План выполнения проекта:

“Обработка данных РНК-секвенирования с
целью изучения функций белка Pcbp1”

Студент: Потапенко Евгений
Научные руководители: Оксана Станевич, Евгений Бакин

Введение:

В лаборатории молекулярной биологии стволовых клеток института цитологии РАН были получены ЭСК с нокаутом по гену *Pcbp1. Эмбрионы образующиеся из таких клеток гибнут на 8ой день развития. Это может свидетельствовать об участии Pcbp1 в переходе из исходного плюрипотентного состояния в позднее праймированное. Поэтому материалом для исследования профилей экспрессии явились ЭСК мыши в исходном плюрипотентном состоянии(Naive) и раннем праймированном состоянии (EpiLC).*

Для исследования были подготовлены библиотеки кДНК по 3 реплики каждого из 4 типов клеток: Naive и EpiLC клетки дикого типа и они же с нокаутом по гену Pcbp1, было произведено их секвенирование и получены необходимые для исследования наборы ридов (Таблица 1).

Sample	Work_number	Number of reads
S1 Native	1	42 380 298
S2 Native	2	41 535 143
S3 Native	3	50 376 749
P1-1 Native	4	45 108 795
P1-3 Native	5	47 062 337
P1-22Native	6	43 063 892
S1 Epi LC	7	50 987 437
S2 Epi LC	8	35 868 628
S3 Epi LC	9	39 986 664
P1-1 Epi LC	10	55 765 949
P1-3Epi LC	11	47 364 318
P1-22Epi LC	12	46 110 750

Таблица 1. Таблица используемых образцов с присвоенным номером для анализа и числом полученных ридов.

Цель проекта: *Изучение профиля экспрессии генов эмбриональных стволовых клеток мыши при нокаутировании гена Pcbp1.*

***Выполнение работы**:*

Данные полученные после секвенирования были загружены на веб-платформу Galaxy, был использован общедоступный сервер на **usegalaxy.org** для анализа данных. Были проанализированы полученные наборы ридов при помощи программы **FastQC** Galaxy Version 0.72+galaxy1 [1]. Затем на основе полученных данных был произведен их процессинг при помощи **Trimmomatic** Galaxy Version 0.38.0 [2] с параметрами -SE SLIDINGWINDOW:4:20 HEADCROP:13. Затем выполнено выравнивание на референсный геном мыши (сборка GRCm38/mm10) при помощи программы **RNA STAR** Galaxy Version 2.6.0b-1 [3] с аргументами -- outFilterMultimapNmax 1, -- outReadsUnmapped Fastx, -- outSAMtype BAM SortedByCoordinate, -- twopassMode Basic. Затем проанализированы результаты при помощи программы **MultiQC** Galaxy Version 1.7 [4]. Для более детальной оценки полученных выравниваний были использованы программы из пакета RseQC: **ReadDistribution** Galaxy Version 2.6.4.1, **Gene Body Coverage(BAM)** Galaxy Version 2.6.4.3 [5]. Результаты объединены при помощи MultiQC в один отчет. Затем был произведен подсчет количества ридов ассоциированных с каждым геном при помощи утилиты **featureCounts** Galaxy Version 1.6.4+galaxy1 [6]. Полученая таблица была обработана при помощи DEseq2 [7]. В работе также были использованы такие пакеты как: limma [8], edger2 [9], gage [10], pathview[11], clusterProfiler[12] и др.

В результате получены таблицы дифференциально экспрессированных генов для клеток Naive и EpiLC нокаутных по гену *pcbp1*. Также получены схемы сигнальных путей наиболее затронутых нокаутом данного гена.

В дальнейшем планируется разбор и детальный анализ генов вовлеченных в сигнальные пути ответственные за поддержание плюрипотентности эмбриональных стволовых клеток мыши.

References

[1] Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[2] Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

[3] Dobin, Alexander and Davis, Carrie A. and Schlesinger, Felix and Drenkow, Jorg and Zaleski, Chris and Jha, Sonali and Batut, Philippe and Chaisson, Mark and Gingeras, Thomas R. (2012). STAR: ultrafast universal RNA-seq aligner. In *Bioinformatics, 29 (1), pp. 15â 21.* [doi:10.1093/bioinformatics/bts635][Link]

[4] Ewels, Philip and Magnusson, MÃ¥ns and Lundin, Sverker and KÃ¤ller, Max (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. In *Bioinformatics, 32 (19), pp. 3047â 3048.* [doi:10.1093/bioinformatics/btw354] [Link]

[5] Wang, Liguo and Wang, Shengqin and Li, Wei (2012). RSeQC: quality control of RNA-seq experiments. In *Bioinformatics, 28 (16), pp. 2184â 2185.* [doi:10.1093/bioinformatics/bts356][Link]

[6] Liao, Y. and Smyth, G. K. and Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. In *Bioinformatics, 30 (7), pp. 923â 930.* [doi:10.1093/bioinformatics/btt656][Link]

[7] Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)

[8] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.

[9] Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140

[10] Luo, W., Friedman, M., Shedden K., Hankenson, K. and Woolf, P GAGE: Generally Applicable Gene Set Enrichment for Pathways Analysis. BMC Bioinformatics, 2009, 10:161

[11] Luo, W. and Brouwer C., Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics, 2013, 29(14): 1830-1831, doi: 10.1093/bioinformatics/btt285

[12] Guangchuang Yu, Li-Gen Wang, Yanyan Han and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology 2012, 16(5):284-287