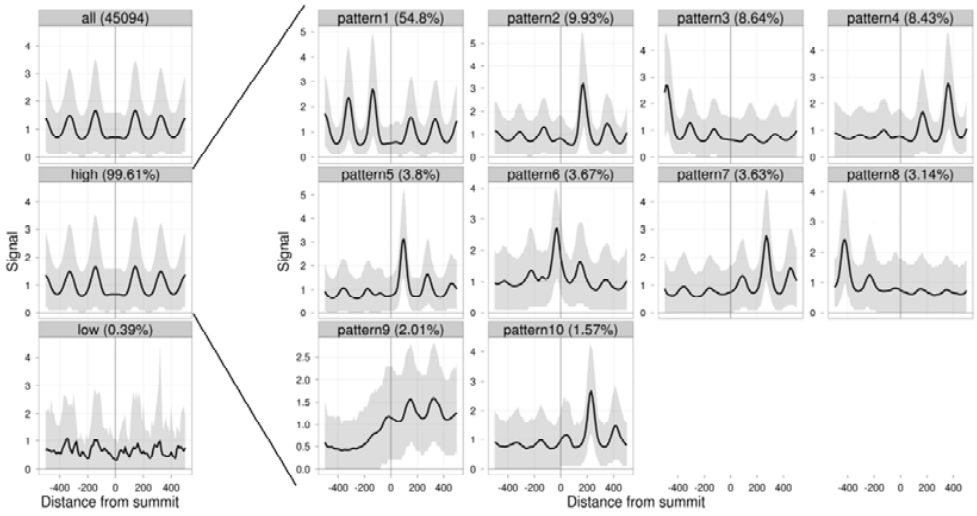
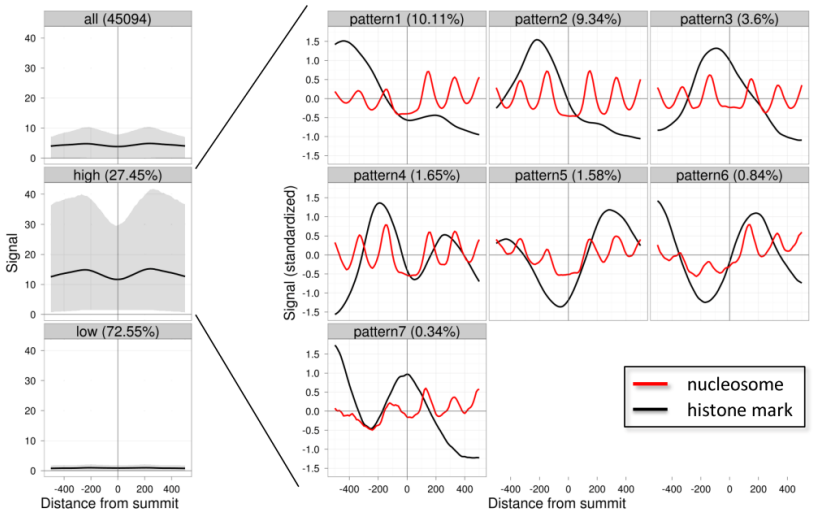
doi:10.1038/nature11247

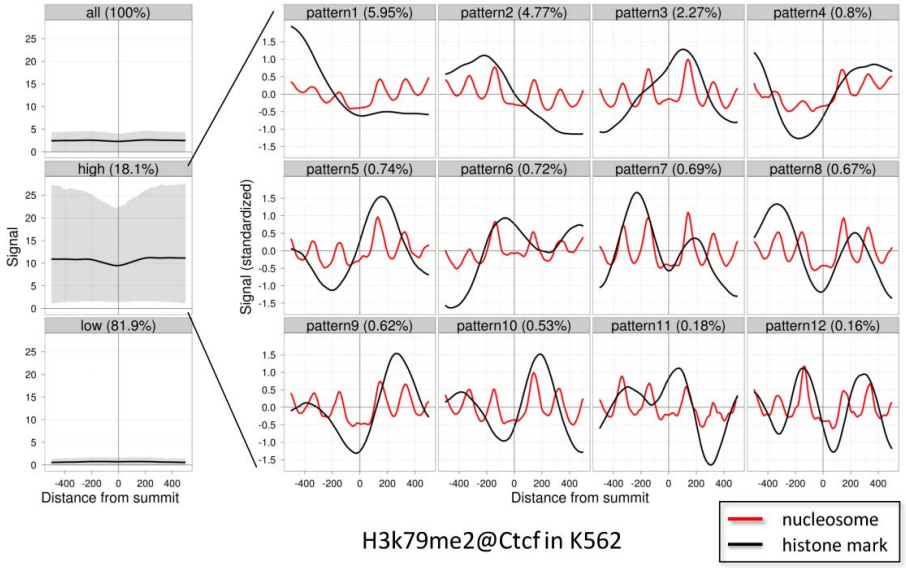


Supplementary Figure 1, section E: nucleosomes@CTCF in K562
We ran the CAGT pipeline using nucleosome sequencing data (MNase-seq) as the target mark around CTCF sites in K562. We see that 99.6% of the profiles belong to the high signal category indicating that almost all CTCF peaks have significant nucleosome occupancy and well-positioned nucleosomes flanking them. However, we also observe distinct diverse shapes of nucleosome positioning, the largest cluster (pattern 1) being largely symmetric and the remaining clusters showing various asymmetric patterns.



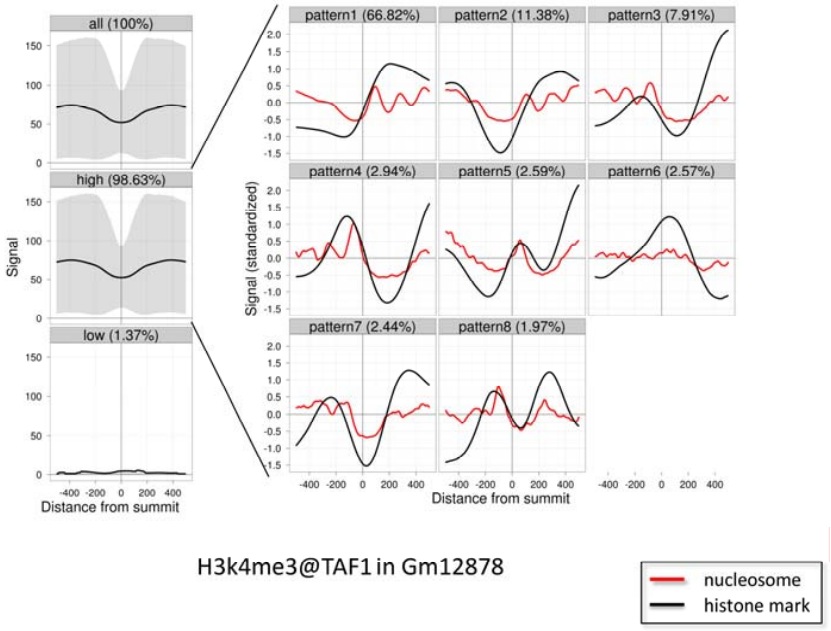
Supplementary Figure 2, section E: H3K9ac@CTCF and corresponding nucleosomes occupancy in K562

We used CAGT to analyze H3k9ac profiles at CTCF peaks in K562. We observe that only 27.45% of all CTCF peaks are enriched for H3k9ac. We then wanted to analyze the relationship of shape patters of H3k9ac to corresponding nucleosome occupancy profiles. Rather than averaging profiles that belong to each shape cluster using the original scale, we standardize each H3k9ac profile and then compute the median of the standardized profiles that belong to each shape cluster (shown as black lines in the pattern subplots). We then extract nucleosome occupancy profiles for all peaks that belong to each H3k9ac pattern cluster; standardize them; flip/reverse profiles around peaks whose corresponding H3k9ac profiles were flipped in the CAGT analysis and compute the median nucleosome occupancy profile. We observe that while the H3k9ac profiles are largely asymmetric, the nucleosome occupancy profiles show strong symmetry with well-positioned nucleosomes on either side of the CTCF. This indicates that histones on either side of the CTCF site tend to be differentially marked with H3k9ac.



