Integrating RNA-seq and assay for transposase-accessible chromatin by sequencing (ATAC-seq) predicts functionally-relevant chromatin regions

Collin B. Merrill[1], Austin B. Montgomery[2], Miguel A. Pabon[2], Aylin R. Rodan[2,3,4] and Adrian Rothenfluh[1,2,4,5,*]

[1]Department of Psychiatry University of Utah, Salt Lake City, UT 84108, USA.

[2]Molecular Medicine Program, University of Utah, Salt Lake City, UT 84112, USA.

[3]Division of Nephrology, Department of Internal Medicine, University of Utah, Salt Lake City, UT 84112, USA.

[4]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA.

[5]Department of Neurobiology, University of Utah, Salt Lake City, UT 84112, USA.

*Corresponding author:

Adrian Rothenfluh, adrian.rothenfluh@hsc.utah.edu

ABSTRACT

Gene regulation is critical for proper cellular function. Next-generation sequencing technology has revealed the presence of regulatory networks that regulate gene expression and essential cellular functions. Studies investigating the epigenome have begun to uncover the complex mechanisms regulating transcription. Assay for transposase-accessible chromatin by sequencing (ATAC-seq) is quickly becoming the assay of choice for epigenomic investigations. Integrating epigenomic and transcriptomic data has the potential to reveal the chromatin-mediated mechanisms regulating transcription. However, integrating these two data types remains challenging. We used the insulin signaling pathway as a model to investigate chromatin regions and gene expression changes using ATAC- and RNA-seq in insulin-treated *Drosophila* S2 cells. We show that insulin causes widespread changes in chromatin accessibility and gene expression. Then, we attempted to integrate ATAC- and RNA-seq data to predict functionally-relevant chromatin regions that control the transcriptional response to insulin. We show that using differential chromatin accessibility can predict functionally-relevant genome regions, but that stratifying differentially-accessible chromatin regions by annotated feature type provides a better prediction of whether a chromatin region regulates gene expression. In particular, our data demonstrate a strong correlation between chromatin regions annotated to distal promoters (1-2 kb from the transcription start site). To test this prediction, we cloned candidate distal promoter regions upstream of luciferase and validated the functional relevance of these chromatin regions. Our data show that distal promoter regions selected by correlations with RNA-seq are more likely to control gene expression. Thus, correlating ATAC- and RNA-seq data can home in on functionally-relevant chromatin regions.

INTRODUCTION

Gene regulation is essential to the development and maintenance of life. Gene regulatory networks describe the interplay between regulatory regions, such as promoters and enhancers, and expression of their target genes (MacNeil and Walhout, 2011). Deciphering how specific regulatory regions control gene transcription can provide insights into biological processes such as cell type differentiation (Tokusumi et al., 2012; Goode et al., 2016), responses to addictive substances (Morozova et al., 2011), and myriad other cell functions.

The vinegar fly, *Drosophila melanogaster*, has long been a favorite model organism of geneticists. *Drosophila* model systems provide many advantages over mammalian systems, particularly in the quest to investigate the function of gene regulatory networks (Wilczynski and Furlong, 2010). In addition to a quick life span and easy husbandry, the *Drosophila* model system has a wealth of genetic tools which allows for a targeted and precise approach to control gene expression and identify genomic regulatory regions (Jenett et al., 2012; Dionne et al., 2018).

The advent of new sequencing techniques has led to a greater understanding of how transcription is controlled. RNA-seq has provided a broader and more detailed picture of complex transcriptional states and responses (Duarte et al., 2016; Petruccelli et al., 2020). Epigenomic assays such as chromatin immunoprecipitation (ChIP-seq), DNAse-seq, and assay for transposase-accessible chromatin by sequencing (ATAC-seq) can interrogate chromatin status and identify transcription factor binding sites (Jothi et al., 2008; Song and Crawford, 2010; Buenrostro et al., 2013). ChIP-seq can pinpoint specific histone markers and transcription factor binding sites, but has proven too specific to find many previously unknown regulatory regions because of highly-specific antibody-mediated pulldown (Quillien et al., 2017). DNAse-seq and ATAC-seq use cleavage enzymes to identify untargeted, genome-wide open chromatin regions (also known as ‘peaks’). However, ATAC-seq is quickly becoming more popular because it requires fewer cells, does not require harsh chemicals, and does not take as much time (Buenrostro et al., 2013). As an increasing number of these and other next generation

sequencing assays are used to probe gene regulation, a data bottleneck is forming. While there have been incredible bioinformatics advances in both analyzing these data and using assay results to make biological predictions, there remains a gap between the computational analyses and biological validation of these sequencing techniques. Without a plan to systematically and efficiently biologically validate data-driven mechanisms, the terabytes of data collected may not yield relevant information regarding gene regulatory networks.

Here, we sought to harness ATAC-seq and RNA-seq to identify genomic regions that drive gene expression in the *Drosophila* genome. We used the insulin signaling pathway as a model because the insulin receptor activates multiple downstream signaling pathways (Kido et al., 2001; Puig et al., 2003; McNeill et al., 2008), resulting in widespread changes to the chromatin state (Kulkarni et al., 2016) and gene expression (Post et al., 2018). *Drosophila* S2 cells were treated with insulin and then assayed by ATAC- and RNA-seq. By integrating these two sequencing datasets, we aimed to more effectively and efficiently determine the regulatory elements which control gene expression. Cloning identified putative regulatory regions into S2 cells proved to be a technically straightforward way to identify regulatory elements which are important in cellular insulin responses. Moreover, we established this as a valuable process to validate bioinformatic data and investigate the functional relevance gene regulatory networks.

MATERIALS AND METHODS

Cell culture

Drosophila S2 cells (Drosophila Genomics Resource Center, Bloomington, IN, USA) were cultured in Schneider's Drosophila Medium (ThermoFisher, Waltham, MA, USA) supplemented with 10% fetal bovine serum (ThermoFisher) at 25 °C. Cells were cultured in Schneider's medium without FBS for 24 h before experiments. Then, cells were incubated with 10 μM insulin (Sigma Aldrich, St. Louis, MO, USA) or vehicle (25 mM HEPES, pH 8.2) for 4 h at 25 °C.

ATAC-seq

S2 cells were incubated with 3 μM DAPI for 10 min. 60,000 cells per sample were sorted using a BD FACS Aria flow cytometer (BD Biosciences, San Jose, CA, USA). The cells were collected into ice-cold PBS (pH 7.4). After sorting, the samples were washed once with ice-cold PBS and centrifuged at 500 g for 5 min at 4 °C. ATAC-seq libraries were prepared as previously described (Buenrostro, 2015). Briefly, 50 μL lysis buffer (10 mM Tris-HCl 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% NP40) was added to each sample, and the sample was centrifuged at 500 g for 10 min at 4 °C. The supernatant was removed, and the nuclei pellet was tagmented using a Nextera DNA Library Prep kit (Illumina, Inc., San Diego, CA, USA) as previously described. Then, the tagmented DNA was purified using a Qiagen MinElute PCR Purification Kit (Qiagen, Germantown, MD, USA). The purified DNA was PCR amplified for 5 cycles using a Nextera DNA Library Index kit (Illumina) and Phusion HF Master Mix (New England BioLabs, Inc., Ipswich, MA, USA) with the following protocol: 72 °C for 5 min, 98 °C for 30 sec, and 5 cycles of 98 °C for 10 sec, 63 °C for 30 sec, and 72 °C for 1 min. A 5-μL aliquot of the pre-amplified reaction was analyzed by qPCR using SsoFast EvaGreen Supermix (Bio-Rad Life Science, Inc., Hercules, CA, USA) and an Applied Biosystems 7900HT qPCR instrument (ThermoFisher) using the following protocol: 1 cycle of 98 °C for 30 sec and 40 cycles of 98 °C for 10 sec, 63 °C for 30 sec, and 72 °C for 1 min. Then, the pre-amplified PCR mixture was amplified for another 10 cycles (corresponding to 1/3 maximum fluorescence) using the same thermocycling parameters. After amplification, the libraries were purified using AMPure XP beads (Beckman Coulter Life Sciences, Indianapolis, IN, USA). Libraries were sequenced on an Illumina HiSeq 2500 instrument using 50-bp single-end reads.

