In vitro DNA-binding profile of transcription factors: methods and new insights

Jinke Wang, Jie Lu, Guangming Gu and Yingxun Liu
The State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, People’s Republic of China
(Correspondence should be addressed to J Wang; Email: wangjinke@seu.edu.cn)

Abstract
The DNA-binding specificity of transcription factors (TFs) has broad impacts on cell physiology, cell development and in evolution. However, the DNA-binding specificity of most known TFs still remains unknown. The specificity of a TF protein is determined by its relative affinity to all possible binding sites. In recent years, the development of several in vitro techniques permits high-throughput determination of relative binding affinity of a TF to all possible k bp-long DNA sequences, thus greatly promoting the characterization of DNA-binding specificity of many known TFs. All DNA sequences that can be bound by a TF with various binding affinities form their DNA-binding profile (DBP). The DBP is important to generate an accurate DNA-binding model, identify all DNA-binding sites and target genes of TFs in the whole genome, and build transcription regulatory network. This study reviewed these techniques, especially two master techniques: double-stranded DNA microarray and systematic evolution of ligands by exponential enrichment in combination with parallel DNA sequencing techniques (SELEX-seq).

Introduction
Transcription factors (TFs) are central to almost every fundamental cellular process (Latchman 2008, Ladunga 2010) and account for ~5–10% of genes in eukaryotes (Reece-Hoyes et al. 2005, Adryan & Teichmann 2006, Ho et al. 2006, Lee et al. 2007). In mammalian TFs, approximately more than 700 were identified to be DNA-binding TFs (Messina et al. 2004, Lee et al. 2007); they bind with the TF binding sites (TFBSs) in the genome and regulate the expression of their target genes. Differential gene expression is achieved in part by the interaction of these DNA-binding regulatory TFs with various TFBSs.

DNA-binding specificity of TFs plays essential roles in cell physiology, cell development and organism evolution. However, the DNA-binding specificity of most known TFs still remains unknown. The specificity of a TF protein is determined by its relative affinity to all possible binding sites. In recent years, the development of several in vitro techniques permits high-throughput determination of relative binding affinity of a TF to all possible k bp-long DNA sequences. These high-throughput techniques greatly promoted the characterization of DNA-binding specificity of many known TFs.

All DNA sequences that can be bound in vitro by a TF with various binding affinities form its in vitro DNA-binding profile (DBP). The in vitro DBP provides an in vitro sequence-recognition profile of a TF. The in vitro DBP is important in the generation of an accurate DNA-binding model (such as position weight matrix, PWM), identification of all DNA-binding sites and target genes of TFs in the whole genome, and construction of transcription regulatory network. The in vitro DBP also has promising applications in transcription therapy, a therapeutic strategy using TFs as targets for disease therapy (Redell & Tewary 2005, Frank 2009, Li & Sethi 2010). Therefore, the studies on in vitro DBP of TFs attract increasing attention in this field.

At present, the in vitro DBP is mainly generated by using two high-throughput methods, double-stranded DNA (dsDNA) microarray and systematic evolution of ligands by exponential enrichment in combination with parallel DNA sequencing techniques (SELEX-seq).

dsDNA microarray

The double stranded DNA (dsDNA) microarray was first reported by Bulyk et al. (1999), which was also named protein-binding microarray (Berger et al. 2008, Badis et al. 2009). In principle, dsDNA microarray contains tens of thousands of dsDNA molecules in a small area on a glass slide, which can be used to detect the binding of a TF protein to these dsDNA molecules in a high-throughput format (Berger & Bulyk 2009).
The whole process of the dsDNA microarray experiment is schematically described in Fig. 1. The prerequisite of dsDNA microarray studies is to prepare the high-density dsDNA microarrays. However, no commercial dsDNA microarray chips can be purchased at present. Therefore, the high-density dsDNA microarray is manufactured in three steps. The first step is to design DNA probes, especially complex probes (Berger et al. 2006, Mintseris & Eisen 2006, Philippakis et al. 2008). The second step is to authorize one of the biotechnical corporations, such as Affymetrix, Nimblegen, and Agilent, to manufacture high-density, single-stranded DNA (ssDNA) microarray. The third step is to convert ssDNA microarray.
into dsDNA microarray by using several approaches, such as constant primer elongation (Bulyk et al. 1999), hairpin primer elongation (Wang et al. 2003b), or hairpin formation (Warren et al. 2006).

The greatest advantage of dsDNA microarray is that the binding interaction of a certain TF with all possible sequence variants of a given length can be simultaneously detected in a single assay. In recent years, dsDNA microarrays containing all possible 8–10 bp DNA duplexes have already been used to study in vitro DBPs of TFs and low–molecular weight ligands. For example, dsDNA microarray with all possible 8 bp sequences was used to determine the binding preferences of the majority of mouse homeodomains (168) (Berger et al. 2008) and the sequence-recognition properties of an engineered small molecule, PA1 (polyamide engineered to target a specific DNA sequence) (Warren et al. 2006). DsDNA microarray with all possible 9 bp sequences was used to profile the DNA-binding spectrum of yeast TFs CBF-1 and CBF1/DREB1B and rice TF OsNAC6 (Kim et al. 2009). DsDNA microarray containing all possible 10 bp sequences was used to characterize the binding specificities of five TFs, from yeast (CBF1 and RAP1), worm (CEH-22), mouse (Zif268), and human (OCT1) (Philippakis et al. 2008). Similarly, dsDNA microarray with every possible 10 bp sequences was used to study sequence-recognition profile of TF AP2 (De Silva et al. 2008). These dsDNA microarray studies yielded a comprehensive binding profile across the entire sequence space of a binding site and collected high-content data of ‘specificity landscape,’ which simultaneously displays the affinity and specificity of a million-plus DNA sequences to DNA-binding molecules (Carlson et al. 2010).

