Genomics of Regulatory Elements: A special focus on the HoxA Gene Cluster and Experimental Techniques.

Akal Sethi*[1]

[1]Department of Biochemistry, McGill University

ABSTRACT

Introduction: The regulation of gene expression plays a pivotal role in maintaining proper biological functions as well as creating cell diversity. Many regulatory elements are cis-acting and act over thousands of base pairs to affect the expression of their target gene. The development of novel techniques has provided a wealth of information into spatial chromatin organization, with the potential to unravel the mechanism of relatively uncharacterized regulatory elements, such as repressors and insulators. **Discussion**: This review will highlight examples of genomic regulation, such as Beta-globin loci and the HoxA cluster, as well as discuss advantages and disadvantages of the currently employed experimental techniques. Moreover, genomic regulation as it pertains to human disease will also be discussed.

*Corresponding author:
sethi.akal@gmail.com
Received: 2 January 2012
Revised: 2 March 2012

INTRODUCTION

Almost every component of molecular and organismal biology encompasses some sort of regulation of gene expression, often in temporal and tissue-specific manners. The organization of complex biological systems requires substantial regulatory information and capabilities. Embryonic development in any species requires the coordinated expression of a multitude of genes and their corresponding gene products to create and structure subtle gradients that lead to proper apportioning of limbs, nerves, and organs. It was postulated over 40 years ago that the level of genomic regulation was proportional to the complexity of the organism *(1)*. However, the current tools to search for elements involved in regulation are still primitive in nature and a method to easily identify genomic regulatory elements in a regular and reliable fashion has yet to be established *(1,2)*. Some regulatory elements such as promoters can be easily identified by their dutiful position at the 5' end of genes and thus thousands have been determined by standard techniques such as chromatin immunoprecipitation and cDNA sequencing *(3)*. Although promoters are relatively simple to study, other regulatory elements such as enhancers, repressors, insulators and barriers are more difficult to elucidate due to their lack of defined positions.

Although the theory of cis (intrachromosomal) regulation was favored before the human genome was sequenced, the ideas of a trans (interchromosomal) approach or a combination of the two were also put forward. Analyses of multiple loci under long range control yielded significant insight into how exactly genes were regulated in cis. One of the most characterized loci is the globin locus that lies upstream of the beta-globin gene. According to van Assendelft et al., this locus regulates beta-globin in a manner independent of the chromosomal integration site *(4)*. Understanding of locus control regions led to a greater understanding of enhancers via a combined approach of transgenic mouse assays and bacterial artificial chromosome-mediated transgenics. In the transgenic mouse assays, regions of interest were fused with a LacZ reporter gene, which allows for visualization of gene expression in temporal and spatial arrays confirming predetermined expression patterns. Deletion analysis identified enhancers spanning hundreds of base pairs within thousands of base pairs *(5,6)*. Bacterial artificial chromosomes (BAC's), originally developed to help sequence the human genome, can be used in transgenic assays to span kilobases of genomic DNA. In such instances, several BAC's can be paired together in a quantitative fashion allowing for a representation of the genomic DNA and thus provide an ideal control for experiments on genomic DNA *(7)*. Such studies culminated in the discovery of enhancers of the Bmp5 gene. These enhancers operate at large distances from the promoter (>200kbp) and were among the first distal acting regulatory elements to be discovered *(7)*. It is vital to develop a greater understanding of long-range regulation as the amount of evidence for position-effect diseases is staggering *(8)*.

These studies, conducted before full sequencing of the human genome, indicated many intrachromosomal regulatory activities that were further corroborated by more modern comparative techniques. Combinations of other vertebrate genome studies and high-throughput screens have identified novel distal regulatory sequences and are slowly coloring in the dark areas of genomic structure. This review will discuss the different classes of regulatory elements that have thus far been clearly characterized as well as the novel techniques that are providing new insights into the foundations and architecture of genomic regulation. Additionally, how regulation leads to human disease will be a point of focus. More specifically, the role of Hox gene regulation in potential malformations and disease will be discussed in greater detail, and this will serve as a primary example throughout the review.

