Comprehensive deletion scan of anti-CRISPR AcrIIA4 reveals essential and dispensable domains for Cas9 inhibition

Authors

Annette B Iturralde[1,2], Cory A Weller[1,3], Simone M Giovanetti[1], Meru J Sadhu[1]

[1] Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA

[2] Present address: Biomedical Sciences Graduate Program, University of Virginia, Charlottesville, Virginia, USA

[3] Present address: Center for Alzheimer's and Related Dementias, National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA

Email: meru.sadhu@nih.gov

Running title

Systematic deletion scan of AcrIIA4

Abstract

Delineating a protein's essential and dispensable domains provides critical insight into how it carries out its function. Here, we developed a high-throughput method to synthesize and test the functionality of all possible in-frame and continuous deletions in a gene of interest, enabling rapid and unbiased determination of protein domain importance. Our approach generates precise deletions using a CRISPR library framework that is free from constraints of gRNA target site availability and efficacy. We applied our method to AcrIIA4, a phage-encoded anti-CRISPR protein that robustly inhibits SpCas9. Extensive structural characterization has shown that AcrIIA4 physically occupies the DNA-binding interfaces of several SpCas9 domains; nonetheless, the importance of each AcrIIA4 interaction for SpCas9 inhibition is unknown. We used our approach to determine the essential and dispensable regions of AcrIIA4. Surprisingly, not all contacts with SpCas9 were required, and in particular, we found that the AcrIIA4 loop that inserts into SpCas9's RuvC catalytic domain can be deleted. Our results show that AcrIIA4 inhibits SpCas9 primarily by blocking PAM binding, and that its interaction with the SpCas9 catalytic domain is inessential.

Significance

For decades, researchers have determined the functionally important parts of proteins by deleting protein segments and seeing which ones affect protein function. This provides critical information about how the protein works. Here, we developed a high-throughput method to generate and test deletions in a protein of interest. Our method uses a CRISPR library approach, and can generate thousands of precisely

programmed deletions in a single protein. We used it to test the effects of deletions in the phage anti-CRISPR protein AcrIIA4. Previous studies have shown that AcrIIA4 binds Cas9 using several interfaces, but the individual importance of these interfaces was unclear. We find that AcrIIA4 acts by blocking PAM binding, while its interaction with Cas9's catalytic domain is dispensable.

Keywords

AcrIIA4 / anti-CRISPR / deletion / MAVE / SpCas9

52 Introduction

53

54 CRISPR systems consist of RNA-guided endonucleases that cleave nucleic acid
55 sequences specified by the variable spacer sequence of a bound guide RNA (gRNA)
56 (1). CRISPR systems are widespread in bacteria and archaea, where they can provide
57 adaptive immunity against viruses, plasmids, and other mobile genetic elements
58 (MGEs) through gRNAs that recognize foreign DNA or RNA and target their cleavage.
59 The adaptability of CRISPR endonucleases to cleave specific DNA sequences or
60 otherwise recruit proteins to specific genomic sites is used in myriad ways in molecular
61 biology research, predominantly featuring the type IIA Cas9 protein from *Streptococcus*
62 *pyogenes*, referred to as SpCas9. gRNA-bound SpCas9 cleaves double-stranded DNA
63 if its sequence matches the gRNA's 20-nucleotide (nt) spacer sequence followed by the
64 protospacer motif (PAM) sequence of NGG.

65 To counter the native role of CRISPR in host defense, MGEs encode a variety of
66 anti-CRISPR proteins, or Acrs (2). Acrs have diverse strategies to block CRISPR
67 cleavage, including through disruption of gRNA function, DNA binding, or DNA
68 cleavage. Acrs have been repurposed technologically to regulate CRISPR, including to
69 restrict CRISPR activity to a particular tissue type or to reduce cytotoxicity associated
70 with CRISPR (3, 4). One of the most well-characterized Acrs is AcrIIA4, a potent
71 inhibitor of SpCas9 identified from a *Listeria monocytogenes* prophage (5). AcrIIA4 is a
72 compact 87-amino acid protein composed of a central three-stranded β-sheet flanked
73 by three α-helices (6–8), as well as a 3_{10} helix in the β1-β2 loop that deforms upon
74 SpCas9 binding (9). AcrIIA4 occupies the DNA-binding interfaces of several domains of

75 SpCas9 (Fig. 1A). AcrIIA4 binds the SpCas9 C-terminal domain (CTD), which binds the
76 PAM sequence of target DNA, using its β3 strand, α3 helix, and β2-β3, β3-α2, and α2-
77 α3 loops. Residues in AcrIIA4's α1-β1 and β2-β3 loops interact with the SpCas9
78 topoisomerase-homology (TOPO) domain, which aids in PAM binding. Finally, the β1-
79 β2 loop of AcrIIA4 occupies the active site of SpCas9's RuvC domain, which cleaves
80 the nontarget DNA strand. However, the individual importance of these interactions to
81 SpCas9 inhibition has not been determined.

82 To precisely determine which domains of AcrIIA4 are essential for anti-CRISPR
83 function, we generated nearly all possible in-frame deletions of all sizes in AcrIIA4 and
84 tested their impact on SpCas9 inhibition. Functional domains of proteins have been
85 mapped for decades by determining the effects of deleting segments of the protein (10).
86 Augmenting such approaches to assay all possible in-frame deletions enables full and
87 precise delineation of a protein's essential and dispensable domains in an unbiased
88 fashion (11), but creating all possible sizes of deletions across a gene scales
89 quadratically with gene size and quickly amounts to thousands of deletions,
90 necessitating a high-throughput approach. CRISPR technology can be used for high-
91 throughput genetic interrogation, by synthesizing a library of DNA molecules encoding
92 gRNAs and/or repair templates and using it to generate genetic variants in a pool of
93 cells that contains CRISPR machinery (12–14). Here, we developed a high-throughput
94 pooled CRISPR approach to comprehensively determine the effects of deletions in a
95 protein and applied it to AcrIIA4. We find that the AcrIIA4 regions responsible for
96 blocking the PAM-interacting domains of SpCas9 are essential, whereas the segment
97 involved in blocking the RuvC catalytic domain is dispensable.

98

99 Results

100

101 We developed an approach for comprehensive deletion mutagenesis that uses
102 CRISPR in the budding yeast *Saccharomyces cerevisiae* (Fig. 1B). We used
103 synonymous variants to generate two alleles of *acrIIA4* with maximum DNA sequence
104 divergence while reflecting yeast codon usage, resulting in 61% DNA identity. We
105 inserted the two *acrIIA4* alleles in tandem into the yeast genome at the *HO* locus,
106 placing them after a galactose-inducible promoter and placing a synthetic Cas9 target
107 site and the *URA3* gene from *Kluyveromyces lactis* between them. To generate an
108 *acrIIA4* allele carrying a deletion between any pair of codons n_1 and n_2, we direct
109 CRISPR cleavage at the synthetic target site between the *acrIIA4* copies and provide a
110 repair template with homology to the first acrIIA4 copy up to codon n_1, followed by
111 homology to the second copy starting from codon n_2. Thus, the use of this repair
112 template during homology-directed repair produces a single fusion *acrIIA4* gene that
113 has the region from codon n_1 to codon n_2 deleted. The fusion event also deletes the
114 CRISPR target site and *URA3*, precluding further cutting and allowing for the selection
115 of cells with successful deletions through their ability to grow on media containing 5-
116 fluoroorotic acid (5-FOA), which is toxic to *S. cerevisiae* cells expressing *URA3* (15).
117 There are 3,741 possible unique in-frame *acrIIA4* deletions. To generate a
118 comprehensive library of deletions we used large-scale array-based synthesis to
119 synthesize a pool of all possible deletion-directing repair templates, composed of 100 nt
120 matching the first *acrIIA4* copy followed by 100 nt matching the second. We cloned the

121 repair template oligos into three replicate 2-micron plasmid libraries for high-copy
122 maintenance in yeast cells. Massively parallel sequencing of the repair templates from
123 two of the resulting plasmid libraries revealed 3,573 (95%) repair templates were
124 represented in at least one of the replicate plasmid pools. The plasmid pools also
125 encoded a gRNA targeting the novel sequence inserted between the *acrIIA4* genes,
126 which was designed to have at least six mismatches with any other SpCas9 target site
127 in the *S. cerevisiae* genome. Note that as the gRNA target site is constant, this
128 approach is not constrained by the density of gRNA target sites in the *acrIIA4* gene, nor
129 is there any aspect of variability in gRNA efficacy.

130 We transformed the repair template library into the *acrIIA4-URA3-acrIIA4* yeast
131 strain. The strain also carried two copies of SpCas9, one under a *MAL62* promoter and
132 the other under a *GAL*L promoter, allowing SpCas9 to be activated by either maltose or
133 galactose (16–18). Both promoters are repressed by glucose. Transformed cells were
134 selected on glucose plates and then re-plated on maltose to induce the maltose-
135 responsive SpCas9 to generate deletion-carrying *acrIIA4* alleles. Resulting colonies
136 were then grown on 5-FOA plates to select cells that had deleted *URA3* and thus were
137 carrying the designed *acrIIA4* deletion alleles.

138 We sequenced the *acrIIA4* alleles present in the 5-FOA-resistant population by
139 massively parallel sequencing. We recovered 3,305 distinct programmed deletions
140 across the three replicates, 88% of the total possible deletions (Fig. 1C). Deletions were
141 recovered across the entire range of sizes.

