Biophysical Journal
Article

Biophysical Society

Coevolutionary Couplings Unravel PAM-Proximal Constraints of CRISPR-SpCas9

Yi Li,[1,2] José A. De la Paz,[3] Xianli Jiang,[3] Richard Liu,[1] Adarsha P. Pokkulandra,[4] Leonidas Bleris,[1,2,3,*] and Faruck Morcos[1,2,3,*]

[1]Department of Bioengineering, [2]Center for Systems Biology, [3]Department of Biological Sciences, and [4]School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, Texas

ABSTRACT The clustered regularly interspaced short palindromic repeats (CRISPR) system, an immune system analog found in prokaryotes, allows a single-guide RNA to direct a CRISPR-associated protein (Cas) with combined helicase and nuclease activity to DNA. The presence of a specific protospacer adjacent motif (PAM) next to the DNA target site plays a crucial role in determining both efficacy and specificity of gene editing. Herein, we introduce a coevolutionary framework to computationally unveil nonobvious molecular interactions in CRISPR systems and experimentally probe their functional role. Specifically, we use direct coupling analysis, a statistical inference framework used to infer direct coevolutionary couplings, in the context of protein/nucleic acid interactions. Applied to *Streptococcus pyogenes* Cas9, a Hamiltonian metric obtained from coevolutionary relationships reveals, to our knowledge, novel PAM-proximal nucleotide preferences at the seventh position of *S. pyogenes* Cas9 PAM (5′-NGRNNNT-3′), which was experimentally confirmed by in vitro and functional assays in human cells. We show that coevolved and conserved interactions point to specific clues toward rationally engineering new generations of Cas9 systems and may eventually help decipher the diversity of this family of proteins.

SIGNIFICANCE The CRISPR/*Streptococcus pyogenes* Cas9 system, which consists of *S. pyogenes* Cas9 protein and a synthetized guide RNA, has revolutionized the field of genome engineering in the past decade. Assays based on directed evolution have been used to improve the system, but these screening-based approaches are inefficient and labor intensive. Further optimization of this technology in both efficiency and specificity using rational engineering is paramount. Herein, we apply a statistical inference framework, direct coupling analysis, to computationally probe nonobvious molecular interactions in CRISPRsystems, and we experimentally confirm our results.

INTRODUCTION

The CRISPR (clustered regularly interspaced short palindromic repeats) system, an immune response mechanism against bacteriophage infections in bacteria and archaea, has revolutionized the field of genome editing and spurred myriads of applications critically relevant to agriculture, biomanufacturing, and human health (1–8). The CRISPR constituents include Cas9, an endonuclease protein, and a single-guide RNA (sgRNA or CRISPR RNA and trans-activating CRISPR RNA in native systems). A prominent property of the CRISPR mechanism is that for Cas9 to bind to DNA, the target sequence in the genomic DNA must be immediately followed by a Cas9-specific protospacer adjacent motif (PAM) (9). As an example, the presence of a 5′-NG(A/G)-3′ PAM is essential for editing using the Cas9 protein variant from *Streptococcus pyogenes* (SpCas9).

Direct coupling analysis (DCA) (10), a global inference method, generates a joint probability distribution over the sequence space of proteins and nucleic acids. By estimating parameters, from multiple sequence alignments (MSAs), describing both the local fields and pairwise interactions (couplings) of this distribution, it is possible to disentangle molecular connectivity within amino acids in proteins or across molecular complexes. Previous studies have shown that coevolutionary coupling parameters could reveal cooperative effects between amino acid residues and thus accurately predict protein structure (11–16), dynamics (17–21),

Submitted June 5, 2019, and accepted for publication September 30, 2019.

*Correspondence: bleris@utdallas.edu or faruckm@utdallas.edu

Editor: Monika Fuxreiter.

https://doi.org/10.1016/j.bpj.2019.09.040

and protein-protein interactions (22–28) as well as mutational landscapes for both proteins and nucleic acids (29–33).

In this report, we investigated proteins and nucleic acid sequences in the CRISPR-SpCas9 family to unravel their properties and discover critical functional components based on coevolutionary relationships. More specifically, we extended the application of DCA to infer important interactions between amino acids and nucleotides and identified residues within the PI (PAM-interacting) domain of protein SpCas9, which may be critical in PAM recognition and interaction. We further analyzed the Hamiltonian of the global probability distribution (the collection of parameters) of amino-acid-nucleotide interactions on the CRISPR/SpCas9 system to uncover novel constrains of PAM sequences for the integrity of the CRISPR functionality.

MATERIALS AND METHODS

DCA and Hamiltonian analysis

We consider a modified version of the DCA implementation described in (10) by Morcos et al. The input sequences for DCA (273 sites in total) consist of the homologous sequences of the PI domain (positions 1–263) and the putative PAM nucleotide sequences (positions 264–273). The sequences were arranged in M-rows matrix ($M = 406$) as follows:

$$A = (A)_i^a; \qquad a = 1, \cdots, M; \qquad i = 1, \cdots, L, \tag{1}$$

and coded using a 21-letter alphabet (numbers from 1 to 21) accounting for the 20 amino acid and an extra symbol for alignment gaps in the PI domains (positions 1–263), in addition to a four-letter alphabet accounting for four nucleotide species for the PAM sequences (positions 264–273). Both single-site and pair weighted frequency counts were empirically determined according to the expressions as follows:

$$f_i(x) = \left(M_{eff} + \lambda\right)^{-1}\left(\lambda / q + \sum_{\alpha=1}^{M} \delta_{X,A_i^\alpha} / m^\alpha\right) \text{and}$$

$$f_{ij}(x,y) = \left(M_{eff} + \lambda\right)^{-1}\left(\lambda / q^2 + \sum_{\alpha=1}^{M} \delta_{X,A_i^\alpha}\delta_{Y,A_j^\alpha} / m^\alpha\right), \tag{2}$$

where q accounts for the alphabet size, and δ_{X,A_i^α} denotes the Kronecker symbol. A pseudocount λ is considered to guarantee finite couplings (10,34). m^α is a bias-correcting term based on sequence similarity defined in (10), and the effective number sequence M_{eff} is the sum of inverses of such weights with the sequence identity threshold θ as 0.1.

Subsequently, a global probability distribution was derived from the maximal entropy principle (10,35) constrained to have marginal probabilities corresponding to the weighted frequency counts introduced in Eq. 2, yielding a mathematical form similar to a Boltzmann distribution, $P(\vec{x}) \propto exp(-H(\vec{x}))$, where the vector $\vec{x}$ represents a specific sequence and the Hamiltonian function is defined as follows:

$$H(\vec{x}) = -\sum_{1 \le i < j \le L} e_{ij}(x_i, x_j) - \sum_{1 \le i \le L} h_i(x_i). \tag{3}$$

Here, the $e_{ij}(x_i, x_j)$ coupling describes the interaction between positions i and j with corresponding residues or bases x_i and x_j, respectively. Similarly, the local field $h_i(x_i)$ describes single-position contribution from position i, which encodes residue x_i. Coupling parameters were inferred through a mean-field approximation. For each pair of sites, weighted according to their derived couplings, a message-passing algorithm was used to infer local fields to match marginal probabilities to the empirical single-site frequencies of a given pair. Afterward, all these local fields are averaged to obtain L approximate values for $h_i(x_i)$ (10). Because the exponential in the probability distribution is a monotonic function, we subsequently focused directly on the Hamiltonian values and proposed that the more negative a Hamiltonian value is, the more favorable the interactions between the amino acid sequence and the PAM nucleotide sequence. To study the global sequence configuration of the PAM sequences for SpCas9, we introduced an interaction-Hamiltonian distribution by fixing the protein component of the sequence (positions 1–263) to SpCas9 and considered variations only on the PAM sequences (positions 264–273) as follows:

$$H_i = h_{N+i}(r) + \sum_{j=1}^{N} e_{i+N,j}\left(r, (Cas9)_j\right) + \sum_{j>N, j\neq i+N,\ a\in\{A,C,G,T\}} e_{i+N,j}(r,a). \tag{4}$$

For the first term (local fields for the PAM sequence), the index i represents the position in the PAM sequence, and r represents the corresponding nucleotide at that position. For the second term (couplings between the SpCas9 amino acid residues and the PAM nucleotides), N represents the length of SpCas9 protein sequence ($N = 263$), and j represents the position of the amino acid residue within the SpCas9 sequence. For the third term (couplings between two different PAM nucleotides), j represents the position of the nucleotide within the PAM sequence. The sum of the three terms will yield a global functional landscape for the SpCas9-PAM interactions, accounting for both single-site and pairwise contributions in a given sequence.

Cell culture and transient transfection

HEK293 cells were acquired from the American Type Culture Collection (catalog number CRL-1573; Manassas, VA) and maintained at 37°C, 100% humidity, and 5% CO_2. The cells were grown in Dulbecco's modified Eagle medium (catalog number 11965–1181; Invitrogen, Carlsbad, CA) supplemented with 10% fetal bovine serum (catalog number 26140; Invitrogen), 0.1 mM minimum essential medium nonessential amino acids (catalog number: 11140–050; Invitrogen), and 0.045 units/mL of penicillin and 0.045 units/mL of streptomycin (penicillin-streptomycin liquid; catalog number 15140; Invitrogen). To pass the cells, the adherent culture was first washed with phosphate buffered saline (Dulbecco's; Mediatech, Manassas, VA; catalog number 21-030-CM), then trypsinized with trypsin-EDTA (0.25% trypsin with EDTAX4Na; catalog number 25200; Invitrogen), and finally diluted in fresh medium.

For transient transfection, ~300,000 cells in 1 mL of complete medium were plated into each well of 12-well culture-treated plastic plates (catalog number 665180; Griener Bio-One, Kremsmünster, Austria) and grown for 16–20 h. All transfections were then performed using 1.75 μL of jetPRIME (Polyplus Transfection, Illkirch-Graffenstaden, France) and 75μL of jetPRIME buffer. The transfection mixture was then applied to the cells and mixed with the medium by gentle shaking.

In vitro cleavage assay using SpCas9

The sgRNAs were synthesized in vitro (target A: primer P17; target B: primer P18; target C: primer P19) using the EnGen sgRNA Synthesis Kit

(catalog number E3322S; New England Biolabs, Ipswich, MA) and subsequently purified using the Monarch RNA Cleanup Kit (catalog number T2040; New England Biolabs). For in vitro SpCas9 cleavage assay, purified SpCas9 (catalog number M0386; New England Biolabs) was first incubated with sgRNA at room temperature for 10 min to form the CRISPR complex. Next, DNA targets were added to the CRISPR reaction mixtures at a ratio of CRISPR:DNA (mol:mol) = 2:1 and incubated at 37°C for 15 min. The reaction mixtures were then treated with 1 μL Proteinase K (catalog number P8107S; New England Biolabs) at room temperature for 10 min. Subsequently, the digested products were subjected to agarose gel electrophoresis. The assay was repeated three times, and the pixel counts of the 997-bp and 702-bp bands were extracted for quantification.

