–26	2.7
nucleic acid binding	4.0e–14	1.6
sequence–specific DNA binding RNA polymerase II transcription factor activity	1.2e–10	4.0
transcription regulatory region DNA binding	8.5e–08	2.7
regulatory region DNA binding	1.4e–07	2.6
transcription regulatory region sequence–specific DNA binding	2.7e–07	4.1
chromatin binding	3.4e–06	2.7

Figure S10 – Related to Fig. 5: Gene ontology analyses of promoter hotspots preserved in mitosis and across many murine tissues. Subsets of G1E+GATA1 promoter hotspots were subjected to analysis using GREAT, which assigns each promoter hotspot to the single nearest gene and identifies GO terms enriched over the background set consisting of all G1E+GATA1 promoter hotspots. Only the top 10 GO terms in the molecular function category that are enriched $\geq$1.5 fold (p-value $<$0.05 after Bonferroni correction for multiple testing) are listed. **a)** G1E+GATA1 promoter hotspots with mitosis-to-interphase ratio of $>$0.85 were used as target regions. **b)** G1E+GATA1 promoter hotspots present across $\geq$10 cell or tissue types were used as target regions. **c)** G1E+GATA1 promoter hotspots with mitosis-to-interphase ratio of $>$0.85 and are present across $\geq$10 cell or tissue types were used as target regions.

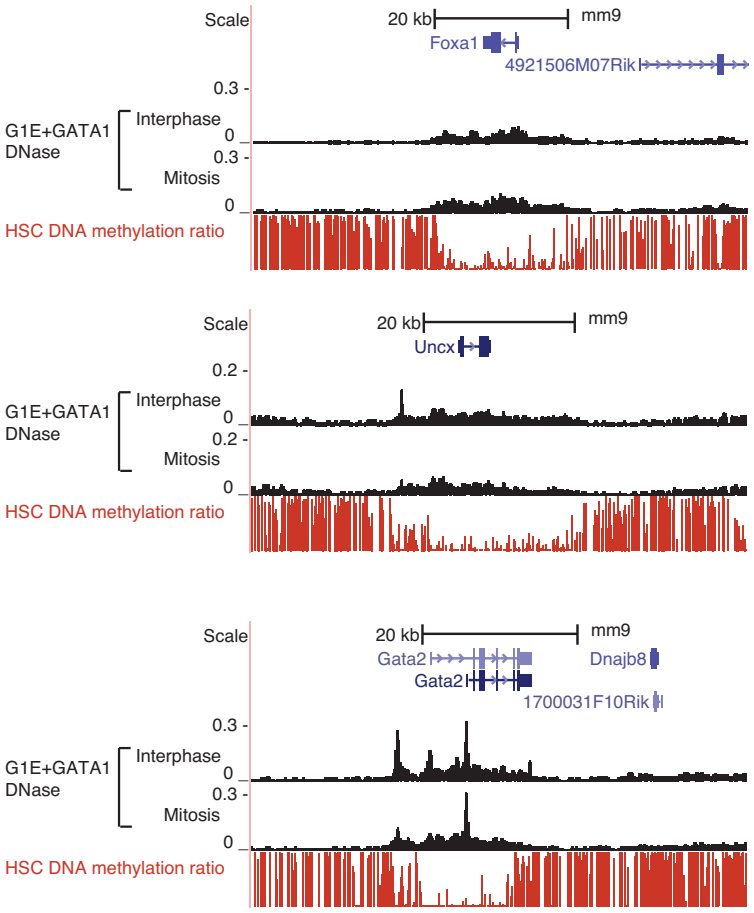

Figure S11 – DNase sensitivity hotspots preserved in interphase and mitosis approximate the borders of the top DNA methylation canyons. Several representative loci marked by DNA methylation canyons in mouse HSCs are shown, along with their DNase sensitivity profiles in interphase and mitosis for G1E+GATA1.

