**Supplementary note**

## Analytical concepts in the analysis of chromatin conformation capture (3C) data and NG Capture-C data.

As a chromosome is a linear polymer the frequency with which any point contacts the rest of the polymer shows an inverse relationship with distance that approximates an inverse square relationship.

- The shape of the distribution is approximately symmetrical but will be affected by the characteristics of individual loci, such as the presence of neighboring active genes or binding sites for structural proteins such as CTCF.
- The specific characteristics of the slope of this relationship are dependent on the cutting frequency of the enzyme used, the depth at which the data is sampled and consequently the window size used to sample the distribution (if required).
- NG Capture-C clearly shows this characteristic distribution when inactive loci are analyzed (See Figure SA1 panel a).
- NG Capture- C uses frequent cutting 4-cutter enzymes and generates data at the highest possible resolution (per restriction fragment or half restriction fragments). Therefore no windowing is required in its analysis.
- At this resolution, the majority of the skew for distance for NG-Capture-C data falls within an ~100kb window of the viewpoint, peaks within an ~ 20k window, and is minimal over the remaining bulk of the chromosome (Figure SA1, panel b).

## Analytical approaches to the analysis of chromatin conformation capture (3C) data and NG Capture-C data.

This general property must be accounted for in some way when determining significant interactions in 3C data. Capture-C data is most similar to a multiplexed form of 4C or 3Cseq data. The approaches commonly used to analyze such data involve the modeling of the underlying propensity of a capture point to interact at a given distance to generate a generalized model. This general model is then used to normalize the empirically derived data from specific loci to call significant interactions.

- The best of such models sample many instances of real data from individual loci to produce a generalized model or generate a model specific to each viewpoint.
- A general model when applied to all analyzed loci does not take into account the specifics of any particular loci, such as chromatin structure, other active genes or the distribution of important structural proteins such as CTCF.

- The output is highly dependent on the tool used and the parameters employed and requires expert manual optimization.
- Such approaches are at their most variable close to the viewpoint where small variations in models or parameters have large effects on the output, due the necessary strictness of normalization required at these distances (Figure SA1 panel a).

**Analysis of NG Capture-C data using distance normalization approaches.**

As NG Capture-C generates data in standard next generation sequencing (NGS) formats (SAM and BAM file) which are easily converted to count based files, these data are straightforward to analyze with existing packages. However, it is important to understand the underlying principles which the approach uses to do any normalization, to judge if this is appropriate for Capture-C data (see appendix on Gothic analysis).

We have tested the analytical performance of two commonly used tools for 4C and 3Cseq data interrogation using NG Capture-C data. We chose the tools, r3C-seq and FourCseq, as they use models generated from the sampling of real data and have several thoughtfully integrated features for both the use of replicates and for comparative analysis. Importantly both tools use differing models to account for distance, a reverse power law and monotonically declining model in the case of r3C-seq and FourCseq respectively.

We tested the tools on well-characterized test loci; $\alpha$ globin (Figures SA2 and SA3); $\beta$ globin (Figures SA4 and SA5) and *Slc25a37* (*Mitoferrin 1*; Figures SA6 and SA7). Of these three loci $\alpha$ and $\beta$ globin are used as gold standards in the 3C field due to the depth of the functional knowledge concerning their regulation. Importantly we ran these analyses on default parameters to simulate the output of these tools at uncharacterized loci.

Both tools performed well in the $\beta$ globin and *Slc25a37* loci, significantly and specifically identifying the known regulatory elements (Figures SA2-SA7). However, both tools under-called the most proximal known elements in the $\alpha$ globin locus to a variable degree; R4 in the case of r3C-Seq and R3, Rm and R4 in the case of FourC-seq (Figures SA2, SA3).

Of course as these elements are known interacting elements in these loci the default parameters may be slackened to now include these closely situated elements. However, the need for such manual optimization makes this suboptimal as a general approach for the identification of regulatory elements in uncharacterized loci.

Importantly, distance normalization did not additionally call any of the weak, long-range interactions along the length of the chromosome as significant using these parameters (Figures SA8 - SA11). This is unsurprising due to the large differences in scale between these long-range interactions and those of the regulatory elements, combined with the minimal corrections imposed by the models at these distances (see *cis* panels Figures SA8 - SA11).

## Identification of regulatory interactions by comparative analysis between tissues.

In light of the detection bias of proximal interactions by these existing approaches it would be of great value to have a complementary approach for the identification of regulatory elements, based on a different premise.

Regulatory elements are known to significantly increase interaction frequency with their cognate promoters in tissues in which the gene (or genes) are active as compared to inactive (Figure SA12). In this way the propensity of the promoter to interact with its surrounding can be determined empirically and at high resolution for each inactive locus, rather than estimated via a generalized model and used to subtract from the active locus, revealing the acquired regulatory interactions.

This observation has been incorporated into mature tools for the analysis of 3C data such as r3C-seq and FourC-seq and these tools significantly detect fold change enrichment at the known elements in our analyses (Figures SA2 – SA7), however the stringency of the distance models removes them from the final output at the $\alpha$ globin locus under default parameters (Figures SA2 and SA3).

NG Capture-C is uniquely suited to use comparative analysis to identify regulatory interactions, due its inherent ability to multiplex more than 6 replicates in a single capture reaction. In this way data from triplicates of biological replicates can be generated from both active and inactive tissues in a single capture reaction, generating highly reproducible and comparable data in a high throughput manner.

These data can be analyzed by tools designed for general analytical approaches, such as Deseq2 (which is also used by FourCseq), to detect significant changes between active states. We tested this approach in our test loci and showed that it identified all known regulatory elements in these loci, including those proximal interactions under-called by the other approaches, specifically and at very high levels of statistical significance (Figure 3 in the main text).

## Appendix

### Correct choice of analysis tools for NG Capture-C data.

Although NG Capture-C data can be loaded into most existing analysis tools it is important to understand the logic employed by any particular tool before using it. The HI-C tools HiCUP and GOTHiC have previously been used to analyze data generated by Capture Hi-C (original Capture-C applied to Hi-C libraries) and it may be assumed that this is also a valid approach for the analysis of NG Capture-C data. However, the basic premise of the model used to call significant interactions in GOTHiC appears to be only valid for data enriched by Hi-C or Hi-C in conjunction with weak sequence specific enrichment.

This is because the model assumes a large component of the data will be the weak "all against all" background characteristic of Hi-C datasets and a further component will be the enriched interactions which will form a binomial distribution. The tool then uses this to call significant interactions by the binomial testing of this assumed data structure.

However, NG Capture-C is the enrichment of a 3C (rather than Hi-C) library by extremely efficient sequence specific enrichment where up to 50 % of the library now comes from specified capture points. Hence the background that remains cannot be assumed to be random "all against all" rather it is the skewed background left by sequence specific enrichment. When tested on NG Capture-C data the p and q values appear to simply follow the bulk enrichment of signal depending on the window size used (Figures SA13 and SA14). Additionally as GOTHiC neither corrects for distance nor has the inherent ability to use replicates or do comparative analysis it would appear not to be suitable for NG Capture-C analysis.

Rather, considering the similarity between NG Capture-C data and 4C or 3C-seq our analysis would suggest that a comprehensive pipeline to identify regulatory elements from NG Capture-C data would consist of comparative analysis between informative tissues using the DEseq2 or similar approach (see main text) combined with appropriate analysis tools such as r3C-seq or FourC-seq.

