DISSERTATION

# CHARACTERIZATION AND APPLICATION OF A NOVEL COMPOSITE NANOMATERIAL COMPRISED OF POROUS PROTEIN CRYSTALS AND SYNTHETIC DNA

Submitted by

Julius D. Stuart

Department of Chemistry

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2022

Doctoral Committee:

Advisor: Christopher D. Snow

Alan J. Kennan
Matthew P. Shores
Susan K. De Long

ABSTRACT

CHARACTERIZATION AND APPLICATION OF A NOVEL COMPOSITE NANOMATERIAL COMPRISED OF POROUS PROTEIN CRYSTALS AND SYNTHETIC DNA

Composite nanomaterials are systems comprised of multiple components boasting enhanced properties over those exhibited by the individual constituents when isolated. Such systems are highly tunable, allowing one to vary component types (e.g., polymer, metal, ceramic) for influencing performance in various contexts. Moreover, composite nanomaterials can be further modified using biofunctionalization for use in biological settings. Composite nanomaterials have been tested in applications including, but not limited to, textile, defense, food, energy and biomedical engineering. A sub-domain within composite nanomaterials involves porous protein crystals soaking, or, separately, encapsulating various guest molecules. Porous protein crystals are ordered, insoluble assemblies forming a network of nanopores capable of allowing inward diffusion of guest molecules. Moreover, recombinant protein variants can be engineered for probing guest molecule binding to host crystal nanopores further highlighting the tunability of this novel composite nanomaterial.

The goal of this work is to evaluate a novel composite nanomaterial comprised of host porous protein crystals and guest double stranded DNA. We show that guest DNA loads into host crystals predominantly along the axial nanopores. Equilibrium adsorption isotherm results suggest guest DNA unbinds from host crystals relatively slowly. Computational modeling and

Fluorescence Recovery After Photobleaching (FRAP) studies suggest intra-nanopore guest diffusion is attenuated relative to bulk diffusion.

We also show that porous protein crystals loading with synthetic DNA barcodes can be used for tracking mosquitoes. Fluorescently labeled crystals can be ingested by mosquito larvae and adults, followed by detection using fluorescence confocal microscopy. Crystal-bound DNA can be liberated from host crystals by incubation with solution containing deoxynucleotide triphosphates (dNTPs). Previously ingested barcode-loaded crystals can be recovered using standard molecular biological techniques.

Lastly, we show a DNA barcode sequence construction strategy that is modular, economical and scalable. Computational sequence design and scoring allowing identification of top candidates for experimental validation. Analysis of next-generation sequencing datasets informs barcode construction specificity while simultaneously reinforcing the multiplexing capabilities boasted by modular DNA barcodes.

ACKNOWLEDGEMENTS

Thank you to all Snow Lab members for your unending support and friendship.

Dr. Christopher D. Snow
Dr. Thaddaus Huber
Dr. Luke Hartje
Dr. Ann E. Kowalski
Dr. Abby Orun
Dr. Ning Zhao
Dr. Gayani Dedduwa-Mudalige
Ashlyn Chen
Alec Jones
Jacob Deroo
Camden Meyer
Sergei Driga
Lauren Beatty
Tyler Sweet

Thank you to all collaborators from whom I have been fortunate to learn and further develop as a researcher.

Dr. Rebekah C. Kading
Dr. Matt Kipper

Thank you to my committee members for your continued guidance and support.

Dr. Alan Kennan
Dr. Matthew Shores
Dr. Susan De Long

Funding
Colorado State University GAUSSI fellowship 2018-2019 (NSF grant DGE-1450032)
National Institutes of Health R21 AI146740
National Institutes of Health/NCATS Colorado CTSA Grant Number UL1TR002535

In Chapter 2, I thank Dr. Brian Munsky and Luke Hartje for computational modeling consultation and technical assistance. I thank Dr. Matt Kipper for data analysis consultation and

technical assistance. I thank Dr. Charles Henry and Zach Call for laser cutting consultation and technical assistance.

In Chapter 3, I thank Dr. Mark D. Stenglein for NGS consultation and technical assistance. I thank Dr. Daniel B. Sloan and Gus Waneka for instrumentation training and access for size selection during NGS library preparation. I thank Dr. Susan K De Long and Maria Irianni Renno for qPCR training and instrument access. I thank all staff from the Genomics Shared Resource at the University of Colorado Cancer Center (CU Anschutz) where sequencing was conducted.

In Chapter 4, I thank Dr. Mark D. Stenglein for NGS consultation and technical assistance. I thank Dr. Daniel B. Sloan and Gus Waneka for instrumentation training and access for size selection during NGS library preparation. I thank Dr. Joshua Chan and Parsa Ghadermazi for Opentrons OT-2 Liquid Handler training and instrument access. I thank all staff from the Genomics Shared Resource at the University of Colorado Cancer Center (CU Anschutz) where sequencing was conducted.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER 1: POROUS PROTEIN CRYSTALS AND SYNTHETIC DNA AS A NOVEL COMPOSITE NANOMATERIAL

## 1.1 Composite Nanomaterials

Composite nanomaterials (CNs) describe systems comprised of more than one material type with a single or multiple components existing on the nanoscale (1). The appeal of CNs results from such systems possessing enhanced properties over those exhibited by the individual components in isolation (2). CNs have been evaluated in various industries including, but not limited to, textile (3, 4), defense (5), food (6, 7), energy (8, 9), and biomedical engineering (2). The relatively small size of CNs contributes to their performance testing over a broad range of environments, ranging from fabrics to biological contexts. Also, CNs are highly tunable assemblies, allowing one to vary system constituents (e.g., metals, ceramics, polymers) followed by performance evaluation for optimization, boasting their versatility in synthesis and testing (1). Biofunctionalization of CNs further permits testing and use in biomedical contexts, such as pathogen detection (10).

Incorporation of DNA within CNs has yielded systems containing synergistic properties for diverse applications (11). Previous examples of DNA containing CNs include cluster-induced silicification on DNA origami scaffolds for providing precise, customized three-dimensional assemblies for nanotechnological applications (12). Gold nanoparticles capped with thymine-rich, FAM labeled DNA was used for mercury (Hg2+) detection (13). Silver nanoclusters and guanine-rich DNA aptamers were combined for cocaine detection yielding a limit of detection (LOD) of 0.1 μM (14). Regulated singlet oxygen generation for photobynamic therapy, a noninvasive, alternative cancer treatment, was achieved with carbon nanotubes and α-

thrombin aptamers (15). DNA containing CNs have also been used for cell transfection using calcium phosphate nanoparticles coated with plasmid DNA encoding enhanced green fluorescent protein (16).

### 1.2 Protein Crystal-Based Composite Nanomaterials

A specialized sub-domain with CNs involves systems using porous protein crystals for loading and adsorbing guest molecules (17). Protein crystals are large, insoluble, macromolecular, ordered assemblies that grow following nucleation in supersaturated solution and contain a network of variable-diameter solvent-filled pores (18). Once used as a purification technique, protein crystals are widely used for structure determination using x-ray diffraction (XRD) (19). However, despite spontaneous assembly in high-salt growth solution, protein crystals dissolve in low-salt, aqueous environments revealing a prominent limitation toward crystal performance evaluation in isotonic solutions (20). Cross-linking, or bioconjugation, describes a procedure for overcoming solubility constraints by leveraging protein functional groups and creating covalent bonds between neighboring sites throughout protein crystals (21). Following cross-linking, porous protein crystals retain overall topology despite transfer to low salt solution conditions (22).

### 1.3 Host-Guest Systems

Previous examples of protein crystal-based CNs include soaking guest small molecules, gold nanoparticles, and proteins, separately, into host crystals followed by attempted structure determination of loaded guests using XRD (23-25). Guest enzymes within host protein crystals containing 13nm diameter pores retain activity despite immobilization (26). Cross-linked mesoporous enzyme crystals containing metal complexes demonstrate enhanced stability

relative to solution-based enzymes (27). Mesoporous materials are defined as having pore diameters in the range of 2-50nm (28). In contrast to diffusion-mediated guest diffusion into existing host crystals, guest-encapsulated crystals form by simply adding guest to the high salt crystal growth solution revealing an alternative pathway for guest inclusion within host hemocyanin crystals containing pores approximately 11nm in diameter(29).

While previous works have described protein crystal-based CNs, less work has been communicated on CNs comprised of both porous protein crystals and guest DNA. Hashimoto et al. soaked cross-linked hemocyanin crystals with solutions containing fluorophore labeled double stranded DNA of increasing lengths (10, 20, 50, and 200 base pairs) (29). All DNA lengths loaded into host crystals as confirmed by confocal microscopy. However, timelapse imaging of guest DNA loading into host crystals was not recorded, presumably because the study of diffusion kinetics fell beyond the target scope of biomacromolecule encapsulation. Information regarding nucleic acid loading kinetics would inform the length of time requisite for saturating host crystals with guest DNA. Moreover, the retrieval of guest DNA *following* crystal-loading was not explored. Recovery of the DNA sequences from the host crystals will be critical for some applications (e.g., prior to downstream processing via polymerase chain reaction (PCR) amplification and/or sequencing).