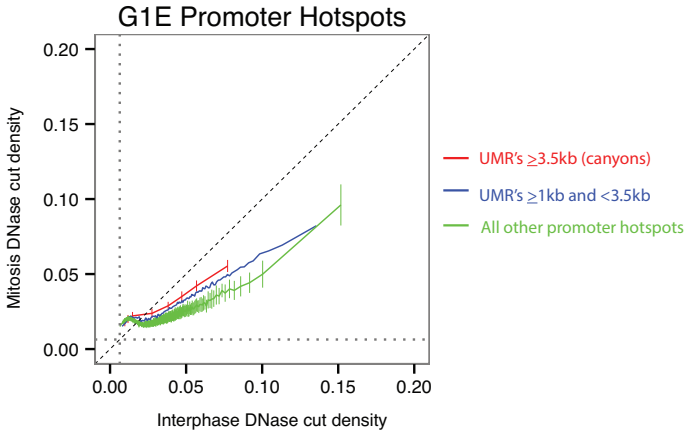

Figure S12 – G1E promoter hotspots overlapping DNA methylation canyons exhibit high mitotic accessibility, compared to promoter hotspots matched for interphase accessibility. The same analysis as in Fig. 5E (for G1E+GATA1) was performed on G1E promoter hotspots.

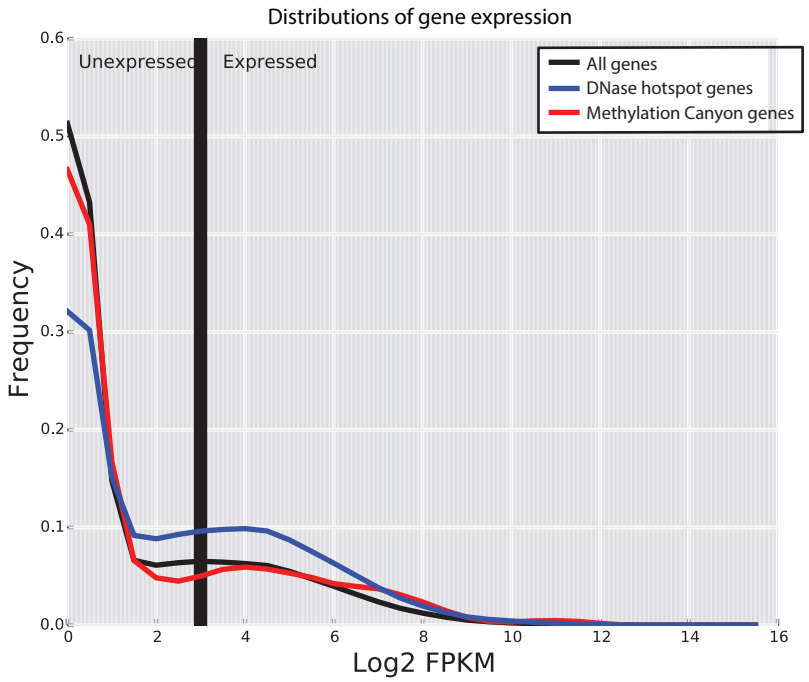

Figure S13 – Related to Fig. 5: Genes residing in DNase hotspots and DNA methylation canyons include both expressed and silent genes. Distributions of the RNA-seq FPKM of genes (T Mishra, C Morrissey, C Keller, B Giardine, E Heuston, S Anderson, V Paralkar, M Pimkin, M Weiss, D Bodine, et al., submitted) are shown for all genes, genes overlapping at least one DNase hotspot, and genes overlapping DNA methylation canyons (UMR $\geq$3.5kb). The vertical black line indicates a threshold (log2 FPKM = 3) for dividing genes into expressed versus non-expressed.

