Adv Biochem Eng Biotechnol
DOI: 10.1007/10_2016_43

ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks

Giulio Pavesi

Abstract The first step in the definition of transcriptional regulatory networks is to establish correct relationships between transcription factors (TFs) and their target genes, together with the effect of their regulatory activity (activator or repressor). Fundamental advances in this direction have been made possible by the introduction of experimental techniques such as Chromatin Immunoprecipitation, which, coupled with next-generation sequencing technologies (ChIP-Seq), permit the genome-wide identification of TF binding sites. This chapter provides a survey on how data of this kind are to be processed and integrated with expression and other types of data to infer transcriptional regulatory rules and codes.

Keywords ChIP-Seq, RNA-Seq, Transcription factors, Transcription regulation

Contents

G. Pavesi (✉)
Department of Biosciences, University of Milan, Via Celoria 26, 20133 Milan, Italy
e-mail: giulio.pavesi@unimi.it

1 Introduction: Chromatin Immunoprecipitation and Next-Generation Sequencing

The introduction of *next-generation sequencing* (NGS) technologies has opened up new avenues for every type of genetic and genomic research [1, 2]. One of the fields in which the impact of NGS has been more relevant is perhaps the study of gene regulation at the transcriptional level, and the subsequent analysis steps such as the construction of regulatory networks.

It is essential for the definition of transcription regulatory networks to establish correct relationships between regulators such as transcription factors (TFs) and the genes they regulate [3], together with the effect of the activity of the TFs (activator or repressor) [4]. A fundamental step forward in this direction has been made possible by lab techniques enabling the large-scale identification of TF-DNA binding sites on the genome, with experiments simply impossible to perform just a few years ago.

Chromatin is a complex of DNA and proteins that forms chromosomes within the nucleus of eukaryotic cells. *Chromatin Immunoprecipitation* (ChIP) [5] is a technique enabling the extraction from the cell nucleus of a specific protein-DNA chromatin complex, including DNA binding proteins such as TFs. The different steps of a ChIP experiment are summarized in Fig. 1. First of all, the DNA-bound proteins are cross-linked, that is, fixed to the DNA. The cross-linked chromatin is usually sheared by sonication, providing fragments of 300–1,000 base pairs (bps) in length. Then a specific antibody that recognizes only the protein (TF) of interest is employed, and the antibody, bound to the TF which in turn is bound to the DNA, permits the selective extraction and isolation of the chromatin complex. At this point, DNA is released from the TF by reverse-crosslinking and purified, and the result is a DNA sample enriched in regions corresponding to the genomic locations of the sites that were bound in vivo by the TF (or, in general, the DNA-binding protein) studied. The experiment is performed on thousands of cells at the same time so as to have a quantity of DNA suitable for further analysis and to have enough "enrichment" in the sample, that is, enough copies of each of the DNA regions bound by the TF, to discriminate them from experimental noise.

The next phase is quite logically the identification of the DNA regions themselves – and of their corresponding location in the genome. The introduction of "tiling arrays" had permitted for the first time the analysis of the DNA extracted on a whole-genome scale (ChIP on Chip [4, 6]) by using probes designed to cover the sequence of a whole genome, or a subset of genomic regions of interest (such as with promoter arrays). The introduction of NGS technologies has enabled this type of experiment to move one step further by providing at reasonable cost perhaps the simplest solution: to identify the DNA extracted by the cell by immunoprecipitation, sequence the DNA itself (ChIP Sequencing, or ChIP-Seq [5, 7]).

Without delving into technical details, given a double-stranded DNA fragment derived as just described, sequencing determines the nucleotide sequence on either strand, moving from the $5'$ to $3'$ direction, or both strands simultaneously (paired-

