MDPI

Article

Detecting differential transcription factor activity from ATAC-seq data

Ignacio J. Tripodi[1,2], Mary A. Allen[2], Robin D. Dowell[1,2,3,*]

1 Computer Science, University of Colorado, Boulder;
2 BioFrontiers Institute, University of Colorado, Boulder;
3 Molecular, Cellular and Developmental Biology, University of Colorado, Boulder
* Correspondence: robin.dowell@colorado.edu; Tel.: +1-303-492-8204

Version May 6, 2018 submitted to Molecules

Abstract: Transcription factors are managers of the cellular factory, and key components to many diseases. Many non-coding single nucleotide polymorphisms affect transcription factors, either by directly altering the protein or its functional activity at individual binding sites. Here we first briefly summarize high throughput approaches to studying transcription factor activity. We then demonstrate, using published chromatin accessibility data (specifically ATAC-seq), that the genome wide profile of TF recognition motifs relative to regions of open chromatin can determine the key transcription factor altered by a perturbation. Our method of determining which TF are altered by a perturbation is simple, quick to implement and can be used when biological samples are limited. In the future, we envision this method could be applied to determining which TFs show altered activity in response to a wide variety of drugs and diseases.

Keywords: transcription factor; perturbation; RNA-seq; DNase I cleavage; ATAC-seq; open chromatin; motif; DAStk

1. Introduction

Transcription factors (TFs) are the managers of the cellular factory, controlling everything from cellular identity to response to external stimuli[1]. Because of their central importance in interpreting the genome, millions of people are affected by mutations residing within TFs[2], causing a wide variety of symptoms (see Table 1). For example, over half of all cancers have a mutation in the TF TP53[3].

Table 1. Examples of diseases caused by mutations in a transcription factor.

Mutated TF	Disease/Symptoms
RUNX1	Familial platelet disorder with associated myeloid malignancy[4]
GRHL3	Cleft Palate[5]
MITF	Deafness[6]; Waardenburg syndrome (hearing loss)[7]
LMX1B	Nail-patella syndrome[8] (poorly developed nails and kneecaps)
TFAM	Mitochondrial DNA depletion syndrome[9]
NKX2-5	Congenital heart disease[10]
TBX5	Holt-Oram syndrome[11] (impared development of the heart and upper limbs)
MAF	Congenital cataract[12] (severe visual impairment in infants)
TCF4	Pitt-Hopkins syndrome[13] (intellectual disability and developmental delay, breathing problems, recurrent seizures)

Moreover, most disease-causing mutations are found in regulatory regions[14,15], e.g. enhancers, which are dense with TF binding sites[16]. A startling 60-76.5% of disease-associated single nucleotide

polymorphisms (SNPs) are in enhancers[17–20], which are short regulatory regions densely bound by TFs[21]. In fact, the well known program HaploReg now lists all TFs that bind over each SNP, a useful piece of information for understanding the impact of a SNP[22].

The relationship between many diseases and transcription factors has led to tremendous interest in global investigations of transcription factor activity. To decipher transcription factor activity requires understanding the two major functions of a transcription factor: binding to DNA and modification of transcription. Transcription factors bind to specific DNA sequences, a TF recognition motif. A number of techniques have been utilized to identify and characterize these recognition motifs[23]. However, because most genomic instances of the motif are not actually bound, having the recognition motif is insufficient. Protein-DNA interactions can be measured genome wide using chromatin immunoprecipitation followed by sequencing (ChIP-seq)[23–25]. Unfortunately, numerous lines of evidence indicate that not all binding events influence transcription[26–28]. Conceptually, this is akin to saying that just because someone is standing in a lab (TF binding) it does not imply they are actually conducting an experiment (altering transcription). Therefore, distinct assays are necessary to identify the locations where a TF is bound to DNA and determine whether that DNA binding leads to altered transcription nearby. A number of high throughput assays are available to interrogate these two key functions.

Extensive attention has focused on determining where in the genome transcription factors bind[23, 29,30]. The ENCODE project alone included approximately 2000 TF ChIP-seq experiments, including 180 TFs in K562 (myeloid leukemia) cells alone[29]. Large regulation project such as ENCODE and Roadmap Epigenomics have been invaluable to our understanding of TF binding. However, there are an estimate 1600 TFs in the human genome and many do not have a reliable antibody for ChIP-seq[23]. Even when antibodies are available, individual transcription factors can have distinct profiles of binding locations across cell types and conditions. Consequently, the cost of individually profiling every TF in each cell type is enormous, much less across different conditions[31]. Finally, if the effect of a particular perturbation is unknown, profiling assorted TFs by ChIP is prohibitively expensive.

An alternative approach to detecting individual protein-DNA binding locations is to infer a large collection of binding events via DNA footprinting[32–34]. Dense mapping of DNase I clevage sites identifies small regions protected from cleavage by the presence of a bound transcription factor[32,33]. While early footprinting studies identified a large repertoire of previously un-characterized motifs protected from cleavage, suggesting many novel transcription factors[34], subsequent work indicates these regions likely reflect sequence based cleavage bias of the DNase I enzyme[35]. Additionally, it is also now clear that most TFs (80%) do not show a measurable footprint[36], thereby limiting the effectiveness of this approach.

Despite these limitations, DNA footprinting assays uncovered a distinct function for transcription factors: altering DNA accessibility. When chromatin accessibility data is considered in the context of known TF sequence motifs[37–40], one can reasonably infer transcription factor binding profiles[41,42]. When accessibility profiles are then compared to ChIP in the context of perturbations, transcription factors could be classified as "pioneer" or "settler" depending on whether they open chromatin or require accessible, exposed DNA to bind[42]. Whether alterations of local chromatin accessibility reflect a byproduct of the TF's DNA binding or its altering of transcription remains unclear.

Altering transcription is the second major function of transcription factors[23]. Because TFs alter transcription, some of the earliest studies of TFs as regulators were based on expression data. For nearly twenty years large compendiums of expression data have been utilized to infer gene regulatory networks[43,44]. Typically these approaches search for modules, collections of co-regulated genes across distinct conditions. Identification of nearby TF recognition motifs[45,46] or co-regulated transcription factors[43] link particular TFs to the module of genes they regulate. For instance, ISMARA (Integrated System for Motif Activity Response Analysis)[47] models gene expression in terms of TF sequence motifs within proximal promoters. Gene regulatory network methods have been instrumental for understanding large scale regulatory networks, but are inherently limited by the fact that they

depend on steady state expression data. Steady state expression assays (microarray or RNA-seq) reflect not only transcription but also RNA processing, maturation and stability. Hence, they are an indirect readout on the effect of perturbations to transcription factors. Additionally, they are generally incapable of reliably detecting small changes at short time points without an impractical number of replicates[48].

