4-1-2022

An Intersection of Computational Biology and Functional Genomics to identify Transcriptional Gene Enhancers and Their Role in Cancer

Naureen Aslam Khattak
Indiana State University

Rusty Allen Gonser
Indiana State University

An Intersection of Computational Biology and Functional Genomics to identify Transcriptional Gene Enhancers and Their Role in Cancer

Naureen Aslam Khattak[1,2], Rusty Allen Gonser[1,2]

[1]Department of Biology, Indiana State University, Terre Haute, Indiana

[2]The Center of Genomic Advocacy (TCGA), Indiana State University, Terre Haute, Indiana

naslam@sycamores.indstate.edu, rusty.gonser@indstate.edu

Abstract

Despite the critical role of gene regulation in cell development and differentiation, the major challenge remains to identify the *cis-regulatory modules* (CRMs). Mainly, these *CRMs* include enhancers, promoters and insulators that governs the spatiotemporal gene regulation. The gene regulatory networks are highly dependent on their *CRMs* and mostly consist of DNA motifs and epigenetic landmarks. The recent advancements in high-throughput sequencing techniques and comparative genomics analysis accelerate the discovery of enhancers, however the major obstacles are to identify the genome-wide location of these CRMs, their dynamic nature of interactions, and cis*/tran*s location which could be hundred to thousands base pairs away from the target gene location. The goal of this literature review is to provide an insight into the *CRMs* specifically enhancers, how they modulate gene expression, mutations that converts normal cell into a disease-state such as cancer. Also, this embedded review article is focused on the use of computational strategies coupled with the biochemical assays to predict functional gene enhancers. The computational strategies such as window clustering, probabilistic modeling, phylogenetic footprinting and discriminative modeling are briefly discussed to scan and locate the putative gene enhancers. Besides theses, biochemical techniques such as ChIP-seq, DNA footprinting, and deletion mapping are briefly reviewed in *Drosophila* to predict functional gene enhancers and dissecting gene regulatory networks. In addition, this review article may help

bench scientists to incorporate bioinformatics tools with biochemical techniques to scan, locate and verify gene enhancer regions within a cell. With best of our knowledge, this is a first-time effort to combine insilico, in vitro and in vivo techniques to explore the connections between CRMs and gene regulation.

Keywords

Gene Enhancers, Cis regulatory modules, ChIP-seq, Deletion Mapping, Cancer, Gene Expression. Cancer, Computational Biology.

Introduction

An interesting question in developmental biology is deciphering how multiple cues are integrated to determine where and under what conditions a specific gene is expressed. Despite all cells in an organism having the same genetic makeup, only a subset of these genes is expressed in each type of cell, thereby providing each cell type with a unique identity. The difference in gene expression is derived by regulatory regions of DNA called enhancers, which bind with specific proteins called transcription factors (TFs) to regulate gene expression. The challenge is still on identifying the multiple signals, which coordinate and communicate with each other to drive the expression of a particular gene. The integration of multiple signals from different Transcription Factors (TFs) mediates the spatial and temporal gene expression in a tightly controlled environment. Thus, a specific gene expressed in a specific tissue at a specific time is highly dependent on TFs. Any mutation these TFs or their associated proteins results in an up and down regulation of gene expression which leads to multiple disorders such as different type of cancer [1]

The TFs and their binding sites are essential for a normal gene activity and controls the rate of gene transcription. Mainly, enhancer (*CRMs* modules) consists of single or multiple binding sites for a variety of TFs that modulate a gene expression either in a direct or indirect way. For direct transcriptional control, an activator (TFs) binds to an enhancer region without any additional

support to initiate gene transcription. However, the indirect transcriptional control acts through cofactors/additional TFs to regulate a gene activity [2]

Gene Enhancers

The enhancer consists of sequence-specific DNA binding sites known as Transcription Factor Binding Sites (TFBS) along with some other signals necessary to regulate a gene expression. The average size of an enhancer is a few hundred to a few thousands base pairs (bp) long. These enhancers make a loop-like structure and recruit TFs to regulate a gene expression as depicted in **figure 1**. Generally, enhancers are divided into two broad categories known as short-range (proximal) and long-range (distal) enhancers. Super-enhancers are also found in the mammalian genome where multiple enhancers are present with an array of TFs bound to these enhancers [3].

Mechanism of Action

Enhancers acts in a cooperative manner or in a stand-alone mode. Single or multiple TFs can bind to an enhancer region. If a TF binds to an enhancer and serve as a docking site for another TFs to activate or repress gene expression, it reflects cooperative or indirect mode of action, whereas the direct mode of action (stand-alone) shows a TF directly binds to the target site and act as an activator/repressor without any additional support as shown in **figure 2** [2]. In addition to direct and indirect mode of action, it also depends upon the (i) binding occupancy of each TF, and (ii) their orientations with respect to Transcription Start Site (TSS) and enhancer category (proximal vs. distal enhancer) in driving gene expression (**figure 3**)

Transcription Factors (TFs): TFs are prerequisite for an efficient gene transcription. This binds on enhancers regions and drives gene expression. Most of the transcription factors reported so far consist of DNA binding domain and activation domain. A brief detail of each TF is below

Helix-loop-Helix motif (HLH)

HLH TFs consist of basic amino acids that contact with DNA and neighboring regions to mediate dimmer formation. Dimmer is formed due to flexibility of loop which allows folding and packing against the other helix (**figure 4**). These TFs play a role in cell development and differentiation. MyoD1 is an example of HLH TF that binds to E2A protein [3]

Helix-turn-Helix motif (HTH)

The HTH binding motif consists of a pair of α helices separated by a light turn. The second α helix lies in the major groove of DNA where it contacts with DNA bases, whereas the first α helix make contact with DNA backbone as shown in **figure 5** [4]

Zinc Finger domain

Zinc finger domains are mainly responsible for inducing growth and differentiation. These TFs were first identified in the Xenopus model organism. The residues cysteine and histidine coordinate with zinc ions and form a zinc finger-like projection. Zinc fingers consist of an α helix and β sheet held together by zinc ions (**figure 6**). Typically, a finger motif has the following sequence: Cys-X2 or $_{4}$-Cys-X_{12}-His-$X_{3\text{-}5}$-His. Several proteins having zinc finger binding sites have been identified such as TF SP1[4]

Leucine zipper motif

The conserved sequence of Leucine zipper motif was first discovered in eukaryotic proteins and has a critical function in cell differentiation and development. Leucine zipper contains 4-5 leucine located seven residues apart found in the basic amino acid region. These two regions spread over 60-80 residues, together constitute bZIP domain. The basic region is held together by dimerization of adjacent zipper regions, when hydrophobic faces of two zippers interact in

parallel orientation, the leucine zipper part stabilizes the protein dimmer as shown in **figure 7** [4]. Different types of TFs are summarized in **table 1.**

Table 1: Transcription Factors Domains and their Function

TFs Domain	Role	Function containing domain	Gene ID (NCBI)	OMIM	Chr location
Homeobox	DNA binding	Numerous *Drosophila* homeotic genes related genes in other organisms such as *Cad, Abd-A/B*	CG1759 CG10325 CG11648	600297 142951 142956	2L 3R 3R
Cysteine-histidine zinc finger	DNA binding	*TFIIIA, Kruppel, Spl*	2971 9314	600860 602253	13 9
Cysteine-cysteine zinc finger	DNA binding	Steroid-thyroid hormone receptor family	*NA	*NA	*NA
Leucine Zipper	Protein Dimerization	*C/EBP, c-fos, c-jun, GCN4, c-myc*	2353 3725	164810 165160	14 1
Helix-loop-Helix	Protein Dimerization	*c-myc, Drosophila daughterless, MyoD, E12, E47*	4609 4654	190080 159970	8 11
Proline-rich region	Gene Activation	*Yeast GCN4, GAL4, steroid-thyroid*	V XVI	856709 855828	--
Glutamine-rich region	Gene Activation	*SP1*	6667	189906	12
Amphipathic acidic alpha-helix	Gene Activation	*CTF/NF1*	4763	613113	17

*NA = Not available

Enhancers and Their Associated Diseases

Based on the critical importance of enhancers in gene regulation (activation or inhibition of gene expression), it is not surprising that any change in the enhancer itself or its associated factors can

result in disease. For a detailed review, please see the Jaret M. Karnuta1 and Peter C. Scacheri, 2018 [7].