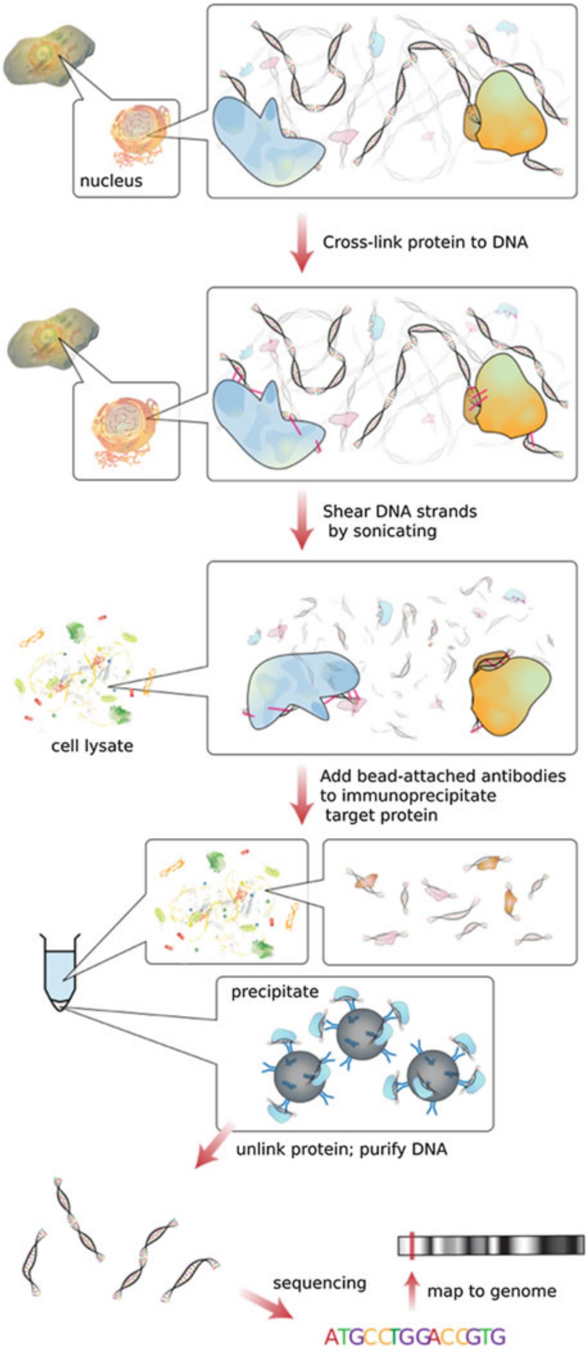
nucleus
Cross-link protein to DNA
Shear DNA strands by sonicating
cell lysate
Add bead-attached antibodies to immunoprecipitate target protein
precipitate
unlink protein; purify DNA
sequencing
map to genome
ATGCCTGGACCGTG

Fig. 1 Chromatin immunoprecipitation workflow (adapted from Wikipedia)

end sequencing). For technical limitations, current NGS platforms can determine the sequence of only a fragment of each region, usually ranging from 50 to 150 bps. Thus, the output is a huge collection of millions of short sequences (called *reads*), which mark the beginning of either or both strands of a DNA region of the sample. The overall number of sequence reads obtained varies from experiment to experiment, and depends on several factors such as the TF involved, sample preparation, experiment replicates, and so on. Suffice it to say that it usually ranges from a few to dozens of millions of short sequence reads.

Once the sequencing has been completed, computational analysis of the data determines which were the DNA regions enriched in the sample (see Fig. 2). First of all, the reads are aligned or "mapped" on the genome to determine their original

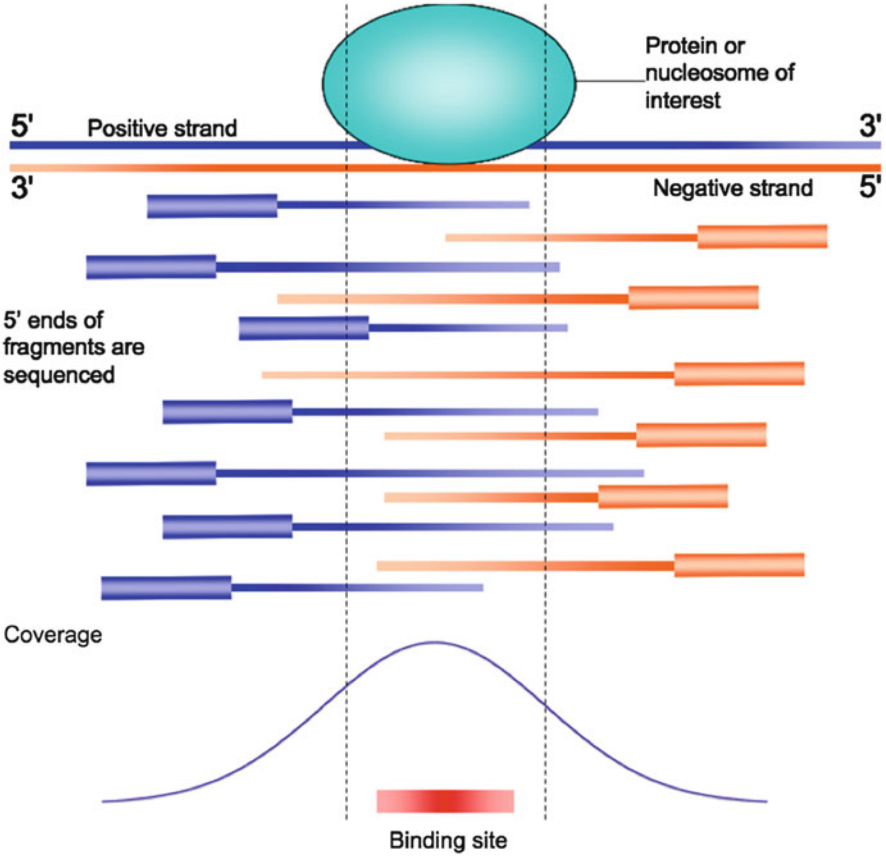
Protein or nucleosome of interest
5'
Positive strand
3'
3'
Negative strand
5'
5' ends of fragments are sequenced
Coverage
Binding site

