Multiplexed analysis of chromosome conformation at vastly improved sensitivity.

James O.J. Davies[1], Jelena M. Telenius[1,2], Simon McGowan[2], Nigel A. Roberts[1], Stephen Taylor[2], Douglas R. Higgs[1] and Jim R. Hughes[1]

[1]Medical Research Council (MRC) Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, Oxford University, Oxford, UK.

[2]Computational Biology Research Group, Weatherall Institute of Molecular Medicine, Oxford University, Oxford, UK.

Abstract

Current methods for analysing chromosome conformation in mammalian cells are either insensitive and low resolution or low throughput. Since available methods are both expensive and relatively difficult to perform and analyse they are not widely used outside of specialised laboratories. Here we have re-designed the Capture-C method producing a new approach, called next generation (NG) Capture-C. This produces unprecedented levels of sensitivity and reproducibility, which can be used to analyse any number of genetic loci and/or many samples in a single experiment. NG Capture-C is straightforward to perform, requiring only standard reagents and access to conventional next generation sequencing platforms. Importantly, high-resolution data can be produced on as few as 100,000 cells and SNPs can be used to generate allele specific tracks. The method should therefore greatly facilitate the task of linking SNPs identified by genome wide association studies with the genes they influence. The complete and detailed protocol presented here, with new publicly available tools for library design and data analysis, will allow most laboratories to analyse chromatin conformation at levels of sensitivity and throughput that were previously impossible.

Introduction

Progress in our ability to annotate regulatory elements in the genome and determine their potential function has been driven by technological advances, such as RNA-seq[1], ChIP-seq[2,3], DNase-seq[4] and ATAC-seq[5]. However, an outstanding challenge is to understand the mechanisms by which regulatory elements control specific gene promoters at a distance (10s to 1000s kb). Using conventional chromosome conformation capture (3C), it is possible to analyse in detail the interactions between enhancers, silencers, boundary elements and promoters at individual loci at high resolution[6-11]. Recently, we have shown, using a high-throughput approach (Capture-C), that it is possible to interrogate *cis*-interactions, at hundreds of loci at high resolution in a single experiment[12]. Such approaches are of immediate value in defining interactions within the chromosomal landscapes of many loci, and identifying the genes and the functional effects of SNPs that are associated with complex diseases, the majority of which lie in intergenic *cis*-acting regulatory elements[13-15].

The original Capture-C protocol[12] used oligos synthesized on a microarray (Agilent SureSelect) with a minimum design of 40,000 oligos, irrespective of the number of desired viewpoints, so the cost per sample is very high for small designs. Laboratories more often want to analyse a much smaller subset of regions in several different samples. Furthermore, the sensitivity possible with this design did not readily allow for the analysis of very long-range *cis*-interactions or *trans*-interactions and did not provide quantitative estimates of mega base scale chromosomal interactions.

To address these limitations we redesigned the Capture-C protocol to use biotinylated DNA oligos in solution so that each set of capture oligos can be designed specifically to capture from one to many hundreds of regions, in a single experiment and designs can be easily expanded by addition of new oligos to existing pools. Importantly, independent 3C libraries (e.g. from different cell types or different stages of development) can now be captured and processed in a single tube making separately indexed samples directly comparable. This greatly increases throughput and allows meaningful subtractive analysis of chromosome conformation in different cell types.

Using this approach, we have defined the smallest number of cells required to identify robust interactions and shown how allele specific interaction profiles can be generated from SNP containing regions.

RESULTS

Experimental workflow and overview of the method.

3C libraries are made using standard methods similar to the protocol for *in situ* HiC [16] (Fig. 1a, Supplementary methods). Prior to oligonucleotide capture, the 3C libraries are sonicated to 200bp followed by the addition of Illumina paired-end sequencing adaptors. This random sonication generates unique fragments prior to ligation of adaptors and PCR amplification. This is a significant advantage of Capture-C compared to 4C and 5C because PCR duplicates due to over-amplification of specific fragments can be removed bioinformatically. This allows the number of unique ligation junctions present in the 3C library to be quantified accurately (Fig. 1b).

Three factors influence the number of unique interactions that can be determined from each viewpoint in a 3C library. First, a theoretical maximum of only four interactions can be detected from each region of interest per cell (one from each end of the captured viewpoint fragment on each allele), so the number of cells contributing to the library determines the maximum number of interactions that can be detected. Secondly, the hybridisation efficiency of the capture probe is important, and this is largely dictated by the underlying sequence. Thirdly, the efficiency of the assay and depth of sequencing required, is determined by the proportion of non-specific background fragments contaminating the library. These fragments, originate from non-captured DNA.

To maximise the number of unique interactions defined, the NG Capture-C protocol has been optimized to analyse 3C material containing eight times more ligation junctions than the previous protocol. This has been achieved by minimising losses during the addition of sequencing adaptors to the 3C library, and mixing material from two parallel library preparations; allowing a total input of 10μg 3C library to be used. This at least doubles the complexity of the material used for the hybridisation reaction. In addition the amount of this material used in the hybridisation reaction has been increased four fold (from 500ng to 2μg).

To reduce the uncaptured, non-specific fragments in the Capture-C library following oligonucleotide capture we implemented two changes to the workflow. First, the library design was simplified so that only single 120bp biotinylated DNA oligonucleotides (rather than 3-5 overlapping oligos) are used to capture each end of the target restriction fragment. In addition, the capture probes are designed to include the restriction sites (Supplementary Fig. 1a) maximizing the capture of informative junction fragments. This makes additional steps, such as biotin fill in, unnecessary, which reduces losses in library complexity, which is a critical component of sensitivity particularly at low cell numbers. [17,18]. Crucially, a second, sequential round of capture

was introduced, which markedly reduces the background of uncaptured material and reduces the need for hugely deep and prohibitively expensive sequencing.

Initially, we tested a minimal design containing probes to only the *Hba-a1* and *Hba-a2* promoters, equivalent to a 4C-seq analysis. We found that a single oligonucleotide capture step enriches the targets ~ 5-20,000 fold over the capture site. Despite this the captured DNA from this single region of interest only makes up less than 1% of the sequenced reads; the remainder being uncaptured background (Fig. 2a i). In the NG Capture-C protocol the use of two sequential oligonucleotide capture steps results in up to 1,000,000 fold enrichment compared to uncaptured 3C library so that captured material now makes up approximately 50% (rather than 1%) of the sequenced material (Fig. 2a i and ii). This second capture step increases the number of PCR cycles (to 34 rather than 20) and the number of PCR duplicates (Fig. 2a iii and iv) sequenced because the library complexity (i.e. the number of interactions available to capture) limits the number of unique interactions that could be sequenced. The greatly improved enrichment means that the depth of sequencing is no longer limiting. However, using the Capture-C protocol, PCR duplicates are easily and efficiently excluded bioinformatically. This is demonstrated by the fact that no differences are seen in the interaction profiles (Fig. 2b) and there is little change in the GC content or read length (Supplementary Fig. 1b) of the sequenced material when comparing single captured and double captured libraries.

To demonstrate the scalability of the approach, we went on to combine the *Hba-a1* and *Hba-a2* capture probes together with capture probes for the *Hbb-b1* and *Hbb-b2* (the adult beta globin genes) and *Slc25A37* (*Mitoferrin 1*). The alpha and beta globin genes are amongst the most extensively characterized genes and their regulatory interactions have been interrogated by almost every 3C based method to date [7,9,10,12,19-22] so they provide important controls for the validation of any new methodology. The interaction profiles of all three genes were almost identical in the biological replicates (Supplementary Fig. 2) and match the previously determined patterns of interactions for these control genes (Fig. 2b, Supplementary Figs. 3-5). However, importantly, for the same depth of sequencing the double capture greatly increases the sensitivity of the profile 30 fold (Fig. 2b, Supplementary Figs. 6&7 Supplementary Table 1). To further demonstrate the power of the approach next we scaled up to a 35 gene design and increased the number of samples analysed in a single experiment, capturing seven pooled indexed libraries in a single assay. The efficiency that results from the double capture step, allowed us to sequence these 245 interaction profiles using a single Illumina HiSeq run (177 million reads).

After normalization of individual profiles for the total number of unique interactions across the genome from each viewpoint in each sample, the genome-wide correlation of

the two replicates for all genes exceed an R^2 value of 0.97, showing exceptional levels of correlation across biological replicates (Supplementary Fig. 1c). The coefficient of variation falls significantly (CV < 50%) when more than 10 normalised interactions mapped to any individual restriction fragment (Supplementary Fig. 1d). Thus ligation junctions present at 1 part in 10,000 in the 3C library can be detected reproducibly, since the data are normalised to a total of 100,000 unique interactions across the genome. Furthermore, the pattern of both short-range interactions (Supplementary Fig. 2&3) and long-range *cis*-interactions (Supplementary Fig. 6&7) are highly reproducible. In addition, NG Capture-C produces a more comprehensive profile than existing 4C-seq[10] (Supplementary Fig. 6&8) and does so regardless of restriction enzyme fragment size (Supplementary Fig. 1e).

A set of tools for design and analysis of Capture-C experiments has been developed (Fig. 1b). An online tool to generate oligonucleotide design for multiple targets can be found at http://apps.molbiol.ox.ac.uk/CaptureC/cgi-bin/CapSequm.cgi and analysis scripts are available via github (https://github.com/telenius/captureC/releases). The depth of data obtained following double capture allows unique interactions to be reported for each individual restriction fragment or half fragment (Supplementary Fig. 8b), which is the highest possible resolution for such experiments; there is no requirement to integrate data by using a moving window.

In summary, the substantial increase in signal also allows multiple 3C libraries (e.g. from different cell types or replicates) to be indexed and pooled prior to capture. This greatly increases the throughput of the assay, and importantly allows biological replicates and different experimental conditions to be processed and analysed together, removing sources of experimental variation.

Identification of tissue-specific regulatory elements using comparative analysis of chromosome conformation in different cell types

At present, there is no ideal way to consistently call all significant interactions from chromosome conformation data. Sequences from any capture point will interact with the surrounding genome, in a distance dependant manner, whether it is active or inactive (supplementary analysis section). Therefore, current analysis of 3C data typically includes approaches to normalise interaction data taking into account the distance from the viewpoint. In practice, the outputs from such approaches are highly dependent on the normalisation model and input parameters used. With all such approaches there is a tendency to under call *cis*-interactions with genuine regulatory sequences lying close to the capture point, where normalisation is most stringent.

The reproducibility of profiles generated from NG Capture C has enabled us to test a complementary approach to identify regulatory interactions by comparing different cell types. Substractive analysis of normalised data from erythroid and non-erythroid (ES) cells successfully identified all known regulatory elements in well characterised test loci (Fig. 3&4, Supplementary Fig. 3-5.) and in the same data identified similar interactions in the other less well characterised loci in the capture design which included clinically significant genes (*CD47* Supplementary Fig. 9) and complete regulatory networks (*Myc*, *Sox2*, *Oct4* (also known as *Pou5f1*), *Klf4* and *Nanog*) (Supplementary Figs. 10-14). Interestingly interactions with regulatory elements were identified over 1Mb from the capture point, that are consistent with previously reported high resolution Hi-C data (Supplementary Fig. 10&12).

This subtractive analysis also uncovered fine details of tissue-specific regulation of genes that are active in two or more cell types. For example, the *Pnpo* gene encodes Pyridoaxime 5'-phosphate Oxidase which is a rate limiting enzyme in the metabolism of vitamin B6 to produce pyridoxal 5'-phosphate[23]; an essential cofactor in the heme synthetic pathway in all cell types. This gene is specifically up-regulated in mouse erythroid cells under the influence of an erythroid-specific enhancer (HS-26).[12] Comparison of ES and erythroid data precisely and specifically identified HS-26 at a resolution sufficient to distinguish it from the promoter of a neighbouring gene (*Cdk5rap3*) located only ~1kb away (Supplementary Fig. 15).

Subtractive analyses of NG Capture C data not only showed new interactions in the specific cell type under investigation but also identified new patterns of interaction in the cell type used for comparison. For example analysis of the *Tal1* locus revealed one pattern of interaction in ES cells and another in erythroid cells; these cells acting as reciprocal controls for each other (Supplementary Fig. 16). It is important to note that

as this approaches relies on changes between active states its goal is to find regulatory elements rather than constitutive structural interactions.

The substractive profiles can be additionally statistically interrogated using common approaches for the differential analysis of NGS count based data, such as the Bioconductor package DEseq2[24] and we compare the effectiveness of this approach at identifying known regulatory elements with two tools commonly used for 3C analysis (Fig. 3, Supplementary Figs. 3-5 and supplementary analysis section). FourCseq[25] and r3C-seq[26] were used as they also use replicates and comparative analysis but additionally normalise for genomic distance using different models. We tested all approaches on the well-characterized test loci; α globin; β globin and *Slc25A37*, using default parameters to simulate the output at uncharacterized loci. Of these three loci α and β globin are used as gold standards in the 3C field due to the depth of the functional knowledge of their regulation. We show that these tools call the known elements in the β globin and *Slc25A37* loci, but variably miss the most proximal elements in the α globin locus, unlike the comparative approach which calls all of the known elements in each locus (see supplementary analysis section).

Reproducible megabase scale *cis* and *trans*-interactions can be identified

Previous Capture-C data did not readily identify weak long-range interactions. The increased sensitivity of NG Capture-C enabled us to investigate such interactions and, importantly, evaluate their relative strength compared to local interactions. Analysis of interaction frequencies across the whole length of the chromosome containing a captured region of interest shows that interactions with the entire chromosome are not easily seen when viewed on the same scale as interactions with the more local regulatory elements . However, reproducible, low level (<100 fold) *cis*-interactions are detected with many other active regions of the chromosome (Supplementary Fig. 7). Similar patterns of general interactions can also been seen in *trans* but these are a further 10 fold weaker than the long-range *cis* interactions (Supplementary Figs. 17-19). The patterns of *trans* interactions become visible when the threshold for any interaction is reduced to fewer than 250 interactions per 100kb and these interactions have similar distributions independent of the gene promoter used as the view point (Supplementary Figs. 18&19) and are correlated with gene density and the number of active promoters, enhancers and CTCF sites (Supplementary Fig. 20). This is of particular interest in the case of the alpha and beta globin genes as they have been reported to interact with each other in erythroid cells. Some have suggested that these interactions are frequent[27] whereas others have shown them to be rare[28]. Now, with the sensitivity and robust quantitation provided by the NG Capture-C approach, the *trans*-interaction between

these two genes appears to be rare (~1000 fold less than local *cis*-interactions) and on the same scale as those with most other active regions of the genome. This increased sensitivity allows interactions to be detected that are unlikely to be functional but this allows us to be confident that all functional interactions quantifiable by 3C approaches can be detected.

In summary we have shown that significant and reproducible albeit weak very long-range interactions also exist in *cis* and *trans.* Furthermore, all analytical approaches tested discriminate appropriately between these weak interactions and the much stronger interactions with known regulatory elements (See supplementary analysis).

Robust interaction profiles can be generated from small cell numbers

When analyzing human primary tissues cell numbers are often a limiting factor and so the NG Capture-C protocol was further adapted to allow interaction profiles to be determined for small numbers of cells (see Supplementary Methods). The 3C library preparation on smaller numbers of cells did not alter digestion efficiency and the amount of DNA extracted per cell was constant despite reducing cell numbers. The preparation of material for the hybridisation reaction was optimised for the reduced DNA content of the 3C libraries. Reducing the number of cells to 100,000 resulted in a reduction in the number of interactions from an average of 137,000 (when cell number was not limiting) to 19,000. However, the interaction profiles with the known enhancers at the alpha and beta globin loci were virtually unchanged compared to the analyses where the cell numbers were not limited (Supplementary Fig. 21) although weak, long-range interactions became difficult to determine reproducibly (Supplementary Fig. 6).

Generation of SNP specific interaction profiles

SNPs underlying GWAS traits are frequently present on one allele but not the other and may affect the regulatory interactions of the affected allele. Therefore, it would be of great value to generate allele-specific interaction tracks. Due to the greatly improved depth of signal provided by the NG Capture-C protocol it was possible to distinguish separate alleles when a SNP is included within a capture point and hence sequenced (Fig. 4a). These allele-specific interaction profiles show that over 95% of the strain-specific SNPs in *cis* are in phase with the captured strain specific SNP, showing interactions with the sister chromatid are relatively rare (Fig. 4b). In this case, the interaction profiles were very similar probably because none of the SNPs were of functional importance (Supplementary Fig. 22).

This type of analysis also applies to non-allelic SNPs. The parologous *Hba-a1* and *Hba-a2* genes in mouse are almost exact copies which differ at only a few base positions, one of which lies near the 5' capture point for this gene (Fig. 4c). This allows separate interaction profiles for the *Hba-a1* and *Hba-a2* genes to be generated (Fig. 4c). The 5' *Hba-a1* interacts with the proximal regulatory elements (HS-12 and R4) more frequently than *Hba-a2* interacts with them. Interestingly, *Hba-a2* has very similar interactions compared to *Hba-a1*, with the MCSR1 and R2 regulatory elements, which are thought to have stronger enhancer function than the other elements[29,30].
Together these data show that using NG Capture C it is possible to analyse allele specific interactions and *cis*-interactions between a regulatory element and two or more duplicated non-allelic paralogues.

Discussion

NG Capture-C was developed to generate a completely flexible assay allowing researchers to analyse interactions involving single or many genes and multiple samples, simply and cheaply. NG Capture-C is able to detect interactions present 1 in 5-10,000 cells, which is far exceeds the current reasonable limit of detection by fluorescence *in-situ* hybridisation (FISH)[31].

The investigation of gene regulation is not only limited by the number of genes or elements that can be interrogated, but also by the number of replicates, conditions, cell types and genetic variants that can be easily analysed. The huge increase in signal of NG Capture-C allows for the simultaneous capture of multiple samples in a single reaction, greatly increasing the throughput and economy of the assay. In practice this allows complete networks of important genes, such as those encoding the Yamanaka pluripotency factors[32] (*Myc, Sox2*, *Oct4*, *Klf4*, Supplementary
Figs. 10-14) to be analysed simultaneously in multiple cell types. The data are compatible with standard analytical tools and their reproducibility and comparability between active and inactive states of NG Capture-C provides a complementary approach to the statistical identification of regulatory elements. This complementary approach identifies all known regulatory elements at well characterised test loci at levels of resolution previously not possible. Importantly, mindful of the current challenges in the analysis of GWAS and regulatory variants, the NG Capture-C method has been optimized to be effective at smaller cell numbers (~100,000) and to generate SNP-specific interaction profiles.

It is important to note that unlike most other high resolution chromosome conformation methods, NG Capture-C provides sufficient depth of data that the output is

expressed as “raw counts” per fragment; there is no need to integrate interactions via a moving window[10,12,16]. Furthermore, the sensitivity provided by double capture together with the ability to remove PCR duplicates means that the interaction data faithfully represent all interactions within the library allowing researchers to make estimates of relative quantitation between weak and strong interactions. The complete and detailed protocol presented here, with new publically available tools for library design and data analysis are intended to allow any laboratory to perform chromatin conformation capture analysis of the highest quality and at levels of throughput that were previously impossible.

Accession codes

All of the Capture-C data sets are available at the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) accession code GSE67959.

Acknowledgements

James Davies would like to thank the Wellcome Trust for funding his work (Wellcome Trust Clinical Research Training Fellowship ref 098931/Z/12/Z). The work was also supported by a Wellcome Trust Strategic Award (reference 106130/Z/14/Z) and the Medical Research Council (MRC Core Funding and Centenary Award reference 4050189188).

We would like to thank E. Repapi for statistical advice and L. Hanssen, M. Oudelaar, D. Jeziorska, B. Graham, M. Kassouf, M. Suciu, H. Long, S. Pasricha, V. Buckle, T. Milne, T. Fulga, T. Sauka-Spengler and A. Drakesmith for their critique of the manuscript. We would like to thank Supat Thongjuea for discussions on analysis.

Author Contributions

J.D. performed the experiments; analysed all the data and wrote the manuscript. J.R.H. designed the experiments, assisted with the bioinformatic analysis and wrote the manuscript. N.A.R. assisted with the experiments. J.M.T., S.M. and S.T. assisted with the bioinformatics analysis and prepared the software for public release. D.R.H wrote the manuscript.

List of Figures and figure legends

Figure 1. Overview of the method

a. Experimental workflow. 3C libraries are made using a very similar method to the protocol for *in situ* Hi-C: namely, formaldehyde crosslinking of live cells (1); restriction enzyme digestion of chromatin (optimized for a four cutter restriction enzyme (e.g. Dpn II)) (2); ligation (3); de-crosslinking and DNA extraction (4). In order to prepare the 3C library for oligonucleotide capture the material is sonicated, which randomly generates ~200bp fragments (5). Sequencing adaptors are then ligated and different indices are added by ligation-mediated PCR (6). Differently indexed samples can then be pooled (7) prior to hybridization with biotinylated oligonucleotides, which allows a single capture reaction to be performed on multiple samples. The captured sequences are then pulled down using streptavidin beads (9) and the material is PCR amplified off the beads using the P5&7 sequences in the sequencing adaptors (10). Steps 8-10 are then repeated. This results in very significant further enrichment; up to 3,000,000-fold over the baseline uncaptured 3C library. The material is then sequenced using either Illumina Miseq (150bp, paired end) or Hiseq (100bp, paired end). Note that the clustering on the flow cell uses the same PCR primers as all of the other PCR steps in the protocol and that 35 cycles are used for clustering compared to 34 cycles in the entire double capture protocol.

b. Data analysis. 1. The raw data are taken in FASTQ format. 2. Initially, the paired end reads are reconstructed into single sequences using the central area of overlap to align the sequences. This is possible for 95% of the reads because the material is sonicated into 200bp fragments, which are then sequenced with 300bp reads (150bp paired end). 3. Next, each read is split *in silico* using the restriction enzyme recognition sequence. This ensures that the reported ligation junctions contain the correct restriction enzyme cut sequence. This splits the reads into its component restriction fragments and the read name is used to link sets of fragments from the same read. 4. Reads that do not contain a sequence that maps inside the captured viewpoint restriction fragment are discarded. 5. Reads that are not unique (based on the sonicated ends) are removed. 6. Interactions are only reported when the entire sequenced read is unique and when one component of a read pair maps completely within a captured fragment and the other maps outside all of the capture fragments and proximity exclusion regions in the experiment. The proximity exclusion zones are normally set at 1kb on either side of the captured viewpoint fragment. This is done to prevent undigested material being reported as interacting and to prevent interactions being falsely reported from fragments that could be captured by two different oligonucleotides. The data are then filtered to remove regions with problematic mappability due to copy number differences[33] and mis-mapped reads from the proximity exclusion region. Due to the depth of the sequence data obtained following double capture, unique interactions can be reported for each individual restriction fragment or half fragment (Supplementary Fig 2b), which is the highest possible resolution for such experiments; there is no requirement to integrate data by using a moving window.

