Published online 10 October 2018

Nucleic Acids Research, 2019, Vol. 47, No. 2 ***e8***
doi: 10.1093/nar/gky890

Fast and global detection of periodic sequence repeats in large genomic resources

Hideto Mori[1,2,3], Daniel Evans-Yamamoto[1,2,3], Soh Ishiguro[1,2,3], Masaru Tomita[2,3,4] and Nozomu Yachie[1,2,3,5,6,*]

[1]Synthetic Biology Division, Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo 153-8904, Japan, [2]Institute for Advanced Biosciences, Keio University, Tsuruoka 997-0035, Japan, [3]Systems Biology Program, Graduate School of Media and Governance, Keio University, Fujisawa 252-0882, Japan, [4]Department of Environment and Information Studies, Keio University, Fujisawa 252-0882, Japan, [5]Department of Biological Sciences, School of Science, The University of Tokyo, Tokyo 113-0033, Japan and [6]PRESTO, Japan Science and Technology Agency (JST), Tokyo 153-8904, Japan

Received July 23, 2018; Revised September 18, 2018; Editorial Decision September 19, 2018; Accepted September 20, 2018

ABSTRACT

Periodically repeating DNA and protein elements are involved in various important biological events including genomic evolution, gene regulation, protein complex formation, and immunity. Notably, the currently used genome editing tools such as ZFNs, TALENs, and CRISPRs are also all associated with periodically repeating biomolecules of natural organisms. Despite the biological importance of periodically repeating sequences and the expectation that new genome editing modules could be discovered from such periodical repeats, no software that globally detects such structured elements in large genomic resources in a high-throughput and unsupervised manner has been developed. We developed new software, SPADE (Search for Patterned DNA Elements), that exhaustively explores periodic DNA and protein repeats from large-scale genomic datasets based on *k*-mer periodicity evaluation. With a simple constraint, sequence periodicity, SPADE captured reported genome-editing-associated sequences and other protein families involving repeating domains such as tetratricopeptide, ankyrin and WD40 repeats with better performance than the other software designed for limited sets of repetitive biomolecular sequences, suggesting the high potential of this software to contribute to the discovery of new biological events and new genome editing modules.

INTRODUCTION

Significant roles of repetitive DNA and protein sequences have been widely reported in both eukaryotes and prokaryotes. Transposable DNA elements are thought to be among the most important evolutionary driving factors that have been expanding within and between species' genomes via their copy-and-paste or cut-and-paste mechanisms (1,2). These repetitive elements induce large-scale genomic rearrangements and transcriptional regulation. Nonmobile short tandem repeat DNA sequences are also key elements inducing structural genome evolution in prokaryotic species. Tandem DNA repeats induce mispairing and slippage between repetitive units during DNA replication and drive genomic contraction and expansion (3). Intramolecular crossover DNA recombination is also promoted between tandem repeat regions of the genome (4). Some of these events are known to be reversible and lead to genomic phase variation, allowing cells and species to rapidly adapt to changing environments without having to undergo irreversible mutations (5).

Repetitive protein domains often serve as structural binding modules that stably interact with biopolymers. They are widely involved in protein folding and interactions in various biological processes. Tetratricopeptide repeats (TPRs) (6) and ankyrin (ANK) (7) repeats are large protein repeat families that are conserved from prokaryotes to eukaryotes. The repeat units of TPRs and ANK repeats are 34 and 33 amino acids (aa) long, respectively, and both are composed of a helix–turn–helix structure. These repetitive domains have been reported to mediate interactions with other proteins and RNAs and play important roles in cell cycle control, transcriptional regulation, translational inhibition, and protein translocation (6,7). WD40 repeat is another large protein repeat family found in both prokaryotic and eukaryotic species, but its functions are particularly well known in eukaryotes (8). A WD40 repeat is composed of seven-bladed β-propellers, where each propeller is around 40 aa long, involving four anti-parallel β-sheets, and serves as a scaffold for protein interaction. Accordingly,

*To whom correspondence should be addressed. Tel: +81 3 5452 5242; Fax: +81 3 5452 5241; Email: yachie@synbiol.rcast.u-tokyo.ac.jp

WD40 proteins coordinate multi-protein complex formation and underlie diverse biological functions such as signal transduction, transcriptional regulation, cell cycle control, chemotaxis, autophagy and apoptosis (8,9).

The structural code for proteins in general remains largely unclear and there have been major challenges in engineering these repeat protein modules to develop synthetic binding reagents for biomedical and nanotechnology applications (10–12). Most protein repeat sequences are 'imperfect' or 'degenerated,' where each repetitive unit contains variable amino acid residues and the degrees of repeat imperfectness vary widely. Some of these variable residues determine binding to specific biomolecules and deciphering this code for DNA binding has been extremely beneficial in the development of genome editing technologies (13,14). Several transcriptional regulators involve tandem protein repeats with specific periodicities to wrap around the double-stranded DNA helix. Cys2His2 zinc fingers (C2H2 ZNFs) are the most common DNA-binding motif found in eukaryotic transcription factors (15). C2H2 ZNFs represent periodic protein repeats that make tandem contact to targeting DNA sequence. The repeat unit size ranges from 28 to 30 aa and the variable amino acid residue pattern in each unit defines its binding to a specific DNA triplet (16). Similarly, transcription activator-like effectors (TALEs) of the type III secretion system encoded in the plant pathogenic bacteria of the *Xanthomonas* genus also have repeating domains (17). They are virulence proteins that bind to the host plant genomic DNA and hijack its gene expression system. The periodicity of the repeat unit ranges from 33 to 35 aa, where the combination of two variable amino acid residues at the 12th and 13th positions of the repeat sequence has a one-to-one relationship with a specific mononucleotide. By fusing DNA cleavage domains such as FokI endonuclease to C2H2 ZNFs and TALEs, the genome editing tools zinc finger nucleases (ZFNs) and TALE nucleases (TALENs), respectively, have been developed, both of which enable highly specific targeted DNA cleavage. Other effector proteins have also been fused to C2H2 ZNFs and TALEs to regulate gene expression and chromosomal structures in various organisms (18,19).

The CRISPR–Cas systems have become the most widely used genome editing technologies in recent years (20). As indicated by their name, clustered regularly interspaced short palindromic repeats (CRISPRs) are widely encoded in prokaryotic genomes (21). The unique characteristics of these CRISPRs and CRISPR-associated (Cas) proteins in bacterial and archaeal immunity have been rapidly identified recently (22). In the immunization process, a fragment of defined length from invading phage or plasmid DNA is incorporated into the 5′ end of a CRISPR locus with a constant motif sequence. Accordingly, the periodic interspaced repeats of CRISPRs have been derived by continuous cycles of this immunization process. In the immunity process, an RNA originating from the immunized DNA is transcribed and processed and guides Cas protein(s) to its complementary sequence of exogenous DNA for cleavage and degradation. Harnessing different Cas proteins and RNAs involved in the immunization/immunity processes of different CRISPR-type families, various genome editing technologies have been established (20,23,24). Cas9 with double-stranded DNA cleavage activity from the type II CRISPR system has been widely used for targeted gene disruption and targeted fragment knock-in in various organisms including mammals. Similar to ZFNs and TALENs, nuclease-deficient Cas9 (dCas9) or mutant Cas9 nickase (nCas9) fused to effector proteins such as transcription factors, deaminases, and fluorescent proteins have been used for various applications such as gene silencing (25), activation (26), base editing (27), and chromosomal labeling (28).