Fig. 2 Schematic view of the result of a ChIP-Seq experiment on a genomic region bound by a TF. DNA is fragmented at random by sonication, and thus the ends of sequenced DNA fragments map on different positions on the genome. Each fragment is assumed to be the 5′ of a 200–300 bps region, and therefore extended. The resulting signal plot ("coverage") shows a typical "peak" shape. The actual DNA sequence bound by the TF should be located in correspondence of the point of maximum of the coverage plot (*bottom*)

position, using one of the several tools available for this task [8]. It is common, at this stage, to have mismatches in the alignment, that is, sequence reads differ from the reference genome sequence usually in single nucleotides. This is for both biological (sequence polymorphisms) and technical (sequencing errors) reasons. Thus, alignment is usually performed allowing for two or three substitutions per read, with no insertions or deletions. In addition, a non-negligible number of sequence reads align at multiple positions, that is, correspond to repetitive regions of the genome. Although originally these were discarded from further processing, it has indeed been shown that TFs can bind repetitive elements of the genome [9]. Thus, reads mapping at multiple positions should also be considered in the remainder of the analysis, for example also keeping those that map at most in ten different positions.

Once read mapping is complete, regions bordered by reads on both ends (on opposite strands) in numbers high enough to represent a “significant enrichment” and not sample contamination or random noise are singled out. This latter step should be performed with respect to a “control” experiment, aimed at producing “random DNA” and thus a random background model. In other words, if “random” genomic DNA was included in immunoprecipitated samples, another experiment producing only “random” DNA from the same type of cell should give the opportunity to filter the results from false positives and artifacts. The control experiment can be performed in different ways by using an antibody not specific for any TF or, if possible, by using a cell in which the gene encoding the TF studied has been “knocked out,” or its expression “knocked down” in order to remove the immunoprecipitated protein from cells [10].

An ideal example of enriched region is shown in Fig. 2. A “true positive” should correspond to a genomic region bordered by several reads on both strands, and the reads on the two ends should be at a distance “typical” of experiments of this kind, that is, a few hundred bps. By plotting the number of reads falling in each genomic position, the region should be comprised between two “peaks,” one made by reads on the positive strand and one on the negative. Each read mapped on the genome can also be extended by the estimated length of the immunoprecipitated DNA fragments. The latter, following a size-selection step before sequencing, is usually about 200 bps. The result is a signal plot estimating how many times each nucleotide of the genome is covered by an “extended read.” Then a “significantly enriched” region should correspond to a peak in the signal plot, usually located in the middle of the region itself. As in experiments such as ChIP-Seq enrichment is essential to obtain reliable results, single-end sequencing is preferred over paired-end, which would produce at the same cost exactly one half of the sequences, and thus less enrichment.

