Methods For Final Project

ChIP-seq

Raw data were cleaned by fastp[1] with default parameter. Clean reads were aligned to GRCh38 with Bowtie2[2]. PCR replicates are removed by SAMtools[3]. MACS2[4] was used to call peaks with parameter: "-- nomodel", "--extsize 200" and "--pvalue 0.001".Motif analysis were used HOMER[5] with parameter: "size 200", "-len 8,10,12,14". Peaks were annotated with ChIPseeker[6] R package. Then we classified each peak as either TSS-proximal or TSS-distal depending on its distance (< or > +/-3 kb, respectively) from TSS. For H3K4me3 ChIP-seq, the distance is defined as +/-1.5kb from TSS. Differential analysis with MAnorm[7] with default parameter.

All the *bam* files were converted to *bigwig* files by deepTools[8] "bamCoverage" with "--normalizeUsing RPKM", "--effectiveGenomeSize 2913022398". The visualization of *bigwig* files were using IGV[9] software.

RNA-seq

Raw data were cleaned by Trimmomatic[10] with "HEADCROP:15". For PDX, reads were aligned to GRCm39 with HISAT2[11]. Host reads were removed by ngs-disambiguate[12]. For MCF10DCIS cells, reads were aligned to GRCh38 with STAR[13]. Expression counts were estimated by featureCounts[14]. Exons were reads were summarized as annotated in Homo_sapiens.GRCh38.106.gtf with default options. EdgeR[15] were used for DEG analysis. Prior to normalization using the Trimmed Mean of M (TMM) method, genes whose

expression was lower than 0.1 Count Per Million (0.5 for the MCF10DCIS) in more than three samples were filtered out. DEGs are filtered with following criteria: log2 fold-change ≥ |0.6|, false discovery rate (FDR) < 0.05 and expression > 0.5 RPKM (1 RPKM in the MCF10DCIS cell line) in all sample in at least one condition. Heatmap was generated by pheatmap R package. The Gene Ontology and KEGG enrichment analysis were generated by clusterProfiler[16]. GSEA analysis was also generated by clusterProfiler.

TCGA data were download by TCGAbiolinks[17] R package. Kaplan–Meier analysis were generated by survminer and survival R packages. Hallmark gene set was downloaded from https://www.gsea-msigdb.org/gsea/msigdb/.

Reference

1. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England) 34, i884-i890 (2018).
2. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-359 (2012).
3. Danecek, P. et al. Twelve years of SAMtools and BCFtools. GigaScience 10(2021).
4. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). Genome Biology 9, R137 (2008).
5. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell 38, 576-589 (2010).
6. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics (Oxford, England) 31, 2382-2383 (2015).

7. Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S.H. & Waxman, D.J. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biology 13, R16 (2012).
8. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research 44, W160-W165 (2016).
9. Robinson, J.T. et al. Integrative genomics viewer. Nature Biotechnology 29, 24-26 (2011).
10. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) 30, 2114-2120 (2014).
11. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology 37, 907-915 (2019).
12. Ahdesmäki, M.J., Gray, S.R., Johnson, J.H. & Lai, Z. Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples. F1000Research 5, 2741 (2016).
13. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) 29, 15-21 (2013).
14. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England) 30, 923-930 (2014).
15. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) 26, 139-140 (2010).
16. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics : a Journal of Integrative Biology 16, 284-287 (2012).
17. Colaprico, A. et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research 44, e71 (2016).