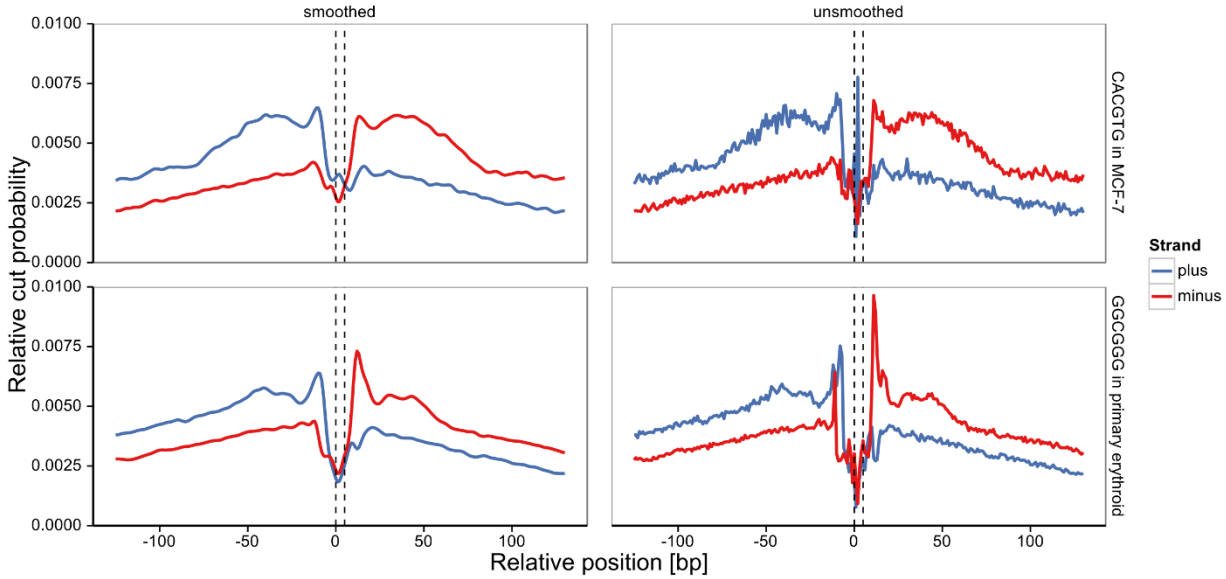
Supplemental Material

Schwessinger et al.

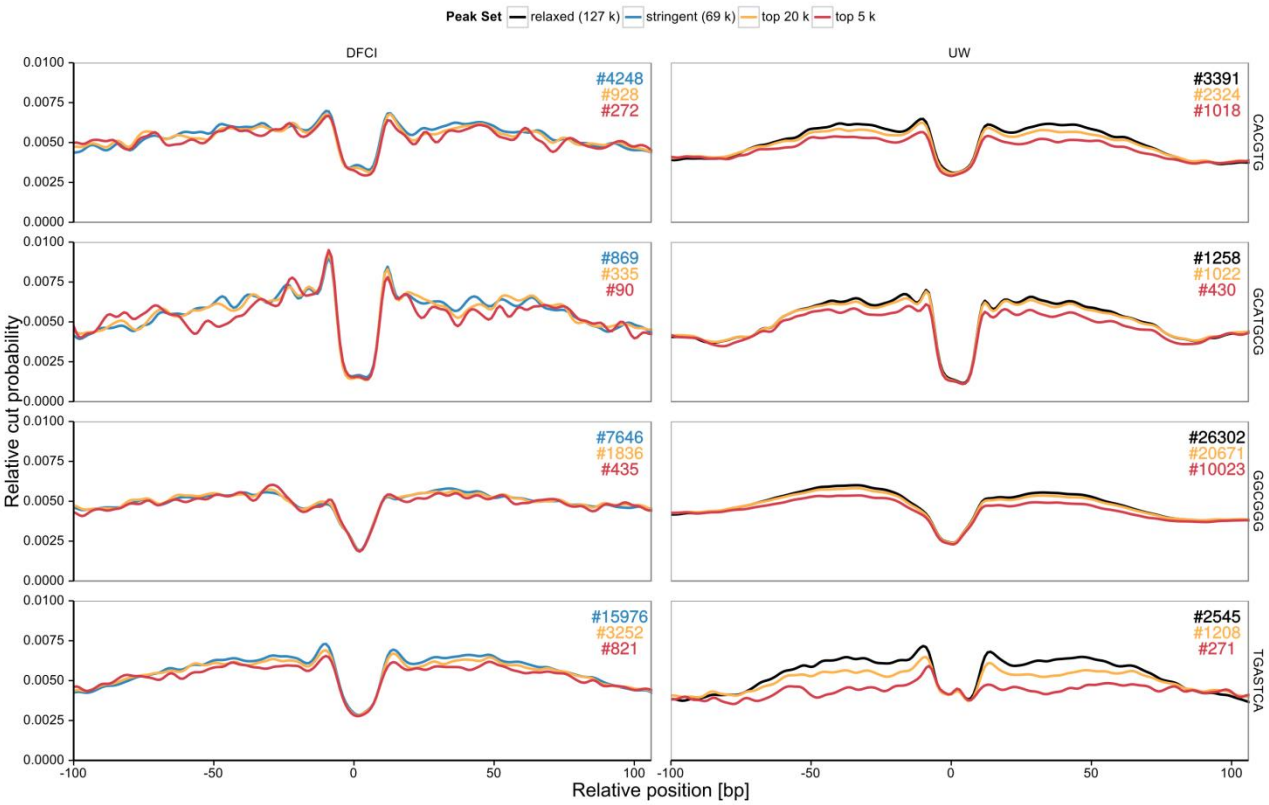
Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell and tissue-specific DNase footprints

Table of Contents

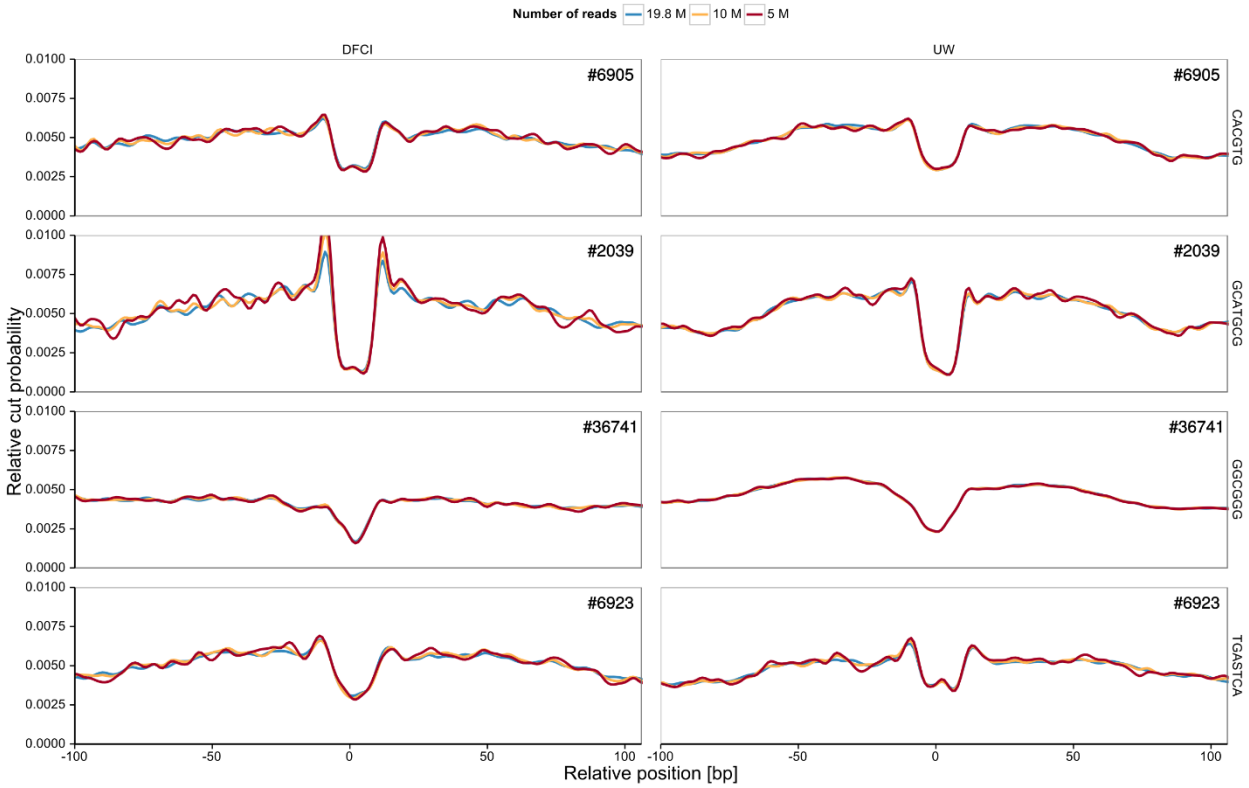
Supplemental Figures



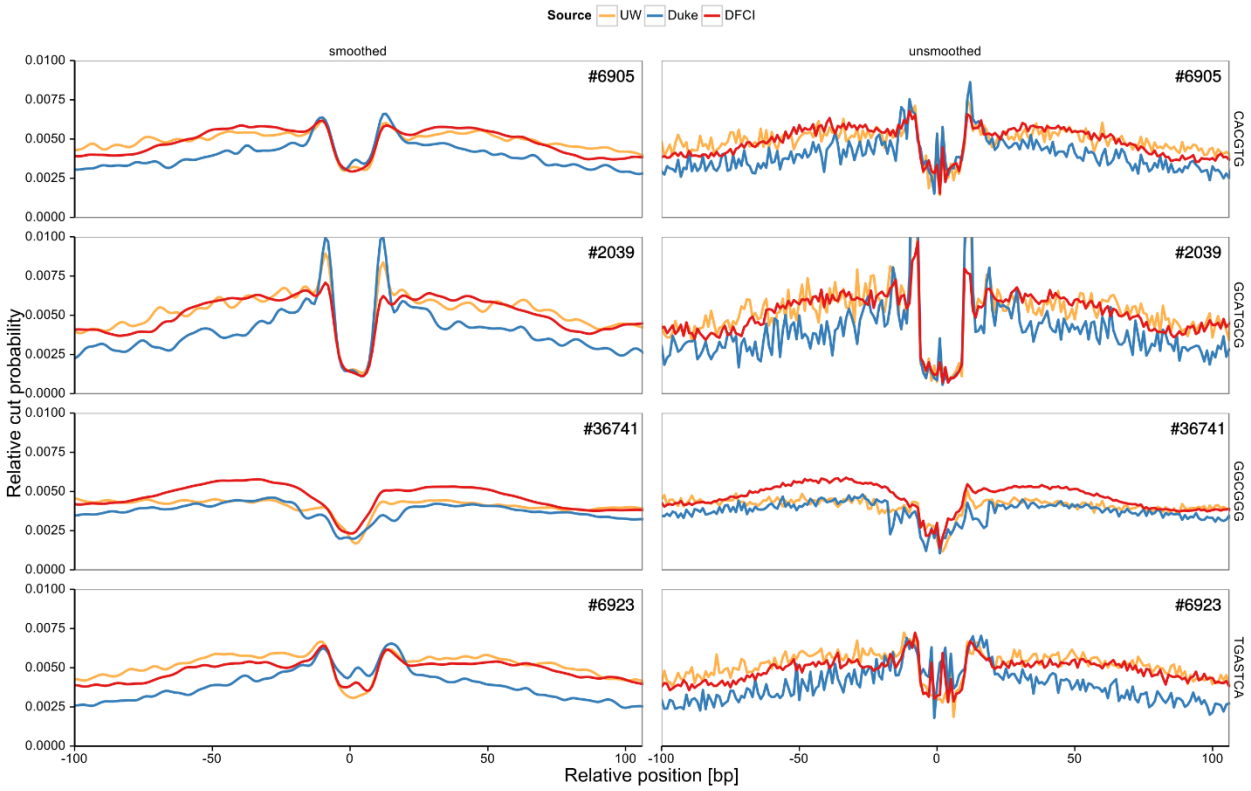
Supplemental Figure S1. DNase-seq strand imbalance in average, *k*-mer based profiles. It has been established that DNase I footprints exhibit a significant strand imbalance[1]. This effect can be observed in the average, *k*-mer based profiles as well, as shown for two *k*-mers associated with strong average footprints: CACGTG and GGCGGG.



