Article

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

Changes in chromatin accessibility are not concordant with transcriptional changes for single-factor perturbations

Karun Kiani[1], Eric M Sanford[2], Yogesh Goyal[3,4,5] & Arjun Raj[3,6,*]

Abstract

A major goal in the field of transcriptional regulation is the mapping of changes in the binding of transcription factors to the resultant changes in gene expression. Recently, methods for measuring chromatin accessibility have enabled us to measure changes in accessibility across the genome, which are thought to correspond to transcription factor-binding events. In concert with RNA-sequencing, these data in principle enable such mappings; however, few studies have looked at their concordance over short-duration treatments with specific perturbations. Here, we used tandem, bulk ATAC-seq, and RNA-seq measurements from MCF-7 breast carcinoma cells to systematically evaluate the concordance between changes in accessibility and changes in expression in response to retinoic acid and TGF-β. We found two classes of genes whose expression showed a significant change: those that showed some changes in the accessibility of nearby chromatin, and those that showed virtually no change despite strong changes in expression. The peaks associated with genes in the former group had lower baseline accessibility prior to exposure to signal. Focusing the analysis specifically on peaks with motifs for transcription factors associated with retinoic acid and TGF-β signaling did not reduce the lack of correspondence. Analysis of paired chromatin accessibility and gene expression data from distinct paths along the hematopoietic differentiation trajectory showed a much stronger correspondence, suggesting that the multifactorial biological processes associated with differentiation may lead to changes in chromatin accessibility that reflect rather than driving altered transcriptional status. Together, these results show many gene expression changes can happen independently of changes in the accessibility of local chromatin in the context of a single-factor perturbation.

Keywords chromatin accessibility; gene regulation; multi-omics integration; RNA-seq and ATAC-seq concordance; signal response

Subject Category Chromatin, Transcription & Genomics

DOI 10.15252/msb.202210979 | Received 15 February 2022 | Revised 11 August 2022 | Accepted 19 August 2022

Mol Syst Biol. (2022) 18: e10979

Introduction

Transcription factors regulate gene expression by binding to specific DNA sequences, facilitating transcription through the recruitment and activation of the transcriptional machinery. Deciphering the combinatorial logic underlying which transcription factors bind to what portions of DNA and in what contexts are a central challenge in creating a complete model of transcriptional regulation. Sequencing-based methods have enabled the measurement of transcript levels for all genes and the putative binding profiles of transcription factors across the genome. However, the precise mapping between changes in these putative binding profiles and the changes in transcriptional activity remains the subject of debate.

A key component of decoding the relationship between transcription factor activity and the resultant changes in transcription is the measurement of transcription factor binding to DNA. Recently, the combination of biochemical binding assays with sequencing-based readouts has led to a cornucopia of methods for making such measurements. One workhorse method is chromatin immunoprecipitation sequencing (ChIP-seq), which characterizes the binding of transcription factors and other DNA-protein interactions genome-wide (Barski *et al*, 2007; Robertson *et al*, 2007; Ma & Zhang, 2020) by using immunoprecipitation of proteins that bind to chromatin and subsequently sequencing the coprecipitated DNA. However, ChIP-seq is limited in that each experiment can only interrogate the binding profile of one transcription factor at a time.

An alternative approach that circumvents that issue is the measurement of changes in the accessibility of DNA to infer changes in the binding of all transcription factors at once. Accessible regions of DNA (i.e., those regions depleted of nucleosomes) represent only 3% of the genome but often participate in the regulation of gene

1 Genetics and Epigenetics, Cell and Molecular Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
2 Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
3 Department of Bioengineering, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, USA
4 Department of Cell and Developmental Biology, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA
5 Center for Synthetic Biology, Northwestern University, Chicago, Illinois, USA
6 Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
*Corresponding author. Tel: +1 215 821 7394; E-mail: arjunrajlab@gmail.com

expression (Weintraub & Groudine, 1976; Wu *et al*, 1979; Lee *et al*, 2004; Thurman *et al*, 2012). These regions can be detected genome-wide by combining the enzymatic activity of nucleases with high-throughput sequencing using techniques such as DNase I hypersensitive site sequencing (DNase-seq; Boyle *et al*, 2008) and assay for transposase accessible chromatin with sequencing (ATAC-seq; Buenrostro *et al*, 2013). The interpretation of these accessibility methods leans heavily on the assumption that changes in regulatory factor binding are reflected in changes in chromatin accessibility. Certainly, there are many examples in which the correspondence between changes in accessibility strongly corresponds to changes in transcriptional output. For instance, the summation of ChIP-seq signal for 42 transcription factors mapped by encoding in K562 chronic myelogenous leukemia cells paralleled the signal from accessible sites revealed by DNase-seq (Thurman *et al*, 2012). Moreover, computational methods to infer transcription factor footprints from accessibility measurements have been shown to recapitulate ChIP-seq binding well (Pique-Regi *et al*, 2011). Accessibility methods can also be used to look for changes in accessibility across various perturbations and cell types. Changes in accessibility generally seem to correspond to changes in transcription in the sense that large changes in transcriptional output are reflected in broad changes in the accessibility of several loci in the surrounding chromatin (González *et al*, 2015a; de la Torre-Ubieta *et al*, 2018).

However, it is unclear how well these accessibility-based methods capture the activity of all transcription factors. It is possible that some transcription factors' binding and activity do not result in corresponding changes in accessibility and vice versa. Such a lack of correspondence could manifest itself as a lack of correlation between changes in accessibility and changes in transcription. Given the underlying assumption that a change in transcription must be mediated by the change in some transcription factor activity, then such a lack of correspondence would suggest that changes in the activity of transcription factors could change expression without changing accessibility near its binding site. Indeed, previous work has demonstrated that the glucocorticoid receptor binds almost exclusively to pre-existing accessible chromatin prior to small-molecule stimulation (John *et al*, 2011) and that activator protein 1 (AP-1) establishes these binding patterns for the glucocorticoid receptor by maintaining chromatin accessibility (Biddie *et al*, 2011). Similarly, the lineage-defining transcription factor Foxp3 binds to preformed accessible sites established by its structural homolog, Foxo1, to establish regulatory T cell identity (Samstein *et al*, 2012). Of note, this process of regulatory T cell specification via Foxp3 is considered a "late differentiation" process, as the precursor cell state, the mature naive $CD4^+$ T cell is considered mature. While reports from the literature generally show a strong correspondence (González *et al*, 2015a; Ampuja *et al*, 2017; de la Torre-Ubieta *et al*, 2018; Starks *et al*, 2019), it is worth noting that the comparisons in such studies are often across rather different cell types. In such cases, it is possible that the changes in accessibility are not driven by regulation *per se* but rather reflect the consequences of sequential exposure to multiple regulatory factors that characterize the differentiation process. Such accessibility changes could, in principle, signify the reinforcement of genes that are already transcriptionally active genes or could even just appear around actively transcribed genes without any functional role. Disentangling such possibilities could be revealed with the use of single-factor perturbations that more directly affect an individual pathway; however, few such data are available.

Here, we used tandem bulk RNA-seq and ATAC-seq data from MCF-7 breast carcinoma cells exposed to multiple doses of retinoic acid or TGF-β to determine the degree of concordance between changes in chromatin accessibility and changes in gene expression. We demonstrate that while some differentially expressed genes have a high concordance between gene expression and chromatin accessibility changes, many other genes are differentially expressed without changes in their local chromatin accessibility. We show that these results hold across multiple parameters and definitions of accessibility change and that it does not depend on the type of transcription factor per se. We evaluated another published dataset of hematopoietic differentiation, which has much deeper and multifactorial differences, that showed much stronger concordance. We finally compare differences in pre-existing accessibility between concordant and non-concordant genes prior to single-factor perturbation. Our results provide a systematic evaluation of the concordance between changes in gene expression and local chromatin accessibility.

Results

Genome-wide expression and chromatin accessibility changes reflect known biology of two perturbations

To measure the correspondence between changes in chromatin accessibility and changes in gene expression, we used MCF-7 breast carcinoma cells due to their previously described transcriptional responses to all-*trans* retinoic acid (Hua *et al*, 2009; referred to from here on as retinoic acid) and transforming growth factor beta (TGF-β; Mahdi *et al*, 2015). We used paired, bulk accessibility (ATAC-seq) and expression data (RNA-seq) from these cells (Sanford *et al*, 2020a; Data ref: Sanford *et al*, 2020b) collected 72 h after continuous exposure to three different doses of each signal (Fig 1A). We chose this timescale because previous work with MCF-7 cells showed more transcriptional changes at 72 h compared with 24 h after exposure to retinoic acid (Hua *et al*, 2009), and chromatin accessibility changes may not be detectable until 24 h after perturbation (Ramirez *et al*, 2017a).

Differential gene expression and differential peak accessibility analysis showed a dose-dependent response to both signals compared with ethanol control (Fig 1A and Appendix Fig S1A, bar plots). The ethanol "vehicle" controls comprise three different densities of cells, and the transcriptomes of control conditions globally were similar regardless of cell density (Appendix Fig S1B). To confirm that global gene expression and chromatin accessibility patterns were similar between signals and dosages, we performed a principal component analysis. For both RNA-seq and ATAC-seq data, all samples exposed to the same signal or ethanol control clustered together, indicating that their gene expression and chromatin accessibility were more similar to each other than to other conditions, supporting the quality of these data.

