Expanding the scope of phage-assisted continuous protein binding selections

G4 progress update submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy

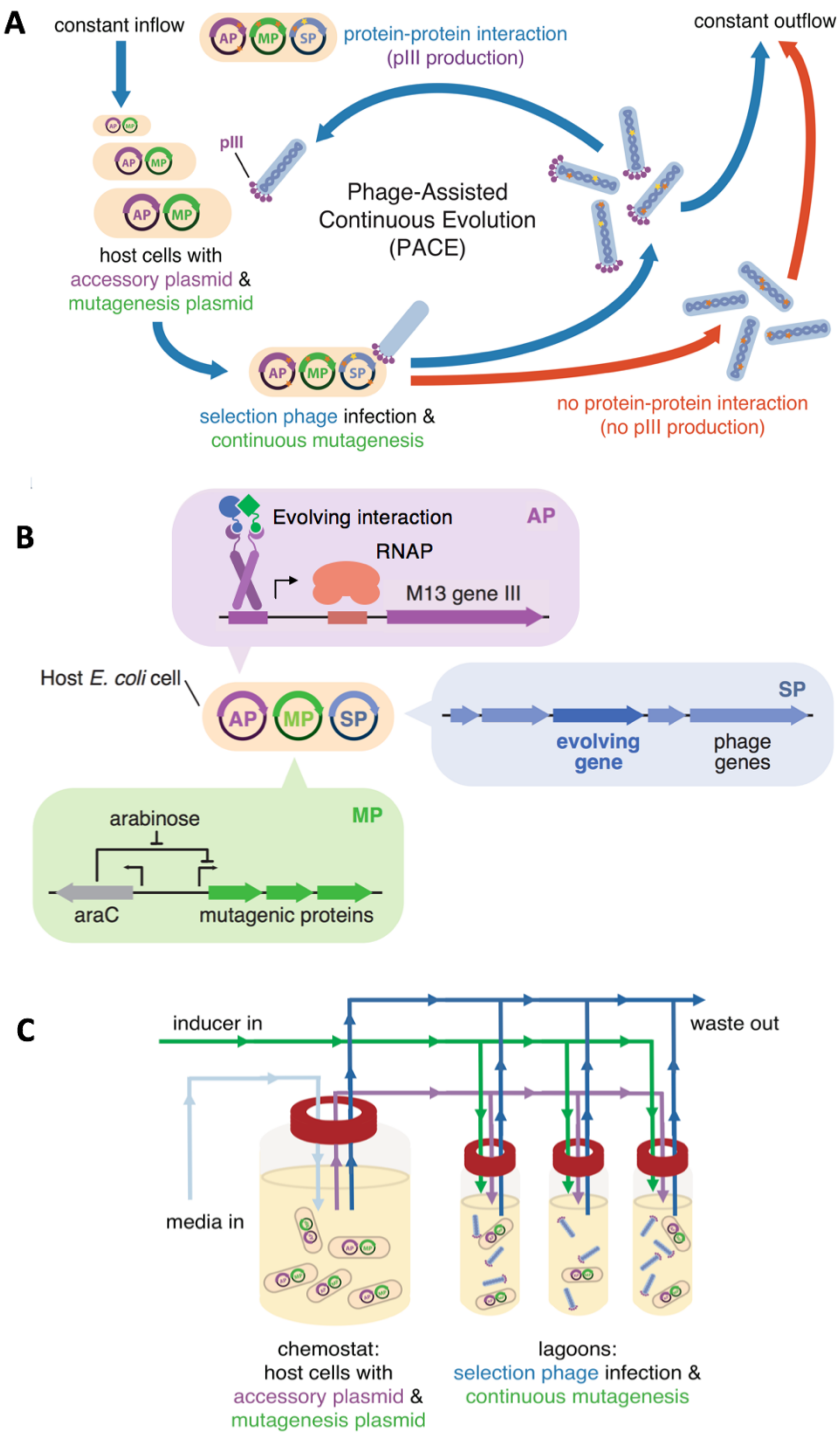
Mary Morrison
Molecules, Cells, and Organisms
Department of Molecular and Cellular Biology,
Harvard University

May 10, 2019
1:00-3:00PM
Rm. 3013, 75 Ames St.
Broad Institute of Harvard & MIT

Mary Morrison

David R. Liu

I. A periplasmic phage-assisted continuous evolution platform for the rapid evolution of antigen/antibody interactions



I have developed a system for the rapid phage-assisted continuous evolution (PACE) of disulfide-rich immunoprotein interactions in the bacterial periplasm. The system, termed periplasmic PACE or pPACE, makes use of the *E. coli* CadCBA regulon to couple phage fitness to a periplasmic binding event. I have employed pPACE to improve the binding of a periplasmic homodimer and of an antibody to its cognate antigen. The platform demonstrates the rapid ability to distinguish between a wild-type antibody and a nonbinding point mutant, and further to correct that point mutant.

Fig. 1: Phage-Assisted Continuous Evolution. A-B) Bacterial 2-hybrid selection schema & general PPACE system overview (Figure courtesy of A. Badran). C) Physical layout of PACE setup[1]. Host cells are maintained in constant culture in a chemostat. The chemostat supplies lagoons that contain evolving phage.

In PACE, an evolving protein of interest (POI) is encoded by M13 selection phage in place of essential phage gene III. The desired biological function of the POI drives gene III transcription in the host, permitting the assembly of infective phage (Fig. 1).[2]

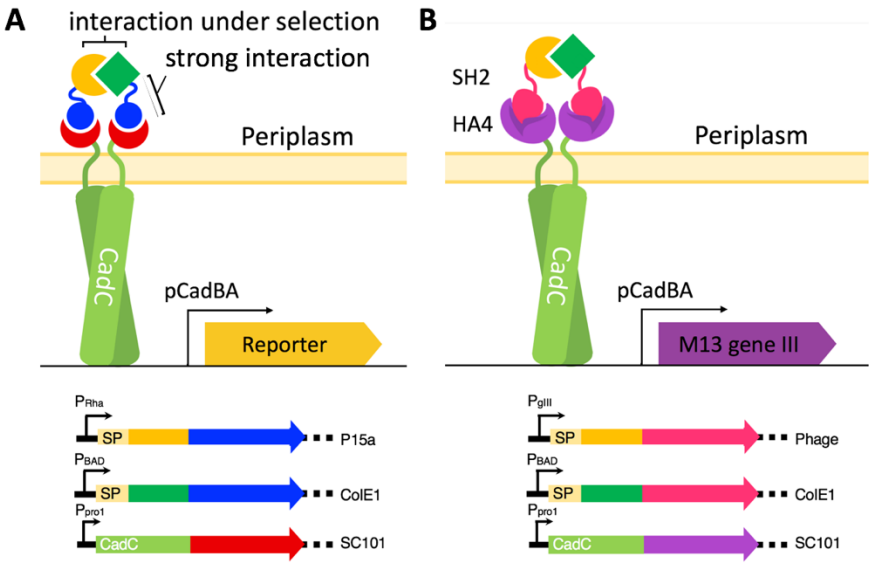
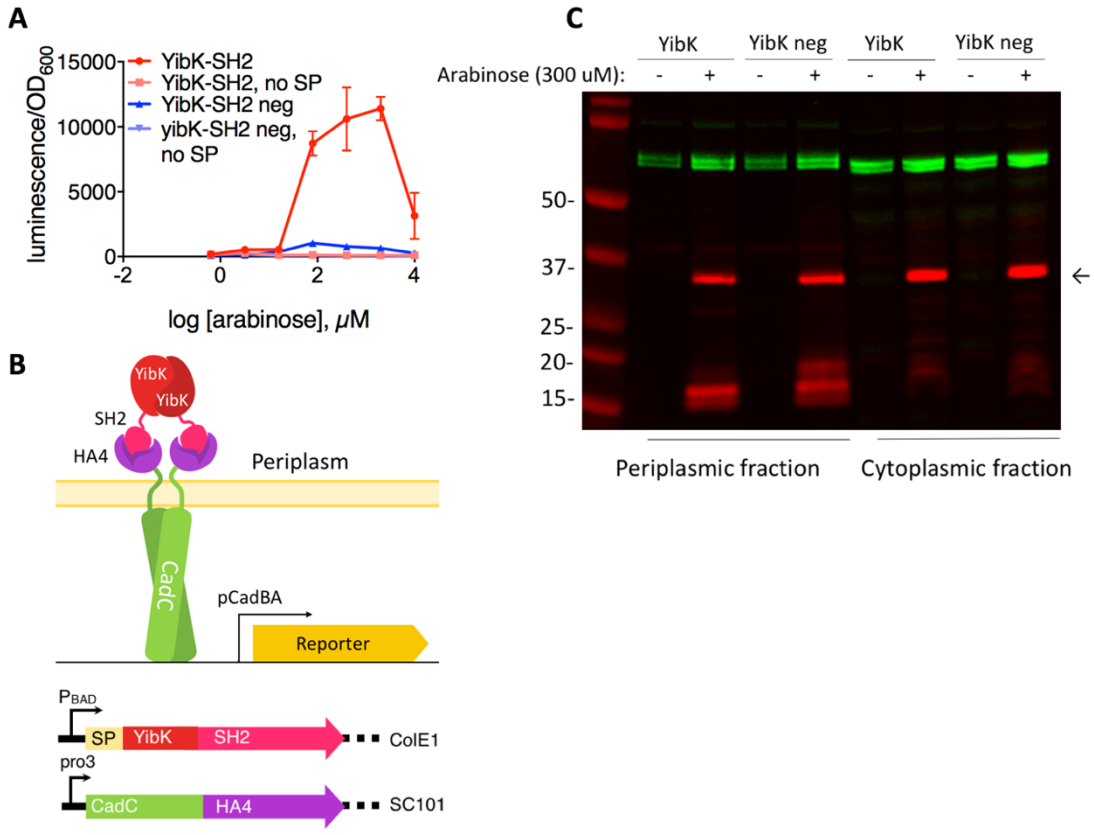


Fig. 2: Four-part complex design. A) Symmetric selection schematic for plasmid-based, phage-free transcriptional assays. B) Asymmetric Selection schematic for PPACE. HA4 fibronectin monobody binds withsub-nanomolar affinity to human protein SH2.[3]

Transmembrane protein CadC is involved in acidic stress regulation in *E. coli.* The periplasmic sensory domain of CadC dimerizes under acidic stress, activating transcription from pCadBA through a cytoplasmic DNA-binding domain.[4] After extensive optimization of promoter, dosing and host strain, I have reprogrammed the CadCBA operon to drive transcription of gene III in response to a periplasmic binding event (Fig. 2).