142 Next, we determined the ability of the *acrIIA4* deletion alleles to inhibit SpCas9
143 activity. We generated a plasmid expressing a gRNA that robustly targets the essential

144 gene *ARB1* (14); usage of this gRNA for CRISPR cutting is lethal. To identify *acrIIA4*
145 deletion alleles that maintained anti-CRISPR function, we transformed the *ARB1*-
146 targeting gRNA plasmid into our pool of cells with programmed *acrIIA4* deletions and
147 induced SpCas9 and *acrIIA4* expression on galactose plates. Approximately 30% as
148 many colonies survived CRISPR induction compared to glucose plates without CRISPR
149 expression. We used massively parallel sequencing of the *acrIIA4* genomic locus to
150 identify the *acrIIA4* alleles present in survivors. Read counts were normalized against
151 the count of each allele from the uninduced glucose condition (Fig. 2A).
152 We observed three distinct clusters of AcrIIA4 deletions that maintain function, in the
153 β1-β2 loop, the β3-α2 loop, and the end of the α3 helix (Fig. 2A, Supplementary Fig. 1).
154 Given the contacts between the β1-β2 loop and the catalytic site of the RuvC domain,
155 the loop's tolerance of deletions was notable. The loop is eight amino acids long; we
156 found that AcrIIA4 retained anti-CRISPR function even when six of these amino acids
157 were deleted (Fig. 2B). As this comprises the major RuvC-interacting surface, we
158 propose that RuvC binding is dispensable for AcrIIA4's anti-CRISPR function.
159 AcrIIA4 makes contacts with the SpCas9 CTD through residues in the β3 strand, α3
160 helix, and β2-β3, β3-α2, and α2-α3 loops. In CRISPR targeting, arginines Arg1333 and
161 Arg1335 of the SpCas9 CTD engage the two guanines in the NGG PAM sequence.
162 Residues 67-70 in the α2-α3 loop and α3 helix of AcrIIA4, which includes three
163 negatively charged residues, interact with the positively charged arginines (6).
164 Interestingly, this segment of AcrIIA4 tolerated single amino-acid deletions, but not any
165 larger deletions (Fig. 2C). We propose that this interaction is required for SpCas9
166 inhibition and can tolerate the removal of a single negative charge from AcrIIA4, but no

167 more. Supporting the importance of this interaction, we see that Asn39 in the β2-β3
168 loop, which also contacts Arg1335, was not dispensable. In contrast, the β3-α2 loop,
169 which contacts a distal component of the SpCas9 CTD, was a hotspot for tolerated
170 deletions and could be shortened by up to four amino acids (Fig. 2D). This suggests
171 that the critical CTD contacts are precisely the ones at the PAM-interacting interface.
172 AcrIIA4 makes relatively fewer contacts to SpCas9's TOPO domain, which is also
173 involved in PAM binding. TOPO residues that interact with the PAM are contacted by
174 the AcrIIA4 residues Asp14 and Asn36 (6), both of which were partially dispensable
175 (Fig. 2E). We noted that Asn36 is adjacent to Asn35 in the β2-β3 loop, and while each
176 asparagine was individually partially dispensable, simultaneous deletion of both was not
177 tolerated. Thus, binding of TOPO by AcrIIA4 appears to be important for AcrIIA4
178 function, with functional redundancy buffering the loss of individual contacts.
179 We observed a large number of tolerated deletions towards the end of the C-
180 terminal helix α3 (Fig. 2E), with up to six terminal residues being dispensable. This
181 segment does not make any contacts with SpCas9 and is largely separate from the rest
182 of AcrIIA4. Consistent with the C-terminus being mutable, a C-terminal GFP fusion does
183 not disrupt anti-CRISPR activity (19). In contrast, deletions of seven or more terminal
184 residues were not tolerated. These deletions remove Ile80, which contributes to the
185 hydrophobic core of AcrIIA4 along with Leu6, Ile10, and Phe55. Intriguingly, we
186 observed that three six-residue deletions encompassing Ile80 retained partial function in
187 all three replicates: Δ78-83, Δ79-84, and Δ80-85 (Supplementary Fig. 2). These
188 deletions move Leu86 to position 80, which we propose partially reconstitutes Ile80's
189 role in AcrIIA4 stability.

190 We were curious whether the dispensable segments we identified in AcrIIA4 were
191 ever absent in naturally occurring AcrIIA4 homologs. We identified 124 homologs of
192 AcrIIA4 through a BLASTp search of the NCBI sequence database (20). Aligning these
193 to AcrIIA4 revealed that 48 had gaps relative to AcrIIA4, falling into six unique patterns
194 (Fig. 3A). Strikingly, most observed alignment gaps were found in regions we identified
195 as dispensable. These included gaps in the β1-β2 and β3-α2 loops and truncations of
196 the C-terminus. The largest gap was seen in several homologs that had a β1-β2 loop
197 five amino acids shorter than in AcrIIA4, consistent with our observation that the AcrIIA4
198 β1-β2 loop could be shortened by up to six amino acids while maintaining anti-CRISPR
199 function.

200 One alignment gap was present in a non-dispensable segment in the AcrIIA4
201 homolog from *Listeria innocua* (NCBI ID WP_185342258.1, hereafter referred to as
202 $AcrIIA4_{Lin}$), which appeared to have a one-amino acid shortening of the indispensable
203 C-terminal end of the β2 strand, in addition to a five-amino acid shortening of the
204 dispensable β1-β2 loop. We generated an AlphaFold model (21) of the $AcrIIA4_{Lin}$
205 protein structure (Fig. 3B) and found that the β2 strand in the model was five amino
206 acids long, the same length as in AcrIIA4, while the modeled β1-β2 loop in $AcrIIA4_{Lin}$
207 was a full six amino acids shorter than in AcrIIA4, consistent with our observed deletion
208 tolerances.

209 We compared our results to three computational approaches for predicting the
210 effects of deletions in AcrIIA4. First, we used AlphaFold to predict structures for all
211 AcrIIA4 deletions of 10 amino acids or fewer, and then used FoldX to calculate the
212 differences in protein stability between the predicted deletion-containing structures and

213 the wild-type structure (22). We found many similarities between the predicted stability
214 effects and our results, such as deletions being more tolerated in loops and at the C-
215 terminus, and the tolerance of the 6-amino acid deletions that replaced Ile80 with Leu86
216 (Fig. 4A, B). The main departure was that many more large deletions were predicted to
217 be stable than we found to be functional, especially in the β1 strand and the β2-β3 and
218 α2-α3 loops. This could suggest these deletions form stable but nonfunctional proteins,
219 or it could reflect bias in AlphaFold towards predicting structures to be stable (23). Next,
220 we predicted AcrIIA4 deletion functionality using ESM1b, which predicts a sequence's
221 likelihood to exist given the "language" of known protein sequences (24). The
222 functionality of a variant can be predicted by comparing the likelihood of the variant
223 sequence to the likelihood of the wild-type sequence. For deletions of single amino
224 acids in AcrIIA4, ESM1b also generally predicted higher deletion tolerance in loops and
225 at the C-terminus (Fig. 4C), though it did not match the experimental data as closely as
226 the AlphaFold-based stability predictions did. The performance of ESM1b decreased for
227 larger deletions, which it often predicted would be more likely than the wild-type AcrIIA4
228 sequence. Also, no deletions had scores lower than -7, which is the recommended
229 cutoff for calling a variant deleterious (24). Perhaps the poor performance of ESM1b
230 stems from AcrIIA4 being a small protein that does not have homology to any described
231 protein domains, which may lead to its sequence not following patterns learned from
232 other protein sequences. Last, we used PROVEAN, which predicts the impact of protein
233 variants by their consequences on alignments to homologous protein sequences (25).
234 Using a PROVEAN score threshold of -5 for loss-of-function, all deletions were
235 predicted to be deleterious aside from those at the C-terminus (Fig. 4D). A more

236 permissive threshold of -10 predicted functionality for some additional small deletions,
237 often in loops but also in the α1 helix and the β strands (Supplementary Fig. 3C). We
238 note that the recommended PROVEAN score threshold for loss-of-function is -2.5 (26).
239 Overall, we found that the three computational predictors of deletion effect had different
240 sensitivities and accuracies, and that predicting the effects of larger deletions is a
241 particular challenge. None of the three predictors captured the increased dispensability
242 of the RuvC-interacting residues relative to the residues that contact Cas9's PAM-
243 interacting domains.
244
245
246 Discussion
247
248 AcrIIA4 is a robust inhibitor of SpCas9. It physically interacts with SpCas9 through
249 an extensive interface that involves several distinct SpCas9 domains, including domains
250 involved in PAM recognition (CTD and TOPO) and DNA cleavage (RuvC), though the
251 importance of each individual interaction to SpCas9 inhibition was unknown. To
252 elucidate the mechanism of AcrIIA4 inhibition of SpCas9, we comprehensively deleted
253 segments of AcrIIA4 and assayed their function. We found that the interaction with the
254 RuvC domain, mediated by the β1-β2 loop of AcrIIA4, is dispensable, as this eight-
255 amino acid loop could be shortened by up to six amino acids without loss of anti-
256 CRISPR activity. Our finding that AcrIIA4 does not need to occupy the RuvC active site
257 to inhibit SpCas9 is consistent with the fact that AcrIIA4 robustly inhibits catalytically
258 inactive Cas9 from binding DNA, which necessarily precedes DNA cleavage (8). Still, it

259 remained possible that AcrIIA4 binding to the RuvC domain contributed to the binding
260 energy required to block DNA binding. Based on our findings, binding to Cas9's PAM-
261 interacting domains is sufficient.

