Maps of *cis*-Regulatory Nodes in Megabase Long Genome Segments are an Inevitable Intermediate Step Toward Whole Genome Functional Mapping

Lev G. Nikolaev*, Sergey B. Akopov, Igor P. Chernov and Eugene D. Sverdlov

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 16/10 Miklukho-Maklaya, 117997, Moscow, Russia

Abstract: The availability of complete human and other metazoan genome sequences has greatly facilitated positioning and analysis of various genomic functional elements, with initial emphasis on coding sequences. However, complete functional maps of sequenced eukaryotic genomes should include also positions of all non-coding regulatory elements. Unfortunately, experimental data on genomic positions of a multitude of regulatory sequences, such as enhancers, silencers, insulators, transcription terminators, and replication origins are very limited, especially at the whole genome level. Since most genomic regulatory elements (e.g. enhancers) are generally gene-, tissue-, or cell-specific, the prediction of these elements by computational methods is difficult and often ambiguous. Therefore, the development of high-throughput experimental approaches for identifying and mapping genomic functional elements is highly desirable. At the same time, the creation of whole-genome map of hundreds of thousands of regulatory elements in several hundreds of tissue/cell types is presently far beyond our capabilities. A possible alternative for the whole genome approach is to concentrate efforts on individual genomic segments and then to integrate the data obtained into a whole genome functional map. Moreover, the maps of polygenic fragments with functional *cis*-regulatory elements would provide valuable data on complex regulatory systems, including their variability and evolution. Here, we reviewed experimental approaches to the realization of these ideas, including our own developments of experimental techniques for selection of *cis*-acting functionally active DNA fragments from large (megabase-sized) segments of mammalian genomes.

Received on: January 17, 2007 - Revised on: February 22, 2007 - Accepted on: February 27, 2007

Key Words: Physical mapping, promoter, enhancer, insulator, S/MAR, transcription factor, replication origin, epigenomics, CpG methylation, chromatin, histone modification, DNase I hypersensitivity.

1) INTRODUCTION

Now that the sequences of many genomes have been completed [1-3], the understanding of gene regulatory mechanisms that determine phenotypic diversity is a major challenge of the coming post-genome era. The common consensus is that mammalian genomes contain 20 to 30 thousand of protein-coding genes. However, there is growing evidence that most of genomic DNA is also transcribed (for recent review see [4]). Many of non-coding transcripts play important role in genome regulation, adding one more dimension to the complexity of genome functioning. In genomes of multicellular eukaryotes, all this multitude of genes and non-coding transcribed areas are interconnected to produce immensely complex patterns of gene expression in order to define numerous specialized cell types. Gene expression is regulated at various levels: linear arrangement of genes and regulatory DNA elements, DNA modification, chromatin structure, etc. up to nuclear compartmentalization. Every aspect of gene regulation is controlled by complex cellular systems, and these systems must interact to define a coherent regulatory network [5]. It should be admitted that despite many spectacular discoveries concerning organiza-

tion of particular functional systems (see [6] for review), we are still far from full understanding of the mechanisms governing such cell-specific ensembles. Even in relatively simple cases, e.g. in attempts to identify genomic *cis*-regulatory elements, we face serious difficulties and have to spend much time and efforts to overcome them. This can be illustrated by the fact that for the moment only less than 100 enhancers and silencers are identified in the human genome, although the theoretically predicted number of them exceeds 100,000 [7].

So, what strategy should be chosen to achieve full understanding of the whole genome regulation in all its complexity? The recent euphoria about the emergence of new powerful techniques like DNA microarrays, which seemed to open the way to whole genome analysis of gene expression and its profiling in different cell types and conditions, has largely faded due to technical problems, poor reproducibility and some limitations (see [8] and references therein). On this background, genome strategies limited to detailed analysis of individual functional domains or isolated long genomic stretches and aimed at their comprehensive functional understanding seem to be more realistic. Subsequent integration of such fragmentary data might allow unraveling complex systems of functional interactions between genomic elements.

There can be two approaches to such a partial analysis, which we will call top-down and bottom-up strategies. In a top-down strategy, a certain stage of a process, like devel-

*Address correspondence to this author at the Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Human Genes Structure and Functions, 16/10 Miklukho-Maklaya, 117997, Moscow, Russia; Tel: +7-495-330-7029; Fax: +7-495-330-6538; E-mail: lev@humgen.siobc.ras.ru

opment of the organism or cell differentiation, is analyzed in an attempt to identify all components characteristic of or interacting during this stage. Then, a diagram of the interactions is constructed showing a so called “connectivity” of the components (reviewed in [9]). The most advanced example of this strategy is the construction of so called Genomic Regulatory Networks (GRNs). GRNs are constructed by combining data on gene expression and *cis*-regulatory elements. They provide direct visualization of functional interconnections among regulatory genes and signaling components resulting in cell-specific patterns of gene activity. The first GRN was constructed for endomesoderm specification in the pregastrular sea urchin embryo, *Strongylocentrotus purpuratus* ([10], reviewed in [11]).

An organism's development includes a succession of stages, and at each subsequent stage new subsets of specialized cells that implement specific functions are formed. The cell specification occurs due to the expression of certain sets of genes in subsets of cells in each region of the developing organism. In turn, the changes in transcriptional patterns are triggered either by external or internal signals provided by maternal molecules inherited from the egg cytoplasm. The signals are received by *cis*-regulatory modules - enhancers, silencers or insulators. Each of them is typically a DNA stretch of 300 or more base pairs in length and contains about 10 or more binding sites for at least four transcription factors [9]. The transcription factors interacting with their target *cis*-elements provide specific input signals that switch the module on or off. A *cis*-module thus forms a complex (node) of transcription factors and cofactors characteristic of a certain class of embryonic cells at a certain time. All these interacting elements can be considered a functional module whose function is switching on a set of new regulatory genes. A multitude of interacting functional nodes form regulatory networks. For each subset of cells at each developmental stage its own regulatory network(s) is formed.

A key feature of *cis*-nodes is their ability to use AND logic [10], when two or more different transcription factors and cofactors must be bound to the node simultaneously in order to activate transcription. A unique specificity of a *cis*-node is determined by a combination of protein-binding sequences.

A number of GRNs were reported in several model systems [12]. The inputs and outputs from each gene of a network can be summarized in circuit diagrams [13]. Analysis of circuit diagrams provides testable predictions and new insights into complex developmental processes [10]. An increasing number of gene regulatory networks are being determined as whole genome studies become more prevalent. More recent studies have led to the elucidation of networks underlying mesoderm specification in the frog *Xenopus laevis* (reviewed in [14]), dorsoventral patterning of the *Drosophila* embryo (reviewed in [15]), vulva differentiation in *Caenorhabditis elegans* [16], B cell differentiation in the mammalian immune system [17], and the segmentation of the *Drosophila* embryo [18]. The GRN approach looks very productive, however, it should be remembered that it is based on whole genome screening of genes expressed during the process under study and therefore suffers from disadvantages characteristic of the microarray technique [8].