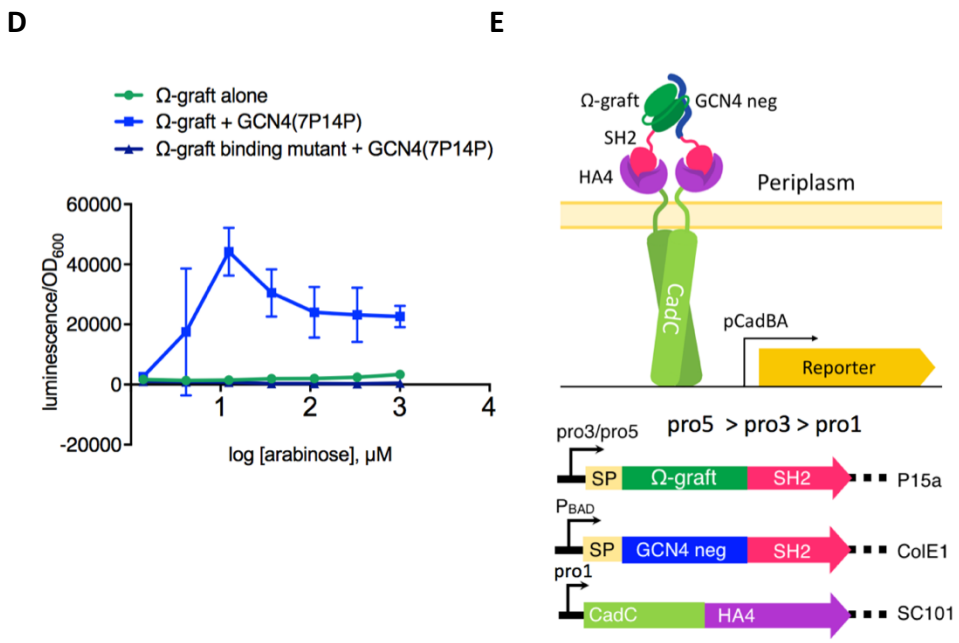
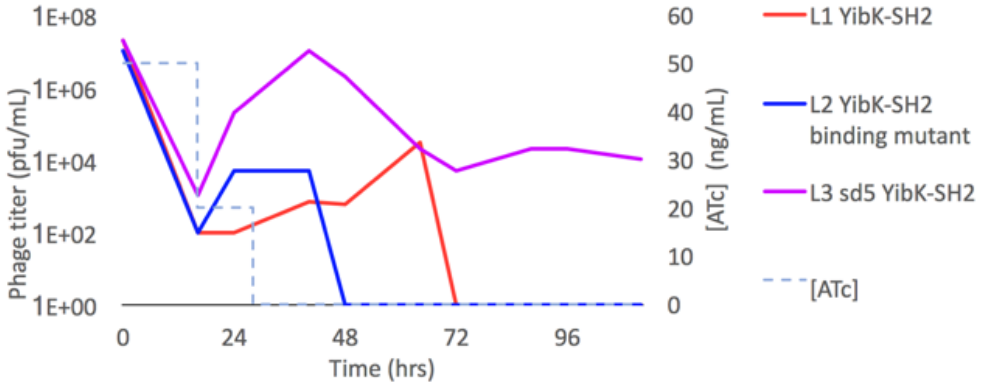


Fig. 3: Test of four-part complex activity and localization. A) YibK and YibK neg (mutant V139R) expressed as –SH2 fusions +/- signal peptide (SP) according to schema shown in B). C) Western blot of YibK-SH2 periplasmic extraction. YibK-SH2: 38kDa (Red: α-SH2. Green: α-GroEL). Arrow indicates YibK-SH2 (34kDa). D) Transcriptional activation assay with GCN4 driven by arabinose. GCN4(7P14P) does not homodimerize. Neg: Ω-graft mutant.[13] E) Schematic of heterodimeric transcriptional assay shown in D).

I carried out pPACE on wild-type (reported kd <1pM) and binding mutant (reported kd = 360uM)[5] forms of homodimeric protein YibK (Fig. 10A). W145C is believed to form an intermolecular disulfide with nearby unpaired cysteine 83 (Fig. 10D). The selection platform was then applied to YibK binding mutant V139R (Kd=360μM) in a non-continuous serial dilution format, and was able to identify both a compensatory point mutation believed to replace a hydrophobic interaction with a salt bridge, and an additional interface mutation R146C, believed to result in an intramolecular disulfide.

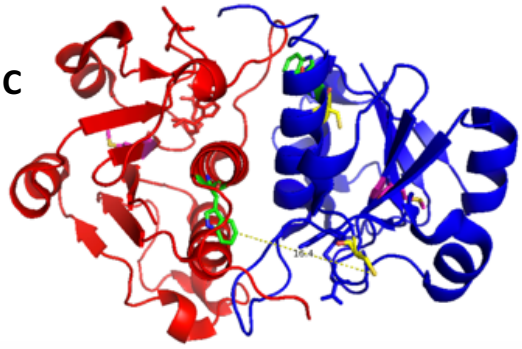
A



B

YibK point mutations						
Phage	H30	R43	K92	M98	R104	W145
L3-112-1				V		C
L3-112-2		C	N			
L3-112-3				V		C
L3-112-4	Y			V	C	
L3-112-5				V		C
L3-112-6				V		C
L3-112-7				V		C
L3-112-8				V		C
L3-112-9				V		C
L3-112-10				V		C

C



D

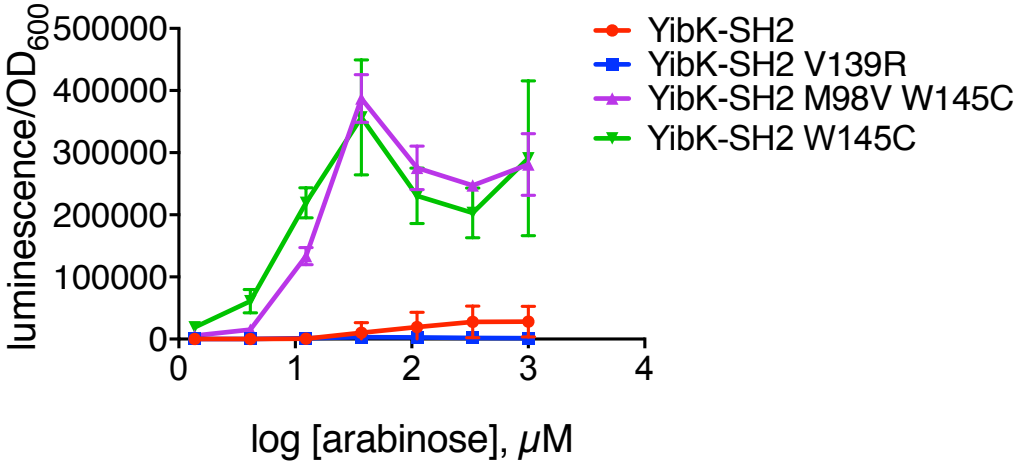
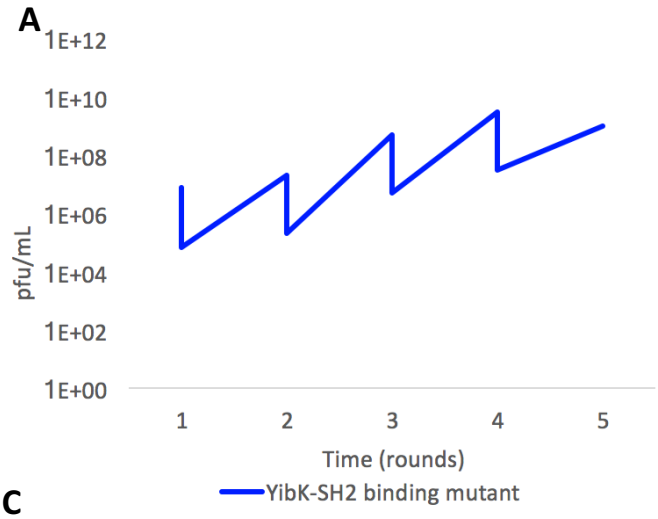
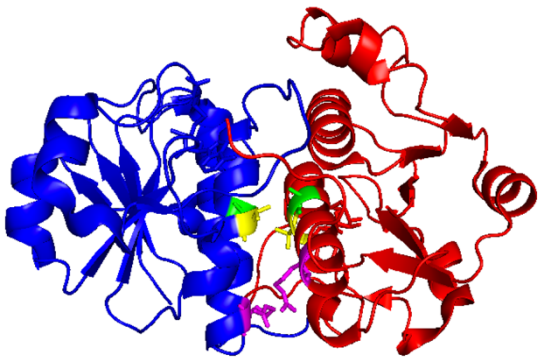
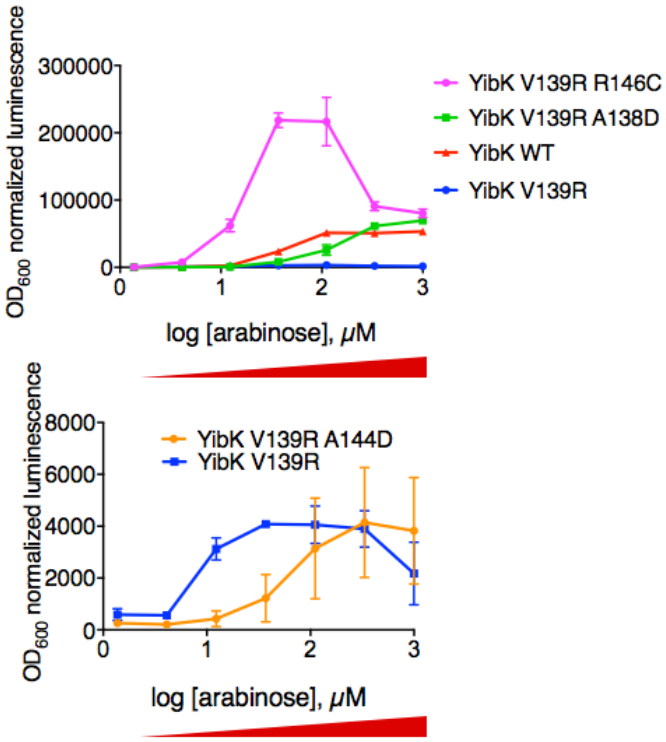