262 Though the physical interactions between AcrIIA4 and SpCas9 are largely mediated
263 by the loops between secondary structural elements, we observed that none of the α
264 helices and β strands of AcrIIA4 were fully dispensable. This likely reflects
265 intramolecular interactions necessary for AcrIIA4 stability and functional shape,
266 matching the observation that evolutionarily observed indels rarely occur in α helices
267 and β strands (27). Consistent with this hypothesis, we observed that partial deletions of
268 helices α1 and α3 were tolerated. These helices are at the very start and end of the
269 protein, and the tolerated deletions were not near the AcrIIA4 core, so deletion of these
270 segments might be less likely to disrupt AcrIIA4's structure. Indeed, AlphaFold-predicted
271 structures of these deletions were calculated to have comparable stability to the WT
272 structure (Fig. 4B). Deletion scans of other proteins have similarly found higher
273 tolerance for deletions at the N- and C-termini of domains (28).

274 Interestingly, we found several segments that could not be deleted despite being
275 composed of smaller dispensable segments, as has been observed previously in other
276 proteins (29). This could occur through functional redundancy, as we propose for amino
277 acids 67-70, which has several negatively charged residues and can tolerate single
278 amino acid deletions, but no further. Alternatively, a segment could have no direct role
279 in protein function, but could link functional components that need to be a minimum
280 distance apart. For instance, while multiple deletions of six amino acids were tolerated
281 in the eight-amino acid β1-β2 loop, deletions of seven amino acids were not. We

282 hypothesize that such deletions disrupt the relative positioning of strands β1 and β2,
283 which may affect AcrIIA4 functionality while maintaining stability, based on AlphaFold's
284 stability predictions (Fig. 4B). The ubiquity of non-additive deletion effects highlights the
285 utility of comprehensively testing deletions, which can reveal functional redundancy and
286 dispensable segments that could be missed with a low-throughput approach.
287 A deep mutational scanning analysis of AcrIIA4 has been reported (30),
288 measuring the ability of all possible nonsynonymous substitutions in AcrIIA4 to inhibit
289 transcriptional repression mediated by catalytically inactive SpCas9 (CRISPRi). Though
290 the two assays could have different sensitivities to Acr function, we see broad
291 agreement in that positions that tolerated deletion also tolerated nonsynonymous
292 variants, and positions that were intolerant of nonsynonymous variants did not tolerate
293 deletion. However, positions that did not tolerate deletions often tolerated
294 nonsynonymous variants, especially in α helices and β strands. This could reflect roles
295 that secondary structures play in the structural integrity of AcrIIA4 or in the proper
296 placement of the Cas9-interacting elements, as suggested above; such roles may be
297 relatively indifferent to the biochemical properties of the secondary structures' amino
298 acid side chains. Another difference between these two analytical approaches is that
299 deletion scanning additionally looks at the effects of deleting multiple amino acids at
300 once, which allows for the analysis of functional redundancy and domain dispensability.
301
302 Comprehensive scanning of deletion effects has many applications in addition to
303 mapping proteins' essential and dispensable segments. It can also be used in genetic
304 engineering to produce improved or novel protein function, as indels have been shown

305 to have larger effects on protein function than missense variants (31). Similarly,
306 deletions are an important source of evolutionary novelty (32, 33), so deletion scanning
307 is useful to understand evolutionary trajectories. Deletion scanning can also be used in
308 clinical genetics to characterize potential disease-causing deletion alleles.
309 Understanding the effects of deletion variants is essential for interpreting the risk of
310 genetic disease from genome sequences: after single-nucleotide variants, deletions are
311 the most common type of variant on ClinVar (34), and each individual person is
312 predicted to harbor approximately 200 in-frame insertions or deletions relative to the
313 human reference genome (35).

314 Our strategy for deletion scanning involved chimerizing two copies of *acrIIA4*
315 following CRISPR cutting between them. A previous study used CRISPR to generate
316 cuts within a single copy of yeast *SGS1*, which were repaired using repair templates
317 that directed 20-amino acid deletions across the gene body (36). By cutting between
318 two copies of our target gene, we did not require PAM sequences close to the regions
319 being deleted, allowing us to comprehensively characterize deletions of small to large
320 size. In addition, our strategy uses a constant gRNA, which avoids issues arising from
321 variable gRNA targeting efficiency. Finally, by concurrently deleting a *URA3* gene
322 between the target gene copies, we could select for cells carrying deletion alleles.

323 Deletion scans have also been carried out using non-CRISPR approaches. One
324 strategy is to use oligonucleotide synthesis technology to directly generate deletion
325 alleles, rather than to generate repair templates that direct deletions following CRISPR
326 cutting. As oligonucleotides synthesized economically at large scale are currently limited
327 to lengths of 300 nt or fewer, to date this approach has been applied to short genes,

328 segments of genes, or small deletions within larger genes (11, 28, 29, 33). Another
329 strategy is to introduce restriction sites into the gene of interest, cleave, and then self-
330 anneal to produce truncated gene products. For deletion scanning methods that use this
331 framework (31, 37), a challenge is that novel sequences are inserted at the deletion
332 junction for most or all of the deletion alleles.

333 An inherent challenge to the concept of deletion scanning is that the number of
334 potential deletions scales quadratically with the size of the protein, which could
335 theoretically be impractical for a large gene even when using a high-throughput
336 approach. We note that the oligonucleotide library synthesis product we used can
337 produce up to 244,000 sequences in a single order. With our approach, that would allow
338 comprehensive deletion scans of proteins up to 700 residues, encompassing most
339 proteins. More importantly, our approach allows specific deletions to be programmed,
340 such as deletions under a certain size or deletions within predicted domains or
341 secondary structure elements. Screening select deletions would lead to a linear scaling
342 with protein length and thus prevent the number of tested deletions from becoming
343 unmanageable.

344 We observed that the repair template plasmid pool had lower representation of
345 repair templates targeting deletions that ended around codons 50-70 of acrIIA4
346 (Supplementary Fig. 4A), leading to a moderate depletion of those deletions from our
347 set of recovered deletion alleles (Fig. 1C). We suspect that the DNA sequence around
348 codons 50-70 of the second acrIIA4 copy is challenging either for the oligonucleotide
349 synthesis or for the library cloning. The representation of deletions in this range was
350 sufficient for our analyses, but if a more complete representation of deletions in that

351 range had been desired, we would expect it to be possible through targeted
352 amplification of those repair templates from the oligonucleotide library, resynthesis of
353 that subset of oligonucleotides, or re-coding of that region using synonymous codons.
354 Indeed, we were able to fully remedy a similar issue that arose in our initial cloning
355 using synonymous re-coding of a segment of the first acrIIA4 copy (see Methods).
356 For applications other than mapping essential domains, such as determining the
357 evolutionary or pathogenic potential of genetic variants, it can be interesting to test
358 insertions as well as deletions (11, 33). Our method could be easily adapted to
359 additionally generate internal duplicated segments, which are the most common form of
360 naturally occurring insertion mutations (38), by having repair templates in which the start
361 point in the second gene copy comes before the end point of the first copy.
362 Homologous proteins often have divergent functions, and chimeras between
363 them can map the causative sequence differences. For instance, the yeast GTPases
364 Sec4 and Ypt1 have different functions in the secretory pathway that were mapped to
365 differences in loops 2 and 7 (39, 40). Chimeras have also been used to study orthologs
366 between species, such as the human and macaque homologs of the innate immune
367 protein TRIM5α, whose differences in HIV restriction were mapped to a loop in the
368 SPRY domain (41–44). Our method could be readily adapted to generate chimeras
369 between two different genes, rather than to make deletions. Generating and
370 characterizing chimeras in high throughput would enable rapid fine-resolution
371 determination of specific amino acids causing functional differences between homologs.
372

373

374 References

375 1. M. Jinek, *et al.*, Structures of Cas9 Endonucleases Reveal *RNA*-Mediated
376 Conformational Activation. *Science* **343**, 1247997 (2014).
377 2. A. R. Davidson, et al., Anti-CRISPRs: Protein Inhibitors of CRISPR-Cas Systems.
378 *Annu. Rev. Biochem.* **89**, 1–24 (2020).
379 3. J. Lee, et al., Tissue-restricted genome editing in vivo specified by microRNA-
380 repressible anti-CRISPR proteins. RNA 1421–1431 (2019).
381 4. C. Li, et al., HDAd5/35++ Adenovirus Vector Expressing Anti-CRISPR Peptides
382 Decreases CRISPR/Cas9 Toxicity in Human Hematopoietic Stem *Cell*s. *Mol. Ther. -*
383 *Methods Clin. Dev.* **9**, 390–401 (2018).
384 5. B. J. Rauch, et al., Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. Cell **168**,
385 150-158.e10 (2017).
386 6. D. Dong, et al., Structural basis of CRISPR–SpyCas9 inhibition by an anti-CRISPR
387 protein. *Nature* **546**, 436–439 (2017).
388 7. H. Yang, D. J. Patel, Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4
389 Targeting SpyCas9. *Mol. Cell* **67**, 117-127.e5 (2017).
390 8. J. Shin, et al., Disabling Cas9 by an anti-CRISPR DNA mimic. *Sci. Adv.* **3**, e1701620
391 (2017).
392 9. I. Kim, et al., Solution structure and dynamics of anti-CRISPR AcrIIA4, the Cas9
393 inhibitor. *Sci. Rep.* **8**, 3883 (2018).
394 **10**. S. Munro, H. R. B. Pelham, Use of peptide tagging to detect proteins expressed
395 from cloned genes: deletion mapping functional domains of Drosophila hsp70. *The*
396 *EMBO Journal* **3**, 3087–3093 (1984).

