DNA Research **20**, 325–337, (2013)
Advance Access publication on 11 April 2013
doi:10.1093/dnares/dst013

High-Resolution Mapping of *In vivo* Genomic Transcription Factor Binding Sites Using *In situ* DNase I Footprinting and ChIP-seq

Onuma Chumsakul[1], Kensuke Nakamura[2], Tetsuya Kurata[3], Tomoaki Sakamoto[3], Jon L. Hobman[4], Naotake Ogasawara[1], Taku Oshima[1,*], and Shu Ishikawa[1,*]

Graduate School of Biological Sciences, Nara Institute of Science and Technology, 8916-5, Takayama, Ikoma, Nara 630-0192, Japan[1]; Department of Life Science and Informatics, Maebashi Institute of Technology, 460-1, Kamisadori, Maebashi-City, Gunma, Japan[2]; Plant Global Education Project, Graduate School of Biological Sciences, Nara Institute of Science and Technology, 8916-5, Takayama, Ikoma, Nara 630-0192, Japan[3] and School of Biosciences, The University of Nottingham, Sutton Bonington Campus, Sutton Bonington, Loughborough, Leicestershire LE12 5RD, UK[4]

*To whom correspondence should be addressed. Tel. +81-743-72-5431 (S.I. and T.O.).
Fax. +81-743-72-5439 (S.I. and T.O.). Email: shu@bs.naist.jp (S.I.); taku@bs.naist.jp (T.O.).

Edited by Dr Katsumi Isono
(Received 23 February 2013; accepted 22 March 2013)

Abstract

Accurate identification of the DNA-binding sites of transcription factors and other DNA-binding proteins on the genome is crucial to understanding their molecular interactions with DNA. Here, we describe a new method: Genome Footprinting by high-throughput sequencing (GeF-seq), which combines *in vivo* DNase I digestion of genomic DNA with ChIP coupled with high-throughput sequencing. We have determined the *in vivo* binding sites of a *Bacillus subtilis* global regulator, AbrB, using GeF-seq. This method shows that exact DNA-binding sequences, which were protected from *in vivo* DNase I digestion, were resolved at a comparable resolution to that achieved by *in vitro* DNase I footprinting, and this was simply attained without the necessity of prediction by peak-calling programs. Moreover, DNase I digestion of the bacterial nucleoid resolved the closely positioned AbrB-binding sites, which had previously appeared as one peak in ChAP-chip and ChAP-seq experiments. The high-resolution determination of AbrB-binding sites using GeF-seq enabled us to identify bipartite TGGNA motifs in 96% of the AbrB-binding sites. Interestingly, in a thousand binding sites with very low-binding intensities, single TGGNA motifs were also identified. Thus, GeF-seq is a powerful method to elucidate the molecular mechanism of target protein binding to its cognate DNA sequences.

Key words: GeF-seq; ChIP-seq; AbrB; *Bacillus subtilis*

1. Introduction

Genome-wide mapping of the *in vivo* DNA-binding sites of transcription factors or other DNA-binding proteins either by Chromatin Immunoprecipitation coupled with microarray (ChIP-chip)[1] or by the recently developed ChIP coupled with high-throughput sequencing (ChIP-seq) method have become widely used techniques in protein–DNA interaction research.[2–5] The resolution of the DNA-binding sites determined by ChIP-seq was a dramatic improvement on the resolution that was possible using ChIP-chip, because of the higher resolution of high-throughput sequencing compared with oligonucleotide arrays. However, for both techniques, the DNA fragments, co-purified with the target protein (ChIP-DNA), are generated by sonication and generally fall within the size range of 100–500 bp. These sonicated fragments are

often much longer than the actual protein-binding site and, thus, the sequence tags of the ChIP-DNA distribute in broad regions around the actual binding sites. In addition, as only the terminal sequences of ChIP-DNA fragments can be obtained by high-throughput sequencing, piled ChIP-seq tags on the forward (+) and reverse (−) strands usually show bimodal peaks.[6,7] To overcome these problems and determine the actual protein-binding sites to within a few 10 bp, algorithms for the processing of ChIP-seq data have been proposed, although the results obtained by them are still predictive.[6–9] Thus, more precise experimental mapping methods are required to determine the exact binding sites of DNA-binding proteins using ChIP-seq technology.

Recently, the ChIP-exo method, which trims the 5′-region of the protein-unbound region of ChIP-DNA by the use of 5′–3′ lambda (λ) exonuclease, has been developed, and this method demonstrated an improvement in resolution in determining the DNA-binding sites of target eukaryotic proteins through the determination of the edge positions of protein-bound genomic sequences.[10] In contrast to DNA exonucleases, DNase I preferentially cleaves endogenous DNA regions that are not protected by bound proteins and, thus, has been employed for *in vitro* footprinting to precisely determine the DNA-binding sites of DNA-binding proteins.[11] Using DNase I digestion, Vora *et al.*[12] proposed a method, designated *in vivo* protein occupancy display (IPOD), which visualizes the *in vivo* binding profile of total DNA-binding proteins on genomic DNA.[12] In this method, genomic DNA cross-linked with total proteins and extracted from formaldehyde-treated cells was digested with DNase I, and the DNase I-resistant DNA fragments were purified by phenol extraction and mapped using a tiling array.

We report here a novel method designated as Genome Footprinting by high-throughput sequencing (GeF-seq; *in vivo* GeF-seq). This method combines *in situ* DNase I digestion of bacterial genomic DNA with a modified ChIP-chip method (ChAP-chip, Chromatin Affinity Precipitation-chip) we previously developed.[13] Unlike IPOD, GeF-seq can visualize the binding profile of a specific target protein at a resolution seen at the *in vitro* footprinting level. We evaluated the resolution achieved using the GeF-seq method by examining the binding profile of the *Bacillus subtilis* transition state regulator, AbrB, in comparison with results obtained by ChAP-chip and a modified ChIP-seq method (ChAP coupled with high-throughput sequencing) utilizing sonication to fragment the genomic DNA. AbrB represses the expression of many genes during exponential growth, and we have demonstrated using ChAP-chip that AbrB binds to hundreds of sites throughout the entire *B. subtilis* genome during exponential growth.[14] AbrB is a small protein (10.4 kDa), having a unique structure. The N-terminal domains of two AbrB molecules form a single DNA-binding domain, and AbrB forms a tetramer having a stable DNA-binding ability, via both N-terminal and C-terminal interactions. Structural modelling of AbrB bound to the target sequence indicated that the AbrB tetramer would interact with ∼20 bp sequences,[15] whereas *in vitro* footprinting studies detected a wider range of binding regions from 25 to 80 bp, suggesting that a higher order structure of the AbrB tetramer may be involved in DNA binding at some sites on the chromosome.[16–18] We previously proposed that AbrB binds to bipartite TGGNA motifs based on the *in vivo* AbrB-binding regions determined by ChAP-chip analysis,[14] which is in accordance with a motif identified by the *in vitro* SELEX method.[17] However, the consensus sequence was detected in a small number of AbrB-binding regions, and the consensus DNA-binding sequence for AbrB is not completely understood at present.

We demonstrate here that, by mapping the sequences of short DNA fragments co-purified with AbrB after *in situ* DNase I digestion of the genomic DNA, the AbrB-binding profile could be visualized with a resolution comparable with that of *in vitro* footprinting. Importantly, the BiPad web server for modelling bipartite sequence elements[19] automatically detected consensus sequences for AbrB binding in >95% of the experimentally determined binding sites. Moreover, highly accurate DNA-binding site information obtained by GeF-seq enabled us to obtain a comprehensive view of the correlation between AbrB-binding signals and cognate recognition sequences; AbrB not only binds to bipartite motifs in sequences with high binding signals, but also to single-sequence motifs in sequences with low signals. These results demonstrate the usefulness of the GeF-seq method.

