HOT or not: examining the basis of high-occupancy target regions

Katarzyna Wreczycka[1],[†], Vedran Franke[1],[†], Bora Uyar[1], Ricardo Wurmus[1], Selman Bulut[2], Baris Tursun[2] and Altuna Akalin [1],*

[1]Bioinformatics and Omics Data Science Platform, Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin, Germany and [2]Gene Regulation and Cell Fate Decision in *C. elegans*, Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin, Germany

ABSTRACT

High-occupancy target (HOT) regions are segments of the genome with unusually high number of transcription factor binding sites. These regions are observed in multiple species and thought to have biological importance due to high transcription factor occupancy. Furthermore, they coincide with housekeeping gene promoters and consequently associated genes are stably expressed across multiple cell types. Despite these features, HOT regions are solely 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 HOT regions are likely to be ChIP-seq artifacts and they are similar to previously proposed 'hyper-ChIPable' regions. Using ChIP-seq data sets for knocked-out transcription factors, we demonstrate presence of false positive signals on HOT regions. We observe sequence characteristics and genomic features that are discriminatory of HOT regions, such as GC/CpG-rich k-mers, enrichment of RNA–DNA hybrids (R-loops) and DNA tertiary structures (G-quadruplex DNA). The artificial ChIP-seq enrichment on HOT regions could be associated to these discriminatory features. Furthermore, we propose strategies to deal with such artifacts for the future ChIP-seq studies.

INTRODUCTION

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is now a standard method to quantitatively assay the binding sites of a DNA binding protein in the genome. Large scale projects such as ENCODE (1) and modENCODE (2) used this technology to find the binding sites of hundreds of proteins in multiple species. With more binding site data available, it has become apparent that certain parts of the genome harbour high frequency of protein-DNA binding events. These regions are called high-occupancy target (HOT) regions and they are observed in multiple species (3,4). HOT regions are associated with housekeeping genes and are enriched in binding events without canonical motifs (5). HOT regions are thought to have biological importance due to high number of binding sites observed, but previous reports failed to assign a clearly distinctive function that would explain the requirement for the exuberant number of bound transcription factors.

In this study, we aim to gain a deeper understanding of the nature of HOT regions and the genomic features associated to them. First, we wanted to investigate the features that are common to HOT regions across species. To date, there has been no cross-species comparison of HOT regions in terms of sequence features. The sequence features that are shared across species can provide a mechanistic insight into HOT region formation, and enable prediction of HOT regions in other species. With the sequence analysis and subsequent integrative analysis, we primarily aim to uncover the rationale behind the propensity of HOT regions to have unusual number of binding events, many of which are motifless binding events (transcription factors binding to a region without the known motif) (5). For us, the plausible explanations for motifless binding are a combination of 1) interaction of transcriptions factors (TFs) where only a handful of them are actually binding to DNA 2) existence of weak binding sites where TFs bind to non-canonical motifs in a weak manner 3) regions with high-affinity for chromatin immunoprecipitation called 'hyper-ChIPable' regions (7). Many of the HOT regions are shown to bind hundreds of proteins based on ChIP-seq experiments (4). Detection of hundreds of proteins occupying an individual HOT region could be explained by extensive protein interaction networks between transcription factors and cofactors, where

only a few factors directly bind to DNA. However, only a handful of such interactions were experimentally validated (3). Therefore, we seek additional explanations for existence of HOT regions in the genome and their association with motifless binding.

For a better understanding of what creates the HOT regions, we investigated nucleotide sequence features (motifs, *k*-mer content, etc.) of HOT regions across species. We built species-specific machine-learning models to learn discriminative sequence features for HOT regions. We showed that HOT regions are associated with certain sequence features that are shared across species. In order to investigate the potential technical biases causing the occurrence of HOT regions, we analyzed ChIP-seq experiments for knocked-out transcription factors in mice. Previously, Teytelman et al. (7) showed that highly expressed loci in yeast give rise to false-positive peaks when they did ChIP-seq experiments for proteins that did not have the corresponding gene in the genome. We set out to examine if such a technical bias could be the driving force behind HOT regions given that motif-less binding is prominent on those regions. As a result of this analysis, we observed false positive signals on HOT regions. Finally, we investigated the association of HOT regions with RNA:DNA hybrids called R-loops. The two classes of regions share similarity in sequence features, and are associated with gene promoters and open chromatin regions (8). We demonstrated association of HOT regions with R-loops. This paper presents a new rationale that explains the apparent high-occupancy of TFs for at least some of the HOT regions. With a better understanding of HOT regions provided here and other potential sources of bias associated with false-positive ChIP-seq peaks, now researchers can avoid these pitfalls and obtain less noisy data by additional computational analysis.

MATERIALS AND METHODS

ChIP-seq data for definition of HOT regions

Analyses of human TF binding sites were performed using the UCSC human hg19 reference genome, mouse mm9, *Drosophila melanogaster* dm3 and *Caenorhabditis elegans* ce10. ChIP-seq files in narrowPeak format were downloaded from the ENCODE (www.encodeproject.org) and modENCODE (data.modencode.org) portals. Human TF binding sites in narrowPeak format were downloaded from the UCSC Uniform track http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/. One hundred sixty-six human TFs, 42 murine, 42 *D. melanogaster* and 83 *C. elegans* were obtained.

Defining HOT regions

For a given set of ChIP-seq peaks per species, we determined the summits of the peaks. Following that, we calculated the density of the summits over the genome using 500 bp sliding windows. We calculated the local maxima of the density vector for each chromosome. We made sure local maxima of the density vector are the only maxima in 2000 bp surrounding of the maxima for human and 1000 bp for other species. This is necessary to remove sub-optimal maxima around the real maxima. 2000 bp threshold was specifically applied for human datasets due to high number of experiments creating multiple local maxima around the real maxima. We then ranked these maxima based on the density scores, which is effectively the number of overlapping ChIP-seq peaks and represents the TF occupancy. These density scores are referred to as TF occupancy throughout the text. We used 99th percentile threshold to define the HOT regions. This is in line with previous methods (9). HOT regions were called using only the regulatory peak sets (no RNA polymerase datasets were included). The regions that are not selected as HOT regions are binned according to their TF occupancy percentiles (number of ChIP-seq peak counts) and used as controls in follow-up analyses. Scripts for all analysis are publicly available at https://github.com/BIMSBbioinfo/HOT-or-not-examining-the-basis-of-high-occupancy-target-regions.

Assigning HOT regions to genes, expression and open chromatin analysis

Distance from HOT regions to the nearest transcription start sites was analysed using GREAT (10). Expression values of genes associated with HOT regions across tissues were obtained from the Expression Atlas EBI database (www.ebi.ac.uk/gxa) and fantom5 CAGE expression (11) (Supplementary Figure S2A) and from the RoadmapEpigenomics (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz, Figure 1D). Accessible sites were obtained from the ENCODE DNAse-seq K562 cell line (ENCFF248FIZ).

Sequence analyses of HOT regions

Extraction of 2-, 3- and 4-mers, CpG frequencies (sum of observed G and C divided by length of genomic region equal to 2000), GC skew, the observed/expected ratios for CpG on HOT regions was computed using scripts written in R version 3.3.1 and BSgenome package. We used the following genome assemblies: mm9, hg19, ce10, dm3. The observed/expected ratios for CpG were calculated according to the formula $(O/E)_{CpG} = [f(CG)/f(C)f(G)] \times$ width(genomic_region), where f denotes the observed frequency of the given mono- or di-nucleotides. *De novo* motifs were found using R package motifRG (12).