Nascent transcription assays (GRO-seq, PRO-seq) directly profile RNA associated with engaged cellular polymerases[49,50]. Consequently, nascent assays are a direct readout on changes to transcription induced by perturbations[21,51]. Interestingly, an additional feature of nascent transcription data is the identification of short unstable transcripts immediately proximal to sites of transcription factor binding[52–57]. Importantly, these transcripts, now known as eRNAs can be employed as markers of TF activity[58]. The change in patterns of eRNA usage genome-wide relative to TF recognition motifs allows one to determine which transcription factors are altered by a perturbation with no a priori information. Unfortunately, nascent transcription protocols[49,50] are onerous, time consuming, and require large numbers of cells as input. Consequently, these experimental assays are predominantly used on cultured cell lines and not yet widely adopted. Therefore, we sought a simpler, easy to use approach to inferring differential transcription factor activity.

The Assay for Transposase-Accessible Chromatin followed by sequencing (ATAC-seq), a method for identifying regions of open chromatin, is particularly attractive because it is quick, easy, inexpensive, and deployable in small cell count samples. Additionally, recent work has shown that changes in chromatin accessibility can inform on TF activity. Specifically, BagFoot[36] combined footprinting with differential accessibility to identify TFs associated with altered chromatin accessibility profiles in the presence of a perturbation. They predominantly focused on DNase I hypersensitivity data, but also examined a small number of ATAC-seq datasets. Here we seek to confirm and extend their results in two ways. First, we ask whether an alternative approach – namely the motif displacement statistic[58], developed initially for nascent transcription analysis, could infer differential TF activity from ATAC-seq datasets. Second, we sought to construct an easy-to-use pipeline specific to the analysis of differential ATAC-seq analysis.

2. Results

We introduce a tool, Differential ATAC-seq toolkit (DAStk), developed with simplicity and ease of implementation in mind, focused around inferring changes in TF activity from ATAC-seq data. Using nascent transcription data we had previously developed the motif displacement score (MD-score) a metric that assesses TF associated transcriptional activity. As such, the MD-score reflects the enrichment of TF sequence motif within an small radius (150 bp) of enhancer RNA (eRNA) origins relative to a larger local window (1500 bp)[58]. While ATAC-seq does not directly provide information on eRNAs, most sites of eRNA activity reside within open chromatin[59]. Therefore we utilize the midpoint of called ATAC-seq peaks (rather than the eRNA origin) as a frame of reference for calculating MD-scores. Then, given two distinct biological conditions, we compare the ratio of MD-scores across the conditions and identify statistically significant changes by a two-proportion Z-test. Using public ATAC-seq data from a variety of human and mouse cell lines (IMR90, H524, NJH29, ZHBTC4) and perturbations (nutlin, doxycycline, tamoxifen), we assessed changes in accessibility over all putative TF sequence recognition motifs (for all motifs within the HOCOMOCO database[38]).

Given our familiarity with TP53 activation[55,60], we first examined this approach on ATAC-seq data gathered before and 6 hours after Nutlin-3a exposure on IMR90 cells[61]. Nutlin-3a is an exquisitely specific activator of TP53. As expected, we found that TP53 displayed the most significant change (p-value $< 10^{-5}$) in MD-score (Figure 1A, in red) of all motifs within the HOCOMOCO database[38]. Relaxing the p-value cutoff (p-value $< 10^{-4}$) we subsequently identified altered activity in TP63 and TP73 (Figure 1A, in maroon), likely reflecting the fact that these two proteins have nearly identical sequence recognition motifs as TP53.

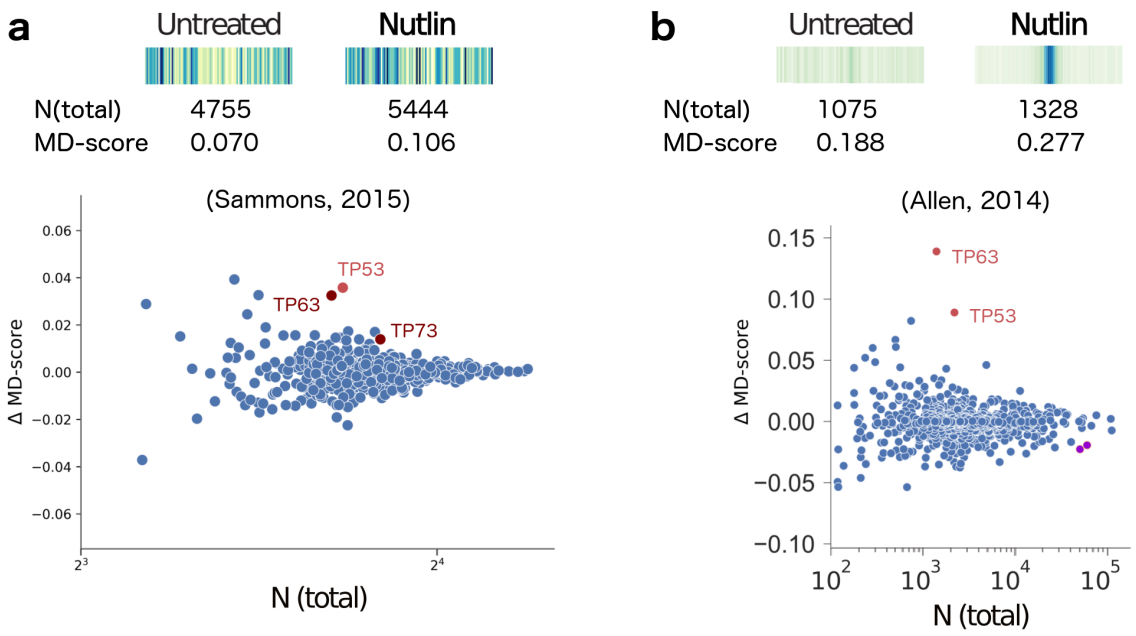

Figure 1. (**a**)Top: The motif displacement distribution as heatmap (increasingly dark blue indicates more instances of motif), MD-score and the number of motifs within 1.5kb of an ATAC-seq peak before and after stimulation with Nutlin-3a (e.g. Nutlin)[61] for TP53, the transcription factor known to be activated. Bottom: For all motif models (each dot), the change in MD-score following perturbation (y-axis) relative to the number of motifs within 1.5kb of any ATAC-seq peak center (x-axis). Red/maroon points indicate significantly increased MD-scores (p-value $< 10^{-5}$, $< 10^{-4}$, respectively). (**b**) Similar analysis obtained from nascent transcription data[55], where MD-scores are measured relative to eRNA origins. Purple dots indicate significantly decreased MD-scores. Figure adapted from Azofeifa et al [58].

