bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CoBRA: Containerized Bioinformatics workflow for Reproducible ChIP/ATAC-seq Analysis - from differential peak calling to pathway analysis

Xintao Qiu[1,2,#], Avery S. Feit[2,3,#] , Ariel Feiglin[4,#], Yingtian Xie[1], Nikolas Kesten[1], Len Taing[1,5], Joseph Perkins[1], Ningxuan Zhou[1], Shengqing Gu[5], Yihao Li[2], Paloma Cejas[1,2], Rinath Jeselsohn[2], Myles Brown[1,2], X. Shirley Liu[1,5], Henry W. Long[1,2]*

1 Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, 02215, USA
2 Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, 02215, USA
3 Albert Einstein College of Medicine, The Bronx, NY, 10461 USA
4 Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, USA.
5 Department of Data Sciences, Dana Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA.
*Correspondence: Henry_Long@dfci.harvard.edu
[#]Equal contributors

KEYWORDS: ChIP-seq, ATAC-seq, Snakemake, Docker, workflow

Word Counts:
ABSTRACT: 118
ARTICLE: 3593

REFERENCES: 27
FIGURES: 4
TABLES: 1
SUPPLEMENTARY FIGURES: 3
SUPPLEMENTARY TABLES: 0

Running Title: “CoBRA: Containerized analysis for ChIP/ATAC-seq”

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

ChIP-seq and ATAC-seq have become essential technologies used as effective methods of measuring protein-DNA interactions and chromatin accessibility. However, there is a need for a scalable and reproducible pipeline that incorporates correct normalization between samples, adjustment of copy number variations, and integration of new downstream analysis tools. Here we present CoBRA, a modularized computational workflow which quantifies ChIP and ATAC-seq peak regions and performs unsupervised and supervised analysis. CoBRA provides a comprehensive state-of-the-art ChIP and ATAC-seq analysis pipeline that is usable by scientists with limited computational experience. This enables researchers to gain rapid insight into protein-DNA interactions and chromatin accessibility through sample clustering, differential peak calling, motif enrichment, comparison of sites to a reference DB and pathway analysis.

Code availability: https://bitbucket.org/cfce/cobra

Introduction

ChIP-seq and ATAC-seq have become essential components of epigenetic analysis. They are employed extensively in the study of protein-DNA interactions and chromatin accessibility respectively. Chromatin immunoprecipitation sequencing (ChIP-seq) is a high-throughput technology that provides unique insights into protein function by mapping genome wide binding sites of DNA-associated proteins. Further, Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) is a high-throughput technology that is imperative in the assessment of genome-wide chromatin accessibility. While numerous pipelines for analyzing ChIP-seq and ATAC-seq data have been reported in the literature [1–8] there remains a strong need for pipelines that can be run by users who have less experience utilizing computational biology tools. Comparisons between ChIP and ATAC-seq experiments can provide insight into differences in protein occupancy, histone marks and chromatin accessibility (Figure 1a), however, existing analysis pipelines can lack useful components necessary in the analysis. There is a need for better normalization between samples, adjustment of copy number variations, integrating new downstream analysis tools such as cistromeDB Toolkit [9] and integrating epigenetic data with RNA-seq.

In this work we developed CoBRA (Containerized Bioinformatics workflow for Reproducible ChIP/ATAC-seq Analysis), a modularized computational workflow which quantifies ChIP and ATAC-seq peak regions and performs unsupervised and supervised analysis. The pipeline provides sample clustering, differential peak calling, motif enrichment and clustering, comparison of sites to a reference DB and pathway analysis. In addition, it provides clear, high-quality visualizations for all results.

CoBRA uses Snakemake [10], a workflow management system to create the computational pipeline. Using Snakemake system enables the reproducibility and scalability of CoBRA. This framework allows for the addition or replacement of analysis tools as well as for the parallelization of computationally intensive processes. To make CoBRA portable, the workflow and its software dependencies are available as a Docker container, which can be used on any machine with Docker installed. This includes local servers, high-performance clusters, and cloud-based machines. Docker will automatically download all required software dependencies because the container encapsulates all of the supporting software and libraries, eliminating the possibility of conflicting dependencies.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CoBRA therefore provides solutions to challenges inherent in many bioinformatics workflows: it is portable, reproducible, scalable and easy to use. It is open source (https://bitbucket.org/cfce/cobra) and well documented online including step-by-step tutorials that go through all three case studies presented in this paper (https://cfce-cobra.readthedocs.io). This combination of features enables researchers to gain rapid insight into protein-DNA interactions and chromatin accessibility with comprehensive state-of-the-art ChIP and ATAC-seq analysis.

Methods

Overall design

The CoBRA pipeline is implemented using the snakemake workflow management system [10] and is described via a human-readable, Python-based language. This allows CoBRA to scale to server, cluster, grid and cloud environments, without the need to modify the workflow. For ChIP-seq and ATAC-seq experiments, CoBRA provides both unsupervised and supervised analyses (Figure 1b). It does not include preprocessing ChIP and ATAC-seq quality control steps as this is best handled within other, specialized pipelines [11].

Further, CoBRA is distributed as a Docker container, which can be used on any machine as long as Docker is installed. Docker containers provide a tool for packaging bioinformatics software. It encapsulates all of the supporting software and libraries, eliminates the possibility of conflicting dependencies, and facilitates the installation of required software. With the built-in snakemake reference rule, CoBRA automatically downloads all needed reference files if they have not been downloaded. As a result, CoBRA is reproducible, portable and easy to deploy. Users specify analysis parameters in a simple human readable configuration file (Supplementary Fig1b). A separate file contains metadata about the samples being analyzed (cell line, treatment, time point, etc.) as well as a specification of the differential comparisons to be performed by the pipeline. This metadata file is in CSV format and can be easily modified in any standard text editor or Excel.

Unsupervised analysis

The pipeline calculates the Reads per Kilobase per Million Mapped Reads (RPKM) using bed files and bam files provided by the user to normalize for sequencing depth and peak size. The RPKM table is filtered through the removal of sites that have low RPKM across multiple samples. Quantile Normalization (default), Z-score, Log transform are available options to normalize the count matrix. To visualize the similarities between samples in the experiment, sample-sample correlation, principal component analysis (PCA) and sample-feature plot are automatically generated by the pipeline.

The sample-sample correlation plot illustrates the similarity between all of the samples on a pairwise basis. It also provides the clustering result based on the Pearson correlation coefficient, r, where distance = 1 - r. The user can opt for using spearman correlation as well as selecting other distance methods (euclidean, manhattan, canberra, binary, maximum, or minkowski) by simply changing the configuration file. The resulting correlation plot helps to determine whether the different sample types can be separated, i.e., samples of different conditions are expected to be more dissimilar to each other than replicates within the same condition. User provided metadata are used to automatically annotate samples in all unsupervised plots.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Further, CoBRA will produce a principal component analysis (PCA) plot depicting how samples are separated in the first two principal components (those with the largest variance) and samples will be automatically color-coded by all user provided annotations. The PCA plot helps the user to determine if any patterns exist between the samples and if outliers are present. Finally, CoBRA will generate a Sample-Feature heatmap. The heatmap illustrates the clustering of samples based on correlation on the horizontal axis and clustering of peaks on the vertical axis. Peaks on the vertical axis can be clustered by hierarchical or k-means clustering. The sample-feature heatmap elucidates patterns of peaks across samples and identifies the clusters that are enriched in a subset of samples.

