Raw sequence to target gene prediction: An integrated inference pipeline for ChIP-seq and RNA-seq datasets

Nisar Wani[1,2,#] and Khalid Raza[1,*]

1. Department of Computer Science,Jamia Millia Islamia , New Delhi, India., Govt. Degree College Baramulla, University of Kashmir, J&K India.
{kraza}@jmi.ac.in

Abstract. Gene expression patterns determine the manner whereby organisms regulate various cellular processes and therefore their organ functions.These patterns do not emerge on their own, but as a result of diverse regulatory factors such as, DNA binding proteins known as transcription factors (TF), chromatin structure and various other environmental factors. TFs play a pivotal role in gene regulation by binding to different locations on the genome and influencing the expression of their target genes. Therefore, predicting target genes and their regulation becomes an important task for understanding mechanisms that control cellular processes governing both healthy and diseased cells.In this paper, we propose an integrated inference pipeline for predicting target genes and their regulatory effects for a specific TF using next-generation data analysis tools.

Keywords: NGS, pipeline, target gene , prediction, ChIP-seq,RNA-seq, regulatory potential

1 Introduction

Omics technologies are key drivers of the data revolution that has taken place in the life sciences domain from last few decades. These technologies enable unbiased investigation of biological systems at genomic scales. Using high throughput Next Generation Sequencing (NGS) methods, genome-wide data is collected from cells, tissues and model organisms (Raza & Ahmad, 2016). These data are key to investigate biological phenomena governing different cellular functions and also help biomedical researchers to better understand the disease etiologies which have not been previously explored. NGS protocols such as, ChIP-seq and RNA-seq are to generate datasets from where we can obtain genome-wide binding map of TFs and epigenetic signatures(Park, 2009; Furey, 2012) and can also measure the gene expression abundance within the cell for the whole genome (Costa, Angelini, De Feis, & Ciccodicola, 2010; Wang, Gerstein, & Snyder, 2009; Ozsolak & Milos, 2011).

Numerous efforts have been put forth to uncover the interplay between genomic datasets obtained from ChIP-seq and RNA-seq for gene regulation studies of individual TFs (Wang. et al., 2013) or mapping Transcription Regulatory

Networks as in (Wade, 2015). Revealing such interaction between these data has significant biomedical implications in various pathological states as well as in normal physiological processes (Yue et al., 2014).Therefore, there is a compelling need to integrate these data to predict the pattern of gene expression during cell differentiation (Kadaja et al., 2014) and development (Comes et al., 2013) and to study human diseases such as , cancer as outlined in (Portela & Esteller, 2010).

The aim of this study is to integrate genome-wide protein DNA interaction (ChIP-seq) and transcriptomic data (RNA-seq) using a multi-step bioinformatics pipeline to infer the gene targets of a TF which serve as building blocks of a transcriptional regulatory network.We have developed a Perl script that implements this multistage pipeline by integrating tools in the same order as depicted in Fig. (1). The choice of tools for each stage is a consequence of thorough literature study among the set of tools available in their respective domains.Our implementation is a partially automated system that requires supervision at the time of quality control of raw reads, but progresses smoothly onwards without any manual intervention to integrate the two datasets and generate TF-specific gene targets.

2 Related Literature

Software tools and methods exist that predict and analyze gene targets by processing ChIP-seq data. A distinguishable group of peak callers such as , CisGenome (H. Jiang, Wang, Dyer, & Wong, 2010), BayesPeak (Spyrou, Stark, Lynch, & Tavaré, 2009), Model-based Analysis of ChIP-seq (MACS) (Zhang et al., 2008),Peakseq (Rozowsky et al., 2009), SICER (Zang et al., 2009) are some of the widely used tools that identify TF-binding sites. These peak callers identify the target genes either by looking for peaks in promoter region or assign a proximal nearest gene in the vicinity of peaks. However, with most TFs ChIP-seq data having peaks in and around the promoter regions is very less. Also predicting targets using nearest peak is not always reliable.TIP (Cheng, Min, & Gerstein, 2011) is another tool that builds a probabilistic model to predict gene targets, but does not take into account gene expression data. Certain databases such as JASPAR (Sandelin, Alkema, Engström, Wasserman, & Lenhard, 2004), TRED (C. Jiang, Xuan, Zhao, & Zhang, 2007) etc. identify target genes for a selected set of TFs based on the motif analysis of the promoter regions using Position Weight Matrices (PWMs). A recent study (Essebier, Lamprecht, Piper, & Boden, 2017) combines multiple approaches to predict target genes.

