HOT or not: redefining the origin of high-occupancy target regions

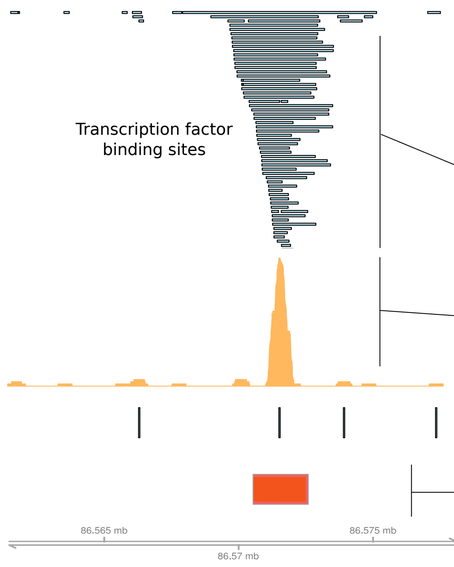
Katarzyna Wreczycka, Vedran Franke, Bora Uyar, Ricardo Wurmus, Altuna Akalin

Max-Delbrück-Center for Molecular Medicine, BIMSB, Berlin-Buch, Germany

High-occupancy target (HOT) regions are segments of the genome with unusual enrichment of transcription factor binding sites. These regions are observed in multiple species and thought to have biological importance due to high density of transcription factor occupancy. Furthermore, they coincide with house-keeping gene promoters and the associated genes are stably expressed across multiple cell types. Despite these features, HOT regions are solemnly defined using ChIP-seq experiments and shown to lack canonical motifs for transcription factors that are thought to be bound there. Although, ChIP-seq experiments are the golden standard for finding genome-wide binding sites of a protein, they are not noise free. Here, we show that these regions are likely to be ChIP-seq artifacts and they are similar to previously proposed `hyper-ChIPpable regions`. Using ChIP-seq data sets for knocked-out transcription factors, we demonstrate enrichment of the knocked-out factors on HOT regions. We observe sequence characteristics and genomic features that are unique to HOT regions that are in turn statistically associated with the artificial ChIP-seq enrichment. Furthermore, we propose strategies to deal with such artifacts for future ChIP-seq studies.

Wreczycka, Franke et al. bioRxiv 2017 http://biorxiv.org/content/early/2017/03/05/107680

Definition of HOT regions



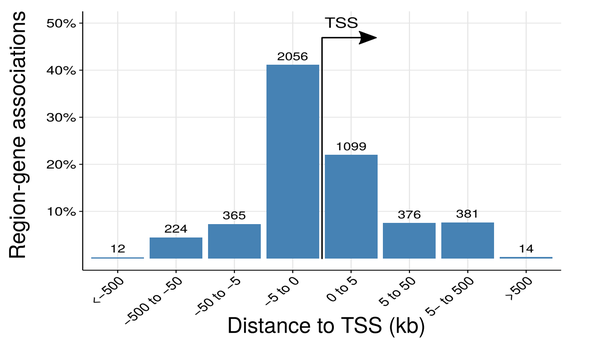
HOT regions are highly occupied by transcription factor binding sites[1].

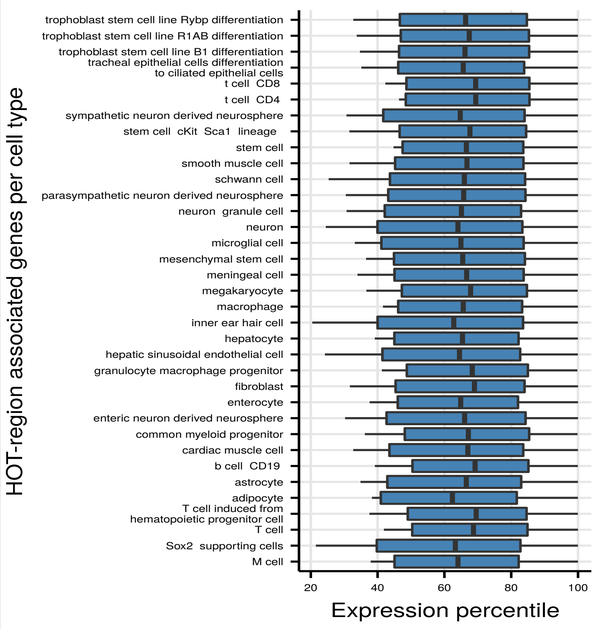
1. Overlap transcription factors (TFs) ChIP-seq peaks onto genome
2. Smooth density of TF occupancy
3. Define local maxima
4. Find maxima with coverage >=99th percentile and extend them to 2kb around centers