Since periodically repeating DNA and protein sequences have diverse and important roles in biology, a simple and optimistic hypothesis could be proposed that new genome editing modules can be discovered from other periodic repeats in large-scale genomic resources. However, there is no universal software that captures various types of periodic repeats from large-scale genomic datasets in an unsupervised manner (Table 1). For example, RepeatMasker is one of the most commonly used tools to detect interspersed DNA repeats and low-complexity DNA sequences (29). However, this software screens only DNA sequences against a database of reported elements and does not evaluate repeat periodicity. Previous software programs developed for *de novo* searches of repetitive biomolecular sequences also have certain limitations. Tandem Repeat Finder is one of the first types of software to screen tandem and low-complexity DNA repeats without prior knowledge (30), but is incapable of capturing highly degenerated or interspaced DNA sequences or protein repeats. RECON (31) and RepeatScout (32) also screen only DNA sequences, focus only on interspersed repeats regardless of periodicity, and exclude tandem or low-complexity repeats. PRAP captures both tandem and interspersed repeats, but screens only DNA sequences (33). Although the recently developed software XSTREAM (34) and T-REKS (35) search for both tandem and highly degenerate repeats from DNA and protein sequences, both are ineffective at capturing interspersed or interspaced repeats including CRISPRs. With the recent interest in genome editing, several software packages such as CRISPRFinder (36), CRISPRDetect (37) and AnnoTALE (38) have been developed to capture genome-editing-associated sequences. However, such specialized software does not have the potential to discover novel genome editing modules.

The previously developed software focuses on limited types of repeat sequences for specific biological targets, but it seems that any software that combines the abilities of the previous software packages for any type of repetitive sequence would give an ambiguous and large set of sequences, which would require substantial effort for further curation and validation. However, none of the above-mentioned software screens repetitive sequences based on sequence periodicity that commonly appears in many significant biological processes. This could be a strong constraint in screening to obtain a set of biomolecular sequences with high potential for expanding our biological knowledge and developing new biotechnologies. Accordingly, we have been motivated to develop simple and fast software called SPADE (Search for Patterned DNA Elements) that globally captures such periodically repetitive biomolecular sequences in large genomic datasets mainly based on a simple evaluation of *k*-mer periodicity.

Table 1. Comparison of different software tools for capturing repetitive biomolecular sequences

		Protein or DNA		Repeat type				
Software	Method	DNA	Protein	Tandem	Degenerate	Interspersed	Periodicity	Comments
RepeatMasker	Supervised	*Yes*	*No*	*Yes*	*Yes*	*Yes*	*No*	
CRISPRFinder	Supervised	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	CRISPR only
CRISPRDetect	Supervised	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	CRISPR only
AnnoTALE	Supervised	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	TALE only
Tandem Repeat Finder	*De novo*	*Yes*	*No*	*Yes*	*No*	*No*	*No*	
RECON	*De novo*	*Yes*	*No*	*No*	*Yes*	*Yes*	*No*	
RepeatScout	*De novo*	*Yes*	*No*	*No*	*Yes*	*Yes*	*No*	
PRAP	*De novo*	*Yes*	*No*	*Yes*	*Yes*	*Yes*	*No*	
XSTREAM	*De novo*	*Yes*	*Yes*	*Yes*	*Yes*	*No*	*No*	
T-REKS	*De novo*	*Yes*	*Yes*	*Yes*	*Yes*	*No*	*No*	
SPADE	*De novo*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	This study

MATERIALS AND METHODS

Genomic resources

The GenBank files for the 7006 complete prokaryotic genomes (downloaded on 31 March 2017) and the human reference genome version GRCh38.p10 were downloaded from the NCBI RefSeq genomes FTP server (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/).

SPADE screening phase 1: detection of highly repetitive regions

In SPADE, each entry sequence is first scanned by a sliding window of a given size w to roughly detect highly repetitive regions (HRRs). Let W_i and k_i be the sliding window at sequence position i of the entry sequence and its left-most k-mer sequence, respectively. At every sliding window position i, the number of k_i within W_i (n_i) is cumulatively counted for every position in the left-most k_i sequence region. In the same sliding window, 1 is also counted for every position of the other k_i sequence regions. Let c_i be the resulting cumulative k-mer score at position i of the entry sequence. After scanning the query sequence with the sliding window, cumulative k-mer peak areas (CKPAs) for which the peak heights are s or more are extracted. From the sequence regions for all of the CKPAs ($c > 0$), the broadest possible regions consisting of multiple gaps of size g or less were extracted as highly repetitive regions (HRRs). We adopted $w = 1000$, $k = 10$, $s = 20$, and $g = 300$ for nucleotide sequence and $w = 300$, $k = 3$, $s = 6$, and $g = 50$ for protein sequence as default parameters of the software.

SPADE screening phase 2: evaluation of periodicity

Let h be the size of a given HRR. For each HRR surrounding region of size $h \pm m$, SPADE generates a position-period matrix (PPM) using a similar sliding window of size h. Let W_i and k_i be the sliding window at sequence position i and its left-most k-mer sequence, respectively. When multiple k_i sequence regions are detected in W_i, the number of k_i regions (n_i) is cumulatively counted for all of the corresponding row–column cells of the first two k-mer regions, where row represents distance between two identical k-mers and column represents sequence position. When $n_i > 2$, from the second k_i sequence regions, this procedure is iteratively repeated except that the number added to each cell is 1. After scanning by the sliding window, the highest peak period d in the column sum distribution of the resulting PPM is identified. All values in a sub-PPM of rows from $[d \times 0.8]$ to $[d \times 1.2]$ are then added up and divided by that from 1 to half the column size of the PPM to produce the periodicity score. HRRs with periodicity scores of p and more are redefined as periodic repeat regions (PRRs). We set $m = 1000$ and $P = 0.5$ for nucleotide sequence and $m = 300$ and $P = 0.3$ for protein sequence as default parameters.

SPADE screening phase 3: identification of repetitive motifs

From each PRR with sequence period d, the k-mer sequence that has contributed the most to the sequence periodicity is extracted as k_{seed}. When multiple k-mer sequences are extracted as the k-mers contributing the most, the left-most k-mer in the PRR is selected as k_{seed}. Starting from all of the k_{seed} sequences found in the PRR, SPADE obtains sequence fragments of size d. The extracted sequences are then aligned by multiple sequence alignment using MAFFT version 7.22 (39) to identify their consensus sequence motif. For each sequence position of the alignment result, the information content of appearing letters (b, bit) and the frequency of alignment gaps (f) are calculated using the Python WebLogo 3.6.0 package (40). After removing positions with f of more than q from the alignment result, letter consistency l of every position is calculated by $b \times f$. The positions of the alignment result are then treated as circular since they are for periodic repeats and punctuated by removing the longest continuous nonconsensus region ($l < u$) of more than r letters. When this punctuation does not happen, the sequence alignment result is linearized as it was before. To map the repeat motif to the PRR sequence, a representative sequence is obtained by taking the most frequent letter in each position of the alignment result. When the representative sequence is shorter than k-mer, the identical sequence regions are scanned in the PRR and annotated as repeat units. Otherwise, the representative sequence is mapped using BLAST+ version 2.6 (41) with the blastn-short (for nucleotide) or blastp-short (for protein) option and alignment length threshold of 50% to the query length or E-value of 0.01 or less. The hit regions in the PRR are then used to construct a sequence logo profile using Python WebLogo 3.6.0 package (40). From the sequence logo profile, a repeat motif

sequence is generated by the most frequent letters, where a highest letter frequency of less than 60% is masked with '*'. $q = 0.5$, $u = 0.8$, and $r = 5$ were set as default parameters of the software.

Protein secondary structure prediction

For each visualized protein repeat motif sequence, the confidence score for α-helix structure or β-sheet structure was calculated using PSIPRED version 3.3 (42). For each PRR detected by SPADE, PSIPRED was initially used to predict all possible secondary structure motifs with the confidence score at each amino acid residue position. We then calculated the average confidence score for each motif at every position in the repeat sequence unit.

