Integration Analysis of RNA-seq and ATAC-seq

John David Lin

March 22, 2019

Introduction

RNA-seq analysis examines the gene expression levels in samples and identifies differentially expressed genes (DE genes). From the change of expression levels in the DE genes, we can infer their activities are potentially related to the condition. ATAC-seq analysis examines chromatin accessibility to infer transcription factor binding site and nucleosome binding region, and to identify differentially bound regions (DB regions). [1] RNA-seq and ATAC-seq both analyze gene-level activities but in different perspectives. I hypothesize that a gene with increasing gene expression level corresponds to increasing chromatin accessibility since greater openness in a region allows for higher transcription activity. Based on the hypothesis, I expect to observe a positive correlation between the results from both datasets or evidence that indicates relationships between those two. The goal of this project is to integrate the results of RNA-seq and ATAC-seq to unveil subtle trends or interactions. I use three integration methods to analyze the relationship between the two datasets in different perspectives. Correlation analysis identifies the correlation between the two datasets and is a baseline for comparing with two other methods. Clustering-based integration analysis is performed by iClusterPlus. [2] iClusterPlus is a clustering-based integration tool that "performs pattern discovery that integrates diverse data types: binary (somatic mutation), categorical (copy number gain, normal, loss), and continuous (gene expression) values". [3] The authors explain that it "uses generalized linear regression for the formulation of a joint model, with respect to a common set of latent variables that we propose represents distinct driving factors (molecular etiology and genetic pathways)". Multiple co-inertia analysis is a multivariate statistical method for multiple omics datasets, that "facilitate input and analysis of more than two omics datasets." [4]

Methods

RNA-seq and ATAC-seq

The RNA-seq workflow I use is the RNAseq123 package in Bioconductor, without any substantial modification. [5] The ATAC-seq workflow is more complex and consists of two parts, primary and secondary analysis. The objectives of primary analysis are alignment and peak calling. It uses the ENCODE ATAC-Seq pipeline developed by the lab of Anshul Kundaje at Stanford, which performs adapter detection, read trimming, alignment with Bowtie2, and filtering. The subsequence steps, peak calling with MACS2,

and blacklist filtering, were carried out using the CCBB ATAC-seq pipeline developed by Amanda. [6] The objective of the secondary analysis is differential binding and consists of 2 parallel pipelines, DiffBind [7] and CSAW [8], both of which are implemented in Jupyter notebooks by Amanda.

According to the post *ATAC-Seq Analysis at CCBB* by Amanda, while "DiffBind is a Bioconductor package that takes in peak set(s) and bams and determines differentially bound peaks based on their read counts using either DESeq2 or edgeR", "CSAW is a Bioconductor package that performs differential binding analysis WITHOUT using peak calls (instead, it counts reads over sliding windows and aggregates those windows that are above background into differentially bound regions)." [6] In other words, DiffBind and CSAW perform differential binding analysis using different methods, peak based, and region based, respectively.

Correlation Analysis

The first integration method is correlation analysis, which takes in differentially bound regions (DB regions) from ATAC-seq (DiffBind and CSAW) and differentially expressed genes (DE genes) from RNA-seq. We compare two datasets at a time, plotting the log fold change (logFC) of one on the x-axis and logFC of the other on the y-axis to create a scatterplot. The plot is fitted with regression lines and lowess lines to identify the trend or relationship. I make three separate comparisons: 1. DiffBind DB regions against DE genes. 2. CSAW DB regions against DE genes. 3. DiffBind DB regions against DB regions from CSAW. The outputs are three scatterplots, one from each comparison. The scatterplots can be subset for closer examination and pathway analysis. The plots are primarily served as a baseline of comparisons with other integration methods.

Integrative Clustering Analysis

The second integration method is integrative clustering analysis based on the Bioconductor package iClusterPlus. iClusterPlus takes in datasets from different platforms. The input I use from RNA-seq is gene expression data, e.g. counts or logCPM. The input I used from ATAC-seq is peak counts data, e.g. binding scores, counts, or logCPM.

For the ATAC-seq data ready for DiffBind, I perform the following processing steps. First, I extract the binding scores and counts matrix produced by dba.count() function in DiffBind. Second, I annotate the matrix with Annotatr [9] to include gene symbols. Third, I merge the samples and separate counts and binding scores as different matrices. Finally, I produce the logCPM matrix using the counts matrix. The binding scores are calculated from peak counts by DiffBind.

With the inputs ready, I then perform the analysis using iClusterPlus workflow described in the vignette. The first step is model tuning which allows iClusterPlus to select lambda values for the best model. The next step is model selection that "select the best sparse model with the optimal combination of penalty parameters." [2] Then, it generates a series of heatmaps for results visualization. Finally, the feature selection step allows us to extract the top features (genes) of the datasets for examination and downstream analysis such as enrichment analysis or pathway analysis.

I make four separate analysis comparing different datasets: 1. RNA-seq read **counts** against ATAC-seq peak **binding scores**. 2. RNA-seq read **counts** against ATAC-seq peak **counts**. 3. RNA-seq **logCPM** against ATAC-seq peak **binding scores**. 4. RNA-seq **logCPM** against ATAC-seq **logCPM**. Although I create four sets of comparisons, only number 2 and 4 go through the entire iClusterPlus analysis and pathway analysis. While it will be interesting to compare counts with binding scores or logCPM with binding scores, the comparisons with the same metrics make more sense. The results section only include the results from comparisons 2 and 4. I select Gaussian as the distribution type for the parameter of clustering for logCPM comparison, and Poisson for counts comparison.

Multiple Co-Inertia Analysis

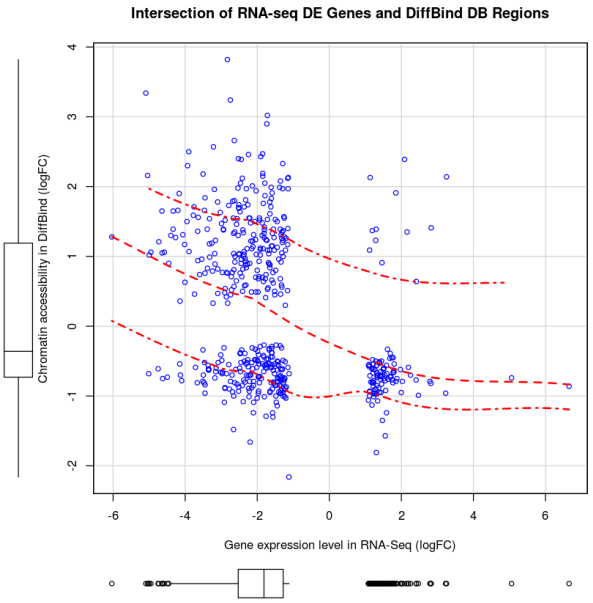
The third integration method is multiple co-inertia analysis based on the Bioconductor package omicade4. [4] It also takes in datasets from different platforms. The inputs for co-inertia analysis are identical to the inputs for iClusterPlus, undergone the processing steps described in the section above. I perform multiple co-inertia analysis as described in the vignette to make four separate comparisons. The comparison groups are identical to the ones in iClusterPlus and described in the section above. The outputs are plots that show the sample space, variable (gene) space, absolute eigenvalues, and pseudo-eigenvalues space, of both datasets integrated.