2. Materials and methods

2.1. Bacterial strain

Bacillus subtilis strain OC001 expressing C-terminal 2HC (12 histidines plus a chitin-binding domain)-tagged AbrB (AbrB-2HC) was used throughout.[14]

2.2. ChAP-chip and ChAP-seq

ChAP-chip data for AbrB binding on the *B. subtilis* genome were taken from our previous report.[14] DNA fragments for ChAP-seq analysis were prepared, as previously described.[13,14] Construction of the DNA library for Illumina sequencing was as described below except for the size of the DNA fragments used: ∼250 bp fragments, corresponding to ∼150 bp DNA

fragments isolated by ChAP without adapter sequences, were selected for PCR enrichment.

2.3. In situ *DNase I digestion of genomic DNA*

The GeF-seq method is schematically illustrated in Fig. 1A. To cross-link protein–DNA complexes, 400 ml of OC001 (*abrB-2HC*) cells grown to the exponential phase in Luria-Bertani medium at 37°C were treated with formaldehyde as previously described.[14] To hydrolyze the cell wall without osmotic burst, cells were treated with 5 mg/ml lysozyme in 3 ml of isotonic sucrose-malate-magnesium buffer (0.02 M maleic acid, 0.5 M sucrose, and 0.02 M $MgCl_2$, pH 6.5 adjusted with NaOH)[20] in the presence of 1 mM phenylmethylsulfonyl fluoride (PMSF). After 20-min incubation at 37°C with mixing, cells were collected by centrifugation at 6000 *g* for 5 min at 4°C. Cells were resuspended in 0.5 ml of a buffer containing 0.1 M Tris–HCl (pH 7.5), 0.2 M NaCl, 1% (v/v) Triton X-100, 0.1% (w/v) Na-deoxycholate, 0.2% (w/v) Brij 58, and 20% (v/v) glycerol.

To determine suitable conditions for *in situ* DNase I digestion of genomic DNA, four samples of OC001

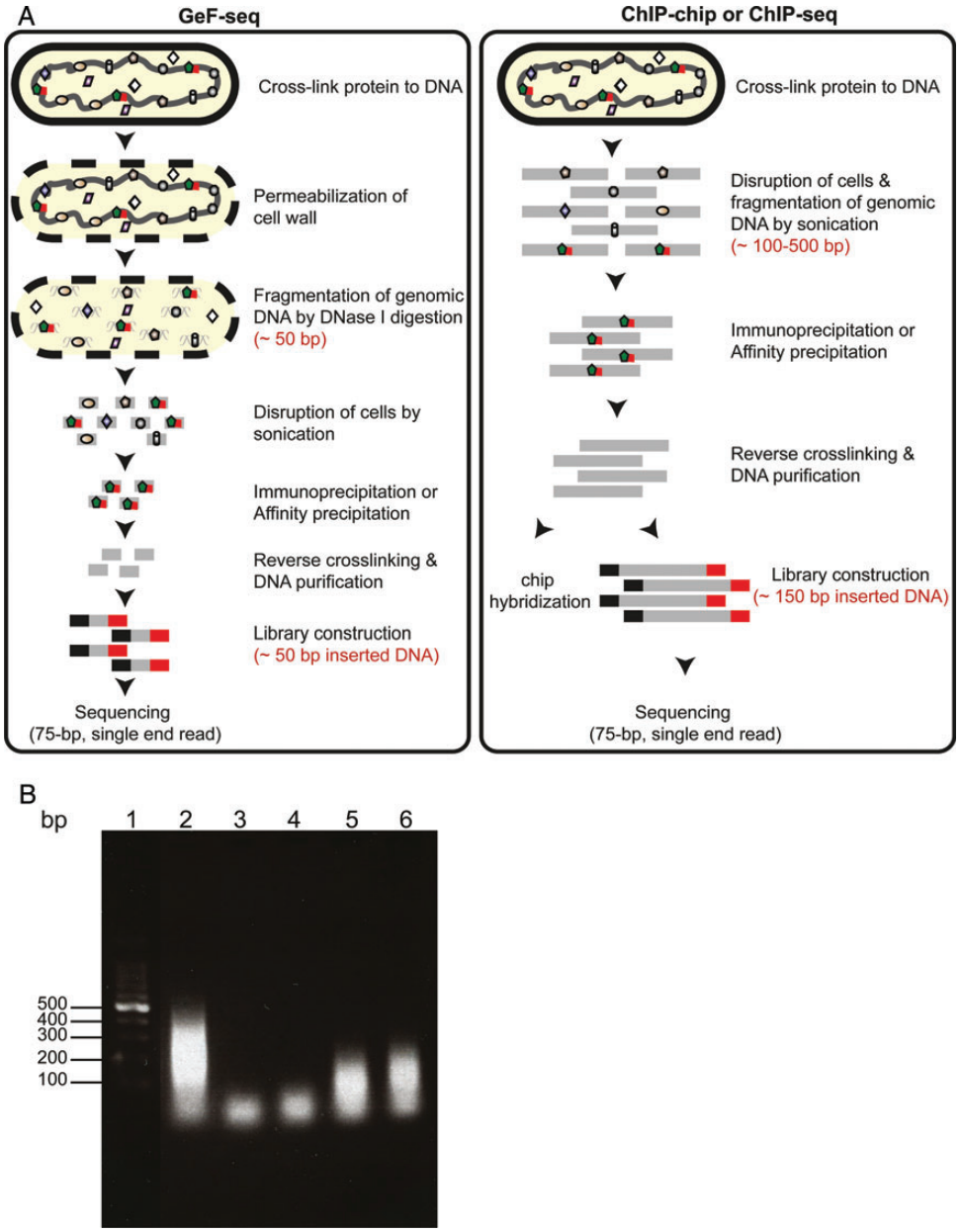

Figure 1. (A) Schematic workflow for GeF-seq (see Materials and methods for detail). For comparison purposes, the ChIP-chip and ChIP-seq methods are also illustrated. (B) *In situ* DNase I digestion. The size of DNA fragments digested by various concentrations of DNase I (1, 0.6, 0.4, and 0.2 U/ml, Lanes 3–6, respectively) and analysed by gel electrophoresis. DNA fragments generated by sonication are run alongside the DNase I digested DNA for comparison purposes (Lane 2).

cells were prepared as described above and mixed with 10 μl of RNase A (10 mg/ml) and 50 μl of a solution containing 100 mM $MgCl_2$ and 50 mM $CaCl_2$. DNase I digestion was started with the addition of 0.5, 0.3, 0.2, and 0.1 units (U) of DNase I (corresponding to a final concentration of 1, 0.6, 0.4, and 0.2 U/ml) (Takara) and incubated at 37°C with shaking (230 rpm) for 30 min. The reaction was terminated by urea denaturation upon the addition of 3 ml of urea-Triton buffer [0.1 M 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (pH 7.5), 0.01 M imidazole, 8 M urea, 0.5 M NaCl, 1% Triton X-100, 10 mM β-mercaptoethanol, and 1 mM PMSF] instead of ethylenediaminetetraacetic acid, which severely inhibits protein purification by Dynabeads TALON (invitrogen). The samples were then sonicated on ice using an Astrason Ultrasonic Processor XL (Misonix) for 10 min (4 s 'on' and 10 s 'off', at output level 5). After centrifugation to remove cell debris, 30 μl of the supernatant was mixed with 70 μl of M-wash buffer (0.1 M Tris–HCl, pH 7.5, 1% sodium dodecyl sulfate, 0.01 M dithiothreitol) and incubated at 65°C overnight to reverse the cross-linking. After the removal of proteins by phenol–chloroform–isoamyl alcohol treatment, DNA was recovered by ethanol precipitation in the presence of glycogen, resuspended in 50 μl of nuclease-free water and run on a 2% agarose gel (Fig. 1B). Treatment with 0.5 units of DNase I (1 U/ml) generated DNA fragments <100 bp in size, and incubation with higher amounts of DNase I resulted in a decrease in the amount of DNA detected by agarose gel electrophoresis (data not shown). Thus, we selected 0.5 units (1 U/ml) of DNase I for further analysis.