Evaluation of performance for detecting CRISPRs

From the entire periodic DNA repeats detected by SPADE, we extracted CRISPR candidates with interspace sizes of 25–60 bp and repeating periods of 58–81 bp. The interspace size parameters and the minimum threshold for the repeating period (interspace size plus repetitive sequence size) were set with reference to the CRISPRFinder screens for CRISPR candidates with interspace size being 25–60 bp and repetitive sequence size being 23–55 bp, but our maximum threshold for the repeating period was defined empirically based on the reported RefSeq CRISPRs. Region overlap agreement (ROA) between two given regions was calculated by dividing the size of the overlapping region by the combined size of the two regions. Recall and precision of the recapturing RefSeq CRISPRs were evaluated for each ROA threshold.

Evaluation of performance for detecting tandem protein repeats

From the 7006 prokaryotic genome resources, we screened the positive reference set (PRS) proteins for TALE, TPR, ANK repeat and WD40 repeat families using HMMER version 3.1 with the Pfam domain signatures of PF03377, PF00515, PF00023 and PF00400, respectively. The PRS proteins for the C2H2 ZNF family were screened from the human reference genome version GRCh38.p10 using the Pfam domain signature of PF00096. Every PRS protein was required to contain three or more of the corresponding Pfam domain copies mapped with an *E*-value of less than 1.0e–10, and we obtained 331, 26 289, 4428, 2672 and 4084 PRS proteins for TALE, TPR, ANK repeats, WD40 repeats, and C2H2 ZNF, respectively (Supplementary Table S4). A total of 100 000 randomly picked prokaryotic proteins and the entire human proteome were screened for Pfam-A domain families version 31.0. Among those that do not have more than one copy of any Pfam domain with an E-value of less than 1.0e–10, we randomly selected 10 000 prokaryotic proteins and 10 000 human proteins as negative reference sets ProNRS10K and HuNRS10K. The performance of the software programs SPADE, XSTREAM and T-REKS was estimated using the recall of PRS proteins and the false positive rate (FPR) in ProNRS10K (for prokaryotic protein repeats) or HuNRS10K (for human protein repeats). The positive likelihood ratio (PLR) was calculated by dividing recall by FPR. Each software was used with its default parameters. Similar analysis was also performed by restricting the detected repeat unit sizes to within the range of expected sizes for different repeat families (34 ± 5 aa, 34 ± 5 aa, 33 ± 5 aa, 42 ± 5 aa, and 28 ± 5 aa for TALE, TPR, ANK repeats, WD40 repeats, and C2H2 ZNF, respectively). Note that, owing to the size filtering, FPRs varied for different repeat families, even when the same negative reference set was used. These measurements were also repeated with PRS proteins prepared using different criteria.

RESULTS

Overview of SPADE

We implemented SPADE to efficiently screen periodically repeating sequences as follows (Supplementary Figure S1). The software first automatically extracts multiple sequence entries from an input file (GenBank or FASTA format) and identifies the sequence type (DNA or protein) for each entry. Each entry sequence is scanned by a sliding window to count *k*-mers and highly repetitive regions are extracted. The sequence periodicity of each highly repetitive region is then evaluated based on a position-period matrix that cumulatively plots the distance between the same neighboring *k*-mers and their sequence positions (see 'Materials and Methods'). From each periodic sequence region, the periodic sequence units are queried for a multiple alignment to identify repetitive motifs. A representative motif sequence is then aligned back to the entry sequence to annotate the periodically repeating units. Finally, the annotations for the detected periodic repeats are added to the input information and output in the GenBank format with options to visualize *k*-mer density, position-period matrix, repetitive unit loci with neighboring genes, and motif sequence logo for each periodic repeat.

Periodic repeats in a CRISPR-encoding genome

Using SPADE, we exhaustively searched for periodic DNA and protein sequences in the 7006 complete prokaryotic genomes that were available in the NCBI RefSeq database. The default parameter set was used for the entire analysis of this study. In the *Streptococcus thermophilus* LMD-9 genome, 7 periodic DNA repeats and 27 periodic protein repeats were detected, including 2 previously annotated CRISPR loci (Figure 1A, Supplementary Table S1). The repeat periods of the annotated CRISPRs were both 66 bp and their detected repeat motif sequences were identical to the reported motifs (Figure 1B). We also found an unannotated interspaced repeat region containing four repeats with a period of 72 bp, in which the repeat motif and interspace sequences were all 36 bp long (Figure 1C). While type II-A Cas genes were found in the neighboring regions of the reported CRISPRs, a type III-A Cas gene cluster was found in the adjacent region of the unannotated repeat, suggesting a functional type III-A CRISPR system in this genome.

The other periodic DNA repeats were all short tandem repeats with a period size of 1–7 bp that were commonly found in prokaryotic genomes (Figure 1D). Among the 27 periodic protein repeats, 24 were short tandem repeats with periodicity of 10 aa or less. The other three included