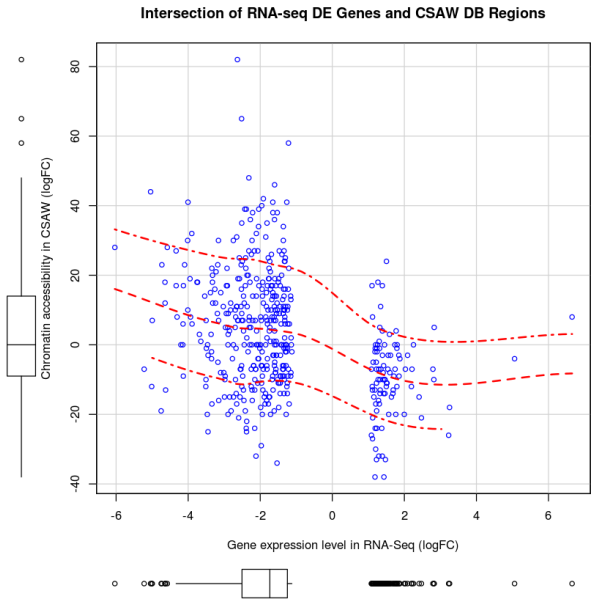
Pathway Analysis

The feature selection step in iClusterPlus allows us to select the top significant genes for pathway analysis. These gene are annotated by WebGestaltR [10] with the union of the ATAC-seq and RNA-seq input datasets as reference genes. The databases to query include geneontology_Biological_Process, geneontology_Molecular_Function, geneontology_Biological_Process_noRedundant, pathway_KEGG, pathway_Panther, pathway_Reactome, pathway_Wikipathway, network_Kinase_target, network_miRNA_target, network_PPI_BIOGRID, network_Transcription_Factor_target, disease_Disgenet, disease_OMIM, drug_DrugBank, and phenotype_Human_Phenotype_Ontology.

Results

Correlation Analysis





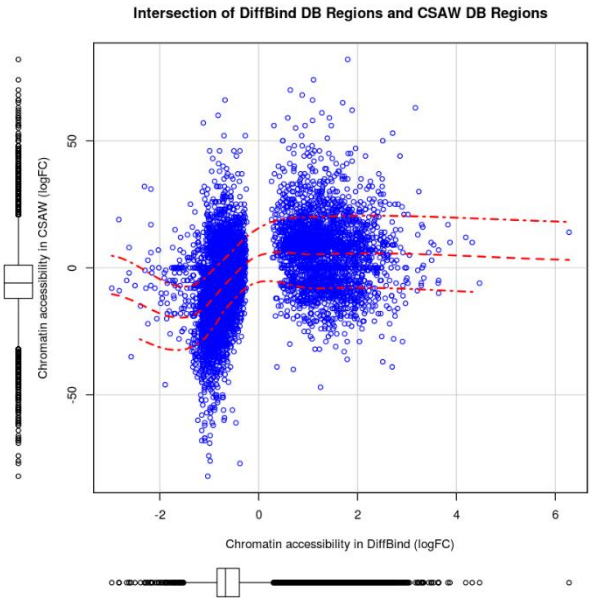


Figure 1: Scatterplots generated from correlation analysis. There are three red curves in each plot. The top and bottom red curves are the variance smoothers. The middle red curve is the mean smoother. The top left plot intersects RNA-seq and DiffBind. The top right plot intersects RNA-seq and CSAW. The bottom left plot intersects DiffBind and CSAW.

Hypothetically, increasing chromatin accessibility correlates to higher transcription activity, therefore higher increasing gene expression, and vice versa. We expect to see the majority of data points in the first (+, +) and third quadrant (-, -). However, in the top left plot, there are very few data points in the first quadrant although some in the third quadrant. The majority of data points fall into the second, third and fourth quadrant. The red regression lines are negative-slope curves that indicate a slightly negative correlation between gene expression and its accessibility. A similar distribution of data points is also observed in the top right scatterplot, which agrees with the top left scatterplot. The top two plots, which compare DiffBind against RNA-seq and CSAW against RNA-seq, agree on a negative correlation between gene expression and chromatin accessibility. The bottom left plot is a sanity check on the data. It compares genes' accessibility detected by DiffBind and CSAW and shows a slightly positive correlation. This relationship is expected because DiffBind and CSAW both detect differential binding and should report similar results.

Integrative Clustering Analysis

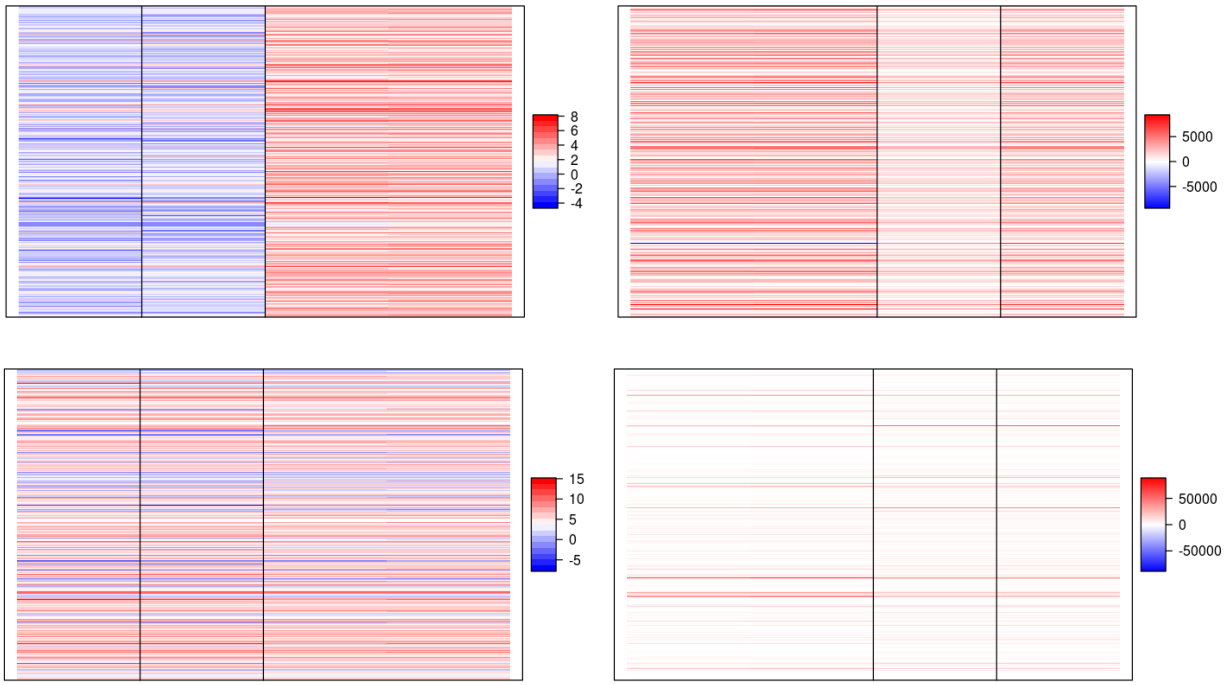


Figure 2. Heatmaps output by iClusterPlus. The left heatmaps are produced with the integration of RNA-seq **logCPM** and ATAC-seq **logCPM** inputs. The right heatmaps are produced with the integration of RNA-seq read **counts** and ATAC-seq peak **counts** inputs. The top heatmaps are the results of clustering of **RNA-seq genes** with their respective metrices. The bottom heatmaps are the clustering results of ATAC-seq **peaks** with their respective metrices.