## Figure SA1 The relationship between interaction frequency and genomic distance in NG Capture-C data.

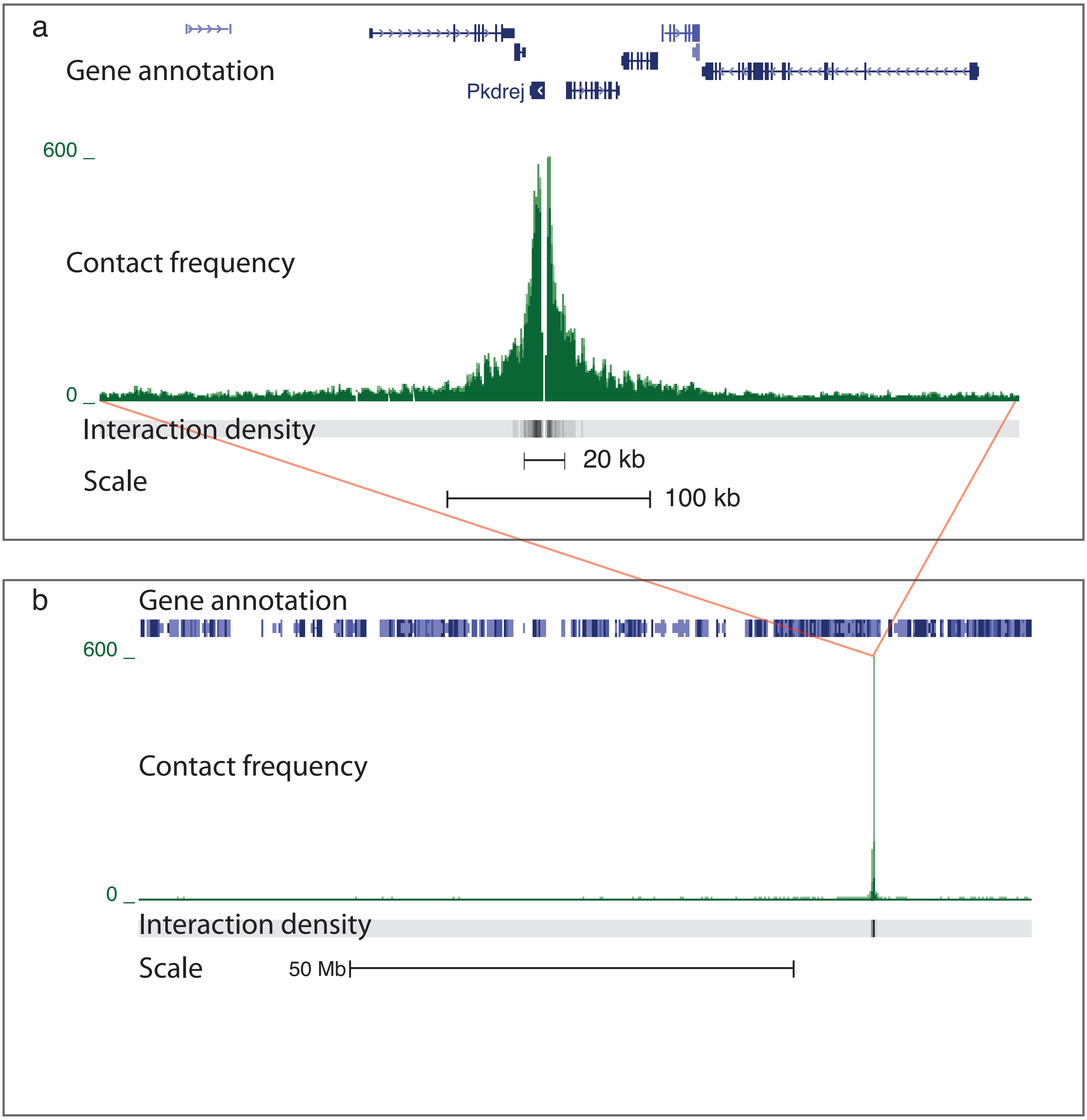

**Figure SA1**
Panel (a) shows the symmetrical distribution of interaction frequencies (per Dpn II fragment) from the inactive promoter of the sperm specific gene *Pkdrej* in eryrthoid tissues, decreasing as an approximate inverse square of distance (green). Below this is a bar showing interaction density colored in grey-scale.
Panel (b) shows the same data as panel (a) on the same color scheme but now over the whole chromosome (chr15).
The red lines show the location of the region covered in panel (a) on panel (b).
The data is plotted in mouse genome build mm9, and the positions of USCS genes are shown as violet boxes in the top of each panel.

# Figure SA2 Comparative analysis of the α globin locus in erythroid and ES cells using r3Cseq.

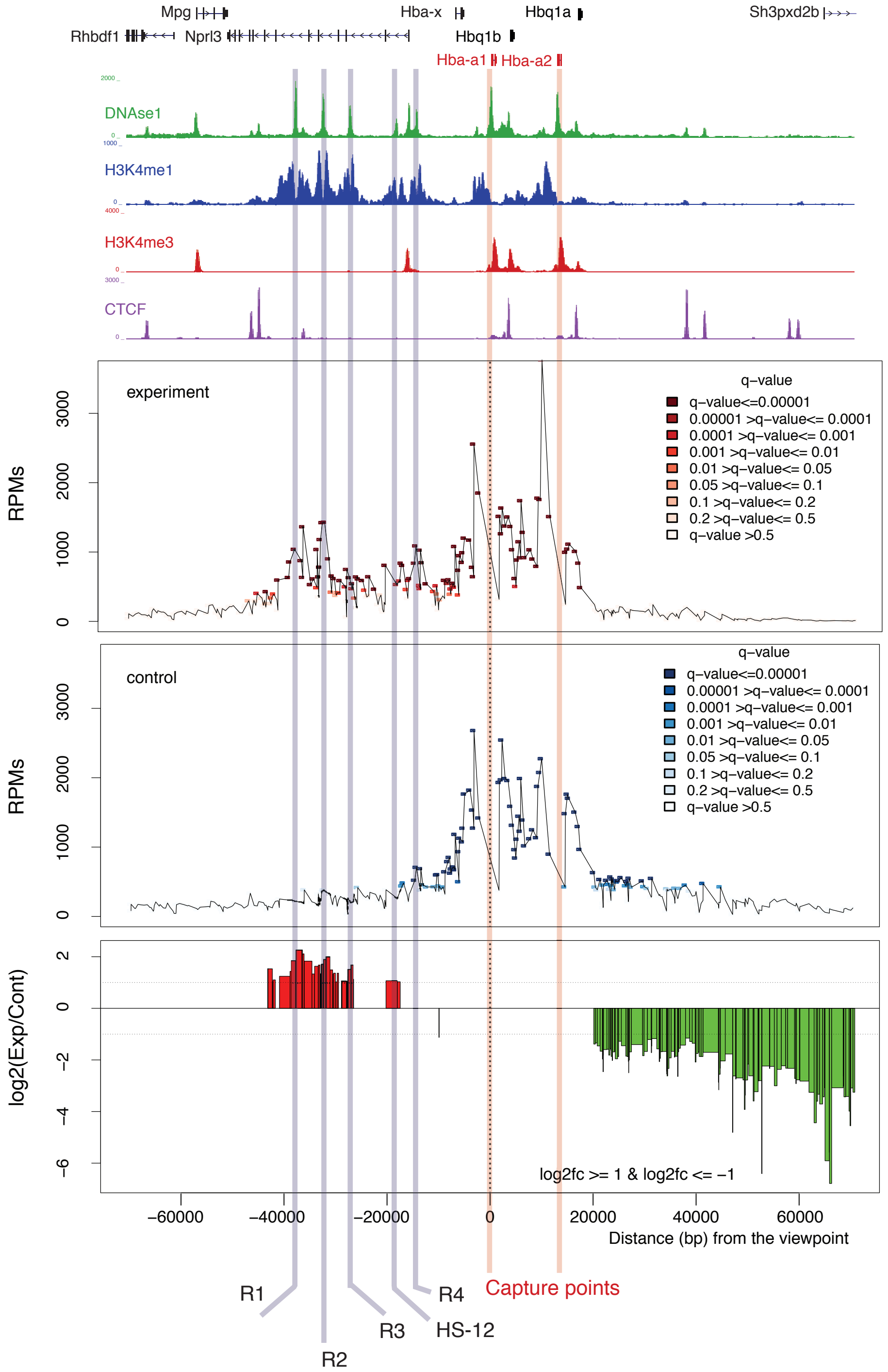

**Figure SA2**

The output of the comparative analysis of r3Cseq package for the mouse α globin locus in erythroid and ES cells, run on default parameters.

The top panel shows the chromatin landscape of the locus in erythroid cells in mouse mm9 genome, showing the capture points and characterized elements within the locus highlighted by red hatched and grey solid lines, respectively, and their correspondence with DNAse-seq, H3K4me1 and H3K4me3 and CTCF binding. The positions of UCSC genes are shown as red and black boxes in the top of the panel. These data are scaled to the output of r3Cseq covering a 70kb window upstream and downstream of the viewpoint.

The two panels below this show the RPM values as calculated by the r3Cseq tool in "experiment" (erythroid cells) and "control" (ES cells). The colour coded ranges of calculated q-values are shown to the right of each panel.

The bottom panel shows the log2 ratio of the RPM values for regions which pass the default q-value filter of 0.05. Regions enriched in erythroid cells are coloured red and regions enriched in ES cells are displayed in green.

Distance (bp) from the viewpoint is shown below this in increments of 20kb, positive values 3' of the viewpoint and negative values 5'. The position of capture points and characterized elements are annotated below the red and the grey solid orientation lines, respectively.