Fluorescence microscopy

Microscopy was performed 48 h posttransfection. The live cells were grown on 12-well plates (Greiner Bio-One) in the complete medium. Cells were imaged using an Olympus IX81 microscope (Tokyo, Japan) in a precision-controlled environmental chamber. The images were captured using a Hamamatsu ORCA-03 cooled monochrome digital camera (Hamamatsu, Japan). The filter sets (Chroma Technology, Bellows Falls, VT) are as follows: ET436/20× (excitation) and ET480/40 m (emission) for cyan fluorescent protein (CFP) and ET560/40× (excitation) and ET630/75 m (emission) for mKate. Data collection and processing was performed in the software package Slidebook 5.0. All images within a given experimental set were collected with the same exposure times and underwent identical processing.

Flow cytometry

48 h posttransfection, cells from each well of the 12-well plates were trypsinized with 0.1 mL 0.25% trypsin-EDTA at 37°C for 3 min. Trypsin-EDTA was then neutralized by adding 0.9 mL of complete medium. The cell suspension was centrifuged at 1000 rotations per minute for 5 min, and after removal of supernatants, the cell pellets were resuspended in 0.5 mL phosphate buffered saline buffer. The cells were analyzed on a BD Biosciences LSRFortessa flow analyzer (San Jose, CA). CFP was measured with a 445-nm laser and a 515/20 band-pass filter and mKate with a 561-nm laser, 610 emission filter, and 610/20 band-pass filter. For data analysis, 100,000 events were collected. A forward scatter/side scatter gate was generated using an untransfected negative sample and applied to all cell samples. Next, a CFP-positive gate (CFP arbitrary unit > 1000) was generated using one of the transfected samples and applied to all other cell samples. The mKate and CFP readings from untransfected HEK293 cells were set as baseline values and were subtracted from all other experimental samples. The normalized mKate values (mKate/CFP) were then calculated. All experiments were performed in triplicates.

RESULTS

Computational analysis and prediction of PAM-proximal constraints

To perform the DCA and Hamiltonian analysis between the PI domains and the PAM sequences (data acquisition workflow; Fig. S1), we used HMMER (36) to construct MSAs, with the PI domain (aa 1102–1358) of SpCas9 (UniProt: Q99ZW2) as the reference sequence. In total, 468 homologous sequences were identified and aligned (263 amino acids in length, gap included; Table S1). Next, for each of the 468 identified PI homologs, the names of the host organisms (the genus name and the species epithet) were recorded, except those for which the species information is missing (Fig. 1).

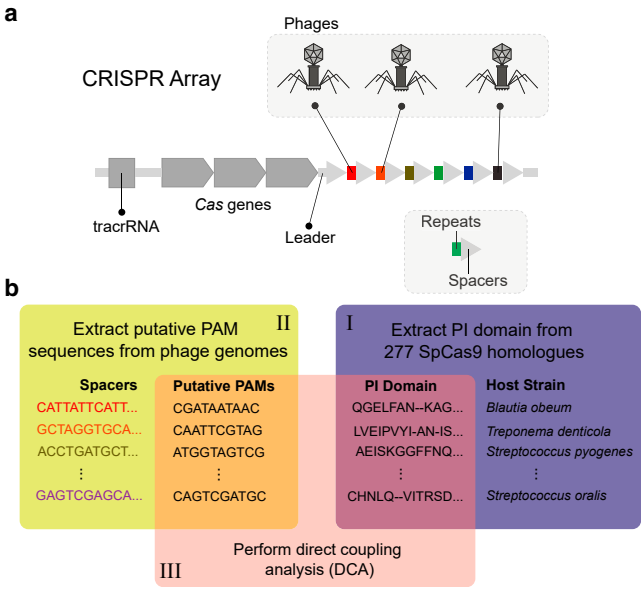


FIGURE 1 A schematic illustration of the CRISPR-associated systems and the procedures for identifying the PAM (protospacer adjacent motif) sequences and their partner Cas9 homologs. (*a*) The CRISPR locus typically consists of three components: *cas* genes, a leader sequence, and a repeat-spacer array. (*b*) The spacer sequences were first predicted and then blasted against the phage genomes to extract the protospacer sequences. The 10-nucleotide sequences immediately downstream of the protospacers were subsequently extracted as the putative PAM sequences, which were then conjugated with their corresponding PI domains for direct coupling analysis (DCA). To see this figure in color, go online.

In total, 186 unique host organisms were identified (Table S2). To identify the CRISPR "spacer" repeat sequences, we first extracted the genome sequences using the National Center for Biotechnology Information (NCBI) database (179/186 host organisms' genome sequences available). Subsequently, PILER-CR (37), which demonstrated both high sensitivity and high specificity in locating CRISPR repeats, was applied to the genome sequences (length of the repeats: 16–64 nucleotides (nts); length of the spacer: 8–64 nucleotides; minimal number of repeats: 3). In total, 4221 distinct CRISPR spacer sequences were identified from 148 host organisms (38 host organisms do not contain identifiable repeats; Table S3).