MotifRG is a discriminative motif analysis tool which searches for overrepresented motifs in a positive set when compared to a negative set. Sequences from HOT regions were used as the positive set, while the negative set was constructed by sampling equal number of sequences as there are HOT regions from non-HOT regions. Motif analysis was performed on separately on HOT regions from *Homo sapiens*, *Mus musculus*, *D. melanogaster* and *C. elegans*. MotifRG was run with the following parameters: start.width =
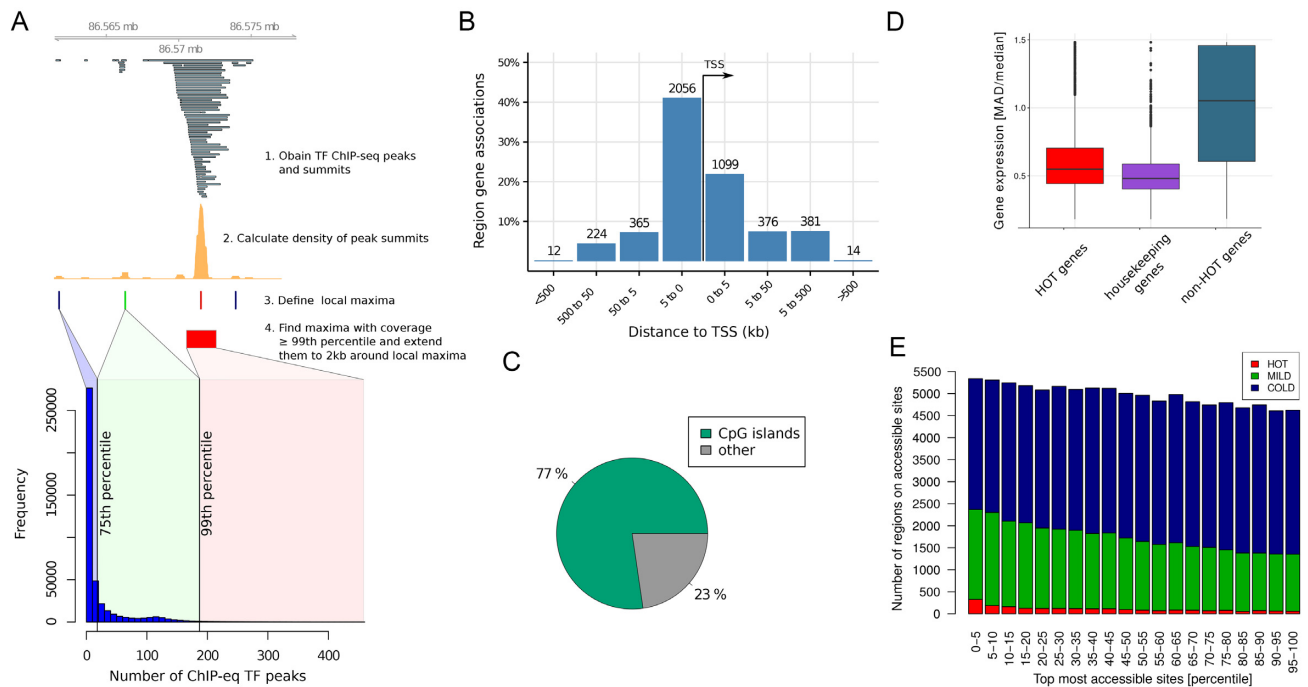
Figure 1. Features of HOT regions. (**A**) Schematic workflow of HOT region definition. The barplot indicates number of ChIP-seq peaks in HOT (red), MILD (green) and COLD (blue) regions. (**B**) HOT regions are located mostly close to transcription start sites and are promoter associated. The figure shows binned orientation and distance between HOT regions and the nearest genes. Associations precisely at 0 refers to the transcription start site of the nearest gene. (**C**) Most HOT regions overlap with CpG islands. (**D**) Gene expression variation on HOT regions. Variation of expression of genes associated with HOT regions is as low as housekeeping genes, and expression is less variable than non-HOT genes. Median absolute deviation (MAD) and median was calculated for each gene across 57 human cell lines and tissues from the Roadmap Epigenomics database. (**E**) Chromatin accessibility on HOT regions. DNase-seq peak set from K562 cell line was ranked according to their signal value. On X axis are percentiles of DNA-seq peaks according to their ranks, on Y axis percent of HOT regions that overlap DNA-seq peaks.

6, both.strand = TRUE, mask = TRUE, enriched.only = TRUE.

Elastic net construction and PCA for discrimination of HOT regions

For training, we used HOT regions defined as regions with TF occupancy percentiles higher than 0.995 for hg19 and 0.99 for other organisms. In order to use as a control, regions with TF occupancy lower than 85th percentile, were sampled matching the number of selected HOT regions. HOT and control regions for mm9 and hg19 were CpG sampled, in order to ensure that the ratio of HOT and control regions that overlap CpG islands were the same. The genomic coordinates of the CpG islands were downloaded from the UCSC Table browser (cpgIslandsExt table). All models had the following set of features: CpG frequencies, ratio of observed versus expected CpGs, GC skew, and 2-, 3-, 4-mers. Feature matrix was standardized prior to training. Models trained and tested for the same species were trained using 10-fold cross-validation. Variable importance scores were calculated for each species-specific model as an absolute value of the model coefficients, which were then normalized to a scale from 0 to 100. Average relative importance was calculated as an average of variable importance scores of all models. Area under the ROC curve (AUC) was computed to measure the accuracy of the models. We used elastic net function from the glmnet R package ([13]). For the

PCA, we used top 10 features ranked by the average relative importance. Using these same features for all species, we calculated PCA and plotted the the color coded scatter plot on principal components for each species. For illustration purposes, we sampled the same number of 'COLD' regions as the number of 'HOT' regions.

Data processing and visualization of KO ChIP-seq, DRIP/RDIP-seq and G4 ChIP–seq samples

Fastq files of KO ChIP-seq experiments (See Supplementary Table S1) were downloaded from the European Nucleotide Archive database (ENA). All fastq files have single-end reads and were uniquely mapped into murine genome version mm9 using Bowtie 1.1.12 with parameters: -p 3 -S -k 1 -m 1 –tryhard -I 50 -X 650 –best –strata –chunkmbs 1000. The bbduk program from the BBMap software 35.14 was used for adapter, quality trimming and filtering with parameters: minlength = 20 qtrim = r trimq = 20 ktrim = r k = 25 mink = 11 ref = 'bbmap/resources/truseq.fa.gz' hdist = 1. Two of KO ChIP-seq samples NFAT1_P+I and NFAT1_None did not pass bbduk tests and were excluded from the further analysis. The FastQC 0.11.3 program was used for quality control. Conversion from SAM to BAM file format, sorting and indexing BAM files was done using samtools 0.1.19, conversion from BAM to BED file formats and then BED to BedGraph file formats using Bedtools-2.17.0, from BedGraph to BigWig file format using Bed-

GraphToBigWig v4. The same pipeline was used for DRIP-seq and RDIP-seq samples, and G4 ChIP-seq (due to lack of detected adapters, bbduk argument 'ref' that indicates a path to adapters was omitted).

The R package genomation ([14](#)) was used for calculating fold enrichment of KO ChIP-seq samples and plotting heatmaps. Fold enrichment of KO ChIP-seq samples was defined as $\log_2$ of IP signal divided by control per base pair. Out of total 25 KO ChIP-seq samples, 15 are positively and 9 are negatively associated with TF occupancy scores. Heatmaps were binned on x-axis into 50 bins, average for each bin was taken, and winsorized to limit extreme values <0.5 and >0.99 percentile. Some of KO ChIP-seq samples are conditional knockouts using Cre-lox recombination system (See Supplementary Table S1). Enrichment presented as boxplots of KO ChIP-seq, DRIP/RDIP-seq, G4-ChIP-seq samples (Figure [3](#)B and [4](#)A, B, D) was calculated as the logarithm base 2 of the number of reads from IP sample overlapping HOT regions (normalized for library size and multiplied by counts per million) divided by the number of the reads from the control sample overlapping HOT regions normalized in the same way. If control was not available then IP with RNaseH treatment was treated as control. For visualisation purposes, windows on heatmaps were sampled: 3000 windows from 0 to 75th TF occupancy percentile, 3000 from 75th to 99th, and 3000 from 99th to 100th percentile.