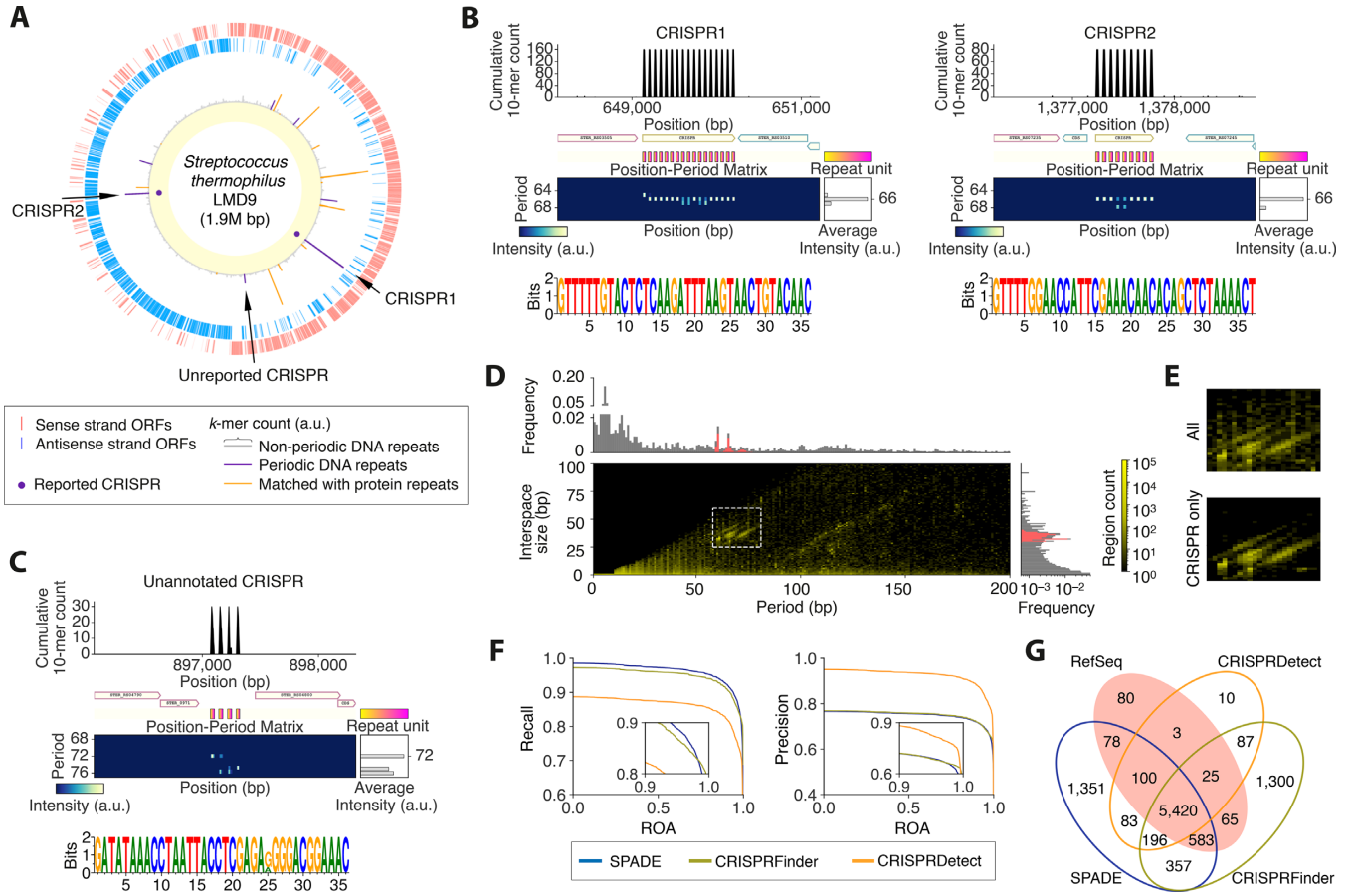

Figure 1. CRISPRs detected by SPADE. (**A**) Circular genome map of the *S. thermophilus* LMD-9 genome. From the outer side, it represents genes encoded on the sense and antisense strands of the genome, cumulative *k*-mer counts with the annotations of periodic DNA and protein repeats, and the previously reported CRISPR loci. (**B**) Previously reported CRISPRs detected by SPADE. Each periodic repeat region is visualized along with cumulative *k*-mer count, neighboring genes, positions of repeat unit sequences and position-period matrix of the surrounding genomic region, and its motif sequence is represented by sequence logo. Each periodic repeat unit is represented by a gradient box where color indicates relative position in each repeat unit sequence. (**C**) A novel CRISPR found in the *S. thermophilus* LMD-9 genome. (**D**) Period-interspace size distribution of the entire periodic DNA repeats captured in the 7006 RefSeq prokaryotic genomes. Magenta bars in the probability density distributions represent CRISPRs reported in the RefSeq dataset. Dashed white line box represents DNA repeats further screened as CRISPR candidates. (**E**) Enlarged view of the dashed white line box in (**D**) and distribution of the RefSeq CRISPRs in the same area. (**F**) Precision and recall in predicting RefSeq CRISPRs by SPADE, CRISPRFinder and CRISPRDetect along with region overlap agreement (ROA) thresholds. (**G**) Venn diagram for DNA repeats detected by SPADE, CRISPRFinder and CRISPRDetect with ROA of ≥50% and their agreement with RefSeq CRISPRs.

a peptidoglycan-binding protein (three repeats with a 17-aa period) and a subtilisin-like serine protease (three repeats with a 32-aa period), both of which were annotated to involve protein repeats, and a nucleotide exchange factor (four repeats with a 14-aa period), which was annotated to involve two α-helices.

Performance in detecting CRISPRs

We then measured the performance of SPADE in detecting CRISPRs, the annotation criteria of which are standardized in the NCBI prokaryotic genome annotation pipeline (43). From the entire 161 465 periodic DNA repeats detected in the 7006 prokaryotic genomes, we obtained 8168 genomic regions with a repeat period size and interspace size of 58–81 and 25–60 bp, respectively (Supplementary Table S2). These parameters were partly derived from CRISPRFinder, the most commonly used tool for CRISPR annotations in recent genomic resources (36,44,45), and partly defined empirically based on the reported CRISPRs (see 'Materials and Methods'). We confirmed that the distribution of period-interspace size combinations for the defined parameter space had good agreement with that for the 6354 reported CRISPRs in the RefSeq database (Figure 1D and E). We then compared the performance of SPADE with the commonly used CRISPR annotation software tools CRISPRFinder and CRISPRDetect in capturing RefSeq CRISPRs. In the same genomic datasets, 8033 regions were detected by CRISPRFinder and 5924 regions were detected by CRISPRDetect (Supplementary Table S2). Precision and recall were decreased along with region overlap agreement (ROA) with reported RefSeq CRISPR regions for all of the software tools but recalls by SPADE were the highest for ROA of up to 98% where the recalls by CRISPRDetect were markedly lower than those by the other tools (Figure 1F). On the other hand, CRISPRDetect outperformed pre-

cision of capturing RefSeq CRISPRs while the precisions of SPADE and CRISPRFinder were similar (Figure 1F).

At 50% ROA, SPADE, CRISPRFinder and CRISPRDetect captured 6181, 6093 and 5548 RefSeq CRISPR regions, respectively, where 5420 were captured by all of the software tools (Figure 1G). We defined 5548 RefSeq CRISPR regions captured by CRISPRDetect with a minimized false positive rate as a high confidence gold standard CRISPR set. Amongst this set, 5520 and 5445 regions were captured by SPADE and CRISPRFinder, respectively. Given that precisions and recalls of CRISPRFinder were higher than those by SPADE for ROA of more than 98%, we concluded that SPADE was slightly better than CRISPRFinder at roughly capturing CRISPRs, but not at the single-base resolution. In sum, although SPADE was not specifically designed for CRISPR annotation, its performance for capturing CRISPRs with simple size thresholds was at least comparable to the most commonly used CRISPR prediction software tools.

Periodic repeats in a TALE-encoding genome

In the *Xanthomonas oryzae* pv. *oryzae* (*Xoo*) PXO83 genome encoding TALE genes and TALE pseudogenes (38), 49 DNA repeats and 194 protein repeats were detected by SPADE (Figure 2A and B, Supplementary Table S3). All of the reported TALEs were recaptured with a repeating period of 34 aa and variable residues at the 12th and 13th amino acid residues and two α-helices in each repeat unit, which were all consistent with the reported features of TALE. We also detected an annotated large CRISPR locus where a highly constant motif of 31 bp was repeated periodically 86 times each with an interspace sequence of around 34 bp (Figure 2C).

Among the other 46 periodic DNA repeats, 40 were short tandem repeats with a period of 10 bp or less, including 25 heptamer repeats that were previously suggested to contribute to phase variation in the *Xanthomonas* genus (46). Three short tandem DNA repeats were found in intergenic regions, one with a period of 12 bp and two with a period of 14 bp. Another short tandem DNA repeat region was found in the middle of an ABC transporter-encoding gene with a period of 16 bp, which is relatively prime to 3, the protein coding frame size (Figure 3A), and another longer sequence with a period of 60 bp was also found to encode a hypothetical gene in less than half of its region (Figure 3B). Furthermore, we found a large periodic DNA region from the genomic position of 3 559 997 to 3 563 142 (3144 bp long) with an average period of ∼787 bp (Supplementary Figure S2). Following a transposase-encoding gene, this region involved three different hypothetical genes, each of which was in a different repeat unit. Interestingly, all of the three repeats partially overlapping with protein-coding regions were found to be widely conserved in the *Xanthomonas* genus with different numbers of repeats, but the coding gene architecture had markedly diverged evolutionarily (Figure 3C–E), indicating that phase variations of protein-coding patterns for these regions rapidly occurred after speciation by genomic contraction and expansion via the repetitive sequences.

Except for TALEs, ∼88% (156 out of 178) of protein repeats were composed of short tandem repeats with a repeat unit size of 10 aa or less (Supplementary Table S3). The other repetitive proteins included three chemotaxis-associated proteins with different periods of 27, 46 and 90 aa, a DNA topoisomerase I, a TolB-like protein known to involve non-WD40 β-propellers, and transporters and a hypothetical protein involving six repeats with a large unit size of 215 aa. Notably, another type III secretion system effector protein of the *Xanthomonas* host infection process was found to have repetitive peptide units, suggesting another function of pathogenic periodic protein structure in hijacking the host plant system (Supplementary Figure S3).