# Figure SA3 Comparative analysis of the $\alpha$ globin locus in Erythroid and ES cells using FourCSeq

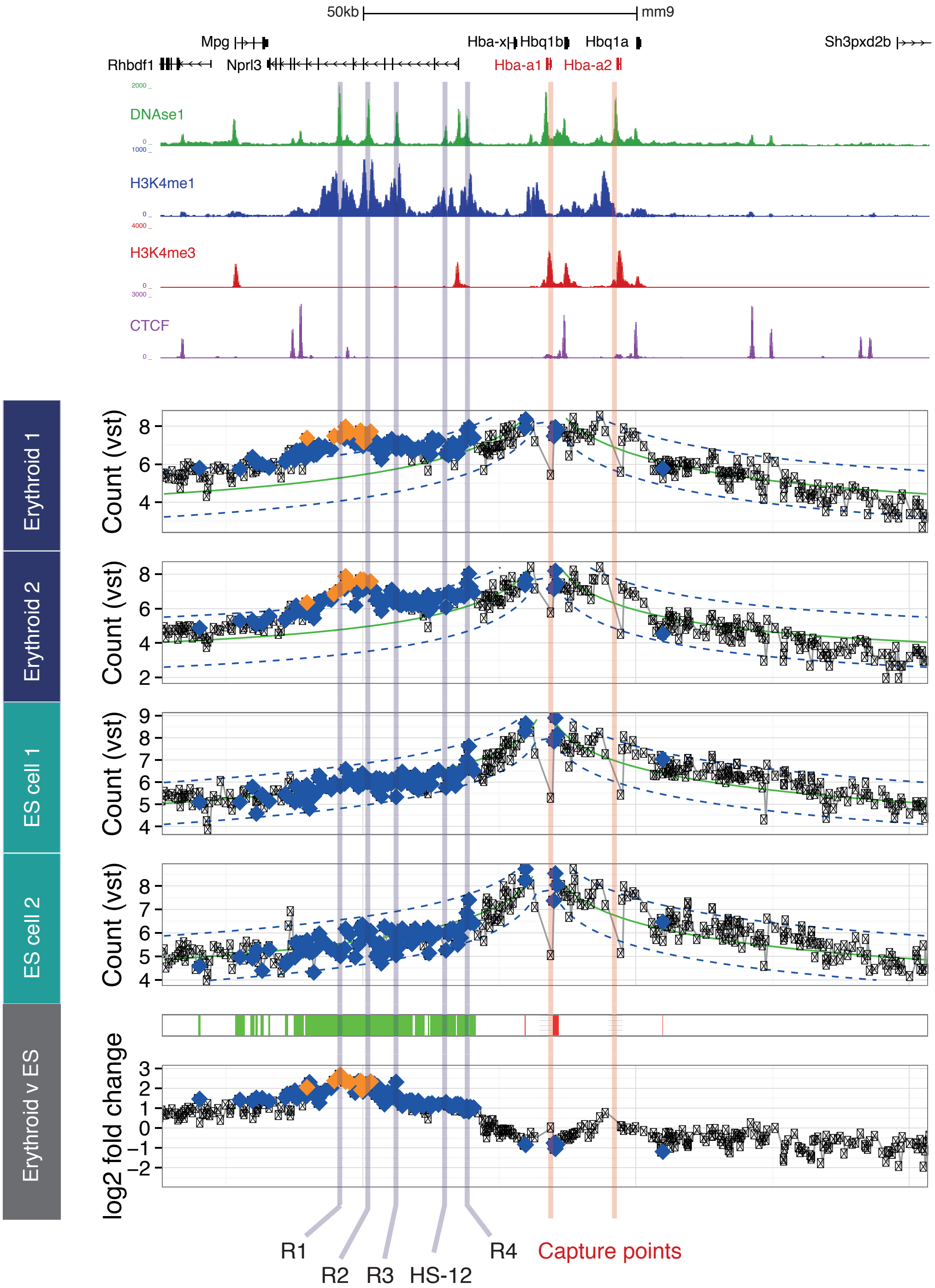

**Figure SA3**
The output of the comparative analysis of FourCSeq package for the mouse $\alpha$ globin locus in erythroid and ES cells, run on default parameters.
The top panel shows the chromatin landscape of the locus in erythroid cells in mouse genome mm9 build, showing the capture points and characterized elements within the locus highlighted by red hatched and grey solid lines, respectively, and their correspondence with DNAse-seq, H3K4me1, H3K4me3 and CTCF binding. The positions of UCSC genes are shown as red and black boxes in the top of the panel. These data are scaled to the output of FourCSeq covering a 70kb window upstream and downstream of the viewpoint.
The subsequent 4 panels show the output from two erythroid replicates (Erythroid 1 + 2) and two ES replicates (ES cell 1+2). Interactions which show significance after Z-score correction are colored red, and interactions, which show significant fold enrichment between the two conditions are colored blue. Interactions significant for both attributes are colored orange.
Below this is a color code bar which shows when interactions are stronger in the first two datasets (green) or the converse (red). The final panel shows the log2 fold change with statistical significance colored as above.
The position of the capture points and the characterized elements are annotated below the red and grey solid orientation lines, respectively.

# Figure SA4 Comparative analysis of the $\beta$ globin locus in erythroid and ES cells using r3Cseq.

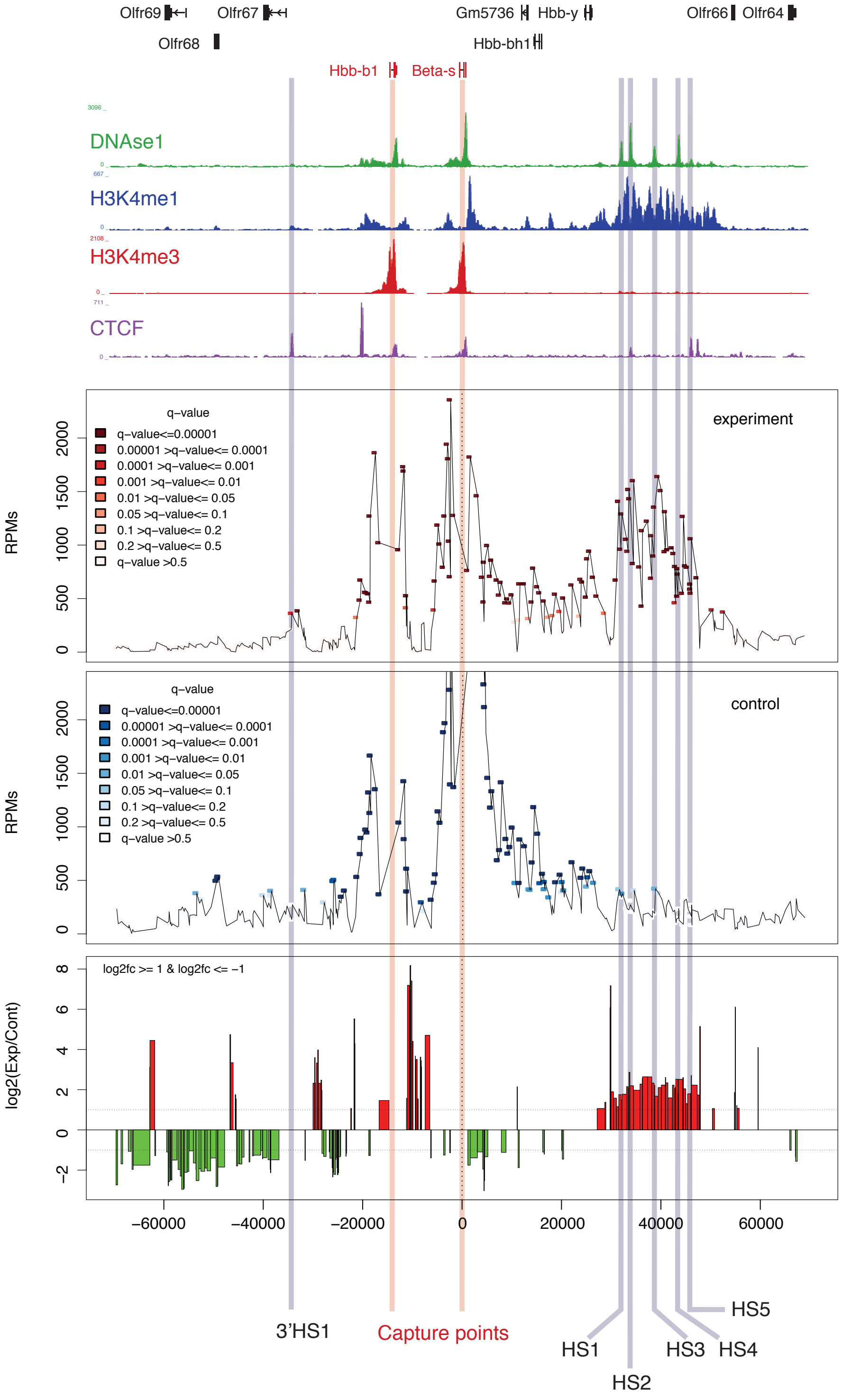

Figure SA4

The output of the comparative analysis of r3Cseq package for the mouse b globin locus in erythroid and ES cells, run on default parameters. The top panel shows the chromatin landscape of the locus in erythroid cells in mouse genome build mm9, showing the capture points and characterized elements within the locus highlighted by red hatched and grey solid lines, respectively, and their correspondence with DNAse-seq, H3K4me1 and H3K4me3 and CTCF binding. The positions of UCSC genes are shown as red and black boxes in the top of the panel. These data are scaled to the output of r3Cseq covering a 70kb window upstream and downstream of the viewpoint.

The two panels below this show the RPM values as calculated by the r3Cseq tool in “experiment” (erythroid cells) and “control” (ES cells). The colour coded ranges of calculated q-values are shown to the left of each panel. The bottom panel shows the log2 ratio of the RPM values for regions which pass the default q-value filter of 0.05. Regions enriched in erythroid cells are coloured red and regions enriched in ES cells are displayed in green.

Distance (bp) from the viewpoint is shown below this in increments of 20kb, positive values 3’ of the viewpoint and negative values 5’. The position of capture points and characterized elements are annotated below the red and the grey solid orientation lines, respectively.