GRNs represent functional links between regulatory modules and transcription factor encoding genes located in various parts of the genome. In contrast, the bottom-up approach is aimed at mapping all *cis*-regulatory modules in certain DNA regions [6]. This approach was put forward by the International ENCODE (the ENCyclopedia Of DNA Elements) project launched by The National Human Genome Research Institute [19] (<http://genome.ucsc.edu/ENCODE/>; <http://www.genome.gov/10005107>). The goal of the ENCODE project initial phases is to identify a suite of approaches that will allow the comprehensive identification of all the functional elements in the human genome.

Our own research uses a similar strategy. We are trying to construct a comprehensive map of all *cis*-regulatory modules within a 1 Mb long DNA segment of human chromosome 19 densely occupied by genes. These genes are expressed either in a coordinated fashion or independently, and we are trying to correlate the gene expression with regulatory modules located near the genes. To achieve this goal, we are developing a variety of techniques allowing identification of *cis*-regulatory modules within long DNA stretches.

Below, we will review recently developed high-throughput experimental techniques for mapping genomic regulatory sequences and epigenetic markers, as well as some results of such a mapping. All these methods have one key step in common, namely a construction of a genomic sublibrary greatly enriched in the DNA fragments containing the elements in question. When this sublibrary is prepared, the rest is a pure technical problem, and the mapping can be performed using various techniques, like microarrays, SAGE-like methods, mass spectrometry, high-throughput sequencing etc. The approaches described below can be subdivided into two groups. The first group includes the approaches that can be called structural. They are based on either specific interactions of DNA and protein(s) (identification of protein binding sites, ChIP), proteins and antibodies (ChIP), DNA and subcellular structures (identification of S/MARs), or structural peculiarities of DNA (identification of replication origins) etc. The second group comprises function-based approaches, usually reporter gene assays (identification of promoters, enhancers, or insulators). It should be noted that the second group of methods is not suitable for whole-genome mapping due to intrinsically limited cell transfection efficiency.

II) IDENTIFICATION AND MAPPING CLASSICAL REGULATORY ELEMENTS

IIa) Promoters

RNA polymerase II promoters are usually located closely upstream of genes. This fact greatly facilitates their identification and mapping, providing the location and transcription start sites of genes are known. About 10000 human transcripts were aligned with the human genome sequence, and genomic sequences upstream of putative transcription start sites were selected as putative promoters. 152 likely promoters were tested for functional activity in transient transfection assay, and 81% of them displayed promoter activity [20]. Further development of this approach allowed the identification of 387 fragments located within a 30-Mbp long human genome sequence and possessing promoter activity in

at least one of 16 cell lines tested [21]. A method for identification of promoters irrespective of their position relative to genes was also proposed and used to identify internal (intronic) and alternative promoters [22]. The 300-600 bp long fraction of *Rsa* I-digested genomic DNA was cloned into a self-inactivating retroviral vector upstream of the promoter-less green fluorescent protein (GFP) reporter gene. GFP+ cells were sorted out in a flow cytometer, and selected inserts tested for promoter activity, sequenced and mapped. However, the method is unsuitable for genome-wide analysis due to technical limitations.

An alternative method of structure-functional identification of promoters uses chromatin immunoprecipitation (ChIP-on-chip technique) [23]. RNA polymerase II and one of the TFIID subunits used as the immunoprecipitation target allowed to identify and map promoters active in a given type of cells and thus to characterize their tissue specificity. In a follow-up study, a genome-wide identification of promoters was performed. More than 10000 active promoters were revealed in a single cell type; about 87% of them were mapped within 2.5 Kbp from annotated 5'-ends of known mRNAs [24]. Although tissue specificity of promoters was not studied at the whole-genome level, some data indicate that most promoters are active in many cell types (see [21], supplemental Table D).

IIb) Enhancers/Silencers

In contrast to promoters, positions of enhancers/silencers are not directly connected with positions of genes under their control, and these elements can function over long distances (for review see [25; 26]). Moreover, a single promoter can be regulated by several enhancers with different tissue specificities. This makes the problem of identifying enhancers/silencers and, especially, their association with the corresponding genes, very complicated. The only published approach is similar to that described above for promoters and employs the ChIP-on-chip technique with antibodies to putative enhancer-binding proteins in combination with an *in silico* approach [27]. Nevertheless, the problem of identification and mapping of enhancer/silencer elements remains generally unresolved. Moreover, application of this method is limited by our knowledge of enhancer-binding proteins, and the method can be used only for identification of regulatory elements for which at least one protein component of the regulatory complex is known.

IIc) Mapping of Loop-Domain Chromatin Structure Elements - S/MARs and Insulators

There is abundant evidence that DNA in eukaryotic cells is organized into loop domains that represent basic structural and functional units of chromatin. Two classes of sequences were supposed to be border elements that form and support the loop domain structure - insulators (reviewed in [28-30]) and scaffold/matrix attachment regions (S/MARs), defined as sequences capable of specific binding to nuclear matrix *in vitro* (reviewed in [31]).

A technique [32] based on selection and cloning of genomic restriction fragments capable of specific binding to isolated nuclear matrix *in vitro* was proposed and found to be applicable to the isolation and mapping of S/MARs from

whole human chromosomes. The technique was used for identifying S/MARs within a 1 Mbp long multigenic *FXVD5-COX7A1* region of human chromosome 19 [33]. A library of putative S/MARs of this locus from the Jurkat cells was prepared and shown to contain ~20 independent S/MARs. 16 of them were isolated, sequenced and assigned to certain positions within the locus. A majority of the S/MARs identified (11 out of 16) were located in intergenic regions, suggesting their role in delimitation of chromatin domains. These 11 S/MARs divided the locus into 10 domains ranging from 6 to 272 kb with an average domain size of 88 kb. The remaining five S/MARs were found within intronic sequences of the *APLP1*, *PRODH2*, *MAG* and *NPHS1* genes. The same technique was used for S/MARs identification within the breast cancer loss of heterozygosity region on human chromosome 16q22.1 [34]. In combination with the whole genome microarrays, this approach can be also used for genome-wide identification of S/MARs.

At least two different approaches were proposed for identification and mapping of insulators. Earlier, the multifunctional transcription factor CTCF was shown to be required for insulator function [35]. This feature was used in the ChIP-on-chip technique to identify more than 200 CTCF-binding DNA fragments in the mouse genome [36]. It was demonstrated that a considerable fraction of these fragments displayed insulator-like properties when assayed using an episome-based test. Although a powerful tool for isolation of CTCF binding sites, this approach, as well as others based on the ChIP-on-chip technology, is limited and does not allow to detect insulator sequences that do not bind the required antigen [37].