To validate that changes in gene expression were consistent with the known biology of these signaling pathways, we performed an over-representation analysis on the upregulated genes in response to high dose retinoic acid or TGF-β against curated gene sets from the molecular signatures database (Liberzon *et al*, 2011, 2015). The

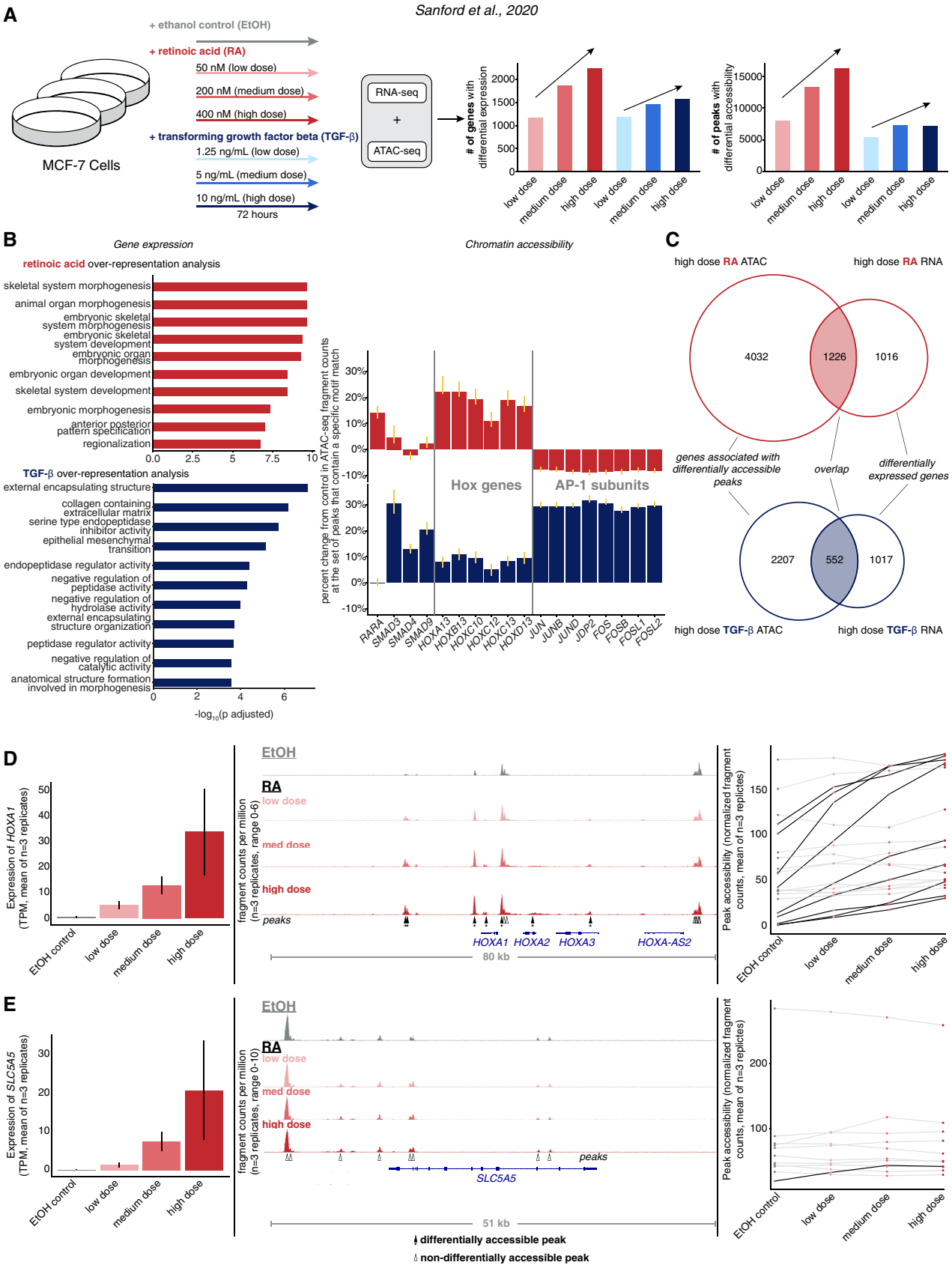
A
Sanford et al., 2020
+ ethanol control (EtOH)
+ retinoic acid (RA)
50 nM (low dose)
200 nM (medium dose)
400 nM (high dose)
+ transforming growth factor beta (TGF-β)
1.25 ng/mL (low dose)
5 ng/mL (medium dose)
10 ng/mL (high dose)
72 hours
MCF-7 Cells
RNA-seq
+
ATAC-seq
# of genes with differential expression
# of peaks with differential accessibility
low dose
medium dose
high dose
B
Gene expression
retinoic acid over-representation analysis
skeletal system morphogenesis
animal organ morphogenesis
embryonic skeletal system morphogenesis
embryonic skeletal system development
embryonic organ morphogenesis
embryonic organ development
skeletal system development
embryonic morphogenesis
anterior posterior pattern specification
regionalization
TGF-β over-representation analysis
external encapsulating structure
collagen containing extracellular matrix
serine type endopeptidase inhibitor activity
epithelial mesenchymal transition
endopeptidase regulator activity
negative regulation of peptidase activity
negative regulation of hydrolase activity
external encapsulating structure organization
peptidase regulator activity
negative regulation of catalytic activity
anatomical structure formation involved in morphogenesis
-log10(p adjusted)
Chromatin accessibility
percent change from control in ATAC-seq fragment counts at the set of peaks that contain a specific motif match
Hox genes
AP-1 subunits
RARA
SMAD3
SMAD4
SMAD9
HOXA13
HOXB13
HOXC10
HOXC12
HOXC13
HOXD13
JUN
JUNB
JUND
JDP2
FOS
FOSB
FOSL1
FOSL2
C
high dose RA ATAC
high dose RA RNA
4032
1226
1016
genes associated with differentially accessible peaks
overlap
differentially expressed genes
2207
552
1017
high dose TGF-β ATAC
high dose TGF-β RNA
D
Expression of HOXA1 (TPM, mean of n=3 replicates)
EtOH control
low dose
medium dose
high dose
EtOH
RA
low dose
med dose
high dose
peaks
fragment counts per million (n=3 replicates, range 0-6)
HOXA1
HOXA2
HOXA3
HOXA-AS2
80 kb
Peak accessibility (normalized fragment counts, mean of n=3 replicates)
E
Expression of SLC5A5 (TPM, mean of n=3 replicates)
EtOH
RA
low dose
med dose
high dose
peaks
fragment counts per million (n=3 replicates, range 0-10)
SLC5A5
51 kb
differentially accessible peak
non-differentially accessible peak
Peak accessibility (normalized fragment counts, mean of n=3 replicates)

Figure 1.

Figure 1. Changes in gene expression can occur with or without concordant changes in chromatin accessibility in response to signal.

A Schematic of signal response experiments in MCF-7 cells from Sanford *et al* (2020a). Briefly, cells were treated with either ethanol vehicle control (gray) or three different doses of retinoic acid (shades of red) or TGF-β (shades of blue). After 72 h of continuous exposure, bulk RNA-seq and ATAC-seq were performed on samples. We show the number of differentially expressed genes and differentially accessible peaks for each dose of each condition compared with ethanol vehicle control.

B Validation that changes in gene expression and chromatin accessibility reflects known biology of perturbations. Left: over-representation analysis of differentially upregulated genes in response to high dose retinoic acid (red) or TGF-β (blue). Top 10 gene sets for each signal by $-\log_{10}$ FDR-adjusted *P*-value are shown. Right: motif enrichment analysis of differentially accessible peaks for selected motifs of transcription factors related to signaling pathways of these signals. Y-axis shows the percentage change of ATAC-seq signal at motif-containing peaks relative to ethanol vehicle control samples. For each condition, we pooled together replicates for all three doses. Error bars represent bootstrapped confidence intervals.

C Overlap between changes in gene expression and changes in chromatin accessibility in response to high dose retinoic acid (top) or high dose TGF-β (bottom). Of the genes that were differentially expressed (right circle of Venn diagram), we looked at the overlap (shaded) of how many of the genes also had at least one differentially accessible peak assigned to it using the "nearest" approach (left circle). We performed Fisher's exact test to show the probability of the joint values of genes with overlapping changes in expression and chromatin accessibility in our data compared with all possible combinations.

D Expression and accessibility change of *HOXA1* in response to increasing doses of retinoic acid. Left: Expression (TPM, average of $n = 3$ biological replicates) in response to increasing dose of retinoic acid (error bars represent SEM). Middle: track view of *HOXA1* locus with accessibility in fragments per million and peaks and differential peaks annotated. Right: quantification of peak accessibility (normalized fragment counts, average of $n = 3$ biological replicates) within a 50 kilobase window of *HOXA1* locus with peaks that are differentially accessible between ethanol vehicle control and high dose retinoic acid conditions marked with black lines.

E Expression and accessibility change of *SLC5A5* in response to increasing doses of retinoic acid. Left: Expression (TPM, average of $n = 3$ biological replicates) in response to increasing dose of retinoic acid (error bars represent SEM). Middle: track view of *SLC5A5* locus with accessibility in fragments per million and peaks and differential peaks annotated. Right: quantification of peak accessibility (normalized fragment counts, average of $n = 3$ biological replicates) within a 50 kilobase window of *SLC5A5* locus with peaks that are differentially accessible between ethanol vehicle control and high dose retinoic acid conditions marked with black lines.

top 10 gene sets based on false discovery rate (FDR)-adjusted P-values were processes canonically associated with retinoic acid (*morphogenesis*, *organ development*, *anterior-posterior patterning*) and TGF-β (*extracellular matrix*, *endopeptidase activity*), respectively (Fig 1B). Gene set enrichment analysis (Subramanian *et al*, 2005) showed that genes that were differentially expressed in response to high dose retinoic acid were significantly enriched for genes associated with skeletal system morphogenesis, and genes that were differentially expressed as a result of exposure to high dose TGF-β were significantly enriched for genes associated with epithelial-to-mesenchymal transition (Appendix Fig S1B). Thus, the differentially expressed genes generally reflected the known biology of the signals the cells were exposed to.

We next wondered if the changes in chromatin accessibility in response to signal were associated with the activity of specific transcription factors, in particular, those associated with the biology of these signaling pathways. We used a modified version of the chromVAR package along with its curated database of transcription factor motifs, cisBP, to identify the transcription factors with the largest predicted change in activity (Schep *et al*, 2017). We used the set of differential peaks to determine the set of the top 150 transcription factors with the greatest magnitude of change. These included the binding motifs of transcription factors that are canonical effectors of retinoic acid (RAR-α, HOXA13) and TGF-β signaling (SMAD3, SMAD4, and SMAD9). For each of these transcription factor motifs, we calculated a motif enrichment score for each condition based on the bias-uncorrected deviation score from chromVAR. The motif enrichment score represents the percentage change in ATAC-seq fragment counts in all peaks that contain a given transcription factor's motif (Fig 1B). For example, the enrichment score of 28% for SMAD3 in the TGF-β condition meant that peaks containing the SMAD3 motif on average saw a 28% increase in fragment counts after exposure to TGF-β. We pooled together the low, medium, and high doses for each condition together in order to decrease the variability of motif enrichment scores estimates. Thus, our data recapitulated expected changes in accessibility, presumably due to the activity of transcription factors well-known to be activated by the signals used. Thus, of the changes in accessibility we did detect, they made sense based on a model of transcription factor activity leading to changes in accessibility. However, it was still possible that the activity of many transcription factors was not captured by changes in accessibility.