397 11. M. Seuma, B. Lehner, B. Bolognesi, An atlas of amyloid aggregation: the impact of
398 substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation. *Nat.*
399 *Commun.* **13**, 7084 (2022).
400 **12**. O. Shalem, et al., Genome-Scale CRISPR-Cas9 Knockout Screening in Human
401 Cells. Science 343, 84–87 (2014).
402 13. G. M. Findlay, et al., Accurate classification of BRCA1 variants with saturation
403 genome editing. Nature **562**, 217–222 (2018).
404 14. M. J. Sadhu, et al., Highly parallel genome variant engineering with CRISPR–Cas9.
405 *Nat. Genet.* **50**, 510–514 (2018).
406 15. J. D. Boeke, F. L. Croute, G. R. Fink, A positive selection for mutants lacking
407 orotidine-5′-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance.
408 *Mol. Gen. Genet. MGG* **197**, 345–346 (1984).
409 16. A. W. Gibson, et al., Constitutive Mutations of the Saccharomyces cerevisiae MAL-
410 Activator Genes MAL23, MAL43, MAL63, and mal64. *Genetics* **146**, 1287–1298 (1997).
411 17. R. L. Finley, H. Zhang, J. Zhong, C. A. Stanyon, Regulated expression of proteins
412 in yeast using the MAL61–62 promoter and a mating scheme to increase dynamic
413 range. Gene **285**, 49–57 (2002).
414 18. C. A. Weller, et al., Highly complete long-read genomes reveal pangenomic
415 variation underlying yeast phenotypic diversity. *Genome Res* gr.277515.122 (2023).
416 https://doi.org/10.**11**01/gr.277515.122.
417 19. E. M. Basgall, et al., Gene drive inhibition by the anti-CRISPR proteins AcrIIA2 and
418 AcrIIA4 in Saccharomyces cerevisiae. *Microbiology*+ **164**, 464–474 (2018).
419 20. E. W. Sayers, et al., Database resources of the National Center for Biotechnology

420 Information. *Nucleic Acids Res.* **52**, D**33**–D43 (2023).

421 21. J. Jumper, et al., Highly accurate protein structure prediction with AlphaFold.

422 Nature **596**, 583–589 (2021).

423 **22**. J. Schymkowitz, et al., The FoldX web server: an online force field. Nucleic Acids

424 Res. 33, W382–W388 (2005).

425 23. G. R. Buel, K. J. Walters, Can AlphaFold2 predict the impact of missense mutations

426 on structure? *Nat. Struct. Mol. Biol.* **29**, 1–2 (2022).

427 24. N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, V. Ntranos, Genome-wide

428 prediction of disease variant effects with a deep protein language model. Nat. Genet.

429 **55**, 1512–1522 (2023).

430 25. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, A. P. Chan, Predicting the Functional

431 Effect of Amino Acid Substitutions and Indels. *Plos One* **7**, e46688 (2012).

432 26. Y. Choi, A. P. Chan, PROVEAN web server: a tool to predict the functional effect of

433 amino acid substitutions and indels. *Bioinform. (Oxf., Engl.)* **31**, 2745–7 (2015).

434 27. S. Pascarella, P. Argos, Analysis of insertions/deletions in protein structures. *J.*

435 *Mol. Biol.* **224**, 461–471 (1992).

436 28. M. Topolska, A. Beltran, B. Lehner, Deep indel mutagenesis reveals the impact of

437 insertions and deletions on protein stability and function. *bioRxiv* 2023.10.06.561180

438 (2023). https://doi.org/10.1101/2023.10.06.561180.

439 29. C. B. Macdonald, et al., DIMPLE: deep insertion, deletion, and missense mutation

440 libraries for exploring protein variation in evolution, disease, and biology. *Genome Biol.*

441 **24**, 36 (2023).

442 30. T. Stadelmann, et al., A deep mutational scanning platform to characterize the

443 fitness landscape of anti-CRISPR proteins. bioRxiv (2021).
444 31. S. Emond, et al., Accessing unexplored regions of sequence space in directed
445 enzyme evolution via insertion/deletion mutagenesis. Nat. Commun. 11, 3469 (2020).
446 32. K. R. McCarthy, et al., Recurrent deletions in the SARS-CoV-2 spike glycoprotein
447 drive antibody escape. Science **371**, 1139–1142 (2021).
448 33. J. Tenthorey, et al., Indels allow antiviral proteins to evolve functional novelty
449 inaccessible by missense mutations. (2024).
450 https://doi.org/10.1101/2024.05.07.592993.
451 34. M. J. Landrum, et al., ClinVar: improving access to variant interpretations and
452 supporting evidence. *Nucleic acids Res.* **46**, D1062–D1067 (2017).
453 35. R. M. Durbin, et al., A map of human genome variation from population-scale
454 sequencing. Nature **467**, 1061–1073 (2010).
455 36. X. Guo, et al., High-throughput creation and functional profiling of DNA sequence
456 variant libraries using CRISPR–Cas9 in yeast. *Nat Biotechnol* **36**, 540–546 (2018).
457 37. A. Shams, et al., Comprehensive deletion landscape of CRISPR-Cas9 identifies
458 minimal RNA-guided DNA-binding modules. Nat. Commun. 12, 5664 (2021).
459 38. S. B. Montgomery, et al., The origin, evolution, and functional impact of short
460 insertion–deletion variants identified in 179 human genomes. *Genome Res.* **23**, 749–
461 761 (2013).
462 39. B. Dunn, T. Stearns, D. Botstein, Specificity domains distinguish the Ras-related
463 GTPases Ypt1 and Sec4. Nature **362**, 563–565 (1993).
464 40. P. Brennwald, P. Novick, Interactions of three domains distinguishing the Ras-
465 related GTP-binding proteins Ypt1 and Sec4. Nature 362, 560–563 (1993).

466 41. M. Stremlau, M. Perron, S. Welikala, J. Sodroski, Species-Specific Variation in the
467 B30.2(SPRY) Domain of TRIM5α Determines the Potency of Human Immunodeficiency
468 Virus Restriction. *J. Virol.* **79**, 3139–3145 (2005).
469 42. M. W. Yap, S. Nisole, J. P. Stoye, A Single Amino Acid Change in the SPRY
470 Domain of Human Trim5α Leads to HIV-1 Restriction. *Curr. Biol.* **15**, 73–78 (2005).
471 43. S. L. Sawyer, L. I. Wu, M. Emerman, H. S. Malik, Positive selection of primate
472 TRIM5α identifies a critical species-specific retroviral restriction domain. *Proc National*
473 *Acad Sci* **102**, 2832–2837 (2005).
474 44. D. Perez-Caballero, T. Hatziioannou, A. Yang, S. Cowan, P. D. Bieniasz, Human
475 Tripartite Motif 5α Domains Responsible for Retrovirus Restriction Activity and
476 Specificity. J. Virol. 79, 8969–8978 (2005).
477 45. D. M. Becker, V. Lundblad, Introduction of DNA into yeast cells. *Curr. Protoc. Mol.*
478 *Biol.* **Chapter 13**, Unit13.7 (2001).
479 46. D. Mumberg, R. Muller, M. Funk, Regulatable promoters of Saccharomyces
480 cerevisiae: comparison of transcriptional activity and their use for heterologous
481 expression. Nucleic Acids Res 22, 5767–5768 (1994).
482 47. K. A. Curran, et al., Short Synthetic Terminators for Improved Heterologous Gene
483 Expression in Yeast. *Acs Synth Biol* **4**, 824–832 (2015).
484 48. M. Valencia, et al., NEJ1 controls non-homologous end joining in Saccharomyces
485 cerevisiae. Nature **414**, 666–9 (2001).
486 49. M. Mirdita, et al., ColabFold: making protein folding accessible to all. *Nat. Methods*
487 **19**, 679–682 (2022).
488 50. J. Jumper, et al., Highly accurate protein structure prediction with AlphaFold.

Nature 596, 583–589 (2021).

51. L. Gerasimavicius, X. Liu, J. A. Marsh, Identification of pathogenic missense mutations using protein stability predictors. Sci. Rep. 10, 15387 (2020).

494 **Methods**

495 **Generation of base strains and plasmids**

496 All strains, plasmids, primers, and oligonucleotides are included in Supplementary
497 Tables 1-5.

498 All plasmids were transformed into NEB 5-alpha Competent *Escherichia coli* cells (New
499 England Biolabs) and purified using Qiaprep Spin Miniprep Kits or Plasmid Plus Maxi
500 Kits for plasmid libraries (QIAGEN). All PCR products were purified using QIAquick
501 PCR Purification Kits or MinElute PCR Purification Kits (QIAGEN). Plasmid and
502 genomic sequences were verified through Sanger sequencing (Eurofins Genomics,
503 Louisville, KY, USA). All yeast transformations were done according to standard lithium
504 acetate transformation procedures (45).