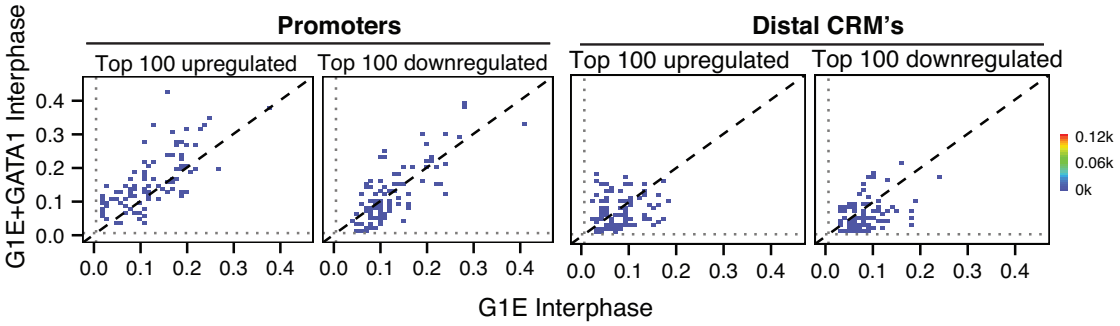

Figure S14 – Related to Fig. 6: GATA1-induced differential expression is associated with only mild promoter accessibility changes and does not explain variations in accessibility of the nearest distal CRM. The top 100 upregulated and top 100 downregulated genes between G1E and G1E+GATA1 were identified using RNA-seq data (T Mishra, C Morrissey, C Keller, B Giardine, E Heuston, S Anderson, V Paralkar, M Pimkin, M Weiss, D Bodine, et al., submitted). Scatter plots of mitosis versus interphase DNase read densities are shown for promoter and distal CRM peaks. Graph conventions are similar to Fig. 6B.