**1.4 Research Specific Aims**

This work describes a novel CN system comprised of porous protein crystals and DNA via: (1) biophysical characterization (Chapter II), (2) mosquito tagging with DNA barcode loaded crystals (Chapter III), and (3) modular DNA barcode design (Chapter IV).

In Chapter II, we show time-base loading of fluorescently labeled DNA into host crystals. Guest DNA loading occurs predominantly along the crystal nanopores. Four computational models are described for predicting guest loading. Guest DNA intra-nanopore diffusion at equilibrium is attenuated relative to guest solution diffusion. Pore-mediated nucleic acid transport has applications in both biological (e.g., nuclear pore complex) and synthetic (e.g., nanopore sequencing) settings.

In Chapter III, we show the performance of DNA-barcoded loaded crystals for tracking mosquitoes. Barcode loaded crystals are ingested by mosquito larvae and adult mosquitoes. Barcodes can be displaced from host crystals by deoxynucleotide triphosphate (dNTP) incubation. Barcode detection is performed using PCR, qPCR and next-generation sequencing (NGS). Disease spreading insect tracking has applications in virology and epidemiology.

In Chapter IV, we describe an economic, scalable approach to DNA barcode sequence design. Numerous modular DNA barcodes are assembled from smaller block variants. Computational sequence design and scoring was performed using Python and a nucleic acid secondary structure prediction program. Analysis of next-generation sequencing (NGS) data informs specificity of barcode design while supporting the high multiplexing capabilities boasted from modular DNA barcodes. Design and deployment of DNA barcodes has applications in inventory management and forensic science.

# CHAPTER 2: CHARACTERIZATION OF GUEST DNA TRANSPORT AND ADSORPTION WITHIN HOST POROUS PROTEIN CRYSTALS[1]

## 2.1 Overview

Nucleic acid transport through protein-based pores is a well characterized phenomenon due, in part, to advancements in nanopore sequencing (30). A less studied area is nucleic acid transport through extended protein-based channels where the additional surface area and increased contact time allow for the study of prolonged binding interactions. Porous protein crystals composed of "CJ" a putative isoprenoid-binding protein from *Campylobacter jejuni* represent a favorable, highly-ordered material for studying DNA transport and binding/un-binding along protein-based channels. These crystals adopt a hexagonal prism shape and contain a densely packed hexagonal array of axial 13-nm diameter nanopores that run from the top to the bottom of the crystal. After crosslinking, the crystals are easily manipulated for experimentation. An adsorption isotherm between host crystals and guest double-stranded 8 base pair DNA (8mer) revealed a high equilibrium adsorption constant of 207 Liters/gram and a maximum guest binding capacity of $3.13 \times 10^{-5}$ mol/mL. Fluorescence confocal microscopy revealed guest DNA loading into host crystals predominately along the major axial crystal nanopores. Four different computational models based on the finite volume method were assessed to model the transport process for guest 8mer dsDNA loading into empty host crystals

[1] The work presented in this chapter is described in the following manuscript: J.D. Stuart, S. Chen, C.D. Snow, Characterization of Guest DNA Transport and Adsorption within Host Porous Protein Crystals. *J Phys Chem B* (In preparation). **Author contributions**: C.D.S., J.D.S., S.C. designed research; J.D.S., S.C. performed research; J.D.S., S.C. analyzed data; J.D.S., S.C. wrote the paper; C.D.S., J.D.S., S.C. edited the manuscript.

in terms of fundamental parameters such as the intra-pore diffusion constant. The diffusion of DNA at equilibrium was also assessed via fluorescence recovery after photobleaching (FRAP) datasets resulting in a predicted intra-pore diffusion coefficient of $1.37 \times 10^{-9}$ $cm^2/sec$, two orders of magnitude less than the calculated bulk solution diffusion coefficient using the Stokes-Einstein equation.

**2.2 INTRODUCTION**

Tough crosslinked protein matrices such as CJ crystals that can uptake and potentially protect otherwise vulnerable nucleic acid cargo may find diverse applications including edible barcodes (31) or the delivery of therapeutic RNA. However, to enable the rational design of new materials in this family, it is essential to obtain an ability to quantitatively model transport processes such as guest nucleic acid uptake and release. Nucleic acid pore-mediated transport is a phenomenon in both biological and synthetic settings (30, 32). For example, mRNA enters the cytosolic space via the nuclear pore complex preceding eukaryotic translation (33). Also, single-stranded DNA traverses a protein-delimited nanopore during nanopore sequencing (30). Nucleic acid transport through porous media is also commonly encountered in both bench and applied research settings including, but not limited to, separating DNA based on size and characterizing ground aquifers, respectively (34, 35).

Conventional porous media possess several limitations frustrating the study of nucleic acid transport. Specifically, one has little control, if any, on the pore length and shape traversed by the nucleic acid and electrochemical composition of the porous environment. In contrast, the solvent channels of a protein crystal are very well defined and highly repetitive. A better

understanding of DNA transport within highly porous protein crystals may shed light on comparable biophysical systems.

Porous protein crystals represent a novel media capable of providing intimate knowledge of the pore environment for studying guest DNA transport (23). The "CJ" porous protein crystals employed herein are derived from cj0420 (UNIPROT Q79JB5), a putative isoprenoid-binding protein from *Campylobacter jejuni.* While previously shown to be 1 of 5 putative glycoproteins that exhibit upregulation in a *C. jejuni* strain lacking the transcriptional regulator DksA-like protein (36), the appealing aspect of CJ lies in the crystals formed by the protein. The modified protein, which lacks the native N-terminal signal sequence (MKKVLLSSLVAVSLLSTGLFA) (37), and is augmented with a C-terminal histag, forms transparent

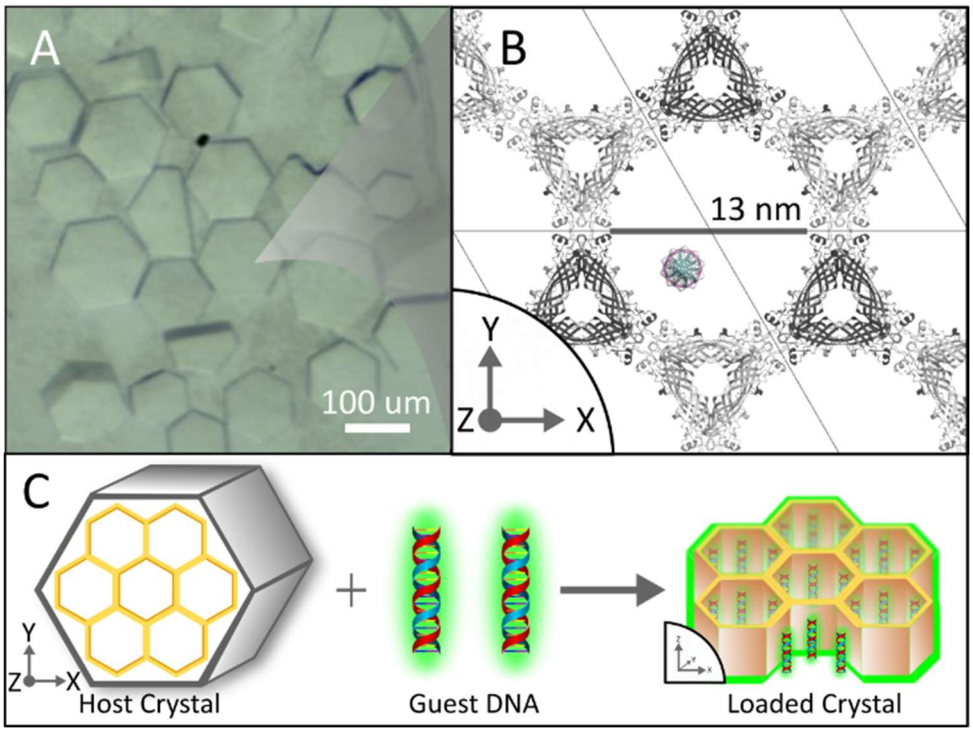



**Figure 2.1. Porous Protein Crystals. A)** Transparent, hexagonal porous protein crystals comprised of "CJ" an isoprenoid binding protein derived from *Campylobacter jejuni*. **B)** PyMol schematic of crystal topology, highlighting the parallel array of 13nm diameter axial nanopores spanning the crystal height. To illustrate scale, the center pore contains a dsDNA molecule. **C)** Experimental overview showing the mixing of fluorophore labeled guest DNA with host crystals for observing crystal uptake and adsorption with confocal microscopy.

hexagonal prism crystals comprised of an array of axial nanopores that are 13 nm in diameter (fig. 2.1) (24). Due to their crystalline nature, x-ray diffraction (XRD) can be used to determine the ordered structure components of the nanopore environment with atomic resolution. Moreover, given that the nanopores are lined with amino acid sidechains, recombinant protein variants can be crystallized to alter the physicochemical composition of the nanopore walls and observing the subsequent influence on guest DNA loading (23). Together, these characteristics make porous protein crystals an ideal porous media for studying nucleic acid pore-mediated transport and adsorption.