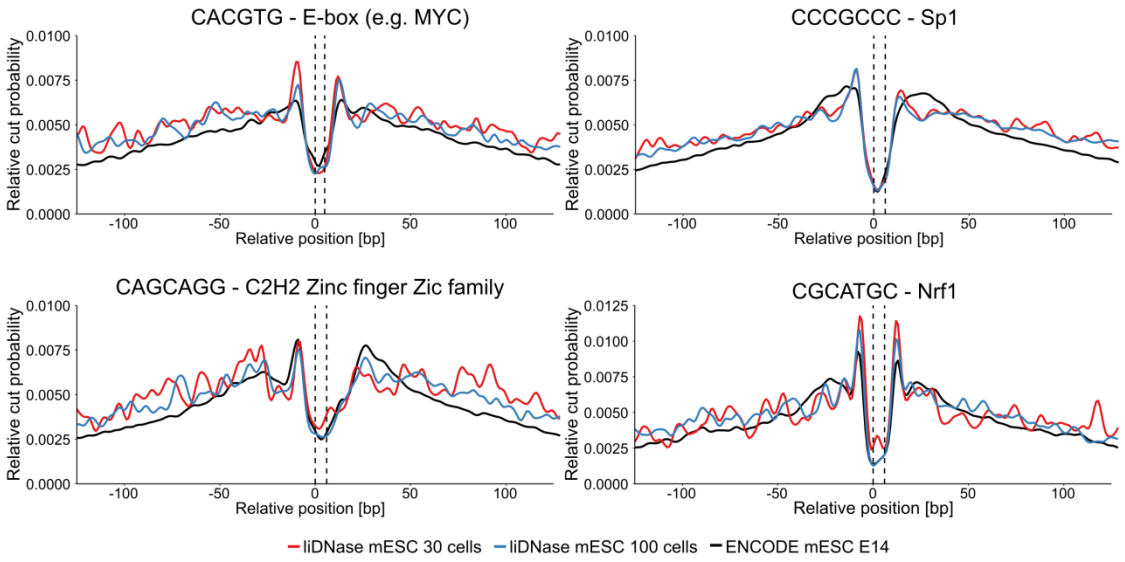
Supplemental Figure S2. Impact of peak number and *k*-mer occurrence on average footprints. Shown are the smoothed relative DNase I cut probability profiles over four *k*-mers associated with housekeeping transcription factors: CACGTG = E-Box (e.g. MYC:MAX); GCATGCG = NRF1; GGCGGG = E2F and three zinc finger Krüppel-related factors (e.g. SP1, E2F4) and TGASTCA = NFE2. DNase-seq data were derived from two different publicly available sources and were processed using down-sampled, equal read depth alignments and different peak sets (relaxed, stringent and top 20 k and top 5 k peaks of the respective set). Consistent with the reduction of peak number the number of *k*-mers occurrences in peaks dropped. The average DNase I footprint shapes are consistent for a decreasing number of peaks and consequently *k*-mer occurrences.



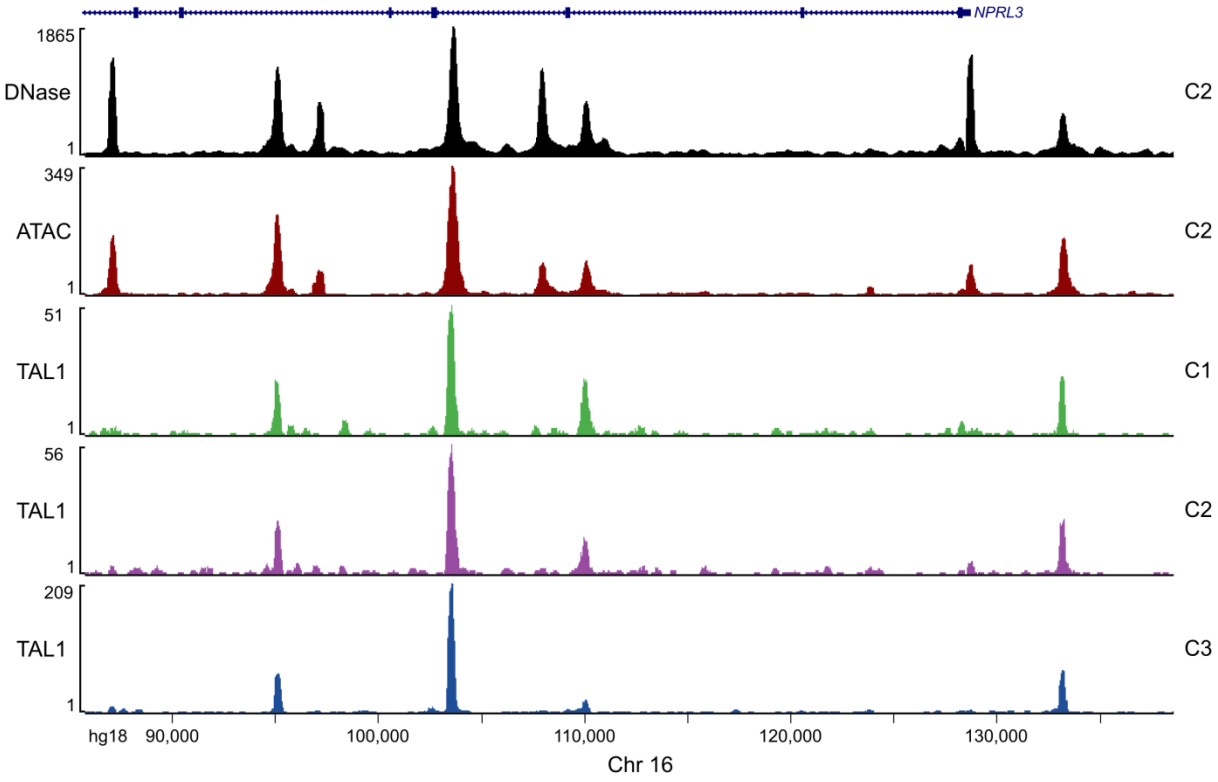
Supplemental Figure S3. Impact of sequencing depth on average footprints. Shown are the smoothed relative DNase I cut probability profiles over four *k*-mers associated with housekeeping transcription factors: CACGTG = E-Box; GCATGCG = NRF1; GGCGGG = E2F, SP1 and TGASTCA = NFE2. Data were derived from two different publicly available sources and were processed using down sampled alignments of 19.8 M (shared read depth), 10 M and 5 M reads and a union peak set.



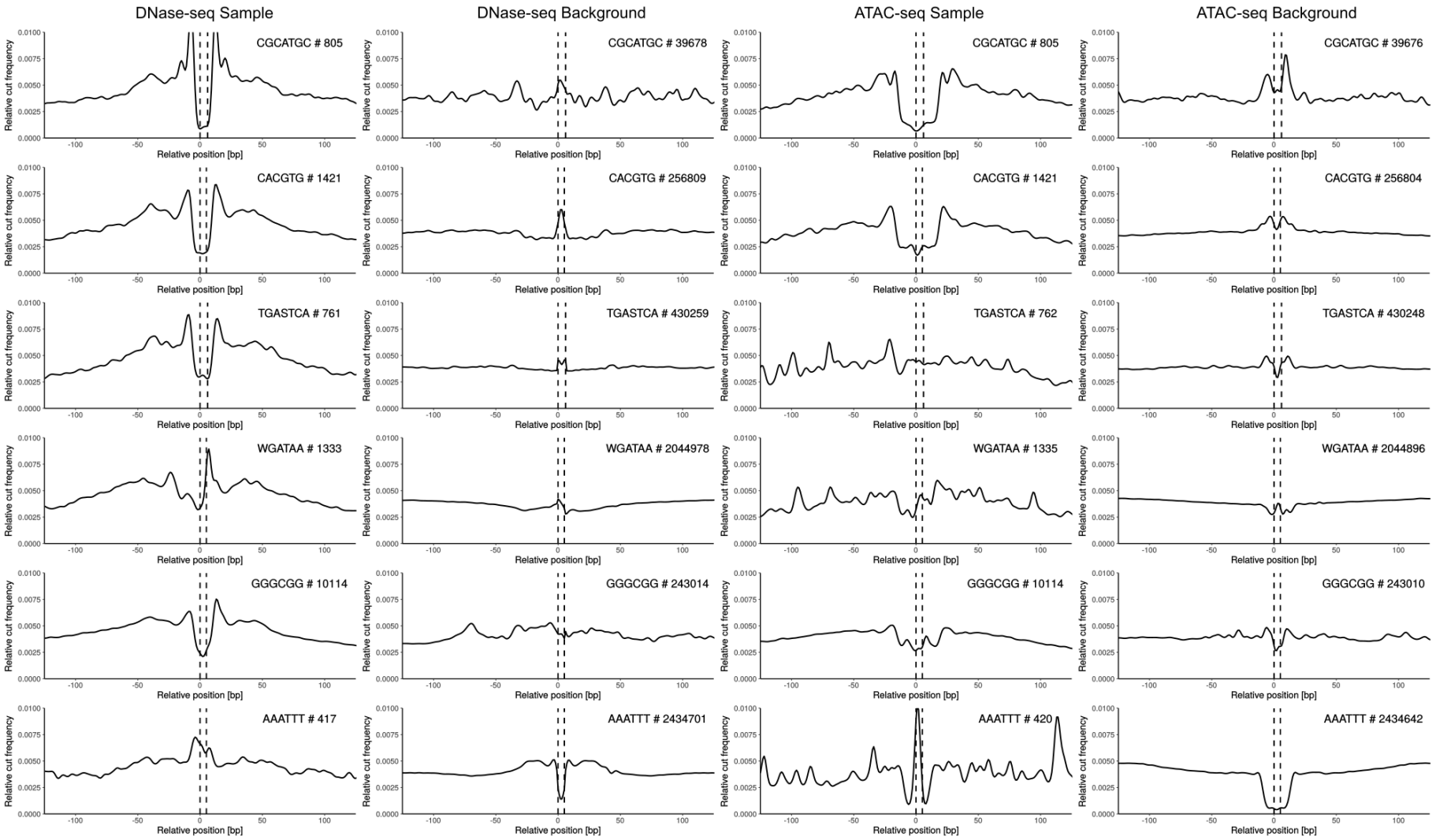
Supplemental Figure S4. Impact of DNase-seq source and protocols on average footprints. Shown are the smoothed and unsmoothed relative DNase I cut probability profiles over four *k*-mers associated with housekeeping transcription factors: CACGTG = E-Box; GCATGCG = NRF1; GGCGGG = E2F, SP1 and TGASTCA = NFE2. Data were derived from three different publicly available sources and were processed using down sampled, equal read depth and a union peak set. Occurrences of the respective *k*-mer were therefore equal and are indicated in black. The general average footprint shape is conserved while the height of the shoulder regions can differ depending on the data source and DNase-seq protocol used DNase-seq protocol.