# Figure SA5 Comparative analysis of the β globin locus in erythroid and ES cells using FourCSeq

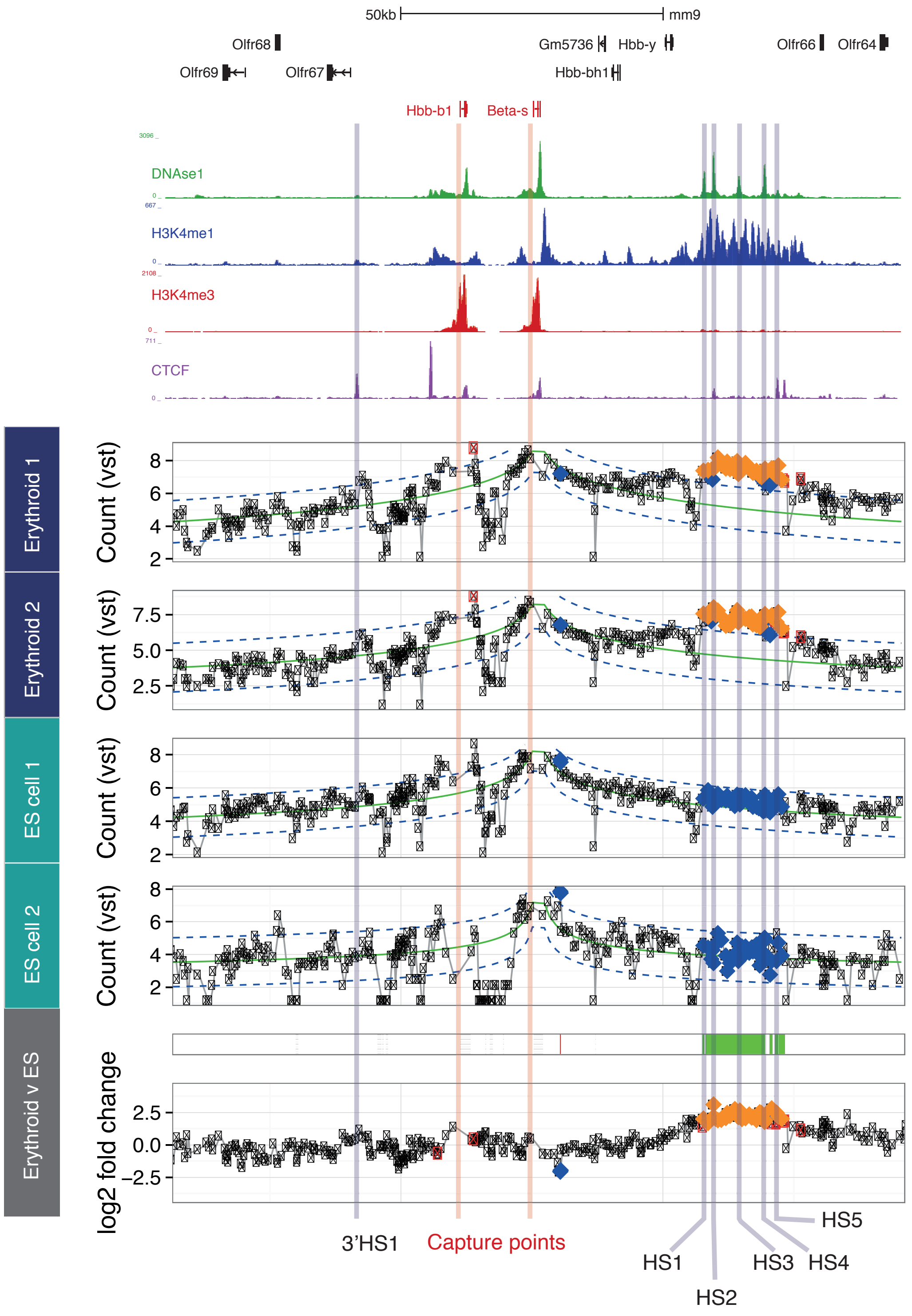

**Figure SA5**
**The output of the comparative analysis of FourCSeq package for the mouse β globin locus in erythroid and ES cells, run on default parameters.**
**The top panel shows the chromatin landscape of the locus in erythroid cells in mouse genome build mm9, showing the capture points and characterized elements within the locus highlighted by red hatched and grey solid lines, respectively, and their correspondence with DNAse-seq, H3K4me1, H3K4me3 and CTCF binding. The positions of UCSC genes are shown as red and black boxes in the top of the panel. These data are scaled to the output of FourCSeq covering a 70kb window upstream and downstream of the viewpoint.**
**The subsequent 4 panels show the output from two erythroid replicates (Erythroid 1 + 2) and two ES replicates (ES cell 1+2). Interactions which show significance after Z-score correction are colored red, and interactions, which show significant fold enrichment between the two conditions are colored blue. Interactions significant for both attributes are colored orange.**
**Below this is a color code bar which shows when interactions are stronger in the first two datasets (green) or the converse (red). The final panel shows the log2 fold change with statistical significance colored as above.**
**The position of the capture points and the characterized elements are annotated below the red hatched and grey solid orientation lines, respectively.**

# Figure SA6   Comparative analysis of the *Slc25a37* (*Mitoferrin 1*) locus in erythroid and ES cells using r3Cseq.

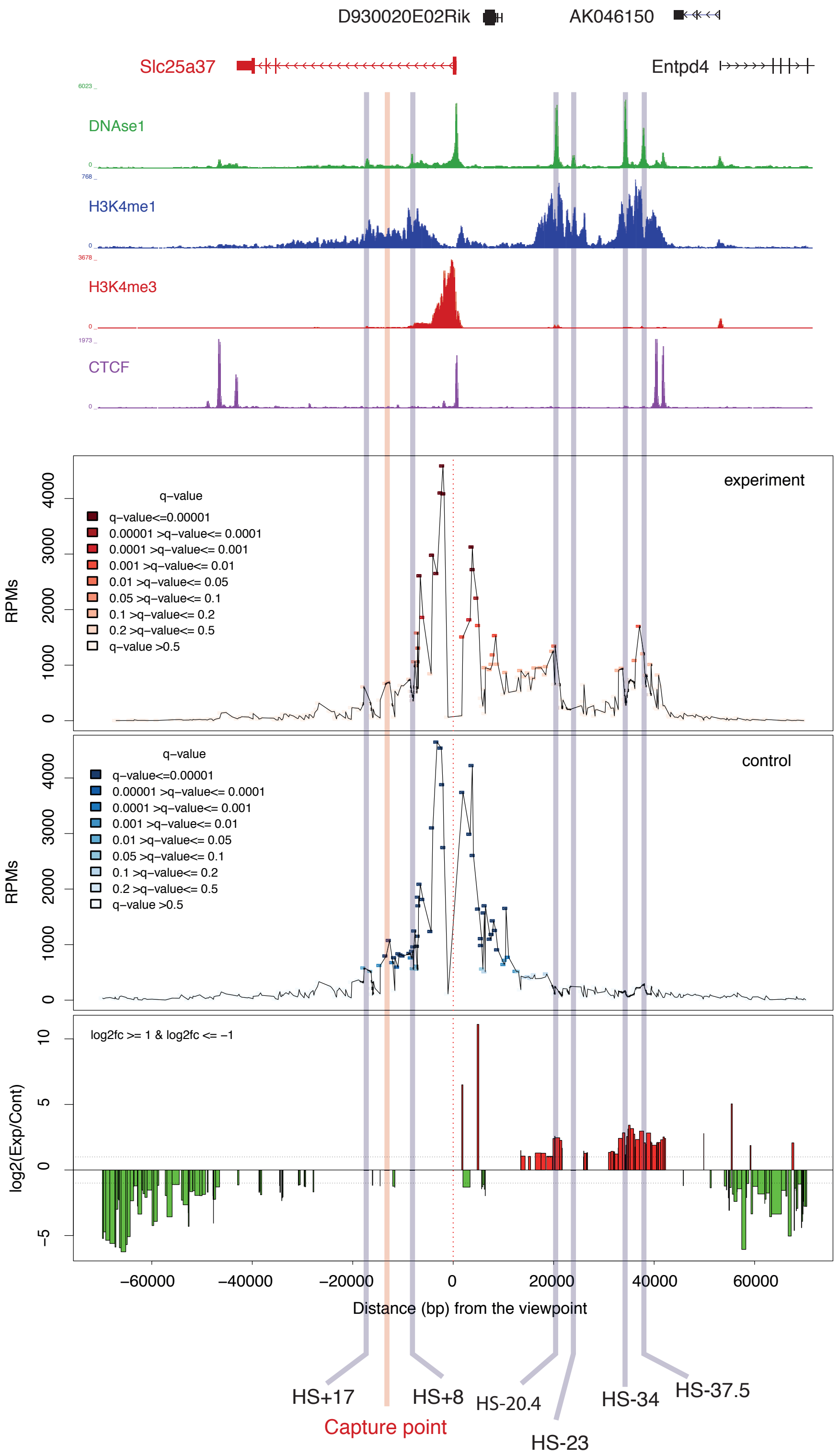

**Figure SA6**

The output of the comparative analysis of r3Cseq package for the mouse *Slc25a37* (*Mitoferrin 1*) locus in erythroid and ES cells, run on default parameters.

The top panel shows the chromatin landscape of the locus in erythroid cells in mouse genome build mm9, showing the capture point and characterized elements within the locus highlighted by red and grey solid lines, respectively, and their correspondence with DNAse-seq, H3K4me1 and H3K4me3 and CTCF binding. The positions of RefSeq genes are shown as red and black boxes in the top of the panel. These data are scaled to the output of r3Cseq covering a 70kb window upstream and downstream of the viewpoint.

The two panels below this show the RPM values as calculated by the r3Cseq tool in “experiment” (erythroid cells) and “control” (ES cells). The colour coded ranges of calculated q-values are shown to the left of each panel. The bottom panel shows the log2 ratio of the RPM values for regions which pass the default q-value filter of 0.05. Regions enriched in erythroid cells are coloured red and regions enriched in ES cells are displayed in green.