The relationship between changes in chromatin accessibility and gene expression varies on a gene-by-gene basis

We next wondered whether genes that were differentially expressed were more likely to have differentially accessible peaks nearby, i.e., was there concordance between gene expression and chromatin accessibility changes at the level of individual genes? To initially characterize the extent of concordance between these data, we looked at the overlap between genes that were differentially expressed in response to high dose signal and genes with differentially accessible peaks nearby after exposure to signal (Fig 1C). We assigned each accessible peak to the nearest transcriptional start site ("nearest approach"). Using this approach, the majority of genes had fewer than 20 peaks assigned to them (Appendix Fig S2A). and found that of the over 2,000 genes upregulated in response to high dose retinoic acid, more than half of them had at least one differential peak (irrespective of the direction of peak change) assigned to its transcriptional start site (P-value $< 2.2 \times 10^{-16}$, Fisher's exact test). Similarly, a third of the genes whose expression was upregulated in response to TGF-β had differential peaks assigned to them (P-value $< 2.2 \times 10^{-16}$, Fisher's exact test). For differentially expressed genes in response to high dose retinoic acid or TGF-β, approximately 75 and 81% of genes had all peaks either not differentially accessible or differentially accessible in the same direction of gene expression changes, respectively (Appendix Fig S2B). Thus, using the "nearest" approach, genes that are differentially expressed are more likely than random chance to have a nearby peak that is differentially accessible in response to retinoic acid or TGF-β.

While using this overlap-based approach showed correspondence between genes that are differentially expressed and their nearby peaks in response to signal, aspects of the nature of the concordance of these changes were not captured by this analysis. For example, the overlap-based method counted all differentially accessible genes

that had at least one differentially accessible peak assigned to them as concordant but did not take into account the proportion or degree to which those nearby peaks change. Moreover, we did not take into account the relationship between the directionality of changes in gene expression and chromatin accessibility. The underlying assumption at the basis of this relationship is that when peaks become more accessible than the nearby gene increases its expression, and the overlap-based approach does not take this correspondence of the direction of change into account. To better characterize these facets of concordance, we first individually examined the changes in chromatin accessibility nearby two genes whose expression was upregulated in response to retinoic acid.

After optimizing parameters for calling peaks and determining differentially accessible peaks (Appendix Fig S3), we found that while a large number of peaks are differentially accessible near the *HOXA1* locus, very few peaks are differentially accessible near the *SLC5A5* locus (Fig 1D and E, track view middle, accessibility plot, right; Appendix Fig S4A and B). *HOXA1* and *SLC5A5* induction are associated with exposure to retinoic acid (Schmutzler *et al*, 1997; Kogai *et al*, 2000; Glover *et al*, 2006), and both genes showed a dose-dependent increase in expression in response to retinoic acid (Fig 1D and E leftmost panels; Appendix Figs S4A and B). Therefore, genes with high expression change in response to signal can show a large degree of accessibility changes or show very little accessibility changes, suggesting that changes in transcription factor activity may or may not be reflected in changes in accessibility.

Chromatin accessibility changes are less concordant with large changes in gene expression in signaling compared with hematopoietic differentiation

Next, we evaluated the concordance between accessibility and gene expression genome-wide while also factoring in the directionality of changes and the relative proportion of peaks that are changing on a gene-by-gene basis. As a point of comparison, we used previously published gene expression and chromatin accessibility data from hematopoietic differentiation (González *et al*, 2015a; Data ref: González *et al*, 2015b) that demonstrated that large changes in gene expression were typically associated with gains or losses (depending on the direction of expression change) of cell type-specific enhancers when comparing the expression and accessibility of hematopoietic stem and progenitor cells (HSPCs) to monocytes.

Before using this dataset as a comparison to ours for measuring concordance between chromatin accessibility and gene expression changes, we verified that the hematopoietic differentiation data was similar to our own by a variety of metrics. First, we wanted to compare whether the number of differentially expressed genes and differentially accessible peaks between HSPCs and monocytes in the hematopoietic differentiation data was similar to the numbers from MCF-7 cells exposed to retinoic acid or TGF-β. We found that both HSPC and monocyte populations had greater than 2,000 genes that were specifically expressed in their respective cell types compared with the approximately 2,000 and 1,500 genes differentially expressed in MCF-7 cells in response to high dose retinoic acid and TGF-β, respectively (Fig 1A). Moreover, HSPC and monocyte populations had more than 6,000 differentially accessible peaks (Appendix Fig S5A) compared with the approximately 15,000 and 6,000 differentially accessible peaks in MCF-7 cells in response to high dose retinoic acid and TGF-β, respectively (Fig 1A).

Next, we annotated the location of peaks based on where in the genome they were located relative to gene bodies and quantified what proportion of peaks fell into annotation categories such as promoter, intergenic, exonic, intronic, etc. ATAC-seq peaks from MCF-7 cells had a larger proportion of peaks at gene promoters (within 3 kilobases upstream or downstream of the transcription start site) whereas a greater proportion of the DNase I hypersensitive sites in the HSPC and monocyte populations were from distal intergenic regions compared with promoters (Appendix Fig S5B). This finding could be the result of inherent differences in the assays or could reflect biological differences. Moreover, the MCF-7 data had a greater proportion of peaks located at gene promoters, which could in principle bias our results toward having a larger degree of concordance because accessibility changes at promoters were more strongly correlated with gene expression changes than distal accessible. Despite this bias, our data demonstrate less concordance.

Given the different assays used to determine genome-wide chromatin accessibility, we realigned the DNase-seq data to the hg38 reference and examined the peaks at a "housekeeping gene" (*GAPDH*), hematopoietic differentiation-specific genes (*CD34*, *CD14*) and retinoic acid and TGF-β-related genes (*DHRS3*, *SERPINA11*) to spot-check that the accessibility data were similar. Indeed, there were similar accessibility profiles for *GAPDH*, and appropriate differences in accessibility given the cell type of signal for the other sites, indicating the accessibility data were comparable (Appendix Fig S6A–E). Moreover, to look at similarities in accessibility genome-wide, we calculated the intersection of the consensus peak sets from hematopoietic differentiation and MCF-7 signal response datasets, which included both peaks that were differentially accessible and those that were not. We observed that approximately 55% of peaks from hematopoietic differentiation data (DNase-seq) overlapped with peaks from the MCF-7 signal response dataset (ATAC-seq). These results show that the datasets do not have systematic qualitative differences in either expression or accessibility, enabling us to compare the degree of concordance across these two systems.

In the original analysis of hematopoietic differentiation, the authors found that regulatory complexity (defined as the number of accessible regions closest to a gene's transcriptional unit) was an important discriminating factor for whether changes in accessibility corresponded to changes in expression, with areas of high complexity showing more correspondence than those of low complexity. Hence, we similarly grouped genes from our MCF-7 dataset into high and low complexity for our comparisons. We categorized genes with more than 7 peaks assigned to them using the "nearest approach" as "high complexity," while genes with 7 or fewer peaks were categorized as having "low complexity" (Fig 2A, top panel). The cutoff for loci complexity was calculated by taking a tertile-based approach (González *et al*, 2015a) and calling any number of peaks above the highest tertile cutoff as high and any peak below that as low complexity (Fig 2B, solid line, lower plot). Because high complexity genes on average had higher levels of expression in the hematopoietic differentiation data, we sought to determine whether there was any difference in expression between high and low complexity genes in our MCF-7 data. The median expression of high complexity loci was similarly higher than low complexity loci in

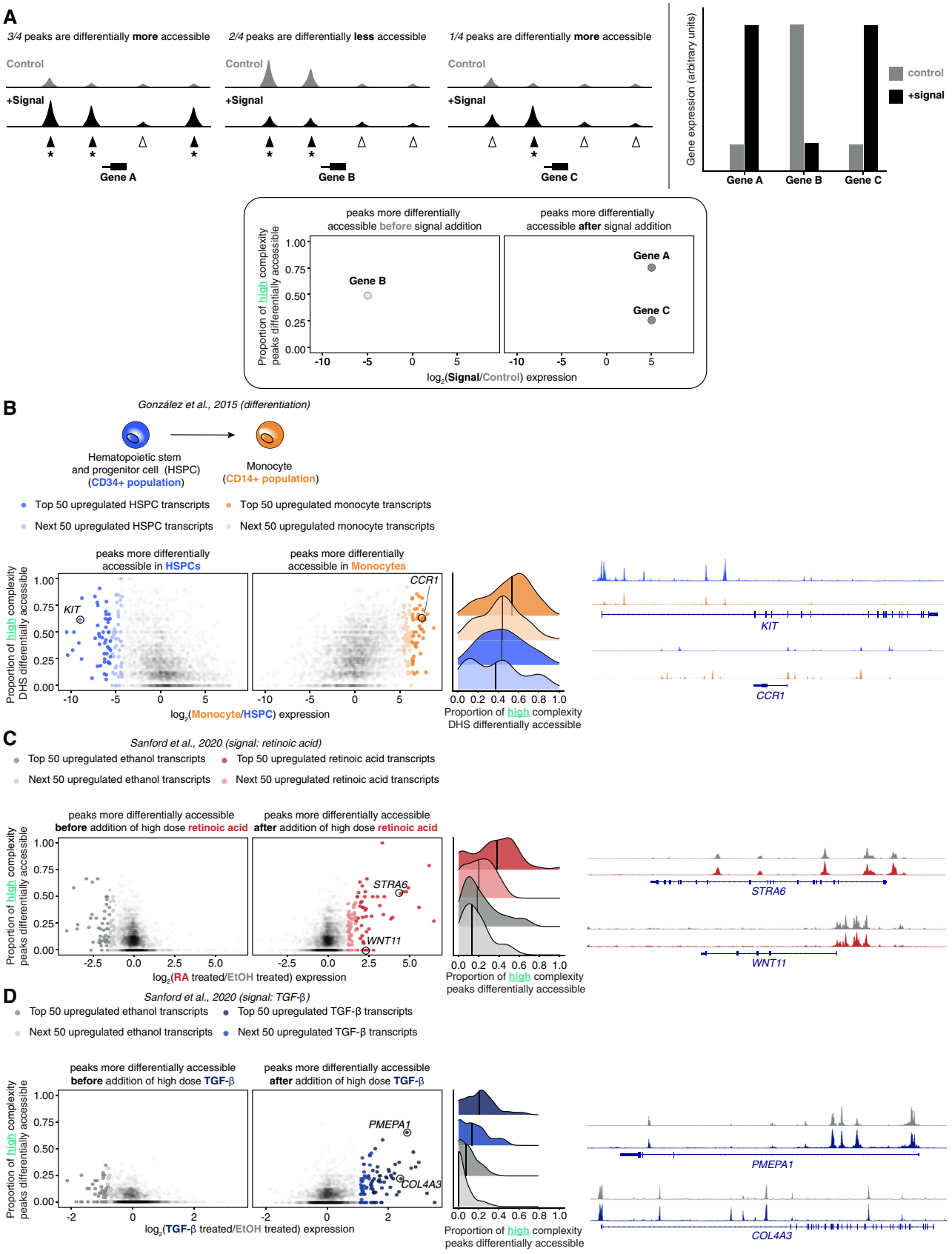

Figure 2.