The CRISPR spacer sequences are hypothesized to derive from infecting phages and subsequently help protect the receiving cells from future infection (38). Because all 141 identified host organisms belong to bacteria kingdom, but not archaea, we performed NCBI nucleotide BLAST (Basic Local Alignment Search Tool) search for each of the CRISPR spacer against either Caudovirales (tailed phage, NCBI taxon identifier (taxid): 28883), Corticovirus (NCBI taxid: 10660), or Plasmaviridae (NCBI taxid: 10472). It is worth noting that 196 spacers from 63 distinct host organisms yielded statistically significant phage protospacer results (NCBI database: nr; e-value cutoff: 0.02; hitlist_size: 1), possibly because of the extreme diversity of bacteriophages and the

incompleteness of phage genome databases (39). Because for SpCas9 and its homologs, the PAM is located immediately downstream from the protospacer sequence, we subsequently extracted such 10-nucleotide PAM sequences (named as 10-PAM) from the identified protospacers (Fig. 1 *b*; Table S4).

We subsequently concatenated the putative 10-PAM sequences with the aligned PI domains. In cases in which multiple Cas9/PI variants or PAM sequences exist within the same host organism, each of the PI domains was associated with each of the PAM sequences (38). In total, 406 conjugated PI-PAM entries were generated (Table S5; positions 1–263: aligned PI domains; positions 264–273: PAM sequences) and were subsequently subjected to DCA for PI amino acid residue-PAM nucleotide pairs.

To study the functional variability of PAM sequences with Cas9, we first performed a naïve sequence alignment of the 196 identified 10-PAM sequences (Table S4), which appears to be promiscuous because nucleotides A or T were preferred through positions 2–10. More importantly, this simple alignment (Fig. S2) failed to capture the known preference of G at position 2, as well as G or A at position 3, for the SpCas9's PAM sequence (40).

Next, we implemented a Hamiltonian analysis to explore the global landscape of all possible PAM sequences interacting with SpCas9 protein (detailed in Materials and Methods, DCA and Hamiltonian Analysis). It is worth noting that compared to the simple alignment, this, to our knowledge, new approach takes advantage of the information extracted from all naturally occurring Cas9-PAM interactions and thus provides a more comprehensive and evolutionarily relevant prediction of PAM sequence functionality for SpCas9.

To test the robustness of our Hamiltonian analysis, we fixed the protein sequence (positions 1–263) as the PI domain of SpCas9. For the 10-PAM, we sampled either 4^5 (1024) of random sequences (5′-NNNNNNNNNN-3′), or according to currently known SpCas9 PAM pattern (5′-NG(A/G)-3′), we sampled 4^5 (1024) random sequences starting with 5′-TGG-3′ sequence (5′-TGGNNNNNNN-3′). Next, we performed bootstrapping (sample size: 1024; repetitions: 2000 times) on the difference of the means of Hamiltonians from the above two sampling subpopulations. As shown in Fig. S3, the subpopulations starting with 5′-TGG-3′ showed significantly lower Hamiltonians (Z-test, p-value < 0.05), which implied more favorable binding to the PI domain of SpCas9 protein.

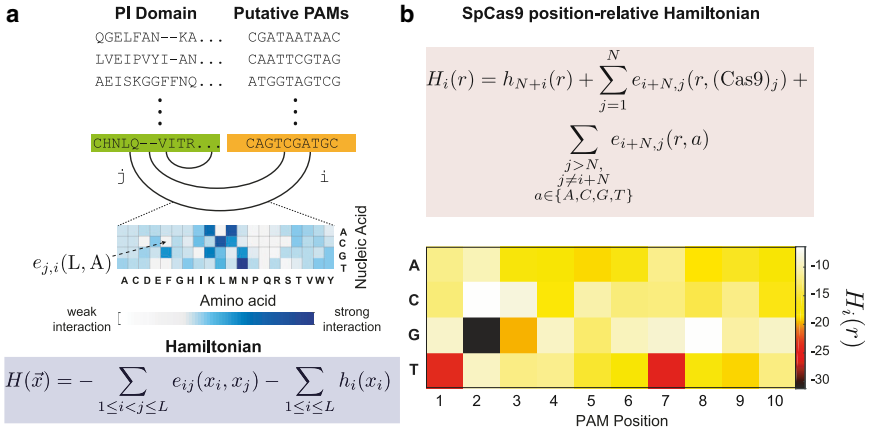
Next, to explore additional positions within the 10-PAM, we fixed the protein sequence (positions 1–263) as SpCas9 and evaluated the Hamiltonians for all possible 10-nucleotide PAM sequences by summing up the local fields for the PAM sequence, the couplings between the SpCas9 amino acid residues and the PAM nucleotides, as well as the couplings between two different PAM nucleotides (Fig. 2 *a*; Eq. 3). As shown in Fig. 2 *b*, the Hamiltonian analysis further confirmed the preference of G at position 2, as well as A/G at position 3. An unexpected observation from the Hamiltonian landscape (which uses the MSAs (Table S5) input to infer the values of DCA parameters) of all amino-acid-nucleotide sequences reveals a strong preference of T at PAM position 7 (Fig. 2 *b*), indicating that the PI-interacting properties of SpCas9 PAM sequences may not yet be fully explored.