Performance in detecting TALEs and C2H2 ZFNs

As C2H2 ZNFs are the most widely used transcription factors in the human genome, we also examined whether SPADE can capture human C2H2 ZNFs. When a C2H2 ZNF encoded on human chromosome 7p22.1 was scanned by SPADE, 20 degenerative repeats of ∼28 aa were detected with two cysteine and two histidine residues conserved at specific positions, like typically reported C2H2 ZNFs (Figure 4). We then assessed the performance of SPADE in detecting TALEs and C2H2 ZNFs. Using the protein domain search software HMMER, we obtained positive reference sets (PRSs) for TALE and human C2H2 ZNF from the prokaryotic genomic dataset and the human proteome, respectively, so each PRS protein contained three or more of the corresponding Pfam motifs (see 'Materials and Methods'). We also prepared 10 000 prokaryotic proteins and 10 000 human proteins that did not have any Pfam motif more than once as negative reference sets (NRSs) ProNRS10K and HuNRS10K, respectively (see 'Materials and Methods'). Using SPADE, repetitive sequences of any period were detected in 328 out of 331 TALE PRS proteins (99.1%) and 3079 out of 4084 human C2H2 ZNF PRS proteins (75.4%), while 192 ProNRS10K proteins (1.9%) and 1269 HuNRS10K proteins (12.7%) were positive (Figure 5A, Supplementary Table S4). When the detected positives were filtered by maximum repeat unit size per protein (maxRUSPP) to be within ±5 aa from the expected average repeat unit size (34 aa for TALE and 28 aa for C2H2 ZNF), the recall of TALE PRS stayed the same (99.1%) and the recall of human C2H2 ZNF PRS was 58.9%, while the false positive rate (FPR) of TALE estimated using ProNRS10K and the FPR of C2H2 ZNF estimated using HuNRS10K were greatly decreased to 0.03% and 0.21%, respectively (Figure 5B and C). This simple size limitation improved positive likelihood ratios (PLRs) of the prediction from 51.6 to 3303.1 (64.0-fold) for TALE and from 5.9 to 280.7 for human C2H2 ZNF (47.2-fold).

Comparison with other software capturing tandem protein repeats

SPADE successfully detected the other degenerate tandem protein repeats widely spread in prokaryotes, including TPRs, ANK repeats, and WD40 repeats (Figure 5D–F). The secondary structure prediction of these degenerated repeats also properly captured their reported structural

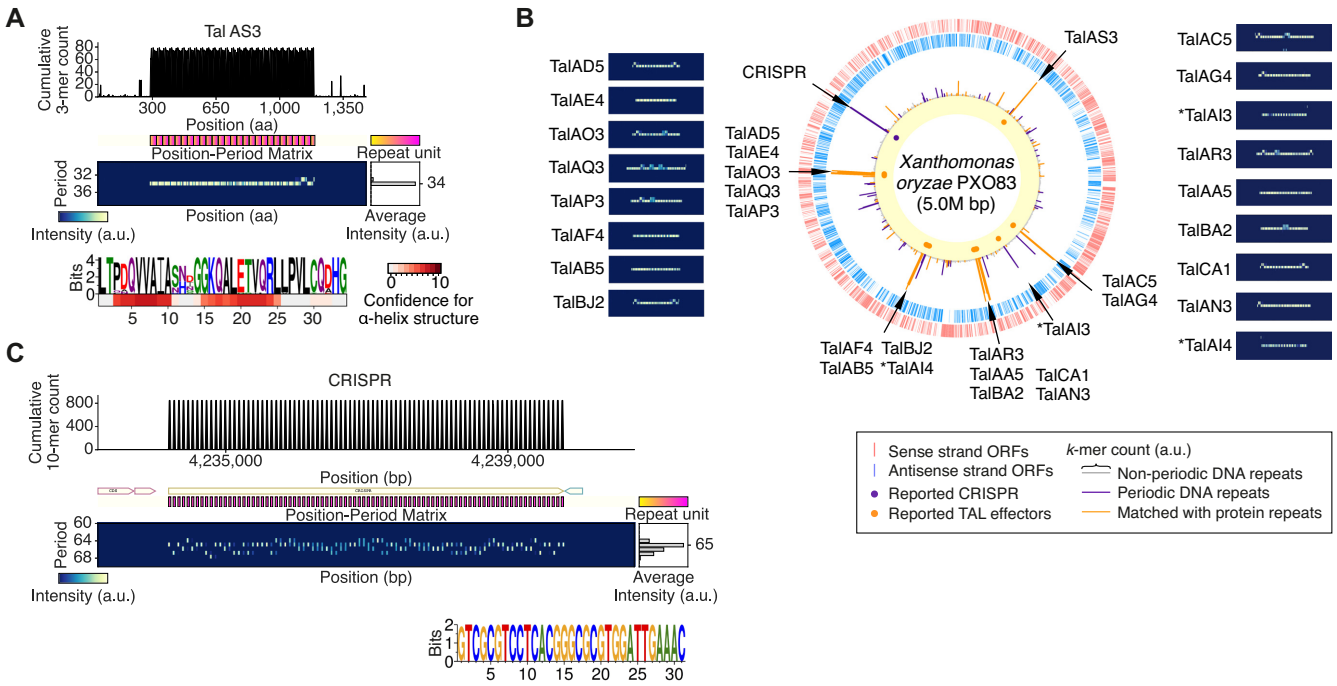

Figure 2. TALEs and a CRISPR detected in the *Xoo* PXO83 genome. (**A**) Example (TalAS3) of TALEs detected by SPADE. For details, see Figure 1B. The heat map under the repeat motif sequence logo represents confidence scores for α-helical structure at each amino acid residue. (**B**) Circular map representation of the *Xoo* PXO83 genome along with position-period matrices for all of the TALE genes annotated in the RefSeq database. '*' denote the ones annotated as TALE pseudogenes by AnnoTALE. For details, Figure 1A. (**C**) CRISPR locus of the *Xoo* PXO83 genome detected by SPADE.

motifs. In each of the repeat sequence motifs identified for a TPR and an ANK repeat, both of which have been reported to have helix–turn–helix structures, we observed two α-helical loops (Figure 5D and E). Four β-strands were also captured in a repeat sequence motif of WD40, consistent with its β-propeller structure (Figure 5F). As XSTREAM (34) and T-REKS (35) have been widely used to explore tandem protein repeats in an unsupervised manner in recent studies (47,48), we next performed a benchmark comparison of SPADE, XSTREAM, and T-REKS in detecting TPRs, ANK repeats, and WD40 repeats, in addition to TALEs and human C2H2 ZNFs. For TPRs, ANK repeats, and WD40 repeats, PRSs were prepared as described above for TALE. ProNRS10K and HuNRS10K were again used as NRSs for detecting repeats in prokaryotic and human protein families, respectively.

T-REKS performed the best in recall for detecting repetitive sequences regardless of repeat unit size, except for WD40, in which SPADE performed the best (Figure 5A, Supplementary Table S4). However, T-REKS also demonstrated the highest FPRs in both ProNRS10K and HuNRS10K datasets. When the overall prediction performance was estimated by PLR, SPADE performed the best in every repeat type (between 1.02-fold and 3.39-fold compared with the second-best software XSTREAM for all repeat types). We also found that the maxRUSPPs detected by SPADE were distributed with peaks at 34, 33 and 42 aa for TPR, ANK repeats, and WD40 repeats, respectively, all of which were the reported typical unit sizes for these protein repeats (Figure 5B). This was not the case for all of the repeats detected by XSTREAM and T-REKS. XSTREAM captured wider ranges of repeat unit sizes for every repeat type and T-REKS tended to capture shorter tandem repeats for the subpopulation of positive reference proteins for TPRs, ANK repeats, and WD40 repeats. Filtering the detected positives by maxRUSPP to be within ±5 aa from the expected average repeat sizes, the recall performance of SPADE was the best for all repeat types, whereas the FPRs of the three software packages were all minimized to below 0.005 in all of the repeat types (Figure 5C, Supplementary Table S4). (Note that the performances could not be compared using PLR as many FPRs for different protein families were zero.) These observations were maintained when the positive reference protein sets were prepared differently (Supplementary Figure S4).