At present, dsDNA microarray has already been applied to profile in vitro DBPs of many DNA-binding molecules. For example, the high-density dsDNA microarray has been successfully used to profile in vitro DBPs of TFs (Berger et al. 2006, Allevy et al. 2009, Zhu et al. 2009), and the studies on in vitro DBPs of TFs and other DNA-binding molecules have made tremendous progress (Warren et al. 2006, Maerkl & Quake 2007, Puckett et al. 2007, Keles et al. 2008, Bonham et al. 2009, Hauschild et al. 2009, Kim et al. 2009, Bolotin et al. 2010, Carlson et al. 2010). In 2009, the fine in vitro DBPs of 104 mouse TFs were successfully profiled by using dsDNA microarray technology (Badis et al. 2009). So far, in vitro DBPs of 406 TFs have already been profiled by using dsDNA microarray technology and stored in the UniPROBE database (Newburger & Bulyk 2009, Robasky & Bulyk 2011).

Many important parameters of DNA/TF interaction can be extracted from in vitro DBPs, among which the most important is DNA-binding specificity of TFs. DNA-binding specificity of TFs has broad impact on cell physiology, cell development and in evolution (Stormo 2000, Bulyk 2003). However, DNA binding specificity of most known TFs still remains unknown. For example, the DNA-binding specificity of only a small fraction of ~1400 human TFs is known. With the advent of dsDNA microarray technology, characterization of DNA-binding specificity of TFs rapidly progressed (Bulyk et al. 2001, Berger & Bulyk 2006, Warren et al. 2006). For example, the DNA-binding specificity of 104 known and predicted mouse TFs from 22 different DNA-binding domain (DBD) structural classes found in metazoan TFs were determined by using the universal dsDNA microarray technology (Badis et al. 2009). For the vast majority of these TFs, this is the first time it was possible to obtain their high-resolution binding specificity data. A merit of characterizing DNA-binding specificity of TFs with dsDNA microarray is that the binding specificity of all TFs, regardless of structural class or species of origin, can be effectively solved by this method, even if no initial information about the binding site is available. In addition, by giving complete information about all possible sequences, dsDNA microarray can provide a true picture of the sequence specificity.

Another important information that can be derived from dsDNA microarray experiments is the relative binding affinity of a TF to all possible sequences of a given length (Philippakis et al. 2008, Berger & Bulyk 2009). Determination of the relative binding affinity of different TFs to their various DNA-binding sites is fundamentally important for a comprehensive understanding of gene regulation. The whole profile of DNA-binding affinity of a TF protein to all possible DNA sequence is useful for identifying all functional TFBSs of a TF especially those TFBSs with low binding affinity. The in vivo studies revealed that both high- and low-affinity TFBSs had biological function, and the biologically important TFBSs are often not of maximal affinity (Jiang & Levine 1993, Tuupanen et al. 2009). A dsDNA microarray study demonstrated that the TFBSs in a wide affinity range were conserved and associated with regulatory function, and besides high-affinity TFBSs, numerous moderate- and low-affinity TFBSs were under negative selection in the mouse genome (Jaeger et al. 2010). The TFBSs with low and medium affinity are indispensable to the construction of the most accurate binding site models in bioinformatics (Roulet et al. 2002). The binding affinity data can also be used to evaluate the regulatory capability of a binding site to its target genes in vivo. In addition, a complete profile of DNA-binding affinity of a TF has promising biomedical applications. For example, DNA sequences with high affinity can be developed as drugs for transcription therapy, such as TF decoys (short duplex oligonucleotide containing DNA–binding site, which can be bound by a TF) (Mann & Dzau 2000, Tomita et al. 2007).

The dsDNA microarray also permits the discovery of subtle preferences of a TF to various DNA sequences, additivity between adjacent nucleotides or interdependencies among different positions in TFBSs, and functional polymorphism in TFBSs. A compact and universal dsDNA microarray can be used to rapidly determine the relative binding preferences of any TF from any organism (Berger et al. 2006). The complete reference tables of all possible binding sites on dsDNA microarray are important for comparing protein–binding preferences for various DNA sequences. For example, the universal dsDNA microarrays provided a complete reference...
table of the binding preference of a TF for each gapped and ungapped 8 bp sequence variant (Badis et al. 2009). The dsDNA microarray can also be used to find the interdependence between nucleotides in TFBSs. For example, dsDNA microarray found 19 clear cases of ‘position interdependence’ TFs, which exhibited strong interdependence among the nucleotide positions of their binding sites (Badis et al. 2009). DsDNA microarray study revealed that position interdependence occurred on a broad scale and had important implications (Badis et al. 2009). It was also found that interdependent nucleotide positions were not always adjacent to each other (Badis et al. 2009). For example, Myb exhibited strong interdependence at positions separated by one nucleotide, with preference for binding either AACCAGTCA or AACTGCCA (Badis et al. 2009). Position interdependencies frequently spanned more than just dinucleotides. For example, estrogen-related receptor α had a strong preference for binding either CAAGGTTCA or AGGGGTCA, but not CAGGGTCA or CCGGGTCA (Badis et al. 2009). DsDNA microarray study revealed that nucleotides of TFBSs exert interdependent effects on the binding affinities of TFs (Bulyk et al. 2002). The extensive existence of position interdependence in TFBSs suggests that it is important to consider the position interdependence in making accurate TFBS models because commonly used TFBS models assumed mononucleotide independence. The additivity between adjacent nucleotides in TFBSs was also found by dsDNA microarray (Benos et al. 2002, Bulyk et al. 2002); this suggests that the additive models are very useful for the identification of TFBSs in genomes (Benos et al. 2002).