On the contrary, some earlier studies used gene expression data for predicting target genes. (Qian et al., 2003) use Support Vector Machines (SVM) to discover relationships between the TFs and their targets; (Honkela et al., 2010) identify targets with time-series expression data by creating a linear activation model based on Gaussian process.

3 Materials and Methods

The proposed pipeline operates on raw NGS data, ChIP-seq & RNA-seq. After preprocessing the raw sequences it yields differential gene expression and peak information of genes from these data sets. Both these datasets are integrated to yield target genes for the ChIPred TF.A working description of the proposed pipeline is presented below.

3.1 Datasets

NGS data is primarily accessed from Sequence Read Archive (SRA) at NCBI (Leinonen. & Sugawara., 2010) and European Nucleotide Archive (ENA) at EBI (Leinonen et al., 2010). Raw sequences in the form of FASTQ files are freely available for download for a variety of cell types, diseases, treatments and conditions. Besides the public databases, a number of projects and consortia offer public access to their data repositories.For example, ENCODE(Consortium et al., 2004) is a publicly funded project that has generated large sets of data for a variety of cell lines, tissues and organs. Raw as well as pre-processed data can be accessed and freely downloaded from ENCODE data portal. Another publicly funded research project, The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) also provides datasets for a variety of cancer types. For the current study we have downloaded ChIP-seq data of MCF7 breast cancer cell line from ENCODE experiment *ENCFF580EKN* and RNA-Seq data of transcriptomic study *PRJNA312817* from Eurpean Nucleotide Archive. The RNA-seq experiment contains 30 samples of time course gene expression data from MCF7 cell line subjected to estrogen stimulation.

3.2 Pipeline Workflow

Molecular measurements within the NGS data exist in the form of millions of reads and are stored as FASTQ files. Information within the raw files is hardly of any value and needs extensive pre-processing before this data can be analyzed. The pre-processing task is a multi-step process and involves the application of a number of software tools. In this section, we present a detailed NGS pipeline that describes necessary steps from pre-processing of RNA-Seq and ChIP-seq data to target genes regulated by ChIPred TF.The pipeline has been implemented using a Perl script by integrating various NGS data processing & analysis tools.Fig. (1) is a graphical depiction of the proposed inference pipeline.

Quality Control

Almost all sequencing technologies produce their outputs in FSATQ files. FASTQ has emerged as the de facto file format for data exchange between various bioinformatics tools that handle NGS data. FASTQ (Cock, Fields, Goto, Heuer, & Rice, 2009) format is a simple extension to existing FASTA format; the