Fig. 4: YibK pPACE. PACE was carried out with regular sampling, at a media flow rate of 1 vol/hr with free pIII ('drift') under the control of [ATc]. A) Phage titers over time. B) Consensus and single YibK mutations from PPACE. Ten phage were sampled and sequenced at 112 hrs. C) M98 (magenta) and W145 (green) in the YibK homodimer. In the righthand YibK domain (blue), free cysteines are indicated in yellow. W145 is 16.4 angstroms away from the nearest cysteine, located in a flexible loop region. D) YibK-SH2 constructs with consensus mutations were subcloned into a plasmid under P_{BAD} for transcriptional binding assays.





C



D

	Point mutations					
	YibK V139R					
Phage	A96	A138	A144	R146	A153	V159
2-day3-1			D	C		
2-day3-2			D	C		
2-day3-3			D	C		L
2-day3-4			D	C	T	
2-day3-5			D	C		
2-day3-6			D	C		
2-day3-7		D	D	C		
2-day3-8	C					

Fig. 5: YibK batch culture to correct negative point mutation. Phage were passaged daily through fresh cells at mid-log. A) Phage titers over time. B) A138 (green) and R146 (magenta) in the YibK homodimer. V139R is shown in yellow. C) Consensus and single YibK mutations from batch culture. Eight phage were sampled and sequenced on day 3. D) YibK-SH2 consensus mutations were subcloned into a plasmid under P_{BAD} for transcriptional binding assays. **Continuous flow PACE to determine whether A138D rises in frequency still needed; Expected by July.**

Fig. S1: PLACEHOLDER: Nonreducing and reducing PAGE to assess formation of disulfide bonds. Expected by July.

When overexpressed, single-chain variable fragment antibodies swap heavy and light chains, causing homodimerization and bypassing the selection.[6] To maintain weak antibody expression, I implemented a split intein system. In this system, the signal peptide directing insertion of antibodies in the periplasmic space is split in half, with the N-terminal half provided by a plasmid in the host cell.

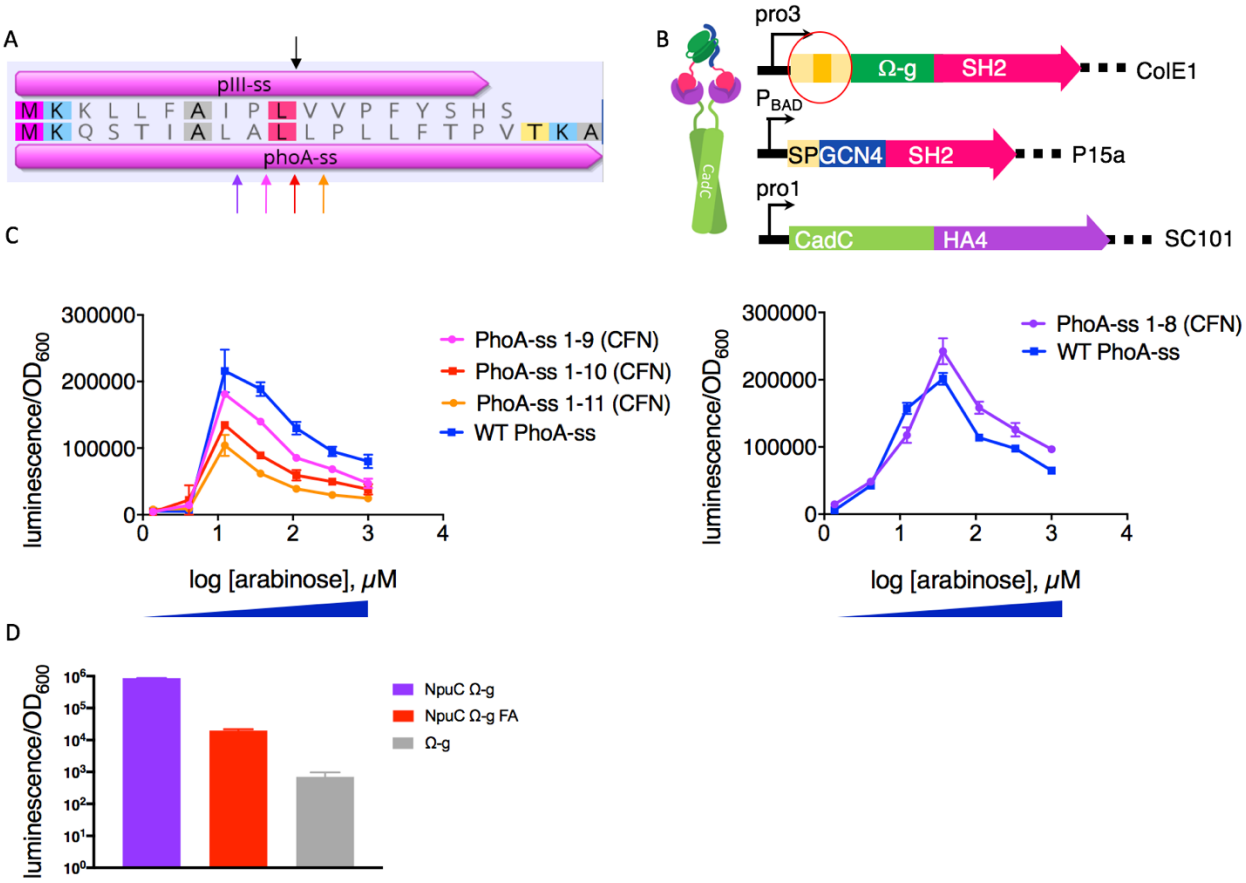


Fig. S1: Antibody intein-gated periplasmic export system development. A) The Npu intein CFN scar was inserted into the phoA signal sequence at each site indicated by an arrow. B) Schematic used for transcriptional assays. C) Transcriptional assays to assess periplasmic binding of antibodies with intein scars insignal sequences. D) Npu intein C-terminal fragment is required to link halves of split phoA signal peptide for periplasmic antibody:antigen binding.

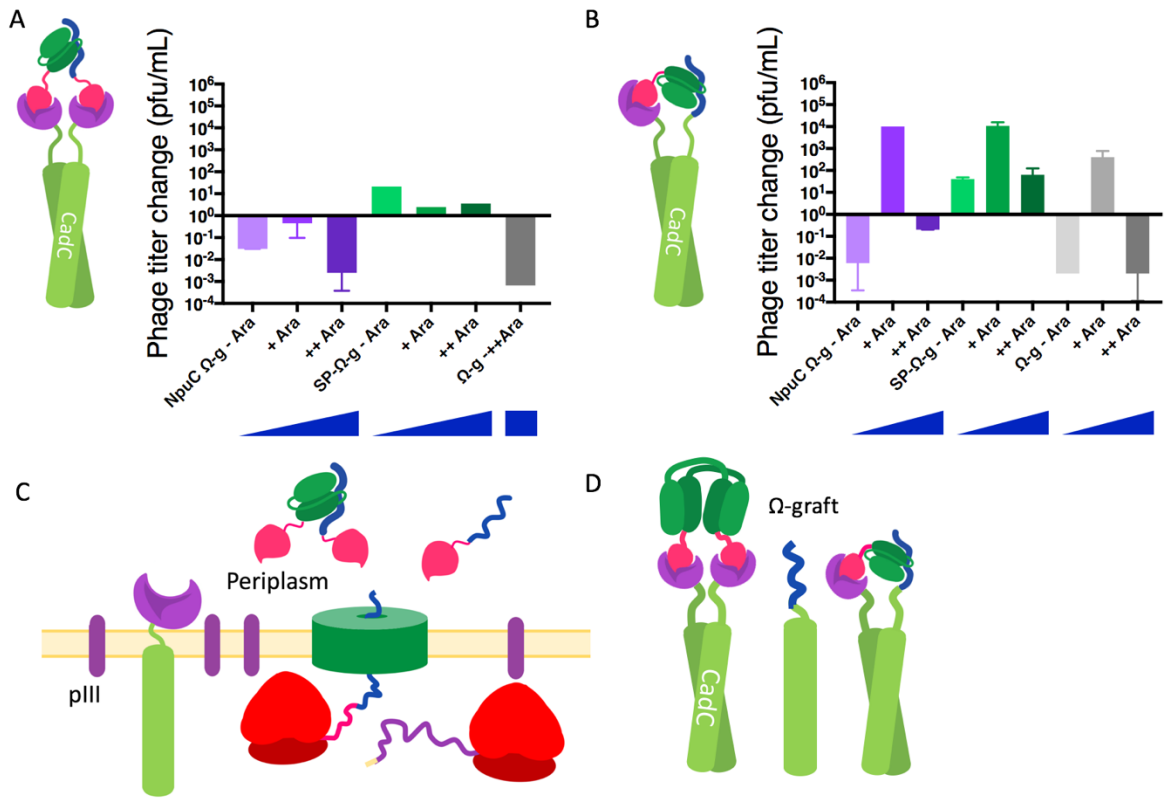


Fig. S2: Modification to the CadC selection architecture improves fitness and signal-to-noise. A-B) Selection architecture and with GCN4 soluble and linked to SH2 (3-hybrid) or with GCN4 tethered to transmembrane CadC (2-hybrid). In overnight batch culture enrichment tests of phage fitness, phage fail to show GCN4-dependent fitness unless GCN4 is CadC-linked. C) A model explaining decreased phage fitness under schema A): competition with pIII for SEC translocon is greater when GCN4 export is driven by SEC-pathway signal peptide PelB than when driven by CadC internal signal sequence. D) Antibody domain-swapping is more likely to cause homodimeric activation if some CadC does not bear HA4.

