Published online 20 October 2009 Nucleic Acids Research, 2009, Vol. 37, No. 22 **e151**
doi:10.1093/nar/gkp802

Bind-n-Seq: high-throughput analysis of *in vitro* protein–DNA interactions using massively parallel sequencing

Artem Zykovich, Ian Korf* and David J. Segal*

Genome Center, University of California, Davis, CA 95616, USA

Received June 16, 2009; Revised August 3, 2009; Accepted September 14, 2009

ABSTRACT

Transcription factor–DNA interactions are some of the most important processes in biology because they directly control hereditary information. The targets of most transcription factor are unknown. In this report, we introduce Bind-n-Seq, a new high-throughput method for analyzing protein–DNA interactions *in vitro*, with several advantages over current methods. The procedure has three steps (i) binding proteins to randomized oligonucleotide DNA targets, (ii) sequencing the bound oligonucleotide with massively parallel technology and (iii) finding motifs among the sequences. *De novo* binding motifs determined by this method for the DNA-binding domains of two well-characterized zinc-finger proteins were similar to those described previously. Furthermore, calculations of the relative affinity of the proteins for specific DNA sequences correlated significantly with previous studies ($R^2 = 0.9$). These results present Bind-n-Seq as a highly rapid and parallel method for determining *in vitro* binding sites and relative affinities.

INTRODUCTION

In the human genome, there are more than 700 predicted C2H2 zinc-finger transcription factors (1), but only ~10% of these have known binding motifs (2). There are a few extant technologies for identifying protein–DNA interactions. ChIP-chip and ChIP-seq can find the *in vivo* genomic binding sites (3). Although highly informative in many cases, these methods are limited by the availability of highly specific antibodies (e.g. many of the C2H2 zinc-finger proteins are related and there may be cross-reactivity), as well as the number of transcription factors and accessible binding sites are available in any particular cell type under any particular environmental condition. A high-throughput protein-binding microarray (PBM) approach has been described in which proteins bind double-stranded oligonucleotides on a microarray (4). While powerful, this technology is limited by the number of features that can be placed on the array. The complete catalog of 10-mers (10^6 features) is approximately the limit for array technology today. However, many DNA-binding proteins have recognition sites longer than 10 bp due to multimerization, heterologous binding partners, or, in the case of some zinc-finger proteins, multiple potential binding sites per protein (5). While arrays will continue to increase in the number of features, a 15-mer (~1 billion features) is far beyond the current capacity. Yeast and bacterial two- and one-hybrid systems have also been described (6–8). These systems have the advantage of *in vivo* selection, with stringency that can be experimentally manipulated. In principle, libraries of target sites up to 15 bp long (10^9 sequences) could be surveyed; however, usage of libraries larger than 10^7 has not been reported. In addition, sequencing throughput is typically low (20–70 sequences). A higher throughput variation required an even smaller library size (9). Cyclical Amplification and Selection of Targets (CAST) and Systematic Evolution of Ligands by Exponential Enrichment (SELEX) (10–12) can be used to find the preferred binding sequences *in vitro*. However, CAST generally involves several rounds of amplification and purification for each protein and is therefore labor-intensive. In addition, the CAST process tends to select for a few high-affinity binding sites. The most accurate binding site models require many low- to medium-affinity sequences (13). Serial Analysis of Gene Expression (SAGE) can be applied to reduce the cloning burden and the cost required to obtain large numbers of sequences (13).

In this study, we introduce Bind-n-Seq, a new high-throughput method for *in vitro* analysis of protein–DNA interactions that takes advantage of next-generation short-read sequencing technology. The concept is simple: proteins are incubated with random oligonucleotides, the

*To whom correspondence should be addressed. Tel: +1 530 754 9134; Fax: +1 530 754 9658; Email: djsegal@ucdavis.edu
Correspondence may also be addressed to Ian Korf. Tel: +1 530 754 4989; Fax: +1 530 754 9658; Email: ifkorf@ucdavis.edu

bound oligonucleotides are sequenced and motifs are extracted from the sequences. Unlike CAST, multiple rounds of binding and amplification are not required. In addition, many binding reactions can be assayed in parallel with bar coded oligonucleotides. Unlike microarrays, Bind-n-Seq is not limited to 10-bp binding sites. In this study, we use a 21-bp binding region. We show the utility of Bind-n-Seq by analyzing the DNA-binding domains (DBDs) of two well-characterized C2H2 zinc-finger proteins, Zif268 and Aart. We found that we were able to obtain ~100 000 reads per sample while simultaneously analyzing 28 samples, and could identify the canonical binding site for each protein with as few as 10 000 reads. The fold of enrichment was found to be proportional to the relative affinity of the protein for a previously described set of sequences. These results show that Bind-n-Seq is a powerful and cost-effective method for studying protein–DNA interactions.