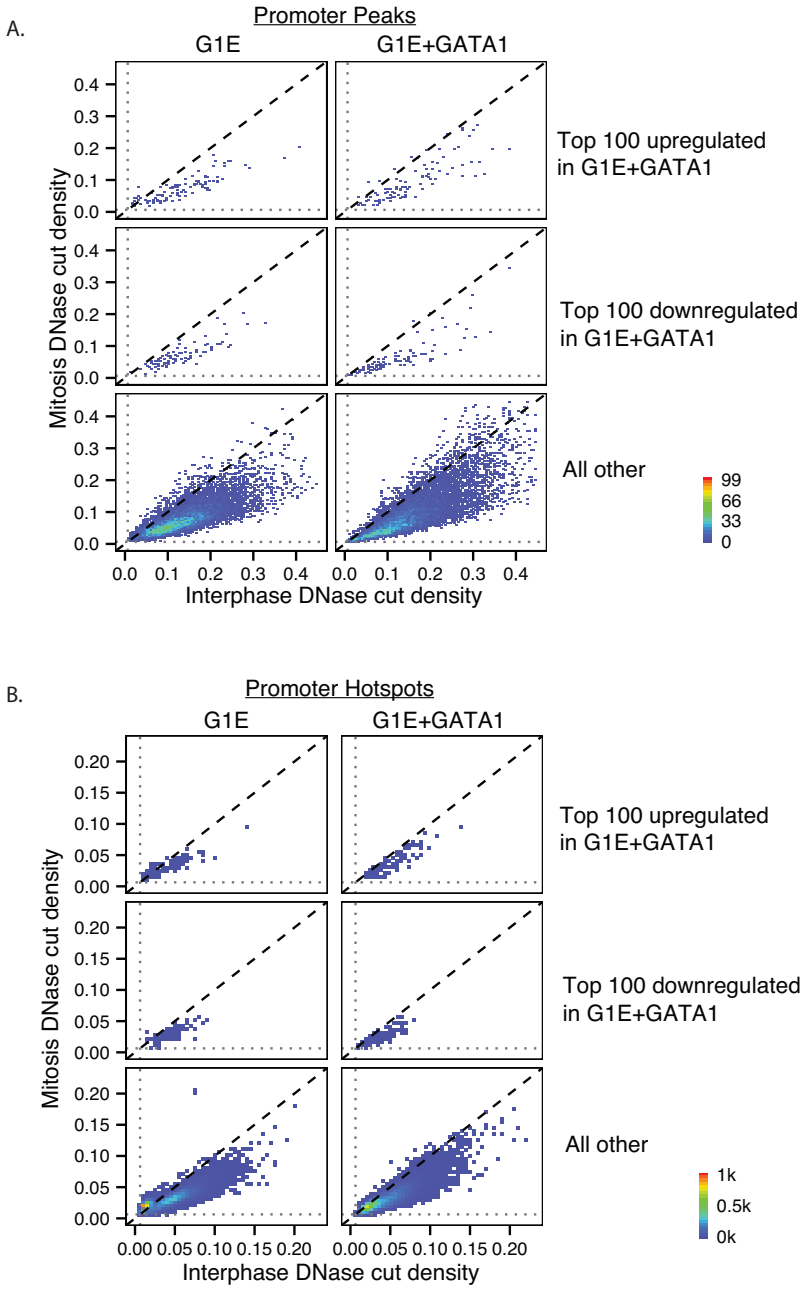

Figure S15 – Preservation of mitotic accessibility are indistinguishable between promoters of genes whose expression are unchanged by GATA1, versus those with strong inducible gene expression changes. a) Based on fold change of gene expression from G1E to G1E+GATA1, promoter peaks are divided into those that are among the top 100 most unregulated, top 100 most down-regulated, and all other promoter peaks. The mitosis DNase cut density versus the interphase DNase cut density are shown separately for G1E and G1E+GATA1. b) Same information as in a) are shown for promoter hotspots.

10,460 Gata1 occupancy sites		
8,831 interphase-only (I-Gata1)	**527 interphase & mitosis (IM-Gata1)**	**1,102 mitosis-only (M-Gata1)**
71.1% overlap DNase hotspots	84.1% overlap with DNase hotspots	21.6% overlap with DNase hotspots
18.7% overlap with DNase peaks	45.2% overlap with DNase Peaks	5.8% overlap with DNase peaks

Figure S16 – Related to Fig. 6: Overlap between GATA1 binding sites and DNAse hotspots and peaks. The union of all DNase hotspots or peaks across all samples were overlapped with GATA1 binding sites categorized based on ChIP-seq peak-calling from Kadauke et al. (2012).

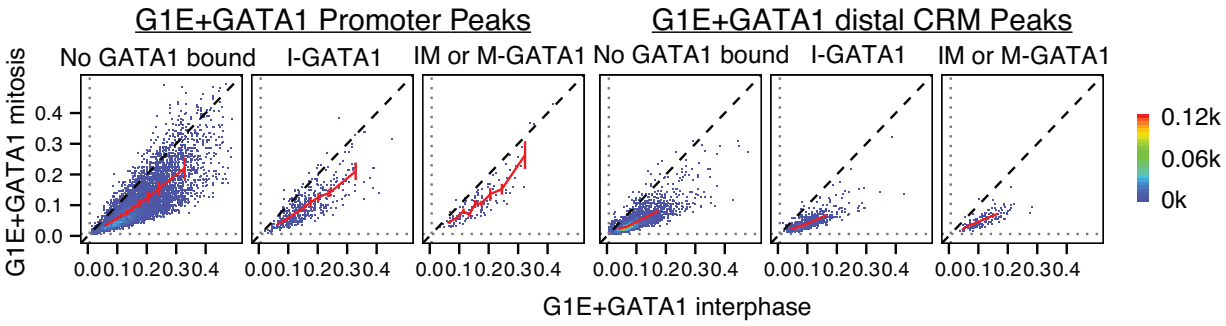

Figure S17 – Related to Fig. 6: GATA1 mitotic occupancy does not contribute significantly to site-specific variations in mitotic accessibility DNase peaks at promoters and distal CRMs are grouped by their overlap with subcategories of GATA1 binding sites, and shown in binned 2D density plots of mitosis DNase cut density versus interphase DNase cut density. Error bars for the moving mean denote SEM of biological replicates (n=3); some may be too small to see.