Enhancers and their Role in Cancer

Disease can appear if a TF is up or down regulated (ectopic activation of gene expression) or becomes active at the wrong time or in the wrong place. The mis-regulation of gene expression plays a major role in the development of certain types of cancers. For instance, over-expression of proto-oncogenes results in 'cancer-causing' oncogenes, particularly the gene responsible for cell growth. A few examples are the growth factors and their receptors (*erbA, fos, myb, and myc*) that encode TFs necessary for growth turned into oncogene if there is any mutation associated with these factors or their genes. Thus, the conversion of these protooncogenes into oncogenes, which can occur either by mutation (over-expression or under-expression) corresponds to a difference in gene regulation pattern which results in cancer [8,9]. Table 2 briefly describes TFs, mutations, and cancer types. A detailed consists of mutation in enhancers regions and their effect or organism type is provided in the supplementary data files **(Table 3**). The mutations include insertion/deletion, translocation, inversion, duplications, and point mutations with phenotype defect, NCBI gene ID, chromosome location and OMIM record are provided below.

Table 2: Different type of cancers and their association with enhancer malfunctioning

Cancer types	Mutations	Gene ID (NCBI)	Chr	MIM record	Ref.
Breast, prostate	*FOXA1*	3169	14	602294	10
Lung, AML	*RAD21*	5885	8	606462	11
Burkitt's lymphoma	*IGH/MYC* translocation	4609	8	190080	12

Breast, lung	*GATA3*	2625	10	131320	13
Transitional cell carcinoma	*NIPBL*	25836	5	608667	14
Urothelial, bladder, breast, head, and neck	*CTCF*	10664	16	604167	15
B-cell lymphoma, lung	*EZH2*	2146	7	601573	16
Bladder, glioblastoma, lung, urothelial	*STAG2*	10735	X	300826	17
Bladder, AML, lung	*SMC3*	9126	10	606062	18
Bladder, lung, urothelial, and breast	*MLL2/MLL3/ MLL4*	8085	12	602113	19

*Chr: Chromosome

*OMIM: Online Mendelian Inheritance in Man (database for human disease)

An example is *fos* and *jun* TFs. These are normally synthesized transiently in response to growth promoting signals and act to activate the genes encoding specific proteins required for cellular growth. If for any reason these proteins are continuously expressed (over-expression), they act to promote the continuous growth in the absence of growth factors and are capable of transforming normal cells into cancer cells. In contrast to above examples, in some cases the TFs failed to regulate correctly (at correct time and place) results in an inappropriate gene activity. This indicates that, as with other cellular processes, gene expression is subject to complex regulatory mechanisms, the failure of which can be as devastating as the failure of the basic process

Computational Strategies for Enhancer Prediction

Computational search algorithms are widely used to identify the enhancers regions or the hotspots areas where an enhancer or their associated TFs may be existing. These search

algorithms either used experimental data to get a fine-tuned matching for *CRMs* or based on mathematical or statistical models to get a reasonable prediction. The most accurate strategy for predicting *CRMs* varies and depends upon the question of interest. For instance, among all different types of prediction tools, which predictive tool stands high for identifying the *CRMs*? Are they time- efficient? Are there any false positive and negative results? Did they use experimental data or based on putative annotation? What is the limitation of the algorithms in terms of predicting short vs. long enhancers or in other words what is the input size limitation of the tool? For details, please refer to Su, J., Teichmann, S. A., & Down, T. A. [23]. The *insilico* methods used for enhancers predictions are shown in **figure 8** are roughly classified into four categories briefly described below.

1. **Window clustering** involves significant clustering of high densities of binding sites within a sequence window.
2. **Probabilistic modelling** consists of identifying sequences that resemble a statistical model of a binding site cluster more than a model of background DNA.
3. **Phylogenetic footprinting** searches for high density regions of binding sites conserved between closely related species.
4. **Discriminative modelling** seeks to identify set of signals on regulatory regions that can maximize the differences between regulatory regions and non-regulatory regions. Many methods are hybrids of two or more strategies.

Window Clustering: The literal meaning of the word "clustering" is to **"**group together" based on similar properties. The same approach is utilized to predict the *CRMs* in the genome based on statistical analyses. The method uses high density TFBS and groups them together based on statistical observation. These significant clusters are then the hotspot for finding the TFs. The (i)

MSCAN [24] (ii) MCAST [25] and (iii) CisPlusFinder [26] tools are based on clustering methods. The input data for these tools consist of motif library against single genome and multiple sequences alignment respectively as shown in **figure 9.**

Probabilistic Modelling: This computational approach utilizes the Hidden Markov Model (HMM) to generate a set of *CRMs* sequences which are based on a combination of a set of TFBS. Common tools include ClusterBuster [27], Stubb, StubbMS [26], MoprhMS [28], CisModule [28], and MultiModule [29]. The difference between these tools is mentioned below

1) The ClusterBuster, Stubb, and CisModule stands on Multiple Sequences Alignment (MSA)
2) StubbMS and Morphs MS are diverse in their first step of execution. StubbMs provides fixed alignment by using Lagan [23]. On the other hand, MorphMS sums up (using probability) all the possible alignments based on their binding sites.
3) CisModule and MultiModule predict *CRMs* in a single step. The CisModule follows Bayesian inference to find the binding sites and location of the *CRMs*, however, the Multimodule uses the same strategy but adds the comparative genomic information to complete the analyses.

Some of the tools

present in window clustering (CisPlusFinder) and Probabilistic Modelling (StubbMS, MorphMs and MultiModule) also follow the Phylogenetic fingerprinting methods based on Multiple sequence alignment approach [23]

Discriminative Modelling: In this method for instance HexDiff [26], the input consists of a hexamer or set of nucleotides (6-mer) with high frequency and differentiate between *CRMs* and non-*CRMs*.

Phylogenetic Footprinting: This method uses phastCons score [27] which serves as independent control and takes sequence conservation as an input. The phastCons score considers the evolutionary distance between species which are followed by the Hidden *Markov Model*. Different bioinformatics tools are now available which provide information about regulatory elements as well as TFs binding sites and target genes of regulatory elements. A brief overview of different bioinformatics software and databases are available in **table 4**. A brief list of databases used for identifying Transcription Factors (TFs) and their binding sites are provided in t**able 5**.