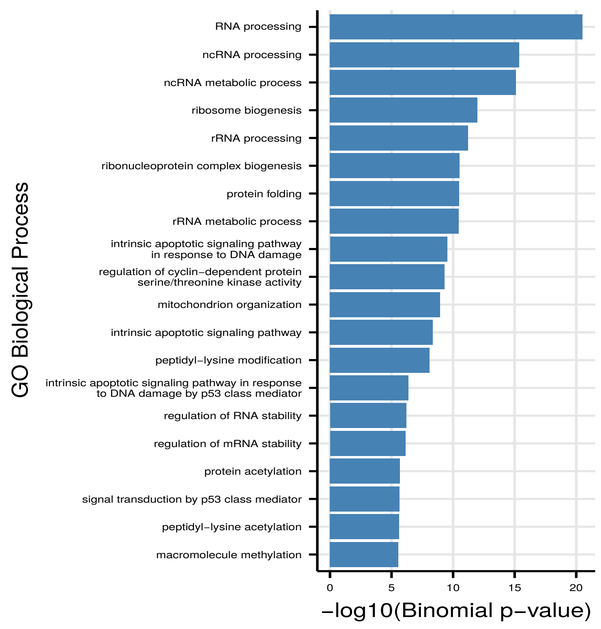
HOT regions are enriched in promoters of house-keeping genes

The majority of HOT regions (80%) are in close proximity to TSS (within 5 kb) of genes.

Genes associated with HOT regions are stably expressed across 35 cell lines.



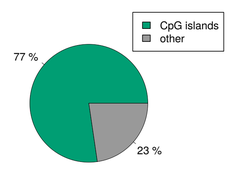


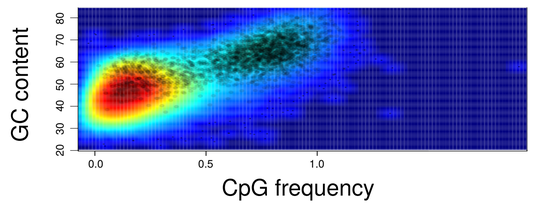


The Gene Ontology (GO) revealed that genes that overlap with HOT regions are enriched with housekeeping GO terms and pathways.

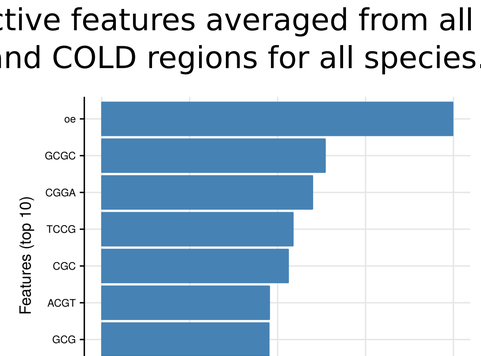
Sequence characteristics of HOT regions

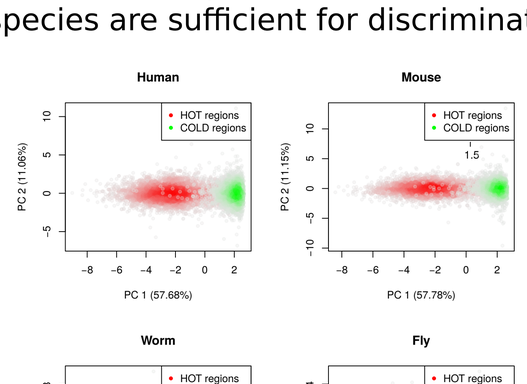
HOT regions are G+C rich and have high CpG frequency.