Figure 2. Signaling shows less concordance between highly differentially expressed genes and chromatin accessibility changes compared with hematopoietic differentiation data for high complexity genes.

A Schematic demonstrating how data from chromatin accessibility and gene expression data are used to create a proportion-based concordance plot.

B Concordance between expression and accessibility changes between hematopoietic stem and progenitor cells and monocytes. Left: plot showing changes in gene expression in $CD34^+$ hematopoietic stem and progenitor cells (blue) and $CD14^+$ monocytes (orange) from González *et al*, 2015a (schematic, top). For the plots, each dot is a gene, and on the *x*-axis is a log2 fold change in expression and on the *y*-axis the proportion of differentially accessible DHSs for each associated gene. The top 100 most highly expressed genes in hematopoietic stem and progenitor cells and monocytes are colored in shades of orange and blue, respectively. Middle: density plot of the distribution of the proportion of high complexity DHS associated with the top 100 expressed genes in $CD34^+$ hematopoietic stem and progenitor cells and $CD14^+$ monocytes with median value marked by a vertical black line. Right: example tracks DNase I sequencing data for KIT and CCR1 (marked on plot on left).

C Concordance between expression and accessibility changes between cells exposed to ethanol vehicle control and high dose retinoic acid. Left: plot showing changes in gene expression and chromatin accessibility between ethanol vehicle control and high dose retinoic acid. Each dot is a gene, and on the *x*-axis is the log2 fold change in expression and on the *y*-axis the proportion of differentially accessible ATAC-seq peaks for each gene. The top 100 most highly expressed genes in ethanol vehicle control and high dose retinoic acid are colored in shades of gray and red, respectively. Middle: density plot of the distribution of the proportion of high complexity ATAC-seq peaks associated with the top 100 expressed genes in ethanol vehicle control and high dose retinoic acid with median value marked by a vertical black line. Right: example ATAC-seq tracks of STRA6 and WNT11.

D Concordance between expression and accessibility changes between cells exposed to ethanol vehicle control and high dose TGF-β. Left: plot showing changes in gene expression and chromatin accessibility between ethanol vehicle control and high dose TGF-β. Each dot is a gene, and on the *x*-axis is the log2 fold change in expression and on the *y*-axis the proportion of differentially accessible ATAC-seq peaks for each gene. The top 100 most highly expressed genes in ethanol vehicle control and high dose TGF-β are colored in shades of gray and blue, respectively. Middle: density plot of the distribution of the proportion of high complexity ATAC-seq peaks associated with the top 100 expressed genes in ethanol vehicle control and high dose retinoic acid with median value marked by a vertical black line. Right: example ATAC-seq tracks of *PMEPA1* and *COL4A3*.

response to both exposure to high dose retinoic acid (23.30 versus 13.27 TPM) and high dose TGF-β (24.06 vs. 13.05 TPM) (Appendix Fig S7C, P-value $< 2.2 \times 10^{-16}$ for both, Kolmogorov–Smirnov test) demonstrating that high complexity genes are more highly expressed as in the hematopoietic differentiation data. Despite this difference in expression, the distributions of peak widths for peaks of high and low complexity genes were similar (Appendix Fig S7D).

We began our analysis by focusing on the high complexity genes. To determine the concordance between gene expression changes and chromatin accessibility changes, we used the "nearest approach" to assign peaks to genes. For each gene, we compared the $\log_2$ of the fold change in expression between conditions versus the proportion of peaks that were differentially accessible in the same direction (i.e., peaks that increase in accessibility for genes that increase in expression after exposure to signal and vice versa). We observed that for hematopoietic differentiation, the 100 most highly expressed high complexity genes in the HSPC and monocyte populations had a high proportion of peaks, which were differentially accessible in the concordant direction, reproducing the conclusions of González *et al* (2015a) that large changes in expression were consistently associated with concordant changes in chromatin accessibility (Fig 2C). Next, we used this approach on our data to compare expression and accessibility changes between ethanol vehicle control and high dose retinoic acid or TGF-β. For both signals, we observed two distinct groups of genes within the top 100 most differentially expressed genes. One group of genes ("accessibility-concordant genes") behaved similarly to those in the hematopoietic differentiation data, demonstrating a concordance between expression and accessibility changes (Fig 2C and D). However, the other group of genes ("accessibility-nonconcordant genes") had large expression changes with little to no peaks nearby changing in accessibility, creating a skew in the distribution toward a lower proportion of peaks being differentially accessible in a concordant manner compared with the hematopoietic differentiation data (Fig 2B–D, density plots).

Adjusting the minimum peak coverage parameter changes the number of differential peaks and the proportion of differential peaks that change in the corresponding direction of expression. We wondered if a lower minimum coverage threshold changed the qualitative result we noticed before and thus conducted the same analysis using a lower minimum peak coverage threshold for determining differential peaks (see methods). We observed that a similar pattern occurred in high complexity genes with this set of parameters (Appendix Fig S8).

González *et al* (2015a) showed that for some low complexity genes, large changes in expression were not accompanied by concordant changes in accessibility (González *et al*, 2015a). We similarly wanted to confirm whether this decreased correspondence was the case in our data in response to retinoic acid and TGF-β. Using the same approach as before, we compared the $\log_2$ of the fold change in expression of low complexity genes to the proportion of peaks with differential accessibility in the concordant direction. The hematopoietic differentiation and signaling data for low complexity all qualitatively had genes whose expression increased without concordant changes in accessibility (Appendix Fig S9A–C). The distribution of the proportion peaks that was differentially accessible in the concordant direction for the top 100 up and downregulated genes was roughly uniform when comparing HSPCs to monocytes (Appendix Fig S9A, density plot on right). By comparison, the distribution was skewed toward more genes having a lower proportion of peaks being differentially accessible in the concordant direction in response to signals in MCF-7 cells, especially in the case of TGF-β (Appendix Fig S9B and C, density plots on right). However, it is also possible that these differences may owe to systematic differences between the datasets, given that the hematopoietic differentiation data have twice as many differentially accessible peaks as the MCF-7 data. Thus, while both the signaling in MCF-7 and hematopoietic data demonstrated large gene expression changes without concordant changes in chromatin accessibility with low complexity genes, a greater proportion of genes did so in the signaling data.

Peaks nearby genes with high concordance have lower accessibility prior to exposure to signal

We wondered what the differences were between genes that were differentially expressed and had large accessibility changes versus those that were differentially expressed and had low accessibility changes. First, for high dose retinoic acid and TGF-β, we split genes into four groups based on whether they were differentially expressed and the proportion of peaks assigned to them using the "nearest" method that was differentially accessible in the appropriate direction. These four groups were (i) genes with differentially upregulated expression and concordant accessibility changes (ii) genes with differentially upregulated expression nonconcordant accessibility changes (iii) genes with differentially downregulated expression and concordant accessibility changes, and (iv) genes with differentially downregulated expression and nonconcordant accessibility changes (Fig 3A and B). We quantified the distribution of peak complexity across these groups and observed that genes with any concordant peaks for both conditions in both directions tended to have a higher degree of locus complexity (Appendix Fig S10A and B).

We first asked whether the change in accessibility between these two gene groups was due to differences in the pre-existing accessibility of peaks for these genes. Indeed, we found the baseline accessibility of peaks for genes with concordant increases in expression and accessibility in ethanol vehicle conditions was lower than those of peaks of genes that increase in expression without a commensurate change in chromatin accessibility (Fig 3C). This relationship was also recapitulated for concordant peaks that increase in expression and accessibility in response to high dose TGF-β (Fig 3D). Similarly, when comparing genes that are differentially downregulated in expression a similar pattern holds true in the opposite direction (Fig 3C and D; Appendix Fig S10C and D). One explanation may be that genes whose nearby chromatin was already accessible were permissive toward the action of the appropriate transcription factors to modulate expression. An alternative explanation is that the ATAC-seq assay itself had saturated in its ability to measure chromatin accessibility. By contrast, the difference in accessibility decreased between genes with a low proportion of peaks that were differentially accessible and genes with a high proportion of accessible peaks after exposure to signal (Appendix Fig S10C and D). Thus, the difference in the proportion of accessible peaks nearby the two groups of genes was partially explained by the pre-existing chromatin accessibility.

Multiple approaches to integrating chromatin accessibility and gene expression changes show a low degree of concordance during signaling

Finally, we measured to what degree the change in the accessibility of chromatin nearby a gene is reflected in the change in gene expression. Because linear distance is not always a good predictor of what accessible regions interact with what genes, we used multiple approaches to assign peaks to genes. First, we used the "nearest approach" to create a one-to-one mapping between accessible sites and genes by assigning them to the nearest transcriptional start site (Li *et al*, 2012; Nair *et al*, 2021), again comparing our signaling dataset to the hematopoietic differentiation dataset. Because many genes have multiple peaks assigned to them, we used two methods for collapsing peak values per gene: either the median accessibility of peaks across genes or the maximum (Fig 4A, schematic). We observed a stronger correlation between accessibility and expression changes in differentiation data (median approach Pearson's $r = 0.34$, maximum approach Pearson's $r = 0.26$) than in MCF-7 in response to signal (retinoic acid: median approach Pearson's $r = 0.27$, maximum approach Pearson's $r = 0.10$; TGF-β: median approach Pearson's $r = 0.27$, maximum approach Pearson's $r = 0.10$; Fig 4A, right side).

Next, we used a window-based approach where there was the possibility of many-to-one mapping of peaks to genes. We assigned all peaks within a 100 kilobase window (Sanford *et al*, 2020a) in order to maximize the number of differential peaks assigned to a gene (Appendix Fig S11A and B). Similar to the "nearest" approach, we collapsed values using median accessibility change across all peaks assigned to a gene and maximum accessibility per gene (Fig 4B, schematic) We observed a similar effect using this approach where there was a stronger correlation between change in accessibility and change in expression between HSPC versus monocyte versus MCF-7 cells exposed to signal (Fig 4B). Of note, the correlation coefficients were similar between both methods of assigning peaks. Additionally, we used the window-based method to subset promoter-proximal peaks (i.e., within 1.5 kilobase pairs up or downstream from the transcription site) and distal peaks (greater than 20 kilobase pairs from the transcriptional site). This approach similarly did not demonstrate a strong relationship in the concordance between chromatin accessibility changes and gene expression changes in response to retinoic acid or TGF-β (Appendix Fig S12A and B). We also used a variable window approach by restricting our analyses to all peaks within the same topologically associating domain (TAD), which also did not demonstrate a strong correlation between changes in accessibility and changes in gene expression (Appendix Fig S12C).