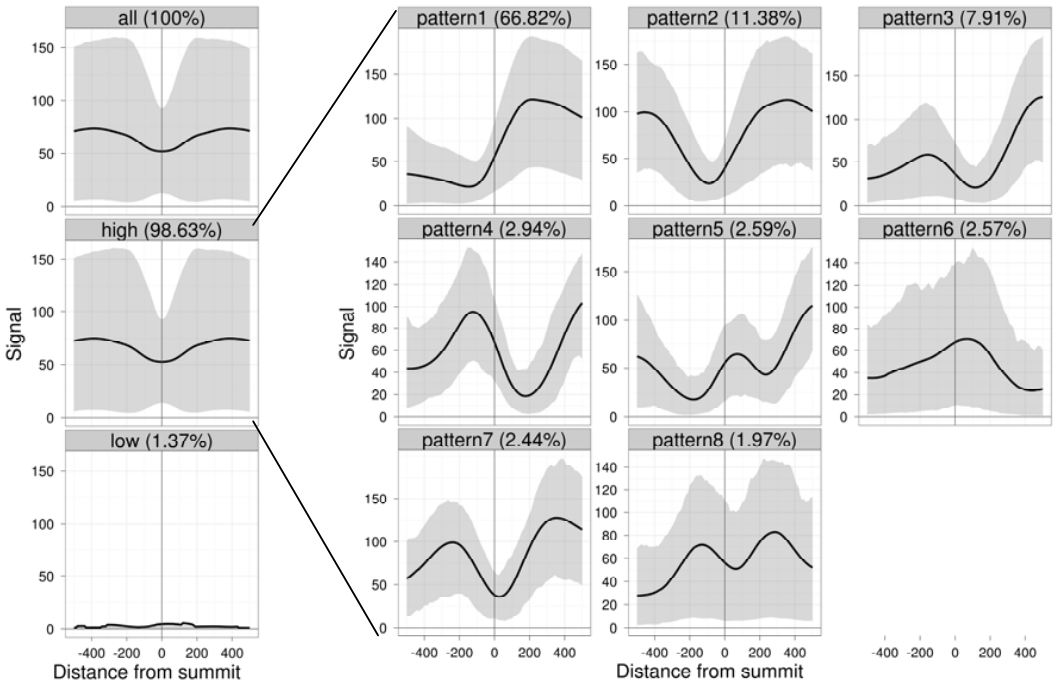
Supplementary Figure 3, section E: H3K79me2@CTCF and corresponding nucleosome occupancy

We used CAGT to analyse patterns of H3k79me2 at CTCF peaks in K562. Only 18% of the CTCF peaks are enriched for H3k79me2. We observe asymmetric shapes of H3k79me2 but symmetric positioning and occupancy of nucleosomes. Only ~50% of these peaks are within 5Kbp of TSS and > 80% are within GENCODEv7 gene boundaries (which is in concordance with the enrichment of H3k79me2 in actively transcribed domains). Hence, proximity to TSSs does not entirely explain the pattern asymmetry of H3k79me2.

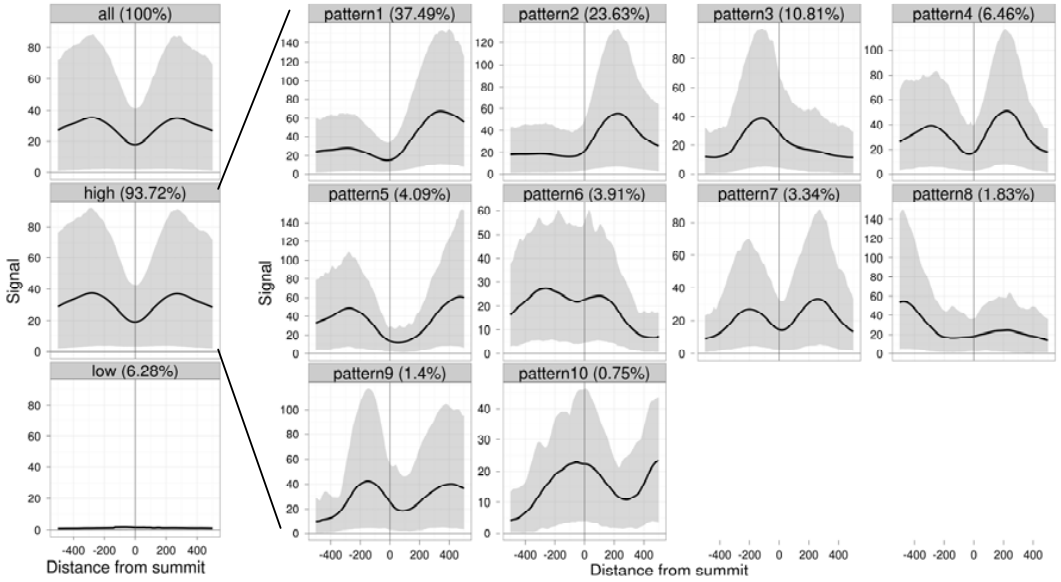


Supplementary Figure 4, section E: H3k4me3@TAF1 and corresponding nucleosome occupancy in GM12878

For TSS-proximal TFs such as TAF1 and for the histone mark H3k4me3 which is typically enriched at active promoters, we observe that the asymmetry of H3k4me3 is strongly correlated with asymmetry of corresponding nucleosome occupancy patterns.



Supplementary Figure 5, section E: H3k4me3@TAF1 in GM12878
CAGT analysis of H3k4me3 patterns at TAF1 peaks in GM12878 shows that a majority of high signal profiles are asymmetric.



Supplementary Figure 6, section E: H3k27ac@GATA1 in K562.