The in vitro DBPs can also be used together with computational models to identify the polymorphisms that affect TF binding and disease predisposition (Tuapanen et al. 2009).

The in vitro DBP data produced by dsDNA microarray can be used to identify TFBSs and target genes of TFs in genome. For example, the in vitro DBPs were used together with computational models to identify target genes of mammalian TFs, such as Tfs (Hallikas & Taipale 2006, Hallikas et al. 2006). A dsDNA microarray was successfully used to identify the genome-wide binding sites and target genes of yeast TFs Abf1, Rap1, and Mig1 (Mukherjee et al. 2004). Based on the in vitro DBPs obtained through dsDNA microarray and SELEX-seq described below, the most accurate binding site models, such as PWM, can be built. These models are very helpful for identifying binding sites and target genes of TFs in genomes. By searching the sequences corresponding to these models in genomes with TFBS identification search engines, binding sites and target genes of TFs in the whole genome can be identified. At present, numerous such TFBS identification search engines have been developed, such as position-specific scoring matrix (Stormo 2000), dictionary model (Sabatti et al. 2005), artificial neural network (Workman & Stormo 2000), hidden Markov model (Marinescu et al. 2005, Drawid et al. 2009), Bayesian network (Chen et al. 2010), and P-Match (Chekmenev et al. 2005).

Accumulation of TFBSs and target genes is indispensable for construction of transcription regulatory network. Although in vivo approaches, such as ChIP-chip (ChIP coupled with DNA microarray chip) (Ren et al. 2000) and ChIP-seq (ChIP coupled with parallel DNA sequencing) (Robertson et al. 2007), have generated in vivo DBPs for many TFs, however, because ChIP-based methods identify TFBSs in a particular cell at the time point of formaldehyde cross-linking, different cell types may need to be cultured in an indeterminate number of different conditions (such as stimulation) to determine all the biologically relevant DNA-binding sites of a given TF. In contrast, dsDNA microarray is an in vitro technology that does not depend on any certain cells and cultivating conditions; therefore, it can exhaustively identify all possible DNA targets that can be bound by a TF, and thus making comprehensive in vitro DBPs of TFs. These in vitro DBPs can be used to identify all potential binding sites of a particular TF in genome. In addition, dsDNA microarray has the capability of identifying DNA-binding sites of all TFs from any species, regardless of the level to which its genome has been characterized; however, ChIP-based methods can only be used to investigate TFs from species which genomes have already been characterized.

A limitation of in vitro techniques like dsDNA microarray in identification of TFBSs is that they cannot determine if or when the identified binding sites are utilized in vivo. Therefore, determination of in vivo relevance of in vitro identified TFBSs remains a great challenge at present. However, some pioneering investigations have already been performed in this field. For example, an in vitro study of DNA-binding specificity of yeast TFs Abf1, Rap1, and Mig1 revealed that in addition to previously identified targets, Abf1, Rap1, and Mig1 bound to 107, 90, and 75 putative new target intergenic regions, respectively, and many of them were upstream of previously uncharacterized open reading frames (Mukherjee et al. 2004). Comparative sequence analysis indicated that many of these newly identified sites are highly conserved across five sequenced sensu stricto yeast species. Therefore, these newly identified sites should be functional in vivo binding sites that may be used in a condition-specific manner. This study reveals that dsDNA microarray can find a large number of binding sites that cannot be found by ChIP-based methods. The DNA-binding specificities of the ETS (E-26) family determined with in vitro dsDNA microarray can be confirmed by in vivo ChIP-seq technology (Wei et al. 2010). It was also found that even relatively small differences in in vitro binding specificity of a TF contributed to site selectivity in vivo (Wei et al. 2010). Recently, dsDNA microarray was used in an integrated approach to identify target genes of human hepatocyte nuclear factor 4α (Bolotin et al. 2010). Comparison of the dsDNA microarray data with ChIP-based data may provide insights into the usage of individual TF-binding sites in vivo (Mukherjee et al. 2004, Warren et al. 2006).
Our studies revealed that the predicted TFBSs and target genes with data from dsDNA microarray experiments provide a valuable blueprint to high-efficiency identification of functional DNA-binding sites and target genes of TFs. In recent years, our laboratory pursued studies of the in vitro DBPs based on the dsDNA microarray technique. We developed three methods for preparing unimolecular (hairpin) dsDNA microarray (Wang et al. 2003a,c, 2005) and used the prepared hairpin–dsDNA microarray to detect the binding of NF-κB to large numbers of DNA sequences (Wang et al. 2003b). We found that NF-κB bound to some mutated DNA sites with high affinity. We speculated that if these sites exist in human genome, they may be potential DNA-binding targets of NF-κB, and the genes neighboring these sites may be the target genes of NF-κB. Through a genome-wide search of the human genome, we found that these sites were indeed distributed in the human genome, we thus predicted these sites to be putative DNA targets of NF-κB and, correspondingly, predicted that the genes neighboring these sites were putative target genes of NF-κB. Through a literature search, we found that some predicted target genes have been identified as the functional NF-κB target genes by the previous experimental studies, such as NFKB2, NFKBIA, BCL2, and VEGFC. At the same time, we verified some selected typical disease-related genes, such as STAT1, MIA-53, HFE-625, and LTBP-1, which were not reported to be target genes of NF-κB, with ChIP-chip and gene expression profile.