Distance (bp) from the viewpoint is shown below this in increments of 20kb, positive values 3’ of the viewpoint and negative values 5’. The position of capture points and characterized elements are annotated below the red hatched and the grey solid orientation lines, respectively.

## Figure SA7 Comparative analysis of the *Slc25a37* (*Mitoferrin 1*) locus in erythroid and ES cells using FourCSeq

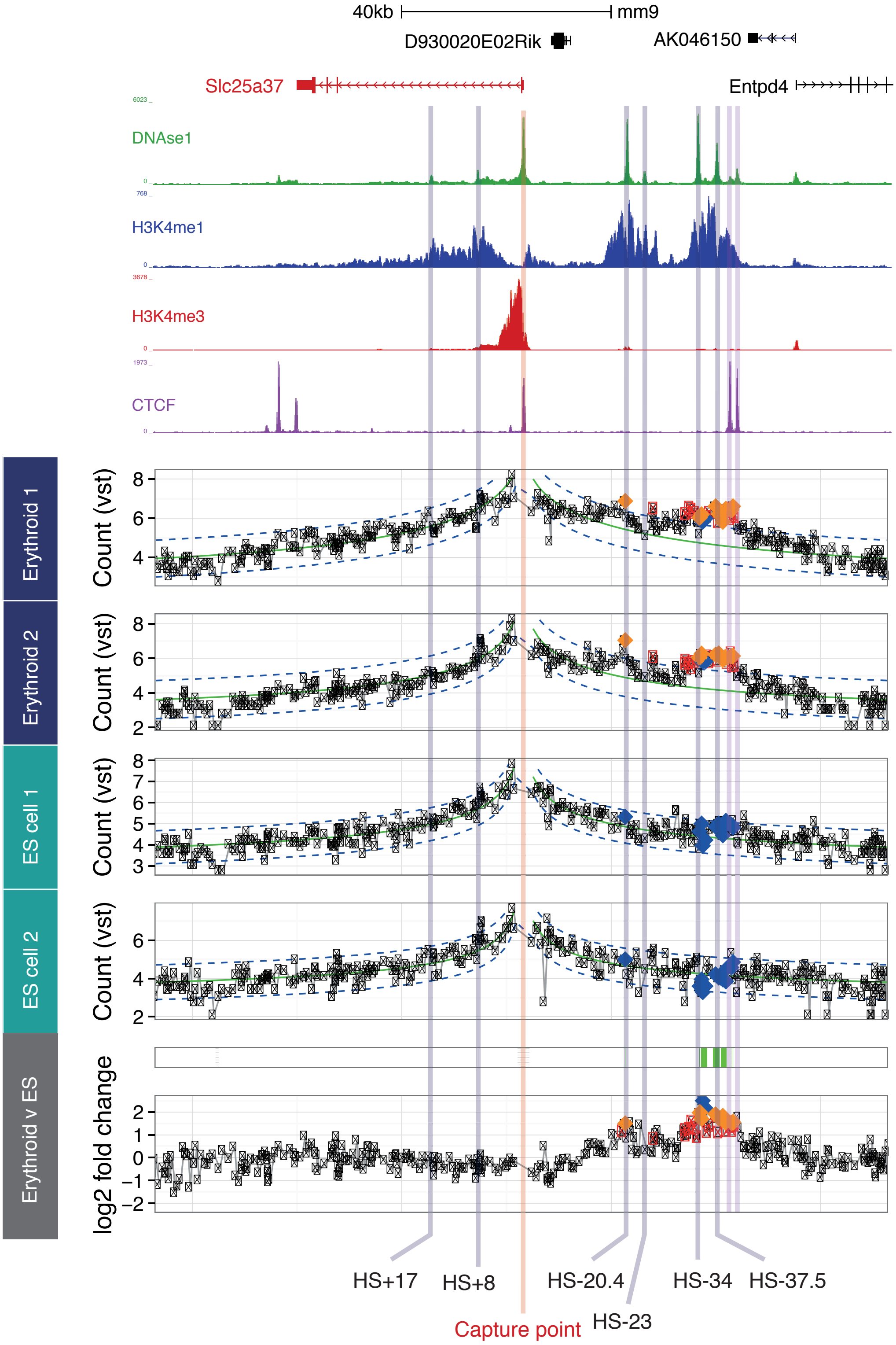

**Figure SA7**

**The output of the comparative analysis of FourCSeq package for the mouse *Slc25a37* (*Mitoferrin 1*) locus in erythroid and ES cells, run on default parameters.**

**The top panel shows the chromatin landscape of the locus in erythroid cells in mouse genome build mm9, showing the capture points and characterized elements within the locus highlighted by red hatched and grey solid lines, respectively, and their correspondence with DNAse-seq, H3K4me1, H3K4me3 and CTCF binding. The positions of UCSC genes are shown as red and black boxes in the top of the panel. These data are scaled to the output of FourCSeq covering a 70kb window upstream and downstream of the view-point.**

**The subsequent 4 panels show the output from two erythroid replicates (Erythroid 1 + 2) and two ES replicates (ES cell 1+2). Interactions which show significance after Z-score correction are colored red, and interactions, which show significant fold enrichment between the two conditions are colored blue. Interactions significant for both attributes are colored orange.**

**Below this is a color code bar which shows when interactions are stronger in the first two datasets (green) or the converse (red). The final panel shows the log2 fold change with statistical significance colored as above.**

# Figure SA8 Genome-wide analysis of statistically significant interactions of the α globin promoters using r3Cseq.

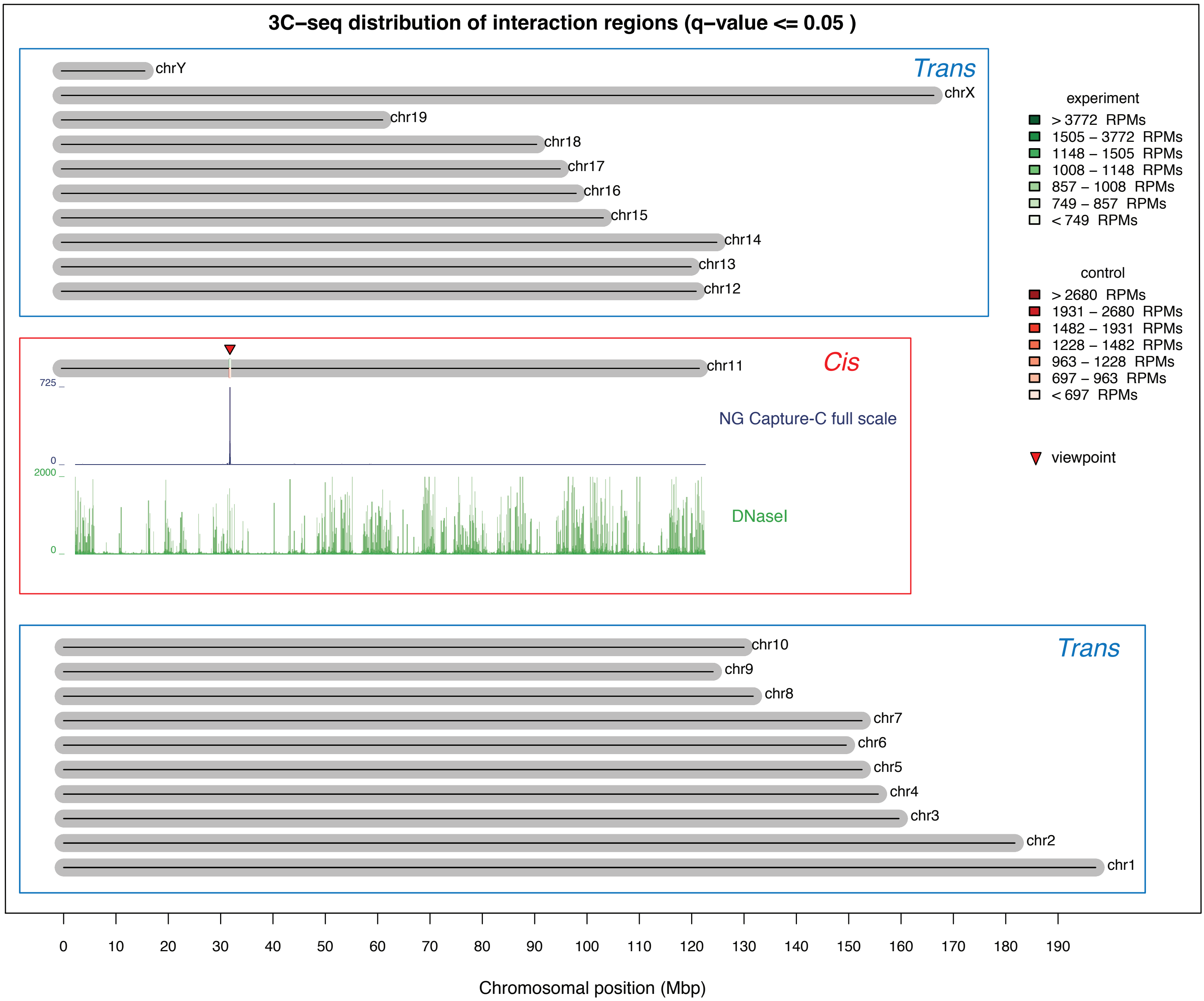

**Figure SA8**

**The output of the genome-wide comparative analysis from the r3Cseq package for the mouse α globin locus in erythroid and ES cells, run on defaults parameters. The distribution of statistically significant interactions are plotted as colored bars on the chromosome ideograms in *trans* (blue boxes) and in *cis* (red box). The keys to color coding of q-values in erythroid ("experiment") and ES cells ("control") are shown to the right of the figure. The distribution of unique interactions from the viewpoint (colored triangle) along the *cis* chromosome is plotted with the distribution of active elements as assessed by DNase-seq. Significant interactions can only be detected in close proximity to the viewpoint even when weighted for distance.**