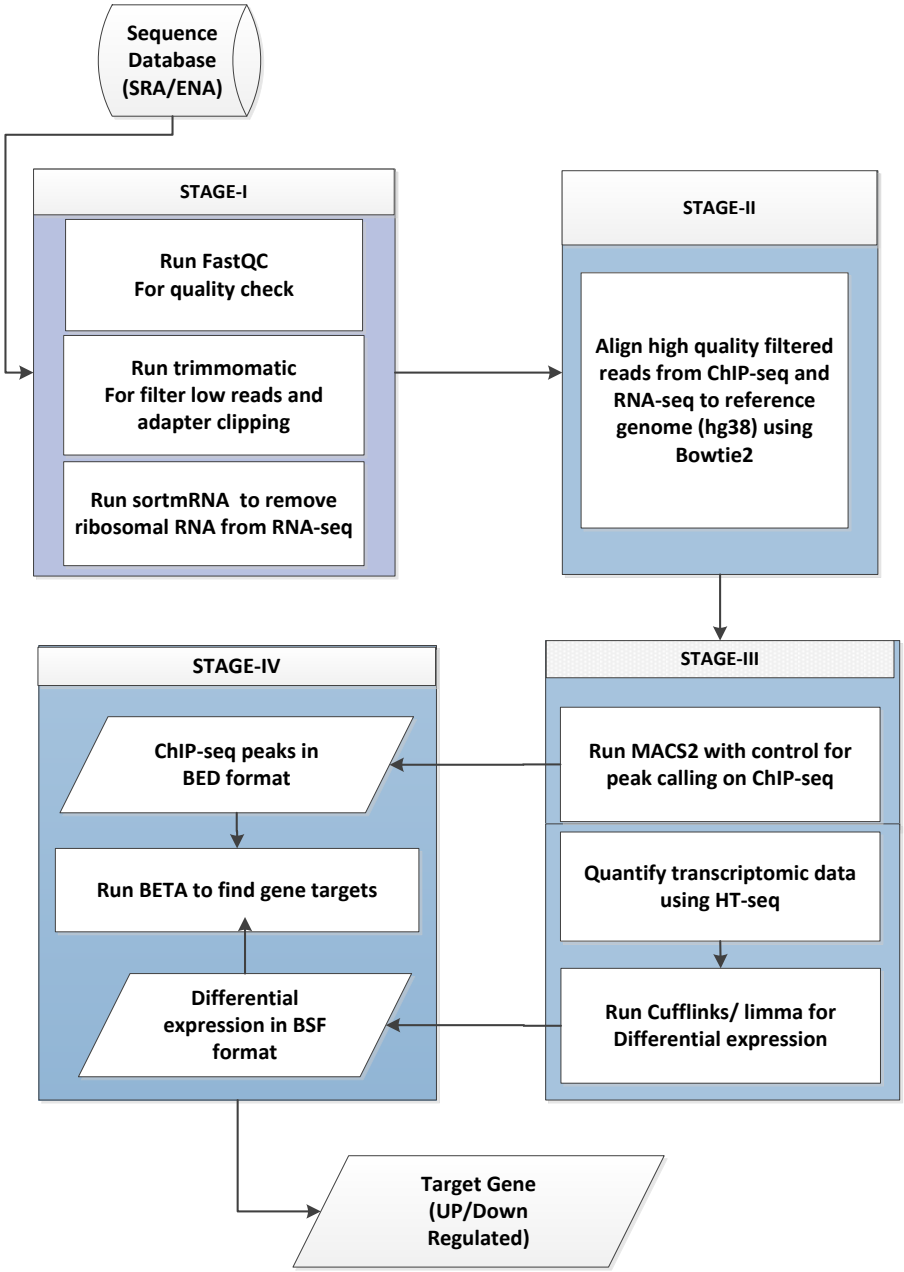


Fig. 1: Proposed inference pipeline for target gene prediction

files are plain ASCII text files with the ability to store both nucleotide sequences along with a corresponding quality score for each nucleotide call.

Base sequence qualities are usually interpreted in terms of Phred quality scores. Phred quality scores Q are defined as a property which is logarithmically linked to error probabilities P of called bases and can be computed as shown in equation(1).

$$Q = -10log_{10}P \tag{1}$$

Phred's error probabilities have been shown to be very accurate (Ewing, Hillier, Wendl, & C., 1998), e.g. if Phred assigns a quality score of 10 means that 1 out of 10 base calls is incorrect, a score of 20 depicts that 1 in 100 bases has been called incorrectly.Usually, a Phred score >= 20 is considered as acceptable read quality, otherwise read quality improvement is required. If read quality is not improved by trimming, filtering, and cropping, there may be some error during library preparation and sequencing. FASTQC (Andrews et al., 2010) is a Java based tool that is used to assess the quality of the reads produced by Next Generation DNA Sequencers. Low quality reads are excised from the FASTQ files to improve the quality of the reads. Various tools are available that can be used to trim bases with poor Phred scores i,e. Phred score less than 20.

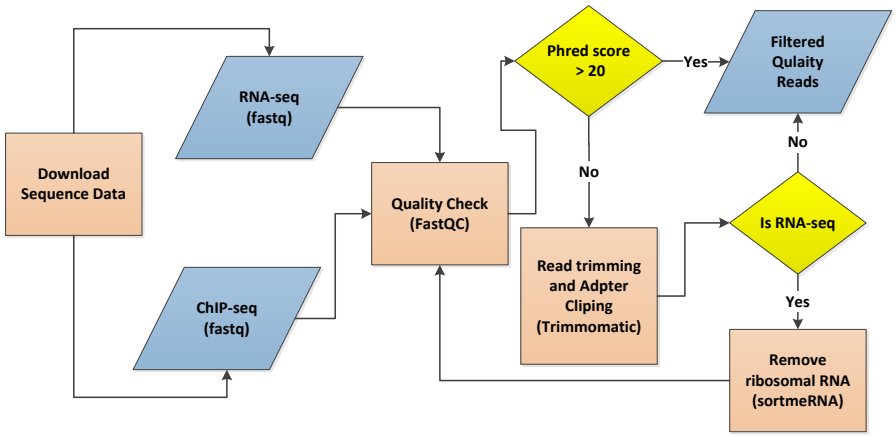


Fig. 2: Flowchart showing quality control check, adapter clipping and trimming steps of raw reads