The Ω-graft scFv binds GCN4(7P14P) with Kd = 500pM.[12] Mutant L230F F231A has 0.013% binding activity vs. the Ω-graft.[13] In phage, the best dynamic range was achieved by expressing GCN4 antigen directly bound to CadC. This architecture may be more sensitive because transmembrane CadC does not compete with pIII for recognition of an N-terminal SEC pathway signal peptide, because a 2-hybrid interaction is more direct than a 3-hybrid, because the inner membrane represents a smaller interaction space than the bulk soluble periplasm, and/or because, critically, replacing >50^% of the CadC-HA4 molecules on the inner membrane with CadC-GCN4 favours heterodimeric antibody:GCN4 interactions over domain-swapping of antibodies (Fig. S2).

I tested the ability of this system to select for wild-type Ω-graft antibody over LF binding mutant Ω-graft in both continuous flow and overnight batch culture. I seeded each population with a 1000:1 ratio of LF mutant to WT. In continuous flow, the WT mutant fixed in the population by 12 hours, and in batch culture, the same effect was achieved following one 1:100 dilution, as determined by restriction digest assay and by PCR and sequencing of the phage pool.

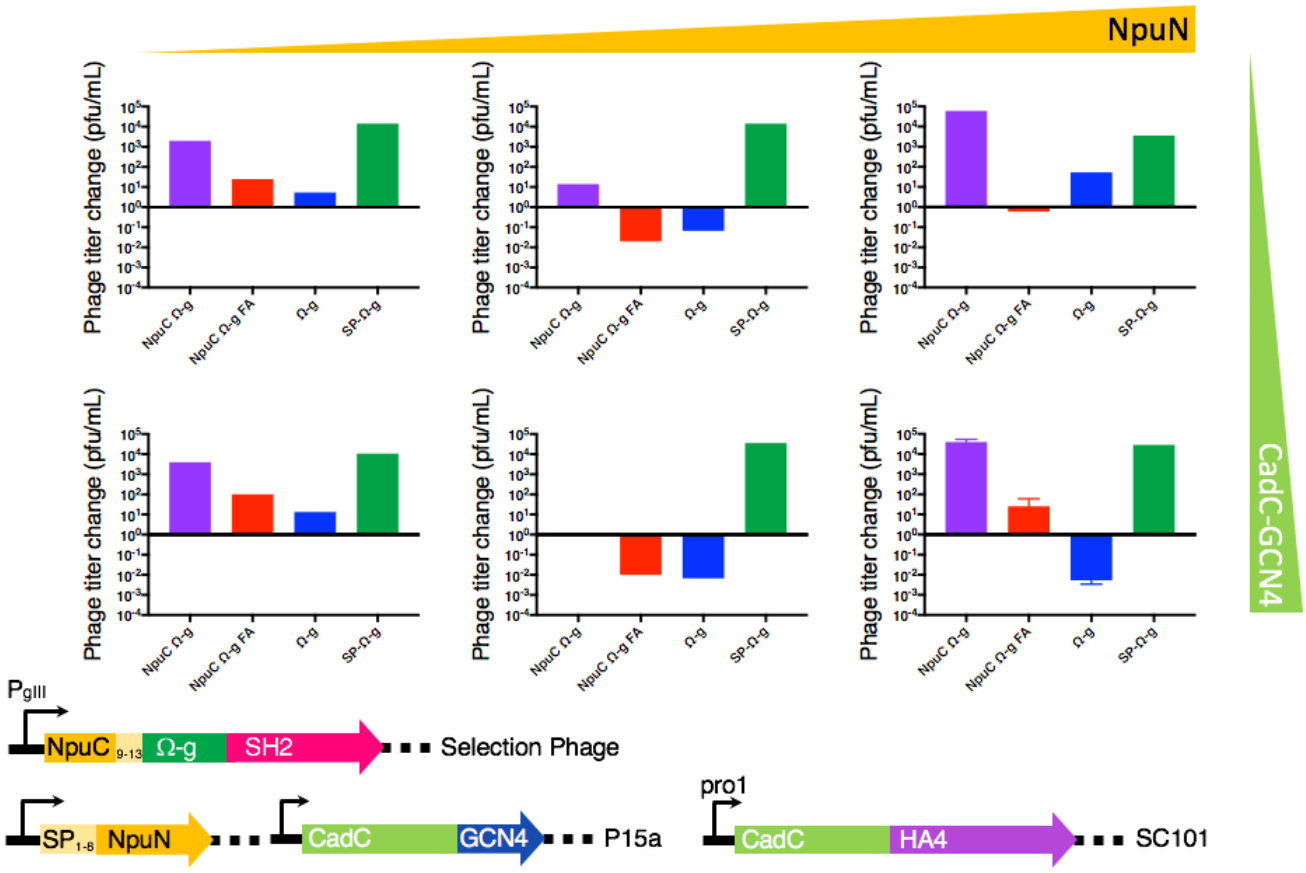


Fig. S3: Optimization of selection component dosing for PACE. Phage titer following an overnight enrichment in batch culture is a proxy for performance in PACE. Wedges indicate dosing changes of NpuN-PhoA signal peptide residues 1-8 (thus antibody export) and CadC-GCN4 across experiments. Gene cassette schematic shown below. FA: nonbinding Ω-g-SH2 point mutant. Pos: PhoA-sp Ω-g-SH2. Neg: PhoA-sp$_{9\text{-}13}$ Ω-g-SH2.

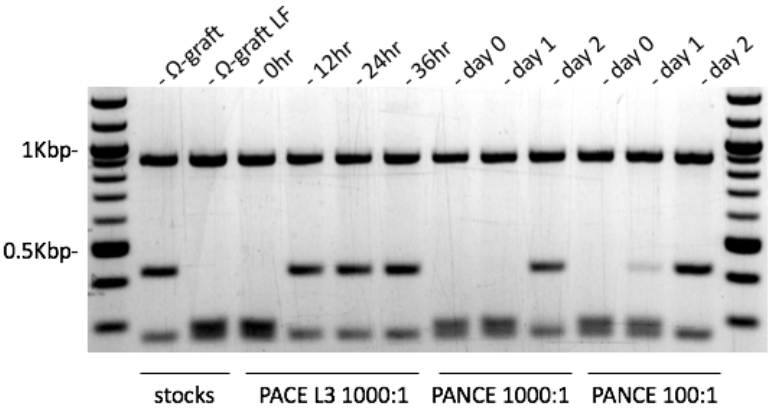
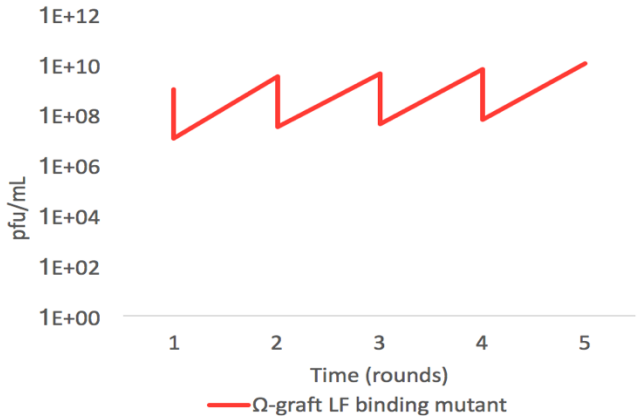


Fig. 6: Competition assays in continuous flow and batch culture. Phage were cleaved with HinfI, which cleaves WT form of Ω-g and not the LF point mutant, resulting in cleavage of 430bp band into 280bp and 150bp (not visible) bands. **This assay needs to be repeated, as the template had a cryptic HinFI**

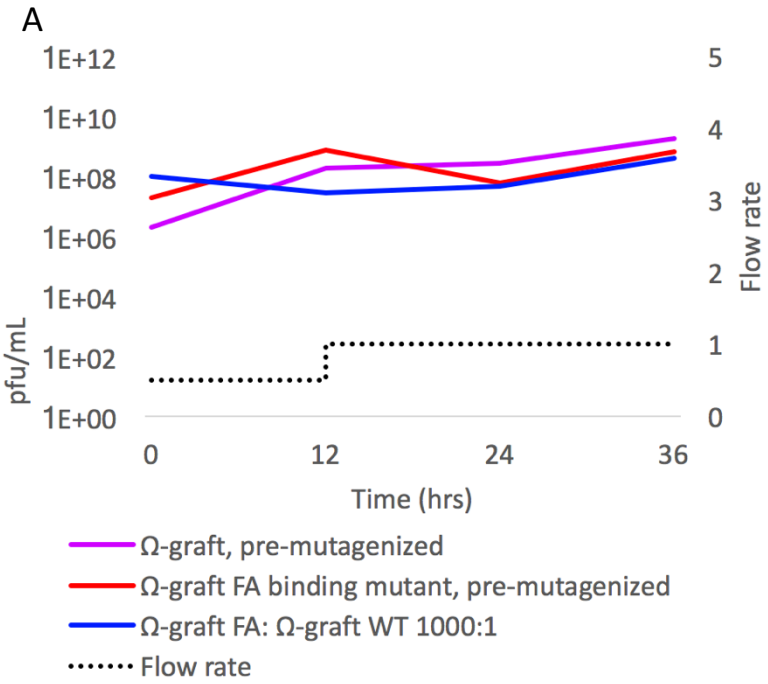
site due to a silent mutation, resulting in the 180bp and 900bp bands seen in all samples. Expected by June.

