Identification and mapping of DNA binding proteins target sequences in long genomic regions by two-dimensional EMSA

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

BioTechniques 41:90-96 (July 2006)
doi 10.2144/000112197

Specific binding of nuclear proteins, in particular transcription factors, to target DNA sequences is a major mechanism of genome functioning and gene expression regulation in eukaryotes. Therefore, identification and mapping specific protein target sites (PTS) is necessary for understanding genomic regulation. Here we used a novel two-dimensional electrophoretic mobility shift assay (2D-EMSA) procedure for identification and mapping of 52 PTS within a 563-kb human genome region located between the FXYD5 and TZFP genes. The PTS occurred with approximately equal frequency within unique and repetitive genomic regions. PTS belonging to unique sequences tended to group together within gene introns and close to their 5' and 3' ends, whereas PTS located within repeats were evenly distributed between transcribed and intragenic regions.

INTRODUCTION

The publication of the human genome sequence (1,2) and sequences of other metazoan genomes greatly facilitated positioning and analysis of various genomic functional elements and first of all coding sequences (3,4). At the same time, a complete functional annotation of sequenced eukaryotic genomes is supposed to include positions of all noncoding regulatory elements. Unfortunately, experimental data on genomic positions of a multitude of regulatory sequences, like enhancers, promoters, transcription terminators, and replication origins, are very limited, especially at the whole genome level. In general, most genomic regulatory elements (e.g., enhancers) are gene-, tissue-, or cell-specific, and prediction of these elements by computational methods is difficult and not always reliable. Therefore, the development of high-throughput experimental approaches to identification and mapping of genomic functional elements is highly desirable.

Specific binding of nuclear proteins to target DNA sequences is a major

mechanism of genome functioning and regulation in eukaryotes (5), that makes identification and mapping of specific protein target sites (PTS) necessary for understanding genomic regulation. To date, several approaches to unbiased mapping of PTS have been proposed and used. The most widely used is a chromatin immunoprecipitation-on-a-chip (ChIP-on-chip) technique that allowed to map target sites for the NF- κ B (6) and CREB (7) transcription factors across human chromosome 22 and the Sp1, c-Myc, and p53 factors across human chromosomes 21 and 22 (8). Another experimental approach named DamID (9) was recently used for mapping GAGA (10), Myc, Max, and Mad/Mnt target sites across the whole *Drosophila* genome (11). It should be noted that the both techniques are applicable only to mapping binding sites of well-characterized transcription factors.

Computational identification of PTS is, in turn, strongly limited by the lack of experimental data necessary for development of algorithms and validation of the results (12,13). A general approach should include

identification of a whole set of specific PTS and grouping them according to their functional role and interactions with other regulatory units. The result should be a protein binding map of extended genomic regions or even whole genomes, ideally a dynamic map depending on cell origin, environmental conditions, and other factors.

Recently, we proposed experimental approaches for identification and mapping of nuclear matrix binding regions (S/MARs) (14) within a 1-Mb human chromosome 19 locus between the *FXYD5* and *COX7A1* markers (15). The locus contains 45 Reference Sequence (16) genes expressed with different tissue specificities and therefore could be a good model for the study of the mammalian genome regulatory network. Here we present an approach for high-throughput identification and mapping of a multitude of PTS within a given genomic region. Using this approach, we mapped 52 sequences capable of specifically binding Jurkat cell nuclear proteins within a 563-kb long *FXYD5-TZFP* human chromosome 19 region, a fragment of the *FXYD5-COX7A1* locus mentioned above.

MATERIALS AND METHODS

Basic Protocols

Growth and transformation of *Escherichia coli* cells, preparation of plasmid DNA, agarose gel electrophoresis, electrophoretic mobility shift assay (EMSA), and other standard manipulations were performed as described (17).

Cells and Nuclear Extract

Jurkat cells (acute T cell leukemia, TIB-152; ATCC, Manassas, VA, USA) were grown in suspension at 37°C and 5% CO₂ in RPMI-1640 supplemented with 10% fetal calf serum, up to approximately 2 × 10⁶/mL. Nuclear extract was isolated as previously described (18) with modifications (19).

Preparation of a Short-Fragment Library

DNA of cosmids R30072, R28588, F19410, R30879, F24108, F16632, R26667, F12426, R28461, F14121, R31396, F25451, R31076, R28052, and P1-derived artificial chromosome (PAC) PC28130 (kindly provided by A. Olsen, Lawrence Livermore National Laboratory, Livermore, CA) was isolated and digested with restriction endonucleases *Sau*3A and *Csp*6I, ligated with the library primer 5'-ACTTGAGCTCGAGTATCCATGAACA-3', and PCR-amplified with the same primer as described previously (14,20).