2.4. *Affinity purification of DNA fragments bound to AbrB*

AbrB–DNA complexes were affinity-purified from the clarified DNase I-treated cell lysate, using Dynabeads TALON as described previously,[13,14] but with the following modification: after protein–DNA complexes were purified and reverse cross-linked by heat treatment at 65°C overnight, proteins were removed using two phenol–chloroform–isoamyl alcohol extractions, and DNA fragments were recovered by ethanol precipitation in the presence of glycogen.

2.5. *Sequencing of DNA fragments co-purified with AbrB*

The DNA library for sequencing by the Illumina Genome Analyzer IIx (GAIIx) was generated using the NEB Next DNA Sample Prep Reagent kit (New England BioLabs) according to manufacturer's instructions for 'Preparing Samples for Sequencing Genomic DNA' (Illumina) with the following modification; after ligation of the adapters to the DNA fragments, the ligated product was run on a 2% [Tris-acetate-EDTA (TAE)] low-range agarose gel (Biorad) at 50 V for 2.5 h in TAE buffer and the region of the gel ~150 bp (although the DNA was not visible on the gel), corresponding to ~50 bp fragments without adapter sequences, was excised. The DNA fragments were then purified using a QIAquick Gel Extraction kit (Qaigen) and amplified using 14 cycles of PCR, to obtain at least 1 fmol of DNA library. The amount of DNA was determined by an Agilent 2100 Bioanalyzer using the High-Sensitivity DNA Kit (Agilent). The sequence of the library was then determined by 75-bp single-ended sequencing using the Illumina GAIIx sequencer according to the manufacturer's instructions.

2.6. *Mapping of read sequences and normalization of tag counts*

A total of 10 369 855 read sequences obtained from the Illumina GAIIx were mapped on the reference genome (*B. subtilis* str. 168, NC_000964.3), and the mapping results were visualized using the mpsmap and psmap softwares (http://metalmine.naist.jp/maps/gefseq), respectively.[21] Because DNA fragments of ~50 bp (without adapter sequences for PCR amplification) were selected in the sample preparation process to obtain complete sequences of the ChAP-DNA fragments, most of the reads reached into the adapter sequence attached to the 3′-end of ChAP-DNA. Thus, unlike general Illumina™ sequencing results obtained by following the instruction manual, most of the read sequences consisted of ~50 bp of ChAP-DNA sequence followed by the adapter sequence, and both of these sequences varied in length. Since mapping of such different lengths of sequence containing the unmappable adapter sequence was not possible using a standard sequence mapping/assembly program, we utilized the property of mpsmap that maps different length sequences to the best chromosomal location, while allowing up to a specified number of mismatches without a gap. In this study, the read sequences were initially mapped allowing a maximum of 35 mismatches, and the adapter sequences were finally removed. As a result of the first mapping, 9 685 519 (93%) of the read sequences were uniquely mapped to the reference genome. (Thus, the genomic regions encoding the 10 rRNA operons were not included in the present analysis.) Then, to remove the adapter sequences, the starting positions were assigned to seven or more bases allowing a two-base mismatch matched with 5′-end of the primer sequence (AGATCGGAAGAGCTCGTATGCCGTCTTCTGCT

TGA) in the 3′-region of the read sequences. In addition, mapped sequences (without adapter sequences) with >2 bp mismatches against the reference sequence were removed, and 8 571 055 (83%) of the read sequences remained for further analysis. Finally, in order to normalize the difference in the local copy number of genomic DNA, counts of mapped reads at each nucleotide position along the genome sequence were linearly scaled by using the *oriC*/*terC* ratio (5.15), estimated by sequencing and mapping of whole genomic DNA fragments digested by DNase I, to define the AbrB-binding signals (Supplementary Fig. S1).

Results shown in Supplementary Fig. S1 suggested that there was preferential digestion of AT-rich genomic sequences by DNase I. However, mapping results of the distribution of ChAP-DNA sequences suggested that the preferentiality apparently did not affect the quantitative estimation of the AbrB-binding profile (Supplementary Fig. S2).

2.7. *Detection of protein-binding regions*

Most of the read sequences were mapped on distinct regions along the genome surrounded by regions where ends of the read sequences accumulated (Fig. 3C). We used this feature to define the AbrB-binding sites. To estimate the end points of the genomic sequences in the read sequences more precisely, we reanalysed them so that adapter sequences at the 3′-ends of the sequences could be subtracted from the genomic DNA they had been attached to during generation of the library for sequencing. We assigned sequences as 'adapter sequences' when five bases at the 3′-end of the read sequence were identical to the adapter primer sequence and the following sequences matched to the primer sequence with no more than two bases mismatched. The accumulation profile of the 3′-ends thus determined across the genome sequence was similar to that of 5′-ends, which was defined as the first base of the read sequences (Supplementary Fig. S3), strongly suggesting that the procedure to estimate the 3′-ends of read sequences was reliable.

Then, the left ends of the read sequences relative to the reference genome sequence were defined as a sum of the 5′-ends of read sequences mapped on the plus strand and the 3′-ends of read sequences mapped on the minus strand, whereas the right ends were defined as a sum of the 3′-ends of read sequences mapped on the plus strand and the 5′-ends of read sequences mapped on the minus strand (Supplementary Fig. S3). We counted the numbers of left and right ends mapped to each