Table 4: Bioinformatics software and databases for predicting *cis-regulatory modules* in genome

Tools	**Principle**	**Input**	**Website**
MSCAN	Window Clustering	Single Genome	http://www.cisreg.ca/cgi-bin/mscan/ MSCAN
MCAST			http://alternate.meme-suite.org
CisPlusFinder			http://jakob.genetik.uni-koeln.de/ bioinformatik/people/nora/nora.html
ClusterBuster	Phylogenetic footprinting	MSA	http://zlab.bu.edu/cluster-buster/
Stubb			http://stubb.rockefeller.edu/
StubbMS			http://stubb.rockefeller.edu/
MorphMS			http://veda.cs.uiuc.edu/Morphalign/

		Motif Library	supplement/
CisModule			http://www.stat.ucla.edu/~zhou/CisModule/
MultiModule			http://www.stat.ucla.edu/~zhou/MultiModule/index.html
EEL	Window Clustering		http://www.cs.helsinki.fi/u/kpalin/EEL/
RP	Discriminative Modeling	CRM annotations	http://www.bx.psu.edu/projects/rp/
HexDiff	Discriminative Modeling		http://www.ics.uci.edu/~bobc/hexdiff.html
PhylCRM	Phylogenetic Footprinting	Single Genome	http://the_brain.bwh.harvard.edu/PhylCRM/
EMMA	Phylogenetic Footprinting/ Probabilistic Modeling	Motif Library	https://www.bioinformatics.nl/cgi-bin/emboss/emma

**For more details on these methods, please refer to Su, J., Teichmann, S. A., & Down, T. A. (2010).

With the availability of the microarray expression data analysis, several tools are published to predict CRMs in tissue and stage specific manner for example LRA [28], Cluster Scan [29] Composite Module Analyst [30], Module Miner [31]. The methods which predict TFBs based on user-defined dataset includes but not limited to Module Scanner [64], TargetExplorer [32], and CisModScan [33]. These computer-based tools are unable to find out novel CRM instead of that they look for the binding sites within a defined sequence.

197 **Table 5: Databases for identifying the Transcription Factors (TFs) and their binding sites**

Database	Acronym	Principle	Model Organism	Website
Transcription factor prediction database	DBD	Database of predicted TFs in completely sequenced genomes. Superfamily, Pfams and Hidden Markov model libraries-based prediction	*B. Subtilis* *C. Elegans,* *D. melanogaster* *E. coli,* *H. Sapiens,* *M. Musculus* *S. cerevisiae.*	www.transcriptionfactor.org
Transcription factor 2 DNA	TF2DNA	Predict TFs binding motifs using experimental and theoretical data source	*E. coli, C. Elegans,* *D. melanogaster,* *M. Musculus*, *S. cerevisiae,* *H. Sapiens,*	http://www.fiserlab.org/tf2dna_db/
JASPAR	JASPAR CORE	A curated, non-redundant set of profiles, derived from published and experimentally defined TF binding sites for eukaryotes uses position weight matrices (PWM)	Eukaryotes *Vertebrata*, *Nematoda*, *Insect*, *Plantae*, *Fungi*, *Urochordata*	http://jaspar.genereg.net
Gene Transcription Regulation Database	GTRD	Database of TFs identified by ChIP-seq experiments for human and mouse.	*M. Musculus*, *H. Sapiens.*	http://gtrd.biouml.org
TRANSFAC	TRANSFAC	a database on TFs and their DNA binding sites	Eukaryotic transcription factors	http://genexplain.com/transfac/#section0
AnimalTFDB	AnimalTFDB	Annotations from the NCBI Entrez Gene and Ensembl	Animal Genomes	http://bioinfo.life.hust.ed

		databases, including basic information, gene phenotypes, homologous genes, and Gene Ontology (GO) Classification of transcription cofactors; (iii) TF binding sites information; (iv) the GWAS phenotype related information of human TFs.		u.cn/AnimalTFDB/

198

199 **Biochemical Techniques for Enhancer Prediction using Transgenic Animal Models**

200 As mentioned earlier, the TFs are specific proteins that binds to the enhancer region and regulate
201 a gene activity. The DNA footprinting, Deletion mapping and Chromatin Immunoprecipitation
202 (ChIP-seq) are frequently used techniques in laboratory to identify the location and binding
203 occupancy of TFs on the enhancer regions in genome as shown in **figure 12**. Although these
204 techniques stand on the same basic principle of identifying TFBS, steps used in these methods
205 are entirely different and are comparable. The overall principle and steps used in these strategies
206 are given below:

207 **Deletion Mapping: TFBS Identification Techniques**

208 This technique takes the advantage of deleting various parts the promoter region of a gene and
209 measured the transcription activity. In this method, mutant gene (depending upon the deletion
210 results in either (i) increase in the transcription activity (ii) decrease in transcription activity or
211 (iii) in some cases no /little effect on the gene regulation. In general, a few nucleotide deletions
212 have little or no effect on the gene transcription, however if the deletion hit the regions which is
213 important for the binding of the TFs, then it will show subsequent decrease in the transcriptional
214 activity. On the other hand, if the deletion occurred in the region which is responsible for the

repression of the gene activity, then it might result in increased gene transcription because the repressor binding site is no longer available to inhibit the gene expression as shown in figure 13. The different scenarios are listed below

i. If the deletion occur/falls in the regions which is important for TF binding (prevents binding of TFs) to activate transcription the level will decrease, the transcriptions (38%)

ii. Deletion of regions also might increase transcription in a case that TF binds to a region that inhibit transcriptions, but deletion of that inhibiting region will increase transcription (114%)

iii. Deletion of other regions might have no or little effect [35]

DNA footprinting: TFBS Identification Techniques

DNA footprinting techniques benefit from the action of nuclease enzymes. The nuclease is the class of enzymes that degrade DNA by breaking down the DNA phosphodiester bond. Based on this fact, if the DNA sequence is treated with nuclease such as DNase I, the free DNA will be digested easily whereas the DNA that is bound to a protein, also known as protected DNA, will remain intact. After digestion process, the bound protein is removed and the DNA sequences specific for the TFs are identified **(figure 14**) [35]

Chromatin Immunoprecipitation by Sequencing (ChIP-seq)

Chromatin Immunoprecipitation is most widely used technique to identify the genome-wide location of TFs. The overall steps involve in this technique are (i) Isolation of cultured cells or a particular tissue (ii) Cross-linking of DNA (iii) DNA sonication (iv) Immunoprecipitation using antibody and (v) identification of DNA-protein bound sequences either by sequencing (ChIP-seq) or DNA hybridization (ChIP-chip). The critical steps are the cross-linking of DNA within the cell. Mostly, formaldehyde is used to cross-link the TFs to the DNA sites at which they are

bound in the living cells. After that, DNA is sonicated into different fragments. The DNA fragmentation is entirely depending upon the experiments needs, sequencing system, and the basic purpose of doing the ChIP-seq. For instance, on average 200-300 base pair (bp) is preferable size for DNA fragmentation which will then proceed for immunoprecipitation. However, in some cases, it might be around 400-500 bp.

The rule of thumb is to have a decent DNA fragment neither too short, nor too long. In both cases, it will affect the results. The shorter DNA fragment can lose the TFBS, and longer DNA fragment might give false positive results. A proactive approach is to check the DNA fragments on gel to get a clear idea of the size. For the immunoprecipitation, controls (positive and negative) are crucial to get a successful result along with at least 2-3 biological replicates. After that, based on the selective techniques i.e., ChIP-seq or ChIP-chip, further steps are carried out. The former technique is followed by sequencing method (next generation sequencing) and later apply DNA hybridization concept (DNA labeling with fluorophore or radiolabeled isotopes) to identify the DNA sequences (figure 16) The sequencing data will be analyzed to find the peaks. The peaks are the area with probability of TFs bound on the genome. For that, raw sequencing reads are mapped on the reference genome (FASTQ file format; raw sequence data) and then the reads are aligned to the genome by using alignment Softwares such as Bowtie2. The next step is the peak detection (MACS software) and visualization via computational tools such as Integrated Genome Viewer (IGV) visualization. The selection of alignment software and visualization is entirely depended upon the user (**figure 17**) [35]

Conclusion and Future Directions

Recent advances accelerate the discovery of the *CRMs,* but many questions remain unanswered. The understanding of the molecular mechanism that governs direct and indirect interactions of TFs to dissect the genetic role of TFs is still under investigation. Although computational

methods are widely employed to predict potential candidates to save time and resources, the use of experimentally verified data is still under way to get maximum confidence. Apart from that, the bench-side scientist is paying more attention to produce high-quality, unbiased, reproducible datasets which can increase the success of training datasets used for *in silico CRMs* prediction. However, to gain a more complete picture of the role of these factors both in normal cellular function and in disease processes, their function to promote or inhibit gene transcription, the consequences of mutations in these *CRMs* regions, and more specifically the functionality of distal enhancers need to be examined.