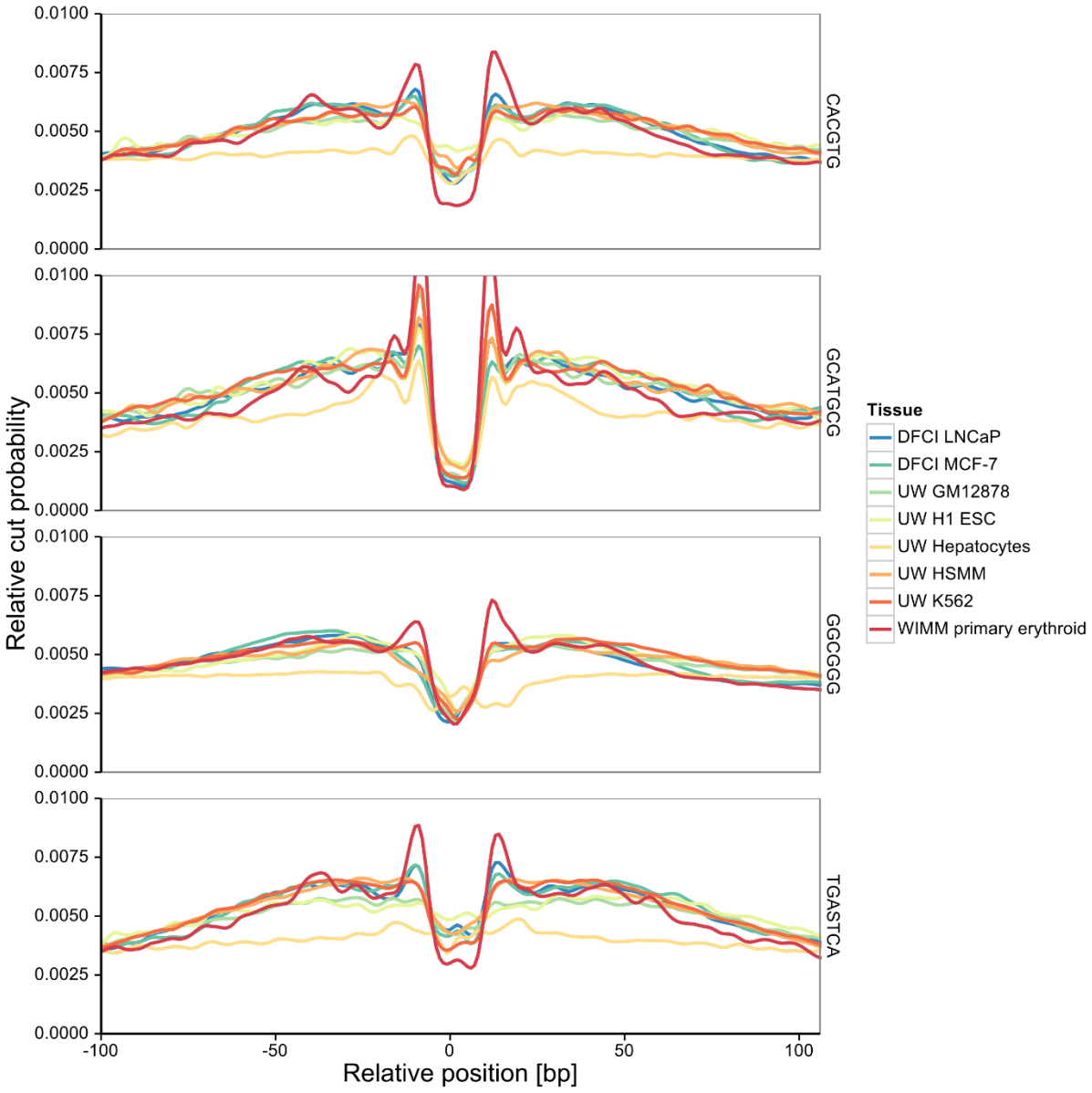
Supplemental Figure S5. Analysing low input DNase-seq (liDNase-seq) data with Sasquatch. Sasquatch is able to analyse low input DNase-seq proposed from Lu et al. 2016[2]. For comparison, example profiles of key transcription factors, derived from 30 and 100 cell liDNase-seq data and from the mouse ENCODE DNase-seq repository are shown. Data derived form liDNase-seq exhibit more footprint flanking noise but resolve the centric footprints comparable to the standard ENCODE data.



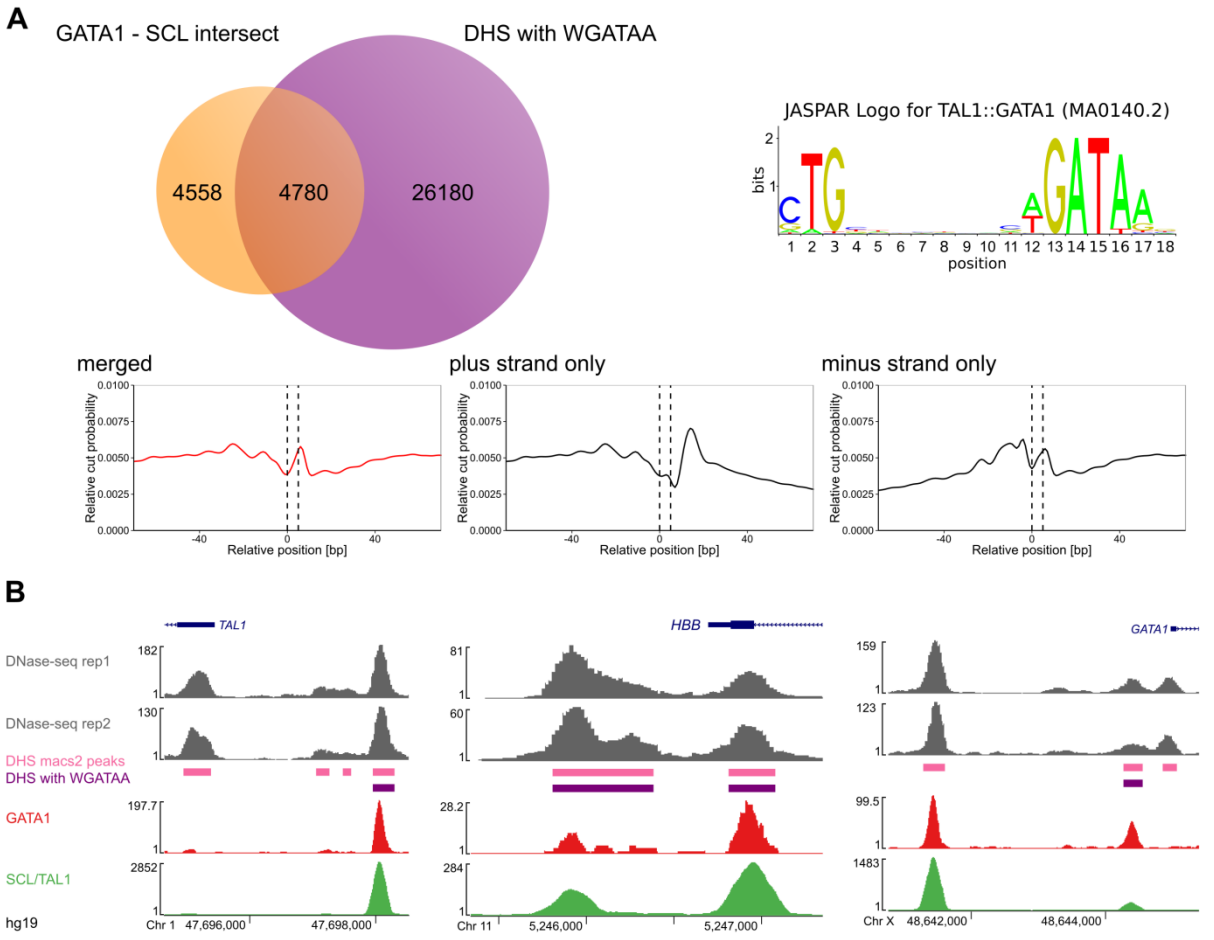
Supplemental Figure S6. DNase-seq, ATAC-seq and TAL1 ChIP-seq of three individuals over the *NPRL3* locus. DNase-seq and ATAC-seq of the same individual show similar sensitivity profiles. The TAL1 ChIP-seq shows distinct signal enrichment in previously characterised regulatory elements harbouring GATA-TAL1 binding sites.



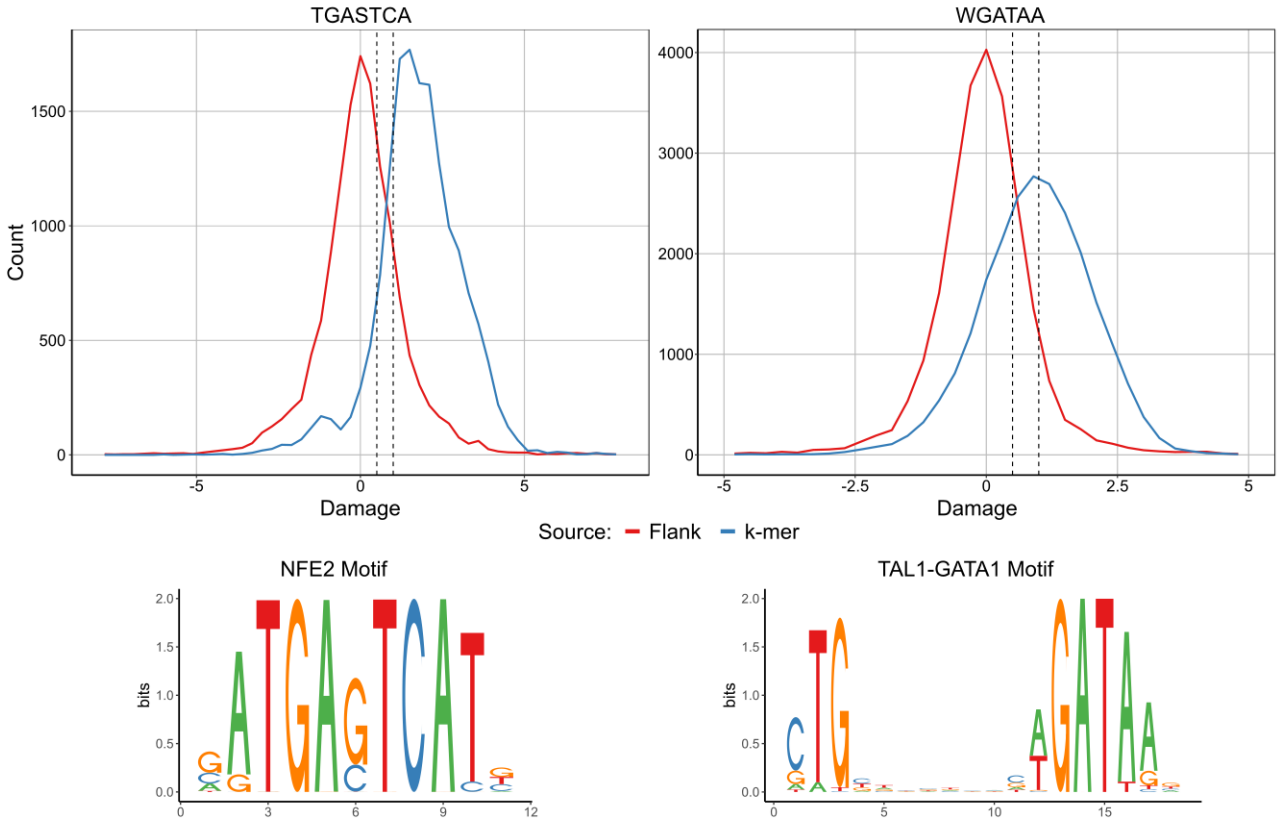
Supplemental Figure S7. Comparison of Sasquatch for DNase-seq and ATAC-seq. DNase-seq and ATAC-seq Sasquatch profiles from primary erythroid samples and de-proteinized DNA background digestions. Exemplary *k*-mers referring to TF known to be bound and one background repeat were selected. Respective occurrence of *k*-mers is indicated by #. ATAC-seq data resolve some but not all footprints, lack diversity in footprint shape and show stronger biases in the background.



