AIMS Biophysics, 2(4): 773-793.
DOI: 10.3934/biophy.2015.4.773
Received date 12 August 2015,
Accepted date 07 October 2015,
Published date 19 November 2015

http://www.aimspress.com/

Review

Exploring function of conserved non-coding DNA in its chromosomal context

Delores J. Grant [1], Leighcraft A. Shakes [2], Hope M. Wolf [2], Derek C. Norford [2], and Pradeep K. Chatterjee [2,3,*]

[1] Department of Biological and Biomedical Sciences and Cancer Research Program, Julius L. Chambers Biomedical/Biotechnology Research Institute, North Carolina Central University, Durham, USA
[2] Julius L. Chambers Biomedical/Biotechnology Research Institute, North Carolina Central University, Durham, USA
[3] Department of Chemistry and Genomics Program, North Carolina Central University, 700 George Street, Durham, NC 27707, USA

* **Correspondence:** Email: pchatterjee@nccu.edu; Tel: +919-530-7017.

Abstract: There is renewed interest in understanding expression of vertebrate genes in their chromosomal context because regulatory sequences that confer tissue-specific expression are often distributed over large distances along the DNA from the gene. One approach inserts a universal sensor/reporter-gene into the mouse or zebrafish genome to identify regulatory sequences in highly conserved non-coding DNA in the vicinity of the integrated reporter-gene. However detailed mechanisms of interaction of these regulatory elements among themselves and/or with the genes they influence remain elusive with the strategy. The inability to associate distant regulatory elements with the genes they regulate makes it difficult to examine the contribution of sequence changes in regulatory DNA to human disease. Such associations have been obtained in favorable circumstances by testing the regulatory potential of highly conserved non-coding DNA individually in small reporter-gene-containing plasmids. Alternative approaches use tiny fragments of chromosomes in Bacterial Artificial Chromosomes, BACs, where the gene of interest is tagged *in vitro* with a reporter/sensor gene and integrated into the germ-line of animals for expression. Mutational analysis of the BAC DNA identifies regulatory sequences. A recent approach inserts a sensor/reporter-gene into a BAC that is also truncated progressively from an end of genomic insert, and the end-deleted BAC carrying the sensor is then integrated into the genome of a developing animal for expression.

The approach allows mechanisms of tissue-specific gene expression to be explored in much greater detail, although the chromosomal context of such mechanisms is limited to the length of the BAC. Here we discuss the relative strengths of the various approaches and explore how the integrated-sensor in the BACs method applied to a contig of BACs spanning a chromosomal region is likely to address mechanistic questions on interactions between gene and regulatory DNA in greater molecular detail.

Keywords: GROMIT strategy; functional non-coding DNA; scanning BACs with enhancer-traps; regulation of amyloid precursor protein gene expression in zebrafish and Humans; lox-Cre recombination in BACs

1. Introduction

A tiny fraction of the DNA in our chromosomes, around 1%, actually codes for proteins and are represented as genes. As one might expect, the sequence of this DNA is mostly conserved among vertebrates because the proteins encoded by these genes perform largely similar functions in the cell. A far more surprising conclusion arose from the massive worldwide effort to sequence the entire genome of humans and several other vertebrates. They suggest a significant fraction of the remaining DNA in our chromosomes that do not code for proteins, so called non-coding DNA, is also highly conserved among vertebrates [1–4]. Some of these highly conserved non-coding sequences between human and fish are more conserved than even a few protein coding sequences between the two very divergent species [2]. Much of this highly conserved non-coding DNA in vertebrates is thought to regulate the expression of genes important during development of the embryo [2–7]. Expression of such genes is restricted to one or a few specific tissues in the developing animal, with the regulatory sequences often located tens of thousands of base pairs away from the coding sequences of the genes they control. Thus understanding the regulation of vertebrate genes in their chromosomal context is important to decipher the underlying mechanisms of their tissue specific expression.

Expression of a gene in vertebrates can be regulated at multiple levels such as; when the primary transcript from the DNA in chromosomes is made; during the complex multi-step processing of this primary transcript to generate mRNA; while translating that mRNA to make proteins in ribosomes; among the numerous post-translational processing steps of the proteins to their mature forms; and during the ultimate destruction of proteins after their useful life in the cell. We focus our attention here to regulation of gene expression at the transcriptional level, where the cellular machinery needs to interact with the DNA existing as nucleoprotein complexes termed chromatin in the chromosomes of a cell.

Much of the DNA in chromosomes of cells exists in a highly compacted state and appears inaccessible for expression. In eukaryotes this compaction starts from repetition of a primary unit, the nucleosome, comprised of 147 bp of double stranded DNA wrapped one and three-quarter turns around a complex consisting of two copies of four different histone proteins, H3, H4, H2A and H2B [8,9]. Inter-nucleosomal DNA is condensed by binding histone H1 or H5, while the nucleosome-bound DNA is further compacted by numerous scaffolding proteins that facilitate the formation of higher orders of tertiary structure shown schematically in Figure 1. Shape-specific

protein-bridging interactions appear to facilitate the long-range pairing of segments of chromatin [10,11]. In prokaryotes, HU proteins package the DNA which appears more accessible to the transcription machinery, but the higher order packaging of the nucleoprotein is less well known.

Based on staining patterns of eukaryotic nuclei, the DNA in chromosomes is classified as heterochromatin which comprises largely of gene-poor regions, while the gene-rich regions exist as euchromatin. These domains may occupy discrete territories in the nucleus, and appear to be non-randomly organized [12]. Numerous biochemical studies that include cutting and rejoining DNA domains in close proximity in the intact cross-linked nucleus also substantiate this non-random organization [13]. The highly compacted DNA in chromatin is mostly inaccessible to the cellular machinery responsible for transcribing and expressing the genes encoded in them. For transcribing, the DNA needs to be de-condensed to a more accessible form and cells have evolved elaborate pathways to convert this "inactive state" chromatin to a transcriptionally active one through a variety of biochemical modifications to the histones and/or the DNA. This is illustrated in Figure 1, schematically. Chromatin activation often involves protein complexes recognizing specific sequences on the DNA, binding to them and initiating biochemical modifications of the histones and sometimes the DNA. The modifications facilitate nucleosomes bound to gene control regions to fall off or assume a more open configuration. This event in turn enables general transcription factors of the basic transcription machinery, and RNA polymerase, to bind to the basal promoter of a gene and initiate transcription.

Cells in multicellular organisms are genetically homogeneous, but structurally and functionally heterogeneous. The heterogeneity can be traced to differential expression of genes in different tissues, starting during development and retained through cell division [14]. Stable alterations such as these are termed "epigenetic", because they are heritable in the short term but do not involve mutations of the DNA itself. Molecular mechanisms that mediate epigenetic regulation of gene expression include DNA methylation and several types of histone modifications. They constitute very important mechanisms for gene activation in cells undergoing differentiation [15]. For example, a chromatin region with a gene exhibiting tissue-specific expression can have its histones methylated to different extents by specific histone methylases to propagate the "off" state in condensed inactive chromatin, indicated by filled red circles in Figure 1. Upon activation, in response to an external influence, specific de-methylases help remove the methyl groups on the histones to de-condense the chromatin and enable expression of the gene [16,17,18]. Similarly, histone acetylases help put acetyl groups onto specific histones; and these often serve as markers of transcriptionally active gene regions of chromatin, indicated by the filled blue stars [19]. An intricate balance between histone acetylases and de-acetylases fine tunes transcriptional activity of genes.

The DNA between multiple control regions, bound with sequence-specific transcription-enhancing proteins, is not devoid of nucleosome binding, as illustrated in Figure 1. It is likely that distances along the DNA between sites that bind transcription factors, enhancing and/or repressing have evolved along with sequences, to optimize protein-protein and protein-DNA interactions critical during transcription. Appropriate phasing of binding sites of regulatory proteins, as well as of nucleosomes, along the DNA would facilitate these interactions as suggested and extensively studied [11,20,21]. Thus considering the complexity of interactions mediating tissue-specific gene expression, one wonders whether such protein-protein-DNA contacts can be duplicated in small plasmids that are often constructed from joining just the DNA-binding sites of these regulatory proteins in an attempt to express the gene out of its chromosomal context. It also highlights the

importance of sequences flanking the actual binding sites of regulatory proteins studied earlier [22,23,24]. Thus it is probably not surprising to find the functional characteristics of activator sequences to be influenced by the environment in which they find themselves [25]. The real challenge from the very onset of "recombinant DNA technology" has been the expression of vertebrate genes in a manner representative of their endogenous counterpart. Endogenously, genes exist in their chromosomal contexts with respect to not only sequences surrounding them over large distances along the DNA but also packaged in their chromatin constitutive state.