The future directions may be emphasized on predicting *CRMs* by using a combinatorial methodology of conserved intra and inter-specific motifs (evolutionary signatures), epigenetic marks, landscape of histone modifications, novel and new experimental techniques, more specific and sensitive bioinformatics tools to detangle the e mystery behind these *CRMs* and their function in shaping gene expression level.

Abbreviation

ChIP-seq: Chromatin Immunoprecipitation sequencing

CRMs: Cis-Regulatory Modules

TFs: Transcription Factors

TFBs: Transcription Factors Binding Sites

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

All authors are agreed to publish this review article

Availability of data and materials

Not applicable: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Competing interests

The authors have no conflict of interest to declare.

Funding

We would like to thank The Center of Genomic Advocacy at Indiana State University and Sigma-Xi grant to support this project.

Authors' contributions

NAK and RAG conceived the ideas. RAG designed methodology. NAK led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgements

We greatly acknowledge The Center of Genomic Advocacy at Indiana State University to provide necessary support to complete this manuscript.

References

1) Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009; 461:199–205.
2) Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. *Nat Rev Genet*. 2013;14(4):288–295. doi:10.1038/nrg3458
3) David S. Latchman. Transcription factors: an overview. Current Status Review. Int. J. Exp. Path. (1993) 74, 417-422
4) Pabo, C. O., & Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annual review of biochemistry*, *61*, 1053–1095. https://doi.org/10.1146/annurev.bi.61.070192.005201

5) Radovick, S., Nations, M., Du, Y., Berg, L. A., Weintraub, B. D., & Wondisford, F. E. (1992). A mutation in the POU-homeodomain of Pit-1 responsible for combined pituitary hormone deficiency. *Science (New York, N.Y.)*, *257*(5073), 1115–1118. https://doi.org/10.1126/science.257.5073.1115

6) Crossley, M., & Brownlee, G. G. (1990). Disruption of a C/EBP binding site in the factor IX promoter is associated with haemophilia B. *Nature*, *345*(6274), 444–446. https://doi.org/10.1038/345444a0

7) Karnuta, J. M., & Scacheri, P. C. (2018). Enhancers: bridging the gap between gene control and human disease. *Human molecular genetics*, *27*(R2), R219–R227. https://doi.org/10.1093/hmg/ddy167

8) Radovick, S., Nations, M., Du, Y., Berg, L. A., Weintraub, B. D., & Wondisford, F. E. (1992). A mutation in the POU-homeodomain of Pit-1 responsible for combined pituitary hormone deficiency. *Science (New York, N.Y.)*, *257*(5073), 1115–1118. https://doi.org/10.1126/science.257.5073.1115

9) Crossley, M., & Brownlee, G. G. (1990). Disruption of a C/EBP binding site in the factor IX promoter is associated with haemophilia B. *Nature*, *345*(6274), 444–446. https://doi.org/10.1038/345444a0

10) Karnuta, J. M., & Scacheri, P. C. (2018). Enhancers: bridging the gap between gene control and human disease. *Human molecular genetics*, *27*(R2), R219–R227. https://doi.org/10.1093/hmg/ddy167

11) Cox, P. M., & Goding, C. R. (1991). Transcription and cancer. *British journal of cancer*, *63*(5), 651–662. https://doi.org/10.1038/bjc.1991.151

12) Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M., Miller, C. A., Welch, J. S.,

Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., & Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, *502*(7471), 333–339. https://doi.org/10.1038/nature12634

13) Adams, E. J., Karthaus, W. R., Hoover, E., Liu, D., Gruet, A., Zhang, Z., Cho, H., DiLoreto, R., Chhangawala, S., Liu, Y., Watson, P. A., Davicioni, E., Sboner, A., Barbieri, C. E., Bose, R., Leslie, C. S., & Sawyers, C. L. (2019). FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. *Nature*, *571*(7765), 408–412. https://doi.org/10.1038/s41586-019-1318-9

14) Cheng H, Zhang N, Pati D. Cohesin subunit RAD21: From biology to disease. Gene. 2020 Oct 20; 758:144966. doi: 10.1016/j.gene.2020.144966. Epub 2020 Jul 17. PMID: 32687945.

15) Nishijima, A., Noguchi, Y., Narukawa, K., Takano, H., & Oshikawa, G. (2019). *[Rinsho ketsueki] The Japanese journal of clinical hematology*, *60*(10), 1425–1430. https://doi.org/10.11406/rinketsu.60.142

16) Takaku, M., Grimm, S. A., & Wade, P. A. (2015). GATA3 in Breast Cancer: Tumor Suppressor or Oncogene?. *Gene expression*, *16*(4), 163–168. https://doi.org/10.3727/105221615X14399878166113

17) Gao, D., Zhu, B., Cao, X., Zhang, M., & Wang, X. (2019). Roles of NIPBL in maintenance of genome stability. *Seminars in cell & developmental biology*, *90*, 181–186. https://doi.org/10.1016/j.semcdb.2018.08.005

18) Debaugny, R. E., & Skok, J. A. (2020). CTCF and CTCFL in cancer. *Current opinion in genetics & development*, *61*, 44–52. https://doi.org/10.1016/j.gde.2020.02.021

19) Lue, J. K., & Amengual, J. E. (2018). Emerging EZH2 Inhibitors and Their Application in Lymphoma. *Current hematologic malignancy reports*, *13*(5), 369–382. https://doi.org/10.1007/s11899-018-0466-6

20) Aquila, L., Ohm, J., & Woloszynska-Read, A. (2018). The role of STAG2 in bladder cancer. *Pharmacological research*, *131*, 143–149. https://doi.org/10.1016/j.phrs.2018.02.025

21) Hill, V. K., Kim, J. S., & Waldman, T. (2016). Cohesin mutations in human cancer. *Biochimica et biophysica acta*, *1866*(1), 1–11. https://doi.org/10.1016/j.bbcan.2016.05.002

22) Dhar, S. S., Zhao, D., Lin, T., Gu, B., Pal, K., Wu, S. J., Alam, H., Lv, J., Yun, K., Gopalakrishnan, V., Flores, E. R., Northcott, P. A., Rajaram, V., Li, W., Shilatifard, A., Sillitoe, R. V., Chen, K., & Lee, M. G. (2018). MLL4 Is Required to Maintain Broad H3K4me3 Peaks and Super-Enhancers at Tumor Suppressor Genes. *Molecular cell*, *70*(5), 825–841.e6. https://doi.org/10.1016/j.molcel.2018.04.028

23) Su, J., Teichmann, S. A., & Down, T. A. (2010). Assessing computational methods of cis-regulatory module prediction. *PLoS computational biology*, *6*(12), e1001020. https://doi.org/10.1371/journal.pcbi.1001020

24) Alkema, W. B., Johansson, O., Lagergren, J., & Wasserman, W. W. (2004). MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic acids research*, *32*(Web Server issue), W195–W198. https://doi.org/10.1093/nar/gkh387

25) Grant, C. E., Johnson, J., Bailey, T. L., & Noble, W. S. (2016). MCAST: scanning for cis-regulatory motif clusters. *Bioinformatics (Oxford, England)*, *32*(8), 1217–1219. https://doi.org/10.1093/bioinformatics/btv750

26) Sinha, S., van Nimwegen, E., & Siggia, E. D. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics (Oxford, England)*, *19 Suppl 1*, i292–i301. https://doi.org/10.1093/bioinformatics/btg1040

27) Frith, M. C., Li, M. C., & Weng, Z. (2003). Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic acids research*, *31*(13), 3666–3668. https://doi.org/10.1093/nar/gkg540