Previous work analyzing guest molecular transport through host crystals includes using confocal laser scanning microscopy (CLSM) for capturing anisotropic guest fluorescein transport into host lysozyme crystals (38-40). Additionally, species ranging from amino acids (41) and entire proteins (25) to enzymatic substrates (42, 43) have been soaked in porous protein crystals. Ligand soaking into crystals was previously described as a scaffold-assisted structure determination tool in the context of drug discovery and optimization (44). Germeia et al. showed that a relatively simple Fickian diffusion model could fit the observed rate of a variety of small molecules diffusing into a variety of host protein crystals (45). Fluorescence confocal microscopy has been used for visualizing guest loading into host crystals (38), in addition to quantifying intra-crystal guest concentration and pH (46). Guest diffusion within porous microparticles has also been described using Fluorescence Recovery after Photobleaching (FRAP) (47) and guest loading into host crystals exhibits hindered diffusion due to adsorption as previously described for proteins within porous agarose gel (48, 49). A common technique for probing molecular diffusion, FRAP involves photobleaching a fluorescent region internal to a

sample, followed by monitoring the gradual increase of fluorescence of the bleached region due to inward diffusion of neighboring unbleached species (50).

Recently, the rate of diffusion and binding kinetics of gold nanoparticles into host CJ protein crystals was described (51). Guest nanoparticle diffusion rates within the crystal nanopores were found to be reduced by guest accumulation within host crystals over time, and were attenuated relative to bulk diffusion. While previous work has described guest molecular diffusion through protein crystals, to our knowledge, no work has examined guest nucleic acid transport within porous protein crystals. The difference in electrochemical composition between gold nanoparticles and DNA, particularly throughout the solvent exposed surface area for the two, would likely result in different loading and simulation results.

This work communicates the characterization of the biophysical interactions governing guest nucleic acid transport into host porous protein crystals. An adsorption isotherm was performed and fit with the Langmuir model for determining an equilibrium adsorption constant. Competing models such as the Freundlich model provided an inferior fit (Fig. S2.1). Fluorescence confocal microscopy was used to image the uptake of guest nucleic acids into empty host crystals. Four different models (Fig. 2.2) were evaluated for modeling the uptake transport of guest DNA into empty crystals by calculating the difference (sum of squared deviations) between the observed guest fluorescence and predicted guest fluorescence as a function of time and position within the crystal. To evaluate the equilibrium transport of DNA within already loaded crystals, we compared simulated transport for bleached and unbleached DNA to experimental Fluorescence Recovery After Photobleaching (FRAP) results.

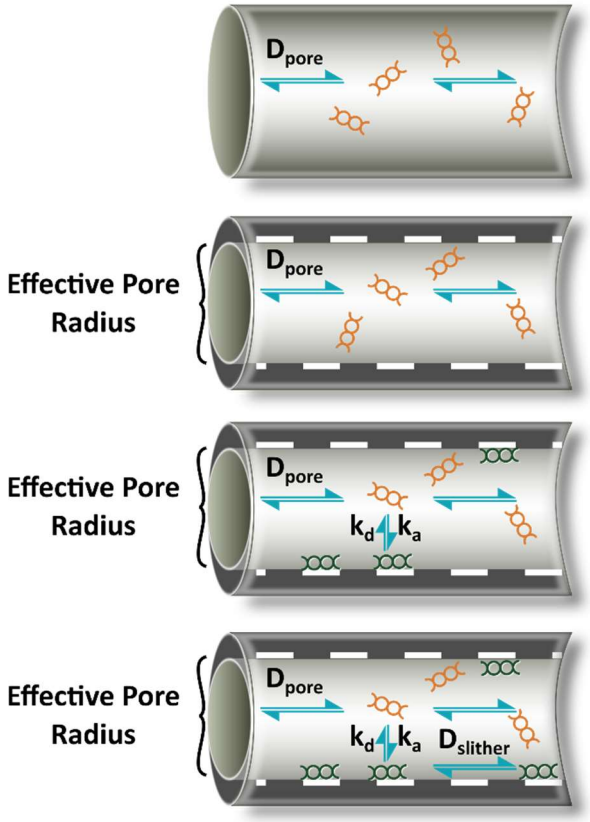



**Figure 2.2. FVM Models. A)** Model 1 schematic employing a single fit variable, Dpore. **B)** Model 2 schematic employing 2 fit variables: $D_{pore}$, pore reduction. **C)** Model 2 schematic employing 4 fit variables: $D_{pore}$, pore reduction, $k_a$, and $k_d$. **D)** Model 4 employing 5 fit variables: $D_{pore}$, pore reduction, $k_a$, $k_d$, and $D_{slither}$.

## 2.3 THEORY

The basis for transport simulations is mass conservation. The finite volume method (FVM) is locally conservative in that the flux leaving a given volume is the same as that entering a neighboring volume, making FVM beneficial for numerical simulation of conservation laws(52). Similar to the finite element method (FEM), the overall control volume is divided into a finite number of non-overlapping elements(53). Herein, we employed the FVM, written in a custom Python script (available upon request), for modeling the loading of guest DNA into host crystals, as observed by fluorescence confocal microscopy. A series of 4 FVM models were tested for accurate prediction of intra-crystal transport.

As shown in table 2.1, fixed parameters for modeling guest loading include the guest radius, $r_g$, determined using dynamic light scattering (fig. S2.5). The initial pore radius, $R_{p0}$, is known from crystal unit cell measurements for the CJ protein (PDB ID 5w17). The nanopore length, $L_p$, is estimated from confocal datasets with the image plane parallel to nanopores. The Langmuir adsorption equilibrium constant, K, along with the estimated max guest binding capacity,$q_{max}$, was obtained from isotherm experiments. The bulk solution concentration was calculated from fluorescence intensity using a standard curve (fig. S2.4). Guest loading occurred at a fixed temperature, T, of 22 °C. The mass transfer coefficient, $k_m$, was adopted from previous literature.

**Table 2.1.** Fixed and Fit Parameters for Finite Volume Models

| Fixed Parameters | Description | Value | Model | | | |
|---|---|---|---|---|---|---|
| $r_g$ | guest radius | 8.1e-8 cm | | | | |
| $R_{p0}$ | length in r-direction | 6.5e-7 cm | | | | |
| $L_p$ | length in z-direction | 7.5e-3 – 25.8e-3cm | | | | |
| K (=$k_a/k_d$) | Langmuir adsorption equilibrium constant | 4.2 $cm^3/mol$ | | | | |
| $q_{max}$ | maximum adsorbed guest molecule concentration | 3.13e-5 $mol/cm^3$ | | | | |
| $C_{sol}$ | bulk solution concentration | 1.1e-10 – 2.0e-10mol/mL | 1 | 2 | 3 | 4 |
| T | temperature | 293.15 K | | | | |
| $D_0$ | bulk solution diffusion coefficient | 3e-7 $cm^2/sec$ | | | | |
| **Fit Parameters** | **Description** | **Unit** | | | | |
| $D_{pore}$ | pore diffusion coefficient | $cm^2/sec$ | | | | |
| $R_p$ | effective pore radius | -- | | | | |
| $k_a$ | adsorption rate constant | $cm^3/(mol \cdot sec)$ | | | | |
| $k_d$ | desorption rate constant | $sec^{-1}$ | | | | |
| $D_{slither}$ | bound species diffusion coefficient along pore walls | $cm^2/sec$ | | | | |

Model 1 (Fig. 2.2A) is quite simple. Transport is modeled as a pure axial diffusion process with a single diffusion constant $D_{pore}$. This model therefore assumes that all guest DNA are mobile, and subsumes any more complicated binding and unbinding behavior into an empirical diffusion constant. The Model 1 fitting equation is:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial z}\left(D_{pore}\frac{\partial C}{\partial z}\right) \quad (1)$$

where $D_{pore}$ corresponds to the intra-pore diffusion coefficient, $C$ is the local concentration of the mobile guest, $z$ is the position along the axial pore axis, and $t$ is time. Notably, most of the parameters needed for the simulation are fixed and are estimated from prior calculations. Fixed and fit parameters for Model 1 are shown in Table 2.1. Notably, the only fit variable for Model 1 is the pore diffusion coefficient.

Model 2 (Fig. 2.2B), expands upon Model 1 by including dynamic pore diameter reduction during guest loading. Specifically, a high local concentration of guest molecules could impede guest diffusion, effectively reducing the volume that is available for free diffusion. To model an effective pore radius Model 2 incorporates the following equations adapted from Hartje et al.(51, 54):

$$R_p = \sqrt{(R_{p0^2} - \left(4 * \pi * \frac{r_{g^3}}{3}\right) * q * \ R_{p0^2} * 6.022e23)} \quad (2)$$

$$\lambda = R/R_p \quad (3)$$

$$D_{pore} = \ D_0(1 + 1.125\lambda(ln\lambda) - 1.56034\lambda + 0.528155\lambda^2 + 0.270788\lambda^5 + 1.91521\lambda^3 - 2.819030\lambda^4 \quad + 1.10115\lambda^6 - 0.435933\lambda^7 \quad (4)$$

where $R_p$ is the effective pore radius, $R_{p0}$ is the initial pore radius, $R$ is the guest radius, $D_p$ is the pore diffusion coefficient, and $D_0$ is the bulk solution diffusion coefficient calculated with the Stokes-Einstein equation. Equation 2 calculates the effective pore radius throughout the pore length over time based on the local concentration of adsorbed guest. The effective pore radius is used in equation 3 to calculate lambda, λ, the quotient between the guest molecule radius and the host crystal pore radius. The pore diffusion coefficient is calculated using lambda from equation 4 as previously described(54).Fixed and fit parameters for Model 2 are shown in Table

2.1. The fit variables for Model 2 include the intra-pore diffusion coefficient $D_{pore}$ and the effective pore radius, $R_p$.