RNA-seq

Total RNA was isolated from S2 cells using a PureLink RNA purification kit (ThermoFisher). Then, rRNA was removed from each sample with a Ribo-Zero rRNA Removal kit (Illumina). RNA

libraries were constructed using a NEBNext Ultra II RNA Library Kit for Illumina and NEBNext Multiplex Oligos for Illumina, Primer Set 1 (New England Biolabs). Libraries were sequenced on an Illumina HiSeq 2500 instrument using 50-bp single-end reads.

ATAC-seq data analysis

ATAC-seq Fastq files were aligned to the dm6 genome assembly (http://ftp.flybase.net/releases/FB2018_06/dmel_r6.25/fasta/) using Novocraft Novoalign with the following settings: --NonC -o SAM -r Random. SAM files were processed to BAM format, sorted, and indexed using Samtools (Li et al., 2009). BAM files were reads per million-normalized and converted to bigWig files using the Bio-ToolBox 'bam2wig.pl' program (https://github.com/tjparnell/biotoolbox/blob/master/scripts/bam2wig.pl). Peak calling was performed on the bigWig files by utilizing the Multi-Replicate Macs ChIPseq pipeline (https://github.com/HuntsmanCancerInstitute/MultiRepMacsChIPseq) with the following settings: --dupfrac 0.2 --size 200 --cutoff 2 --peaksize 300 --peakgap 200. Called peaks were annotated in R with the ChIPseeker package (Yu et al., 2015). Count data for called peaks was collected from processed BAM files using the Bio-ToolBox 'get_datasets.pl' program (https://metacpan.org/pod/distribution/Bio-ToolBox/scripts/get_datasets.pl). The count data was then used to identify differentially accessible regions with the R package DEseq2 (Anders and Huber, 2010).

RNA-seq data analysis

RNA-seq fastq files were aligned to the BDGP6 genome assembly using the STAR aligner (Dobin et al., 2013) with the following settings: --twopassMode Basic --outSAMtype BAM SortedByCoordinate --outWigType bedGraph --outWigStrand Unstranded -- clip3pAdapterseq AGATCGGAAGAGCACACGTCTGAACTCCAGTCA. The resulting sorted BAM files were indexed using Samtools (Li et al., 2009). FeatureCounts was used to collect count data for BDGP6 genes using the following command: -T 16 -s 2 (Liao et al., 2014). Count data for all

replicates and experimental conditions were combined into a single count matrix in R. The count matrix was subsequently used to identify differentially expressed genes with the R package DEseq2 (Anders and Huber, 2010).

Integration analysis of ATAC-Seq and RNA-Seq datasets

The ATAC-seq peak data were compared to the RNA-seq data to determine how chromatin accessibility influenced gene expression. The raw RNA-seq and ATAC-seq counts for each sample were compared using the gene annotation of the ATAC-seq peak and the assigned RNA-seq gene. The raw count value was averaged by experimental condition and genomic assay type. Then, the RNA-seq and ATAC-seq datasets were compared using the annotated genes and the $\log_2$ fold change values for each peak/gene in the respective genomic assay. ATAC-seq peaks with an FDR < 0.1 and genes detected by RNA-seq with an FDR < 0.05 were used to compare the differentially accessible peaks and differentially expressed genes. Pearson correlation analysis was performed between the $\log_2$ fold change values of the genomic assays and between the average raw count values of the genomic assays (controlling for the experimental condition).

Plasmid construction and transformation

A multiple cloning site (MCS) was cloned into the backbone pDEST VanGlow-GL vector (Janssens et al., 2017). Then, we removed the *miniwhite* cassette using the restriction enzymes AflII (3L137, 3R131, X216, 3L114, 2L796, 3L981, 2L220, 2R177, 2L225, and FOXO TFBS) or SmaI and PmlI (X950, 2L846, and 2R111) (New England Biolabs). The digested plasmids were incubated with T4 ligase (New England BioLabs) for 15 min at room temperature and purified by 1% gel electrophoresis. The plasmids were linearized using AvrII and PacI (sites contained in MCS) (New England BioLabs).

Genomic DNA was purified from S2 cells using a Monarch Genomic DNA Purification kit (New England BioLabs). Candidate chromatin peak sequences and 100-200 bp flanking sequences (Cusanovich et al., 2018) (Supplementary File 1) were amplified using Phusion High-Fidelity PCR MasterMix (New England BioLabs) (primer sequences are listed in Supplementary File 2) and a C1000 thermocycler (Bio-Rad Life Science). The peak sequences were amplified as follows: 98 °C for 5 min, followed by 35 cycles of 98 °C for 30 sec, 52-68 °C gradient for 30 sec, 72 °C for 4 min, and a final incubation for 5 min at 72 °C. The amplified fragments were purified on 1% agarose gels and extracted using a Monarch Gel Purification kit (New England BioLabs) and cloned into linearized VanGlo-GL-MCS plasmid using NEBuilder HiFi Assembly master mix (New England BioLabs). The assembled plasmids were transformed into DH5α cells and grown overnight. Clones were screened by restriction digestion using EcoRI-HF (New England BioLabs). Sequences were confirmed by Sanger sequencing at GeneWiz (South Plainfield, NJ, USA). Confirmed plasmids were transformed into S2 cells using TransIT Insect Transfection Reagent (Mirus Bio, Madison, WI, USA). Transformed cells were grown for 48 h before use in experiments.

Luciferase assays

Transformed S2 cells were serum starved for 24 h and treated with insulin or vehicle as described above (*Cell culture*). Then, luciferase activity was assayed using a Luciferase Reporter Substrate Assay Kit-Firefly (Abcam, Cambridge, MA, USA). Luminescence was detected with a BioTek Synergy HTX microplate reader (BioTek Instruments, Winooski, VT, USA) and analyzed using Gen5 2.01.17 software (BioTek Instruments).

Statistical analysis

Relative luminescence data was analyzed using GraphPad Prism 9.0 software. Statistical differences were tested using Student's t-tests with at least three biological replicates. Heatmaps

were generated using the R package ComplexHeatmap (Gu et al., 2016). PCA plots were created using the R package pcaExplorer (Marini and Binder, 2019). Correlation plots were produced with the R package ggpubr (Kassambara, et al., 2020).

RESULTS

ATAC-seq

9726 high-confidence peaks were identified in the insulin-exposed S2 cells, while 9560 high-confidence peaks were identified in the control S2 cells. After merging the control and experimental peak sets, 10269 peaks remained. PCA showed that the largest variance arose from insulin treatment, as expected (Fig. 1A). We plotted the Log2 fold change for each peak and generated a heatmap from this data to visualize the overall differences in accessibility after insulin treatment. In general, peaks that were more closed in the vehicle-treated samples were more open in the insulin-treated samples (Fig. 1B). Because ATAC-seq provides an unbiased view of chromatin accessibility, we evaluated the feature distribution of the entire peakset (Fig. 1C). Many regulatory sequences in *Drosophila* are located within promoter, intronic, or distal intergenic regions (Roy et al., 2010), so we were particularly interested in the peaks annotated to these feature types. As expected, most of the peaks were annotated as promoters [59.29%, ≤ 1 kb from transcription start site (TSS), 9.48% 1-2 kb from TSS, 68.77% total]. 15.09% of the peaks were annotated to an intronic region (5.87% in first intron, 9.22% in other introns) and 9.38% of peaks were labeled as distal intergenic sequences. Then, we examined whether the peak regions with increased or decreased accessibility after insulin treatment had a similar feature type distribution. In the upregulated peak set [$\log_2$ fold change > 0, false discovery rate (FDR) < 0.05], 63.48% of peaks were annotated to promoter regions, 15.04% were annotated as introns, and 12.01% were identified as distal intergenic regions. In the downregulated peak set ($\log_2$ fold change < 0, FDR < 0.05), 74.21% of peaks were annotated to the promoter region, while 15.05% were annotated as introns, and 6.68% of peaks were labeled as distal intergenic.