505 A plasmid carrying the two *acrIIA4* alleles was synthesized by Genewiz (MSp513). The
506 sequence of the second *acrIIA4* allele matched the sequence used by Basgall, et al.
507 (19). The sequence of the first *acrIIA4* allele was generated using custom python scripts
508 in which each codon was selected to share as few bases as possible with the
509 corresponding codon of the second *acrIIA4* allele. When multiple codons met that
510 criterion, the codon was randomly selected from the available codons with weights
511 matching codon usage in the *S. cerevisiae* genome. The first *acrIIA4* allele was
512 preceded by the *GAL*L promoter (46) and the second was followed by the T_{synth8}
513 terminator (47). The *acrIIA4* genes were flanked by 120-bp homology arms to the *HO*
514 locus. This Acr construct was amplified using PfuUltraII fusion HS DNA Polymerase
515 (Agilent Technologies) with the primers ANx87 and ANx88 and was transformed into

516 strain MS262 (MSy15), which had *NEJ1* deleted to eliminate non-homologous end-
517 joining as a repair pathways for DSBs (48) in favor of homology-directed repair
518 (Supplementary Table 1). To increase transformation efficiency, during the
519 transformation we directed a double-stranded DNA break at the *HO* locus by co-
520 transforming the Acr construct with a plasmid constitutively expressing SpCas9 from the
521 *TEF* promoter, acquired from Addgene (plasmid #: 162606; designated MSp181 in
522 Supplementary Table 2), and a PCR construct encoding a gRNA targeting the yeast *HO*
523 locus. The gRNA construct was generated by amplifying the targeting sequence from
524 pMS58 (MSp20) with primers ANx84 and ANx85 (Supplementary Table 3). Positive
525 transformants of the *acrIIA4* construct into the *HO* locus were selected by growth on
526 plates lacking uracil and leucine (CSM -Ura -Leu; Sunrise Science Products) and
527 confirmed by Sanger sequencing. A sequence-verified clone, in which MSp181 had
528 been spontaneously lost, was designated MSy261.

529 MSy261 was used to perform the first deletion scan replicate experiment, as described
530 in the section “Creation of acrIIA4 deletion alleles”. However, plasmids in the repair-
531 template plasmid library (whose generation is described in the section “Design of repair
532 templates”) that contained the sequence of codons 26-35 of the first copy of *acrIIA4*
533 were strongly depleted from the plasmid library for unclear reasons, presumably related
534 to that sequence posing a challenge either for oligonucleotide synthesis or library
535 cloning. We thus generated a second strain with these codons re-coded, MSy262,
536 which was used for the second and third deletion scan replicate experiments
537 (Supplementary Fig. 4B,C). This was done by amplifying two segments of the *acrIIA4*
538 construct on MSy261, upstream and downstream of codons 26-35, using Herculase II

539 Fusion DNA Polymerase (Agilent) and introducing the alternate sequence by including it
540 in the primers used for amplification (ANx80-83). To make MSy262, the two amplicons
541 were transformed into MSy14, which carries a galactose-inducible SpCas9 plasmid,
542 with the *HO*-targeting gRNA described above, and immediately plated onto CSM -Leu -
543 Ura galactose plates.

544 Our strategy of introducing deletions into *acrIIA4* involved cutting with SpCas9 between
545 the *acrIIA4* copies. For that step, we generated a maltose-inducible SpCas9 plasmid to
546 induce SpCas9 expression while maintaining repression of the galactose-inducible
547 *acrIIA4*, to prevent the anti-CRISPR activity of acrIIA4 from inhibiting the generation of
548 deletions. We inserted a DNA construct containing the *MAL63* transcription factor gene
549 and *MAL62* promoter sequence, ordered from Genewiz (MSp514), upstream of the
550 SpCas9 gene on MSp14. We named the resulting plasmid MSp500, and it has been
551 deposited on Addgene as plasmid # 227704. The sequence of the *MAL* genes was
552 taken from the genome sequence of YJM145 (MSY28) (18). Amplification of the *MAL*
553 locus sequence was done with KAPA Hi-Fi Hotstart DNA Polymerase (and primers
554 ANx16.5 and 17.5), and inserted into MSp14 via Gibson assembly (New England
555 Biolabs, Ipswich, MA, USA) following a digestion of the plasmid with SpeI-HF (New
556 England Biolabs).

557 MSy261 was transformed with a plasmid carrying a galactose-inducible SpCas9, p415-
558 GalL-Cas9-CYC1t (Addgene plasmid #:107922; MSp6) and transformants were
559 selected on CSM -Ura -Leu glucose plates. One colony was picked and transformed
560 with MSp500, with transformants selected on CSM -Ura -Leu + clonNAT (100ug/mL)
561 glucose plates (MSy263). MSy262, already carrying a galactose-inducible SpCas9, was

562 also transformed with MSp500 and plated on CSM -Ura -Leu + clonNAT (100ug/mL)
563 glucose plates to produce MSy264.

564

565 **Design of repair templates**

566 Custom python scripts were used to generate repair template sequences for directing all
567 possible in-frame deletions within *acrIIA4*. In brief, to generate the repair template for
568 the deletion from the *m*th to the *n*th codon, we appended the 100 nt up to the *m*th codon
569 of the first *acrIIA4* allele to the 100 nt after the *n*th codon of the second *acrIIA4* allele.
570 For deletions close to the start of *acrIIA4*, their repair templates included sequence from
571 the *GAL*L promoter preceding the first copy, and for deletions close to the end, their
572 repair templates included the *acrIIA4* terminator and the sequence of the *HO* locus
573 flanking the insertion site. We included the start codon and stop codon as deletable
574 codons. We also designed repair templates that generated a clean chimera of the two
575 *acrIIA4* alleles into a single *acrIIA4* without deletion, for 85 repair templates. Once all
576 sequences were created, 15-nt primer sites were added to the beginning and end of the
577 strings, making all sequences 230 characters long. Sequences were then ordered as an
578 oligonucleotide pool through Agilent (Santa Clara, California, USA). This process was
579 performed once to generate the repair templates used for replicate 1, and then again
580 with the modified sequence of the first *acrIIA4* allele to generate the repair templates
581 used for replicates 2 and 3 (Supplementary Tables 4 and 5). For the purpose of this
582 experiment, we proceeded with data analysis of 3,741 sequences, excluding deletions
583 of the start or stop codons, as well as repair templates that generated clean chimeras.

584 MSp218 was used as the backbone of the three replicate repair template plasmid pools,
585 and was generated from pMEL16 (MSp50), a *HIS3* SpCas9 gRNA plasmid acquired
586 from Jean-Marc Daran (Addgene plasmid #:107922). For cloning in the repair
587 templates, we engineered EcoRI and BamHI sites downstream of the gRNA gene. In
588 addition to carrying the repair templates, we designed MSp218 to encode a gRNA
589 sequence targeting a sequence inserted between the two *acrIIA4* alleles,
590 ACGTATCGGTACATGCGCAT. This sequence was picked to have equal
591 representation of each of the four bases, as well as to have at least six mismatches to
592 any other SpCas9 target sites in the yeast genome. MSp218 has been deposited on
593 Addgene as plasmid # 227698.

594 We amplified the repair template oligonucleotide pools using primers that additionally
595 added 12 nt of random DNA on each end to serve as barcodes as well as homology
596 arms for Gibson assembly (ANx23 and 24); PCR was performed with PfuUltra II Fusion
597 High-fidelity DNA Polymerase for 15 cycles. MSp218 was double digested using EcoRI-
598 HF and BamHI-HF (New England Biolabs) on the cloning site, and the oligo amplicons
599 were inserted via Gibson assembly. 50uL of Gibson assembly were transformed into
600 1000uL of 5-alpha competent *E. coli* cells for each replicate (New England Biolabs).
601 Replicate 1 produced 100,000 colonies, Replicate 2 produced 100,000 colonies, and
602 Replicate 3 produced 80,000 colonies. Colonies were washed off plates and plasmid
603 libraries were isolated with the Qiagen Plasmid Plus Maxi kit.

604

605 **Creation of *acrIIA4* deletion alleles**

606 The repair template plasmid pools were transformed en masse (2.05 µg of Replicate 1;
607 3.5 µg of Replicates 2 and 3) into approximately 10^9 cells of MSy263 or MSy264 as
608 appropriate, and plated on CSM -Ura -Leu -His +clonNAT glucose plates. Replicate 1
609 produced 260,000 colonies, Replicate 2 produced 300,000 colonies, and Replicate 3
610 produced 250,000 colonies. Cells were then harvested by washing the plate with
611 glucose -H,L,U +clonNAT liquid media, and a 100-mL culture of these cells was then
612 started at an optical density (OD_{600}) of 0.002 using the same media. The culture was left
613 to grow overnight at 30ºC until it reached an OD_{600} of 1. Approximately 2×10^9 cells
614 were then plated over multiple CSM Maltose -His -Leu +cloNAT plates to induce
615 SpCas9 from MSp500, resulting in 77,000 colonies for Replicate 1, 66,000 colonies for
616 Replicate 2, and 80,000 colonies for Replicate 3. Cells were harvested by washing
617 plates with YPD, and approximately 4×10^7 cells were replated onto 5-fluoroorotic acid
618 (5-FOA) -Leu -His plates. On 5-FOA plates, Replicate 1 produced 1.6×10^6 colonies,
619 Replicate 2 produced 1×10^6 colonies, and Replicate 3 produced 1×10^6 colonies.
620

621 **Characterization of *acrIIA4* deletion allele function**

622 We generated a plasmid carrying a gRNA that directs CRISPR cleavage of the essential
623 gene *ARB1* by inserting an *ARB1*-targeting gRNA sequence into pLK88 (MSp11), a
624 *URA3* high-copy plasmid from Leonid Kruglyak. An oligonucleotide containing the gRNA
625 sequence was ordered from Integrated DNA Technologies (ANx86) and inserted by
626 Gibson assembly into MSp11 following a double digestion with SphI-HF and BstEII-HF
627 (New England Biolabs). This plasmid was named MSp501 and has been deposited on
628 Addgene as plasmid # 227705.