Interestingly, Nutlin-3a has also been analyzed using nascent transcription data albeit in a different cell line (HCT116) at a shorter time point (1 hour)[55]. The MD-score analysis of the nascent data[58] obtained very similar results (Figure 1B). Unfortunately, a direct comparison of individual genomic loci between the two data sets is not feasible because they used different cell lines and drug exposure times. However, a couple of interesting observations concerning the overall MD-score trends are none-the-less noteworthy. First, the co-localization of the TP53 motif with ATAC-peak midpoints is far less striking than the co-localization of motifs with the eRNA origins (observed in the heatmap histograms). This observation, combined with the relative lower magnitude of ΔMD-scores (y-axis) suggests that the eRNA origin (obtained in nascent transcription) is a far more precise method of localizing and detecting changes in TF activity. Second, despite this lack of precision, ATAC-seq correctly identifies TP53 as the most dramatically altered MD-score whereas the best scoring motif with nascent transcription is TP63. Why this discrepancy exists is unclear, but given the relative similarity of these two motifs it may simply be coincidental.

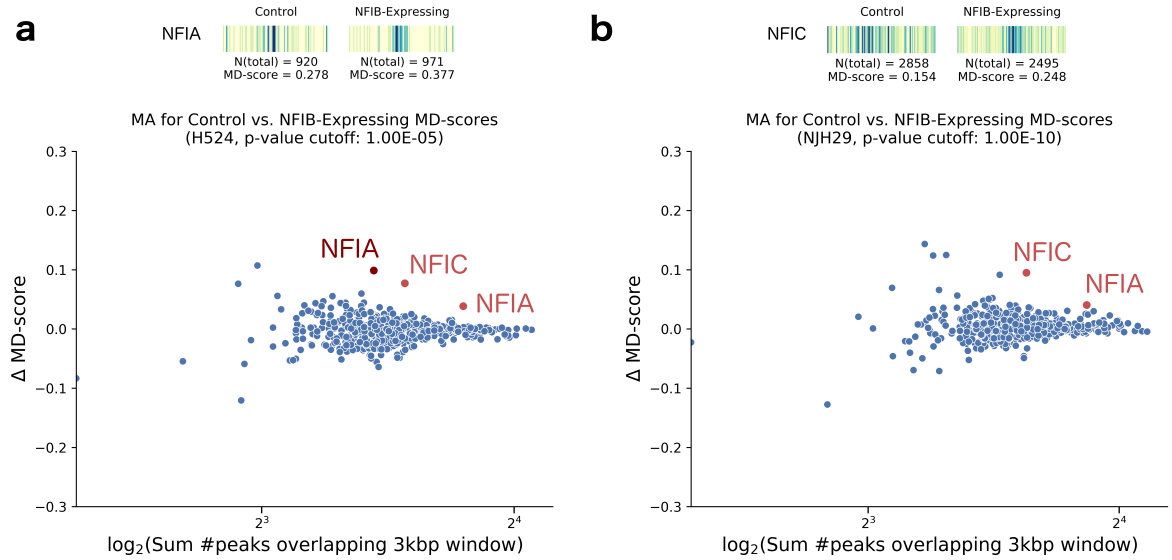

Figure 2. (**a**) Top: Motif displacement distribution as heatmap (increasingly dark blue indicates more instances of motif), MD-score and the number of motifs within 1.5kb of an ATAC-seq peak in control and NFIB-induced H524 cells with doxycycline, for the upregulated TF NFIA. Bottom: For all motif models (each dot), the change in MD-score following perturbation (y-axis) relative to the number of motifs within 1.5kb of any ATAC-seq peak center (x-axis). Red/maroon points indicate significantly increased MD-scores (p-value < 10-5, < 10-4, respectively). (**b**) Equivalent analysis performed on NJH29 cells, displaying a motif displacement distribution of the NFIC TF upregulation. We note that in the doxycycline-treated cells, most ATAC-seq peaks are located closer to the motif center than on the control cells.

We next analyzed differential ATAC-data gathered by Denny et. al to examine whether Nfib promotes metastasis via increasing chromatin accessibility. For this question, they examined two human small cell lung carcinoma (SCLC) cell lines (H524 and NJH29), profiling by ATAC-seq before and four hours after doxycycline treatment. Using the MD-score approach, we detect changes in TF activity for multiple members of the NFI family (Figure 2A,B). An increase in NFIA (two different motifs) and NFIC was detected in both cell types (p-value $< 10^{-5}$ for H524s; p-value $< 10^{-10}$ for NJH29s). As further confirmation of the NFI signal, we tested one of their mouse samples (KP22 cells) and found an increase of NFIA (p-value $< 10^{-5}$), consistent with the human results. We next asked whether our results were sensitive to the particular peaks utilized. To this end we sub-sampled peaks from the NJH29 data and re-ran our analysis. Both NFIA and NFIC are detectable as significant (p-value $< 10^{-10}$) even when using only half of the ATAC-seq peaks, suggesting the signal is reasonably robust.