On the other hand, the same region should not appear – at least with the same number of bordering reads or with the same height of the central peak – in the control experiment. Given the shape of the enriched regions as shown in Fig. 2, this part of the analysis is usually referred to as “peak calling,” that is, identifying all the “peak shaped” regions whose enrichment can be considered to be statistically significant. From the introduction of ChIP-Seq experiments, several different methods for peak calling have been introduced, all following the above considerations but differing in the statistical approaches employed in the definition of significant enrichment. The latter is computed according to the overall number of reads that can be associated with a candidate peak, their distribution on the two DNA strands, and the height of the peak summit. These values are in turn compared to background expected values that might or might not be derived from a control experiment. In a quite ample literature, a few methods have emerged over the years as de facto standards, such as, for example, MACS [11, 12], SPP [13], and PeakSeq [14], which have been employed in the large scale analysis of hundreds of ChIP-Seq experiments performed in the framework of the ENCODE project [15, 16].

The output of peak-calling is a list of genomic regions, likely to be bound by the TF studied in vivo, with p-values and false discovery rates (FDRs) associated with each one. Thus, not only is a “yes/no” output provided but also an estimate of the probability of each region to be considered a false positive call, and hence an estimate of its actual enrichment in the sample. The latter can be employed to restrict, for example, downstream analyses only to the “most likely” or “most significantly enriched” candidates (e.g., only those for which the estimated FDR is under a given threshold). In addition, the “summit” point of each region is usually included in the output, that is, the genomic coordinate of the single base pair where the signal plot associated with the peak is maximum (see Fig. 2). As the actual point of contact with DNA of the TF or the complex investigated should be present in all the regions extracted, the latter should be close to the summit point, which can thus be used to approximate the binding site of the TF within the region for downstream analyses.

2 Finding Transcription Factor Binding Sites

The actual DNA region bound by a TF usually ranges in size from 8–10 to 16–20 bps [3]. TFs bind the DNA in a sequence-specific fashion, that is, they recognize sequences that are similar but not identical, differing in a few nucleotides from one another. As peak regions bound by a TF identified through ChIP-Seq are usually several hundreds of bps long, further processing is needed to identify the actual binding sites within them. Motif discovery or enrichment tools can be employed for this task [17, 18]. The general idea is that the regions identified by the ChIP-Seq, should contain a subset of oligos appearing in all or most of the sequences (thus allowing for experimental errors and the presence of false positives in the set) similar enough to one another to be instances of sites recognized by the same TF. The same set of similar oligos should also not appear with the same frequency and/or the same degree of similarity in a set of sequences selected at random or built at random with a generator of “biologically feasible” DNA sequences [19]. This set of similar and over-represented oligos collectively build a *motif* recurring in the input sequences, describing the binding specificity of the TF itself. Instances of the motif within the enriched regions can then be used to identify the actual binding sites within them. A motif enrichment analysis might also be useful for the identification of additional motifs enriched within the regions which could correspond to binding sites for additional TFs binding DNA in close proximity to the one investigated [20], and thus likely to co-associate with it forming regulatory modules.

3 Associating Binding Sites with Target Genes

The results of ChIP-Seq experiments provide a map of the binding sites on the genome for the TF investigated, but obviously no information regarding genes whose transcription is affected by each of the binding sites. For building regulatory networks it is therefore essential to associate each region with one or more "target" genes.

The first logical step is to single out binding sites located within promoters. There is no unique definition of what constitutes the "promoter" of a gene or of its size. It is usually described as a region of a few hundred or thousand base pairs located upstream of its transcription start site (TSS). ChIP-Seq experiments performed on histone modifications, however, revealed that active promoters have a very precise chromatin signature, that is, a pattern of modifications such as H3K4me3 or H3K9ac covering a few nucleosomes upstream and downstream of the TSS itself [21]. Hence, even if it narrows down the number of binding sites that can be assigned to promoters, it is advisable not to define a region too broad around TSSs as "promoter" and avoid going beyond 1 kbp upstream or downstream of the TSS. Indeed, TF binding regions outside these "core promoters" (e.g., within the first intron or further than 1 kbp upstream of the TSS) exhibit a different chromatin signature, with modifications such as H3K27ac or H3K4me1 that are indicators of distal "enhancer" or "silencer" regions but not of promoters.

Associating distal binding sites, not close to TSSs, with the "right" target genes is perhaps the hardest part of this type of analysis. Even factors usually associated with promoters and TSSs such as NF-Y [9, 22] have the majority of their binding sites located in distal regulatory regions. Thus, restricting the analysis only to binding sites located in promoters has the effect of missing several target genes regulated by the binding of the TF to distal elements; on the other hand, associating a distal regulatory element with the wrong gene produces wrong data.