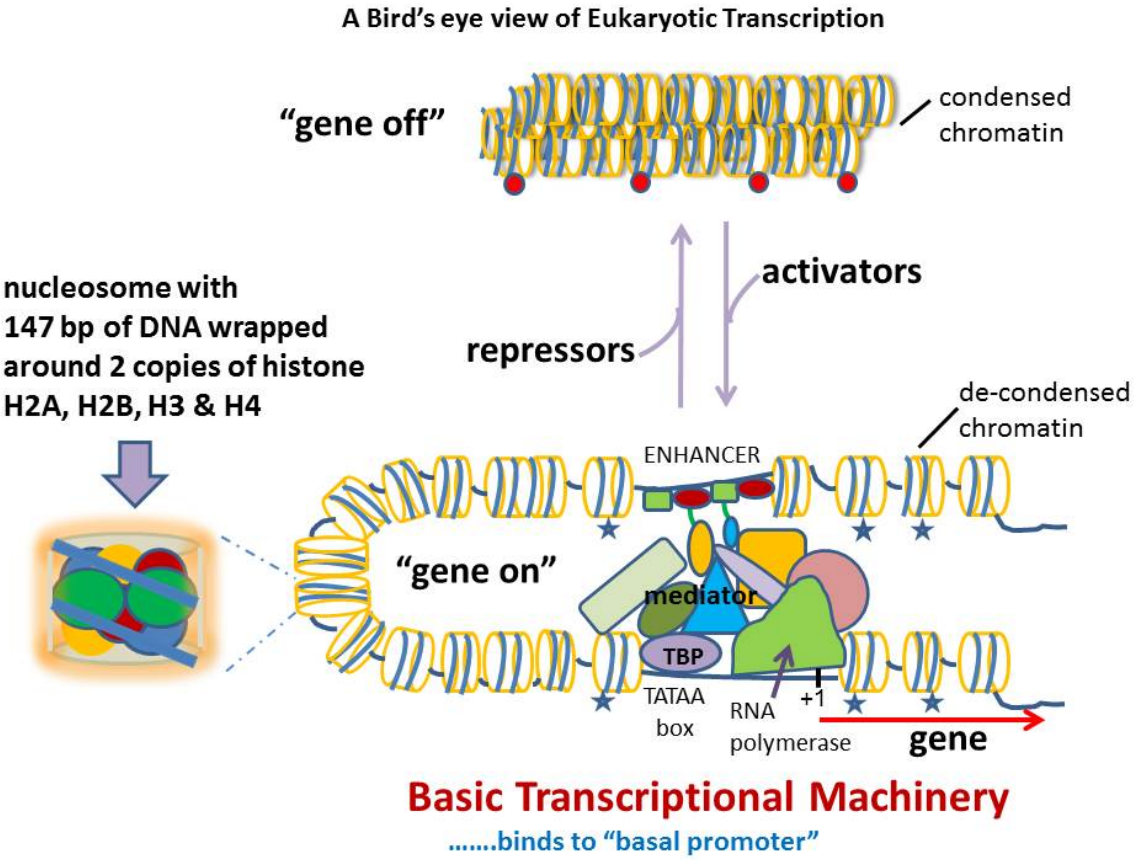

Figure 1. The nucleosome is shown schematically on the left. It contains a core of histone proteins, two each of histones H2A, H2B, H3 and H4, shown as spheres colored green, red, orange and blue. Double-stranded DNA, 147 bp, shown in blue, wraps around this core of histone proteins. The highly compacted "condensed chromatin" is acted upon by "activators" to "de-condense" it and turn on gene transcription. Highly complex protein-protein interactions (proteins indicated by the different colored shapes) mediated with or without DNA are illustrated. Specific sequences on the DNA in open regions of chromatin recruit and position "general transcription factors" along with the basic transcriptional machinery. Histones of nucleosomes in "gene on" regions of de-condensed chromatin are de-methylated and often modified by histone acetylation (indicated by filled blue stars). Histones in "gene off" regions found in condensed chromatin are methylated and de-acetylated (shown by filled red circles).

2. Strategies for Functional Identification of Non-coding Regulatory DNA

2.1. Piecemeal and "out-of-context" approaches

Analysis of DNA sequence of entire genomes from numerous organisms indicates that vertebrate genomes have expanded compared to lower eukaryotes such that regulatory sequences of genes have often become separated by large distances along the DNA from their coding sequences. This exacerbates the problem of expressing vertebrate genes in their chromosomal contexts owing to inadequate technology available. The easy way out was to first identify potential gene-regulatory sequences by their high degree of homology among vertebrate genome sequences, from species as divergent as the human and zebrafish, and then test them individually for their regulatory activity [2–7]. Thus the regulatory role of large sets of highly **c**onserved **n**on-coding DNA **e**lements, also referred to as CNEs, have been identified by their influence on the expression of an easily detectable reporter-gene. Such a reporter-gene cannot be expressed and remains undetectable when there are no transcription-enhancing CNE-sequences around. However in close proximity of a CNE on the same DNA molecule, the reporter gene expresses itself and is easily detectable. Reporter genes such as the Green Fluorescent Protein (GFP) gene, or the *lacZ* gene that is easier to detect after expression in opaque tissue, was fused to the CNE in a small circular plasmid DNA and expressed in the developing animal. The small plasmid DNA was injected into fertilized eggs and expressed in one of two ways: either transiently without integrating it into the genome of the zebrafish [2–5], or after integration into the germ-line of the mouse [6]. In both cases however, the regulatory function of the non-coding DNA was determined individually, and devoid of the context of all other regulatory DNA in the chromosome. Many of the CNEs identified this way are found to regulate expression of genes critical to development of the embryo, and are sometimes found mutated in human diseases.

2.2. Using the Entire Genome Approach

The past decade has seen a renewed emphasis on analyzing the function of gene regulatory sequences in their chromosomal environment than in isolation [26–30]. A reporter-gene probe is inserted into the genome of the animal and regulatory characteristics of the DNA surrounding the inserted reporter-gene probe are analyzed. For example, the identification and functional characteristics of highly conserved non-coding DNA across vertebrate species have been explored recently using the "Genome Regulatory Organization Mapping with Integrated Transposons (GROMIT)", strategy [29,30]. In this approach, the *Sleeping Beauty* transposon, a mobile genetic element, is inserted randomly into the mouse genome by injecting the transposon DNA into fertilized eggs. The *Sleeping Beauty* transposon functions only in vertebrates. Like other transposons, it can carry exogenous DNA and integrate it into the genome of an animal. Its "cargo" includes the *LacZ* reporter-gene (termed as "sensor" in the report and schematically shown in Figure 2A). The reporter-gene/universal-sensor comprises of a gene that senses the regulatory environment of the DNA where it is inserted. It has a short **b**asal **p**romoter (BP) with no activity by itself but expresses, and is very sensitive to, transcription enhancing regulatory influences in its chromosome-inserted vicinity [28,29,30]. Therefore, this system helps determine the net regulatory input acting on a given genomic position where the *Sleeping Beauty* transposon is inserted. Because the reporter-gene/sensor

is incorporated covalently into the mouse genomic DNA, it measures the integrated regulatory activity of all elements (activating and repressing) acting on that position; and most importantly in their chromosomal contexts. This distinguishes GROMIT from other studies that employ reporter-gene assays using small plasmids, as described in the previous paragraph. The latter were used to test individual CNEs, chosen by cross-species genome sequence comparisons, in isolation. Additionally, the CNE-reporter gene fusions were in small plasmids, and in the case of mouse or frog or zebrafish embryos, integrated at random locations in the genome [6,7,31]. Methods similar in concept to GROMIT but using different transposable elements have been described in earlier reports using the zebrafish system [26,27,31].