Two-Dimensional EMSA

For two-dimensional EMSA (2D-EMSA), a pool of short DNA fragments of a 563-kb *FXRD5-TZFP* region of human chromosome 19 was radioactively labeled by PCR with the library primer and purified as described previously (21). The 2D-EMSA was performed generally as described (21), but instead of purified DNA binding protein, 2.5 µg Jurkat cell nuclear extract protein was added to the initial EMSA reaction.

The gel was then autoradiographed overnight, the area containing PTS (see Figure 1A) was excised, cut into small pieces, and DNA was eluted and precip-

itated as described (21). To increase the specificity of selection, the above 2D-EMSA procedure was repeated twice. Finally, 2 µL polyacrylamide gel eluate were PCR-amplified (20 cycles of 94°C for 20 s, 60°C for 60 s, and 72°C for 90 s) and cloned in a pGEM®-T vector (Promega, Madison, WI, USA) according to the manufacturer's recommendations. White colonies (184) were selected and arrayed on a 96-well microplate. The selected clones were checked by PCR, and those lacking inserts or producing more than one PCR product (double inserts) were discarded.

One-Dimensional EMSA

For EMSA, inserts of individual clones were labeled by PCR as described above and purified by electrophoresis in a 5% polyacrylamide gel. EMSA was done essentially as described above with a 50,000 counts per minute (cpm) probe, 1 µg nuclear extract protein, and 1 µg poly(dI-dC)*poly(dI-dC). For competition experiments, an excess of an unlabeled probe was added.

Sequencing, Computer Analysis, and Mapping

Sequencing was done with a ABI PRISM® BigDye™ Terminator v. 3.1 kit using an ABI PRISM 3100-Avant™ automated sequencer (all from Applied Biosystems, Foster City, CA, USA). The sequences obtained were mapped by comparison with those deposited in GenBank® using the BLAST (22) server at the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov/BLAST). The data were further analyzed using the University of California, Santa Cruz (UCSC) Human Genome Browser (genome.ucsc.edu) (23).

RESULTS

EMSA is one of the most widely used methods to explore interactions between DNA and nuclear proteins. The EMSA approach was initially proposed for quantifying interactions between DNA and purified proteins (24,25) and was later adapted for crude cellular or

nuclear protein extracts (26). Recently, we proposed a two-dimensional variant of EMSA (2D-EMSA) that allowed us to identify and map binding sites of the CTCF transcription factor within a 1-Mb human genome region (21). Here we present a modification of 2D-EMSA that allows one to obtain and clone DNA fragments capable of binding nuclear extract proteins of given cells, with the pattern of the fragments probably being

Figure 1. The principle and results of two-dimensional electrophoretic mobility shift assay (2D-EMSA). (A) General scheme of 2D-EMSA. DNA-protein complexes were initially separated in a nondenaturing one-dimension polyacrylamide gel, and after disruption of the complexes, the DNA fragments released were separated in a two-dimension sodium dodecyl sulfate (SDS)-containing gel. The area of spots corresponding to target DNA sequences is outlined by the dashed line. (B) The result of 2D-EMSA with nuclear extract from Jurkat cells and DNA fragments representing the *FXRD5-TZFP* region of human chromosome 19. (C) Electrophoretic comparison of input DNA fragments with protein target sites (PTS) selected by 2D-EMSA. The most pronounced bands are marked by arrows.

Figure 1A presents the principle of 2D-EMSA. A pool of short DNA fragments covering the genome region of interest is prepared and labeled with ^{32}P . This pool is mixed with nuclear extract prepared from a specific cell line and containing DNA binding proteins characteristic of this cell type. Then, the DNA-protein complexes are separated in a nondenaturing first-dimension gel, as in the conventional EMSA. The resulting gel strip that contains DNA-protein complexes and free DNA fragments is localized by autoradiography, excised from the gel, and incubated in sodium dodecyl sulfate (SDS)-containing buffer to disrupt DNA-protein complexes. The strip is then loaded onto a second-dimension gel (the same as for the first dimension but with 0.1% SDS), and the gel was run and autoradiographed. Figure 1B shows the results of the 2D-EMSA. The diagonal spots in the resulting gel represent the fragments with approximately the same electrophoretic mobility in both dimensions (i.e., the fragments that did not bind to proteins). The fragments bound to proteins in the first dimension are retarded, but their mobility is restored in the second dimension due to dissociation of DNA-protein complexes. Therefore, the spots corresponding to the fragments initially bound by nuclear proteins are located below the diagonal, and the area

Construction and Properties of a Nuclear PTS Library

After two rounds of 2D-EMSA, the area outlined by a dashed line (Figure 1B) was cut out, and the DNA was eluted. To qualitatively estimate the efficiency of the selection procedure, the DNA was labeled and separated in a single 40-cm denaturing polyacrylamide gel side-by-side with the initial short-fragment library (Figure 1C). As seen from the figure, the two patterns of fragments obtained were substantially different.