The heatmaps show the clustering results of gene expression data and peak counts data, both in logCPM or counts. The top left plot is produced with RNA-seq logCPM and shows significant difference in enrichment between Ctrl and Ox samples. The bottom left is produced with ATAC-seq logCPM and shows a variety of positively and negatively enriched genes. The top right plot is produced with RNA-seq read counts and shows uniformly positively enrichment across both Ctrl and Ox samples. The bottom right is produced by ATAC-seq peak counts and shows little enrichment across both sample. The black lines in the heatmaps group the samples into clusters. With multiple samples, iClusterPlus identifies which samples are closely related. However, because I only use two samples, the clustering results are trivial.

- "Abcc8"
- "Angpt4"
- "Ccl9"
- "Csf1"
- "Csf3"
- "Dmkn"
- "Draxin"
- "Foxj1"
- "Gm14124"
- "Itga2"
- "Lcn2"
- "Lgals3bp"
- "Ndrg4"
- "Neb"
- "Npy"
- "Nxf3"
- "Pax8"
- "Pfkl"
- "Plcb1"
- "Prdm11"
- "Tns3"
- "Wnt3"
- "Wt1"
- "Zfp14"

List 1. The intersection of genes between RNA-seq and ATAC-seq identified by iClusterPlus after integrative clustering analysis. These genes are produced with RNA-seq **logCPM** and ATAC-seq **logCPM** as inputs.

There is a total of 16688 unique peak-associated genomic regions that exist in all four samples, identified by DiffBind from ATAC-seq, and 11163 genes with read coverage, identified by RNA-seq analysis. These genes are inputs for iClusterPlus. After the integrative analysis of RNA-seq **logCPM** against ATAC-seq **logCPM**, iClusterPlus identifies 767 significant genes from ATAC-seq and 371 from RNA-seq. List 1 is the intersection of these two sets, with 24 genes. These genes are significant in both ATAC-seq and RNA-seq. The analysis of RNA-seq **counts** and ATAC-seq **counts** starts from the same data as inputs. After the analysis, iClusterPlus identifies 4172 genes from ATAC-seq and 2791 from RNA-seq. The intersection of these two sets contains 709 genes and listed in supplementary list 1. The list of genes, produced with RNA-seq **counts** and DiffBind **counts** as inputs, are too long for this section, therefore included as a supplementary material.

Databases	Number of Pathways (logCPM inputs)	Number of Pathways (counts inputs)
comp_geneontology_Biological_Process	0	83
comp_geneontology_Biological_Process_noRedundant	3	18
comp_geneontology_Molecular_Function	2	1
comp_network_Transcription_Factor_target	0	6
comp_pathway_KEGG	0	3
comp_pathway_Panther	0	3
comp_pathway_Reactome	0	9
comp_pathway_Wikipathway	0	4

Table 1. Pathway analysis by WebGestaltR on the results of iClusterPlus. The table shows the databases and the number of pathways annotated. The second column is the result of integrating RNA-seq read **logCPM** and ATAC-seq peak **logCPM** data. The third column is the result of integrating RNA-seq read **counts** and ATAC-seq peak **counts** data.

Table 1 compares the number of detected pathways between and counts inputs. Counts inputs has higher total number of pathways because of the larger number of intersection genes after clustering analysis.

Comp_geneontology_Biological_Process_ noRedundant		**Comp_geneontology_Molecular_Function**	
geneSet	**description**	**geneSet**	**description**
GO:0050900	leukocyte migration	GO:0030545	receptor regulator activity
GO:0051271	negative regulation of cellular component movement	GO:0048018	receptor ligand activity
GO:0040013	negative regulation of locomotion		

Table 2. Pathway analysis by WebGestaltR on the results of iClusterPlus. iClusterPlus integrates the RNA-seq **logCPM** and ATAC-seq **logCPM** data. The table shows the gene sets and their pathway annotation.

WebGestaltR identifies five significant pathways. The pathways from GO Biological Process relate to the movement of cells and cellular components, which suggest that the treatment samples have significance changes in genes that control movements, induced by the condition. The pathways from GO molecular function are related to receptor activities. The results of the pathway analysis on the integration of RNA-seq **counts** and DiffBind **counts** are too long and included as supplementary table 1.

Multiple Co-Inertia Analysis

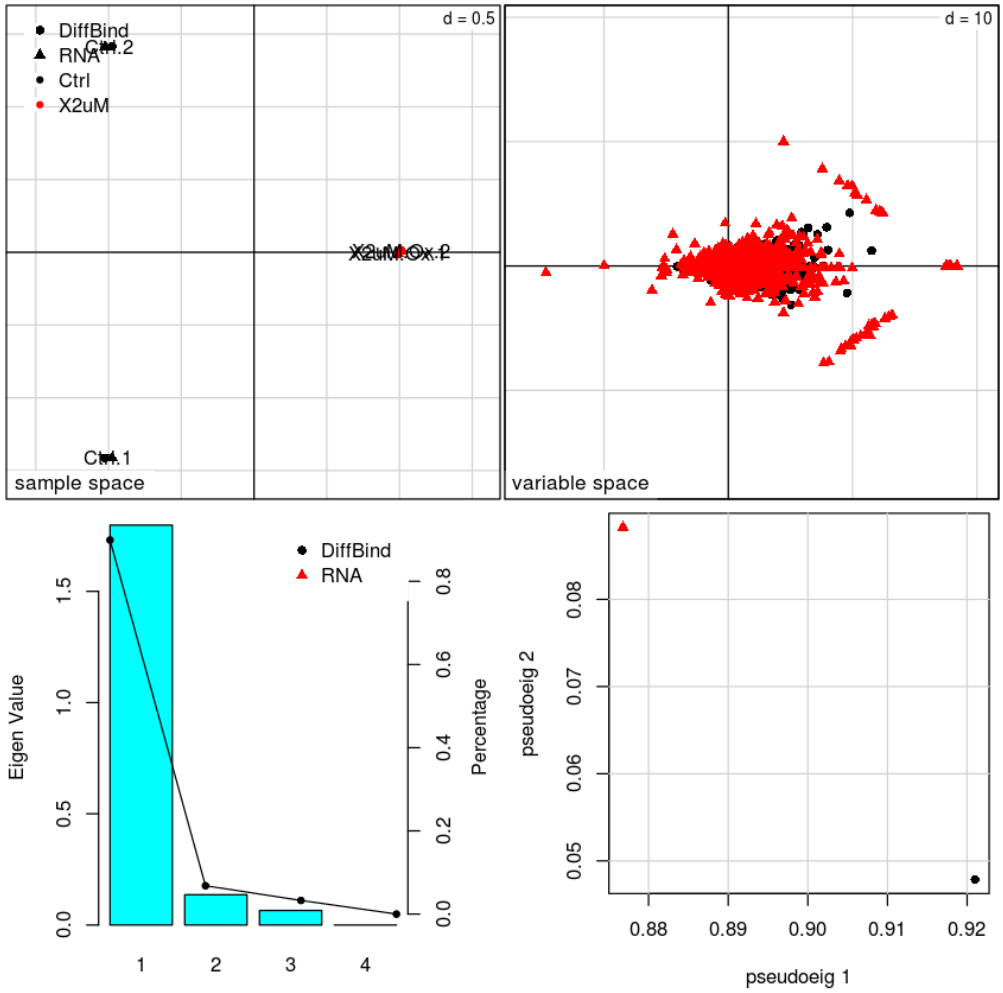


Figure 3. The output of multiple co-inertia analysis is four plots: sample space, variable (gene) space, absolute eigenvalues, and pseudo-eigenvalues space. This set of plots are produced by *omicade4* as the integration of RNA-seq **logCPM** and ATAC-seq **logCPM** inputs.