Almost half of over 500 insertions isolated using the GROMIT strategy were considered independent, the remainder ascribed to the anomalous activity, characteristic of *Sleeping Beauty* transposon, known as 'local hopping' [28,29]. Little over a hundred mouse transgenic lines were established for analyses of tissue-specific transcription enhancing sequences, also known as enhancers. These analyses of the mouse genome revealed that the activities of enhancers are distributed over large intervals along the DNA, forming broad regulatory landscapes, which extend far away from genes. Histone de-methylases could play a role in generating such landscapes of regulatory activity [15–18]. Substantial interplay between enhancers was also observed using this approach [30]. These findings are in sharp contrast to those determined earlier for isolated single enhancer elements, CNEs, in reporter assays, where potential activities may have been silenced or repressed due to their loss of context. The different conclusion from the two approaches was expected, and highlights the results of recent findings: a detailed analysis of the tissue-specificity of an isolated enhancer of the zebrafish *appb* gene, tested out of its native context, was determined to be quite different from that in the context of its own gene [25,32].

2.3. Drawbacks of Entire Genome Approaches: Traditional Enhancer-Trapping/integrated-reporter/ integrated-sensor/GROMIT strategies

Traditional enhancer-trapping/integrated-reporter/sensor strategies, including GROMIT, have no doubt offered high-throughput screening formats to analyze highly conserved non-coding DNA functionally in their chromosomal contexts, helped identify many tissue-specific enhancers and isolate numerous reporter-gene expressing transgenic lines in the animal systems used for such studies. However a potential drawback of the method arises from the requirement that the trap/ sensor be injected into the fertilized egg at the one-cell stage, or its equivalent, to integrate it into the germ-line. These strategies rely on expression of the reporter-gene/ sensor from integrated DNA copies in the germ-line. It is unclear how much of the genome is accessible for insertion of the sensor/trap at this stage of development of the embryo and constitutes a potential hurdle. It is therefore not surprising to note that although sequence comparisons of genomes across species have suggested the existence of between 1400 and 3100 highly conserved sequence elements [2,5,6], those actually analyzed either by enhancer-trapping in zebrafish, or the mouse using GROMIT range between 95 [27] and 165 [29], respectively.

Ideally, one would want insertions of the *Sleeping Beauty* transposon, carrying the reporter gene/sensor, to be completely free of location-bias and random on the DNA. It is also important to remember that one can have only one insertion of the reporter gene/sensor/enhancer-trap per genome in order to keep the analysis unambiguous, thus precluding saturation with sensor insertions for

complete coverage of difficult-to-insert locations on the genome. Consequently regulators of tissue-specific gene expression in conserved non-coding sequences have been identified in only a small fraction of the genome in animal systems studied, and vast regions appear refractory to probing by this approach, as also suggested in the previous paragraph. Additionally, the 50 bp globin minimal promoter may not interact with all regulatory elements in non-coding DNA in a manner reminiscent of the endogenous gene: ideally one would want to have the same basal promoter of the gene(s) that is/are influenced by one or more of the regulatory elements as in the animal. Lastly, approaches such as these do not lend themselves easily to identifying and/or addressing questions on mechanisms of how multiple enhancers/repressors, in domains that may be discontinuous along the DNA, act in concert to restrict expression of the gene in a particular tissue. Detailed mechanistic interpretations of enhancer(s) function are thus not possible from these low-resolution analyses on the entire genome.

2.4. Bacterial Artificial Chromosomes, (BACs), to the rescue [33]

In this report we highlight an alternate way, using essentially the same tools, for example, insertion of reporter-gene/sensor into fragments of chromosomes instead of the entire genome of the animal, such as insert the sensor into ~ 300 kbp of contiguous DNA sequence from the chromosome of an animal cloned and faithfully propagated in bacteria as Bacterial Artificial Chromosomes**,** (BACs). Because the DNA in chromosomes of animals is very large, in the order of several millions of base pairs in length, it becomes quite difficult to carry out biochemical experiments with it. Thus the entire DNA of an organism has been fragmented into overlapping ~ 300 kbp pieces, with high redundancy, and cloned in vectors that can propagate these large pieces of DNA very efficiently and with high fidelity in bacteria. Such propagation of large pieces of an organism's DNA in bacteria occurs extra-chromosomally and as a library of several hundred thousand individual clones, each of which carries a unique 300 kb piece of DNA from a chromosome of the organism. Such libraries are called BAC libraries. BAC libraries have been constructed from the DNA of numerous organisms and are available commercially.

Insertion of a reporter-gene/sensor can readily be made into BAC DNA using a transposon that works in bacteria. The sensor-integrated BAC DNA is then purified and injected into the fertilized egg for integration into the genome of the animal. The strategy readily allows analyses of mechanisms of regulation by conserved non-coding DNA (illustrated in Figure 2B, 2C). Not only is the actual methodology far less resource-intensive, as all of the insertions are performed in a bacterial host, but it also has the potential for much greater coverage of the genome because individual BACs, in contrast to the entire genome of the animal, can be made to have at least one insertion of reporter gene/sensor using the bacterial transposon Tn10. Because the approach is amenable to additional mutational analyses (see reference [25] for details), it helps to functionally characterize transcription enhancing sequences at a much higher resolution and address questions on mechanisms of gene enhancer interactions in far greater detail. Additionally, the approach allows flexibility in selecting the basal promoter driving the reporter gene/sensor from knowing all gene(s) in the BAC that are likely to be influenced by the regulatory DNA.

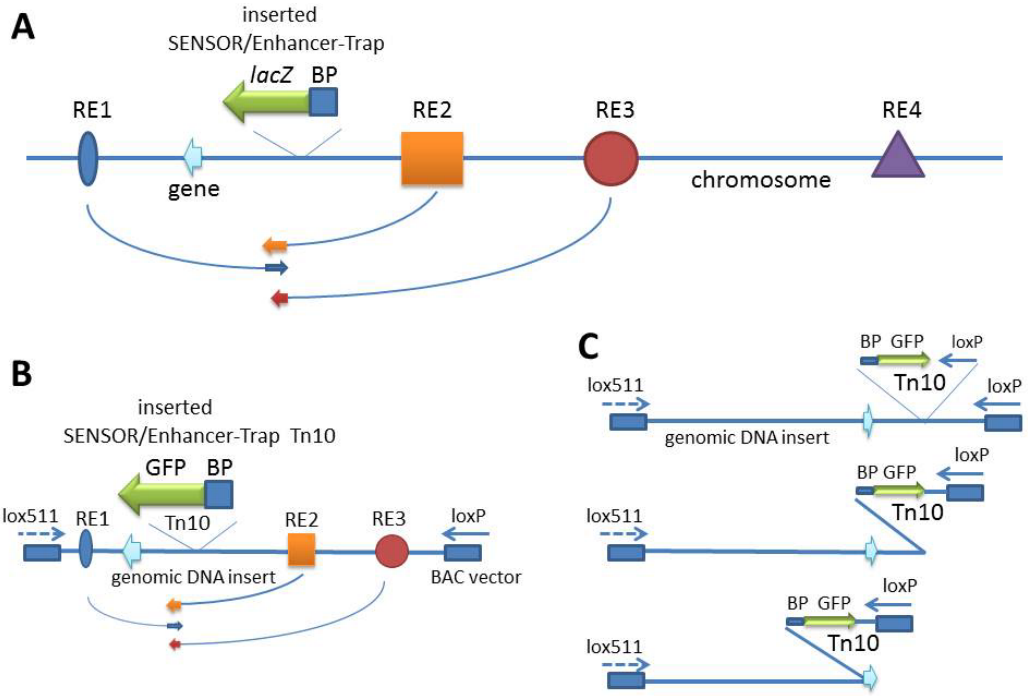

Figure 2. Panel A: The "Genome Regulatory Organization Mapping with Integrated Transposons (GROMIT)", strategy as outlined in references [29,30]. A basal promoter (BP) from the human β-globin gene is fused to a promoter-less *lacZ* gene, the combination serves as the reporter-gene/sensor. This reporter-gene/sensor cassette is carried by the *Sleeping Beauty* transposon and inserted into the DNA of the mouse genome. The inserted BP-*lacZ* gene is not expressed unless acted upon by transcription activating sequences in conserved non-coding DNA and collinear with it. RE1, RE2, RE3 and RE4 represent Regulatory Elements in non-coding DNA upstream, (RE2-RE4), or downstream, (RE1), of the gene of interest, shown as thick/fat arrowhead in light blue. One or more of these RE's may regulate the gene, either by enhancing or repressing, or both. Each RE site on DNA may be bound specifically by a complex array of proteins, through protein-DNA and Protein-Protein interactions. Expression of the integrated BP-*lacZ* gene in mouse embryos thus scores for activator sequences interacting with the reporter/sensor. The site of integration of *Sleeping Beauty* transposon carrying reporter/sensor in the mouse genome can be determined. Panel B: The integrated-reporter-gene/sensor in the BAC approach is shown schematically. A promoter-less Green Fluorescent Protein (GFP) gene is fused to a basal promoter (BP) of the gene(s) being evaluated in the set of BACs and serves as reporter-gene/sensor. This is fused to the DNA of a Tn10-loxP transposon (a jumping gene, mobile DNA element specific to bacteria). The Tn10 transposon is then inserted this time into the BAC DNA in its bacterial host. The BAC DNA with the integrated-reporter/sensor is then expressed by injecting it into fertilized zebrafish eggs as described in reference [50]. The BP-GFP gene inserted into BAC DNA is not expressed or detectable as green fluorescence unless acted upon by transcription activating sequences elsewhere in the BAC DNA. Expression of BP-GFP gene therefore scores for interacting activator sequences on