Supervised analysis

A common question being asked is what the differential sites are (TF binding/ histone modification/ chromatin accessibility) between sample groups. Several tools (DEseq2, edgeR, Limma) currently available can be applied to analyze differential sites, most of which are derived from RNA-seq count analysis. However, there are differences between the RNA-seq and ChIP-seq count analysis. In RNA-seq experiments, most reads are in the exome, where reads can be normalized by the total number of reads mapped to all genes. In contrast, most ChIP-seq reads are outside of peaks. The FRiP score (fraction of reads in peaks) typically ranges from 1 to 40 percent [11]. Reads in peaks are only a portion of total reads that have been sequenced. Therefore, all reads need to be normalized by the total number of uniquely mapped reads to account for sequence depth. CoBRA uses the bam file to calculate sequencing depth. It utilizes sequencing depth as a scale factor in differential peak calling by DESeq2 (although the user can specify reads in peaks for scaling if specifically required). This is an essential step in differential peak calling. The default scale factor utilized by DESeq2 to normalize the data is the total number of reads mapped to peaks. This method can result in the calling of false positive differential peaks. Using sequencing depth as the scale factor ensures that reads are normalized for experimental variation and not biological variation between samples.

Multiple comparisons can be done within a single run. For each comparison, the number of differential peaks for two adjusted p-value cutoffs and two fold-change cut-offs will be displayed in a summary chart. Further, the bigwig files are used to plot the peak intensity of the differential peaks in a heatmap using deepTools2 [12].

The differentially enriched regions from DEseq2 for each comparison are subsequently run through HOMER [13] for motif enrichment analysis. Motif enrichment analysis is a fundamental approach to look for transcription factor motifs that might be enriched in peaks of interest. We use Homer in the pipeline to look for known and de novo motifs that are enriched in the differential peak regions compared to GC matched, randomly selected genome background. In addition, we utilize a motif clustering algorithm to organize various motifs by similarity making the output (Supplementary Figure 3) easier to evaluate for distinct results. By mapping the peaks to the nearest gene, CoBRA uses GSEA pre-ranked analysis to investigate the pathways that are enriched and depleted for both up and down peaks.

The up and down-regulated sites are also automatically compared to a comprehensive database of ChIP/ATAC and DNase data [9; 14]. This Cistrome Toolkit analysis determines the most similar samples in terms of genomic interval overlaps with the differential sites. The toolkit is particularly useful to identify the major transcription factors related to the differential perturbations. In addition, it can be useful in the identification of the potential biological source (cell line, cell type and tissue type) of the regions of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

In order to illustrate the utility of CoBRA, we applied it to three projects with components that illustrate the different capabilities of our workflow: a GR ChIP-seq data set from the ENCODE project, H3K27ac ChIP-seq data from colon cancer cell lines, and an ATAC-seq experiment on HL-60 promyelocytes differentiating into macrophages. Each example demonstrates some key functions of the CoBRA pipeline.

Case Studies

Example 1: Normalizing GR ChIP-seq data in a dose-response experiment

We downloaded publicly available glucocorticoid receptor (GR) ChIP-seq data (GSE32465) from a lung adenocarcinoma cell line (A549) at 3 different concentrations of dexamethasone, a potent GR agonist. In an analysis of this dataset [15], it was found that the number of Glucocorticoid Receptor (GR) binding sites increases with increasing dexamethasone concentration. In the experiment, samples were treated with 0.5nM, 5nM, or 50nM of dexamethasone. CoBRA's unsupervised analysis showed that the sample replicates cluster tightly together. Similarities and differences between samples are illustrated by the correlation between treatments vs within treatment in the dendrogram at the top of sample-sample heatmap (Figure 2a), as well as the principal component plot (Supplementary Figure 1).

While unsupervised analyses are useful, the advantage of the CoBRA pipeline is its ability to accurately call differential peaks accounting for a variety of factors. We applied DESeq2 to assess the differences in peak binding for samples treated with 50nM of dexamethasone versus samples treated with 0.5nM of dexamethasone. Utilizing DESeq2's default scale factor method which normalizes the data using the total number of reads in peaks, differential peaks are called (Figure 2b) where they are clearly not present (Figure 2c left). A group of peaks at the bottom of the figure 2c exhibit similar binding intensity, however, they are considered downregulated in 50nM treatment in the DEseq2 result.

DEseq2 by default normalizes all samples by total reads in the read count table. In RNA-seq, most reads are in the exome, where reads can be normalized by the total number of reads mapped to all genes. In contrast, in the GR ChIP-seq experiment, samples treated with 50nM dexamethasone exhibit much greater GR binding and the FRiP score is higher than samples treated with 0.5nM (9.3 vs 0.9). Therefore, DESeq2's normalization method decreases the peak intensity in the 0.5nM treated samples because the FRiP scores are higher in the 50nM sample resulting in false positive differential peaks (Figure 2c right). In CoBRA, we use a scaling factor dependent on the sequencing depth of each sample. This eliminates the false positive downregulated peaks called by DESeq2 using the default scaling factor (Figure 2c right and Figure 2b). Furthermore, more real differential gained peaks have been successfully identified with CoBRA's scaling method.

An additional feature of CoBRA is that it automatically analyzes the differential peaks to provide additional insight into their origin and identify similar systems in the literature. In one analysis it determines the most similar ChIP-seq data that is available in a large, curated database of ChIP and ATAC data - cistrome.org [14]. For the gained GR binding sites in the dexamethasone treatment, the result from the Cistrome Toolkit [9] clearly shows that the *NR3C1* in lung tissue is the most similar ChIP-seq in the cistrome database (Figure 2d). CoBRA provides a list of GEO accession numbers corresponding to all ChIP-seq data with similarity to the differential peak set. Using these identifiers, ChIP seq data of interest can be

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

downloaded for further investigation from Cistrome DB[14]. While obviously correct in this simple case, this tool can provide unique insight into gained or lost sites such as identifying which transcription factor potentially binds to a differential peak set after a perturbation and in investigating similar cellular systems. In addition, CoBRA performs a de novo motif analysis on differential sites which can help to identify potential transcriptional regulators enriched in our differentially accessible chromatin elements. In this example the top cluster has all hormone receptor motifs enriched in the upregulated peaks.

Example 2: Correcting for Copy Number variation in H3K27ac ChIP-seq

We further illustrate the advantages of the CoBRA pipeline utilizing data from colorectal cancer cell lines. Microsatellite Instable (MSI) and Microsatellite Stable (MSS) are two classes used to characterize colorectal cancers. To analyze these cell lines, we selected six publicly available datasets from several experiments: three MSI samples and three MSS samples [16-20] (GSM1866974, GSM2265670, GSM1224664, GSM1890746, GSM2058027, GSM1890746).