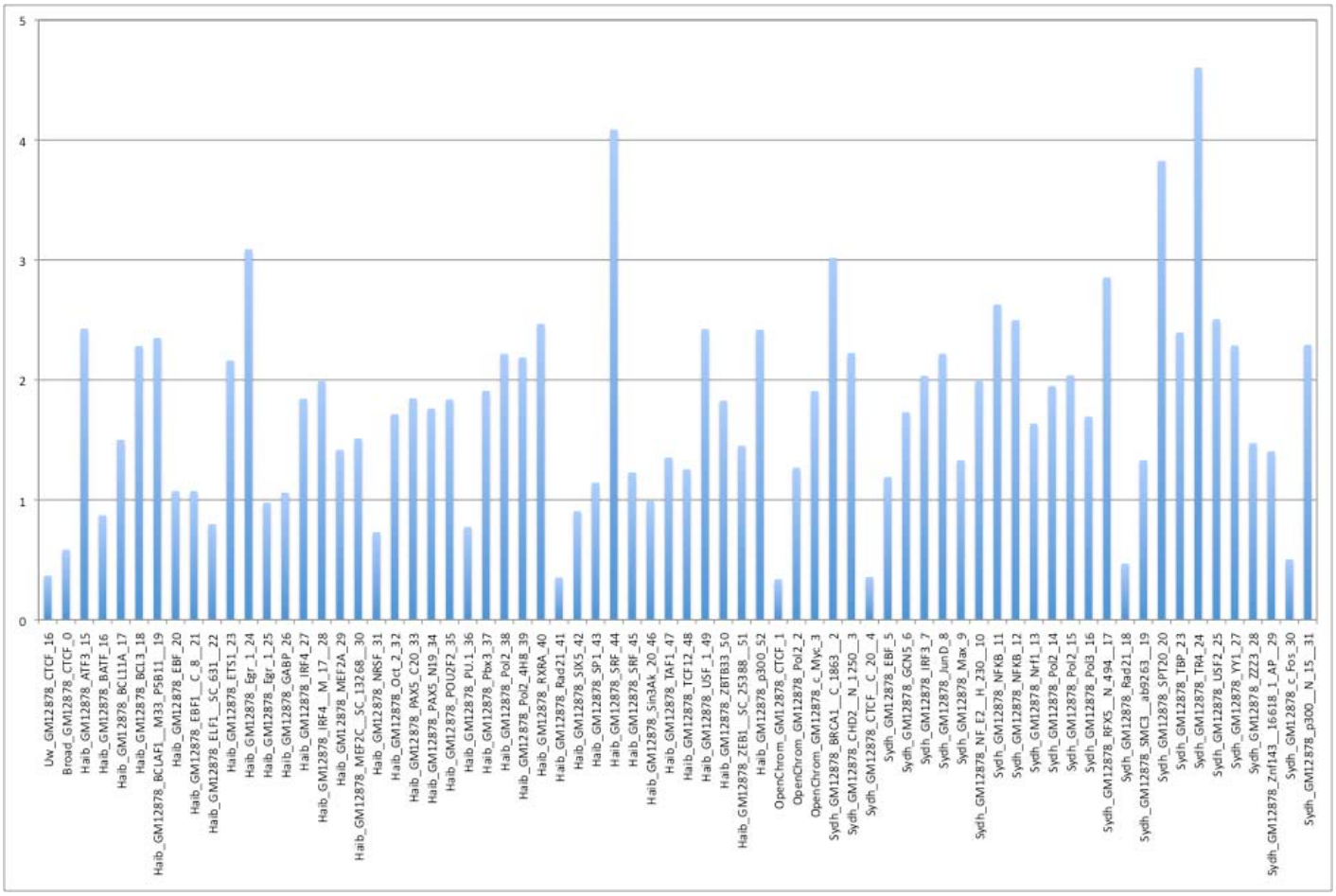
CAGT analysis of H3k27ac at GATA1 peaks in K562 shows that while 94% of the peaks show significant enrichment of H3k27ac, most of these profiles are highly asymmetric. Only 21% of the peaks are proximal to annotated GENCODEv7 TSSs and yet > 70% of the high signal profiles of H3k27ac around GATA1 sites are highly asymmetric. Hence, proximity to TSSs cannot be the only factor driving asymmetry of chromatin patterns.

Cluster	Profiles	Profile with motif	Flipped profiles	Motif (+) and no_flip	Motif (+) and flip	Motif (-) and no_flip	Motif (-) and flip	pval
1	5658	2416	3015	517	715	604	580	0.999
2	1607	674	808	197	178	133	166	0.0155
3	885	288	417	71	58	86	73	0.3898
4	593	252	325	49	80	60	63	0.9454
5	313	131	155	27	36	41	27	0.9658
6	308	110	140	23	34	26	27	0.7660

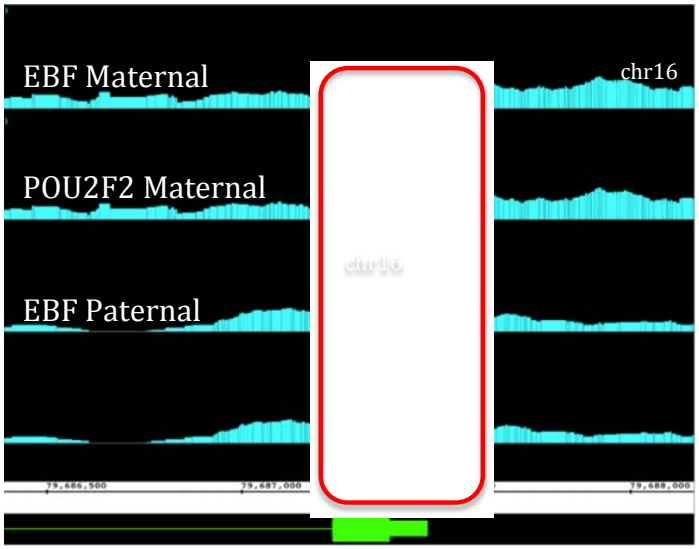
Supplementary Table 1, section E:
Contingency table for correlation of profile flipping and strandedness of motif hits for shape groups of H3k27me3@CTCF in H1hesc. The columns from left to right are as follows: "Cluster" - the pattern shape id. "Profiles" – number of high signal profiles that belong to the cluster. "Profiles with motif" – number of high signal profiles in each cluster with bonafide motif hits. "Flipped profiles" – number of high signal profiles that were flipped in each cluster. "Motif (+) and no_flip" – number of high signal profiles with motif hits on the + strand that were not flipped. "Motif (+) and flip" – number of high signal profiles with motif hits on the + strand that were flipped. "Motif (-) and no_flip" – number of high signal profiles with motif hits on the - strand that were not flipped. "Motif (-) and flip" – number of high signal profiles with motif hits on the - strand that were flipped.

Cluster	Profiles	Profile with motif	Flipped profiles	Motif (+) and no profile flip	Motif (+) and profile flip	Motif (-) and no profile flip	Motif (-) and profile flip	pval
1	1044	78	555	27	23	14	14	0.2823
2	658	47	294	8	8	17	14	0.5034
3	301	18	179	5	6	4	3	N/A
4	180	13	90	2	4	4	3	N/A
5	114	7	55	2	2	3	0	N/A
6	109	7	45	3	2	1	1	N/A
7	93	8	53	6	2	0	0	N/A

Supplementary Table 2, section E:
Contingency table for correlation of profile flipping and strandedness of motif hits for shape groups of H3k27ac@GATA1 in K562. The column labels are the same as Supplementary Table 1, section E.



Supplementary Figure 1, section K: In this figure we plot the percentage of maternal and paternal specific peaks (i.e. peaks not detected using the reference genome GRCh37) for ChIP-Seq datasets in GM12878. We see that on average ~2% of peaks are either maternal or paternal specific when using the maternal or paternal haplotypes for NA12878 as a reference.