Trimmomatic (Bolger, Lohse, & Usadel, 2014) is a Java based open source tool used for trimming illumina FASTQ data and removing adapters. Additionally RNA extracted using NGS does include non-coding RNA molecules besides the coding ones. These non-coding RNAs are usually the ribosomal RNA. For quantifying the gene expression patterns using RNA-seq, it is essential that these non-coding RNAs be filtered from the existing reads. SortMeRNA (Kopylova,

Noé, & Touzet, 2012) is a very efficient and accurate tool that is used to filter out the ribosomal RNA from the metatranscriptomic data.A diagrammatic flow of qulaity check and trimming is shown in Fig.(2).

Mapping sequence reads to a reference genome

With required read quality achieved after trimming and adapter clipping , the next step is to align the short reads to a reference sequence. The reference sequence is our case is human genome assembly hg38/hg19 but it can also be a reference transcriptome, or a de novo assembly incase a reference sequence is not available. There are numerous software tools that have been developed to map reads to a reference sequence. Besides the common goal of mapping these tools vary considerably from each other both in algorithmic implementation and speed. A brief account can be found in (Flicek & Birney, 2009)

In this pipeline we used Bowtie2 genome aligner (Langmead & Salzberg, 2012), because it is a memory-efficient and an ultrafast tool for aligning sequencing reads. Bowtie2 performs optimally with read lengths longer than 50bp or beyond 1000bp (e.g. mammalian) genomes. It builds an FM Index while mapping reads to keep its memory footprint small and for the human genome it is typically around 3.2 GB of RAM.

Expression quantification of RNA-Seq

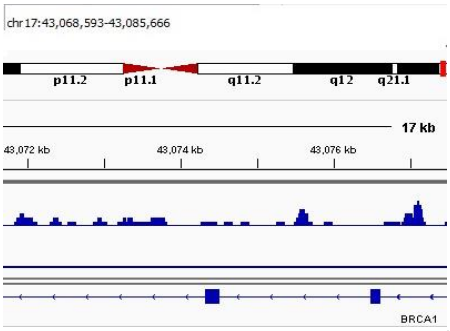
Although raw RNA-seq reads do not directly correspond to the gene expression, but we can infer the expression profiles from the sequence coverage or the mapping reads that map to a particular area of the transcriptome. A number of computational tools are available to quantify the gene expression profiles from RNA-seq data. Software tools, such as Cufflinks, HTSeq, IsoEM, and RSEM are freely available. A comparative study of these tools is presented in (Chandramohan, Wu, Phan, & Wang, 2013).

Despite clear advantages over microarrays, there are still certain sources of systematic variations that should be removed from RNA-seq data before performing any downstream analysis. These variations include between sample differences, such as sequencing depth and within sample differences e.g. gene length, GC content etc. In order to circumvent these issues and exploit the advantages offered by RNA-seq technology, the reads/ kilobase of transcript per million mapped reads (RPKM) normalizes a transcript read count by both its length and the total number of reads in the sample (Pepke, Wold, & Mortazavi, 2009). For data that has originated from the paired-end sequencing, a similar normalization metric called FPKM (fragments per kilobase of transcript per million mapped reads) is used. Both RPKM and FPKM use similar operations for normalizing single end and paired end reads (Conesa et al., 2016).

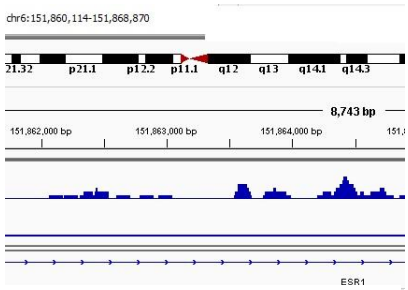
Counts per million (CPM) is another important metric provided by limma package to normalize gene expression data. Once the normalized expression estimates are available, we can obtain a differential expression of gene lists across the samples or conditions using limma voom (Ritchie et al., 2015) provided by R bioconductor.