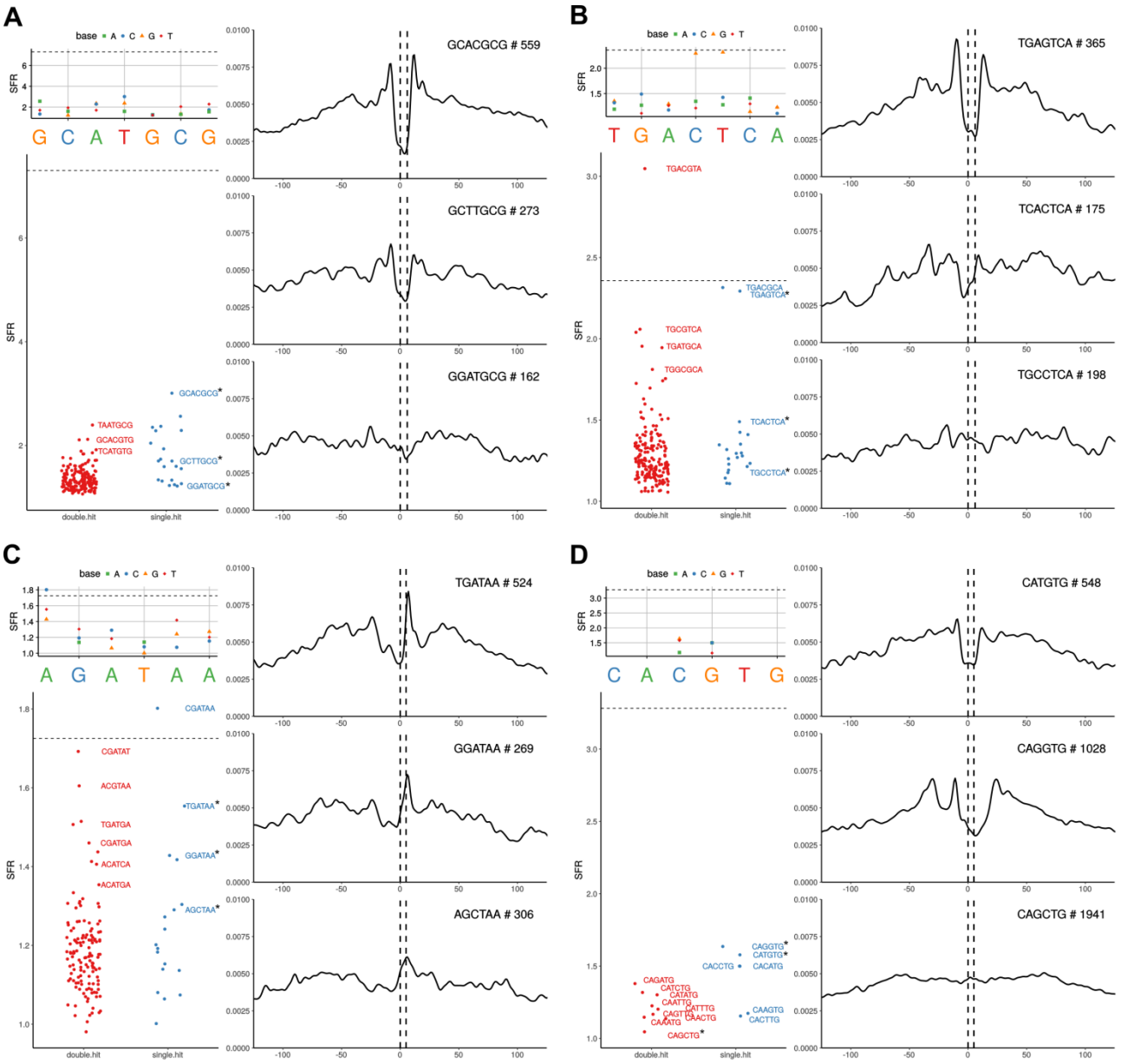
Supplemental Figure S8. Common transcription factor associated *k*-mers across different tissues and data sources. Shown are the smoothed relative DNase I cut probability profiles over four *k*-mers associated with housekeeping transcription factors: CACGTG = E-Box; GCATGCG = NRF1; GGCGG = E2F, SP1 and TGASTCA = NFE2. Data include publicly available and in-house data and cover different tissues. While the shoulder strength and shape may depend on the data source and protocol used, the general shapes of the footprints are consistent.



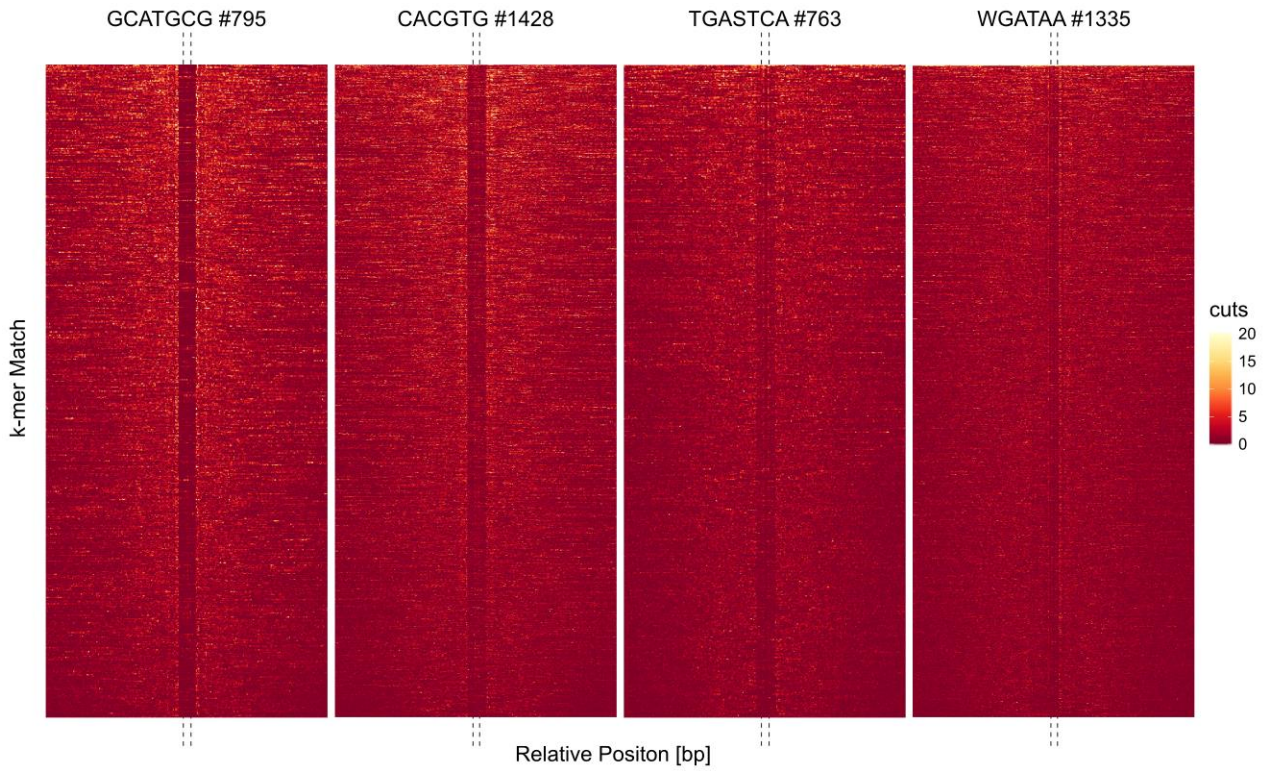
Supplemental Figure S9. Overlap of GATA1, TAL1 and DHS containing WGATAA matches in K562 cells. To demonstrate the co-occurrence of GATA1 and TAL1 in DHS, we retrieved aligned GATA1 and TAL1 ChIP-seq and DNase-seq data on the K562 cell line from ENCODE (SYDH TFBS: wgEncodeEH000638, wgEncodeEH001824 and UW DNase I HS: wgEncodeEH000484). Peaks were called using MACS2 with default settings and intersections were determined using BEDTools[3]. (A) We found 9338 GATA1 and TAL1 intersecting peaks of which more than half intersected with DHS that contained at least one WGATAA match. The JASPAR logo for TAL1 GATA1 co-binding is displayed on the right. This frequent co-binding is also reflected in the average footprint profiles over the *k*-mer WGATAA, where the average footprint is clearly extended upstream of the WGATAA match centre. Due to strand-imbalance, this pattern is primarily visible on the plus strand. (B) Examples of overlapping WGATAA containing DHS with GATA1 and TAL1 peaks.



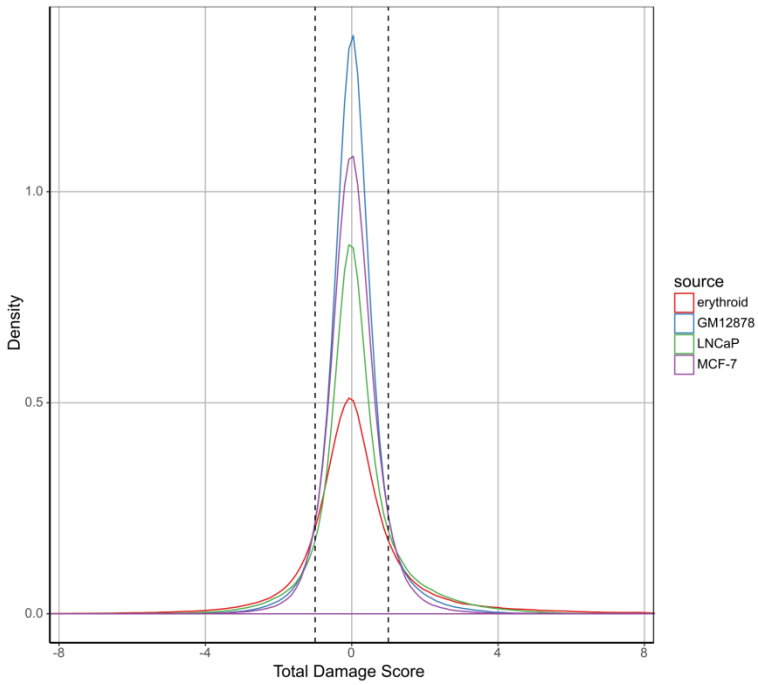
Supplemental Figure S10. Estimating specificity of damage scores on a TF basis. To estimate the specificity of the damage scores, we extracted the sequences around all matches to two *k*-mers within hypersensitive. Both *k*-mers are associated with TF binding in erythroid tissues (NFE2 and GATA1 respectively). We then simulated all possible mutations, assuming that changes within the *k*-mers itself are enriched for true binding changes, while mutations in the 3 bp directly flanking them would be enriched for negative controls that do not change binding. The information content of the respective JASPAR motifs supports this assumption (TAL1-GATA1 co-binding motif). The associated damage scores form distinct distributions, while the separation depends on the footprint strength of the underlying factor. Dashed lines indicate our empirically derived stringent and relaxed thresholds of 1.0 and 0.5.



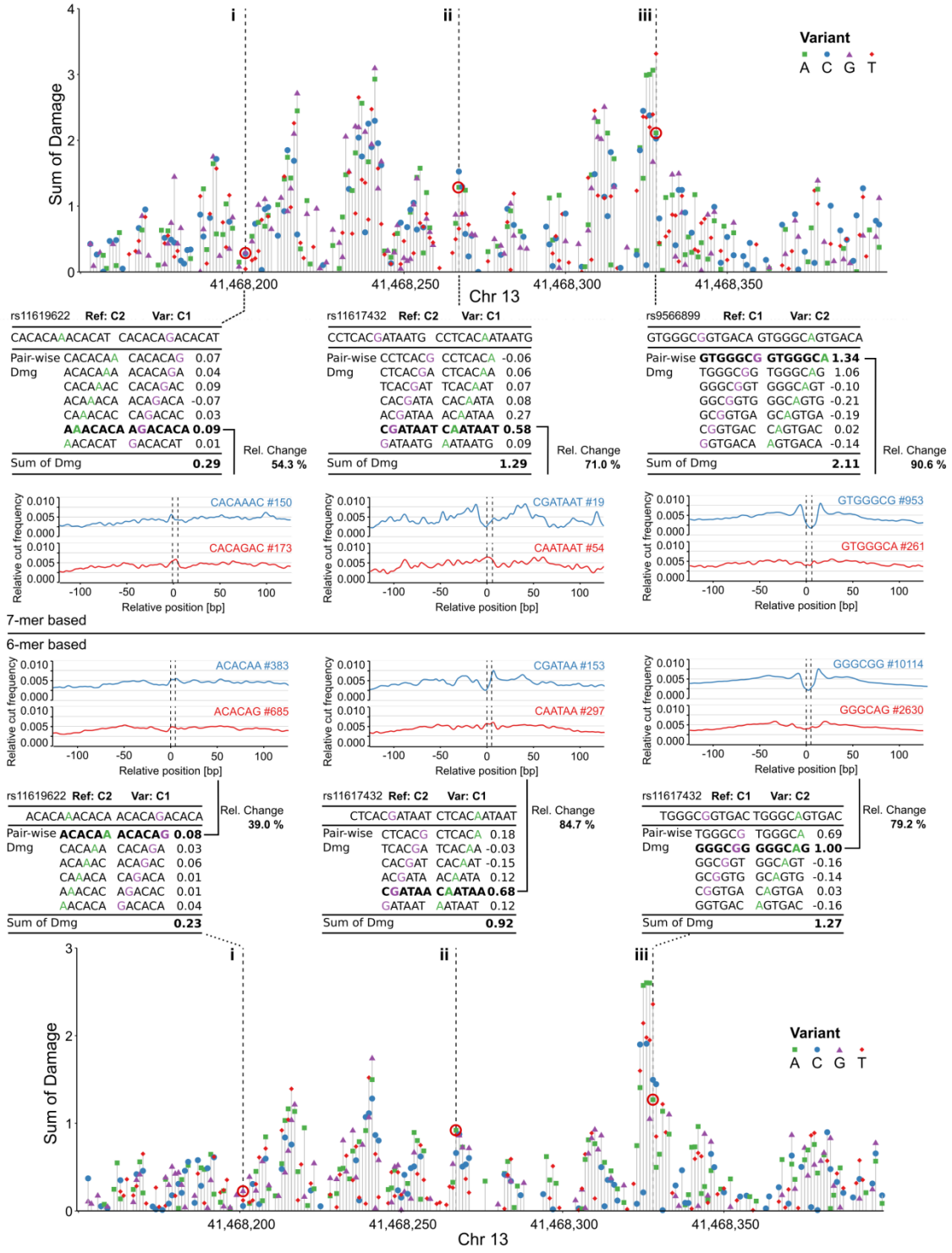
Supplemental Figure S11. Simulating SFR changes associated with *k*-mer changes. Four *k*-mers known to be associated with TF binding in erythroid tissues were mutated by single and double base pair changes and the corresponding SFRs and exemplary profiles visualised. Dashed lines indicate the SFR of the original *k*-mer queried. The SFRs for single changes were also plotted at their relative position to visualize base importance. A) NRF1 B) NFE2 C) GATA1 D) Only canonical E-boxes were probed to visualise changes in factor occupancy.