TYPES OF DISTAL REGULATORY ELEMENTS

There have been several types of regulatory elements identified. The word "identified" is apt since techniques have proven their existence, rather than elucidated their function. Although the promoter is not a distal regulatory element in the majority of instances, its vital role in gene expression warrants its mention. As Noonan and McCallion phrase it, "the promoter is the fulcrum around which transcription pivots," in other words, the promoter is the site where basal transcriptional units such as polymerases and the overall holoenzyme form *(2)*. The promoters determine the direction and orientation of transcription as described by Maston et al. *(9)*. The long-range regulatory elements discussed in this review have been typified by their responses to certain biological assays and thus can be interpreted as broad categories rather than specifically determined elements.

Elements that positively regulate transcription are known as enhancers. Enhancers are position independent as they may be distal or proximal to the promoter of the target gene. Although the exact manner in which enhancers up-regulate transcription is still unknown, one sensible model proposes that enhancers attract transcription factors that together promote the assembly of transcription machinery at the promoter of the target gene. In this proposed model, the chromatin loops together bringing the enhancer and the promoter into proximal contact despite being base pairs away in the linear genome *(10)*. A prime example of enhancer capability is the enhancer of the Sonic Hedgehog (Shh) gene, a Hox gene. Shh is one of the genes responsible for creating the posterior-anterior axis in developing limbs in vertebrate embryos. Deletion analysis in the enhancer, approximately 1 MB from the Shh promoter in an intron of the LMBR1 gene, caused errors in Shh expression in the anterior of the limb and led to disease and malformation in the form of preaxial polydactyly *(11,12)*. Complete deletion of the enhancer resulted in a loss in expression and degeneration of limbs *(13)*. The data obtained from the Shh enhancer shows the importance of enhancers in development.

Elements that negatively regulate transcription are known as negatively-acting elements. To date, most of the regulatory elements identified are enhancers because they are easy to find with clear and straightforward biological assays. These assays include the transgenic insertion of reporter genes that can be analyzed by luciferase assays, fluorescent microscopy for GFP tagged products, or bright field microscopy of beta-galactosidase stained tissues. One of the more intriguing facts about regulatory elements such as enhancers and negatively-acting elements is their variability when exposed to different environmental stimuli. Stress, diet, hormones, temperature and lack of nutrients can all vary regulatory elements' actions through cellular signals *(14)*. This point is important to consider as it indicates that the results seen from assays measuring regular element activity have an impact on the function performed by that element. Under these circumstances, the binding of certain

transcription factors can cause various enhancers to be bound by repressor proteins and thus function as negatively-acting elements *(2,15,16)*.

Insulators, or barriers, are positioned such that the adjacent genes do not interfere with each other's expression. There are two mechanisms by which insulators can separate regulation: barrier activity or enhancer blocking (EB) activity, both of which are measured by synthetically designed assays *(17)*. Barrier activity is the ability of a sequence to create definitive borders between regions of euchromactin and heterochromatin, while EB activity limits the positive regulatory ability of enhancers by acting in a position-dependent manner. The first vertebrate insulator to be thoroughly studied and understood was the HS4 sequence in the chicken beta-globin locus that shows both barrier and EB activity in assays involving reporter genes. In this case, a protein that binds insulator sequences, the CCCTC-binding factor (CTCF), aids in EB activity *(18,19)*. It is worth mentioning that CTCF also mediates expression within the same gene locus. In the mouse beta-globin locus, CTCF binding along with the joining of certain regulatory elements yields an active chromatin hub ACH. Deletion of CTCF or its binding sites led to instability of the ACH *(20)*. The study of mouse and chicken globin loci, in combination with other vertebrate globin loci has led to a general consensus that the binding of CTCF is nearly mandatory for enhancer blocking activity in vertebrates. On the other hand, vertebrate barrier activity still seems to be relatively independent of CTCF binding suggesting that vertebrate barrier and EB activity act through independent mechanisms. However, it is still clear that insulators can be comprised of enhancer blockers, barriers, or both.