IgG samples corresponding to the following antibody ENCAB000AOJ were downloaded from ENCODE. Samples marked with 'extremely low read depth' were removed from the analysis. Samples which belong to the same biosample term id were pooled together. Signal was visualized in a region of ±1 kb around HOT (>99th percentile), MILD (between 99th and 75th percentile), and COLD regions (below 75th percentile). Prior to visualization, the reads were extended to 200 bp in a stranded fashion, and the signal was normalized to per million reads.

Methylation dynamics for HOT regions

Methylation over the regions of interest is extracted from the Roadmap Epigenomics Consortium Whole-genome Bisulfite sequencing data sets ([15](#)). Our regions of interest consist of HOT regions, non-HOT regions (regions with lower TF occupancy), and CpG islands not associated with HOT regions (non-HOT CGI). For each region of interest, we extracted overlapping methylation value for each cell type and calculated the mean methylation value per region. We plotted the distribution of mean methylation values for each set of regions: HOT regions, non-HOT regions (binned into different TF occupancy levels), and non-HOT CGI. For each cell type, we calculated the interquartile range and median methylation values for HOT regions and non-HOT CGI. Next, we plotted the distributions of medians and interquartile ranges across cell types as boxplots to compare the methylation dynamics for HOT regions to non-HOT CGI.

PFAM domains and human protein–protein interactions

Reviewed UniProt ([16](#)) human protein sequences (as of 31 August 2015) were scanned for occurrence of PFAM HMM (both PFAM-A and PFAM-B) models using HMMER3 ([17](#)). The HMM scanning detected 9511 types of PFAM domains in 19 275 proteins.

PFAM ([18](#)) entries for single-stranded DNA-binding domains were collected from the PFAM database by combining all members of the following PFAM clans: OB (for OB-fold domains), KH (for K-homology domains), RRM (for RRM-like domains) and sPC4-like (for the Whirly domain). The collection of these four clans contains 90 different types of PFAM domains.

Human protein–protein interaction data was downloaded from the iRefWeb database ([19](#)). In order to dissect which protein–protein interactions of TFs are direct (or relatively direct physical interactions of proteins), the interactions were filtered for the following criteria: (i) interactor A (uidA) is from taxa:9606 and interactor B (uidB) is from taxa: 9606; (ii) interaction type between uidA and uidB is one of 'MI:0915 (physical association)', 'MI:0407 (direct interaction)', 'MI:0403 (colocalization)', 'MI:0914 (association)', or 'MI:0191 (aggregation)'; (iii) both uidA and uidB have an ID mapped to UniProt accessions.

RESULTS

HOT regions exist in multiple species and cover transcription start sites (TSS) of stably expressed genes across cell types

HOT regions are observed in multiple species—human ([3](#),[20](#)), *D. melanogaster* ([21](#)), yeast ([22](#)) and *C. elegans* ([20](#),[23](#)). Based on density of ChIP-seq peaks used as a measure of TF occupancy, we defined HOT regions in human, mouse, worm (*C. elegans*) and fly (*D. melanogaster*) (see Figure [1](#)A and Materials and Methods). Our method detected 4324 HOT regions in human, 2638 in mouse, 422 in *C.elegans* and 408 in *D. melanogaster*, out of 428498 regions with at least one peak in human, 245250 in mouse, 40921 in *C. elegans*, and 37853 in *D. melanogaster*. These examined regions along with their TF occupancy percentiles (percentiles from ChIP-seq peak count distribution) as well as HOT regions are accessible via UCSC track hub (https://bimsbstatic.mdc-berlin.de/hubs/akalin/HOTRegions/hub.txt).

Human HOT regions were based on 159 different transcription factors. HOT regions are composed of highly ranked peaks (high ChIP signal) (Supplementary Figure S1A), and most of the regions bind more than half of the TFs (Supplementary Figure S1B). Interestingly, we observe that the better the quality of the antibodies used, the more HOT regions a given TF is found to occupy (Supplementary Figure S1B).

HOT regions are typically located at promoters. The majority of HOT regions (80%) are in close proximity to TSS (within 5 kb) (Figure [1](#)B). In human and mouse, they are mostly associated with CpG islands (Figure [1](#)C). Genes associated with HOT regions are stably expressed across cell types and tissues, with variability similar to housekeeping genes (Figure [1](#)D). Gene expression levels for those genes are generally above the median level of expression for all genes in the respective cell types (Supplementary Figure S2A). Gene Ontology (GO) ([10](#)) analysis revealed a variety of biological processes highly represented in HOT region-associated genes such as RNA processing, ncRNA pro-

cessing, ncRNA metabolic process and ribosome biogenesis (Supplementary Figure S2B), which is in line with the findings reported by Xie et al. (3). Additionally, although we observe a marginal association between HOT regions and chromatin accessibility, chromatin accessibility alone is not sufficient to explain HOT region formation (Figure 1E). Therefore, having knowledge that a region is highly accessible provides no information on whether the region is HOT, likewise, having information that a region is HOT provides no information on how accessible the region is.

To summarize, consistent with the published features of HOT regions (3,4), our findings confirm that genes associated with HOT regions are mostly housekeeping genes—they are required for the maintenance of basic cellular functions and are constitutively expressed.

HOT regions have specific *k*-mer content compared to control regions

We analysed sequence characteristics of HOT regions in human, mouse, *C. elegans* and *D. melanogaster*, since sequence characteristics may be shared across species and that could explain the existence of HOT regions in multiple species. For this purpose, we built machine-learning models that can discriminate HOT regions from non-HOT or so called 'COLD' regions using sequence features. The machine-learning model is primarily used for identifying sequence features that are predictive of HOT regions. We used 2, 3 and 4 bp long *k*-mer frequencies. In addition, we used GC content and CpG observed/expected ratio (O/E ratio). CpG islands are a frequent feature of HOT regions in human and mouse. Although, *C. elegans* and *D. melanogaster* do not have CpG islands, CpG enrichment could be important at least for *C. elegans,* for which HOT regions are enriched for CpG dinucleotides (20). We built a predictive model of 'hotness' of genomic regions using a penalized multivariate regression method (24). We built four species-specific models using normalized feature matrices as inputs. We had high accuracy for all models: cross-validation AUC between 0.82 and 0.94 for all the models. The top 10 feature importance averaged across species shows that CpG and GC rich *k*-mers along with CpG O/E ratio are the most important predictors for all the models (see Figure 2A for feature importance across all models and Figure S3A for individual models). The most predictive features averaged from all species are sufficient for discriminating HOT and COLD regions for all species. Although, localized CpG and GC spikes across genomes of *C. elegans* and *D. melanogaster* are not common, we can discriminate HOT and COLD regions across all four species using the same GC/CpG rich top features. We used principal component analysis to visualize the discrimination between HOT and COLD regions using the top features (Figure 2B). To determine whether there are higher order sequences which differentiate between HOT and non-HOT regions, we performed discriminative *de novo* motif analysis on HOT regions from all four species. The resulting motifs were short (5–6 bp with high information content), GC and CpG dominant (Supplementary Figure S3B). The motifs partially matched binding sites of known transcription factors which bind GC rich sequences, such as SP1.

