UNIVERSITY OF CALGARY

Advanced Computational Approaches to ‘Omics Data Set Analyses

by

Tyler Kolisnik

A THESIS

SUBMITTED TO THE CUMMING SCHOOL OF MEDICINE

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF BACHELOR OF HEALTH SCIENCES HONOURS

Bachelor of Health Sciences
Cumming School of Medicine
University of Calgary
Calgary, AB

March 2015

Abstract

Tyler Kolisnik, Bachelor of Health Sciences in Bioinformatics. University of Calgary, 2015. Advanced Computational Approaches to 'Omics Data Set Analyses.

Primary Supervisor: Dr. Mark Bieda

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a relatively new but widely used method for obtaining genome scale localization data on a ChIP target, such as a transcription factor or histone modification. The declining cost of sequencing has led to an enormous expansion in the use of this approach. Gene expression microarray analysis is a commonly used method for obtaining information on gene expression levels within a cell. While each of these approaches are highly useful on their own, in combining the two meaningful results regarding transcription factor function within the genome can be deduced. We present a series of 25 R scripts for ChIP-seq analyses, gene expression analyses and combined ChIP-seq and gene expression analyses. These R scripts were constructed as part of a larger project to produce software suites for ChIP-seq and gene expression data analysis. We present data analysis results derived from the use of these scripts with comparisons to current literature and expected results, which serve to confirm the functionality of our programs. In addition, we perform and present a novel pathway analysis for gene expression data for corticosteroid exposure in human lung tissue samples. In sum, our software allows the user to uncover a wealth of information about ChIP targets, including binding motifs, transcription start site coverage, genomic binding locations, and the expression levels of the genes at which the target is present or absent. The YesWorkflow management system was employed to provide visual diagrams of the R scripts, in order to encourage user-friendliness. These scripts were designed with an ideology of reproducible research in mind, and as such they should be fully comprehensible, modifiable and extensible by bioinformaticians in this area of research

Acknowledgements

The completion of this thesis would not have been possible without the help and patience of several important people. Firstly, I would like to acknowledge my supervisor, Dr. Mark Bieda for his expert level guidance and teaching that made this thesis possible. Working in the Bieda lab has been an amazing experience that nurtured my understanding of bioinformatics and science in general. Dr. Bieda has always emphasized the importance of learning above all else, and I know that the information he has taught me about computer science and bioinformatics will stay with me throughout my career and life. I would also like to thank Nathan Cormier for his advice and collaboration, and for making my time in the Bieda lab more enjoyable.

Thanks to Dr. Robert Newton for providing me with gene expression data to be analyzed and trusting me with finding results for him.

Thanks to Dr. Bertram Ludäscher and the YesWorkflow team for allowing my R scripts and subsequently generated YesWorkflow diagrams to be used as the main visual examples for the YesWorkflow project paper and IDCC 2015 presentation.

I would also like to thank my Research in Progress group members, and my preceptors, Dr. Fabiola Aparicio-Ting and Dr. Jonathan Lytton, whose feedback helped me guide my work towards a successful thesis.

I would like to thank the University of Calgary and the Alberta Children's Hospital Research Institute, for providing me with an outstanding education that prepared me for this thesis, and a lab space to work in. Additionally, I would like to thank my friends and fellow honours thesis students whose determination and perseverance pushed me to strive for excellence.

Lastly, I would like to thank my parents, Teresa Gallik and Steven Kolisnik, who have loved and supported me throughout my education, and without whom none of my successes would be possible

Dedication

This thesis is dedicated to my late grandparents, Albert and Genevieve Gallik.

Table of Contents

List of Tables

List of Figures

Introduction

Advanced Computational Approaches to 'Omics Data Set Analyses

Modern biological research has progressed to an era where experiments conducted often result in the production of data sets too large for analysis by hand. Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) and Gene Expression Analysis by microarray are two common experimental methods that produce these large data sets. Given the size of the human genome at 3 billion base pairs, searching and mapping sequence reads from ChIP-seq's next generation sequencing (NGS) data requires extensive computational power and state of the art algorithms. The data sets produced by NGS often lead to files that are in the tens of gigabytes size range, as they contain millions of lines of information that must be statistically analyzed. Gene expression analysis is another commonly conducted experimental method, it is performed on microarrays that can contain tens of thousands of probe sets that need to be linked to their corresponding known genes. This leads to large data sets of gene expression data that can be correlated with gene ontology terms and pathways, as well as be linked with other downstream analyses. The analysis of ChIP-seq and gene expression data often involves numerous complex steps which can be quite cumbersome to even an experienced user, this complexity increases in an order of magnitude when data from the two experiments are combined.

In addition to the vast increases in output data sets generated, with the cost of these experiments getting lower and lower yearly, the total number of data sets requiring analyses has increased exponentially. Furthermore, these experiments are often conducted on numerous different kinds of platforms that output data sets of different kinds, causing compatibility issues to arise. With this in mind, the need for universal scripts that can perform analyses of big-data

generating experiments using the most widely-accepted modern techniques is obvious.

Kepler pipelines have been previously used to design workflows for microarray data processing by Stropp et al. 2012 (1, 27), and in 2015 Cormier et al. (under submission) aim to present a series of pipelines for ChIP-seq analysis (1, 7). While this platform has many user-friendly advantages, it is not a commonly used standard within the field. We present a series R scripts to go along with Cormier et al's Kepler workflows for ChIP-seq analysis, as well as R scripts for independent and paired analyses of Gene Expression and ChIP-seq data (Table 1). R scripts are a very popular standard for scripts in the field of bioinformatics. These R scripts have been designed with modifiability and extensibility in mind and alongside them we present the corresponding YesWorkflow diagrams to increase user-friendliness.

Background Information

Chromatin-Immunoprecipitation followed by sequencing (ChIP-seq) is a common method for obtaining information about how the proteome and genome interact with one another. This method was invented in 2007 by Barski et al. and was proven to be more sensitive and specific than the previous methods of Chromatin-Immunoprecipitation followed by microarray analysis (ChIP-chip). As a result, by 2009 the Barski et al. was the third most cited biology paper in the world and it continues to hold an important place in the field today (4). The importance of ChIP-seq lies in what it allows us to learn about protein-DNA interactions. Specifically, ChIP-seq allows us to characterize all of the locations in the genome where these proteins are present, identify their relative position to nearby genes, and characterize the typical binding motifs they associate with. From this information we are able to make predictions about the structure and function of the DNA-binding protein.

Gene expression patterns determine cellular behavior and identity, gene expression

microarrays and their analysis subsequently give us insight into these cellular characteristics. Gene expression analysis by use of microarrays has been in wide use for many years, with the most common arrays being produced by Affymetrix. These arrays use probes or probe sets which hybridize to fluorophore labeled cDNA or cRNA and fluorophore detection allows us to measure gene expression levels (9).

Most modern laboratories send their ChIP-seq and Gene Expression data away to a hired bioinformatician or analysis company, which may charge thousands of dollars for their services. While there are free to use tools for the analysis of ChIP-seq and gene expression data on the web, such as Galaxy (11), these are oftentimes confusing to non-bioinformaticians and there are limitations where extensibility and modifiability are concerned. In addition to this, uploading and downloading large data files to the internet often adds unnecessary lag time when many of the analyses could be conducted in-house on a standard laboratory computer. The series of R scripts we present, as well as Cormier et al's (under submission) pipelines, will allow for a complete integrated software suite that allows non-bioinformaticians, and bioinformaticians alike the ability to process and customize ChIP-seq and gene expression data analysis (7).

The R scripts we present were constructed with an idea of flexibility in mind, taking into account that a user may want to analyze data all at once, or in a stage-by-stage process. As such, numerous pipelines for combined and partial analyses were developed. This allows a user options in choosing which parts of the suite they wish to include or exclude. Many of our scripts are integrated with external software programs, namely, Multiple EM for Motif Elicitation (MEME), Hypergeometric Optimization of Motif EnRichment (HOMER), and Model-Based Analysis of ChIP-seq (MACS) (3, 12, 30). All of these external programs have been have been established as reliable standards within the field, and are used in many of the latest ChIP-seq

analysis studies (3, 19). In combining these external programs with R code, and R Bioconductor packages we were able to produce 25 scripts for ‘omics data set analyses.

The R scripts we produced are outlined in Table 1. The first 20 scripts listed are for partial and complete ChIP-seq analyses that mirror the Kepler workflows produced by Cormier et al (7). These allow for all steps of ChIP-seq analysis prior to read mapping, including peak calling, peak annotation, motif analysis, and custom UCSC genome browser track creation. The subsequent two scripts are for gene expression analysis from both .CEL files (which need to be normalized) and from already normalized data. These produce pathway diagrams, lists of differentially expressed genes, and Gene Ontology (GO) statistics. The final three scripts are implicated in combining results from ChIP-seq analysis with gene expression data. We present a script (CombineGEoneChIP.R) for combining ChIP-seq data from one target in a cell type with gene expression data from an untreated sample of that cell type. This allows for the determination of: 1) What genes does the ChIP-seq target affect; and 2) What are the transcription levels of those genes, relative to other genes in the cell. We also present scripts (CombineGEoneChIP.R. GetDiffPeaks.R) for combining two ChIP-seq samples (experiment, control) for a target, with gene expression data from the same two samples in order to determine: 1) What genes are inhabited by the target in both samples; 2) What are the expression levels of those genes in both samples; and 3) What are the differences in gene expression levels. We can then subsequently infer potential functionality of the ChIP target by associating it with the up and down regulation of particular genes or pathways.

In order to verify the accuracy of these scripts, we were able to compare our results with expected results, and with the results of reputable studies. Studies by Cormier et al. 2015 used GATA1 and H3K27me to verify their Kepler pipelines for ChIP-seq successfully, and while our

R scripts follows a nearly identical approach, we chose to verify them through analyzing two new datasets for H3K27ac data in K562 cells and H3K27ac data in MCF-7 cells and comparing our results with expected results from the literature (7). The R script for gene expression analysis from .CEL files (GEAnalysisFromCEL.R) was previously verified by us and is presented in a poster presentation, by comparison to data from a study by Beier et al. (5, 18). In this thesis, this script was adapted to produce GEAnalysisFromNormalized.R and was used to produce de novo results from a gene expression microarray experiment. Lastly, the R scripts for combined gene expression with differential ChIP-seq were verified by a direct comparison to the literature.

Methods

The R language for Statistical Computing

The primary scripts for this project were designed using the R language. This language was chosen partly due to my previous experience in R, but largely because it is widely used in the field of bioinformatics for the analysis of gene expression and ChIP-seq data, making it a standard within the field. R allows for the easy manipulation of large datasets and allows the user access to thousands of built-in statistical and plotting functions. In addition to this, there are external R packages that have been developed, such as the many seen on the R/Bioconductor website, which can be easily implemented to fulfill other data analysis needs. It should be noted that these programs were designed around the latest R version, 3.1.2 (Pumpkin Helmet) and compatibility issues may arise if using another version (24). At present the Bioconductor packages used in the creation of these R scripts allow for the analysis of data from 532 different organisms (8, 21).

ChIP-seq Analyses

Raw sequence read data for H3K27ac binding in MCF-7 and K562 cells were obtained

from the ENCODE projects data at NCBI GEO: GSM945854, GSM733656, respectively (6). Raw sequence read data for glucocorticoid receptor (GR) binding in dexamethasone exposed (2 samples) and control (2 samples) ATt-20 cells were obtained from the Sequence Read Archive (SRA) SRX034872, SRX034871, SRX034870, SRX034868, respectively (20). The raw sequence reads for MCF-7, K562 were mapped to hg19, and those for AtT-20 were mapped to mm9 using Cormier et al.'s 2015 (under submission) pipeline (7). Next, these mapped reads were ran through the script "fullPipelineScript.R", and results were obtained.

Combined Gene Expression Analysis with ChIP-seq data from one sample

Gene expression data for untreated MCF-7 was obtained from GEO: GSE15805. These .CEL files, along with the ChIP-annotated gene data from the corresponding MCF-7 ChIP-seq experiment were inputted into the R script CombineGEoneChIP.R, which normalized the .CEL file data using a Robust Multiarray Average (RMA) and annotated the probesets with gene names and expression levels. It subsequently combined this list with the list of annotated genes from the MCF-7 ChIP-seq experiment in order to determine gene expression levels at all of the genes showing ChIP target presence or absence, and outputted a bar plot and histogram comparing the two, as well as performing a t-test in order to determine the statistical significance of the differences between the average expression levels in these two gene groups (Figures 8, 9). Analysis of K562 cells followed an identical procedure, gene expression data for untreated K562 cells was also obtained from GSE15805, and results are shown in Figures 10, 11.

Differential Peak and Gene Finding between two ChIP-seq analyses

Mapped sequence reads and peaks from the result of the ChIP-seq analyses of GR binding in dexamethasone exposed (2h @ 100nM) and unexposed AtT-20 cells were inputted into the GetDiffPeaks.R script. This script makes use of the MAnorm program to locate

differential peaks were calculated using the GetDiffPeaks.R script which employs MAnorm (26). The resultant list of peaks unique to dexamethasone exposed, unique to unexposed, and common to both was inputted into the R script annotatePeaks.R which mapped these peaks with nearby genes.