1 Supplemental Methods

1.1 Processing of reads from DNase-seq libraries

Raw reads from DNase-seq libraries were first groomed using FASTQ Groomer on Galaxy (Giardine, 2005; Blankenberg et al., 2001; Goecks et al., 2010). This program verifies that each base call has a corresponding quality value, and that the quality value is in the Sanger, Phred+33 format. Groomed reads were then trimmed as only the first twenty bases of each read are genomic DNA. The trimmed reads were then mapped to mouse mm9 genome using Bowtie (Langmead et al., 2009) using the parameters -m = 4 , -k = 1, and -best, thus allowing the 20 basepair reads to map to at most 4 locations, but reporting only the single, best alignment. This option was chosen to allow reads to map in duplicated regions, while still excluding reads to highly repetitive DNA. The mapped reads from Bowtie were then filtered to retain only those reads successfully mapped to the mm9 genome.

1.2 Algorithm for DNase hotspot and peak detection and DNase cut density quantitation

Several algorithms have been developed for the detection of DNase hypersensitive sites (DHS) (Madrigal and Krajewski, 2012), but they produce only partially concordant results (Koohy et al., 2014). To enhance the accuracy of our DHS detection, we required that our DHS's must be called by two of the major publicly available methods, DNase2Hotspots (Baek et al., 2011) and F-Seq (Boyle et al., 2008), as described below.

1.2.1 DNase2Hotspots

Using DNase2Hotspots as described in (Baek et al., 2011), we initially called hotspots using mapped reads pooled from biological triplicates for each experimental condition. After testing a range of z-score thresholds, we decided on a final stringent threshold of 0% FDR that still captured many mildly DNase-sensitive areas. From within hotspots that meet the 0% FDR threshold, we called DNase peaks further enriched from the surrounding hotspot, using z-score $>$20 as the threshold.

1.2.2 F-Seq

Reads pooled from biological triplicates for each experimental condition were converted to bed files that were used as input to F-Seq (Boyle et al., 2008), which generates smoothed signals based on density of reads mapped, normalized by total mapped reads in the library. From these pooled reads we used F-Seq to identify a set of top 100,000 read-enriched regions.

For visualization in browser tracks, genome-wide signal tracks produced by F-Seq for individual replicates as well as the reads pooled from replicates. Note that F-Seq signal is linearly proportional to RPKM. The tracks are available for viewing on the PSU Genome Browser via the links provided in the Data Access section.

1.2.3 Obtaining DNase cut densities from the final set of hotspots and peaks that meet the criteria of both DNase2Hotspots and F-Seq

To reduce false positive rates, we filtered for hotspots obtained from DNase2Hotspots that overlap by at least 1bp with one of the top 100,000 enriched regions identified by F-Seq. The final set of hotspots is defined as the union of the F-Seq filtered sets of hotspots identified in each experimental condition (G1E interphase and mitosis, and G1E+GATA1 interphase and mitosis), and the final set of peaks was obtained analogously by taking the union of peaks identified. Additionally, we filtered out hotspots that intersect a small set of Penn State University "black-listed" regions previously found to exhibit high number of reads mapped that are artifacts. "DNase cut density" or "DNase read density" in the main text and figures refer to F-Seq smoothed signals (which are linearly proportional to RPKM) from each experimental condition contained within these final sets of individual hotspots and peaks.

1.2.4 Considerations for normalization and quantitative interpretation of DNase read densities

Within-sample and inter-sample technical variations are issues that must be considered for quantitative interpretations of DNase-seq results.