Supplemental Figure S12. Evidence for DNase I cut protection of TF associated *k*-mers within DHS. Shown are DNase I cut profiles over 250 bp surrounding all matches of *k*-mers within DHS. The *k*-mers are associated with TFs known to be active in erythroid tissues (from left to right: NRF1, E-box, NFE2, GATA1). The number of matches is marked by # and the matches were sorted for total cuts. Dashed lines indicate the actual *k*-mer positions. DNase I cuts were capped at 20 cuts for color scale purposes.

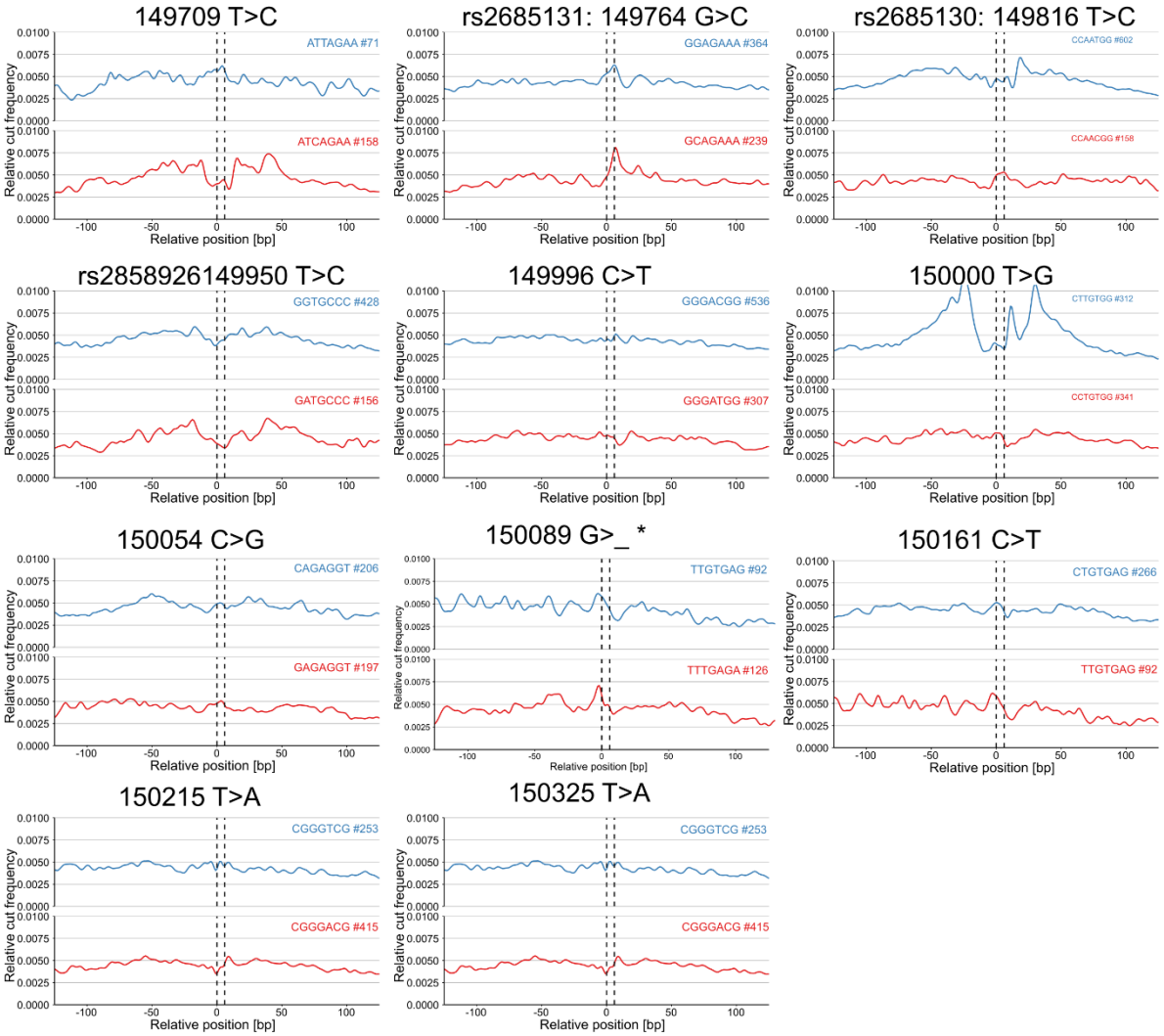


Supplemental Figure S13 Distribution of damage scores within DHS. Across three tissues, all DHS on Chr 16 were in silico mutated and the damage score of every possible substitution of every base was calculated. This is not a background distribution because it includes all positive, factor bound sequences within those sides as well as the negative sequences that are not bound within open-chromatin context. We found the majority of mutations to not alter TF binding, while the flanks of the distribution that indicate altered binding potential are larger depending on the DNase-seq data quality with respect to the resolution for identifying average footprints. Dashed lines indicate the -1.0 and 1.0 stringent damaging cut off we empirically derived.

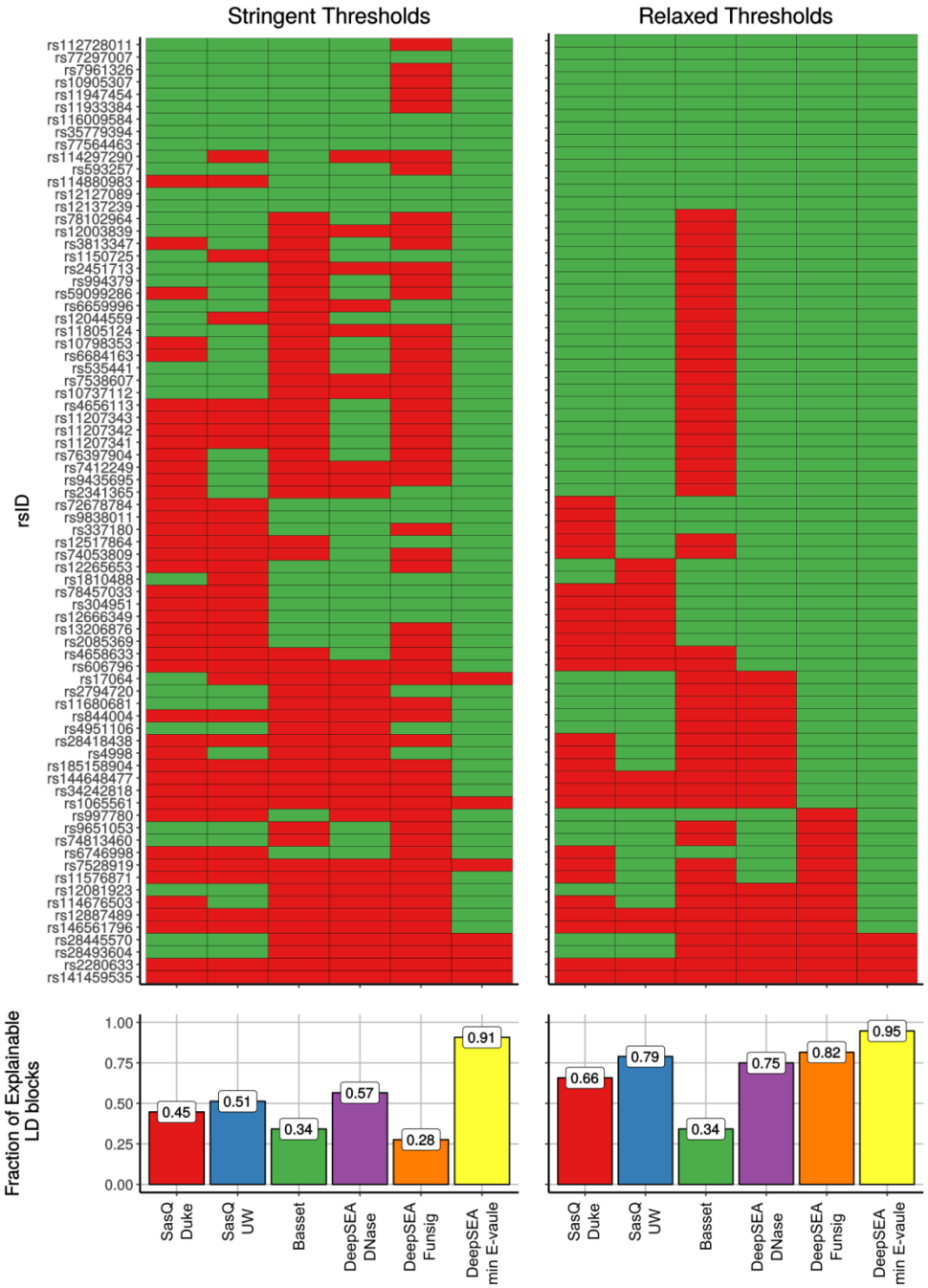


Supplemental Figure S14. Comparison of different *k*-mer sizes for estimating the impact of sequence variants. Detailed footprinting damage prediction including the respective genotype for individuals C1 and C2 at every SNP, detailed pair-wise damage table and comparison of 7-mer and 6-mer based analysis. By default we used a *k*-mer size of 7 bp to predict the impact of sequence variation. When detecting profiles derived from a very low number of *k*-mer occurrences, for example in variant **ii**, we fall back to query the variant on a 6-mer basis which are associated with higher occurrences. Overall, total damage scores on a 6-mer basis tend to be lower and thus offer a decreased discriminatory power, while general trends are conserved.

chr	pos	id	reference.seq	alternative.seq	total.dmg	re.change
chr16	149709	.	CATTATTAGAAAA	CATTATCAGAAAA	-1.411	0.706
chr16	149764	rs2685131	CAGAAGGAGAAAG	CAGAAGCAGAAAG	-0.954	0.924
chr16	149816	rs2685130	TTCCAATGGCTTG	TTCCAACGGCTTG	0.48	0.554
chr16	149950	rs2858926	TGAGAGGTGCCCA	TGAGAGATGCCCA	-0.587	0.501
chr16	149996	.	CCGGGACGGCTTG	CCGGGATGGCTTG	-0.122	0.331
chr16	150000	.	GACGGCTTGTGGG	GACGGCCTGTGGG	0.720	0.787
chr16	150054	.	AGGTTTCAGAGGT	AGGTTTGAGAGGT	0.331	0.522
chr16	150089	.	CAGGTTGTGAGAG	CAGGTTTGAGAG	0.041*	0.033*
chr16	150161	.	CACAGGCTGTGAG	CACAGGTTGTGAG	-0.321	0.464
chr16	150215	.	CCCGGGTCGGCTT	CCCGGGACGGCTT	-0.246	0.706
chr16	150325	.	CCCGGGTCGGCTT	CCCGGGACGGCTT	-0.246	0.706



Supplemental Figure S15. Sasquatch analysis and footprint profiles of all SNPs found close to r_SNP and/or present in family members of the patient. We found no other variant then the r_SNP to have a striking damaging potential. *For the single base-pair deletion we adapted Sasquatch to calculate a pseudo-damage score, by summing up all SFR values across the reference and variant sequence, normalizing them for a single *k*-mer window, calculating the difference and extrapolating that difference to a full 13 bp comparison window by multiplying it by 7.