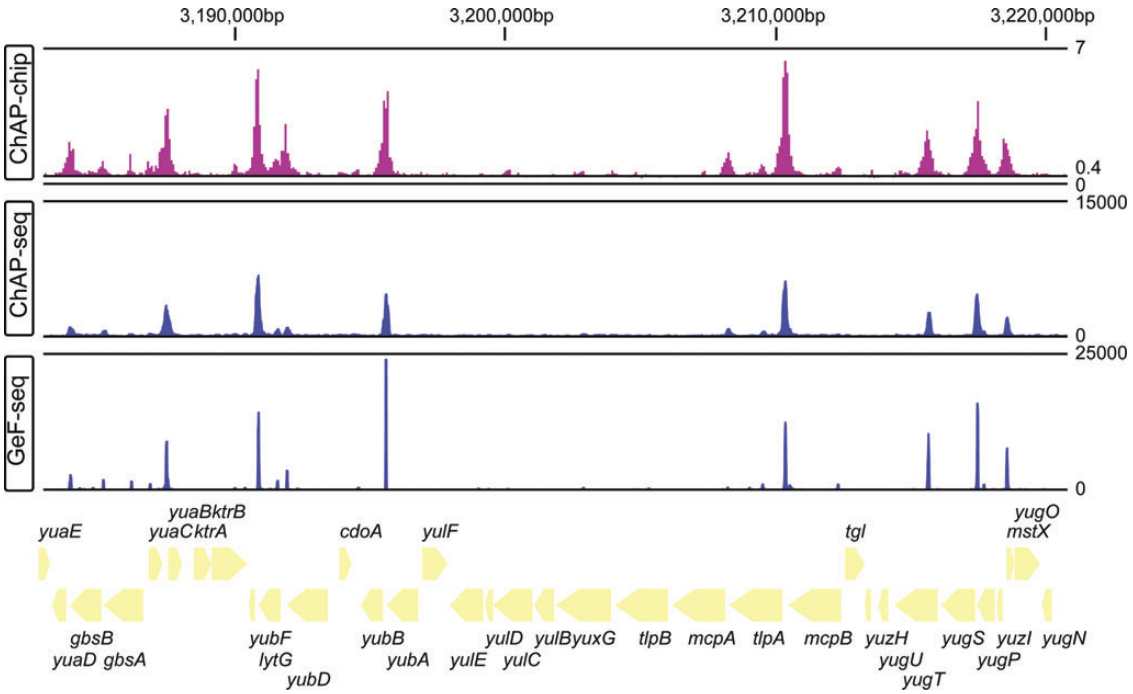

Figure 2. Example of AbrB-binding profiles detected by ChAP-chip, ChAP-seq, and GeF-seq on the *B. subtilis* genome. The distribution of AbrB-binding sites on the *B. subtilis* genome is shown in the indicated genome region at the top of the figure. For the ChAP-chip results, previously published data were used. The top, middle, and bottom lines indicate binding intensities of 7, 0.4, and 0, respectively, as defined by the previous report.[14] For ChAP-seq and GeF-seq, protein-binding intensities (number of reads after normalization of the local copy number of the genomic DNA) are presented as a sum of binding intensities on the plus and − strands. The top lines in both the ChAP-seq and GeF-seq columns indicate binding intensities of 15 000 and 25 000, respectively, and the bottom lines indicate binding intensities of 0. The gene organization is schematically shown at the bottom.

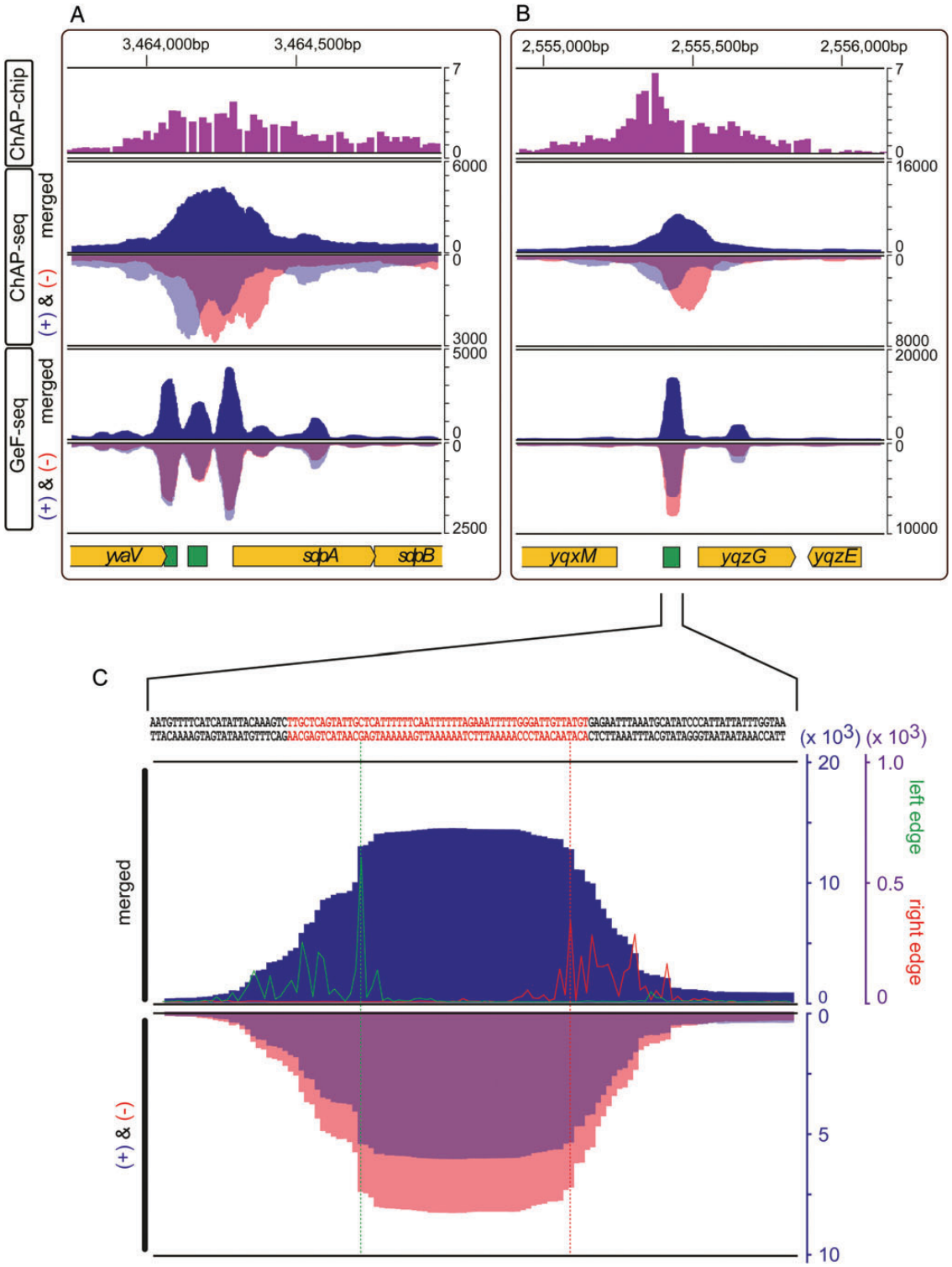

Figure 3. Comparison of AbrB-binding sites determined by GeF-seq, ChAP-Seq, and ChAP-chip. (A) A typical example of AbrB-binding regions that are merged into one peak in ChAP-Seq and ChAP-chip, but separately detected by GeF-seq. For the ChAP-chip results, previously published data were used. For ChAP-seq and GeF-seq, binding intensities on the forward (+) and reverse (−) strands are inversely shown by transparent blue and red peaks, respectively (thus, overlapping regions appear purple) in the respective inverted orientation (+ and −) and their summed binding intensities are shown by a dark blue peak in the upper part of the figure (merged). The gene organization is schematically shown at the bottom. (B) A typical example of AbrB-binding regions, where ChAP-seq detected two peaks on the plus and minus strands with a gap, but GeF-seq detected two peaks at the same position (see Results for details). (C) Resolution of detection within an AbrB-binding region of the chromosome by GeF-seq analysis. The intergenic region between *yqxM* and *yqzG* in (B) is magnified. The number of left and right ends is depicted by green and red lines, respectively, at each base. The scale for the protein-binding intensity and the number of left and right ends are indicated on the right inside and outside, respectively, and the highest positions on the left and the right edges are marked by vertical dotted lines whose boundaries are defined as the protein-binding region. The AbrB-binding region, which was previously determined by *in vitro* footprinting,[18] is also shown by green boxes in (A) and (B) and red letters in the DNA sequence on the top in (C).

nucleotide, and positions with ≥10 ends of read sequences and with the highest number of ends within ±30 bp windows were determined in 1 bp steps, as candidates for the left and the right boundaries of the DNA-binding sequences. Then, we extracted regions surrounded by a pair of possible left and right boundaries positioned within a range from 25 to 80 bp (considering the *in vitro* AbrB footprinting results), and regions, where AbrB-binding signal intensities exceeded a threshold value at more than half of the nucleotides between them, were extracted as AbrB-binding sites. In this study, we first extracted AbrB-binding sequences using the signal intensity corresponding to the top 10th percentile of all nucleotides across the genome as the threshold (Supplementary Fig. S4). At some regions, different combinations of boundaries surrounding the overlapping sequences satisfied the criteria. In such cases, the innermost sequences were selected as AbrB-binding sequences for further analysis. Finally, the average of the AbrB-binding signals within the individual binding sequence was calculated, as the AbrB-binding signal intensities of each binding site.