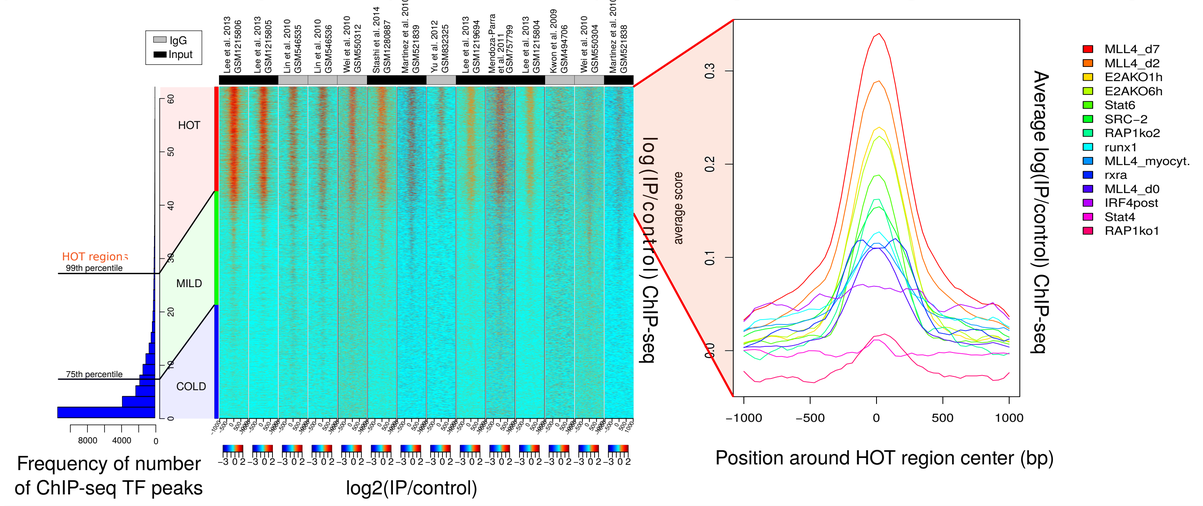
Statistical analysis using Elastic Net (regularized regression method) analysis of human, mouse, fly and worm HOT regions of observed/expected ratios for CpG occurence, 2,3,4 k-mers, CpG frequency and GC skew. HOT regions share common low-level sequence features across species. The most predictive features averaged from all species are sufficient for discriminating HOT and COLD regions for all species.



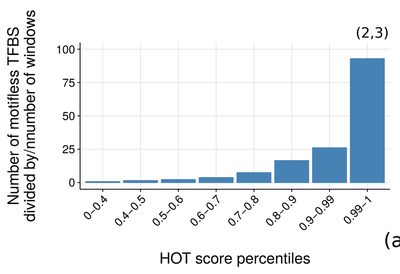


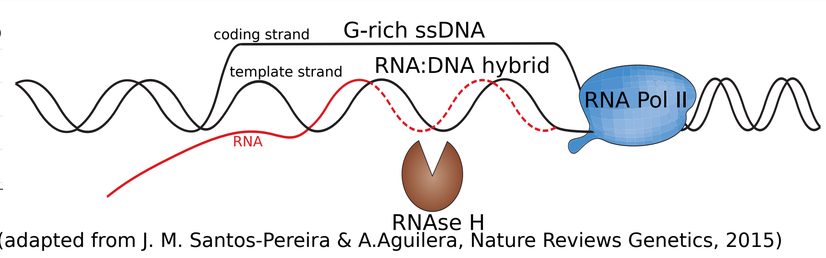
False positive signal of knock-out ChIP-seq experiments on HOT regions

14 out of 22 publicly available ChIP-seq experiments with knock-out of the genes that encodes target proteins show enrichment even though the chipped protein shouldn't be present in the analysed sample. Such false positive signal is the highest on HOT regions.