Within-sample variations between genomic regions can result from intrinsic biases in sequence preferences for DNase I cleavage , and/or sequencing biases. Our analyses of changes in DNase cut density changes, which compare the read density changes within the same region across conditions, are not influenced by these types of biases as they should affect the same genomic region equally across conditions.

We accounted for inter-sample variations arising from differences in sequencing depth, which are relatively minor (Table 1), by using the smoothed signal obtained from F-Seq, which are linearly proportional to RPKM. Furthermore, F-Seq signals from individual biological replicates indicated that the cross-replicate variances (shown as SEM by error bars in the main figures where relevant) are minimal, compared to the effects we drew conclusions upon, so we did not perform additional normalization procedures for the analyses in the figures presented. All trends presented are qualitatively the same if analyses are performed on quantile-normalized read densities (data not shown).

1.3 ChromHMM genome segmentation and promoter vs. distal CRM classification

ChromHMM (Ernst and Kellis, 2012) was applied on the ChIP-seq data of five histone modifications in five mouse hematopoietic cell types to learn a multivariate HMM model for segmentation of mapped genome in each cell type. Specifically, the ChIP-seq mapped reads were first pooled from replicates for each of the five histone modifications (H3K4me3, H3K4me1, H3K36me3, H3K9me3 and H3K27me3) in the five mouse cell types (G1E, G1E+GATA1, erythroblasts, megakaryocytes, and CH12). These mapped reads were first processed by ChromHMM into binarized data in every

200 bp window over the entire mapped genome, with ChIP input reads as the background control. Then the model was learned from the binarized data in all five cell types, giving a single model with a common set of emission parameters and transition parameters, which was then used to produce segmentations for G1E and G1E+GATA1 based on the most likely state assignment of the model (segmentations for the other cell types are not used for this study). We tried models with different numbers of states and selected a nine-state model as it appeared that all nine states have clearly distinct emission properties, while the interpretability of distinction between states in models with additional states was less clear. Emission probabilities for histone modifications for each state are shown in Fig. S8. Browser views of chromatin states are shown in Fig. S3.

The following definitions were used for the classification of CRMs. Promoter hotspots and peaks are defined as those that overlap a previously annotated TSS, or those for which coverage by State 5 in either G1E or G1E+GATA1 occupies >90% of the bases within the hotspot or peak. Distal CRM hotspots or peaks are those that do not overlap a known TSS, and for which the coverage by State 2, State 3, and State 4 in either G1E or G1E+GATA1 occupies >90% of the bases within the hotspot or peak.

1.4 Tissue distribution of DNase hotspots, Gene Ontology analyses, and overlap with large DNA hypomethylation domains

DNase hotspots from our data sets were intersected (defined as at least 1bp overlap) with a master list of DHS's from 45 cell or tissue types from the Mouse ENCODE consortium (J Vierstra, E Rynes, R Sandstrom, R Thurman, M Zhang, T Canfield, P Sabo, R Byron, R Hasen, A Johnson, et al., submitted), and the number of cell or tissue types in which a given DNase hotspot overlaps at least one DHS called was obtained. The list of 45 cell or tissue types used in this analysis are the following: mPATSKI, mESC_F4WT, mCD19, E14, mKidney, m3134, mA20, mRetina, mGER, mFatPad, mESC_FIKO, CH12, mCJ7, mThymus, mB, mMuscle_Skeletal, mMesoderm, mGenitalFatPad, mCD1, mFibroblast, mHindlimbsBuds, mfLiver, mCerebrum, mTN, ZhBTc4, mfLiver_F5, mfLiver_F2, mfLiver_F3, mfLiver_F1, mLung, mForelimbBuds, mLiver, mSpleen, mIntestine_Large, mATr, mfLiver_DLCR, mTR, mBrain, m416B, mCerebellum, mfBrain, MEL, mATn, mHeart, mNIH_3T3.