In the absence of further information, this step usually follows the "nearest neighbor rule": a distal binding site is associated with the closest TSS on the genome. If the binding site is within a gene body (the transcribed region of a gene) then it is attributed to the gene itself. Given a reference annotation providing the genomic coordinates of genes that can be retrieved from any genome browser [23, 24], this analysis can be performed with in-house developed scripts, or with tools such as HOMER [25] or GREAT [26]. On the other hand, as a typical ChIP-Seq experiment returns several thousands of bound regions, associating every peak with the closest TSS results in a very sizable portion of the annotated genes to be considered targets of the TF investigated. Hence, further criteria are employed to reduce their number, usually by establishing a threshold on the distance from the TSS of the binding sites. For example, in the large-scale analysis performed in the Roadmap Epigenomics project [21], an enhancer region was associated with the closest gene if its TSS was located at less than 30 kbp from the enhancer itself. Otherwise, no association was defined.

Modern experimental techniques based on immunoprecipitation and NGS, the most relevant being ChIA-PET or ChIA-Seq experiments [27, 28], have enabled light to be shed on this aspect too. The ChIA-PET (or -Seq) method combines ChIP-Seq methods and Chromosome conformation capture techniques such as 3C [29] for the identification of long-range chromatin regulatory interactions [30]. The immunoprecipitation is performed against a protein usually found in complexes connecting enhancers to the respective TSSs, such as p300, to be pulled down together with all the DNA regions bound to it. Before sequencing, linker sequences are incorporated onto the free ends of the DNA fragments tethered to the protein complexes. To build connectivity of the DNA fragments, the linker sequences are ligated by nuclear proximity ligation. The resulting DNA sequences is thus formed by both the enhancer and the promoter, connected by a linker sequence. Application of NGS paired-end sequencing produces sequence pairs coming from each of the two connected regions. Subsequent mapping on the genome finally results not in single peaks but in "paired" peaks, located at different positions of the genome, where reads in one peak are found to be paired in sequencing with reads in the other. Paired peaks correspond to pairs of genomic regions connected by the protein complex immunoprecipitated. These experiments thus enable the identification of unique, functional chromatin interactions between distal and proximal regulatory transcription-factor binding sites and the promoters of the genes with which they interact. Remarkably, their application has revealed the serious limitations of the application of the "nearest neighbor" rule introduced before: for example, in mouse stem cells only about one-third of the long-distance enhancer-promoter interactions have been shown to be associated with the gene nearest to the enhancer [31]. An enhancer located within a transcribed region can also regulate a distal gene. Finally, a sizable number of the enhancers (about 30% of the total) were even associated with genes located on different chromosomes. All in all, then, in the absence of long-distance interaction data, all the enhancer-promoter associations should be taken with a pinch of salt.

4 Assessing TF Activity from Expression Data

TFs can have the effect of both activating and repressing the transcription of target genes. Thus, the activity of any TF can be assessed by performing experiments in which the expression of the TF is limited or, vice versa, amplified. Then the activity of the TF on target genes can be measured by identifying those genes that change their expression level as a consequence of the TF inactivation or over-expression.

Before the introduction of genome-wide techniques such as ChIP-Seq this was indeed the method of choice for the identification of putative target genes for TFs. It is, however, important to stress the fact that this approach, alone, might also identify genes that are not direct targets. In other words, the TF directly affects the expression of a subset of differentially expressed genes; some of the direct targets can in turn regulate further genes, also found to be differentially expressed, and so on.

Before the advent of NGS technologies, expression studies were usually performed with oligonucleotide microarrays. Then the application of NGS to RNA (RNA-Seq) was shown to be able not only to reconstruct and assemble whole transcriptomes, but also to provide a reliable quantification of the expression level of each gene [32, 33].

One of the key advantages of RNA-Seq over microarrays is that they enable one to identify and reconstruct the single alternative transcripts of the same gene, as well as estimate their expression level. This, in turn, has revealed alternative splicing and alternative transcript production to be ubiquitous features of eukaryotic genes [34]. From the viewpoint of transcription regulation it is worth mentioning that alternative promoters and transcription start sites have emerged as a widespread feature. This is a very important point in the association between TF binding and promoters, as a TF-gene association could be missed if the alternative promoter bound by the TF is not included in the analysis. For a TF binding only one of the alternative promoters of a gene, its effect on gene transcription should be assessed only for the corresponding transcripts. Techniques such as Cap Analysis Gene Expression (CAGE [35]), coupled with NGS sequencing [36], enable one to identify more reliably alternative TSSs and the relative transcription level.