ChIP-seq for knock-out transcription factors have enrichment in HOT regions

Upon observing common low-level sequence features of HOT regions across species, we investigated whether potential technical biases in ChIP-seq could at least partially explain false positive signals on HOT regions. Previous studies suggest that even if the ChIP-ed protein does not exist in the analysed sample, highly expressed loci might give rise to false-positive peaks in yeast (7) and *D. melanogaster* (21). In order to address this question with a more comprehensive collection of datasets, we downloaded all available experiments where the ChIP-ed transcription factor was not physically present in the cell, as the gene encoding the transcription factor was 'knocked-out'. This set consists of 43 ChIP-seq experiments for knock-out (KO) transcription factors (KO ChIP-seq), where only 24 experiments have a control experiment in the form of input DNA or mock-IP (See Supplementary Table S1 for accession numbers and details). These experiments are carried out by different labs, which reduces the lab-specific bias for KO generation and ChIP-seq experiments. More than half of the KO ChIP-seq experiments show a clear signal enrichment (measured as IP/control) over HOT regions. KO ChIP-seq experiments with strong enrichment on HOT regions are shown in Figure 3A and experiments without signal enrichment are shown in Figure S5A. The signal is absent from regions which do not have extreme enrichment of TF binding events. Pooling all of the available signal enrichment for the KO ChIP-seq experiments with strong enrichment on HOT regions also shows the trend where signal enrichment on average is higher for HOT regions (Figure 3B shows signal enrichment of HOT regions and other control regions binned based on their TF occupancy percentiles). We examined real ChIP-seq peaks from the wild-type experiments and we observed that KO and WT ChIP-seq scores have a strong correlation on HOT regions. The magnitude of the correlation between WT and KO signal strength indicates that in most cases all WT peaks overlapping HOT regions represent a potential bias in ChIP experiments (Supplementary Figure S4). We have also checked the possibility that the signal in the KO experiments might originate from pulldown of highly related proteins—from paralogous transcription factors containing similar epitopes. Out of 24 proteins used in the KO experiments, only seven have known paralogues; eliminating the possibility of this confounding variable (Supplementary Table S3).

In addition, we noticed that some KO ChIP-seq experiments used IgG 'mock' ChIP-seq as control. The IgG ChIP-seq experiments should ideally control for unspecific binding that could potentially cause a false positive signal, and yet more than half of KO ChIP-seq experiments that have IgG ChIP-seq as control show signal enrichment on HOT regions (see Figure 3A). Following up on this, we wanted to see whether the HOT regions show an enrichment of signal in IgG control experiments. We downloaded available IgG control experiments from ENCODE, where antibodies from the same vendor was used in multiple cell types (results shown in Figure S5B). HOT regions showed a consistent enrichment in multiple IgG experiments, however, the
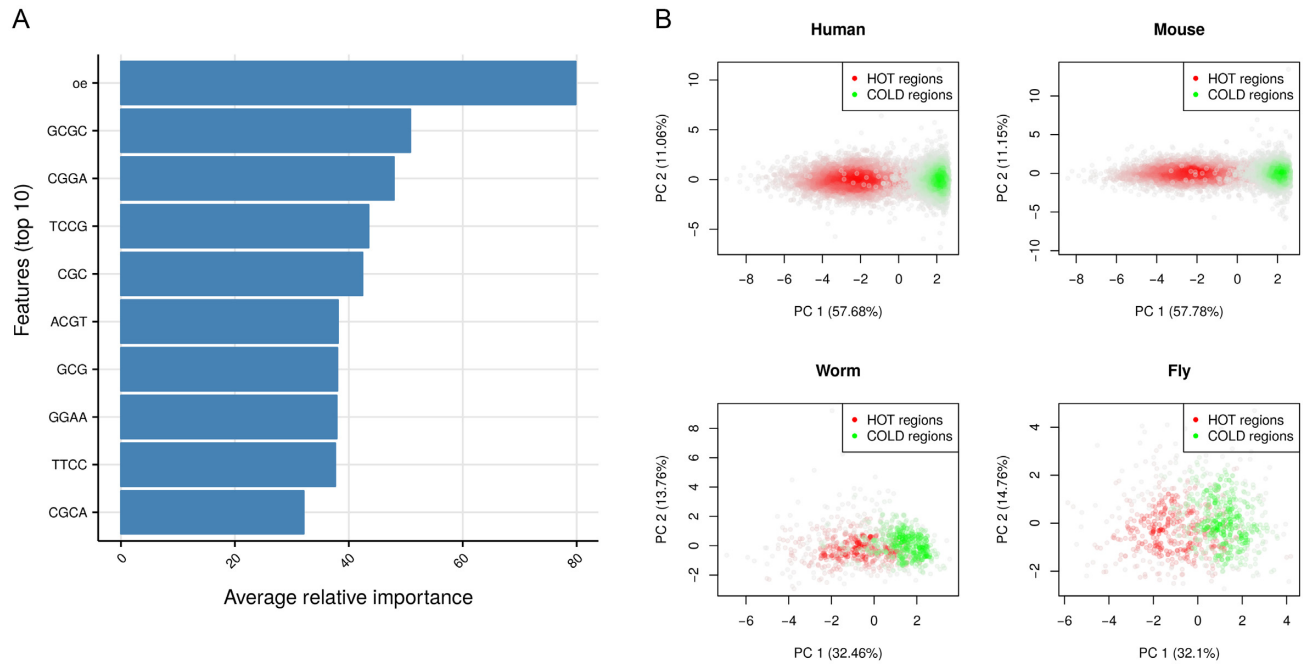
Figure 2. *k*-mer properties of HOT regions. (**A**) Top 10 features ordered in relative importance averaged across species. Importance scores are scaled to 0–100 scale for each species then averaged. Top feature [oe] means observed/expected ratio for CpG dinucleotides. (**B**) Principal component analysis using top 10 features shown in A. PCA is carried out for human, mouse, worm and fly separately. Scatter plots using first two principal components are shown, each dot represent HOT and COLD regions. The scatter plot is colored based on density of points, the more dense the points the darker the color.
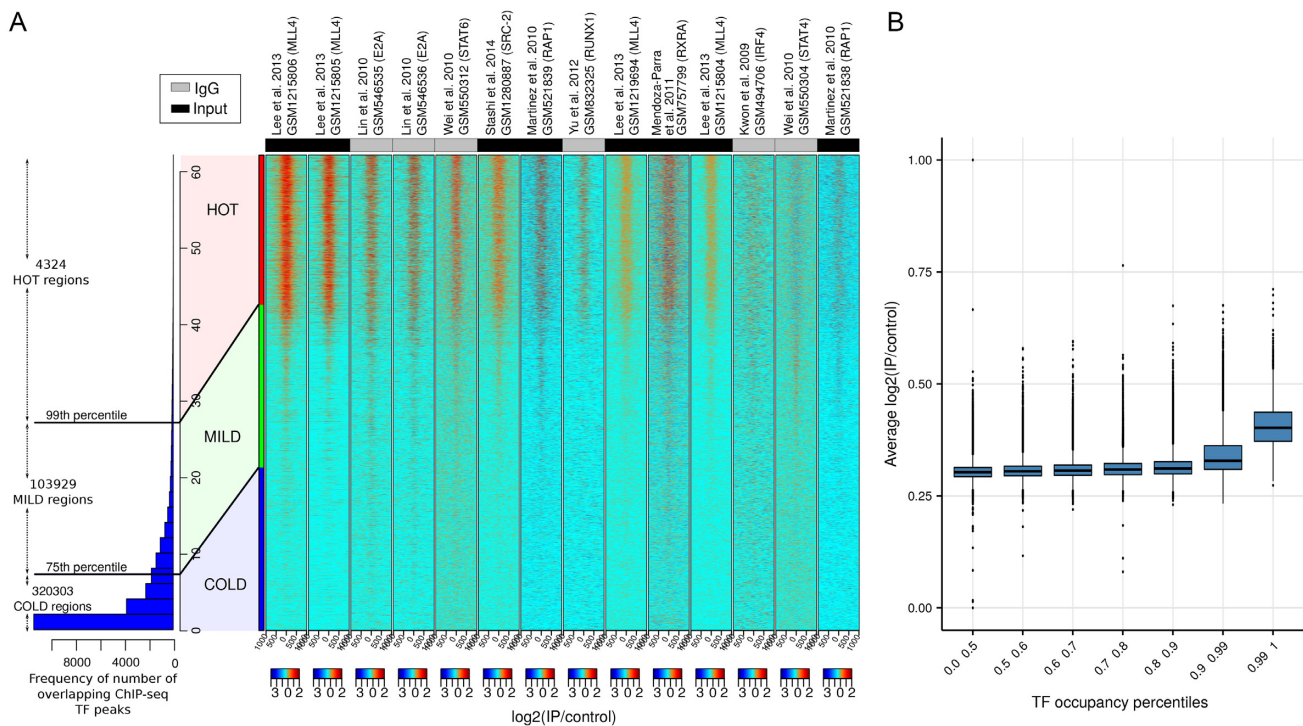
Figure 3. (**A**) Heatmaps show KO ChIP-seq experiments with signal on HOT regions. The barplot shows the distribution of the number of ChIP-seq peaks over all bound genomic regions. Regions overlapping >26 peaks sites (top 1 percentile) are labeled as HOT (4324 regions). The values in the heatmaps are log2(KO – IP/control). The status of the control sample input DNA or IgG is color coded on the top of the heatmaps. (**B**) Boxplots show distribution of average log2(IP/control) values for different sets of TF occupancy percentile bins. The log2(IP/control) values for each region in A is averaged across KO ChIP-seq experiments. The rightmost boxplot represents HOT regions, TF occupancy >99th percentile, and the rest represent control region sets with different TF occupancy percentiles.