MATERIALS AND METHODS

Protein purification

The coding regions for the DBDs were subcloned into the BamHI/HindIII sites of pMAL-c2X (New England Biolabs). This vector enables bacterial expression of the proteins as fusions with the maltose binding protein (MBP). The N-terminal MBP domain improves the solubility of the expressed proteins and allows for rapid one-step purification over amylose resin (New England Biolabs). Proteins were over-expressed in BL21 (DE3) *Escherichia coli* (Invitrogen) after isopropyl β-D-1-thiogalactopyranoside (IPTG) induction (0.3 mM) at an OD_{600} of 0.7–1.0 for 2 h at 37°C. Cells were pelleted and resuspended in 5 ml of Zinc Buffer A [ZBA; 10 mM Tris (pH 7.5), 90 mM KCl, 1 mM $MgCl_2$, 90 μM $ZnCl_2$, 5 mM DTT] and 50 μg/μl RNAse A. Following sonication, proteins in clarified lysates were applied to an amylose resin column, washed with ZBA and eluted with 3 ml of ZBA + 10 mM maltose. Because we intended to again use amylose resin to capture the proteins in our binding reactions, free maltose needed to be removed. Samples in Slide-A-Lyzer cassettes (Pierce) were dialyzed in 0.5 l of ZBA + 5 mM DTT for 2 h at 23°C with stirring, followed by fresh buffer for 16 h at 4°C. Concentration and purity was assessed by UV absorption (NanoDrop) and Coomassi-stained polyacrylamide gel electrophoresis with sodium dodecyl sulfate (SDS–PAGE) with bovine serum albumen (BSA) standards. Purified protein was stored in ZBA + 30% glycerol solution at −20°C until use.

Binding reactions

The 93-mer oligonucleotides containing Illumina primer-binding sites, a 2-nt AA leader, a 3-nt bar code and 21-nt random region, were synthesized (Sigma). Templates were made double-stranded by primer extension in a 25 μl reaction containing 0.88 μM of one template, 88 μM reverse primer, 1×TaqPro Complete (2.0 mM Mg^{2+}, Denville) at 95°C for 2 min, 63°C for 1 min, 72°C for 4 min, then 4°C. To initiate binding reactions, an additional 25 μl volume was added to achieve a ‘total’ concentration of 0.12 mg/ml herring sperm DNA (Sigma), 100 μM $ZnCl_2$, 5 mM DTT, 1% BSA and the indicated concentrations of KCl and purified binding protein (see Results section). Reactions were incubated for 2 h at 23°C.

Resin-based enrichment

To prepare the amylose resin, 50 μl of packed resin was washed twice (by pelleting and resuspension) with 500 μl of water, then twice with the appropriate wash buffer [10 mM Tris (pH 8.5), 100 μM $ZnCl_2$, 1 mM $MgCl_2$, 5 mM DTT and the indicated concentration of KCl]. A 50 μl binding reaction was added to the 50 μl of prepared resin, mixed and incubated for 30 min at 23°C with gentle mixing for every 10 min. The mixture was then washed three times using the appropriate wash buffer for 10-min incubations. Protein–DNA complexes were eluted by a 10-min incubation in 50 μl of elution buffer [10 mM Tris (pH 8.5), 10 mM maltose], pelleting the resin, then carefully transporting the supernatant to a new tube. An addition round of pelleting and transport resulted in cleaner samples with less resin contamination.

In the second run of Bind-n-Seq (see Results section), four additional parameters were examined. Long wash (lw): washed six times using 1 ml of buffer with 10-min incubations. Extra round of selection (+r): output DNA from the first round was amplified by polymerase chain reaction (PCR) using Illumina primers Pr4-f and Pr3-r (95°C for 15 s, 63°C for 15 s and 72°C for 30 s). The reaction was monitored periodically until 10 μl produced a visible band on agarose gel (~11 cycles). This material was used as input for a subsequent round of enrichment. Ficoll (f): purified proteins were stored in 30% ficoll as an alternative to glycerol (14). Salt concentration: 200 mM of KCl was examined because insufficient sequences were obtained when 500 mM was used in the first run.