629 Yeast cells with *acrIIA4* deletion alleles were washed off 5-FOA plates using CSM -Leu
630 -His glucose liquid media. A 50-mL culture was started at OD_{600} of 0.01 using the same
631 liquid media, and it was left to grow overnight at 30ºC. Once grown overnight, the
632 culture was used to start a new culture in YPD starting at OD_{600} of 0.125, allowed to
633 double twice, and then transformed with MSp501. In Replicate 1, 1.5 µg of MSp501 was
634 transformed into approximately 10^9 cells; in Replicates 2 and 3, 2 µg were transformed
635 into approximately 10^9 cells. Transformations were plated on CSM -Leu -His -Ura
636 glucose plates, producing 300,000 colonies in Replicate 1, 600,000 colonies in
637 Replicate 2, and 800,000 colonies in Replicate 3. Plates were then scraped once more
638 and 1×10^6 cells were plated on CSM -Leu -His -Ura galactose plates to simultaneously
639 induce the *acrIIA4* deletion alleles and SpCas9 to target *ARB1*. An equal number of
640 cells were plated on CSM -Leu -His -Ura glucose plates as controls. On glucose,
641 Replicate 1 produced 1×10^6 colonies, Replicate 2 produced 900,000 colonies, and
642 Replicate 3 produced 1×10^6 colonies. On galactose, Replicate 1 produced 300,000
643 colonies, Replicate 2 produced 280,000 colonies, and Replicate 3 produced 370,000
644 colonies.
645

646 **Data collection and sequencing**

647 Yeast genomic DNA was extracted from $4x10^7$ cells (per library) harvested off of the 5-
648 FOA plates, galactose plates, and final glucose plates using the DNeasy Blood & Tissue
649 Kit (QIAGEN). We then amplified the chimeric *acrIIA4* locus from the genomic DNA,
650 using primers containing overhangs for attaching Illumina adapters. The forward primers
651 for the genomic DNA annealed upstream of the first *acrIIA4* allele and the reverse

652 primers annealed downstream of the second allele, such that they would amplify all
653 chimeric *acrIIA4* alleles. Additionally, yeast plasmid DNA was extracted from $5x10^7$ cells
654 harvested off of galactose and final glucose plates of replicates 2 and 3 using Qiaprep
655 Spin Miniprep Kit (QIAGEN). We amplified the 200-bp repair template region from the
656 initial *E. coli* plasmid pools of replicates 1 and 2 and the yeast plasmid extractions from
657 replicates 2 and 3. Genomic primers (ANx25-34) and plasmid primers (ANx35-44) were
658 staggered to maintain sequence diversity on the Illumina flow cell. All PCRs were
659 performed using KAPA HiFi Hotstart ReadyMix DNA Polymerase for 20 cycles. A
660 second PCR was performed to attach the Nextera adapter index primers using KAPA
661 HiFi Hotstart ReadyMix DNA Polymerase for 10 cycles, after which samples were
662 prepared for sequencing according to Illumina Miseq sequencing protocols.In total, four
663 Miseq runs using Illumina Miseq 500 cycle V2 kits were performed: once for sequencing
664 the genomic reads and *E. coli* plasmid pool of replicate 1, once for sequencing the *E.*
665 *coli* plasmid pool of replicate 2, and once for sequencing the genomic and plasmid
666 reads of replicates 2 and 3.

667

668 **Data analysis**

669 PEAR software (Version 0.9.8) was used to process and align forward and reverse
670 Miseq reads. Custom python and bash scripts were used to match Miseq sequencing
671 reads of the genomic and plasmid DNA amplicons to the possible *acrIIA4* deletion
672 alleles and generate read counts of perfectly matched reads to each deletion in each
673 library. For the individual glucose and galactose genomic libraries, counts were first

674 normalized by dividing the number of reads corresponding to a specific deletion by the
675 total amount of reads in that library. We then divided all galactose values by their
676 corresponding glucose value to calculate the enrichment of *acrIIA4* deletions in cells
677 performing a lethal CRISPR edit. The analysis of deletion tolerance primarily utilized the
678 read counts of deletion alleles amplified from the yeast genomic DNA. Analyzing
679 replicate-replicate correlation plots between the three replicates, we found good
680 agreement (r of 0.65, 0.68, and 0.68) when we capped gal/glu ratios at 30
681 (Supplementary Fig. 5A), which is slightly higher than the ratio for WT AcrIIA4. This
682 indicated that variation above 30 was due to noise rather than biological significance,
683 probably arising from variation in PCR amplification. We thus used this cap in
684 downstream analyses. We also found higher correlations values between replicates for
685 deletions of 10 or fewer residues (r of 0.79, 0.79, and 0.76) (Supplementary Fig. 5B),
686 which we inferred was due to noise introduced by occasional spontaneous escaper
687 mutants carrying large nonfunctional acrIIA4 deletions to high frequency. To minimize
688 the effects of such escaper mutants, when merging the data from the three replicates
689 for display in Fig. 2, we took the median of the galactose/glucose read count values for
690 a given deletion across the replicates. A deletion was only counted as observed in a
691 given replicate if it had at least five reads in glucose. If a deletion was not observed in
692 one replicate, the lower of the remaining two values was taken. If there were no counts
693 recorded for a given deletion in all three replicates, it was represented by a blank pixel
694 on the heat map. For comparison, we also took the mean of the galactose/glucose read
695 count values across replicates, and found that the dispensability of each analyzed
696 domain was unaffected (Supplementary Fig. 6). We also performed a parallel analysis

697 using the plasmid reads, and found that repair templates were enriched following
698 CRISPR induction corresponding to the genomic AcrIIA4 deletion alleles we found to
699 maintain anti-CRISPR function (Supplementary Fig. 7).

700 To determine the statistical significance of the domains that appeared to be dispensable
701 (β1-β2 loop, β3-α2 loop, and the end of the α3 helix), we calculated the mean of the
702 functional scores in the smallest triangular area encompassing the entirety of each
703 domain, for each replicate. We then acquired the mean of an equal number of randomly
704 selected data points as were contained in each triangular area and repeated this 1,000
705 times (Supplementary Fig. 1) for comparison by the Monte Carlo method.

706 To find naturally occurring AcrIIA4 homologs with deletions relative to the canonical
707 AcrIIA4, we performed a BLASTp search on the NCBI BLASTp server using standard
708 parameters. We used custom R scripts to identify homologs with deletions, and
709 manually verified that the deletions did not result from incomplete sequencing.
710 ColabFold (version 1.5.5) (49), an open-source version of AlphaFold2 (50), was used to
711 generate the structural models of AcrIIA4 homologs and deletion alleles through
712 ChimeraX; of the predicted structures for each sequence, the highest confidence
713 structure was used (restricting to relaxed structures for the predicted structures for
714 deletion alleles). We then used the “stability” command in FoldX (Version 5.0) (22) to
715 calculate the Gibbs free energy of folding (ΔG) for each deletion allele’s predicted
716 structure. We subtracted the calculated ΔG for the wild-type AcrIIA4 structure from the
717 ΔG value for each deletion to obtain ΔΔG values. We thresholded ΔΔG values at a
718 maximum deleteriousness of 1.5, as that has previously been recommended as the
719 threshold for destabilizing mutations (51). The analysis was not substantially changed

720 when a much more permissive threshold of 15 was used (Supplementary Fig. 3B). For
721 prediction of protein function using ESM1b, we used scripts from Brandes et al. for all
722 deletions up to 10 amino acids in length (24). For prediction of protein function for all
723 possible deletions using PROVEAN (25), we used the precise software version
724 combinations of psiblast and blastdbcmd from ncbi-blast-2.4.0+, cd-hit version 4.8.1,
725 PROVEAN version 1.1.5 and the RefSeq non-redundant database v4.

726 All statistical measurements and visualizations were done with custom R scripts
727 (Versions 4.2.2 and 4.3.2) and R package ggplot2 (Versions 3.4.3 and 3.5.0) through
728 RStudio. Crystal structure visualizations were generated with the PyMOL molecular
729 visualization system, Versions 2.5.0 and Version 3.0.0 Open-Source, Schrödinger, LLC.

730

731 **Data Availability**

732 - Python and R scripts:

733 https://github.com/annetteig/AcrIIA4-Functional-Domains

734

735 **Acknowledgements**

736 We thank Bryan Thurtle-Schmidt, Elena Zehr, two anonymous reviewers, and all
737 members of the Sadhu lab for helpful discussion. We thank Jean-Marc Daran, Lars
738 Steinmetz, and Leonid Kruglyak for strains and plasmids. This work utilized the
739 computational resources provided by the National Institutes of Health (NIH) HPC
740 Biowulf Cluster (http://hpc.nih.gov). This work was supported by the Intramural

Research Program of the National Human Genome Research Institute, NIH (1ZIAHG200401).