We also wondered if the correlation between the extent of chromatin accessibility changes and gene expression changes would be different at the two lower doses. We used both the median and maximum peak value per gene while assigning peaks to genes using the nearest and window approaches. We observed a similarly weak correlation as high dose signal using all methods at both low and medium doses (Appendix Fig S11C and D). Consequently, the correlation between the magnitude of change in gene expression and chromatin accessibility was modest across the range of doses of signals.

To see whether peaks in specific genomic regions (promoters, parts of the gene body, downstream and intergenic areas) had unique relationships between change in chromatin accessibility and change in gene expression, we subsetted our correlation analysis. We annotated peaks using ChiPseeker (Yu *et al*, 2015) to categorize them as being at promoters, within the gene body (5′ UTR, 3′ UTR, intronic, and exonic sequences), downstream of the gene end, or at intergenic sequences. We used peaks assigned to genes using the "nearest" approach and took the median change in accessibility per gene. The strongest correlation between changes in accessibility and gene expression across sets of comparisons was at promoter peaks (Fig 4C). While promoter correlation is quantitatively stronger, the overall qualitative conclusion remains the same. We also quantified changes in intronic reads and compared them with changes in

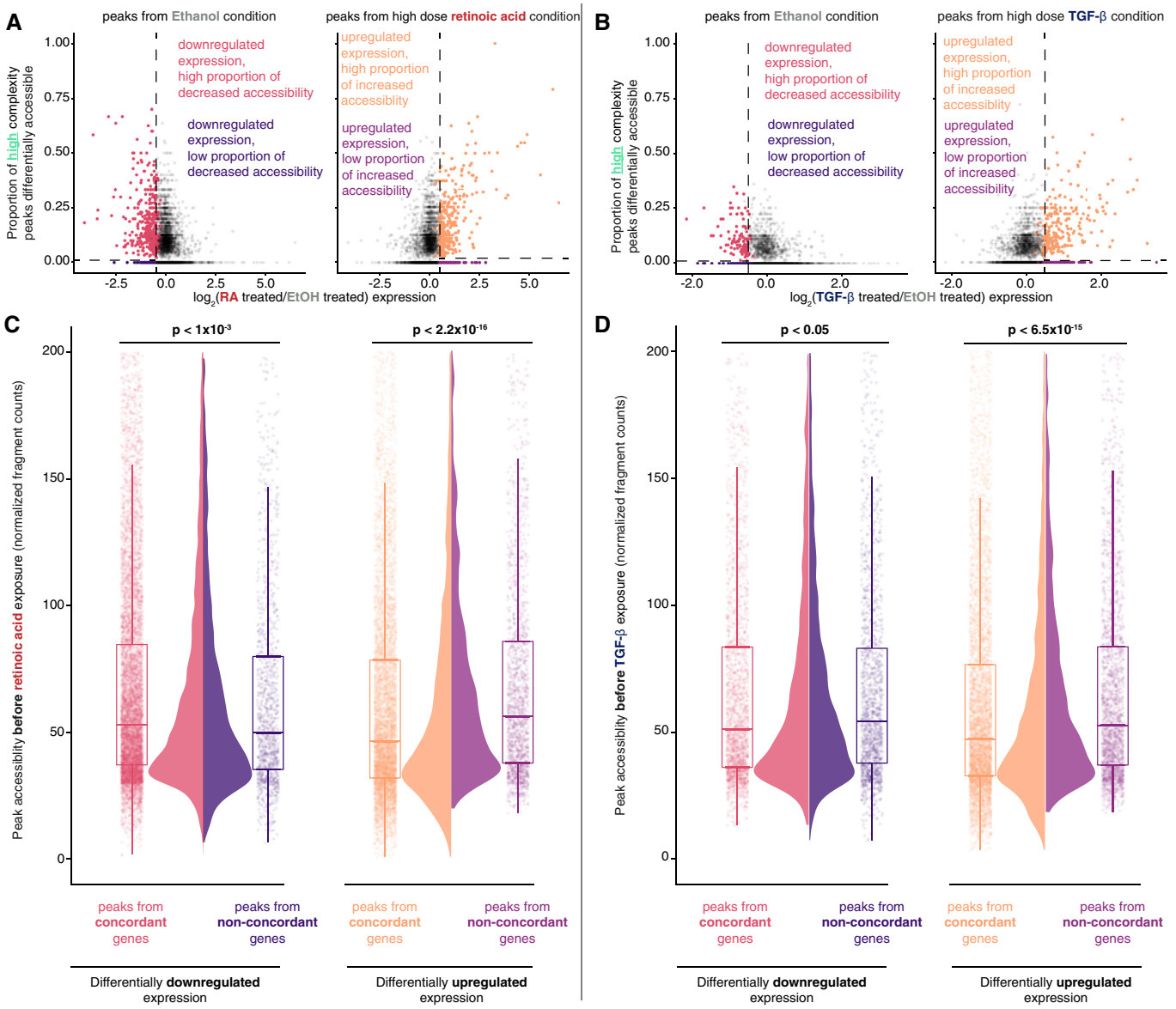

Figure 3. Separation of differentially expressed genes in response to signal into high and low concordance groups shows differences in pre-existing accessibility.

A Categorization of differentially expressed genes in response to high dose retinoic acid based on the direction of expression change and proportion of peaks differentially accessible in the same direction.

B Categorization of differentially expressed genes in response to high dose TGF-β based on the direction of expression change and proportion of peaks differentially accessible in the same direction.

C Differential accessibility in ethanol vehicle control conditions prior to addition of high dose retinoic acid. Accessibility of every peak assigned using the "nearest" approach for gene groups from (A) in ethanol vehicle control conditions. Each peak must be present in at least the majority of n = 3 biological replicates to be used for analysis. *P*-values represent the probability of these data or more extreme under the null hypothesis that the distribution of peak accessibilities was drawn from the same probability distribution via the Kolmogorov–Smirnov test. Box and whisker plot: central band—median, box: 25th and 75th percentiles also known as the interquartile range (IQR), whiskers: 25th percentile—1.5*IQR and 75th percentile + 1.5*IQR.

D Differential accessibility in ethanol vehicle control conditions prior to addition of high dose TGF-β. Accessibility of every peak assigned using the "nearest" approach for gene groups from (B) in ethanol vehicle control conditions. Each peak must be present in at least the majority of n = 3 biological replicates to be used for analysis. *P*-values represent the probability of these data or more extreme under the null hypothesis that the distribution of peak accessibilities was drawn from the same probability distribution via the Kolmogorov–Smirnov test. Box and whisker plot: central band—median, box: 25th and 75th percentiles also known as the interquartile range (IQR), whiskers: 25th percentile—1.5*IQR and 75th percentile + 1.5*IQR.

exonic reads in order to determine whether there is a stronger relationship between more nascent RNA changes and accessibility changes. However, the quantitative relationship was no better than those using other methods (Appendix Fig S13). Thus, despite using a variety of approaches for both assigning peaks to genes and collapsing the accessibility of all peaks for a given gene to a single

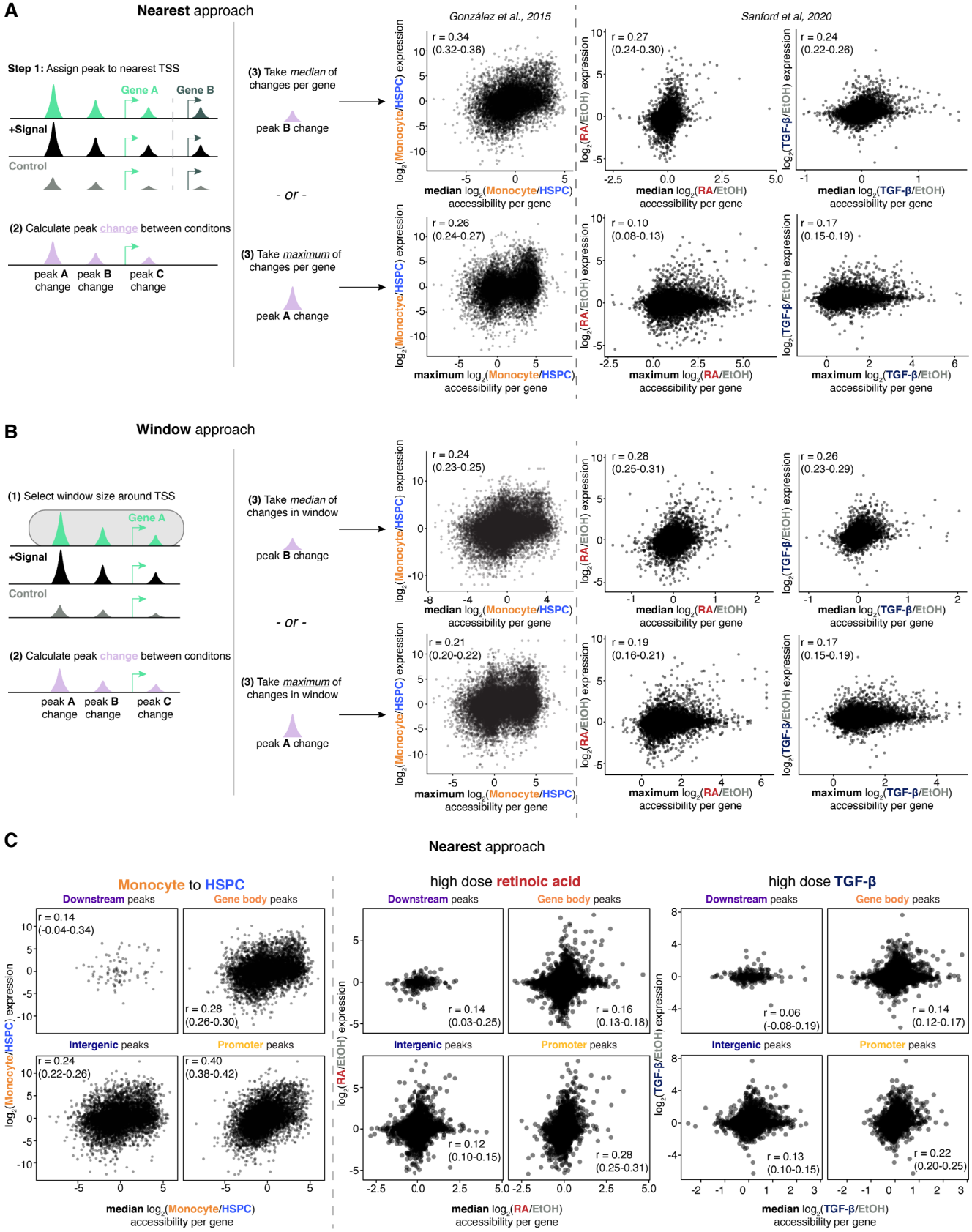

Figure 4.