These results suggest that insulin treatment may affect chromatin associated with gene regulation in S2 cells.

RNA-seq

In addition to analyzing how insulin affects chromatin accessibility in S2 cells, we also examined whether insulin treatment affects gene expression. We treated S2 cells with insulin and examined gene expression changes by RNA-seq. In this dataset, we detected 10,287 genes. PCA indicated that the largest variance between the samples was generated by insulin treatment (Fig. 2A). Heatmap analysis showed that, in general, genes that were upregulated in the insulin-treated samples were downregulated in the control samples, and vice versa (Fig. 2B).

Differential Chromatin Accessibility and Gene Expression

In the ATAC-seq peakset, 773 peaks were significantly differentially accessible (FDR < 0.1) between the insulin-exposed and the control samples. 364 of the peaks were more accessible upon insulin exposure, while 409 peaks were less accessible after exposure to insulin. We examined the feature distribution of significant differentially accessible peaks and observed that 64.78% of peaks were annotated to promoter regions (58.64%, ≤ 1 kb from TSS; 6.14% 1-2 kb from TSS). 12.96% were annotated to introns (6.14%, first intron; 6.82%, other introns) and 11.82% were annotated to distal intergenic regions. When we examined the Log_2 fold changes of the differentially accessible peaks (stratified by feature type), we observed that each feature type contained peaks that were either more- or less-accessible for each feature. Further, peaks that were more accessible in one treatment group (ex. introns after insulin) were less accessible in the other treatment group (introns in control samples). We also examined the significant differentially-expressed genes from the RNA-seq dataset. In this peakset, 3612 genes were differentially expressed (FDR < 0.05) between the insulin-exposed and control samples. 2056 genes were upregulated after insulin exposure, while the remaining 1560 were downregulated.

Further, genes that were significantly upregulated after insulin treatment were downregulated in the control group (Fig. 3C).

Integration of ATAC-seq and RNA-seq results

To determine if the changes in open chromatin regions correlated with gene expression changes, we compared the ATAC-seq and RNA-seq data. In the RNA-seq dataset, 8421 peaks were annotated to a gene. Interestingly, only 287 genes were associated with an annotated differentially accessible peak in the ATAC-seq dataset. However, several ATAC-seq peaks were annotated to different features or regions associated with the same gene in many cases. Because of this overlap, we reasoned that there may be a correlation between the ATAC-seq and RNA-seq peaks annotated to the same genes. Thus, we used Pearson correlation analysis to determine whether the raw counts from each ATAC-seq or RNA-seq peak were correlated. We observed a significant correlation between the log mean counts from each peakset ($p < 2.2e-16$), though the correlation was weak ($R = 0.1$) (Fig. 4A). When we stratified the count data by feature type (Fig. 4B), we only observed a significant correlation between ATAC-seq peaks annotated to promoters ≤ 1 kb from the TSS and the same genes from the RNA-seq data ($p < 2.2e-16$). However, as with the overall counts, the correlation was weak ($R = 0.2$). Importantly, the significant correlations likely arise from the presence of numerous data points. These results show that the correlation between RNA-seq and ATAC-seq data are driven by accessibility in the promoter.

We next investigated the association between the ATAC-seq and the RNA-seq Log2 fold changes after insulin treatment. For all peaks in this merged dataset, Pearson correlation analysis showed a significant, but weak, correlation ($R = 0.089$, $p = 3.6e-16$) (Fig. 5A). Performing a similar analysis after stratifying by feature type (Fig. 5B) showed similarly weak correlations (R from -0.096 to 0.18) which were significant for introns ($p = 0.0034$), promoters ≤ 1 kb from the TSS ($p = 2.2e-16$), and promoters 1-2 kb from the TSS ($p = 4.8e-07$).

We next asked if there were significant correlations between significantly different ATAC- and RNA-seq peaks. When we examined the relationships between only the significantly different peaks, we observed that *R* increased by over 300% (*R* = 0.32, *p* = 3e-8) for all ATAC- and RNA-seq $\log_2$ fold changes. This stronger correlation appeared to be driven by ATAC-seq peaks annotated as proximal (≤1kb) and distal (>1kb) promoters, with *R* = 0.33 (*p* = 6e-6) and 0.66 (*p*=0.00013), respectively.

Functional testing of significant differentially-accessible ATAC-seq peaks

ATAC-seq provides an unbiased interrogation of the chromatin state. We hypothesized that ATAC-seq peaks could be used to develop tools to investigate chromatin-mediated gene regulation. First, we selected four ATAC-seq peaks with the largest log2 fold change and cloned these sequences into a VanGlow-GL-MCS vector (Supplementary Table 2). Then, S2 cells were transfected with these vectors for 48 h, starved for 18 h, and treated with 10 μM insulin or vehicle for 4 h. Of the four tested plasmids, only one showed significantly increased luciferase activity after insulin treatment: 3L114 (*p* = 0.016) (Fig. 7A). Introns often contain regulatory regions (Roy et al., 2010), so we next chose the four intron regions with the largest $\log_2$ fold changes and performed similar luciferase assays. Again, only 1 peak region gave significantly increased luciferase activity: X216 ($p < 0.05$) (Fig. 7B). Because distal promoters were the most strongly correlated with gene expression in our previous analyses, we examined whether this correlation analysis could predict whether ATAC-seq peak sequences would drive gene expression. We chose four peaks with the highest $\log_2$ fold change from distal promoter regions that were significantly more open after insulin and cloned those sequences into the VanGlow-GL-MCS plasmid. Of the tested peaks, 2 produced significantly increased luciferase activity after insulin treatment: 2L225 (*p* = 0.0033) and 2R111 ($p < 0.03$) (Fig. 7C). Interestingly, when we examined whether the genes associated with the cloned peaks were up- or downregulated, we observed that while each region was more accessible in the ATAC-seq dataset, the associated genes from

the RNA-seq dataset were either downregulated or not detected. These results suggest that correlating the $\log_2$ fold changes between ATAC-seq and RNA-seq analyses can somewhat predict whether a region identified by ATAC-seq will drive gene expression. However, because ATAC-seq does not provide information regarding whether a chromatin-associated region, especially if the region is distant from the TSS (such as a distal enhancer or repressor), these predictions must be functionally validated.

DISCUSSION

Next-generation sequencing has enabled comprehensive, in-depth investigations into the role of chromatin structure, chromatin-mediated gene regulation, gene expression, and gene networks. These methods have generated a huge amount of data and have predicted the function of many genes and downstream cellular processes. However, the biological insights predicted by deep-sequencing assays do not always lead to functional investigations. Further, these biological insights must be functionally validated before their full potential can be realized. To address this issue, we examined the relationship between data collected using ATAC-seq and RNA-seq to identify correlations between chromatin and downstream gene regulation. Then, we used the observed relationships to determine whether correlations between chromatin and gene regulation can identify functionally-relevant associations between open chromatin regions and downstream gene expression.

Insulin signaling is an excellent model pathway to examine the relationship between chromatin and gene expression. Activating the insulin receptor causes the activation of several pathways, including the PI3K (Dekanty et al., 2005) and Ras/ERK (Zhang et al., 2011) pathways. These pathways have various effects on the chromatin state (Mouchel-Vielh et al., 2011; Sánchez-Alegría et al., 2018) and gene expression (Garofalo, 2002). Further, insulin receptor signaling is involved in several cellular processes including cell growth, protein synthesis, and gluconeogenesis (Goberdhan and Wilson, 2003). Thus, activating insulin signaling provided an

excellent way to identify broad chromatin and gene expression changes, which allowed us to integrate these physiological changes and determine whether chromatin regions that become more open after insulin signaling could predict chromatin-mediated gene regulation.