Gene Expression Analyses with paired ChIP-seq and gene expression data from two samples

Gene expression data for dexamethasone exposed (2h @ 100 nM) and control AtT-20 cells was obtained from GEO: GSE26189. Along with this gene expression data, the list of genes unique to dexamethasone exposed AtT-20 cells from the differential peak and gene finding between two ChIP-seq analyses was inputted into the R script CombineGEtwoChIP.R. This script normalized the gene expression levels using the RMA method, and matched the list of genes with the corresponding expression levels in each of the two cell types. The difference in gene expression levels were taken and plotted into the histogram (Figure 12).

Important Methods for ChIP-seq Analysis:

These scripts employ various external programs in order to conduct a thorough analysis of ChIP-seq data. The program MACS is used for determination of peaks from the mapped reads data (30). The program Homer is used in order to calculate enrichment at TSS for the production of the TSS region coverage heatmap (Figure 4) and the TSS region coverage plot (Figure 5) (12). The program MEME is also employed for discovering the top binding motifs of the ChIP target (Figures 7 A-D) (3).

Gene Expression Analysis from Normalized Microarray Data

Normalized Gene expression data for budesonide exposed vs unexposed human lung biopsy tissue samples were obtained from the lab of Dr. Robert Newton (University of Calgary). This data was inputted into the R script GEAnalysisFromNormalized.R which combined them

with the Gage and Pathview R/Bioconductor packages to produce lists of genes which were up regulated, down regulated, and tables of up and down regulated pathways, (Table 6, 7) and pathway diagrams.

Affy/SimpleAffy

In order to read in gene expression data from affymetrix based platforms, the R packages Affy and SimpleAffy were employed. These packages allowed for high-level analysis of data from affymetrix arrays at the probe level. These packages allowed us to combine data from multiple .CEL files and subsequently normalize it through a Robust Multi-array Average normalization (RMA), before combining the data into a meaningful data structure called an expression set for easy manipulation. (9, 23).

Gene Ontology analysis with GOStats

Gene Ontology (GO) is a bioinformatics tool that allows us to link genes with attributes through a database. Specifically, it allows us to group genes into GO categories which fall under three domains: Cellular Component (CC), Molecular Function (MF), and Biological Process (BP). Furthermore the specific GO term associated with the gene or group of genes is given, which allows us to quickly associate them with a function. While a simple gene expression analysis can tell us what genes are up and down regulated in a cell type, it is not until GO analysis is performed that we know what those genes do (2, 8). Where ChIP-seq alone is concerned, GO analysis can associate functional terms with the genes inhabited by the ChIP target, giving hints at what processes the target controls. A sample of GO analysis results is given in Table 5. GOStats is the R/Bioconductor package that allows for easy linking to the GO database so that Gene Expression data can be quickly linked to these GO terms from an R script (2, 8).

Pathway analysis with KEGG and Pathview

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database that allows for the integration of gene information with genomic pathways, as well as chemical and systemic functional information. In addition, the KEGG database provides intricate and highly-detailed diagrams of thousands of biological pathways. Pathview is the R/Bioconductor package that serves as the tool for linking gene and gene expression information into KEGG pathway diagrams. In particular, it allows us to take our gene expression data from an experiment and determine what pathways in a cell are upregulated or down regulated, and view a graphical representation of the pathway along with all of its genes, color-coded in association with their expression levels, an example is shown in Figures (13, 14). In addition to these graphical representations, tables are outputted that show lists of the up and down regulated pathways and their associated p values (Tables 6, 7) (15, 16, 21).

YesWorkflow

YesWorkflow implementation allowed for workflow diagrams of the R scripts to be generated so that a user who wishes to make use of them can view the diagram and quickly ascertain the inputs, outputs and overall steps of the script. This software allows in-code comments directly imbedded into R scripts (or scripts of any programming language) to be converted directly into outputted diagrams (typically in png, jpeg or pdf format). We felt that it was important to maintain the user-friendliness found in the Kepler pipelines within the R scripts, and YesWorkflow allowed us to accomplish this. Examples of YesWorkflow diagrams are shown in Figures 16-20, and S2. Our R script GEAnalysisFromNormalized.R was introduced at the IDCC ’15 conference and included in the subsequent IJDC paper (Figure 21) as an example of proper use of in-code comments for the generation of a meaningful YesWorkflow

diagram in a bioinformatics script (22).

Data Sources

ChIP-seq reads:

Most ChIP-seq data used in these analyses was taken from the ENCODE project, which is believed to be a reputable source for high-quality data (6). ChIP-seq data from MCF-7 cells was taken from ENCSR000EWR (GEO: GSM945854) SRA Accession: SRX504925, SRX504926 (control). ChIP-seq data for K562 cells from ENCSR000AKP (GEO: GSM733656) SRA Accession: SRR227385, SRR227386 (control). ChIP-seq data from AtT-20 cells was taken from SRA SRP004871, specifically: SRX034872 (control), SRX034871 (control), SRX034870, SRX034868. The SRA Toolkit was used to convert all .sra files to .fastq files for analysis (20).

Gene Expression Microarray Data:

Gene Expression Data for MCF-7 cells were obtained from GEO: GSE15805, specifically, the vehicle controls: GSM993575, and GSM 993576. Gene Expression Data for K562 cells were obtained from GEO: GSE15805, specifically, the vehicle controls: GSM993551, GSM993552, GSM993553, GSM993554, GSM993555, GSM993556, and GSM993557. Gene Expression Data for Human Lung Biopsies for Budesonide-Exposed and Control samples were obtained from the Lab of Dr. Robert Newton at the University of Calgary. Gene Expression data for treated and AtT-20 cells was obtained from GEO: GSE26189, specifically: GSM642875 (2h), GSM642876 (2h), GSM642877 (control), GSM642878 (control).

Arrays Tested:

Gene Expression Data Analysis was tested on the common arrays shown in table 8. The scripts are compatible with all arrays made by Affymetrix for which there are R/Bioconductor annotation packages. At present this allows for analysis of data from over 532 different species.

This includes high-density and classic gene arrays, as well as exon/RNA arrays. Given the universality of the Affymetrix array and R/Bioconductor annotation package design, despite having only tested a subset of commonly used arrays we are confident that the pipeline will easily handle alternate array types.

Results

The overarching result of this thesis are the 25 R scripts that allow for ChIP-seq analysis, gene expression analysis, and combined ChIP-seq and gene expression analysis, shown in Table 1. In verifying the functionality, reliability and reproducibility of these scripts, numerous results were obtained for the studies analyzed.

H3K27ac up regulates gene expression in both MCF-7 and K562 cells: A comparison of ChIP-seq and gene expression data using CombineGEoneChIP.R

The R script fullPipelineScript.R was able to successfully take in mapped sequence read data for H3K27ac in MCF-7 cells and perform a complete ChIP-seq analysis. Peak calling was successful, and resultant peak statistics were calculated (Table 2), showing that 23623 peaks were found and 1.641% of the genome is covered with peaks. These peaks were then annotated to determine what genes they correspond to within a range of 5000bp upstream and 2000bp downstream, this gene list is later used for combining results with gene expression analysis (Tables 3 and 4). Gene ontology analysis was performed (Table 5) to determine which functional terms are likely controlled by H3K27ac, a likely important one being the nucleus. A UCSC browser track was created from these mapped reads, and it was uploaded to the UCSC genome browser for comparison with H3K27ac ChIP-seq data from 6 different ENCODE experiment cell types (Figure 1) (6, 25). The MCF-7 data shows a high correlation with the ENCODE data, with well-defined peak and non-peak areas that match. Figures 2 and 3 show example peaks

automatically generated by the script. Figure 2 demonstrates a high-scoring peak, indicative of likely H3K27ac presence in that particular genomic location. Figure 3 shows a region devoid of peaks, indicating there is no H3K27ac presence in this region. In order to determine the behavior of H3K27ac around transcription start sites, Figures 4 and 5 were generated. Figure 4 is a heatmap of all the locations of all the mapped reads relative to the TSSs, within a +/- 5000 base pair range. The abundance of these reads in a line above 0 on the x axis show that H3K27ac exhibits TSS-centric behavior, which is something we know from the literature (17, 28, 29). Figure 5 demonstrates the average read coverage (tag density per base pair per TSS) around each TSS. This gives a more clear view as to where this modification is found specifically, in regions directly surrounding base pair 0 of the transcription start sites. These results have been confirmed in literature, a study in 2012 by Koike et al generated a nearly identical figure of the average TSS coverage of H3K27ac throughout 6 circadian rhythm cycles in mouse liver cells (Figure 6) (17). The final step in the ChIP-seq analysis was the production of the binding motifs (Figure 7), however, the usefulness of binding motifs for histone modifications is a contested topic, and this step is likely more useful for ChIP-seq analysis on transcription factors.

The annotated gene list was obtained and combined with gene expression data through the R script CombineGEoneChIP.R to produce Figures 8 and 9. Figure 8 demonstrates a bar plot of the average gene expression level of all of the genes in a control type MCF-7 cell where H3K27ac was present (7.6 in Log_2 terms) versus the average gene expression levels of all of the genes where the H3K27ac was not found to be present (5.8 in Log_2 terms). A t-test was performed on these two data sets. The resulting p value was found to be $p = 2.2x10^{-16}$ for a difference in the means of -1.507566 with a 95% confidence interval of (-1.579433, -1.435699). This provides us with evidence against the null hypothesis that there would be no difference

between the means gene expressions, and allows us to conclude with high certainty that genes with H3K27ac presence have higher expression levels than those that do not. Figure 9 shows an alternative depiction of these altered expression levels, in the form of a multi-histogram, where genes of H3K27ac present and not present categories were grouped into sections of 0.5 by their expression levels ranging from 1-14, and plotted against the relative frequency of these genes. This figure shows that at higher expression levels (6.5 and up), the relative abundance of genes with H3K27ac peaks is significantly higher than that of genes without H3K27ac, and at lower expression levels (6 and under), the converse is true.

In order to further establish the correlation of heightened gene expression levels and H3K27ac presence, we examined ChIP-seq and gene expression data from K562 cells. We performed all of the same analyses as in the MCF-7 samples, and the resultant bar plot and histogram were produced. Figure 10 shows the bar plot for average gene expression of genes where H3K27ac is present compared to the average gene expression of genes where H3K27ac is not present in K562 cells. This plot showed an average gene expression level of 7.3 (in Log_2 terms) where H3K27ac was present and 5.5 (in Log_2 terms) where not present. A t-test was conducted and the p value found to be $2.2x10^{-17}$, with a difference in the means of -1.2896981 with a 95% confidence interval of (-1.358675,-1.220721). Again, this provides us with evidence against the null hypothesis that there is no difference in the average gene expression at genes where H3K27ac is present vs not present. The multi-histogram in Figure 11 shows a similar pattern to that in Figure 9, at higher gene expression levels (7 and above) higher relative frequencies of genes with H3K27ac presence are found than those without H3K27ac presence, and at lower expression levels (6.5 and below) higher relative frequencies of genes without H3K27ac presence are found than those with H3K27ac. Studies by Yang et al. and Zentner et al.

have established H3K27ac as a modification that increases gene expression levels, and this coincides with our result (28, 29). Given that these scripts were able to reproduce near identical results for characterizing the function of H3K27ac in two different cell types, and that this result coincides with the literature in that H3K27ac is a gene upregulating modification, we have high confidence that these scripts are performing accurately.

Dexamethasone exposure has no effect on gene expression levels of GR-binding genes: A Cross-Comparison of paired ChIP-seq and gene expression data using GetDiffPeaks.R and CombineGEtwoChIP.R

The R scripts GetDiffPeaks.R and CombineGEtwoChIP.R which we designed allowed for cross-comparing paired ChIP-seq data by determining the locations of peak intersections and the locations of unique peaks in each ChIP sample. These were then able to be mapped to genes to determine which genes the ChIP target is present at in both ChIP samples, and which genes are uniquely affected by each ChIP-target (Table 9). In order to test this we compared ChIP-seq data from a study by John et al. where ChIP-seq and gene expression data from *Mus musculus* AtT-20 cells treated with 100 nM dexamethasone for 2 hours and untreated controls were cross-compared (14). *GetDiffPeaks.R* found 5396 peaks unique to the treated cells, 1000 peaks unique to the untreated cells and 266 peaks common to both. The 5396 peaks unique to treated cells as our peaks of interest for this study, as these demarcate the locations where GR receptor binding was "gained" in the genome, following dexamethasone exposure. These contrasted the 3242 peaks found by John et al, however they designed their own peak-calling algorithm and used a very sensitive False-Discovery Rate (FDR) of 0% (14). Our peak calling program uses MACS which empirically estimates the FDR for each peak by comparing it to the control (30). Differences between Peak Callers is a highly debated topic in bioinformatics, and there is

currently no peak calling program that has been deemed superior to all others, however MACS performs competitively among the best (19, 30). We found that the called peaks correlated with 490 genes. Of these 490 genes we were able to obtain gene expression data for 118 true genes with official symbols. Our results for the difference in gene expression across these 118 genes in exposed and unexposed AtT-20 cells are shown in Figure 21. There was no visible differences between gene expression levels in the exposed and unexposed cells. Log2 fold gene expression level changes ranged from a 1 fold increase to a 1 fold decrease, with 67% of the genes showing a fold decrease of 0.5. From this we are unable to establish a relationship between gene expression level changes and GR binding. John et al. were able to correlate their peaks to 282 genes, pseudo-genes and genes with no official symbol (14). They found similar results, stating that "we found no clear relationship between glucocorticoid receptor occupancy patterns and transcriptional activation of nearby genes", indicating that for this data analysis we have obtained the same results (14).