We have developed a function-based approach for direct selection of insulators from mixtures of genomic fragments [38] (Fig. 1). This approach exploits the most characteristic feature of insulators, their ability to prevent transactivation of promoters by their cognate enhancers when located between these elements [39], and includes a positive-negative selection procedure with the use of the herpes simplex virus thymidine kinase gene (*HSV-tk*). Importantly, the approach is suitable for revealing insulator sequences in constructions stably integrated into the genome. Using this technique, several potential insulators within the *FXVD5-COX7A1* locus of human chromosome 19 were identified and mapped.

Controversial data were obtained on interrelations of S/MARs and insulator elements. At least in some cases, sequences capable of binding nuclear matrix were reported to possess insulator activity [40]. Also, two different human S/MARs from the apolipoprotein B and alpha 1-antitrypsin loci were active in a *Drosophila white* gene position effect assay [41].

IId) Replication Origins

No attempts of mapping sequences functioning as replication origins in higher eukaryotes were made so far, apparently due to the absence of appropriate selection methods. However, a promising origin-trapping approach was recently published [42] and used to produce a library of DNA fragments containing replication origins. The approach is based on trapping of the replication bubble in solidifying agarose. When 15 randomly selected clones were analyzed by a strin-

Fig. (1). General scheme of selection and mapping of insulator sequences [38]. PGK-pr — the human phospho-glycerate kinase promoter, *Neo^R* — the neomycin phosphotransferase gene, CMV-Enh — the cytomegalovirus enhancer, CMV-mPr — a cytomegalovirus minimal promoter, HSV-*tk* — the herpes simplex virus thymidine kinase gene. The vector contains a neomycin-resistance marker for positive selection of cells with the integrated plasmid, as well as a negative selection marker, HSV-*tk*, whose expression sensitizes cells to ganciclovir. The insertion of a DNA fragment with enhancer-blocking (insulator) properties in between the promoter and enhancer of HSV-*tk* significantly inhibits expression of this gene in cells stably transfected with the construct and imparts resistance to ganciclovir to these cells. The short fragments of a library representing the genome locus of interest were cloned as inserts in between the HSV-*tk* promoter and enhancer. The ganciclovir-resistant cells were collected, their genomic DNAs isolated, the inserts of integrated plasmids rescued by PCR with 1L and 1R primers, cloned and analyzed. With kind permission of Springer Science and Business Media.

gent two-dimensional gel replicon mapping method, all were shown to correspond to replication origins active in early S-phase. Hopefully, this approach will help to map replication origins in near future.

IIe) Protein Binding Sites

Specific binding of nuclear proteins to target DNA sequences is a major mechanism of genome functioning and regulation in eukaryotes and governs the functioning of all above mentioned regulatory elements [6; 43]. That makes identification and mapping specific protein target sites (PTSs) necessary for understanding genomic regulation. Several approaches to unbiased mapping of PTSs have been proposed so far. The most widely used is the ChIP-on-chip technique which allowed researchers to map target sites for the NF-kappaB [44] and CREB [45] transcription factors across human chromosome 22, and the Sp1, c-Myc, and p53 factors across human chromosomes 21 and 22 [46]. Another experimental approach named DamID [47; 48] was recently used for mapping GAGA [49], and Myc, Max, and Mad/Mnt target sites across the whole *Drosophila* genome [50]. It should be stressed that both techniques have one major disadvantage - they need antibodies to the transcription factors

in question and therefore can be applied only to mapping binding sites of well characterized transcription factors.

To overcome this disadvantage, we have recently proposed a two-dimensional electrophoretic mobility shift assay (2D-EMSA) technique that allowed us to select, clone and map many PTSs in one experiment [51; 52]. The 2D-EMSA is unbiased with respect to DNA-binding proteins and can therefore select PTSs of yet unidentified factors (Fig. 2). Moreover, the results of 2D-EMSA strongly depend on the protein composition of nuclear extracts used, and thus might characterize the pool of nuclear proteins of cells or tissues, i.e. tissue specificity of transcription factors binding. Using 2D-EMSA, we have identified and mapped 52 PTSs within a 563 kb human genome region. Only a rather small fraction (~15%) of the identified PTSs were found in promoter regions (5'-regions) of genes. Most (70%) of the PTSs were located within intronic sequences or in intergenic regions. One may speculate that these PTSs belong to *trans-acting* regulatory elements, like enhancers, insulators or locus control regions.

Most genomic regulatory elements that determine structural and functional properties of metazoan organisms are tissue- and/or cell-specific. This tissue specificity is in turn

Fig. (2). General scheme of two-dimensional EMSA (2D-EMSA) [51]. DNA-protein complexes were formed by incubation of a labeled library of short fragments representing the genome locus of interest with nuclear extract proteins and initially separated in a non-denaturing 1st dimension polyacrylamide gel (EMSA dimension). After disruption of the complexes, the DNA fragments released were separated in a 2nd dimension SDS-containing gel. Fragments that do not bind proteins have roughly the same mobility in both dimensions, and their spots form the diagonal. In contrast, the fragments retarded in the first dimension due to protein binding produce spots located below the diagonal. The area of these spots corresponding to target DNA sequences outlined by the dashed line was excised, DNA fragments eluted, cloned and analyzed.

determined by tissue-specific binding of proteins, in particular transcription factors. The absence or presence of a specific transcription factor complex with the corresponding *cis*-element might be the result of the absence or presence of this transcription factor in the cell, post-translational modification of the factor allowing or preventing the binding, or modification (e.g. methylation) of the *cis*-element itself. Therefore, the DNA sequences capable of tissue-specific binding of regulatory proteins play an indispensable role in genomic regulatory networks and their large-scale identification, cloning and mapping are of great importance for the understanding of functional activity of the human and other genomes. Taking into account a huge quantity of *cis*-regulatory elements in genomes, the problem of characterization of the tissue specificity of protein-DNA binding is extremely important. However, this problem cannot be solved using conventional approaches.

III) IDENTIFICATION AND MAPPING OF GENOME EPIGENETIC ELEMENTS

While the number of "classical" regulatory elements in eukaryote genomes can be estimated as tens of thousands,

the number of epigenetic elements is by orders of magnitude higher. This makes the task of their genome-wide mapping too laborious and necessitates its simplification. The task can be simplified in different ways: one can (i) map all or most of the elements within a relatively small portion of the complete genome (the partial mapping approach discussed above); (ii) "simplify" the genome by selecting the regions of interest, like coding sequences or CpG islands; and (iii) use methods of comparative genomics, such as differential display or subtractive hybridization to obtain information about differences in the location of epigenetic markers in various tissues or organisms.

IIIa) Mapping of Methylated CpG Sites

Methylation of cytosine residues in CpG dinucleotides plays a key role in the regulation of gene expression in higher eukaryotes [53]. Numerous approaches have been proposed to assess CpG methylation (reviewed in [54; 55]). The number of potential CpG methylation sites is equal to the number of CpG dinucleotides in the genome, so the approaches used usually included simplification. Below we will briefly discuss several recently reported methods of

high-throughput identification and mapping of methylated CpG sites.