Model 3 (Fig. 2.22C), further expands upon Model 2 by including the possibility that guest molecules can bind to the host crystal and become immobile. While this model now includes adsorption ($k_a$) and desorption ($k_d$) rate constants, we only add a single fitting parameter since the $k_a/k_d$ ratio is constrained to be equal to the experimentally measured Langmuir isotherm equilibrium constant. Model 3 uses the following equation:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial z}\left(D_{pore}\frac{\partial C}{\partial z}\right) - k_a C(q_{max} - q) + k_d q \quad (5)$$

where the parameters are the adsorption rate constant ($k_a$), the desorption rate constant ($k_d$), the current local adsorbed guest concentration ($q$), and the predicted maximum adsorbed guest molecule concentration ($q_{max}$., Other fixed parameters for Model 3 are shown in Table 2.1. Notably, Model 3 was previously successfully used to model the uptake of guest gold nanoclusters into empty host CJ crystals (51). All parameters are not completely independent as the equilibrium adsorption constant was used to determine the adsorption and desorption rate constants ($K_L = k_a/k_d$).

Model 4 (Fig. 2.2D), further expands on Model 3 by considering that adsorbed guest might still undergo transport. Essentially Model 4 posits that there may be two mobile species with varying nanopore diffusion rates. To this end, Model 4 introduces an additional diffusion constant $D_{slither}$, to allow bound guest to move along the nanopore axis. Model 4 uses the following equation:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial z}\left(D_{pore}\frac{\partial C}{\partial z}\right) + \frac{\partial}{\partial z}(D_{slith}\quad\frac{\partial q}{\partial z}) - k_a C(q_{max} - q) + k_d q \tag{6}$$

Fixed parameters for Model 4 are shown in Table 2.1. The fit variables for Model 4 include pore diffusion coefficient ($D_p$), effective pore radius ($R_R$), adsorption rate constant ($k_a$), desorption rate constant ($k_d$), and the bound species diffusion coefficient ($D_{slither}$).

For comparing to the predicted nanopore diffusion coefficient, the bulk solution diffusion coefficient was calculate using the Stokes-Einstein equation:

$$D = \frac{k_B T}{6\pi\eta r} \tag{7}$$

where $k_b$ is the Boltzmann constant, T is temperature, η is the solution viscosity, and r is the guest molecule hydrodynamic radius determined using dynamic light scattering (fig. S2.5).

## 2.4 RESULTS

### 2.4.1 Adsorption Isotherm

An adsorption isotherm experiment was performed on a collection of porous protein crystals by first measuring the side length and height of individual crystals as shown in Figure 2.3A, allowing calculation of the crystal volume (fig. 2.3B) and the resultant crystal mass. Crystals were incubated with FAM-labeled guest double-stranded DNA (5' - CGCTGGCG – 3'), 8mer, and the change in supernatant DNA concentration pre- and post- incubation was recorded for increasing guest DNA concentrations. The ratio of adsorbed DNA mass to estimated scaffold crystal mass ($q$) was plotted as a function of $C$, the final equilibrium external guest DNA concentration. The adsorption isotherm results are shown in Figure 2.3C. The adsorption isotherm results were consistent with a highly favorable interaction between guest DNA and

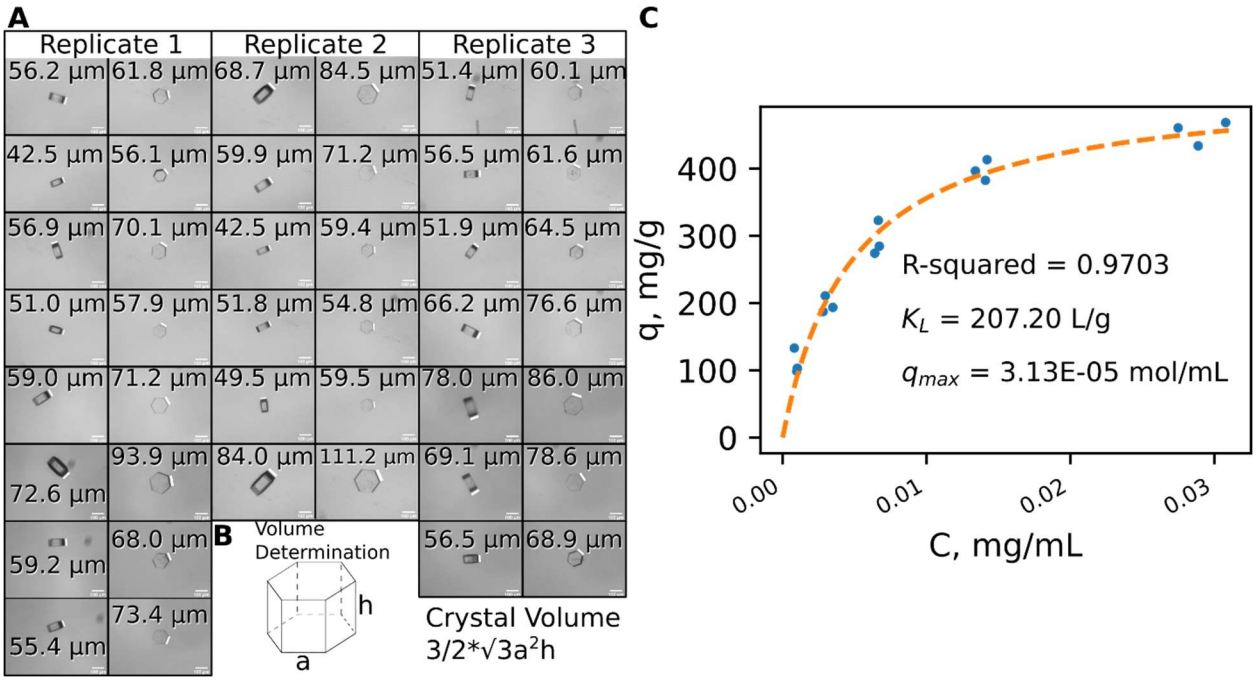



**Figure 2.3. Adsorption Isotherm. A)** Images of individual crystals taken for measuring crystal side length and height. **B)** The crystal dimensions were used, along with the unit cell volume provided by the PDB file (2fgs) to estimate total crystal mass. **C)** Adsorption isotherm results plotted (circle) were well fit by a Langmuir curve (dashed line).

host crystals. Specifically, the dataset was fit using a Langmuir model yielding an $R^2$ of 0.9703. Competing models (e.g. Freundlich) had a slightly inferior fit yielding an $R^2$ of 0.9507 (Fig. S2.1). Ultimately, the monolayer adsorption implied by the Langmuir model may also be more plausible than the multilayer adsorption of the Freundlich model considering the electrostatic repulsion of the guest DNA under the buffer conditions in use (Tris-EDTA buffer only) and considering the confined nanopore environment. The calculated equilibrium adsorption constant was 207 Liters/gram and the maximum binding capacity was 3.13 x $10^{-5}$ mol/mL, with the latter corresponding to approximately 43 oligonucleotides per unit cell. While the maximum guest binding capacity is on the same order as previously reported for gold nanoparticles (7.16 x $10^{-5}$ mol/mL), the equilibrium adsorption constant reported here is two orders of magnitude greater than the previously reported value for gold nanoparticles (7.51 L/g) (51). Previously, the adsorption of xanthene dyes within crosslinked lysozyme crystals resulted in adsorption

constants ranging from 3 x $10^4$ – 1.06 x $10^6$ L/mol and maximum binding capacity values ranging from 5.8 x $10^{-3}$ mol/kg – 1 x $10^{-1}$ mol/kg (55). The elevated equilibrium adsorption constant suggests the rate of guest DNA unbinding from host crystals occurs much slower than guest binding.

### 2.4.2 Confocal Microscopy

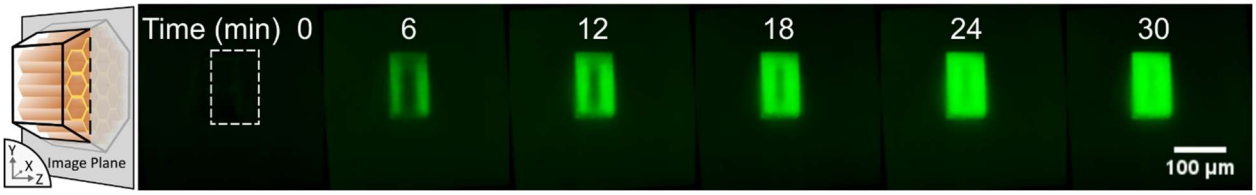



**Figure 2.4. Fluorescence Microscopy.** Unloaded crystals were oriented on the side such that the parallel nanopores ran left-to-right in the image plane. Initially exhibiting low background fluorescence, fluorophore-labeled DNA was added to the crystal solution. Over time, DNA accumulated within the crystals in a symmetrical fashion that started at the crystal edges containing nanopore openings and gradually increased toward the crystal center.