Experimental validation and functional analysis

To determine if the seventh position of PAM indeed plays a role in the CRISPR/SpCas9 system, we designed a fluorescence reporter plasmid UBC-mKate-PEST and incorporated a target sequence 5′-AGTCAAGGCACTCTTGCCTA-3′ (target A) immediately downstream of the ATG start codon of mKate-PEST transcript (Fig. 3 *a*). For the 10-PAM sequence (5′-CGGTGNNNTG-3′), we fixed nucleotides at positions 1–6 as CGGTGG and nucleotides at positions 8–10 as GTG, whereas for position 7, four constructs were prepared using nucleotides A, G, C, or T.

In vitro cleavage assay using SpCas9

To test the effect of the seventh position of PAM in vitro, we isolated UBC-mKate-PEST transcripts by PCR using



$$H_i(r) = h_{N+i}(r) + \sum_{j=1}^{N} e_{i+N,j}(r,(\mathrm{Cas9})_j) + \sum_{\substack{j>N,\\ j\neq i+N\\ a\in\{A,C,G,T\}}} e_{i+N,j}(r,a)$$

$$H(\vec{x}) = -\sum_{1\le i<j\le L} e_{ij}(x_i,x_j) - \sum_{1\le i\le L} h_i(x_i)$$

FIGURE 2 Hamiltonian analysis of the SpCas9 PAM sequences reveals a preference of T nucleotide at position 7. (*a*) All parameters for the local fields and couplings were estimated using a mean-field approximation algorithm. These global probability parameters were then used for evaluating the Hamiltonians for aligned PI-PAM sequences. (*b*) The global landscape of SpCas9 PAM sequences was analyzed from the point of view of a position-relative Hamiltonian, which measures the contributions of all other positions to each PAM position. To see this figure in color, go online.

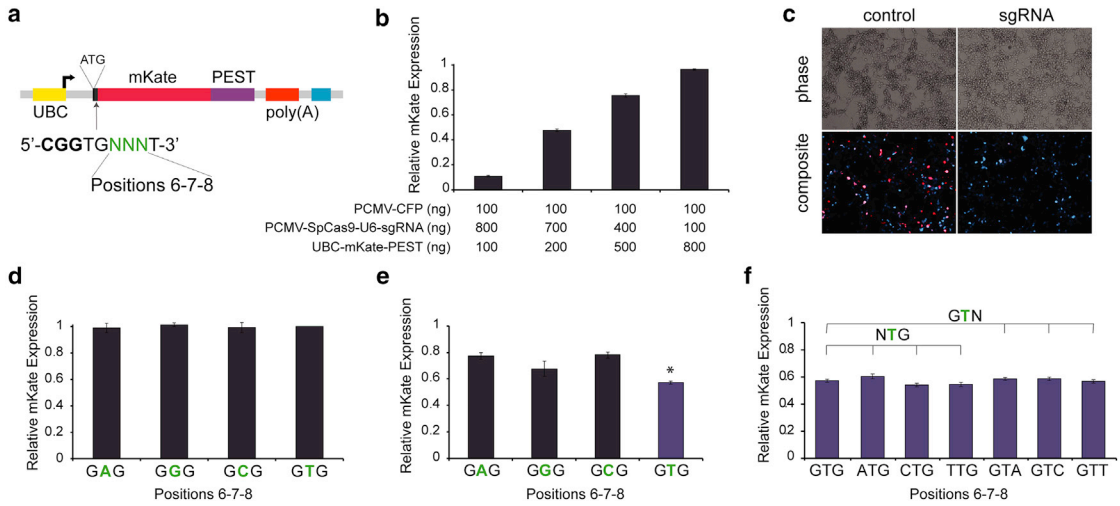


FIGURE 3 A fluorescence reporter plasmid assay confirms a preference of T nucleotide at position 7. (*a*) A schematic illustration of the mKate-PEST-based fluorescence reporter plasmid for the CRISPR/SpCas9 system. (*b*) Flow cytometry analysis of the editing efficiencies of the CRISPR/SpCas9 system at various CRISPR/SpCas9 (ng): reporter plasmid (ng) ratios. (*c*) Fluorescence microscopy images showed that the CRISPR/SpCas9 complex completely suppressed the expression of mKate-PEST at the ratio of CRISPR/SpCas9 (800 ng): reporter plasmid (100 ng). (*d*) Flow cytometry analysis showed that the nucleotide variations at PAM position 7 did not affect the expression levels of mKate-PEST. (*e*) Flow cytometry analysis of the PAM sequences differing at the seventh position confirmed the preference of T nucleotide at PAM position 7 for the CRISPR/SpCas9 system. (*f*) Flow cytometry analysis of the PAM sequences differing at the sixth and eighth positions showed no nucleotide preferences at PAM positions 6 and 8 for the CRISPR/SpCas9 system. Two-sample *t*-tests assuming unequal variances were performed in (*d*)–(*f*), and * indicates *p*-values <0.05 when comparing T nucleotides to the other three nucleotides at position 7. To see this figure in color, go online.

primers P15 and P16 (Table S6). The target DNA fragments (997 bp) were designed so that upon treatment of the CRISPR/SpCas9 system (41), two bands (702 bp, 295 bp) were produced. For in vitro cleavage assay, the DNA targets were incubated with SpCas9/sgRNA complex at a 1:2 ratio (mol:mol) for 15 min at 37°C. As shown in Fig. S4, the CRISPR/SpCas9 complex displayed higher editing efficacies when the seventh position of PAM was T compared to G or C. Having a T in the seventh position yields an increase in efficacy of 20% with respect to C, which yields the lowest efficacy and an increase of 15% with respect to G. This result agrees with the computational prediction that the seventh position proximal to the PAM may have an effect on editing properties.