Analysis of gene expression data from human lung tissue samples exposed to budesonide versus controls illustrated affected pathways and highlighted novel target genes

After normalized gene expression data from budesonide exposed human lung biopsy tissue samples and controls was placed into the R script GEAnalysisFromNormalized.R gene expression results for pathway analysis were obtained. Table 6 illustrates the list of pathways found to be up regulated by budesonide exposure, notably, considering budesonide is a steroid hormone, the steroid hormone biosynthesis pathway (highlighted in blue in the table), is an expected result. Table 7 shows the list of pathways found to be down regulated by budesonide exposure. Figure 13 shows one of the KEGG pathway diagrams automatically generated by this R script, it illustrates the steroid hormone biosynthesis pathway, with shades of red that indicate

the relative levels of up regulation within each gene in the pathway. Figure 14 shows a similar diagram, however this is for the down regulated oxidative phosphorylation pathway, where the shades of red indicate the relative levels of down regulation for each gene.

Overall Program Results / User-Friendliness Results

For the 20 R scripts produced (Table 1, 1-20) that match Kepler pipelines by Cormier et al. 2015 (under submission), no YesWorkflow diagrams were deemed necessary as the Kepler pipelines themselves function as descriptive schematics of the program, Figure 15 shows a sample Kepler pipeline where these features can be observed. For the five remaining R scripts (Table 1, 21-25), YesWorkflow diagrams were produced (Figures 16-20) to complement the code with a schematic of inputs, outputs, and key steps. In addition the YesWorkflow diagrams, in order to encourage extensibility and modifiability of these scripts, several key coding principals were adhered to in their production:

- The Block Approach: Code within R scripts was divided into clear blocks, or sections, using comment characters in order to provide a clear visual division between segments of the code that perform different tasks.
- Descriptive Variable Names: Variable names were designed to be meaningful, and hint at their contents without being too cumbersome.
- Descriptive comments: Maximal description was used in in-code comments to avoid any confusion in regards to initial script set up, correct parameter usage, code section function, and outputs obtained.

Discussion

By these results, our specific aims to produce a series of R scripts for ChIP-seq analysis, gene expression analysis, and combined ChIP-seq and gene expression analysis are believed to

be successful. To reiterate the results: Our comparison of gene expression data and ChIP-seq data for H3K27ac binding in MCF-7 and K562 produced the result that H3K27ac binding up regulates gene expression, as was expected from the literature. In addition to this, we were able to produce mapped reads and TSS region coverage plots that highly correlated with literature samples for the same histone modification, showing that this histone modification has a tendency to be localized around TSSs. In our next study, cross comparison of ChIP-seq data from dexamethasone exposed and control AtT-20 cells showed that dexamethasone exposure had no effect on gene expression levels of GR-binding genes, a result mirrored in the study reproduced (14). Lastly, our final study result predicted de novo gene and pathway up and down regulation in human lung tissue cells, while certain results obtained were expected, the complete interpretation of these results will be documented in a future study by Dr. Robert Newton.

What sets our series of scripts apart from other analysis platforms is the emphasis on downstream analyses that combine data from two experimental platforms. In addition to this, with the YesWorkflow module and best coding practices, we present these scripts with the hope that users will modify them, extend them, or add our analyses as extensions to their own pipelines. Our results have successfully confirmed the roles of the highly characterized ChIP-seq targets, GR and H3K27ac, and in future steps studies of uncharacterized ChIP targets can take place to uncover their roles.

Workflow and Script Limitations

While we try our best to prevent them, all studies and software programs come with limitations that must be accounted for. These are outlined as follows:

Array Types

The gene expression workflows and scripts are only compatible with Affymetrix arrays

and only exon and gene microarrays. The arrays that have been tested and confirmed to work with these scripts are outlined in table 8.

Operating System Limitations

At present, the programs and R scripts use commands that are only supported by Unix operating systems. These will not work on Windows based systems. The programs have all been tested on Kubuntu 14.04 LTS (Trusty Tahr).

Computational Limitations

The programs were designed to be used on a standard laboratory computer in the year 2015. As a result it is recommended that the computer being used have at least 8 GB of RAM and at least 20 GB of free storage space to accommodate the size of the input and output files.

User-friendliness limitations

While a strong effort was made to accommodate for user-friendliness, the user must have basic computer skills and should be familiar with working in a Unix environment, installing programs, and inputting parameters. The user should also be familiar with the steps of the experiment design and the appropriate ChIP-seq/gene expression analysis approaches for that design. In addition they should ensure they are familiar with the YesWorkflow diagrams, so that they comprehend the correct inputs and outputs. While this program analyses experimental data, it does not interpret the results for you, and this may result in human errors. In addition, if the data inputted is from a poorly done experiment, this program will likely be unable to provide the user with meaningful results.

Depreciation Limitations

Given the fast pace of technological advancement in the field of bioinformatics, it is likely that certain parts of this program will begin to depreciate within the next few years. This

program was designed for the Linux operating system Kubuntu 14.04 LTS (Trusty Tahr) that will be supported until 2019. The most likely parts of this program to depreciate are the R packages used within the scripts, this may be able to be overcome by backdating R to the version this program suite was designed with, 3.1.2 (Pumpkin Helmet) or by forcing the out of date R packages to load.

Conclusions

In analyzing the ChIP-seq data from H3K27ac in MCF-7 cells, H3K27ac in K562 cells, GR in AtT-20 cells and gene expression data from Dr. Robert Newton and the matched ChIP-seq experiments, we have found that our scripts in large produce results that match the expected results from the literature. In doing this, we have shown reliability and proven the effectiveness of the presented scripts. Subsequently, in finding results that match the literature, we also verify the literature. In addition to this, we have used our scripts to find new results which will be presented in a later publication by Dr. Robert Newton, University of Calgary. These results follow a theme of reproducible research, an ideology that data and software must be openly published so that others can verify their claims and build upon them. By including the YesWorkflow management system, and by using best coding practices, we have designed scripts that will be comprehensible, modifiable, extensible and distributable.

All in all, the resultant scripts from this research collectively contribute to the publicly available bioinformatics resources, and given the way that they have been designed, and the results they have produced they may be a valuable resource for advancements and discoveries within the field of bioinformatics.

Tables:

Table 1. List of R scripts, descriptions and corresponding figures.

Rscript Name	Description	Corresponding Figure(s) and Table(s)
Scripts with matching Kepler Workflows		
fullPipelineScriptnoHomer.R	Takes in .sam mapped sequence read data and performs all of the below analyses which are not dependent on homer or output from homer	Figure 1, 7, Tables 2-7.
fullPipelineScript.R	Takes in .sam mapped sequence read file and performs all of the below analyses on the data	Figures 1-5, 7 Tables 2-7.
fullPipelineScript_startfromPeaks.R	Takes in a MACS peak file and its corresponding .bam mapped sequence read file and performs all of the below analyses on the data (does not create its own peak file)	Figures 1-5, 7 Tables 2-7.
startFromPeaks_noBAM.R	Takes in a MACS peak file and performs all of the below analyses which do not require the mapped sequence data.	Table 2 Figures 5,6
runMacs.R	Uses MACS peak caller to identify genomic regions with enhanced sequence read coverage. Takes in .bed sequece files and produces .bed peak files	Table 3
peakStats.R	Reads the peak files created by MACS and calculates descriptive statistics about the peaks.	Table 2
annotatePeaks.R	Uses an annotation database (user specified) to map peaks to their most proximal gene (within a set distance). Uses a MACS peak file as input.	Table 4
TSSdensityPlot.R	Uses .bed mapped sequence reads as input and the external program Homer to calculate the average sequence read density relative to TSSs and graphs it.	Figure 5, 6
heatmapGenerator.R	Uses .bed mapped sequence reads as input. Homer calculates the read coverage relative every TSS and R is used to display a heatmap of this data.	Figure 4
distanceToTSS.R	Takes in MACS peak file as input. Calculates how many peaks are within several distance intervals to their closest TSS.	None
peakExamples.R	Creates graphs of the top scoring peaks from a MACS peak file as well as graphs of regions without peaks as a comparison.	Figure 2, 3
gagePathway.R	Uses the KEGG database to determine which biological pathways are likely influenced by the studied transcription factor. Uses the output from annotatePeaks as input.	Figures 13, 14
GOenrichment.R	Creates a table of Gene Ontology Statistics.	Table 5
motifDiscovery.R	Uses DREME to perform de novo motif discovery on the input data set. Takes a list of peak sequences in .fasta format	Figures 7 A-D
makeUCSCfile.R	Creates a compressed .bedgraph file from a mapped sequence read .bam file which can uploaded directly to the UCSC genome browser	Figure 1, S1
extendReads.R	Takes in a .bed mapped reads file and creates a new file which extends the reads to a specified size in a direction appropriate for the strand.	None

IndexBAM.R	Sorts and indexes a .bam file which allows it to be used in peakExamples and makeUCSCfile.	None
getPeakSequences.R	Takes a MACS .bed peak file as input. Uses a specified reference genome to extract the genomic sequences for each peak. Creates a .fasta format file.	None
SAMtoBED.R	Converts a .sam mapped reads file and creates a .bam and .bed file.	None
BAMtoBEDTK.R	Converts a .bam file to a .bed file.	None
Scripts with matching YesWorkflow Diagrams		
GEAnalysisFromCEL.R	Performs a normalized Gene Expression analysis starting from .CEL files.	Tables 4 – 7, Figures 13, 14.
GEAnalysisFromNormalized.R	Performs a gene expression analysis from a tab-delimited input file of normalized expression data.	Table 6, 7, Figures 13, 14
CombineGEoneChIP.R	Combines gene expression data from a control of a cell type with ChIP-seq data for the same cell-type to determine the effect of the ChIP-target on gene expression at its binding locations.	Figure 8, 9, 10, 11
GetDiffPeaks.R	Prior to running CombineGEtwoChIP.R this must be run to calculate the differential peaks between the two ChIP-samples.	Table 9
CombineGEtwoChIP.R	Combines two sets of gene expression data (control, exp) with two sets of ChIP-set data (control, exp) for a target to determine the difference in target binding and the difference in gene expression at the targets binding locations.	Figure 12

Table 2. Peak Statistics Output from Analysis of ChIP-seq of H3K27ac data in MCF-7 cells.

Total number of peaks:	23623
Min peak length (bp):	388
Max peak length (bp):	8170
Median peak length (bp):	1843
Mean peak length (bp):	2179.238
Peak length standard deviation:	1321.272
% of genome with peaks:	1.641
number of peaks longer than 6 kb:	487
% of peaks larger than 6 kb:	2.062
Number of peaks per chromosome:	
chr1	2293
chr2	1937
chr3	1459
chr4	907
chr5	1236
chr6	1314
chr7	1215
chr8	997
chr9	1055
chr10	1127
chr11	1018
chr12	1324
chr13	448
chr14	866
chr15	812
chr16	977
chr17	1167
chr18	283
chr19	1142
chr20	676
chr21	327
chr22	447
chrX	576
chrY	19
chrM	1

Table 3. Cropped Output of Mapping Peaks to Genes from R scripts for ChIP-seq analysis of H3K27ac in MCF-7 cells.

chromosome	peak start	peak end	peak name	score	TSS	TSS_start	TSS_end
chr1	893796	894668	MACS_peak_3	127.86	893918	891918	898918
chr1	893796	894668	MACS_peak_3	127.86	894679	892679	899679
chr1	893796	894668	MACS_peak_3	127.86	895967	890967	897967
chr1	893796	894668	MACS_peak_3	127.86	896829	891829	898829
chr1	893796	894668	MACS_peak_3	127.86	897461	892461	899461
chr1	900879	902649	MACS_peak_4	149.64	901877	896877	903877
chr1	900879	902649	MACS_peak_4	149.64	901877	896877	903877
chr1	900879	902649	MACS_peak_4	149.64	901877	896877	903877
chr1	934211	936997	MACS_peak_5	175.48	935552	933552	940552
chr1	948243	949860	MACS_peak_6	184.07	948847	943847	950847
chr1	956229	957930	MACS_peak_7	75.16	955503	950503	957503
chr1	1066954	1069363	MACS_peak_14	156.01	1072397	1067397	1074397

strand	entrezID	UCSC_ID	Gene_Symbol	Full Gene Name
-	26155	uc001aby.4	NOC2L	nucleolar complex associated 2 homolog (S. cerevisiae)
-	26155	uc001abz.4	NOC2L	nucleolar complex associated 2 homolog (S. cerevisiae)
+	339451	uc001aca.2	KLHL17	kelch-like family member 17
+	339451	uc001acb.1	KLHL17	kelch-like family member 17
+	339451	uc001acc.2	KLHL17	kelch-like family member 17
+	84069	uc001acd.3	PLEKHN1	pleckstrin homology domain containing, family N member 1
+	84069	uc001ace.3	PLEKHN1	pleckstrin homology domain containing, family N member 1
+	84069	uc001acf.3	PLEKHN1	pleckstrin homology domain containing, family N member 1
-	57801	uc001aci.2	HES4	hes family bHLH transcription factor 4
+	9636	uc001acj.4	ISG15	ISG15 ubiquitin-like modifier
+	375790	uc001ack.2	AGRN	agrin
+	254099	uc001acv.3	LINC01342	long intergenic non-protein coding RNA 1342

Table 4. Cropped Output List of Genes from R script from ChIP-seq analysis of H3K27me3 in MCF-7 cells.