Supplementary Figure 2, section K:
In this figure we show an example of a binding site on chromosome 16 that is present using reads mapped to the paternal allele for both POU2F2 and EBF however not for the maternal allele for GM12878. In the region of the binding site (indicated in red) there are however no sequence variants that differ between the maternal and paternal haplotypes. The difference is due to a maternal specific insertion on chromosome 1 that causes the reads in this vicinity to not be uniquely mapping on the maternal allele.

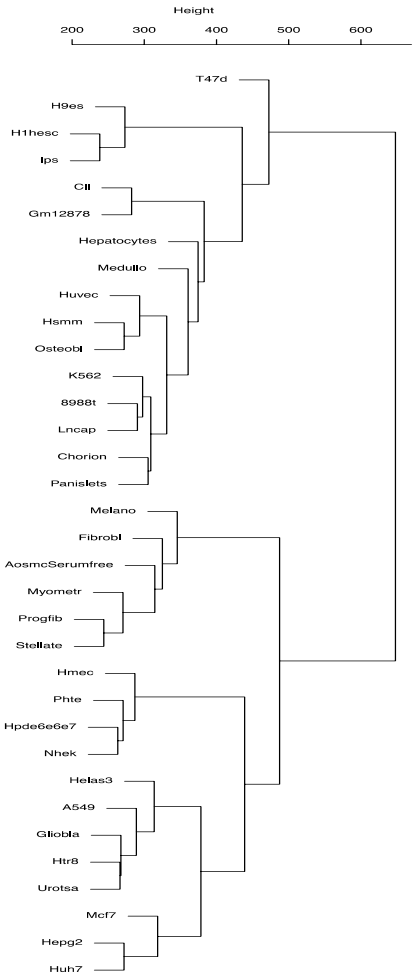
Supplementary Table 1, section K:

			Genes		Pseudogenes	NC RNAs	Binding Sites
			all transcripts	at least one transcript			Total
All variants in NA12878	Variants overlapping an Annotation	Homozygous	4109	7010	7033	6209	159002
		SNPs	4037	6845	6322	5552	143050
		indels	71	163	698	649	15887
		deletions	1	2	13	8	65
		Heterozygous	6722	11427	12728	10350	259597
		SNPs	6611	11211	11476	9554	234801
		indels	102	204	1232	769	24428
		deletions	9	12	20	27	368
	Variants having a potentially functional effect	Homozygous	1897	3353	NA	NA	7131
		SNPs	1873	3276	NA	NA	6234
		indels	24	77	NA	NA	897
		Heterozygous	3149	5547	NA	NA	11909
		SNPs	3100	5462	NA	NA	10161
		indels	49	85	NA	NA	1748
Rare variants in NA12878	Variants overlapping an Annotation	Homozygous	24	366	101	62	2247
		SNPs	24	364	22	36	505
		indels	0	1	78	26	1737
		deletions	0	1	1	0	5
		Heterozygous	29	279	540	397	19235
		SNPs	16	246	347	308	12185
		indels	13	32	192	84	7009
		deletions	0	1	1	5	41
	Variants having a potentially functional effect	Homozygous	12	23	NA	NA	471
		SNPs	12	22	NA	NA	351
		indels	0	1	NA	NA	120
		Heterozygous	218	372	NA	NA	616
		SNPs	212	353	NA	NA	364
		indels	6	19	NA	NA	252

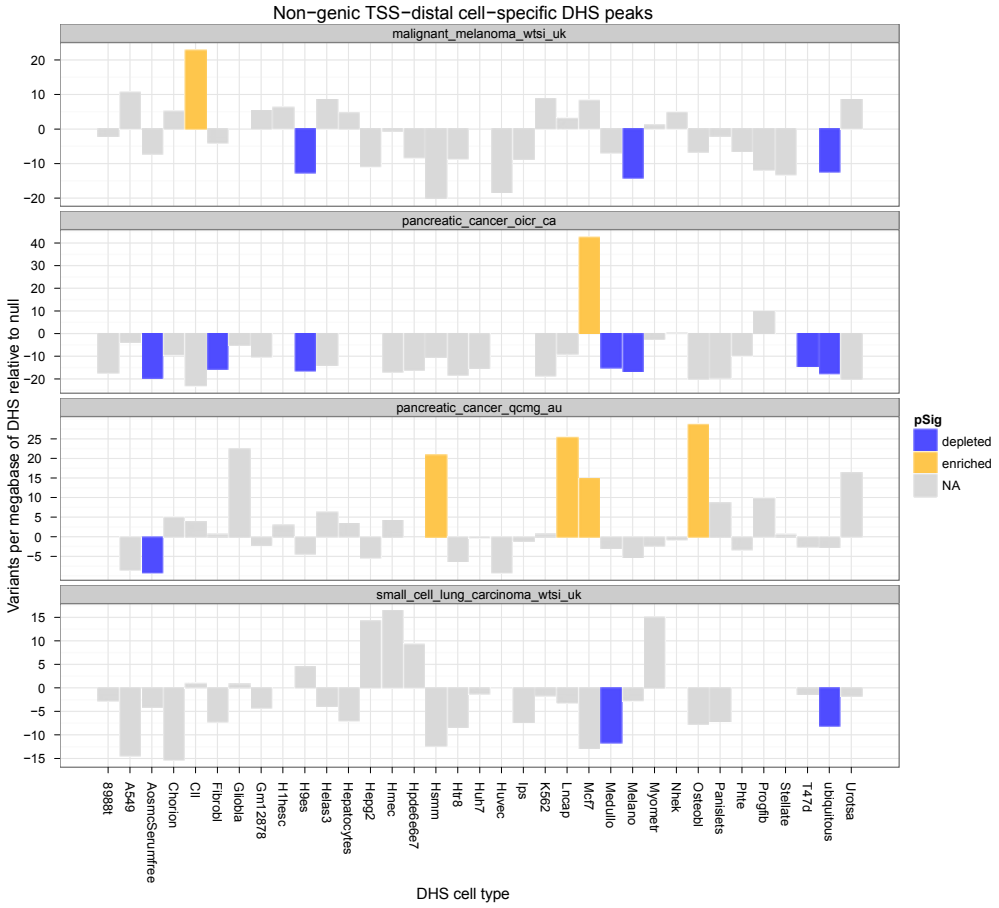
In this table we list the variant that overlap annotation detected by ENCODE. Annotations are either GENCODE genes, pseudogenes, non-coding RNAs or TF binding sites. Variants for NA12878 (from the 1000 Genome Project) are subdivided into all variants and rare variants as well as those that simply overlap an annotation and those that are likely to have a functional effect on an annotation: either causing loss of function (premature stop, frame shift or disrupting a splice site) of a non-synonymous SNP for the case of a protein coding genes or overlapping a identified TF binding motif with a binding site.