Figure 2. Comparison of single and double oligonucleotide capture

3C material generated from erythroid cells was captured using a single set of oligonucleotides designed to the alpha globin promoters (Supplementary Table). Since the two copies of the gene are virtually identical interaction profiles are generated from both genes simultaneously. After the first oligonucleotide capture step some of the material was sequenced using the Illumina MiSeq. The remaining library was used as input for a second round of oligonucleotide capture and the resulting material was then sequenced.

a. Comparison of the enrichment (to scale) resulting from the single and double capture. (i) Single capture results in 5-20,000 fold enrichment but this only results in around 0.3% of the reads containing a sequence that maps to the captured fragment. (ii) Double capture increases the enrichment markedly; producing up to 3,000,000 fold enrichment. This dramatically increases the percentage of reads containing a restriction fragment that maps to the capture region from 0.3% to 48.6%. The number of unique interactions is increased around 30-fold following double capture (from 10,832 to 327,787) (iii & iv) because the library complexity now becomes the limiting factor.
b. Comparison of the raw informative interactions count per restriction enzyme fragment for single and double capture. The red vertical lines denote the location of captured viewpoints. The light blue lines highlight the five well described regulatory elements in the mouse (R1, R2, R3, R4 and HS-12). This shows that double capture does not alter the significantly alter the local interaction profile but it has 30-fold increased sensitivity.

Figure 3. High-resolution identification of regulatory element by comparative analysis between active and inactive states

Top panel shows the overlaid normalized mean Capture-C profiles from eythroid (genes active in red) and ES cells (genes inactive blue) at three erythroid specific loci alpha globin, beta globin and *Slc25A37* (*Mitoferrin 1*) in (erythroid n=4 and ES cells n=3). These data were generated along with the profiles for another 32 gene promoters simultaneously from seven samples in a single capture reaction (making a total of 245 interaction profiles from one oligonucleotide capture reaction). The Y-axis denotes the mean number of unique interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide.

The captured viewpoint fragments are highlighted in red and the interactions with the well-known enhancers as annotated by DNAseI hypersensitivity are highlighted as black hatched lines. The differential track (Δ Capture-C) shows that interactions with the local erythroid enhancers are clearly and specifically increased in erythroid cells when the genes are active. Below this DESeq2 analysis of the differential enrichment (minus $\log_{10}$ adjusted p values) mapped across the three loci shows the highly significant enrichment of the known regulatory interactions.

Figure 4. SNP specific interaction profiles

a. This shows a density plot of the reads mapping to the captured restriction fragment (the *Tal-1* promoter fragment is shown). SNPs under the captured promoter allowed us to generate allele specific interaction profiles in F1 crosses between

C57BL/6 and CBA/J mice (see also Supplementary Figure 17). In the example locus the SNP rs252622560 has been used to separate interactions from the two different alleles.

b. This shows a graphical representation of the % of SNPs in phase in the interacting reads compared with the strain of the captured allele in *cis*. This demonstrates that the chromosome predominately interacts with itself in *cis* rather than its sister chromatid.

c. SNP specific NG Capture-C has been used to generate specific interaction profiles for *Hba-a1* and *Hba-a2* parologous genes. A single nucleotide difference between the two genes allows generation of specific tracks (see inset). *Hba-a1* is the more active of the two genes, producing around 70% of the total mRNA. Comparison of the two biological replicates shows that the SNP specific profiles are highly reproducible. The Δ Capture-C track shows the difference of the mean *Hba-a1* and *Hba-a2* profiles. This reveals that that the *Hba-a1* gene preferentially interacts with the enhancers, particularly proximal HS-12 and R4 elements. The *Hba-a2* gene interacts much more strongly with the chromatin between the two genes and interestingly it interacts with the most distal enhancer (R1) to a very similar degree to the *Hba-a1* gene.

Supplementary material

Online Methods

Preparation of 3C libraries

Single cell preparations of erythroid cells were made by gently dissociating cells from the spleen of a mouse treated with phenylhydrazine (40mg/g body weight x3 doses 12h apart; sacrificed on day 5). Phenylhydrazine causes haemolytic anaemia and marked erythroid expansion in the spleen so that 80% or more of cells are erythroid cells (as defined by CD71+ ter119+). The cells were passed through a 40μm cell strainer to remove clumps. For ter119 selection, cells were stained with ter119-phycoerythrin (PE) and purified using anti-PE MACS beads (Miltenyi Biotec) prior to fixation with formaldehyde. Mouse E14 ES cells were trypsinised and washed once prior to fixation.
Each aliquot of 10^7 cells was resuspended in 10ml of RPMI with 10% FCS in a 15-ml conical centrifuge tube. 549μl 37% (vol/vol) formaldehyde was added to each aliquot to make an overall concentration of 2% (vol/vol). A 10 minute incubation was performed at room temperature on a roller mixer. The crosslinking reaction was then quenched with 1.5ml cold 1M glycine and the sample was centrifuged immediately for 5 min at 220g in a precooled centrifuge at 4°C. The supernatant was discarded and the pellet was gently resuspended in 10ml cold Phosphate Buffered Saline (PBS). The cells were centrifuged again (5 min 220g 4°C) and the supernatant discarded. The pellet was resuspended in 5ml cold lysis buffer and incubated on ice for 20 min. The nuclei were centrifuged (5 min 500g 4°C) and the supernatant carefully removed. Multiple aliquots can be snap frozen using liquid N_2 or dry ice and ethanol and stored for several months at -80°C. Cells were resuspended in 1ml water (MilliQ or Sigma) and Dounce homogenised (45 strokes; 5ml Dounce homogeniser) on ice. The sample was pelleted (5min 22,000g 4°C) and resuspended to a total of 650μl water (Milli-Q or Sigma). Three reactions were set up for each sample in 1.5ml Eppendorf Safe-Lock microcentrifuge tubes. Each digestion reaction was made up of 200μl cell suspension; 80μl of x10 restriction enzyme buffer; 10μl SDS 20% (vol/vol) and water to make a final volume of 800μl after the later addition of 66μl Triton X-100 and restriction enzyme. A control reaction to check for nonspecific digestion (final volume 200μl) was also set up in a 1.5ml Eppendorf tube. This included 50μl cell suspension; 40μl x10 restriction enzyme buffer; 2.5μl SDS 20% (vol/vol) and 111μl water, making a total of 200μl after the addition of Triton X-100. All reactions were placed on a thermomixer (Eppendorf) for 1h at 37°C shaking at 1400 r.p.m. 66μl Triton X-100 was added to each of the digestion reactions and 16μl to the control reaction followed by an incubation of 1h on the thermomixer (37°C 1400 rpm). Three aliquots of 500U restriction enzyme were added to each digestion reaction several hours apart. The samples were incubated on

the thermomixer (37°C 1400 rpm) for 16-24h after the initial dose of restriction enzyme.
100μl was removed from each digestion reaction and pooled to make a control to assess digestion. The DNA was extracted from the two controls using a standard phenol/chloroform extraction (including proteinase K and RNAse steps).
The restriction enzyme in the digestion reactions was heat inactivated by incubating at 65°C for 20 mins. The samples were then cooled on ice and 500μl water (Sigma); 133μl x10 ligation buffer and 8μl high concentration T4 DNA ligase (Thermoscientific, 30U/μl) was added to each digestion reaction. The samples were then agitated at 1400 r.p.m. overnight using the thermomixer at 16°C.
To decrosslink the ligated material, 5μl Proteinase K (Thermoscientific >600U/ml) was added to each reaction and incubated at 65°C overnight. The three reactions were pooled in a 15ml conical centrifuge tube; 30μl RNAse (Roche) was added prior to an incubation at 37°C for 30 min. DNA was purified from the reaction using a phenol choloroform extraction ((4ml) of phenol/chloroform/isoamyl alcohol (25:24:1) / 4ml chloroform). The DNA was precipitated in a large volume to improve removal of DTT. The 4ml sample was placed in a 50ml tube with 7ml water (Milli-Q); 1.5 ml 2M sodium acetate and 35ml 100% ethanol. The samples were frozen (-80°C for at least 2h) and centrifuged at 20,000 g for 30min at 4°C. The pellet was then washed with 10ml 70% ethanol dry and dried room temperature prior to being reconstituted in PCR grade water. This '3C library' can be stored at -20°C for several months.

3C library controls

To determine the efficiency of digestion and ligation and check for non-specific digestion 10μl of each control and 5μl of the 3C library was run on a 1% (wt/vol) agarose gel. The digestion efficiency was also checked using qPCR with primers designed across one of the restriction enzyme digestion sites at the alpha globin promoter (DpnII digestion control) and another primer set that lies close to the other end of the same restriction fragment (*Hba-a1&2* control primer). Digestion efficiencies were always in excess of 70% for libraries used for analysis. The concentration of DNA in the 3C library was determined using Qubit (BR).

Real time primers for assessing digestion
DpnII digestion control forward primer GTGTCACCAAAACCAGCTCA
DpnII digestion control reverse primer CCTGGAATCCTTTGGCTCAAG
DpnII digestion control Taqman probe GGGCAGCTAAGATGCAAGTC

Hba-a1&2 control forward primer TGGAGGGCATATAAGTGCTACTTG
Hba-a1&2 control reverse primer TGCTTTTGTCTTCCCCAGAGA
Hba-a1&2 control Taqman probe TGCAGGTCCAAGACACTTCTGATTCTGACA

Addition of Sequencing Adaptors

5μg of 3C library was sonicated to 200bp using a Covaris S220 Focussed ultrasonicator (6 cycles of 60s: duty cycle 10%; intensity 5; cycles per burst 200). The degree of sonication was confirmed using an Agilent Bioanalyser or Tapestation (DNA 1000). Illumina Truseq indexed sequencing adapters were added using NEBnext reagents (E6000 / E6040 / E7335 / E7500). This involved end repair, addition of overhanging A bases, ligation of adapters and PCR to add the indices. The DNA was cleaned up between reactions using Ampure XP beads at a 1:1.8 ratio for all clean up steps to minimize the selection of larger fragments and losses of material were minimised. 6-8 cycles of PCR were used when addition the Truseq indices using the Agilent Herculase II PCR kit. Generally 1.5-2μg of adapter ligated material was generated, however, to maximize library complexity the library preparation was usually done in duplicate (to use 10μg of input material) and the samples were pooled. The libraries were analysed using an Agilent Bioanalyser or Tapestation (DNA 1000) both pre and post the PCR and addition of sequencing adaptors as this allowed the DNA losses (and library complexity) to be assessed prior to amplification.

Biotinylated DNA oligonucleotides (IDT ultramers or Sigma long synthesis) were reconstituted to a concentration of 2.9μM. This allowed different oligonucleotides to be mixed in equimolar quantities so that 4.5μl of the resulting library would always contain a total of 13pmol of oligonucleotide pool. These oligonucleotides are vastly in excess in the hybridization reaction so that contamination with very small quantities can results in significant capture, which can lead to spurious results. We recommend that oligonucleotides for different experiments are ordered from the manufacturer and handled separately as contamination can occur during the manufacturing process.

Oligonucleotide Capture

1.5-2μg of adapter ligated material was placed in a 1.5ml microcentrifuge tube with 5μg COT DNA from the appropriate species; 1000pM Nimblegen HE Universal blocking oligo and 1000pM Nimblegen HE Index specific blocking oligo (corresponding to the Illumina TS index used). The sample was then dried using a vacuum centrifuge (50-60°C) until no liquid remained. The residue was dissolved in 7.5μl Nimblegen Hybridization Buffer and 3μl Nimblegen Hybridization Component A followed by denaturation at 95°C for 10 minutes. Concurrently 4.5μl of the biotinylated capture oligonucleotide library (total 13pM) was heated to in a 0.2ml PCR tube to 47°C in a PCR block. After 10 minutes the 3C library and blocking oligonucleotides were added to the preheated biotinylated oligonucleotides at 47°C. The hybridization reaction was incubated in a PCR machine at 47°C for 64-72h (with a heated lid at 57°C).

The Nimblegen SeqCap EZ Wash Buffers (I, II, III, Stringent and Bead Wash Buffers) were prepared and where necessary preheated to 47°C using the thermomixer. 100μl

M270 streptavidin beads were aliquoted into a 1.5ml microcentrifuge tube and allowed to warm to room temperature for 30 min. Two washes with 200μl Bead Wash Buffer were performed, using a DynaMag device to capture the beads and allow the supernatants to be discarded. After the final wash the hybridization reaction was added directly to the beads and mixed thoroughly by pipetting up and down and vortexing. The samples were put into the thermomixer at 47°C and mix at 500 rpm for 45 minutes. After 45 minutes 100μl of Wash Buffer I, heated to 47°C, was added and the samples were mixed by vortexing for 10 seconds. The tube was placed in a DynaMag device and the liquid discarded once it became clear. 200μl Stringent Wash Buffer, heated to 47°C, was added and mixed before incubating at 47°C for 5 minutes. The tube was then put into a DynaMag device and the liquid was discarded once it became clear. This step was repeated twice so that two washes were performed with Stringent Wash Buffer. 200μl of Wash Buffer I was added to the sample at room temperature and it was mixed by vortexing for 2 mins. The tube was then returned to the DynaMag device and the liquid discarded once it had become clear. 200μl of Wash Buffer II was added and mixed by vortexing for 1 minute. Then the tube was returned to the DynaMag device and the liquid discarded. The beads were then resuspended in 200μl of Wash Buffer III and the sample was mixed by vortexing for 30 seconds. The tube was replaced in the DynaMag device and the liquid discarded once it became clear. The beads were resuspended in 40μl of PCR grade water (the beads can be stored at -15 to -25°C at this point). The captured material was PCR amplified directly from the beads using either the SeqCap EZ Post-Capture LM PCR Master Mix and Post LM-PCR oligos (x18 cycles) or the newer Kappa master mix supplied in the SeqCap EZ accessory kit v2 (x14 cycles). An Ampure-XP bead clean up was then performed and the captured material removed from the beads using 30μl PCR grade water (Sigma). The captured material was assessed using the Agilent Bioanalyser or Tapestation.

This material was then used as input for the second round of oligonucleotide capture. The hybridization reaction was set up as for the first capture although less input material was used. 75% of the material up to a total of 2μg was used for the second hybridisation reaction since it is likely that thousands of copies of each captured ligation junction are present in the library by this point. For the second round of capture the material was only hybridized for 24h rather than 64-72h. The bead washes and PCR amplification of the material were identical to the first capture.

Following the second capture the mass of captured material was assessed using the Agilent Bioanalyser or Tapestation and Qubit. A 4nM solution (the concentration required for loading the Illumina MiSeq) was made using the size of the fragments assessed by the Bioanalyser or Tapestation and the concentration measured by the Qubit. Oligonucleotide capture enrichment can be determined by real time PCR, using the *Hba-a1&2* control primers above and a standard curve of genomic DNA to compare to the concentration of the input material determined by Qubit.

Multiplexed library capture

Multiple samples can be captured simultaneously by labelling them with different index adaptors and mixing prior to the oligonucleotide hybridization. In order to maintain library complexity, for the first capture, 1-2 μg from each sample was pooled in an exact 1:1 stoichiometry. It is important to do this precisely as the percentage of reads obtained from each sample will be directly related to the amount of DNA mixed. 5μg COT DNA and 1000pmol of universal TS HE blocking oligonucleotides were added for each sample and 1000pmol of the index specific blocking oligonucleotide was added for each sample. The mixture can then either be split into multiple identical hybridization reactions each of the same volume of a single sample or one large hybridization reaction can be made. The hybridisation, streptavidin bead capture and wash protocols were followed as outlined above, except that the volumes were adjusted appropriately when larger volume captures were undertaken. The PCR reactions were performed using the same volumes as for a single capture (multiple reactions were performed in parallel). For the second capture, the material was pooled from all of the PCRs and a single second capture was performed on this material. It is possible to use a single volume capture at this point because the library should contain thousands of copies of each captured read and so it is unlikely significant complexity will be lost during the second capture.

Sequencing

A 4nM solution of the libraries was made using the fragment size from the bioanalyser or tapestation and the overall concentration measured by the Qbit. The concentration can also be confirmed using real time PCR (SYBR green) with the P5 and P7 sequences on the adaptors. The majority of material was sequenced using the Illumina Miseq (300bp V2 chemistry), which produced 10-20 million 150bp paired end reads depending on the cluster density. One larger experiment was sequenced on the Illumina HiSeq producing 100bp paired end sequences.

Adaptations for reduced cell numbers

The 3C library preparation was performed as above with the following adaptations: a) the volume of the digestion reaction was reduced to 200μl for 3 million cells or less and 50μl for 500,000 cells or less. When less than 1 million cells were used to save material the two control samples were omitted and digestion efficiency was assessed on the ligation reaction using real time PCR. The CT value for the ligation reaction is nearly identical to the digestion control because the probability of the fragment ligating back to its original position appears negligible compared to the proportion of undigested material. The entire library preparation was performed in a single 1.5ml Eppendorf tube to minimize losses. The phenol-chloroform extraction was performed as above except that the DNA precipitation was performed in a smaller volume (x3 volume 100%

ethanol; 1/10th volume NaOAC 2M; 1µl glycogen (Invitrogen) as carrier). All of the material was sonicated to 200bp and sequencing adaptors were added using the NEBnext Ultra DNA library prep reagents (E7370). Additional PCR cycles were used to compensate for the smaller quantities of DNA (10 cycles for 500,000 cells / 12 cycles for 100,000 cells). Following this the material underwent a double oligonucleotide capture as outlined above.

Data Analysis

Initially the adaptor sequences are removed from the reads in the raw FASTQ files using Trim_galore (a wrapper tool around Cutadapt and FastQC; Babraham Institute http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). This is particularly necessary when using 150bp paired end sequencing because some of the reads are shorter than 150bp and the sequence will extend into the adaptor. The paired end reads were then reconstructed into single reads, where possible using FLASH with interleaved output settings [34]. These two steps can be omitted when shorter reads are used that do not have an area of central overlap and instead a file of these reads is generated with the paired end reads interleaved in strict order (read 1 FASTQ followed by read 2). An *in silico* restriction enzyme digestion of the reads was then performed using the script (DpnII2E.pl https://github.com/telenius/captureC/releases) with the name of the read being used to keep a record of each sub-fragment. The resulting FASTQ file of sub-fragments was then aligned using bowtie1 (using P1, M2, best and strata settings). Fragments that result from non-specific ligation and do not contain the restriction cut sequence will not be mapped to the genome by bowtie 1 and are therefore discarded. It is important that the reads are in strict order for the subsequent analysis, which can be achieved either by sorting based on the name or using one processor for the alignment.

The resulting sam file is then analysed with the main script CCanalyser2.pl (https://github.com/telenius/captureC/releases). This classifies the sub-fragments as either being: a) "capture" if they are contained within the capture fragment; b) "proximity exclusion" if they are inside the defined proximity exclusion coordinates (usually 1kb on either side of the capture fragment) or c) "reporter" if they are outside of all of the capture and proximity exclusion regions in the entire experiment. PCR duplicates were excluded by removing reads that had the same start and end coordinates of each sub-fragment. For long-range *cis* and *trans* analysis the start and stop coordinates of the interacting read itself also had to be unique. This more stringent filter was used to remove PCR duplicates because occasionally sequencing errors in the captured restriction fragment allowed PCR duplicates to appear unique. Unique interactions were only reported when the read was unique and there were one or more "reporter" and a single "capture" sub-fragment defined from a single read.

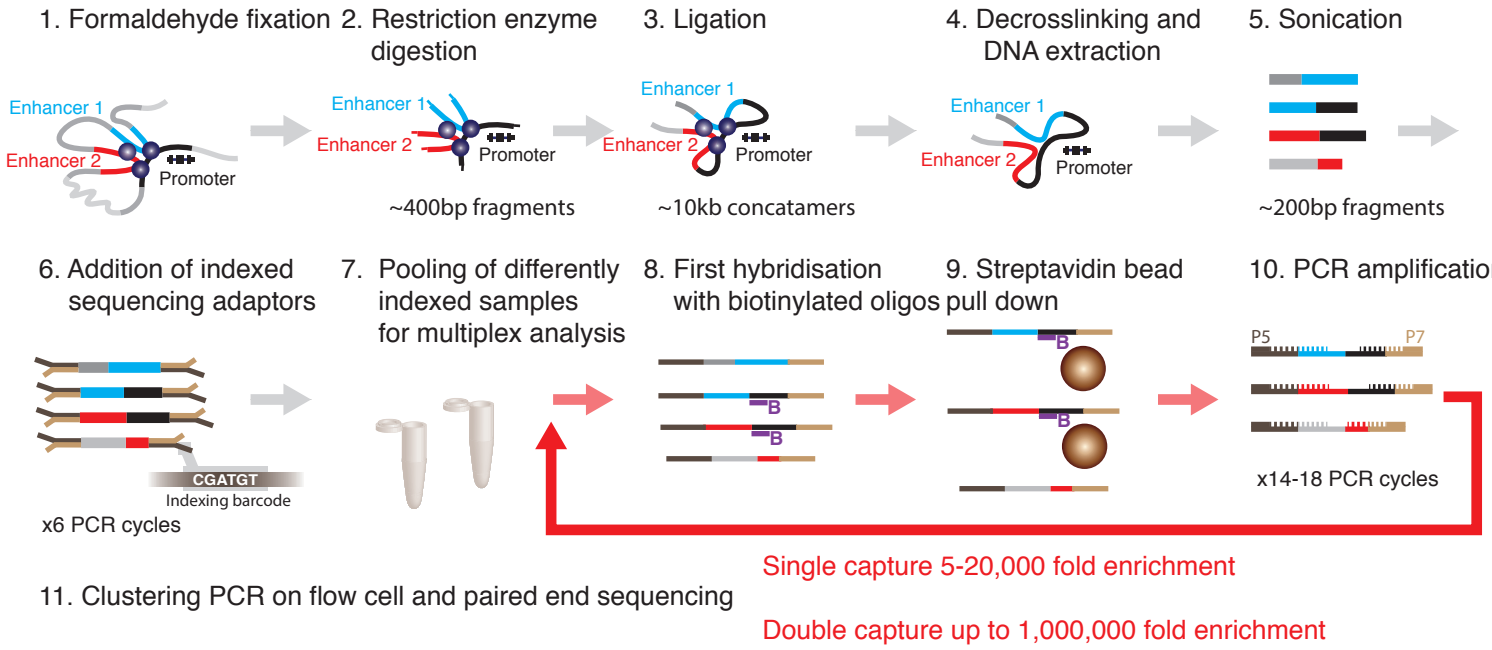
CCanalyser2.pl can map the reads either to the whole restriction enzyme fragment or, to give the maximum resolution possible, they can be mapped to the half fragment based on the mid point of the read and restriction fragment.
CCanalyser2.pl is also capable of creating SNP specific tracks, in which a specific base has to be present at a specific position in the capture fragment for the data to be included.
The data are then filtered to remove regions with problematic mappability due to copy number differences and mismapped reads from the proximity exclusion region. The latter was achieved by mapping the sequence of the proximity exclusion zone back to the genome using BLAT. Restriction fragments outside of 2Mb from the viewpoint (this was chosen so that gene duplications, such as *Hba-a1&2* were not excluded) were excluded if the proximity exclusion zone mismapped to them. The read count per fragment was normalized to the total number of reads in the track to give the number interactions per 100,000 interactions in the whole track using R. These data were subsequently converted to a format suitable for viewing in the UCSC genome browser (http://genome.ucsc.edu/)[35,36].
Statistical analysis was performed using DESeq2[24]. Unnormalised raw counts per restriction fragment were used for this analysis and restriction fragments with no reads mapping to them were excluded from the analysis.