The sample space plot shows the relatedness among samples represented by their location in space. The Ox samples are very closely related. A similar representation applies for variable space that shows the relatedness of the genes. The majority of the genes from ATAC-seq (labeled DiffBind) overlap in space with the genes from RNA-seq. The bottom left plot shows the eigenvalue for each eigenvector. The first eigenvector has the highest eigenvalue that contributes to the majority of the variance. The bottom right plot "shows the pseudo-eigenvalues space of all datasets, which indicates how much variance of an eigenvalue is contributed by each dataset." [4]

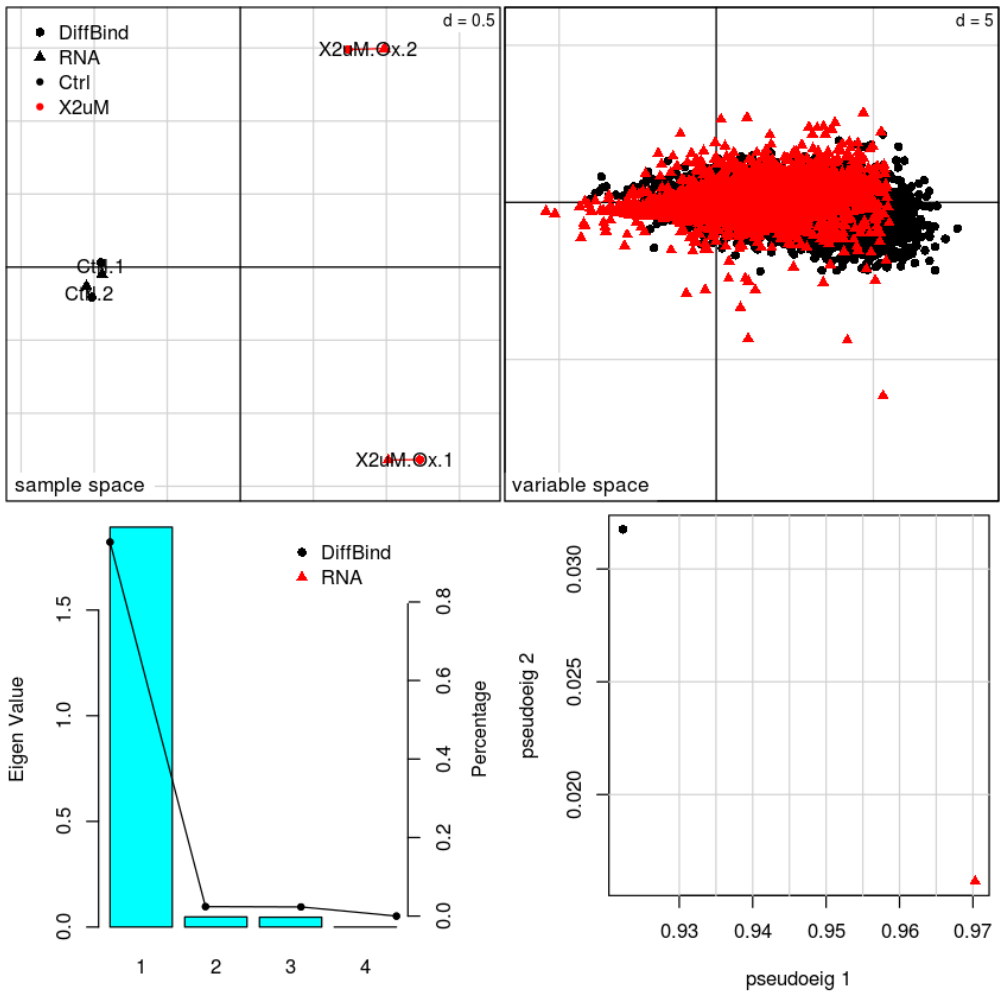


Figure 4. The output of multiple co-inertia analysis is four plots: sample space, variable (gene) space, absolute eigenvalues, and pseudo-eigenvalues space. This set of plots are produced by *omicade4* as the integration of RNA-seq **counts** and ATAC-seq **counts** inputs.

The sample space shows that the Ctrl samples are closely related. Similar to figure 3, the majority of the genes from ATAC-seq (labeled DiffBind) overlap in space with the genes from RNA-seq. The first eigenvector also has the highest eigenvalue. The analysis also generates plots with other axes showing information in different dimensions. Those figures are included as supplementary figure 1 and 2.

Discussion

Correlation Analysis

As my hypothesis states, increasing chromatin accessibility correlates to higher transcription activity, therefore higher increasing gene expression, and vice versa. However, the top plots in figure 1 tell a different story. The distribution of plots and the negative-slope regression lines suggest a negative correlation. Although the plots suggest such trend, they are still inconclusive and require improvements.

Compared to the total of 15161 DB regions identified by DiffBind and 20977 identified by CSAW, only 612 DE genes are identified by RNA-seq, resulting in a total of 521 genes in each intersection. Although the purpose of RNA-seq is to identify a small set of highly enriched DE genes for downstream analysis, this subset is perhaps too small to tell the whole story of the relationship between RNA-seq and ATAC-seq. Therefore, I might require to lower the selection threshold in RNA-seq to allow more genes in correlation.

Increasing the sample size also allows more genes in the correlation. Because this study only compares the samples Ctrl and Ox, such observed correlation only pertains to the selected samples and not the entirety of ATAC-seq and RNA-seq. However, increasing the sample size can also introduce larger variance in the data. Therefore more stringent filtering criteria are required.

The filtering criteria in the plots can be improved for consistency. For instance, the top right plot shows DiffBind filters out genes below a threshold of logFC, along the x-axis. Whereas the right plot shows no filtering step in CSAW. Such step can be added downstream of CSAW for improving consistency.

The logFC of CSAW has drastic different ranges than DiffBind or RNA-seq results. The top two plots show that the logFC of RNA-seq results is in the range from -6 to +6 and that of DiffBind is in the range from -2 to 4, which are reasonable logFC ranges. Surprisingly, the logFC of CSAW lies in the range from -40 to 80, which is an order of 10 log fold change greater than DiffBind and RNA-seq ranges. This issue will be further investigated by re-examining how CSAW calculates the logFC.

Integrative Clustering Analysis

iClusterPlus shows a variety of heatmaps (figure 2) in the integrative analysis. However, it seems like its plotHeatmap() function misses out on labeling of the plots, e.g. sample labels and feature labels. I will further examine the function to find whether such features exist.

iClusterPlus performs integrative clustering analysis to group samples into clustering, as shown in figure 2. This feature will be useful for large sample size since it identifies samples which are closely related. The clustering results in my case is trivial because there are only two samples with two replicates each. In further research, I will include more samples from the Shalapour dataset to generate more meaningful clustering results.

