**Methods:**

**Overlapping between Top2b, Ctcf and Rad21 (Slide No.7)**

Overlapping of binding regions of Top2b, Ctcf and Rad21 are determined using bedtools. Peaks overlapping by at least 1 base pair will be merged and annotated by the factors that co-occupy. (For example, if a Ctcf peak and a Top2b peak overlap by at least 1 bp, these two peak regions will be merged and annotated as Ctcf-Top2b) Numbers of different peak overlapping are indicated in the 3-way venn diagram.

For each factor, its original peak regions were extracted and annotated by the overlapping factors. Its binding intensity ( reads per million mapped reads across each peak region, reads are extended to 150bps ) at different peak regions is calculated and plotted as boxplot.

**Conservation of Ctcf sites (Slide No.8)**

Raw fastq files of liverCtcf ChIP-seq data of human, rat and dog was obtained from Schmidt et. al. 2012, E-MTAB-437. Short sequencing reads were aligned to Ensembl 70 genomes (GRch37.68, rnor5.0, cfam3.1). All the mouse data in this study were aligned to mm9. Ctcf sites were then lifted over to mm10 in order to match Ensembl 70. Compara was used to determine evolutionary conservation levels of mouse Ctcf peaks. Peaks unique to mouse are labelled as “mouse only”, peaks shared between mouse and rat are labelled “rodents only”, the rest are labelled as “beyond rodents”.

RepeatMasker (4.0.5) was used to find repeat sequences within Ctcf peak regions. Proportion of CTCF peaks have at least one SINE/B2 elements were calculated. These peaks were extracted and their conservation levels (assessed as previously described) were plotted.

**Triple sites analysis (Violin plot and bar plots, Slide No. 9)**

Regions which are merged by overlapping Ctcf, Rad21 and Top2b peaks represent genomic regions that are co-occupied by the three factors, and are refereed to as “triple sites”. Each triple site was scanned with the Ctcf core motif using RSAT matrix-scan with the command “matrix-scan -v 1 -quick -i -m -matrix\_format transfac -origin start -bginput -markov 1 -2str -uth pval 0.0004 -return pval”. If multiple motifs are found within one peak region, only the motif with the highest weight is kept.

The Ctcf core motif is 20bps and centre of the motif is defined as the 10th bp from 5' end of the motif.

For each Ctcf motif, the genomic distance between Ctcf core motif centre and peak summits of that Ctcf peak and nearest peak summits of Rad21 and Top2b are calculated as “genomic position of the summit – genomic position of the motif centre”. The distribution of these distances were plotting as the left panel of the violin plot. However, the motif has a directionality and can be found on both strands. To take that into account, if the Ctcf core motif is found on the reverse strand (labelled as R), the previously calculated distance will be reversed to truly reflect the positional order of peak summits and motif centre. If the distance is positive, the peak summit is found at the downstream (3') side of the motif centre and vice versa. After correcting for Ctcf motif direction, the distribution of the distances were plotted again as the right panel.

For each peak, the order of Rad21, Top2b peak summits and Ctcf motif centre are listed according to the distances calculated before and after correcting for Ctcf motif direction. For example, in a peak region, the distance of Rad21 peak summit and Top2b peak sumits to Ctcf motif center is -20bp and 10bp respectively, this peak will be annotated as “Rad21, motif, Top2b”. Number of Ctcf peaks with different factor direction annotations were plotted as bar plot.

**Explanation of files in folder:**

﻿**Reads**:

Contains read files for Top2b, Ctcf and Rad21 as indicated in the files names. Reads are stored in bam format. Reads which are not uniquely mappable or fall into ENCODE mm9 blacklist regions are removed.

**Peaks**:

Contains peak files for the corresponding factors. Peaks were called with MACS2 and stored in ENCODE narrowpeak format. MACS2 parameters are “ -q 0,01 --keep-dup all”. Peaks were further filtered with SignalValue >= 5.

**Top2b\_Ctcf\_Rad21\_3way\_overlaps\_ori\_peaks.txt**

Peaks of 3 factors were merged using “bedtools merge”. Merged regions are reported with original information of the all peaks that are merged. An simple annotation of which factors each region contains is also reported

* chr: chromosome
* start: start of the merged region
* stop: end of the merged region
* peak\_info: original peak information
  + information for each peak is separated by “;”
  + factor: chromosome- peak start- peak stop: summit position: signalValue (reported by MACS2): qvalue (reported by MACS2)
* overlapping\_factors: factors that co-occupy this region

**Ctcf\_peak\_anno\_with\_orientation\_conservation.txt**

Contains information for each Ctcf peak identified in the experiment.

* Chr: chromosome
* peak\_start: start of the Ctcf peak
* peak\_end: stop of the Ctcf peak
* peak\_info:
  + the original information of the Ctcf peak, same format as in the other file.
  + factor: chromosome- peak start- peak stop: summit position: signalValue (reported by MACS2): qvalue (reported by MACS2)
* Ctcf\_motifs:
  + contains Ctcf core motif scan results of that peak region. Found motifs are separated by “;” if multiple motifs are found in one peak region
  + For each found motif: chromosome- start of the motif – stop of the motif – center of the motif – direction of the motif – weight (as reported by RSAT matrix-scan) - (-log10(pvalue)) of the motif (as reported by RSAT matrix-scan “sig” column)
* Ctcf\_motif\_best:
  + Format is same as Ctcf\_motifs column, only the motif with the highest weight is kept in this column and used for other analysis.
* Composition:
  + What factors co-occupy this Ctcf peak region
* factors\_order:
  + For “triple sites” (Ctcf peaks co-occupied by Top2b and Rad21), how Top2b and Rad21 peak summits (as reported by MACS2) and Ctcf motif center is ordered ( 5'-3' direction; Ctcf motif direction is taken into the account).
* Conservation:
  + Evolutionary conservation level of that Ctcf peak as defined by compara. For digits represents: dog, human, mouse, rat. 1 means the mouse peak is conserved in the corresponding species while 0 means the peak is not conserved. X means the orthologous regions is not found in the Compara multi-species alignment. All the X were treated as 0 in later analysis.

**Ctcf\_meme\_motif1.transfac:**

Ctcf core motif which was used for matrix-scan.