Peak Calling Early pre-processing steps of ChIP-seq data resemble that of RNA-seq. Beginning with the quality check of raw reads, read trimming and adapter elimination, filtered high quality reads are then mapped to reference genome using Bowtie2 as described above.

In order to identify the genomic locations where the protein of Interest (POI) has attached itself to DNA sequences, the aligned reads are subjected to a process known as Peak Calling. Software tools that predict the binding sites where this protein has bound itself by identifying location within the genome with significant number of mapped reads (peaks) are called Peak Callers. A detailed description of various ChIP-seq peak callers is presented in (Pepke et al., 2009). Although a number of tools are available, but for this study have used the most efficient and open source tool called Model based Analysis of ChIP-seq (MACS) from (Zhang et al., 2008). Nowadays an upgraded version of MACS called MACS2 is commonly used for this purpose.A MACS2 algorithm does process aligned ChIP-seq bam files both with control and without control samples. Mapped reads are modelled as sequence tags (an integer count of genomic locations mappable under the chosen algorithm). Depending upon the type of



a) showing FOXA1 peaks for BRCA1 when visualised in IGV.



b) FOXA1 peaks for ESR1 gene in IGV.Peaks have been called using MACS2.

Fig. 3: FOXA1 peaks for BRCA1 and ESR1

protein being ChIPed, different types of peaks are observed when viewing the information in a genome browser, such as Integrated Genome Viewer (IGV). Most of the TFs act as point-source factors and result in narrow peaks, factors such as histone marks generate broader peaks, and proteins such as RNA Pol II can give rise to mixed peaks (both narrow as well as broad). Fig. 3(a) & Fig. 3(b) show the FOXA1 peaks for BRCA1 and ESR1 as viewed in IGV.

4 Results

The proposed pipeline first generates a list of differentially expressed genes (DEGs) from the normalized expression profiles, thereby identifying the gene

activity for both factor-bound and factor-unbound conditions. These DEGs are then integrated with the binding information from stage-III of the pipeline.Both these intermediate data sets are passed as input to stage-IV that employs Binding and expression target analysis (BETA) for prediction process; it calculates binding potential derived from the distance between transcription start site and the TF binding site, thereby modeling the manner in which the expression of genes is being influenced by TF binding sites. Using contributions from the individual TF binding sites, we obtain a cumulative score of overall regulatory potential (probability of a gene being regulated by a factor) of a gene. The percentage of up and down regulated genes is shown in Fig.(4)

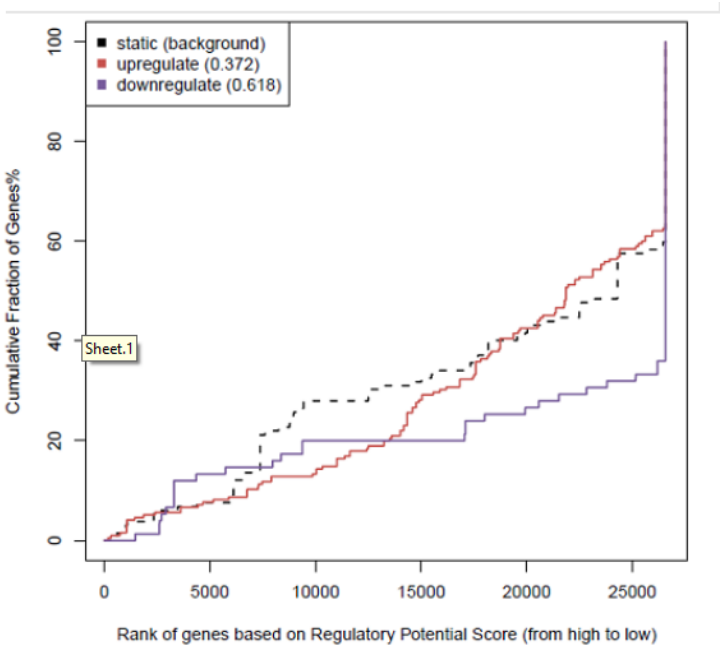


Fig. 4: FOXA1 predicted genes activation & repression function

During the process of target prediction each gene receives two ranks, one from the binding potential R_{bp} and other from differential expression R_{de}.Both these ranks are multiplied to obtain rank product $R_p = R_{bp}$ x R_{de}. Genes with more regulatory potential and more differential expression are more likely to be as real targets.Table 1 shows a list of up and down regulated genes.