28) Zhou, Q., & Wong, W. H. (2004). CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(33), 12114–12119. https://doi.org/10.1073/pnas.0402858101

29) Zhou, Qing; Wong, Wing Hung (2007). Coupling hidden Markov models for the discovery of Cis -regulatory modules in multiple species. Ann. Appl. Stat. 1, no. 1, 36--65. doi:10.1214/07-AOAS103. https://projecteuclid.org/euclid.aoas/1183143728

30) Palin, K., Taipale, J. & Ukkonen, E. (2006). Locating potential enhancer elements by comparative genomics using the EEL software. *Nat Protoc* 1, 368–374 https://doi.org/10.1038/nprot.2006.56

31) King, D. C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., & Hardison, R. C. (2005). Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome research*, *15*(8), 1051–1060. https://doi.org/10.1101/gr.3642605

32) Chan, B.Y., Kibler, D. (2005). Using hexamers to predict cis-regulatory motifs in Drosophila. *BMC Bioinformatics* 6, 262. https://doi.org/10.1186/1471-2105-6-262

33) Warner, J. B., Philippakis, A. A., Jaeger, S. A., He, F. S., Lin, J., & Bulyk, M. L. (2008). Systematic identification of mammalian regulatory motifs' target genes and functions. *Nature methods*, *5*(4), 347–353. https://doi.org/10.1038/nmeth.1188

34) Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21054/

35) Haseloff J, Siemering KR (2005) The uses of green fluorescent protein in plants. In: Green Fluorescent Protein. Wiley, pp 259–284. doi:10.1002/0471739499.ch12

36) Jefferson RA, Kavanagh TA, Bevan MW. GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J*. 1987;6(13):3901–3907.

37) Goldsbrough AP, Tong Y, Yoder JI, Tong Y (1996) *Lc* as a non-destructive visual reporter and transposition excision marker gene for tomato. Plant J 9:927–933

38) Bischof J, Maeda RK, Hediger M, Karch F, Basler K. An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. *Proc Natl Acad Sci U S A*. 2007;104(9):3312–3317. doi:10.1073/pnas.0611511104

39) A.H. Brand, N. Perrimon. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development 1993, 118: 401-415

40) Koen J. T. Venken, Hugo J. Bellen. Transgenesis upgrades for *Drosophila melanogaster* Development 2007 134: 3571-3584; doi: 10.1242/dev.00568

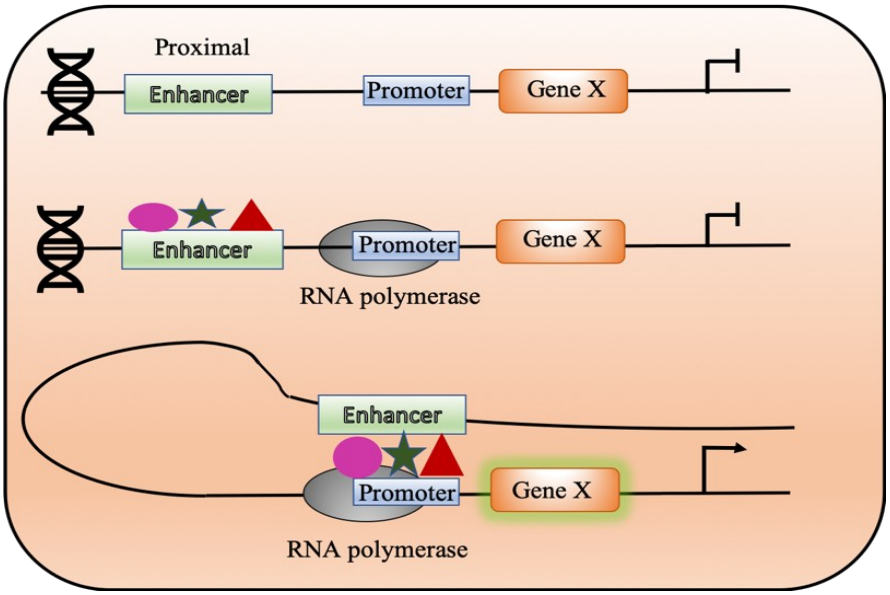
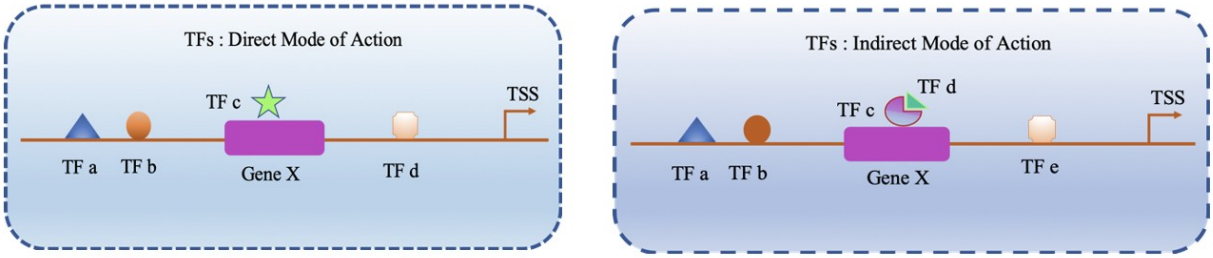


Figure 1: The 3C loop-formation of enhancer to activate/repress a gene. In first case, the proximal enhancer is present near to the gene promoter site but there is no TFs attach to the enhancer. In second case, three different TFs (circle, star, and triangle) binds to the enhancer region but unable to activate the gene expression. In this scenario, the TFs are most likely the repressor proteins which inhibit the gene expression. In third case, the TFs binds and results in a loop-like formation which enables enhancer region to come into the proximity of the promoter and RNA polymerase to activate the gene expression.



(A) **(B)**

Figure 2: A simple schematic of the direct and indirect mode of action of a TF. (A) The direct mode requires only a single TF to bind and regulate gene expression. (B) On the other hand, the

indirect mode may have multiple TFs or a single TF with multiple binding sites with other co-factors or TFs to regulate gene expression.

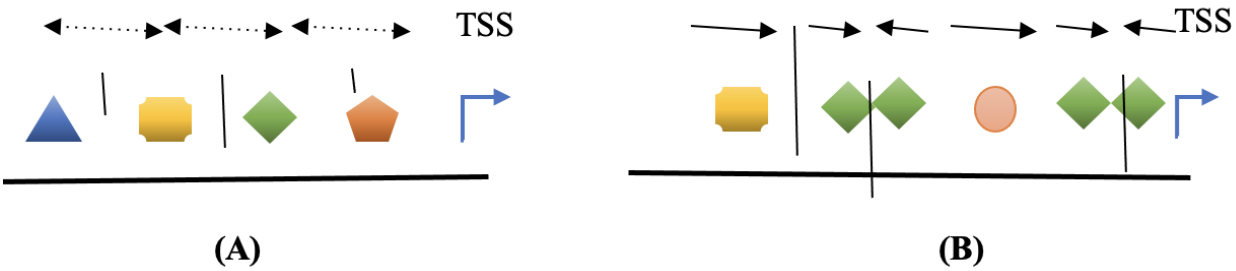


Figure 3: **(A)** shows the distance between each TFBs whereas **(B)** focuses on the orientation of TFBs. The vertical bar represents the binding affinity of each of the TF. The smaller vertical bar shows less binding affinity and vice versa.