Supplemental Figure S16. Benchmarking Sasquatch against deep learning approaches using bQTL SNPs in LD blocks. Shown are the most significant bQTLs across five TFs identified in lymphoblastoid cell lines[4]. The bQTLs were imputed and grouped into high LD-blocks. For each SNP in the LD-block the potential impact was predicted based on GM12878 data using: 1) Sasquatch (two DNase-seq sources Duke and UW) 2) Basset and 3) DeepSEA where the DNase predictor, the functional significance score and the smallest e-value across all GM12878 predictors was used. Visualised is the fraction of explainable LD-blocks per tool. Stringent and relaxed thresholds were used respectively.

Supplemental Methods

Cell source, culture, preparation and DNase-seq protocol

Human primary erythroid stem cell progenitors were isolated from peripheral blood, using CD34 coupled magnetic beads. Cells were expanded for 7 days in low Epo (0.5 IU/ml) conditions and then transferred for differentiation in high Epo medium (3.0 IU/ml). On day 13, cells were washed in cold PBS, and counted. 50 million cells were used for the DNase-seq protocol.

DNase-seq protocol was performed as previously published[5]. DNA was purified using a phenol chloroform extraction and the optimal digests were selected for library preparation using NEB Library Preparation kit. Libraries were amplified and bar-coded using the NEB Next 2xMastermix (NEB) and TruSeq oligos. A size selection step for small fragments was performed. DNase-seq library profiles were visualized using D1000 tape on the Tapestation (Agilent) and quantified using the universal library quantification kit (KAPA Biosystems). Samples were sequenced on Illumina platform using: 150 bp paired end reads (MiSeq), 75bp paired end reads (HiSeq) or 40/75 bp (NextSeq) paired end reads. For the background libraries, 100 ng of genomic DNA was incubated with DNase I for 3 minutes and then libraries were created to test sequence for sequence bias.

ATAC-seq protocol

ATAC-seq was performed as previously published[6]. Nuclei of 70000 lysed cells were isolated and transposition with Tn5 transposase (Nextera, Illumina) was performed for 30 minutes at 37 °C. The DNA was extracted using a MinElute kit (Qiagen). Libraries were then amplified and barcoded using the NEBNext 2xMastermix (NEB) and the custom ATAC-seq primers published by Buenrostro and colleagues. ATAC-seq libraries profiles were visualized using D1000 tape on the Tapestation (Agilent) and quantified using the universal library quantification kit (KAPA Biosystems). For the ATAC-seq background 100 ng of genomic DNA was incubated with the Tn5 transposase and the library was following the protocol described above.

Estimating intrinsic DNase I sequence bias

To correct for DNase I sequence bias, 6-mer based weighting factors were calculated according to the DNase I sequence preferences. For estimating these propensities, deproteinized genomic DNA was digested with DNase I and sequenced according to the protocol described above. Sequencing was performed at high depth (119 million mapable reads for the human and 133 million for the mouse background). DNase I cuts were mapped and average cut profiles of all possible 6-mers within the mapable genome were recorded as described above. The relative sequence cut probability of each 6-mer was than calculated as the recorded number of cuts at the 4th base position of each 6-mer divided by the number of recorded 6-mer occurrences along the entire mapable genome. The relative cut probabilities were shifted to achieve a median cut probability of 1 and relative weights were calculated to correct the relative sequence cut probability of every 6-mer to 1. A sequence context of 6 was chosen because it has been shown to capture the DNase I sequence bias sufficiently. An ATAC-seq background was estimated using the same procedure.

Retrieving and processing public available data

We retrieved DNase-seq data from three different publicly available sources (DFCI = Dana-Farber Cancer Institute[7], GSE51915; Duke = ENCODE, DNase I HS, Duke University, UW = ENCODE, DNase I HS, University of Washington). ENCODE[8] data were downloaded as aligned reads and called peaks were retrieved from the repository. DFCI data were retrieved as raw data and mapping and peak calling was performed using our in-house pipeline as described above, with the additional

step of flashing (FLASH v1.2.8[9]) and remapping unmapped reads. Peak calls for ENCODE data were downloaded. For DFCI data a manually curated SeqMonk workflow like peaks were called. Although ENCODE peak calls and peak calls derived from our in-house pipeline or MACS2[10] differ strikingly in terms of stringency, the results are comparable because Sasquatch is robust towards peak call stringency (Supplemental Fig. S2-6). When available, replicates were processed separately and their cut profiles were merged afterwards by averaging the *k*-mer occurrences and DNase I cut profiles.

Processing liDNase-seq data

Mouse ESC, liDNase-seq raw data from 30 and 100 cells input (GSM2029801, GSM2029802) were downloaded and processed using our in-house DHS pipeline as described above. Peaks were called using MACS2 with default settings and *k*-mer based cut profiles were calculated as described above. Complementary, mENCODE mESC E14 (wgEncodeEM003417, GSM1014154) data were retrieved and merged as described for the human ENCODE data. Overlay profiles of TF core sequences across the three data types were plotted using Sasquatch's core functions.

Comparison against JASPAR motifs

Comparison against JASPAR motif database was performed utilizing the R packages Biostrings[11] and TFBSTools[12] (v1.6.1). Position weight matrices (PWM) data were retrieved from the JASPAR 2016[13] (v1.0.0) R package distribution (using all PWM versions). Each *k*-mer was padded with "N"s and scanned against the PWM data using the searchSeq function, reporting up to six highest matching factors with relative scores >= 0.8.

Estimating evidence of *k*-mer protection from DNase I cutting

Sequence matches to the *k*-mers (WGATAA, TGASTCA, GCATCGC, CACGTG) within hypersensitive sites were extracted and the surrounding DNase I cuts were recorded. These cut profiles per match were sorted for total number of cuts within the window and plotted while capping the maximum reads per base to 20 for colour scaling.

SFR distribution over *k*-mer changes

To assess the changes in SFR associated with single and double base pair changes, the SFR of exemplary factor bound *k*-mers (AGATAA, GCATGCG, TGACTCA) and of every possible single or double base pair changed *k*-mer was calculated. To visualise multiple factors of related *k*-mers, the SFR of a common E-box motif *k*-mer (CACGTG) and of every possible canonical E-box *k*-mer following CANNTG) was calculated.

Estimating the distribution of damage scores within hypersensitive sites

For estimating the general distribution of damaging scores within hypersensitive sites, the respective DHS peak calls of four tissues on Chromosome 16 were used and every possible mutation of every base within those was simulated and the damaging score calculated. Note that this is not a pure negative background as it contains negative SNPs in the DHS background as well as all SNPs within true binding sites.

Estimating the damage score distribution and specificity on a factor basis

In the context of binding site rich open-chromatin regions sampling a true negative background to set the boundaries of what can be classed as strong losses or gains in footprints, is difficult. Therefore, we devised a strategy to approximate true negative and true positive variants using the very well characterised motifs two TFs (NFE2 and GATA1. We extracted the surrounding

sequences for all k-mer matches in open-chromatin sites in erythroid cells. As it has been shown that the ability of these motifs to mediate binding is dependent on the centre 6 and 7 bases respectively, we can assume that SNPs lying directly within the k-mer would be enriched for binding altering variants, while SNPs directly flanking those k-mers would be more likely to be neutral. The sequences surrounding all *k*-mer matches (WGATAA, TGASTCA) were extracted. To approximate a set of SNPs affecting and a set of non-affecting transcription factor binding, every possible mutation within the *k*-mer and within the 3 bps flanking the *k*-mer up and downstream respectively was simulated and the damaging score (Supplemental Fig. S10).

Estimate impact of DNase-seq protocol, sequencing depth, peak calling and *k*-mer occurrence

To compare the influence of different DNase-seq protocol, MCF-7 DNase-seq data from three different sources (DFCI: GSM1255280**;** Duke: GSM816627; and ENCODE: GSM1024767), each with its respective DNase-seq protocol details were retrieved and pre-processed as described above. Alignments were down sampled using SAMtools[14] to match the dataset with the lowest read number (~19.8 M reads). Merged union peaks were derived from all three peak calls using BEDTools *merge* (*-d 10*). DNase I cut profiles per dataset were then processed with Sasquatch using the union peaks and down sampled alignments. To assess the impact of sequencing depth, data were processed using the respective peak calls, the down sampled alignments and further down sampled alignment versions with 10 M and 5 M reads respectively. To assess the impact of peak calling stringency and *k*-mer occurrences, each down sampled alignment was processed with each dataset's respective initial peak call (in-house pipeline output for DFCI or uniquely processed ENCODE peaks for Duke and UW). In addition, cut profiles were calculated using only the top 20,000 or 5,000 peaks, after sorting for signal strength or p-value respectively.

Supplemental References

1. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41,** e201 (2013).
2. Lu, F. *et al.* Establishing Chromatin Regulatory Landscape during Mouse Preimplantation Development. *Cell* **165,** 1375–1388 (2016).
3. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
4. Tehranchi, A. K. *et al.* Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell* **165,** 730–741 (2016).
5. Hosseini, M. *et al.* Causes and consequences of chromatin variation between inbred mice. *PLoS Genet.* **9,** e1003570 (2013).
6. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–8 (2013).
7. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **11,** 73–8 (2014).
8. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
9. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27,** 2957–2963 (2011).

10. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).

11. H, P., Aboyoun P, G. R. & S, D. Biostrings: String objects representing biological sequences, and matching algorithms. R and DebRoy S (2017). Biostrings: String objects (2017).

12. Tan, G. & Lenhard, B. Sequence analysis TFBSTools : an R / Bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32,** 4–5 (2016).

13. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44,** gkv1176 (2015).

14. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

15. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492,** 369–75 (2012).

Supplemental Tables

Supplemental Table S1 Sasquatch analysis of SNPs associated with differential TAL1 binding.

ID	Chr	Pos hg18	Sequence	kmer Ref	kmer Var	SFR Ref	SFR Var	Total Damage	Rel. Change	JASPAR
rs73185392	chr13	41494435	AGCCCCG/AATAATA	AGCCCCG	AGCCCCA	2.297	1.097	3.07	0.925	GATA2=0.98; GATA5=0.95; GATA3=0.94;
rs1356334	chr2	236034216	CTGGTGG/AGAATTA	CTGGTGG	CTGGTGA	2.718	1.239	2.786	0.861	HIC2=0.93; E2F6=0.92; TEAD3=0.9;
rs9566899	chr13	41468327	GTGGGCG/AGTGACA	GTGGGCG	GTGGGCA	2.736	1.164	2.112	0.906	SP1=0.93; SP1=0.91; ZNF354C=0.9;
rs11617432	chr13	41468266	CCTCACG/AATAATG	CGATAAT	CAATAAT	1.916	1.265	1.287	0.71	GATA2=0.98; GATA3=0.94; GATA5=0.94;
rs28508084	chr15	96337645	AACTAAA/GTCAGGC	AACTAAA	AACTAAG	1.452	1.213	0.923	0.392	GATA3=0.9; EN1=0.89; GATA2=0.89;
rs198415	chr1	11824178	TGGTAT/CTAGTCCC	GTATTAG	GTATCAG	1.601	1.249	0.853	0.306	GSC2=0.93; OTX1=0.92; TEAD3=0.92;
rs9532971	chr13	41493929	GCCTCAT/CAGAGGT	ATAGAGG	ACAGAGG	1.463	1.08	0.754	0.574	EMX2=0.88; LHX9=0.88; NOTO=0.87;
rs1469713	chr19	19389805	GCAGACA/GTCTGGG	GACATCT	GACGTCT	1.746	1.301	0.754	0.596	GATA2=0.96; MEIS1=0.94; TCF4=0.94;
rs11619622	chr13	41468200	CACACAA/GACACAT	AAACACA	AGACACA	1.196	1.089	0.292	0.191	FOXL1=0.96; FOXP3=0.96; FOXD2=0.96;
rs112342010	chr2	119649696	ACAGCCG/AGCTGGC	AGCCGGC	AGCCAGC	1.183	1.069	0.192	0.292	TFAP2A=0.91; MEIS1=0.86; TCF3=0.84;
rs9929936	chr16	73494733	GCCAACT/ATTCTCT	AACTTTC	AACATTC	1.367	1.101	0.039	0.314	SPI1=0.93; SOX10=0.93; FOXD2=0.89;
rs2180414	chr6	157802774	GTGATAA/CCCTCTC	TGATAAC	TGATACC	1.341	1.123	-0.296	0.08	GATA5=0.96; GATA2=0.96; GATA3=0.96;
rs12205172	chr6	6425000	CGGCTAC/AAGTTAT	CGGCTAC	CGGCTAA	1.187	1.434	-0.382	0.376	BSX=0.92; RHOXF1=0.91; BARX1=0.91;
rs1469712	chr19	19389820	ACTCAGC/ATAAAGG	CAGCTAA	CAGATAA	1.181	1.78	-1.294	0.768	RHOXF1=0.94; NRL=0.89; OTX1=0.86;
rs3759283	chr12	8125378	AGTTTAC/TCTTTGG	TTACCTT	TTATCTT	1.239	1.655	-1.533	0.636	FOXD2=0.96; FOXL1=0.91; FOXO3=0.91;
rs9937638	chr16	73494722	AGCTTTC/ATCTGCC	TTCTCTG	TTATCTG	1.178	1.756	-1.916	0.765	SOX10=0.93; SPIB=0.88; SPI1=0.88;

Supplemental Table S2: List of 100, 1000 genomes version 3 imputed and DHS intersected SNPs with Sasquatch predictions. Population indicates the population background based on which the respective variant was identified as a proxy SNP for one of original 75 SNP[15] (orig 75).