enrichment was weak and showed variability, which was dependent on the cell type (Supplementary Figure S5C).

HOT regions are associated with R-loops and G-quadruplex DNA

We next investigated the associations of HOT regions with other GC rich features of the genome. One such feature that shares the same type of annotation with HOT regions such as CpG islands is R-loops. An R-loop is a nucleic acid structure that is composed of an RNA–DNA hybrid and a displaced single-stranded DNA (25). Their formation and stabilization are associated with GC content and CpG islands (26) and G-quadruplexes (27). R-loops exist across a broad spectrum of species from bacteria to high eukaryotes (28) and are shared across mammals (8,28). R-loop accumulation is a source of replication stress, genome instability, chromatin alterations, or gene silencing. They are associated with cancer and a number of genetic diseases (25).

R-loops can be detected genome-wide using a method called RNA–DNA immunoprecipitation followed by sequencing (DRIP-seq). It involves immunoprecipitation and sequencing of DNA fragments using the RNA–DNA hybrid specific S9.6 antibody (29), which was developed by extensively testing for specificity to RNA–DNA hybrids (30). We analysed publicly available DRIP-seq datasets to investigate R-loop enrichment on HOT regions (8,31,32) (See Supplementary Table S1 for accession numbers). We observed R-loop enrichment on HOT regions in every analyzed cell line, compared to other region sets, binned based on their TF occupancy percentiles (Figure 4A). We observed this enrichment even when the DRIP-seq experiments with RNAseH treatment were used as controls. The RNAseH treatment removes R-loops and subsequent DRIP-seq experiment results in depleted signal for R-loops. This shows that the S9.6 antibody binds specifically to R-loops and does not show additional interactions with other forms of DNA and DNA-binding proteins. In addition, we also observed DRIP-seq enrichment on HOT regions of *C. elegans* (Figure 4B). These results suggest that R-loops across different species overlap with HOT regions.

R-loops usually colocalize with G-quadruplex DNA (G4-DNA) which is a tertiary structure of single-stranded DNA (25). These structures can form on the displaced single-stranded G-rich DNA on the opposite side of the R-loop. We calculated the enrichment of G4-DNA on HOT regions using G4-ChIP-seq experiments (33), which are shown to enrich for G4-DNA specifically. We observed enrichment of G4-DNA signals on HOT regions, which is consistent with R loop localization on HOT regions (Figure 4D).

In addition, we would also expect to see R-loops in hyper-ChIPable regions in yeast, originally defined in yeast by Teytelman *et al.* (7). Indeed, we see enrichment of DRIP-seq signal (34) on published hyper-ChIPable regions (Figure 4C).

Since R-loops are associated with HOT regions, occupying TFs must be able to bind RNA–DNA hybrids or single-stranded DNA (ssDNA). Therefore, we checked if the TFs assayed by ENCODE have ssDNA binding or RNA–DNA hybrid binding domains, or such GO term an-

notations. Out of 165 studied TFs in human, only two of them contain at least one of the ssDNA binding domains: BACH1 contains a 'DUF1866' domain (from the RRM clan) and E2F6 contains a BRCA-2_OB3 domain (from the OB clan). Furthermore, none of the 165 TFs have an annotation of the GO term 'single-stranded DNA binding' (GO:0003697). When considering the direct interaction partners of these TFs, 31 out of 165 TFs (18.8%) have at least one direct interaction partner with an ssDNA-binding domain and 11 out of 165 TFs (6.7%) have at least one direct interaction partner with the GO term annotation for ssDNA binding. On the other hand, Cauli_VI domain that mediates the binding of RNASEH1 to RNA/DNA hybrids, is annotated only for two proteins in the whole proteome (RNASEH1 and Ankyrin repeat and LEM domain-containing protein 2 (ANKLE2)) and none of the human proteins have the associated GO term 'DNA/RNA hybrid binding (GO:0071667)' (according to the reviewed UniProt sequence annotations). Therefore, we could not detect any association of TFs or TFs' interaction partners with RNA/DNA hybrid binding function.

HOT regions have stable hypo-methylation across cell types

We investigated the CpG methylation dynamics over HOT regions, using base-pair resolution methylation data across multiple human cell types. Since most of the HOT regions are associated with CpG islands and genes with above average expression levels, we would expect low methylation over HOT regions (35). In addition, hypo-methylated CpGs are prevalent in R-loops, and the formation of R-loops are proposed to be protecting the R-loop associated loci from de novo DNA methylation (25). Consistent with these information, we observed hypo-methylation in HOT regions, compared to controls. The median methylation levels for HOT regions was similar to the median methylation levels of CpG islands not associated with HOT regions (non-HOT CGI) (see Figure 4E for an example cell line, see Figure S6 for all analyzed cell lines). Interestingly, non-HOT CGI had higher variation of methylation than the HOT regions despite the median methylation for both sets being low. This was a trend evident in all the cell types examined. Across the cell types, non-HOT CGI had 3–4 times higher methylation variation than HOT regions. This indicates that although HOT regions are associated with CpG islands, they are different from non-HOT CpG islands in their methylation dynamics and they maintain low levels of methylation across different cell types (see Figure 4F).