It is necessary to point out that in identification of functional binding sites in the genome dependent on in vitro DBP data, the binding context of TFs, such as epigenetic modification of DNA, nucleosome position, chromosome structure, allosteric effects, and fluctuating cellular conditions, should be taken into consideration. A recent review pointed out that TFs’ selection of specific DNA response element (REs) in the presence of degenerate sequences cannot be viewed only from the standpoint of DNA sequence variability and TF-binding affinity under steady-state conditions (Pan et al. 2009). It was proposed that the fluctuating cellular conditions should be a key factor in the TFs’ selection of specific binding sites among the numerous similar binding sites present in the genome, because they lead to dynamic changes in the ensemble of protein (and DNA) conformational states via allosteric effects (Pan et al. 2009). This proposition is supported by the studies on regulatory diversity within the p53 transcriptional gene network selectivity in p53-dependent transcription (Espinosa 2008), which revealed that the p53-dependent transcriptional program is remarkably flexible, as it varies with the nature of p53-activating stimuli, the cell type, and the duration of the activation signal. These studies demonstrate that although the differential affinity of TF to various DNA-binding sites is a major factor in functional control, other factors are also important. Another recent review about the mechanisms of TF selectivity to TFBSs outlined that the recognition of selective binding site sequence and TF activation involve three major factors: the cellular network, protein and DNA as dynamic conformational ensembles, and the tight packing of multiple TFs and co-regulators on stretches of regulatory DNA (Pan et al. 2010). It was also revealed that the selective binding of p53 is achieved via a chromatin-dependent mechanism, but not through modulation of its binding affinity to certain REs (Millau et al. 2010). It was proposed that the formation of stress-specific p53 binding patterns is due to chromatin and chromatin remodeling, rather than the modulation of sequence-specific p53 binding affinity. It was revealed that several features, including but not limited to, the epigenetic landscape of the locus, p53 posttranslational modifications, the nature of the p53 RE, and p53-interacting partners, function in concert to determine the target promoter selectivity and the specificity of the p53 transcriptional response (Beckerman & Prives 2010).

It was demonstrated that DNA-induced allosteric effects on TFs play a critical role in TFs’ selection of specific binding sites. It was revealed that the TFBS can act as allosteric effectors to determine the TFs’ conformation. The selective gene transcription is not only mediated by TFs binding to TFBSs but TFs may also be modified in an allosteric manner by TFBSs themselves to generate the pattern of regulation that is appropriate to an individual gene (Lefstin & Yamamoto 1998). For example, the differential interaction of the DBD of estrogen receptor (ER) with the A2 and pS2 estrogen-responsive elements (EREs) brings about global changes in ER conformation (Wood et al. 1998). The conformational changes in ER induced by individual ERE sequences lead to the association of the receptor with different TFs and assist in the differential modulation of estrogen-responsive genes in target cells. The allosteric effects of DNA sites on the configuration of TF Pit-1 played essential role in control of differential expression of GH in different cells, somatotrope and lactotrope (Scully et al. 2000). It was also revealed that DNA-binding sites could allosterically modulate the transcriptional regulatory activity of glucocorticoid receptor (GR; Gronemeyer & Bourguet 2009). The transcriptional regulatory activity of the GR does not correlate with the affinity with which it binds to different GR-binding sites (GBSs), but rather with the sequence of the GBS, because the conformation of the GR’s domain relevant to transcriptional regulatory activity was determined by the sequence of the GBS to which GR was bound (Meijssing et al. 2009). GR-binding sequences, differing by as little as a single bp, differentially affect GR conformation and regulatory activity. Therefore, it was proposed that DNA is a sequence-specific allosteric ligand of GR that tailors the activity of the receptor toward specific target genes. The ability of specific DNA sequences to allosterically regulate the transcriptional regulatory activity of GR provides a mechanism to achieve gene-specific regulatory activity, by which GR finely tunes its target gene network (Meijssing et al. 2009)

In addition to its great values in basic biological research described above, the in vitro DBPs of TFs also have promising biomedical applications. For example, in vitro DBPs of TFs can be used to guide the design and selection of artificial TFs
(Gommans et al. 2005, 2007, Klug 2005), small molecules of TF mimics (Kwon et al. 2004, Xiao et al. 2007, Block et al. 2009, Rodriguez-Martinez et al. 2010, Kushal et al. 2011), and TF decoys (Penolazzi et al. 2007, Tomita et al. 2007), which can be developed as drugs for transcription therapy. The in vitro DBPs can also accelerate the creation of precision-tailored DNA therapeutics (Carlson et al. 2010).