MSS tumors are one of the most highly mutated tumor types [21] and typically exhibit a high number of copy number alterations. Without adjustment, a differential peak caller will rank peak loci with high copy number gain in MSS as being the most differential compared to MSI. These genetic differences, while important, can obscure important epigenetic differences between MSI and MSS. In order to observe differential peaks other than those called as a result of the presence of CNV, copy number variation adjustment was conducted on all samples. For this example the copy number was called using the ChIP-seq data itself with CopywriteR [22] but can also be done with qDNAseq [23] using the input control if available. Any other source of CNV data can also be used when put in a standard igv format. This CNV adjustment alters the differential peaks called by DESeq2. In the case of the MSS vs. MSI comparison, many peaks at the 8q region of the chromosome are being called significantly differential (Figure 3a) but, following CNV correction, the number of differential peaks in this region significantly decreased (Figure 3b).

Gene Set Enrichment analysis is performed on the ranked list of genes produced by CoBRA. Without CNV adjustment, GSEA can indicate greatest enrichment in gene sets solely related to amplification. As a result, it is challenging to assess the true epigenetic differences between the two colorectal cancer types. For instance, the gene set 'NIKOLSKY_BREAST_CANCER_8Q12_Q22_AMPLICON' includes genes up-regulated in non-metastatic breast cancer tumors with amplification in the 8q22 region. Without adjustment for copy number variation, this gene set is significantly enriched (Figure 3c). It is the 3th ranked gene, with a normalized enrichment score of -2.84 and an adjusted p-value less than 0.0001. With CNV adjustment, this gene set is far less enriched (Fig. 3c). It is the 468th ranked gene set and has a normalized enrichment score of -1.32 and an adjusted p-value of 1.

After CNV correction, the Hallmarks GSEA analysis shows that the MSI cell line has enrichment in the following pathways: TNFA signaling via NFKB, TFG beta signaling, and Inflammatory response (Figure 3d). This is consistent with the literature[24-25] in reference to colon cancer with MSS tumors exhibiting more inflammatory signaling.

Example 3: Unsupervised analysis of time series ATAC-seq data

In this example, we illustrate the efficacy of CoBRA's analysis of ATAC-seq experiments by following the chromatin accessibility profile of differentiating cells [26]. In this experiment researchers utilized a five-day time course (0hr, 3hr, 24hr, 96hr, and 120hr) to profile accessible chromatin of HL-60 promyelocytes differentiating into macrophages (GSE79019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The CoBRA output includes a principal component analysis (PCA) plot (Figure 4a) that demonstrates the temporal differentiation of the macrophages; the early time point is on the left side while the late time point is on the right. Furthermore, the output includes a sample-feature heatmap utilizing k-means (k=3) clustering (Figure 4b) that further illustrates the dramatic differences in open chromatin profiles. The three clusters show clear differences in open chromatin between the early (cluster 1), intermediate (cluster 2), and late stage (cluster 3) time points.

CoBRA automatically performs a de novo motif analysis on each of the three clusters of accessible sites to identify motifs of potential transcriptional regulators enriched in differentially accessible chromatin elements. This analysis identified many transcription factor motifs enriched in each cluster (Figure 4c). Motifs for PU.1, RUNX and MYB were enriched in cluster 1, which exhibits a decrease in accessibility during myeloid differentiation. It is likely that a depletion of PU.1, RUNX and MYB occupancy occurs at these elements during cellular commitment. In addition, we observe the EGR and MAF motifs in clusters 3 suggesting a gain of EGR and MAF occurs at these elements during macrophage differentiation. The motif analysis for cluster 2 also identified chromatin element NFKB and NFE2 as being active during differentiation and depleted in the latter stages. All of these findings are consistent with the results from published papers [26].

Finally, ChIP-seq and ATAC-seq data is often generated in parallel with RNA-seq on the same samples. An extension to CoBRA can take the differential expression gene list from RNA-seq analysis tools such as VIPER [27] and highlight differentially expressed genes that also exhibit differential chromatin accessibility. The volcano plot in Figure 4d is a visualization of the genes differentially expressed during macrophage differentiation and highlights those genes that also have nearby opening chromatin during differentiation. Genes near open chromatin during differentiation are more likely to be upregulated. This profile that combines chromatin accessibility with gene expression can provide insight to potentially identify major transcriptomic elements driving differentiation.

Discussion

The case studies that we have presented highlight typical use cases for CoBRA. The first example is accurate identification of differential peaks using appropriate normalization of ChIP-seq data. Some methods fail to normalize correctly in calling differential peaks when the FRiP score is impacted by perturbations. CoBRA reduces false positives and identifies more true differential peaks by correctly normalizing for sequencing depth.

The second example demonstrates how CoBRA can be used to account for amplification due to copy number variation present in experimental samples. This is an important feature, as copy number variation can drive the greatest differences between some tumor samples and obscure other biological changes to the cistrome that occur as a result of treatment or other experimental conditions. After adjustment of CNV, differential peaks called by DESeq will not be affected by amplification between samples, allowing biologists to better understand whether differences are caused by changes in the genetic or epigenetic landscape.

The third example illustrates how CoBRA can be applied to ATAC-seq experiments. Unsupervised analyses can identify changes in the chromatin accessibility over time with treatment, and clustering provides insight into similarities and differences between samples and the investigation of the transcription factor motif enrichment in each cluster.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The application of CoBRA to these experiments demonstrate the broad capabilities of the workflow in analyzing ChIP-seq or ATAC-seq experiments. While other workflows used to analyze ChIP or ATAC experiments exist, they lack some of the features present in CoBRA (table 1). Additionally, the highly modular Snakemake framework allows for rapid integration of new approaches or replacement of existing tools. Modules can be added simply by adding a new Snakemake "rule" and adding a flag in the config file (Supplementary Figure 2a-c) to turn the analysis on. Further, CoBRA's "rules" can be composed of tools written in R, Python, or shell script. The framework allows for great flexibility because each module can be evaluated in its own environment using different tools (e.g. Python 2.7 and Python 3 based software).

The methods for installing, deploying, and using CoBRA along with a detailed tutorial are provided in the documentation available online(https://cfce-cobra.readthedocs.io/). The workflow was designed to work with Docker, which allows the user to automatically download all required software dependencies, eliminating the possibility of conflicting dependencies. This makes CoBRA easy for those with limited computational training to install and run the workflow. Furthermore, the user does not need to prepare any reference files, as CoBRA automatically downloads all needed reference files. As a result, CoBRA is portable, reproducible and easy to deploy.

Pipelines Features	**CoBRA**	**Diffbind**	**HMCan-diff**	**ChIPcomp**	**Deeptools**	**esATAC**	**OPENANNO**
Sample-sample Correlations	✓		✓		✓	✓	
Sample-feature Clustering	✓	✓			✓		
PCA analysis	✓	✓					
Normalize based on Sequencing Depth	✓	✓					
CNV Correction for Differential Peak Calling	✓		✓				
Motif Analysis	✓					✓	

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pathway Analysis	✓					✓	
Package Easy Update	✓	✓		✓		✓	✓
Easy Support New Species	✓						
Docker Containerized	✓						
Annotate Peak Regions with Public ChIP-seq database	✓						✓
Step-by-Step Tutorial with Multiple Case Studies	✓	✓			✓	✓	

Table 1. A comparison of the features of CoBRA with other available pipelines.