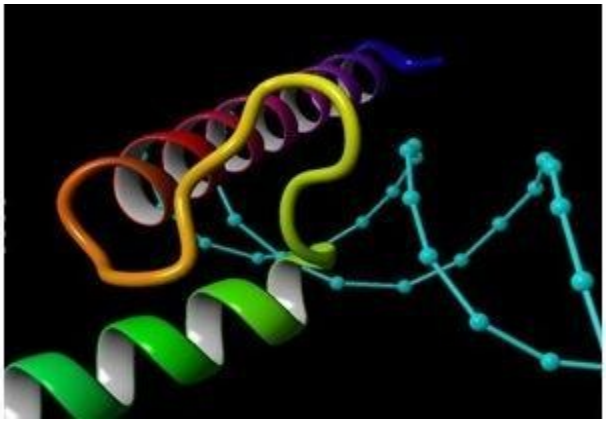


Figure 4: Helix-loop-helix

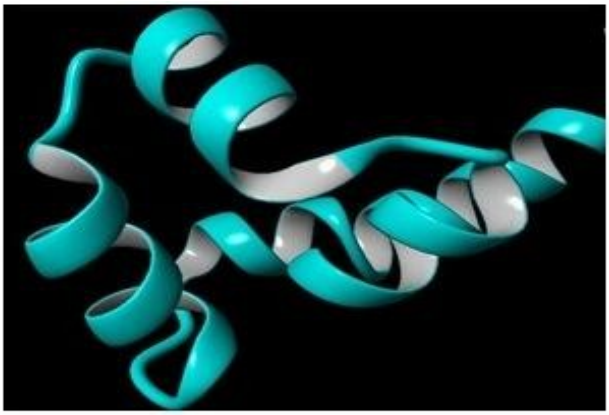


Figure 5: Helix turn Helix

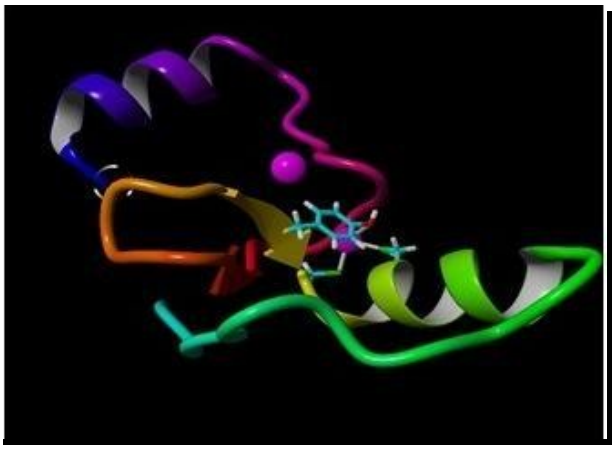


Figure 6: Zinc finger motif

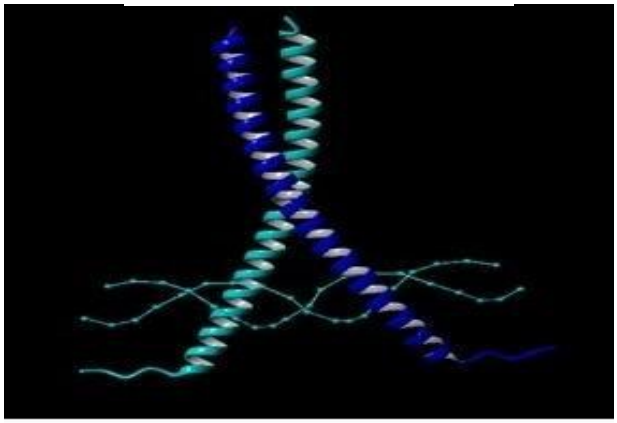


Figure 7: Leucine zipper motif

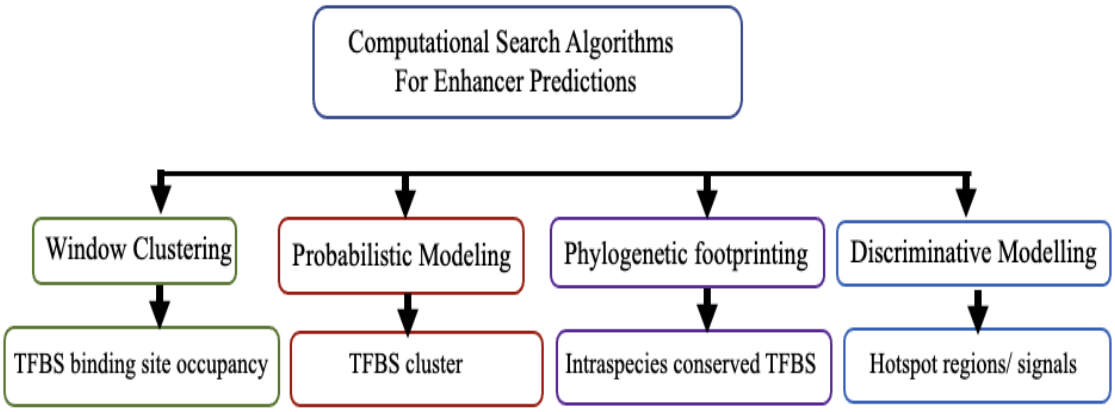


Figure 8: The common strategies used in computational search algorithm for identifying the *CRMs* regions in genome.

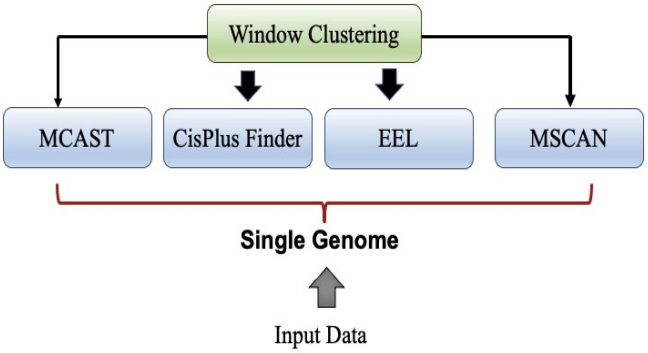


Figure 9: Window Clustering for Enhancer

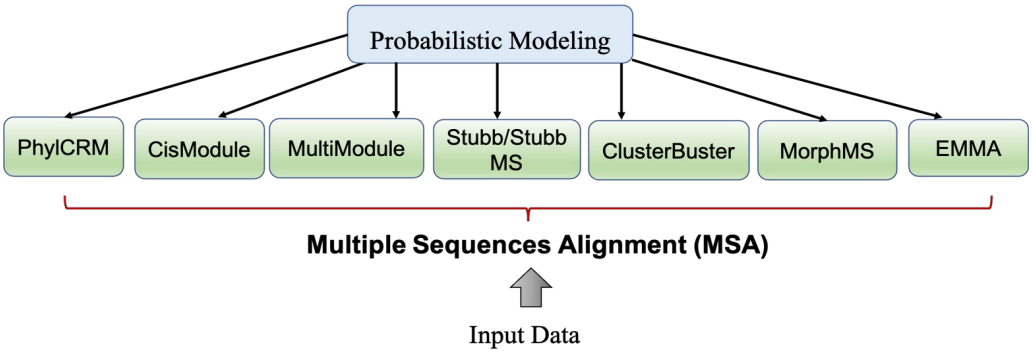


Figure 10: Different Softwares under the category of Probabilistic Modeling approach

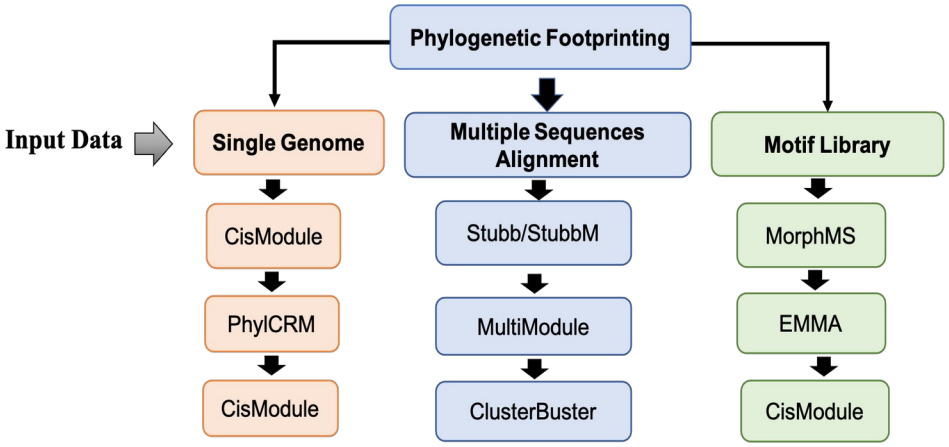


Figure 11: Phylogenetic footprinting approach with the input data type and Software for enhancer predictions