ATAC-seq identified widespread chromatin changes after insulin treatment. In general, chromatin regions that became more open in insulin-treated cells were less open in control cells, and vice versa. These results suggest that insulin widely alters chromatin accessibility. When we annotated the peak regions detected by ATAC-seq in control and insulin-treated cells, we observed that most of the regions were annotated as promoters. Since promoters are required for gene expression, it is no surprise that many regions with altered chromatin accessibility are directly involved in gene expression. Interestingly, numerous chromatin regions were annotated as intronic or distal intergenic regions, which often contain regulatory sequences (Roy et al., 2010), indicating that insulin signaling likely activates distant regulatory regions that are involved in gene expression.

Because of the widespread chromatin accessibility changes we observed after insulin treatment, we also examined how insulin altered gene expression in S2 cells. Similar to the ATAC-seq results, we observed that genes with increased expression after insulin treatment were downregulated in vehicle-treated cells, and vice versa. When we examined the differentially-accessible peak regions identified by ATAC-seq and differentially-expressed genes identified by RNA-seq, this observation was even more pronounced. Additionally, when we stratified the differentially-accessible chromatin peaks, we observed that each feature type (promoter ≤ 1kb, intron, distal intergenic region, etc.) contained regions with increased and decreased accessibility after insulin treatment. Similar to the feature distribution in all chromatin peak regions, most of the peaks were annotated as promoters, with numerous peaks annotated as distal promoters and intron regions. These results suggest that insulin causes significant changes to chromatin accessibility and downstream gene expression.

Integrating ATAC- and RNA-seq datasets remains challenging. Several tools have been developed to address this problem, with varying levels of success. We reasoned that using normalized ATAC- and RNA-seq counts may identify correlated chromatin regions and genes. While we observed a significant correlation when we examined all chromatin regions annotated to genes detected by RNA-seq, the correlation was extremely weak. When we stratified the ATAC-seq dataset by feature type, we only observed a significant correlation for promoters within 1 kb of the TSS. Importantly, ATAC-seq provides an unbiased examination of chromatin regions. In the overlapping dataset, we observed that several chromatin peaks were annotated to the same gene, which increased the overall number of the combined ATAC- and RNA-seq counts. Thus, the significant correlations observed for all counts and for promoters within 1 kb of the TSS were likely an artifact of the sample size, even when corrected for multiple comparisons. These data indicate that normalized counts are not an effective way to identify correlated chromatin regions and associated genes.

We performed a similar analysis using the $\log_2$ fold change of chromatin peaks and genes after insulin treatment. Similar to the normalized count results, we observed a significant correlation between log2 fold changes between the ATAC- and RNA-seq datasets, but the correlation was even weaker than for the normalized counts. When we stratified the ATAC-seq dataset by annotated feature type, we observed that only promoter regions were significantly correlated with gene expression. While the correlations we observed between promoters and genes were weak, they were stronger than those observed for the normalized count data. When we examined chromatin peaks with significant differential accessibility and differentially-expressed genes, we observed a much stronger correlation between ATAC- and RNA-seq log2 fold changes. Stratifying the chromatin peaks by feature type revealed moderate correlations between proximal promoters (≤ 1 kb from TSS) and intron regions and their associated genes. We also observed a strong correlation between distal promoters (1-2 kb from TSS) and their associated genes. In *Drosophila*, intronic regions often contain regulatory sequences (Roy et al.,

2010), so increased chromatin accessibility associated with introns is one mechanism to control gene expression (Duret, 2001). Additionally, distal promoter regions may include regulatory regions such as enhancers or repressors that are critically involved in regulating gene expression (Gisselbrecht et al., 2020). Therefore, it is not surprising that we observed a strong correlation between distal promoter regions and expressed genes. These results suggest that chromatin peaks with significantly more open accessibility may predict increased expression of the associated gene.

To validate this hypothesis, we functionally validated chromatin peaks that were significantly more open. We selected the peaks based on log2 fold changes, peaks associated with features known to contain regulatory sequences (introns), and peaks from a highly correlated feature (distal promoters). In peaks chosen by the largest log2 fold changes in all feature types and in all introns, we observed 25% of the cloned peaks induced significantly different luciferase activity, suggesting that selecting chromatin peaks using only large log2 fold changes or a particular feature type is not an effective way to validate the function of those chromatin regions. In contrast, we observed that 50% of peaks chosen from the feature type with the highest correlation between the ATAC- and RNA-seq datasets had increased luciferase activity. These results suggest that correlating significantly more accessible chromatin peaks within a particular feature type to associated gene expression is an effective way to identify which sequences can be successfully validated. Interestingly, each of the chromatin peaks that showed significantly increased luciferase activity were more accessible in the ATAC-seq dataset, but the associated genes were either not detected or were downregulated in the RNA-seq dataset. These results suggest that the chromatin regions we tested may be repressors, though this remains to be confirmed.

Collectively, our results indicate that while correlating annotated chromatin peaks identified by ATAC-seq and differentially-expressed genes detected by RNA-seq can home in functionally-relevant chromatin regions. However, because annotated chromatin peaks may be one of

several regulatory sequences (Roy et al., 2010; Gisselbrecht et al., 2020), these chromatin regions must still be functionally validated.

ACKNOWLEDGMENTS

This work was also supported by the University of Utah Flow Cytometry Facility, the University of Utah Genomics Core Facility, the High Throughput Sequencing Core at the Huntsman Cancer Institute, and the National Cancer Institute through Award Number 5P30CA042014-24. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

FUNDING

This study was supported by grants from the National Institute on Drug Abuse (Grant #5R21DA049635-02, to A.R.) and the National Institute on Alcohol Abuse and Alcoholism (Grant #5R01AA019526-10, to A.R.).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

Anders S, Huber W (2010) Differential expression analysis for sequence count data.

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.

Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, Pliner HA, Christiansen L, Qiu X, Steemers FJ, Trapnell C, Shendure J, Furlong EEM (2018) The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature

555:538–542.

Dekanty A, Lavista-Llanos S, Irisarri M, Oldham S, Wappner P (2005) The insulin-PI3K/TOR pathway induces a HIF-dependent transcriptional response in Drosophila by promoting nuclear localization of HIF-α /Sima. J Cell Sci 118:5431–5441.

Dionne H, Hibbard KL, Cavallaro A, Kao JC, Rubin GM (2018) Genetic reagents for making split-GAL4 lines in Drosophila. Genetics 209:31–35.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.

Duarte FM, Fuda NJ, Mahat DB, Core LJ, Guertin MJ, Lis JT (2016) Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation.

Duret L (2001) Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet 17:172–175.

Garofalo RS (2002) Genetic analysis of insulin signaling in Drosophila. Trends Endocrinol Metab 13:156–162.

Gisselbrecht SS, Palagi A, Kurland J V., Rogers JM, Ozadam H, Zhan Y, Dekker J, Bulyk ML (2020) Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. Mol Cell 77:324-337.e8.

Goberdhan DCI, Wilson C (2003) The functions of insulin signaling: Size isn’t everything, even in Drosophila. Differentiation 71:375–397.

Goode DK, Obier N, Vijayabaskar MS, Lie-A-Ling M, Lilly AJ, Hannah R, Lichtinger M, Batta K, Florkowska M, Patel R, Challinor M, Wallace K, Gilmour J, Assi SA, Cauchy P, Hoogenkamp M, Westhead DR, Lacaud G, Kouskoff V, Göttgens B, Bonifer C (2016) Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Dev Cell 36:572–587.

Gu Z, Eils R, Schlesner M (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32:2847–2849.

Janssens DH, Hamm DC, Anhezini L, Xiao Q, Siller KH, Siegrist SE, Harrison MM, Lee CY (2017) An Hdac1/Rpd3-Poised Circuit Balances Continual Self-Renewal and Rapid Restriction of Developmental Potential during Asymmetric Stem Cell Division. Dev Cell 40:367-380.e7.

Jenett A, Rubin GM, Ngo TTB, Shepherd D, Murphy C, Dionne H, Pfeiffer BD, Cavallaro A, Hall D, Jeter J, Iyer N, Fetter D, Hausenfluck JH, Peng H, Trautman ET, Svirskas RR, Myers EW, Iwinski ZR, Aso Y, DePasquale GM, Enos A, Hulamm P, Lam SCB, Li HH, Laverty TR, Long F, Qu L, Murphy SD, Rokicki K, Safford T, Shaw K, Simpson JH, Sowell A, Tae S, Yu Y, Zugates CT (2012) A GAL4-Driver Line Resource for Drosophila Neurobiology. Cell Rep 2:991–1001.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36:5221–5231.