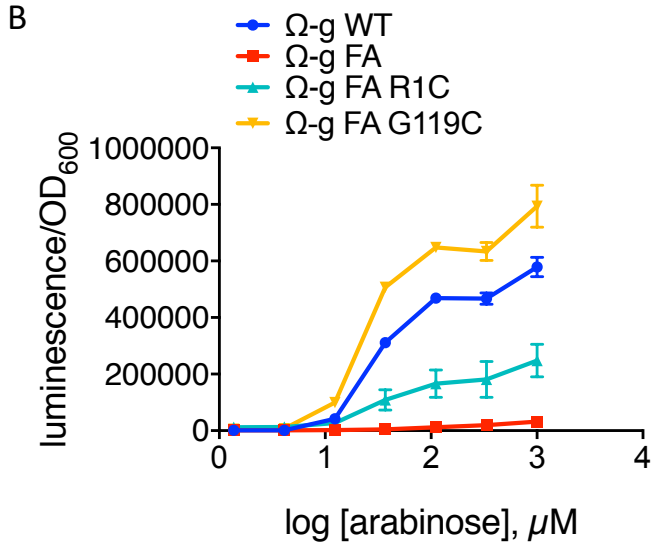


	Ω-graft				
	R1	R112	G115	F231	A232
d5-1		C		L	
d5-2			C		
d5-3	C	C		L	
d5-4				L	
d5-5		C		L	
d5-6		C		L	
d5-7			C		
d5-8		C		L	

Fig. S4: Batch culture of LF point mutant. Phage were passaged daily through fresh cells at mid-log. A) Phage titers over time. B) Consensus and single Ω-g mutations from batch culture. Eight phage were sampled and sequenced on day 5.

In batch culture, phage were able to correct mutation Ω-graft L230F by the 4th passage. In continuous flow with higher dosing of CadC-HA4, L230F reverted in a subset of the population but did not dominate the lagoon. Cysteine mutations were observed at position R1 or in the G-rich linker between antibody heavy and light chains in both batch and continuous experiments. While the split intein system reduces the propensity for antibody homodimerization, it does not outright eliminate it, allowing phage to partially cheat the selection by forming disulfide bonds between linkers. It is unlikely





that these disulfide bonds are a complete solution, as signal appears to still be GCN4-dependent in plasmid-based transcription assays.

C	Point mutations									
	SP	Ω-graft								
	P12	R1	S11	R112	G116	G119	G123	L224	F231	A232
L2-1				C					L	
L2-2					C					
L2-3	L	C						S		
L2-4	L	C						S		
L2-5		C	I							
L2-6		C	I							
L2-7	L	C						S		
L2-8							C			
L3-1		P							L	F
L3-2		P							L	F
L3-3		P							L	F
L3-4		P							L	F
L3-5		P							L	F
L3-6		P				C			L	F
L3-7		P							L	F
L3-8		P							L	F

Fig. 7: Ω-graft pPACE. PACE was carried out without ongoing mutagenesis. A) Phage titers over time. B) Ω-graft-SH2 constructs with consensus mutations were subcloned into a plasmid under P_{BAD} for transcriptional binding assays. C) Consensus and single mutations from pPACE. Eight phage were sampled and sequenced at 36 hrs. **Pending: Later timepoints and remainder of the conserved mutations tested in transcriptional binding assays.**

Fig. S5: PLACEHOLDER: Development of obligate CadC heterodimer. Expected within 2-3 weeks.

Fig. 8: PLACEHOLDER: pPACE with continuous mutagenesis and decreased CadC-HA4 expression to determine whether disulfide issue persists. Expected by June. Alternatively, pPACE with obligate CadC heterodimer. Will replace Fig. 7 if issue is resolved. Expected by July.

I am working on reducing the fitness benefit of homodimerization by decreasing CadC-HA4 expression. However, a more elegant and complete fix would be the development of an obligate heterodimer form of CadC. The DNA-binding domains of CadC behave as monomers in solution, binding independently to two head-to-head TTAXXXT motifs in the pCadBA promoter. Transcription appears to require both binding events, and can be ablated by substituting alanine at any of 3 positions that form sequence-specific side-chain:base interactions. I predict that an orthogonal CadC DNA-binding domain and promoter binding site can be created by replacing one DNA-binding domain and its TTAXXXT motif with the closely related DNA-binding domain of the E. coli protein PhoB and its 7-mer

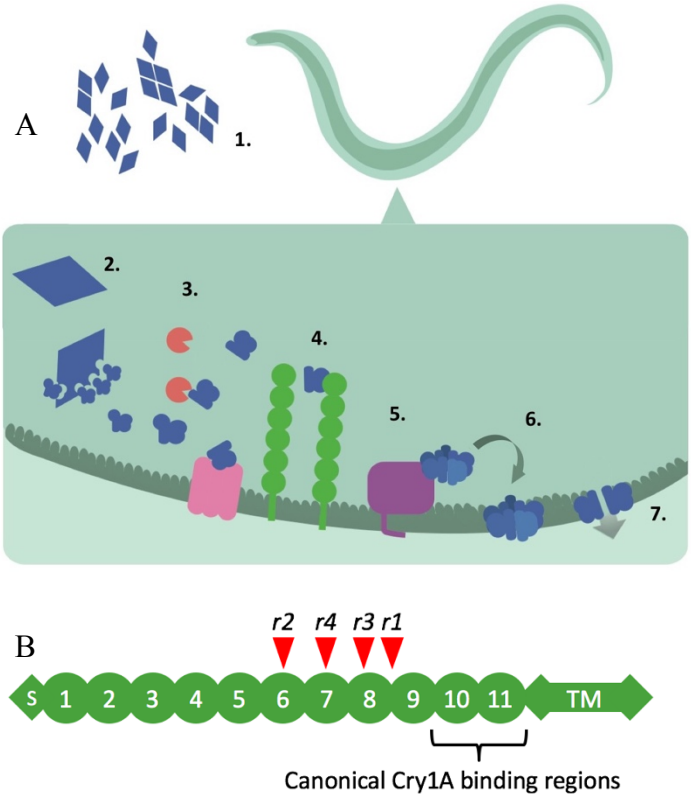
pho box recognition motif. These experiments are currently underway and results are expected by next week.

Fig. 8: PLACEHOLDER: pPACE with clinically relevant antibody and determination of evolved binding kinetics. Expected by 2020.

The ongoing goal of this project is the evolution of a clinically relevant antibody or antigen. Most of my candidates are from phage display experiments, ensuring they retain binding as ScFvs. One set of candidates are antibodies Ab-C1 and Ab-C2, raised against peptide epitopes of oncogene HER2, which show weak binding to HER2+ carcinoma cells in flow cytometry. The reported EC-50 is 1-5nM in the case of the full IgG antibody, and thus likely to be weaker in ScFv format.[7]

Another possible candidate is the P26 antibody, which targets the influenza haemagglutinin stalk, an exposed and slow-evolving region of HA that can be expressed solubly in *E. coli* in various engineered forms. P26 binds the HA trimer in the stalk region with EC-50 12nM in IgG format. Tighter binding should be achievable by evolving the scFv form against stalk monomers.

II. i. Evolution of Cry toxins to combat crop pests



The 3-domain Crystal or Cry toxins have potent and highly specific activity against toxin-specific subsets of invertebrate species[8] (Fig. II-1).

Fig. II-1: Cry toxin mechanism of infection. A)1. Crystalline toxin inclusions are first ingested by the invertebrate host. **2.** Host-specific proteases and intestinal pH contribute to protoxin solubilization and **3.** processing to the active toxin. **4.** Toxin domain II binds primary and secondary receptors on the brush border membrane via domain II loop regions. **5.** Domain I, a bundle of α-helices, oligomerizes to form a pre-pore. **6.** The pre- pore inserts into the membrane. **7.** The membrane loses integrity **8.** Collapse of the epithelium leads to septicemia and death[8-10]. **B) Lepidopteran cadherin schematic.** N-terminal signal sequence (S), cadherin repeats (CRs) 1-11, membrane-proximal region, transmembrane region (TM), and cytoplasmic domain. Cry1Ac resistance-linked mutations in *P. gossypiella* cadherin-1 have been reported: *r1*,

r3, and *r4* are deletions of 8, 42, and 5 amino acids, respectively, and *r2* is an early stop codon.[11,12]

Cry1Ac has been used in agriculture to control populations of the cotton pest moth *Pectinophora gossypiella*, commonly known as Pink Bollworm (PBW), since 1996. In India and China, resistance to Cry1Ac is increasingly widespread (Fig. 1B)[11]. With support from my collaborators, Jeff Fabrick and Bruce Tabashnik at the University of Arizona, I am interested in enhancing Cry1Ac toxicity to susceptible and resistant Pink Bollworms, either by enhancing binding to current receptors or discovering *de novo* receptors.

Cry1Ac intoxicates Pink Bollworm via a cadherin receptor, PgCad1. Cry1Ac binds most strongly to membrane-proximal cadherin domains 10-11 of this gene, with some binding to cadherin domains 7-9[11]. Four PgCad1 resistance alleles have been described, termed *r1-r4*. Of these alleles, all but r2 affect cadherin domains 6-9, suggesting a role for domains 6-9 in addition to domains 10-11.