The fluorescent tag (FAM) on the ds8mer allowed guest loading to be recorded using fluorescence microscopy. Figure 2.4 displays a timelapse of a single replicate host crystal during guest loading. Loading datasets for additional replicates are shown in Figure S2.2. The crystals, initially non-fluorescent, displayed an increase in fluorescence that was first detected at the edges of the crystals. As loading progressed, the fluorescence migrated towards the crystal interior. As expected, the confocal loading datasets clearly show that guest loading into host crystals was anisotropic, with guest loading dominated by z-axis diffusion along the 13-nm diameter nanopores. Crystals became nearly saturated after only 30 minutes of incubation.

### 2.4.3 Modeling of Guest Loading

To estimate the biophysical parameters governing guest loading into host crystals such as the nanopore diffusion coefficient and the kinetic rate constants of binding and unbinding, a total of 4 computational models were employed for fitting confocal microscopy loading datasets. The

models, numbered 1-4, increase in complexity by incorporating additional molecular phenomena and fitting parameters. Model 1, represents guest uptake as a simple diffusive process, with a single empirical diffusion constant to represent the entire transport process. The resulting empirical $D_{\underline{pore}}$ may be significantly underestimate the intra-nanopore diffusion rate if the true process involves guest immobilization via binding. The second model, Model 2, again estimates the guest nanopore diffusion coefficient but also considers the possibility that the effective nanopore diameter may be reduced during loading. It is not surprising, therefore, that the estimated intrapore diffusion constant for Model 2 was higher than Model 1 (to compensate for a constricted pore). Model 3, was previously used for simulating gold nanocluster loading into host crystals and introduces interconversion with an immobile guest

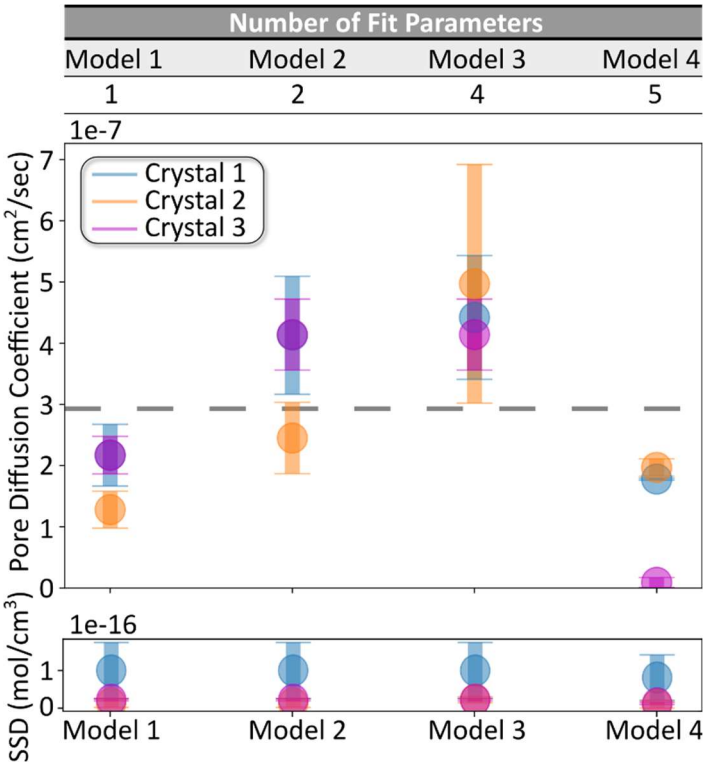



**Figure 2.5. Modeling of Guest Loading.** The number of fit parameters for each of the 4 FVM models employed (**top**). The predicted pore diffusion coefficient for each of the three crystals among all models (**middle**). Circle denotes the average of 3 crystal image widths used for analysis. Error bars denote the standard deviation. Dashed line represents the calculated diffusion coefficient using the Stokes-Einstein equation. The sum of the squared deviations from the model and data for all models (**bottom**).

population via adsorption ($k_a$) and desorption ($k_d$). Finally, Model 4, introduces a fourth parameter, $D_{slither}$, that considers that a strongly bound guest population might still translocate along crystal nanopore while bound. In other words, Model 4 postulates that the observed transport might reflect relatively unhindered diffusion ($D_{pore}$) for guest DNA that is in the middle of the nanopores and not directly interacting with the host crystal, simultaneously with slower transport for DNA that maintains more direct interactions via a likely combination of electrostatics and hydrogen bonding with the scaffold crystal proteins.

We proceeded to identify the fit parameters for each model that resulted in the best fit to the experimental data (the lowest SSD). Figure 2.5 shows that the best fit parameters for all four models resulted in comparable deviations with experiment (SSD values were $4.52 \times 10^{-17}$, $4.52 \times 10^{-17}$, $4.37 \times 10^{-17}$, and $3.44 \times 10^{-17}$ mol/cm$^3$ for Models 1-4, respectively). It is somewhat surprising that the additional fitting parameters did not allow Models 2, 3, or 4 to greatly improve upon the fit quality for Model 1. It is also important to assess whether the best-fit models are physically reasonable. The estimated intra-pore diffusion coefficient ($D_{pore}$) from each of the models employed are shown in figure 2.5. Models 2 and 3 result in values greater than the calculated bulk solution diffusion coefficient represented by the dashed gray line. Given that the guest diffusion is likely to be attenuated within the nanopores due to the high local concentration of guest DNA and host protein, the overestimation in pore diffusion by Models 2 and 3 was treated as artifactual. In contrast, both models 1 and 4 resulted in pore diffusion coefficients less than the bulk solution value. Given that models 1 and 4 yielded similar results despite model 4 being more complex, model 1, the simplest model, was chosen for downstream modeling of guest diffusion at equilibrium.

The data fitting process, for all models, began with specifying a range of initial guest DNA diffusion coefficients, followed by modeling, calculating the sum of squared deviations (SSD) between the experimental data and model, and then plotting the SSD as a function of the pore diffusion coefficient shown for model 1 (Fig. 2.6A). One challenge with the experimental data is that the edges of the crystal can have a fluorescence intensity that is artificially high due to internal reflection effects. Since the same transport model should work throughout the host