NOC2L
KLHL17
PLEKHN1
HES4
ISG15
AGRN
LINC01342
MIR200B
SDF4
B3GALT6
UBE2J2
ACAP3
PUSL1
CPTP
DVL1
MXRA8

Table 5. Gene Ontology Analysis for ChIP-seq on H3K27ac in MCF-7 cells.

GOID	Term	Pvalue	OddsRatio	ExpCount	Count	Size
GO:0044446	intracellular organelle part	4.54E-058	1.907668441	1458.770113	1858	3254
GO:0005737	cytoplasm	1.26E-056	2.081917244	1060.04034	1402	2528
GO:0090304	nucleic acid metabolic process	1.09E-054	1.813774996	1821.395731	2230	3672
GO:0005634	nucleus	1.18E-054	1.802738167	1733.121393	2144	3779
GO:0005730	nucleolus	2.49E-045	3.348196806	329.1007945	503	661

Table 6. KEGG Pathway Analysis results for pathways up regulated in budesonide exposed cells. Expected result highlighted in blue.

KEGG Pathways Upregulated in Budesonide Exposed Cells		
Pathway Code	Pathway Name	p Values
hsa04960	Aldosterone-regulated sodium reabsorption	0.005150403
hsa04910	Insulin signaling pathway	0.014410475
hsa04920	Adipocytokine signaling pathway	0.018992523
hsa04070	Phosphatidylinositol signaling system	0.025607065
hsa04610	Complement and coagulation cascades	0.02639603
hsa04150	mTOR signaling pathway	0.030314186
hsa04630	Jak-STAT signaling pathway	0.0327606
hsa04510	Focal adhesion	0.060022625
hsa04660	T cell receptor signaling pathway	0.061256631
hsa00140	Steroid hormone biosynthesis	0.065122596
hsa04640	Hematopoietic cell lineage	0.065615908
hsa04722	Neurotrophin signaling pathway	0.065793386

Table 7. KEGG Pathway Analysis results for pathways down regulated in budesonide exposed cells.

KEGG Pathways Downregulated in Budesonide Exposed Cells		
Pathway Code	Pathway Name	p Values
hsa04740	Olfactory transduction	2.00E-05
hsa00010	Glycolysis / Gluconeogenesis	0.001693544
hsa00240	Pyrimidine metabolism	0.011232358
hsa03050	Proteasome	0.024994643
hsa00830	Retinol metabolism	0.029858978
hsa00480	Glutathione metabolism	0.030013278
hsa00190	Oxidative phosphorylation	0.047604523
hsa04623	Cytosolic DNA-sensing pathway	0.050920848
hsa03030	DNA replication	0.06673399
hsa04622	RIG-I-like receptor signaling pathway	0.070012672

Table 8. Compatibility-Tested Array Types by Probe Density. Note: Other Affymetrix Gene and Exon arrays will work, these are just the ones that have been proven to work.

Array Type Tested	**Probeset Density**
[HuEx-1_0-st] Affymetrix Human Exon 1.0 ST Array [GENCODE v10]	~4 probes per exon ~40 probes per gene
[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	Probes per gene ~11
[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array]	Probes per gene: ~26
[MoGene-1_0-st] Affymetrix Mouse Gene 1.0 ST Array	Probes per gene ~11
[MoEx-1_0-st] Affymetrix Mouse Exon 1.0 ST Array	~4 probes per exon ~40 probes per gene

Table 9. Differentially Expressed Peaks from R scripts GetDiffPeaks.R and annotatePeaks.R in a comparison of ChIP-seq data targeting the glucocorticoid receptor (GR) from AtT-20 cells exposed to dexamethasone for 2 hours compared to ChIP-seq data for unexposed controls. Highlighted in blue are peaks unique to cells treated with glucocorticoids. Highlighted in green are peaks that both have in common. Highlighted in red are peaks unique to the unexposed controls.

chr	start	end	description	#raw_read_1	#raw_read_2	M_value_rescaled	A_value_rescaled	-log10(p-value)
chr1	3351027	3351347	unique_peak1	16	0	4.076492986	2.038246493	5.41853992
chr1	3406280	3407243	unique_peak1	17	7	1.157987345	3.578993673	1.38887104
chr1	3504986	3505326	unique_peak1	14	0	3.898039991	1.949019996	4.81647993
chr1	5097178	5097657	unique_peak1	10	1	2.455832529	2.227916265	2.8342087
chr3	154387547	154387941	common_peak1	7	5	0.416830439	2.79337772	1.10484139
chr4	3016602	3018363	common_peak1	104	97	0.057737098	6.643578393	0.32831267
chr4	154282699	154283075	common_peak1	134	11	3.445799288	5.307862145	26.2402037
chr5	114623028	114623250	common_peak1	20	47	-1.207192799	4.981366101	2.51566815
chr1	188839967	188840078	unique_peak2	0	4	-2.284926242	1.179464974	1.10720997
chr1	189127059	189127112	unique_peak2	0	4	-2.284926242	1.179464974	1.10720997
chr10	7351216	7351356	unique_peak2	0	5	-2.547960648	1.310982177	1.32905872
chr10	16122354	16122456	unique_peak2	0	5	-2.547960648	1.310982177	1.32905872

Figures:

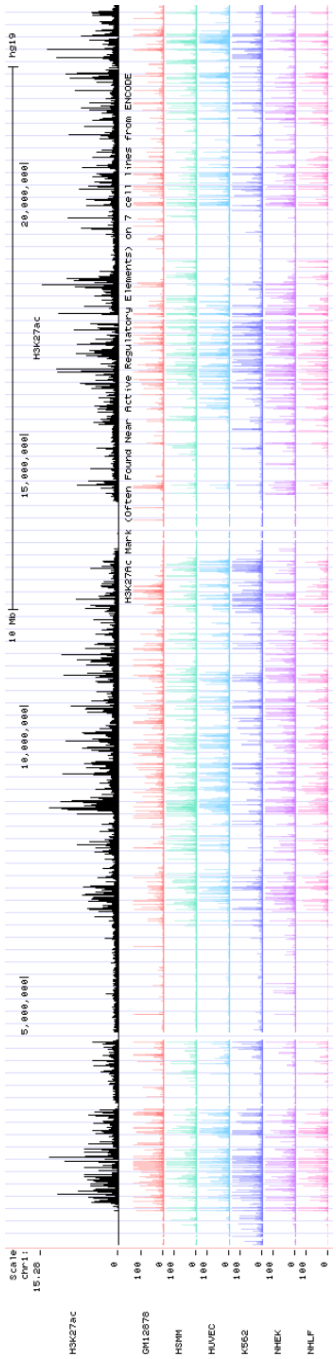


Figure 1.

The UCSC Genome Browser showing our custom-generated H3K27ac track for MCF-7 cells (shown in black) in comparison to H3K27ac peaks in other human cell types from the ENCODE project (25).

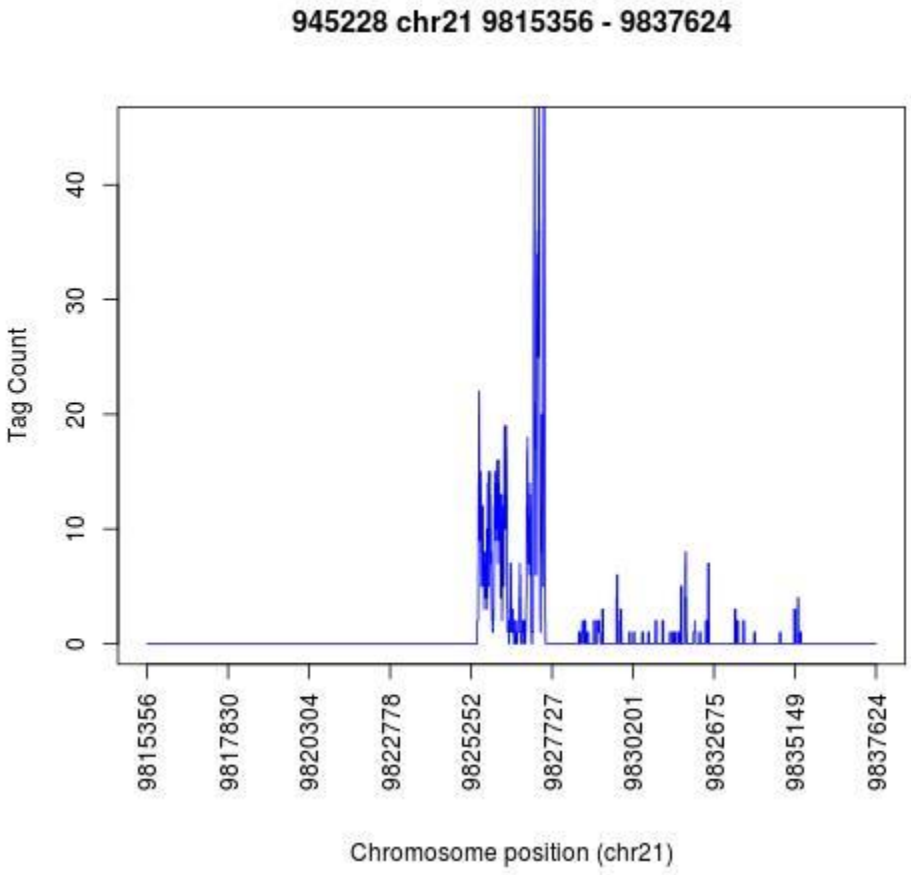


Figure 2.

A coverage graph from a MCF-7 cell chromosomal region with a high-scoring H3K27ac peak.

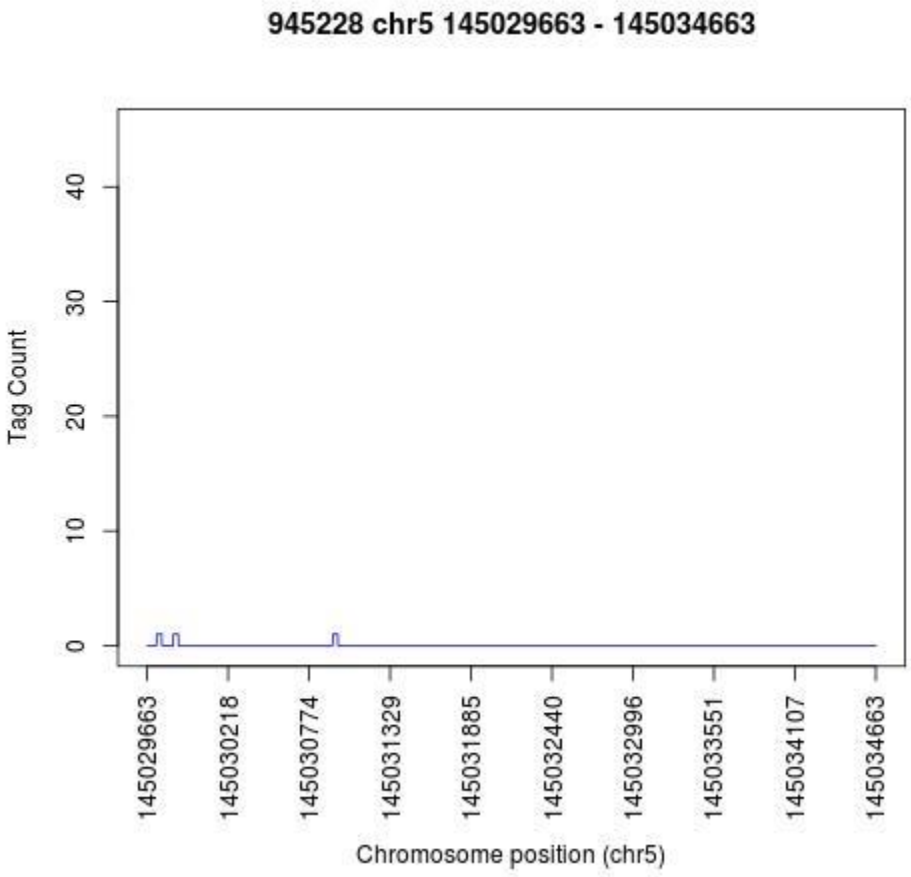


Figure 3.