Figure 4. Multiple approaches to quantifying peak accessibility show a low correlation between gene expression changes and accessibility changes in signaling.

A "Nearest" approach to assigning peaks to genes shows less concordance in signaling compared with hematopoietic differentiation. Left: schematic showing "nearest" approach where peaks are assigned to the nearest transcriptional site and change in accessibility (purple) on a per-gene basis is calculated by either median change in accessibility (top row) or maximum peak change (bottom row). Right: scatterplots showing the change in peak accessibility (median or maximum) versus $\log_2$ fold change in expression on y-axis for hematopoietic differentiation data from González *et al* (2015a) (left column) and for high dose retinoic acid and high dose TGF-β (right two columns). Pearson's correlation coefficients were reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

B "Window" approach to assigning peaks to genes shows less concordance in signaling compared with hematopoietic differentiation. Left: schematic showing "window" approach where all peaks within a certain window of the transcriptional start site are assigned to that gene and the change in accessibility (purple) on a per-gene basis is calculated by the median change in accessibility (top row) or the maximum change in accessibility (bottom row). Right: scatterplots showing the change in peak accessibility (median or maximum) using "window" approach with a 100 kilobase window versus $\log_2$ fold change in expression on y-axis for hematopoietic differentiation data from González *et al* (2015a) (left column) and for high dose retinoic acid and high dose TGF-β (right two columns). Pearson's correlation coefficients were reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

C Using the "nearest" approach to look for correlation between accessibility and gene expression changes based on annotations of peak location. First two columns showing correlation for hematopoietic differentiation data from González *et al* (2015a), and right four columns showing correlation for high dose retinoic acid and high dose TGF-β, respectively. Pearson's correlation coefficients were reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

value, we failed to appreciate a strong relationship between changes in accessibility and changes in gene expression.

Finally, we wondered if peaks that contained the motifs of transcription factors that are associated with retinoic acid and TGF-β signaling only (as opposed to all peaks) would show a stronger correlation between the changes in chromatin accessibility and gene expression. We annotated peaks with a log-likelihood score of a given motif being found in that peak and subsetted those peaks with a nonzero log-likelihood score to examine the correlation between changes in accessibility and gene expression. Using this approach, we examined log-likelihood scores for motifs associated with retinoic acid signaling (RARA-α, HOXA13, and FOXA1) and motifs associated with TGF-β (SMAD3, SMAD4, and SMAD9). We observed that focusing on peaks annotated with peaks we would *a priori* expect to be involved in modulating gene expression in response to signal showed a limited correlation between changes in chromatin accessibility and changes in gene expression (Fig 5A and B). We also used this approach to look for peak-expression correlations for transcription factors downstream of IL-1α (NFκB1, REL, and RELA, of which NFκB1 has pioneer activity) and similarly found little correlation between chromatin accessibility and gene expression changes (Appendix Fig S14B). These results suggest that particular transcription factors show no more concordance between peak changes and expression changes, even for pioneer transcription factors. To exclude the possibility that "subregions" within unchanging peaks could be facilitating transcription factor binding, we measured RAR-α and SMAD footprints within these peaks. The number of reads between control and exposure conditions did not change for these footprints, indicating that there are no more accessible subregions that could mediate transcription factor binding within peaks that were not differentially accessible (Appendix Fig S15A and B).

Discussion

Here, we integrated tandem, genome-wide chromatin accessibility, and transcriptomic data to characterize the extent of concordance between them in response to inductive signals. We demonstrated that while certain genes have a high degree of concordance of change between expression and accessibility changes, there is also a large group of differentially expressed genes whose local chromatin remains unchanged. By comparison, data from cell types along the hematopoietic differentiation trajectory had a much higher degree of concordance between genes with large gene expression changes and chromatin accessibility changes.

What might explain the lack of concordant changes in chromatin accessibility? One explanation could be that pre-existing chromatin accessibility dictates the *de novo* binding of transcription factors, but that the binding of transcription factors to those regions does not result in further changes to accessibility. Such effects have been reported in the context of glucocorticoid signaling, in which the glucocorticoid receptor almost exclusively binds to chromatin that is already accessible in response to dexamethasone (John *et al*, 2011). Indeed, we demonstrated that genes that lacked concordance between changes in chromatin accessibility and gene expression were more likely to have nearby chromatin that was already accessible (Fig 3C and D). It is possible that in MCF-7 cells, the transcriptional effects of RA and TGF-β do not lead to a significant change in the activity of pioneer transcription factors, which are able to bind directly to condensed or inaccessible chromatin to facilitate its opening (Zaret, 2020). Also, implicit in our approach is the assumption that an increase in accessibility is associated with an increase in expression, which is not necessarily the case if a genomic locus becomes accessible to a repressive factor or a bound repressive factor is displaced by a nucleosome.

We looked at MCF-7 cells exposed to retinoic acid and TGF-β because these two signals induce a robust transcriptional response through distinct mechanisms. RAR-α remains bound to DNA and interacts with transcriptional activators in response to retinoic acid binding, while SMAD family members require TGF-β to bind to surface receptors to translocate to the nucleus. Yet, despite these differences, we observed that many genes changed expression independent of changes in chromatin accessibility for both signals. It is, however, possible that signaling molecules that exert their effects through very different types of transcription factors may have a different profile of concordance between changes in accessibility and gene expression. It is possible that other types of factors in a different context (e.g., different cell lines) may yield a stronger correspondence (Appendix Fig S14). Indeed, some acute stimuli can lead to more concordance (Appendix Fig S16A). Potential reasons for this difference are the ability of transcription factors such as NFκB to rapidly decondense heterochromatin to quickly mediate inflammatory responses (Jurida *et al* 2015a; Weiterer *et al*, 2020a). Further, systematic studies of a number of signaling pathways and timescales will be required to make a complete determination of the degree of concordance in various contexts.

Our data characterized molecular changes resulting from a single input (retinoic acid or TGF-β) in a clonal cell line, whereas the majority of work reported a stronger concordance between simultaneous measurements of accessibility and transcription compared with entirely different cell types or cells undergoing a directed differentiation protocol. What we have observed in the case of a single perturbation applied to cells that are not thought to change type *per se* is increased or decreased transcription with less concomitant nearby change in accessibility. How can one reconcile these observations? One possibility is that if we were to leave the signal on for longer, or combine it over time with the effects of several other signals, that we eventually would observe many further changes in accessibility proximal to a gene, concordant with the aforementioned results from comparisons between cell types. Whatever the source, these further

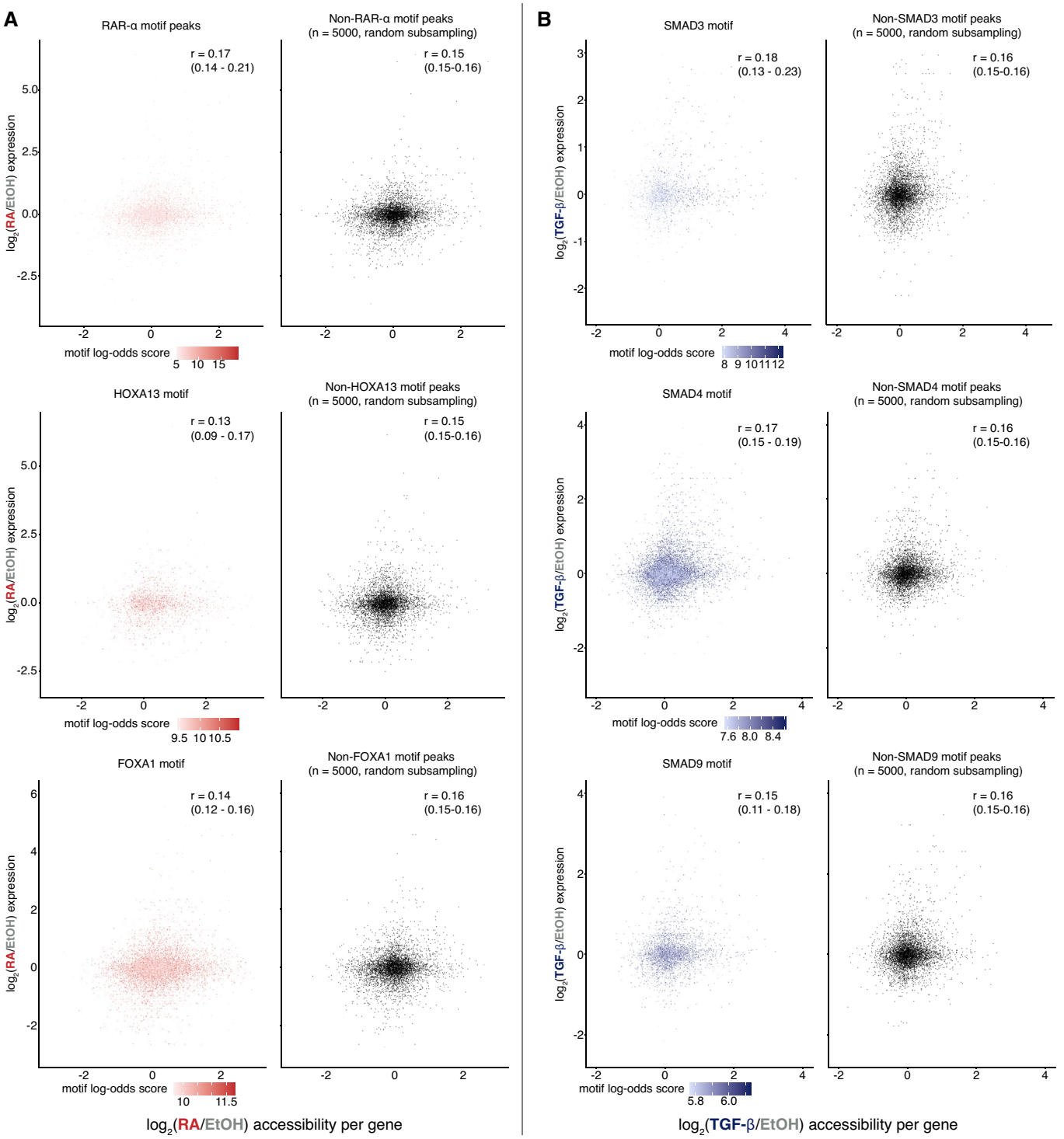

Figure 5.

