Methods

ChIP-exo data processing

Raw ZNF ChIP-exo reads were obtained from GEO accession GSE78099[1].

The read trimming, genome browser track generation, and other processing scripts used can be found at `https://github.com/georgimarinov/GeorgiScripts`

Reads were aligned against the `hg38` genome as 1×36mers using Bowtie[2] (version 1.0.1) with the following settings: `-v 2 --k 2 -m 1 -t` –best— `--strata` i.e. only retaining uniquely mapping reads.

Strand-specific forward- and reverse- 5'-end read count tracks were then generated using the `make5primeWigglefromBAM-NH.py` script with default settings plus the addition of the `-stranded +` or `-stranded -` options, then converted to `bigWig` files using the `wigToBigWig` program in the UCSC Genome Browser Utilities.

Initial peak calling was carried out using version 2.1.0 of the MACS2 peak caller[3] with the following settings, specific to ChIP-exo:

```
macs2 callpeak
-t ChIP-exo.36mers.unique.bam
-n ChIP-exo.MACS-2.1.0
-g hs -f BAM --keep-dup all --shift -75
--extsize 150 --nomodel &
```

To identify robustly reproducible peaks, the IDR framework[4] was then applied, as follows.

Where replicates were available, IDR was run on individual replicates and on pooled pseudoreplicates (generated by randomly dividing in two a merged BAM file of both replicates), as previously described[6?] . For most other datasets, aligned reads were split into individual pseudoreplicates, generated from the BAM file for the single available replicate.

Relaxed peak calls for input into IDR were then generated as follows for the merged files, for the individual replicates and for the pooled and individual pseudoreplicates:

```
macs2 callpeak
-t ChIP-exo.36mers.unique.bam
-n ChIP-exo.MACS-2.1.0.p1e-1
-g hs -f BAM --keep-dup all --shift -75
--extsize 150 --nomodel --to-large -p 1e-1
```

The top 300,000 peaks were retained as follows while filtering out peaks on the mitochondrial chromosome:

```
cat ChIP-exo.MACS-2.1.0.p1e-1_peaks.narrowPeak
| egrep -v chrM | sort -k 8nr,8nr | head -300000
| awk 'BEGIN{OFS="\t"}{$4="Peak_"NR ; print $0}'
| gzip -c >
ChIP-exo.MACS-2.1.0.p1e-1_peaks.sorted.gz
```

Version 2.0.4 of the IDR package (`https://github.com/kundajelab/idr`) was then run as follows:

```
idr --samples
pseudoRep1.peaks.sorted.gz
pseudoRep1.peaks.sorted.gz
--peak-list ChIP-exo.peaks.sorted.gz
--input-file-type narrowPeak --output-file
ChIP-exo.ind-pseudoReps.IDR
--rank p.value --soft-idr-threshold 0.05
--plot --use-best-multisummit-IDR
```

Peaks above an IDR value of 0.05 were then filtered out. The remaining set was intersected against the ENCODE[7] set of "blacklisted" regions[8] for `hg38` to remove likely artifacts, and to arrive at a final set of peaks.

References

1. Imbeault M, Helleboid PY, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**(7646):550–554
2. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3):R25.
3. Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**(9):1728–1740.
4. Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5**:1752–1779.
5. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**(9):1813–1831.
6. Marinov GK. 2017. Identification of Candidate Functional Elements in the Genome from ChIP-seq Data. *Methods Mol Biol* **1543**:19–43.
7. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, Halow J, Van Nostrand EL, Freese P, Gorkin DU, Shen Y, He Y, Mackiewicz M, Pauli-Behn F, Williams BA, Mortazavi A, Keller CA, Zhang XO, Elhajjajy SI, Huey J, Dickel DE, Snetkova V, Wei X, Wang X, Rivera-Mulia JC, Rozowsky J, Zhang J, Chhetri SB, Zhang J, Victorsen A, White KP, Visel A, Yeo GW, Burge CB, Lcuyer

E, Gilbert DM, Dekker J, Rinn J, Mendenhall EM, Ecker JR, Kellis M, Klein RJ, Noble WS, Kundaje A, Guig R, Farnham PJ, Cherry JM, Myers RM, Ren B, Graveley BR, Gerstein MB, Pennacchio LA, Snyder MP, Bernstein BE, Wold B, Hardison RC, Gingeras TR, Stamatoyannopoulos JA, Weng Z. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**(7818):699–710.

8. Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**(1):9354.