DISCUSSION

No software program have been developed that can universally screen for periodic DNA and protein repeats; the only available software tools are those that screen for reported motifs or certain types of periodic repeats. Nevertheless, the performance of SPADE capturing CRISPRs was on par with the commonly used CRISPR prediction software tools and outperformed XSTREAM and T-REKS in the sensitivity for capturing various tandem protein repeats, regardless of the degree of consensus in the repeat unit motifs. SPADE also captured TALEs and ZNFs in a highly specific and unsupervised manner, indicating its potential to contribute to the discovery of new genome editing modules from large genomic and/or metagenomic resources. This is supported by the fact that we found that a non-TALE

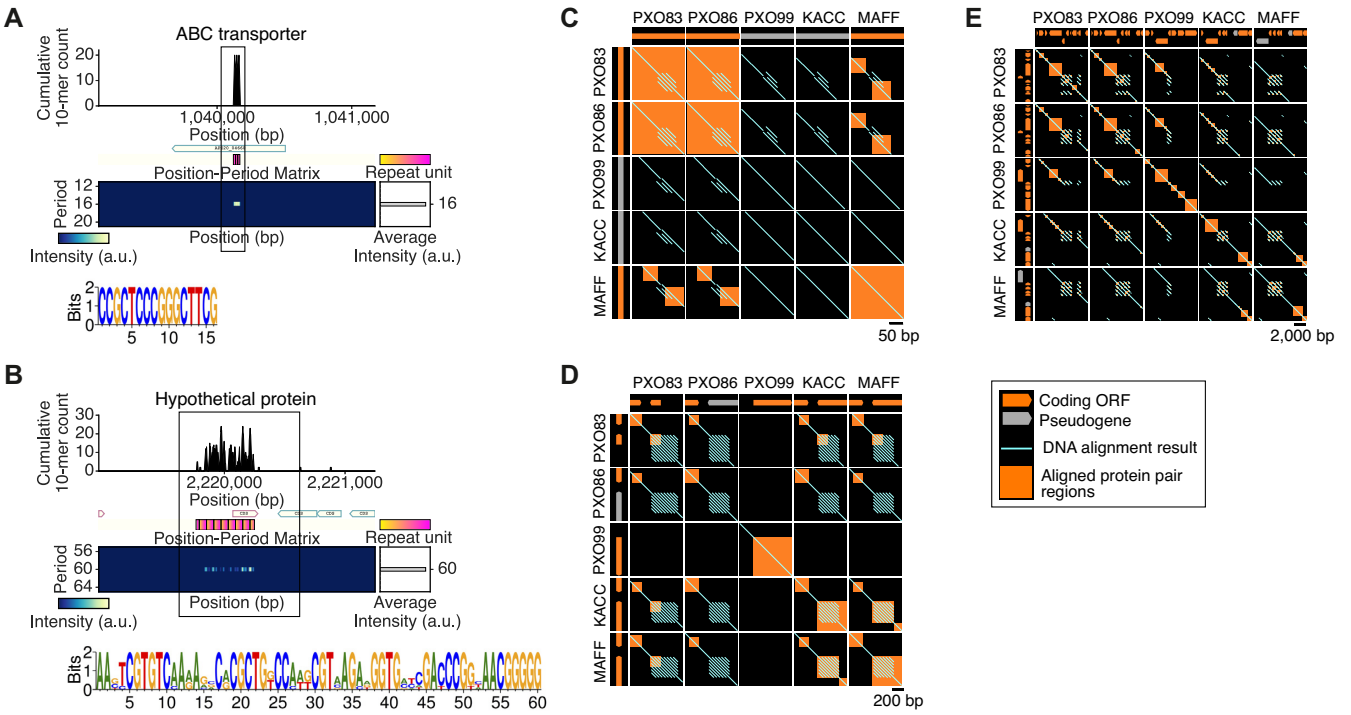

Figure 3. Comparative analysis of the repetitive DNA regions determined to overlap with protein-coding regions in the *Xanthomonas* genus. (**A**) An ABC transporter-coding region involving a DNA repeat of 16-bp period. Line box denotes the genomic regions used for the comparative genomic analysis presented in (**C**). (**B**) A hypothetical protein-coding region overlapped with a DNA repeat of 60-bp period. Line box denotes the genomic regions used for the comparative genomic analysis presented in (D). (C–E) Sequence alignment analysis of the repetitive DNA sequences and their protein-coding patterns amongst the *Xoo* PXO83, PXO86, PXO99, KACC10331, and MAFF311018 genomes. With protein/pseudogene-coding structures under the label of each genome, the within- and between-genome alignment results are represented by light blue lines for DNA and orange boxes for proteins, respectively. (**C**) For an ABC transporter-coding region in the *Xoo* PXO83 genome. (**D**) For a hypothetical protein-coding region overlapping with a DNA repeat of 60-bp period in the *Xoo* PXO83 genome. (**E**) For a large DNA repeat region with a repeat unit size of 787 that overlaps with three different hypothetical genes in the *Xoo* PXO83 genome.

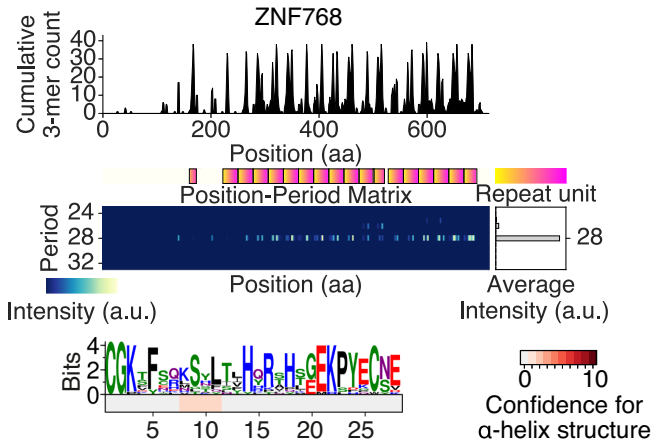

Figure 4. Periodic features of human ZNF768 detected by SPADE.

type III secretion system protein of *Xanthomonas* host infection machinery had periodic repeats like TALEs and ZFNs (Supplementary Figure S3). We also captured bacterial homologs of pentatricopeptide repeats (PPRs) that are involved in translational regulation in plants (Supplementary Figure S5). As the binding code of PPR to RNA has recently been deciphered, it has been suggested as a potential programmable RNA editing and regulating module (49).

The majority of the periodic repeats detected in the 7006 prokaryotic genomes still need further investigation. We detected many short tandem DNA and proteins repeats. In particular, tandem heptamer DNA repeats were the most abundant in intergenic regions of a wide range of prokaryotic species (Figure 1D). However, there has been no clear clue about the function of this globally existing prime number periodicity in genomic DNA. We also found various interspaced repeats that had clear sequence periodicities with no CRISPR annotation or neighboring Cas gene. They included many tRNA operons in various prokaryotes, as reported previously (Supplementary Figure S6), but the others remain to be explored. Genomic expansion and contraction have been thought to occur at the tandem repeat sequences, leading to phase variation. Even after excluding corresponding protein repeats, the repeat periods of both tandem and interspaced DNA repeats showed particular abundance for these in multiples of three. Furthermore, some genes were encoded in part of a repeat unit of a large tandem repeat region (Supplementary Figure S2). As seen in the *Xanthomonas* genus (Figure 3E), these findings suggest the roles of tandem repeats in *de novo* gene birth or gene death. We also found many tandem DNA repeats within (or partially within) protein-coding regions, some of which