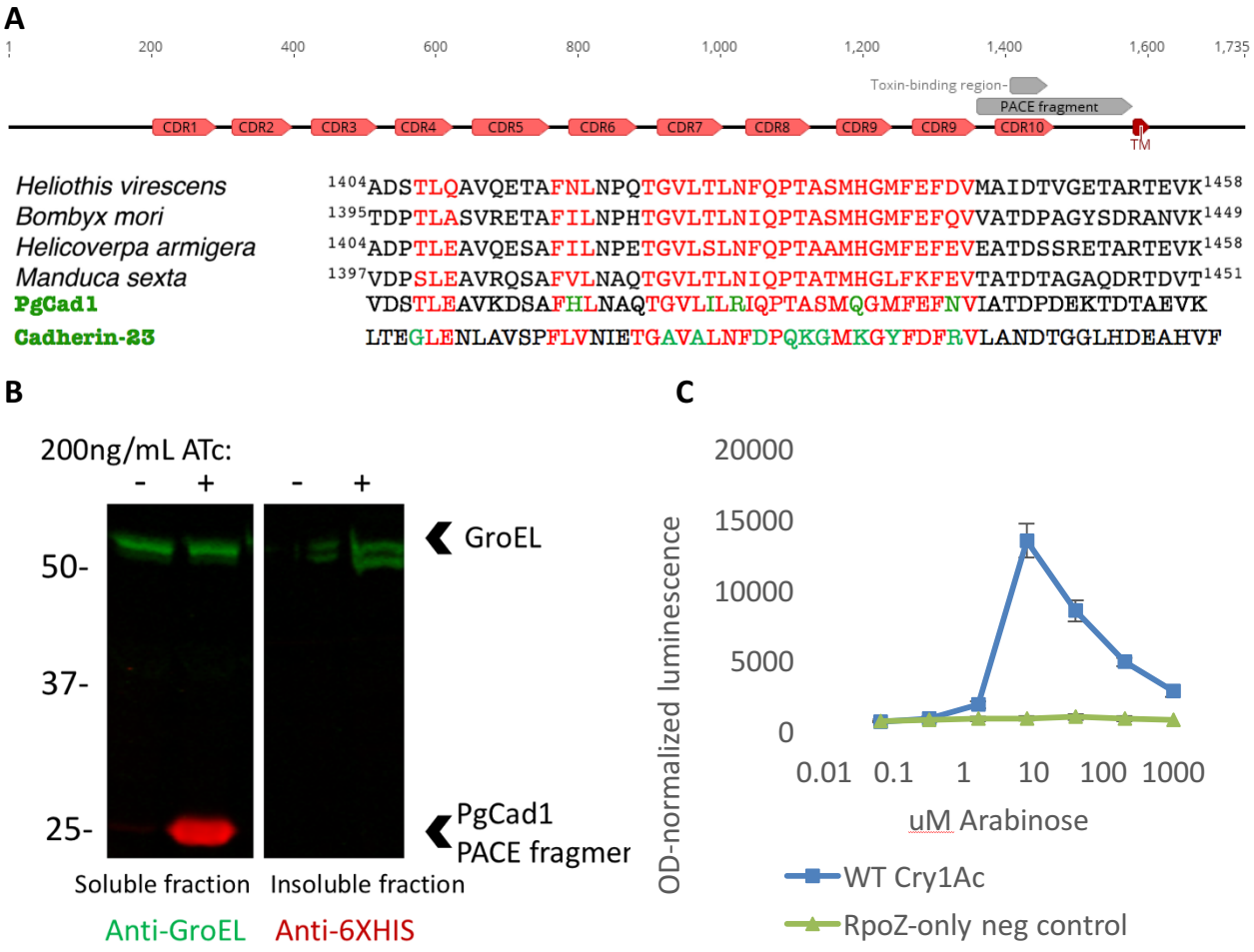
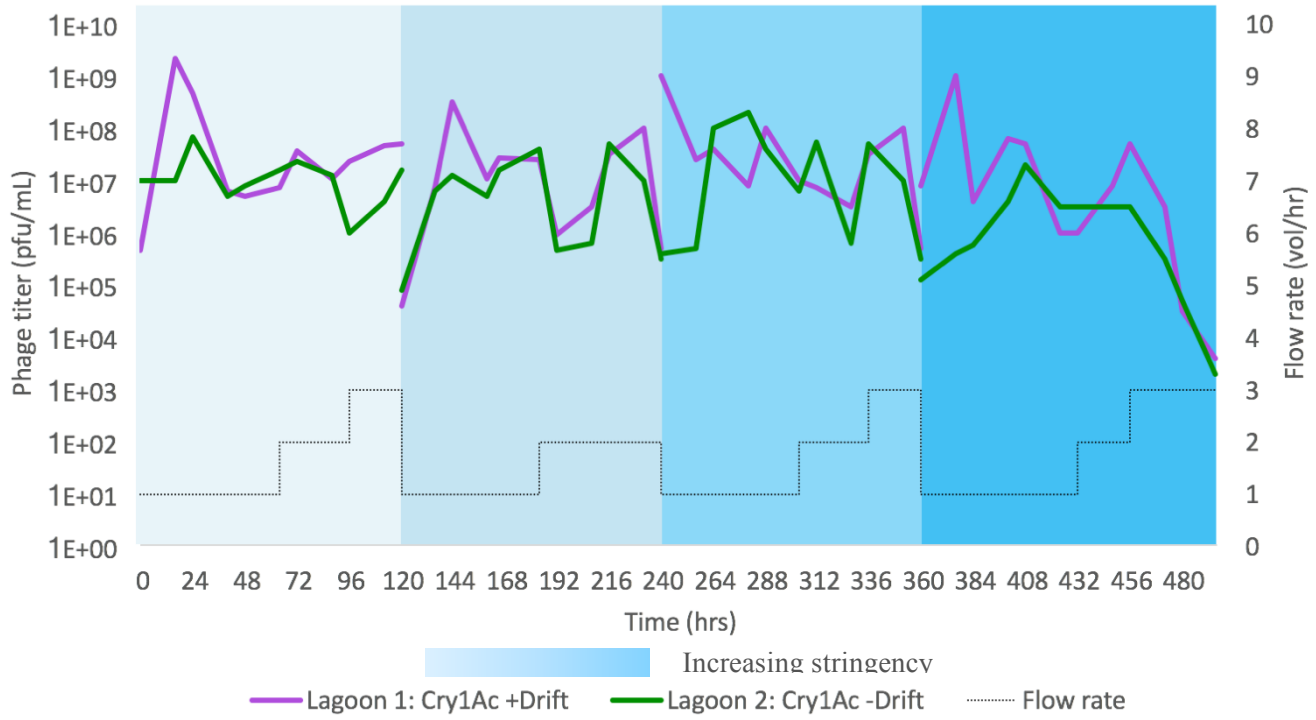


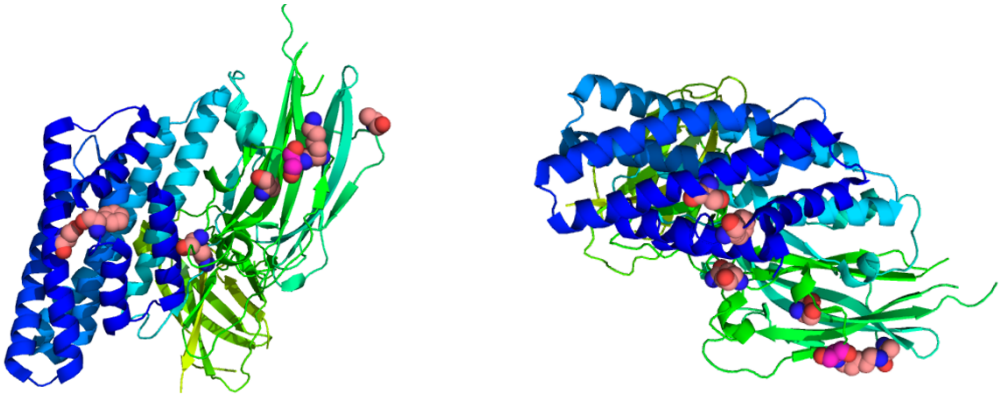
Fig. II-2: PgCad1 target fragment selection for PACE. A) Annotated PgCad1 gene with toxin-binding region and fragment selected for PACE annotated, and toxin-binding region of PgCad1 and Cadherin-23, aligned with toxin-binding regions of known insect Cry1Ac receptors[2]. B) Western blot of his-tagged PgCad1 PACE fragment (25kDa) C) Bacterial 2-hybrid transcriptional activation assay using luciferase as a reporter. RpoZ-Cry1Ac fragment is under a PBAD promoter.

I carried out PACE to evolve Cry1Ac binding to domain 11 of PgCad1. Most mutations mapped to one surface of the protein, and many were clustered in the receptor-binding loops of Domain II; S402L is highly exposed on domain II loop 4.