DISCUSSION

HOT regions are locations in the genome with remarkably high occupancy of transcription factors. They are formed by the combination of topmost ranking peaks from hundreds of ChIP experiments. HOT regions are mostly associated with promoters of stably expressed genes. They are located in open chromatin regions, however, DNA accessibility does not explain their formation. We showed that the low-level sequence features, such as GC rich and CpG containing *k*-mers, are shared across HOT regions of different species. Most interestingly, we demonstrated that HOT re-
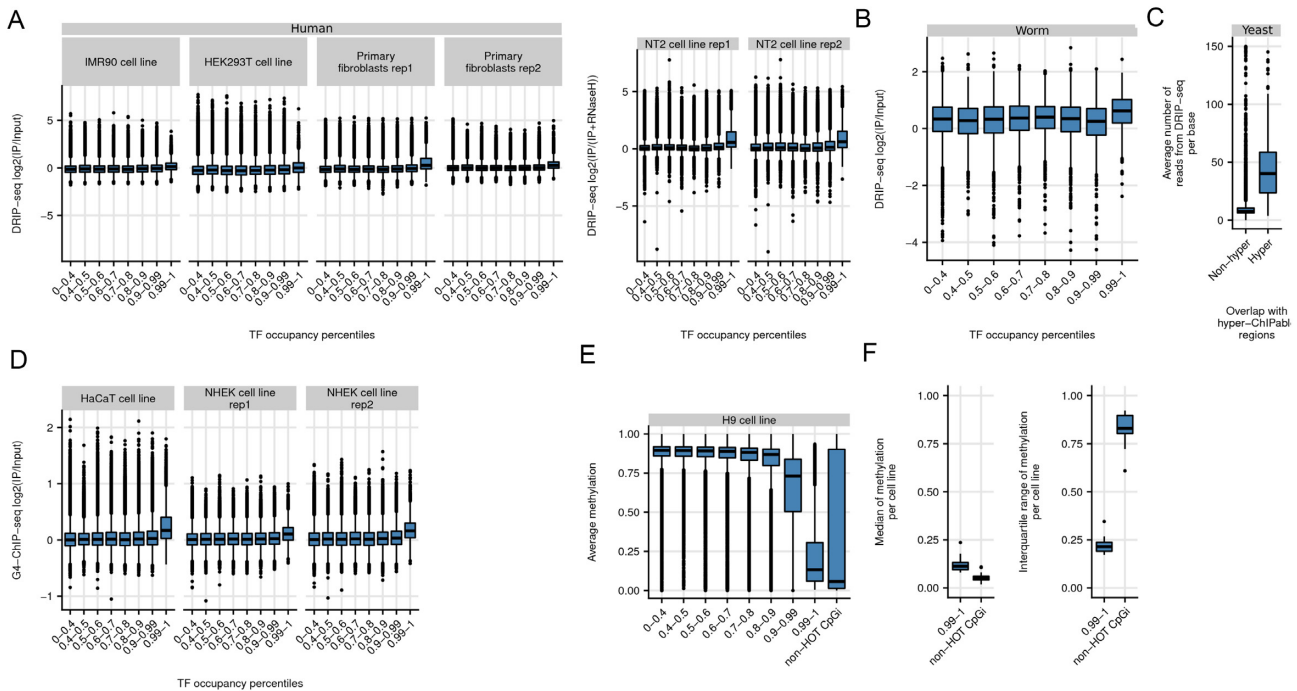
Figure 4. R-loops are associated with HOT regions. The boxplots show DRIP-seq log2(IP/control) for HOT regions and control regions binned based on their TF occupancy percentile in (**A**) Various human cell lines and in (**B**) worm. Boxplots show DRIP-seq read count per base-pair for hyper-ChIPable regions and all other genes as controls. (**C**) HyperChIP-able regions in yeast are enriched in R-loops. (**D**) HOT regions are enriched with G-quadruplex DNA (G4-ChIP-seq). Boxplots show $\log_2$(IP/control) for HOT regions and control regions binned based on their TF occupancy percentile. (**E**) HOT regions are hypo-methylated in comparison to controls in H9 cell line. Boxplots show distributions of methylation for HOT regions (rightmost boxplot) and control regions binned based on their TF occupancy percentile. (**F**) Left boxplot shows distributions of methylation medians across cell types for HOT regions and CpG islands that are not associated with HOT regions (non-HOT CpGi). Right boxplot shows distributions of methylation IQRs (interquartile ranges) across cell types for HOT regions and non-HOT CpGi.

gions are specifically enriched with false positive signals, using KO transcription factor ChIP-seq. These false positive signals are antibody dependent since KO ChIP-seq experiments show variable intensity of signals on HOT regions. The traditionally suggested controls, such as IgG ChIP-seq, can not reliably control for these artifacts. We showed that HOT regions associate with R-loops, in multiple organisms, as well as G-quadruplex DNA structures. Our results support the view that the peaks observed on HOT regions might be produced by the unspecific enrichment in multiple ChIP-seq experiments, rather than by the pull-down of specific transcription factors.

There might be many causes for the persistent false positive signal on HOT regions. The ChIP-seq signal consists of the signal from actual binding events and the noise. The noise is usually attributed to sequencing depth, library preparation, but most importantly to antibody specificity ([36]). The observed false positive signal could be obtained through pull-down of non-target proteins; this would however require that all experimentally used antibodies cross-react with a small set of proteins which constitutively bind GC rich promoters in multiple cell lines—a scenario which is highly improbable. The degree of overlap of HOT regions with R-loops suggests another hypothesis—that the antibodies cross-react directly with polynucleotide epitopes present in the HOT regions ([37]). R-loops are formed during transcription of GC rich, hypomethylated regions, where
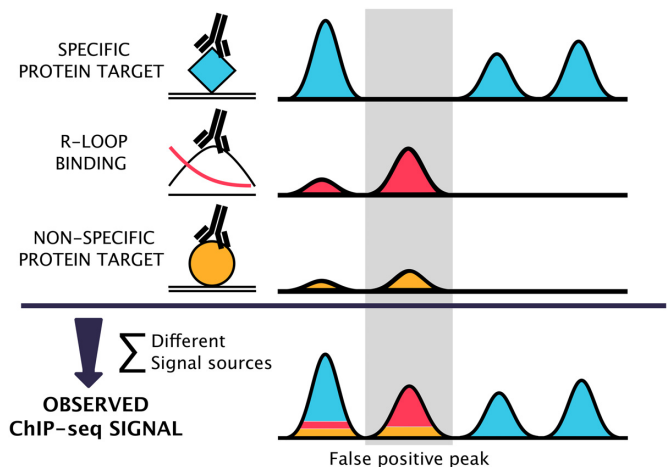

Figure 5. The observed ChIP signal arises from a combination of different signal sources. The signal in a ChIP experiment originates from an antibody binding to the intended target protein (blue), and nonspecific antibody binding—either to the non-target proteins (orange) or directly to polynucleotide structures, such as R-loops (red). The error (orange + red) is not proportional to the signal from the targeted protein, rather, it depends on sequence properties, antibody properties and expression characteristics of individual genomic regions. The combination of different noise profiles result in a subset of ChIP-seq peaks being false positives.

the nascent RNA strand displaces one of the DNA strands, forming an RNA:DNA Watson-Crick base pairing with the complementary strand. Such displacement causes R-loop prone regions to contain multiple polynucleotide structures: double stranded DNA, single stranded DNA, RNA:DNA hybrids, single stranded RNA (reviewed in (25)), G quadruplex complexes (38,39), etc., all of which can be bound by antibodies with a range of affinities (38,40–43). Anti-DNA antibodies are abundant in the serum of normal animals immunized with protein fragments (44–53), and are frequently polyspecific (48,54–62)—they can bind both polynucleotide and non-polynucleotide (e.g. peptide, phospholipid) epitopes (43). A recent study (63) has shown that anti-5methylcytosine antibodies nonspecifically enrich short tandem repeat sequences. Abundance of epitopes in constrained genomic regions, along with the fact that the HOT regions are associated with CpG islands of housekeeping genes (which are ubiquitously expressed and form R-loops in many cellular systems), and the promiscuity of antibodies, provide a simple explanation for the ubiquity of enrichments observed on HOT regions in various ChIP-seq experiments. Serum of non-immunized, healthy animals usually contains a low percentage of anti-DNA binding antibodies. This could explain why the IgG samples, when used as controls, show a signal on HOT regions, but the intensity of the signal is much lower than from antibodies produced by deliberate immunization. The recommended experimental methods for ascertaining antibody specificity (64) control almost exclusively for binding of antibodies to non-target proteins, so the direct interaction of antibodies with polynucleotide epitopes might be an underappreciated source of false positives in ChIP-seq experiments (Our model summarized in Figure 5). The signal on HOT regions could additionally arise by direct binding of TFs to single-stranded DNA (ssDNA) or RNA–DNA hybrids. Based on the current protein domain annotations, few to none of the TFs have such capabilities.