the same BAC DNA. Panel C: Positioning of BP-GFP sensor integrated into BAC DNA. The loxP Tn10 transposon carrying the BP-GFP gene is first inserted into BAC DNA and then a Cre-mediated recombination between the loxP in Tn10 and the one endogenous to the BAC deletes the DNA between the end of genomic-insert and the site of insertion of Tn10. Thus the integrated-reporter/sensor is placed at the new end of genomic-insert created by the deletion of DNA between the loxP sites [50].

3. Potential Hurdles to Exploring Regulation of Gene Expression in Higher Vertebrates

Regulation of gene expression in higher vertebrates is complex. Genes can be regulated by a variety of mechanisms/pathways such as regulatory protein binding to specific sequences, non-coding RNA binding, splicing elements and sequences that mark chromatin structure, histone modifications in chromatin, etc. We confine our discussion here only to those DNA sequences that could participate in a subset of the mechanisms outlined, including binding regulatory proteins that direct expression of a gene often in a tissue- and time-specific manner.

As noted earlier, genomes of vertebrates have expanded during evolution, separating many transcription factor binding sites from one another, and from the start sites of genes, by large distances along the DNA. It prevents *cis*-acting regulatory sequences to be housed with the gene(s) they influence in traditional small plasmids. This makes the much larger sized BACs, which faithfully propagate approximately 300 kb of DNA, an ideal vehicle for accurate expression of the genes contained in them, because they represent tiny pieces of the animal's chromosome and are likely to house both the gene as well as the regulatory element(s) in many cases [2]. However they are quite difficult to manipulate at the bench because of their size. Entirely different technology was therefore developed in the past decade and a half to modify and engineer BACs for a variety of studies. These procedures differ from earlier recombinant DNA technology developed over a quarter century ago for "small DNA" in that they use the reciprocal exchange of sequences between two double stranded DNA molecules in a concerted manner, termed DNA recombination, to alter sequence in a BAC, instead of the traditional 'cut-and-paste' mechanisms used for small plasmids. These recombination procedures can be divided broadly into two categories: 1) those that require sequence homology between a vector and the target sequence in BAC DNA, and 2) those that do not. As one might expect, choosing a DNA-sequence to alter in a BAC is a requirement for homology driven recombination because sequences flanking the target sequence in BAC need to be homologous between targeting plasmid and BAC DNA. It can introduce a degree of bias when a sequence is chosen to be altered for subsequently testing a hypothesis. Changing all sequence-segments in a BAC, one at a time, overcomes such bias, but the process then becomes quite tedious. One needs also to look out for potential undesirable rearrangements within a BAC when it comes from the genome library of a higher vertebrate because such DNA has a high content of repetitive sequences, which can rearrange during targeting using homologous recombination. Insertions of the bacterial transposon Tn10 on the other hand, are not sequence-specific [34], and allow random changes to be executed in DNA in BACs. Our approach to modifying BAC DNA uses both Tn10 insertions as well as site-specific recombination such as that of unique 34 bp DNA sequences, called loxP sites, by the recombinase enzyme protein Cre of bacteriophage P1.

4. Modifying BACs Using Homologous Recombination

Libraries of ~ 300 kb fragments covering the entire genome of an animal, constructed as BACs, are propagated in a bacterial host that has been rendered recombination deficient [35,36,37]. It was necessary to render the host recombination deficient because the genomes of higher vertebrates, including mammals, have a high content of sequence repeats and BACs containing such DNA in them are prone to rearrangements internally through recombination in the bacterial host. Thus procedures based on homologous recombination require reintroduction of the function into the bacterial host in order to alter DNA in the BAC. Initially *E. Coli RecA* gene was reintroduced and numerous sequence modifications to large DNA in BACs were engineered [38,39,40]. These have been reviewed earlier [41]. Soon thereafter recombination genes from phage λ were introduced into BAC clones, and the red α, β, γ genes reconstituted homologous recombination activity to carry out sequence alterations in BACs [42–45]. The phage λ recombination machinery required shorter lengths of homology to bring about the sequence changes in BACs. Using one of these homologous recombination procedures, a reporter gene/sensor can be integrated at any desired location in the BAC DNA. Influence of non-coding regulatory DNA in the vicinity of the sensor in the BAC is then analyzed by introducing, and subsequently expressing the modified BAC DNA in mouse or zebrafish after integrating it in the genome. The methodology is most useful when clues for altering a sequence in a BAC are available from cross-species sequence homology comparisons. However, non-coding DNA can sometimes be conserved for function without an obvious similarity of sequence between species in many developmentally regulated genes [46–49]. In such cases it is difficult to choose which sequence to test, and a random unbiased BAC modification strategy might be preferable.

5. Sensor/Enhancer-Trap Integrated in BAC Strategy: Comparison with GROMIT Strategy

An entirely different strategy similar to GROMIT conceptually, but using BACs instead of the entire genome, has been developed and tested in the zebrafish system. The zebrafish analogue of the **A**myloid **P**recursor **P**rotein (APP) gene in humans is the *appb* gene, and this was used for developing the methodology. Thus several BACs from a zebrafish genome library and containing the zebrafish *appb* gene were used [25,50]. Because zebrafish embryos develop outside of the mother's womb and are transparent during a large part of their early development, the sensor used in our approach comprised of a promoter-less Enhanced Green Fluorescent Protein (EGFP) gene cassette. A 50 bp β-globin gene promoter was used in the GROMIT technology as a common basal promoter to interact with all conserved non-coding DNA influencing all gene-environments. While we can use such a common "basal promoter" in the scaled up version using pools of BACs, there exist attractive alternatives such as a) using a mixture of DNA sequences, approximately hundred base pairs in length, comprising all the basal promoters of genes contained in the BAC pool, b) constructing a weighted average hybrid sequence, for example, a consensus sequence from several non-identical sequence cassettes, to reflect the variety of gene promoters in the pool, or c) to focus on a single gene in the BAC and use the basal promoter, (BP), of the gene itself, as was done for the zebrafish *appb* gene [25,50].

The major difference in strategy between the GROMIT approach and ours is the host in which the respective transposons, *Sleeping beauty* or Tn10, are integrated into DNA: while the GROMIT

uses mouse ES cells, our approach uses the bacteria *E. coli* containing the BAC DNA. The differences in cost between these procedures are several orders of magnitude.

6. Sensor Integrated into BAC-Approach Applied to zebrafish *appb* Gene-containing BACs

A bacterial transposon Tn10 was used instead of the vertebrate transposon *Sleeping Beauty* to deliver and integrate the enhancer-trap/sensor into the *appb*-BAC DNA. The Tn10 was constructed to include the *appb* basal promoter fused to the promoter-less EGFP gene, and the small transposon Tn10-containing plasmid (~ 7 kb) was introduced into the BAC DNA-containing bacterium by the simple calcium chloride and heat-shock transformation protocol. The transposase enzyme protein gene in Tn10 is regulated by an IPTG-inducible *tac*-promoter. Upon induction with IPTG, the Tn10 containing the basal promoter-sensor fusion integrates into the BAC DNA. The methodology is described in detail elsewhere [51]. The Tn10 transposon DNA also contains a 34 bp sequence in which 13 bp inverted-repeats flank an 8 bp central core sequence, known as a loxP sequence, in addition to the basal promoter-sensor fusion. The genomic DNA insert in BAC is flanked by a loxP and a mutant lox511 sequence [36,37]. The loxP site in the inserted-Tn10 can recombine with the loxP located in BAC in the presence of Cre-recombinase enzyme to delete the intervening DNA, illustrated in Figure 2C. Thus the integrated sensor can be placed accurately at the newly created end generated by the deletion of genomic DNA insert in the BAC. A series of random single insertions of the Tn10 in individual BAC DNA molecules can thus generate a library of BAC DNA with progressively deleted ends, each of which carries an integrated-sensor at the new end of genomic DNA. Libraries of such progressive truncations from either end of genomic insert DNA can be made quite easily and are described elsewhere [52,53]. DNA from end-truncated BACs can be analyzed by **F**ield **I**nversion Gel **E**lectrophoresis (FIGE), an example of which is shown in Figure 3. Sequence of the newly created DNA-end helps determine the location of the integrated-sensor on the BAC DNA and its relationship to the zebrafish *appb* gene [25,50].