Gel shift-based enrichment

A 50 μl binding reaction was loaded on a 5% tris-borate-ethylenediaminetetraacetic acid (TBE) [9 mM Tris (pH 8.3), 9 mM boric acid, 0.2 mM ethylenediaminetetraacetic acid (EDTA)] polyacrylamide gel (Bio-Rad). Control binding reactions (Zif268 protein with 6-carboxy-fluorescein (FAM)-labeled Zif268 oligonucleotide targets) were also loaded to indicate the mobility of the protein–DNA complexes. The gel was run for 30 min at 120 V in 0.5× TBE. Gels were imaged without drying on a Storm 860 (Molecular Dynamics). Bands containing protein–DNA complexes were excised and the DNA eluted overnight at 23°C in 400 μl of 0.5 M ammonia acetate, 10 mM magnesium acetate and 0.1% SDS. The liquid was transferred to a new tube and the DNA precipitated by the addition of 800 μl 100% ethanol at −80°C for 20 min, and centrifugation for 20 min at 4°C. The pellet was washed with 500 μl of ethanol, pelleted, dried and resuspended in 50 μl of elution buffer.

Output DNA normalization and quantification

Real-time PCR (RT-PCR) was used to determine the number of cycles required to amplify all output samples to equal concentrations that would be sufficient for sequencing. Samples were analyzed on a DNA Engine Opticon 2 System (MJ Research) using 20 μl reactions containing 1 μl output DNA, 0.5 μM each of primers Pr4-f and Pr3-r and the SYBR Green PCR Master Kit (Applied BioSystems) (50°C for 2 min, read, 90°C for 10 min, 40 cycles of 95°C for 15 s, 63°C for 1 min, read every cycle, melting curve 40–95°C, hold for 3 s and read every 0.7°C). A standard curve was generated using template DNA at 300, 30, 3 and 0.3 nM.

Bioinformatics

All Perl scripts and raw sequence files are freely available at http://korflab.ucdavis.edu/BnS. Sequencing reads were filtered and sorted with htsAnalysis.pl. Filters included: only A, C, G, T letters allowed; valid bar code and constant regions and unique random regions.

An overview of the bioinformatics processing of motif finding is shown in Supplementary Figure S1. The pipeline included several Perl scripts (names ending in .pl) and multiple em for motif elicitation (MEME) (15). Motif finding was performed with MEME 3.5.7 with parameters: -dna -revcomp -nmotifs 5. To produce intermediate motifs, we added -minw 9 -maxw 10 for Zif268 and -minw 10 -maxw 15 for Aart. To produce the final motifs we set -mod oops or -mod zoops and $P < 1e^{-100}$. Reads were scored for the presence of a motif using motif_mapper.pl. The threshold score corresponded to a P-value of 0.001 in a simulation of 1 million random 21-mers.

The relative affinities for 15 sequences (10-mers) were calculated with seqMapper.pl. Fold enrichment score was calculated by as the ratio of reads that contained a particular 10-mer sequence in protein-containing reads to no-protein control reads.

Sequencing

Sequencing was performed with the Illumina Genome Analyzer. Both runs employed version 1 sequencing reagents. The first run was on a GA I, pipeline 0.3 and the second run was on a GA II, pipeline 1.0.

RESULTS

Experimental design

An overview of Bind-n-Seq is shown in Figure 1. In each binding reaction, a purified protein was mixed with double-stranded Bind-n-Seq target oligonucleotides (DNAs), which contained a bar code, a binding region consisting of 21 random bp and flanking primer-binding sites. A randomized region of 21 bp contains 4.4×10^{12} members (4^{21}). Each binding reaction contained a 3-fold over-representation of each possible 21-mer, corresponding to 22 pmol or 440 ng of single-stranded 93-mer oligonucleotides (see 'Materials and Methods' section). Hence, each binding reaction contained more than 10^7 copies of each possible 10-mer or more than 10^2 copies of each possible 18-mer. Double-stranded DNAs were created by primer extension. After incubation, the protein–DNA complexes were separated from unbound and low-affinity DNAs and then the bound DNAs were released and quantified. The DNAs from several different experiments were combined in approximately equal concentrations and massively parallel sequencing was performed using an Illumina Genome Analyzer. After sequencing, bar codes were used to sort the DNAs into their respective experiments. Common sequence patterns (motifs) were identified using a combination of MEME (15) and custom Perl scripts.