Kido Y, Nakae J, Accili D (2001) The Insulin Receptor and Its Cellular Targets 1 . J Clin Endocrinol Metab 86:972–979.

Kulkarni MM, Kulkarni MM, Sopko R, Sun X, Hu Y, Nand A, Villalta C, Moghimi A, Yang X, Mohr SE, Hong P, Asara JM, Perrimon N (2016) An Integrative Analysis of the InR/PI3K/Akt Network Identifies the Dynamic Response to Insulin Signaling. Cell Rep 16:3062–3074.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Liao Y, Smyth GK, Shi W (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930.

MacNeil LT, Walhout AJM (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res 21:645–657.

Marini F, Binder H (2019) PcaExplorer: An R/Bioconductor package for interacting with RNA-seq principal components. BMC Bioinformatics 20:1–8.

McNeill H, Craig GM, Bateman JM (2008) Regulation of neurogenesis and epidermal growth factor receptor signaling by the insulin receptor/target of rapamycin pathway in drosophila.

Genetics 179:843–853.

Morozova T V, Mackay TFC, Anholt RRH (n.d.) Transcriptional Networks for Alcohol Sensitivity in Drosophila melanogaster.

Mouchel-Vielh E, Rougeot J, Decoville M, Peronnet F (2011) The MAP kinase ERK and its scaffold protein MP1 interact with the chromatin regulator Corto during Drosophila wing tissue development. BMC Dev Biol 11:1–14.

Petruccelli E, Brown T, Waterman A, Ledru N, Kaun KR (2020) Alcohol Causes Lasting Differential Transcription in Drosophila Mushroom Body Neurons.

Post S, Karashchuk G, Wade JD, Sajid W, De Meyts P, Tatar M (2018) Drosophila Insulin-Like Peptides DILP2 and DILP5 Differentially Stimulate Cell Signaling and Glycogen Phosphorylase to Regulate Longevity. Front Endocrinol (Lausanne) 9:245.

Puig O, Marr MT, Ruhf ML, Tjian R (2003) Control of cell number by Drosophila FOXO: downstream and feedback regulation of the insulin receptor pathway.

Quillien A, Abdalla M, Yu J, Ou J, Zhu LJ, Lawson ND (2017) Robust Identification of Developmentally Active Endothelial Enhancers in Zebrafish Using FANS-Assisted ATAC-Seq. Cell Rep 20:709–720.

Roy S, Ernst J, Kharchenko P V., Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A, Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova AA, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, van Baren M, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SCR, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B,

Russell S, Cherbas P, Graveley BR, Lewis S, Micklem G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M, Booth B, Comstock CLG, Dobin A, Drenkow J, Dudoit S, Dumais J, Fagegaltier D, Ghosh S, Hansen KD, Jha S, Langton L, Lin W, Miller D, Tenney AE, Wang H, Willingham AT, Zaleski C, Zhang D, Acevedo D, Bishop EP, Gadel SE, Jung YL, Kennedy CD, Lee OK, Linder-Basso D, Marchetti SE, Shanower G, Nègre N, Grossman RL, Auburn R, Bellen HJ, Chen J, Domanus MH, Hanley D, Heinz E, Li Z, Meyer F, Miller SW, Morrison CA, Scheftner DA, Senderowicz L, Shah PK, Suchy S, Tian F, Venken KJT, White R, Wilkening J, Zieba J, Nordman JT, Orr-Weaver TL, DeNapoli LC, Ding Q, Eng T, Kashevsky H, Li S, Prinz JA, Hannon GJ, Hirst M, Marra M, Rooks M, Zhao Y, Bryson TD, Perry MD, Kent WJ, Lewis SE, Barber G, Chateigner A, Clawson H, Contrino S, Guillier F, Hinrichs AS, Kephart ET, Lloyd P, Lyne R, McKay S, Moore RA, Mungall C, Rutherford KM, Ruzanov P, Smith R, Stinson EO, Zha Z, Artieri CG, Malone JH, Jiang L, Mattiuzzo N, Feingold EA, Good PJ, Guyer MS, Lowdon RF (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science (80-) 330:1787–1797.

Sánchez-Alegría K, Flores-León M, Avila-Muñoz E, Rodríguez-Corona N, Arias C (2018) PI3K signaling in neurons: A central node for the control of multiple functions. Int J Mol Sci 19:3725.

Song L, Crawford GE (2010) DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 5:pdb.prot5384.

Tokusumi Y, Tokusumi T, Shoue DA, Schulz RA (2012) Gene regulatory networks controlling hematopoietic progenitor niche cell production and differentiation in the Drosophila lymph gland. PLoS One 7:41604.

Wilczynski B, Furlong EEM (2010) Challenges for modeling global gene regulatory networks during development: Insights from Drosophila. Dev Biol 340:161–169.

Yu G, Wang LG, He QY (2015) ChIP seeker: An R/Bioconductor package for ChIP peak

annotation, comparison and visualization. Bioinformatics 31:2382–2383.

Zhang W, Thompson BJ, Hietakangas V, Cohen SM (2011) MAPK/ERK signaling regulates insulin sensitivity to control glucose metabolism in Drosophila. PLoS Genet 7:1002429.