ID	Chr	Pos hg18	Ref	Var	Population	Sequence Ref	Sequence Var	k-mer Ref	k-mer Var	SFR Ref	SFR Var	Total Damage	% Change	Class
rs608662	chr6	139839444	G	A	AS,CEU	ACTTGCGCGCAGG	ACTTGCACGCAGG	GCGCAGG	ACGCAGG	2.135	1.28	4.138	0.753	strong loss
rs9890212	chr17	27169890	G	C	AS,CEU	TGCCGGGGGCGTC	TGCCGGCGGCGTC	GGGGGCG	GGCGGCG	3.113	1.538	3.571	0.745	strong loss
rs11575895	chr17	43971785	A	G	AS,CEU	CCCACTAGTGGCC	CCCACTGGTGGCC	CCACTAG	CCACTGG	3.188	1.351	3.502	0.84	strong loss
rs1369312	chr15	66084407	G	T	AS	AATCCCGCACGCG	AATCCCTCACGCG	CGCACGC	CTCACGC	2.567	1.212	3.269	0.865	strong loss
rs3181215	chr17	27077331	A	G	CEU	TGGCCTAGATGTT	TGGCCTGGATGTT	CTAGATG	CTGGATG	3.027	1.234	2.219	0.885	strong loss
rs34731408	chr10	45961020	C	T	AS,CEU	TCCTCCCCCTACC	TCCTCCTCCTACC	CCCCCTA	CCTCCTA	1.958	1.419	2.13	0.563	strong loss
rs71496621	chr10	46076348	T	C	AS,CEU	AGAGCTTAGCCCC	AGAGCTCAGCCCC	TAGCCCC	CAGCCCC	2.339	1.279	2.088	0.792	strong loss
rs2923444	chr8	42397004	C	T	AS,CEU	AGCACCCGCCCGG	AGCACCTGCCCGG	CCCGCCC	CCTGCCC	2.782	1.128	2.044	0.928	strong loss
rs17577024	chr17	44186252	G	A	AS,CEU	GTCCAAGTATACA	GTCCAAATATACA	GTATACA	ATATACA	1.927	1.202	1.866	0.782	strong loss
rs76029685	chr10	46168346	G	A	AS	TGCCGGGGCGCCA	TGCCGGAGCGCCA	CGGGGCG	CGGAGCG	2.012	1.259	1.748	0.744	strong loss
rs131806	chr22	50963965	C	G	AS	TCCCTGCGCTCTG	TCCCTGGGCTCTG	CCTGCGC	CCTGGGC	2.09	1.169	1.533	0.845	strong loss
rs769236	chr4	122745038	C	T	AS,CEU	GCCCGCCGGAGCG	GCCCGCTGGAGCG	GCCCGCC	GCCCGCT	1.707	1.093	1.483	0.868	strong loss
rs2047866	chr15	76136194	C	T	CEU	GTGAGGCGCCCGG	GTGAGGTGCCCGG	AGGCGCC	AGGTGCC	1.554	1.104	1.329	0.813	strong loss
rs75820736	chr10	46076236	C	A	AS,CEU	TGTGCGCTAGGAC	TGTGCGATAGGAC	TGTGCGC	TGTGCGA	2.239	1.23	1.109	0.815	strong loss
rs643381	chr6	139839423	C	A	AS,CEU	CCCCCCCCAGGGC	CCCCCCACAGGGC	CCCCCCC	CCCCCAC	1.892	1.529	1.085	0.407	strong loss
rs242561	chr17	44026548	T	C	AS,CEU	CGATTCTGCTGAG	CGATTCCGCTGAG	TCTGCTG	TCCGCTG	1.725	1.165	1.082	0.773	strong loss
rs55714296	chr17	44137189	A	T	AS,CEU	CCATTCACCTGCC	CCATTCTCCTGCC	CACCTGC	CTCCTGC	1.682	1.121	1.052	0.822	strong loss
rs9369425	chr6	43810974	G	A	CEU	TGGAACGTTGTAT	TGGAACATTGTAT	ACGTTGT	ACATTGT	1.768	1.276	0.94	0.641	weak loss
rs10758656	chr9	4852599	A	G	AS,CEU	GTAGATAAGGTGC	GTAGATGAGGTGC	AGATAAG	AGATGAG	1.877	1.146	0.874	0.834	weak loss
rs41384744	chr17	44137070	A	G	AS,CEU	GGAACTAAGAGAG	GGAACTGAGAGAG	CTAAGAG	CTGAGAG	1.486	1.127	0.856	0.739	weak loss
rs17577052	chr17	44186301	T	C	AS,CEU	AACTACTAACAAG	AACTACCAACAAG	CTACTAA	CTACCAA	1.549	1.25	0.792	0.545	weak loss
rs113417378	chr17	44270809	A	G	AS,CEU	GCGAGCAAGCGGG	GCGAGCGAGCGGG	GCAAGCG	GCGAGCG	1.592	1.207	0.789	0.651	weak loss

rs737092	chr20	55990405	T	C	AS,CEU,orig75	TAAGTGTCTGGCT	TAAGTGCCTGGCT	GTCTGGC	GCCTGGC	1.307	1.101	0.679	0.673	weak loss
rs4737010	chr8	41630447	G	A	AS,CEU	GCCAGCGACCACC	GCCAGCAACCACC	GACCACC	AACCACC	1.931	1.27	0.67	0.71	weak loss
rs4926725	chr1	47679366	A	T	CEU	ATGGCTACAGCAG	ATGGCTTCAGCAG	ATGGCTA	ATGGCTT	1.4	1.182	0.596	0.544	weak loss
rs6592965	chr7	50427982	G	A	AS,CEU	TGAGAGGGAATGG	TGAGAGAGAATGG	GAGAGGG	GAGAGAG	1.387	1.118	0.571	0.695	weak loss
rs2696633	chr17	44270059	A	T	AS,CEU	CGGGGGATTTTTC	CGGGGGTTTTTTC	GGGATTT	GGGTTTT	1.542	1.17	0.559	0.685	weak loss
rs7199443	chr16	67841129	T	G	AS,CEU	CAGAGATGAGGGT	CAGAGAGGAGGGT	GAGATGA	GAGAGGA	1.275	1.038	0.51	0.861	weak loss
rs1175550	chr1	3691528	A	G	AS,CEU,orig75	GCCTAGATTGGGC	GCCTAGGTTGGGC	CTAGATT	CTAGGTT	1.232	1.047	0.504	0.8	weak loss
rs11865131	chr16	163667	G	A	AS,CEU	TCCTGTGGGGGTG	TCCTGTAGGGGTG	TGTGGGG	TGTAGGG	1.538	1.31	0.497	0.424	neutral
rs2271176	chr4	122791601	G	C	AS,CEU	GCCTAGGTCCTGG	GCCTAGCTCCTGG	CTAGGTC	CTAGCTC	1.459	1.187	0.476	0.593	neutral
rs6808837	chr3	141217954	T	C	AS,CEU	TCTAGATTAAGCT	TCTAGACTAAGCT	TCTAGAT	TCTAGAC	1.243	1.672	0.45	0.638	neutral
rs439558	chr17	43717803	T	C	AS,CEU	CACAAGTGCTGGA	CACAAGCGCTGGA	TGCTGGA	CGCTGGA	1.63	1.159	0.44	0.748	neutral
rs11670503	chr19	4458063	A	G	CEU	AGGCCAACTGAAC	AGGCCAGCTGAAC	CAACTGA	CAGCTGA	1.573	1.131	0.409	0.771	neutral
rs79337279	chr10	46153540	G	T	AS	TTCTTAGGTGGAA	TTCTTATGTGGAA	TAGGTGG	TATGTGG	2.426	1.54	0.403	0.621	neutral
rs12126653	chr1	3773815	C	G	AS	CCACCCCAGGCCG	CCACCCGAGGCCG	CCACCCC	CCACCCG	1.48	1.229	0.384	0.522	neutral
rs11085824	chr19	13001547	A	G	AS,CEU	GCTAAGATCGCCC	GCTAAGGTCGCCC	AGATCGC	AGGTCGC	1.4	1.193	0.373	0.517	neutral
rs4926524	chr1	47679258	C	T	CEU	CAACTGCGGCCCA	CAACTGTGGCCCA	AACTGCG	AACTGTG	1.438	1.186	0.35	0.576	neutral
rs4490057	chr17	76375095	A	G	AS,CEU	TGCTCCATTATCG	TGCTCCGTTATCG	TCCATTA	TCCGTTA	1.393	1.173	0.321	0.56	neutral
rs1546723	chr6	109625879	G	A	AS,CEU	GGGAGCGGAGTGG	GGGAGCAGAGTGG	GCGGAGT	GCAGAGT	1.729	1.16	0.273	0.78	neutral
rs140491	chr22	21922364	T	C	AS,CEU	CCGATCTGAGGGC	CCGATCCGAGGGC	GATCTGA	GATCCGA	1.282	1.164	0.215	0.418	neutral
rs12718598	chr7	50428445	T	C	AS,CEU,orig75	GCCCCATGTCGTC	GCCCCACGTCGTC	CCCATGT	CCCACGT	1.239	1.726	0.214	0.671	neutral
rs9901219	chr17	37558369	T	G	AS,CEU	CCAGGATCCTGTA	CCAGGAGCCTGTA	AGGATCC	AGGAGCC	1.259	1.054	0.2	0.79	neutral
rs77127734	chr11	73018413	T	A	CEU	AAGTTTTCACAAC	AAGTTTACACAAC	TTTCACA	TTACACA	1.419	1.223	0.17	0.467	neutral
rs2285089	chr22	32898291	G	A	AS,CEU	AAAAGCGGAGACT	AAAAGCAGAGACT	AGCGGAG	AGCAGAG	1.129	1.47	0.035	0.725	neutral
rs12609866	chr19	33184412	C	T	AS,CEU	CCACAGCGGAAGG	CCACAGTGGAAGG	CAGCGGA	CAGTGGA	1.085	1.186	0.031	0.54	neutral
rs2572207	chr15	66070693	C	T	AS,CEU,orig75	GTAACACTGTATA	GTAACATTGTATA	CTGTATA	TTGTATA	1.059	1.347	0.015	0.829	neutral
rs8113575	chr19	13030280	G	A	AS,CEU	TTGGGAGCAGGAC	TTGGGAACAGGAC	TTGGGAG	TTGGGAA	1.264	1.174	-0.02	0.34	neutral