# Figure SA9 Genome-wide analysis of statistically significant interactions of the β globin promoters using r3Cseq.

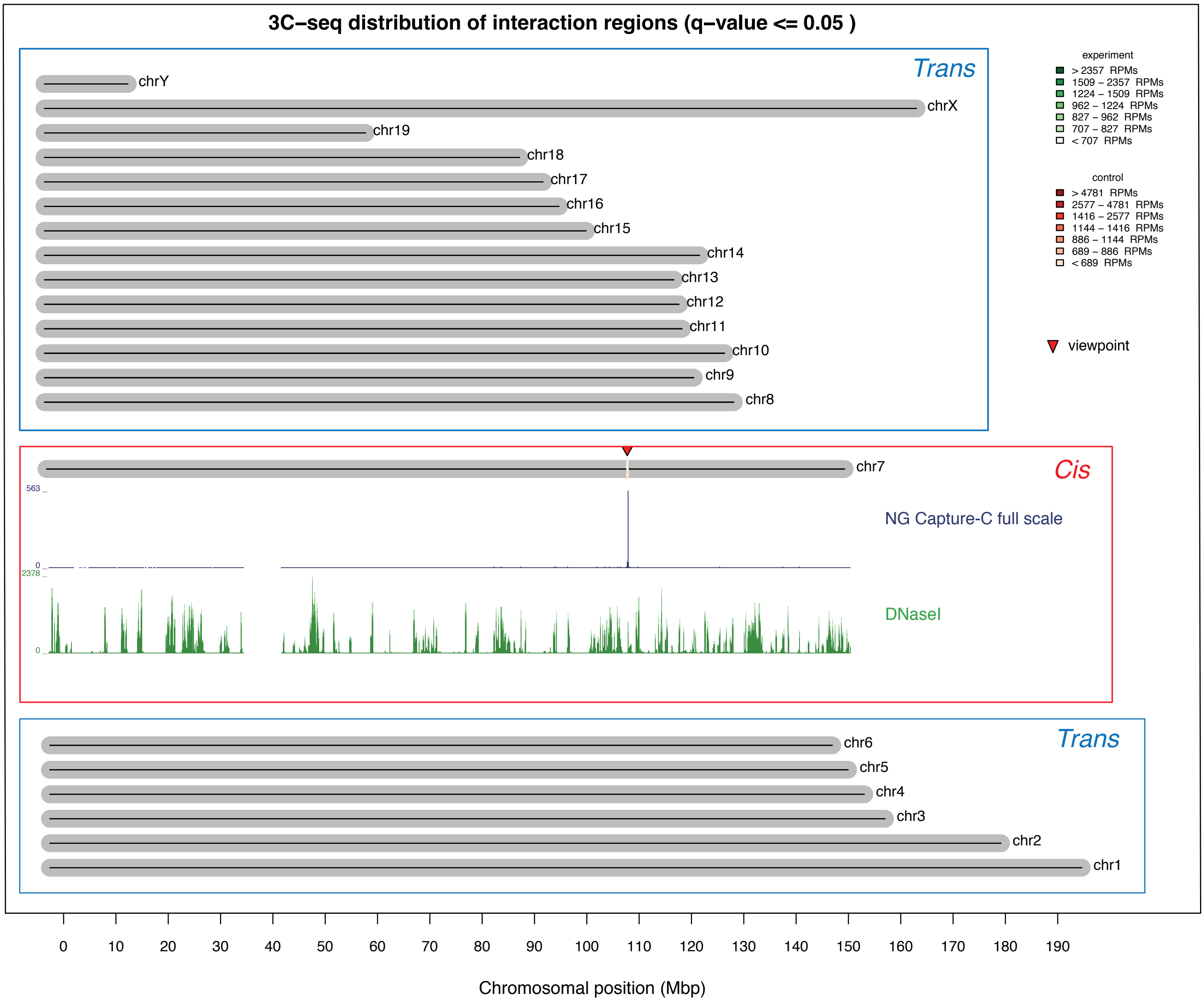

**Figure SA9**
**The output of the genome-wide comparative analysis from the r3Cseq package for the mouse β globin locus in erythroid and ES cells, run on defaults parameters. The distribution of statistically significant interactions are plotted as colored bars on the chromosome ideograms in *trans* (blue boxes) and in *cis* (red box). The keys to color coding of q-values in erythroid ("experiment") and ES cells ("control") are shown to the right of the figure. The distribution of unique interactions from the viewpoint (colored triangle) along the *cis* chromosome is plotted with the distribution of active elements as assessed by DNase-seq. Significant interactions can only be detected in close proximity to the viewpoint even when weighted for distance.**

# Figure SA10 Genome-wide analysis of statistically significant interactions of the *Slc25a37* (***Mitoferrin 1***) promoters using r3Cseq.

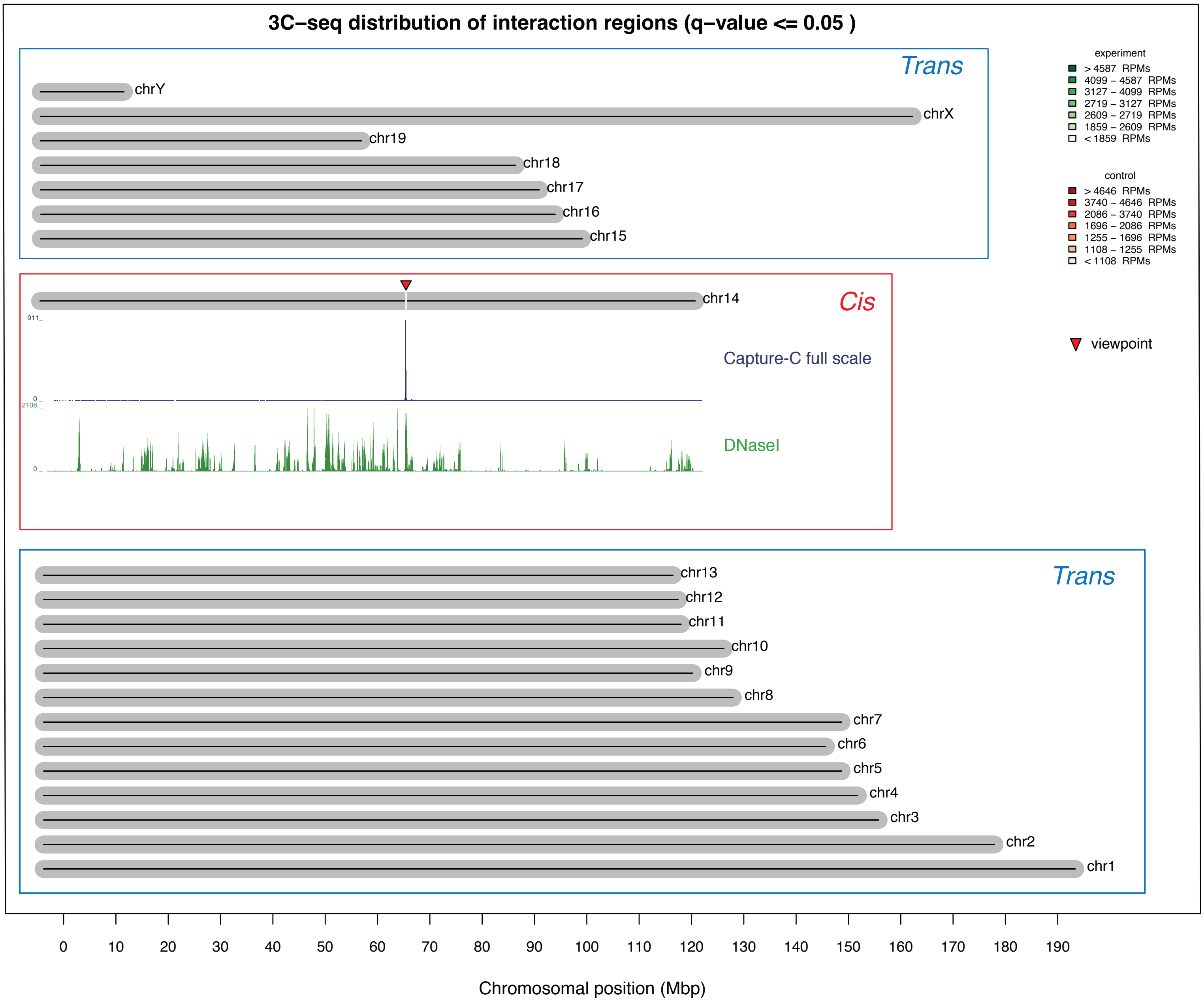

**Figure SA10**

**The output of the genome-wide comparative analysis from the r3Cseq package for the mouse *Slc25a37* (*Mitoferrin 1*) locus in erythroid and ES cells, run on defaults parameters. The distribution of statistically significant interactions are plotted as colored bars on the chromosome ideograms in *trans* (blue boxes) and in *cis* (red box). The keys to color coding of q-values in erythroid ("experiment") and ES cells ("control") are shown to the right of the figure. The distribution of unique interactions from the viewpoint (colored triangle) along the *cis* chromosome is plotted with the distribution of active elements as assessed by DNase-seq. Significant interactions can only be detected in close proximity to the viewpoint even when weighted for distance.**