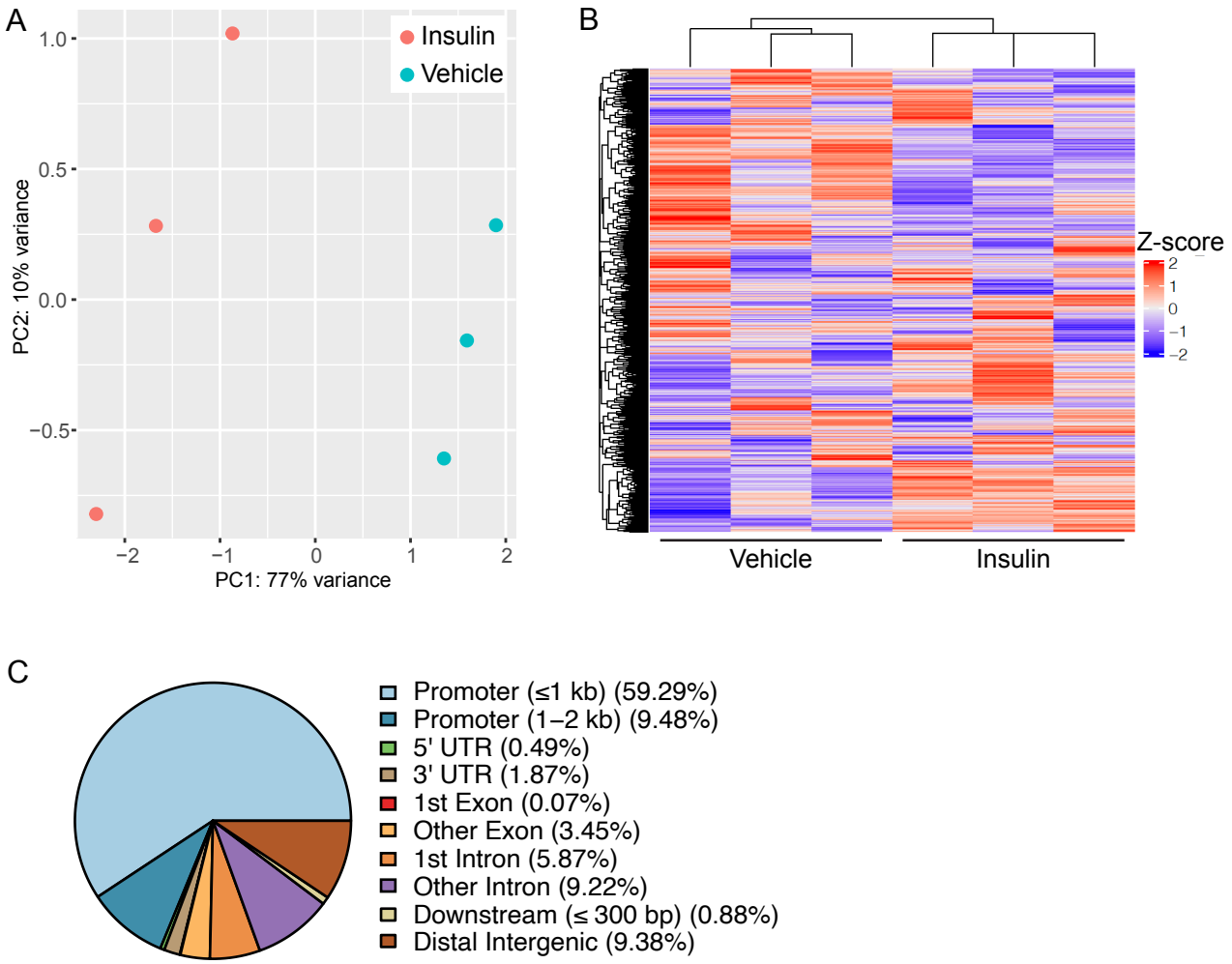

Figure 1. ATAC-seq elucidates the chromatin landscape in S2 cells after insulin treatment. A) Principal component analysis of S2 cells after treatment with insulin or vehicle. B) Heatmap of all chromatin peaks detected by ATAC-seq in S2 cells. The colors represent the z-scores calculated for each row in the heatmap. Each row represents a detected chromatin peak. Red indicates a more-accessible peak and blue indicates a less-accessible peak. C) Proportions of each genomic feature type in all annotated chromatin peaks analyzed by ATAC-seq.

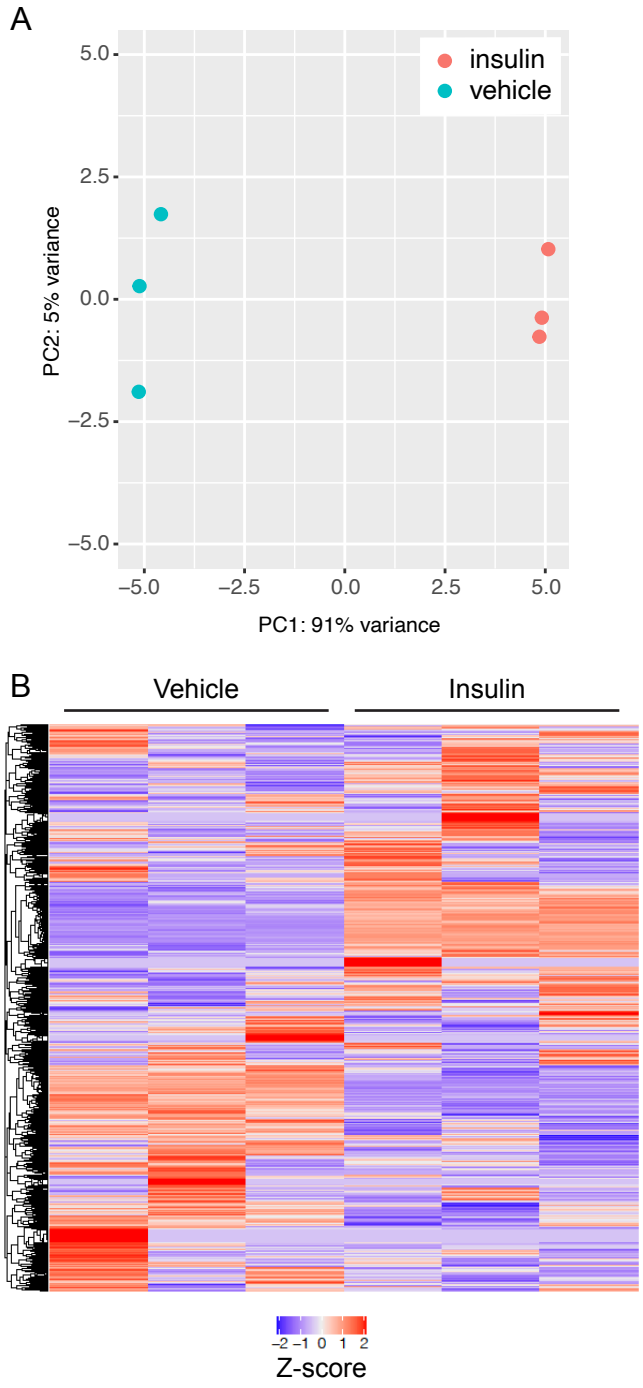

Figure 2. RNA-seq identifies gene expression in S2 cells after treatment with insulin. A) PCA of genes identified by RNA-seq in vehicle- and insulin-treated S2 cells. B) Heatmap of expressed genes identified by RNA-seq in S2 cells after treatment with vehicle or insulin. Each row represents an individual gene, and the values in each row were scaled to the row mean. Red indicates increased expression and blue indicates decreased expression.