The DNA fragments obtained in two rounds of 2D-EMSA were PCR-amplified, and the resulting library of potential target sites of nuclear DNA binding proteins was cloned and arrayed.

Number of PTS in the Genome

The protein binding fragments were sequenced, and their sequences compared with the Human Genome Database. In total, 78 target sequences belonging to the *FXYD5-TZFP* region were identified. The remaining 20 sequences were mapped to other human genome regions or belonged to the *E. coli* genome. A comparison of the 78 sequences with each other revealed 52 unique sequences. These sequences were mapped across the *FXYD5-TZFP* region of human chromosome 19 (Figure 3), and their precise positions in the corresponding genome context are presented in Supplementary Table S1 (available online at www.BioTechniques.com).

To estimate the number of independent clones in the library, we assumed that the frequency of occurrence of cloned sequences in library samples (see Supplementary Table S1) fit the Poisson distribution. Accordingly, the Poisson curve was adjusted by the least-squares method to fit the data obtained and used to calculate the library size as (the number of selected clones)/ q , where q is a parameter of the Poisson distribution. The estimated number was found to be about 120. Therefore, 52 PTS found here may represent about a half of potential PTS of this region identifiable by 2D-EMSA in Jurkat cells.

The possible number of protein binding sites in a genomic region of a given length can be evaluated. It can be assumed that expression of each gene is controlled by several transcription factors, so this number should depend on the number of both active and silenced genes in the region. Recent

Figure 2. Verification of DNA-binding properties of the selected protein target sites (PTS). (A) Electrophoretic mobility shift assay (EMSA) analysis of eight randomly selected DNA fragments (PTS) obtained by two-dimensional EMSA (2D-EMSA). (B) EMSA competition analysis of three fragments capable of nuclear protein(s) binding. Visual disappearance of DNA-protein complexes after addition of an unlabeled probe (10- and 20-fold molar excess) indicates specificity of DNA-protein binding. NE, nuclear extract.

Figure 3. Map of the *FXYD5-TZFP* locus. Genes are designated by blue horizontal arrows with arrowheads indicating the direction of transcription. Vertical arrows designate locations of unique (red arrows) and repeat-containing (green arrows) nuclear protein target sites (PTS). The identified genes are as follows (from left to right): *FXYD5*, FXYD domain-containing ion transport regulator 5; *LISCH7*, liver-specific transcription factor; *USF2*, upstream transcription factor 2, *c-fos* interacting; *HAMP* (*Leap-1*), liver-expressed antimicrobial peptide; *MAG*, myelin-associated glycoprotein; *CD22*, CD22 antigen; *GPR40*, 41, 43, G protein-coupled receptors 40, 41, and 43; *UNQ46*, gene with unknown function; *BX648076*, gene with unknown function; *GAPDS*, glyceraldehyde-3-phosphate dehydrogenase, testis-specific; *NIFIE14*, seven transmembrane domain protein; *ATP4A*, ATPase, H⁺/K⁺ exchanging, α polypeptide; *BC064390*, gene with unknown function; *ETV2*, *ETS* variant gene 2; *COX6B*, cytochrome c oxidase subunit VIb; *UPK1A*, uroplakin 1A; and *TZFP*, testis zinc-finger protein.

estimations for the complete human genome (3000 Mb) were 12,000 Sp1, 25,000 c-Myc, 1600 p53 (8), and about 12,000 NF- κ B target sites (6). We have identified at least 52 nuclear PTS in the *FXYD5-TZFP* region of 563 kb in length (over 270,000 if extrapolated to the whole genome) in a single cell type, which should represent a considerable fraction of all PTS. Taking into account the calculations above, the actual number could be twice as large.

PTS Map

All identified PTS (Supplementary Table S1) were subdivided into two almost equal groups—unique and repeat-containing, the members of which contain over 25% repeated sequences. The repeats were most often represented by Alu or long-interspersed nuclear elements (LINEs) retroelements, as well as long terminal repeats (LTRs) of human endogenous retroviruses. A complete list and the positions of the repeats can be found in Supplementary Table S2. The finding that half of the protein binding sites were embedded in repeats suggests an important role of repeated elements, especially retroelements, in gene regulation.