The recombinase protein Cre is provided to the host cell by infecting the bacteria containing sensor-integrated BAC with the bacteriophage P1. The P1 phage head also packages the modified BAC DNA within the cell and transports it out of the bacterium upon lyses of the cell. This requires the loxP site on the BAC, and the steps in the recombination process are described and illustrated in detail schematically elsewhere [54]. Although the steps in the recombination process appear complicated, the actual experiment follows a single tube procedure [51].

The protocol allows scaling up and as many as twelve BACs can be processed simultaneously. The limitation arises primarily from the differences in growth rates of BACs which puts them out of phase for conducting subsequent steps of the protocol. Alternatively, BACs of similar growth rates can be pooled together. Such parallel processing of BACs that are part of a contig spanning a genome region would closely resemble the GROMIT strategy after the sensor-integrated BACs are expressed in the appropriate host. The concept is schematically illustrated in Figure 3, panel on right.

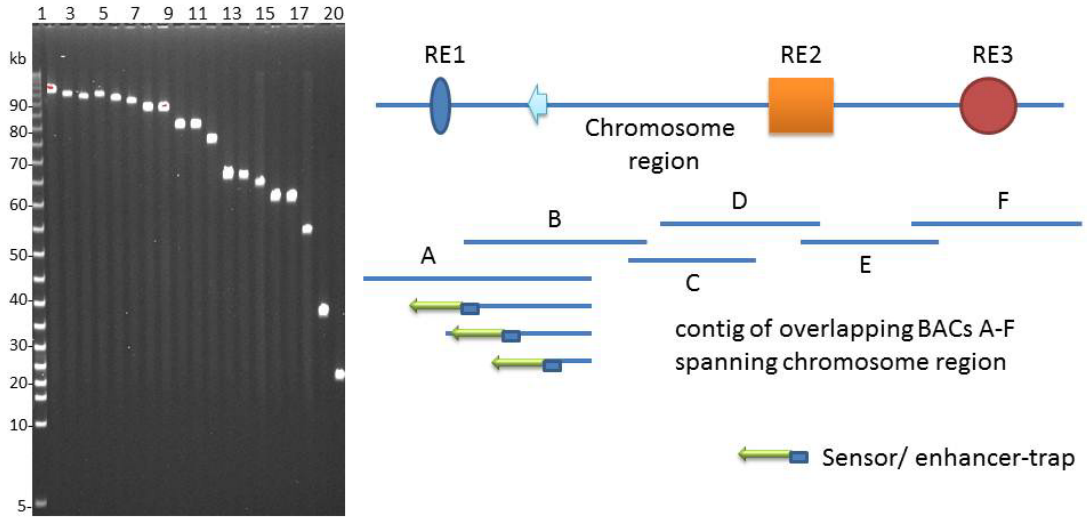

Figure 3. Panel on left shows an ethidium bromide stained gel containing DNA of BAC clones from an end-deletion library where each clone carries *appb*-BAC with integrated-sensor/enhancer-trap at loxP-end and iTol2kan at lox511 end (see references [25,53] for details). Panel on right shows schematically the integrated-sensor in BAC strategy applied to a contig of BACs spanning a chromosome locus. RE1, RE2 and RE3 represent Regulatory Elements in non-coding DNA surrounding the gene of interest, arrowhead in light blue.

7. Pros and Cons of Expressing BACs with Integrated-Sensors Versus Integrating Sensors into the Genome

The GROMIT strategy expresses sensors integrated into the genome and is ideally suited to register the sum total of regulatory influence of highly conserved non-coding DNA surrounding the site of integration of the sensor. Not only are the gene and regulatory DNA studied as part of the chromosome, their contexts with respect to all other regulatory elements influencing the gene of interest at the location are almost preserved [26–30]. This contrasts sharply with reporter assays and also somewhat from the BACs with integrated sensors. In reporter assays, highly conserved non-coding DNA (CNEs), were fused individually to the reporter gene GFP in a small plasmid and expressed transiently in zebrafish without integrating it into the genome [2,4,5,31]. Because embryos are not transparent and develop internally in the mother, expression in the mouse was performed by integrating randomly into the genome a small plasmid containing the CNE fused to a *lac Z* gene [6]. Both these reporter assay-based methods analyze function of the CNEs individually and out of, or inappropriate, context of all other CNEs influencing the gene. In contrast BACs with integrated sensors/enhancer-traps can analyze the influence of all CNEs regulating the gene, but only to the extent that they are available in the length of DNA in the BAC. The random insertion of the BAC DNA into the appropriate host genome, which is a requirement in this approach, can alter the super long-range context of the sensor to influences by elements outside of the BAC DNA. The problem could be circumvented somewhat by protocols where all such BACs are able to integrate at a common site in the genome for better comparison purposes. Such targeting/guiding strategy is available for small plasmid vectors using efficient site-specific transgenesis with the PhiC31

integrase system [55,56], or more generally with the Cre-loxP system [57,58]. The BAC integrated-sensor approach in addition is able to analyze in great mechanistic detail the individual contributions of multiple CNEs which might be discontinuous along the DNA that influences the gene, as demonstrated earlier [25]. Thus a fruitful approach might use the GROMIT strategy to unearth genome regions rich in regulatory DNA and then use the BAC integrated-sensor approach to conduct detailed analyses with BACs from a contig spanning that chromosome locus, as illustrated schematically in Figure 3. For chromosome regions inaccessible by the GROMIT approach, using the BAC strategy remains the only alternative. The detailed analyses using BACs should not be cumbersome because the BACs with integrated-sensor required to scan a region functionally are generated as a library in a single experiment from a chosen BAC clone.

8. Exploring Regulatory Mechanisms by Expressing BACs with Integrated-Sensors in the zebrafish

Overlapping BACs from a zebrafish library, containing the *appb* gene in the desirable orientation with respect to loxP and with ~ 100 kb of sequence flanking the gene at both ends, were obtained from BAC/PAC Resources, CA. End-deletion libraries were made from several of these BACs first with the loxP Tn10 transposon, Tn-US [50], that had as its cargo, a promoter-less EGFP gene fused to a basal promoter of the zebrafish *appb* gene, designated as the sensor/enhancer-trap. This sensor was placed in front of a loxP sequence in the Tn10 transposon. The purpose of generating end-deletions with the Tn-US and expressing end-deleted *appb* BACs with integrated Tn-US-sensor at the new end was to functionally scan for sequences upstream of the *appb* gene potentially capable of enhancing its expression.

The purified DNA from a set of clones from a deletion library was analyzed for size and end-sequenced to locate the position of integrated-sensor on the zebrafish genome. The electrophoretic system most suitable for analyzing BAC DNA of this size is the Field Inversion Gel Electrophoresis (FIGE) system, and an example of a collection of DNA from BACs with integrated-sensors is shown in Figure 3. Specific BAC clones were chosen from such analyses, and the BAC DNA injected into fertilized zebrafish eggs for expression.