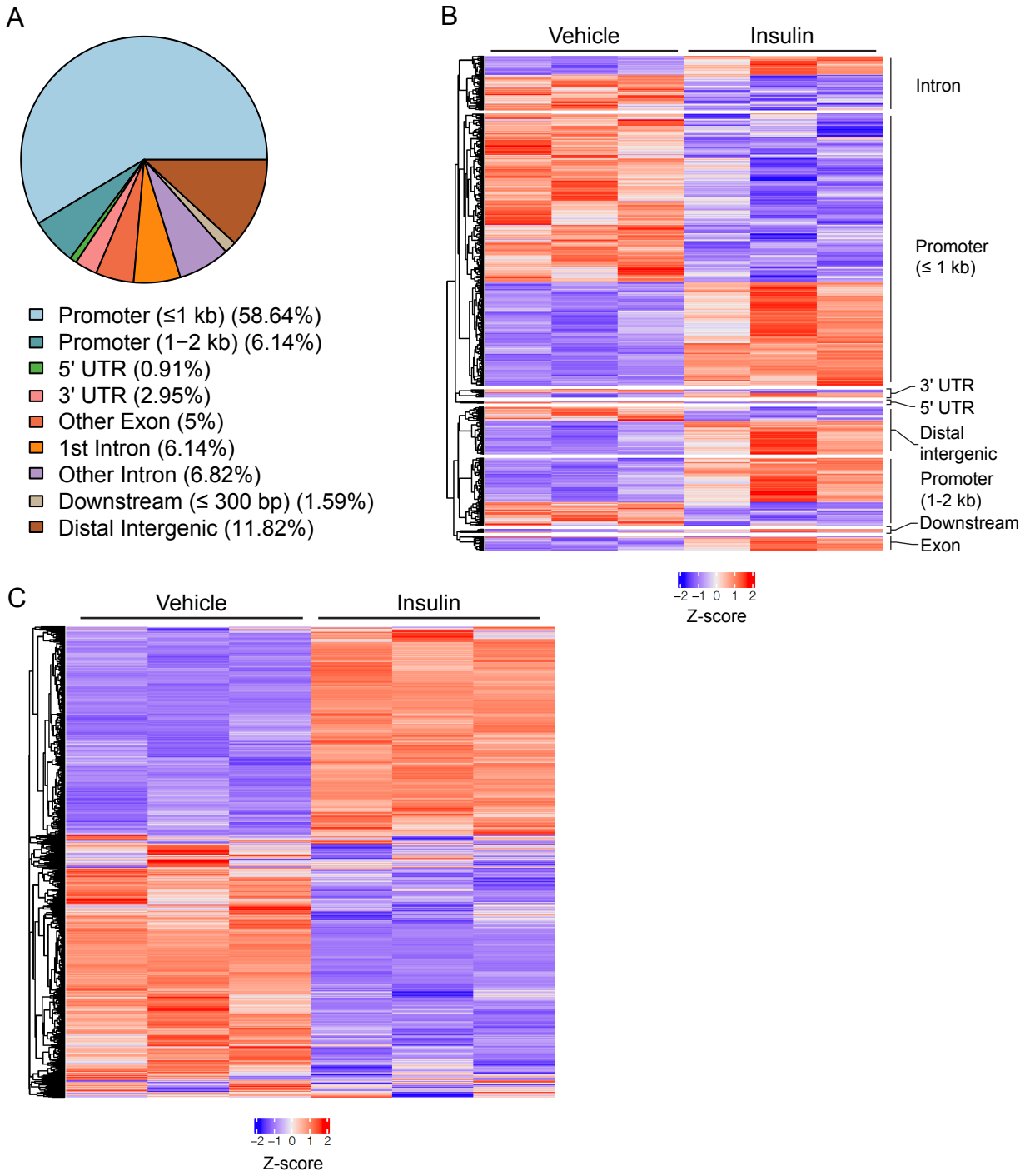

Figure 3. Insulin induces differential chromatin accessibility and gene expression in S2 cells. A) Proportions of genomic features annotated to differentially-accessible chromatin peaks analyzed by ATAC-seq. B) Heatmap of differential chromatin accessibility in significantly different chromatin peaks, stratified by feature type. Each row represents an individual chromatin peak. The values in each row were scaled to the row mean. Red indicates more-accessible chromatin regions and blue indicates less-accessible regions. C) Heatmap of differential gene expression in S2 cells after treatment with vehicle or insulin. Each row represents a differentially-expressed gene. The values in each row are scaled to the row mean. Red indicates an upregulated gene and blue indicates a downregulated gene.

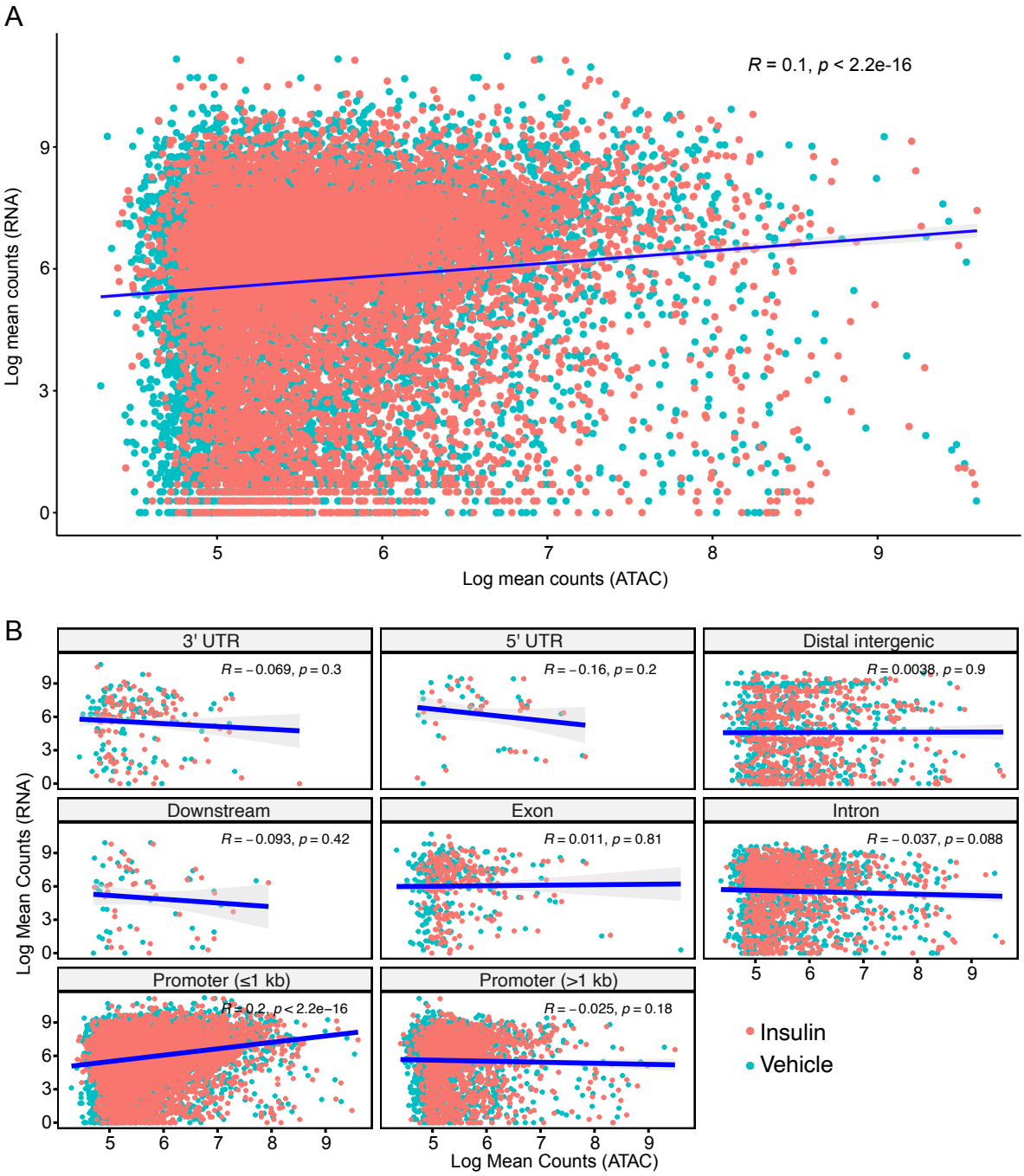

Figure 4. Normalized counts do not efficiently identify chromatin peaks for functional validation. A) Genes identified by RNA-seq were overlapped with chromatin peaks annotated to the same genes. The normalized ATAC- and RNA-seq counts were log scaled and plotted for Pearson correlation analysis. B) Overlapping ATAC- and RNA-seq counts from (A) were stratified by genomic feature. Pearson correlation analysis was used to identify correlations between ATAC- and RNA-seq counts by feature.