Methylation-Sensitive-Representational Difference Analysis (MS-RDA)

The methylation-sensitive restriction enzyme *Hpa* II was used to prepare amplicons of genomic DNA from two different cell lines that were then compared by subtractive hybridization [56; 57]. This method allows identification of the genomic regions differentially methylated in two different cell lines or tissues. Apart from this simplification, the number of methylated CpGs to be analyzed was lowered by considering only CpGs within *Hpa* II sites and also due to intrinsic imperfection of the original RDA approach [58].

Methylation-Sensitive Restriction Landmark Genomic Scanning (RLGS-M)

Methylation-sensitive restriction landmark genomic scanning (RLGS-M) is based on the use of methylation-sensitive restriction enzymes with long recognition sites as methylation landmarks [59]. After digestion and radioactive labeling, DNA fragments are resolved by two-dimensional agarose gel electrophoresis (after the first dimension the fragments are digested with a different restriction enzyme). With the methylation-sensitive restriction endonuclease *Not* I whose recognition sites are located predominantly within CpG islands, the resulting RLGS profile reflects the methylation status and the number of copies of the methylated sequences. This method is suitable for comparison of CpG islands methylation in different (e.g. normal and tumorous) tissues [60].

Bisulfite Modification of DNA

In this approach bisulfite-treated DNA, where unmethylated, but not methylated, cytosine residues are converted into thymine residues, is used as template for PCR amplification. Due to differential methylation status, the amplified product may thus contain DNA fragments with altered nucleotide sequences. The PCR product obtained is fluorescently labeled and used as a probe for hybridization to arrayed oligonucleotides that can discriminate between methylated and unmethylated sites. Using this approach, methylated CpG sites from regulatory regions of some genes located both within and outside of CpG islands were identified and mapped for a number of normal and cancerous tissues [61-63].

Bisulfite approach in combination with bisulfite sequencing and mass-spectrometry was also used in the "Epigenome Project" [64]. This project is aimed at genome-wide mapping and analysis of methylation status and its functional significance. At the pilot stage of the project, the analysis of methylation of the human major histocompatibility complex was accomplished [64].

In a recent paper [65], bisulfite-modified bases were identified using methylation-specific primers and randomly ordered bead arrays. Methylation profiles of 1536 specific CpG sites in 371 genes were analyzed in normal and cancerous lung tissues, as well as in several cell lines.

Immunoprecipitation of Methylated DNA

An immunoprecipitation analysis with the use of antibodies to methylated cytosines in combination with microarrays

was proposed and employed to generate methylation profiles of all human chromosomes at 80-kb resolution and for a large set of CpG islands [66].

Methylated-CpG Island Recovery Assay (MIRA)

MIRA is based on the ability of methyl-CpG-binding domain proteins, such as methyl-CpG-binding domain protein-2 (MBD2), to specifically bind methylated DNA sequences. Sonicated genomic DNA was incubated with immobilized MBD2 in the presence of the methyl-CpG-binding domain protein 3-like-1 (MBD3L1), a binding partner of MBD2 that increases the affinity of MBD2 for methylated DNA fragments [67]. The recovered DNA fragments were then used as a PCR template. This procedure preferentially recovers massively methylated regions of CpG islands. MIRA was later combined with microarray hybridization that allowed identification of a number of genes whose regulatory regions are methylated in primary lung cancers and melanomas [68].

Non-Methylated Genomic Sites Coincidence Cloning (NGSCC)

This approach [69] includes two stages (Fig. 3). At the first stage, genomic DNA fragments containing non-methylated recognition sites of the methylation-sensitive restriction endonuclease are selectively PCR-amplified using the PCR suppression technology [70]. At the second stage, a novel coincidence cloning procedure is used to select from this pool of fragments and clone the fragments of a specific genomic region. The method was applied to the analysis of hypomethylated CpG patterns along a one megabase human chromosome 19 region in normal testis and in seminoma tissues. Analysis of expression of genes located within hypomethylated regions showed that the expression of some genes was correlated with hypomethylation of the region, while other genes were expressed regardless of their methylation status. In combination with the rapid identification of genomic splits (RIDGES) technique [71], NGSCC allows determination of the methylation patterns of long DNA stretches.

Methylation Profiling Using *McrBC* Endonuclease

To analyze the methylation profile, sheared genomic DNA was treated with the methylation sensitive *McrBC* endonuclease, and the fraction depleted in methylated CpGs was isolated by agarose gel electrophoresis [72; 73]. This and a control untreated fraction were used for hybridization with microarray. The DNA methylation landscape of the human genome was monitored at more than 21,000 sites, including 79% of the annotated transcription start sites. The analysis revealed a complex pattern of DNA methylation along each autosome, with a trend of increasing density with approaching to telomeres. 2% of CpG islands were densely methylated, and 17% had significant levels of 5^mC irrespective of their relation to transcription start sites.

Identification of Methylated CpG Sites by Mass Spectrometry

A high-throughput identification of bisulfite-converted bases with the use of peptide nucleic acid (PNA) library hybridization and matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry was re-

Fig. (3). Summary of the non-methylated genomic sites coincidence cloning (NGSCC) procedure [69; 71]. Stage 1. Selective amplification of A-H and A-S fragments based on PCR suppression effect [70]. Methylated sites in genomic DNA are denoted by H (*Hpa* II recognition sites), counterpart non-methylated sites in cosmid DNA are denoted by S (*Sau*3A I recognition sites), cleavage sites of another frequent cutter enzyme (*Alu* I) are denoted by A. Suppression adapters for A and H/S are denoted by filled and empty rectangles, respectively. Stage 2. Selective amplification of sequences common between the whole genome and the region under study. The adapters at the ends of the selected fragments are removed, and new adapters ligated. The selectively amplified fragments of the genomic and cosmid DNAs are then mixed, denatured and annealed.

cently described [74]. This method is highly sensitive and provides reliable quantitation of the methylation level.

IIIb) Open and Closed Regions of Chromatin

Peculiarities of chromatin structure within and near transcribed regions have led to the concept of "open" (decondensed, active) and "closed" (compact, inactive) chromatin regions and their participation in regulation of gene activity

(reviewed in [75]). Accordingly, several approaches to high-throughput mapping of open and closed chromatin regions were proposed.

Mapping of Specific Chromatin Regions within Isolated Nuclei

Chromatin fiber structure across the human genome was analyzed using micrococcal nuclease digestion of isolated

nuclei and separation of compact and open chromatin regions by sucrose gradient centrifugation [76]. Distribution of compact and open chromatin regions within the genome was further investigated by hybridization with metaphase chromosomes and whole-genome microarrays with total genomic DNA as a reference. Open chromatin fibers correlate with regions of the highest gene density, but generally no correlation was found with the gene expression level.

Another direct chromatin fractionation scheme utilizes differential solubility of compact (H1-histone rich) and open chromatin. Total chromatin was treated with micrococcal nuclease and divided into three fractions - soluble in 5 mM MgCl₂, soluble in EDTA, and insoluble. Alternatively, the chromatin was fractionated by its accessibility to DNase I. The fractions obtained were hybridized to microarrays containing coding sequences of about 22000 genes, and the chromatin condensation data were compared to the expression data obtained using the same microarrays. It was found that expression is inversely proportional to the level of chromatin accessibility, however, this correlation was rather weak [77].