No expression of EGFP from any of the end-deleted sensor-integrated BACs was observed [50]. The result indicated additional sequences downstream of the *appb* gene were essential for its expression. The rationale for such a conclusion arose from the fact that sequences downstream of *appb* were absent from these BACs with integrated sensor, because those sequences get eliminated due to the direction of making deletions (see Figure 4). It indicated a requirement of sequences downstream of *appb* for the gene's expression. Thus sequence fragments from within intron 1 of *appb*, for a start, were fused downstream of the EGFP cassette in Tn10 to produce a composite basal-promoter-EGFP-intron 1-enhancer cassette, (BP-EGFP-IE), as cargo, and end-deletions of the BAC were generated with this new Tn10. DNA from specific BAC clones from the new BP-EGFP-IE enhancer-trap library expressed the *appb* gene in neurons of zebrafish embryos when injected into fertilized eggs. Expression of a large number of BACs from this library led to the following conclusion: as long as the BP-EGFP-IE sensor-integrated BAC also contained sequences around 31 kb upstream of the *appb* gene, expression was in neurons, which is the tissue where the endogenous *appb* gene is expressed [59]. However if the end-deletions in the BAC extended to beyond that location, expression was specifically in the notochord. Thus sequences in the −31 kb region of *appb*

contain an additional regulatory element critical for expression of the *appb* gene in neurons [50]. The hypothesis was confirmed by expression of a neuron-expressing-BAC DNA that was again truncated from the opposite lox511-end to delete the −31 kb region of *appb* [25]. Expression was also in the notochord if DNA from the Tn10 transposon plasmid alone, carrying the basal-promoter-EGFP-intron 1-enhancer (BP-EGFP-IE) cassette as cargo and no BAC DNA, was injected into eggs, as expected [50]. The results can be summarized and illustrated as follows in Figure 4.

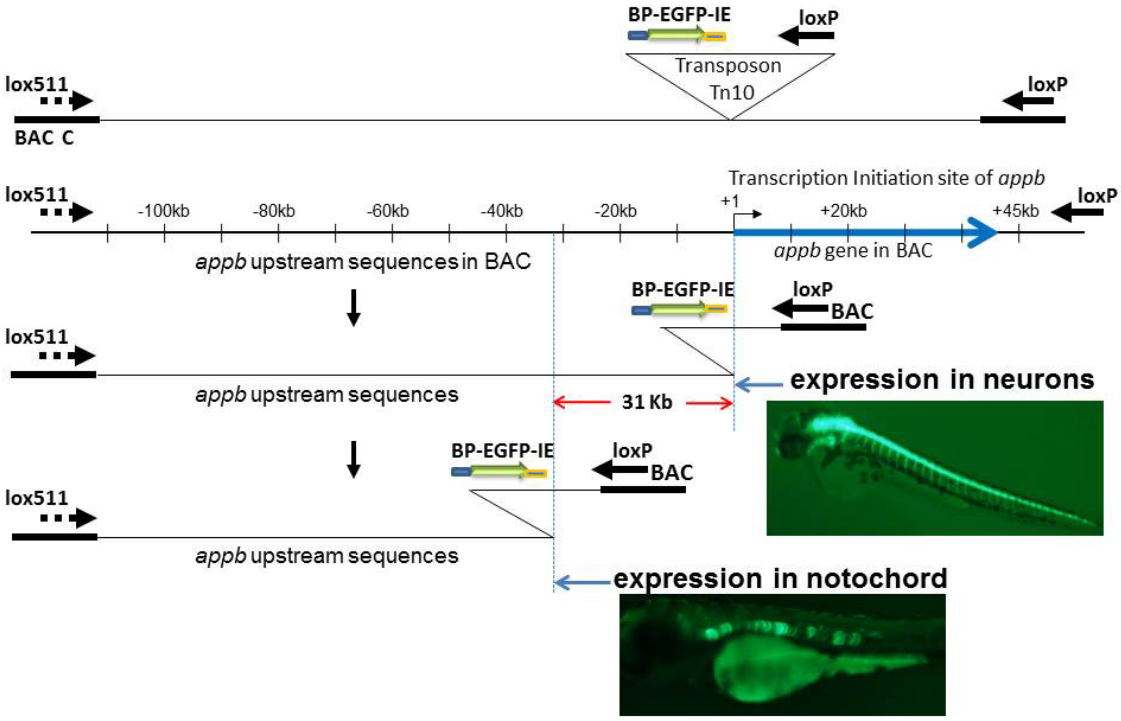

Figure 4. Schematic representation of results from the integrated-sensor in BAC strategy to explore regulation of the zebrafish *appb* gene. A composite sensor comprising of a promoter-less EGFP gene flanked by 0.35 kb DNA immediately upstream of *appb* gene to serve as basal promoter (BP), and 0.8 kb DNA containing the intron 1 enhancer (IE), was placed in front of the loxP sequence in the Tn10 transposon. Insertion of this composite sensor BP-EGFP-IE-loxP-Tn10 into *appb* BACs followed by Cre-recombination of the loxP sites generated libraries of BACs with DNA progressively deleted from the loxP end and containing the BP-EGFP-IE sensor at the newly created end [50]. After characterization of the DNA by size and end-sequence, the DNA from suitable BACs from the library was injected individually for expression into fertilized zebrafish eggs. Expression analysis of a large number of sensor-integrated *appb* BACs indicates that sequences around ~ 31 kb of DNA immediately upstream of the *appb* transcription start site is required for neuronal expression of *appb*: in its presence expression is neuronal (fluorescent picture inset), while in its absence expression switches to the notochord of zebrafish (fluorescent picture inset). Injection of plasmid DNA containing enhancer-trap transposon alone, without the BAC, also produces this notochord expression pattern [25,50]. Figure adapted from [25].

9. Regulatory Elements Important to Neuronal Expression of *appb* in zebrafish Embryos

The cargo sequence in the Tn10 transposon plasmid was subjected to mutational analyses to define the minimal intron 1 enhancer of the zebrafish *appb* gene. It is a convenient way to alter sequences of part of a BAC quickly and was relatively easy to do because the Tn10 plasmid is small. The mutated intron 1 enhancer, (IE), of the BP-EGFP-IE cassette was reintroduced into the *appb* BAC through the making of end-deletions by the loxP-Tn10, and the resulting BACs analyzed for expression in zebrafish embryos. Analysis of expression patterns of a large number of such mutated sensor-integrated BACs indicated that the putative binding sites of two previously known transcription factors, E4BP4/NFIL3 and Forkhead (fkd) were critical for function of the intron enhancer. These sites specifically bound the DNA-binding protein domains of E4BP4/ NFIL3 and Forkhead expressed in *E.Coli* in binding assays *in vitro*. Both these sites are over-represented throughout the *appb* gene region of zebrafish. Additionally the enhancing element at −31 kb of the *appb* gene, which was identified earlier, was also found to contain a triplet of E4BP4/NFIL3 sites [25].

It is also important to note that in addition to the long-range enhancers, upstream and downstream of the *appb* transcription start site that are required to confer tissue-specificity of expression of the gene, and are described here, there are several other transcription factor binding sites identified within the basal proximal promoter region of the Amyloid Precursor Protein (APP) gene from numerous vertebrate species during the past three decades [60–68]. These would be represented here in the zebrafish *appb* gene by the approximately 300 bp sequence immediately upstream of the transcription start site, for example in the BP-fragment of the enhancer-trap, that is common to all reporter-gene/ sensors used in our study [25,50].

10. Clues to Regulation of the APP Gene in Humans

The gene coding for the Amyloid Precursor Protein, (APP), in humans plays a central role in Alzheimer's Disease, (AD), and spans a DNA region approximately four times as large as its counterpart in zebrafish *appb*. Sequence comparisons across vertebrate species in the gene region failed to find any significant similarities. Understanding the regulation of such genes is a challenge because it is difficult to choose sequence segments, which when mutated, would alter function and thus prove its significance. As noted earlier, homology-based recombination strategies to alter sequence in a BAC requires choosing a sequence to target, and the task of covering all regions surrounding, as well as intervening, the coding sequences of the gene in this way becomes arduous. On the other hand, the transposon based strategy allows random truncations to be introduced from a fixed end of the genomic insert DNA in the BAC, with all truncated BACs generated in a single experiment, making all sites equally important for probing.

Despite the lack of overall similarity in the DNA regions, the human APP gene region also has an overabundance of E4BP4/NFIL3 sites. The nucleosomes bound to many of these putative E4BP4/NFIL3 -binding sites were modified by the histone acetylation marker H3K9Ac in chromatin immune-precipitation (ChIP) assays in the human cell-line SHSY5Y that expressed the APP gene. Such marking is indicative of the E4BP4/NFIL3 -binding sites existing in a transcriptionally active chromatin state [69], as concluded in the previous section for zebrafish *appb* gene. These results

together suggest the sites most likely are participating to transcribe the APP gene in the human cell-line SHSY5Y [25].