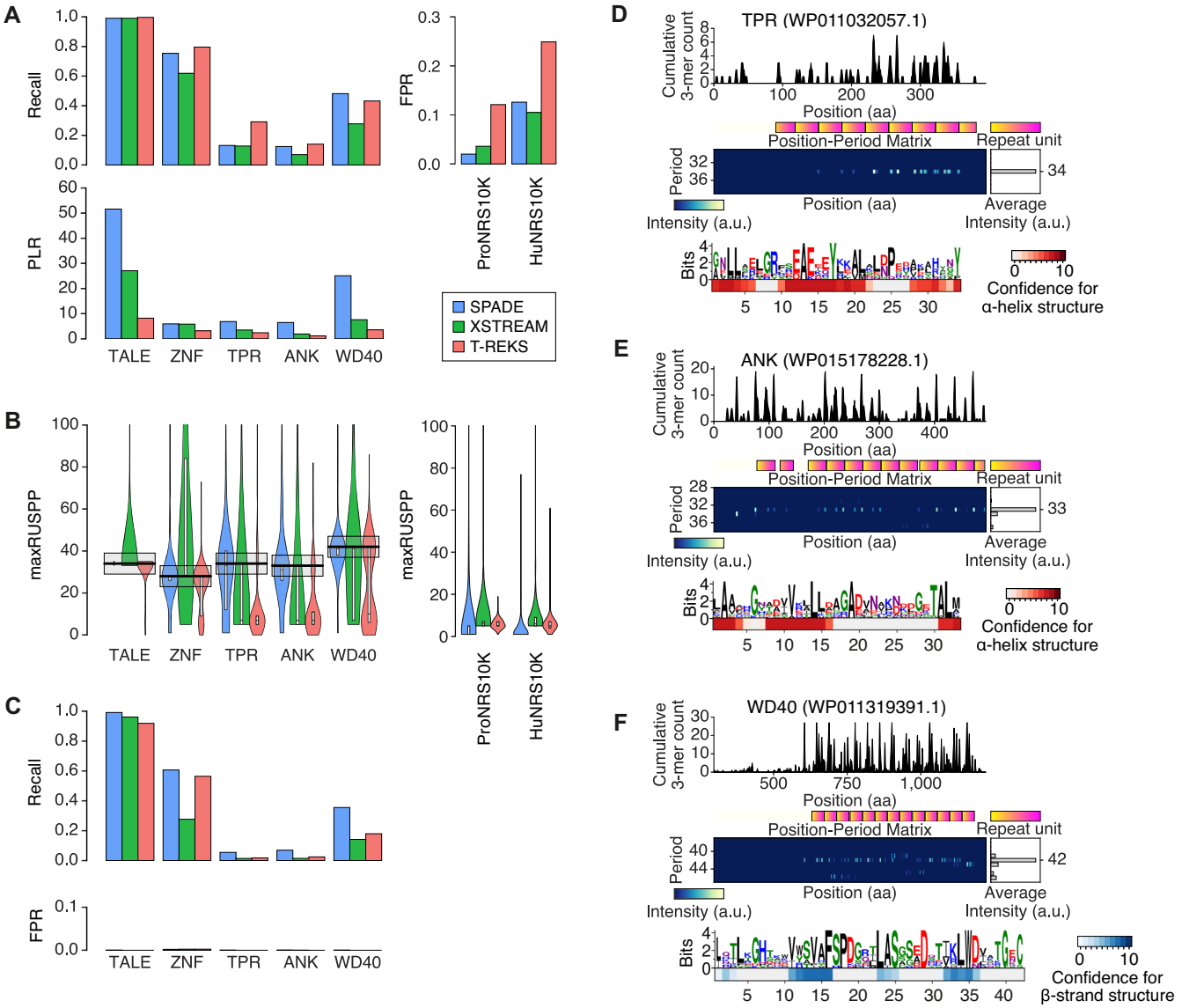

Figure 5. Comparison of the performance of SPADE, XSTREAM, and T-REKS to detect tandem degenerate protein repeats. (**A**) Recalls, false positive rates (FPRs), and positive likelihood ratios (PLRs) of SPADE, XSTREAM, and T-REKS in capturing TALEs, ZNFs, TPRs, ANK repeats, and WD40 repeats. (**B**) Distribution of maximum repeat unit sizes per protein (maxRUSPPs) detected by each software for each protein family. Each bold black bar and each gray box denote the reported typical repeat unit size for the corresponding protein category and the ±5 aa range from the reported repeat unit size, respectively. (**C**) Recalls and FPRs of the different software after filtering maxRUSPPs of the detected positives to be within ±5 aa from the expected repeat unit size. From the maxRUSPP distributions of ProNRS10K (negative control for prokaryotic protein repeat families) and HuNRS10K (negative control for ZNFs), FPRs for different protein repeats of different expected repeat unit sizes were estimated. (D–F) Example protein repeats detected by SPADE for TPR (**D**), ANK repeat (**E**), and WD40 repeat (**F**). The heat map under each repeat motif sequence logo represents the confidence scores for α-helical structure (red heat map) or β-sheet structure (blue heat map) at each amino acid residue position.

were indicated to have contributed to functional phase variation of protein-coding patterns (Figure 3A–D).

The default parameter set of SPADE robustly captured most of the important biological sequences tested in this study with higher precision than did the other software. We further analyzed various kinds of simulated repeats using SPADE and confirmed that the default *k*-mer parameters for DNAs and proteins performed the best and precisely captured the periodicities of various degenerate tandem and interspaced repeats with up to ∼10% and ∼30% sequence noise for DNA and proteins, respectively, regardless of the repeat unit size and interspace size (Supplementary Figure S7).

SPADE was implemented using Python and can be executed with MAFFT and BLAST on MacOS X and Linux operating systems, and on Windows Subsystem for Linux (WSL) of Windows 10. It automatically detects the input file and sequence types and outputs results in the GenBank file format, which can be further input into other software programs, with various visualizations as represented in the figures. Accordingly, here we propose that SPADE is fast and user-friendly software based on a simple algo-

rithm to globally capture periodic biomolecular sequences. Although we mainly focused on measuring the performance of this software predominantly using prokaryotic genomes in this study, further wide-ranging investigations of these periodically repeating sequences together with screening of eukaryotic and metagenomic resources could lead to the discovery of new biological events and genome editing tools.

DATA AVAILABILITY

SPADE is open-source software under GNU General Public License v3.0, which is available at https://github.com/yachielab/SPADE.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the members of the Yachie laboratory at The University of Tokyo and Keio University for useful comments and discussions throughout the course of this study. We also sincerely appreciate Hirotada Mori for sharing his laboratory space at the Nara Institute of Science and Technology (NAIST).

FUNDING

New Energy and Industrial Technology Development Organization (NEDO) Genome Editing Program; Japan Society for the Promotion of Science (JSPS) 18K19777 (to N.Y.); Japan Science and Technology Agency (JST) PRESTO program 10814 (to N.Y.); Japan Agency for Medical Research and Development (AMED) PRIME program 17gm6110007 (to N.Y.); The Naito Foundation (to N.Y.); The Nakajima Foundation (to N.Y.); The Takeda Foundation (to N.Y.); SECOM Science and Technology Foundation (to N.Y.); TTCK fellowships (to H.M., D.E.-Y., S.I.); Mori Memorial Foundation (to H.M.); Yamagishi Student Project Support Program (to D.E.-Y.) of Keio University; JSPS DC1 Fellowship (to S.I.). Funding for open access charge: Research Budget.
Conflict of interest statement. None declared.