A



B

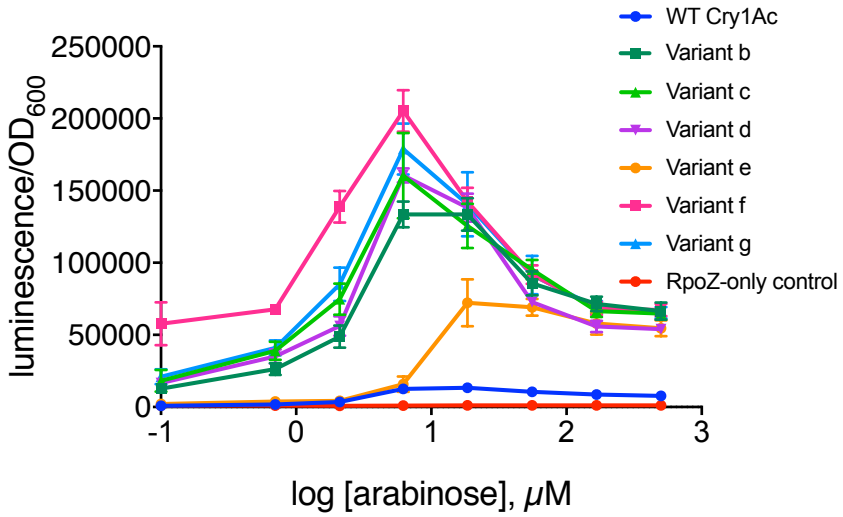


C	Cry1Ac mutations									
Phage	Time (hrs)	Lagoon	F67	S102	A343	A344	T360	K401	S402	Q415
1	488	1	I				I	R	L	
2	488	1	I				M	R	L	
3	488	1	I	R			I	R	L	
4	488	1	I	R			I	R	L	
5	480	1	I						L	
6	480	1	I						L	

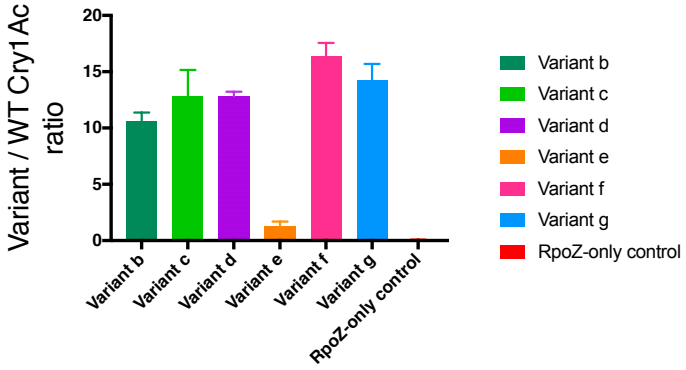
7	480	1	I		S		I	R	L	
8	480	1							L	
1	488	2	I				I	R	L	
2	488	2	I		S		I	R	L	
3	488	2	I		S		I	R	L	
4	488	2	I		S		I	R	L	K
5	480	2							L	K
6	480	2							L	
7	480	2	I				I	R	L	
8	480	2	I		S	S	I	R	L	

Fig. II-3: PgCad1-Cry1Ac PACE: A) To induce free pIII production (drift), lagoons was treated with +/- 500ng/mL anhydrotetracycline (ATc) for the first 16 hours of PACE and 200ng/mL for the following 24 hours PACE. Phage titers from PACE across 4 separate chemostats, indicated by line breaks. B) consensus mutations mapped to the Cry1Ac crystal structure. [13] C) 8 phage per lagoon were sampled from the final leg of PACE. Consensus mutations from both lagoons are colour-coded as shown on the structure in B).

A



B



C

Variant	F67	E89	Y313	Y314	A343	T360	F371	K401	S402	R499
b	I		S	C		I		R	L	
c	I				S	I		R	L	

d					S	I		R	L	
e		I				I		R	L	
f		I		K		S	I		R	L	S
g								C	L	

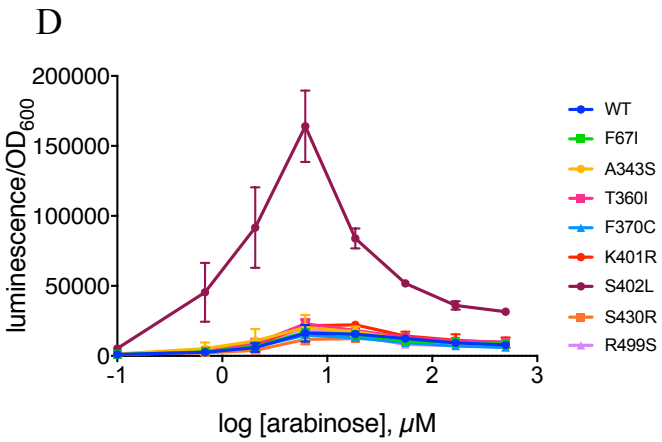


Fig. Fig. II-4: Plasmid-based transcriptional assays of Cry1Ac PACE mutations in isolation and combination. A-C) 6 phage variants were subcloned from the final timepoint of PACE and compared to the wild-type. D) Each of 12 common PACE mutation from all stages of the analysis was subcloned in isolation into the WT Cry1Ac background.

The clear next step is the purification of evolved Cry1Ac for thermostability and binding (thermal shift and dip-and-read biosensor, respectively), but I found that Cry1Ac was extremely intractable to expression in *E. coli*, and have further had significant difficulty in finding commercial means to express the protein in its native *B. thuringiensis* (we do not wish to do so ourselves, as no-one in the lab has *B. thuringiensis* experience and expression requires significant expertise). Currently, one of my collaborators in the Aroian group, Kelly Flanagan, is engaged in the second of two efforts to express the toxin for testing.

Evolved Cry1Ac variants will be sent to Drs. Tabashknik and Fabrick for *in vivo* and *in vitro* characterization. *In vivo* activity will be assayed via 21-day diet incorporation bioassay with resistant and susceptible PBW strains carrying all four resistance alleles. For *in vitro* characterization, brush border membrane vesicles (BBVMs) will be isolated by dissection and homogenization from susceptible and resistant PBWs for dot blot assays visualized with anti-Cry1Ac antibodies[11].

Fig. II-5: PLACEHOLDER: Thermostability and PgCad1 binding data, Cry1Ac WT vs. evolved variants d,f,g. Anticipated by September if expression by collaborators is successful.

Fig. II-6: PLACEHOLDER: Dot blot binding assays, Cry1Ac WT vs. evolved variants d,f,g on isolated Pink Bollworm midgut brush border membrane vesicles. Anticipated by January, depending on collaborators.

Fig. II-7: PLACEHOLDER: Pink Bollworm feeding assay toxicity data, Cry1Ac WT vs. evolved variants d,f,g. Anticipated by January, depending on collaborators.

In the meantime, I am interested in engineering binding to PgCad1 domains 6-9 or to an alternate, paralogous cadherin protein in PBW. Along with PgCad1, one other cadherin mRNA (cadherin-23) has been identified in the midgut transcriptome of PBW.[14] However, the soluble fragments we tested did not bind Cry1Ac in transcription assays, so any such binding activity would need to be evolved *de novo*. PACE is a poor tool for *de novo* binding, so I instead turned to phage display, and generated a 1-million-member phagemid library of fibronectin-3-derived nanobodies. This project has since been passed along to my undergraduate student, Simon Shen, who has succeeded in purifying a phage library of 1E10 pfu and is currently expressing tagged target proteins from projects 2 and 3 for bead-based phage display. It is difficult to predict a schedule for these efforts, as Simon is also a full-time student, but he is fully independent and has made steady progress.

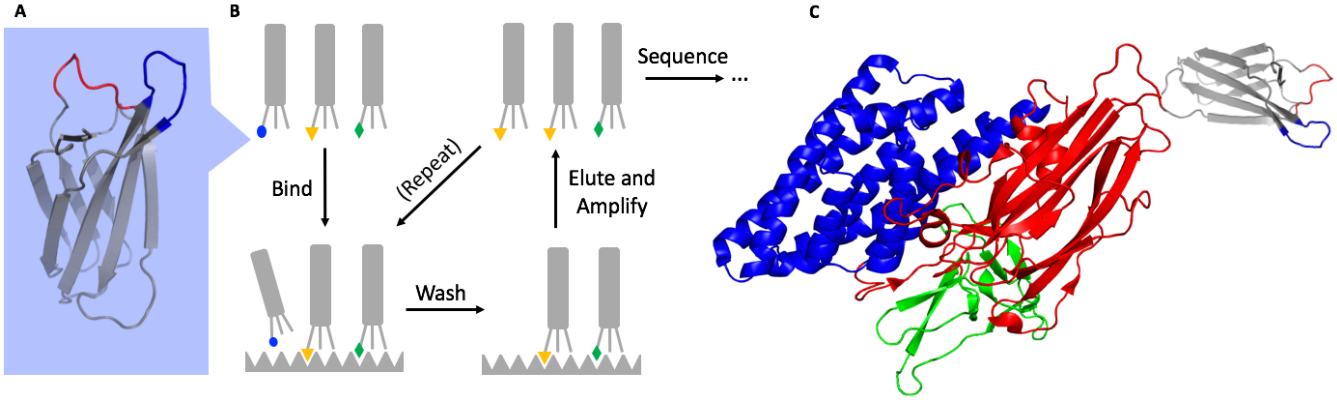


Fig. II-8: Phage display schematic: A) Fibronectin type III (FN3) scaffold. PDB: 3K2M. Five amino acids in the both the BC loop (red) and FG loop (blue) are diversified by site-saturation mutagenesis. **B)** The diversified FN3 are fused to the coat protein pIII and displayed by bacteriophage, which contain the genetic sequence encoding the corresponding FN3 variant. This phage library is added to the target protein immobilized to a solid support. C) Schematic: Display-selected FN3 is grafted into flexible binding loops of Cry1Ac.

Fig. II-9: PLACEHOLDER: Phage display results: FN3 loop insert sequences, ELISA binding intensities and performance in plasmid-based 2-hybrid luciferase assays. Timeline unknown.