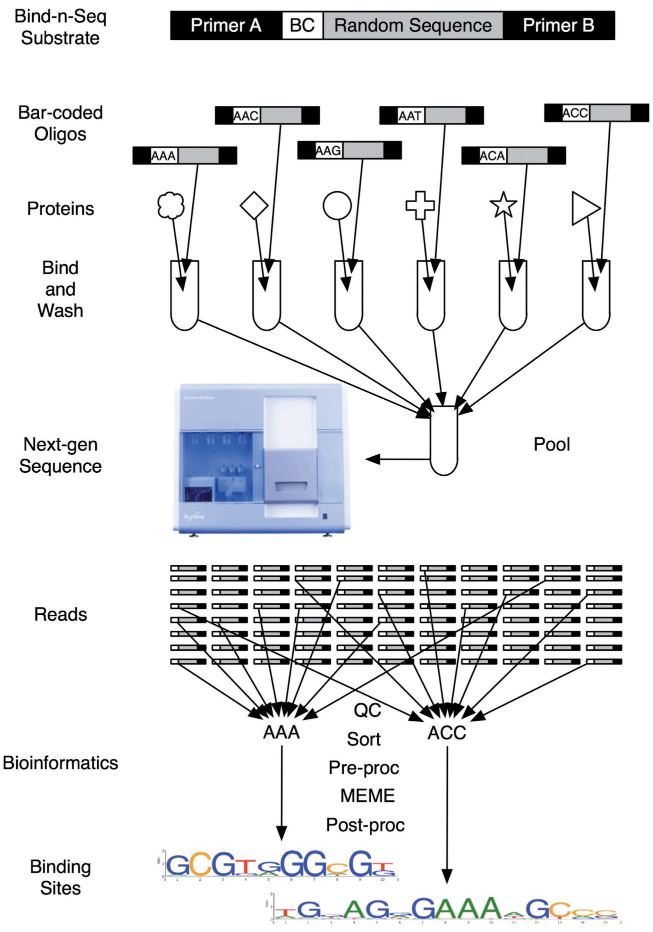

Figure 1. Bind-n-Seq overview. The Bind-n-Seq substrate is an oligo containing constant regions (Primer A and Primer B) a 3-nucleotide bar code (BC) and 21 bp random region. Bar coded oligonucleotides are mixed with various proteins, washed to remove unbound DNA, pooled and sequenced with short read technology. Reads are sorted by their bar codes and processed through several bioinformatics procedures that result in motifs corresponding to the DNA binding sites of each protein.

Massively parallel sequencing has a higher error rate than traditional Sanger sequencing and many of the errors are not simply low frequency point mutations. Regardless of the source of the erroneous reads (possibly technical artifacts or contaminants), some quality control measures are required for any *de novo* sequencing application. In Bind-n-Seq, one can be assured that sequences are derived from a binding reaction by observing the primer-binding site B on the 3′-side (Figure 1). Two fixed nucleotides, AA, in addition to the bar code were used to determine valid sequences on the 5′-side. We used

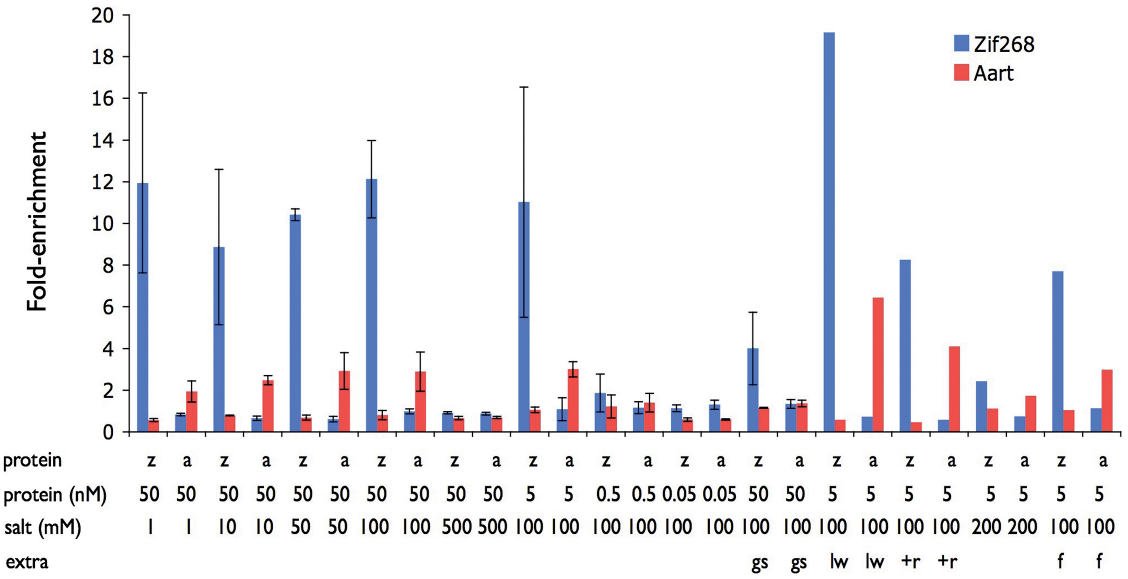

Figure 2. Motif enrichment. The fold-enrichment of known motifs in various binding reactions is shown for Zif268 (blue) and Aart (red). Y-axis: fold-enrichment of a motif in a binding reaction over control (no-protein). Reaction conditions: z, Zif268; a, Aart; protein concentration is shown in nM, salt (KCl) concentration is shown in mM, gs, gel shift; lw, long wash; +r, extra round of selection; f, ficoll. Error bars show the range of values for replicated experiments.

the AA dinucleotide to estimate sequencing error rate near the bar code and found that the error rate was <0.1% at both positions (data not shown). Bar code mis-assignment was therefore quite rare.