In this work, we have focused on regions that show high enrichment in multiple ChIP-seq experiments. Although we provide evidence that HOT regions do not contain several dozens of bound transcription factors, the real extent of detected false positive interactions is probably not limited to HOT regions. With the currently available data, it is not possible to estimate the proportion of an antibody specific error resulting from the enrichment due to the pull-down of non-target proteins vs. the direct binding to polynucleotide epitopes. Examination of the DNA binding properties of monoclonal antibodies, for example with protein binding arrays (65,66), might provide the required data for constructing more precise error models.

Lack of a strong signal over HOT regions in a subset of KO ChIP-seq samples shows that by using stringent antibody validation methods, it is possible to perform highly specific ChIP experiments. A level of prudence is needed though—a lack of signal in a KO ChIP-seq experiment might also be caused by technical conditions such as low number of reads, low library complexity or unsuccessful IP.

Our results, consistent with other recommendations (37,64,67,68), emphasise the need for critical examination and extensive testing of antibodies prior to their experimental usage. Whenever possible, controls in ChIP-seq experiments should be performed by ChIP-ing of protein in a system where the protein is not physically present, as implemented in Knockout Implemented Normalization method (KOIN) (69). If such controls are unfeasible, we provide lists of HOT regions and the ChIP-seq peaks overlapping with those regions should be carefully examined. We would like to encourage a careful, and methodical approach, where the existence of HOT regions is taken into account when performing functional association (i.e. colocalization analysis, functional enrichment) with the binding data—it is important to check whether the statistics are primarily driven by overlaps with HOT regions or not. On top of that, more stringent filtering for ChIP-seq peaks on HOT regions, such as removing peaks without canonical motifs, might be necessary.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
2. Celniker,S.E., Dillon,L.A.L., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
3. Xie,D., Dan,X., Boyle,A.P., Linfeng,W., Jie,Z., Trupti,K. *et al.* (2013) Dynamic trans-acting factor colocalization in human cells. *Cell*, **155**, 713–724.
4. Boyle,A.P., Araya,C.L., Cathleen,B., Philip,C., Chao,C., Yong,C., Gardner,K., Hillier,L.W., Janette,J., Jiang,L. *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**, 453–456.
5. Yip,K.Y., Cheng,C., Bhardwaj,N., Brown,J.B., Leng,J., Kundaje,A., Rozowsky,J., Birney,E., Bickel,P., Snyder,M. and Gerstein,M. (2012) Classification of human genomic regions based on experimentally

determined binding sites of more than 100 transcription-related factors. *Genome Biol.*, **13**, R48.

6. Afek,A., Ariel,A., Schipper,J.L., John,H., Raluca,G. and Lukatsky,D.B. (2014) Protein−DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 17140–17145.

7. Teytelman,L., Thurtle,D.M., Rine,J. and van Oudenaarden,A. (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 18602–18607.

8. Sanz,L.A., Hartono,S.R., Lim,Y.W., Steyaert,S., Rajpurkar,A., Ginno,P.A., Xu,X. and Chédin,F. (2016) Prevalent, dynamic, and conserved R-Loop structures associate with specific epigenomic signatures in mammals. *Mol. Cell*, **63**, 167–178.

9. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, **330**, 1775–1787.

10. McLean,C.Y., Dave,B., Michael,H., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

11. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.

12. Yao,Z., Macquarrie,K.L., Fong,A.P., Tapscott,S.J., Ruzzo,W.L. and Gentleman,R.C. (2014) Discriminative motif analysis of high-throughput dataset. *Bioinformatics*, **30**, 775–783.

13. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

14. Akalin,A., Franke,V., Vlahoviček,K., Mason,C.E. and Schübeler,D. (2015) Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics*, **31**, 1127–1129.

15. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

16. The UniProt Consortium (2016) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

17. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

18. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2011) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

19. Turner,B., Razick,S., Turinsky,A.L., Vlasblom,J., Crowdy,E.K., Cho,E., Morrison,K., Donaldson,I.M. and Wodak,S.J. (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database*, **2010**, baq023.

20. Chen,R.A.-J., Stempor,P., Down,T.A., Zeiser,E., Feuer,S.K. and Ahringer,J. (2014) Extreme HOT regions are CpG-dense promoters in C. elegans and humans. *Genome Res.*, **24**, 1138–1146.

21. Jain,D., Dhawal,J., Sandro,B., Angelika,Z., Tobias,I. and Becker,P.B. (2015) Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res.*, **43**, 6959–6968.

22. Park,D., Daechan,P., Yaelim,L., Gurvani,B. and Iyer,V.R. (2013) Widespread misinterpretable ChIP-seq Bias in Yeast. *PLoS One*, **8**, e83506.

23. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, **330**, 1775–1787.

24. Zou,H., Hui,Z. and Trevor,H. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B Stat. Methodol.*, **67**, 301–320.

25. Santos-Pereira,J.M. and Aguilera,A. (2015) R loops: new modulators of genome dynamics and function. *Nat. Rev. Genet.*, **16**, 583–597.

26. Ginno,P.A., Lim,Y.W., Lott,P.L., Korf,I. and Chédin,F. (2013) GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res.*, **23**, 1590–1600.

27. Skourti-Stathaki,K. and Proudfoot,N.J. (2014) A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev.*, **28**, 1384–1396.

28. Li,X. and Manley,J.L. (2006) Cotranscriptional processes and their influence on genome stability. *Genes Dev.*, **20**, 1838–1847.

29. Ginno,P.A., Lott,P.L., Christensen,H.C., Korf,I. and Chédin,F. (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell*, **45**, 814–825.

30. Lima,W.F., Murray,H.M., Damle,S.S., Hart,C.E., Hung,G., De Hoyos,C.L., Liang,X.H. and Crooke,S.T. (2016) Viable RNaseH1 knockout mice show RNaseH1 is essential for R loop processing, mitochondrial and liver function. *Nucleic Acids Res.*, **44**, 5299–5312.

31. Lim,Y.W., Sanz,L.A., Xu,X., Hartono,S.R. and Chédin,F. (2015) Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi-Goutières syndrome. *Elife*, **4**, e08007.

32. Zeller,P., Padeken,J., van Schendel,R., Kalck,V., Tijsterman,M. and Gasser,S.M. (2016) Histone H3K9 methylation is dispensable for Caenorhabditis elegans development but suppresses RNA:DNA hybrid-associated repeat instability. *Nat. Genet.*, **48**, 1385–1395.

33. Hänsel-Hertsch,R., Beraldi,D., Lensing,S.V., Marsico,G., Zyner,K., Parry,A., Di Antonio,M., Pike,J., Kimura,H., Narita,M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267–1272.

34. Wahba,L., Costantino,L., Tan,F.J., Zimmer,A. and Koshland,D. (2016) S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev.*, **30**, 1327–1338.

35. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.

36. Kidder,B.L., Hu,G. and Zhao,K. (2011) ChIP-Seq: technical considerations for obtaining high-quality data. *Nat. Immunol.*, **12**, 918–922.

37. Weitzmann,M.N. and Savage,N. (1994) Cloning of an antibody binding DNA sequence: pitfalls of DNA/protein immunoprecipitation reactions. *J. Immunol. Methods*, **173**, 7–10.

38. Kalsi,J.K., Martin,A.C., Hirabayashi,Y., Ehrenstein,M., Longhurst,C.M., Ravirajan,C., Zvelebil,M., Stollar,B.D., Thornton,J.M. and Isenberg,D.A. (1996) Functional and modelling studies of the binding of human monoclonal anti-DNA antibodies to DNA. *Mol. Immunol.*, **33**, 471–483.

39. Duquette,M.L., Handa,P., Vincent,J.A., Taylor,A.F. and Maizels,N. (2004) Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev*, **18**, 1618–1629.