We then sought to determine how the ΔMD-score approach compared to the BagFoot[36] at identifying differential TF activity. BagFoot also identified NIFA and NIFC within the SCLC differential ATAC-seq data[36]. However, they additionally claimed HNF6 as potentially altered in the SCLC data. Importantly, Baek et. al. noted that the HNF6 result did not hold when their approach utilized bias corrected data (based on naked DNA digested with Tn5). Given our MD-score approach does not identify HNF6 as altered further supports the idea that this result reflects a data artifact rather than a true biological phenomena. Interestingly, the MD-score approach and Bagfoot obtained nearly identical results on a second differential ATAC-seq dataset. In this case, King and Klose[62] showed BRG1, essential for pluripotency-related chromatin modifications, is required to make chromatin accessible at OCT4 target sites. To this end they treated ZHBTC4 mouse embryonic stem cells (ESCs) with tamoxifen for 72 hours to block BRG1 expression. When compared to the unperturbed mouse ESC control, we observed lowered MD-scores for SOX2, PO5F1 (Oct4) and NANOG in the BRG1-blocked cells (p-value $< 10^{-13}$; Figure 3A), directly confirming the BagFoot findings.

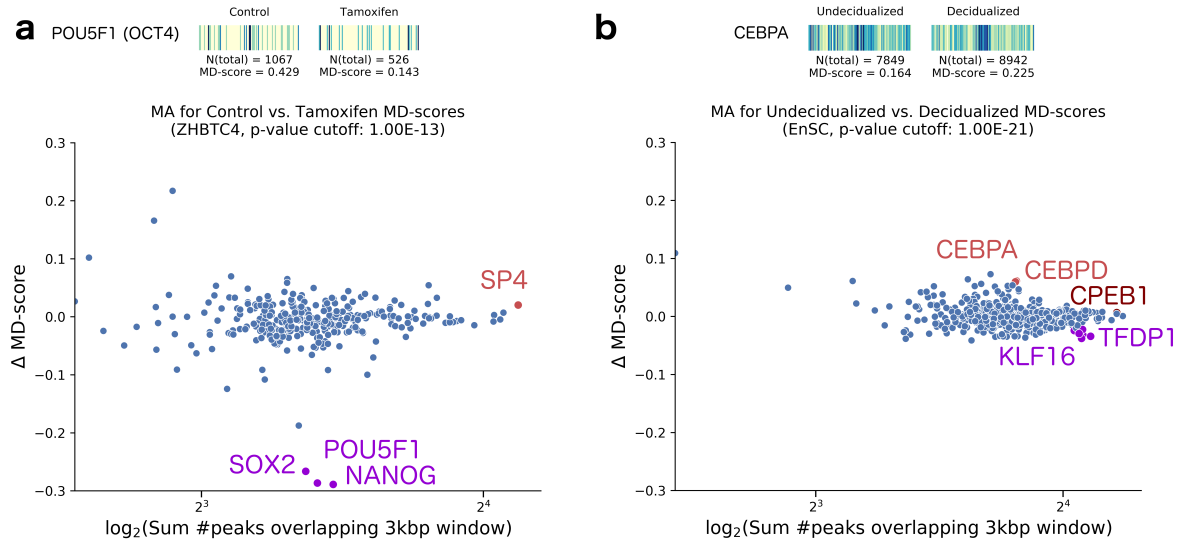

Figure 3. (**a**) Top: Motif displacement distribution as heatmap (increasingly dark blue indicates more instances of motif), MD-score and the number of motifs within 1.5kb of an ATAC-seq peak before and after stimulation with Tamoxifen[62] for the inhibited TF POU5F1, also known as OCT4. We observe that the decreased MD-score reflects not only a smaller number of peaks nearby this motif, but also a sharp decrease in co-localization with the motif. Bottom: For all motif models (each dot), the change in MD-score following perturbation (y-axis) relative to the number of motifs within 1.5kb of any ATAC-seq peak center (x-axis). Red points indicate significantly increased MD-scores (p-value $< 10^{-13}$). Purple dots indicate significantly decreased MD-scores, at the same indicated p-value. (**b**) Equivalent analysis performed on endometrial stromal cells, before and after undergoing a decidualization process[63]. The motif displacement heatmap illustrates ATAC-seq peak distances to CEBPA, the TF expected to be upregulated.

Finally, we also examined a differential ATAC-seq data obtained for decidualized and undecidualized human endometrium cells[63]. Spontaneous decidualization occurs in response to progesterone signalling (i.e. by an implanted embryo at the early stages of pregnancy). Using our MD-score approach, we found the CEBP family of transcription factors had increased activity in decidualized cells, consistent with the author's conclusion (Figure 3B). Additionally, we also found significantly lowered MD-scores for the KLF16 motif (a TF known to be involved in regulatory uterine cell biology[64]) and TFDP1 (a known target to the estrogen receptor $ER\beta$ present in all endometrial cell types[65] of lower activity during the secretory phase, in concert with the decidualization process). In all cases, the magnitude of MD-score alterations were relatively small, and yet the transcription factors uncovered can be linked to the underlying decidualization process.

3. Discussion

We sought to identify changes in TF activity across differential ATAC-seq datasets, as this protocol is inexpensive, simple and requires relatively small cell counts. Here we demonstrate two important results. First, using a simple statistic (the motif displacement score) as a co-localization measure of ATAC-seq peak midpoints to TF sequence motif sites across the genome, we correctly detect changes in TF activity. Second, our approach independently confirms the results obtained by BagFoot[36], as the two analysis techniques are distinct in their approach to quantifying differences in chromatin accessibility across conditions. Arguably, regardless of which analysis technique is preferred — differential ATAC-seq is a relatively simple and inexpensive way to assess for changes in TF activity induced by perturbations.

We believe there are two distinct advantages to the MD-score approach to assessing TF activity. First, the MD-score is calculated relative to a local background window. Consequently it cleanly

accounts for the localized sequence bias observed at promoters and enhancers[58], which likely reduces false positives. Second, the statistic is relatively simple to implement and naturally accommodates multiprocessing for faster computations. DAStk can easily be incorporated at the tail-end of a traditional processing pipeline for ATAC-seq data, in that MD-scores are calculated directly from called peaks and genomic sequence.

Our MD-score statistic was originally developed for analysis of nascent transcription data[58] and focused on enhancer RNA co-localization with motifs. Given most eRNAs originate from areas of open chromatin[21,57,66] and many transcription factors can alter chromatin accessibility[42], it is perhaps unsurprising that differential chromatin accessibility can be used to infer changes in TF activity. However, it remains unclear whether the observed alterations of chromatin reflect a distinct functional activity of transcription factors or are simply a side effect of DNA binding and/or altering transcription. While a careful examination of the two Nutlin-3a datasets (Figure 1) identifies several genomic regions altered uniquely in only one of the two datasets (ATAC-seq or nascent), the lack of matched data makes interpretation of these differences difficult. Do they reflect differences of cell type or distinct functional activities of TP53? A careful comparison of chromatin accessibility and nascent transcription data in the context of a perturbation will be necessary to fully address this question.