Bind-n-Seq reactions

An important goal of our initial binding experiments was to establish reaction conditions that would be stringent enough to provide a significant single-step enrichment of preferred binding sequences, while still allowing enough bound DNA to be recovered for subsequent sequence analysis. One of the attractive features of Bind-n-Seq is that bar coding allows one to combine several binding reactions into a single sequencing reaction. We used this feature to examine several different binding conditions for the DBD of two zinc-finger proteins: the 3-finger Zif268 [aa 349–421, (16)] and the 6-finger engineered Aart (14). We surveyed protein concentrations of 0.05, 0.5, 5 and 50 nM, which cover and exceed the reported K_D values for Zif268 and Aart [6 nM and 50 pM, respectively (14,16,17)]. The proteins were expressed in bacteria as fusions with the MBP, which improves solubility and allows for rapid one-step purification over amylose resin. We compared two enrichment methods: affinity capture of proteins in solution (resin) and separation on a polyacrylamide gel (gel shift) (see 'Materials and Methods' section). For the resin-based enrichment method, amylose resin was added to the binding reactions at equilibrium to capture the proteins, then washed three times with a parameter-specific wash buffer. Buffer salt concentrations ranging from no additional salt (0 mM KCl) to 500 mM KCl were surveyed. For the gel shift method, fluorescently labeled control binding reactions allowed gels to be scanned without drying. We simultaneously examined nine reaction conditions for each protein in the first Bind-n-Seq experiment and four additional reactions for each protein in the second experiment.

Evaluating Bind-n-Seq with known motifs

After quality control procedures (see 'Materials and Methods' section), we obtained a 'clean' data set containing ~100 000 reads for each bar code (Supplementary Tables S1A, S1B and Figure S2). The various reads, separated by their bar codes, were used to compare the success of each binding condition above. Established position weight matrices (PWMs) for Zif268 and Aart (14,18) were used to calculate how many reads contained the expected binding motifs (see 'Materials and Methods' section). It was found that 5 nM protein, 100 mM KCl and the long wash conditions provided the optimal enrichment for both Zif268 and Aart (Figure 2). Under these conditions, the Zif268 expected motif was found in 7.7% of the Zif268-enriched reads compared to 0.42% in the no-protein control. Similarly, 2.0% of the Aart-enriched reads contained the expected motif compared with 0.34% in the control (Supplementary Figure S3).

De novo motif finding with MEME

We used MEME to perform *de novo* motif analysis (see 'Materials and Methods' section and Supplementary Figure S1) on Zif268 and Aart reads under optimal conditions. The computational costs to analyze the complete data set with MEME are prohibitive. Therefore, our overall strategy was to split the reads into five non-overlapping sets of 10 000 sequences, derive intermediate motifs using MEME and merge sequences corresponding to the intermediate motifs in a final application of MEME. For each non-overlapping set, of 10 000 sequences, each run of MEME produced slightly different but highly similar motifs (Supplementary Figure S4).