For Gene Ontology (GO) analysis, the Gene Regions Enrichment of Annotations Tool (GREAT version 2.0.2, McLean et al. (2010)) was used to assign each promoter hotspot to the single nearest gene within 1000kb, and enrichment of GO terms over the a background set of all G1E+GATA1 promoter hotspots was investigated for several sets of target regions: 1) promoter hotspots with the G1E+GATA1 mitosis-to-interphase ratio >85%, 2) promoter hotspots present in ≥10 cell or tissue types, and 3) promoter hotspots G1E+GATA1 mitosis-to-interphase ratio >85% and present in at ≥10 cell or tissue types. For Fig. 5B, gene-GO term associations were obtained from AmiGO2 (`http://amigo.geneontology.org/amigo`), and the fraction of all genes that belong to GO terms "sequence-specific DNA binding transcription factor activity" (GO.0003700) and "signaling receptor activity" (GO.0038023) were plotted for each subset of promoter hotspots examined. For all GO analyses, essentially the same

results were obtained using G1E promoter hotspots (data not shown).

1.5 Association of hotspots and peaks with gene expression

DNase hotspots and peaks were assigned to genes that they overlap, or if they do not overlap a single gene then they are assigned to the single nearest gene using the closest command from bedtools v2.19.0. Hotspots and peaks that overlap multiple genes are excluded from the following analysis. Of the genes whose TSS overlap hotspots and have an RNA-seq (Tejaswini et al., 2014, submitted) log2(FPKM)$>$3 in either G1E or G1E+GATA1, the top 100 up-regulated and top 100 down-regulated genes from G1E to G1E+GATA1 were identified. The promoter and distal CRM hotspots and peaks assigned to these most differentially expressed genes were examined in Fig. S14 and Fig. S15.

1.6 Plotting and graphics

Plots were rendered in R version 3.0.2 (R Core Team, 2013) using the ggplot2 (Wickham, 2009), plyr (Wickham, 2011), and reshape2 (Wickham, 2007) packages. Bins for moving means were obtained using the cut2 function from the Hmisc package (Harrell, 2014).

References

Baek, S., Sung, M.-H., and Hager, G. L., 2011. Quantitative Analysis of Genome-Wide Chromatin Remodeling. pages 433–441. Humana Press, Totowa, NJ.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J., 2001. *Galaxy: A Web-Based Genome Analysis Tool for Experimentalists*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Boyle, A. P., Guinney, J., Crawford, G. E., and Furey, T. S., 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**(21):2537–2538.

Ernst, J. and Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, **9**(3):215–216.

Giardine, B., 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, **15**(10):1451–1455.

Goecks, J., Nekrutenko, A., and Taylor, J., 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, .

Harrell, F., 2014. *Hmisc: Harrell Miscellaneous*. R package version 3.14-4.

Kadauke, S., Udugama, M. I., Pawlicki, J. M., Achtman, J. C., Jain, D. P., Cheng, Y., Hardison, R. C., and Blobel, G. A., 2012. Tissue-specific mitotic bookmarking by hematopoietic transcription factor GATA1. *Cell*, **150**(4):725–737.

Koohy, H., Down, T. A., Spivakov, M., and Hubbard, T., 2014. A comparison of peak callers used for DNase-seq data. *PLoS ONE*, .

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 Biol*, .

Madrigal, P. and Krajewski, P., 2012. Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Frontiers in genetics*, **3**:230.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G., 2010. McLean et al. - 2010 - GREAT improves functional interpretation of cis-regulatory regions. - Nature biotechnology. *Nature Biotechnology*, **28**(5):nbt.1630–9.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Wickham, H., 2007. Reshaping data with the reshape package. *Journal of Statistical Software*, **21**(12):1–20.

Wickham, H., 2009. *ggplot2: elegant graphics for data analysis*. Springer New York.

Wickham, H., 2011. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, **40**(1):1–29.