HOX A CLUSTER

Although CTCF contact is vital to proper expression and transcription for beta-globin loci, not all genes follow this trend, as some require the disintegration of contacts for activity. The HoxA locus is a prime example of such an instance. It is heavily regulated and is thus a cluster of great interest due to its constantly changing genomic architecture and function *(21)*. HoxA is part of the Hox gene family, one that is highly conserved and encodes for transcription factors responsible in the regulation of development *(22)*. Although there are 4 Hox clusters accounting for a total of 39 genes located on separate chromosomes, the HoxA cluster is one of the most characterized and pertinent and is located on chromosome 7. It codes for a total of 11 transcription factors. Hox genes obtained a claim to fame upon the discovery of how, during development, the spatial orientation of the genes on the chromosomes mimic the gene expression itself in the developing limb. In other words, Hox genes found at the 5' end of the cluster would be expressed in the posterior and much later in development than genes located at the 3' end *(23,24)*. The spatial and temporal expression results suggest that chromatin structure as well as regulatory elements play key roles in controlling Hox gene clusters *(25,26)*. Additionally, it has been found that silencing Hox genes is essential, as overexpression can lead to disease (see below). Specifically, during active transcription of HoxA genes, looping of chromatin is absent compared to its palpable existence during transcriptional silencing *(9)*. The HoxA locus provides an example of the variability of regulation when compared to the beta-globin loci in vertebrate tissues.

DETERMINATION OF CHROMATIN ORGANIZATION

Although this review focuses more on modern day techniques, it would be remiss not to mention the technique that provided much of the information regarding spatial chromatin organization. DNA fluorescence in situ hybridization (DNA-FISH) uses complementarity to hybridize a DNA probe to DNA that has been chemically fixed to a glass slide. The DNA probe contains an antigen for a fluorescent antibody, which may be visualized by epifluorescent microscopy. Using a multitude of DNA targets and different fluorescent antibodies, it is possible to determine the position of genomic locations. Although DNA-FISH has low resolution compared to more modern and advanced techniques, it is a highly accepted method in measuring *in vivo* contacts within single cells. Chromosome conformation capture (3C) has become one of the basic and most functional methods in ascertaining the organization of chromatin as well as helping to understand the connection between gene expression and organization *(27,28)*. Traditionally, 3C can be organized into five steps. The first step consists of chemically fixing cells with formaldehyde or some other crosslinking reagent to capture chromatin in its current structure and thus provide a picture of chromosomal architecture. The second step involves digestion with restriction enzymes. The use of enzymes releases pieces of DNA that were close together due to the cross linking agent. The third step consists of a ligation of the fragments favoring the ligation of ones that were cross-linked close together. The fourth step eliminates all bound proteins and contaminants resulting in a library in which products are comprised of fragments that were close together in nuclear space. The frequency of these products is inversely proportional to the ligated fragments distance in linear organization. The final step of 3C quantifies the individual ligated fragments by PCR and gel electrophoresis, or quantitative PCR and melting curve analysis *(29,30)*. Although 3C yields high-resolution results, it is relatively low throughput and covers small genomic domains. However, some of these disadvantages have been remedied in offshoot techniques of 3C.

Chromosome conformation capture carbon copy (5C) is a derivative of the 3C technique, but provides a more high throughput method *(31-34)*. A normal 3C library is generated; however, the difference lies in the way the analysis is conducted after generation of the library. In contrast to 3C, in which analysis is done by PCR gel detection, in 5C, the 3C libraries are transformed into 5C libraries and then studied by microarrays. The transformation from 3C to 5C libraries is mediated by annealing and subsequent ligation of primers that match 3C ligation points thus allowing for quantitative detection of 3C products. These "carbon copies" are amplified by PCR and then analyzed by microarrays. 5C cannot identify contacts without knowledge of the region of interest, as it requires the predicted 3C junctions in order to create a 5C library. However, this method still has significant value due to its high-throughput capability.