**SELEX-seq**

SELEX, also referred to as in vitro selection or in vitro evolution, is an evolutionary process that allows the extraction, from an initially random pool of aptamers, of those molecules capable of binding to the target of interest (Stoltenburg et al. 2007). It was originally developed to screen oligonucleotides of either ssDNA or RNA that can specifically bind to DNA or RNA-binding proteins (Oliphant et al. 1989, Ellington & Szostak 1990, Tuerk & Gold 1990). At present, SELEX is used to develop high-affinity nucleic acid aptamers not only for a wide variety of pure molecules (such as protein) (Park et al. 2009) but also for complex systems such as live cells (Cell-SELEX) (Paul et al. 2009, Avcı-Adalı et al. 2010, Sefah et al. 2010). The screened DNA or RNA aptamers can be applied to basic research and disease diagnosis and treatment (Djordjevic 2007, Marton et al. 2010).

The procedures of the SELEX experiment include in vitro chemical synthesis of a single chain oligonucleotide library, mixing the oligonucleotide library with the target molecules such as RNA-binding protein to form complexes of a target molecule and oligonucleotide, isolation of bound oligonucleotides, and PCR amplification of enriched oligonucleotides to prepare a new library for the next round of the selection process. Through several rounds of repeated screening, the aptamers with high affinity and specificity can be obtained. As for the studies of TFs, the most critical step of SELEX is the isolation of DNA–protein complexes from free DNA. The methods used in this step include gel retardation assay (Tsai & Reed 1998, Tantin et al. 2008), affinity chromatography (Liu & Stormo 2005), filter-binding assay (Alex et al. 1992, Ferraris et al. 2010), and other approaches (Xue 2005, Gopinath 2007, Kim et al. 2010).

In previous studies, SELEX was frequently used for the purpose of characterizing the binding specificity of TFs. In such an experiment, SELEX yielded a library of dsDNA molecules binding to TF proteins, which was then used to generate a computational model, e.g. a position-specific scoring or weight matrix that served to predict binding sites of TFs in regulatory DNA sequences. SELEX has already been used to determine the binding specificity of many TFs, such as Sox2 (Maruyama et al. 2005), Oct4 (Tantin et al. 2008), Nanog (Mitsu et al. 2003), c-Myc (Papoulas et al. 1992), AP2, bHLH, NAC, MYB (Xue 2005), NF-κB (Kunsch et al. 1992), and GKL (Shields & Yang 1998). However, these studies used low-throughput cloning DNA sequencing technology; therefore, only limited numbers of DNA molecules (rarely exceeding 100 sequences) were sequenced. These limited DNA sequences produce low-resolution in vitro DBPs of TFs. To increase the sequencing throughput of the traditional SELEX method, the concatamerization step of serial analysis of gene expression (SAGE) was incorporated with SELEX (Roulet et al. 2002). In SELEX-SAGE, the SELEX-screened dsDNA fragments were first digested with a restriction endonuclease BglIII, then the digested dsDNA fragments with sticky ends were concatemerized and cloned, finally the cloned DNA was sequenced by cloning DNA sequencing technology. SELEX-SAGE can generate large numbers (> 1000) of ligands in a single assay; therefore, this method was called high-throughput SELEX (HTPSELEX), which can originate large volumes of data (Jagannathan et al. 2006). However, SELEX-SAGE was still a method dependent on cloning DNA sequencing technology. Its DNA sequencing throughput is not high enough for generating comprehensive in vitro DBPs of TFs.

In the last two years, a new high-throughput technique named SELEX-seq was developed, which combined the conventional SELEX technique with massively parallel DNA sequencing techniques, such as Illumina SOLEXA. Zykovich et al. (2009) first reported SELEX-seq technique and used it to investigate DNA-binding motif and relative DNA-binding affinity of the TFs Zif268 and Aart; they named the technique bind-n-seq. Jolma et al. (2010) developed an improved SELEX-seq technique and applied it to building in vitro DBPs of 19 TFs. They validated the method by determining binding specificities of TFs belonging to 14 different classes and confirming the specificities for NFATC1 and RFX3 by using ChIP-seq (Jolma et al. 2010). These successful proof-in-principle studies demonstrated the great value of this technique in profiling in vitro DNA-binding spectrums of TFs. This technique is powerful but cost-effective: it can generate over 5 million sequences in a single assay with a cost as low as 1500 dollars. This technique thus provides a true high-throughput method for building comprehensive in vitro DBPs of TFs or other DNA-binding molecules. The whole process of this technique is schematically described in Fig. 2.