A coverage graph of a MCF-7 cell chromosomal region with no H3K27ac peak.

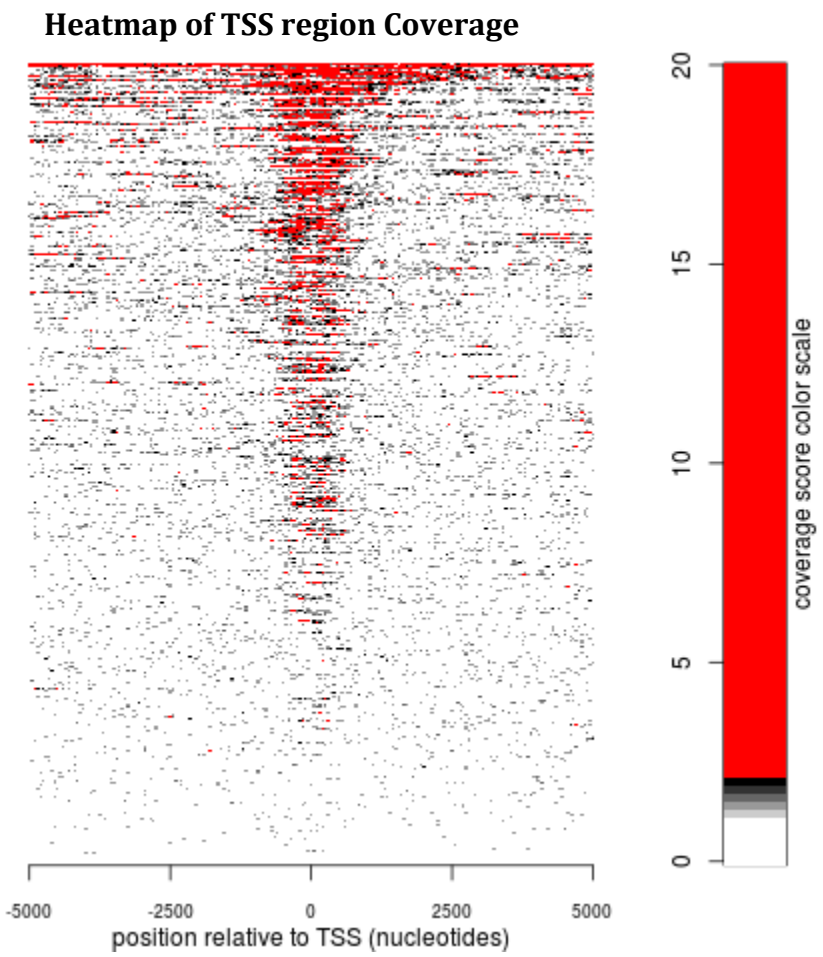


Figure 4.

A heatmap of transcription start site (TSS) region coverage for ChIP-seq of MCF-7 cells with a H3K27ac target. This diagram shows the localization of H3K27ac relative to TSSs within the genome.

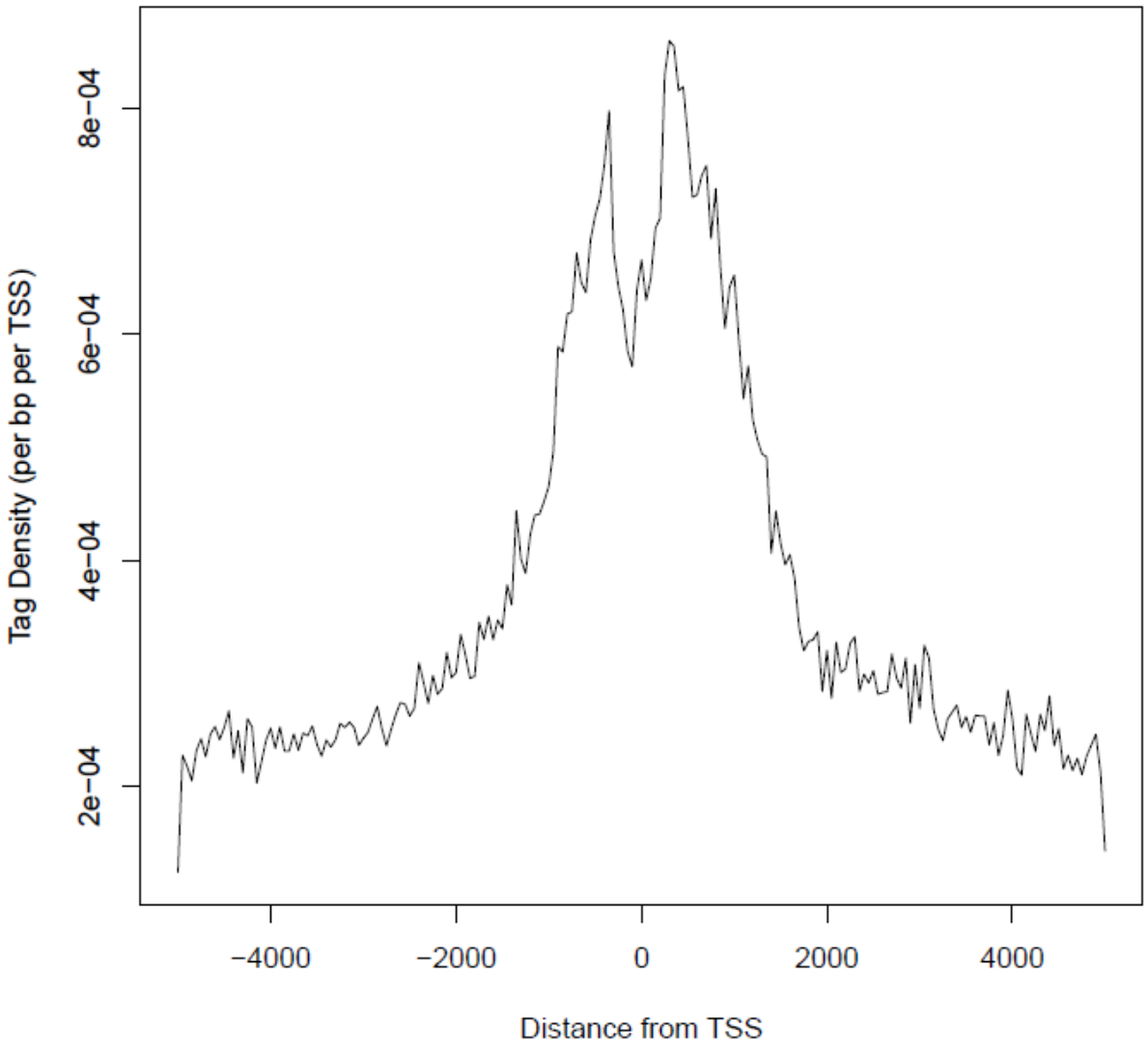


Figure 5.

A TSS density plot from ChIP-seq of MCF-7 cells with an H3K27ac target, showing the average read coverage (tag density per base pair per TSS) around the TSS.

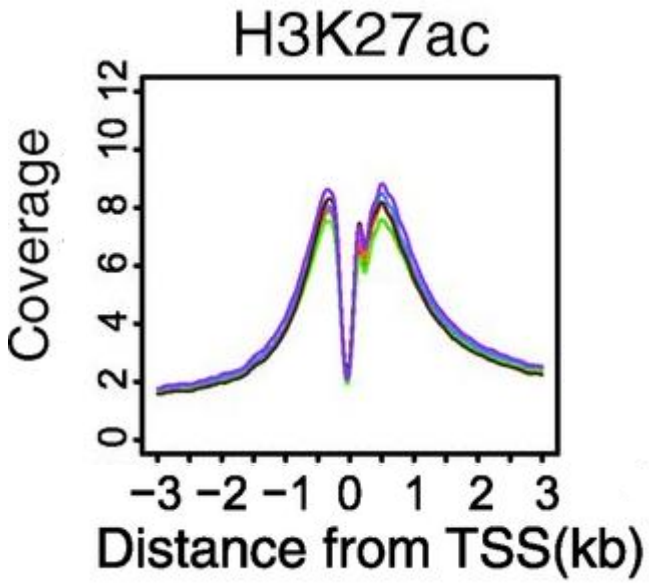
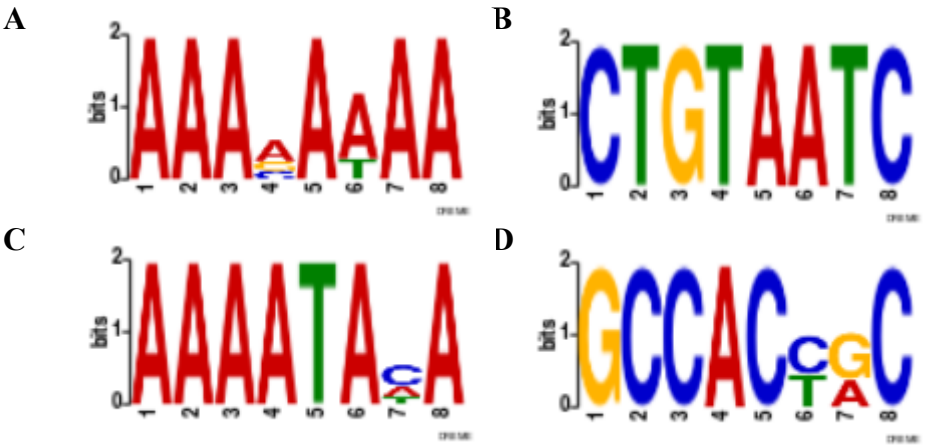


Figure 6.

A TSS density plot from Koike et al. (2012) showing H3K27ac binding in mouse liver cells during 6 circadian rhythm phases.



Figures 7 A-D.

De novo DNA sequence binding motifs found for H3K27ac presence in MCF-7 cells, in alphabetical order from most common to least common.

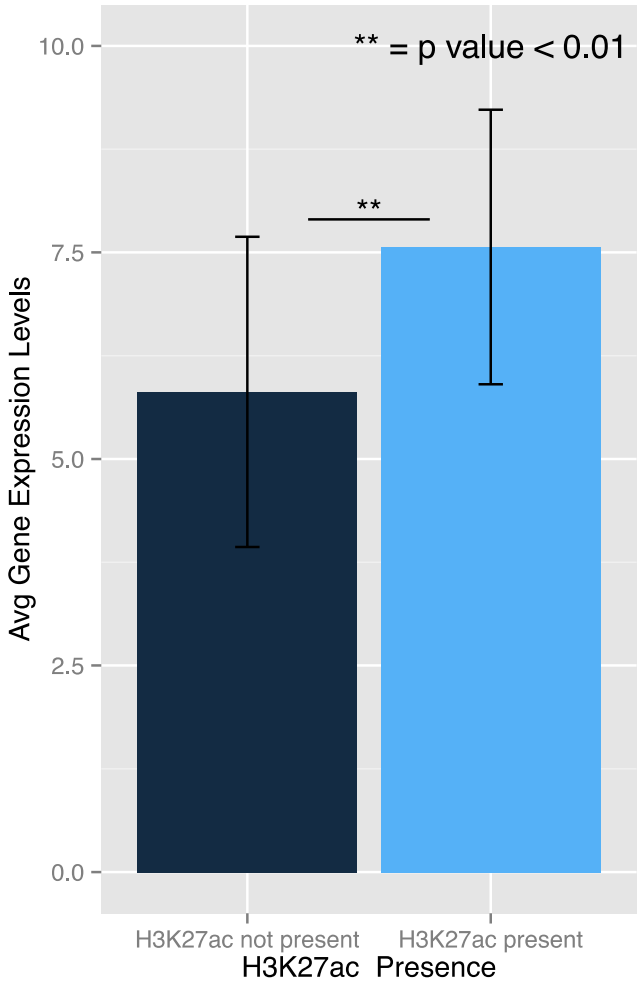


Figure 8.

Combined ChIP-seq and gene expression data result for H3K27ac target in MCF-7 cells. In dark blue gene expression of genes without H3K27ac presence is shown. In light blue gene expression of genes with H3K27ac presence is shown. Actual P value: $2.2x10^{-16}$.