All repeat-containing and unique PTS were placed on a physical map of the *FXYD5-TZFP* region (Figure 3). An interactive map can be obtained with the use of the Human Genome

Browser (genome.ucsc.edu/cgi-bin/hgGateway?db=hg12), assembly of May 2004, and an annotation file in Supplementary Table S3. The distribution of PTS within the region is summarized in Table 1. About 50% of the *FXYD5-TZFP* region is occupied by genes and expressed sequence tag (ESTs), and the rest is occupied by intergenic sequences. The data presented here support the earlier observation (6,8) that PTS are not restricted to the 5' regions of genes, but just as often are localized 3' to or within genes.

All PTS were subdivided into intronic, 5' and 3' sites (located within 1 kb from the 5' or 3' ends of genes, respectively), and intragenic PTS depending on their positions relative to genes and ESTs. As seen from Table 1, unique PTS tend to be located inside or near genes—only 2 of 25 unique sites were found >1 kb distant from the corresponding gene. In contrast, over

50% (14 of 27) of the repeat-containing PTS were located within intragenic regions.

Although some target sequences overlap with exons, none was found within exons. However, as borders of binding sites within target sequences are not known, the existence of exonic binding sites cannot be completely ruled out.

DISCUSSION

Recently, we utilized the 2D-EMSA technique for identification of binding sites for a well-characterized transcription factor. This technique allowed us to find 10 new binding sites of the CTCF protein within the human chromosome 19 *FXYD5-COX7A1* locus (21). In the present work, we applied the 2D-EMSA to a much more complex system that includes a multitude of DNA binding proteins present in the nuclear extract of Jurkat cells.

Table 1. Transcription Factor Binding Sites Statistics

Position	Repeated Sites	Unique Sites	Repeated Plus Unique Sites
3' to gene/EST	3	4	7
5' to gene/EST	2	6	8
Intronic	8	13	21
Intergenic	14	2	16
Total	27	25	52

EST, expressed sequence tag.

The modified 2D-EMSA approach allowed us to select, clone, and map DNA fragments (PTS) capable of specific interaction with nuclear proteins. The method is simple and allows identification of hundreds of protein binding fragments in one experiment. Moreover, the results of 2D-EMSA will strongly depend on the protein composition of the nuclear extracts used and thus might characterize the pool of nuclear proteins of given cells or tissues (i.e., tissue specificity of transcription factors binding). The 2D-EMSA technique is unbiased with respect to DNA binding factors and can therefore select PTS of yet unidentified proteins. At the same time, this approach has some limitations. First, it is based on in vitro DNA-protein interactions and, therefore, leaves aside possible in vivo modifications of this process due to chromatin structure variations, methylation of DNA, and the modification of proteins. In addition, it preferentially detects DNA sites with the highest affinity to target proteins, as low affinity sites will produce relatively small amounts of complexes. For these reasons, some existing in vivo contacts can be missed. However, due to its dynamic nature, chromatin structure can most probably create only transient barriers to the binding of transcription factors (discussed in Reference 27). Our data on the CTCF binding are in line with this hypothesis—all identified in vitro CTCF binding sites did bind CTCF in vivo as verified by ChIP. Interestingly, three of the CTCF binding sites coincided or overlapped with PTS 34, 37, and 43 found in this work. It means that at least some of the proteins that can bind specific DNA sequences in vitro will be able to bind them also within chromatin.

Only a rather small fraction (approximately 15%) of the identified PTS was found in promoter regions (5' regions) of genes. Most (70%) of the PTS were located within intronic sequences or in intergenic regions. One may speculate that these PTS belong to regulatory elements not structurally linked to genes, like enhancers, insulators, or locus control regions.

In conclusion, we propose an approach for identification and cloning of a large number of DNA binding

protein target sequences within long genomic regions in a single experiment. This approach can be a useful instrument of functional genomics, especially in combination with microarray technologies.

ACKNOWLEDGMENTS

The authors thank V.K. Potapov and N.V. Skaptsova for oligonucleotide synthesis and B.O. Glotov for critical reading of the manuscript. The authors would like to express special gratitude to A. Olsen (Lawrence Livermore National Laboratory) for providing cosmid DNAs. This work was supported by the Russian Foundation for Basic Research (project 05-04-48814), President of the Russian Federation grant NSH 2006.2003.4, and the Russian Academy of Sciences Physical and Chemical Biology Program. DNA sequencing was done at the Genome Center (www.genome-centre.narod.ru) under support of the Russian Foundation for Basic Research (grant no. 00-04-55000).