REFERENCES

1. Kazazian,H.H. Jr (2004) Mobile Elements: Drivers of genome evolution. *Science*, **303**, 1626–1632.
2. Levin,H.L. and Moran,J.V. (2011) Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.*, **12**, 615–627.
3. Zhou,K., Aertsen,A. and Michiels,C.W. (2014) The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol. Rev.*, **38**, 119–141.
4. Bichara,M., Wagner,J. and Lambert,I.B. (2006) Mechanisms of tandem repeat instability in bacteria. *Mutat. Res.*, **598**, 144–163.
5. Henderson,I.R., Owen,P. and Nataro,J.P. (1999) Molecular switches — the ON and OFF of bacterial phase variation. *Mol. Microbiol.*, **33**, 919–932.
6. D'Andrea,L.D. and Regan,L. (2003) TPR proteins: the versatile helix. *Trends Biochem. Sci.*, **28**, 655–662.
7. Li,J., Mahajan,A. and Tsai,M.-D. (2006) Ankyrin Repeat: A unique motif mediating Protein−Protein interactions. *Biochemistry*, **45**, 15168–15178.
8. Stirnimann,C.U., Petsalaki,E., Russell,R.B. and Muller,C.W. (2010) WD40 proteins propel cellular networks. *Trends Biochem. Sci.*, **35**, 565–574.
9. Fimia,G.M., Stoykova,A., Romagnoli,A., Giunta,L., Di Bartolomeo,S., Nardacci,R., Corazzari,M., Fuoco,C., Ucar,A., Schwartz,P. *et al.* (2007) Ambra1 regulates autophagy and development of the nervous system. *Nature*, **447**, 1121–1125.
10. Main,E.R.G., Xiong,Y., Cocco,M.J., D'Andrea,L. and Regan,L. (2003) Design of stable α-Helical arrays from an idealized TPR *Motif*. *Structure*, **11**, 497–508.
11. Binz,H.K., Amstutz,P., Kohl,A., Stumpp,M.T., Briand,C., Forrer,P., Grutter,M.G. and Pluckthun,A. (2004) High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol.*, **22**, 575–582.
12. Voet,A.R.D., Noguchi,H., Addy,C., Simoncini,D., Terada,D., Unzai,S., Park,S.-Y., Zhang,K.Y.J. and Tame,J.R.H. (2014) Computational design of a self-assembling symmetrical β-propeller protein. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 15102.
13. Urnov,F.D., Rebar,E.J., Holmes,M.C., Zhang,H.S. and Gregory,P.D. (2010) Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.*, **11**, 636–646.
14. Cermak,T., Doyle,E.L., Christian,M., Wang,L., Zhang,Y., Schmidt,C., Baller,J.A., Somia,N.V., Bogdanove,A.J. and Voytas,D.F. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.*, **39**, e82.
15. Tupler,R., Perini,G. and Green,M.R. (2001) Expressing the human genome. *Nature*, **409**, 832.
16. Scot,A.W., Lena Nekludova,A. and Carl,O.P. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 183–212.
17. Boch,J. and Bonas,U. (2010) Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.*, **48**, 419–436.
18. Gordley,R.M., Gersbach,C.A. and Barbas,C.F. (2009) Synthesis of programmable integrases. *Proc Natl. Acad. Sci. U.S.A.*, **106**, 5053–5058.
19. Schirrmeister,B.E., Antonelli,A. and Bagheri,H.C. (2011) The origin of multicellularity in cyanobacteria. *BMC Evol. Biol.*, **11**, 45.
20. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
21. Makarova,K.S., Wolf,Y.I., Alkhnbashi,O.S., Costa,F., Shah,S.A., Saunders,S.J., Barrangou,R., Brouns,S.J., Charpentier,E., Haft,D.H. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **13**, 722–736.
22. Marraffini,L.A. and Sontheimer,E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.*, **11**, 181–190.
23. Kunne,T., Kieper,S.N., Bannenberg,J.W., Vogel,A.I., Miellet,W.R., Klein,M., Depken,M., Suarez-Diez,M. and Brouns,S.J. (2016) Cas3-Derived target DNA degradation fragments fuel primed CRISPR adaptation. *Mol. Cell*, **63**, 852–864.
24. Shipman,S.L., Nivala,J., Macklis,J.D. and Church,G.M. (2016) Molecular recordings by directed CRISPR spacer acquisition. *Science*, **353**, aaf1175.
25. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–1183.
26. Konermann,S., Brigham,M.D., Trevino,A.E., Joung,J., Abudayyeh,O.O., Barcena,C., Hsu,P.D., Habib,N., Gootenberg,J.S., Nishimasu,H. *et al.* (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, **517**, 583–588.
27. Nishida,K., Arazoe,T., Yachie,N., Banno,S., Kakimoto,M., Tabata,M., Mochizuki,M., Miyabe,A., Araki,M., Hara,K.Y. *et al.* (2016) Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science*, **353**, aaf8729.
28. Chen,B., Gilbert,L.A., Cimini,B.A., Schnitzbauer,J., Zhang,W., Li,G.W., Park,J., Blackburn,E.H., Weissman,J.S., Qi,L.S. *et al.* (2013) Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, **155**, 1479–1491.

29. Chen,N. (2004) Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinformatics*, **5**, doi:10.1002/0471250953.bi0410s05.
30. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
31. Bao,Z. and Eddy,S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.*, **12**, 1269–1276.
32. Price,A.L., Jones,N.C. and Pevzner,P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21**, i351–i358.
33. Chen,G.L., Chang,Y.J. and Hsueh,C.H. (2013) PRAP: an ab initio software package for automated genome-wide analysis of DNA repeats for prokaryotes. *Bioinformatics*, **29**, 2683–2689.
34. Newman,A.M. and Cooper,J.B. (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8**, 382.
35. Jorda,J. and Kajava,A.V. (2009) T-REKS: identification of tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics*, **25**, 2632–2638.
36. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.
37. Biswas,A., Staals,R.H., Morales,S.E., Fineran,P.C. and Brown,C.M. (2016) CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics*, **17**, 356.
38. Grau,J., Reschke,M., Erkes,A., Streubel,J., Morgan,R.D., Wilson,G.G., Koebnik,R. and Boch,J. (2016) AnnoTALE: bioinformatics tools for identification, annotation, and nomenclature of TALEs from Xanthomonas genomic sequences. *Sci. Rep.*, **6**, 21077.
39. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
40. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
41. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
42. Buchan,D.W., Minneci,F., Nugent,T.C., Bryson,K. and Jones,D.T. (2013) Scalable web services for the PSIPRED protein analysis workbench. *Nucleic Acids Res.*, **41**, W349–W357.
43. Tatusova,T., DiCuccio,M., Badretdin,A., Chetvernin,V., Nawrocki,E.P., Zaslavsky,L., Lomsadze,A., Pruitt,K.D., Borodovsky,M. and Ostell,J. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, **44**, 6614–6624.
44. Minot,S., Sinha,R., Chen,J., Li,H., Keilbaugh,S.A., Wu,G.D., Lewis,J.D. and Bushman,F.D. (2011) The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.*, **21**, 1616–1625.
45. Mason,O.U., Hazen,T.C., Borglin,S., Chain,P.S., Dubinsky,E.A., Fortney,J.L., Han,J., Holman,H.Y., Hultman,J., Lamendella,R. *et al.* (2012) Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J.*, **6**, 1715–1727.
46. Rehm,C., Wurmthaler,L.A., Li,Y., Frickey,T. and Hartig,J.S. (2015) Investigation of a quadruplex-forming repeat sequence highly enriched in xanthomonas and nostoc sp. *PLoS One*, **10**, e0144275.
47. Sędzielewska Toro,K. and Brachmann,A. (2016) The effector candidate repertoire of the arbuscular mycorrhizal fungus Rhizophagus clarus. *BMC Genomics*, **17**, 101.
48. Mackinder,L.C.M., Meyer,M.T., Mettler-Altmann,T., Chen,V.K., Mitchell,M.C., Caspari,O., Freeman Rosenzweig,E.S., Pallesen,L., Reeves,G., Itakura,A. *et al.* (2016) A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 5958.
49. Kobayashi,K., Kawabata,M., Hisano,K., Kazama,T., Matsuoka,K., Sugita,M. and Nakamura,T. (2012) Identification and characterization of the RNA binding surface of the pentatricopeptide repeat protein. *Nucleic Acids Res.*, **40**, 2712–2723.