Mapping Regions Accessible to Dam Methylation In Vivo

A high-throughput method for *in vivo* identification of open and closed chromatin regions according to the accessibility of chromatin to the *E. coli* Dam methylase was proposed [78]. The method is based on the ability of Dam methylase to methylate adenine residues in GATC sequences [79] and on a technique similar to NGSCC described above [69]. Using this method, the distribution of open (accessible to Dam methylase) chromatin regions was studied within a 140 kb long human genome segment in HEK-293 cells. Dam methylated sites were found in gene introns, exons, and intergenic regions, and their distribution along DNA was uneven and corresponded to transcriptional activity of genes in the locus. Moreover, DNA regions accessible to Dam methylase apparently coincided with those hypersensitive to DNase I.

Mapping of Chromatin Regions Containing Modified Histones

Post-translational modification of histones was recently recognized as one of the major mechanisms of gene regulation. These modifications form an epigenetic pattern along the genome, or "histone code" [80], and provide binding sites for specific proteins that in turn induce alterations in chromatin structure and gene expression. Acetylation of lysine residues in histones is generally linked with open chromatin and active transcription, whereas methylation of lysine and arginine residues of histones is associated with either chromatin activation or silencing depending on methylation sites [81].

Genome-wide mapping of chromatin regions containing acetylated H3 in resting and activated human T cells was reported [82]. The authors used chromatin immunoprecipitation with anti-diacetylated K9/K14 histone H3 antibodies combined with a SAGE-like approach. It was found that only 1.2% of chromatin is associated with the K9/K14 diacetylated histone H3. Generally high levels of H3 acetylation were observed in gene-rich regions of the genome. High levels of acetylation were also detected in promoter regions. At

the same time, chromatin accessibility of a genetic locus was not caused by generally elevated acetylation. It was concluded, therefore, that chromatin accessibility and gene expression of a genome region are correlated with hyperacetylation of promoters and other regulatory elements, like enhancers, LCRs, and insulators, but not with generally elevated acetylation of the entire region.

The methylation state of histone H3 lysine-9 (K-9) was analyzed in human colon cancer cell line and tissue samples by chromatin immunoprecipitation with the use of anti-H3-K9 methylated and anti-H3-K9 acetylated antibodies and human CpG island microarrays [83]. The analysis demonstrated strong correlation of high-level K9 methylation/acetylation ratio with methylation of CpG islands.

Chromatin immunoprecipitation-on chip was also used to study the histone H3 methylation profile in coding regions of human genes. Antibodies to histone H3 molecules methylated at positions K4, K36, K79, associated with gene activation, and to those methylated at positions K9, K27 and K20, associated with gene repression, were used to map histone modifications in human THP-1 monocyte cells. The correlation between histone H3 methylation at different positions and H3 acetylation at positions K9/K14 was investigated. It was shown that H3-K4Me₂, H3-K4Me₃, H3-K36Me₂, and H3-K79Me₂ modifications were associated with hyperacetylation and active genes, whereas H3-K9Me₂, H3-K9Me₃, H3-K27Me₂, and H4-K20Me₂ were associated with hypoacetylation [84].

Using chromatin immunoprecipitation and direct cloning and sequencing of precipitated DNA fragments, the methylation of histone H4 arginine-3 was studied in HL-60 cells. It was found that all methylation sites identified were located in introns or within 5 kb from 5' ends of specific genes. A significant number of the genes in the proximity of the identified genomic loci marked by H4 arginine-3 methyl epitopes are involved in signaling pathways and developmental processes [85].

IIIc) Mapping of DNase I Hypersensitive Sites

Positions of DNase I hypersensitive sites often mark location of important genome regulatory elements, like promoters, enhancers, insulators etc. [86]. Several high-throughput methods of identification and mapping of these sites were recently proposed.

One approach includes limited digestion of isolated nuclei with DNase I, isolation of the solubilized DNA, blunting the ends, digestion with another enzyme to produce sticky ends, and directional cloning of the resulting fragments. A library of DNase I cleavage sites obtained in such a way from human CD4⁺ T-cells was sequenced, and positions of the cleavage sites (blunt ends) mapped. The inserts of the library clones were found to map more frequently to the genomic regions known to contain regulatory elements, such as regions upstream of genes and regions located within CpG islands or functionally significant sequences that align between mouse and human [87]. In the follow-up report [88], 230,000 DNase I cleavage tags were identified and mapped to the genome having revealed 14,190 clusters of sequences that grouped within close proximity to each other. By using a real-time PCR, it was found that the majority of these clus-

Fig. (4). A draft integrated map of a 160-kb human genome locus containing nine identified genes. Indicated are positions of protein target sites [51], putative insulators [38], nuclear matrix attachment regions (S/MARs) [33], open and closed chromatin regions as revealed by accessibility to Dam methylase [78], and CpG unmethylated sites [69; 71]. The map was prepared with the use of the Human Genome Browser at UCSC [92].

ters represent valid DNase hypersensitive sites. Approximately 80% of these sites uniquely mapped within one or more annotated regions of the genome believed to contain regulatory elements, including regions 2 kb upstream of genes, CpG islands, and highly conserved sequences.

Another research team [89] used subtractive hybridization to prepare a library enriched in DNA fragments containing DNase I cleavage sites. The fragments were mapped to a ~30 Mbp region of human chromosome 21. The DNase I sensitive sites were predominantly found near transcription start sites of known genes and non-gene-associated CpG islands.

The methods of DNase I cleavage sites identification were recently adapted for hybridization with genome microarrays. Being similar in principle, such approaches use somewhat different methods of probe preparation. For example, one of these methods [90] includes biotinylation of DNA fragments formed after treatment with DNase and isolation of these fragments on a streptavidin column. Alternatively, a probe can be prepared by isolation of short DNA fragments containing closely spaced DNase I cleavage sites [91]. Both probes were hybridized to tiled microarrays representing 44 human genome regions of the ENCODE project [19].

IV) DRAFT INTEGRATED MAP OF REGULATORY ELEMENTS

Fig. (4) represents the first attempt to construct an integrated map of regulatory elements in a 160-kb human genome region. The map was constructed using data obtained in [33; 38; 52; 69; 71; 78]. It shows positions of protein binding sites, insulators and S/MAR-elements, and of some epigenetic elements - regions accessible to Dam methylase, as well as non-methylated CpGs. The map is obviously far from complete yet, especially regarding tissue specificity of most regulatory sequences. Nevertheless, it allows to get the first impression about the complexity of genome regulatory networks and the problem of their study.