We further incorporated different sgRNA target sequences (target B (5′-GGGGCCGGTTACTGTGGCTC-3′) and target C (5′-GCTGCCTCCGCTCTACTCAC-3′)) into the UBC-mKate-PEST transcripts to test whether the T preference in the seventh position was sgRNA specific. As shown in Figs. S5 and S6, for both sequences, having a T in the seventh position resulted in the highest editing efficacies. More specifically, for target B, having a T in the seventh position yields an increase in efficacy of 46% with respect to C and an increase of 66% with respect to A. Similarly, for target C, having a T in the seventh position yields an increase in efficacy of 27, 24, and 23% compared to A, C, and G, respectively.

It is worth noting that although the preference of T in the seventh position was observed with all three sgRNA targets, the effects of the other three nucleotides (A, C, G) appeared variable. As an example, for target A, having an A in the seventh position also yields better efficacies with respect to C (17%) or G (12%). However, for target B, having an A in the seventh position resulted in the lowest efficacy. These results indicate that the sgRNA target sequences may play a contextual role in the nucleotide preferences of PAM. Alternatively, because the Hamiltonian analysis was performed using naturally occurring PI and PAM sequences, we hypothesized that the in vitro cleavage assay may not fully capture the dynamics of the interaction between SpCas9/sgRNA complex and its DNA substrates in cells. This motivated a validation using in vivo functional assays.

Functional assay in kidney cells using SpCas9

Because of the generally high editing efficacy of the CRISPR/SpCas9 system, we first selected a relatively high reporter plasmid:CRISPR complex (ng:ng) ratio so that the suppression effects will not be close to 100% for all reporter constructs (Fig. 3, *b* and *c*). To test the effect of the seventh position of PAM in cells, we first transiently transfected each of the four UBC-mKate-PEST constructs containing target A (5′-AGTCAAGGCACTCTTGCCTA-3′) and showed comparable mKate expression levels in HEK293 cells (Fig. 3 *d*). Accordingly, the single-nucleotide variance at position 7 does not interfere with transcription and expression of the reporter fluorescent protein.

Subsequently, we transiently transfected HEK293 cells using 100 ng of PCMV-CFP, 500 ng of the reporter plasmids (GAG, GGG, GCG, and GTG), and 400 ng of either PCMV-SpCas9-U6-sgRNA/active or PCMV-SpCas9-U6-sgRNA/control. As shown in Fig. 3 *e*, although the PCMV-SpCas9-U6-sgRNA/active construct suppressed the expression levels of mKate for all four reporter plasmids, the GTG variant showed the lowest efficiency (57.3% compared to 77.5% for GAG, 67.6% for GGG, and 78.0% for GCG), which is consistent with our theoretical observations and in vitro results.

Finally, using the GTG variant as the baseline, we further generated reporter plasmid variants at positions 6 and 8 (ATG, CTG, and TTG for position 6; GTA, GTC, and GTT for position 8). As shown in Fig. 3 *f*, all variants showed suppression levels of mKate comparable to that of the GTG construct (57.3% compared to 60.5% for ATG, 54.1% for CTG, 54.6% for TTG, 58.6% for GTA, 58.7% for GTC, and 56.8% for GTT), indicating that nucleotides at these two positions have a lower effect on the CRISPR/SpCas9 operation.

DISCUSSION

CRISPR is a simple yet powerful technology for genome editing. The use of the CRISPR systems in higher organisms (including mammalian cells and animal models) is revolutionizing biology and biotechnology (42–44).

A recent study revealed widespread existence of anti-Cas9 antibodies in the sera of human patients (45). More critically, the same study has detected antigen-specific T-cells against Cas9 in the patient's serum, raising the risk of adverse immune responses such as cytokine release syndrome upon the CRISPR treatment (46,47). Therefore, it becomes imperative for the expression level of therapeutic SpCas9 vehicles to be maintained relatively low while having sufficient editing efficacies. We believe that the functional relevance of interactions preserved through evolution, such as the seventh position of SpCas9's PAM sequence, can guide the engineering of CRISPR systems with editing efficiency at relatively low copies.