Expression of BACs with mutated IE-sensors identified a second transcription factor-binding site, Forkhead (fkd), as indispensable in regulating the zebrafish *appb* gene. The role of fkd sites in regulating transcription of APP in humans appears a little more elusive: although only two fkd consensus sites exist in the ~ 400 kb DNA containing the APP gene, there are thirteen additional end-mutated fkd sites with 8 of 9 bases identical (consecutively), within this region (locations shown schematically in Figure 4 of [70]). If such end-mutated sites turn out to be functional in human APP, suggesting these to be indeed evolutionary remains of the Forkhead consensus sites, then it would indicate that potential regulatory pathways for the human APP gene might be operating on a model of conservation of transcription factors rather than overall conservation at the level of DNA sequence in regulatory regions of the two genes. The detailed interactions at the molecular level between various components of the transcriptional machinery in zebrafish and human APP are likely to be somewhat different, presumably due to slight variations in the protein sequences involved in the two cases, although functionally they may be following similar regulatory pathways.

The significance of these findings to Alzheimer ’s disease, (AD), in humans have been noted and discussed in earlier reports [41]. The importance of immunological and inflammatory processes underlying the onset of Alzheimer’s disease in humans is well recognized from clinical studies with patients [71]. It is interesting to note the transcription factor E4BP4/ NFIL3 is also known to be intricately linked with the immune system, where it is required for protecting natural killer (NK) T cells [72]. It is also known to regulate IL-12 p40 expression in macrophages [73]. Thus the characteristics of the disease learned previously from studies of patients with AD can now be correlated with the molecular biology of a gene playing a key role in the disease.

Thus the integrated-sensor in BACs approach has the advantage of analyzing in molecular detail the complex regulation of a gene by more than one regulatory DNA domain. This includes genes such as *appb,* with regions of regulation both upstream and downstream of the coding sequences. The methodology provides an efficient way to generate large numbers of integration-ready sensor-inserted BACs for injection into zebrafish or mouse embryos for expression [25,50].

11. Conclusion

As outlined in a previous section the GROMIT strategy can easily unearth genome regions rich in regulatory DNA, but lacks the ability to analyze detailed mechanisms of interaction of regulatory DNA domain(s) with the genome-integrated sensor. Also, it can identify regulatory function in only a subset of the CNEs in the genome scored earlier in non-coding DNA that is highly conserved among vertebrates. Vast regions of the genome thus appear inaccessible to this approach. Integrating sensors/enhancer-traps into BACs from a contig spanning a chromosomal locus, as illustrated schematically in Figure 3, and expressing them individually in mouse or zebrafish is a viable approach to understanding the regulatory role of non-coding DNA. It has provided detailed mechanistic insight into the key components mediating the expression of the zebrafish *appb* gene, and the possible role of E4BP4/ NFIL3 and the Forkhead family of transcription factors in regulating the gene in humans. A combination of the two approaches, the GROMIT strategy and integrated sensor/enhancer-trap in BACs approach, together might prove most beneficial to this field of research.

Vast amounts of data and information has been generated in recent years in projects such as ENCODE, 1000 Genomes Project, and so on, which have revealed that much of the variation in the human genome lies in noncoding and inter-genic regions suggesting a role in gene regulation and possibly human disease [74,75,76]. The two approaches reviewed here, GROMIT and the integrated-sensor in BACs approach, together, should be ideally suited to validate this huge array of information.

Acknowledgements

The project described was supported by Award Number P20MD000175 from the National Center on Minority Health and Health Disparities (NCMHD) and funds from the North Carolina Biotechnology Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCMHD or the National Institutes of Health. We thank Ms. Rosalind Grays, Camilla Felton, Connie Keys, Crystal McMichael, Jody Lewis, and Darlene Laws for support and encouragement. PKC would like to thank Dr. Ken Harewood for his unwavering support and encouragement throughout the investigation and development of the technology described here.

Conflict of Interest

The authors report no conflict of interests in this research.

References

1. The Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562.
2. Woolfe A, Goodson M, Goode DK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3: e7.
3. Venkatesh B, Yap WH (2005) Comparative genomics using fugu: a tool for the identification of conserved vertebrate cis-regulatory elements. *Bioessays* 27: 100–107.
4. Ahituv N, Prabhakar S, Poulin F, et al. (2005) Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. *Hum Mol Genet* 14: 3057–3063.
5. Shin JT, Priest JR, Ovcharenko I, et al. (2005) Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. *Nucleic Acids Res* 33: 5437–5445.
6. Pennacchio LA, Ahituv N, Moses AM, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444: 499–502.
7. Allende ML, Manzanares M, Tena JJ, et al. (2006) Cracking the genome's second code: enhancer detection by combined phylogenetic foot-printing and transgenic fish and frog embryos. *Methods* 39: 212–219.
8. Berg JM, Tymoczko JL, Stryer L, et al. (2012) *Biochemistry*, 7th ed. WH Freeman & Company, New York.
9. Lodish H, Berk A, Kaiser C, et al. (2004) *Molecular Cell Biology*, 5th edition, WH Freeman and Company.
10. Cherstvy GA, Teif VB (2013) Structure-driven homology pairing of chromatin fibers: the role of electrostatics and protein-induced bridging. *J Biol Phys* 39: 363–385.

11. Beshnova DA, Cherstvy AG, Vainshtein Y, et al. (2014) Regulation of the nucleosome repeat length *in vivo* by the DNA sequence, protein concentrations and long-range interactions. *PLoS Comput Biol* 10: e1003698.
12. Sexton T, Cavalli G (2015) The Role of Chromosome Domains in Shaping the Functional Genome *Cell* 160: 1049–1059.
13. Rao SS, Huntley MH, Durand NC, et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159: 1665–1680.
14. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genetics* 33: 245–254.
15. Kooistra SM, Helin K (2012) Molecular mechanisms and potential functions of histone demethylases**.** *Nat Rev Mol Cell Biol* 13: 297–311.
16. Kouzarides T (2007) Chromatin Modifications and Their Function. *Cell* 128: 693–705.
17. Barski A, Cuddapah S, Cui K, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823–837.
18. Wu L, Wary KK, Revskoy S, et al. (2015) Histone Demethylases KDM4A and KDM4C Regulate Differentiation of Embryonic Stem Cells to Endothelial Cells. *Stem Cell Reports* Jun 24. pii: S2213-6711(15)00159-9. doi: 10.1016/j.stemcr.2015.05.016 [Epub ahead of print].
19. Kurdistani SK, Tavazoie S, Grunstein M (2004) Mapping global histone acetylation patterns to gene expression. *Cell* 117: 721–733.
20. Ganapathi M, Singh GP, Sandhu KS, et al. (2007) A whole genome analysis of 5' regulatory regions of human genes for putative cis-acting modulators of nucleosome positioning. *Gene* 391: 242–251.
21. Hebbar PB, Archer TK (2007) Chromatin-dependent cooperativity between site-specific transcription factors *in vivo*. *J Biol Chem* 282: 8284–8291.
22. Gartenberg MR, Ampe C, Steitz TA, et al. (1990) Molecular characterization of the GCN4-DNA complex. *Proc Natl Acad Sci U S A* 87: 6034–6038.
23. Dalma-Weiszhausz DD, Gartenberg MR, Crothers DM (1991) Sequence-dependent contribution of distal binding domains to CAP protein-DNA binding affinity. *Nucleic Acids Res* 19: 611–616.
24. Ulanovsky L, Bodner M, Trifonov EN, et al. (1986) Curved DNA: design, synthesis, and circularization. *Proc Natl Acad Sci U S A* 83: 862–866.
25. Shakes LA, Du H, Wolf HM, et al. (2012) Using BAC Transgenesis in Zebrafish to Identify Regulatory sequences of the Amyloid Precursor Protein gene in Humans. *BMC Genomics* 13: 451.
26. Kawakami K, Takeda H, Kawakami N, et al. (2004) A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Developmental Cell* 7: 133–144.
27. Ellingsen S, Laplante MA, Konig M, et al. (2005) Large-scale enhancer detection in the zebrafish genome. *Development* 132: 3799–3811.
28. Kokubu C, Horie K, Abe K, et al. (2009) A transposon-based chromosomal engineering method to survey a large cis-regulatory landscape in mice. *Nat Genet* 41: 946–952.
29. Ruf S, Symmons O, Uslu VV, et al. (2011) Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat Genet* 43: 379–386.
30. Symmons O, Spitz F (2013) From remote enhancers to gene regulation: charting the genome's regulatory landscapes. *Philos Trans R Soc Lond B Biol Sci* 368: 20120358.