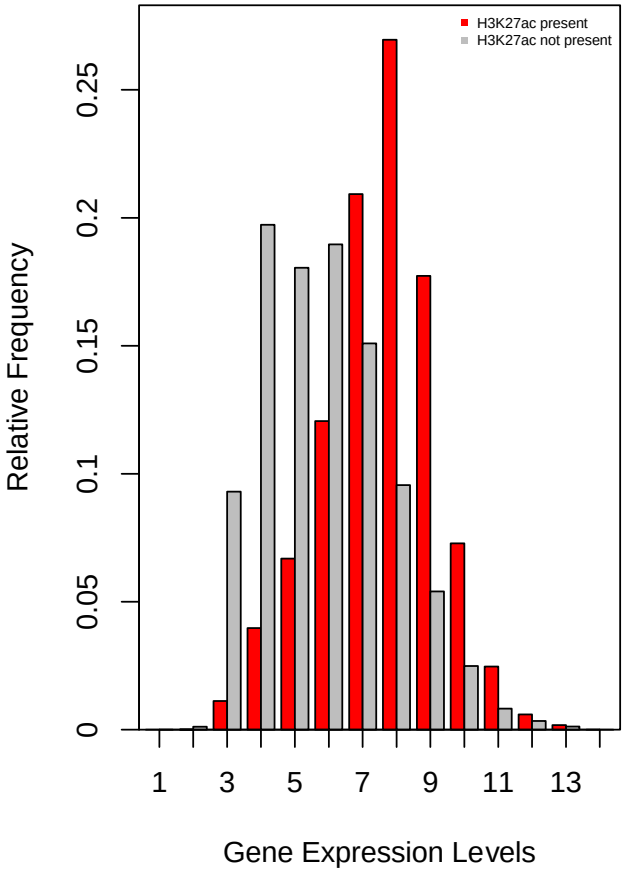


Figure 9.

Combined ChIP-seq and gene expression data result for H3K27ac target in MCF-7 cells. Shown in grey is the relative frequency (normalized number of genes) where H3K27ac was not present over a range of gene expression levels. Shown in red is the relative frequency (normalized number of genes) where H3K27ac was present over a range of gene expression levels.

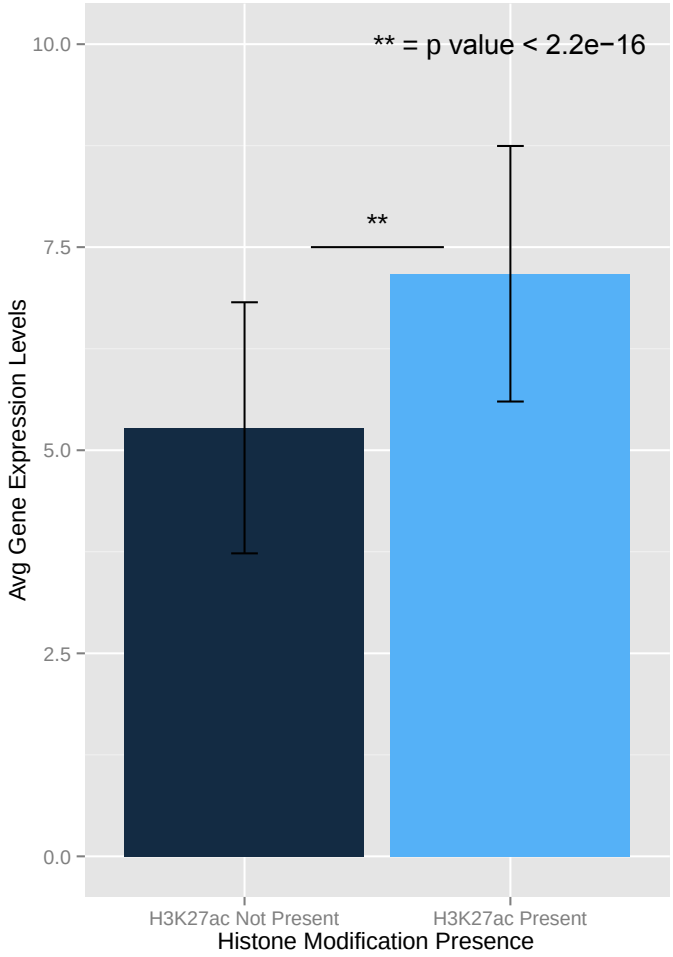


Figure 10.

Combined ChIP-seq and gene expression data result for H3K27ac target in K562 cells. In dark blue gene expression of genes without H3K27ac presence is shown. In light blue gene expression of genes with H3K27ac presence is shown. Actual P value: $2.2x10^{-17}$.

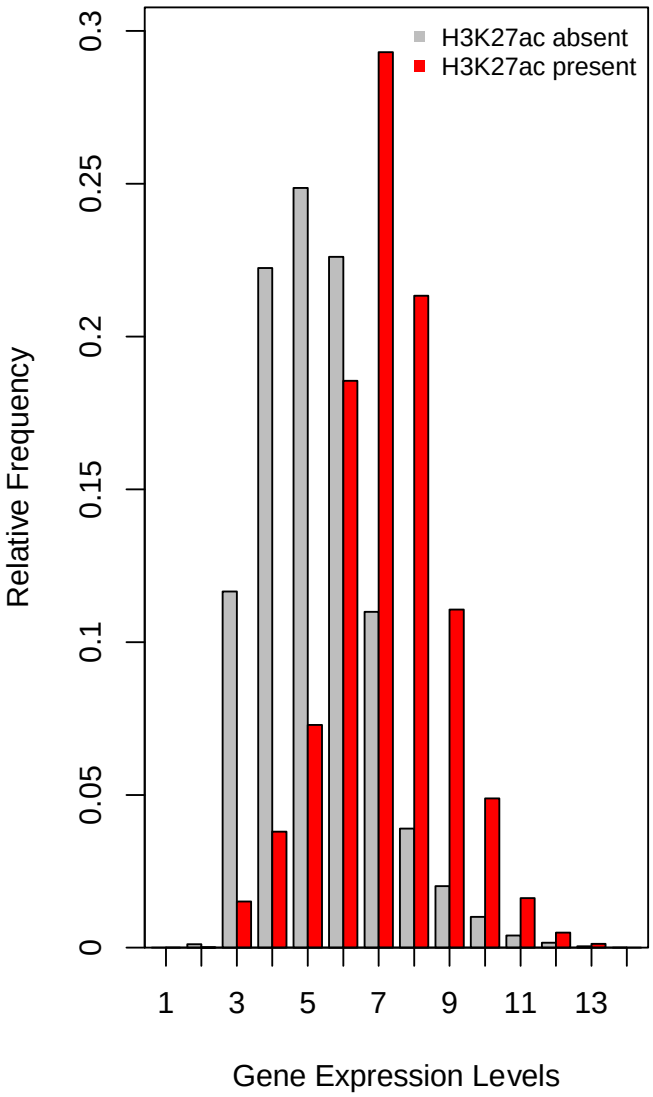


Figure 11.

Combined ChIP-seq and gene expression data result for H3K27ac target in K562 cells. Shown in grey is the relative frequency (normalized number of genes) where H3K27ac was not present over a range of gene expression levels. Shown in red is the relative frequency (normalized number of genes) where H3K27ac was present over a range of gene expression levels.

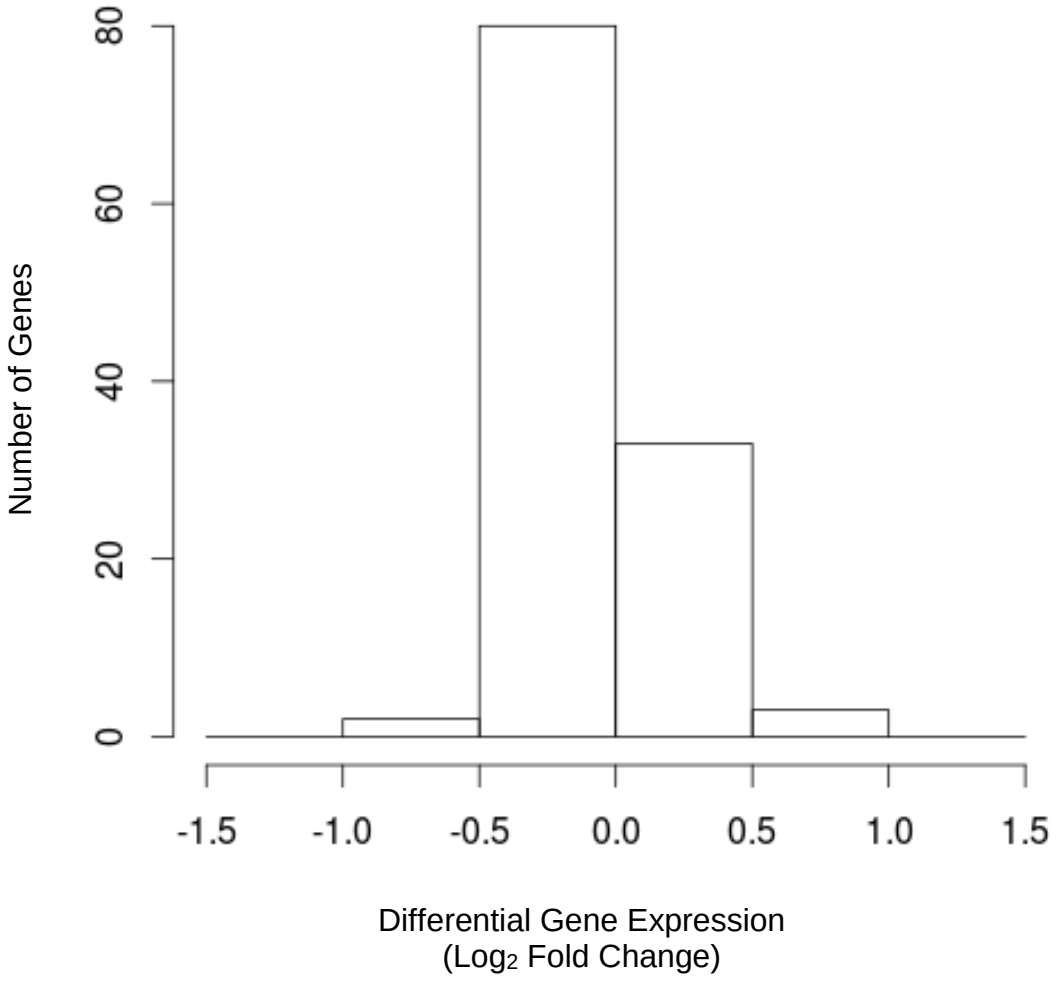


Figure 12.

A histogram showing the difference in gene expression in 118 genes in which dexamethasone exposure was shown to cause GR binding in dexamethasone exposed AtT-20 cells vs. unexposed cells. Dexamethasone exposure was 100 nM for 2 hours.

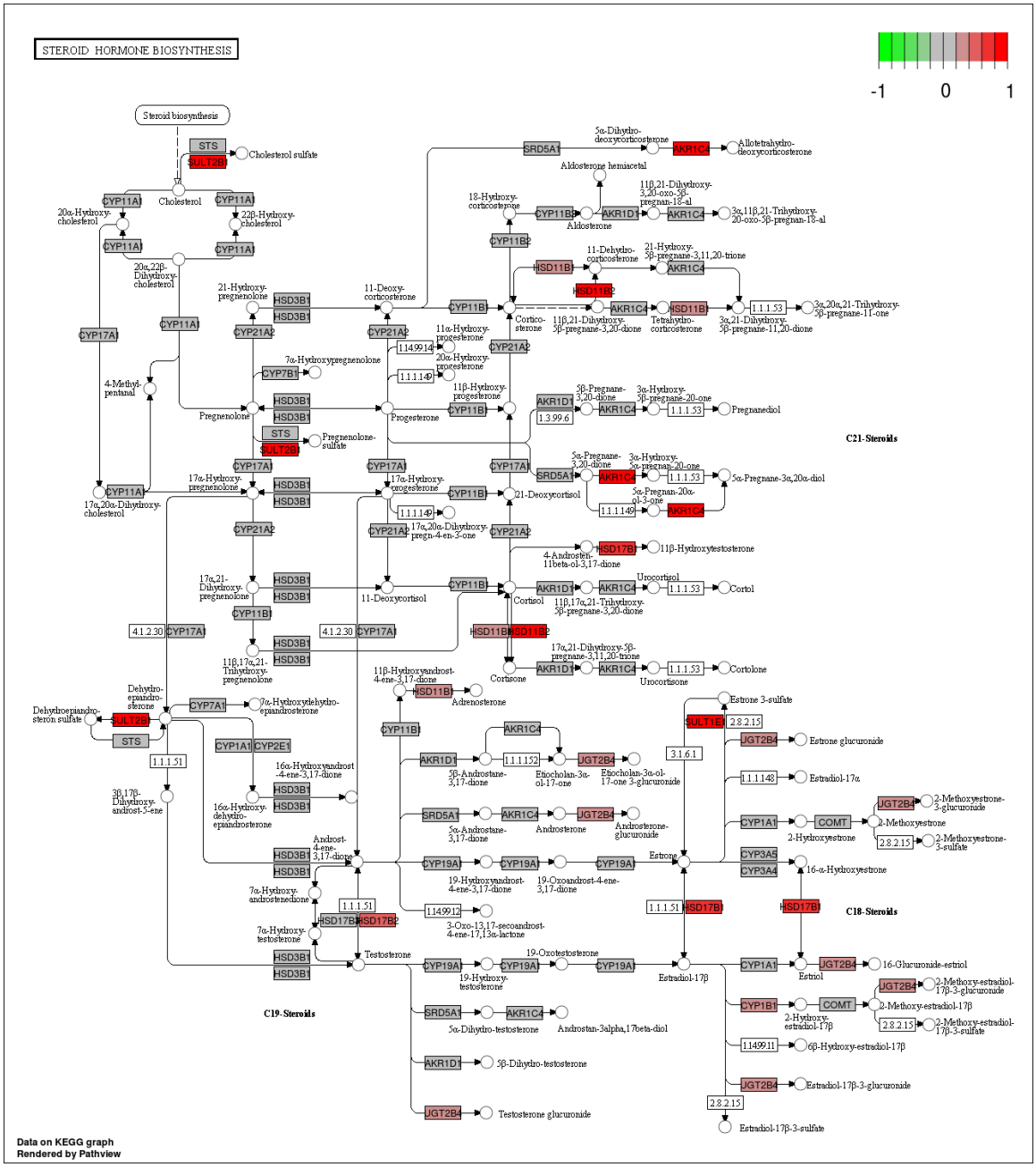


Figure 13.