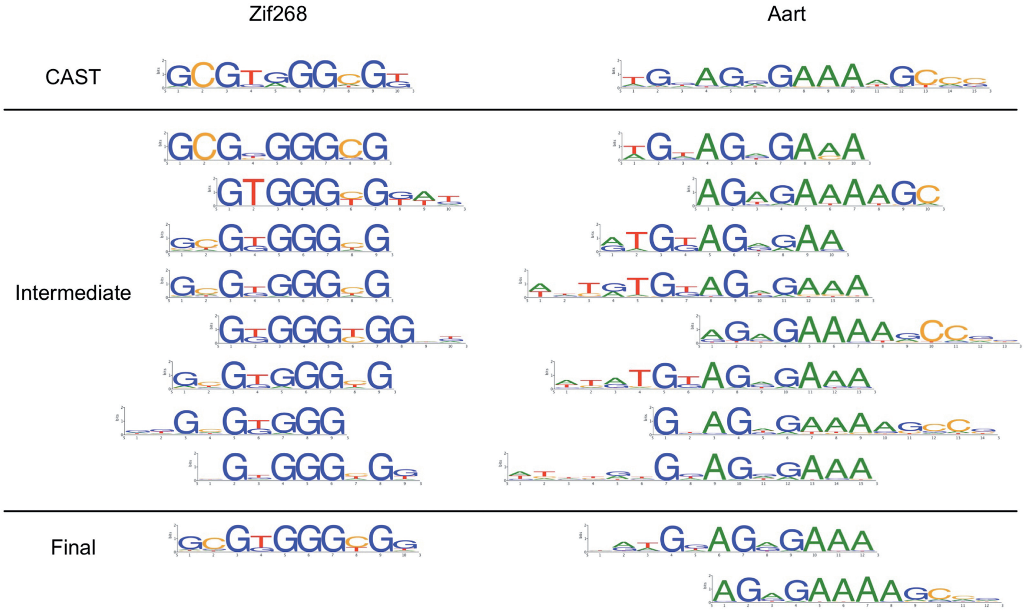

Figure 3. Comparison of CAST and Bind-n-Seq motifs. Intermediate motifs are the result of several non-overlapping sets of reads. Final motifs use reads matching intermediate motifs.

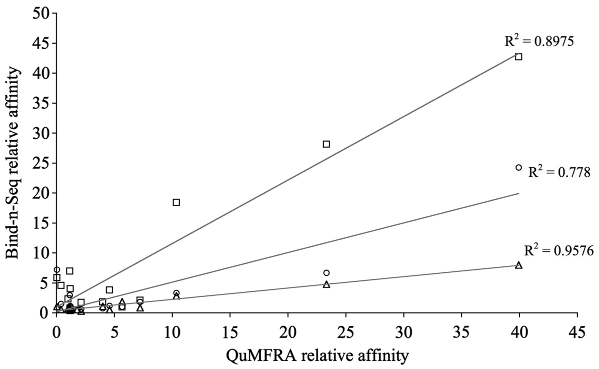

Figure 4. Comparison of relative affinities determined by Bind-n-Seq and QuMFRA. Bind-n-Seq relative affinity is calculated as the fold-enrichment of the 15 sequences (10-mer) (Supplementary Table S2) compared to a no-protein control. All reactions are 5 nm protein and 100 mM salt. Squares are run 1, circles are run 2, triangles are long wash. The relative affinity for the same 15 sequences (10-mer) assessed by QuMFRA is taken from Liu *et al*., 2005.

Among these motifs were low complexity patterns such as poly-A or poly-G/T. These were likely the result of imperfect randomness in the original oligonucleotide library (see 'Discussion' section). To filter off the low complexity patterns, we demanded that all motifs be enriched 4-fold over background (Supplementary Figure S5). To merge the five experiments, we collected all reads that match the intermediate motifs and ran MEME on this subset to arrive at the final motif(s) (Figure 3).

Calculating relative affinities

The relative enrichment of individual *k*-mers should correspond to the affinity of the protein–DNA interaction, since DNA sequences with higher affinity are expected to be more highly enriched than lower affinity sequences. Liu and Stormo (11) reported previously the relative affinity of Zif268 for 15 different DNA sequences using a quantitative multiple fluorescence relative affinity (QuMFRA) assay. To determine if Bind-n-Seq would be a similarly useful method for measuring relative affinities, we examined the fold-enrichment of the same 15 sequences (10-mer) (Supplementary Figure S6) and found an excellent correlation with the results of the previous study ($R^2 = 0.9$ under optimal reaction conditions, Figure 4).

DISCUSSION

Our results demonstrate that Bind-n-Seq can identify reasonable *in vitro* binding site motifs and relative affinity information from a one-step enrichment of an oligonucleotide library containing 1.3×10^{13} 21-mer sequences. Motifs were derived from ~100 000 potential binding sites, which should contain examples of the high-, medium- and low-affinity sequences required for an accurate binding site model (13). The methodology is rapid, and the use of bar coding and massively parallel sequencing allows multiple protein samples to be analyzed simultaneously. This conclusion is supported by the observation that reasonable motifs were identified in each of the non-overlapping sets of 10 000 reads. These results suggest that Bind-n-Seq could be parallelized even further, with 256 samples using four nucleotide bar codes.

Bind-n-Seq was able to determine *de novo* binding motifs that were similar to those described previously for both the 3-zinc finger Zif268 and 6-zinc finger Aart DBDs. These proteins have very different binding characteristics. Zif268 has a G-rich binding motif (18) and binds the sequence 5′-GCG-TGG-GCG-T-3′ with an affinity of about 6 nM (16,17). Aart has a more A-rich binding motif (14) and binds the sequence 5′-ATG-TAG-GGA-AAA-GCC-CGG-3′ with an affinity of 50 pM (14). Both sets of found motifs agree with previous *in vitro* target site selection experiments that not all base pairs seem to be equally involved in the binding of these proteins.