Author contributions

A.B.I., C.A.W., and M.J.S. conceptualized research; A.B.I., C.A.W., S.M.G., and M.J.S. performed formal analysis; A.B.I. performed investigation; A.B.I., C.A.W., and M.J.S. developed methodology; C.A.W. and M.J.S. performed supervision; A.B.I., C.A.W., S.M.G., and M.J.S. prepared visualizations; and A.B.I., C.A.W., and M.J.S. wrote the paper. All authors read and revised the paper.

Disclosure and competing interest statement

The authors declare no competing interests.

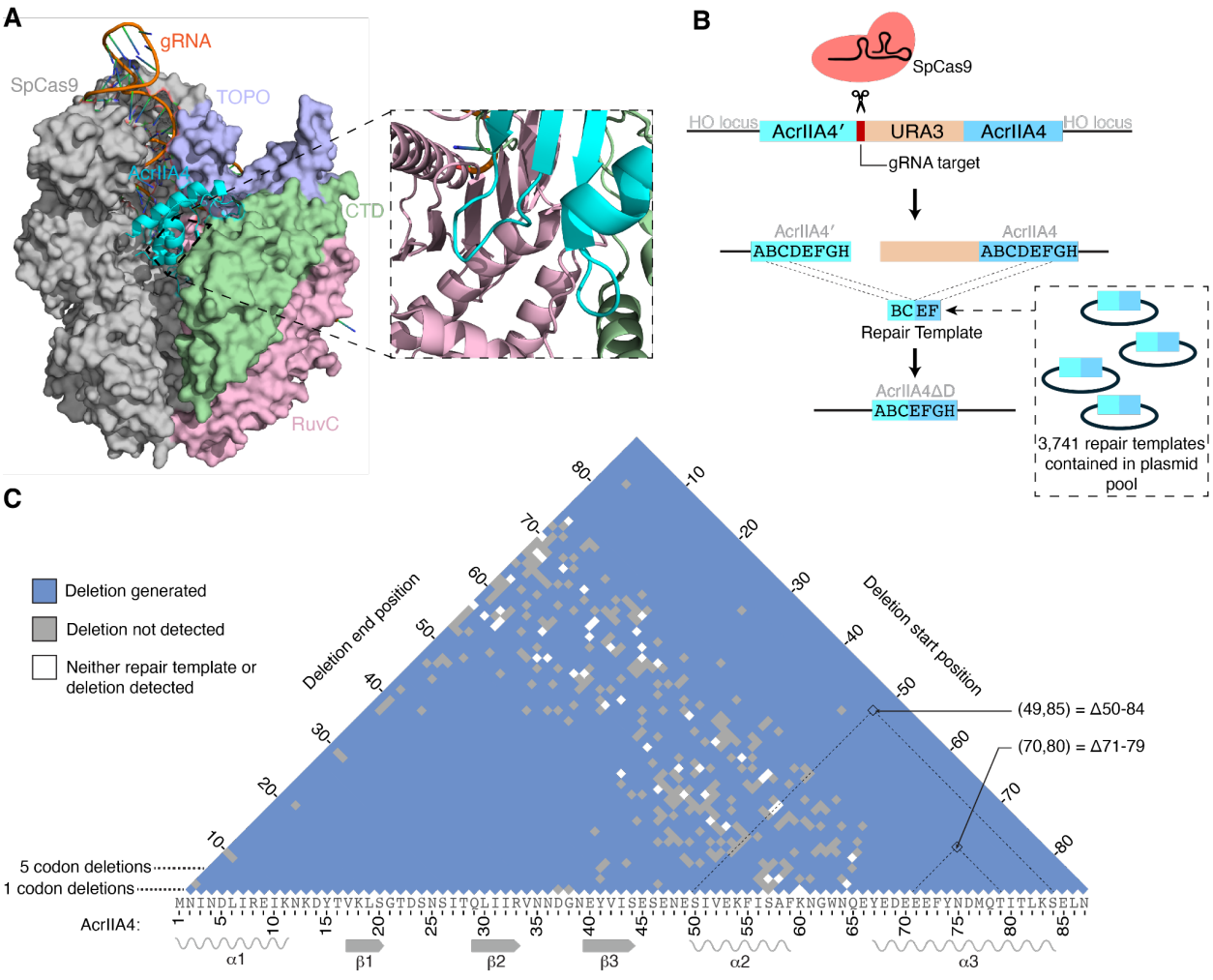


Figure 1. Comprehensive generation of in-frame deletions in AcrIIA4, an inhibitor of SpCas9. (A) Structure of AcrIIA4 bound to SpCas9 (PDB 5VW1) (7). AcrIIA4 (cyan) binds SpCas9's CTD (green), TOPO (lavender), and RuvC (pink) domains. The inset shows the AcrIIA4-RuvC binding interface, which is largely obscured in the larger view by CTD. (B) Strategy for generating all possible in-frame deletions in AcrIIA4 in yeast. CRISPR cleavage using a constant gRNA is followed by repair directed by one of 3,741 deletion-generating repair templates. (C) High-throughput sequencing of AcrIIA4 alleles following deletion introduction identified 87% of the possible in-frame deletions, spanning the range of possible deletion sizes.

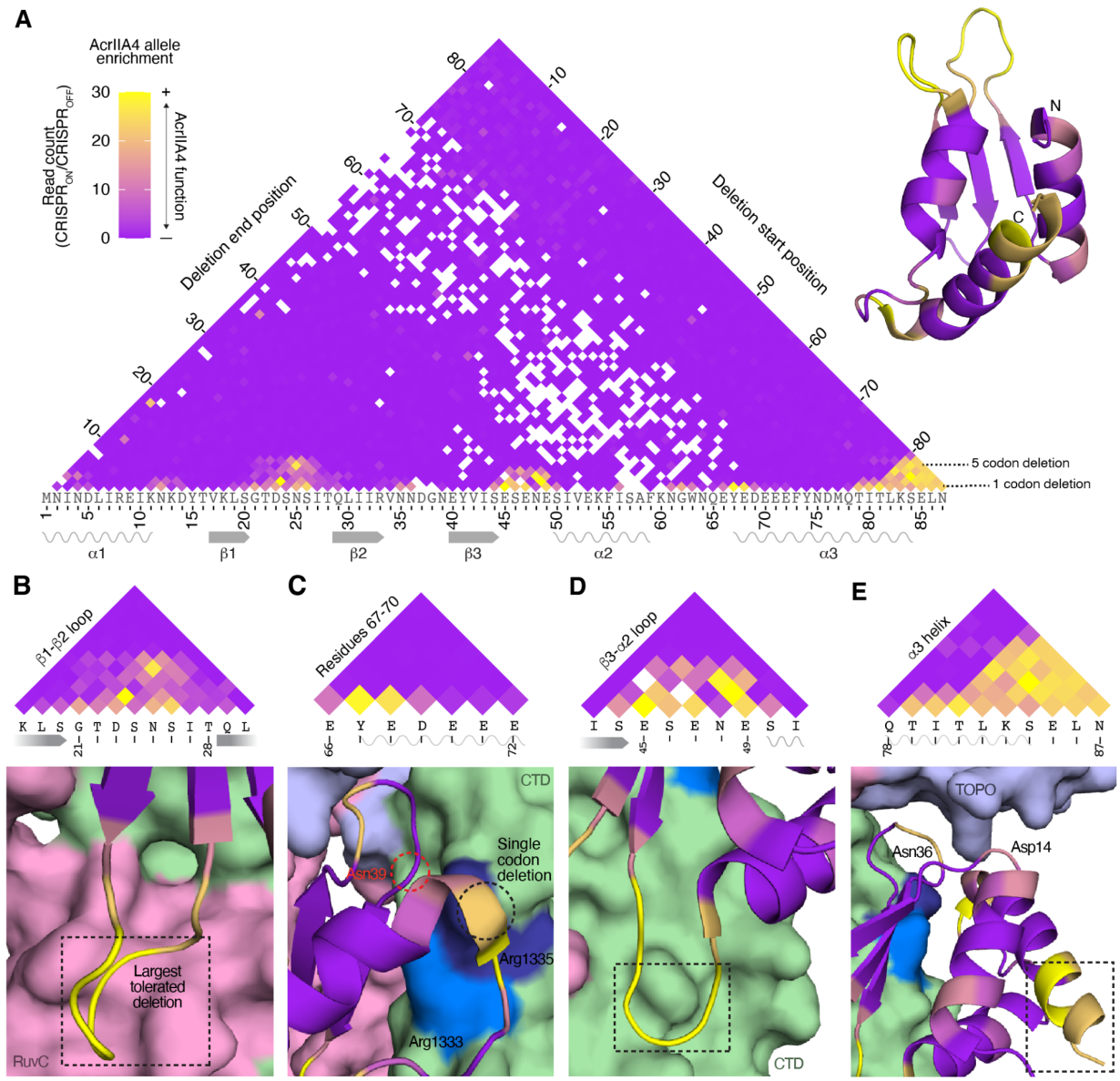
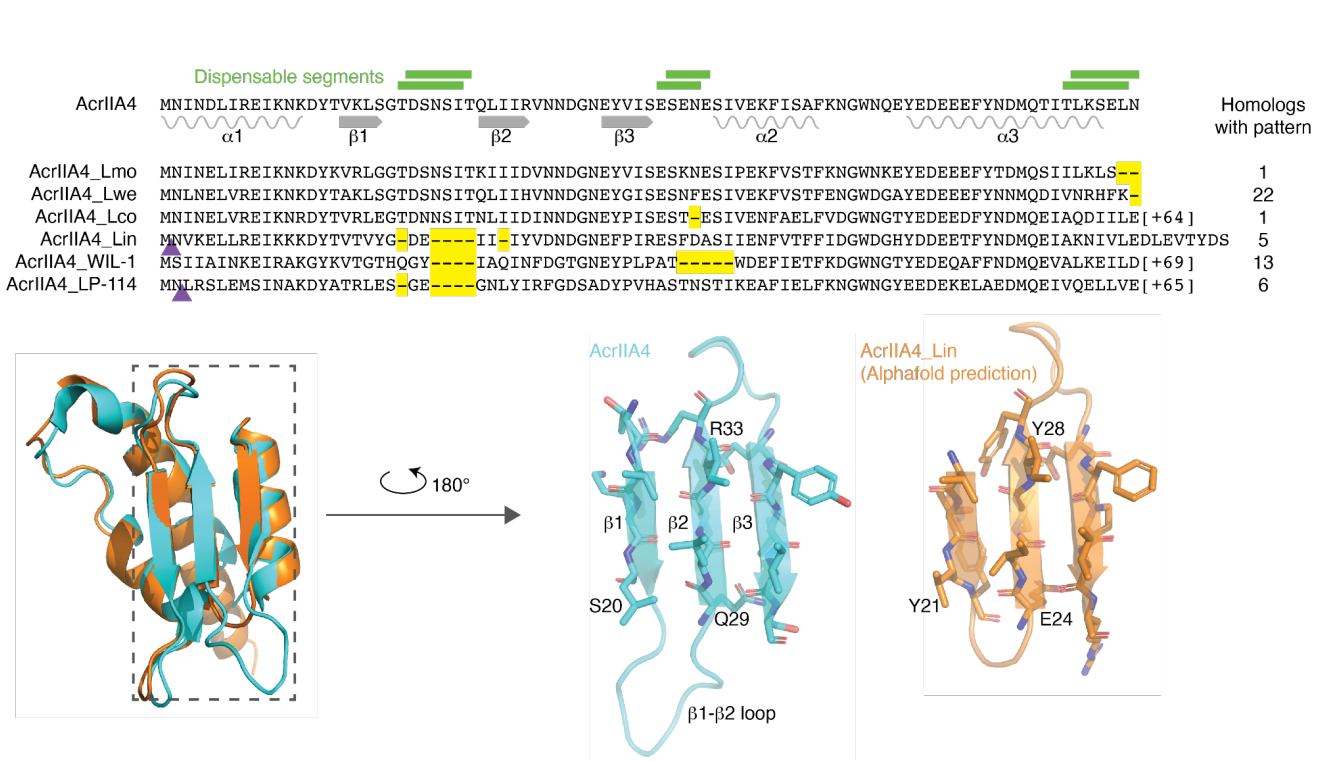