			Genes		Peudogenes	NC RNAs	Binding Sites
			all transcripts	at least one transcript			Total
All variants in NA12878	Annotations overlapping a variant	Homozygous	1634	1843	2364	2264	2452
		SNP	1609	1798	2220	1858	2098
		indel	58	112	452	400	353
		deletion	1	1	13	6	1
		Heterozygous	3511	4022	1798	4876	2410
		SNP	3457	3958	1756	4252	1975
		indel	84	126	405	608	432
		deletion	6	5	20	16	3
	Annotations potentially affected by a variant	Homozygous	420	1300	NA	NA	420
		SNP	100	480	NA	NA	350
		indel	330	850	NA	NA	60
		Heterozygous	690	1870	NA	NA	660
		SNP	270	1110	NA	NA	570
		indel	460	810	NA	NA	80
Rare variants in NA12878	Annotations overlapping a variant	Homozygous	190	210	1190	580	540
		SNP	180	180	350	220	130
		indel	10	50	790	360	410
		deletion	0	10	10	0	0
		Heterozygous	2480	3330	3130	3140	2470
		SNP	2410	3140	2420	2320	1750
		indel	90	200	1310	790	700
		deletion	0	10	10	30	20
	Annotations potentially affected by a variant	Homozygous	100	400	NA	NA	100
		SNP	0	0	NA	NA	0
		indel	100	400	NA	NA	100
		Heterozygous	1000	2200	NA	NA	400
		SNP	600	800	NA	NA	300
		indel	400	1400	NA	NA	100

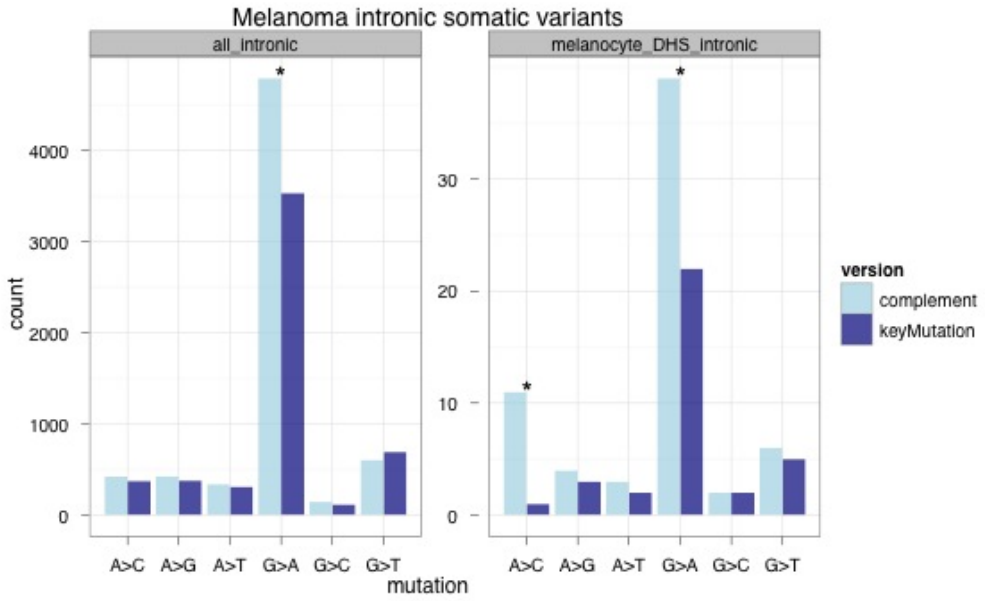
Supplementary Table 2, section K:
This table is the element-centric version of Supp. Table 1, section J. Instead of counting variants that overlap annotations detected by ENCODE by count the number of ENCODE annotations that overlap variants within NA12878.



Supplementary Figure 1, section L:
A hierarchical tree based on DHS signature Euclidean distance among 34 different cell types.



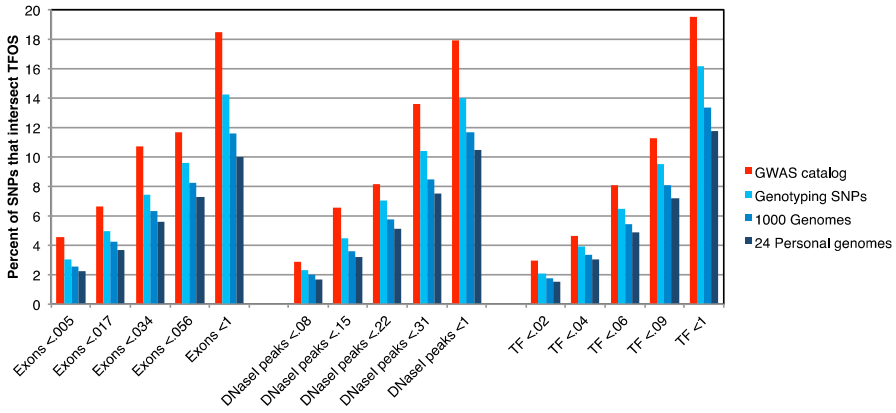
Supplementary Figure 2, section L:
Different sets of cancer somatic variants show different levels of enrichment/depletion relative to intergenic TSS-distal cell-type-specific and ubiquitous DHSs.



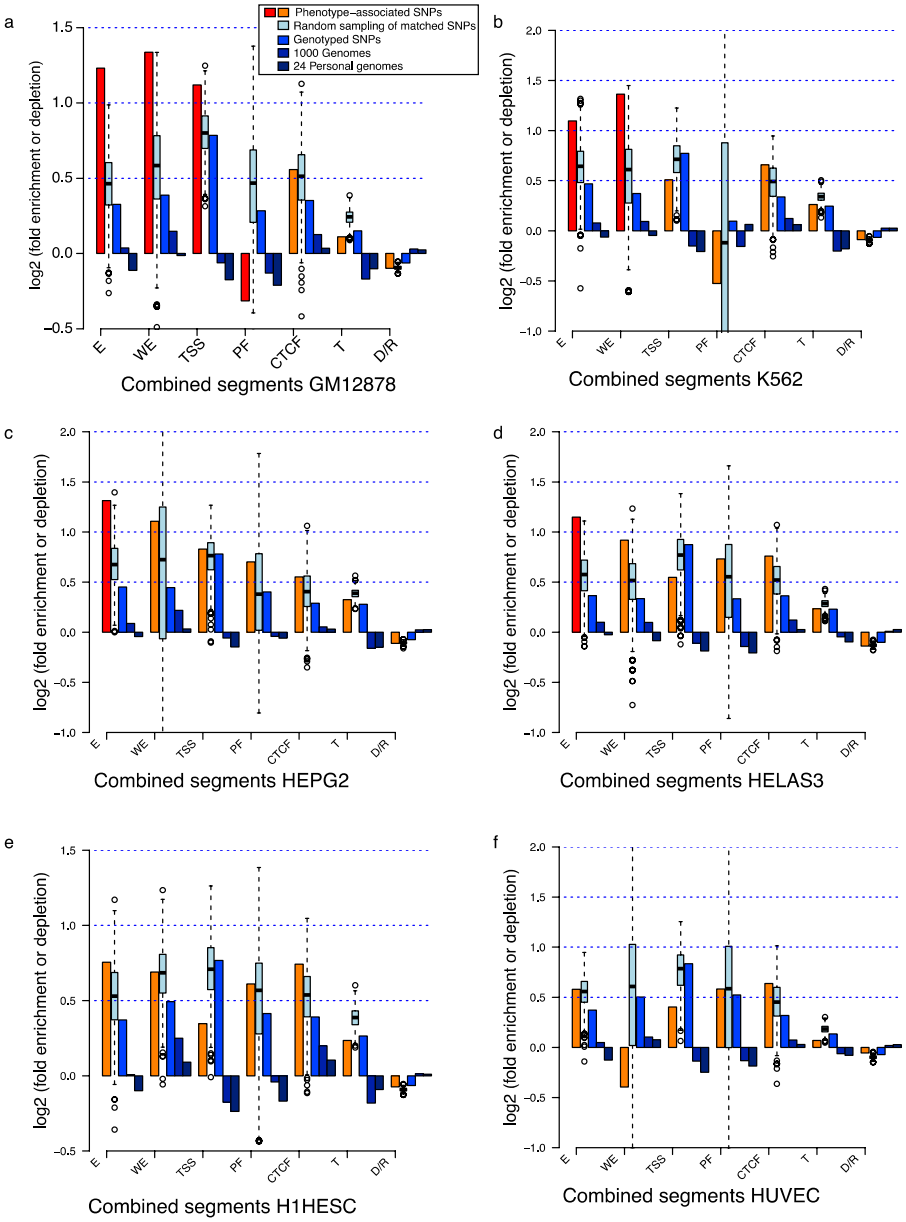
Supplementary Figure 3, section L:
Mutational bias is asymmetric with respect to orientation relative to the transcribed strand. The 12 possible mutations are collapsed into 6 such that the key mutation (A>C, for example) and its complement (T>G) version are represented with different colors. An asterisk (*) represents $P < 0.05$ for a Binomial test on counts for a pair of key and complement mutations. Comparing all intronic variants (left panel) to only those that overlap melanocyte DHSs (right panel) suggests there is a change in the mutational process ($P = 0.06533$; Fisher's exact 2x12 test).