31. Bessa J, Tena JJ, de la Calle-Mustienes E, et al (2009) Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev Dyn* 238: 2409–2417.
32. Chatterjee S, Lufkin T (2012) Regulatory genomics: Insights from the zebrafish. *Curr Top Genet* 5: 1–10.
33. Simon MI (1997) Dysfunctional genomics: BACs to the rescue. *Nat Biotechnol* 15: 839.
34. Gilmore RC, Baker Jr J, Dempsey S, et al. (2001) Using PAC nested-deletions to order contigs and microsatellite markers at the high repetitive sequence containing Npr3 gene locus. *Gene* 275: 65–72.
35. Shizuya H, Birren B, Kim UJ, et al. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci U S A* 89: 8794–8797.
36. Osoegawa K, Woon PY, Zhao B, et al. (1998) An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* 52:1–8.
37. Frengen E, Weichenhan D, Zhao B, et al. (1999) A modular, positive selection bacterial artificial chromosome vector with multiple cloning sites. *Genomics* 58: 250–253.
38. Yang XW, Model P, Heintz N (1997) Homologous recombination based modification in Escherichia coli and germline transmission in transgenic mice of a bacterial artificial chromosome. *Nat Biotechnol* 9: 859–865
39. Gong S, Yang XW, Li C, et al. (2002) Highly efficient modification of bacterial artificial chromosomes (BACs) using novel shuttle vectors containing the R6Kgamma origin of replication. *Genome Res* 12: 1992–1998.
40. Yang Z, Jiang H, Chachainasakul T, et al. (2006) Modified bacterial artificial chromosomes for zebrafish transgenesis. *Methods* 39:183–188.
41. Shakes LA, Wolf HM, Norford DC, et al. (2014) Harnessing Mobile Genetic Elements to Explore Gene Regulation. *Mobile Genetic Elements* 4: e29759.
42. Zhang Y, Buchholz F, Muyrers JP, et al. (1998) A new logic for DNA engineering using recombination in Escherichia coli. *Nat Genet* 20: 123–128.
43. Muyrers JP, Zhang Y, Testa G, et al. (1999) Rapid modification of Bacterial Artificial Chromosomes by ET recombination. *Nucleic Acids Res* 27: 1555–1557.
44. Suster ML, Abe G, Schouw A, et al. (2011) Transposon-mediated BAC transgenesis in zebrafish. *Nat Protoc* 6: 1998–2021
45. Warming S, Costantino N, Court DL, et al. (2005) Simple and highly efficient BAC recombineering using galK selection. *Nucleic Acids Res* 33: e36.
46. Fisher S, Grice EA, Vinton RM, et al. (2006) Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* 14: 276–279.
47. Blow MJ, McCulley DJ, Li Z, et al. (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* 42: 806–810.
48. Kague E, Bessling SL, Lee J, et al. (2010) Functionally conserved cis-regulatory elements of COL18A1 identified through zebrafish transgenesis. *Dev Biol* 337: 496–505.
49. Taher L, McGaughey DM, Maragh S, et al. (2011) Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res* 7: 1139–1149.

50. Shakes LA, Malcolm TL, Allen KL, et al. (2008) Context dependent function of APPb Enhancer identified using Enhancer Trap-containing BACs as Transgenes in Zebrafish. *Nucleic Acids Res* 36: 6237–6248.
51. Chatterjee PK (2015) Directing Enhancer-traps and iTol2 End Sequences to Deleted BAC ends with loxP- and lox511-Tn10 transposons. *"Bacterial Artificial Chromosomes" Methods in Molecular Biology* series. Editor: K. Narayanan, Series editor: J. Walker. (2015) 2nd ed., Humana Press/ Springer, 99–122.
52. Shakes LA, Garland DM, Srivastava DK, et al. (2005) Minimal Cross-recombination between wild type and loxP511 sites *in vivo* facilitates Truncating Both Ends of Large DNA Inserts in pBACe3.6 and Related Vectors. *Nucleic Acids Res* 33: e118.
53. Shakes LA, Abe G, Eltayeb MA, et al. (2011) Generating libraries of iTol2-end insertions at BAC ends using loxP and lox511-Tn10 transposons. *BMC Genomics* 12: 351.
54. Chatterjee PK, Shakes LA, Wolf HM, et al. (2013) Identifying Distal *cis*-acting Gene-Regulatory Sequences by Expressing BACs Functionalized with loxP-Tn10 Transposons in Zebrafish. *Royal Soc Chem Adv* 3: 8604–8617.
55. Mosimann C, Puller AC, Lawson KL, et al. (2013) Site-directed zebrafish transgenesis into single landing sites with the PhiC31 integrase system. *Dev Dyn* 242: 949–963.
56. Kirchmaier S, Höckendorf B, Möller EK, et al. (2013) Efficient site-specific transgenesis and enhancer activity tests in medaka using PhiC31 integrase. *Development* 140: 4287–4295.
57. Araki K, Araki M, Yamamura K (1997) Targeted integration of DNA using mutant lox sites in embryonic stem cells. *Nucleic Acids Res* 25: 868–872.
58. Liu W, Wang Y, Qin Y, et al. (2007) Site-Directed Gene Integration in Transgenic Zebrafish Mediated by Cre Recombinase Using a Combination of Mutant Lox Sites. *Marine Biotechnol* 9: 420–428.
59. Musa A, Lehrach H, Russo VA (2001) Distinct expression patterns of two zebrafish homologues of the human APP gene during embryonic development. *Dev Genes Evol* 211: 563–567.
60. Salbaum JM, Weidemann A, Lemaire HG, et al. (1988) The promoter of Alzheimer's disease amyloid A4 precursor gene. *EMBO J* 7: 2807–2813.
61. Yoshikai SI, Sasaki H, Dohura K, et al. (1990) Genomic organization of the human amyloid beta-protein precursor gene. *Gene* 87: 257–263.
62. Lahiri DK, Robakis NK (1991) The promoter activity of the gene encoding Alzheimer beta-amyloid precursor protein (APP) is regulated by two blocks of upstream sequences. *Brain Res Mol Brain Res* 9: 253–257.
63. Song W, Lahiri DK (1998) Functional identification of the promoter of the gene encoding the Rhesus monkey beta-amyloid precursor protein. *Gene* 217:165–176.
64. Lahiri DK, Song W, Ge YW (2000) Analysis of the 5′-flanking region of the beta-amyloid precursor protein gene that contributes to increased promoter activity in differentiated neuronal cells, *Brain Res Mol Brain Res* 77: 185–198.
65. Lahiri DK, Ge YW, Maloney B (2005) Characterization of the APP proximal promoter and 5'-untranslated regions: Identification of cell-type specific domains and implications in APP gene expression and Alzheimer's disease. *FASEB J* 19: 653–665.
66. Rogers JT, Randall JD, Cahill CM, et al. (2002) An iron-responsive element type II in the 5' untranslated region of the Alzheimer's amyloid precursor protein transcript. *J Biol Chem* 277: 518–528.

67. Shaw KT, Utsuki T, Rogers J, et al. (2001) Phenserine regulates translation of beta-amyloid precursor protein mRNA by a putative interleukin-1 responsive element, a target for drug development. *Proc Natl Acad Sci U S A* 98: 7605–7610.
68. Theuns J, Brouwers N, Engelborghs S, et al. (2006) Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease. *Am J Hum Genet* 78: 936–946.
69. Chakraborty T, Perlot T, Subrahmanyam R, et al. (2009) A 220-nucleotide deletion of the intronic enhancer reveals an epigenetic hierarchy in immunoglobulin heavy chain locus activation. *J Exp Med* 206: 1019–1027.
70. Wolf HM, Nyabera KO, De La Torre KK, et al. (2014) Long Range Gene-Regulatory Sequences Identified by Transgenic Expression of Bacterially-Engineered Enhancer-trap BACs in Zebrafish. *Mol Biol Genetic Eng* 2: 2.
71. Popović M, Caballero-Bleda M, Puelles L, et al. (1998) Importance of immunological and inflammatory processes in the pathogenesis and therapy of Alzheimer's disease. *Int J Neurosci* 95: 203–236.
72. Kamizono S, Duncan GS, Seidel MG, et al. (2009) Nfil3/E4bp4 is required for the development and maturation of NK cells *in vivo*. *J Exp Med* 206: 2977–2986.
73. Kobayashi T, Matsuoka K, Sheikh SZ, et al. (2011) NFIL3 is a regulator of IL-12 p40 in macrophages and mucosal immunity. *J Immunol* 186: 4649–4655.
74. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816.
75. The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52–58.
76. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56–65.

AIMS Press