Predictions from our DCA and Hamiltonian analysis may have eluded the current model of SpCas9-sgRNA-PAM complex. As an example, PI position 63, which corresponds to I1166 within the β19 loop of SpCas9 (48), appeared 6 times in the list of top 40 pairs with highest direct information values (Table S7; coupled with 10-PAM positions 1, 2, 4, 6, 8, and 10) based on our DCA, whereas according to one SpCas9-sgRNA-PAM complex crystal structure (Protein Data Bank (PDB): 4UN3), it is relatively distant from the PI-PAM interface (18.2 Å). Similarly, the same crystal structure indicates that the position 7 of the PAM sequence does not make direct contact with the PI domain of SpCas9. One possible explanation is that x-ray crystallography mainly captures the averaged information of a protein's structure and thus will ignore certain critical properties of the protein dynamics (49). This limitation may be even more prominent in our study as recent studies demonstrated that the SpCas9-sgRNA-DNA complex is highly flexible and state dependent (precatalytic, postcatalytic, and product states) (50,51). Indeed, as shown in Fig. S7 *a*, we calculated the distances between each amino acid residue within the PI domain and each nucleotide within the PAM from 10 currently available SpCas9-sgRNA-PAM complex crystal structures (PDB: 4UN3, 5Y36, 6O0X, 6O0Y, 6O0Z, 6AI6, 6AEB, 6AEG, 5FW2, 5FW3), and the histogram showed that the vast majority (92.5%) of the amino-acid-residue-nucleotide pairs are relatively distant (>12 Å), indicating that in the context of the CRISPR complex, the contact information provided by crystal structures may not fully capture all its potential properties. Moreover, for any such contacts that are relatively close (<12 Å), we analyzed the relative occurrences of each PAM position for each of the crystal structures. As shown in Fig. S7 *b*, in all structures, the positions 2 and 3, which are known to be critical in SpCas9 recognition, had relatively low occurrences, which further elucidates the limitation of relying on the crystal structures for the validation of DCA predictions.

Additionally, our DCA analysis implied a preference of T in the first position (Fig. 2 *b*), which is inconsistent with the current SpCas9 PAM model (5′-NG(A/G)-3′) (3,9). Indeed, using the same in vitro cleavage assay as described in Fig. S4, we fixed the sgRNA target sequence (target A, 5′-AGTCAAGGCACTCTTGCCTA-3′) while altering the nucleotide in the position 1 of the PAM sequences (5′-NGGTGGTGTG-3′). As shown in Fig. S8, all four versions of constructs yielded similar cleavage efficacies. In the particular case, we suspect that the importance of the second position (strong preference to G) reduces the sensitivity to the first position. In any case, these observations highlight the need to validate any DCA (or computational predictions) using experiments.

To conclude, herein, we introduced a coevolutionary framework to probe nonobvious molecular interactions in CRISPR systems. Identifying key amino acid residues using current methods is prohibitively challenging and time-consuming because of the extremely large sequence space involved in protein structure and interactions. We argue that coevolved and conserved interactions reveal key amino acid configurations that can yield novel Cas9-based systems endowed with high-value properties. The rational engineering of a CRISPR-Cas9 system with desired properties will have a transformative impact in genome editing technologies and will facilitate a wide range of applications.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2019.09.040.

AUTHOR CONTRIBUTIONS

Y.L., F.M., and L.B. designed the experiments. Y.L., R.L., and A.P.P. performed the experiments. Y.L., J.A.P., X.J., L.B., and F.M. analyzed the data. Y.L., J.A.P., L.B., and F.M. wrote the article. L.B. and F.M. supervised the project.

ACKNOWLEDGMENTS

This work was funded by the US National Science Foundation Faculty Early Career Development grant 1351354 (L.B.), National Science Foundation 1361355 (L.B.), Cecil H. and Ida Green Chair in Systems Biology Science (L.B.), and The University of Texas at Dallas (L.B. and F.M.).