4. Materials and Methods

4.1. Processing pipeline

Each ATAC-seq dataset was subjected to a standard data processing pipeline. The SRR datasets were converted to FASTQ format using `fastq-dump` v2.8.0 with argument `-split-3`. Paired-ended raw reads were trimmed using `trimmomatic` v0.36 at a fixed length with options `PE -phred33 CROP:36 HEADCROP:6`. After verifying the dataset quality with `FastQC` v0.11.5, the reads were aligned to the hg19 or mm10 reference genome, using `Bowtie` v2.2.9 with arguments `-p32 -X2000`. The resulting SAM files were converted to BAM format using `samtools` v1.3.1 using the `view -q 20 -S -b` arguments and sorted with the `sort -m500G` arguments. Bam files were then converted to BedGraph format for easier processing using `bedtools` v2.25.0 with arguments `-bg -ibam INPUT_BAM_FILE -g GENOME_REFERENCE` and read counts were normalized by the millions mapped. Finally, MACS v2.1.1.20160309 was used to call broad peaks from the ATAC-seq BAM files with arguments `callpeak -n ASSAY_PREFIX -nomodel -format BAMPE -shift -100 -extsize 200 -B -broad`.

The human motif sites calculated in Azofeifa et al[58] for the hg19 reference genome were used for human cells. The motif sites for mouse cells were obtained using FIMO with position weight matrices (PWMs) from HOCOMOCO, with a p-value cutoff of 10^{-6} (arguments `-max-stored-scores 10000000 -thresh 1e-6`.

4.2. Public Datasets

We used samples from the following public GEO datasets for our analysis: GSE58740 (samples SRR1448793 and SRR1448795), GSE81255/GSE81258 (samples SRR3493647, SRR3493653, SRR3493643, SRR3493645, SRR3493626, SRR3493627, SRR3493634, and SRR3493635), GSE87822 (samples SRR4413799 and SRR4413811), and GSE104720 (samples SRR6148318 and SRR6148319).

4.3. DAStk Software

The Differential ATAC-seq toolkit (DAStk) is a collection of scripts publicly available at https://biof-git.colorado.edu/dowelllab/DAStk for download. We used 642 PWMs of human motifs in the HOCOMOCO[38] database (to verify the presence of ATAC-seq peaks nearby), and 427 mouse motifs. TF sequence motifs were mapped to the hg19 or mm10 reference genomes with a p-value cutoff of $10^{-}6$. For each motif, the number of ATAC-seq peaks was accounted for, within a large (1500bp radius) and small (150bp radius) window, to calculate the motif displacement score. The difference between the MD-score in each condition and the number of ATAC-seq peaks nearby (large window)

each motif were used to produce the MA plots. Those motifs with a statistically significant difference in MD-score were labeled, as determined by a z-test of two proportions[58].

Acknowledgments: This work was funded in part by a NSF IGERT grant number 1144807 (IJT, RDD), NIH training grant T15LM009451 (IJT), a Sie Fellowship (MAA) and a NSF ABI DBI-12624L0 (RDD). The authors acknowledge the BioFrontiers Computing Core at the University of Colorado Boulder for providing High Performance Computing resources (NIH 1S10OD012300) supported by BioFrontiers' IT.

Author Contributions: MAA and RDD conceived and designed the experiments; IJT implemented DAStk and performed the experiments; IJT and RDD analyzed the data; all authors contributed to writing the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

TF	Transcription factor
ATAC	Assay for Transposase-Accessible Chromatin
SNP	single nucleotide polymorphisms
ChIP	chromatin immunoprecipitation
eRNA	enhancer RNA
MD-score	motif displacement score
SCLC	small cell lung carcinoma

References

1. Spitz, F.; Furlong, E.E.M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **2012**, *13*, 613–626.
2. Latchman, D.S. Transcription-Factor Mutations and Disease. *New England Journal of Medicine* **1996**, *334*, 28–33.
3. Muller, P.A.J.; Vousden, K.H. p53 mutations in cancer. *Nat Cell Biol* **2013**, *15*, 2–8.
4. Jongmans, M.C.J.; Kuiper, R.P.; Carmichael, C.L.; Wilkins, E.J.; Dors, N.; Carmagnac, A.; Schouten-van Meeteren, A.Y.N.; Li, X.; Stankovic, M.; Kamping, E.; Bengtsson, H.; Schoenmakers, E.F.P.M.; van Kessel, A.G.; Hoogerbrugge, P.M.; Hahn, C.N.; Brons, P.P.; Scott, H.S.; Hoogerbrugge, N. Novel RUNX1 mutations in familial platelet disorder with enhanced risk for acute myeloid leukemia: clues for improved identification of the FPD/AML syndrome. *Leukemia* **2009**, *24*, 242–246.
5. Leslie, E.J.; Liu, H.; Carlson, J.C.; Shaffer, J.R.; Feingold, E.; Wehby, G.; Laurie, C.A.; Jain, D.; Laurie, C.C.; Doheny, K.F.; McHenry, T.; Resick, J.; Sanchez, C.; Jacobs, J.; Emanuele, B.; Vieira, A.R.; Neiswanger, K.; Standley, J.; Czeizel, A.E.; Deleyiannis, F.; Christensen, K.; Munger, R.G.; Lie, R.T.; Wilcox, A.; Romitti, P.A.; Field, L.L.; Padilla, C.D.; Cutiongco-de la Paz, E.M.C.; Lidral, A.C.; Valencia-Ramirez, L.C.; Lopez-Palacio, A.M.; Valencia, D.R.; Arcos-Burgos, M.; Castilla, E.E.; Mereb, J.C.; Poletta, F.A.; Orioli, I.M.; Carvalho, F.M.; Hecht, J.T.; Blanton, S.H.; Buxó, C.J.; Butali, A.; Mossey, P.A.; Adeyemo, W.L.; James, O.; Braimah, R.O.; Aregbesola, B.S.; Eshete, M.A.; Deribew, M.; Koruyucu, M.; Seymen, F.; Ma, L.; de Salamanca, J.E.; Weinberg, S.M.; Moreno, L.; Cornell, R.A.; Murray, J.C.; Marazita, M.L. A Genome-wide Association Study of Nonsyndromic Cleft Palate Identifies an Etiologic Missense Variant in GRHL3. *The American Journal of Human Genetics* **2016**, *98*, 744–754.
6. Smith, S.; Kelley, P.; Kenyon, J.; Hoover, D. Tietz syndrome (hypopigmentation/deafness) caused by mutation of MITF. *Journal of Medical Genetics* **2000**, *37*, 446–448.
7. Tassabehji, M.; Newton, V.E.; Read, A.P. Waardenburg syndrome type 2 caused by mutations in the human microphthalmia (MITF) gene. *Nat Genet* **1994**, *8*, 251–255.
8. Marini, M.; Bocciardi, R.; Gimelli, S.; Di Duca, M.; Divizia, M.T.; Baban, A.; Gaspar, H.; Mammi, I.; Garavelli, L.; Cerone, R.; Emma, F.; Bedeschi, M.F.; Tenconi, R.; Sensi, A.; Salmaggi, A.; Bengala, M.; Mari, F.; Colussi, G.; Szczaluba, K.; Antonarakis, S.E.; Seri, M.; Lerone, M.; Ravazzolo, R. A spectrum of LMX1B mutations in Nail-Patella syndrome: New point mutations, deletion, and evidence of mosaicism in unaffected parents. *Genet Med* **2010**, *12*, 431–439.