It is worth mentioning that the usual measures of transcript level employed are concentration measures. That is, the “expression level” of a transcript or gene is an estimate of the fraction of the RNA sample that can be assigned to it, described by normalized measures such as “reads per kilobase of exon per million reads” (RPKM) or “transcripts per million” (TPM). This, in turn, can produce incorrect conclusions when applied to experiments resulting from TF inactivation or overexpression. Suppose, for example, that a TF acts purely as an activator, targeting 10% of the genes of the genome studied. Upon inactivation of the TF, the transcript level of its target genes is decreased and the rest of the genome remains unchanged. As expression measures used are relative and describe concentration with respect to the overall sample, we observe a marked reduction of the transcript levels for the target genes, but at the same time an increase of the expression estimate of non-target genes, some of which might also finally be “significantly overexpressed” by statistical analysis. Hence, the TF is incorrectly observed to act both as an activator and a repressor. Other than previous knowledge about the TF activity, indicators of the possible presence of this effect for an activator TF are a large majority of genes significantly down-regulated with just a few overexpressed, the latter having very high expression estimates. Vice versa for repressor TFs. In case of doubt, special techniques should be employed in the design and analysis of the expression experiment, as shown, for example, in [37].

5 Mining Available Data

The ever decreasing cost of next-generation sequencing has led to the widespread application of the techniques described in this chapter, such as ChIP- and RNA-Seq. It has indeed become common practice to study simultaneously more than one TF in a given condition in order to have more meaningful results and to identify co-associations and modules of key regulators [38, 39]. The last few years have also witnessed the completion of large-scale general purpose projects in which hundreds of TFs have been tested in several different cell lines. The most relevant example is perhaps the (still ongoing) ENCODE project [40], in which hundreds of human and mouse TFs have been analyzed through ChIP-Seq in several different cell lines, or the modENCODE project for model organisms such as *Drosophila melanogaster* or *Caenorhabditis elegans* [41]. TF ChIP-Seq data are integrated by other data relevant for transcriptional regulation analysis such as chromatin structure, histone modifications, DNA methylation, expression profiles from RNA-Seq and CAGE experiments, and ChIA-PET data for long-distance chromosomal interactions. Analysis of co-occurrence of TF binding sites of the genome revealed that TFs tend to associate, forming distinct co-regulatory modules [15], giving rise to many enriched regulatory network motifs (e.g., noise-buffering feed-forward loops). Hence, any TF should not be viewed as a separate entity whose interactions with other regulatory factors happen only by chance, but should be considered as part of more complex regulatory modules, and the construction of regulatory networks should consider this point.

Other than the deluge of information they contain, these data, or those contained in large repositories such as Cistrome [42], constitute a perfect benchmark set for any bioinformatics or systems biology approach to the study of transcriptional regulation. They can also be retrieved to complement data produced locally. There also exist resources in which data have already been processed, for example tools such as Cscan [43] or Enrichr [44], which already have pre-computed associations between TFs and target genes for hundreds of experiments.

6 Conclusions

The introduction and the creative use of next-generation sequencing technologies have opened new avenues for every aspect of genetic and epigenetic research. Perhaps the field that has benefited most from them is regulation of gene expression at the transcriptional and post-transcriptional level. This chapter provides a brief survey of the experimental and bioinformatic techniques currently employed for the study of transcription factors, summarized in Fig. 3, from the identification of target genes to the characterization of their activity, and all fundamental steps for subsequent studies such as the definition and analysis of transcriptional regulatory networks.