When determining the binding site for a novel protein, the length of the binding site may be unknown. In addition, some polydactyl zinc-finger proteins have been shown to use different combinations of zinc fingers to bind to different sequences (19–22), in which there may be more than one binding site. For these reasons, it is useful to look for multiple long motifs. To derive the final motif for Zif268, we began with a 12 bp site and up to five motifs. This resulted in one motif, but each end of the motif included one non-specific position (Supplementary Figure S7) indicating that the binding site was actually 10 bp. The final motif was then produced by setting the pattern length to 10 bp. For Aart, we began with a 15 bp site and up to five motifs, and this produced three motifs with a few 5′- or 3′-non-specific positions. Reducing the length to 12 bp resulted in two motifs corresponding to either a longer 5′- or 3′-end. Although the Aart CAST-derived pattern is 15 bp, 10–11 bp is apparently enough for Bind-n-Seq. These results are consistent with a recent survey of the *in vitro* DNA binding specificity of 104 murine proteins using the PBM methodology, which found that roughly half of the proteins recognized multiple different sequence motifs (23).

A significant advantage of Bind-n-Seq over other common methods for binding site determination is the large number of potential binding sites recovered in each experiment. This data can then be mined in various ways to gain additional insights into the protein–DNA interaction. In this study, we demonstrated that *k*-mer fold-enrichment can be used to estimate relative affinities. The relative affinity of Zif268 to 15 different 10-mer DNA sequences was calculated with an accuracy similar to that of other quantitative methods (11). As expected, the accuracy was greatest for high-affinity sequences, with low-affinity sequences clustering more generally near the no-protein control (Figure 4). This analysis would not have been possible with the relatively few sequences obtained from a standard SELEX experiment (~20–40 sequences). Compared to PBMs, which can also provide a measure of relative affinity, Bind-n-Seq can sample longer binding sites.

A somewhat surprising finding was that similar experimental conditions (long wash) provided optimal enrichment for both Zif268 and Aart (Figure 2). Also unexpected was that the motif for the lower affinity protein could be recovered more frequently than that of the higher affinity protein; however, this may have been due in part to the difference in binding site length. Shorter sequences appear more frequently in the library than longer ones. The long wash condition proved more stringent than higher salt or one additional round of selection, suggesting that the most critical parameter in these experiments was the dissociation rate of the complexes (k_{off}). The observation that these conditions were useful for proteins differing in affinity by 100-fold provides optimism that these are a limited set of conditions that can be used to analyze many diverse proteins simultaneously. Of course, many proteins bind DNA with affinities lower than those used here and it is possible that their binding motifs could not be determined by this or any other target site selection methodology. However, many low-affinity proteins show cooperativity with neighboring binding factors and the long binding region of the Bind-n-Seq DNAs may be useful for their analysis.

In principle, a variety of sequencing platforms are appropriate for Bind-n-Seq. Short-read sequencing technologies are ideal because the sequenced region is only 35 bp. We found that the longer read 454 technology would require significant software modification to provide binding calls on short oligonucleotide sequences. Another important consideration is the quality of the oligonucleotides. The Bind-n-Seq method relies on the ability to detect protein-dependant binding motifs from a background of random sequences. However, if the background is not perfectly random, motifs could appear to arise due to the background bias. To determine if our oligonucleotide library contained sequence biases, we purchased Bind-n-Seq oligonucleotides from several vendors and sequenced the 21-mer random region to determine if they contained non-random features. All vendors had general compositional biases such that the frequency of each nucleotide was not 25% (Supplementary Figure S8). This was true even with the 'hand mix option'. Sigma oligonucleotides appeared to have a strong neighbor effect, which could be seen from dinucleotide compositions (Supplementary Figure S9). All Sigma oligonucleotides had a low-complexity bias (Supplementary Figure S10), which could explain the low-complexity motifs found by MEME. The Bind-n-Seq reactions in this study employed Sigma oligonucleotides. We do not know if the differences among manufacturers is due to differences in synthesis technology, chemical batches or other factors. Although Bind-n-Seq was able to identify known binding sites in this study, we expect it to perform better with unbiased oligonucleotides. For this reason, we think it would be best to examine the randomness and use the most random oligonucleotides available.

We envision Bind-n-Seq to be useful in variety of basic research and biotechnology settings. One important application is in characterizing transcription factors with unknown binding sites. In the human genome alone, there are more than 700 zinc-finger proteins, most of which have unknown targets. Identifying these targets is not only useful from a discovery perspective, it also brings a wealth of new data to those seeking to model the zinc-finger binding code. A simple-recognition code currently exists but is based on a small number of protein–DNA interactions and is thus incapable of accurately predicting uncharacterized proteins (24–28). An accurate zinc-finger binding code would not only allow the prediction of binding sites, it would also improve the design of engineered domains to target a desired specific sequence (29,30). Bind-n-Seq can also be used to examine the intended and unintended *in vitro* targets of engineered zinc-finger proteins. Although this study focused on zinc-finger proteins because they are the best understood DBD, we expect Bind-n-Seq should be useful for investigating other families of DNA-binding proteins.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank the DNA Technologies and Expression Analysis Cores for carrying out the sequencing on the Illumina Genome Analyzer and pipeline analysis.