Predicted targets from the inference pipeline results have been widely reported in literature.e.g, FOXA1 up regulated BRCA1 and down regulated ESR1, GATA3 and ZNF217 haven been reported in (Baran-Gale, Purvis, & Sethupathy,

Chromosome	RefSeqID	Rankproduct	Gene Symbol	Regulation
chr17	NM_007298	2.19E-04	BRCA1	up
chr12	NM_004064s	2.96E-04	CDKN1B	up
chr5	NM_001193376	6.16E-04	TERT	up
chr7	NM_000492	9.59E-04	CFTR	up
chr2	NM_001204109	9.77E-04	2 BCL2	up
chr10	NM_001002295	4.39E-04	GATA3	down
chr14	NM_138420	7.89E-04	AHNAK	down
chr20	NM_006526	2.72E-03	ZNF217	down
chr6	NM_001122741	5.79E-03	ESR1	down
chr1	NM_001878	5.79E-03	CRABP2	down

Table 1: **Target genes predicted by inference pipeline**

2016).Similarly evidence regarding the role of multiple loci on TERT gene are related to ER(-ve) breast cancer(Bojesen et al., 2013). Many of these predicted targets are well known prognostic biomarkers whose role has been established well in the scientific literature. Once we have a set of target genes, a further downstream analysis of these genes can be done by using gene ontologybased tools such as DAVID to map them with their corresponding biological functions.

5 Discussion & Conclusion

The availability and expansion of ChIP-seq and RNA-seq datasets is fuelling an exponential rise in the number of studies being conducted in the area of integrated computational analysis. The motive behind these research endeavors is to address basic questions about how multiple factor binding is related to transcriptional output within in vivo DNA. The proposed inference pipeline is used to decipher the regulatory relationship between TFs that bind to DNA and their corresponding target genes that they influence resulting in their activation/repression.From RNA-seq and ChIP-seq reads, the pipeline generates one file containing differential expression of genes and the other DNA-binding events in the form of peaks. Both these files are integrated in the final stage to yield targets for the TF/TFs whose peaks file was used.

The inference pipeline presented in this paper extracts target genes and hence the regulatory network for a specific TF that has been ChIPred. In case we are required to build a regulatory network for a set of TFs, we need to input new peak files for every new TF in a loop and record the target genes of this TF and

its regulatory influence in a separate file. In the current study, we considered only TFs and their influence on gene expression. However, a wider study can include multiple TFs, methylation data, histone marks and polymerase loading to improve the efficiency of the proposed pipeline.

Deciphering the transcriptional regulatory relationships and understanding the elements of regulatory mechanisms that control gene expression is a key research area of regulatory biology. Therefore computational integration of factor binding and other genome-wide data, such as gene expression will be sought after to extract functionally important connections of a working regulatory code.

Acknowledgment The author Nisar Wani acknowledges Teacher Fellowship received from University Grants Commission, Ministry of Human Resources Development, Govt. of India vide letter No. F.BNo. 27-(TF-45)/2015 under Faculty Development Programme.

References

Andrews, S., et al. (2010). *Fastqc: a quality control tool for high throughput sequence data.*

Baran-Gale, J., Purvis, J. E., & Sethupathy, P. (2016). An integrative transcriptomics approach identifies mir-503 as a candidate master regulator of the estrogen response in mcf-7 breast cancer cells. *RNA*, *22*(10), 1592–1603.

Bojesen, S. E., Pooley, K. A., Johnatty, S. E., Beesley, J., Michailidou, K., Tyrer, J. P., ... others (2013). Multiple independent variants at the tert locus are associated with telomere length and risks of breast and ovarian cancer. *Nature genetics*, *45*(4), 371–384.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Chandramohan, R., Wu, P.-Y., Phan, J. H., & Wang, M. D. (2013). Benchmarking rna-seq quantification tools. In *Engineering in medicine and biology society (embc), 2013 35th annual international conference of the ieee* (pp. 647–650).

Cheng, C., Min, R., & Gerstein, M. (2011). Tip: a probabilistic method for identifying transcription factor target genes from chip-seq binding profiles. *Bioinformatics*, *27*(23), 3221–3227.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, *38*(6), 1767–1771.

Comes, S., Gagliardi, M., Laprano, N., Fico, A., Cimmino, A., Palamidessi, A., ... others (2013). L-proline induces a mesenchymal-like invasive program in embryonic stem cells by remodeling h3k9 and h3k36 methylation. *Stem cell reports*, *1*(4), 307–321.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... others (2016). A survey of best practices for rna-seq data analysis. *Genome biology*, *17*(1), 13.