A KEGG/Pathview diagram which demonstrates the genes up-regulated in the Steroid Hormone Biosynthesis pathway in human lung tissue samples exposed to budesonide compared to unexposed control samples. The darker shades of red indicate higher levels of up regulation.

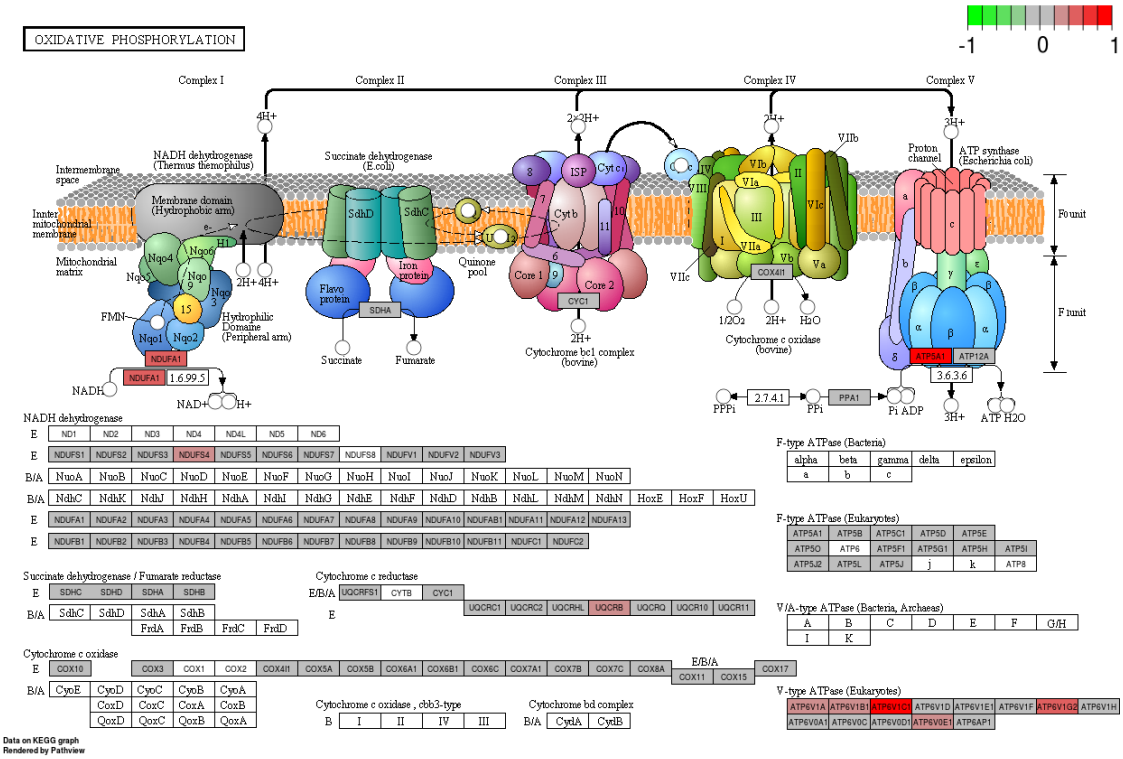


Figure 14.

A KEGG/Pathview diagram which demonstrates the genes down-regulated in the Oxidative phosphorylation pathway in human lung tissue samples exposed to budesonide compared to unexposed control samples. The darker shades of red indicate higher levels of down regulation.

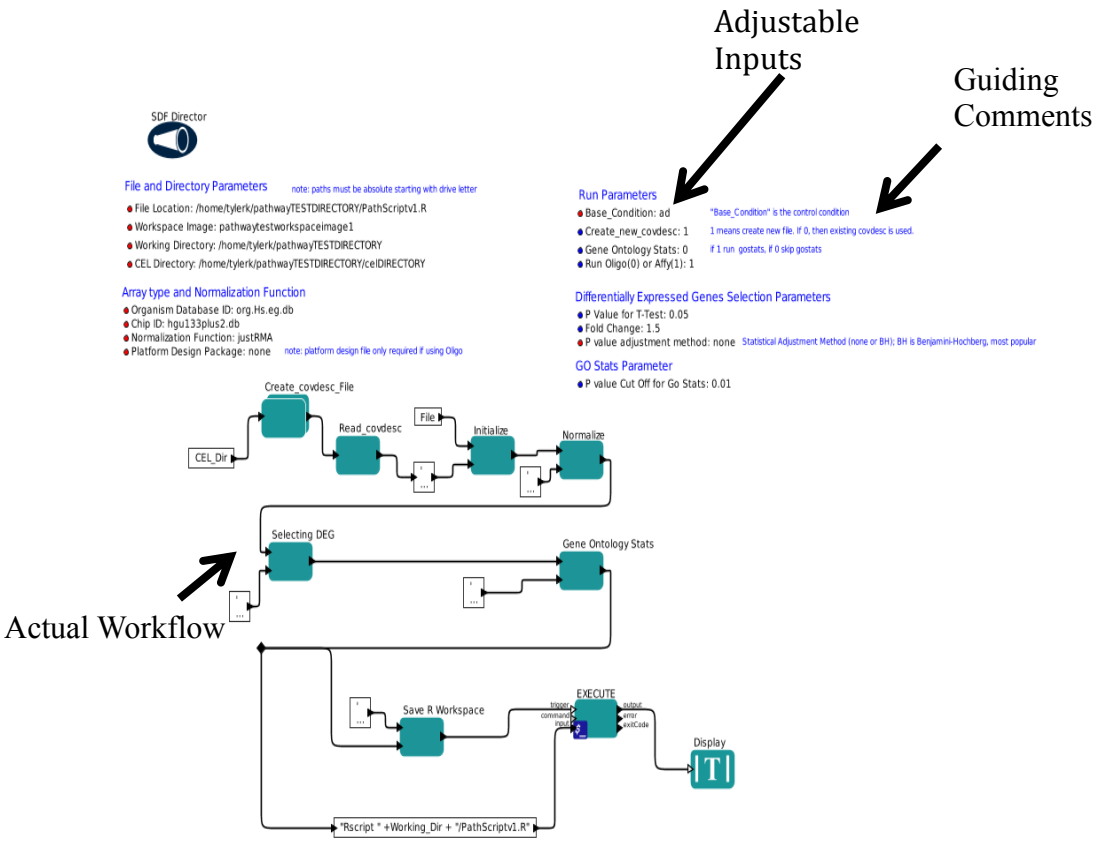


Figure 15.

A sample bioKepler workflow for analyzing gene expression microarray data. This workflow illustrates the flow of data processing much like a YesWorkflow Diagram, while actually processing code as well.

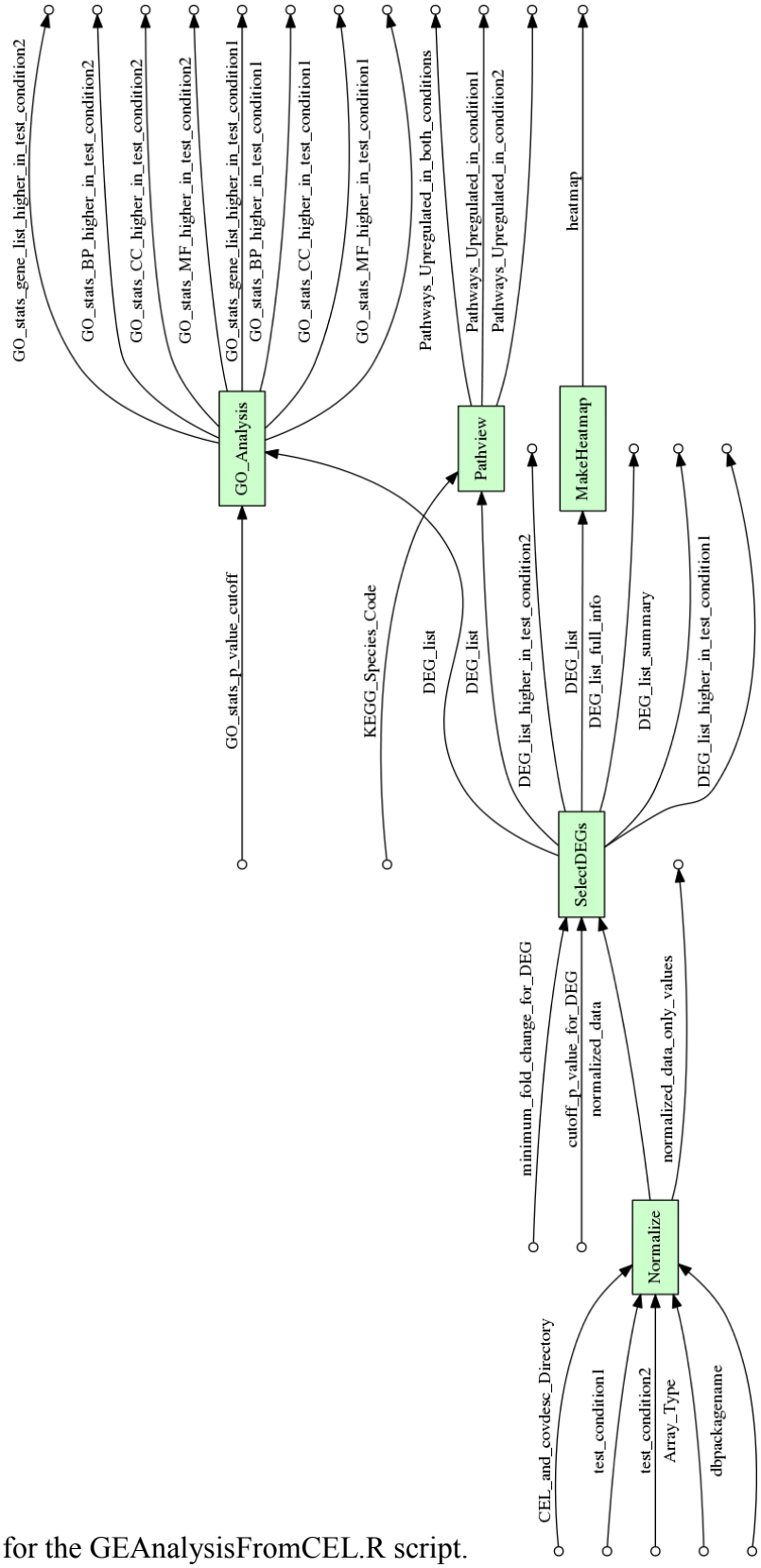


Figure 16.

A YesWorkflow Diagram for the GEAnalysisFromCEL.R script.

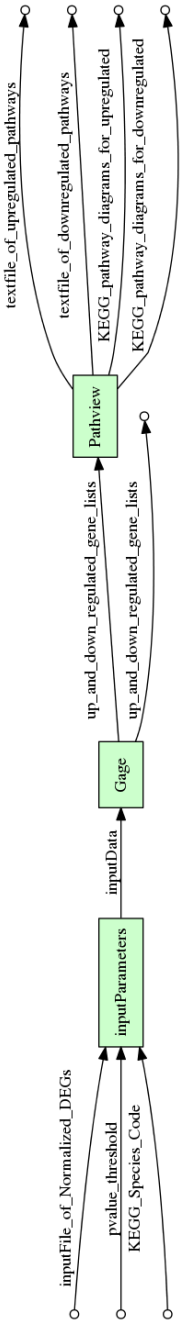


Figure 17.

A YesWorkflow Diagram for the GEAnalysisFromNormalized.R script.