2.8. *Motif analysis*

AbrB-binding DNA motifs were analysed by the BiPad web server (http://bipad.cmh.edu) for modelling bipartite sequence elements.[19] The BiPad program performs multiple local alignment by entropy minimization and cyclic refinement using a stochastic greedy search strategy, and we used the following settings: left half-site, gap range lengths, right half-site, and the iteration cycles were set to 9, 0 or 1, 9, and 500, respectively. To examine the possibility of whether the AbrB-binding motif was discovered by chance, we selected three sets of data, each of which consists of 300 50 bp sequences randomly selected from the *B. subtilis* 168 genome sequence by the RSA tool,[22] and analysed by Bipad.

2.9. *Sequencing data*

Sequencing data in this study have been submitted to the DDBJ Sequence Read Archive (DRA) and the BioProject database under accession code DRA000758 and PRJDB675, respectively.

3. Results

3.1. In vivo *GeF-seq*

To improve the resolution of protein-binding site determination by ChIP-seq or ChAP-seq methodologies, we attempted *in situ* DNase I digestion of the cross-linked bacterial nucleoid to restrict the size of DNA fragments co-purified with the target protein to directly interacting sequences (Fig. 1A). We employed *B. subtilis* AbrB as a model protein, whose binding sites were recently determined by use of the ChAP-chip method to be >600 sites scattered across the genome.[14] Exponentially growing *B. subtilis* OC001 cells expressing C-terminal 2HC (12 histidines plus a chitin-binding domain)-tagged AbrB (AbrB-2HC) were treated with formaldehyde to stabilize the protein–DNA interactions by cross-linking, and the collected cells were treated with lysozyme in isotonic buffer to facilitate an efficient penetration of DNase I into cells. Then, the genomic DNA was fragmented to <100 bp by the DNase I treatment, followed by the affinity purification of the cross-linked AbrB–DNA complexes using cobalt-coated magnetic beads. DNA fragments co-purified with AbrB (ChAP-DNA) were isolated after reversing the cross-linking between proteins and DNA. As we intended to obtain whole sequences of ChAP-DNA to avoid the bimodal distribution of sequence tags, DNA fragments containing ∼50 bp of inserted DNA, after ligation of adapter sequences, were selected to prepare the library for high-throughput sequencing by Illumina GAIIx. It has been demonstrated that AbrB interacts with ∼20 bp sequences[15] and, thus, we also expected that 50 bp fragments would be enough to cover single AbrB-binding sites. Single-ended 75-bp sequencing by the Illumina GAIIx generated 9 685 519 (uniquely mapped) sequence reads. As expected, most of the read sequences (88.5%) contained the adapter sequences for PCR amplification at the 3′-end portion, with an average insert size of 50 bp after removal of them (Supplementary Fig. S5), and insert sequences were mapped on distinct sites on the *B. subtilis* genome. Then, counts of the mapped reads at each nucleotide position along the genome were normalized for differences in the local copy number of genomic DNA, to define the AbrB-binding signals.

3.2. *Comparison of the distribution of AbrB-binding signals determined by GeF-seq, ChAP-seq, and ChAP-chip*

To evaluate the resolution of the GeF-seq method in identifying genomic protein-binding sites, we initially compared the distributions of AbrB-binding signals along the genome as determined by three methods: GeF-seq, ChAP-seq, and ChAP-chip. The distributions of the AbrB-binding signals on the genome determined by GeF-seq and ChAP-seq in the present study were highly consistent with that of ChAP-chip we reported previously.[14] Typical examples of the comparison are presented in Fig. 2, and the complete profiles of the binding signals across the genome obtained by the three methods are available in Supplementary Fig. S2. Close-up views of profiles of AbrB-binding signals (Fig. 3A and B) indicated that,

although the ChAP-seq method improved the resolution of detection of the binding regions compared with ChAP-chip, the GeF-seq method dramatically improved the resolution even when compared with ChAP-seq. Importantly, GeF-seq could resolve the closely positioned binding sites that appear as one peak in the ChAP-seq method, as shown in Fig. 3A.

Using ChAP-seq, binding sites were often detected as two broad peaks on the forward (+) and reverse (−) strands, as previously reported.[6] In contrast, using GeF-seq, the distributions of sequence tags on the plus and − strands overlapped in the middle of the two ChAP-seq peaks (Fig. 3B). Thus, the use of short DNA fragments enabled the conclusive determination of protein-bound regions of DNA without the necessity for the bioinformatic prediction of the binding sites. In addition, AbrB-binding signals at each binding site generally distributed in a trapezoid form, and the ends of the read sequences accumulated at the left and right edges (Fig. 3C). These observations strongly suggested that *in situ* DNase I digestion occurred at the boundaries of protein-binding sites, as observed in *in vitro* DNase I footprinting. Furthermore, the lengths of sequences protected from DNase I digestion (27−80 bp) suggested that these sequences would be interacting with one to three AbrB tetramer(s). We used this feature to define AbrB-binding sequences, as described below.

3.3. *Determination of AbrB-binding sequences*

To automatically extract DNA sequences bound by AbrB from the GeF-seq results, we developed an analytical pipeline as described in Materials and methods. Briefly, we first surveyed pairs of nucleotide positions showing the highest accumulation of ends of read sequences, as candidates for the borders of the protein-binding regions. Then, AbrB-binding signals between them were evaluated using a relaxed threshold value, corresponding to the signal intensity at the top 10th percentile of all nucleotide positions across the genome (Supplementary Fig. S4). This resulted in 5897 possible AbrB-binding sites being detected (Supplementary Table S2), which included not only specific, but also non-specific AbrB-binding, sites. These were extracted and ranked by their average binding signal intensities of nucleotides included in each site. The peak ID was given from 1 to 5897 by their intensity ranked from high to low, respectively.

The top 700 binding sites accompanied by high-binding signal intensities were first examined, because this number was approximately similar to that obtained by previous ChAP-chip analysis (694 binding sites).[14] The length of the AbrB-binding regions determined by the GeF-seq ranged from 27 to 79 bp (Supplementary Fig. S6), which was consistent with the results of *in vitro* footprinting experiments listed in a database of transcriptional regulation in *B. subtilis*[23] and in recent reports.[17,18,24] Among 32 AbrB-binding sites previously determined by *in vitro* DNase I footprinting, our GeF-seq experiment detected 11 AbrB-binding sequences (Fig. 3C and Supplementary Table S1). GeF-seq also detected 10 AbrB-binding sequences within the 5897 possible AbrB-binding sites, although binding intensities were lower than those of the top 700 binding sites (Supplementary Table S1). Thus, we found that these 21 binding sites matched those obtained by *in vitro* DNase I footprinting. These results indicate that our GeF-seq method has the ability to detect protein-bound DNA sequences with a resolution comparable with that of the *in vitro* footprinting method, although differences in boundaries are observed between our GeF-seq result and the *in vitro* footprinting result, which may result from differences in conditions between *in vivo* and *in vitro* experiments.

3.4. *Identification of consensus sequences for the AbrB binding*

In previous ChAP-chip analysis,[14] we found a possible consensus sequence for AbrB binding to be TNCCA−4 bp−TGGNA, which is composed of a pair of two AbrB-binding motifs previously identified by the *in vitro* SELEX method.[17] However, those motifs were detected in a limited number of AbrB-bound sequences. In addition, we found that not only TNCCA−4 bp−TGGNA, but also other bipartite TGGNA motifs, in palindromic or tandem orientation, separated by 4−5 bp were enriched in AbrB-bound DNA sequences on the *B. subtilis* genome.