File name	DCC Accession
wgEncodeOpenChromDnase8988tPk.narrowPeak.gz	wgEncodeEH001103
wgEncodeOpenChromDnaseA549Pk.narrowPeak.gz	wgEncodeEH001095
wgEncodeOpenChromDnaseAosmcSerumfreePk.narrowPeak.gz	wgEncodeEH000601
wgEncodeOpenChromDnaseChorionPk.narrowPeak.gz	wgEncodeEH000595
wgEncodeOpenChromDnaseCllPk.narrowPeak.gz	wgEncodeEH001104
wgEncodeOpenChromDnaseFibroblPk.narrowPeak.gz	wgEncodeEH000583
wgEncodeOpenChromDnaseGlioblaPk.narrowPeak.gz	wgEncodeEH001100
wgEncodeOpenChromDnaseGm12878Pk.narrowPeak.gz	wgEncodeEH000534
wgEncodeOpenChromDnaseH1hescPk.narrowPeak.gz	wgEncodeEH000556
wgEncodeOpenChromDnaseH9esPk.narrowPeak.gz	wgEncodeEH000594
wgEncodeOpenChromDnaseHelas3Pk.narrowPeak.gz	wgEncodeEH000540
wgEncodeOpenChromDnaseHepatocytesPk.narrowPeak.gz	wgEncodeEH001107
wgEncodeOpenChromDnaseHepg2Pk.narrowPeak.gz	wgEncodeEH000537
wgEncodeOpenChromDnaseHmecPk.narrowPeak.gz	wgEncodeEH001101
wgEncodeOpenChromDnaseHpde6e6e7Pk.narrowPeak.gz	wgEncodeEH001106
wgEncodeOpenChromDnaseHsmmPk.narrowPeak.gz	wgEncodeEH000584
wgEncodeOpenChromDnaseHtr8Pk.narrowPeak.gz	wgEncodeEH001105
wgEncodeOpenChromDnaseHuh7Pk.narrowPeak.gz	wgEncodeEH001111
wgEncodeOpenChromDnaseHuvecPk.narrowPeak.gz	wgEncodeEH000548
wgEncodeOpenChromDnaseIpsPk.narrowPeak.gz	wgEncodeEH001110
wgEncodeOpenChromDnaseK562Pk.narrowPeak.gz	wgEncodeEH000530
wgEncodeOpenChromDnaseLncapPk.narrowPeak.gz	wgEncodeEH001097
wgEncodeOpenChromDnaseMcf7Pk.narrowPeak.gz	wgEncodeEH000579
wgEncodeOpenChromDnaseMedulloPk.narrowPeak.gz	wgEncodeEH000574
wgEncodeOpenChromDnaseMelanoPk.narrowPeak.gz	wgEncodeEH000602
wgEncodeOpenChromDnaseMyometrPk.narrowPeak.gz	wgEncodeEH000603
wgEncodeOpenChromDnaseNhekPk.narrowPeak.gz	wgEncodeEH000553
wgEncodeOpenChromDnaseOsteoblPk.narrowPeak.gz	wgEncodeEH001098
wgEncodeOpenChromDnasePanisletsPk.narrowPeak.gz	wgEncodeEH000575
wgEncodeOpenChromDnasePhtePk.narrowPeak.gz	wgEncodeEH001099
wgEncodeOpenChromDnaseProgfibPk.narrowPeak.gz	wgEncodeEH000576
wgEncodeOpenChromDnaseStellatePk.narrowPeak.gz	wgEncodeEH001108
wgEncodeOpenChromDnaseT47dPk.narrowPeak.gz	wgEncodeEH001109
wgEncodeOpenChromDnaseUrotsaPk.narrowPeak.gz	wgEncodeEH001113

Supplementary Table 1, section L:
A list of DCC accessions for DNaseI hypersensitive site data sets used in this analysis. (to locate data sets search with the accession at http://genome.ucsc.edu/cgi-bin/hgFileSearch?db=hg19):