9. Stiles, A.R.; Simon, M.T.; Stover, A.; Eftekharian, S.; Khanlou, N.; Wang, H.L.; Magaki, S.; Lee, H.; Partynski, K.; Dorrani, N.; Chang, R.; Martinez-Agosto, J.A.; Abdenur, J.E. Mutations in TFAM, encoding mitochondrial transcription factor A, cause neonatal liver failure associated with mtDNA depletion. *Molecular Genetics and Metabolism* **2016**, *119*, 91–99.
10. Chung, I.M.; Rajakumar, G. Genetics of Congenital Heart Defects: The NKX2-5 Gene, a Key Player. *Genes* **2016**, *7*, 6.
11. Yi Li, Q.; Newbury-Ecob, R.A.; Terrett, J.A.; Wilson, D.I.; Curtis, A.R.J.; Ho Yi, C.; Gebuhr, T.; Bullen, P.J.; Robson, S.C.; Strachan, T.; Bonnet, D.; Lyonnet, S.; Young, I.D.; Raeburn, J.A.; Buckler, A.J.; Law, D.J.; Brook, J.D. Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nat Genet* **1997**, *15*, 21–29.
12. Narumi, Y.; Nishina, S.; Tokimitsu, M.; Aoki, Y.; Kosaki, R.; Wakui, K.; Azuma, N.; Murata, T.; Takada, F.; Fukushima, Y.; Kosho, T. Identification of a novel missense mutation of MAF in a Japanese family with congenital cataract by whole exome sequencing: A clinical report and review of literature. *American Journal of Medical Genetics Part A* **2014**, *164*, 1272–1276.
13. Sweatt, J.D. Pitt-Hopkins Syndrome: intellectual disability due to loss of TCF4-regulated gene transcription. *Exp Mol Med* **2013**, *45*, e21.
14. Lee, T.I.; Young, R.A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **2013**, *152*, 1237 – 1251.
15. Maurano, M.T.; Haugen, E.; Sandstrom, R.; Vierstra, J.; Shafer, A.; Kaul, R.; Stamatoyannopoulos, J.A. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nature Genetics* **2015**, *47*, 1393–1401.
16. Shlyueva, D.; Stampfel, G.; Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **2014**, *15*, 272–286.
17. Maurano, M.T.; Humbert, R.; Rynes, E.; Thurman, R.E.; Haugen, E.; Wang, H.; Reynolds, A.P.; Sandstrom, R.; Qu, H.; Brody, J.; Shafer, A.; Neri, F.; Lee, K.; Kutyavin, T.; Stehling-Sun, S.; Johnson, A.K.; Canfield, T.K.; Giste, E.; Diegel, M.; Bates, D.; Hansen, R.S.; Neph, S.; Sabo, P.J.; Heimfeld, S.; Raubitschek, A.; Ziegler, S.; Cotsapas, C.; Sotoodehnia, N.; Glass, I.; Sunyaev, S.R.; Kaul, R.; Stamatoyannopoulos, J.A. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **2012**, *337*, 1190–1195.
18. Corradin, O.; Scacheri, P.C. Enhancer variants: evaluating functions in common disease. *Genome Medicine* **2014**, *6*, 85.
19. Mifsud, B.; Tavares-Cadete, F.; Young, A.N.; Sugar, R.; Schoenfelder, S.; Ferreira, L.; Wingett, S.W.; Andrews, S.; Grey, W.; Ewels, P.A.; Herman, B.; Happe, S.; Higgs, A.; LeProust, E.; Follows, G.A.; Fraser, P.; Luscombe, N.M.; Osborne, C.S. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **2015**, *47*, 598–606.
20. Farh, K.K.H.; Marson, A.; Zhu, J.; Kleinewietfeld, M.; Housley, W.J.; Beik, S.; Shoresh, N.; Whitton, H.; Ryan, R.J.H.; Shishkin, A.A.; Hatan, M.; Carrasco-Alfonso, M.J.; Mayer, D.; Luckey, C.J.; Patsopoulos, N.A.; De Jager, P.L.; Kuchroo, V.K.; Epstein, C.B.; Daly, M.J.; Hafler, D.A.; Bernstein, B.E. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **2015**, *518*, 337–343.
21. Lam, M.T.; Li, W.; Rosenfeld, M.G.; Glass, C.K. Enhancer {RNAs} and regulated transcriptional programs. *Trends in Biochemical Sciences* **2014**, *39*, 170 – 182.
22. Ward, L.D.; Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research* **2011**, *40*, D930.
23. Lambert, S.A.; Jolma, A.; Campitelli, L.F.; Das, P.K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T.R.; Weirauch, M.T. The Human Transcription Factors. *Cell* **2018**, *172*, 650–665.
24. Farnham, P.J. Insights from genomic profiling of transcription factors. *Nat Rev Genet* **2009**, *10*, 605–616.
25. Gade, P.; Kalvakolanu, D.V. Chromatin Immunoprecipitation Assay as a Tool for Analyzing Transcription Factor Activity. *Methods in molecular biology (Clifton, N.J.)* **2012**, *809*, 85–104.
26. Whitfield, T.W.; Wang, J.; Collins, P.J.; Partridge, E.C.; Aldred, S.F.; Trinklein, N.D.; Myers, R.M.; Weng, Z. Functional analysis of transcription factor binding sites in human promoters. *Genome Biology* **2012**, *13*, R50.
27. Spivakov, M. Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays* **2014**, *36*, 798–806.
28. Cusanovich, D.A.; Pavlovic, B.; Pritchard, J.K.; Gilad, Y. The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genet* **2014**, *10*, e1004226.

29. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.
30. Consortium, R.E.; Kundaje, A.; Meuleman, W.; Ernst, J.; Bilenky, M.; Yen, A.; Heravi-Moussavi, A.; Kheradpour, P.; Zhang, Z.; Wang, J.; Ziller, M.J.; Amin, V.; Whitaker, J.W.; Schultz, M.D.; Ward, L.D.; Sarkar, A.; Quon, G.; Sandstrom, R.S.; Eaton, M.L.; Wu, Y.C.; Pfenning, A.R.; Wang, X.; Claussnitzer, M.; Liu, Y.; Coarfa, C.; Harris, R.A.; Shoresh, N.; Epstein, C.B.; Gjoneska, E.; Leung, D.; Xie, W.; Hawkins, R.D.; Lister, R.; Hong, C.; Gascard, P.; Mungall, A.J.; Moore, R.; Chuah, E.; Tam, A.; Canfield, T.K.; Hansen, R.S.; Kaul, R.; Sabo, P.J.; Bansal, M.S.; Carles, A.; Dixon, J.R.; Farh, K.H.; Feizi, S.; Karlic, R.; Kim, A.R.; Kulkarni, A.; Li, D.; Lowdon, R.; Elliott, G.; Mercer, T.R.; Neph, S.J.; Onuchic, V.; Polak, P.; Rajagopal, N.; Ray, P.; Sallari, R.C.; Siebenthall, K.T.; Sinnott-Armstrong, N.A.; Stevens, M.; Thurman, R.E.; Wu, J.; Zhang, B.; Zhou, X.; Beaudet, A.E.; Boyer, L.A.; De Jager, P.L.; Farnham, P.J.; Fisher, S.J.; Haussler, D.; Jones, S.J.M.; Li, W.; Marra, M.A.; McManus, M.T.; Sunyaev, S.; Thomson, J.A.; Tlsty, T.D.; Tsai, L.H.; Wang, W.; Waterland, R.A.; Zhang, M.Q.; Chadwick, L.H.; Bernstein, B.E.; Costello, J.F.; Ecker, J.R.; Hirst, M.; Meissner, A.; Milosavljevic, A.; Ren, B.; Stamatoyannopoulos, J.A.; Wang, T.; Kellis, M.; Consortium, R.E. Integrative analysis of 111 reference human epigenomes. *Nature* **2015**, *518*, 317–330.
31. Steinhauser, S.; Kurzawa, N.; Eils, R.; Herrmann, C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in Bioinformatics* **2016**, *17*, 953–966.
32. Hesselberth, J.R.; Chen, X.; Zhang, Z.; Sabo, P.J.; Sandstrom, R.; Reynolds, A.P.; Thurman, R.E.; Neph, S.; Kuehn, M.S.; Noble, W.S.; Fields, S.; Stamatoyannopoulos, J.A. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods* **2009**, *6*, 283–289.
33. Chen, X.; Hoffman, M.M.; Bilmes, J.A.; Hesselberth, J.R.; Noble, W.S. A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics* **2010**, *26*, i334–i342.
34. Neph, S.; Vierstra, J.; Stergachis, A.B.; Reynolds, A.P.; Haugen, E.; Vernot, B.; Thurman, R.E.; John, S.; Sandstrom, R.; Johnson, A.K.; Maurano, M.T.; Humbert, R.; Rynes, E.; Wang, H.; Vong, S.; Lee, K.; Bates, D.; Diegel, M.; Roach, V.; Dunn, D.; Neri, J.; Schafer, A.; Hansen, R.S.; Kutyavin, T.; Giste, E.; Weaver, M.; Canfield, T.; Sabo, P.; Zhang, M.; Balasundaram, G.; Byron, R.; MacCoss, M.J.; Akey, J.M.; Bender, M.A.; Groudine, M.; Kaul, R.; Stamatoyannopoulos, J.A. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **2012**, *489*, 83–90.
35. He, H.H.; Meyer, C.A.; Hu, S.S.; Chen, M.W.; Zang, C.; Liu, Y.; Rao, P.K.; Fei, T.; Xu, H.; Long, H.; Liu, X.S.; Brown, M. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature Methods* **2013**, *11*, 73 EP –.
36. Baek, S.; Goldstein, I.; Hager, G.L. Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. *Cell Reports* **2017**, *19*, 1710 – 1722.
37. Kulakovskiy, I.V.; Medvedeva, Y.A.; Schaefer, U.; Kasianov, A.S.; Vorontsov, I.E.; Bajic, V.B.; Makeev, V.J. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Research* **2013**, *41*, D195–D202.
38. Kulakovskiy, I.V.; Vorontsov, I.E.; Yevshin, I.S.; Sharipov, R.N.; Fedorova, A.D.; Rumynskiy, E.I.; Medvedeva, Y.A.; Magana-Mora, A.; Bajic, V.B.; Papatsenko, D.A.; Kolpakov, F.A.; Makeev, V.J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research* **2018**, *46*, D252–D259.
39. Matys, V.; Kel-Margoulis, O.V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; Voss, N.; Stegmaier, P.; Lewicki-Potapov, B.; Saxel, H.; Kel, A.E.; Wingender, E. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **2006**, *34*, D108–D110.
40. Portales-Casamar, E.; Thongjuea, S.; Kwon, A.T.; Arenillas, D.; Zhao, X.; Valen, E.; Yusuf, D.; Lenhard, B.; Wasserman, W.W.; Sandelin, A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research* **2010**, *38*, D105–D110.
41. Pique-Regi, R.; Degner, J.F.; Pai, A.A.; Gaffney, D.J.; Gilad, Y.; Pritchard, J.K. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research* **2011**, *21*, 447–455.