Hi-C is a technique created to ascertain all long-range DNA contacts at the same time *(35)*. After fixation and digestion as usual in such techniques, Hi-C deviates by filling in the overhangs created by the restriction enzymes, and labels some of the inserted nucleotide with epitopes. Normal ligation is then performed followed by sonification to minimize fragment size. The Hi-C products with epitopes are quarantined by affinity chromatography and then analyzed on high-throughput methods, allowing the mapping of all cis and trans chromosomal contacts.

LONG RANGE REGULATION IN DISEASE

Genomic rearrangements or mutations in regulatory sequences have been found to play roles in human disease *(8)*. Many cis regulatory sequences have been clinically identified by examining the DNA of patients with certain disease phenotypes. By analyzing these abnormalities at the chromosome organizational level, new insights were proposed regarding wild-type chromosomal architecture *(8)*. Mutations that disrupt promoter regulation usually lead to some form of misexpression and thus, most likely a disease *(8)*. There are certain ways in which mutations can lead to disease. For example, genetic evidence may suggest an associated with a certain disease phenotype. Similarly, structural mutations (deletions, insertions, translocations, rearrangements) may be close to a gene vital in the prevention of human disease. Furthermore, the disease phenotype could result from variation within the potential disease and such variation can account for some of the disease risk *(8,9)*. An example of a disease phenotype resulting from an anomaly in long range regulation is the deletion of an enhancer sequence approximately 900 kb upstream of the POU3F4 promoter, which leads to X-linked deafness *(36,37)*. Campomelic dysplasia can be attributed to a mutation in an enhancer of SOX9 and as stated above, preaxial polydactyly results from a deletion in the Shh enhancer *(12,38)*.

It has been found that diseases associated with the Hox cluster arise from a combination of errors in epigenetic mechanisms and errant long-range regulation *(39)*. In the case of HSC, HoxA9 is up-regulated by increased enhancer activity and decreased methylation, which in turn affects the activity of many Hox genes that play a role in acute myeloid leukemia.

CONCLUSION

Although gene regulation is commonly thought of and measured in transcriptional output, the actual transcriptional control is mediated via chromosomal structure and regulatory elements. Significant advancements have been made in the study of regulatory elements such as enhancer sequences, while repressors and insulators are not as well characterized. Hopefully, models of gene regulation such as beta-globin loci and the HoxA cluster, in combination with novel techniques such as 5C and Hi-C, will help construct a blueprint of spatial chromatin organization and identify new regulatory elements. Advances in these regards would assist in the battle against diseases involving developmental malformations and cancer by providing an understanding into the disease mechanism, thus providing a foundation for the development of a cure.

ACKNOWLEDGEMENTS

A special thank you to Dr. Josee Dostie for giving me the opportunity to be in her lab, Dr. Soizik Berlivet for patiently teaching me all the techniques and the theory behind them, and to Jennifer Crutchley, David Wang and James Fraser for answering my incessant questions.

REFERENCES

1. Britten RJ, Davidson EH, *Science*, **165**, 349–5, (1969).
2. Noonan, J. P. and A. S. McCallion, *Annu. Rev. Genomics Hum. Genet*, **11**, 1-23, (2010).
3. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM, *Genome Res*, **16**, 1–10, (2006).
4. Blom van Assendelft G, Hanscombe O, Grosveld F, Greaves DR, *Cell*, **56**, 969–77, (1989).
5. Frasch M, Chen X, Lufkin T, *Development*, **121**, 957–74, (1995).
6. Malicki J, Cianetti LC, Peschle C, McGinnis W, *Nature*, **358**, 345–47, (1992).
7. DiLeone RJ, Marcus GA, Johnson MD, Kingsley DM, *Proc. Natl. Acad. Sci. USA*, **97**, 1612–17, (2000).