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
2. Mikkelsen, T.S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-60 (2007).
3. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**, 651-7 (2007).
4. Hesselberth, J.R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **6**, 283-9 (2009).
5. 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).
6. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-11 (2002).
7. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* **10**, 1453-65 (2002).
8. Noordermeer, D. *et al.* The dynamic architecture of Hox gene clusters. *Science* **334**, 222-5 (2011).
9. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-13 (2012).
10. van de Werken, H.J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods* **9**, 969-72 (2012).
11. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499-506 (2013).
12. Hughes, J.R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* (2014).
13. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* **46**, 136-43 (2014).
14. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-5 (2012).
15. Parker, S.C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A* **110**, 17921-6 (2013).
16. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
17. Jager, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* **6**, 6178 (2015).
18. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res* **25**, 582-97 (2015).
19. Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G. & Higgs, D.R. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J* **26**, 2041-51 (2007).

20. Hughes, J.R. *et al.* High-resolution analysis of cis-acting regulatory networks at the alpha-globin locus. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120361 (2013).
21. Bau, D. *et al.* The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18**, 107-14 (2011).
22. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**, 1348-54 (2006).
23. Kang, J.H. *et al.* Genomic organization, tissue distribution and deletion mutation of human pyridoxine 5'-phosphate oxidase. *Eur J Biochem* **271**, 2452-61 (2004).
24. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
25. Klein, F.A. *et al.* FourCSeq: analysis of 4C sequencing data. *Bioinformatics* (2015).
26. Thongjuea, S., Stadhouders, R., Grosveld, F.G., Soler, E. & Lenhard, B. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Res* **41**, e132 (2013).
27. Osborne, C.S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* **36**, 1065-71 (2004).
28. Noordermeer, D. *et al.* Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat Cell Biol* **13**, 944-51 (2011).
29. Bernet, A. *et al.* Targeted inactivation of the major positive regulatory element (HS-40) of the human alpha-globin gene locus. *Blood* **86**, 1202-11 (1995).
30. Anguita, E. *et al.* Deletion of the mouse alpha-globin regulatory element (HS -26) has an unexpectedly mild phenotype. *Blood* **100**, 3450-6 (2002).
31. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **26**, 11-24 (2012).
32. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-76 (2006).
33. Kowalczyk, M.S. *et al.* Intragenic enhancers act as alternative promoters. *Mol Cell* **45**, 447-58 (2012).
34. Magoc, T. & Salzberg, S.L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-63 (2011).
35. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
36. Raney, B.J. *et al.* Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**, 1003-5 (2014).

Figure 1

a. Overview of experimental work flow



b. Data analysis

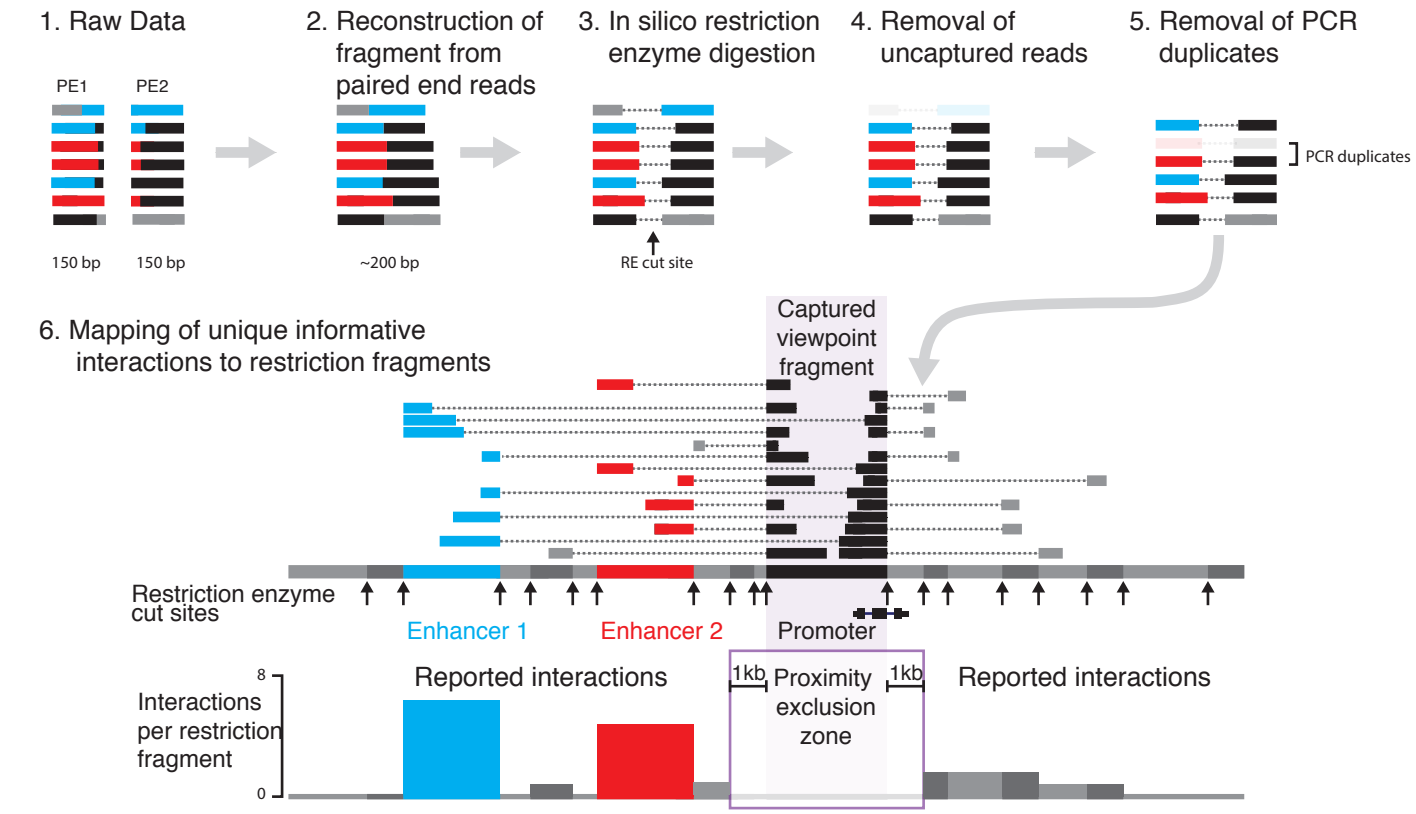


Figure 2

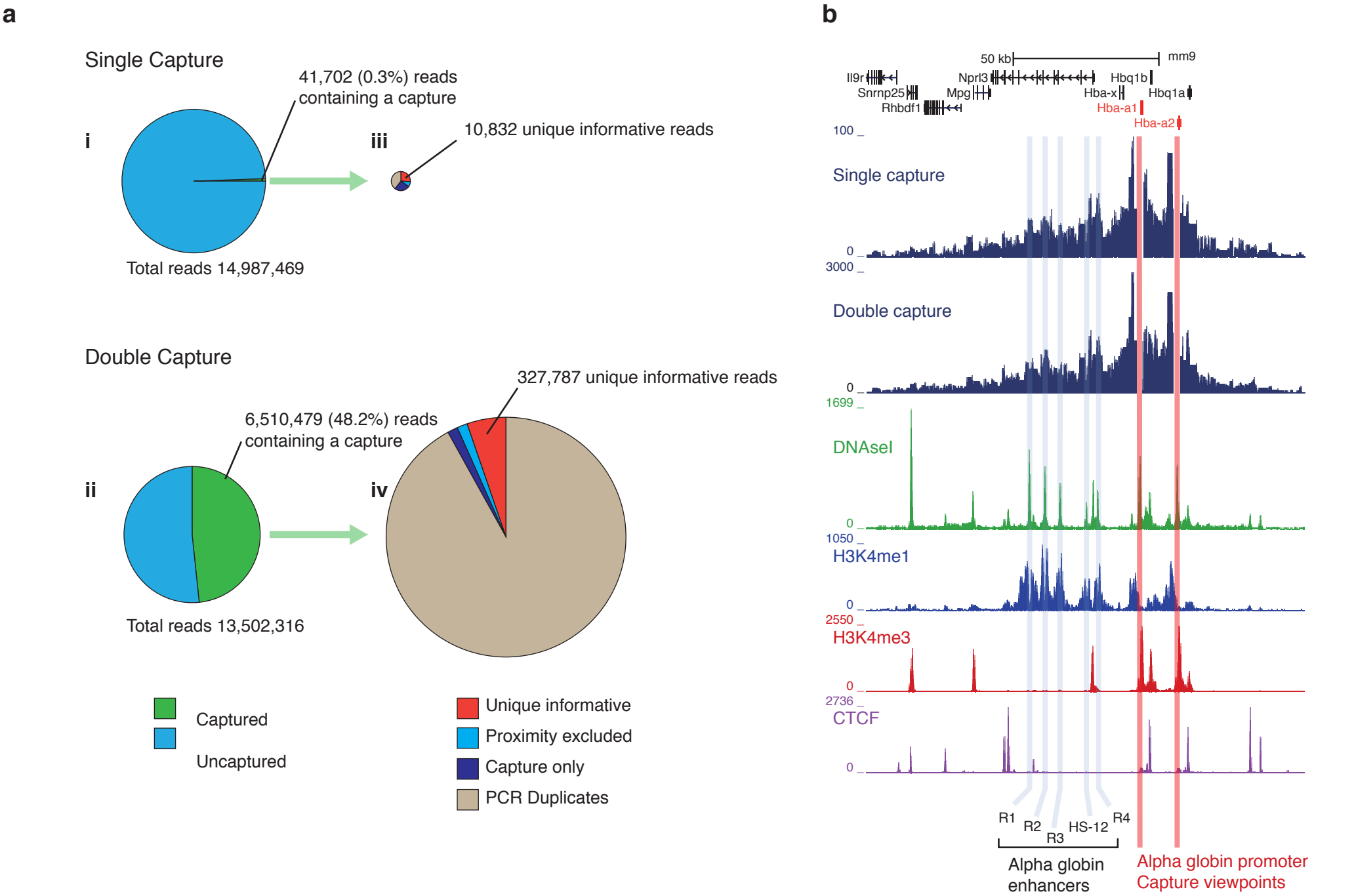

a
Single Capture
41,702 (0.3%) reads
containing a capture
i
iii
10,832 unique informative reads
Total reads 14,987,469
Double Capture
327,787 unique informative reads
6,510,479 (48.2%) reads
containing a capture
ii
iv
Total reads 13,502,316
Captured
Uncaptured
Unique informative
Proximity excluded
Capture only
PCR Duplicates
b
50 kb
mm9
Il9r
Nprl3
Hbq1b
Snrnp25
Mpg
Hba-x
Hbq1a
Rhbdf1
Hba-a1
Hba-a2
100
Single capture
0
3000
Double capture
0
1699
DNAseI
0
1050
H3K4me1
0
2550
H3K4me3
0
2736
CTCF
0
R1
R2
R3
HS-12
R4
Alpha globin
enhancers
Alpha globin promoter
Capture viewpoints

Figure 3

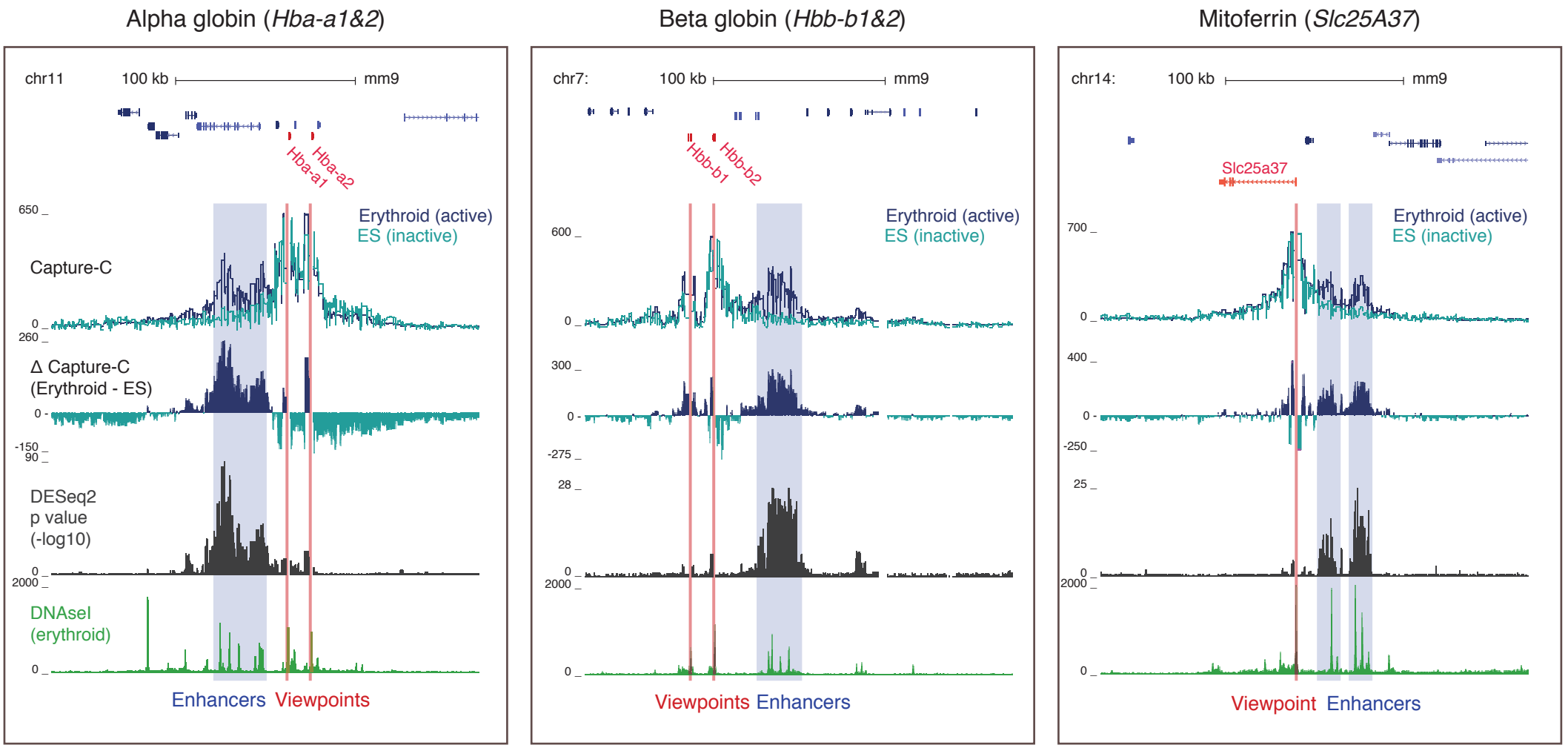
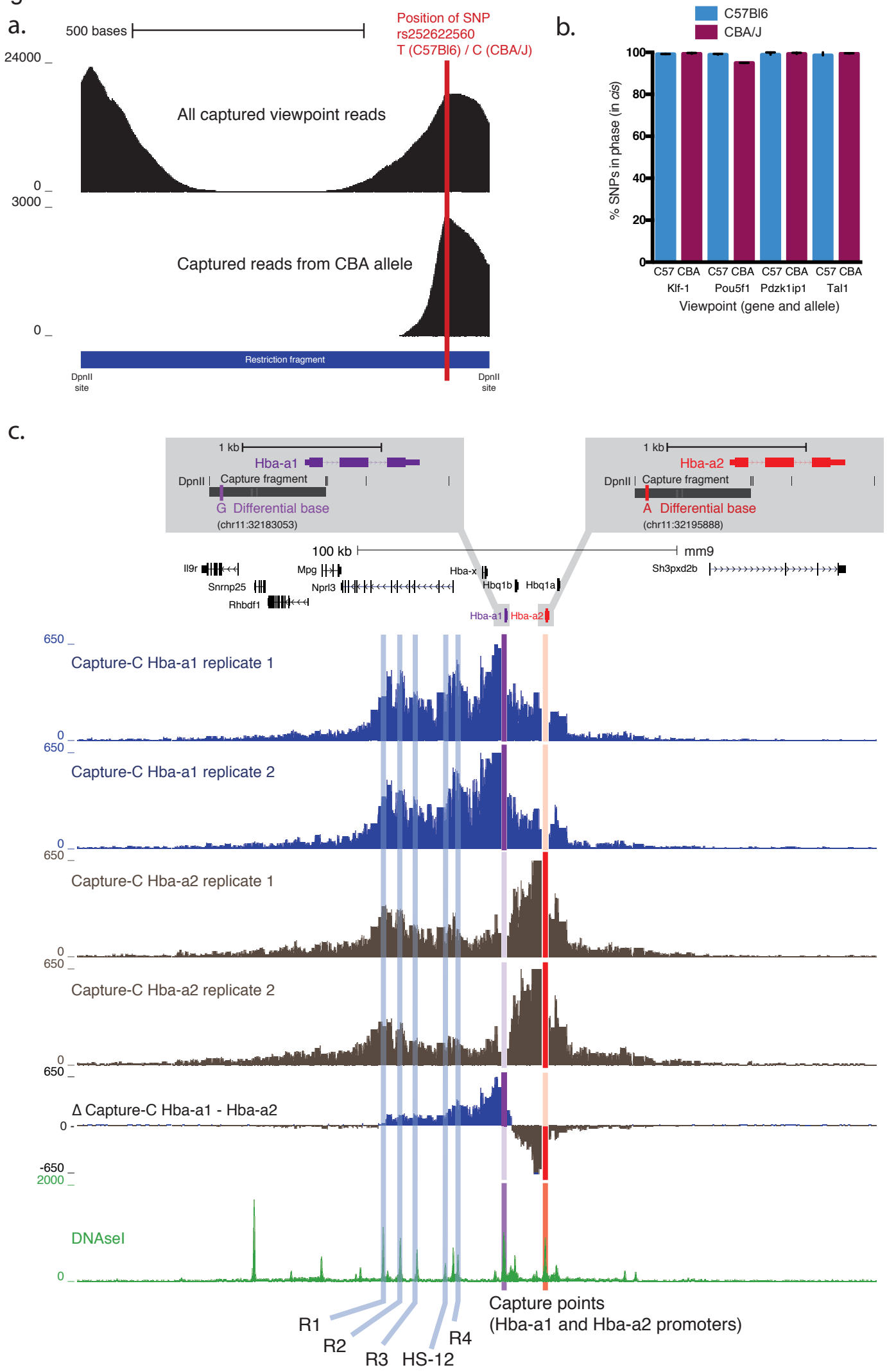
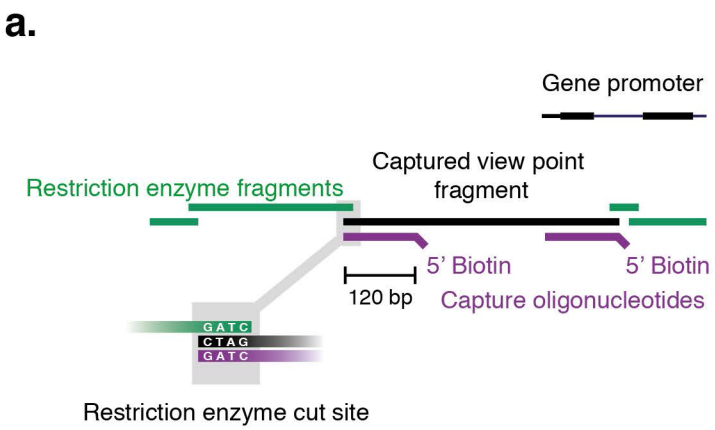
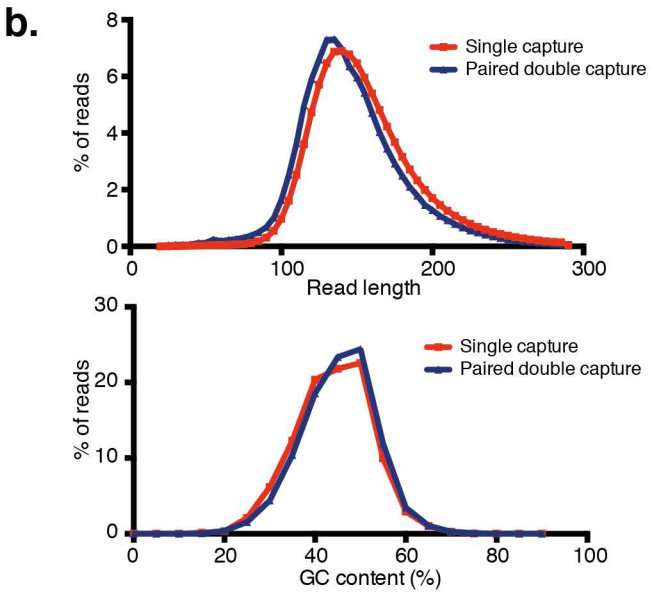

Alpha globin (*Hba-a1&2*)
chr11
100 kb
mm9
Hba-a1
Hba-a2
Erythroid (active)
ES (inactive)
Capture-C
650
0
260
Δ Capture-C
(Erythroid - ES)
0
-150
90
DESeq2
p value
(-log10)
0
2000
DNAseI
(erythroid)
0
Enhancers
Viewpoints
Beta globin (*Hbb-b1&2*)
chr7:
100 kb
mm9
Hbb-b1
Hbb-b2
Erythroid (active)
ES (inactive)
600
0
300
0
-275
28
0
2000
0
Viewpoints
Enhancers
Mitoferrin (*Slc25A37*)
chr14:
100 kb
mm9
Slc25a37
Erythroid (active)
ES (inactive)
700
0
400
0
-250
25
0
2000
0
Viewpoint
Enhancers