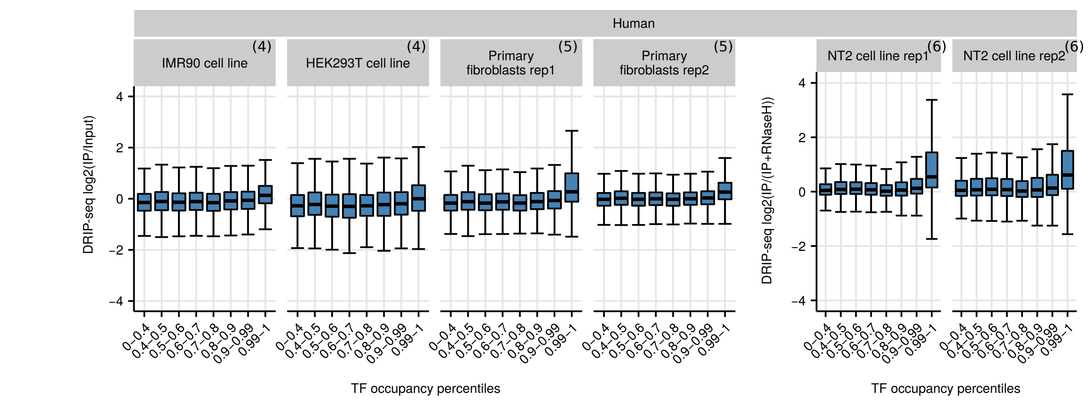
Association of HOT region with R-loops

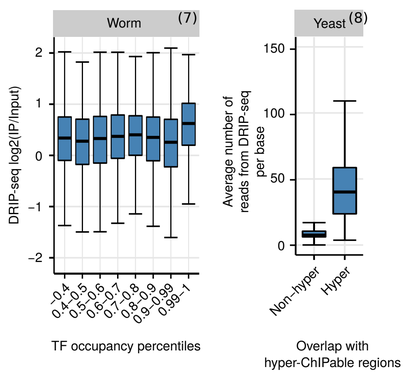




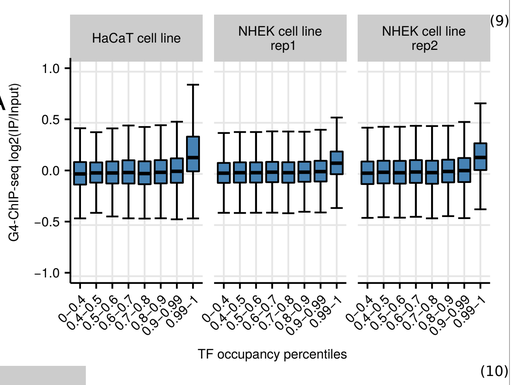
(adapted from J. M. Santos-Pereira & A.Aguilera, Nature Reviews Genetics, 2015)

R-loops are RNA-DNA hybrids and the displaced single-stranded DNA (ssDNA) and can be detected using DRIP-seq. HOT regions show high enrichment of DRIP-seq signal in multiple organisms (human, worm and yeast) and disappear after RNAse-H treatment.

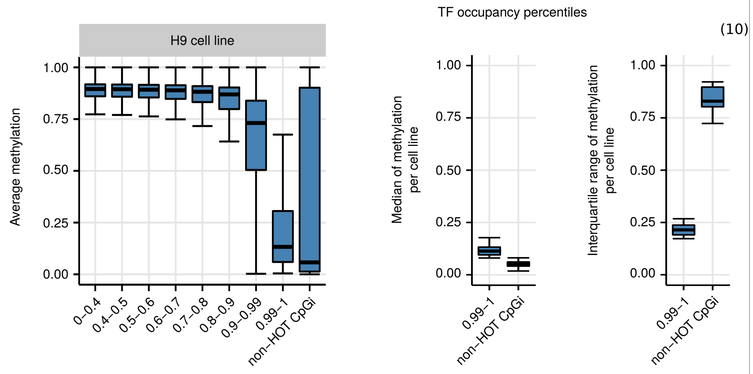




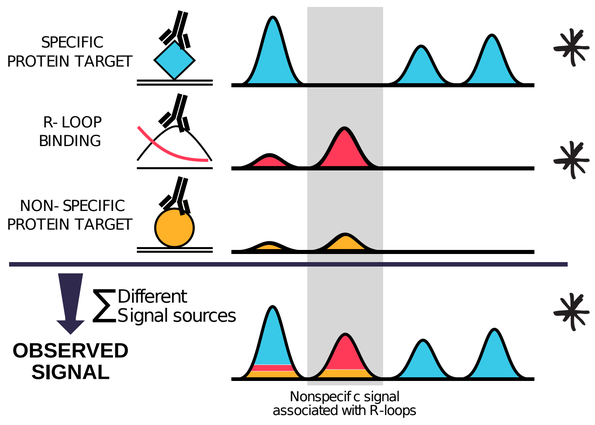
HOT regions are associated with DNA tertiary structures (G-quadruplex DNA)



R loops prevent methylation on HOT regions.



Conclusion and future directions



- HOT region are most likely technical artifacts of ChIP-seq
- We recommend to mask HOT regions for ChIP-seq data analysis or to use KO ChIP-seq as a control
- Including RNAse-H in ChIP-seq protocol prevents R-loops formation and thereby might remove some of HO regions

SYSTEMS BIOLOGY

MDC MAX DELBRÜCK CENTER FOR MOLECULAR MEDICINE IN THE HELMHOLTZ ASSOCIATION

BIMSB THE BERLIN INSTITUTE FOR MEDICAL SYSTEMS BIOLOGY

References:
1) Xie et al. Cell 2013
2) Factorbook, Wang et al. Genome Research 2012
3) ENCODE, Rosenbloom et al. Nucleic Acids Res. 2015
4) Nadel et al. Epigenetics & Chromatin 2015
5) Lim et al. eLife 2015
6) Sanz et al. Mol Cell 2016
7) Zeller et al. Nat Genetics 2016
8) Teytelman et al. Proc. Natl. Acad. Sci. U.S.A 2013
9) Hänsel-Hertsch et al. Nat Genetics 2016
10) RoadmapEpigenomics, Bernstein et al. Nat Biotechnol 2013