Figure 2. Functional characterization of AcrIIA4 deletion alleles. (A) Functionality of AcrIIA4 deletion alleles, determined by their read counts in CRISPR$_{ON}$ conditions normalized against CRISPR$_{OFF}$ read counts. Deletions are colored by their median value across three replicate experiments to minimize the effects of outlier escaper mutants; alleles only present in two replicates are represented by the lower of the two values. Residues in the AcrIIA4 crystal structure (right) are colored by the maximum

771 functionality of deletions encompassing that residue (six-residue deletions or shorter).
772 (B-E) Dashed boxes demarcate the largest tolerated AcrIIA4 deletion in the view. (B)
773 Large dispensable regions within the RuvC-interacting β1-β2 loop. (C) AcrIIA4 residues
774 67-70, which interact with Arg1333 and Arg1335, are tolerant of single amino acid
775 deletions. Asn39, which interacts with Arg1335, is not dispensable. (D) Large
776 dispensable regions within the CTD-interacting β3-α2 loop. (E) Asp14 and Asn36,
777 TOPO-interacting residues, are partially dispensable. The end of the C-terminal helix α3
778 is dispensable; note that the final two amino acids are not resolved in the crystal
779 structure but were also dispensable.



780

781 **Figure 3.** Deletions in naturally observed homologs of AcrIIA4 occur in experimentally
782 determined dispensable segments. (A) Alignment of six AcrIIA4 homologs representing
783 the identified deletion patterns. Yellow dashes signify deletions relative to AcrIIA4, while
784 purple triangles demarcate the positions of insertions. The positions of select segments
785 identified as dispensable in Fig. 2 are illustrated with green lines above the AcrIIA4
786 sequence. From top to bottom, the AcrIIA4 homolog NCBI IDs and source are
787 EAD0751510.1 isolated from *L. monocytogenes*; MBC1462049.1 from *L. welshimeri*,
788 WP_185587529.1 from *L. cossartiae*, WP_185342258.1 from *L. innocua*,
789 YP_010843666.1 from phage WIL-1, and YP_009045111.1 from phage LP-114. (B) The
790 AlphaFold-predicted structure of AcrIIA4_Lin has a full-length β2 strand.

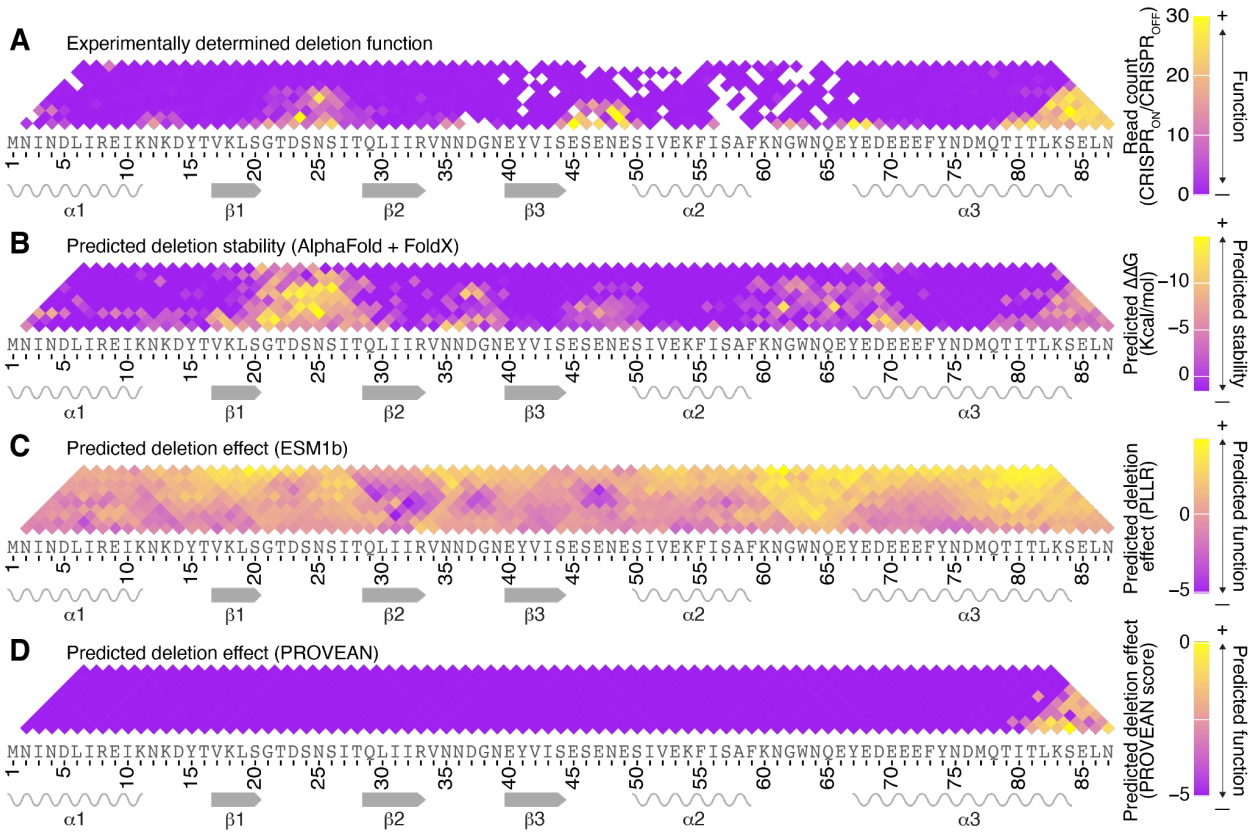


Figure 4. Predictions of the effect of deletions of 10 amino acids or fewer in AcrIIA4. (A) Our experimentally determined deletion effects. (B) The effect of deletions on AcrIIA4 stability predicted using AlphaFold2 to predict protein structures and FoldX to calculate the Gibbs free energy of folding, and then determining the change in free energy from the wild-type AcrIIA4 prediction. (C) Deletion effects predicted by ESM1b, calculated as the pseudo-log likelihood ratio (PLLR) between the wild-type and deletion variants. (D) Deletion effects predicted by PROVEAN.