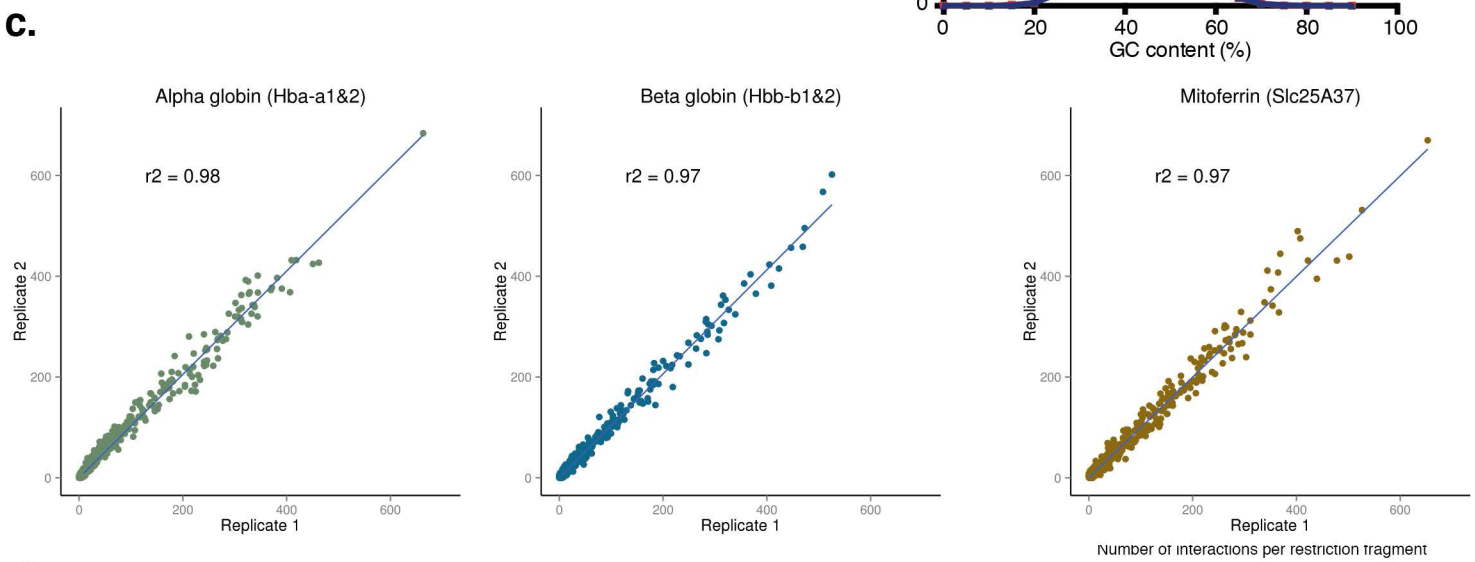
Figure 4

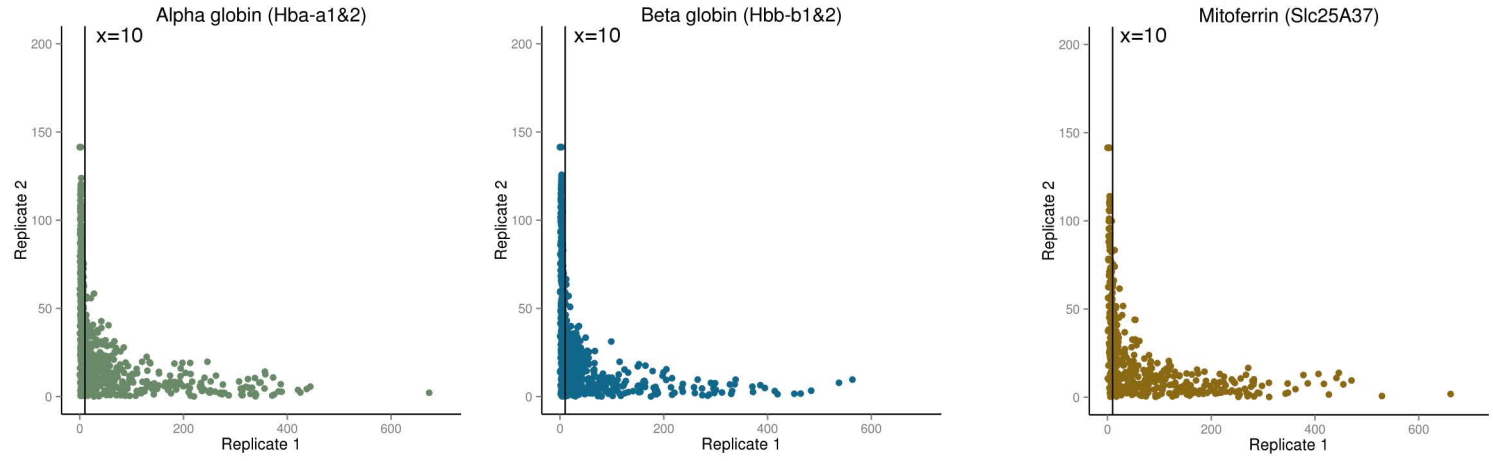

a.
500 bases
Position of SNP
rs252622560
T (C57Bl6) / C (CBA/J)
24000
All captured viewpoint reads
0
3000
Captured reads from CBA allele
0
Restriction fragment
DpnII site
DpnII site
b.
C57Bl6
CBA/J
% SNPs in phase (in cis)
100
80
60
40
20
0
C57 CBA C57 CBA C57 CBA C57 CBA
Klf-1 Pou5f1 Pdzk1ip1 Tal1
Viewpoint (gene and allele)
c.
1 kb
Hba-a1
DpnII
Capture fragment
G Differential base
(chr11:32183053)
1 kb
Hba-a2
DpnII
Capture fragment
A Differential base
(chr11:32195888)
100 kb
mm9
Il9r
Mpg
Hba-x
Sh3pxd2b
Snrnp25
Nprl3
Hbq1b
Hbq1a
Rhbdf1
Hba-a1
Hba-a2
650
Capture-C Hba-a1 replicate 1
0
650
Capture-C Hba-a1 replicate 2
0
650
Capture-C Hba-a2 replicate 1
0
650
Capture-C Hba-a2 replicate 2
0
650
Δ Capture-C Hba-a1 - Hba-a2
0
-650
2000
DNAseI
0
R1
R2
R3
HS-12
R4
Capture points
(Hba-a1 and Hba-a2 promoters)

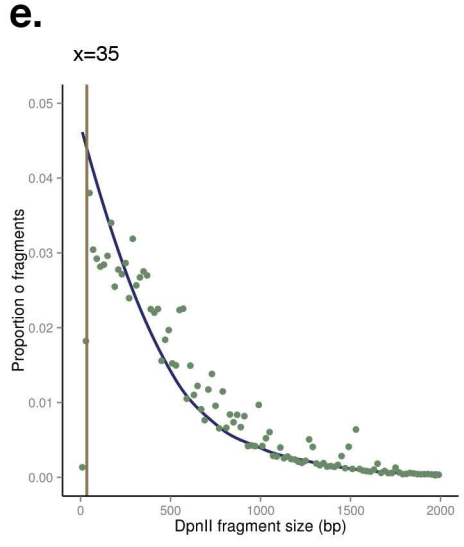
Supplementary Figure 1











Supplementary Figure 1

a. Capture oligonucleotide design. 120bp oligonucleotides are designed to the ends of the desired viewpoint restriction enzyme fragments. The promoters of genes were used exclusively in this study but any non-repetitive restriction fragment in the genome could be used (see Supplementary Table for the oligonucletide sequences used). The oligonucleotides include the restriction enzyme cut sequence (GATC in the case of DpnII) obviating the need for a biotin fill in. The oligonucleotides include a 5' biotin moiety to allow streptavidin bead capture of the target sequence and any interacting partners. The oligonucleotide sequences are screened using the online Capsequm tool for repetitive sequences and short repeats.

b. Comparison of GC content and read length between single and double capture. This shows that despite additional PCR amplification there is virtually no change in GC content in the reported reads and that the reconstructed read lengths remain nearly constant.

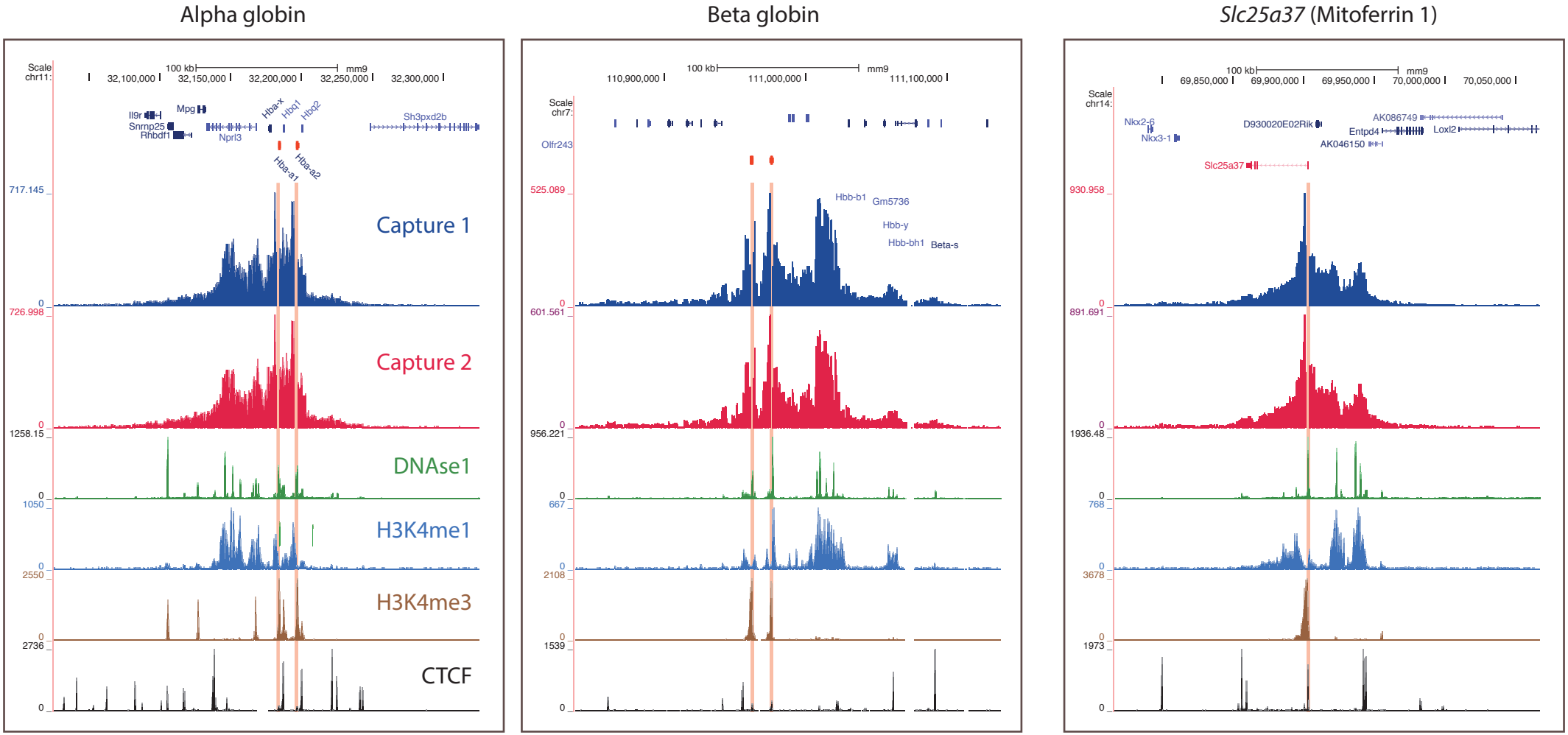
c. Comparison of two biological replicates from three different genes (*Hba-a1&2*, *Hbb1&2* and *Slc25A37* (mitoferrin)) showing very high degrees of correlation. The data has been normalised to a total interaction count of 100,000 across the whole genome.

d. Plot of coefficient of variation against normalised number of unique interactions per restriction fragment. The vertical lines are plotted at 10 interactions showing that the CV drops below 50% when more than 10 unique interactions are detected per restriction fragment. Given that the track is scaled to contain a total of 100,000 interactions this means that we are able to reliably detect interactions present at 1 allele in 10,000.

e. Distribution of unique interactions by DpnII fragment size (green points) compared to the distribution of DpnII fragments across the genome (blue line). This shows that when the DpnII fragment is above 35bp in size (vertical line) the distribution of the interactions follows that of the distribution of fragments across the genome. (Data used from the Capture-C profile for the *Hba-a1&2* promoters).

Supplementary Figure 2

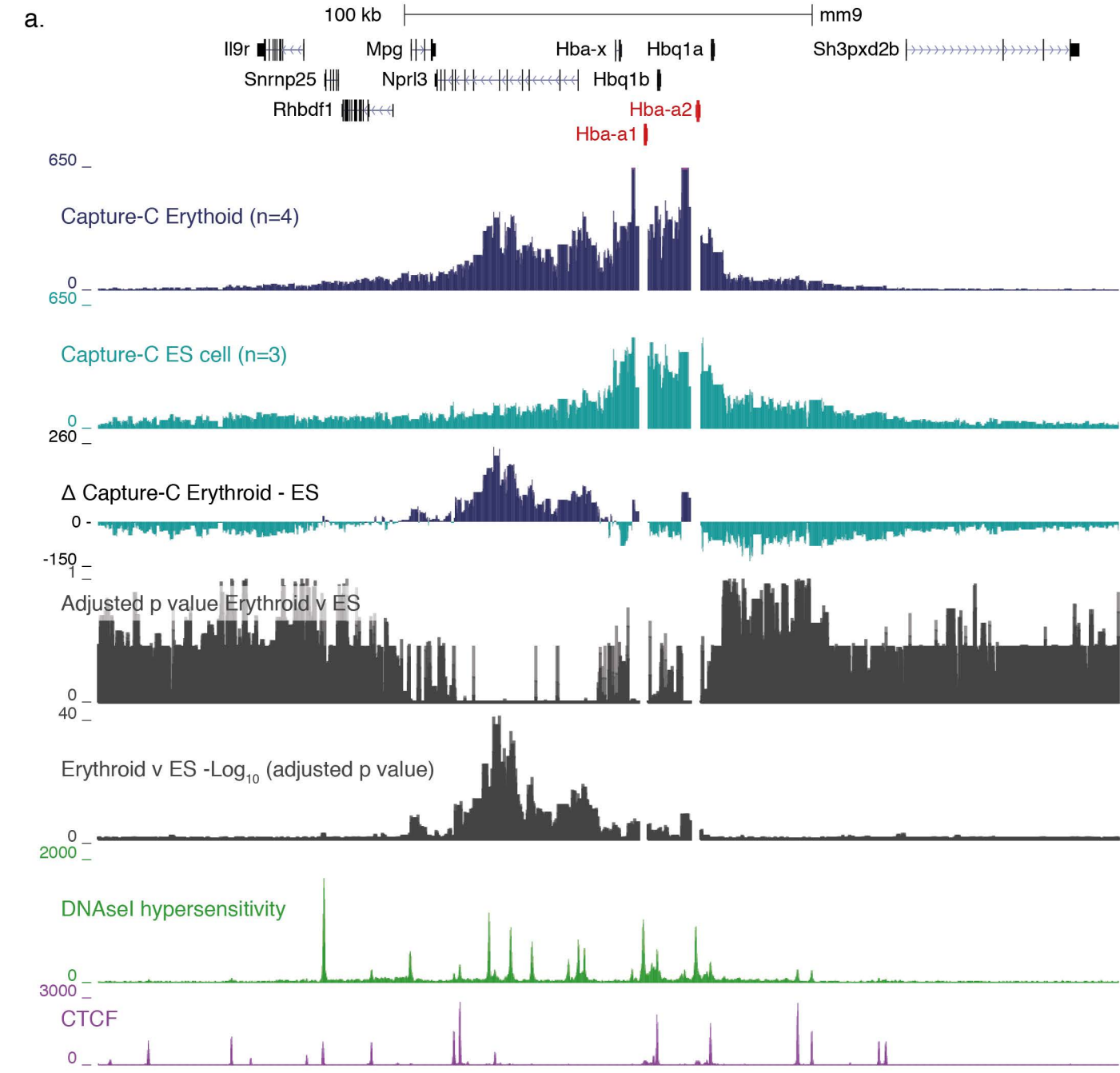
The reproducibilty of NG Capture-C interaction profiles at the resolution of indvidual Dpn II restriction fragments

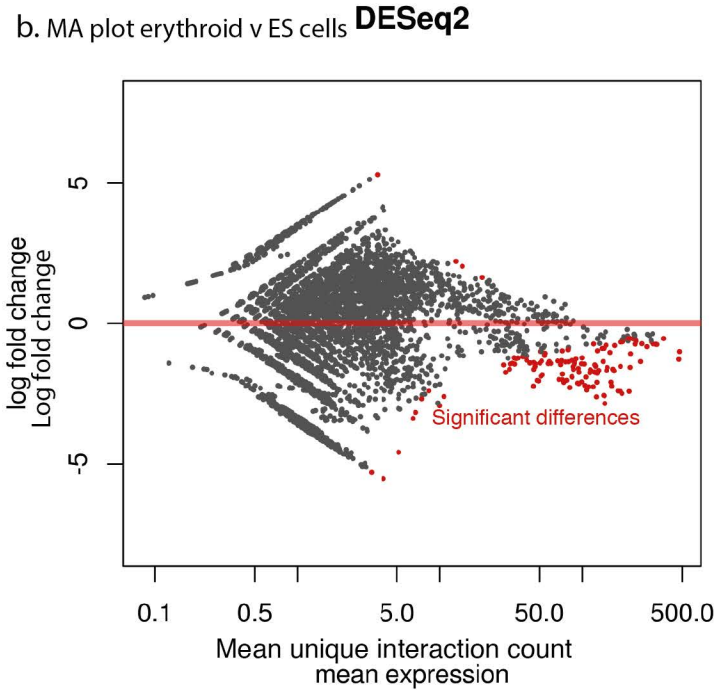


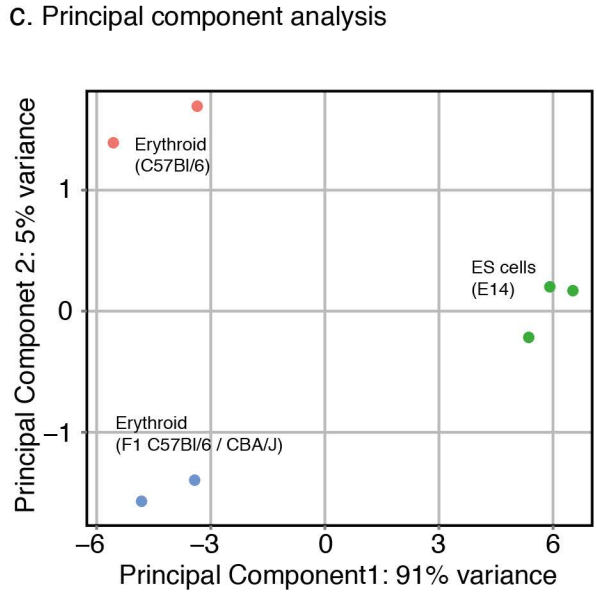
Supplementary Figure 2

Profiles generated from two biological replicate of erythroid cells are shown in blue and red in a 300 kb window. Interaction data is plotted as normaized interactions over individual restriction fragments of the restriction enzyme used to create the 3C library (Dpn II) The genes of the captured promoters are highlighted in red (UCSC gene annotation). The position of the capture fragments are highlighted by red bars across the tracks. Genomic data tracks are shown below the interaction profiles colored for data type, DNAse1 (green); H3K4me1 (light blue); H3K4me3 (brown) and the binding of the CTCF protein (black).

Supplementary figure 3







Supplementary Figure 3

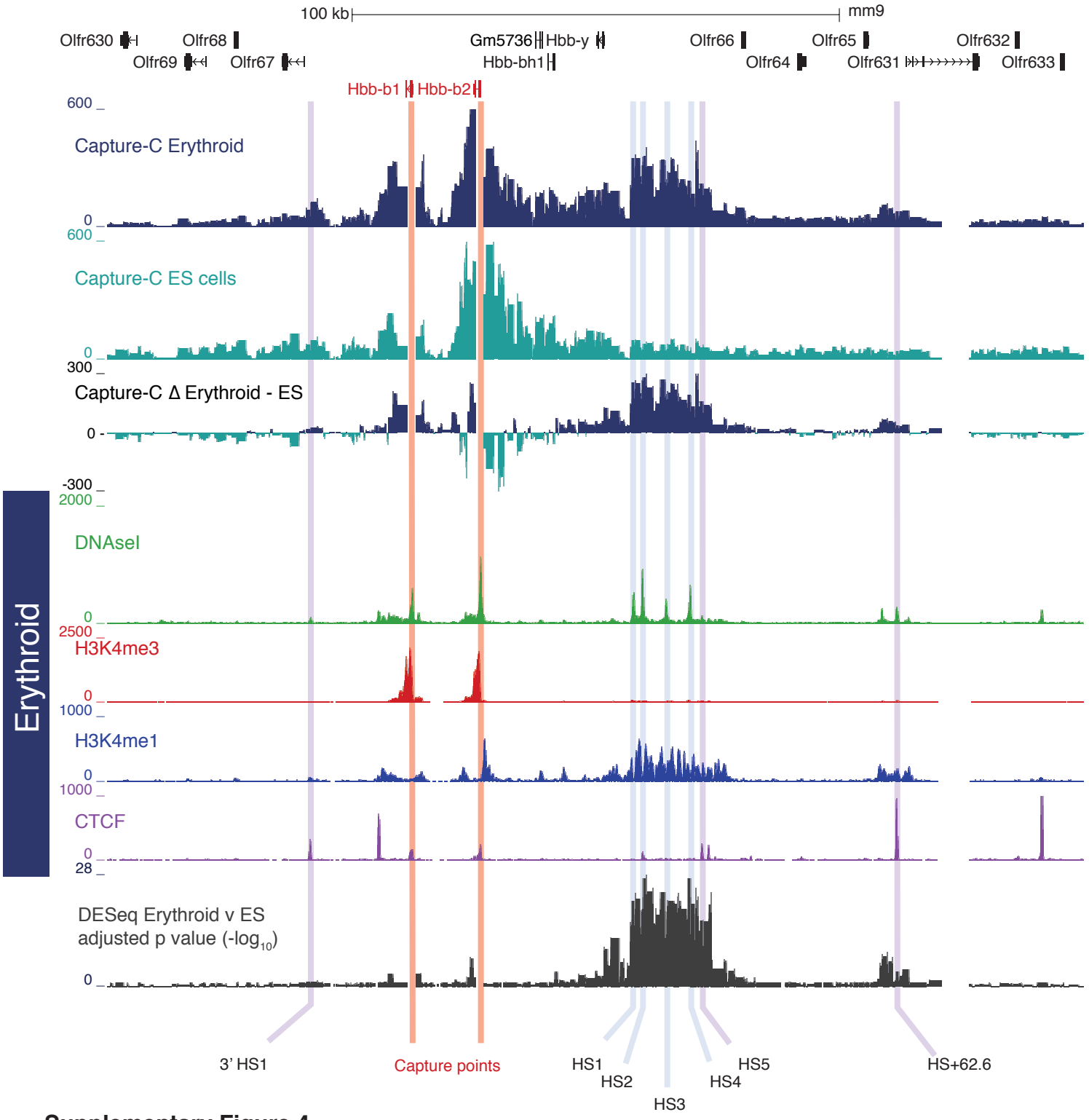
Statistical analysis of the differences between the Capture-C profile in Erythroid (n=4 (2 x C57Bl/6 and 2 x CBA/J)) and ES cells (n=2) at the alpha globin locus. Raw counts of unique interactions mapped to each restriction fragment were analysed using the bioconductor package DESeq2. This is used extensively for tha analysis of RNA-seq data and uses a model based on the negative binomial distribution.

a. Shows the adjusted p values and log transformed adjusted p values mapped across the alpha globin locus. This shows that there are very significant differences between the values for all of the restriction fragments containing regulatory elements with p values well below 10^{-30}.

b. MA plot of erythroid v ES cell data. This shows that the majority of significant differences are with elements that interact more strongly in erythroid cells.

c. Principal component analysis showing that the cell types cluster together with 91% of the variance being between erythroid and ES cells and 5% varience between the data from the different strains of mice.