40. Wang,Y., Mi,J. and Cao,X. (2000) Anti-DNA antibodies exhibit different binding motif preferences for single stranded or double stranded DNA. *Immunol. Lett.*, **73**, 29–34.

41. Braun,R.P. and Lee,J.S. (1986) Variations in duplex DNA conformation detected by the binding of monoclonal autoimmune antibodies. *Nucleic Acids Res.*, **14**, 5049–5065.

42. Jin,H., Sepúlveda,J. and Burrone,O.R. (2004) Specific recognition of a dsDNA sequence motif by an immunoglobulin VH homodimer. *Protein Sci.*, **13**, 3222–3229.

43. Barbas,S.M., Ditzel,H.J., Salonen,E.M., Yang,W.P., Silverman,G.J. and Burton,D.R. (1995) Human autoantibody recognition of DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 2529–2533.

44. Cerutti,M.L., Zarebski,L.M., de Prat Gay,G. and Goldbaum,F.A. (2005) A viral DNA-binding domain elicits anti-DNA antibodies of different specificities. *Mol. Immunol.*, **42**, 327–333.

45. Moens,U., Mathiesen,I., Ghelue,M.V. and Rekvig,O.P. (2002) Green fluorescent protein modified to bind DNA initiates production of anti-DNA antibodies when expressed in vivo. *Mol. Immunol.*, **38**, 505–514.

46. Moens,U., Seternes,O.M., Hey,A.W., Silsand,Y., Traavik,T., Johansen,B. and Rekvig,O.P. (1995) In vivo expression of a single viral DNA-binding protein generates systemic lupus erythematosus-related autoimmunity to double-stranded DNA and histones. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 12393–12397.

47. Voynova,E.N., Tchorbanov,A.I., Todorov,T.A. and Vassilev,T.L. (2005) Breaking of tolerance to native DNA in nonautoimmune mice by immunization with natural protein/DNA complexes. *Lupus*, **14**, 543–550.

48. Deocharan,B., Qing,X., Beger,E. and Putterman,C. (2002) Antigenic triggers and molecular targets for anti-double-stranded DNA antibodies. *Lupus*, **11**, 865–871.

49. Sciascia,S.A., Robson,K., Zhu,L., Garland,M., Grabosch,S., Kelamis,J., Messamore,W., Bradley,T., Sourk,A., Westberg,L. *et al.* (2007) Immunization of nonautoimmune mice with DNA binding domains of the largest subunit of RNA polymerase I results in production of anti-dsDNA and anti-Sm/RNP antibodies. *Autoimmunity*, **40**, 38–47.

50. Marchini,B., Puccetti,A., Dolcher,M.P., Madaio,M.P. and Migliorini,P. (1995) Induction of anti-DNA antibodies in non autoimmune mice by immunization with a DNA-DNAase I complex. *Clin. Exp. Rheumatol.*, **13**, 7–10.

51. Desai,D.D., Krishnan,M.R., Swindle,J.T. and Marion,T.N. (1993) Antigen-specific induction of antibodies against native mammalian DNA in nonautoimmune mice. *J. Immunol.*, **151**, 1614–1626.

52. Tran,T.T., Reich,C.F. 3rd, Alam,M. and Pisetsky,D.S. (2003) Specificity and immunochemical properties of anti-DNA antibodies induced in normal mice by immunization with mammalian DNA with a CpG oligonucleotide as adjuvant. *Clin. Immunol.*, **109**, 278–287.

53. Petrakova,N., Gudmundsdotter,L., Yermalovich,M., Belikov,S., Eriksson,L., Pyakurel,P., Johansson,O., Biberfeld,P., Andersson,S. and Isaguliants,M. (2009) Autoimmunogenicity of the helix-loop-helix DNA-binding domain. *Mol. Immunol.*, **46**, 1467–1480.

54. Lakamp,A.S. and Ouellette,M.M. (2011) A ssDNA aptamer that blocks the function of the anti-FLAG M2 antibody. *J. Nucleic Acids*, **2011**, 720–798.

55. Wun,H.L., Leung,D.T., Wong,K.C., Chui,Y.L. and Lim,P.L. (2001) Molecular mimicry: anti-DNA antibodies may arise inadvertently as a response to antibodies generated to microorganisms. *Int. Immunol.*, **13**, 1099–1107.

56. Pisetsky,D.S., Hoch,S.O., Klatt,C.L., O'Donnell,M.A. and Keene,J.D. (1985) Specificity and idiotypic analysis of a monoclonal anti-Sm antibody with anti-DNA activity. *J. Immunol.*, **135**, 4080–4085.

57. Zhang,W., Dang,S., Wang,J., Nardi,M.A., Zan,H., Casali,P. and Li,Z. (2010) Specific cross-reaction of anti-dsDNA antibody with platelet integrin GPIIIa49-66. *Autoimmunity*, **43**, 682–689.

58. Reichlin,M., Martin,A., Taylor-Albert,E., Tsuzaka,K., Zhang,W., Reichlin,M.W., Koren,E., Ebling,F.M., Tsao,B. and Hahn,B.H. (1994) Lupus autoantibodies to native DNA cross-react with the A and D SnRNP polypeptides. *J. Clin. Invest.*, **93**, 443–449.

59. Caponi,L., Chimenti,D., Pratesi,F. and Migliorini,P. (2002) Anti-ribosomal antibodies from lupus patients bind DNA. *Clin. Exp. Immunol.*, **130**, 541–547.

60. Yasuda,K., Richez,C., Uccellini,M.B., Richards,R.J., Bonegio,R.G., Akira,S., Monestier,M., Corley,R.B., Viglianti,G.A., Marshak-Rothstein,A. *et al.* (2009) Requirement for DNA CpG content in TLR9-dependent dendritic cell activation induced by DNA-containing immune complexes. *J. Immunol.*, **183**, 3109–3117.

61. Kumar,S., Hinks,J.A., Maman,J., Ravirajan,C.T., Pearl,L.H. and Isenberg,D.A. (2011) p185, an immunodominant epitope, is an autoantigen mimotope. *J. Biol. Chem.*, **286**, 26220–26227.

62. Gaynor,B., Putterman,C., Valadon,P., Spatz,L., Scharff,M.D. and Diamond,B. (1997) Peptide inhibition of glomerular deposition of an anti-DNA antibody. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 1955–1960.

63. Lentini,A., Lagerwall,C., Vikingsson,S., Mjoseng,H.K., Douvlataniotis,K., Vogt,H., Green,H., Meehan,R.R., Benson,M. and Nestor,C.E. (2018) A reassessment of DNA-immunoprecipitation-based genomic profiling. *Nat. Methods*, **15**, 499–504.

64. Wardle,F.C. and Tan,H. (2015) A ChIP on the shoulder? Chromatin immunoprecipitation and validation strategies for ChIP antibodies [version 1; peer review: 2 approved]. *F1000Res*, **4**, 235.

65. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.

66. Bulyk,M.L. (2007) Protein binding microarrays for the characterization of DNA-protein interactions. *Adv. Biochem. Eng. Biotechnol.*, **104**, 65–85.

67. Parseghian,M.H. (2013) Hitchhiker antigens: inconsistent ChIP results, questionable immunohistology data, and poor antibody performance may have a common factor. *Biochem. Cell Biol.*, **91**, 378–394.

68. Uhlen,M., Bandrowski,A., Carr,S., Edwards,A., Ellenberg,J., Lundberg,E., Rimm,D.L., Rodriguez,H., Hiltke,T., Snyder,M. *et al.* (2016) A proposal for validation of antibodies. *Nat. Methods*, **13**, 823–827.

69. Krebs,W., Schmidt,S.V., Goren,A., De Nardo,D., Labzin,L., Bovier,A., Ulas,T., Theis,H., Kraut,M., Latz,E. *et al.* (2014) Optimization of transcription factor binding map accuracy utilizing knockout-mouse models. *Nucleic Acids Res.*, **42**, 13051–13060.