Most DNA/TF-binding information produced by dsDNA microarray can also be obtained by SELEX-seq. For example, SELEX-seq can be used to characterize DNA-binding specificity of TFs in ultrahigh resolution and determine the relative binding affinities of TFs to millions of DNA sequences. A recent SELEX-seq study revealed that the enrichment of a sequence in SELEX is proportional to the relative affinity of a TF protein to it (Zykovich et al. 2009). SELEX-seq data can also be used to determine DNA-binding motifs of TFs (Zykovich et al. 2009). For example, the binding motifs of two well-characterized zinc-finger proteins (Zif268 and Aart) were found with SELEX-seq data using the motif-finding program MEME, and the found motifs were similar to those previously derived from the cyclic amplification and selection of targets (Zykovich et al. 2009). The SELEX-seq-generated binding profile of mouse TF
and computational procedure to infer parameters of TF-DNA elicitation (MEME) (Bailey & Elkan 1994). A more detailed experimental some typical algorithms for this purpose, such as multiple EM for motif finding motifs or position weight matrix (PWM) models with reads data are then used to perform subsequent bioinformatics analysis, filtered reads are sorted according to bar code sequence. The qualified regions. If multiple DNA samples are simultaneously sequenced, the letters allowed, valid bar code, and constant regions and unique random parallel DNA sequencing techniques, such as Illumina SOLEXA. The procedures of SELEX-seq include three steps. The first step is to design and synthesize an ssDNA library and then convert ssDNA into dsDNA using primer extension. The ssDNA library contained all possible oligonucleotides in random sequence of length k. The length of k can be determined according to the nucleotide numbers of DNA-binding consensus bound by a TF protein of interest. The second step is to bind TF proteins to randomized dsDNA library and isolate the protein-bound dsDNAs from binding reaction using gel mobility shift assay (also called electrophoresis mobility shift assay, EMSA) (Zykovich et al. 2009), affinity chromatography (Zykovich et al. 2009), or TF protein-coupled microwell plate (Jolma et al. 2010). The isolated dsDNAs are amplified by PCR to prepare a new library for the next round of the selection process. Through several rounds of repeated screening, the TF protein-bound dsDNAs were enriched. To find all sequences that can be bound by a TF protein with various binding affinities, especially sequences with low affinity, the isolated dsDNAs from each round of selection can be separately collected for sequencing (Zykovich et al. 2009). The third step of SELEX-seq is to sequence the bound dsDNAs with massively parallel DNA sequencing techniques, such as Illumina SOLLEXA. The sequencing reads are filtered with filters including only A, C, G, and T letters allowed, valid bar code, and constant regions and unique random regions. If multiple DNA samples are simultaneously sequenced, the filtered reads are sorted according to bar code sequence. The qualified reads data are then used to perform subsequent bioinformatics analysis, including finding motifs or position weight matrix (PWM) models with some typical algorithms for this purpose, such as multiple EM for motif elicitation (MEME) (Bailey & Elkan 1994). A more detailed experimental and computational procedure to infer parameters of TF-DNA interaction from SELEX experiments was described by some studies (djordjevic & sengupta 2006). Full colour version of this figure available via http://dx.doi.org/10.1530/JME-11-0010.

domesoderm (EOMES) was very similar to the dsDNA microarray-derived profile (jolma et al. 2010). The SELEX-seq-generated binding profiles of 18 TFs were generally in good agreement with the existing data; however, some notable differences were seen, including TF POU2F2 and RFX3 (jolma et al. 2010). It was also validated that the binding profiles generated using SELEX-seq method were relevant for the in vivo situation. For example, the enriched sequence motifs of TF RFX3 and NFATC1 in K562 and Jurkat cells from the ChiP-seq peaks using the MEME algorithm revealed a profile that was very similar to that generated using SELEX-seq method (jolma et al. 2010). The broad utility of the SELEX-seq method was highlighted by the profiles for 14 TFs, which belong to 23 major DBD families occupying most major branches of TF-binding specificities (jolma et al. 2010).

SELEX-seq has several significant advantages over dsDNA microarray (Table 1). SELEX-seq combines SELEX with massive parallel DNA sequencing technique, thus possessing advantages of both techniques. SELEX-seq needs no complex design and on-chip synthesis of oligonucleotides and provides an easy cost-effective alternative approach to research in vitro DBPs of TFs beside dsDNA microarray. Along with rapid development of DNA sequencing techniques, most advanced DNA sequencing techniques become affordable and reachable to general researchers, including equipment and commercialized DNA sequencing services. Moreover, SELEX-seq has ultrahigh throughput over dsDNA microarray by using bar code technique (Fig. 3). For example, up to 28 samples were simultaneously analyzed in one sequencing by using 3 nt bar-coded oligonucleotides (zykovich et al. 2009). Due to its significant advantages over dsDNA microarray technology, SELEX-seq may replace dsDNA microarray technology as the master technique in future in vitro DBP studies.

It was noted that in current dsDNA microarray and SELEX-seq studies, many experiments targeted to mammalian TFs employed the purified recombinant TF proteins expressed in bacteria, such as Escherichia coli (mukherjee et al. 2004, Berger et al. 2006, Kim et al. 2009, Zykovich et al. 2009). These recombinant proteins prepared from prokaryote expression may exclude possible post-translational modifications of mammalian TF proteins and omit the potential effects of these modifications on the DNA-binding activity of TFs. However, like many other proteins, TFs are post-translationally modified under different conditions and by different modifiers, such as phosphorylation, hydroxylation, acetylation, ubiquitination, and sumoylation (grove & Wallhout 2008). The post-translational modifications can affect the regulatory activity of a TF, as well as its localization or stability. For instance, the post-translational modifications play essential roles in regulating the activity of EST TF superfamily (tootle & rebay 2005). The phosphorylation was reported to inhibit the DNA binding of Ets1, Er81, and Ern of ETS TF superfamily, but enhance the DNA binding of Sap1, Elk-1, and Elf-1. The acetylation of ERz at
The relative binding affinity is directly reported by fluorescence signal. The affinity to all sequences, including low-affinity sequences, can be found. No PCR is needed in detecting process. The high-density dsDNA microarray may be helpful to find the fine specificity difference between sequences with single nucleotide differences.