Figure 5. Focusing on peaks annotated for biologically relevant transcription factor motifs fails to demonstrate a strong correlation between the magnitude of gene expression and chromatin accessibility changes.

A Peaks annotated for motifs of transcription factors related to retinoic acid biology (RAR-α, HOXA13, FOXA1, left column) showed weak correlation between changes in gene expression and chromatin accessibility in response to high dose retinoic acid. Peaks are colored based on the log-odds of a motif being present in a given peak. Plot of expression and accessibility change for 5,000 randomly sampled peaks lacking the corresponding peak (right column). Pearson's correlation for peaks not having a given motif is for all peaks without that motif, not the 5,000 subsampled peaks. Pearson's correlation coefficients were reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

B Peaks annotated for motifs of transcription factors related to retinoic acid biology (SMAD3, SMAD4, SMAD9, left column) showed weak correlation between changes in gene expression and chromatin accessibility in response to high dose TGF-β. Peaks are colored based on the log-odds of a motif being present in a given peak. Plot of expression and accessibility change for 5,000 randomly sampled peaks lacking the corresponding peak (right column). Pearson's correlation for peaks not having a given motif is for all peaks without that motif, not the 5,000 subsampled peaks. Pearson's correlation coefficients were reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

changes in accessibility do not seem to occur randomly, given that they largely reflect the direction of change in transcription (increased accessibility for upregulation, decreased for downregulation). It may be that these subsequent changes in accessibility do not explicitly change transcription but rather alter the underlying regulatory logic of the gene, i.e., the removal of a signal may not lead to a decrease in the gene's transcription, or the gene's transcription may be sensitized or desensitized to some other set of transcription factors.

Materials and Methods

PCA of RNA and ATAC-sequencing samples

Principal component analysis and visualization of RNA-seq and ATAC-seq samples were performed using raw counts and performing a variance stabilizing transform. Results were visualized using functions from the R DESeq2 package (Love *et al*, 2014).

RNA-sequencing analysis

Initial RNA-sequencing analysis was performed as previously (preprint: Goyal *et al*, 2021). Briefly, reads were aligned to the hg38 assembly using STAR v.2.7.1a and counted uniquely mapped reads with HTSeq v0.6.1 and hg38 GTF file from Ensembl (release 90). We used DESeq2 v1.22.2 in R 3.5.1 using a minimum absolute-value log-fold change of 0.5 and a q value of 0.05. For genes with multiple annotated transcriptional start sites, we used the "canonical" transcription start site from the knownCanonical table from GENCODE v29 in the UCSC Table Browser.

We performed functional over-representation and gene set enrichment analysis (Subramanian *et al*, 2005) of upregulated transcripts in the high dose retinoic acid and high dose TGF-β using clusterProfiler v4.0.5 and enrichplot v1.12.3 (Wu *et al*, 2021). *P*-values for the over-representation analysis were adjusted using a false discovery rate approach. We used the C5 ontology and H hallmark curated gene sets from the Molecular Signatures Database (MSigDB) v7.4 (Liberzon *et al*, 2011, 2015) as reference gene sets to compare our upregulated genes.

ATAC-sequencing analysis

ATAC-seq alignment and peak calling were performed as previously (Sanford *et al*, 2020a). We aligned peaks to the hg38 assembly using bowtie2 v2.3.4.1 and filtered out low-quality alignments with samtools v1.96, removed duplicate read pairs with picard 1.96, and used custom Python scripts along with bedtools v2.25.0 to create alignment files with inferred Tn5 insertion points. We called peaks using MACS2 (Zhang *et al*, 2008) v2.1.1.20160309 with the command, "`macs2 callpeak --nomodel --nolambda --keep-dup all --call-summits -B --SPMR --format BED -q 0.05 --shift 75 --extsize 150.`"

Since we had three biological replicates per condition, we used a majority rule approach to retain only summits that were found in at least two replicates (Yang *et al*, 2014). Using these condition-specific peak files, we used bedtools to create a consensus peak file by merging each individual condition's peak summit file together in a manner that disallowed overlapping peaks. We used bedtools merge command "`bedtools merge -d 50`" to combine features within 50 base pairs of each other into a single peak after testing multiple merge distances. We used the number of ATAC-seq fragment counts at each peak in this merged consensus peak file for differential peak analysis.

We used the custom peak analysis algorithm from Sanford *et al* (2020a) that took advantage of additional ethanol control conditions to estimate the false discovery rate in ethanol controls to then identify differential peaks. Briefly, reads were quantified for each peak in the master consensus file and fragments at each peak were normalized to correct for differences in total sequencing depth using the equation:

$$\frac{\text{samples total reads in peaks}}{\text{mean number of reads in peaks across all samples}}.$$

Next, an estimated false discovery rate was calculated in each cell of a 50 × 50 grid containing 50 exponentially-spaced steps of minimum fold change values (ranging from 1.5–10) and 50 exponentially spaced steps of the minimum number of normalized fragment counts in the condition with the greater number of counts (ranging from 30 to 237 or 10 to 237). The estimated false discovery rate (FDR) was calculated using the equation:

$$\text{estimated FDR} = \frac{(\text{no. of conditions})(\text{est. number of false positive peaks per condition})}{\text{total number of differential peaks in experimental conditions}}.$$

After calculating the estimated FDR in each cell of the 50 × 50 grid, we then pooled together differential peaks contained in any cell with an FDR of less than 0.25%.

We performed motif analysis on our set of differential peaks using chromVAR v1.8.0 (Schep *et al*, 2017), its associated cisBP database of transcription factor motifs, and the motifmatchR package from bioconductor. To decrease the variance of the transcription factor motif deviation scores, we pooled together the different dosages of retinoic acid or TGF-β. The chromVAR code was modified to extract an internal metric that equals the fractional change in fragment counts at motif-containing peaks for a given motif. For the calculation of motif enrichment scores, the motifs we used were derived from four distinct groups of motif recognition sequences: RARA, HOX, SMAD, and AP-1.

Footprint was performed by subsetting on peaks that were not differentially accessible but met the minimum normalized fragment count threshold. These peaks were used as the basis for the HINT (Li *et al*, 2019) subroutine of the Regulatory Genomics Toolbox suite (http://www.regulatory-genomics.org/).

Hematopoietic differentiation data

We used pre-existing RNA- and DNase I-seq data (aligned to genome assembly hg19) of hematopoietic differentiation (Data ref: González *et al*, 2015b) to compare against our data. We used data from the website provided in the paper (http://cbio.mskcc.org/public/Leslie/Early_enhancer_establishment/) to download annotations of peaks (peaksTable.txt), counts of DNase-seq (DNaseCnts.txt), and RNA-seq counts (RNAseqCnts.txt). Counts presented in these data files were quantile normalized and averaged when biological replicates were available. We filtered peaks with "CD14" or "CD34" under the "accessPattern" annotation to choose from peaks that were relevant for comparing HSPCs to monocytes. We used a $\log_2$ fold change of greater than or equal to 2 as a cutoff for assigning differential peaks. We used the pre-existing annotations of genes for each peak for peak-gene mapping. For determining the $\log_2$ fold change in gene expression we discarded genes whose maximum expression value across the two conditions was fewer than 5 quantile-normalized units.

For visualization of this dataset with our own accessibility data, we realigned raw fastq files DNase-seq files to the hg38 assembly using bowtie v2.3.4.1 and filtered out low-quality alignments with samtools v1.1 to generate new bam alignment files. The alignment files were combined using samtools merge in a single .bam file per cell type. Bam files were converted to .bigWig format using deeptools 3.5.1 (Ramírez *et al*, 2016) "`bamCoverage --normalizeUsing CPM`" to create a "consensus" .bigWig for visualization. Peaks for CD34+ and CD14+ samples were made by filtering peaks annotated for these populations in the "accessPattern" column and creating separate .bed files using a custom script. The peak location in these .bed files was then lifted over from hg19 to hg38 using UCSC hgLiftOver. For comparing the overlap of peaks between datasets, we created consensus peak sets across all sample types and used the bedtools intersect function to quantify the proportion of peaks that intersected between the hematopoietic differentiation and signaling data.

Peak annotation

Peaks were annotated using ChIPseeker (Yu *et al*, 2015) to determine the relative proportion of features in the data from González *et al* (2015a) (DNAse-seq) and Sanford *et al*, 2020a (ATAC-seq). For ease of visualization, certain categories like the three promoter categories were collapsed into one. ChiPseeker was also used to identify the nearest transcriptional start site to a gene used for the nearest integration approach described below. For making scatterplots of change in accessibility versus change in expression annotated by peak feature, a custom script was used to combine annotations from ChIPseeker into four categories: downstream, gene body, intergenic, and promoter.

For each of the top 150 most variable transcription factor motifs we identified using differential accessibility analysis, we used the R bioconductor motifmatchR package to annotate both the number of motif matches and a log-likelihood match score for each peak.

IL-1α stimulation and myelocyte differentiation data

Pre-existing RNA- and ATAC-seq data from KB epithelial cells before and after 1 h of stimulation with IL-1α (Data ref: Jurida *et al*, 2015b; Data ref: Weiterer *et al*, 2020b) and differentiation data of cells along the myelocyte lineage (Data ref: Ramirez *et al*, 2017b) were used for further comparison. Both types of data were aligned and normalized as mentioned above and peaks were annotated accordingly.

For the IL-1α data, there was only one biological replicate and the data were not collected concurrently. For the myelocyte data, we compared three replicates of HL-60 promyelocytes and three replicates of monocyte-derived macrophages, which were the result of a directed differentiation over the course of 168 h that involved exposure to vitamin D3 and phorbol-12-myristate-13-acetate (PMA). For determining the $\log_2$ fold change in gene expression we discarded genes whose maximum expression value across the conditions was fewer than 5 TPM. We used a $\log_2$ fold change of greater than or equal to 0.58 as a cutoff for assigning differential peaks.

ATAC-seq footprinting

To examine whether possible "subpeaks" within nondifferentially accessible peaks could explain concordance, be used "`samtools merge`" to combine aligned and filtered .bam files across conditions to make a single .bam file per condition. We then used a custom script to make .bed files of the peaks that were not differentially accessible and used HINT-ATAC (Li *et al*, 2019) from the regulatory genomics toolbox using "`rgt-hint footprinting`," "`rgt-motifanalysis matching`," and "`rgt-hint differential`" commands to measure the difference in reading accessibility for motifs across ethanol control and either high dose retinoic acid or high dose TGF-β conditions.