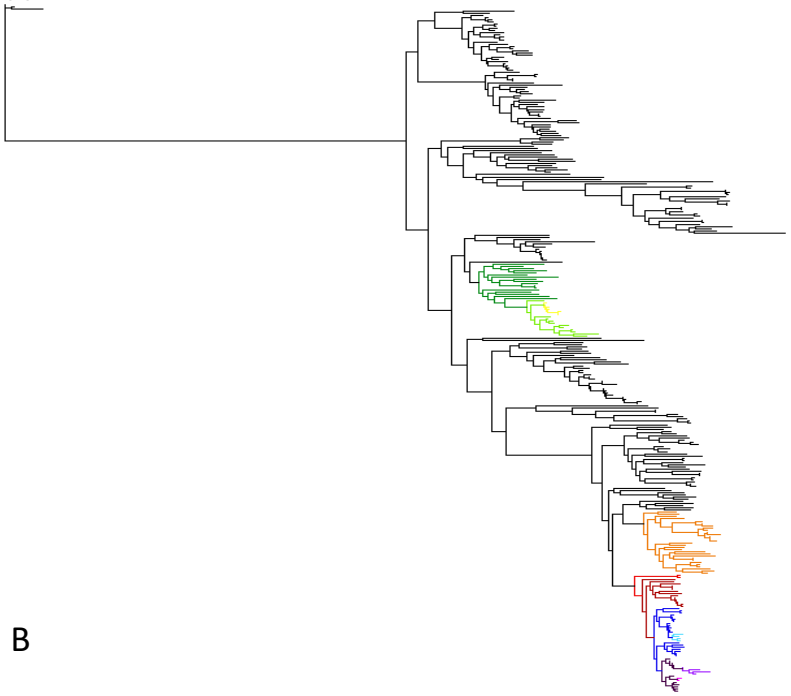
Fig. II-10: PLACEHOLDER: Thermostability, binding and toxicity assays of best FN3 variants from phage display grafted into Cry1Ac Domain II loops. Timeline unknown.

III. Ancestral Reconstruction and Family Shuffle to isolate Cas9 variants with unique properties

As targeted gene therapies enter the clinic, there is now a significant therapeutic need for CRISPR systems with alternate PAM sequences that may target novel sites, and for smaller Cas9 variants for delivery to disease-relevant tissues by viral vectors.

The full diversity of the Cas9 family tree has not been explored in much depth for the purpose of engineering improved Cas9s. While studies of the evolutionary history of Cas9 have been carried out, no efforts have been made to our knowledge to investigate the useful properties of putative ancestral Cas9 sequences, or to generate novel Cas9s through DNA shuffling experiments. I am collaborating with labmate Chris Wilson, an expert in ancestral reconstruction, to broaden the pool of sequences to be shuffled by reaching not only across the tips of the Cas9 gene tree but also back in time through its branches.

A



B

CLADE	BOOT-STRAP	DEFINING FEATURE
65	100	G1218R (Seen in spCas9-NG ARRRR variant with expanded PAM targeting space[15])
66	99	R1335 becomes conserved (Thought to interact with position 3 of NGG PAM; R1335V is key for spCas9-NG ARRRR and VRVRFRR variants[15])
68	100	D1135 to aliphatic residue (Seen in spCas9-NG VRVRFRR and xCas9 variants with expanded PAM targeting space and enhanced cutting activity[15,16])
74	99	Loss of basic residue at position 1218[15]
91	100	E543N (E543D seen in Seen in spCas9-NG VRVRFRR and xCas9 variants[15,16])

96	86	KQYT insert in region 1335-1337 (R1335V and T1337R seen in VRVRFRR variant[15])
105	100	D1135 to aliphatic residue; ILSTNNK insert and loss of basic residue at position 1218[15]
119	100	E1219 to aliphatic residue (key mutation in Cas9 with expanded PAM targeting space and enhanced cutting activity[15,16])
214	95	Conserved S/T at position 409 (S409I seen in xCas9[16]); defines Nme2-like sequences[17]
219	100	E1135 becomes conserved, followed by 24-AA insert; R1335 becomes conserved
230	100	E543T

Fig. III-1: Cas9 gene phylogeny, with nodes of interest colour-coded. A) 312 Cas9 sequences were collected from uniprot and NCBI. A MAFFT alignment was used as a starting point for a more refined Probcons alignment. Cas9 phylogeny was then generated based on maximum likelihood and neighbor-joining. Scale indicates genetic distance, units = substitutions per position. B) Clades of interest were defined by the synapomorphic defining features listed, as determined by alignment with spCas9 variants with relaxed PAM specificities (NGG to NG). Bootstrap values shown as a percentage of 1000 simulations. Clades are colour-coded as in A).

Ancestral sequence reconstruction (ASR) uses modern sequence diversity and Bayesian statistics to infer a probable sequence corresponding to an historical node on a gene tree. Aancestral enzymes are often more thermostable than modern counterparts, and are generally less specialized and thus more tractable to directed evolution towards new functions.[18]

Chris generated a Cas9 gene tree from 312 Cas9 sequences, with sequence selection guided by Koonin *et al.* 2017 (Fig. III-1).[19]

I used polymorphisms at positions known to be important in PAM specificity to define clades of interest. I focused my search on regions of the tree close to *S. pyogenes* Cas9 or *Neisseria meningitides* Cas9, an enzyme of interest for its small size but limited by its lengthy PAM sequence (Fig. III-1). We carried out a Bayesian analysis according to a Markov model to determine the most probable sequence at these nodes. I am currently cloning, and expect to have 12 Cas9 variants fully assembled in their relevant vectors within two weeks.

Following assembly of Cas9 variants, I will characterize their solubility and toxicity, then carry out incremental truncation and family shuffling, and assemble a phage library of shuffled variants. I can bias this library towards small size by simple gel excision.

The greatest bottleneck to high-throughput testing of Cas9 orthologs is the requirement for a cognate guide RNA. These elements have undergone significant drift and differ between species. There is evidence that Cas9 enzymes may be semi-promiscuous with regard to tracrRNA sequence. We thus chose to focus on identifying guides in modern taxa with Cas9s resembling our reconstructed sequences. Most of our sequences have 50%-70% sequence identity with Cas9s from the *Neisseria* or *Streptococcus* genera, for which crRNA sequences but not tracrRNA sequences are well-annotated.

Chris is currently working on an *in silico* pipeline for determining tracrRNA sequences from genomic data alone. As a starting point, I will conduct binding assays using unshuffled ancestral Cas9s and the single guides engineered for *S. pyogenes* and *Neisseria meningitides* Cas9. We will select 5-10 promising tracrRNA sequences and multiplex tracrRNA:crRNA pairs in host cells encoding the one-

hybrid Cas9 binding selection architecture described in Hu *et al.* 2018.[16] I can then select for functional members of the library or for binding to hard-to-access PAMs, or further screen for small yet functional Cas9 variants. Downstream mammalian cell characterization will be carried out by another labmate, Jessie Davis.

Because this project is in early stages, its timeline and scope is difficult to predict and will largely depend on the results of early-stage experiments.

References

1 Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499-503.

2 Badran, A. H. *et al.* Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. *Nature* **533**, 58-63.

3 Wojcik, J. *et al.* A potent and highly specific FN3 monobody inhibitor of the Abl SH2 domain. *Nature structural & molecular biology* **17**, 519-52.

4 Kuper, C. & Jung, K. CadC-mediated activation of the cadBA promoter in Escherichia coli. *Journal of molecular microbiology and biotechnology* **10**, 26-39.

5 Mallam, A. L. & Jackson, S. E. The dimerization of an alpha/beta-knotted protein is essential for structure and function. *Structure (London, England : 1993)* **15**, 111-122.

6 Škerlová, J. *et al.* Optimization of the crystallizability of a single-chain antibody fragment. *Acta Crystallographica. Section F, Structural Biology Communications* **70**, 1701-1706.

7 Rockberg, J., Schwenk, J. M. & Uhlen, M. Discovery of epitopes for targeting the human epidermal growth factor receptor 2 (HER2) with antibodies. *Mol Oncol* **3**, 238-247.

8 Adang, M. J., Crickmore, N. & Jurat-Fuentes, J. L. Diversity of Bacillus thuringiensis Crystal Toxins and Mechanism of Action. **47**, 39-87.

9 Bravo, A., Likitvivatanavong, S., Gill, S. S. & Soberón, M. Bacillus thuringiensis: A story of a successful bioinsecticide. *Insect biochemistry and molecular biology* **41**, 423-431 (2011).

10 Hui, F. *et al.* Structure and glycolipid binding properties of the nematicidal protein Cry5B. *Biochemistry* **51**, 9911-9921 (2012).

11 Fabrick, J. A. & Tabashnik, B. E. Binding of Bacillus thuringiensis toxin Cry1Ac to multiple sites of cadherin in pink bollworm. *Insect Biochem Mol Biol* **37**, 97-106 (2007).

12 Fabrick, J. A. & Tabashnik, B. E. Similar genetic basis of resistance to Bt toxin Cry1Ac in Boll-selected and diet-selected strains of pink bollworm. *PLoS One* **7**, e35658 (2012).

13 Evdokimov, A. G. *et al.* Structure of the full-length insecticidal protein Cry1Ac reveals intriguing details of toxin packaging into in vivo formed crystals. *Protein Science : A Publication of the Protein Society* **23**, 1491-1497 (2014).

14 Tassone, E. E. *et al.* Sequencing, de novo assembly and annotation of a pink bollworm larval midgut transcriptome. *GigaScience* **5**, 28 (2016).

15 Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science (New York, N.Y.)* **361**, 1259-1262 (2018).

16 Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57, (2018).

17 Lee, C. M., Cradick, T. J. & Bao, G. The Neisseria meningitidis CRISPR-Cas9 System Enables Specific Genome Editing in Mammalian Cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **24**, 645-654 (2016).

18 Wheeler, L. C., Lim, S. A., Marqusee, S. & Harms, M. J. The thermostability and specificity of ancient proteins. *Current opinion in structural biology* **38**, 37-43 (2016).

19 Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Current opinion in microbiology* **37**, 67-78 (2017).

I thank Aditya Raguran and Tina Wang for assistance in early stages of PPACE design and in transcriptional assays.