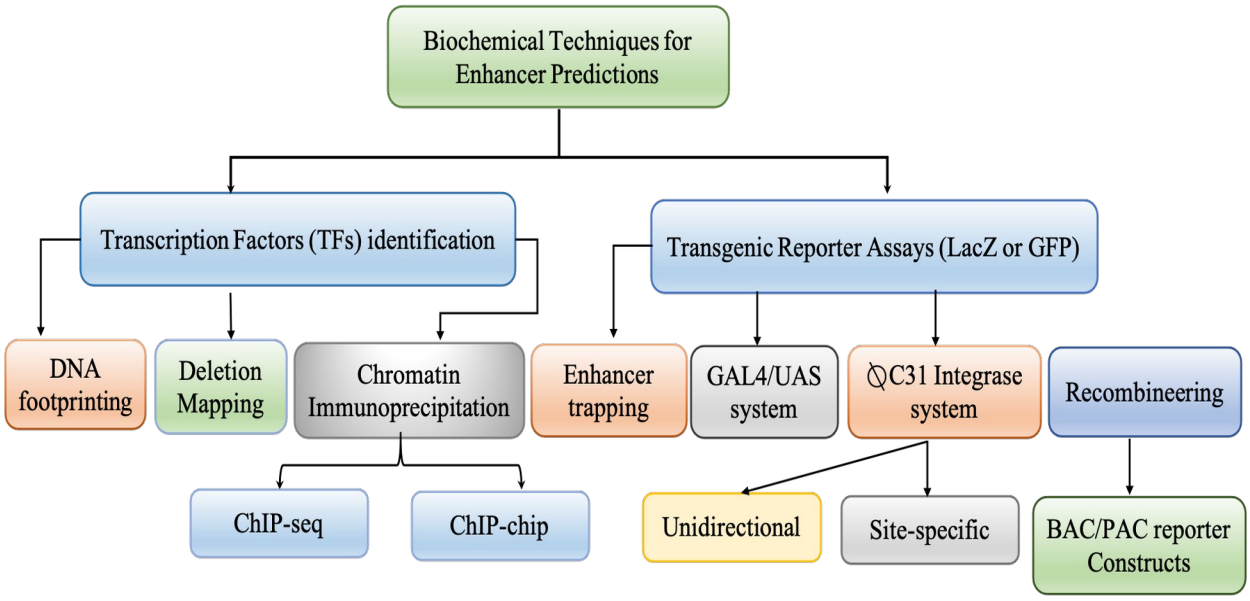


Figure 12: Biochemical methods to identify the Transcription Factor Binding sites and Enhancer regions in genome

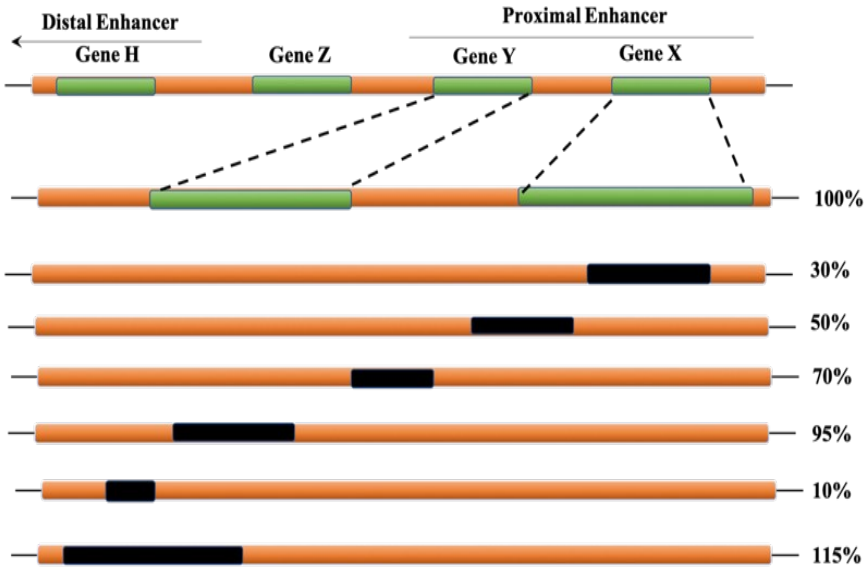


Figure 13: The deletion mapping technique apply to remove regions in gene of interest to see the effect of these deleted regions on gene expression

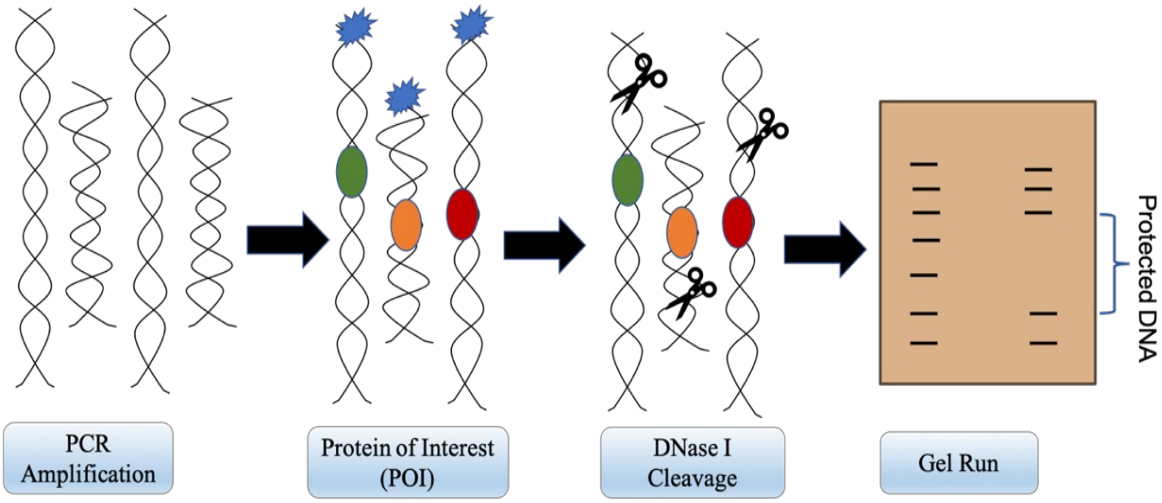


Figure 14: The DNA footprinting strategy to identify DNS-Protein binding sites.