In summary we have developed a new pipeline, CoBRA (Containerized Bioinformatics workflow for Reproducible ChIP/ATAC-seq Analysis), that is fast, efficient, portable, customizable and reproducible. The workflow builds upon the ongoing effort to make computational research reproducible using defined workflows running inside Docker containers. CoBRA allows users of varying levels of technical skill to quickly process and analyze new data from ChIP-seq and ATAC-seq experiments. It is the authors' hope that CoBRA can be a starting point for others to build upon and improve CoBRA as a tool and extend its ability to analyze the cistrome.

Availability of data and software

The dataset(s) supporting the conclusions of this article are all publicly available in the NCBI Sequence Read Archive as referenced in the text.

The software described in this article is publicly available online.

Project name: CoBRA.

Project home page: https://bitbucket.org/cfce/cobra, https://cfce-cobra.readthedocs.io

Archived version: publication.

Operating system(s): UNIX; MacOS.

Programming language: multiple.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Other requirements: Docker, wget, git, miniconda3.

License: GNU GPL.

Any restrictions to use by non-academics: N/A.

CRediT author statement

Xintao Qiu: Methodology, Software, validation, Writing - Original Draft
Avery S. Feit: Methodology, Software, validation, Writing - Original Draft
Ariel Feiglin: Methodology, Software, validation
Yingtian Xie: Validation, Data Curation
Nikolas Kesten: Validation
Len Taing: Software, Validation
Joseph Perkins: Software
Ningxuan Zhou: Validation, Investigation
Shengqing Gu: Validation, Investigation
Yihao Li: Validation, Investigation
Paloma Cejas: Validation, Investigation
Rinath Jeselsohn: Resources, Validation
Myles Brown: Conceptualization, Supervision, Funding
X. Shirley Liu: Supervision, Project Administration
Henry W. Long: Supervision, Funding, Conceptualization, Writing - Review & Editing

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgements

H.W.L. and M.B. acknowledge funding from the National Institutes of Health (USA) grant 2PO1CA163227.

Authors' ORCID IDs

Xintao Qiu 0000-0002-8560-7017
Paloma Cejas 0000-0002-8417-4811
Rinath Jeselsohn 0000-0001-7996-7529
Myles Brown 0000-0002-8213-1658
X. Shirley Liu 0000-0001-8588-1182
Henry W. Long 0000-0001-6849-6629

References

1. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

411 Nature. 2012;481:389–93.

412 2. Ashoor H, Hérault A, Kamoun A, Radvanyi F, Bajic VB, Barillot E, et al. HMCan: a
413 method for detecting chromatin modifications in cancer samples using ChIP-seq data.
414 Bioinformatics. 2013;29:2979–86.

415 3. Allhoff M, Seré K, Chauvistré H, Lin Q, Zenke M, Costa IG. Detecting differential peaks
416 in ChIP-seq signals with ODIN. Bioinformatics. 2014;30:3467–75.

417 4. Shen L, Shao N-Y, Liu X, Maze I, Feng J, Nestler EJ. diffReps: detecting differential
418 chromatin modification sites from ChIP-seq data with biological replicates. PLoS One.
419 2013;8:e65598.

420 5. Shao Z, Zhang Y, Yuan G-C, Orkin SH, Waxman DJ. MAnorm: a robust model for
421 quantitative comparison of ChIP-Seq data sets. Genome Biol. 2012;13:R16.

422 6. Liang K, Keles S. Detecting differential binding of transcription factors with ChIP-seq.
423 Bioinformatics. 2012;28:121–2.

424 7. Wei Z, Zhang W, Fang H, Li Y, Wang X. esATAC: an easy-to-use systematic pipeline for
425 ATAC-seq data analysis. Bioinformatics. 2018;34:2664–5.

426 8. Chen S, Wang Y, Jiang R. OPENANNO: annotating genomic regions with chromatin
427 accessibility. doi:10.1101/596627.

428 9. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded
429 datasets and new tools for gene regulatory analysis. Nucleic Acids Res. 2019;47:D729–35.

430 10. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine.
431 Bioinformatics. 2012;28:2520–2.

432 11. Qin Q, Mei S, Wu Q, Sun H, Li L, Taing L, et al. ChiLin: a comprehensive ChIP-seq and
433 DNase-seq quality control and analysis pipeline. BMC Bioinformatics. 2016;17:404.

434 12. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a
435 next generation web server for deep-sequencing data analysis. Nucleic Acids Res.
436 2016;44:W160–5.

437 13. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations
438 of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for
439 Macrophage and B Cell Identities. Molecular Cell. 2010;38:576–89.
440 doi:10.1016/j.molcel.2010.05.004.

441 14. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a data
442 portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids
443 Res. 2017;45:D658–62.

444 15. Wu D-Y, Bittencourt D, Stallcup MR, Siegmund KD. Identifying differential
445 transcription factor binding in ChIP-seq. Front Genet. 2015;6:169.

446 16. Tak YG, Hung Y, Yao L, Grimmer MR, Do A, Bhakta MS, et al. Effects on the
447 transcriptome upon deletion of a distal element cannot be predicted by the size of the
448 H3K27Ac peak in human cells. Nucleic Acids Res. 2016;44:4123–33.

449 17. Piunti A, Hashizume R, Morgan MA, Bartom ET, Horbinski CM, Marshall SA, et al.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

450 Therapeutic targeting of polycomb and BET bromodomain proteins in diffuse intrinsic
451 pontine gliomas. Nat Med. 2017;23:493–500.

452 18. Maurano MT, Wang H, John S, Shafer A, Canfield T, Lee K, et al. Role of DNA
453 Methylation in Modulating Transcription Factor Occupancy. Cell Rep. 2015;12:1184–95.

454 19. McCleland ML, Mesh K, Lorenzana E, Chopra VS, Segal E, Watanabe C, et al. CCAT1
455 is an enhancer-templated RNA that predicts BET sensitivity in colorectal cancer. J Clin
456 Invest. 2016;126:639–52.

457 20. Rahnamoun H, Lee J, Sun Z, Lu H, Ramsey KM, Komives EA, et al. RNAs interact with
458 BRD4 to promote enhanced chromatin engagement and transcription activation. Nat Struct
459 Mol Biol. 2018;25:687–97.

460 21. Taieb J, Le Malicot K, Shi Q, Penault-Llorca F, Bouché O, Tabernero J, et al. Prognostic
461 Value of BRAF and KRAS Mutations in MSI and MSS Stage III Colon Cancer. J Natl
462 Cancer Inst. 2017;109. doi:10.1093/jnci/djw272.

463 22. Kuilman T, Velds A, Kemper K, Ranzani M, Bombardelli L, Hoogstraat M, et al.
464 CopywriteR: DNA copy number detection from off-target sequence data. Genome Biol.
465 2015;16:49.

466 23. Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, et al. DNA
467 copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome
468 sequencing with identification and exclusion of problematic regions in the genome assembly.
469 Genome Res. 2014;24:2022–32.

470 24. Jung B, Staudacher JJ, Beauchamp D. Transforming Growth Factor β Superfamily
471 Signaling in Development of Colorectal Cancer. Gastroenterology. 2017;152:36–52.
472 doi:10.1053/j.gastro.2016.10.015.