In the present GeF-seq analysis, the lengths of automatically extracted AbrB-binding sequences were restricted to an *in vitro* DNase I footprinting level. Thus, we expected that the large amount of precise information on AbrB-binding sequences might give us a clear view on the consensus AbrB-binding sequence. We then utilized the BiPad web server, a web interface to predict sequence elements embedded within unaligned sequences, to analyse the experimentally derived AbrB-binding sequences. BiPad predicts various pairs of bipartite motifs with different gaps in different orientations as one consensus sequence.[19] BiPad successfully identified a mixture of bipartite TGGNA motifs in 96% (678) of the 700 experimentally identified sequences (Fig. 4A), and we found that we could classify them into six patterns by manually sorting the predicted consensus in each AbrB-binding sequence (Fig. 4B and Supplementary Table S3). As a result, consensus sequences were found to be composed of bipartite TGGNA motifs separated by 4 or 5 bp AT-rich sequences arranged in direct, reverse

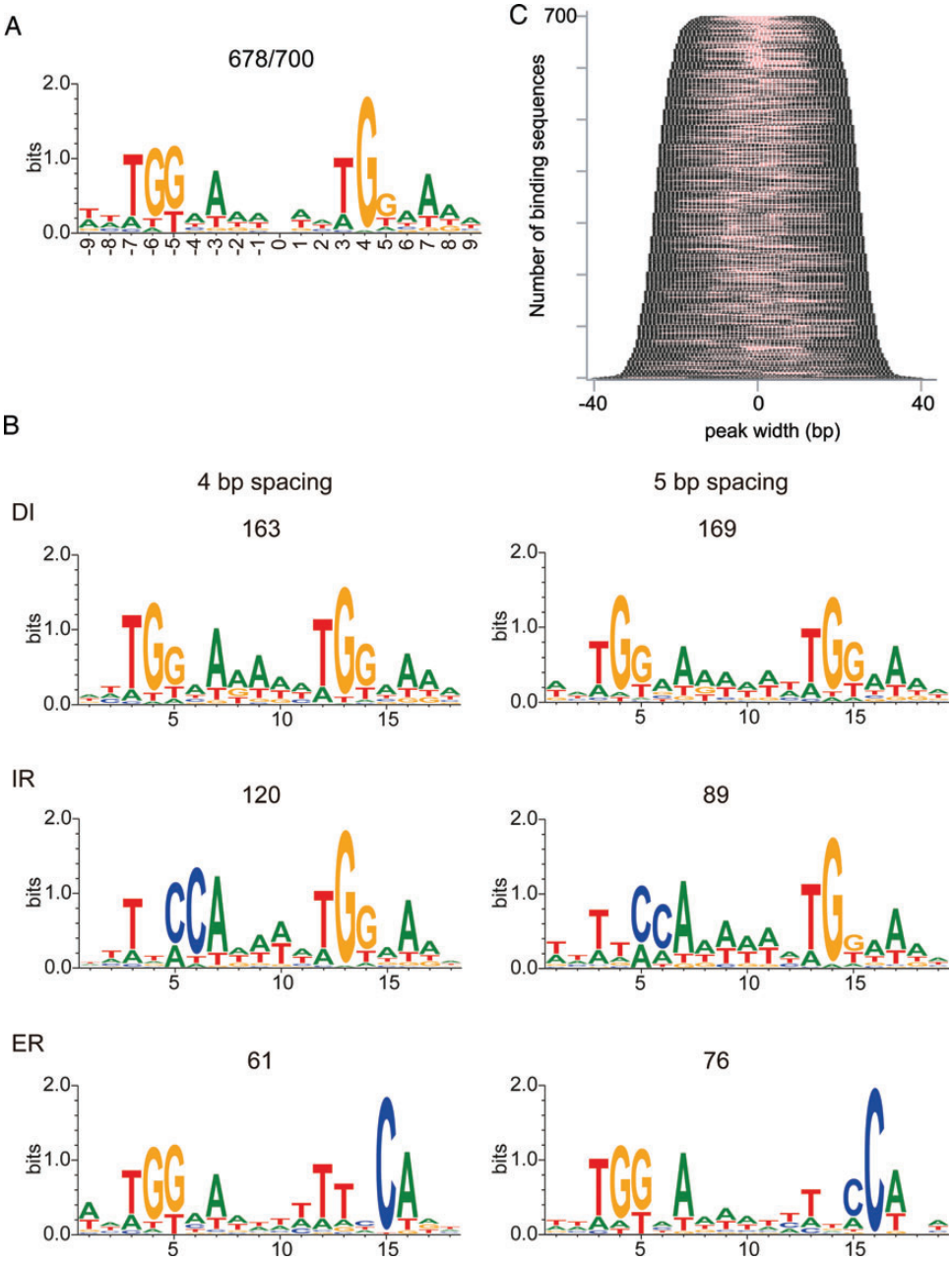

Figure 4. Consensus sequences of AbrB-binding sites identified by BiPad. (A) The top 700 AbrB-binding sequences ranked by their average binding intensity from high to low were analysed. The results of BiPad analysis were plotted regardless of the orientation of the TGGNA/TNCCA motif; direct repeat (DI), reversed direct repeat (RDR), inverted repeat (IR), and everted repeat (ER). The number of input sequences and sequences judged to contain the consensus sequences by the algorithms are indicated. The mixture of four types of bipartite patterns with 4- and 5-bp spacing in (A) were sorted and the graphical logo for each pattern was re-created in (B). The DI and RDR bipartite pattern were plotted in the same logo. The number of sequences belonging to each bipartite pattern is shown. (C) Distribution of bipartite DNA motifs within the 700 binding sequences. The binding sequences were sorted by length, and the motifs identified by Bipad were marked red.

direct, inverted, and everted repeat orientations (Fig. 4B). Importantly, the location of the consensus sequence was usually close to the middle of the experimentally identified binding sequences (Fig. 4C). Thus, we not only confirmed that the AbrB-binding consensus sequence we proposed previously was indeed detectable in almost all of the AbrB-binding sequences with high binding signals, but we also demonstrated that the information on the protein-binding sequences automatically extracted by the GeF-seq analysis enabled us to clearly identify a consensus sequence for protein binding, at least in the case of AbrB.

It should be also noted that any clear consensus sequence was not detected in the remaining 22 sequences, although a degenerate single TGGNA motif was detected (data not shown). Since AbrB

binding to these sequences was clearly detected with high signal intensity (Supplementary Fig. S2 and Supplementary Table S2), this result indicates that AbrB also binds to sequences without bipartite motifs by some mechanism, for example, when the DNA sequence forms structure(s) to fit the AbrB-binding surface.

3.5. *Correlation between AbrB-binding signals and motif discovery*

Here, using 700 AbrB-binding DNA sequences with high GeF-seq AbrB-binding signals, we identified bipartite AbrB-binding motifs across the *Bacillus* genome arranged in any orientation with a 4- or 5-bp spacing, which is consistent with our previous ChAP-chip analysis[14] and the *in vitro* SELEX results reported by Xu and Strauch.[17] These results strongly suggested that, when binding signal intensities are high, the sequences are those specifically recognized by AbrB. We usually use a threshold value to discriminate 'real' protein-binding peaks and possible 'artificial' binding peaks in ChIP-chip and ChIP-seq experiments. However, actually, these threshold values have been operationally defined by researchers; for example, aiming to remove false positives or to remove false negatives, and examination of actual protein binding to extract possible binding sequences has rarely been examined thoroughly. The finding that we could identify AbrB-binding consensus sequences in almost all of the binding sequences accompanied with high binding signals in GeF-seq prompted us to comprehensively examine whether there was conservation of the binding motifs in sequences with lower binding signals.