FUNDING

National Institutes of Health (GM077403). Funding for open access charge: National Institutes of Health award (GM077403).

Conflict of interest statement. None declared.

REFERENCES

1. Tadepally,H.D., Burger,G. and Aubry,M. (2008) Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. *BMC Evol. Biol.*, **8**, 176.
2. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
3. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
4. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. III and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
5. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
6. Meng,X., Brodsky,M.H. and Wolfe,S.A. (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.*, **23**, 988–994.
7. Durai,S., Bosley,A., Abulencia,A.B., Chandrasegaran,S. and Ostermeier,M. (2006) A bacterial one-hybrid selection system for interrogating zinc finger-DNA interactions. *Comb. Chem. High Throughput Screen.*, **9**, 301–311.
8. Joung,J.K., Ramm,E.I. and Pabo,C.O. (2000) A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **97**, 7382–7387.
9. Meng,X., Thibodeau-Beganny,S., Jiang,T., Joung,J.K. and Wolfe,S.A. (2007) Profiling the DNA-binding specificities of engineered Cys2His2 zinc finger domains using a rapid cell-based method. *Nucleic Acids Res.*, **35**, e81.
10. Wright,W.E., Binder,M. and Funk,W. (1991) Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol. Cell Biol.*, **11**, 4104–4110.
11. Liu,J. and Stormo,G.D. (2005) Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. *Nucleic Acids Res.*, **33**, e141.
12. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
13. Roulet,E., Busso,S., Camargo,A.A., Simpson,A.J., Mermod,N. and Bucher,P. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, **20**, 831–835.
14. Segal,D.J., Crotty,J.W., Bhakta,M.S., Barbas,C.F. III and Horton,N.C. (2006) Structure of Aart, a designed six-finger zinc finger peptide, bound to DNA. *J. Mol. Biol.*, **363**, 405–421.
15. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
16. Pavletich,N.P. and Pabo,C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science*, **252**, 809–817.
17. Segal,D.J., Dreier,B., Beerli,R.R. and Barbas,C.F. III (1999) Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. *Proc. Natl Acad. Sci. USA*, **96**, 2758–2763.
18. Swirnoff,A.H. and Milbrandt,J. (1995) DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell Biol.*, **15**, 2275–2287.
19. Filippova,G.N., Fagerlie,S., Klenova,E.M., Myers,C., Dehner,Y., Goodwin,G., Neiman,P.E., Collins,S.J. and Lobanenkov,V.V. (1996) An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol. Cell Biol.*, **16**, 2802–2813.
20. Tsai,R.Y. and Reed,R.R. (1998) Identification of DNA recognition sequences and protein interaction domains of the multiple-Zn-finger protein Roaz. *Mol. Cell Biol.*, **18**, 6447–6456.
21. Hata,A., Seoane,J., Lagna,G., Montalvo,E., Hemmati-Brivanlou,A. and Massague,J. (2000) OAZ uses distinct DNA- and protein-binding zinc fingers in separate BMP-Smad and Olf signaling pathways. *Cell*, **100**, 229–240.
22. Nagaoka,M., Kaji,T., Imanishi,M., Hori,Y., Nomura,W. and Sugiura,Y. (2001) Multiconnection of identical zinc finger: implication for DNA binding affinity and unit modulation of the three zinc finger domain. *Biochemistry*, **40**, 2932–2941.
23. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
24. Cho,S.Y., Chung,M., Park,M., Park,S. and Lee,Y.S. (2008) ZIFIBI: Prediction of DNA binding sites for zinc finger proteins. *Biochem. Biophys. Res. Commun.*, **369**, 845–848.
25. Liu,J. and Stormo,G.D. (2008) Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics*, **24**, 1850–1857.
26. Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.
27. Siggers,T.W. and Honig,B. (2007) Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.*, **35**, 1085–1097.
28. Pabo,C.O. and Nekludova,L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.
29. Camenisch,T.D., Brilliant,M.H. and Segal,D.J. (2008) Critical parameters for genome editing using zinc finger nucleases. *Mini Rev. Med. Chem.*, **8**, 669–676.
30. Cathomen,T. and Joung,J.K. (2008) Zinc-finger nucleases: the next generation emerges. *Mol. Ther.*, **16**, 1200–1207.