473 25. Koi M, Tseng-Rogenski SS, Carethers JM. Inflammation-associated microsatellite
474 alterations: Mechanisms and significance in the prognosis of patients with colorectal cancer.
475 World J Gastrointest Oncol. 2018;10:1–14.

476 26. Ramirez RN, El-Ali NC, Mager MA, Wyman D, Conesa A, Mortazavi A. Dynamic Gene
477 Regulatory Networks of Human Myeloid Differentiation. Cell Syst. 2017;4:416–29.e3.

478 27. Cornwell M, Vangala M, Taing L, Herbert Z, Köster J, Li B, et al. VIPER: Visualization
479 Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis.
480 BMC Bioinformatics. 2018;19:135.

481
482
483

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1. Overview of CoBRA.
a) Biological motivation of CoBRA. Comparisons between ChIP and ATAC-seq peaks in well-designed experiments can provide insight into differences in protein occupancy, histone marks and chromatin accessibility.
b) Overview of the workflow performed by CoBRA. Read counts are quantified and normalized for sequencing depth and CNV for clustering and differential peak calling analysis. The result of differential peak calling is used downstream for motif enrichment, GSEA, CistroneDB toolkit, and BETA analysis.

Figure 2. Example of unsupervised and supervised analysis of differential GR binding in A549 cells.
a) Sample-Sample heatmap depicting clustering and correlation between A549 cells treated with varying concentrations of dexamethasone.
b) Visualization the differences in the GR binding between the 0.5 and 50nM samples, plotted using mean of the peak intensities versus log2(fold change). This illustrates the change in the inferred differential GR binding profile following normalization using scaling factor determined by total reads in peaks (top) and sequencing depth (bottom).
c) Deeptools heatmap illustrating differential peaks called by DESeq2 using default scaling factor by total reads in peaks (left) or using scaling factor determined by sequencing depth (right). A group of peaks at the bottom of the left figure exhibit similar binding intensity, however, they are considered downregulated in 50nM treatment in the default DEseq2 result.
d) Cistrome Toolkit result illustrating publicly available ChIP seq datasets ranked by binding profile similarity to gained GR binding sites with dexamethasone treatment.

Figure 3. Identification of differential sites correcting for copy number.
a) Copy number distribution for an MSS sample on Chromosome 8 (top). Distribution of differentially called peaks with (middle) and without (bottom) CNV adjustment between MSS and MSI cell lines.
b) Significant differential peaks in 8Q prior to CNV correction are highlighted. X axis is the log2 fold change and y axis is -log10 of the adjusted p-value. Left side is without CNV correction, and the right side is with CNV correction.
c) Enrichment plot for NIKOLSKY_BREAST_CANCER_8Q12_Q22_AMPLICON gene set without (left side) and with (right side) CNV adjustment.
d) Enrichment of Hallmarks gene sets after CNV correction based on the differential peak ranking comparing MSS with MSI.

Figure 4. Analysis of ATAC-seq from HL-60 promyelocytes differentiating into macrophages with CoBRA.
a) PCA plot depicting how samples cluster along the first two principal axes.
b) Sample-Feature heatmap created by CoBRA which depicts sample clustering on the horizontal axis and chromatin accessibility clustering on the vertical axis. Cluster 1,2, and 3 represent sites open at early, middle and late differentiation stages respectively.
c) Motifs enriched in early, middle, and late stage differentiation identified by CoBRA.
d) Genes differentially expressed during macrophage differentiation (120hr over 0hr). Those genes that also have nearby differential chromatin changes during differentiation are highlighted.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 1.
a) PCA plot depicting similarity between dexamethasone treated samples.
b) Clustering result of the motif enrichment for sites up with 50nM treatment over 0.5nM.

Supplementary Figure 2.
a) Example of parameter setup in config.yaml file.
b) Example of parameter setup in config.yaml file.
c) Example of sample path setup in config.yaml file.

Supplementary Figure 3. File structure of CoBRA input and output.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