Conserved lysine residues resulted in enhanced DNA-binding activity (Kim et al. 2006). The post-translational modification of the DNA-binding subunits by phosphorylation, acetylation, and ubiquitination regulated the activity of NF-kB (Mattioli et al. 2006, Geng et al. 2009, Moreno et al. 2010). To overcome the limitations of these TF proteins expressed in prokaryotes, the newest in vitro DBP study with SELEX-seq used purified TF proteins expressed in eukaryotes, such as mammalian cells (Jolma et al. 2010). The TF proteins expressed in mammalian cells can thus be used to characterize DNA-binding preferences of proteins requiring post-translational modifications. Another important problem regarding TF protein is that some experiments were performed only with the purified recombinant DBDs, not full length, of mammalian TF proteins expressed in bacteria (Zykovich et al. 2009) or mammalian cells (Jolma et al. 2010). These experiments with DBDs may also have serious limitations, because some regions of a full-length TF protein contribute to the dimerization of members of some TF family or superfamily, and this kind of dimerization to form homodimers or heterodimers is critical to the DNA-binding activity of TFs. For example, the TF E2F often binds its DNA-binding sites with low affinity, however, when dimerized with DP, its DNA-binding activity is greatly enhanced (Tao et al. 1997). Therefore, the full-length TF proteins expressed in mammalian cells should be used in future in vitro DBP studies. In addition, it is important to use TF protein samples combining two different members of a TF family or superfamily in future studies, in order to find a more complete and accurate DNA-binding spectrum of TFs. However, in most current in vitro DBP studies with dsDNA microarray and SELEX-seq, only a single TF protein was used. The binding affinity experiments should be carried out in the presence of other protein factors.

It is worthy to note that besides methods described above, many other high-throughput methods were also developed for in vitro quantifying DNA-binding specificities of TFs in recent years. These methods include oligonucleotide mass tags and mass spectroscopy (Zhang et al. 2007), DIP-chip (Liu et al. 2005), microarray evaluation of genomic aptamers by shift (MEGAshift; Tantin et al. 2008, Ferraris et al. 2010), mechanically induced trapping of molecular interactions (MITOMI; Fordyce et al. 2010), microwell-based competition assay (Wei et al. 2010), transcriptional regulatory sequences (TRSs) interrogating with a flow cytometry and deep sequencing (TRS-FY-DS; Kinney et al. 2010), synthetic saturation mutagenesis (Patwardhan et al. 2009), and bacterial one-hybrid selections (Meng et al. 2005).

Table 1 Advantages and limitations of dsDNA microarray and SELEX-seq techniques

<table>
<thead>
<tr>
<th></th>
<th>dsDNA microarray</th>
<th>SELEX-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput</td>
<td>Low throughput: one chip, one protein. Chip cannot be</td>
<td>High-throughput: the use of bar coding techniques in MPS (Fig. 3) allows</td>
</tr>
<tr>
<td></td>
<td>repeatedly used to detect several proteins</td>
<td>multiple protein samples to be analyzed simultaneously</td>
</tr>
<tr>
<td>Capacity</td>
<td>Limited by the number of features that can be placed</td>
<td>Not limited to 10 bp binding sites. Proteins with long DNA-binding to be</td>
</tr>
<tr>
<td></td>
<td>on the array. The 10^5 feature (all possible 10 bp</td>
<td>investigated. For example, in vitro interactions of two TFs (Zil268 and</td>
</tr>
<tr>
<td></td>
<td>sequences) is the limit for array technology today,</td>
<td>AaT) to all possible 21 bp sequences were detected with this technique.</td>
</tr>
<tr>
<td></td>
<td>but many TFs have binding sites (longer than 10 bp.</td>
<td>In traditional SELEX, the length of random sequences can be up to 35 bp</td>
</tr>
<tr>
<td></td>
<td>About 10^5 features (all possible 15 bp sequences)</td>
<td>Techniques for parallel DNA sequencing are reachable for many biotechnical</td>
</tr>
<tr>
<td></td>
<td>are far beyond the current capacity of DNA microarray</td>
<td>corporations and some researchers. The general researchers can get</td>
</tr>
<tr>
<td></td>
<td>technique</td>
<td>cost-effective and commercial service of DNA sequencing from many</td>
</tr>
<tr>
<td></td>
<td></td>
<td>biotechnical corporations</td>
</tr>
<tr>
<td>Equipment</td>
<td>No commercialized machines. Complex machines for</td>
<td>Machines for parallel DNA sequencing can be purchased and owned by many</td>
</tr>
<tr>
<td></td>
<td>manufacturing high-density DNA microarray are</td>
<td>biotechnical corporations and some researchers. The SELEX processes do not</td>
</tr>
<tr>
<td></td>
<td>possessed by a few corporations, such as Affymetrix,</td>
<td>need any costly and complex machines</td>
</tr>
<tr>
<td></td>
<td>Nimblegen, and Agilent. In addition, detection process with dsDNA microarray need costly</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probe design</td>
<td>Complex and tedious design of millions of probes</td>
<td>Easy design and synthesis of random sequence DNA library.</td>
</tr>
<tr>
<td></td>
<td>synthesized on-chip. With high cost of probe design</td>
<td>No cost of probe design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cost of parallel DNA sequencing is affordable to general researchers.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>This technique is cost effective, which can generate over 5 millions of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sequences in a single assay with the cost as low as 1500 dollars</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>Cost of manufacturing high-density DNA microarray is</td>
<td>The relative binding affinity is calculated with an abundance of sequences</td>
</tr>
<tr>
<td></td>
<td>still prohibitive to general researchers. To</td>
<td>in total sequenced reads. Some low-affinity sequences may be omitted. PCR</td>
</tr>
<tr>
<td></td>
<td>customize a DNA microarray with 1 million features</td>
<td>is used in SELEX and sequencing which may introduce bias to some sequences</td>
</tr>
<tr>
<td></td>
<td>can cost over 6000 dollars in China</td>
<td>or artifacts. SELEX-seq may have difficulty finding the fine specificity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>different between sequences with single nucleotide differences</td>
</tr>
<tr>
<td>Determination of</td>
<td>The relative binding affinity is directly reported</td>
<td></td>
</tr>
<tr>
<td>relative</td>
<td>by fluorescence signal. The affinity to all sequences,</td>
<td></td>
</tr>
<tr>
<td>relative</td>
<td>including low-affinity sequences, can be found. No PCR is needed in detecting process. The</td>
<td></td>
</tr>
<tr>
<td>specificity</td>
<td>high-density dsDNA microarray may be helpful to find</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the fine specificity difference between sequences with single nucleotide differences</td>
<td></td>
</tr>
</tbody>
</table>