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

- Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, et al. 2001. Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature* 409:860-921.
- Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, et al. 2001. The sequence of the human genome. *Science* 291:1304-1351.
- Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M.C. Frith, N. Maeda, R. Oyama, T. Ravasi, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* 309:1559-1563.
- Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308:1149-1154.
- Kadonaga, J.T. 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116:247-257.
- Martone, R., G. Euskirchen, P. Bertone, S. Hartman, T.E. Royce, N.M. Luscombe, J.L. Rinn, F.K. Nelson, et al. 2003. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci. USA* 100:12247-12252.
- Euskirchen, G., T.E. Royce, P. Bertone, R. Martone, J.L. Rinn, F.K. Nelson, F. Sayward, N.M. Luscombe, et al. 2004. CREB binds to multiple loci on human chromosome 22. *Mol. Cell. Biol.* 24:3804-3814.
- Cawley, S., S. Bekiranov, H.H. Ng, P. Kapranov, E.A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116:499-509.
- van Steensel, B. and S. Henikoff. 2000. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* 18:424-428.
- van Steensel, B., J. Delrow, and H.J. Bussemaker. 2003. Genomewide analysis of *Drosophila* GAGA factor target genes reveals context-dependent DNA binding. *Proc. Natl. Acad. Sci. USA* 100:2580-2585.
- Orian, A., B. van Steensel, J. Delrow, H.J. Bussemaker, L. Li, T. Sawado, E. Williams, L.W. Loo, et al. 2003. Genomic binding by the *Drosophila* Myc, Max, Mad/Mnt transcription factor network. *Genes Dev.* 17:1101-1114.
- Bulyk, M.L. 2003. Computational prediction of transcription-factor binding site locations. *Genome Biol.* 5:201.
- Pavesi, G., G. Mauri, and G. Pesole. 2004. In silico representation and discovery of transcription factor binding sites. *Brief. Bioinform.* 5:217-236.
- Chernov, I.P., S.B. Akopov, L.G. Nikolaev, and E.D. Sverdlov. 2002. 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.* 84:590-600.
- Olsen, A.S., A. Georgescu, S. Johnson, and A.V. Carrano. 1996. Assembly of a 1-Mb restriction-mapped cosmid contig spanning the candidate region for Finnish congenital nephrosis (NPHS1) in 19q13.1. *Genomics* 34:223-225.
- Pruitt, K.D. and D.R. Maglott. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* 29:137-140.
- Sambrook, J. and D.W. Russell. 2001. *Molecular Cloning: A Laboratory Manual*. CSH Laboratory Press, Cold Spring Harbor, NY.
- Dignam, J.D., R.M. Lebovitz, and R.G. Roeder. 1983. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* 11:1475-1489.
- Nikolaev, L.G. 1996. Identification and isolation of proteins, recognizing the sequence of the human immunodeficiency virus (HIV-1) enhancer. *Mol. Biol. (Mosk.)* 30:714-720.
- Nikolaev, L.G., T. Tsevegijn, S.B. Akopov, L.K. Ashworth, and E.D. Sverdlov. 1996. Construction of a chromosome specific li-

- brary of human MARs and mapping of matrix attachment regions on human chromosome 19. *Nucleic Acids Res.* 24:1330-1336.
21. **Vetchinova, A.S., S.B. Akopov, I.P. Chernov, L.G. Nikolaev, and E.D. Sverdlov.** Two-dimensional EMSA: identification and mapping of transcription factor CTCF target sequences within an FXYD5-COX7A1 region of human chromosome 19. *Anal. Biochem.* (In press).
22. **Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
23. **Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler.** 2002. The human genome browser at UCSC. *Genome Res.* 12:996-1006.
24. **Garner, M.M. and A. Revzin.** 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res.* 9:3047-3060.
25. **Fried, M. and D.M. Crothers.** 1981. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res.* 9:6505-6525.
26. **Strauss, F. and A. Varshavsky.** 1984. A protein binds to a satellite DNA repeat at three specific sites that would be brought into mutual proximity by DNA folding in the nucleosome. *Cell* 37:889-901.
27. **Morse, R.H.** 2003. Getting into chromatin: how do transcription factors get past the histones? *Biochem. Cell Biol.* 81:101-112.

Received 14 February 2006; accepted 12 April 2006.

Address correspondence to Lev G. Nikolaev, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 16/10 Miklukho-Maklaya, 117997, Moscow, Russia. e-mail: lev@humgen.siocb.ras.ru

To purchase reprints of this article, contact: Reprints@BioTechniques.com