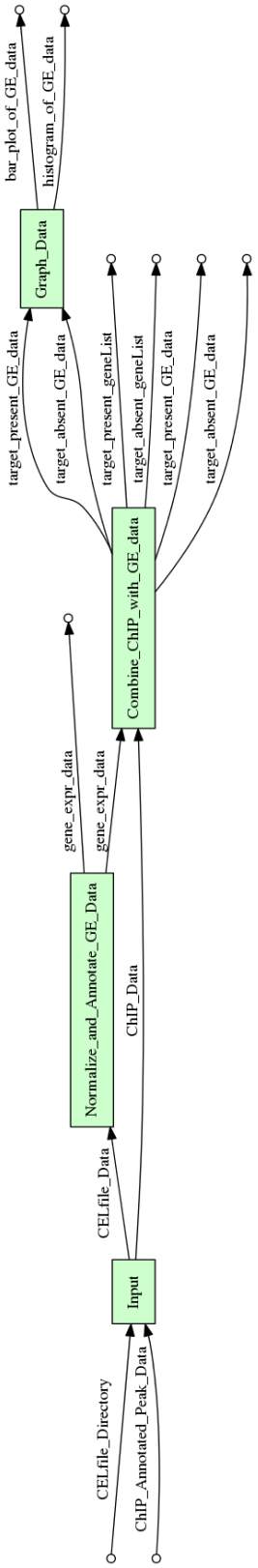


Figure 18.

A YesWorkflow Diagram for the CombineGEoneChIP.R script.

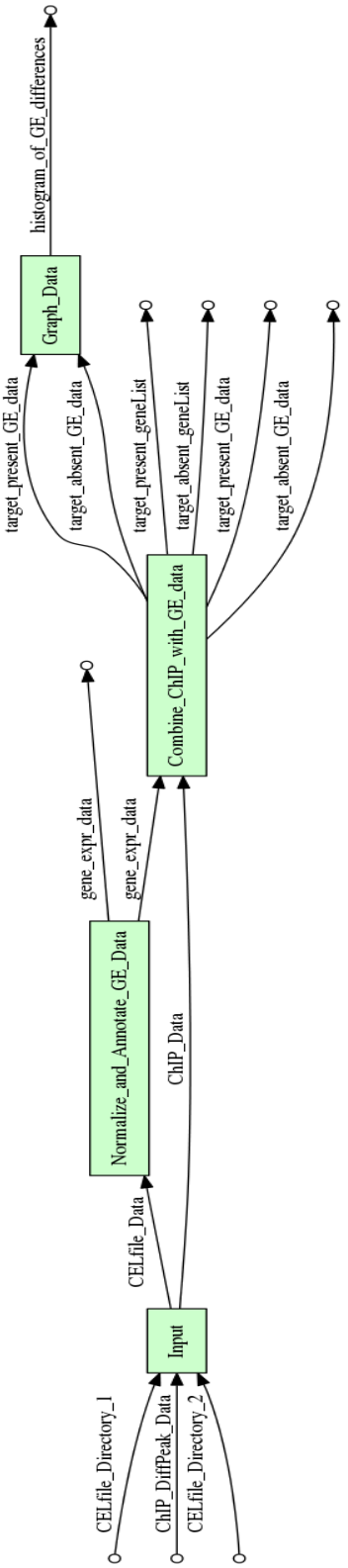


Figure 19.

A YesWorkflow Diagram for the CombineGEtwoChIP.R script.

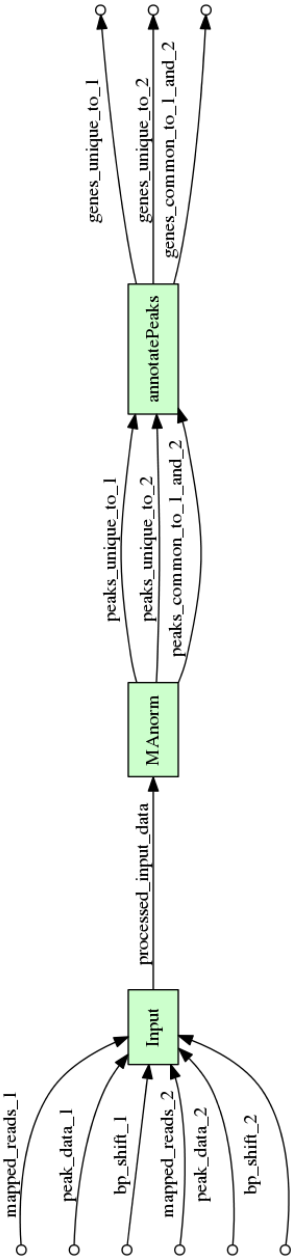


Figure 20.

A YesWorkflow Diagram for the DiffPeaks.R script.

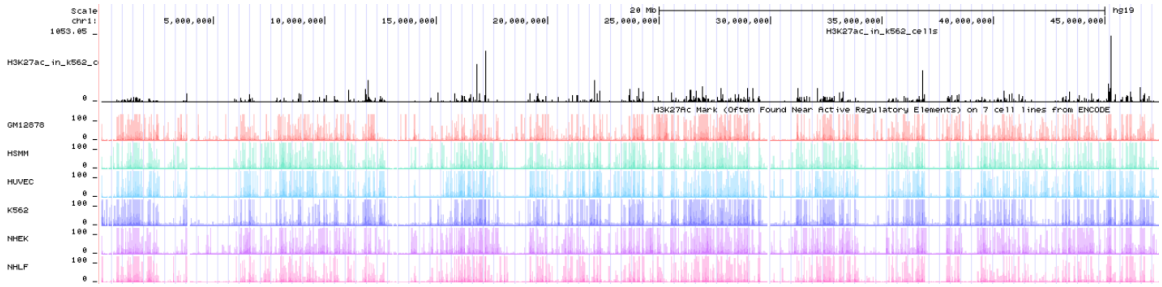
Figure 21. A screen clipping from the International Data Curation Conference 2015 in London, where YesWorkflow was first presented to the world by B. Ludäscher. This figures shows a portion of the R script GEAnalysisFromNormalized.R, and the YesWorkflow comments embedded within the script in order to enhance user-friendliness and produce the YesWorkflow Diagram.

Reference List:

1. Altintas I, Wang J, Crawl D, Li W. Challenges and approaches for distributed workflow-driven analysis of large-scale biological data, in: Proceedings of the Workshop on Data analytics in the Cloud at EDBT/ICDT 2012 Conference, 2012, pp 73-78.
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May; 25(1):25–9.
3. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, et al. MEME SUITE: tools for motif discovery and searching, Nucleic Acids Research, 2009, 37:W202-W208.
4. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007 May 18; 129(4):823–37.
5. Beier D, Hau O, Proescholdt M, Lohmeier A, et al. CD133 (+) and CD133 (-) glioblastoma-derived cancer stem cells show differential growth characteristics and molecular profiles. Cancer Research. 2007 May 1; 67(9):4010-5.
6. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6; 489(7414):57–74.
7. Cormier N, Kolisnik T, Bieda M. Reusable, extensible, and modifiable R scripts and Kepler workflows for comprehensive ChIP-seq analysis (under submission).
8. Falcon S and Gentleman R (2007). Using GOstats to test gene lists for GO term association. Bioinformatics, 23(2), pp. 257-8.
9. Gautier L, Cope L, Bolstad BM and Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004 20(3), pp. 307–315.
10. Glozak MA, Seto E. Histone deacetylases and cancer. Oncogene 2007 Aug 13. 26(37):5420-32.
11. Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010 Aug 25; 11(8):R86.
12. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 2010, 38:576-589.
13. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 4 (2): 249–64.
14. John S, Sabo PJ, Thurman RE, Sung M., Biddie, SC, Johnson TA, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet. 2011 Mar; 43(3):264–8.
15. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000 28, 27-30.
16. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, and Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2000. 42, D199–D205.
17. Koike N, Yoo S-H, Huang H-C, Kumar V, Lee C, Kim T-K, et al. Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals. Science. 2012 Oct 19; 338(6105):349–54.
18. Kolisnik T, Bieda M. A universal microarray data analysis platform designed using bioKepler and the R language for statistical computing. Poster. 2014 Presented at: University of

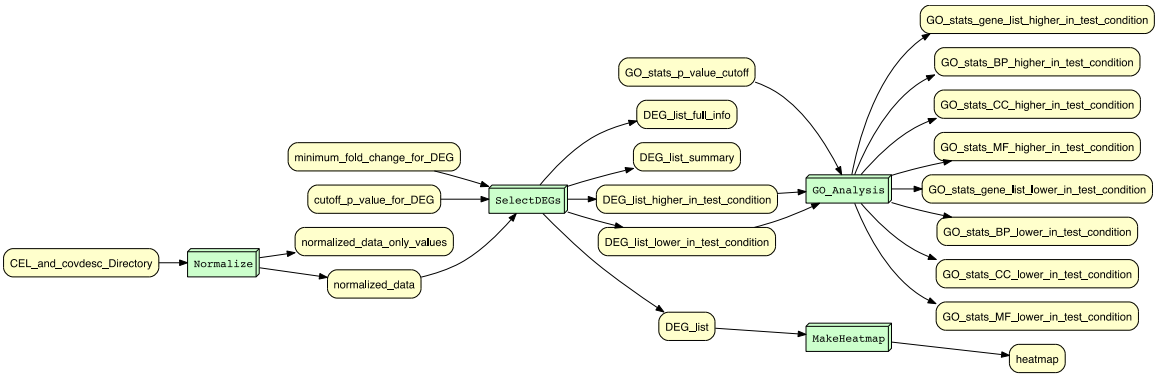
Calgary Undergraduate Research Symposium.
19. Laajala TD, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, Elo LL. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. BMC Genomics. 2009 Dec 18; 10(1):618.
20. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011 Jan; 39(Database issue):D19–21.
21. Luo, Weijun, Brouwer, and Cory. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013. 29(14), pp. 1830-1831.
22. McPhillips T, Song T, Kolisnik T, Aulenbach, S, Ludäscher B, et al. YesWorkflow: A User-Oriented Language-Independent Tool for Recovering Workflow Structure, Provenance, and Semantics from Scripts. International Journal of Digital Curation. 2015, February.
23. Miller CJ. simpleaffy: Very simple high level analysis of Affymetrix data. R package version 2015 2.42.0.
24. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2008.
25. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ. Track Data Hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics. 2014 Apr 1; 30(7):1003-5. Epub 2013 Nov 13.
26. Shao Z, Zhang Y, Yuan G-C, Orkin SH, Waxman DJ. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol. 2012; 13(3):R16.
27. Stropp T, McPhillips T, Ludäscher B, Bieda M. Workflows for microarray data processing in the Kepler environment. BMC Bioinformatics. 2012 May 17; 13(1):102.
28. Yang H, Lan P, Hou Z, Guan Y, Zhang J, Xu W, et al. Histone deacetylase inhibitor SAHA epigenetically regulates miR-17-92 cluster and MCM7 to upregulate MICA expression in hepatoma. Br J Cancer. 2015 Jan 6; 112(1):112–21.
29. Zentner, GE; Henikoff, S. Regulation of nucleosome dynamics by histone modifications. Nature structural & molecular biology. 2013 March 20 (3): 259–66.
30. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008 Sep 17; 9(9):R137.

Appendix A: Supplementary Figures



Supplementary Figure 1.

The UCSC Genome Browser showing our custom-generated H3K27ac track for K562 cells in comparison to H3K27ac peaks in other human cell types from the ENCODE project (25).



Supplementary Figure 2.

A figure showing an alternate style of a YesWorkflow Diagram, creatable through customization options.

Supplementary Figure 3.

Cropped Output of Mapping Peaks to Genes from R scripts for ChIP-seq analysis of dexamethasone exposed AtT-20 cells.

chromosome	peak start	peak end	peak name	score	TSS	TSS_region_start
chr12	31595363	31595696	MACS_peak_1	52	31596477	31594477
chr12	31595363	31595696	MACS_peak_1	52	31596477	31594477
chr2	113883057	113883538	MACS_peak_3	67.51	113878547	113876547
chr15	42510253	42510775	MACS_peak_1	102.44	42508341	42506341
chr15	42510253	42510775	MACS_peak_1	102.44	42508341	42506341
chr15	42510253	42510775	MACS_peak_1	102.44	42508341	42506341
chr8	127095572	127096235	MACS_peak_5	96.07	127093607	127091607

TSS_region_end	strand	UCSC_ID	entrezID	Gene_Symbol	Full Gene Name
31601477	-	uc007ngx.2	11431	Acp1	acid phosphatase 1, soluble
31601477	-	uc007ngw.2	11431	Acp1	acid phosphatase 1, soluble
113883547	-	uc008lpz.1	11464	Actc1	actin, alpha, cardiac muscle 1
42513341	-	uc007vpd.1	11600	Angpt1	angiopoietin 1
42513341	-	uc007vpe.1	11600	Angpt1	angiopoietin 1
42513341	-	uc007vpc.1	11600	Angpt1	angiopoietin 1
127098607	-	uc009nxe.1	11606	Agt	angiotensinogen A8

Supplementary Figure 4.

Cropped Outputted List of Genes from R script from ChIP-seq analysis of dexamethasone exposed AtT-20 cells.

Acp1
Actc1
Angpt1
Agt
Aldoa
Amy1
Anxa4
Aox1
Arf6
Arntl
Asgr1
Atp1a1
Gdf2
Btrc
C1qa
Calca
Runx1t1
Cd24a