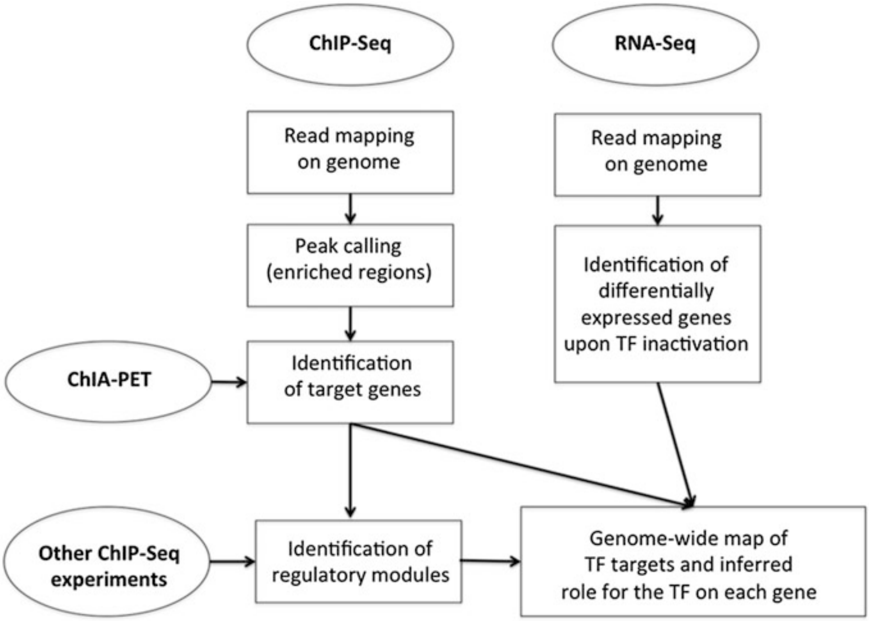
ChIP-Seq
RNA-Seq
Read mapping on genome
Read mapping on genome
Peak calling (enriched regions)
Identification of differentially expressed genes upon TF inactivation
ChIA-PET
Identification of target genes
Other ChIP-Seq experiments
Identification of regulatory modules
Genome-wide map of TF targets and inferred role for the TF on each gene

Fig. 3 Combining different NGS-based experiments for building the regulatory map of a given TF. ChIP-Seq identifies genomic regions bound by the TF, and further processing the corresponding target genes. The latter can in turn be more easily singled out by capturing long-distance interactions with experiments such as ChIA-PET. RNA-Seq experiments assess significant changes of gene expression upon TF inactivation. Different ChIP-Seq experiments performed in the same condition can be combined to identify regulatory modules

References

1. Horner DS, Pavesi G, Castrignano T, De Meo PD, Liuni S, Sammeth M, Picardi E, Pesole G (2010) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinform 11(2):181–197. doi:10.1093/bib/bbp046
2. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24(3):133–141. doi:10.1016/j.tig.2007.12.007
3. Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424 (6945):147–151. doi:10.1038/nature01763
4. Blais A, Dynlacht BD (2005) Constructing transcriptional regulatory networks. Genes Dev 19 (13):1499–1511. doi:10.1101/gad.1325605
5. Collas P, Dahl JA (2008) Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. Front Biosci 13:929–943
6. Pillai S, Chellappan SP (2009) ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. Methods Mol Biol 523:341–366
7. Mardis ER (2007) ChIP-seq: welcome to the new frontier. Nat Methods 4(8):613–614. doi:10.1038/nmeth0807-613
8. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25. doi:10.1186/gb-2009-10-3-r25

9. Fleming JD, Pavesi G, Benatti P, Imbriano C, Mantovani R, Struhl K (2013) NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. Genome Res 23 (8):1195–1209. doi:10.1101/gr.148080.112
10. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6(11 Suppl):S22–S32. doi:10.1038/nmeth.1371
11. Feng J, Liu T, Zhang Y (2011) Using MACS to identify peaks from ChIP-Seq data. Curr Protoc Bioinformatics Chapter 2:Unit 2. 14. doi:10.1002/0471250953.bi0214s34
12. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9 (9):R137. doi:10.1186/gb-2008-9-9-r137
13. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5(9):829–834. doi:10.1038/nmeth.1246
14. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27(1):66–75. doi:10.1038/nbt.1518
15. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489(7414):91–100. doi:10.1038/nature11245
16. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22(9):1813–1831. doi:10.1101/gr.136184.111
17. Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. Nucleic Acids Res 43 (W1):W39–W49. doi:10.1093/nar/gkv416
18. Zambelli F, Pesole G, Pavesi G (2014) Using Weeder, Pscan, and PscanChIP for the discovery of enriched transcription factor binding site motifs in nucleotide sequences. Curr Protoc Bioinformatics 47:2. 11. 11–12. 11. 31. doi:10.1002/0471250953.bi0211s47
19. Zambelli F, Pesole G, Pavesi G (2013) Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Brief Bioinform 14(2):225–237. doi:10.1093/bib/bbs016
20. Zambelli F, Pesole G, Pavesi G (2013) PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. Nucleic Acids Res 41(Web Server issue):W535–W543. doi:10.1093/nar/gkt448
21. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M,

Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M (2015) Integrative analysis of 111 reference human epigenomes. Nature 518(7539):317–330. doi:10.1038/nature14248

22. Ceribelli M, Dolfini D, Merico D, Gatta R, Vigano AM, Pavesi G, Mantovani R (2008) The histone-like NF-Y is a bifunctional transcription factor. Mol Cell Biol 28(6):2047–2058. doi:10.1128/MCB.01861-07
23. Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C, Humphrey J, Kerhornou A, Khobova J, Aranganathan NK, Langridge N, Lowy E, McDowall MD, Maheswari U, Nuhn M, Ong CK, Overduin B, Paulini M, Pedro H, Perry E, Spudich G, Tapanari E, Walts B, Williams G, Tello-Ruiz M, Stein J, Wei S, Ware D, Bolser DM, Howe KL, Kulesha E, Lawson D, Maslen G, Staines DM (2015) Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res. doi:10.1093/nar/gkv1209
24. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Fujita PA, Eisenhart C, Diekhans M, Clawson H, Casper J, Barber GP, Haussler D, Kuhn RM, Kent WJ (2015) The UCSC Genome Browser database: 2016 update. Nucleic Acids Res. doi:10.1093/nar/gkv1275
25. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38(4):576–589. doi:10.1016/j.molcel.2010.05.004
26. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28(5):495–501. doi:10.1038/nbt.1630
27. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, Ariyaratne PN, Mohamed YB, Ooi HS, Tennakoon C, Wei CL, Ruan Y, Sung WK (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol 11(2):R22. doi:10.1186/gb-2010-11-2-r22
28. Paulsen J, Rodland EA, Holden L, Holden M, Hovig E (2014) A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. Nucleic Acids Res 42(18), e143. doi:10.1093/nar/gku738
29. Simonis M, Kooren J, de Laat W (2007) An evaluation of 3C-based methods to capture DNA interactions. Nat Methods 4(11):895–901. doi:10.1038/nmeth1114
30. Li G, Cai L, Chang H, Hong P, Zhou Q, Kulakova EV, Kolchanov NA, Ruan Y (2014) Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application. BMC Genomics 15(Suppl 12):S11. doi:10.1186/1471-2164-15-S12-S11
31. Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, Ngan CY, Lim J, Tai E, Poh HM, Wong E, Mulawadi FH, Sung WK, Nicolis S, Ahituv N, Ruan Y, Wei CL (2013) Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature 504(7479):306–310. doi:10.1038/nature12716
32. Fonseca NA, Marioni J, Brazma A (2014) RNA-Seq gene profiling—a systematic empirical comparison. PLoS One 9(9), e107026. doi:10.1371/journal.pone.0107026
33. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18(9):1509–1517. doi:10.1101/gr.079558.108
34. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456(7221):470–476. doi:10.1038/nature07509

35. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100 (26):15776–15781. doi:10.1073/pnas.2136655100
36. Takahashi H, Lassmann T, Murata M, Carninci P (2012) 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc 7(3):542–561. doi:10.1038/nprot.2012.005
37. Loven J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA (2012) Revisiting global gene expression analysis. Cell 151(3):476–482. doi:10.1016/j.cell.2012.10.012
38. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133(6):1106–1117. doi:10.1016/j.cell.2008.04.043
39. Hutchins AP, Diez D, Takahashi Y, Ahmad S, Jauch R, Tremblay ML, Miranda-Saavedra D (2013) Distinct transcriptional regulatory modules underlie STAT3's cell type-independent and cell type-specific functions. Nucleic Acids Res 41(4):2155–2170. doi:10.1093/nar/gks1300
40. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res 22(9):1798–1812. doi:10.1101/gr.139105.112
41. Brown JB, Celniker SE (2015) Lessons from modENCODE. Annu Rev Genomics Hum Genet 16:31–53. doi:10.1146/annurev-genom-090413-025448
42. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y, Pape UJ, Poidinger M, Chen Y, Yeung K, Brown M, Turpaz Y, Liu XS (2011) Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol 12(8):R83. doi:10.1186/gb-2011-12-8-r83
43. Zambelli F, Prazzoli GM, Pesole G, Pavesi G (2012) Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. Nucleic Acids Res 40 (Web Server issue):W510–W515. doi:10.1093/nar/gks483
44. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14:128. doi:10.1186/1471-2105-14-128