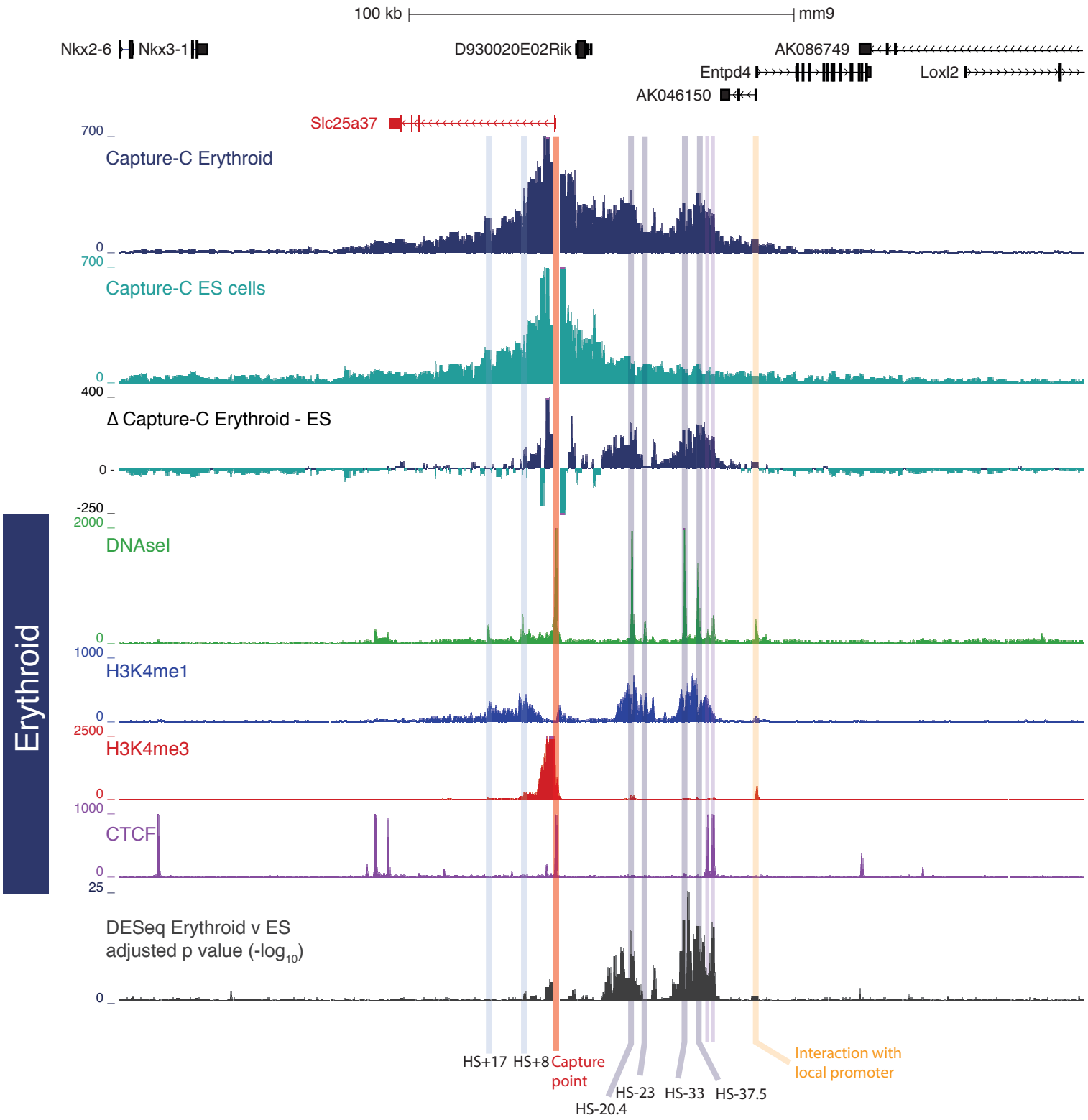
Supplementary Figure 4



Supplementary Figure 4

Capture-C data at the beta globin (*Hbb-b1&2*) locus. There are two copies of the gene and so the interaction profiles are a composite of the profiles from the two promoters. The Δ Capture-C profile clearly picks out interactions with the very well described regulatory eleements (HS1, HS2, HS3, HS4 and HS5). In addition there are clear interactions with the CTCF site at HS+62.6 and an additional erythroid specific site just upstream of this. The DESeq analysis highlights these regulatory elements precisely with p values below 10^{-20} over multiple adjacent fragments.

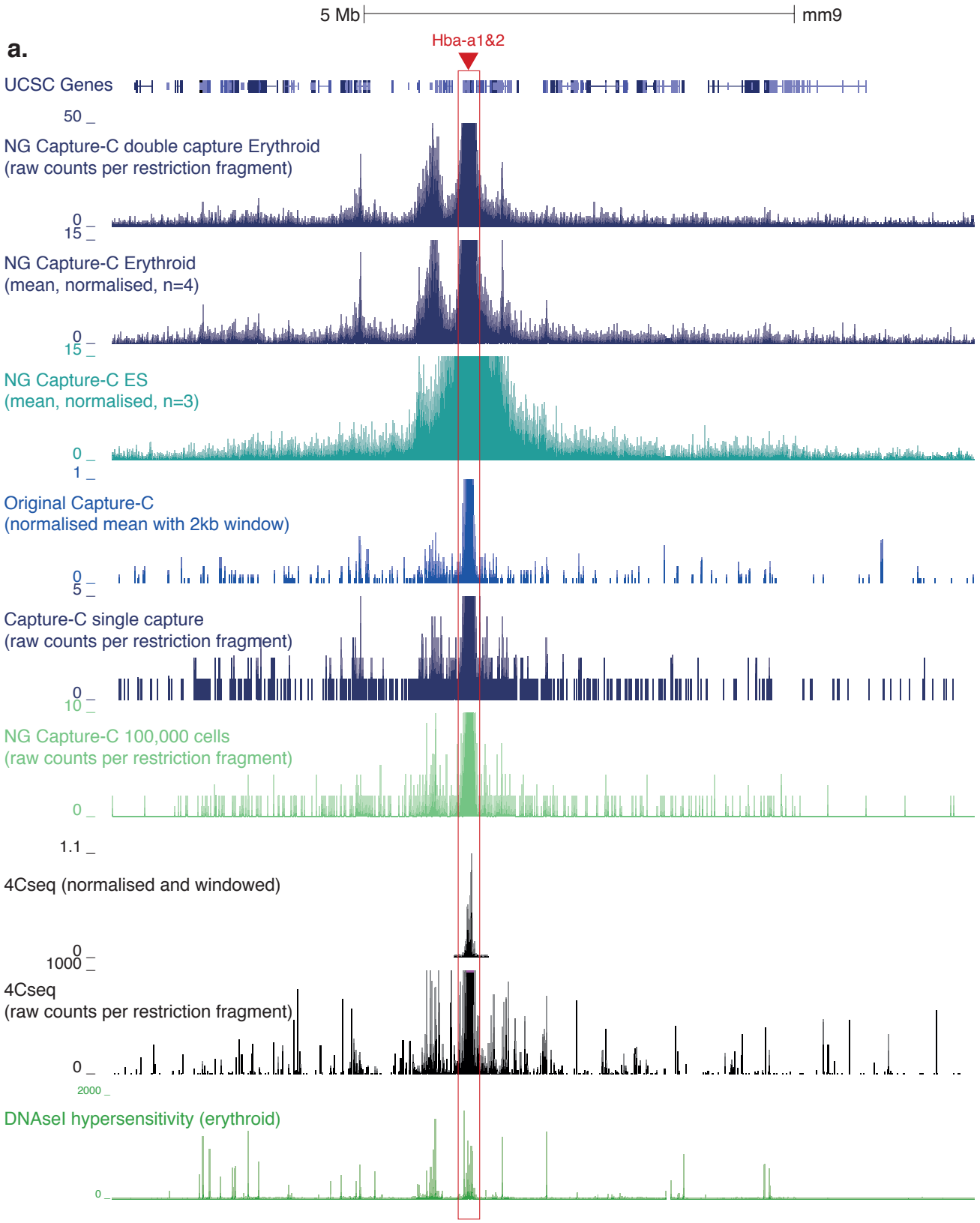
Supplementary Figure 5



Supplementary Figure 5
Capture-C from the mitoferrin (*Slc25A37*) gene promoter demonstrating interactions with the previously identified regulatory elements (Hughes et al., 2014). Note that the differential Capture-C track specifically highlights the two blocks of active regulatory elements and that the DESeq analysis only highlights these interactions as being statistically significant.

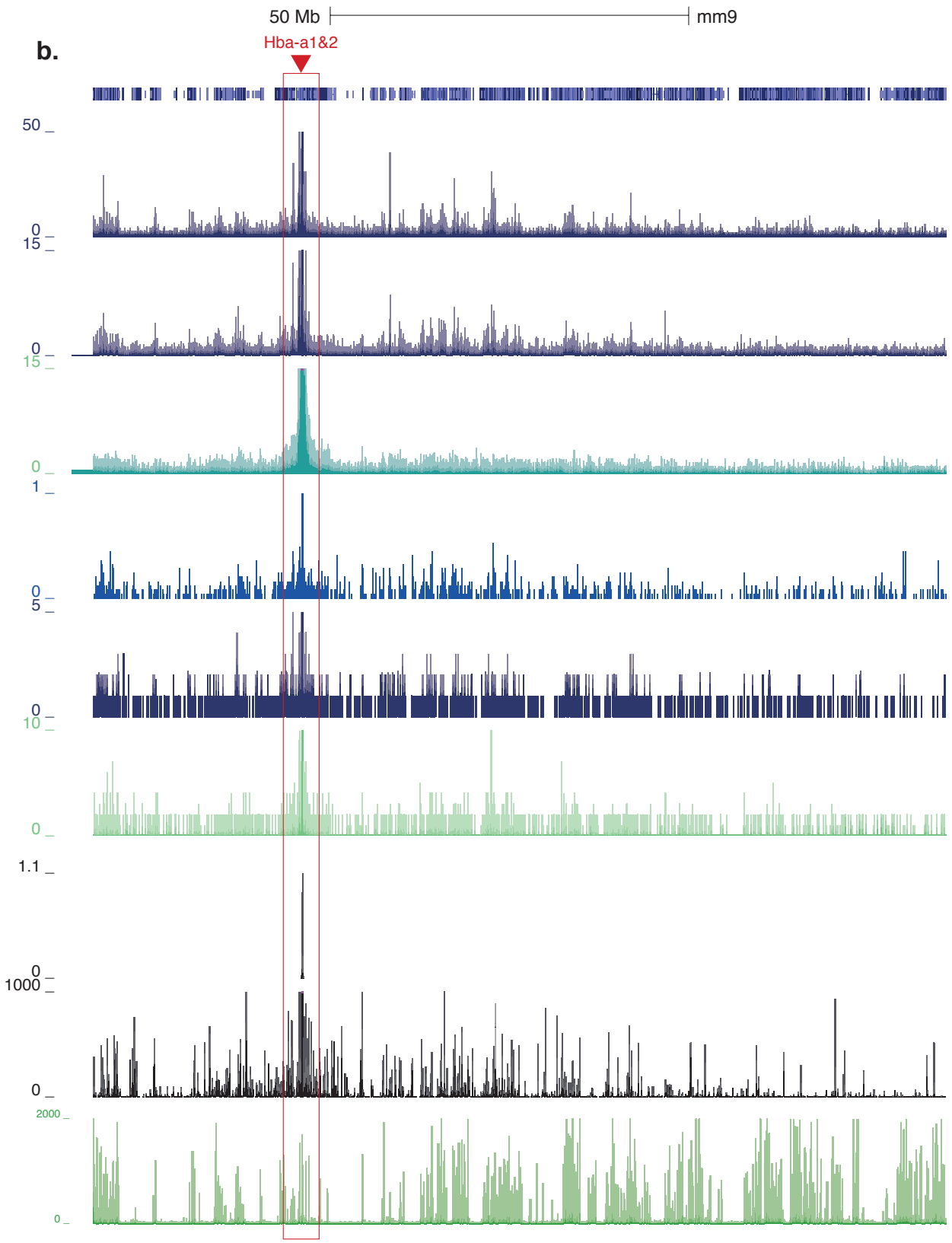
Supplementary Figure 6

10mb region around capture point . All tracks scaled to show weak interactions



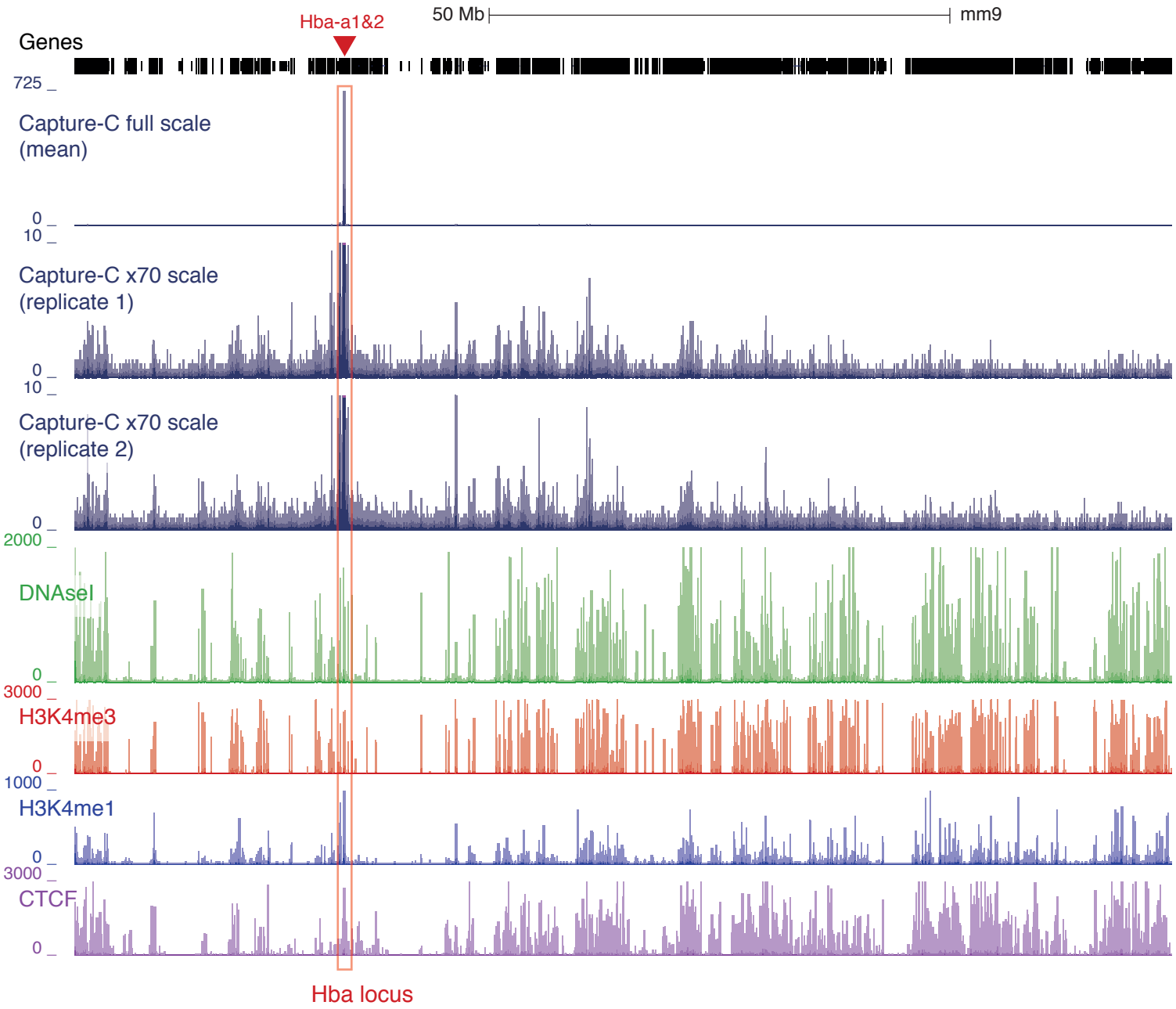
Supplementary Figure 6

Capture-C from the alpha globin promoters (Hba-a1&2) in erythroid cells panel a. shows 10Mb around the alpha globin locus and panel b. the whole of chromosome 11. The red box in panel a. denotes the region shown in Supplementary Fig 2a. and the red box in panel b. shows the region shown in panel a. The top track shows the raw count per restriction fragment from erythroid cells. The second and third tracks shows the normalised counts from 4 erythroid replicates co-captured with 3 replicates in ES cells respectively. These data are corrected for the total number of interactions across the genome to a total of 100,000 so the Y axis of these tracks shows the % x1000 of the interactions with each restriction fragment compared to the number of interactions



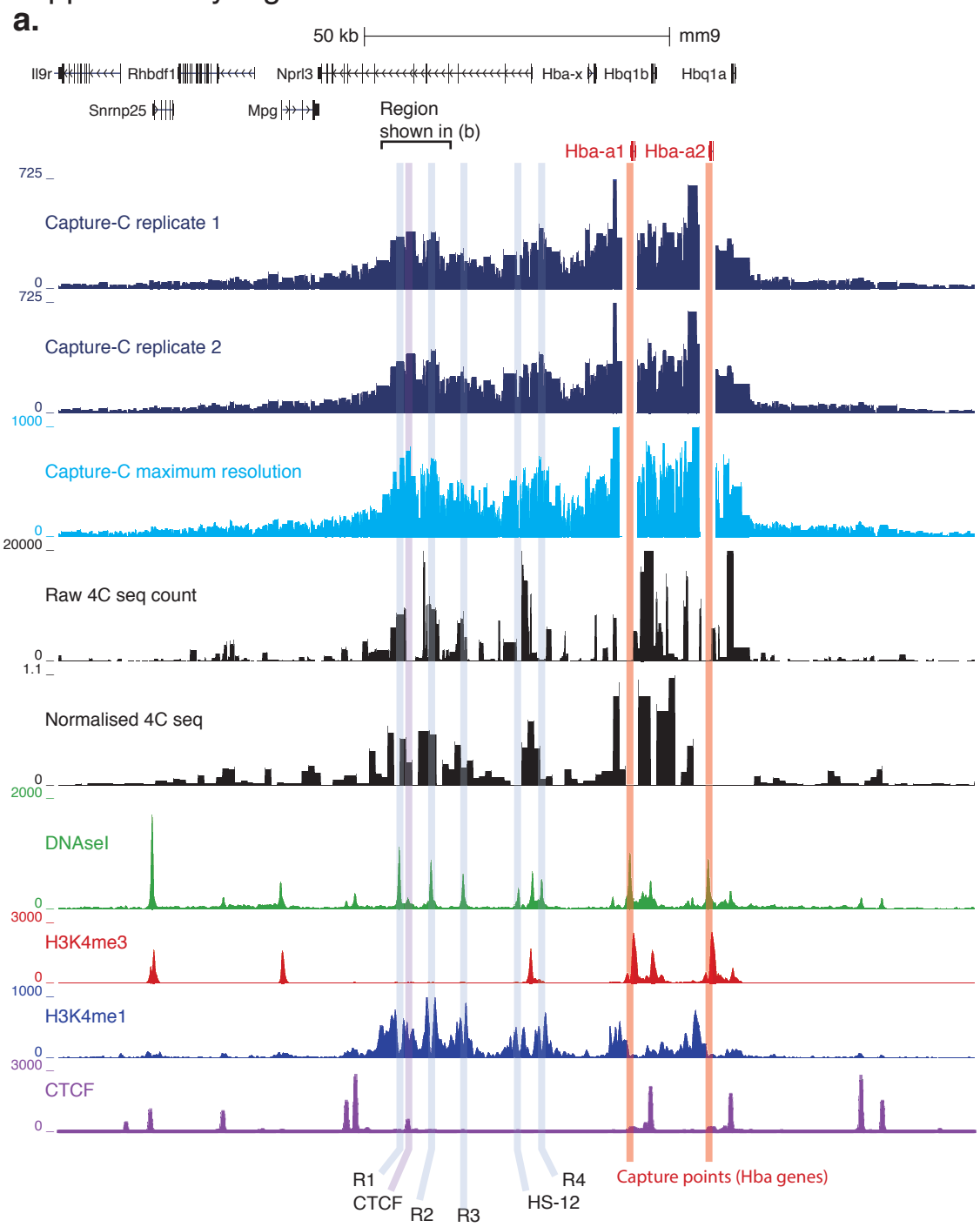
across the whole genome. The next three tracks show data from the original Capture-C experiment (normalised with 2kb window); a single capture experiment from just the alpha globin promoter (raw counts per restriction fragment) and the effects of reducing cell numbers (raw counts). The next two track show data from 4C seq (van de Werken et al., 2012) the normalised data does not extend far outside the locus but the raw counts per restriction fragment can be determined across the whole chromosome. 4C seq also shows very long range interactions but it is difficult to quantify these reliably because it is impossible to differentiate between genuine ligation junctions and PCR duplicates. NG Capture-C clearly has a much higher sensitivity than the original method and single capture.