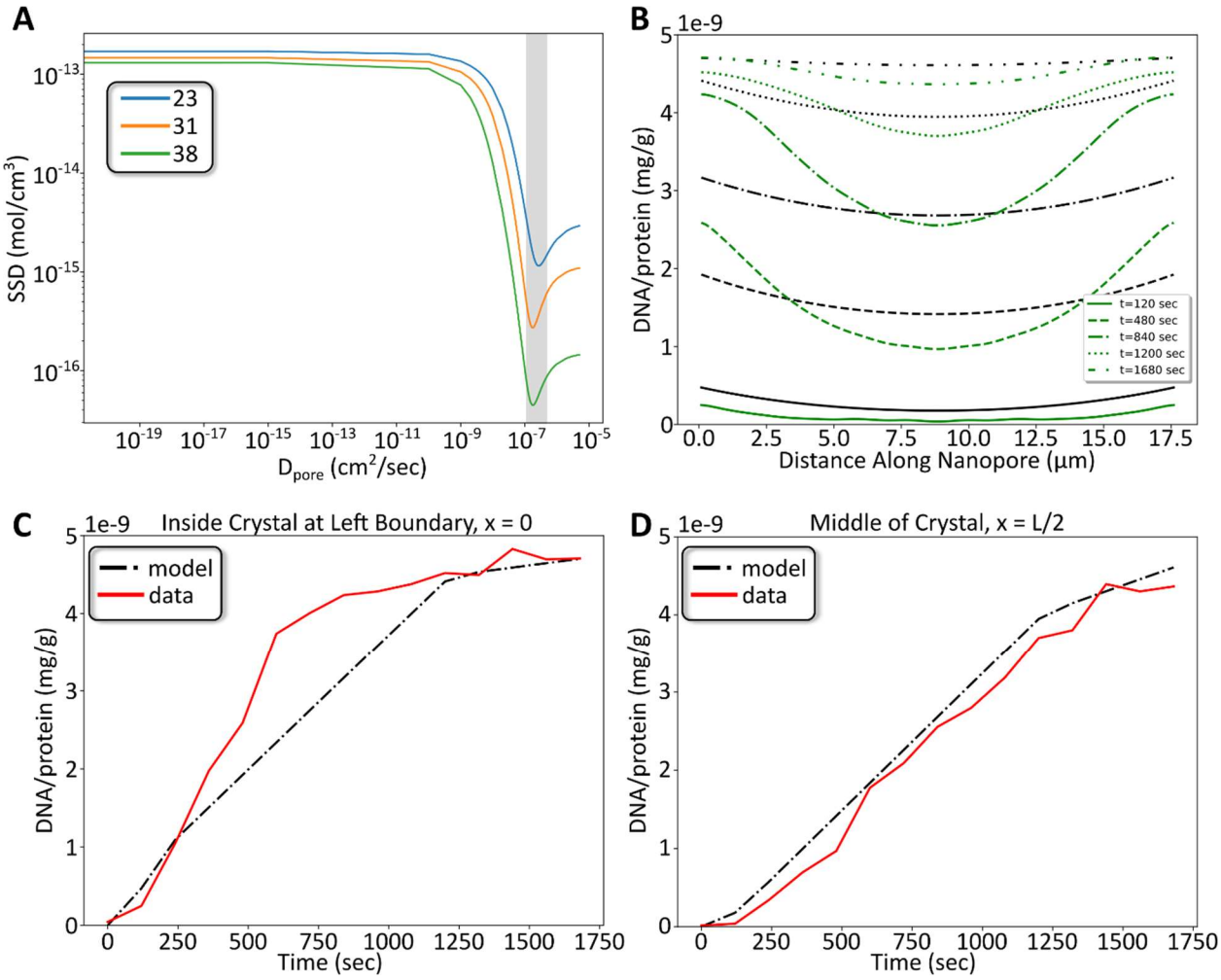



**Figure 2.6. Modeling Results. A)** The sum of squared deviations (SSD) between model 1 and the experimental data plotted as a function of $D_{pore}$ ranging from $10^{-5}$ to $10^{-19}$ for a single replicate. The number (inset) for each curve corresponds to the number of pixels trimmed from confocal loading dataset images prior to modeling. Highlighted region spans $D_{pore}$ range with the minimum SSD. **B)** Experimental guest DNA concentration plotted along with corresponding model curves as a function of distance along crystal nanopores at 5 different timepoints. Green curves represent experimental confocal data and black curves represent model 1. **C)** Adsorbed DNA concentration and model plotted as a function of time at the crystal boundary. **D)** Adsorbed DNA concentration and model plotted as a function of time at the crystal center.

crystal and for any subsection thereof, we decided to fit transport to the interior region of the crystal, and to verify that the extent of crystal edge trimming did not affect the results. Accordingly, modeling was performed on confocal loading images after trimming increasing widths of the image edges (23, 31, and 38 pixel widths). The best pore diffusion coefficient was chosen as that with the lowest SSD value. For Model 1, regardless of image width for model 1, all the calculations converged on virtually the same empirical pore diffusion coefficient with the smallest SSD, yielding an average pore diffusion coefficient ($D_{pore}$) of (1.9 ± 0.6) x $10^{-7}$ $cm^2$/sec. The data and model curves from the minimum SSD Dpore were plotted as a function of distance across the parallel nanopores at various timepoints (fig. 2.6B) displaying an initially growing deviation between the experimental data and model that recedes toward the end of the experiment. The adsorbed guest DNA was then plotted as a function of time at the crystal edge (fig. 2.6C) and crystal center (fig. 2.6D) displaying an elevated agreement between the data and model at the crystal center than at the crystal edge.

**2.4.4 Fluorescence Recovery After Photobleaching (FRAP)**

The data above was collected under non-equilibrium conditions as guest DNA diffused into a previously unloaded crystal. In contrast, to determine the guest pore diffusion coefficient when adsorbed DNA is at equilibrium and at high concentration, fluorescence recovery after photobleaching (FRAP) was performed at 3 different regions on a single crystal loaded with guest 8mer. Following DNA loading, the crystal was transferred to buffer to remove unbound DNA . For each FRAP experiment, a region within the loaded crystal was exposed to 405 nm

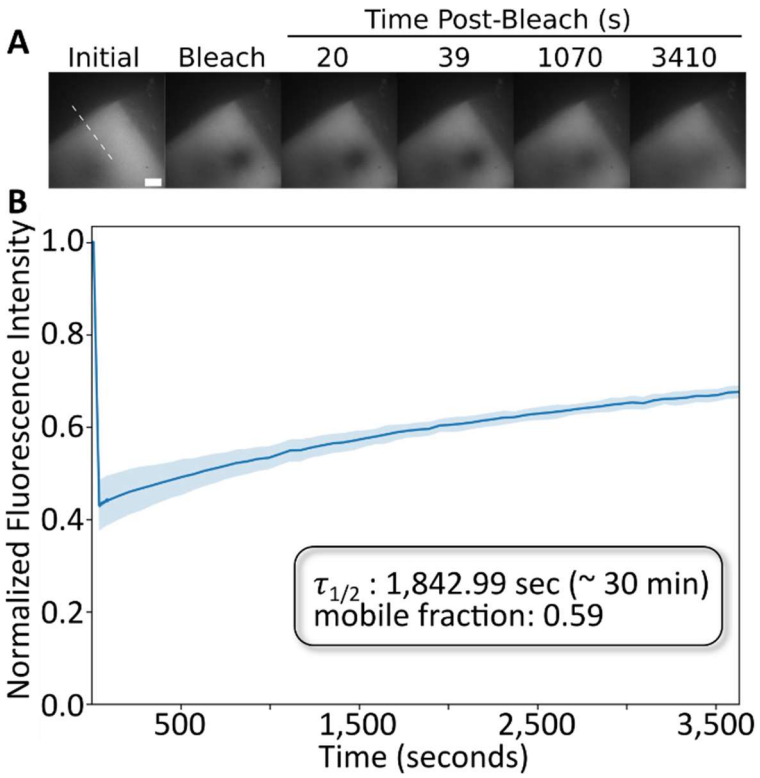



**Figure 2.7. Fluorescence Recovery After Photobleaching. A)** Timelapse imaging of a single crystal pre- and post-bleaching. Each crystal was oriented on its side with the nanopore axis indicated by the dashed line. Scale bar denotes 10 µm. **B)** Fluorescence recovery plotted as a function of time. The blue line represents the average intensity from 3 replicates and the shaded region represents the standard deviation.

light to photobleach DNA conjugated fluorophores, followed by monitoring the gradual increase of fluorescence of the bleached region due to inward diffusion of neighboring unbleached DNA. Immediately following bleaching, a dark circular region was created that gradually accumulated an increase in fluorescence signal due to diffusion of unbleached guest 8mer into the bleached region (fig. 2.7A). Using the online FRAP analysis tool EasyFRAP (56), the calculated average half-time of full recovery ($\tau_{1/2}$) was 1,843 sec ( ~ 30 min) with an estimated mobile fraction of 0.59, indicative of a 0.41 immobile fraction which may suggest that a fraction of the guest DNA is bound much more tightly than the population that is responsible for the observed FRAP recovery (Fig. 2.7).

### 2.4.5 Modeling of Equilibrium Guest Diffusion

Since the more complex models (Models 2,3, and 4) failed to significantly improve upon the quality of the fit, we decided to use Model 1 to model equilibrium pore diffusion. Similar to the transport modelling results, the equilibrium FRAP modelling results are shown in figure 2.8A. Performed with model 1, data fitting resulting in a predicted $D_{pore}$ coefficient of 1.4 x $10^{-9}$ $cm^2$/sec, approximately 2 orders of magnitude lower than the calculated bulk diffusion