Journal of Endocrinology (2011) 210, 15–27 www.endocrinology-journals.org
measure their relative and absolute affinities (Maerkl & Quake 2007). Discover both high- and low-affinity target sequences and is a microfluidics-based approach that can simultaneously detect multiple specific DNA-binding interactions and determine the sequence specificities of DNA-binding proteins (Hallikas & Taipale 2006). This method is suitable for high-throughput screening to identify proteins or small molecules that modulate protein–DNA binding interactions. However, this method requires prior knowledge of one high-affinity binding site for the protein of interest. The methods of TRS-FY-DS and synthetic saturation mutagenesis combined the traditional reporter construct technique that is used to detect the transcription activity of a DNA sequence in cells with new parallel DNA sequencing techniques (Patwardhan et al. 2009, Kinney et al. 2010). The advantage of these two methods is to detect the DNA-binding specificity of a TF to various DNA sequences at the level of transcription activation in true intracellular environment. Therefore, they can reveal the transcription activation functions of different DNA sequences under the complex interaction with TFs and their cofactors.

Summary

Regulatory TFs are a class of sequence-specific DNA-binding proteins that play essential roles in the regulation of gene expression. As more and more TFs of many organisms are identified, the identification of all their functional TFBSs and target genes in the whole genome and construction of transcription regulatory network controlled by them are increasingly of importance. Therefore, two kinds of research become more and more intense in the field of TF-related studies. One is the characterization of in real DNA-binding specificity of TFs via dsDNA microarray and SELEX-seq techniques. The latter attracts increasing attentions due to its comprehensiveness in characterizing DNA-binding specificity and quantifying relative or absolute DNA-binding affinity of TFs. The exhaustive data obtained through these in vitro studies play critical roles in decoding gene regulatory codes and deep understanding complex transcriptional regulatory networks and the mechanism through which TFs control the fine temporal and special expressions of genes. At present, many in vitro methods chaired by dsDNA microarray and SELEX-seq have been developed, and large amounts of DNA-binding data are rapidly accumulated along with practical applications of these methods. At the same time, the corresponding bioinformatics are also promptly developed.
of these experimental and computational approaches will greatly improve future studies on TFs, which have already become the promising leads of genomics, system biology, regulatory biology, and transcription therapy biomedicine.

**Declaration of interest**

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

**Funding**

This study was funded by the National Natural Science Foundation of China (60871014) and Supporting Program of New Century Excellent Talents of Ministry of Education (NCET-08-0110).

**References**


Ladunga I 2010 Computational Biology of Transcription Factor Binding. New York: Springer.


www.endocrinology-journals.org


Pan YP, Tsai CJ, Ma BY & Nussinov R 2009 How do transcription factors select specific binding sites in the genome? Nature Structural and Molecular Biology 16 1118–1120. (doi:10.1038/nsmb1109-1118)


R110. (doi:10.1038/rnb1068)


U885–U837. (doi:10.1038/rnb1068)


Evaluating the binding affinities of NF-kappaB p50 homodimer to the wild-type and single-nucleotide mutant Ig-kappaB sites by the unimolecular dsDNA microarray. *Analytical Biochemistry* **316** 192–201. (doi:10.1016/S0003-2697(03)00049-6)


Received in final form 28 February 2011
Accepted 9 March 2011
Made available online as an Accepted Preprint
9 March 2011