V) CONCLUSIONS

Correct positioning of a multitude of *cis*-acting and epigenetic elements within a sequenced genome is a great challenge that has to be met in order to construct correct functional genomic maps. The scale of the problem (identification and mapping of millions of functional elements for hundreds of different cell types) is enormous. Therefore, we suggest a strategy that consists in functional characterizing relatively short genomic fragments followed by combining these data into a whole genome functional map comprising positions and functional characteristics of regulatory modules. As the first step, we propose to develop a series of novel approaches that would allow identification of regulatory and epigenetic elements within such short fragments and construction of cell type specific genomic DNA sub-libraries enriched in the elements of interest. The libraries will serve as a source of the elements for their subsequent mapping. Moreover, thoroughly characterized short genomic fragments can be good test objects for new methods and can be considered good model systems for better understanding the concerted regulation of large genetic ensembles.

ACKNOWLEDGEMENTS

We are grateful to B. Glotov for critical reading of the manuscript. The work was supported by the Russian Foundation for Basic Research (project 05-04-48814), the Scientific School Program of the Russian Federation President (project NSH 2006.2003.4), program of the Russian Academy of Sciences on Molecular and Cellular Biology, and INTAS (project 01-0279).

REFERENCES

- [1] Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman,

- D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Showkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornisch, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hunkamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J. Initial sequencing and analysis of the human genome. *International Human Genome Sequencing Consortium. Nature* **2001**, 409: 860-921.
- [2] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor, Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkuch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X. The sequence of the human genome. *Science* **2001**, 291: 1304-1351.
- [3] Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyas, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J.P., Von, Niederhausern, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C., Lander, E.S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**, 420: 520-562.
- [4] Carninci, P. Tagging mammalian transcription complexity. *Trends Genet.* **2006**, 22: 501-510.
- [5] Jackson, D. A. The amazing complexity of transcription factories. *Brief. Funct. Genomic Proteomic* **2005**, 4: 143-157.
- [6] Maston, G. A., Evans, S. K. and Green, M. R. Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **2006**, 7: 29-59.

- [7] Pennisi, E. Searching for the genome's second code. *Science* **2004**, 306: 632-635.
- [8] Shields, R. The emperor's new clothes revisited. *Trends Genet.* **2006**, 22: 463.
- [9] Levine, M. and Davidson, E. H. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. USA* **2005**, 102: 4936-4942.
- [10] Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Caestani, C., Yuh, C. H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livi, C. B., Lee, P. Y., Revilla, R., Rust, A. G., Pan, Z., Schilstra, M. J., Clarke, P. J., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L. and Bolouri, H. A genomic regulatory network for development. *Science* **2002**, 295: 1669-1678.
- [11] Oliveri, P. and Davidson, E. H. Gene regulatory network controlling embryonic specification in the sea urchin. *Curr. Opin. Genet. Dev.* **2004**, 14: 351-360.
- [12] Stathopoulos, A. and Levine, M. Genomic regulatory networks and animal development. *Dev Cell* **2005**, 9: 449-462.
- [13] Longabaugh, W. J., Davidson, E. H. and Bolouri, H. Computational representation of developmental genetic regulatory networks. *Dev. Biol.* **2005**, 283: 1-16.
- [14] Koide, T., Hayata, T. and Cho, K. W. Xenopus as a model system to study transcriptional regulatory networks. *Proc. Natl. Acad. Sci. USA* **2005**, 102: 4943-4948.
- [15] Stathopoulos, A. and Levine, M. Whole-genome analysis of Drosophila gastrulation. *Curr. Opin. Genet. Dev.* **2004**, 14: 477-484.
- [16] Inoue, T., Wang, M., Ririe, T. O., Fernandes, J. S. and Sternberg, P. W. Transcriptional network underlying Caenorhabditis elegans vulval development. *Proc. Natl. Acad. Sci. USA* **2005**, 102: 4972-4977.
- [17] Singh, H., Medina, K. L. and Pongubala, J. M. Contingent gene regulatory networks and B cell fate specification. *Proc. Natl. Acad. Sci. USA* **2005**, 102: 4949-4953.
- [18] Ochoa-Espinosa, A., Yucel, G., Kaplan, L., Pare, A., Pura, N., Oberstein, A., Papatsenko, D. and Small, S. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *Proc. Natl. Acad. Sci. USA* **2005**, 102: 4960-4965.
- [19] ENCODE and consortium The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **2004**, 306: 636-60.
- [20] Trinklein, N. D., Aldred, S. J., Saldanha, A. J. and Myers, R. M. Identification and functional analysis of human transcriptional promoters. *Genome Res.* **2003**, 13: 308-312.
- [21] Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L. and Myers, R. M. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* **2006**, 16: 1-10.