Figure SA11 Interactions along the *cis* chromosome for α globin, β globin *Slc25a37* (*Mitoferrin 1*) loci in erythroid cells using FourCSeq

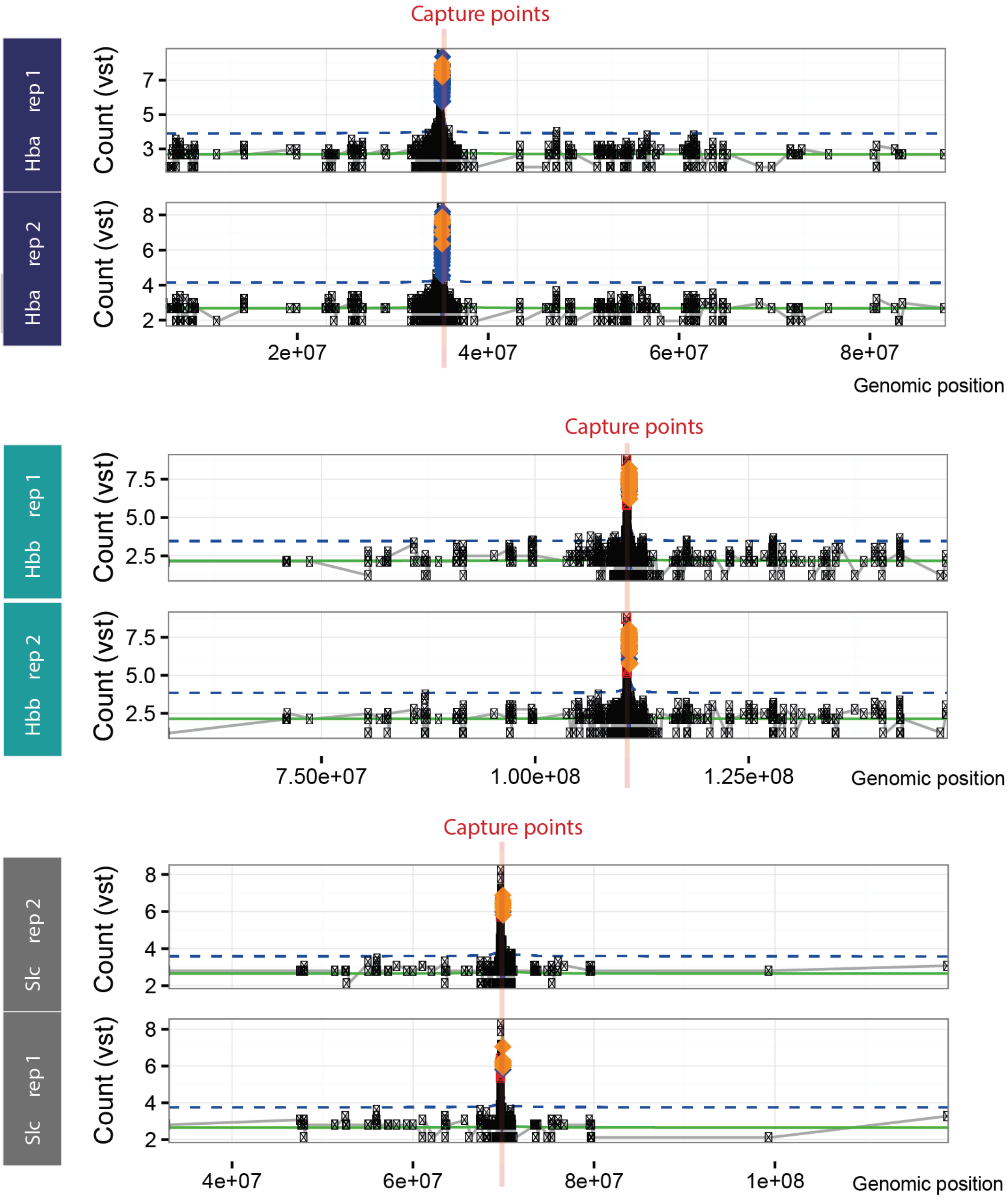

Figure SA11
The output of the comparative analysis of FourCSeq package for the mouse α globin, β globin and *Slc25a37* (*Mitoferrin 1*) loci in erythroid cells, run on default parameters.
The top 2 panels show show the output from two α globin locus erythroid replicates (Hba rep1+rep2). Interactions which show significance after Z-score correction are colored red, and interactions, which show significant fold enrichment between erythroid and ES cells, are colored blue. Interactions significant for both attributes are colored orange. The position of the capture points is shown with red orientation line.
Below this are the corresponding figures for two β globin and *Slc25a37* replicates (Hbb rep1+rep2, Slc rep1+rep2), color scheme as above. The figure demonstrated, that no significant long range interactions with the capture points are seen in FourCSeq analysis.

## Figure SA12  Increases in contact frequency at active and in active regulatory elements

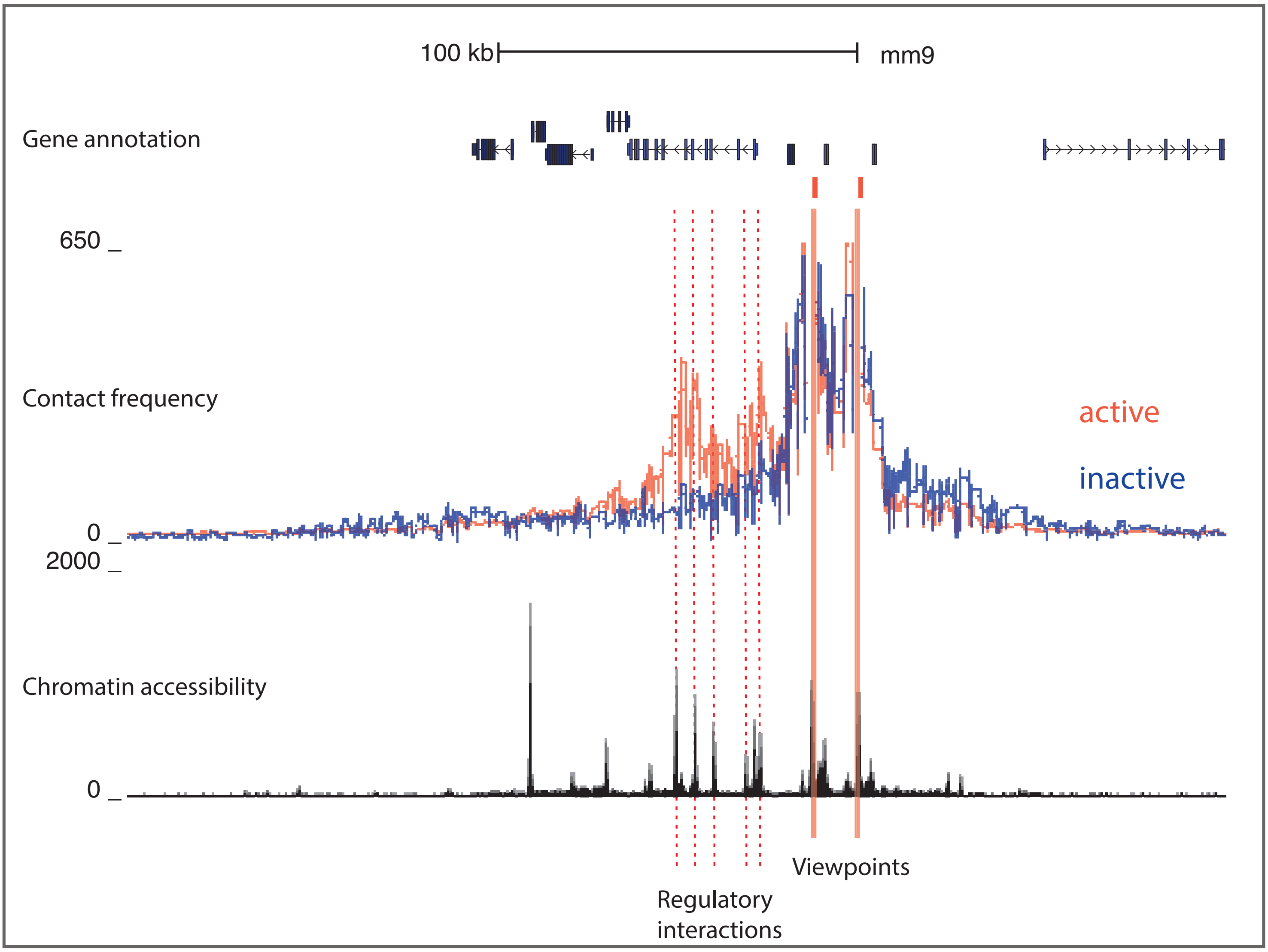

Figure SA12
The increase in normalised contact frequencies associated with regulatory elements (red hatched lines) in an active tissue (red plot) as compared to an inactive tissue (blue plot).   Illustration is based on real data from the $\alpha$ globin promoters (viewpoints shown in red bars) - please see main text.