In this study, I use iClusterPlus to integrate ATAC-seq and RNA-seq with two input types, logCPM and counts. Because RNA-seq has counts-based data, I need a similar metric from ATAC-seq for comparisons, and so I a DBA object from DiffBind that contains peak counts and binding scores. I later transform counts into logCPM for a more conventional and normalized comparisons. While it will be interesting to compare counts with binding scores or logCPM with binding scores, the comparisons with the same metrics make more sense. Further study can investigate how different input types compare by continuing my two other discarded comparisons.

Table 1 shows a drastic difference between the number of pathways detected with logCPM and counts inputs. Although counts inputs have more results, logCPM inputs offer more valid comparisons within samples because it is normalized for sequencing depth. I will develop further analysis using normalization methods that allow comparisons across samples.

Table 2 shows the results of pathway analysis in biological processes and molecular functions. The identified pathway in biological processes are all related to movements of the cell. As this paper suggested [11], Ox might stand for Oxaliplatin and is a molecular therapy for colon cancer. The results make sense because it has “cytotoxic effect mostly through dna damage. Apoptosis of cancer cells can be caused by formation of dna lesions, arrest of dna synthesis, inhibition of rna synthesis, and triggering of immunologic reactions”. As a result, Oxaliplatin suppresses cellular functions in the samples with Ox condition. Receptor related activities are identified in molecular function. To suppress cancer cell activities, Ox might have altered target signaling cascade to change certain gene expression.

Multiple Co-Inertia Analysis

The multiple co-inertia analysis generates a series of plots as the results of integration. In the variable space of figure 3 and 4, I observe that the data points of ATAC-seq and RNA-seq occupy similar location. We can infer a subtle relationship between the two. The rationale for choosing multiple co-inertia analysis over regular co-inertia analysis is its ability to integrate multiple datasets. The original plan is to test

multiple samples from the Shalapour datasets for integration, which I assume that the multiple co-inertia analysis can handle more effectively.

With the same problem like the other two integration methods, multiple co-inertia analysis suffers from the small sample size. It is design to process data with large sample size therefore its functionality is tailored towards that. As a result, the results are difficult to interpret and the plots provide little information of the relationship between these two datasets. Further improvements can be made by including more samples from the Shalapour datasets.

Pathway Analysis

WebGestaltR performs well with annotating pathways, critical to my analysis. I initially included gene set enrichment analysis using GSVA, but then discard it. GSVA is not suitable for the output of iClusterPlus because it should input the whole datasets before the integration (e.g. counts, logCPM, etc.). Therefore, GSVA is appropriate for a parallel analysis rather than an integration method.

Final overview

Overall, the study shows interesting results and demonstrates a potentially subtle relationship between ATAC-seq and RNA-seq. The integrative analysis allows for higher specificity of the result while sacrificing some sensitivity, creating only a small set of annotated pathways. After all, this is just a first step of the attempt to fully integrate ATAC-seq and RNA-seq. There are improvements to be made, e.g. including more samples from the Shalapour datasets, before moving forward. The future of the project will be the integration of multi-omics data, including other data types, e.g. DNA-seq, ChIP-seq, MeRIPseq, etc. The integration of multi-omics data will provide great insights to biomedical research by allowing us to fully understand the hidden messages from these high-dimensional omics data.

Reference

1. Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current protocols in molecular biology, 109, 21.29.1-9. doi:10.1002/0471142727.mb2129s109
2. Mo Q, Shen R (2018). *iClusterPlus: Integrative clustering of multi-type genomic data*. R package version 1.18.0.
3. Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., … Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences of the United States of America, 110(11), 4245-50.
4. Meng C, Kuster B, Culhane A, Gholami AM (2013). "A multivariate approach to the integration of multi-omics datasets." *BMC Bioinformatics*.
5. Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., & Ritchie, M. E. (2018). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Research, 5, ISCB Comm J-1408. doi:10.12688/f1000research.9005.3
6. Post: ATAC-Seq Analysis at CCBB: https://ccb-ucsd.slack.com/files/U03AEJ1E9/FA3MBSN15/Post__ATAC-Seq_Analysis_at_CCBB
7. Stark R, Brown G (2011). *DiffBind: differential binding analysis of ChIP-Seq peak data*. http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf
8. Lun ATL, Smyth GK (2016). "csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows." Nucleic Acids Res., 44(5), e45.
9. Cavalcante RG, Sartor MA (2017). "annotatr: genomic regions in context." Bioinformatics. R package version 1.8.0.
10. Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Research.
11. Alcindor, T., & Beauger, N. (2011). Oxaliplatin: a review in the era of molecularly targeted therapy. Current oncology (Toronto, Ont.), 18(1), 18-25.

Supplementary Materials

Integrative Clustering Analysis

Supplementary table 1. Pathway analysis by WebGestaltR on the results of iClusterPlus. iClusterPlus uses the RNA-seq **counts** and DiffBind **counts** data as inputs.

comp_geneontology_Biological_Process

geneSet	description
GO:0009156	ribonucleoside monophosphate biosynthetic process
GO:0032787	monocarboxylic acid metabolic process
GO:0009201	ribonucleoside triphosphate biosynthetic process
GO:0046394	carboxylic acid biosynthetic process
GO:0016053	organic acid biosynthetic process
GO:0009141	nucleoside triphosphate metabolic process
GO:0009206	purine ribonucleoside triphosphate biosynthetic process
GO:0009124	nucleoside monophosphate biosynthetic process
GO:0009145	purine nucleoside triphosphate biosynthetic process
GO:0006754	ATP biosynthetic process
GO:0009127	purine nucleoside monophosphate biosynthetic process
GO:0009168	purine ribonucleoside monophosphate biosynthetic process
GO:1901292	nucleoside phosphate catabolic process
GO:0009199	ribonucleoside triphosphate metabolic process
GO:0006090	pyruvate metabolic process
GO:0006733	oxidoreduction coenzyme metabolic process

geneSet	description
GO:0009142	nucleoside triphosphate biosynthetic process
GO:0046434	organophosphate catabolic process
GO:0034404	nucleobase-containing small molecule biosynthetic process
GO:0046031	ADP metabolic process
GO:0006753	nucleoside phosphate metabolic process
GO:0009185	ribonucleoside diphosphate metabolic process
GO:0019362	pyridine nucleotide metabolic process
GO:0046496	nicotinamide nucleotide metabolic process
GO:0051186	cofactor metabolic process
GO:0072522	purine-containing compound biosynthetic process
GO:0009117	nucleotide metabolic process
GO:0009166	nucleotide catabolic process
GO:0072524	pyridine-containing compound metabolic process
GO:0072330	monocarboxylic acid biosynthetic process
⋮	⋮
GO:0070814	hydrogen sulfide biosynthetic process
GO:0009167	purine ribonucleoside monophosphate metabolic process
GO:0009126	purine nucleoside monophosphate metabolic process
GO:0019693	ribose phosphate metabolic process
GO:0051188	cofactor biosynthetic process
GO:0090407	organophosphate biosynthetic process

geneSet	description
GO:0019359	nicotinamide nucleotide biosynthetic process
GO:0019363	pyridine nucleotide biosynthetic process
GO:0006163	purine nucleotide metabolic process
GO:0050667	homocysteine metabolic process
GO:0072525	pyridine-containing compound biosynthetic process
GO:0006091	generation of precursor metabolites and energy
GO:0009259	ribonucleotide metabolic process
GO:0009150	purine ribonucleotide metabolic process
GO:0009403	toxin biosynthetic process
GO:0006165	nucleoside diphosphate phosphorylation
GO:0046939	nucleotide phosphorylation
GO:0072521	purine-containing compound metabolic process
GO:0034976	response to endoplasmic reticulum stress
GO:0070813	hydrogen sulfide metabolic process
GO:0007220	Notch receptor processing
GO:0046599	regulation of centriole replication
GO:0006520	cellular amino acid metabolic process
GO:0022613	ribonucleoprotein complex biogenesis
GO:0019318	hexose metabolic process
GO:0036503	ERAD pathway
GO:0042274	ribosomal small subunit biogenesis