rs11248850	chr16	163598	G	A	AS,CEU,orig75	CTTGAGGGAGCAG	CTTGAGAGAGCAG	TGAGGGA	TGAGAGA	1.005	1.225	-0.071	0.976	neutral
rs20549	chr16	67969930	A	G	AS,CEU	TAGATAACGCGTG	TAGATAGCGCGTG	ACGCGTG	GCGCGTG	1.695	2.605	-0.08	0.567	neutral
rs11240734	chr1	203651824	C	T	AS,CEU	TTATGGCTGCTCT	TTATGGTTGCTCT	TTATGGC	TTATGGT	1.757	1.52	-0.1	0.313	neutral
rs7196789	chr16	67927124	C	T	CEU	CGTAGGCTTGTTT	CGTAGGTTTGTTT	CTTGTTT	TTTGTTT	1.172	1.34	-0.135	0.493	neutral
rs2304903	chr15	75315778	A	G	AS,CEU	AGGGCCAGGGTTT	AGGGCCGGGGTTT	CAGGGTT	CGGGGTT	1.151	1.546	-0.139	0.724	neutral
rs6778081	chr3	195819205	C	T	AS,CEU	TTGAGACGGAGTT	TTGAGATGGAGTT	CGGAGTT	TGGAGTT	1.284	1.115	-0.197	0.597	neutral
rs11628273	chr14	65509878	C	T	CEU	CTGGAACGGCCTA	CTGGAATGGCCTA	CGGCCTA	TGGCCTA	1.132	1.317	-0.255	0.583	neutral
rs3811742	chr4	122722693	C	G	AS,CEU	TCGGGTCTCAAGG	TCGGGTGTCAAGG	GTCTCAA	GTGTCAA	1.337	1.684	-0.282	0.507	neutral
rs1476792	chr11	67196237	T	C	AS,CEU	ATTAAGTTCTGAT	ATTAAGCTCTGAT	TTAAGTT	TTAAGCT	1.027	1.202	-0.284	0.866	neutral
rs12611419	chr19	33184369	A	G	AS,CEU	AGCAGCAGCAAGA	AGCAGCGGCAAGA	GCAGCAG	GCAGCGG	1.356	1.118	-0.294	0.668	neutral
rs7206671	chr16	67807146	A	G	AS,CEU	GAGTATAGCACCT	GAGTATGGCACCT	ATAGCAC	ATGGCAC	1.431	1.909	-0.318	0.526	neutral
rs78999882	chr10	46076813	A	T	AS	CTGCACAGTACAA	CTGCACTGTACAA	GCACAGT	GCACTGT	1.253	1.518	-0.357	0.511	neutral
rs11042154	chr11	9029700	G	A	CEU	TATCCAGATGTGT	TATCCAAATGTGT	TATCCAG	TATCCAA	1.18	1.721	-0.365	0.751	neutral
rs11089620	chr22	21922456	C	G	CEU	GGAGCCCGCGCCG	GGAGCCGGCGCCG	CCGCGCC	CGGCGCC	1.55	1.737	-0.377	0.254	neutral
rs6692253	chr1	47680527	G	A	CEU	ACCTGTGAGCACA	ACCTGTAAGCACA	GTGAGCA	GTAAGCA	1.127	1.271	-0.377	0.531	neutral
rs10751450	chr1	203650945	C	T	AS,CEU	CTGTGGCCCTATC	CTGTGGTCCTATC	CCCTATC	TCCTATC	1.181	1.418	-0.394	0.567	neutral
rs2072814	chr22	32870769	C	T	CEU	CGGCTCCAGGAGG	CGGCTCTAGGAGG	TCCAGGA	TCTAGGA	1.158	1.436	-0.424	0.637	neutral
rs12937114	chr17	42325073	A	G	AS,CEU	TTTGGCAGCTGCC	TTTGGCGGCTGCC	TTGGCAG	TTGGCGG	1.719	1.36	-0.441	0.5	neutral
rs7177266	chr15	76195940	C	T	AS,CEU	AATTAGCTGGGCG	AATTAGTTGGGCG	AGCTGGG	AGTTGGG	1.246	1.416	-0.46	0.408	neutral
rs80137870	chr19	33182750	C	T	AS,CEU	CAACGCCAGAGGC	CAACGCTAGAGGC	CCAGAGG	CTAGAGG	1.456	1.772	-0.476	0.409	neutral
rs112035106	chr10	46089925	C	T	AS,CEU	TCCGCTCGGCCCG	TCCGCTTGGCCCG	GCTCGGC	GCTTGGC	1.159	1.369	-0.492	0.569	neutral
rs4737009	chr8	41630405	G	A	AS,CEU,orig75	TTTACCGAGAAAG	TTTACCAAGAAAG	GAGAAAG	AAGAAAG	1.018	1.249	-0.511	0.929	weak gain
rs76491632	chr10	45966598	T	C	AS	TATCCATTATAAA	TATCCACTATAAA	TCCATTA	TCCACTA	1.393	1.875	-0.525	0.551	weak gain
rs7183915	chr15	66097378	A	C	AS	TACCTGAACTCCA	TACCTGCACTCCA	TGAACTC	TGCACTC	1.096	1.276	-0.544	0.651	weak gain
rs17616316	chr14	103822762	C	G	AS,CEU,orig75	CCACTACAAGGAA	CCACTAGAAGGAA	ACTACAA	ACTAGAA	3.307	1.098	-0.566	0.957	weak gain
rs7114009	chr11	73115314	A	G	CEU	AAGCCTAGCGGAT	AAGCCTGGCGGAT	AGCGGAT	GGCGGAT	1.235	1.459	-0.567	0.489	weak gain

rs369043198	chr17	43669800	A	G	AS,CEU	GAACACAGGTCTG	GAACACGGGTCTG	CAGGTCT	CGGGTCT	1.135	1.387	-0.64	0.652	weak gain
rs1541253	chr1	203652040	T	C	AS,CEU	GACAACTACTATC	GACAACCACTATC	ACTACTA	ACCACTA	1.526	2.855	-0.701	0.716	weak gain
rs3752531	chr12	121157851	A	G	CEU	GAGTAGATAGGGG	GAGTAGGTAGGGG	ATAGGGG	GTAGGGG	1.29	1.687	-0.727	0.578	weak gain
rs901683	chr10	45966422	G	A	AS,CEU,orig75	TCTGTGGCTTTTC	TCTGTGACTTTTC	CTGTGGC	CTGTGAC	1.249	1.623	-0.847	0.6	weak gain
rs2974750	chr19	13044544	C	A	AS,CEU	CGGGCCCGGGTGG	CGGGCCAGGGTGG	GGGCCCG	GGGCCAG	1.143	1.52	-0.867	0.725	weak gain
rs2239760	chr12	121163518	C	A	CEU	CGGTTCCTCGCCT	CGGTTCATCGCCT	CGGTTCC	CGGTTCA	1.032	1.561	-0.884	0.944	weak gain
rs12911421	chr15	75287822	C	A	CEU	GGAAGTCTCGCGA	GGAAGTATCGCGA	AGTCTCG	AGTATCG	1.238	1.639	-0.913	0.627	weak gain
rs2238368	chr16	170328	C	T	CEU	GACAGACATTCTA	GACAGATATTCTA	ACAGACA	ACAGATA	1.252	1.639	-0.995	0.605	weak gain
rs1541252	chr1	203651927	T	C	AS,CEU	GTCTACTACTACA	GTCTACCACTACA	ACTACTA	ACCACTA	1.526	2.855	-1.023	0.716	strong gain
rs2075672	chr7	100240296	A	G	AS,CEU,orig75	GAAGGAAGGCATA	GAAGGAGGGCATA	GAAGGCA	GAGGGCA	1.149	1.871	-1.029	0.829	strong gain
rs2303316	chr17	37704217	A	G	AS	TAGTCAAGCAGTA	TAGTCAGGCAGTA	AGCAGTA	GGCAGTA	1.421	2.328	-1.044	0.683	strong gain
rs35968565	chr10	46076196	G	T	AS,CEU	GCTTGAGACAATA	GCTTGATACAATA	AGACAAT	ATACAAT	1.237	1.539	-1.221	0.56	strong gain
rs7547793	chr1	203653544	A	C	AS,CEU	CAGTGGAATGATC	CAGTGGCATGATC	TGGAATG	TGGCATG	1.204	1.508	-1.262	0.599	strong gain
rs589235	chr6	139839960	T	C	AS,CEU	ATCTCCTCTCCCC	ATCTCCCCTCCCC	CCTCTCC	CCCCTCC	1.082	1.52	-1.306	0.842	strong gain
rs79589869	chr17	43930238	C	A	AS,CEU	AATGCTCCTGTGC	AATGCTACTGTGC	TGCTCCT	TGCTACT	1.05	1.312	-1.418	0.839	strong gain
rs11650282	chr17	27090741	C	G	AS,CEU	ATCCTTCTGACAA	ATCCTTGTGACAA	CCTTCTG	CCTTGTG	1.294	1.771	-1.46	0.619	strong gain
rs3747093	chr22	21984379	G	A	AS,CEU	CTGCCCGGGCCCC	CTGCCCAGGCCCC	GGGCCCC	AGGCCCC	1.347	2.352	-1.53	0.744	strong gain
rs75595592	chr10	46039930	A	G	AS,CEU	CCACCAACTAAAG	CCACCAGCTAAAG	CCACCAA	CCACCAG	1.207	2.872	-1.53	0.89	strong gain
rs12459922	chr19	4455862	A	G	CEU	CACCTGACGGCCC	CACCTGGCGGCCC	CACCTGA	CACCTGG	1.203	1.736	-1.68	0.724	strong gain
rs9902953	chr17	27139834	C	A	AS,CEU	GCGATACCTCCCA	GCGATAACTCCCA	GCGATAC	GCGATAA	1.542	2.265	-2	0.572	strong gain
rs11866877	chr16	170044	G	A	CEU	AGATCTGGATAAG	AGATCTAGATAAG	CTGGATA	CTAGATA	1.102	1.696	-2.01	0.854	strong gain
rs8107610	chr19	33186579	C	T	AS,CEU	TGGCTTCGGCAGT	TGGCTTTGGCAGT	CGGCAGT	TGGCAGT	1.219	1.94	-2.093	0.767	strong gain
rs73425119	chr18	43753831	G	C	AS,CEU	AAGCCCGGGGAGC	AAGCCCCGGGAGC	AGCCCGG	AGCCCCG	1.195	2.297	-2.19	0.849	strong gain
rs13069307	chr3	142315074	G	A	AS,CEU	GCGCCCGTCCACC	GCGCCCATCCACC	CGCCCGT	CGCCCAT	1.36	2.565	-2.418	0.77	strong gain
rs10751451	chr1	203650978	C	T	AS,CEU	ATCTTACCGCTCC	ATCTTATCGCTCC	TTACCGC	TTATCGC	1.299	2.427	-2.922	0.791	strong gain
rs62064663	chr17	44080039	T	G	AS,CEU	AAGCCTTGGGGCG	AAGCCTGGGGGCG	TGGGGCG	GGGGGCG	1.737	3.113	-2.943	0.651	strong gain

Supplemental Table S3. Alignment and peak calling details.

ID	In-house DNase-seq mapping pipeline run details	Parameters peak caller
Human		
WIMM_primary_erythroid_Fibach_Fade8	trimming=TRUE; flashing=FALSE; bowtie1 -m 2 --maxins 356 --chunkmb 256 –lanes 1	merge=10; contig=20; depth=40
DFCI_LNCaP_Vehicle	trimming=TRUE; flashing=TRUE; bowtie1 -m 2 --maxins 350 --chunkmb 256 –lanes 1	merge=7; contig=12; depth=12
DFCI_MCF7	trimming=TRUE; flashing=TRUE; bowtie1 -m 2 --maxins 350 --chunkmb 256 –lanes 1	merge=7; contig=12; depth=12
Mouse		
HHMI_liDNase_mESC_100cells	trimming=TRUE; flashing=TRUE; bowtie1 -m 2 --maxins 350 --chunkmb 256 –lanes 1	macs2 default
HHMI_liDNase_mESC_30cells	trimming=TRUE; flashing=TRUE; bowtie1 -m 2 --maxins 350 --chunkmb 256 –lanes 1	macs2 default
WTHG_C57bl6_erythroblasts_term_diff_rep3_2	trimming=TRUE; flashing=TRUE; bowtie1 -m 2 --maxins 350 --chunkmb 256 –lanes 1	macs2 default