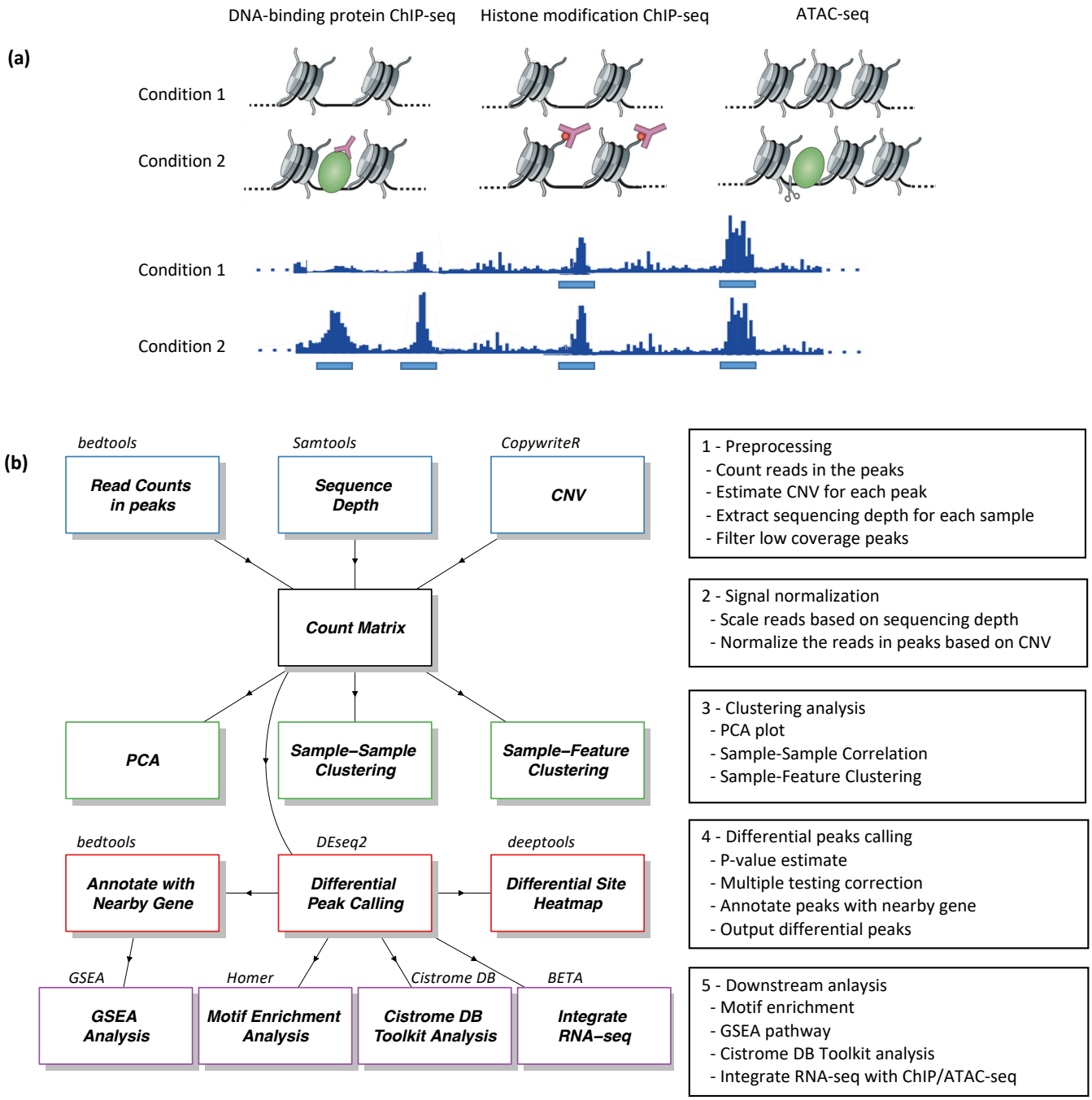


Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

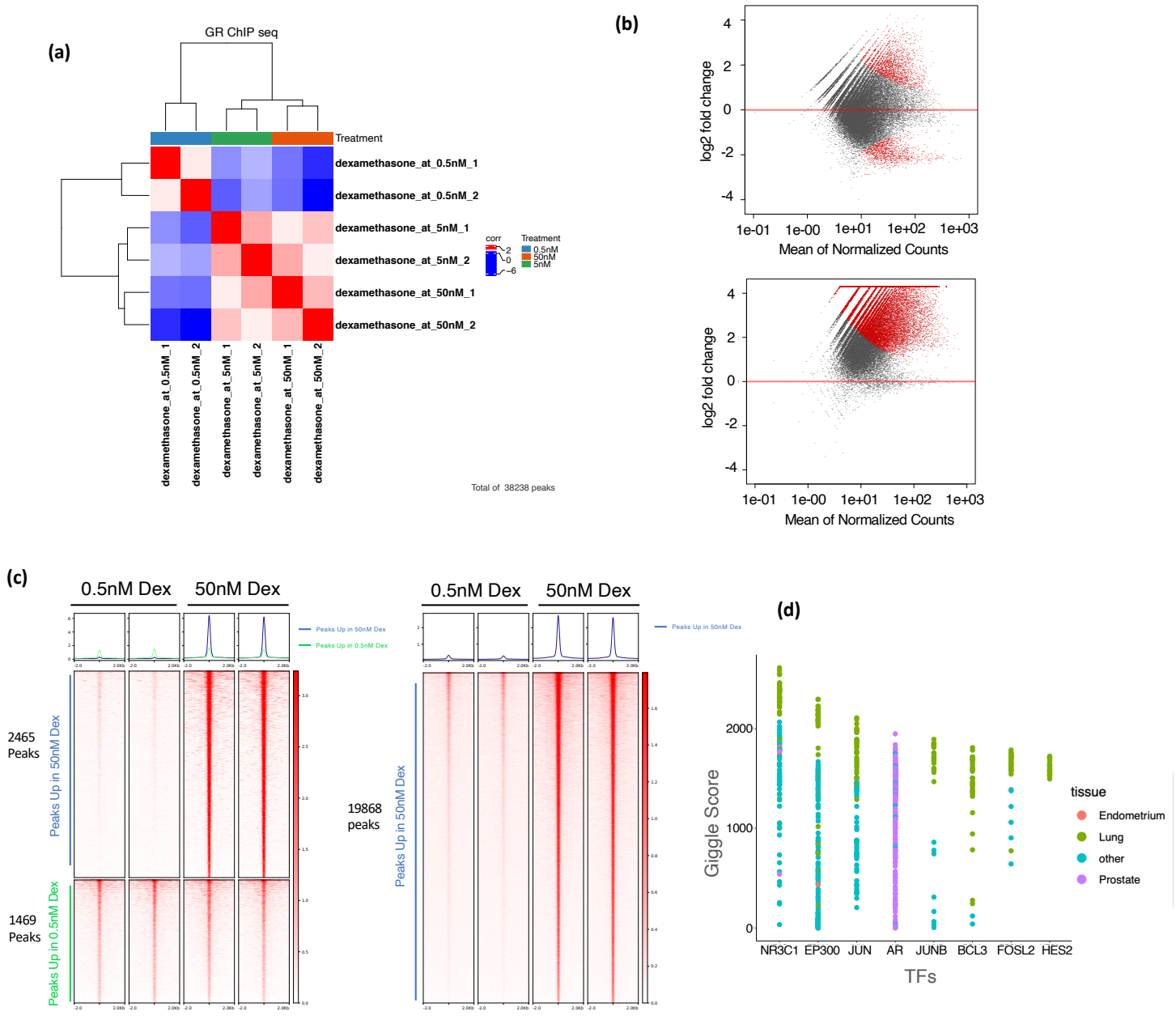


Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

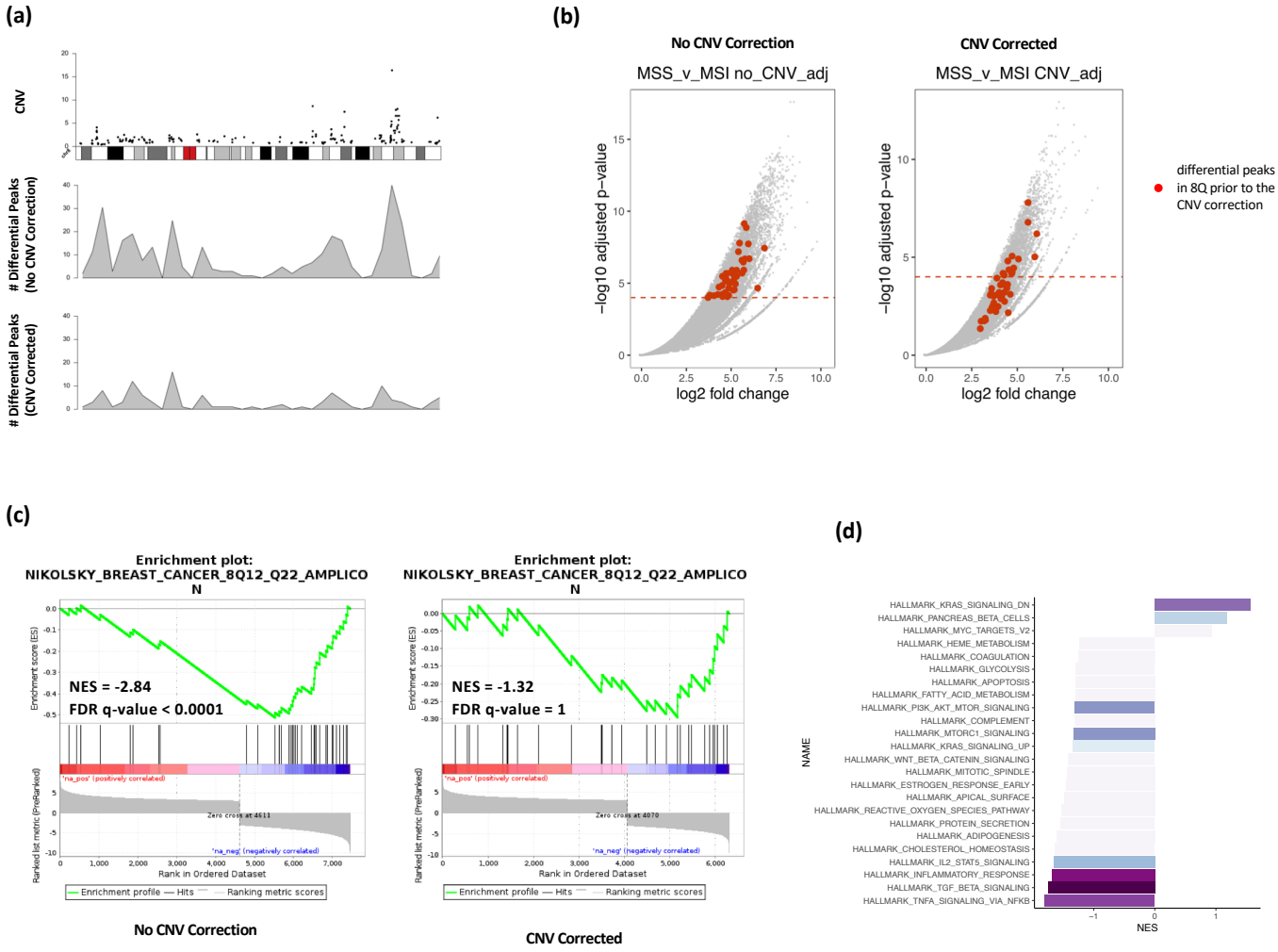


Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

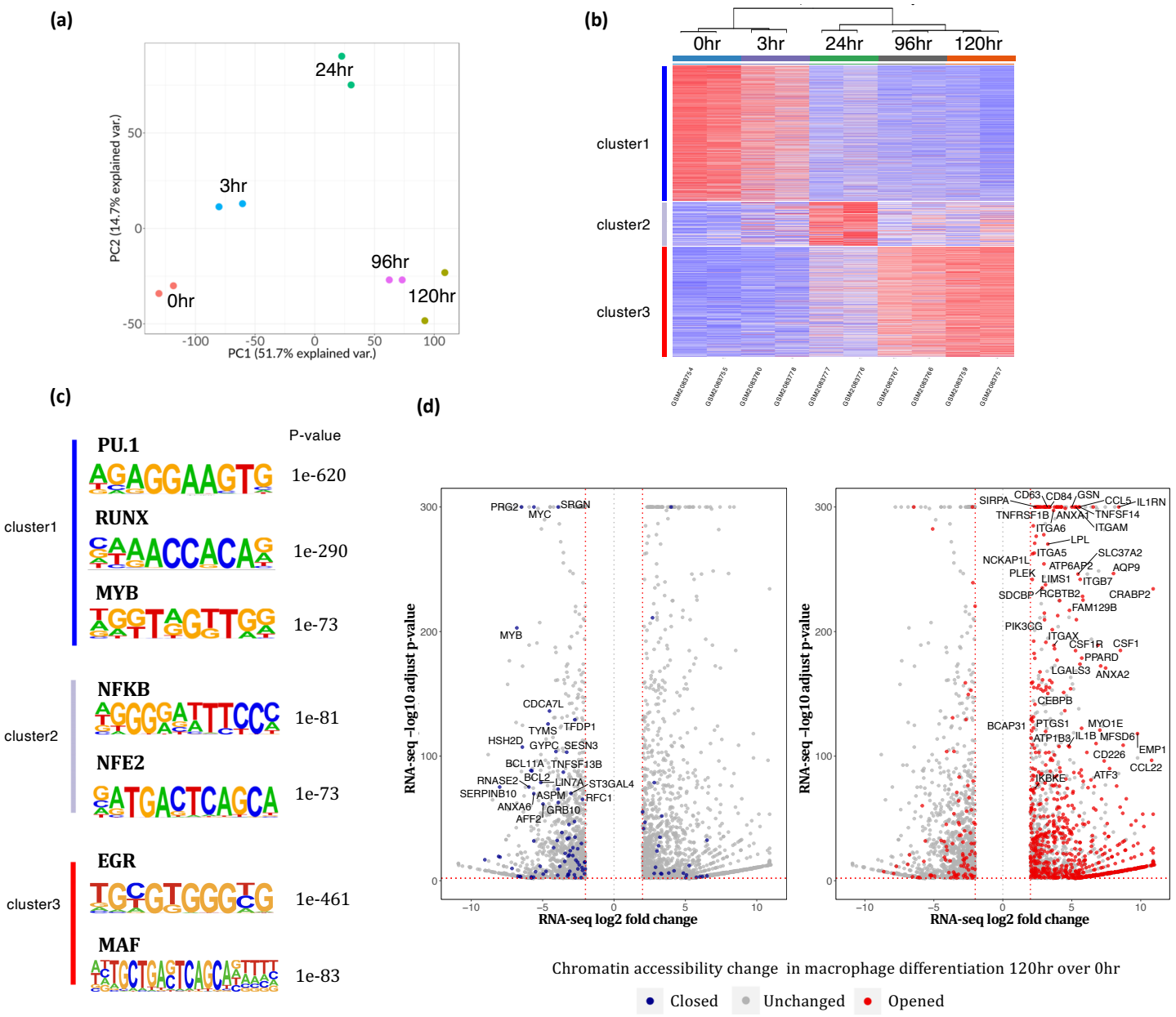
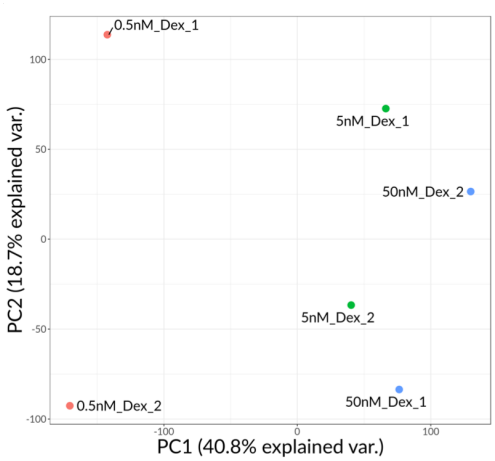


Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(a)



(b)