geneSet	description
GO:0005996	monosaccharide metabolic process
GO:0006413	translational initiation
GO:0030433	ubiquitin-dependent ERAD pathway

comp_geneontology_Molecular_Function

geneSet	description
GO:0003735	structural constituent of ribosome

comp_pathway_KEGG

geneSet	description
mmu03010	Ribosome
mmu01230	Biosynthesis of amino acids
mmu00010	Glycolysis / Gluconeogenesis

comp_pathway_Reactome

geneSet	description
R-MMU-70263	Gluconeogenesis
R-MMU-72613	Eukaryotic Translation Initiation
R-MMU-72737	Cap-dependent Translation Initiation

geneSet	description
R-MMU-191273	Cholesterol biosynthesis
R-MMU-156827	L13a-mediated translational silencing of Ceruloplasmin expression
R-MMU-72649	Translation initiation complex formation
R-MMU-72662	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S
R-MMU-72702	Ribosomal scanning and start codon recognition
R-MMU-70326	Glucose metabolism

comp_pathway_Wikipathway

geneSet	description
WP163	Cytoplasmic Ribosomal Proteins
WP157	Glycolysis and Gluconeogenesis
WP103	Cholesterol Biosynthesis
WP662	Amino Acid metabolism

Supplementary list 1. The intersection of genes between RNA-seq and DiffBind identified by iClusterPlus after integrative clustering analysis. These genes were produced with RNA-seq **counts** and DiffBind **counts** as inputs.

- "A430005L14Rik"
- "Aars"
- "Abcf1"
- "Abcf2"
- "Acaca"
- "Acap3"
- "Acox1"
- "Acsl3"
- "Actr5"
- "Acy1"
- "Adam10"
- "Adamts9"
- "Adamtsl5"
- "Add3"
- "Adipor2"
- "Afmid"
- "Agpat5"
- "Ahdc1"
- "Ahi1"
- "Ahsa2"
- "Ak2"
- "Aldh18a1"
- "Aldh1a3"
- "Aldoa"
- "Aldoc"
- "Alg14"
- "Ankrd13d"
- "Ankrd24"
- "Anxa6"
- "Ap2a2"
- "Ap2s1"
- "Ap5z1"
- "Apba3"
- "Aph1b"
- "Aph1c"
- "Arf2"
- "Arhgap12"
- "Arhgap29"
- "Arhgap31"
- "Arhgef1"
- "Arhgef19"
- "Arhgef5"
- "Arid3a"
- "Arid5a"
- "Arl6ip5"
- "Arpc3"
- "Arsj"
- "Asb4"
- "Asnsd1"
- "Atf3"
- "Atf6b"
- "Atg16l2"
- "Atl3"
- "Atp1a1"
- "Atp5b"
- "Atp5c1"
- "Atp5s"
- "Atp6v0b"
- "AU041133"
- "Aunip"
- "Baz2b"
- "Bcat2"
- "Bclaf3"
- "Bdh1"
- "Bhlha15"
- "Blmh"
- "Blvrb"
- "Brpf3"
- "Brwd3"
- "Btd"
- "Btrc"
- "Bud23"
- "C130036L24Rik"
- "C1galt1c1"
- "C1qbp"
- "C3"
- "Cad"
- "Cage1"
- "Calcoco1"
- "Calr4"
- "Car13"
- "Carm1"
- "Catsperd"
- "Cbfa2t2"
- "Cbs"
- "Cbx1"
- "Ccdc130"
- "Ccdc86"
- "Cct3"
- "Cct7"
- "Cd320"
- "Cd44"
- "Cdc42ep1"
- "Cdk5r1"
- "Cdk6"
- "Cdkl3"
- "Cdkn1a"
- "Cdkn2aip"
- "Cdnf"
- "Cenpj"
- "Cep192"
- "Cep83"
- "Cfap36"
- "Cggbp1"
- "Chac1"
- "Chchd2"
- "Chrm1"
- "Chrnb1"
- "Chst15"
- "Ckap5"
- "Clec16a"
- "Clk1"
- "Clk2"
- "Clu"
- "Cluh"
- "Cnn2"
- "Cnst"
- "Cntnap1"
- "Coa4"
- "Commd4"
- "Copz1"
- "Coq10b"
- "Coq2"
- "Cotl1"
- "Cox7a2l"
- "Cpeb4"
- "Cpm"
- "Crebrf"
- "Creld2"
- "Cs"
- "Ctcf"
- "Cth"
- "Cul7"
- "D10Jhu81e"
- "Dancr"
- "Dapk1"
- "Dazap1"
- "Dcp1a"
- "Dctn4"
- "Dctpp1"
- "Ddt"
- "Ddx39"
- "Ddx41"
- "Ddx54"
- "Dennd5a"
- "Dhcr24"
- "Dhx37"
- "Dhx8"
- "Dixdc1"
- "Dlat"
- "Dlg1"
- "Dmtn"
- "Dmxl1"
- "Dnaaf5"
- "Dnajb11"
- "Dnajb14"
- "Dnajb5"
- "Dnmbp"
- "Dnpep"
- "Dnph1"
- "Dok7"
- "Donson"
- "Dpep1"
- "Dst"
- "Dus1l"
- "Dusp16"
- "Dusp3"
- "Dync2h1"
- "Eaf2"
- "Ech1"