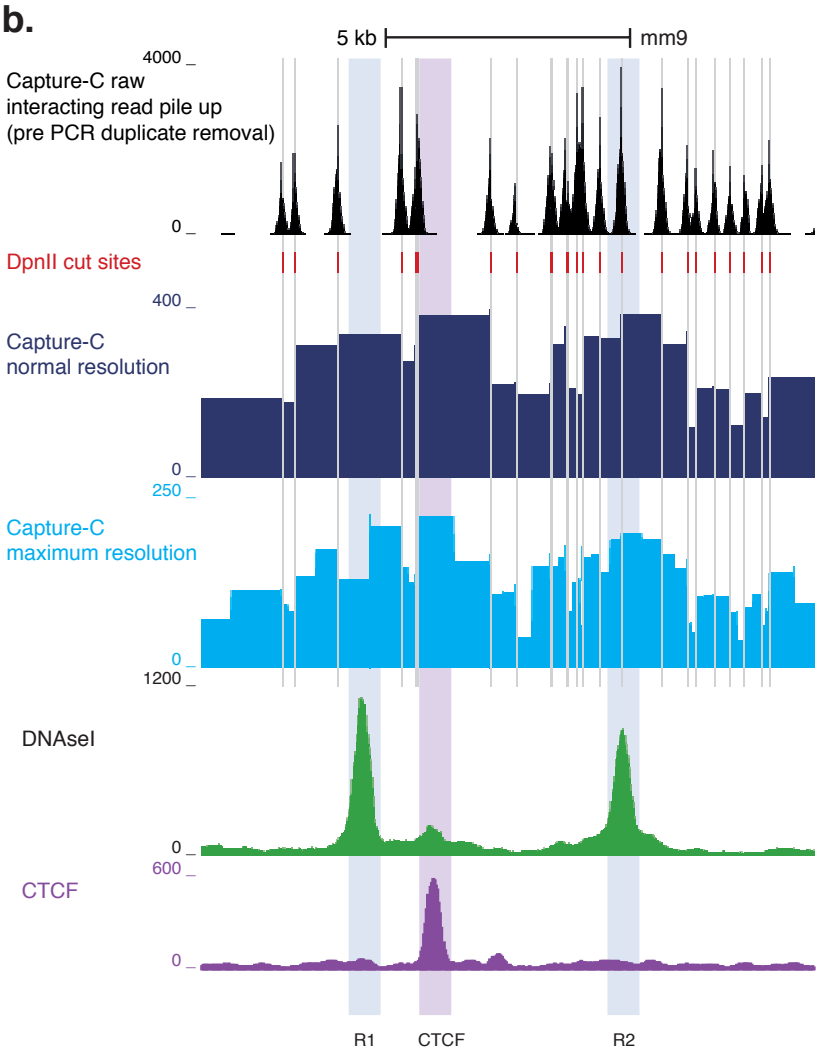
Supplementary Figure 7



Supplementary Figure 7
Capture-C from the alpha globin promoters (*Hba-a1&2*) in erythroid cells showing the whole of chromosome 11. The top track shows the data with the same scale on the Y axis as used for Supplementary Fig. 2a. On the lower two tracks the scale has been changed to show that there are reproducible interactions in cis over several megabases, however, these data points are still reported per Dpn II fragment. These correlate with gene rich regions, active enhancers and promoters as well as CTCF sites.

Supplementary Figure 8

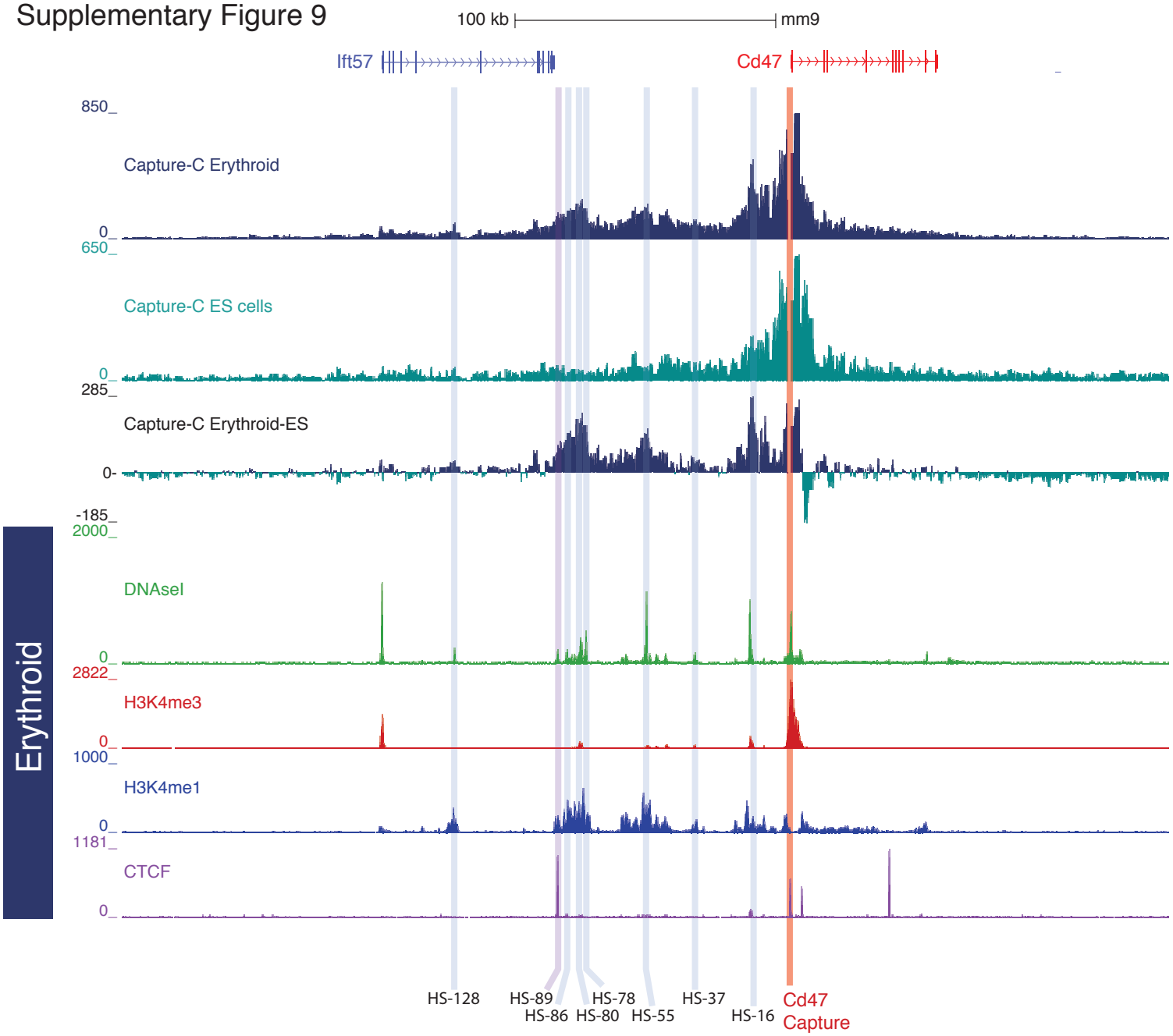




Supplementary Figure 8

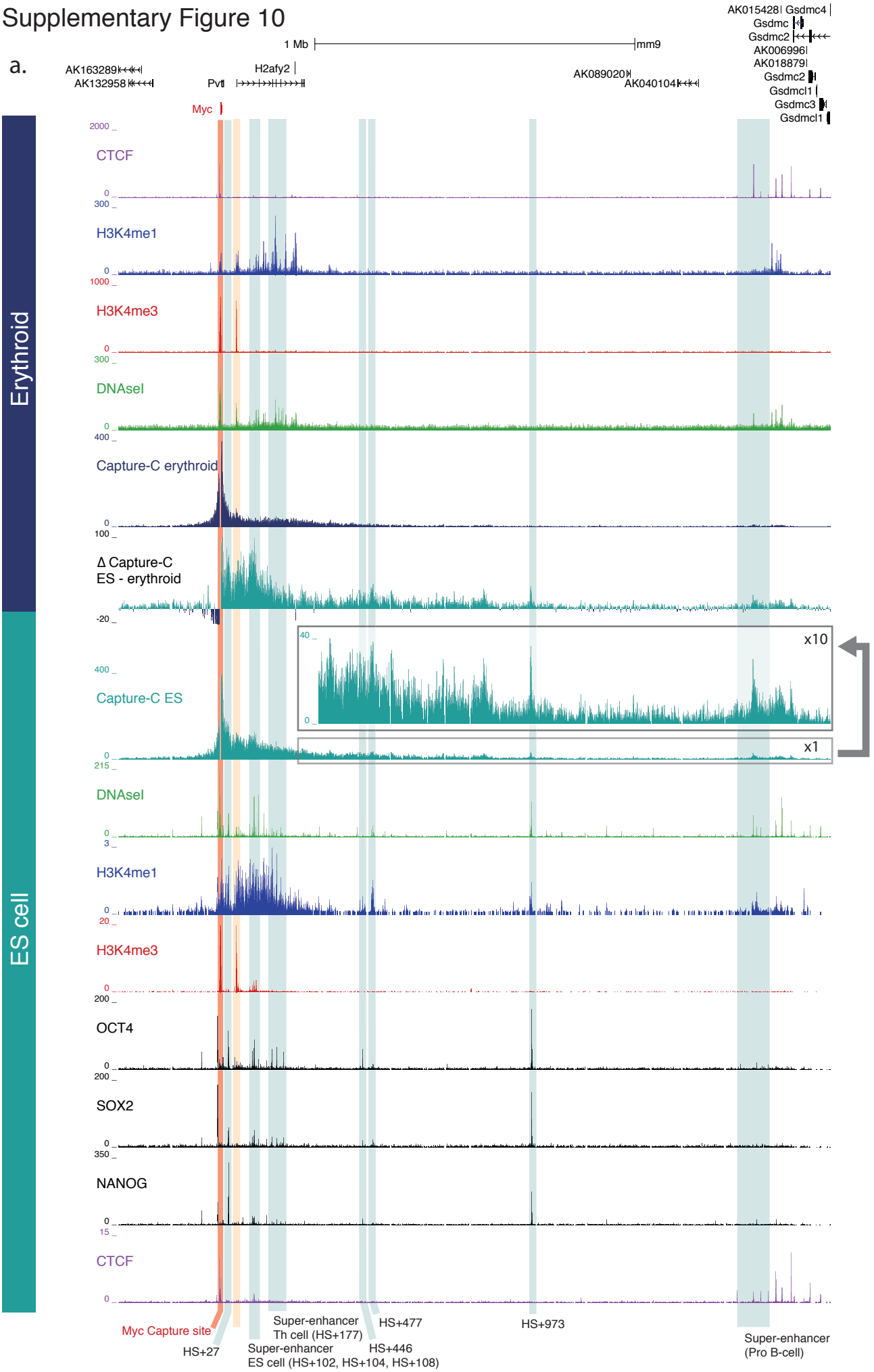
a. Comparison of two biological replicates from the alpha globin promoters (*Hba-a1&2* highlighted in red) with 4C-seq (van de Werken et al., 2012).
The top two tracks are from biological replicates showing the high degree reproducibility between samples. The third track shows the maximum possible resolution for C experiments with interactions mapped to the half of the restriction fragment which contains the midpoint of the interacting read (see (b) for further details).
The raw 4C-seq profile was generated using the Capture-C analysis tools (normal resolution). The PCR duplicate filtering we use for all Capture-C data cannot be performed because the amplicons from any particular location are virtually identical. If this duplicate filtering is applied to the 4C-seq data this results in nearly complete loss of the interaction profile. Below it is the published normalised 4C-seq profile showing the degree of correction necessary to compensate for differences in PCR efficiency.
b. Comparison of Capture-C data mapped to normal and maximum resolution. This shows a close up on the two most important regulatory elements in the alpha globin locus (R1 and R2) (also depicted in (a)). The top profile shows the raw pile up of interacting reads prior to PCR duplicate removal. Note that the interacting reads are almost entirely juxtaposed to the DpnII cut sites (the sharp peaks correspond to the DpnII cut sequence, where the reads on both sides of the junction overlap). Since the reads are so close to the ligation junction and for larger fragments virtually no interacting reads span the middle of the read they can be mapped to the whole restriction fragment (normal resolution) or two the half of the fragment, which contains the midpoint of the read. This potentially provides the maximum possible resolution for 3C experiments.

Supplementary Figure 9

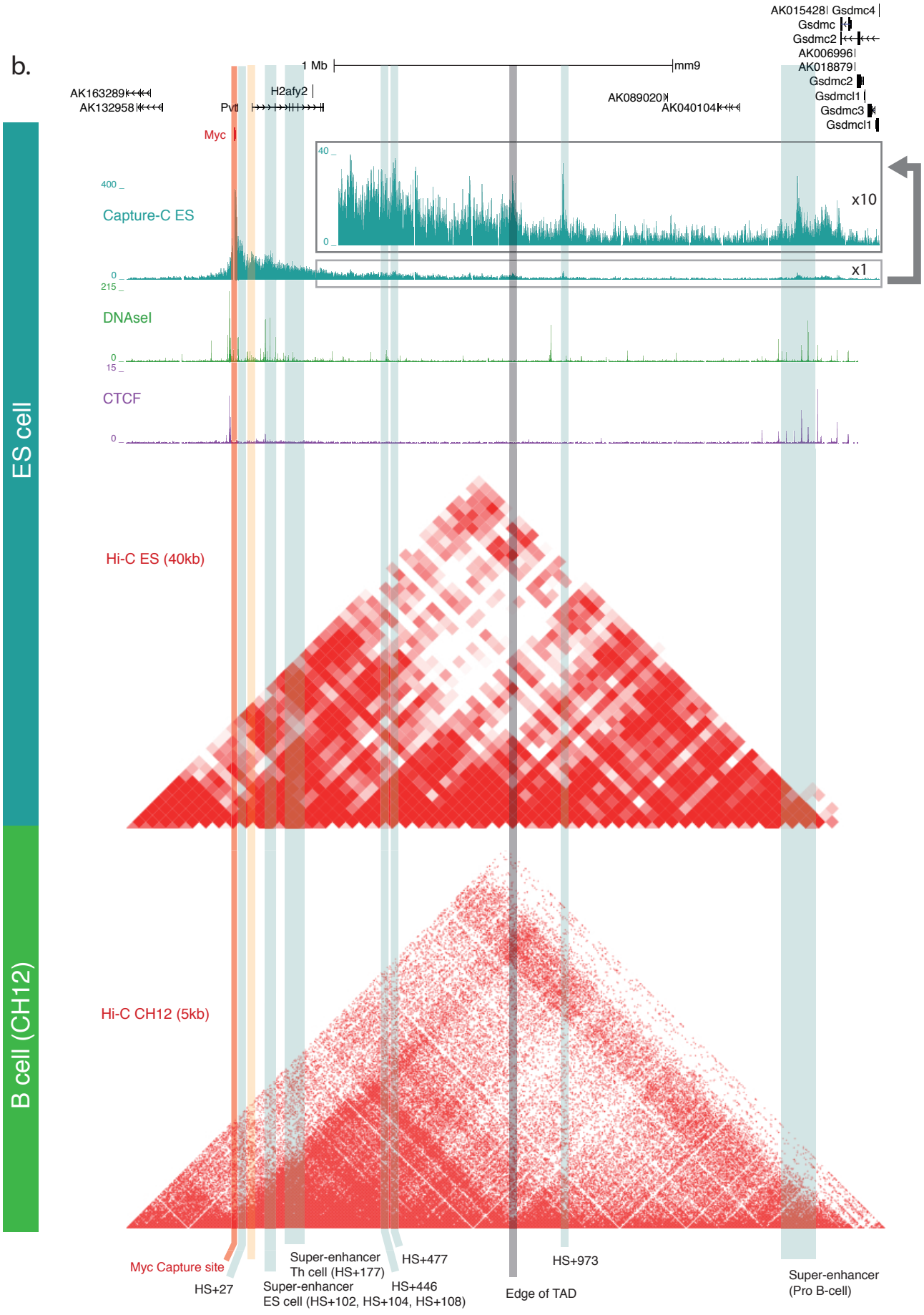


Supplementary Figure 9
Capture-C data from the *Cd47* promoter showing erythroid specific interactions with 8 elements over 128kb. Interestingly comparison of the off and on states shows that there is little change in the interaction profile over the gene itself.

Supplementary Figure 10



Supplementary Figure 10
Capture-C at the *Myc* locus.
a. Shows the comparison between erythroid and ES capture C data from the promoter of the *Myc* gene. The inset box shows the ES cell specific data with a x10 fold change in scale on the Y axis. This complex locus has three different annotated superenhancers (Whyte *et al.*, 2013). Interestingly not only does the promoter interact with the ES cell specific super

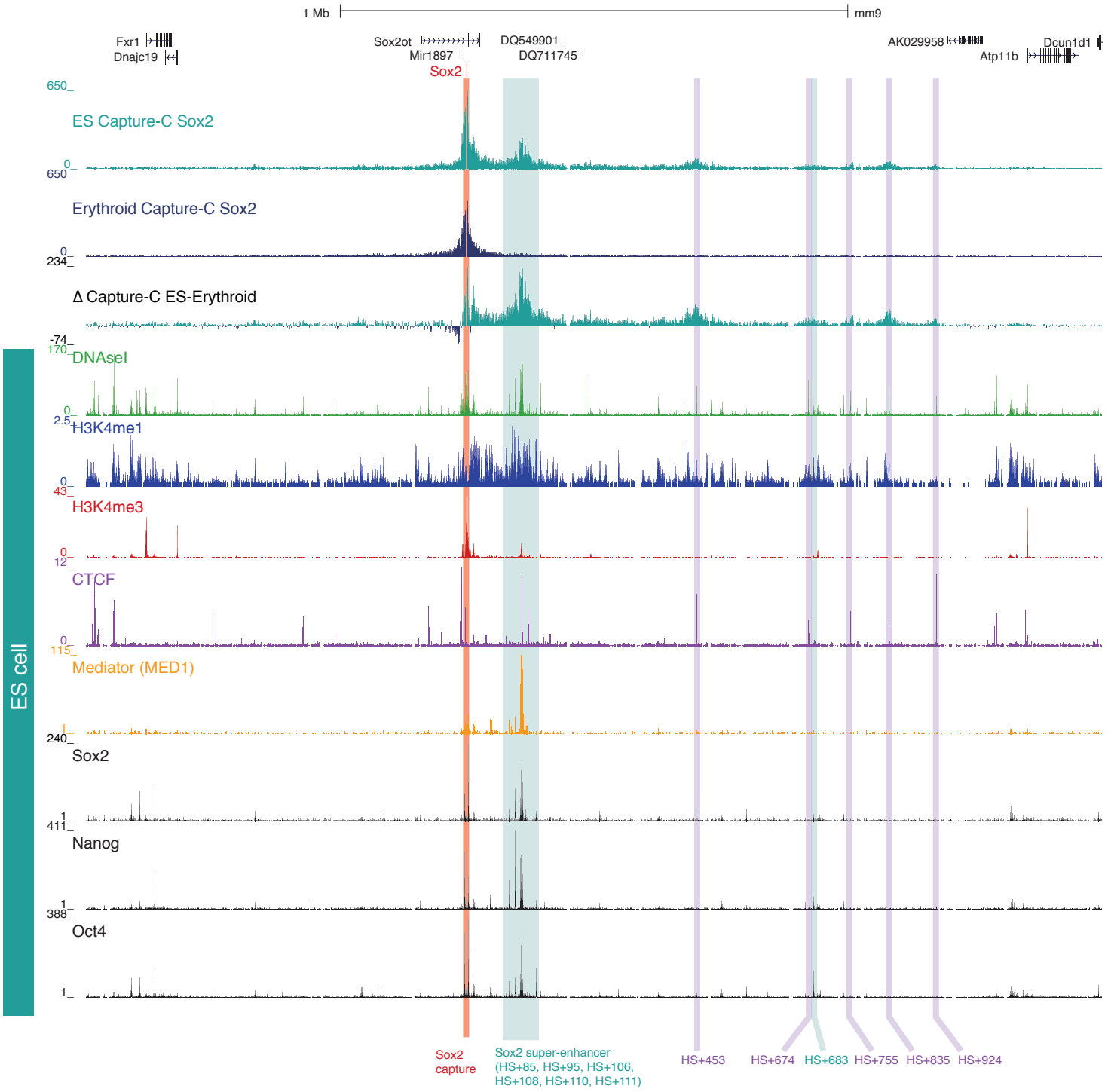


-enhancer but it also interacts with a super-enhancer called in Th cells as well as one described in Pro-B cells, which is well over 1Mb away from the gene. In addition there is a very well localised interaction with an ES cell specific enhancer at HS+973.

Panel b. shows a comparison with between NG Capture-C and two Hi-C data sets from ES cells at 40kb resolution and CH12 cells (a B cell derived line) at 5kb resolution, which is one of the highest resolution Hi-C datasets available at present. These data show that the TAD containing the promoter does not contain all of the regulatory elements. Our data and the high resolution Hi-C data sets show clear interactions extending outside of the TAD with more distal regulatory elements.

Erythroid data: Hughes et al., 2014. ES data: DNAseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259), Sox2 (Young lab GSM1082341), Nanog (Young lab GSM1082342), Oct4 (Young lab GSM1082340). Hi-C data obtained from Yue Lab (http://www.3dgenome.org) ES cells (Dixon et al., 2012), Hi-C CH12 cells (Rao et al., 2014)

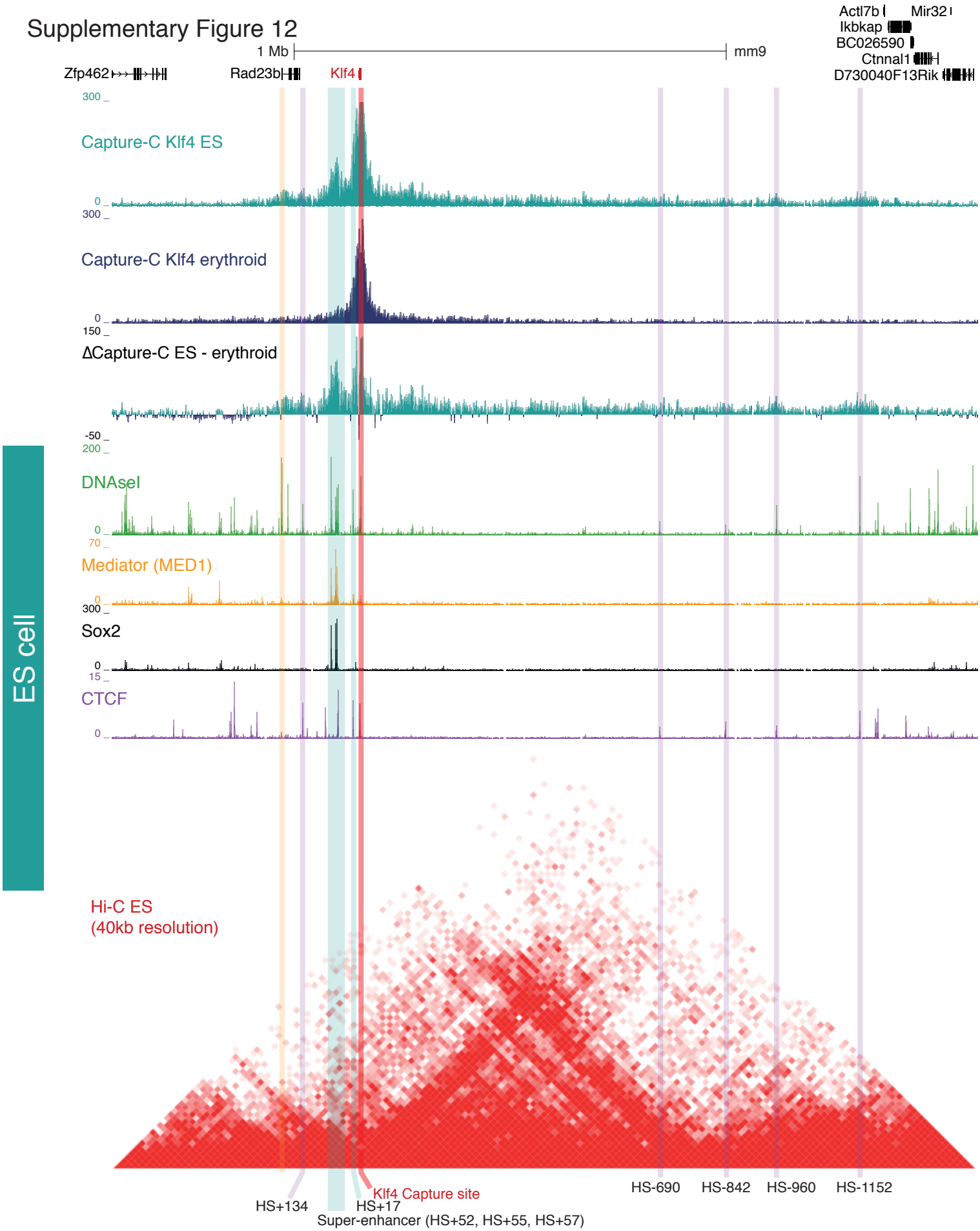
Supplementary Figure 11



Supplementary Figure 11
Capture-C from the Sox2 promoter in ES and erythroid cells. This shows that in addition to interactions with the ES cell super-enhancer identified by Whyte *et al.*, 2013 the gene has several other cell type specific interactions extending nearly 1Mb away from the promoter. These tend to be with CTCF sites (highlighted in purple) although there is an additional potential regulatory element bound by Nanog and Oct4 (HS+683).