Cluster	Motif	Name	PValue	log(PValue)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif
1		GR(NR)/JR3/RAW264.7-GRE-ChIP-Seq(Unpublished)/Homer	1e-2378	-5477.435464	3019.0	44.27%	1521.0	3.63%
		GR(NR)/JR3/A549-GR-ChIP-Seq(GSE32465)/Homer	1e-2142	-4932.236126	2349.0	34.44%	824.5	1.97%
		AR(NR)/LNCAP-AR-ChIP-Seq(GSE27824)/Homer	1e-2031	-4677.025589	2876.0	42.17%	1742.3	4.16%
		PGR(NR)/EndoStromal-PGR-ChIP-Seq(GSE69539)/Homer	1e-1835	-4225.926702	2602.0	38.15%	1540.2	3.67%
		PR(NR)/T47D-PR-ChIP-Seq(GSE31130)/Homer	1e-1225	-2820.686095	5632.0	82.58%	16086.6	38.37%
2		Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1e-1129	-2599.846466	2798.0	41.03%	3627.8	8.65%
		Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	1e-1093	-2518.963523	2590.0	37.98%	3139.8	7.49%
		JunB(bZIP)/Dendritic Cells-JunB-ChIP-Seq(GSE36099)/Homer	1e-1040	-2396.976282	2728.0	40.00%	3729.9	8.90%
		Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-1037	-2389.268204	2975.0	43.62%	4536.5	10.82%
		BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	1e-1027	-2365.228649	2920.0	42.82%	4396.5	10.49%
		Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	1e-1015	-2338.547207	2128.0	31.20%	2151.7	5.13%
		AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-950	-2188.543174	3045.0	44.65%	5166.9	12.32%
		Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	1e-918	-2115.098968	1762.0	25.84%	1544.5	3.68%
		Bach2(bZIP)/OCILy7-Bach2-ChIP-Seq(GSE44420)/Homer	1e-493	-1135.631294	1172.0	17.18%	1285.2	3.07%
		Bach1(bZIP)/K562-Bach1-ChIP-Seq(GSE31477)/Homer	1e-191	-440.281436	392.0	5.75%	343.5	0.82%
		NF-E2(bZIP)/K562-NFE2-ChIP-Seq(GSE31477)/Homer	1e-189	-435.587020	427.0	6.26%	422.3	1.01%
		Nrf2(bZIP)/Lymphoblast-Nrf2-ChIP-Seq(GSE37589)/Homer	1e-168	-387.840346	362.0	5.31%	336.6	0.80%
		NFE2L2(bZIP)/HepG2-NFE2L2-ChIP-Seq(Encode)/Homer	1e-141	-325.917248	322.0	4.72%	320.6	0.76%
		MafA(bZIP)/Islet-MafA-ChIP-Seq(GSE30298)/Homer	1e-76	-175.974922	1475.0	21.63%	5620.6	13.41%

Supplementary Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(a) From config.yaml input file into Cobra

```yaml
#Project Name
#Use in pca, sample-sample, sample-feature plot. Please use "_" to seperate different words
project: ChIP_seq

#enhancer option enhancer/promoter/all
enhancer: all

#Location of metasheet
metasheet: metasheet.csv
ref: "scripts/ref.yaml"


#Assembly is needed when seperate enhancer/promoter, motif finding, nearyby gene
assembly: hg19


#At least mini_num_sample should have RPKM > rpkm_threshold
rpkm_threshold: 1
mini_num_sample: 0


#Scale method for the nomalize counts among samples
#z- z-score
#q- quantile-normalize
#l- log-transform
scale: q

#Fliter metric in feature selection
#sd- Standard deviation
#cov- Coefficient of Variation
#av- mean
filter-opt: cov

#top percent cutoff
filter-percent: 100

#limited of peaks to use for plot
SSpeaks: 20000000
SFpeaks: 20000000

#number of k-means clustering in sample-feature plot
num_kmeans_clust: 6
```

Supplementary Figure 2a

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(a) From config.yaml input file into Cobra

```yaml
#correlation method for sample-sample, sample-feature plot
# "person" or "spearman"
cor_method: 'pearson'

#distance method for sample-sample, sample-feature plot
# "euclidean", "manhattan", "canberra", "binary", "maximum" or "minkowski"
dis_method: 'euclidean'

#DEseq_cut_off - Padj/LG2FC
Padj: 0.05
LG2FC: 0

#DEseq normalize method
#def - normlize by default setting of DEseq2
#depth - normlize by the sequence depth of each sample
nor_method: 'depth'

#Motif analysis - true/false
motif: 'false'

#BAM files sorted? true/false
bam_sort: 'true'

#CNV correction? true/false
CNV_correction: 'false'

#unchanged heatmap
unchanged_heatmap: 'false'

#fastq as input
fastq_in: 'true'

#number of threads used in bwa mem
thread: 8
```

Supplementary Figure 2b

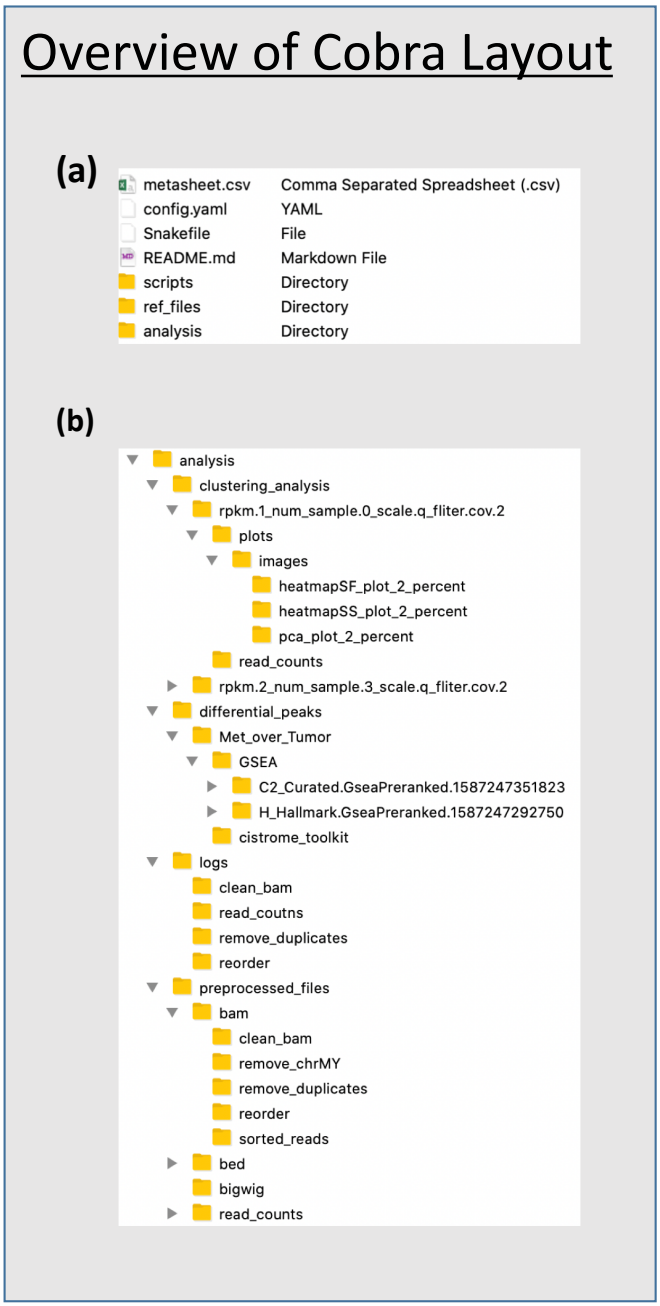
bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(a) From config.yaml input file into Cobra

```yaml
# sample names, e.g. "sample01" "sample02" can be any abitrary string
# HOWEVER, these names must match what is in metasheet.csv
# FOR each sample, define the path to the fastq file

fastq:
   sample1:
      - ./XX1_R1.fastq.gz
   sample2:
      - ./XX2_R1.fastq.gz
   input:
      - ./XX_input_R1.fastq.gz


# bed, bam and bigwig is not needed when fastq_in is true
bed:
   sample1: ./XX1.bed
   sample2: ./XX2.bed

bam:
   sample1: ./XX1.bam
   sample2: ./XX2.bam


bigwig:
   sample1: ./XX1.bw
   sample2: ./XX2.bw

# tab-separated cnv files
cnv:
   sample1: ./XX1.igv
   sample2: ./XX2.igv
```

Supplementary Figure 2c

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.367409; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 3