- [22] Khambata-Ford, S., Liu, Y., Gleason, C., Dickson, M., Altman, R. B., Batzoglou, S. and Myers, R. M. Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. *Genome Res.* **2003**, 13: 1765-1774.
- [23] Kim, T. H., Barrera, L. O., Qu, C., Van Calcar, S., Trinklein, N. D., Cooper, S. J., Luna, R. M., Glass, C. K., Rosenfeld, M. G., Myers, R. M. and Ren, B. Direct isolation and identification of promoters in the human genome. *Genome Res.* **2005**, 15: 830-839.
- [24] Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D. and Ren, B. A high-resolution map of active promoters in the human genome. *Nature* **2005**, 436: 876-880.
- [25] Bondarenko, V. A., Liu, Y. V., Jiang, Y. I. and Studitsky, V. M. Communication over a large distance: enhancers and insulators. *Biochem. Cell Biol.* **2003**, 81: 241-251.
- [26] West, A. G. and Fraser, P. Remote control of gene transcription. *Hum. Mol. Genet.* **2005**, 14 Spec No 1: R101-111.
- [27] Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E. and Taipale, J. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **2006**, 124: 47-59.
- [28] Wei, G. H., Liu de, P. and Liang, C. C. Chromatin domain boundaries: insulators and beyond. *Cell Res.* **2005**, 15: 292-300.
- [29] Brasset, E. and Vaury, C. Insulators are fundamental components of the eukaryotic genomes. *Heredity* **2005**, 94: 571-576.
- [30] Gaszner, M. and Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.* **2006**, 7: 703-713.
- [31] Chernov, I. P., Akopov, S. B. and Nikolaev, L. G. Structure and function of nuclear matrix associated regions (S/MARs). *Russ. J. Bioorg. Chem.* **2004**, 30: 1-11.
- [32] Nikolaev, L. G., Tsevegiyn, T., Akopov, S. B., Ashworth, L. K. and Sverdlov, E. D. Construction of a chromosome specific library of human MARs and mapping of matrix attachment regions on human chromosome 19. *Nucleic Acids Res.* **1996**, 24: 1330-1336.
- [33] Chernov, I. P., Akopov, S. B., Nikolaev, L. G. and Sverdlov, E. D. Identification and mapping of nuclear matrix attachment regions in a one megabase locus of human chromosome 19q13.12: Long-range correlation of S/MARs and gene positions. *J. Cell. Biochem.* **2002**, 84: 590-600.
- [34] Shaposhnikov, S. A., Akopov, S. B., Chernov, I. P., Thomsen, P. D., Joergensen, C., Collins, A. R., Frengen, E. and Nikolaev, L. G. A map of nuclear matrix attachment regions within the breast cancer loss of heterozygosity region on human chromosome 16q22.1. *Genomics* **2007**, 89: 354-361.
- [35] Bell, A. C., West, A. G. and Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **1999**, 98: 387-396.
- [36] Mukhopadhyay, R., Yu, W., Whitehead, J., Xu, J., Lezcano, M., Pack, S., Kanduri, C., Kanduri, M., Ginja, V., Vostrov, A., Quitschke, W., Chernukhin, I., Klenova, E., Lobanenko, V. and Ohlsson, R. The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res.* **2004**, 14: 1594-1602.
- [37] Magdinier, F., Yusufzai, T. M. and Felsenfeld, G. Both CTCF-dependent and -independent insulators are found between the mouse T cell receptor alpha and Dad1 genes. *J. Biol. Chem.* **2004**, 279: 25381-25389.
- [38] Akopov, S. B., Ruda, V. M., Batrak, V. V., Vetchinova, A. S., Chernov, I. P., Nikolaev, L. G., Bode, J. and Sverdlov, E. D. Identification, genome mapping, and CTCF binding of potential insulators within the FXYD5-COX7A1 locus of human Chromosome 19q13.12. *Mamm. Genome* **2006**, 17: 1042-1049.
- [39] Kellum, R. and Schedl, P. A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Mol. Cell. Biol.* **1992**, 12: 2424-2431.
- [40] Sass, A. V., Ruda, V. M., Akopov, S. B., Snezhkov, E. V., Nikolaev, L. G. and Sverdlov, E. D. Regulatory potential of S/MAR elements in transient expression. *Bioorg. Khim.* **2005**, 31: 77-81.
- [41] Namciu, S. J., Friedman, R. D., Marsden, M. D., Sarausad, L. M., Jasoni, C. L. and Fournier, R. E. Sequence organization and matrix attachment regions of the human serine protease inhibitor gene cluster at 14q32.1. *Mamm. Genome* **2004**, 15: 162-178.
- [42] Mesner, L. D., Crawford, E. L. and Hamlin, J. L. Isolating apparently pure libraries of replication origins from complex genomes. *Mol. Cell* **2006**, 21: 719-726.
- [43] Kadonaga, J. T. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* **2004**, 116: 247-257.
- [44] Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T. E., Luscombe, N. M., Rinn, J. L., Nelson, F. K., Miller, P., Gerstein, M., Weissman, S. and Snyder, M. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci. USA* **2003**, 100: 12247-12252.
- [45] Euskirchen, G., Royce, T. E., Bertone, P., Martone, R., Rinn, J. L., Nelson, F. K., Sayward, F., Luscombe, N. M., Miller, P., Gerstein, M., Weissman, S. and Snyder, M. CREB binds to multiple loci on human chromosome 22. *Mol. Cell. Biol.* **2004**, 24: 3804-3814.
- [46] Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K. and Gingeras, T. R. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **2004**, 116: 499-509.
- [47] van Steensel, B. and Henikoff, S. Identification of *in vivo* DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* **2000**, 18: 424-428.
- [48] Orian, A. Chromatin profiling, DamID and the emerging landscape of gene expression. *Curr. Opin. Genet. Dev.* **2006**, 16: 157-164.
- [49] van Steensel, B., Delrow, J. and Bussemaker, H. J. Genomewide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding. *Proc. Natl. Acad. Sci. USA* **2003**, 100: 2580-2585.