ES cell data: DNAseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259), Sox2 (Young lab GSM1082341), Nanog (Young lab GSM1082342), Oct4 (Young lab GSM1082340)

Supplementary Figure 12

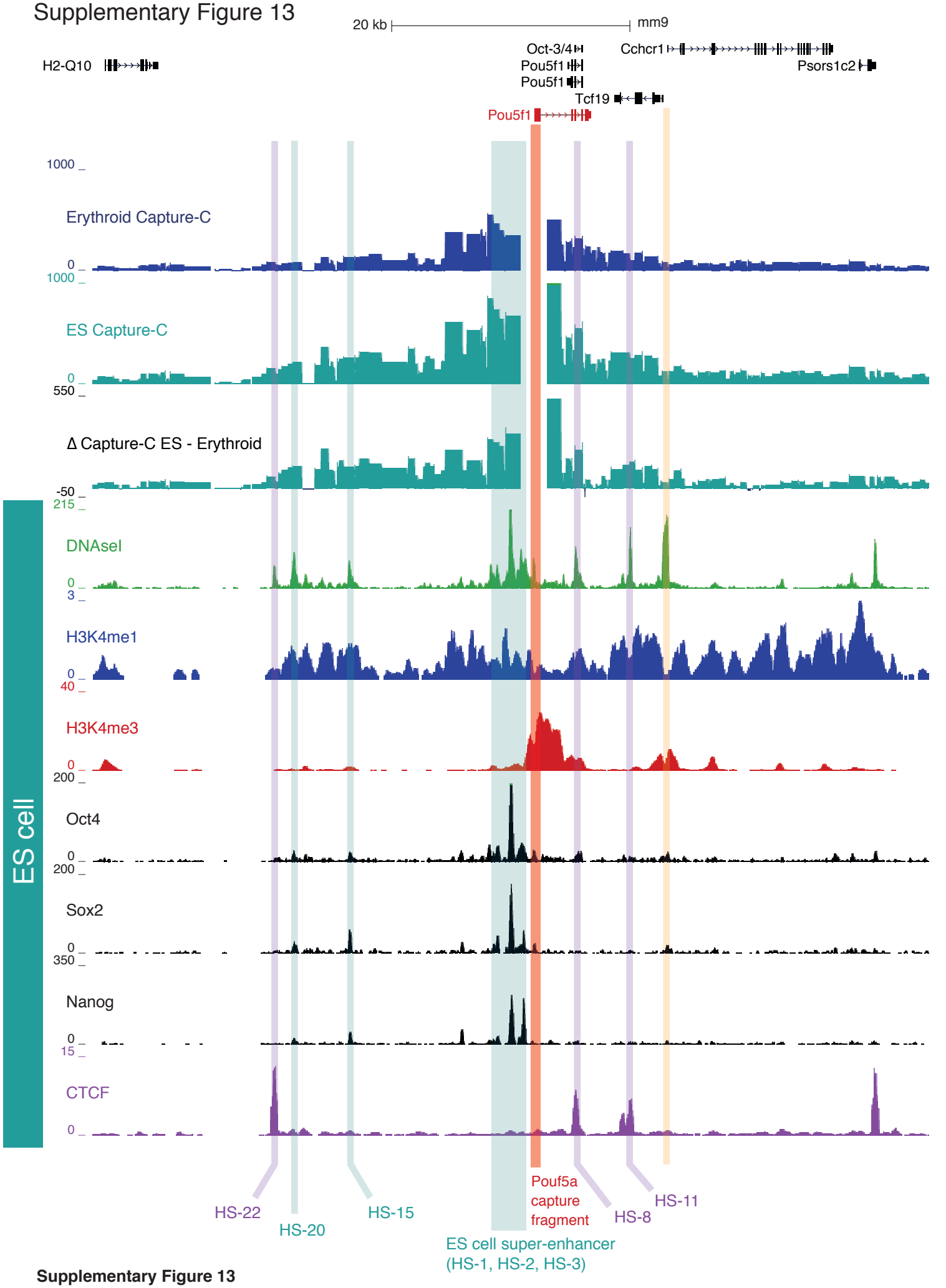


Supplementary Figure 12

Capture-C at the *Klf4* locus in ES and erythroid cells. There is clear interaction with the previously identified ES cell specific group of enhancers between HS+52 and HS+57. In addition there are interactions extending over 1Mb on the other side of the gene with sites bound by CTCF. Interestingly the Hi-C data confirms the domain of interactioin between the *Klf4* promoter and the CTCF sites in the middle of the gene desert.

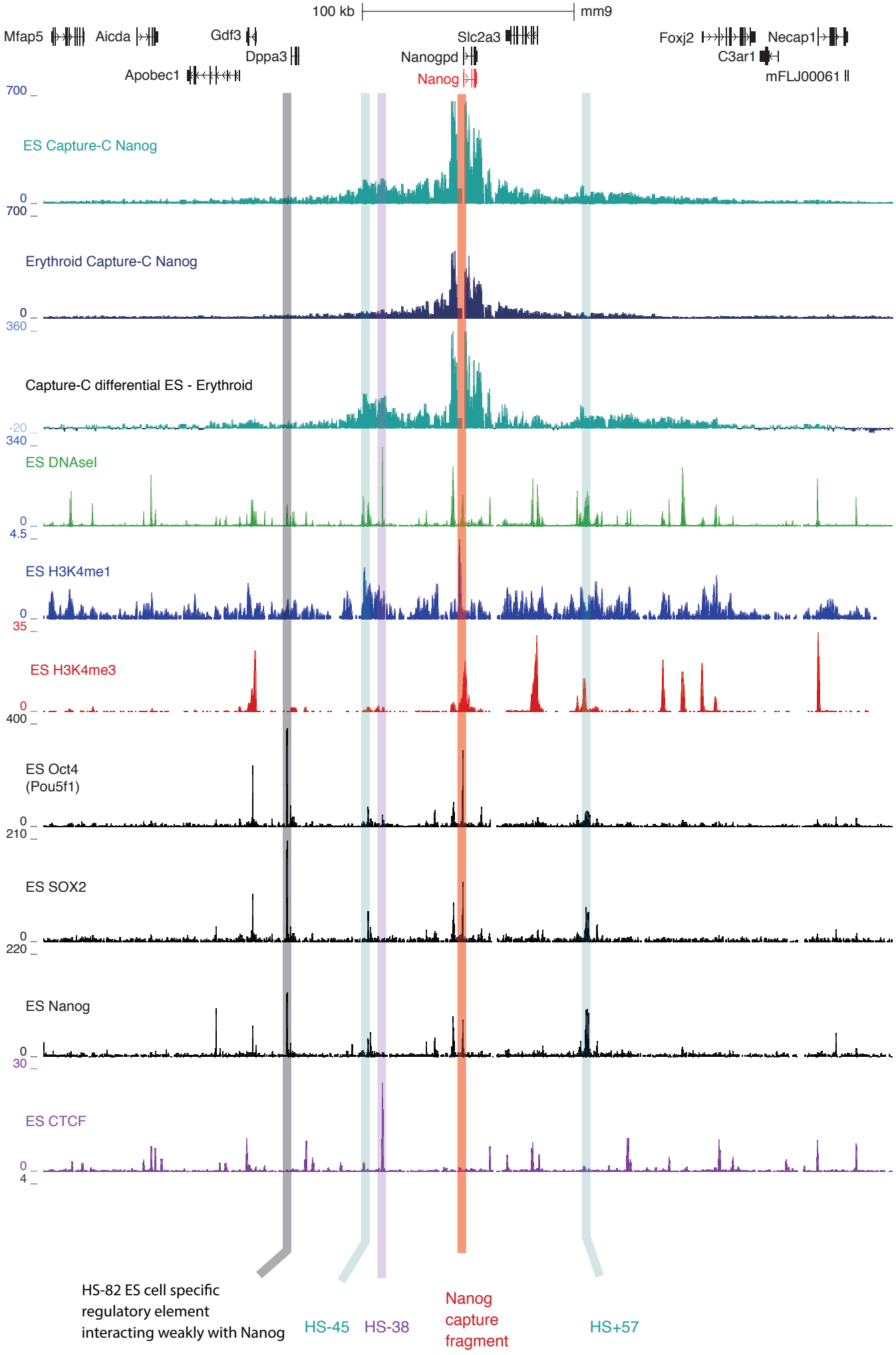
ES cell data: DNAseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259), Sox2 (Young lab GSM1082341), Nanog (Young lab GSM1082342), Oct4 (Young lab GSM1082340). Hi-C data Dixon et al., 2012.

Supplementary Figure 13



Supplementary Figure 13
Capture-C from the the *Oct4* (*Pou5f1*) promoter. There is clear interaction with the regulatory elements adjacent to the promoter, which meet criteria for a super-enhancer in ES cells. In addition the domain of interaction extends a further 22kb upstream to a CTCF binding site at HS-22. There are interactions with two further potential regulatory elements in this domain at HS-15 and HS-20 that are DNAseI hypersensitive and bound by ES cell specific transcription factors.

Supplementary Figure 14

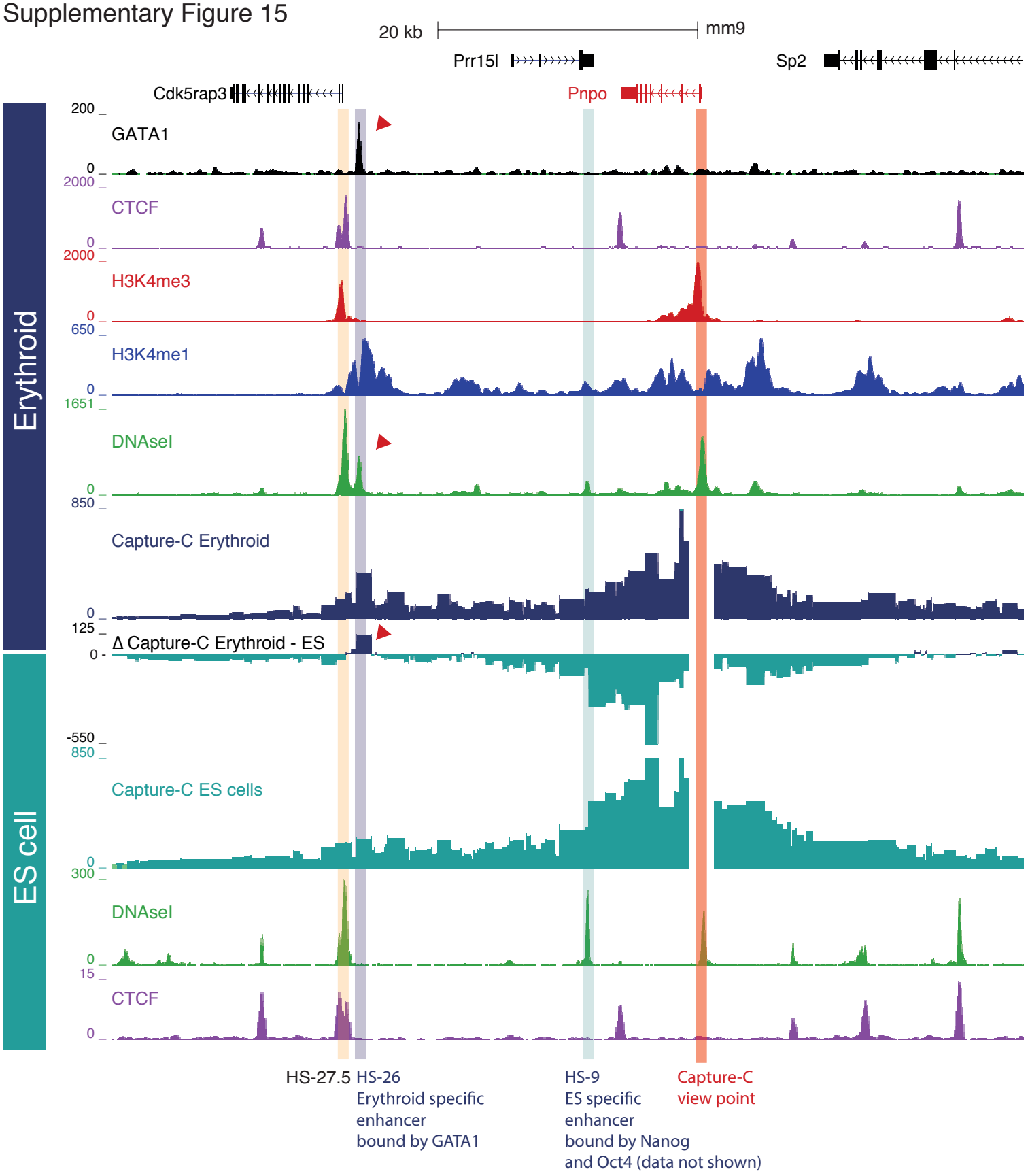


Supplementary Figure 14

Capture-C from the promoter of *Nanog* showing interactions with a regulatory element at HS-45. There is an additional regulatory element (HS-82) that is bound strongly by Oct4, Nanog and Sox2 in ES cells, however, this only interacts weakly with the promoter compared to the other elements and we would predict that it is less likely to be important in regulating the gene than the other elements.

ES cell data: DNAseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259), Sox2 (Young lab GSM1082341), Nanog (Young lab

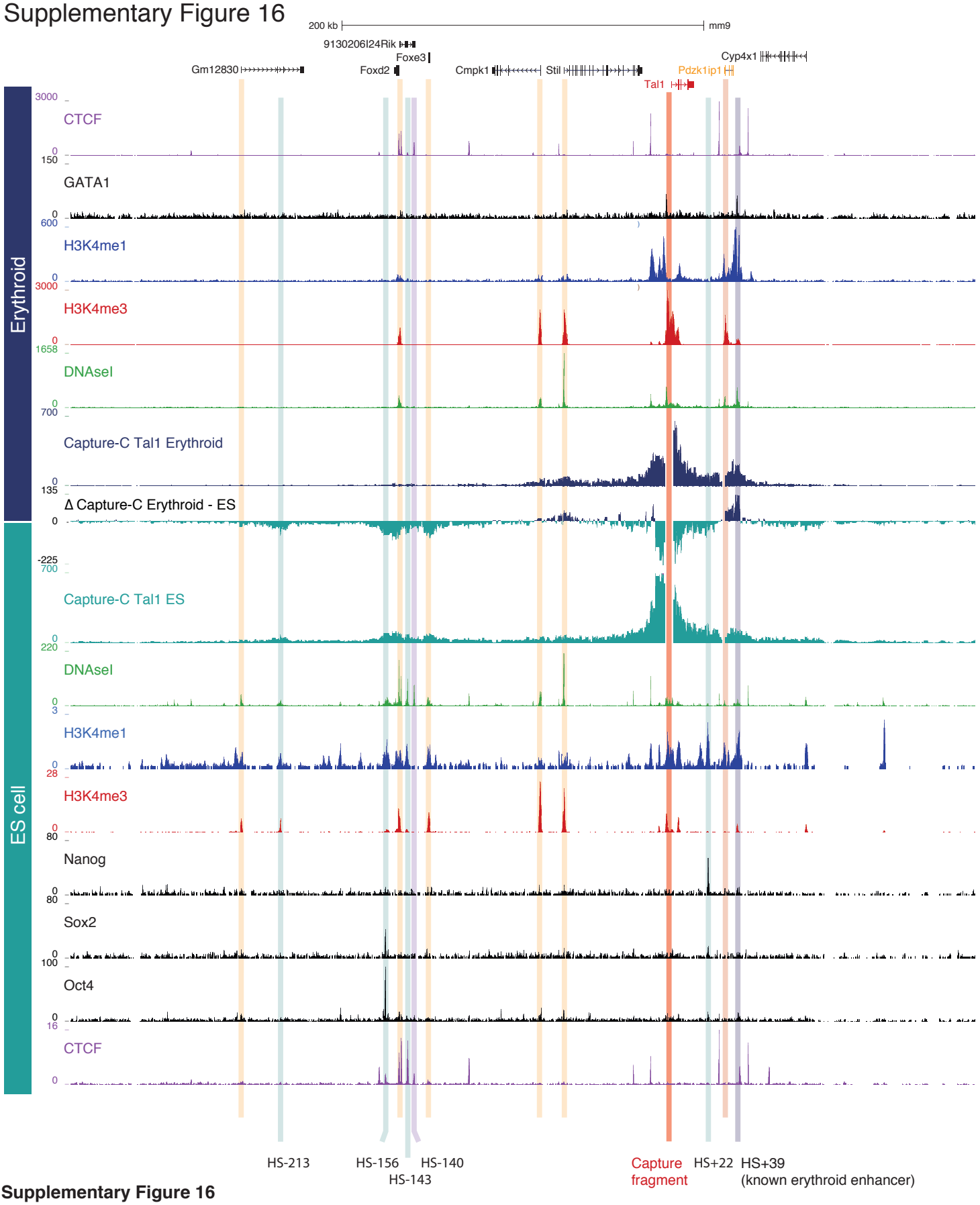
Supplementary Figure 15



Supplementary Figure 15
Capture-C at the *Pnpo* (pyridoxine 5'-phosphate oxidase) locus. This gene is essential for vitamin B(6) metabolism and is ubiquitously expressed. However, vitamin B(6) is important in haem synthesis and the gene is therefore upregulated in erythroid tissues. This is dependent on a regulatory element (HS-26), which becomes DNAseI sensitive in erythroid cells and is bound by GATA1. The Δ Capture-C clearly and very specifically pinpoints this element (red arrow heads), which would otherwise appear to be simply regulating the promoter of the adjacent gene Cdk5rap3.

Erythroid data (Hughes et al., 2014)
ES cell data: DNAseI-seq (ENCODE UW); CTCF (LICR GSM918748)

Supplementary Figure 16

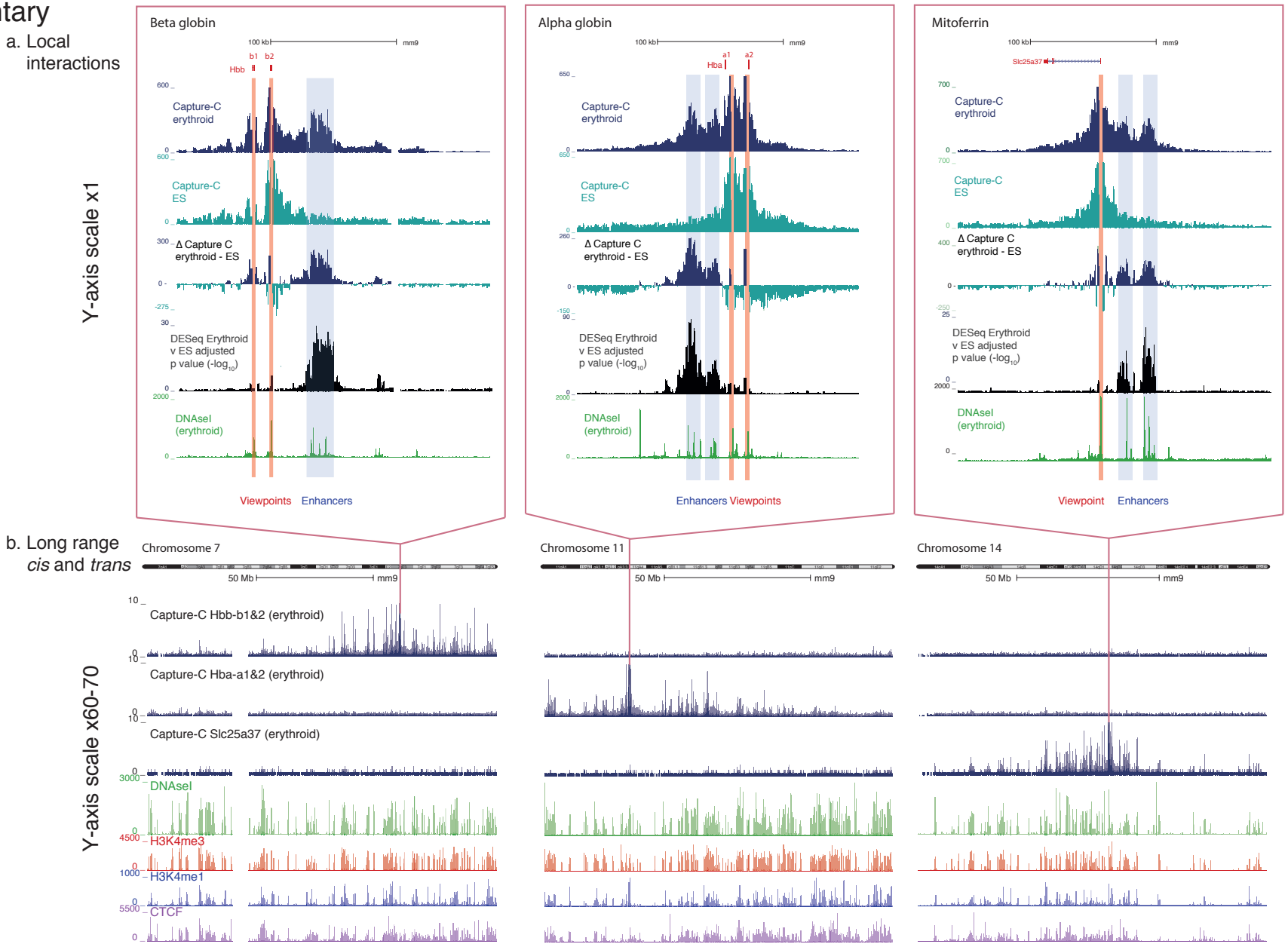


Supplementary Figure 16

Capture-C from the *Tal1* gene in ES and erythroid cells. The Capture-C data helps to describe the potential regulatory elements at this complex locus. Comparative analysis between ES cells and erythroid cells was undertaken using DESeq2. There is a clear interaction with the previously described enhancer at HS+39 (Zhou et al., 2013, Blood http://dx.-doi.org/10.1182/blood-2013-02-483875). In ES cells there are cell type specific interactions with a group of DNaseI sites around 150kb upstream of the gene. This includes the promoter of *Foxe3*, which is H3K4 trimethylated in ES cells and a potential regulatory element bound by Sox2 and Oct4 (HS-156).