REFERENCES

1. Cong, L., F. A. Ran, …, F. Zhang. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science.* 339:819–823.
2. Jinek, M., K. Chylinski, …, E. Charpentier. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science.* 337:816–821.
3. Mali, P., L. Yang, …, G. M. Church. 2013. RNA-guided human genome engineering via Cas9. *Science.* 339:823–826.
4. Gilbert, L. A., M. H. Larson, …, L. S. Qi. 2013. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell.* 154:442–451.
5. Qi, L. S., M. H. Larson, …, W. A. Lim. 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell.* 152:1173–1183.
6. Moore, R., A. Spinhirne, …, L. Bleris. 2015. CRISPR-based self-cleaving mechanism for controllable gene delivery in human cells. *Nucleic Acids Res.* 43:1297–1303.
7. Li, Y., C. M. Nowak, …, L. Bleris. 2018. CRISPR-based editing reveals edge-specific effects in biological networks. *CRISPR J.* 1:286–293.
8. Nowak, C. M., S. Lawson, …, L. Bleris. 2016. Guide RNA engineering for versatile Cas9 functionality. *Nucleic Acids Res.* 44:9555–9564.
9. Li, Y., S. Mendiratta, …, L. Bleris. 2016. Exploiting the CRISPR/Cas9 PAM constraint for single-nucleotide resolution interventions. *PLoS One.* 11:e0144970.
10. Morcos, F., A. Pagnani, …, M. Weigt. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA.* 108:E1293–E1301.
11. Sułkowska, J. I., F. Morcos, …, J. N. Onuchic. 2012. Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. USA.* 109:10340–10345.
12. Dos Santos, R. N., A. J. R. Ferrari, …, L. Martínez. 2018. Enhancing protein fold determination by exploring the complementary information of chemical cross-linking and coevolutionary signals. *Bioinformatics.* 34:2201–2208.
13. Jones, D. T., D. W. Buchan, …, M. Pontil. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics.* 28:184–190.
14. Hopf, T. A., L. J. Colwell, …, D. S. Marks. 2012. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell.* 149:1607–1621.
15. Ovchinnikov, S., H. Park, …, D. Baker. 2017. Protein structure determination using metagenome sequence data. *Science.* 355:294–298.
16. Davtyan, A., N. P. Schafer, …, G. A. Papoian. 2012. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B.* 116:8494–8503.
17. Jana, B., F. Morcos, and J. N. Onuchic. 2014. From structure to function: the convergence of structure based models and co-evolutionary information. *Phys. Chem. Chem. Phys.* 16:6496–6507.
18. Morcos, F., B. Jana, …, J. N. Onuchic. 2013. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. USA.* 110:20533–20538.
19. Sfriso, P., M. Duran-Frigola, …, M. Orozco. 2016. Residues coevolution guides the systematic identification of alternative functional conformations in proteins. *Structure.* 24:116–126.
20. Malinverni, D., A. Jost Lopez, …, A. Barducci. 2017. Modeling Hsp70/Hsp40 interaction by multi-scale molecular simulations and coevolutionary sequence analysis. *eLife.* 6:e23471.
21. Malinverni, D., S. Marsili, …, P. De Los Rios. 2015. Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones. *PLoS Comput. Biol.* 11:e1004262.
22. dos Santos, R. N., F. Morcos, …, J. N. Onuchic. 2015. Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci. Rep.* 5:13652.
23. Cheng, R. R., F. Morcos, …, J. N. Onuchic. 2014. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. USA.* 111:E563–E571.
24. Tamir, S., S. Rotem-Bamberger, …, R. Nechushtai. 2014. Integrated strategy reveals the protein interface between cancer targets Bcl-2 and NAF-1. *Proc. Natl. Acad. Sci. USA.* 111:5177–5182.
25. Ovchinnikov, S., H. Kamisetty, and D. Baker. 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife.* 3:e02030.
26. Hopf, T. A., C. P. Schärfe, …, D. S. Marks. 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife.* 3:e03430.
27. Jiang, X. L., E. Martinez-Ledesma, and F. Morcos. 2017. Revealing protein networks and gene-drug connectivity in cancer from direct information. *Sci. Rep.* 7:3739.
28. Cong, Q., I. Anishchenko, …, D. Baker. 2019. Protein interaction networks revealed by proteome coevolution. *Science.* 365:185–189.
29. Cheng, R. R., O. Nordesjö, …, F. Morcos. 2016. Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. *Mol. Biol. Evol.* 33:3054–3064.
30. Zhou, Q., N. Kunder, …, Z. T. Campbell. 2018. Global pairwise RNA interaction landscapes reveal core features of protein recognition. *Nat. Commun.* 9:2511.
31. Hopf, T. A., J. B. Ingraham, …, D. S. Marks. 2017. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35:128–135.
32. Figliuzzi, M., H. Jacquier, …, M. Weigt. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.* 33:268–280.
33. Tubiana, J., S. Cocco, and R. Monasson. 2019. Learning protein constitutive motifs from sequence data. *eLife.* 8:e39397.
34. Durbin, R., S. R. Eddy, …, G. Mitchison. 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK.
35. Honerkamp, J. 2012. Statistical Physics: An Advanced Approach with Applications. Springer Science+Business Media, Berlin, Germany.
36. Johnson, L. S., S. R. Eddy, and E. Portugaly. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics.* 11:431.
37. Edgar, R. C. 2007. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics.* 8:18.
38. Barrangou, R., C. Fremaux, …, P. Horvath. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science.* 315:1709–1712.
39. Klucar, L., M. Stano, and M. Hajduk. 2009. phiSITE: database of gene regulation in bacteriophages. *Nucleic Acids Res.* 38 (suppl_1):D366–D370.
40. Kleinstiver, B. P., M. S. Prew, …, J. K. Joung. 2015. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature.* 523:481–485.

41. Jamison, B. V., M. W. Thairu, and A. K. Hansen. 2018. Efficacy of in vivo electroporation on the delivery of molecular agents into aphid (hemiptera: Aphididae) ovarioles. *J. Insect Sci.* 18:49.

42. Amoasii, L., J. C. W. Hildyard, …, E. N. Olson. 2018. Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy. *Science.* 362:86–91.

43. Nelson, C. E., C. H. Hakim, …, C. A. Gersbach. 2016. In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. *Science.* 351:403–407.

44. Sánchez-Rivera, F. J., and T. Jacks. 2015. Applications of the CRISPR-Cas9 system in cancer biology. *Nat. Rev. Cancer.* 15:387–395.

45. Simhadri, V. L., J. McGill, …, Z. E. Sauna. 2018. Prevalence of pre-existing antibodies to CRISPR-associated nuclease Cas9 in the US population. *Mol. Ther. Methods Clin. Dev.* 10:105–112.

46. Charlesworth, C. T., P. S. Deshpande, …, M. H. Porteus. 2019. Identification of preexisting adaptive immunity to Cas9 proteins in humans. *Nat. Med.* 25:249–254.

47. Lee, D. W., R. Gardner, …, C. L. Mackall. 2014. Current concepts in the diagnosis and management of cytokine release syndrome. *Blood.* 124:188–195.

48. Nishimasu, H., F. A. Ran, …, O. Nureki. 2014. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell.* 156:935–949.

49. Kuzmanic, A., N. S. Pannu, and B. Zagrovic. 2014. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nat. Commun.* 5:3220.

50. Osuka, S., K. Isomura, …, S. Uemura. 2018. Real-time observation of flexible domain movements in CRISPR-Cas9. *EMBO J.* 37:e96941.

51. Zhu, X., R. Clarke, …, S. Subramaniam. 2019. Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. *Nat. Struct. Mol. Biol.* 26:679–685.