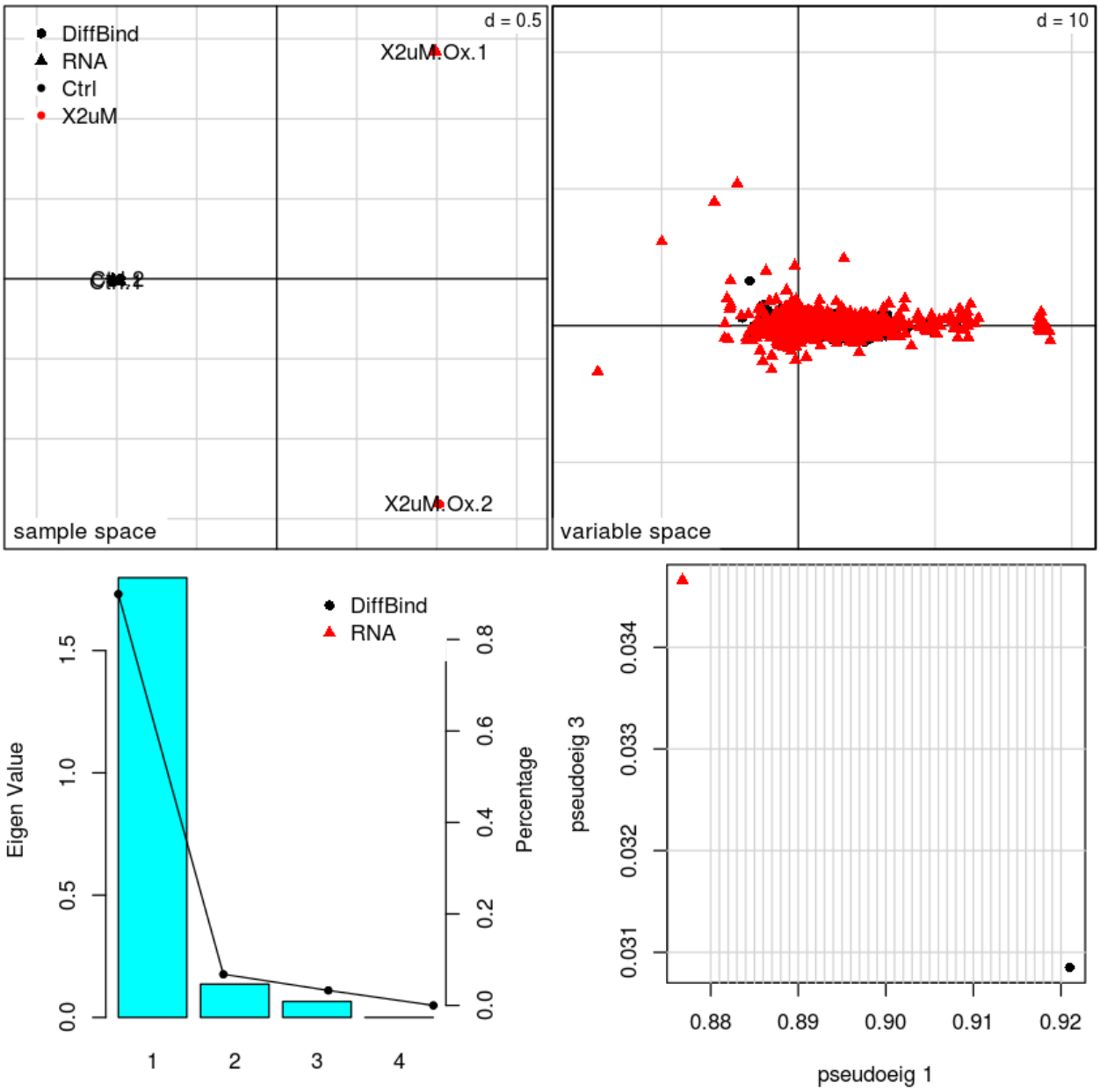
- "Echdc1"
- "Edem1"
- "Eea1"
- "Eef1g"
- "Ehbp1l1"
- "Eif2b2"
- "Eif2b4"
- "Eif3f"
- "Eif3i"
- "Eif3l"
- "Eif4a1"
- "Eif4b"
- "Eif4g1"
- "Elac2"
- "Ell3"
- "Emg1"
- "Eno1"
- "Eno1b"
- "Entpd5"
- "Entpd6"
- "Entpd7"
- "Epb41l2"
- "Epb41l4b"
- "Epha2"
- "Eprs"
- "Ercc8"
- "Erlin2"
- "Erp29"
- "Evpl"
- "Exd2"
- "Exoc4"
- "Exosc4"
- "F8"
- "Faf2"
- "Fam117a"
- "Fam210b"
- "Fam221a"
- "Fam45a"
- "Fbxo21"
- "Fbxo44"
- "Fbxo5"
- "Fbxo6"
- "Fdps"
- "Fez2"
- "Fhl4"
- "Fkbp11"
- "Flot1"
- "Fnip1"
- "Foxred2"
- "Fubp1"
- "Fyn"
- "Gale"
- "Galk1"
- "Galnt11"
- "Galnt2"
- "Ganab"
- "Gapdh"
- "Gcat"
- "Gcc1"
- "Gdi2"
- "Gdpd5"
- "Ggct"
- "Glce"
- "Gm2a"
- "Gnb5"
- "Gnpat"
- "Gnptab"
- "Gorab"
- "Gorasp2"
- "Gpatch2l"
- "Gpatch8"
- "Gpr160"
- "Grcc10"
- "Gstt3"
- "Gtf2e2"
- "H1f0"
- "H2afz"
- "H3f3b"
- "Hacl1"
- "Haus3"
- "Hbegf"
- "Hint2"
- "Hip1r"
- "Hipk4"
- "Hmbs"
- "Hmgb3"
- "Hmgcs1"
- "Hnrnpa1"
- "Hnrnpab"
- "Hnrnpdl"
- "Hnrnpm"
- "Hr"
- "Hsp90ab1"
- "Hspa5"
- "Hspe1"
- "Iars"
- "Id3"
- "Idi1"
- "Ifngr1"
- "Ifrd2"
- "Igfbp5"
- "Il34"
- "Il4ra"
- "Il6st"
- "Ilvbl"
- "Immp1l"
- "Imp3"
- "Impact"
- "Impdh1"
- "Inpp5a"
- "Invs"
- "Iqcb1"
- "Irak2"
- "Irf9"
- "Irs2"
- "Isoc2a"
- "Itih2"
- "Jak2"
- "Jarid2"
- "Jmjd6"
- "Junb"
- "Kat2a"
- "Kctd21"
- "Kdelr2"
- "Kdelr3"
- "Khk"
- "Kifc3"
- "Klf10"
- "Klf11"
- "Klf6"
- "Klhdc3"
- "Klhl12"
- "Kmt2a"
- "Kntc1"
- "Krt14"
- "Krt19"
- "Lactb"
- "Lag3"
- "Lap3"
- "Larp1b"
- "Ldha"
- "Ldlr"
- "Letm1"
- "Lig3"
- "Lims1"
- "Lix1l"
- "Lman2"
- "Lmnb1"
- "Lonp1"
- "Lonp2"
- "Lpcat3"
- "Lpp"
- "Lpxn"
- "Lrpap1"
- "Lrrc56"
- "Lrrc75a"
- "Lrrk1"
- "Lss"
- "Lta4h"
- "Lynx1"
- "Lyst"
- "M6pr"
- "Man1a2"
- "Manf"
- "Map1b"
- "Map1lc3a"
- "Map2k3"
- "Map3k6"
- "Map4k1"
- "Map4k2"
- "Mapk8ip1"
- "Mbnl2"
- "Mbtps1"
- "Mdh2"
- "Me3"
- "Med11"
- "Med13l"
- "Med19"
- "Mettl23"
- "Mfsd11"
- "Mid1ip1"
- "Mief1"
- "Mipep"
- "Mipol1"
- "Mlec"
- "Mlxip"
- "Mlycd"
- "Mob4"