# Figure SA13 Analysis of NG Capture-C data from the α globin locus using the HiCUP and Gothic Hi-C tools over mouse chr11

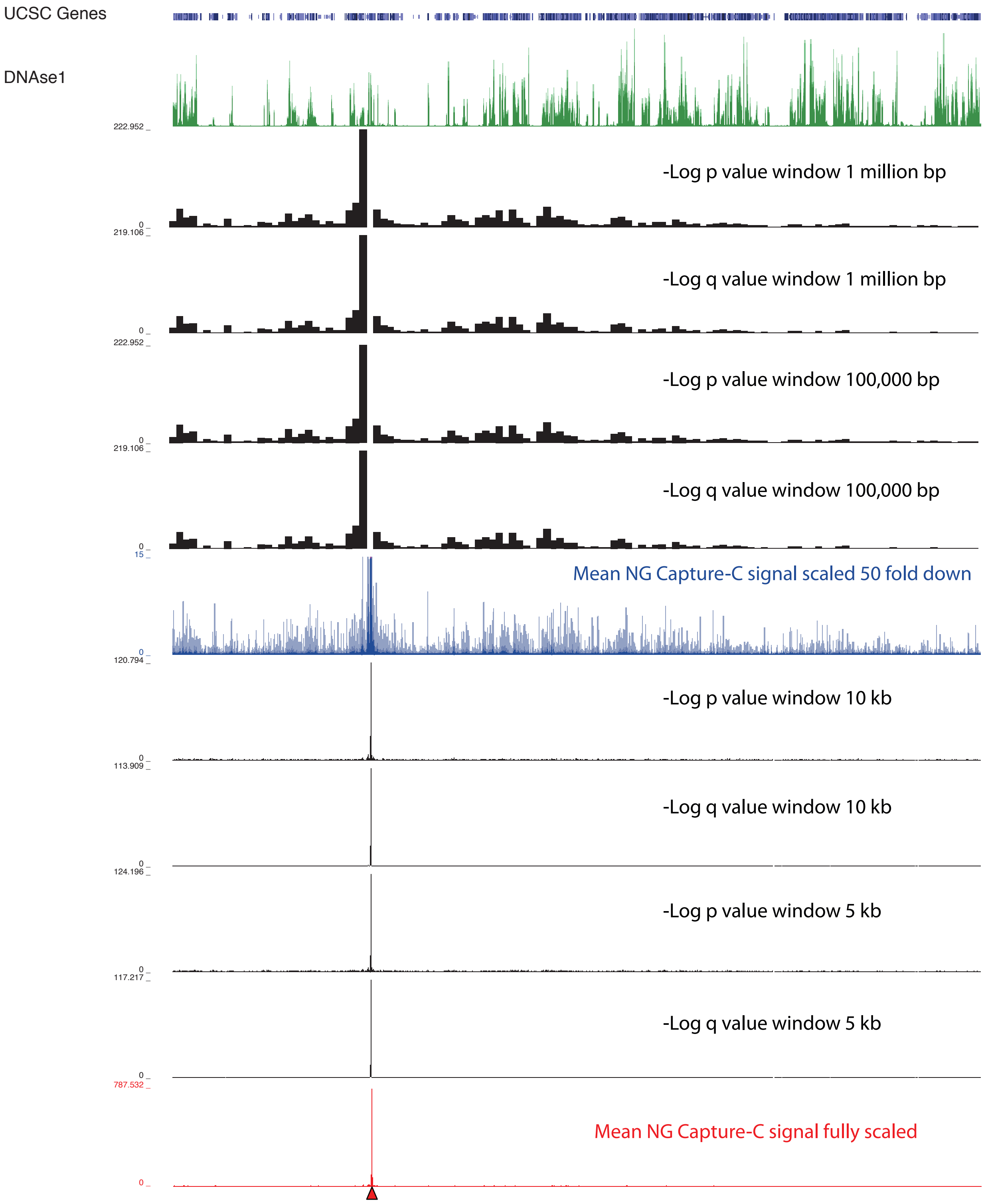

Figure SA13
The output of the genome-wide comparative analysis for chr11 (the *cis*-chromosome) from the GOTHiC Hi-C tool for the mouse α globin locus in erythroid cells.
The GOTHiC tool analysis window decreases from the 1million bp default to 5 kb in incremental steps, plotting the minus Log p- and q-values for each window size, respectively. The top tracks show UCSC genes as violet boxes, and the distribution of active elements as assessed by DNase-seq (in green).
The raw signal (counts per Dpn II-fragment) from the capture experiment (mean of the 3 biological replicates) is shown full scaled (red) and downscaled ~ 50 fold (blue).
The capture points representing the α globin promoters are shown with a red and black triangle.
In large window sizes the p- and q-values can be seen to mirror the low level interaction density (blue graph) and in smaller windows to track the fully scaled density (red graph).

# Figure SA14   Analysis of NG Capture-C data over the $\alpha$ globin locus using GOTHiC

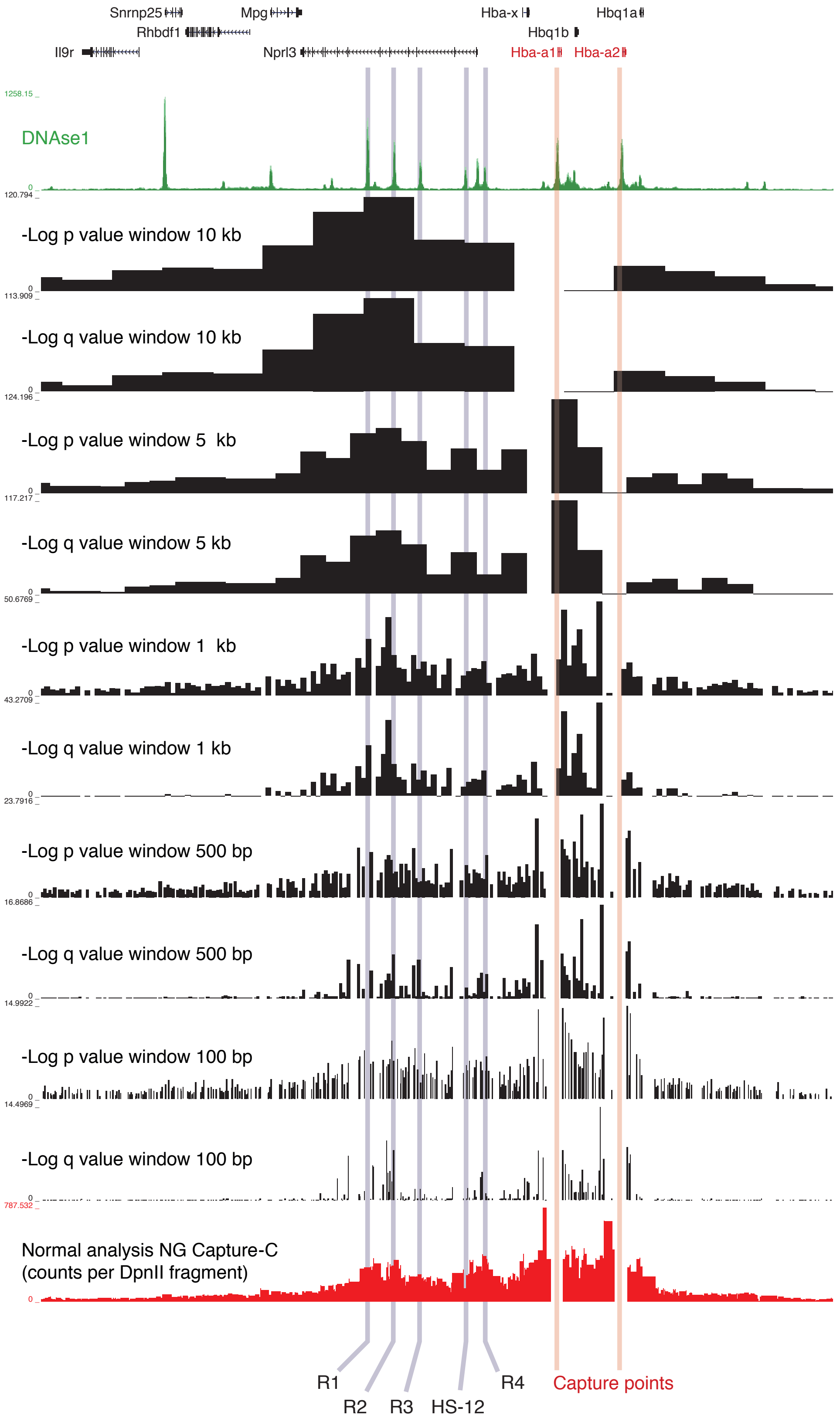

**Figure SA14**

The output of the genome-wide comparative analysis for region close to the capture sites from the GOTHiC Hi-C tool for the mouse $\alpha$ globin locus in erythroid cells.

The top tracks show UCSC genes as red and black boxes, and the distribution of active elements as assessed by DNase-seq (in green).  The raw signal from the capture experiment (mean of the 3 biological replicates) is shown in red.

The position of the capture points and the characterized elements are annotated below the red hatched and grey solid orientation lines, respectively.

The GOTHiC tool analysis window decreases from 10 kb to 100b in incremental steps, plotting the minus Log p- and q-values for each window size respectively. The p- and q-values are seen to follow the capture signal distribution in a window-size dependent manner.