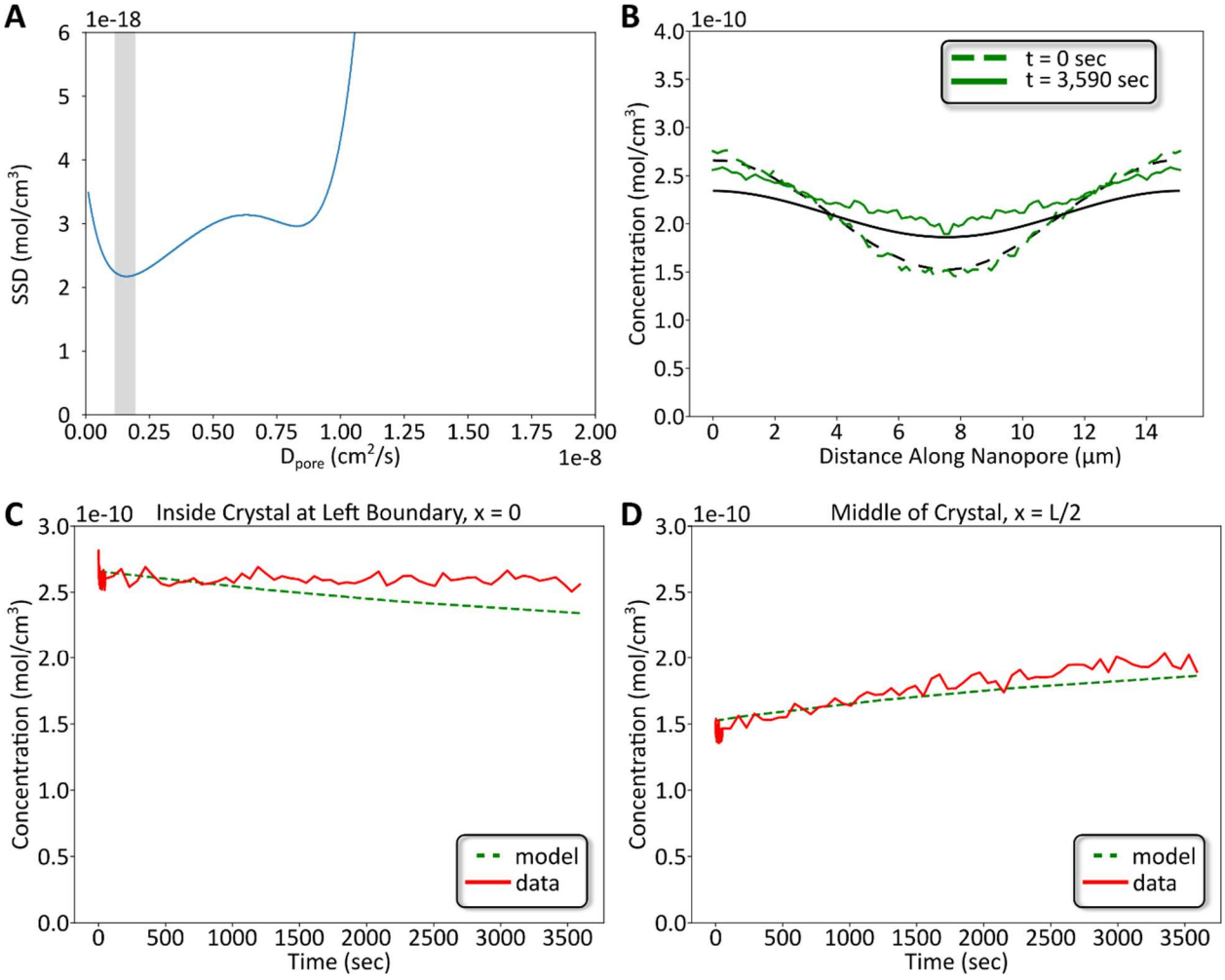



**Figure 2.8. FRAP Model Results. A)** The sum of squared deviations (SSD) between model 1 and experimental FRAP data plotted as a function of Dpore ranging from 0 to 1.25e-8 $cm^2$/s for a single replicate. Highlighted regions spans $D_{pore}$ range with the lowest SSD. **B)** Experimental adsorbed guest DNA concentration plotted along with corresponding model curves as a function of distance along crystal nanopores immediately after photobleaching and after ~1 hour. Green curves represent experimental FRAP data and black curves represent model 1. **C)** Adsorbed DNA concentration and model plotted as a function of time at the crystal boundary. **D)** Adsorbed DNA concentration and model plotted as a function of time at the crystal center.

coefficient of 2.9 x $10^{-7}$ $cm^2$/sec using the Stokes-Einstein equation (eqn. 7). Comparison to the $D_{pore}$ obtained for the uptake of DNA into empty crystals (1.9 x $10^{-7}$ ± 0.6 x $10^{-7}$ $cm^2$/sec) reveals that predicted pore diffusion at equilibrium is slower compared to pore diffusion during guest loading. An alternative approach for solely calculating diffusion coefficients from confocal datasets as previously described (57) yielded a similarly attenuated pore diffusion coefficient from experimental FRAP data (fig. S2.3). Model results display agreement with experimental data suggested by the overlapping time-based curves shown in figure 8B-D. These results demonstrate the versatility of model 1 in estimating both transport and equilibrium pore diffusion.

**2.5 CONCLUSIONS**

The work performed herein communicates the characterization of DNA transport into empty host porous protein crystals and DNA diffusion within host crystals at equilibrium. Adsorption isotherm results suggest guest DNA desorbs from host crystals much more slowly than guest DNA adsorbs. Guest DNA loading into host crystals, observed using confocal microscopy, occurred symmetrically along the major axial nanopores. Among the 4 models tested, the more complex models resulted in fits with unrealistically large intra-pore diffusion (Models 2 & 3) or with no significant improvement to the fit despite the addition of more parameters (Model 4). As expected, modeling of pore diffusion at equilibrium from experimental FRAP datasets resulted in a predicted pore diffusion coefficient (1.4e-9 $cm^2$/sec) that was attenuated relative to the calculated bulk solution diffusion coefficient (2.9e-7 $cm^2$/sec) as well as the estimated intra-pore diffusion coefficient (1.9e-7 $cm^2$/sec) observed during guest uptake into empty nanopores.

As previously reported for gold nanoparticles, the predicted pore diffusion coefficient was reduced by approximately 70 % relative to calculated bulk diffusion. In contrast, the pore diffusion coefficient for guest DNA was reduced by approximately 36 % relative to calculated bulk diffusion. The rate of nucleic acid transport into host crystals was less susceptible to restricted volume effects relative to gold nanoparticles. In addition to particle surface electrochemical composition, the difference in guest particle volume likely influenced the pore transport rate. The guest radius for gold nanoparticles was reported as $1.54 \times 10^{-7}$ cm resulting in an estimated guest volume, assuming a spherical guest shape, of $1.53 \times 10^{-20}$ $cm^3$. Here, the guest nucleic acid radius, obtained via DLS, was $8.1 \times 10^{-8}$ cm, resulting in an estimated guest volume of $2.2 \times 10^{-21}$ $cm^3$, which is approximately 85 % less volume occupied by the guest nucleic acid. For the same number of particles present at the pore entrance, a greater volume is occupied by the gold nanoparticles resulting in a greater hindrance to neighboring nanoparticles in solution for pore access. This elevated hindrance resulted in a greater reduction in pore diffusion for gold nanoparticles relative to nucleic acid as shown in this study. Future work includes examining the influence of guest DNA length and sequence on crystal loading behavior while also optimizing model parameters for enhancing agreement between predicted fit curves and experimental guest loading data.

## 2.6 METHODS

### 2.6.1 Protein expression, Porous Crystal Growth and Crosslinking

Protein expression, crystal growth using sitting-drop vapor diffusion and 1-Ethyl-3-(dimethylaminopropyl)carbodiimide hydrochloride (EDC) crosslinking were performed as previously described (21, 23).

### 2.6.2 Adsorption Isotherm

Several EDC crosslinked crystals (6 – 8) were imaged using Motic software for measuring individual crystal size that allowed for calculation of total crystal volume and mass for each replicate. Crystals were transferred to a Qubit tube containing 200 µL of incrementally increasing FAM labeled 8mer (5' - CGCTGGCG – 3') concentrations: 0.625, 1.25, 2.5, 5, and 10 µM. The solution was covered with 75 µL hexane to reduce solvent evaporation. The guest 8mer concentration in solution was recorded using a Qubit 2 fluorometer. Following incubation and shaking for several days (3 – 4) to achieve equilibrium, the final guest solution concentration was recorded. Approximately half of the guest solution (100 µL) was removed and replaced with a more concentrated guest solution followed by re-initiation of incubation and shaking. Using the decline in solution guest DNA amount during incubation, the crystal adsorbed guest DNA mass was calculated, normalized to total crystal mass and plotted as a function of the equilibrium guest concentration.

### 2.6.3 Confocal Microscopy

Individual 1-Ethyl-3-(3-dimethylaminopropyl) Carbodiimide Hydrochloride (EDC) crosslinked CJ crystals (3 replicates) were transferred to a microwell containing 0.4 % agarose for immobilization during imaging, while also ensuring the crystal remained in the correct orientation, such that the axial nanopores were parallel with the image plane. Following immobilization, 4 µL of 500 nM 8mer in TE buffer (10 mM Tris-HCl, 1mM EDTA·$Na_2$), pH 7.5 was added to the microwell and then sealed with transparent tape. Image acquisition was initiated following addition of guest solution. Images were taken at various planes within the crystals every 2 minutes for 30 minutes. Loading experiments were performed on a Nikon Eclipse Ti

confocal microscope equipped with an Andor iXon Ultra 897U EMCCD camera and observed with a Plan Apo λ 20x objective and 488 nm excitation laser set to 10% power. Images were analyzed using Fiji (58) and custom python scripts for modeling (code available upon request).

### 2.6.4 Modeling of Guest Loading

The finite volume method (FVM) was used to simulate guest loading into host crystals. Written in Python, the code used equilibrium adsorption data and confocal loading datasets as inputs for fitting the parameters shown in Table 1 depending on the model used (code available upon request).

### 2.6.5 Fluorescence Recovery After Photobleaching (FRAP)

Fluorescence Recovery After Photobleaching (FRAP) was conducted on a Nikon Eclipse Ti confocal microscope equipped with a 100X Plan Apo λ oil-immersion objective. FRAP was performed in triplicate on a single crystal previously incubated with 10 µM 8mer in TE buffer, pH 7.5 to achieve equilibration for approximately 24-hrs. Following incubation, the loaded crystal was transferred to fresh TE buffer to remove unbound, excess DNA for approximately 24 hrs. For each replicate, a ~10 µm diameter region was exposed to 100% 405nm light for 30 sec to achieve photobleaching, followed by observation of recovery at 1% 488nm light for 1 hr. Image acquisition settings, or when and how often an image is taken, during FRAP are shown in table 2.2.