42. Sherwood, R.I.; Hashimoto, T.; O'Donnell, C.W.; Lewis, S.; Barkal, A.A.; van Hoff, J.P.; Karun, V.; Jaakkola, T.; Gifford, D.K. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature biotechnology* **2014**, *32*, 171–178.
43. Segal, E.; Shapira, M.; Regev, A.; Pe'er, D.; Botstein, D.; Koller, D.; Friedman, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **2003**, *34*, 166–176.
44. Thompson, D.; Regev, A.; Roy, S. Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution. *Annual Review of Cell and Developmental Biology* **2015**, *31*, 399–428.
45. Tavazoie, S.; Hughes, J.D.; Campbell, M.J.; Cho, R.J.; Church, G.M. Systematic determination of genetic network architecture. *Nature Genetics* **1999**, *22*, 281 EP –.
46. Pilpel, Y.; Sudarsanam, P.; Church, G.M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* **2001**, *29*, 153 EP –.
47. Balwierz, P.J.; Pachkov, M.; Arnold, P.; Gruber, A.J.; Zavolan, M.; van Nimwegen, E. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research* **2014**.
48. Hart, S.N.; Therneau, T.M.; Zhang, Y.; Poland, G.A.; Kocher, J.P. Calculating Sample Size Estimates for RNA Sequencing Data. *Journal of Computational Biology* **2013**, *20*, 970–978.
49. Core, L.J.; Waterfall, J.J.; Lis, J.T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **2008**, *322*, 1845–1848.
50. Kwak, H.; Fuda, N.J.; Core, L.J.; Lis, J.T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science (New York, N.Y.)* **2013**, *339*, 950–953.
51. Liu, Y.; Chen, S.; Wang, S.; Soares, F.; Fischer, M.; Meng, F.; Du, Z.; Lin, C.; Meyer, C.; DeCaprio, J.A.; Brown, M.; Liu, X.S.; He, H.H. Transcriptional landscape of the human cell cycle. *Proceedings of the National Academy of Sciences* **2017**, *114*, 3473–3478.
52. Kim, T.k.; Hemberg, M.; Gray, J.M.; Costa, A.M.; Bear, D.M.; Wu, J.; Harmin, D.A.; Laptewicz, M.; Barbara-Haley, K.; Kuersten, S.; Markenscoff-Papadimitriou, E.; Kuhl, D.; Bito, H.; Worley, P.F.; Kreiman, G.; Greenberg, M.E. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **2010**, *465*, 182–187.
53. Hah, N.; Murakami, S.; Nagari, A.; Danko, C.G.; Kraus, W.L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Research* **2013**, *23*, 1210–1223.
54. Luo, X.; Chae, M.; Krishnakumar, R.; Danko, C.G.; Kraus, W.L. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. *BMC Genomics* **2014**, *15*, 155–155.
55. Allen, M.A.; Mellert, H.; Dengler, V.; Andryzik, Z.; Guarnieri, A.; Freeman, J.A.; Luo, X.; Kraus, W.L.; Dowell, R.D.; Espinosa, J.M. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. *eLife* **2014**, *3*, e02200.
56. Puc, J.; Kozbial, P.; Li, W.; Tan, Y.; Liu, Z.; Suter, T.; Ohgi, K.A.; Zhang, J.; Aggarwal, A.K.; Rosenfeld, M.G. Ligand-Dependent Enhancer Activation Regulated by Topoisomerase-I Activity. *Cell* **2015**, *160*, 367 – 380.
57. Danko, C.G.; Hyland, S.L.; Core, L.J.; Martins, A.L.; Waters, C.T.; Lee, H.W.; Cheung, V.G.; Kraus, W.L.; Lis, J.T.; Siepel, A. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Meth* **2015**, *12*, 433–438.
58. Azofeifa, J.G.; Allen, M.A.; Hendrix, J.R.; Read, T.; Rubin, J.D.; Dowell, R.D. Enhancer RNA profiling predicts transcription factor activity. *Genome Research* **2018**, *28*, 334–344.
59. Azofeifa, J.G.; Dowell, R.D. A generative model for the behavior of RNA polymerase. *Bioinformatics* **2017**, *33*, 227–234.
60. Azofeifa, J.G.; Allen, M.A.; Lladser, M.E.; Dowell, R.D. An Annotation Agnostic Algorithm for Detecting Nascent RNA Transcripts in GRO-Seq. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2017**, *14*, 1070–1081.
61. Sammons, M.A.; Zhu, J.; Drake, A.M.; Berger, S.L. TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity. *Genome Research* **2015**, *25*, 179–188. doi:10.1101/gr.181883.114.
62. King, H.W.; Klose, R.J. The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. *eLife* **2017**, *6*, e22631.

63. Vrljicak, P.; Lucas, E.S.; Lansdowne, L.; Lucciola, R.; Muter, J.; Dyer, N.P.; Brosens, J.J.; Ott, S. Analysis of chromatin accessibility in decidualizing human endometrial stromal cells. *The FASEB Journal* **0**, *0*, fj.201701098R.
64. Daftary, G.S.; Lomberk, G.A.; Buttar, N.S.; Allen, T.W.; Grzenda, A.; Zhang, J.; Zheng, Y.; Mathison, A.J.; Gada, R.P.; Calvo, E.; Iovanna, J.L.; Billadeau, D.D.; Prendergast, F.G.; Urrutia, R. Detailed Structural-Functional Analysis of the Krüppel-like Factor 16 (KLF16) Transcription Factor Reveals Novel Mechanisms for Silencing Sp/KLF Sites Involved in Metabolism and Endocrinology. *Journal of Biological Chemistry* **2012**, *287*, 7010–7025.
65. Tamm, K.; Rõõm, M.; Salumets, A.; Metsis, M. Genes targeted by the estrogen and progesterone receptors in the human endometrial cell lines HEC1A and RL95-2. *Reproductive Biology and Endocrinology* **2009**, *7*, 150.
66. Azofeifa, J.; Allen, M.A.; Lladser, M.E.; Dowell, R. FStitch: A Fast and Simple Algorithm for Detecting Nascent RNA Transcripts. Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; ACM: New York, NY, USA, 2014; BCB '14, pp. 174–183.