Erythroid data: Hughes et al., 2014.
ES cell data: DNaseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259), Sox2 (Young lab GSM1082341), Nanog (Young lab GSM1082342), Oct4 (Young lab GSM1082340).

Supplementary Figure 17

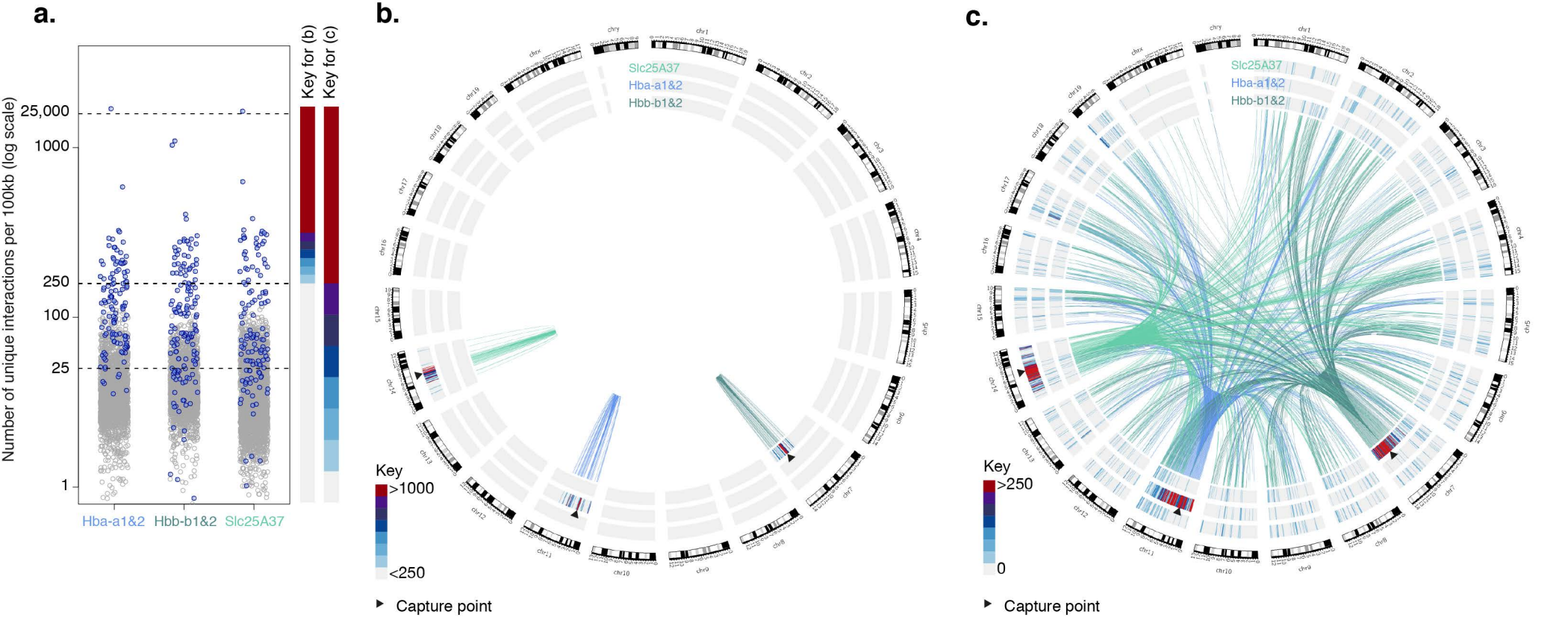


Supplementary Figure 17. Comparison of *cis* and *trans* interactions

a. Top panel shows the normalized mean Capture-C profile at the alpha globin, beta globin and *Slc25A37* (*Mitoferrin 1*) loci in erythroid (n=4) and ES cells (n=3). These data were generated along with the profiles for another 32 gene promoters simultaneously from seven samples in a single capture reaction (making a total of 245 interaction profiles from one oligonucleotide capture reaction). The captured viewpoint fragments are highlighted in red and the interactions with the well known enhancers as annotated by DNAseI hypersensitivity are highlighted in blue. The differential track (Δ Capture-C) shows that interactions with the local erythroid enhancers are clearly and specifically increased in erythroid cells when the genes are active. The Y-axis denotes the mean number of unique interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide. DESeq analysis highlights regulatory elements precisely with p values below 10^{-20} over multiple adjacent fragment.

b. This shows the whole chromosome in cis for the above genes (beta globin (Hbb-b1&2) - chr7; alpha globin (HbA-a1&2) – chr11; Mitoferrin 1 (Slc25a37) – chr14). Note the 60-70-fold change in scale on the Y-axis compared to the top panel. Much weaker interactions can be seen spreading out from the viewpoint. These correlate with active promoters and enhancers as well as CTCF sites as determined by DNAseI hypersensitivity and ChIP-seq for H3K4me3, H3K4me1 and CTCF.

Supplementary Figure 18



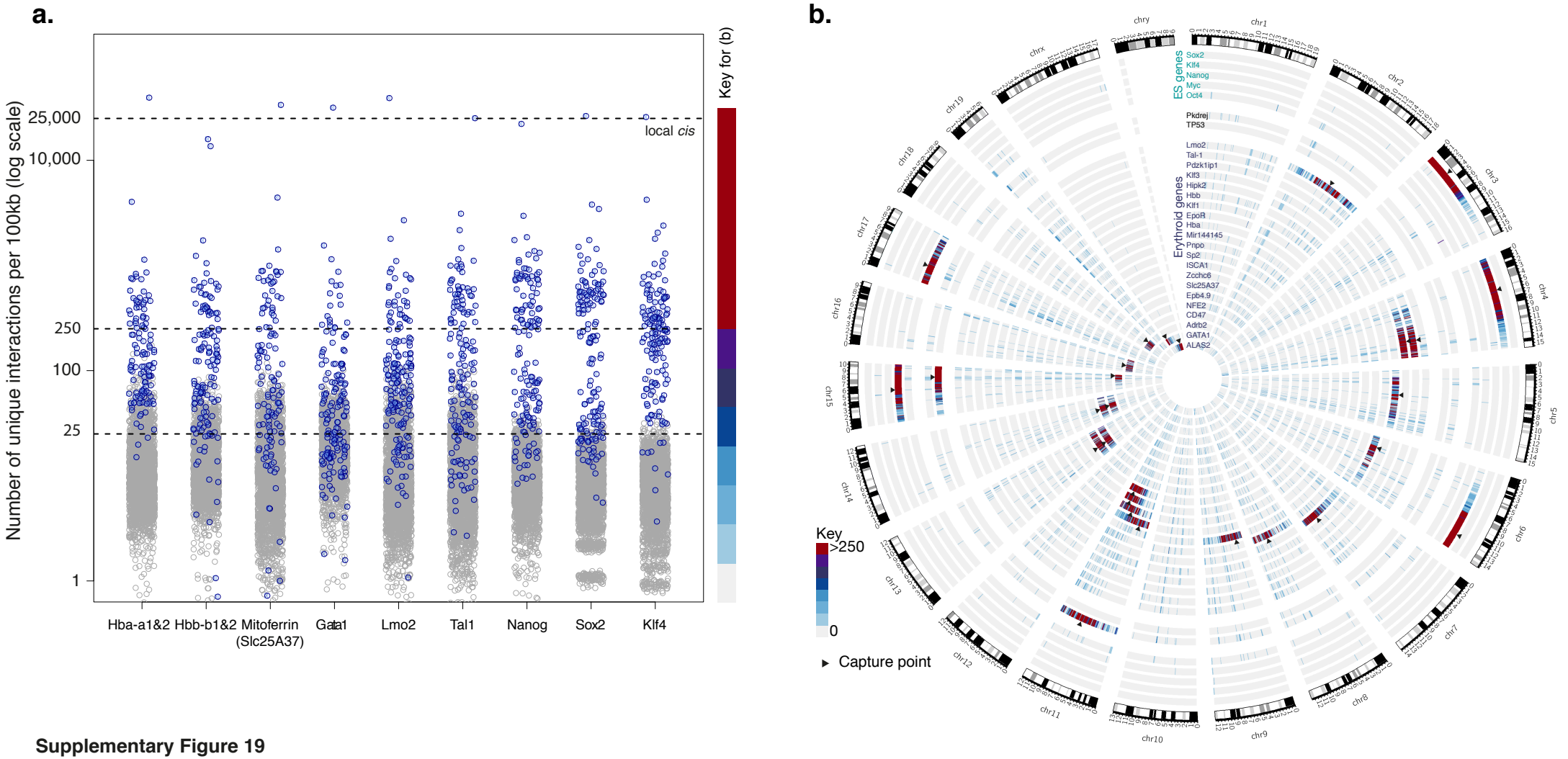
Supplementary Figure 18

a. Scatter plot of numbers of interactions in 100kb bins from three viewpoints. *Cis* and *trans* interactions are denoted by blue and grey circles respectively.

b. Genome wide heatmap (each gene promoter is shown in a separate ring) and looping glyphs of interactions from Mitoferrin (*Slc25A37*), *Hba-a1&2* and *Hbb-b1&2.* A cut off of 250 interactions per 100kb has been used to exclude very weak infrequent interactions. This shows that there are no interactions in *trans* with any of the gene promoters with a frequency of 100 fold less than those of the strong functional local interactions.

c. Genome wide heatmap and looping glyphs of interactions from the same viewpoints as in a., however, no cut off has been used to exclude very infrequent interactions. This shows that there are very weak *trans* from all of the gene promoters. Interestingly these weak contacts have similar distributions in *trans* irrespective of the view point used.

Supplementary Figure 19

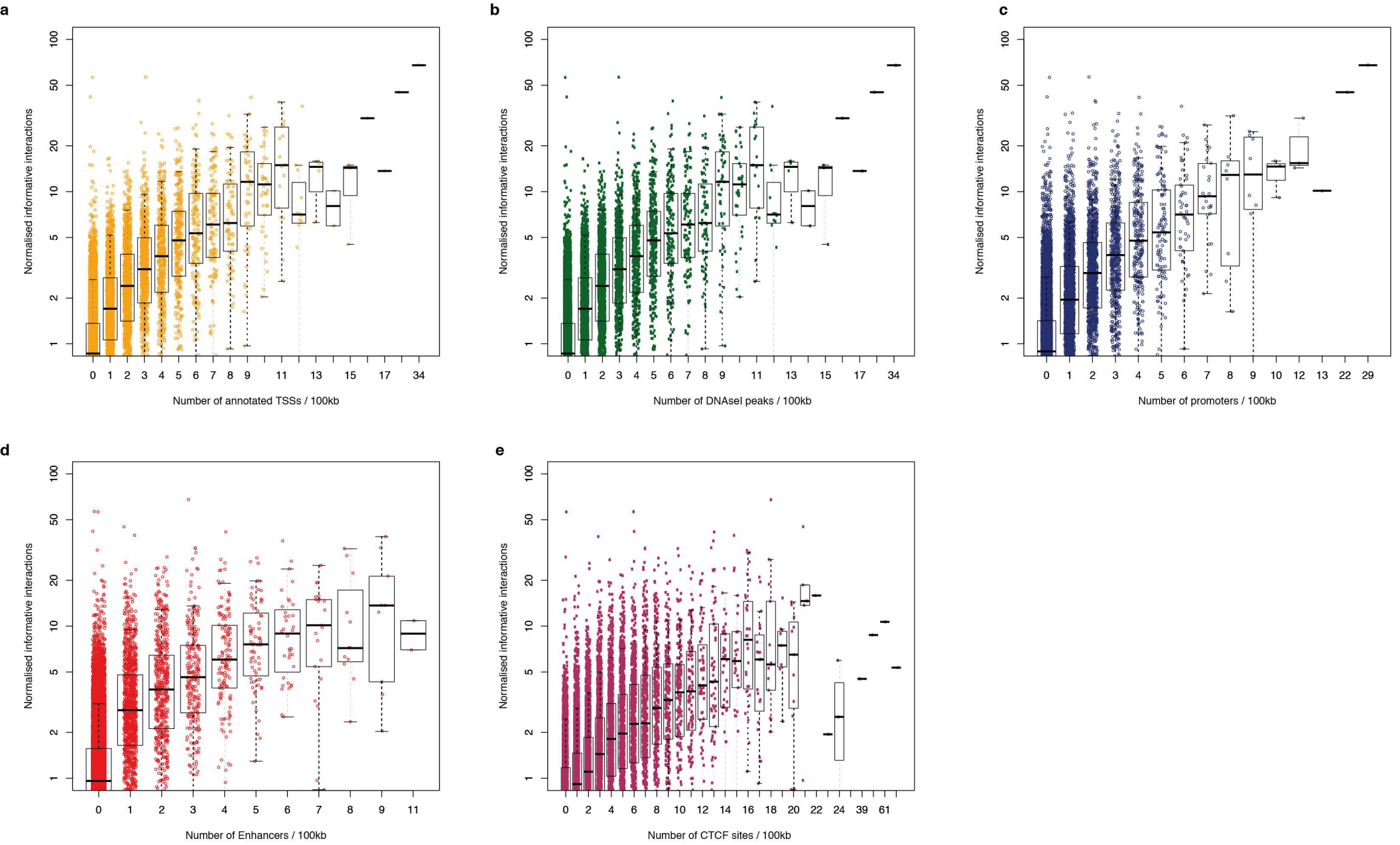


Supplementary Figure 19

a. Scatter plot of the total informative interaction count across the genome in 100kb bins. *Cis* interactions are in blue and *trans* in grey. For all of the genes the local interactions around the viewpoint are around 100-fold stronger than the weak long range *cis* interactions and 1000-fold more frequent than *trans* interactions.

b. Genome wide heat map of interactions from 28 genes. Each ring represents the profile from a single view point, with the captured viewpoint being denoted by the black arrow heads. The heat map shows that strong interactions (red) are only found in *cis* from all 28 genes. In addition weak *trans* interactions can be seen and these tend to have similar distributions from active genes irrespective of the view point.

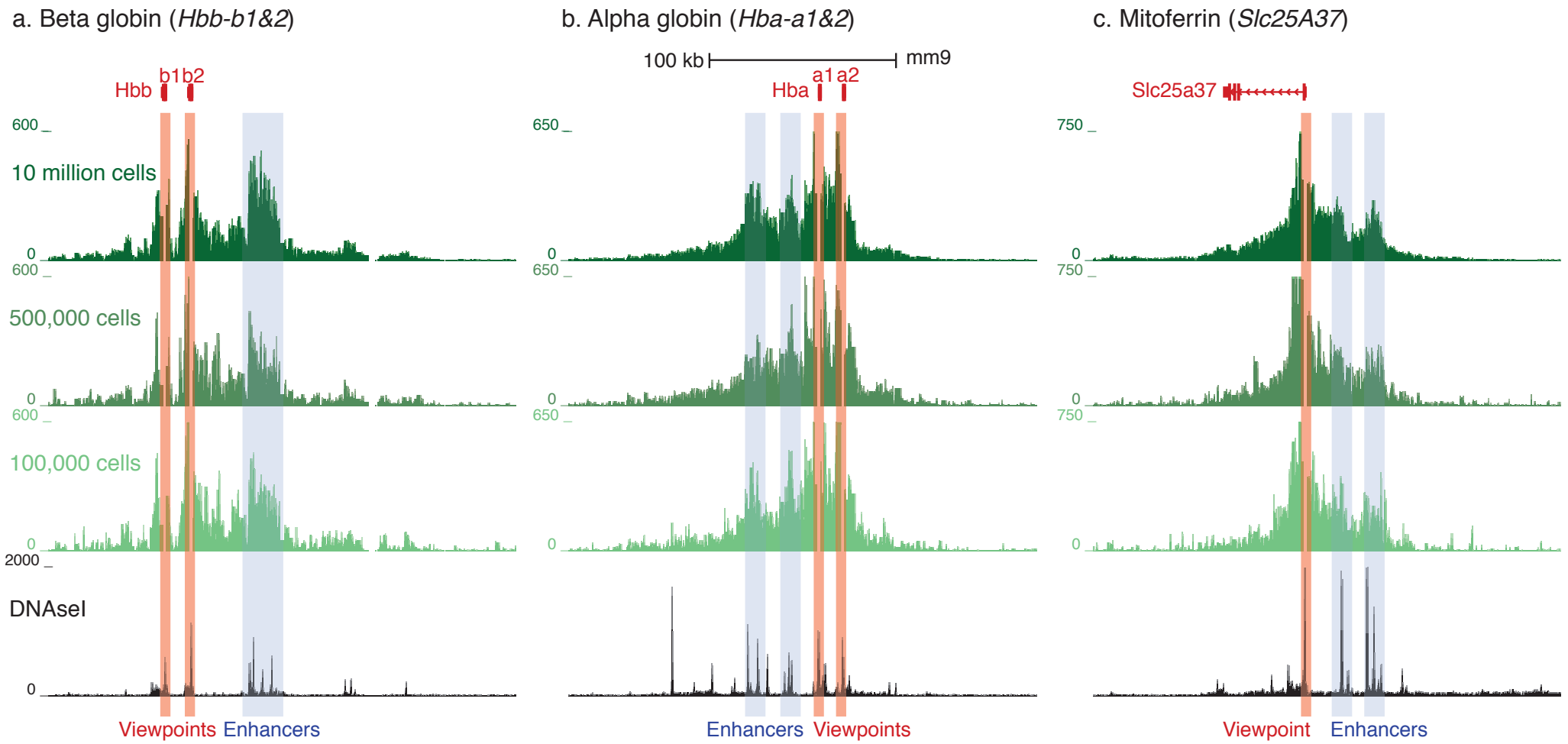
Supplementary Figure 20



Supplemenatary Figure 20

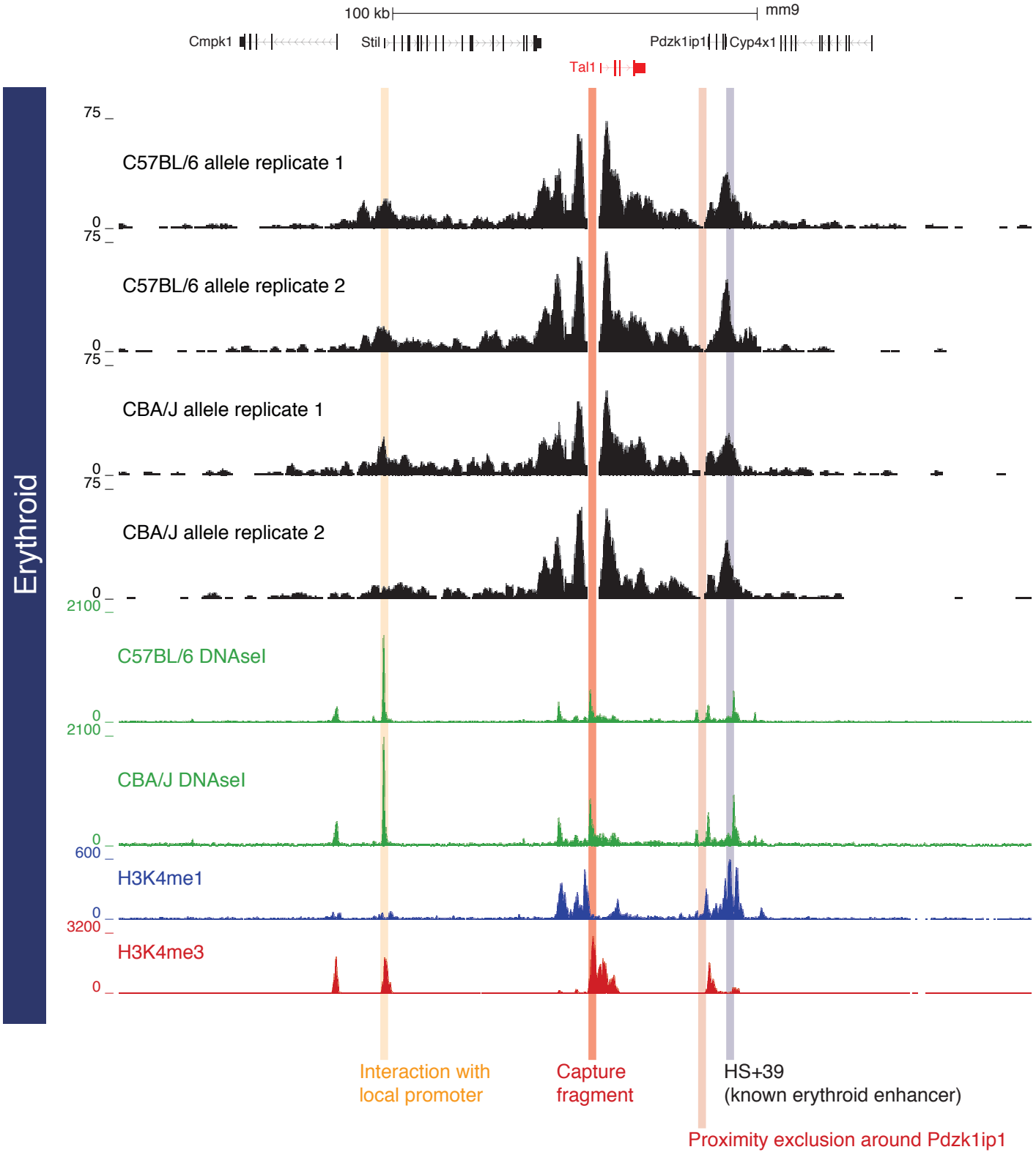
Analysis of *trans* interactions from the alpha globin (*Hba-a1&2*) promoters. The genome was divided into 100kb bins and the number of a. transcription start sites; b. DNAseI hypersensitive peaks; c. enhancers (as determined by DNAseI and H3K4me1); d. promoters (as determined by DNAseI and H3K4me3) and e. CTCF peaks was counted in each bin. There is a correlation between all of these factors and the number of normalised informative interactions.

Supplementary Figure 21



Supplementary Figure 21. Robust interaction profiles generated from small numbers of cells
Comparison of capture-c profiles from the promoters at the beta globin (*Hbb-b1&2*) (a), alpha globin (*Hba-a1&2*) (b) and mitoferrin (*Slc25A37*) (c) loci (highlighted in red) generated from different numbers of cells ranging from 10 million to 100,000. The plots show the number of interactions per restriction enzyme fragment for the mean of two biological replicates, normalized to 100,000 total informative interactions genome-wide. Despite the 100-fold reduction in cell number, the interactions with the main regulatory elements (highlighted in blue) can still be clearly detected at all three genes.

Supplementary Figure 22



Supplementary Figure 22
Allele specific Capture-C data (with a 2kb moving window) from the *Tal-1* locus from F1 erythroid cells (C57BL/6 x CBA/J crosses). Since the gene regulation at this locus is the same in both strains the profiles from the different alleles are very similar.