To this end, we divided the 5897 possible AbrB-binding sequences into 20 sets each containing 300 sequences, according to average AbrB-binding signal intensities (Supplementary Table S2 and Supplementary Fig. S4C), and the consensus sequence for each set was extracted by the Bipad program (Fig. 5). In the three datasets with the top AbrB-binding signal intensities (1–300, 301–600, and 601–900), the bipartite TGGNA motifs with a 4- or 5-bp spacer sequence were detected in almost all sequences (98, 96, and 96% respectively, Fig. 5). In the next group of datasets with lower binding signal intensities (901–1200, 1201–1500, and 1501–1800), although the consensus sequences containing bipartite TGGNA motifs with a 4- or 5-bp spacer sequence could be detected, one half-site became degenerate. Interestingly, in the following 10 sets (from 1801 to 4800), only a single TGGNA motif was detected, whereas, in the remaining four sets with the lowest binding signal intensities (from 4801 to 5897, Fig. 5), the single motif becomes very degenerate. We confirmed that the TGGNA motif was not detected by chance because no motif was detected in similar sets of DNA sequences (300 50-bp sequences) that were randomly extracted from the genome sequence of *B. subtilis* 168 (data not shown). These results strongly suggested that AbrB not only binds stably, with high experimentally derived binding signals, to bipartite TGGNA motifs, but also interacts with single TGGNA motifs in sequences with lower but significant experimentally derived signal intensities (Supplementary Fig. S4).

4. Discussion

We demonstrate here that, by mapping the sequences of ~50 bp fragments co-purified with AbrB after *in situ* DNase I digestion of genomic DNA, *in vivo* AbrB-binding sites could be determined with a resolution comparable with that of *in vitro* footprinting. Furthermore, comprehensive and precise information on the DNA sequences that AbrB binds gave us a clear view of AbrB binding on the *B. subtilis* genome—it would stably bind to bipartite TGGNA motifs, but it also interacted with many single TGGNA motifs on the genome.

In vitro DNase I footprinting has currently been one of the most widely used methods to determine at high resolution the precise DNA sequences bound by transcription factors and other DNA-binding proteins. However, this method is laborious and can be performed against only a few DNA sequence targets in one experiment. In addition, the synthetic conditions under which DNase I foorprinting assays have been conducted risks leading to artifactual results for several reasons, e.g. the use of purified proteins that are not modified as would occur *in vivo* and may not work in the same way, the low-ionic strength of solutions used in *in vitro* footprinting experiments that may allow non-specific DNA–protein interactions, the use of short DNA sequences that may lack the secondary structure of DNA found *in vivo*, experiments conducted at non-physiological temperatures, and the absence of essential effectors, which may impair the specific binding of the protein to the corresponding DNA sequence. In contrast, in the GeF-seq method, DNA–protein interactions in the nucleoid are stabilized in the living cells by formaldehyde treatment, and then DNA digestion is carried out *in situ* to retain the native DNA-binding state of the target protein. Thus, the GeF-seq method identifies the actual DNA-binding sequences of target proteins across the whole genome simultaneously, with minimal risk of artefacts, and at high resolution, which is comparable with that of *in vitro* footprinting. In analysing the resolution of the method, we confirmed that 21 AbrB-binding regions we found using GeF-seq were consistent with *in vitro* footprinting results that have been reported previously (Supplementary

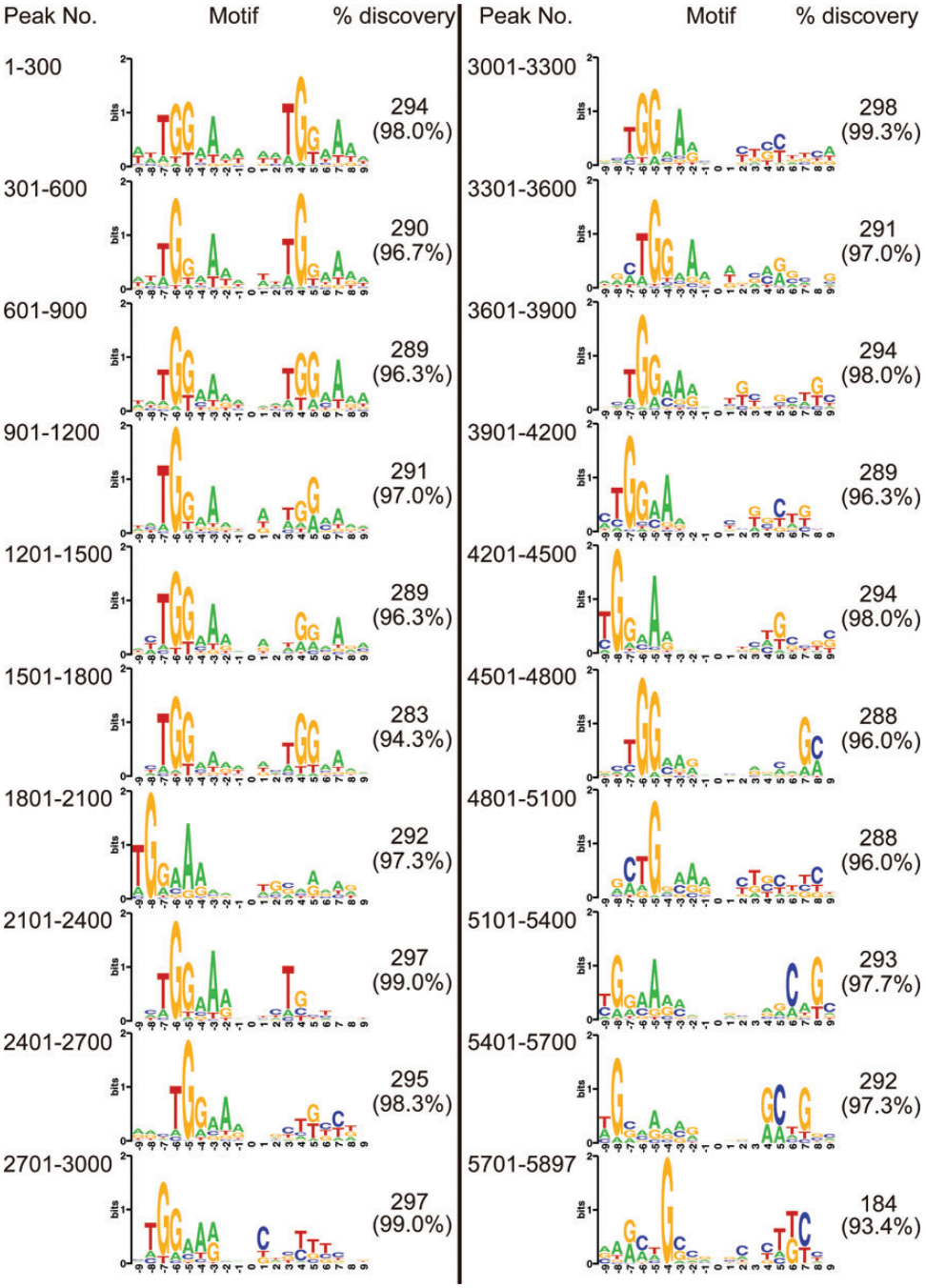

Figure 5. Consensus sequences of the AbrB-binding sequences detected in the top 10th percentile threshold. Each group of 300 signals in the 20 sets of AbrB-binding sequences were ranked from high to low by binding intensities and were analysed by Bipad to detect respective consensus sequences. The number of sequences containing the motif is shown in parallel.

Table S1). It should also be noted that some AbrB-binding sites previously detected by *in vitro* footprinting (Supplementary Table S1) were not detected using GeF-seq. Although it is possible that such regions might be occupied by AbrB under some specific conditions, AbrB binding to these DNA sequences may occur only in *in vitro* experimental conditions.

Here, we detected not only bipartite TGGNA motifs in probable high-affinity AbrB-binding sites, but also single TGGNA motifs in possible low-affinity AbrB-binding sites, that would be generally ignored as non-specific, using the binding-site prediction software. Such low-affinity AbrB binding may not be biologically important, but it is possible that those binding sites may have a role to concentrate AbrB molecules on the nucleoid to increase the chance of finding high-affinity binding sites, which are directly involved in gene regulation.[25] We usually use a

threshold value to discriminate 'real' protein-binding peaks on the genome and possible 'artificial' binding peaks in ChIP-chip and ChIP-seq experiments. However, our results clearly demonstrated that the use of threshold values could discard important information. Our results suggest that comprehensive and precise information on protein-binding sequences obtained by GeF-seq analysis, in combination with the identification of consensus sequences in them, would give us a clear and comprehensive view of protein binding on the genome.