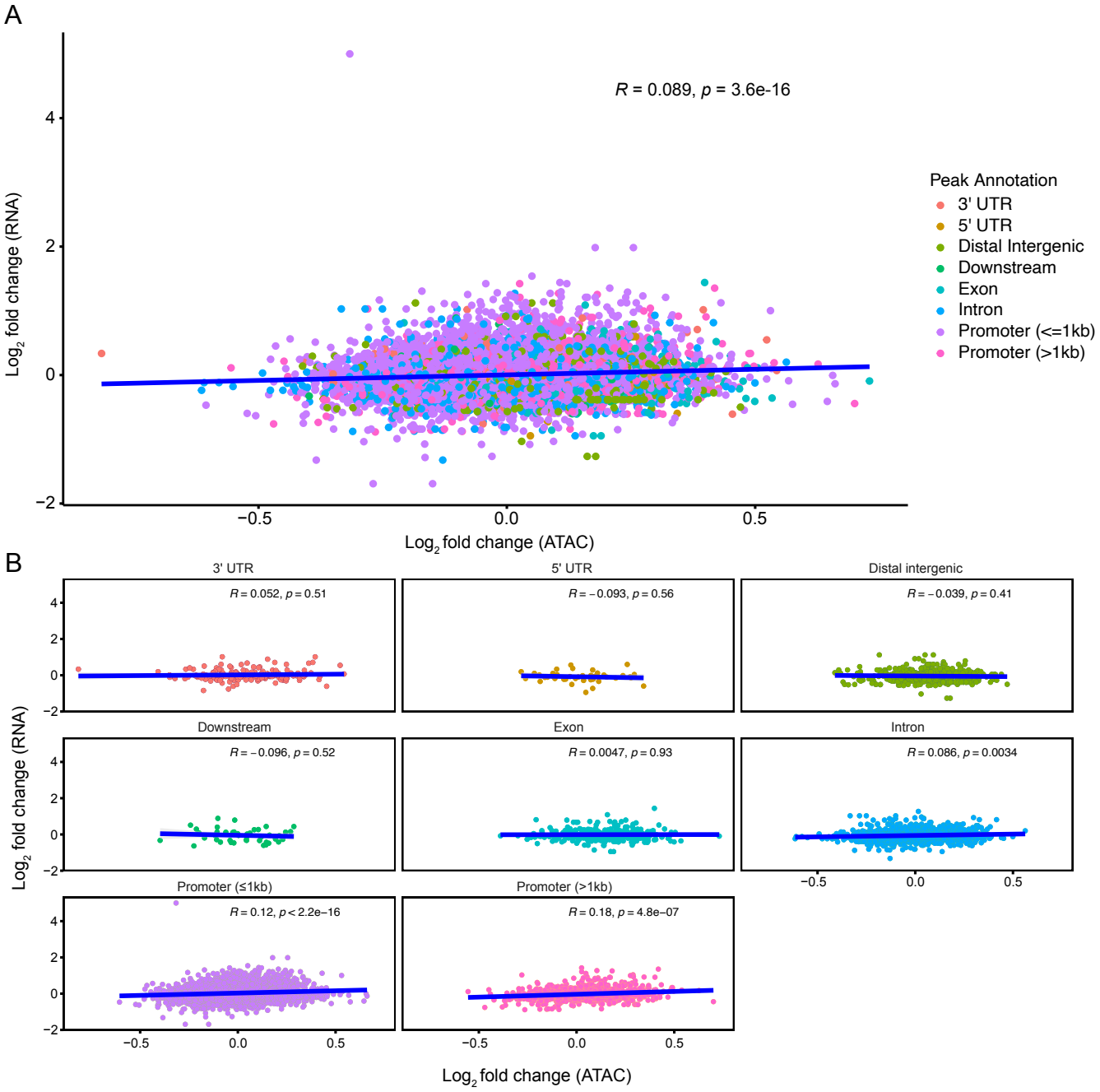

Figure 5. Log2 fold changes identify annotated chromatin peaks for functional validation. A) Chromatin peaks were overlapped with expressed genes. Pearson correlation analysis shows a very weak correlation between log2 fold changes in chromatin accessibility and gene expression. B) Overlapping ATAC- and RNA-seq log2 fold change values from (A) were stratified by genomic feature. Pearson correlation analysis was used to identify correlations between ATAC- and RNA-seq counts by feature.

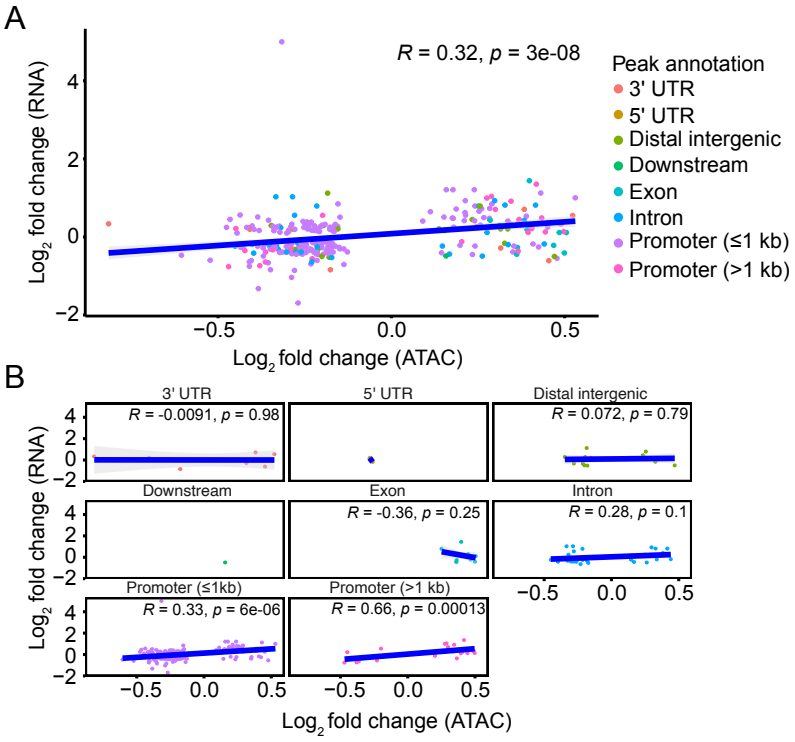

Figure 6. Correlating chromatin peaks with significantly different chromatin accessibility and differential gene expression homes in on chromatin regions for functional validation. A) Log2 fold change values from differentially-accessible chromatin peaks and differentially-expressed genes were analyzed by Pearson correlation analysis. B) Overlapping chromatin peaks and differentially-expressed genes from (A) were stratified by feature type. Significant correlations identify genome features suitable for functional validation.

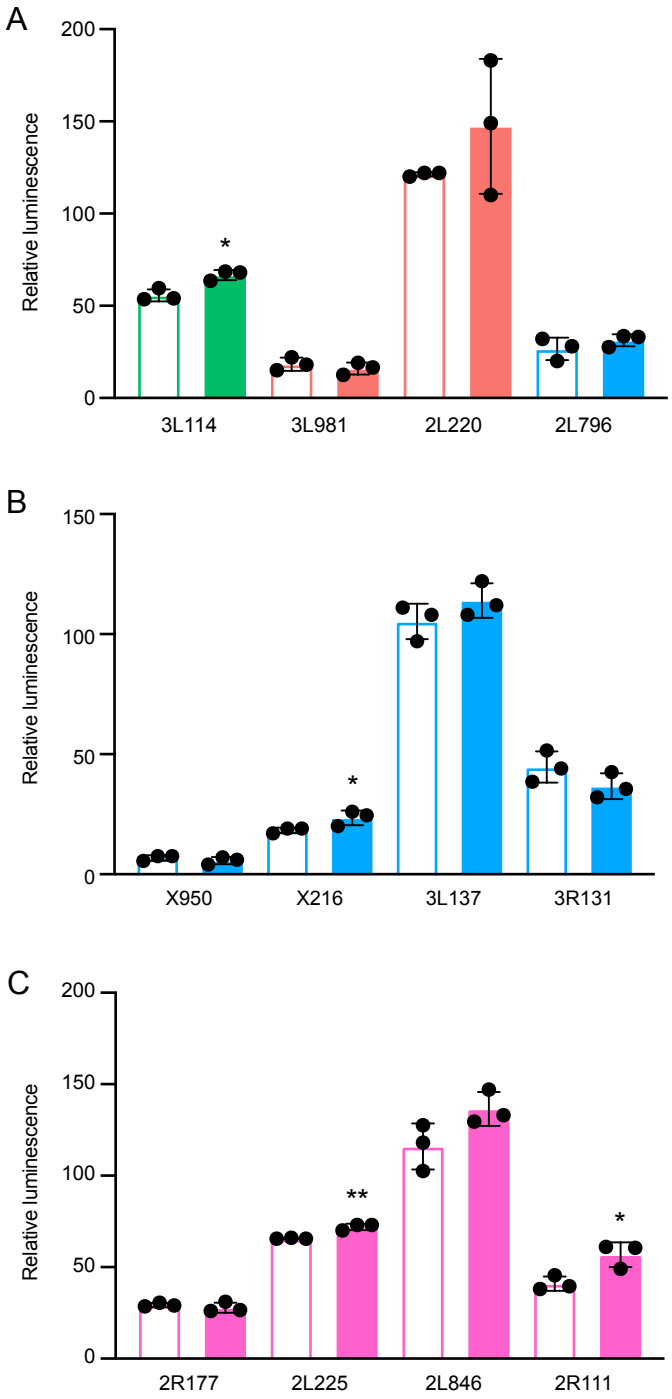

Figure 7. Differentially-accessible chromatin regions induce increased luciferase activity. A) Candidate chromatin peaks selected by largest log2 fold change were cloned upstream of luciferase for functional validation. B) Chromatin peaks annotated as introns with significantly different accessibility were selected for further validation. C) Chromatin peaks were selected from distal promoters (1-2 kb from TSS), which had the highest correlation with differential gene expression. Data represent means ± standard deviation of three biological replicates. Differences were analyzed by Student's t-test. *$p < 0.05$, **$p < 0.01$.