- [50] Orian, A., van Steensel, B., Delrow, J., Bussemaker, H. J., Li, L., Sawado, T., Williams, E., Loo, L. W., Cowley, S. M., Yost, C., Pierce, S., Edgar, B. A., Parkhurst, S. M. and Eisenman, R. N. Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. *Genes Dev.* **2003**, *17*: 1101-1114.
- [51] Chernov, I. P., Akopov, S. B., Nikolaev, L. G. and Sverdlov, E. D. Identification and mapping of DNA binding proteins target sequences in long genomic regions by two-dimensional EMSA. *Bio-Techniques* **2006**, *41*: 90-96.
- [52] Vetchinova, A. S., Akopov, S. B., Chernov, I. P., Nikolaev, L. G. and Sverdlov, E. D. Two-dimensional electrophoretic mobility shift assay: identification and mapping of transcription factor CTCF target sequences within an FXVD5-COX7A1 region of human chromosome 19. *Anal. Biochem.* **2006**, *354*: 85-93.
- [53] Bird, A. DNA methylation patterns and epigenetic memory. *Genes. Dev.* **2002**, *16*: 6-21.
- [54] Fraga, M. F. and Esteller, M. DNA methylation: a profile of methods and applications. *Biotechniques* **2002**, *33*: 632, 634, 636-649.
- [55] Dahl, C. and Guldberg, P. DNA methylation analysis techniques. *Biogerontology* **2003**, *4*: 233-250.
- [56] Ushijima, T., Morimura, K., Hosoya, Y., Okonogi, H., Tatematsu, M., Sugimura, T. and Nagao, M. Establishment of methylation-sensitive-representational difference analysis and isolation of hypo- and hypermethylated genomic fragments in mouse liver tumors. *Proc. Natl. Acad. Sci. USA* **1997**, *94*: 2284-2289.
- [57] Kaneda, A., Takai, D., Kaminishi, M., Okochi, E. and Ushijima, T. Methylation-sensitive representational difference analysis and its application to cancer research. *Ann. N. Y. Acad. Sci.* **2003**, *983*: 131-141.
- [58] Lisitsyn, N. and Wigler, M. Cloning the differences between two complex genomes. *Science* **1993**, *259*: 946-951.
- [59] Hayashizaki, Y., Hirotsune, S., Okazaki, Y., Hatada, I., Shibata, H., Kawai, J., Hirose, K., Watanabe, S., Fushiki, S., Wada, S., Sugimoto, T., Kobayakawa, K., Kawara, T., Katsuki, M., Shibuya, T. and Mukai, T. Restriction landmark genomic scanning method and its various applications. *Electrophoresis* **1993**, *14*: 251-258.
- [60] Costello, J. F., Fruhwald, M. C., Smiraglia, D. J., Rush, L. J., Robertson, G. P., Gao, X., Wright, F. A., Feramisco, J. D., Peltomaki, P., Lang, J. C., Schuller, D. E., Yu, L., Bloomfield, C. D., Caligiuri, M. A., Yates, A., Nishikawa, R., Su Huang, H., Petrelli, N. J., Zhang, X., O'Dorisio, M. S., Held, W. A., Cavenee, W. K. and Plass, C. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat. Genet.* **2000**, *24*: 132-138.
- [61] Adorjan, P., Distler, J., Lipscher, E., Model, F., Muller, J., Pelet, C., Braun, A., Florl, A. R., Gutig, D., Grabs, G., Howe, A., Kursar, M., Lesche, R., Leu, E., Lewin, A., Maier, S., Muller, V., Otto, T., Scholz, C., Schulz, W. A., Seifert, H. H., Schwöpe, I., Ziebarth, H., Berlin, K., Piepenbrock, C. and Olek, A. Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res.* **2002**, *30*: e21.
- [62] Gitan, R. S., Shi, H., Chen, C. M., Yan, P. S. and Huang, T. H. Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. *Genome Res.* **2002**, *12*: 158-164.
- [63] Shi, H., Maier, S., Nimmrich, I., Yan, P. S., Caldwell, C. W., Olek, A. and Huang, T. H. Oligonucleotide-based microarray for DNA methylation analysis: principles and applications. *J. Cell. Biochem.* **2003**, *88*: 138-143.
- [64] Rakyán, V. K., Hildmann, T., Novik, K. L., Lewin, J., Tost, J., Cox, A. V., Andrews, T. D., Howe, K. L., Otto, T., Olek, A., Fischer, J., Gut, I. G., Berlin, K. and Beck, S. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.* **2004**, *2*: e405.
- [65] Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., Thomas, N. J., Wang, Y., Vollmer, E., Goldmann, T., Seifart, C., Jiang, W., Barker, D. L., Chee, M. S., Floros, J. and Fan, J. B. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.* **2006**, *16*: 383-393.
- [66] Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L. and Schubeler, D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **2005**, *37*: 853-862.
- [67] Rauch, T. and Pfeifer, G. P. Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. *Lab. Invest.* **2005**, *85*: 1172-1180.
- [68] Rauch, T., Li, H., Wu, X. and Pfeifer, G. P. MIRA-Assisted Microarray Analysis, a New Technology for the Determination of DNA Methylation Patterns, Identifies Frequent Methylation of Homeodomain-Containing Genes in Lung Cancer Cells. *Cancer Res.* **2006**, *66*: 7939-7947.
- [69] Azhikina, T., Gainetdinov, I., Skvortsova, Y., Batrak, A., Dmitrieva, N. and Sverdlov, E. Non-methylated Genomic Sites Coincidence Cloning (NGSCC): an approach to large scale analysis of hypomethylated CpG patterns at predetermined genomic loci. *Mol. Genet. Genomics* **2004**, *271*: 22-32.
- [70] Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. and Siebert, P. D. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl. Acad. Sci. USA* **1996**, *93*: 6025-6030.
- [71] Azhikina, T., Gainetdinov, I., Skvortsova, Y. and Sverdlov, E. Methylation-free site patterns along a 1-Mb locus on Chr19 in cancerous and normal cells are similar. A new fast approach for analyzing unmethylated CCGG sites distribution. *Mol. Genet. Genomics* **2006**, *275*: 615-622.
- [72] Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., Lavine, K., Mittal, V., May, B., Kasschau, K. D., Carrington, J. C., Doerge, R. W., Colot, V. and Martienssen, R. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **2004**, *430*: 471-476.
- [73] Lippman, Z., Gendrel, A. V., Colot, V. and Martienssen, R. Profiling DNA methylation patterns using genomic tiling microarrays. *Nat. Methods* **2005**, *2*: 219-224.
- [74] Schatz, P., Distler, J., Berlin, K. and Schuster, M. Novel method for high throughput DNA methylation marker evaluation using PNA-probe library hybridization and MALDI-TOF detection. *Nucleic Acids Res.* **2006**, *34*: e59.
- [75] Woodcock, C. L. Chromatin architecture. *Curr. Opin. Struct. Biol.* **2006**, *16*: 213-220.
- [76] Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P. and Bickmore, W. A. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **2004**, *118*: 555-566.
- [77] Weil, M. R., Widlak, P., Minna, J. D. and Garner, H. R. Global survey of chromatin accessibility using DNA microarrays. *Genome Res.* **2004**, *14*: 1374-1381.
- [78] Bulanenkova, S., Snezhkov, E., Nikolaev, L. and Sverdlov, E. Identification and mapping of open chromatin regions within a 140 kb polygenic locus of human chromosome 19 using *E. coli* Dam methylase. *Genetica* **2006**, In press.
- [79] Singh, J. and Klar, A. J. Active genes in budding yeast display enhanced *in vivo* accessibility to foreign DNA methylases: a novel *in vivo* probe for chromatin structure of yeast. *Genes Dev.* **1992**, *6*: 186-196.
- [80] Jenuwein, T. and Allis, C. D. Translating the histone code. *Science* **2001**, *293*: 1074-1080.
- [81] Lam, A. L., Pazin, D. E. and Sullivan, B. A. Control of gene expression and assembly of chromosomal subdomains by chromatin regulators with antagonistic functions. *Chromosoma* **2005**, *114*: 242-251.
- [82] Roh, T. Y., Cuddapah, S. and Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **2005**, *19*: 542-552.
- [83] Kondo, Y., Shen, L., Yan, P. S., Huang, T. H. and Issa, J. P. Chromatin immunoprecipitation microarrays for identification of genes silenced by histone H3 lysine 9 methylation. *Proc. Natl. Acad. Sci. USA* **2004**, *101*: 7398-7403.
- [84] Miao, F. and Natarajan, R. Mapping global histone methylation patterns in the coding regions of human genes. *Mol. Cell. Biol.* **2005**, *25*: 4650-4661.
- [85] Balint, B. L., Gabor, P. and Nagy, L. Genome-wide localization of histone 4 arginine 3 methylation in a differentiation primed myeloid leukemia cell line. *Immunobiology* **2005**, *210*: 141-152.
- [86] Felsenfeld, G. and Groudine, M. Controlling the double helix. *Nature* **2003**, *421*: 448-453.
- [87] Crawford, G. E., Holt, I. E., Mullikin, J. C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E. D., Wolfsberg, T. G. and Collins, F. S. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl. Acad. Sci. USA* **2004**, *101*: 992-997.

- [88] Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G. and Collins, F. S. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **2006**, *16*: 123-131.
- [89] Sabo, P. J., Humbert, R., Hawrylycz, M., Wallace, J. C., Dorschner, M. O., McArthur, M. and Stamatoyannopoulos, J. A. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl. Acad. Sci. USA* **2004**, *101*: 4537-4542.
- [90] Crawford, G. E., Davis, S., Scacheri, P. C., Renaud, G., Halawi, M. J., Erdos, M. R., Green, R., Meltzer, P. S., Wolfsberg, T. G. and Collins, F. S. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods* **2006**, *3*: 503-509.
- [91] Sabo, P. J., Kuehn, M. S., Thurman, R., Johnson, B. E., Johnson, E. M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., Weaver, M., Shafer, A., Lee, K., Neri, F., Humbert, R., Singer, M. A., Richmond, T. A., Dorschner, M. O., McArthur, M., Hawrylycz, M., Green, R. D., Navas, P. A., Noble, W. S. and Stamatoyannopoulos, J. A. Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nat. Methods* **2006**, *3*: 511-518.
- [92] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. The human genome browser at UCSC. *Genome Res.* **2002**, *12*: 996-1006.