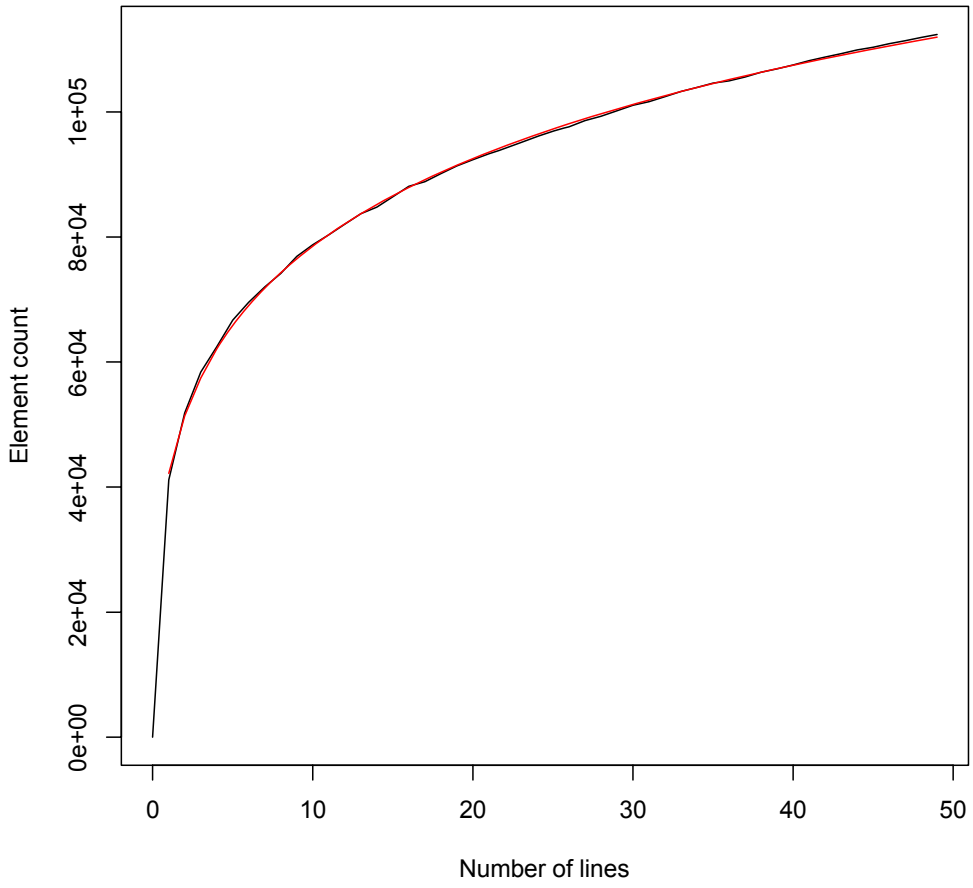


Supplementary Figure 1, section M. Genomic segments enriched for function-associated elements are also enriched for GWAS-identified SNPs. To ascertain whether GWAS SNPs are distributed closer to functional (or predicted functional) regions, the human genome was divided into 500 kb windows, which were then partitioned into quintiles (with equal number of windows in each quintile) based on the fraction of nucleotides in exons (leftmost graphs), DNase hotspots (middle graphs), and TF-bound regions (rightmost graphs). The limiting value for the determinative feature in each quintile is given in the labels on the x-axis. The bars on the y-axis indicate the percent of SNPs in each partition that overlap a TF-bound DNA segment. This is done for four categories of SNPs: GWAS-identified (red), genotyping SNPs from the Illumina 1M array (light blue), SNPs from the 1000 Genomes Project (medium blue), and SNPs extracted from 24 personal genomes (dark blue; see Personal Genome Variants track at http://main.genome-browser.bx.psu.edu[1]). The percent of GWAS-identified SNPs overlapping TF-bound segments is consistently higher than the percent of other SNPs overlapping TF-bound segments for each set of partitions, indicating that the GWAS-identified SNPs are closer to functional regions annotated by ENCODE.



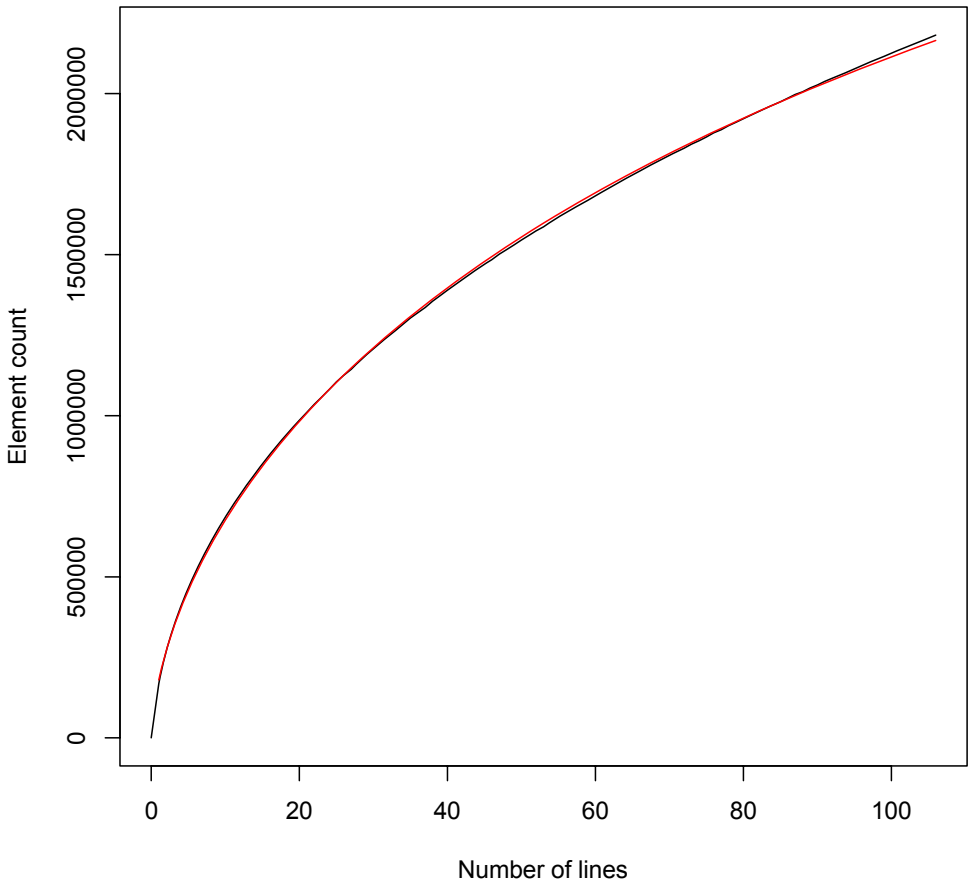
Supplementary Figure 2, section M. Enrichment of Genetic Variants Associated with Complex Traits by GWAS in Segmentations Based on Epigenetic Features. The DNA intervals in the seven classes from combined segmentation (utilizing both chromHMM and Segway) were examined for overlap with GWAS-identified SNPs (NHGRI GWAS SNP catalog June 2011), and levels of enrichment or depletion compared to random expectation were computed. The bars showing the values for GWAS-identified SNPs are colored

red if the enrichment (or depletion) is significant relative to the distribution of results from 1000 matched null sets and orange if otherwise. Each matched null set contains genotyping SNPs matched to the GWAS-identified SNPs by CEU allele frequency, distance from the nearest TSS, and category of genomic location (exon, intron, untranslated region, or intergenic). The distribution of enrichments for the matched null set are shown as light blue bars with bounds at 1.5 times the interquartile range, and any outliers beyond shown as circles. Darker blue bars show levels of enrichment (or depletion) in the segmentation classes for several control SNP datasets: SNPs on the Illumina 2.5M chip as an example of a widely used GWAS SNP typing panel; SNPs from the 1,000 Genomes project; and SNPs extracted from 24 personal genomes (see Personal Genome Variants track at http://main.genome-browser.bx.psu.edu[1]). Each panel shows the results based on segmentations in one of six cell lines. GWAS SNPs are highly enriched in segmentation classes associated with enhancers and transcription start sites across several of the cell types.