Specifically, we clearly demonstrate here that the consensus sequence for the high-affinity AbrB binding is comprised of bipartite TGGNA motifs gapped by a 4- or 5-bp AT-rich sequence arranged in direct, reverse direct, inverted, and everted repeat orientations. This result is consistent with a previous *in vitro* SELEX study,[17] and our informatics analysis showing that various bipartite motifs are enriched in AbrB-binding regions determined by ChIP-chip.[14] Thus, the GeF-seq results reported here show, for the first time, the highly flexible proposed consensus sequences, which are actually recognized by AbrB molecules in *in vivo*. Previous structural modelling of AbrB bound to the target DNA sequence indicated that the AbrB tetramer would interact with ~20 bp sequences,[15] whereas *in vitro* footprinting studies detected a wider range of binding regions from 25 to 80 bp. In this study, GeF-seq also detected a similar range of AbrB-binding regions from 27 to 80 bp in size. When the positions of the bipartite motifs within the binding sequences are depicted, the motifs are usually located in the middle of the binding sequences, but some are not centrally located in the long binding sequences (Fig. 4C, Supplementary Table S3). Interestingly, we observed that the binding region is generally composed of multiple TGGNA motifs almost covering the full length of the sequenced region (data not shown), suggesting that higher oligomers of AbrB may interact with multiple TGGNA motifs.

Here, we have demonstrated that GeF-seq is a powerful tool for helping to understand the *in vivo* distribution of DNA-binding proteins on the genome. However, several issues remain to be explored, in order to fully establish the GeF-seq method. (i) We have not yet examined how DNase I digestion conditions would affect the results, although the results shown in Supplementary Figs S1 and S2 suggest that the GeF-seq results would be robust against changes in DNase I digestion conditions. (ii) We empirically selected criteria to map read sequences and to define protein-binding sites on the genome. Further improvements in the sequence data processing algorithms are desirable to automate this process. (iii) The Bipad program outputs one consensus sequence for each input sequence, and a method to identify multiple motifs in each sequence is desirable. (iv) GeF-seq data suggest that protein-binding signal intensities to the genome should correlate with protein-binding affinities to the cognate target sequences, but this needs to be shown experimentally. (v) GeF-seq has successfully determined protein-binding sites across a bacterial genome, but examination of whether this method is applicable for much larger genomes of higher organisms is necessary.

Acknowledgements: We are grateful to Dr C. Bi for useful advice on motifs analysis using BiPad web server.

Supplementary data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work has been supported by the Advanced Low Carbon Technology Research and Development Program (ALCA) of the Japan Science and Technology Agency (JST) and a UK Royal Society International Joint Project to T.O. and J.L.H. Interactions between the authors' laboratories have been facilitated by a BBSRC/JST Japan Partnering Award.

References

1. Kim, T.H. and Ren, B. 2006, Genome-wide analysis of protein-DNA interactions, *Annu. Rev. Genomics Hum. Genet.*, **7**, 81–102.
2. Barski, A., Cuddapah, S., Cui, K., et al. 2007, High-resolution profiling of histone methylations in the human genome, *Cell*, **129**, 823–37.
3. Robertson, G., Hirst, M., Bainbridge, M., et al. 2007, Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, *Nat. Methods*, **4**, 651–7.
4. Mikkelsen, T.S., Ku, M., Jaffe, D.B., et al. 2007, Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Nature*, **448**, 553–60.
5. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. 2007, Genome-wide mapping of in vivo protein-DNA interactions, *Science*, **316**, 1497–502.
6. Pepke, S., Wold, B. and Mortazavi, A. 2009, Computation for ChIP-seq and RNA-seq studies, *Nat. Methods*, **6**, S22–32.
7. Jothi, R., Cuddapah, S., Barski, A., Cui, K. and Zhao, K. 2008, Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data, *Nucleic Acids Res.*, **36**, 5221–31.
8. Kharchenko, P.V., Tolstorukov, M.Y. and Park, P.J. 2008, Design and analysis of ChIP-seq experiments for DNA-binding proteins, *Nat. Biotechnol.*, **26**, 1351–9.
9. Zhang, Y., Liu, T., Meyer, C.A., et al. 2008, Model-based analysis of ChIP-Seq (MACS), *Genome Biol.*, **9**, R137.

10. Rhee, H.S., and Pugh, B.F. 2011, Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution, *Cell*, **147**, 1408–19.
11. Hampshire, A.J., Rusling, D.A., Broughton-Head, V.J., and Fox, K.R. 2007, Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands, *Methods*, **42**, 128–40.
12. Vora, T., Hottes, A.K., and Tavazoie, S. 2009, Protein occupancy landscape of a bacterial genome, *Mol. Cell.*, **35**, 247–53.
13. Ishikawa, S., Ogura, Y., Yoshimura, M., et al. 2007, Distribution of stable DnaA-binding sites on the *Bacillus subtilis* genome detected using a modified ChIP-chip method, *DNA Res.*, **14**, 155–68.
14. Chumsakul, O., Takahashi, H., Oshima, T., et al. 2011, Genome-wide binding profiles of the *Bacillus subtilis* transition state regulator AbrB and its homolog Abh reveals their interactive role in transcriptional regulation, *Nucleic Acids Res.*, **39**, 414–28.
15. Sullivan, D.M., Bobay, B.G., Kojetin, D.J., et al. 2008, Insights into the nature of DNA binding of AbrB-like transcription factors, *Structure*, **16**, 1702–13.
16. Strauch, M.A. 1995, Delineation of AbrB-binding sites on the *Bacillus subtilis* spo0H, kinB, ftsAZ, and pbpE promoters and use of a derived homology to identify a previously unsuspected binding site in the bsuB1 methylase promote, *J. Bacteriol.*, **177**, 6999–7002.
17. Xu, K. and Strauch, M.A. 1996, In vitro selection of optimal AbrB-binding sites: comparison to known in vivo sites indicates flexibility in AbrB binding and recognition of three-dimensional DNA structures, *Mol. Microbiol.*, **19**, 145–58.
18. Strauch, M.A., Bobay, B.G., Cavanagh, J., Yao, F., Wilson, A. and Le Breton, Y. 2007, Abh and AbrB control of *Bacillus subtilis* antimicrobial gene expression, *J. Bacteriol.*, **189**, 7720–32.
19. Bi, C., and Rogan, P.K. 2006, BIPAD: a web server for modeling bipartite sequence elements, *BMC Bioinformatics*, **7**, 76.
20. Wyrick, P.B. and Rogers, H.J. 1973, Isolation and characterization of cell wall-defective variants of *Bacillus subtilis* and *Bacillus licheniformis*, *J. Bacteriol.*, **116**, 456–65.
21. Nakamura, K., Oshima, T., Morimoto, T., et al. 2011, Sequence-specific error profile of Illumina sequencers, *Nucleic Acids Res.*, **39**, e90.
22. Van Helden, J. 2003, Regulatory sequence analysis tools, *Nucleic Acids Res.*, **31**, 3593–6.
23. Sierro, N., Makita, Y., de Hoon, M. and Nakai, K. 2008, DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information, *Nucleic Acids Res.*, **36**, D93–96.
24. Murray, E.J., Strauch, M.A. and Stanley-Wall, N.R. 2009, SigmaX is involved in controlling *Bacillus subtilis* biofilm architecture through the AbrB homologue Abh, *J. Bacteriol.*, **191**, 6822–32.
25. Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E.G., Berg, O.G. and Elf, J. 2012, The lac repressor displays facilitated diffusion in living cells, *Science*, **336**, 1595–8.