- "Morf4l2"
- "Mpnd"
- "Mpp7"
- "Mpst"
- "Mrpl11"
- "Mrpl12"
- "Mrpl58"
- "Mrps12"
- "Msrb2"
- "Mthfd1"
- "Mthfd1l"
- "Mtss1l"
- "Mturn"
- "Mvk"
- "Mvp"
- "Mxd1"
- "Mxd4"
- "Myc"
- "Myl6b"
- "Myo5c"
- "Myo9a"
- "Naglu"
- "Narf"
- "Nbdy"
- "Ncl"
- "Ncln"
- "Nedd8"
- "Nelfb"
- "Nfatc1"
- "Nfkbie"
- "Ngdn"
- "Ngly1"
- "Nicn1"
- "Nkiras1"
- "Nkrf"
- "Nmral1"
- "Noa1"
- "Npm1"
- "Nr2c2"
- "Nrbp2"
- "Nrip1"
- "Nsd3"
- "Nt5c"
- "Ntmt1"
- "Nub1"
- "Nup210"
- "Nupl2"
- "Oaz1"
- "Ofd1"
- "Ogdh"
- "Osbpl2"
- "Osgin2"
- "Otub1"
- "P4ha2"
- "Pabpn1"
- "Padi2"
- "Paip2b"
- "Pard6g"
- "Parp10"
- "Parp4"
- "Parp8"
- "Pbdc1"
- "Pck2"
- "Pcmtd1"
- "Pcyox1l"
- "Pcyt2"
- "Pdia6"
- "Pdp2"
- "Pelp1"
- "Pgd"
- "Pgghg"
- "Pgp"
- "Phb2"
- "Phf13"
- "Phlda1"
- "Pim3"
- "Pkm"
- "Pla2g15"
- "Plcg2"
- "Plxna3"
- "Plxnd1"
- "Pnkd"
- "Pofut2"
- "Polr1c"
- "Polr2a"
- "Pomt2"
- "Ppfia3"
- "Ppif"
- "Ppm1g"
- "Ppox"
- "Ppp1r35"
- "Ppp1r9a"
- "Ppp2r1a"
- "Ppp4c"
- "Prdx1"
- "Prdx3"
- "Prelid1"
- "Prkcsh"
- "Prkra"
- "Prnp"
- "Prox2"
- "Prps1l3"
- "Prrg2"
- "Prune2"
- "Psap"
- "Psenen"
- "Psma4"
- "Psmc4"
- "Psmd1"
- "Psmd12"
- "Psmd13"
- "Psmd7"
- "Psme1"
- "Psme2"
- "Ptdss1"
- "Ptges"
- "Ptges3"
- "Ptrh2"
- "Pttg1ip"
- "Puf60"
- "Pura"
- "Pxn"
- "Pycr1"
- "Qdpr"
- "Rab11fip3"
- "Rab1b"
- "Ran"
- "Rapsn"
- "Rasa4"
- "Rbm14"
- "Rbm42"
- "Rbm48"
- "Rcn1"
- "Rdm1"
- "Retsat"
- "Rev3l"
- "Rfesd"
- "Rfx5"
- "Rhbdd1"
- "Ribc1"
- "Riok3"
- "Riox1"
- "Rnd1"
- "Rnf181"
- "Rnf217"
- "Rnf26"
- "Rpl10"
- "Rpl12"
- "Rpl13a"
- "Rpl18a"
- "Rpl22l1"
- "Rpl27a"
- "Rpl31"
- "Rpl37"
- "Rpl39"
- "Rpl9"
- "Rps10"
- "Rps14"
- "Rps19"
- "Rps20"
- "Rps24"
- "Rps25"
- "Rps3"
- "Rps4x"
- "Rps6"
- "Rps8"
- "Rps9"
- "Rrm2b"
- "Rrp9"
- "Rsph3b"
- "Rspry1"
- "Rtkn2"
- "Ruvbl1"
- "Ruvbl2"
- "Sars2"
- "Sbk1"
- "Scamp5"
- "Scap"
- "Sdhaf2"
- "Sec11a"
- "Sec23ip"
- "Sec61a1"
- "Selenok"
- "Selenoo"
- "Sema3b"
- "Senp8"
- "Sept1"
- "Serpine1"

- "Serpini1"
- "Setd5"
- "Sf1"
- "Sfpq"
- "Sgk1"
- "Sgpp1"
- "Sgsm3"
- "Sh3bgrl3"
- "Shoc2"
- "Ska2"
- "Slc25a10"
- "Slc25a13"
- "Slc25a3"
- "Slc27a1"
- "Slc29a2"
- "Slc2a1"
- "Slc2a6"
- "Slc31a1"
- "Slc35b1"
- "Slc35b2"
- "Slc37a4"
- "Slc38a2"
- "Slc39a3"
- "Slc39a7"
- "Slc52a2"
- "Slco1a5"
- "Slk"
- "Slx4ip"
- "Smad6"
- "Smarca4"
- "Smim24"
- "Smim8"
- "Smpd3"
- "Smtn"
- "Snap29"
- "Snd1"
- "Sned1"
- "Snn"
- "Snrnp35"
- "Snrnp40"
- "Snrpd2"
- "Snrpg"
- "Sp1"
- "Spata5"
- "Spcs2"
- "Sphk1"
- "Spice1"
- "Spns2"
- "Sppl3"
- "Spr"
- "Sptan1"
- "Sqle"
- "Sqstm1"
- "Srebf2"
- "Srpr"
- "Srsf2"
- "Ssh1"
- "Stap2"
- "Stard4"
- "Stim1"
- "Stk19"
- "Stk38l"
- "Stk4"
- "Stx17"
- "Stxbp2"
- "Syngr2"
- "Synrg"
- "Tanc2"
- "Taok3"
- "Tapbp"
- "Tars"
- "Tbrg4"
- "Tcam1"
- "Tcf19"
- "Tcn2"
- "Tcp1"
- "Tcp11l1"
- "Tecr"
- "Ten1"
- "Tfdp2"
- "Thnsl1"
- "Thop1"
- "Tk2"
- "Tmc4"
- "Tmco4"
- "Tmem101"
- "Tmem129"
- "Tmub1"
- "Tmx2"
- "Tnfaip8"
- "Tns1"
- "Tom1"
- "Tom1l2"
- "Tpi1"
- "Tpp1"
- "Tprn"
- "Trafd1"
- "Trappc1"
- "Trappc6a"
- "Treh"
- "Trim7"
- "Trir"
- "Trnau1ap"
- "Trp53inp1"
- "Trp53inp2"
- "Tubb4b"
- "Tufm"
- "Twsg1"
- "Txn2"
- "Uba7"
- "Ube2n"
- "Ubfd1"
- "Ubqln2"
- "Umad1"
- "Uqcrfs1"
- "Usf3"
- "Usp20"
- "Usp30"
- "Vamp1"
- "Vars"
- "Vcl"
- "Vcp"
- "Vps26a"
- "Wdpcp"
- "Wdr19"
- "Wdr4"
- "Wdr5"
- "Wdr5b"
- "Wdr77"
- "Wdr83"
- "Wdr83os"
- "Xntrpc"
- "Yars"
- "Ypel5"
- "Ywhae"
- "Zan"
- "Zbtb18"
- "Zbtb20"
- "Zbtb32"
- "Zbtb37"
- "Zbtb7a"
- "Zcwpw1"
- "Zdhhc13"
- "Zfp263"
- "Zfp330"
- "Zfp35"
- "Zfp358"
- "Zfp367"
- "Zfp39"
- "Zfp512"
- "Zfp593"
- "Zfp597"
- "Zfp607b"
- "Zfp672"
- "Zfp868"
- "Zfp932"
- "Zfp946"
- "Zfp956"
- "Zgrf1"
- "Zhx1"
- "Zkscan17"
- "Zmat1"
- "Zwilch"

Supplementary figure 1.

The output of multiple co-inertia analysis with different axis from the figures in the results section. This set of plots are produced by *omicade4* as the integration of RNA-seq **logCPM** and ATAC-seq **logCPM** inputs.



Supplementary figure 2.

The output of multiple co-inertia analysis with different axis from the figures in the results section. This set of plots are produced by *omicade4* as the integration of RNA-seq **counts** and ATAC-seq **counts** inputs.