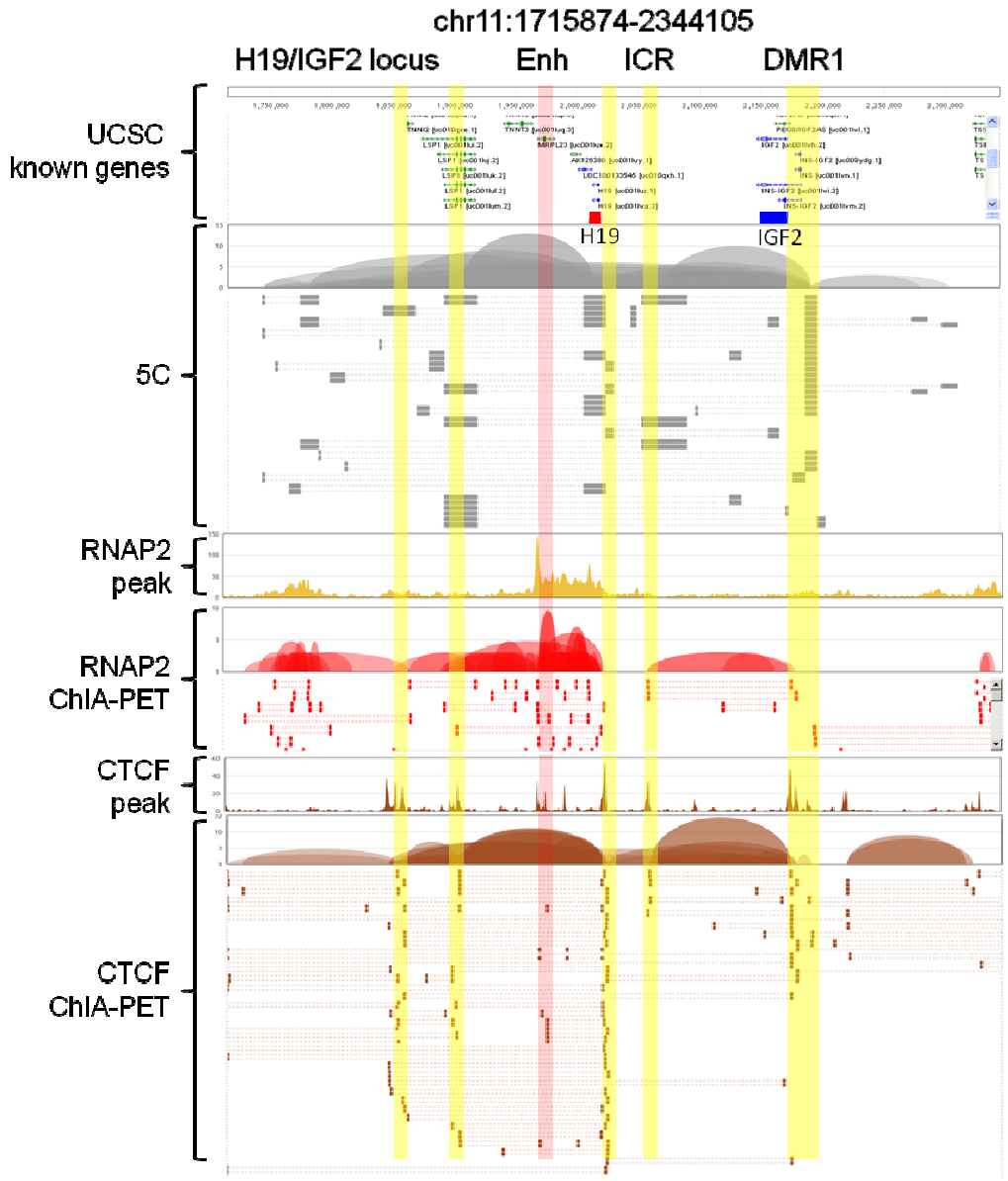
Supplementary Figure 1, section R:

Mean CTCF SPP Optimal IDR peak element count for x cell types after clustering from 1,000 random samples (black line) and fit using the Weibull distribution (red line). Elements are non-overlapping and have maximum length 5000bp.

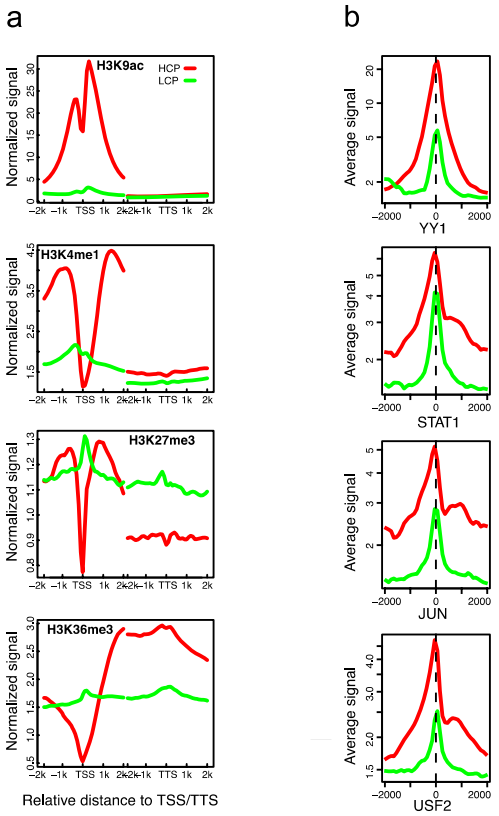


Supplementary Figure 2, section R:

Mean DNase1 element count for x cell types after clustering from 20,000 random samples (black line) and fit using the Weibull distribution (red line). Elements are non-overlapping and have maximum length 5000bp.



Supplementary Figure 1, section Y:
Overlap of 5C and ChIA-PET identified chromatin interactions. An example at the H19/IGF2 locus is shown with 5C interaction tracks, binding peaks and interaction tracks by RNAPII and CTCF ChIA-PET data. Overall, the interaction profiles detected by 5C and ChIA-PET are similar, revealing the sum of interactions captured by 5C and specific interactions by RNAPII and CTCF ChIA-PET. The minor differences between various datasets could reflect some technical deviation such as detection resolution and scope of the methods. The differences between RNAPII and CTCF ChIA-PET might suggest protein factor specificities in interactions.



Supplementary Figure 1, section Z:
Modelling Transciption Levels from Histone Modification and TF-Binding Patterns. Panels a and b show the aggregate behaviour of selected histone marks (Panel a) and TFs (Panels b) over TSSs. In Panel a, H3K9ac, H3K4me1, H3K27me3, and H3K36me3 levels are aggregated over TSSs on the left and transcription termination sites (TTS) on the right in both high CpG TSSs (red) and low CpG TSSs (green). This shows the expected dramatic peaks in activating marks (H3K9ac) over high CpG promoters, with a characteristic dip at the point of initiation, and depletion of H3K4me1 around the immediate TSS region. For the repressive mark H3K27me3, there are higher levels over lower CpG promoters, including the gene body. H3K36me3, which is known to be deposited by the action of transcription elongation, is on the 3' side of TSSs and continues to the TTS. In panel b, four TFs (YY1, Jun, STAT1, and USF2) are shown, again split by high CpG (red) and low CpG (green). They all show a distinct peak over the TSS. Further analysis with respect to other cell lines and RNA measurement types are reported elsewhere[2,3].

References

1 Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943-947, doi:10.1038/nature08795 (2010).

2 Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome research* **22(9)**, doi:10.1101/gr.136838.111 (2012).

3 Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome biology* **13:R53** doi:10.1186/gb‐2012‐13‐9‐r53 (2012).