Consortium, E. P., et al. (2004). The encode (encyclopedia of dna elements) project. *Science*, *306*(5696), 636–640.

Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with rna-seq. *BioMed Research International*, *2010*.

Essebier, A., Lamprecht, M., Piper, M., & Boden, M. (2017). Bioinformatics approaches to predict target genes from transcription factor binding data. *Methods*.

Ewing, B., Hillier, L., Wendl, & C., P., Green. (1998). Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Biology*.

Flicek, P., & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature methods*, *6*, S6–S12.

Furey, T. S. (2012). Chip-seq and beyond: new and improved methodologies to detect and characterize protein-dna interactions. *Nature reviews. Genetics*, *13*(12), 840.

Honkela, A., Girardot, C., Gustafson, E. H., Liu, Y.-H., Furlong, E. E., Lawrence, N. D., & Rattray, M. (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences*, *107*(17), 7793–7798.

Jiang, C., Xuan, Z., Zhao, F., & Zhang, M. Q. (2007). Tred: a transcriptional regulatory element database, new entries and other development. *Nucleic acids research*, *35*(suppl_1), D137–D140.

Jiang, H., Wang, F., Dyer, N. P., & Wong, W. H. (2010). Cisgenome browser: a flexible tool for genomic data visualization. *Bioinformatics*, *26*(14), 1781–1782.

Kadaja, M., Keyes, B. E., Lin, M., Pasolli, H. A., Genander, M., Polak, L., ... Fuchs, E. (2014). Sox9: a stem cell transcriptional regulator of secreted niche signaling factors. *Genes & Development*, *28*(4), 328–341.

Kopylova, E., Noé, L., & Touzet, H. (2012). Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, *9*(4), 357–359.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., ... others (2010). The european nucleotide archive. *Nucleic acids research*, *39*(suppl_1), D28–D31.

Leinonen., R., & Sugawara., H. (2010). The sequence read archive. *Nucleic acids research*, *39*(suppl_1), D19–D21.

Ozsolak, F., & Milos, P. M. (2011). Rna sequencing: advances, challenges and opportunities. *Nature reviews.Genetics*, *12*(2), 87.

Park, P. J. (2009). Chip-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, *10*(10), 669.

Pepke, S., Wold, B., & Mortazavi, A. (2009). Computation for chip-seq and rna-seq studies. *Nature methods*, *6*, S22–S32.

Portela, A., & Esteller, M. (2010). Epigenetic modifications and human disease.

Nature biotechnology, *28*(10), 1057–1068.

Qian, J., Lin, J., Luscombe, N. M., Yu, H., Gerstein, & Mark. (2003). Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, *19*(15), 1917–1926.

Raza, K., & Ahmad, S. (2016). Principle, analysis, application and challenges of next-generation sequencing: a review. *arXiv preprint arXiv:1606.05254*.

Redestig, H., Weicht, D., Selbig, J., & Hannah, M. A. (2007). Transcription factor target prediction using multiple short expression time series from arabidopsis thaliana. *BMC bioinformatics*, *8*(1), 454.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, *43*(7), e47–e47.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., ... Gerstein, M. B. (2009). Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nature biotechnology*, *27*(1), 66–75.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., & Lenhard, B. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, *32*(suppl_1), D91–D94.

Spyrou, C., Stark, R., Lynch, A. G., & Tavaré, S. (2009). Bayespeak: Bayesian analysis of chip-seq data. *BMC bioinformatics*, *10*(1), 299.

Wade, J. T. (2015). Mapping transcription regulatory networks with chip-seq and rna-seq. In *Prokaryotic systems biology* (pp. 119–134). Springer.

Wang., S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., ... Liu, X. S. (2013). Target analysis by integration of transcriptome and chip-seq data with beta. *Nature protocols*, *8*(12), 2502.

Wang, Z., Gerstein, M., & Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57–63.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., ... others (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, *45*(10), 1113–1120.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., ... others (2014). A comparative encyclopedia of dna elements in the mouse genome. *Nature*, *515*(7527), 355.

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., & Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification chip-seq data. *Bioinformatics*, *25*(15), 1952–1958.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., ... others (2008). Model-based analysis of chip-seq (macs). *Genome biology*, *9*(9), R137.