RNA and ATAC data integration

We employed a variety of methods for assigning peaks to genes. In the "nearest" approach, we used annotation from ChIPseeker to assign each peak to the nearest transcriptional start site. With this method, each peak is uniquely mapped to a single gene. In the "window" approach, we used a window of 50 kilobases on either side of the transcriptional start site (100 kilobases in total) to assign peaks to a gene, which could result in a peak being assigned to

multiple genes. To look at promoter-proximal peaks, we used a window of 1.5 kilobases on either side of the transcriptional start site (3 kilobases in total) to assign peaks to a gene. To examine promoter distal peaks, we took all peaks within the 100 kilobase total window and subsequently all peaks within 20 kilobases up or downstream of the transcriptional start site. For creating windows based on topologically associating domains (TADs), we used available MCF-7 TAD data (Barutcu *et al*, 2015) and lifted coordinates from hg19 to hg38 using UCSC hgLiftOver. We then used bedtools intersect to assign out peaks within TADs near transcriptional start sites.

For the initial overlap analysis for Fig 1, peaks were assigned to genes using the "nearest method." Overlap was considered if any of the differentially expressed genes had a differentially accessible peak, regardless of whether the expression and accessibility changes were in the same direction. Further analyses integrating gene expression and chromatin accessibility data took the magnitude and direction of change into account.

Intron/exon analysis

In order to look at the possibility of more nascent transcripts being more concordant with chromatin accessibility changes, we performed a secondary alignment of RNA-seq reads to an intron-formatted .gtf file and quantified counts using HT-seq with the option "`--type intron`." Raw intron and exon counts were quantile normalized and used to create scatterplots to compare the change in counts across conditions versus change in accessibility.

Track visualization

We visualized accessibility data using the web-based version of integrative genomics viewer (IGV) (Robinson *et al*, 2011, 2020). We prepared accessibility data for visualization by taking consensus files and converting them to .bigWig file format with either fragments per million or counts per million normalization. Bed files for identifying peaks were created using custom scripts.

Statistics and software

Unless otherwise stated, we performed analyses using R v4.1.0 with data manipulation and visualization done with tidyverse v1.3.1 (Wickham *et al*, 2019) and ggpubr v0.4.0. We used a Kolmogorov–Smirnov test to compare means. Unless otherwise stated, 95% confidence intervals for Pearson's r were calculated by bootstrapping using 10,000 replicates.

Data availability

The datasets and computer code produced in this study are available in the following databases:

- Raw and processed data: BioStudies (S-BSST886; https://www.ebi.ac.uk/biostudies/studies/S-BSST886).
- Intermediate data types, processing, and plotting scripts: GitHub (https://github.com/kdhkiani/concordancePaper).
- Links to pre-existing data from other studies can be found in Appendix Table S1.

Expanded View for this article is available online.

Acknowledgements
We are greatly indebted to Professor Christina Leslie and Alvaro González for many insightful discussions and for assistance in working with their datasets. We also thank the members of the Raj lab for valuable feedback, especially Ally Coté, Lee Richman, and Ryan Boe. KK acknowledges support from NIH T32 GM008216; EMS acknowledges support from F30 HG010986; YG holds a Career Award at the Scientific Interface from BWF; and AR acknowledges support from NIH Director's Transformative Research Award R01 GM137425, NIH R01 CA238237, NIH R01 CA232256, NIH P30 CA016520, NIH SPORE P50 CA174523, and NIH U01 CA227550.

Author contributions
Karun Kiani: Conceptualization; data curation; software; formal analysis; visualization; methodology; writing – original draft; writing – review and editing. **Eric M Sanford:** Conceptualization; data curation. **Yogesh Goyal:** Conceptualization; methodology. **Arjun Raj:** Conceptualization; resources; supervision; funding acquisition; investigation; methodology; writing – original draft; project administration; writing – review and editing.

Disclosure and competing interests statement
AR receives royalties related to Stellaris RNA FISH probes. All other authors declare that they have no conflict of interest.

References

Ampuja M, Rantapero T, Rodriguez-Martinez A, Palmroth M, Alarmo EL, Nykter M, Kallioniemi A (2017) Integrated RNA-seq and DNase-seq analyses identify phenotype-specific BMP4 signaling in breast cancer. *BMC Genomics* 18: 68

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823–837

Barutcu AR, Lajoie BR, McCord RP, Tye CE, Hong D, Messier TL, Browne G, van Wijnen AJ, Lian JB, Stein JL *et al* (2015) Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol* 16: 214

Biddie SC, John S, Sabo PJ, Thurman RE, Johnson TA, Schiltz RL, Miranda TB, Sung M-H, Trump S, Lightman SL *et al* (2011) Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell* 43: 145–155

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132: 311–322

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10: 1213–1218

Glover JC, Renaud J-S, Rijli FM (2006) Retinoic acid and hindbrain patterning. *J Neurobiol* 66: 705–725

González AJ, Setty M, Leslie CS (2015a) Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat Genet* 47: 1249–1259

González AJ, Setty M & Leslie CS (2015b) Processed data from ENCODE (http://cbio.mskcc.org/public/Leslie/Early_enhancer_establishment/). [DATASET]

Goyal Y, Dardani IP, Busch GT, Emert B, Fingerman D, Kaur A, Jain N, Mellis IA, Li J, Kiani K *et al* (2021) Pre-determined diversity in resistant fates emerges from homogenous cells after anti-cancer drug treatment. *bioRxiv* https://doi.org/10.1101/2021.12.08.471833 [PREPRINT]

Hua S, Kittler R, White KP (2009) Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell* 137: 1259–1271

John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* 43: 264–268

Jurida L, Soelch J, Bartkuhn M, Handschick K, Müller H, Newel D, Weber A, Dittrich-Breiholz O, Schneider H, Bhuju S *et al* (2015a) The activation of IL-1-induced enhancers depends on TAK1 kinase activity and NF-κB p65. *Cell Rep* 10: 726–739

Jurida L, Soelch J, Bartkuhn M, Handschick K, Müller H, Newel D, Weber A, Dittrich-Breiholz O, Schneider H, Bhuju S *et al* (2015b) Gene Expression IOmnibus GSE64224 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64224). [DATASET]

Kogai T, Schultz JJ, Johnson LS, Huang M, Brent GA (2000) Retinoic acid induces sodium/iodide symporter gene expression and radioiodide uptake in the MCF-7 breast cancer cell line. *Proc Natl Acad Sci U S A* 97: 8519–8524

Lee C-K, Shibata Y, Rao B, Strahl BD, Lieb JD (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet* 36: 900–905

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J *et al* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148: 84–98

Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG (2019) Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* 20: 45

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27: 1739–1740

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 1: 417–425

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550

Ma S, Zhang Y (2020) Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq. *Mol Biomed* 1: 9

Mahdi SHA, Cheng H, Li J, Feng R (2015) The effect of TGF-beta-induced epithelial-mesenchymal transition on the expression of intracellular calcium-handling proteins in T47D and MCF-7 human breast cancer cells. *Arch Biochem Biophys* 583: 18–26

Nair VD, Vasoya M, Nair V, Smith GR, Pincas H, Ge Y, Douglas CM, Esser KA, Sealfon SC (2021) Differential analysis of chromatin accessibility and gene expression profiles identifies cis-regulatory elements in rat adipose and muscle. *Genomics* 113: 3827–3841

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* 21: 447–455

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 44: W160–W165

Ramirez RN, El-Ali NC, Mager MA, Wyman D, Conesa A, Mortazavi A (2017a) Dynamic gene regulatory networks of human myeloid differentiation. *Cell Syst* 4: 416–429.e3

Ramirez RN, El-Ali NC, Mager MA, Wyman D, Conesa A & Mortazavi A (2017b) Gene Expression Omnibus GSE79046 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79046). [DATASET]

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A *et al* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4: 651–657

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. *Nat Biotechnol* 29: 24–26

Robinson JT, Thorvaldsdóttir H, Turner D, Mesirov JP (2020) igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *bioRxiv* https://doi.org/10.1101/2020.05.03.075499

Samstein RM, Arvey A, Josefowicz SZ, Peng X, Reynolds A, Sandstrom R, Neph S, Sabo P, Kim JM, Liao W *et al* (2012) Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* 151: 153–166

Sanford EM, Emert BL, Coté A, Raj A (2020a) Gene regulation gravitates toward either addition or multiplication when combining the effects of two signals. *Elife* 9: e59388

Sanford EM, Emert BL, Coté A & Raj A (2020b) Gene Expression Omnibus GSE152749 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152749). [DATASET]

Schep AN, Wu B, Buenrostro JD, Greenleaf WJ (2017) chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* 14: 975–978

Schmutzler C, Winzer R, Meissner-Weigl J, Köhrle J (1997) Retinoic acid increases sodium/iodide symporter mRNA levels in human thyroid cancer cell lines and suppresses expression of functional symporter in nontransformed FRTL-5 rat thyroid cells. *Biochem Biophys Res Commun* 240: 832–838

Starks RR, Biswas A, Jain A, Tuteja G (2019) Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics Chromatin* 12: 16

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545–15550

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B *et al* (2012) The accessible chromatin landscape of the human genome. *Nature* 489: 75–82

de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, Geschwind DH (2018) The dynamic landscape of open chromatin during human cortical neurogenesis. *Cell* 172: 289–304.e18

Weintraub H, Groudine M (1976) Chromosomal subunits in active genes have an altered conformation. *Science* 193: 848–856

Weiterer S-S, Meier-Soelch J, Georgomanolis T, Mizi A, Beyerlein A, Weiser H, Brant L, Mayr-Buro C, Jurida L, Beuerlein K *et al* (2020a) Distinct IL-1α-responsive enhancers promote acute and coordinated changes in chromatin topology in a hierarchical manner. *EMBO J* 39: e101533

Weiterer S-S, Meier-Soelch J, Georgomanolis T, Mizi A, Beyerlein A, Weiser H, Brant L, Mayr-Buro C, Jurida L, Beuerlein K *et al* (2020b) Gene Expression Omnibus GSE134436 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134436). [DATASET]

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J *et al* (2019) Welcome to the tidyverse. *J Open Source Softw* 4: 1686

Wu C, Wong YC, Elgin SC (1979) The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* 16: 807–814

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L *et al* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovations* 2: 100141

Yang Y, Fear J, Hu J, Haecker I, Zhou L, Renne R, Bloom D, McIntyre LM (2014) Leveraging biological replicates to improve analysis in ChIP-seq experiments. *Comput Struct Biotechnol J* 9: e201401002

Yu G, Wang L-G, He Q-Y (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* 31: 2382–2383

Zaret KS (2020) Pioneer transcription factors initiating gene network changes. *Annu Rev Genet* 54: 367–385

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9: R137