8. Kleinjan DA, Lettice LA, *Adv. Genet*, **61**, 339–88, (2008).
9. Maston GA, Evans SK, Green MR, *Annu. Rev. Genomics Hum. Genet,* **7**, 29–59 (2006).
10. Dekker J, Rippe K, Dekker M, Kleckner N, *Science*, **295**, 1306–11, (2002).
11. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, *et al*, *Hum. Mol. Genet*, **12**, 1725–35, (2003).
12. Lettice LA, Horikoshi T, Heaney SJH, van Baren MJ, van der Linde HC, *et al*, *Proc. Natl. Acad. Sci. USA*, **99**, 7548–53, (2002).
13. Sagai T, Masuya H, Tamura M, Shimizu K, Yada Y, *et al*, *Mamm*. Genome, **15**, 23–34, (2004).
14. McGrane MM, *J. Nutr. Biochem*, **18**, 497–508, (2007).
15. Murayama A, Kim M-S, Yanagisawa J, Takeyama K-I, Kato S, *EMBO* J, **23**, 1598–608, (2004).
16. Perissi V, Aggarwal A, Glass CK, Rose DW, Rosenfeld MG, *Cell*, **116**, 511–26, (2004).
17. Gaszner M, Felsenfeld G, *Nat. Rev. Genet*, **7**, 703–13, (2006).
18. Crutchley, J. L., X. Q. D. Wang, *et al, Biomarkers in Medicine*, **4**, 611-629, (2010).
19. Chung JH, Whiteley M, Felsenfeld G, *Cell*, **74**, 505–1, (1993).
20. Splinter E, Heath H, Kooren J *et al, Genes Dev*, **20**, 2349–2354, (2006).
21. Fraser J, Rousseau M, Shenker S *et al*, *Genome Biol*, **10**, R37, (2009).
22. Krumlauf R. *Cell*, **78**, 191–201, (1994).
23. Kmita M, Duboule D, *Science*, **301**, 331–333, (2003).
24. Duboule D, Morata G, *Trends Genet*, **10**, 358–364, (1994).
25. Morey C, DA Silva NR, Perry P, Bickmore WA, *Developmental*, **34**, 909–919, (2007).
26. Chambeyron S, Bickmore WA, *Genes Dev*, **18**, 1119–1130, (2004).
27. Miele A, Dekker J, *Methods Mol. Biol*, **464**, 105–121, (2009).
28. Miele A, Gheldof N, Tabuchi TM, Dostie J, Dekker J, *Current Protocols in Molecular Biology*. (Wiley, Hoboken, 2006).
29. Hagege H, Klous P, Braem C *et al*, *Nat. Protoc*, **2**, 1722–1733, (2007).
30. Abou El Hassan M, Bremner R, *Nucleic Acids Res*, **37**, E35, (2009).
31. Dostie J, Richmond TA, Arnaout RA *et al*, *Genome Res*, **16**, 1299–1309, (2006).
32. Dostie J, Zhan Y, Dekker J, *Current Protocols in Molecular Biology* (Wiley, Hoboken, 2006)
33. Dostie J, Dekker J, *Nat. Protoc*, **2**, 988–1002, (2007).
34. Van Berkum NL, Dekker J, *Methods Mol. Biol*, **567**, 189–213, (2009).
35. Lieberman-Aiden E, Van Berkum NL, Williams L *et al*, *Science*, **326**, 289–293, (2009).
36. de Kok YJ, Merkx GF, van der Maarel SM, Huber I, Malcolm S, *et al*, *Hum. Mol. Genet*, **4**, 2145–50, (1995).
37. de Kok YJ, Vossenaar ER, Cremers CW, Dahl N, Laporte J, *et al*, *Hum. Mol. Genet*, **5**, 1229–35, (1996).
38. Pfeifer D, Kist R, Dewar K, Devon K, Lander ES, *et al*, *Am. J. Hum. Genet*, **65**, 111–24, (1999).
39. Barber, B. A., M. Rastegar, *Annals of Anatomy - Anatomischer Anzeiger* **192**, 261-274, (2010).
40. J. Borrow, A.M. *et al*, *Nat. Genet*, **12**, 159–167, (1996).
41. B. Argiropoulos, R.K. Humphries, *Oncogene*, **26**, 6766–6776, (2007).