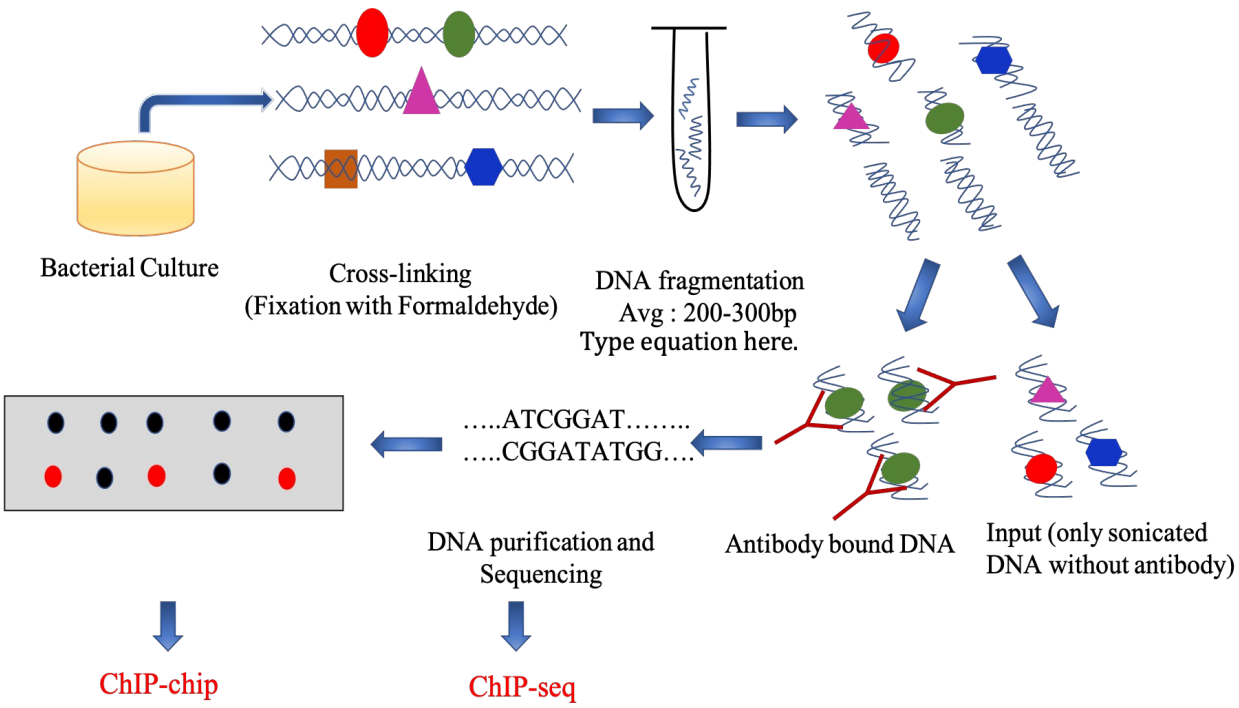


Figure 16: The ChIP -sequence procedure for identifying the TFs binding sites on enhancer.

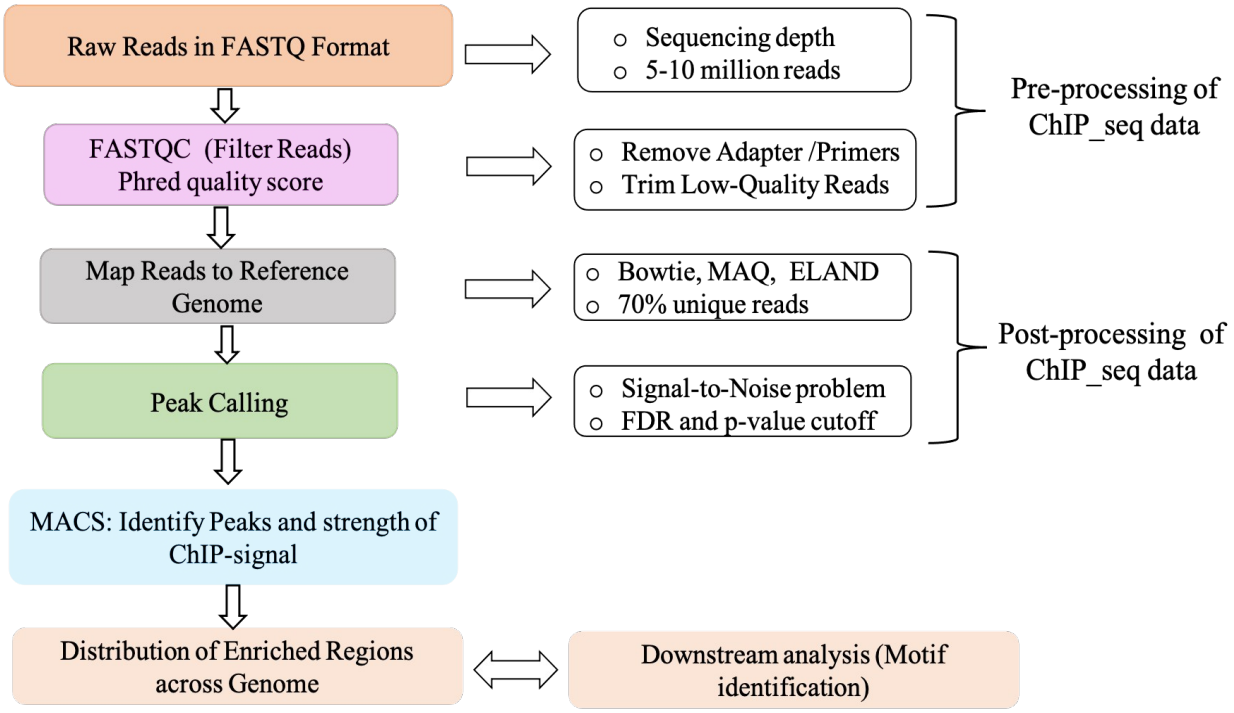


Figure 17: The post ChIP-seq data analysis pipeline

572 **Supplementary Files**

573 **Table 3: Mutations in the Enhancer Regions and their Effect on Organism Phenotype.**

Mutations	Phenotype defect	Enhancer defect	Gene ID (NCBI)	Chr	MIM record
Insertion/ Deletion	X-linked deafness type 3 (DFN3)	Multiple deletions 900 kb from *POU3F4*	5456	X	300039
	Split-hand-split food malformation (SHFM)	7q21.3 deletion affecting enhancer sequences within *DYNC1I1*	1780	7	603772
	Autosomal dominant adult-onset demyelinating leukodystrophy	Deletion eliminating *TAD*, allowing for enhancer adoption of *LMNB1*	395342 4001	24 and 5	150340
	Preaxial polydactyly	13 bp insertion in the zone of polarizing activity regulatory sequence (ZRS) affecting sonic hedgehog (SHH) expression.	64327	7	605522
Translocation	Aniridia Involves 11p13,	Involves 11p13, downstream of *PAX6*	5080	11	607108
	Pierre Robin sequence	1 Mb away from *SOX9*. Abrogates binding of *MSX1* in vitro studies.	6662	17	608160
	Split-hand syndrome	t (2; 7) (p25.1; q22), separates limb enhancers in *DYNC1I1* from *DLX5/6.*	1780	7	603772
Inversion	Limb syndactyly Hand-foot-genital syndrome	Enhancer adoption by SHH induced by a 7q inversion.	6469	7	600725
	Hand-foot-genital syndrome	syndrome Chromosome 7 inversion causing a HOXA13 enhancer	3209	7	142959

		delocalization.			
Duplication	Disorders of sex development (DSD)	16p13.3 duplication of *GNG13* and *SOX8* enhancers. 600 kb upstream of *SOX9*	51764 30812 6662	13, 8 and 9	607298 605923 608160
	Keratolytic Winter Erythema	Duplication of enhancer upstream of *CTSB*.	1508	8	116810
	Haas-type polysyndactyly and Laurin-Sandrow syndrome	Microduplications in SHH limb enhancer ZRS	6469 64327	7	600725 605522
Point Mutations	Holoproscencephaly	460 kb upstream of SHH resulting in loss of SHH brain enhancer-2 activity.	6469	7	600725
	Van der Would syndrome	Mutation in *IRF6* enhancer, abrogating p63 and E47 binding.	3664	1	607199
	Preaxial polydactyly	Various point mutations in the ZRS enhancer (e.g.,295 T>C)	64327	7	605522
	Hirschsprung disease	Common non-coding variant within an enhancer like sequence in *RET* intron 1.	5979	10	164761

574