**Table 2.2.** FRAP acquisition settings

| Phase | Interval | Duration | Loops |
|---|---|---|---|
| 1 | 1 sec | 9 sec | 10 |
| 2 | No Delay | 30 sec | 1 |
| 3 | No Delay | Continuous | 100 |
| 4 | 1 min | 59 min | 60 |

FRAP data analysis was performed as previously described (50). For each FRAP dataset, 3 regions of interest (ROI) were defined using the image analysis software Fiji (58). ROI 1 was

the bleached region, ROI 2 was the entire crystal, and ROI 3 was the dark background not containing the crystal. For each image, the following equation was used for background subtraction and data normalization:

$$\frac{ROI_1(t) - ROI_3(t)}{ROI_2(t) - ROI_3(t)} \; x \; \frac{ROI_2(t_0) - ROI_3(t_0)}{ROI_1(t_0) - ROI_3(t_0)}$$

# CHAPTER 3: MOSQUITO TAGGING USING DNA-BARCODED NANOPOROUS PROTEIN MICROCRYSTALS[2]

## 3.1 Overview

Conventional mosquito marking technology for mark-release-recapture (MRR) is quite limited in terms of information capacity and efficacy. To overcome both challenges, we have engineered, lab tested, and field evaluated a new class of marker particles, in which synthetic, short DNA oligonucleotides (DNA barcodes) are adsorbed and protected within tough, crosslinked porous protein microcrystals. Mosquitoes self-mark through ingestion of microcrystals in their larval habitat. Barcoded microcrystals persist transstadially through mosquito development if ingested by larvae, do not significantly affect adult mosquito survivorship, and individual barcoded mosquitoes are detectable in pools of up to at least 20 mosquitoes. We have also demonstrated crystal persistence following adult mosquito ingestion. Barcode sequences can be recovered by qPCR and next-generation sequencing (NGS) without detectable amplification of native mosquito DNA. These DNA-laden protein microcrystals have the potential to radically increase the amount of information obtained from future MRR studies compared to previous studies employing conventional mosquito marking materials.

---

[2] The work presented in this chapter is described in the following publication: Julius D Stuart, Daniel A Hartman, Lyndsey I Gray, Alec A Jones, Natalie R Wickenkamp, Christine Hirt, Aya Safira, April R Regas, Therese M Kondash, Margaret L Yates, Sergei Driga, Christopher D Snow, Rebekah C Kading, Mosquito Tagging Using DNA-Barcoded Nanoporous Protein Microcrystals. *PNAS Nexus* 10.1093/pnasnexus/pgac190, pgac190 (2022). **Author contributions**: R.C.K., C.D.S., J.D.S., D.A.H., L.I.G. designed research; J.D.S., D.A.H., A.A.J., L.I.G., N.R.W., C.D.S., R.C.K., S.D., M.Y., A.S., T.K., A.R. and C.H. performed research; J.D.S., D.A.H., A.A.J., N.R.W. analyzed data; J.D.S., R.C.K., C.D.S., D.A.H., L.I.G., A.A.J., N.R.W. wrote the paper; J.D.S., A.A.J., N.R.W., C.D.S., and R.C.K. edited the manuscript.

### 3.2 Introduction

The intensity of disease transmission to humans by mosquitoes, or vectorial capacity, depends, in part, on ecological parameters such as the feeding behavior of vectors on relevant vertebrate hosts, vector survivorship, dispersal patterns, and population density (59). Mark-release-recapture (MRR) is a standard approach to gather this epidemiologically significant information on mosquito behavior and ecology directly from field populations (60-62). Mosquito MRR studies have shaped our understanding of vector-borne disease transmission dynamics world-wide, comprising a body of literature of hundreds of studies on over 50 mosquito vectors of human pathogens (62). Among these studies, MRR targeting *Culex* (*Cx.*) *tarsalis* mosquitoes are among the most abundant (62).

Current marking techniques for mosquito dispersal studies include the use of topical fluorescent powders and paints (63), ingestible dyes (64), or larval habitat marking with nonradioactive rubidium chloride (65) and, separately, stable isotopes (66, 67). While fluorescent powders are widely used, they are limited by the difficulty of marking large numbers of mosquitoes, lack of powder retention, and impose negative effects on mosquito behavior and survivorship (63, 68, 69). Mosquitoes reared from larval habitats enriched in stable isotopes (70) or rubidium (65) can be detected via mass spectrometry or x-ray fluorescence spectrophotometry, respectively. However, these methods only provide a handful of distinguishable markers, and detection via mass spectrometry is expensive and training-intensive. As an alternative to physically marking mosquitoes, recent investigations have also employed the use of single nucleotide polymorphisms (SNPs) and spatial genetic tools to estimate geographical dispersal of mosquitoes from the genetic relatedness of individuals (71,

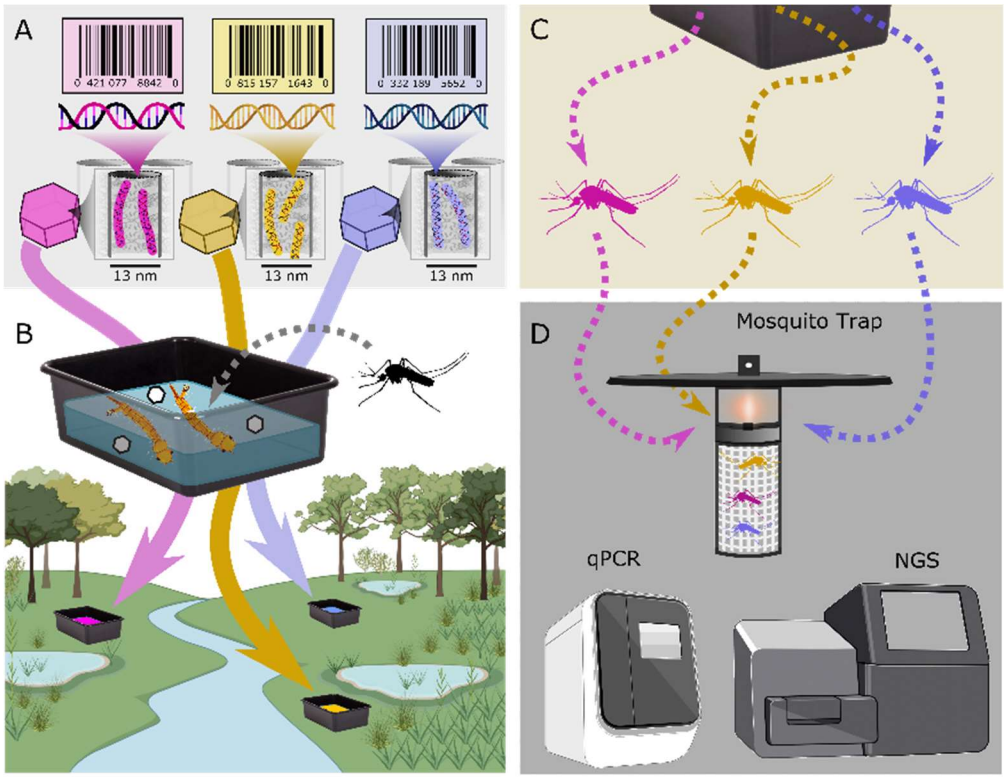



**Figure 3.1. Mark-Release-Recapture strategy. (A)** Synthetic DNA "barcode" sequences are designed, amplified, and loaded into the nanopores of engineered porous protein microcrystals. **(B)** Tubs are filled and placed at specific locations, dosed with specific DNA barcodes, and populated by mosquito larvae. **(C)** Transstadial persistence of the marker sequence in emerging adult mosquitoes allows for **(D)** detection of the recent origin of captured mosquitoes via qPCR or next-generation sequencing.

72). While this approach is innovative and provides valuable insight into myriad mosquito population parameters, real-time field dispersal estimates are indirect, and specific expertise in generating and analyzing these types of data is required.

In this study, DNA barcodes are short (~100 - 200bp) pieces of double-stranded DNA of known sequence representing the material's unique signature detected via sequencing (73). The feasibility of DNA serving as a tracking material has been tested in various applications (74). Surface adsorbed DNA, in the form of silica-encapsulated DNA, has been studied as a tracking material for oils (75), trophic pathways (76), reservoir imaging (73), and aquifer characterization (35). Surface adsorption is suggested to afford nucleic acid resistance to nucleases (77).

Advances in parallel synthesis and sequencing further promote DNA employment as MRR markers (78).

Recently, a hybrid fluorescent dye/DNA tag material was evaluated for mosquito marking (79). While the externally applied DNA tags of variable length remained robustly detectable up to 3 weeks, initial fluorescence-based recapture identification using ultra-violet (UV) light may degrade DNA tags. Here we develop a complementary self-marking strategy via ingestion.

We have identified a promising candidate material in DNA loaded microcrystals that may overcome limitations inherent to conventional mosquito marking materials. Our central hypothesis is that porous protein microcrystals can carry and protect DNA barcodes following ingestion, and therefore be used to study movement patterns of field-collected mosquitoes (Fig. 3.1). Microcrystals composed of an isoprenoid binding protein "CJ", from *Campylobacter jejuni*, feature an array of 13-nm diameter pores suitable for DNA uptake (Fig. 3.2). Barcode DNA was designed using sequences not found in reference DNA sequence databases (Fig. S3.1) (80, 81). We then optimized the loading and recovery of DNA barcodes from cross-linked protein microcrystals and demonstrated persistent marking and barcode recovery from adult mosquitoes following oral ingestion as larvae. Furthermore, we have demonstrated that host microcrystals can confer some protection upon guest DNA from conditions that degrade free DNA. By overcoming previous limitations on marker diversity and encoding time and location data into a single barcode, this strategy allows data collection on vector movement patterns with a level of resolution that has been previously impossible.

### 3.3 Results

#### 3.3.1 DNA Loading and Controlled Release from Host Microcrystals

Following crosslinking, protein microcrystals exhibited negligible background fluorescence when imaged under 488 nm light (Fig. 3.2A). Following incubation in solution containing fluorescently labeled 15bp double-stranded DNA (15mer), microcrystals appeared markedly

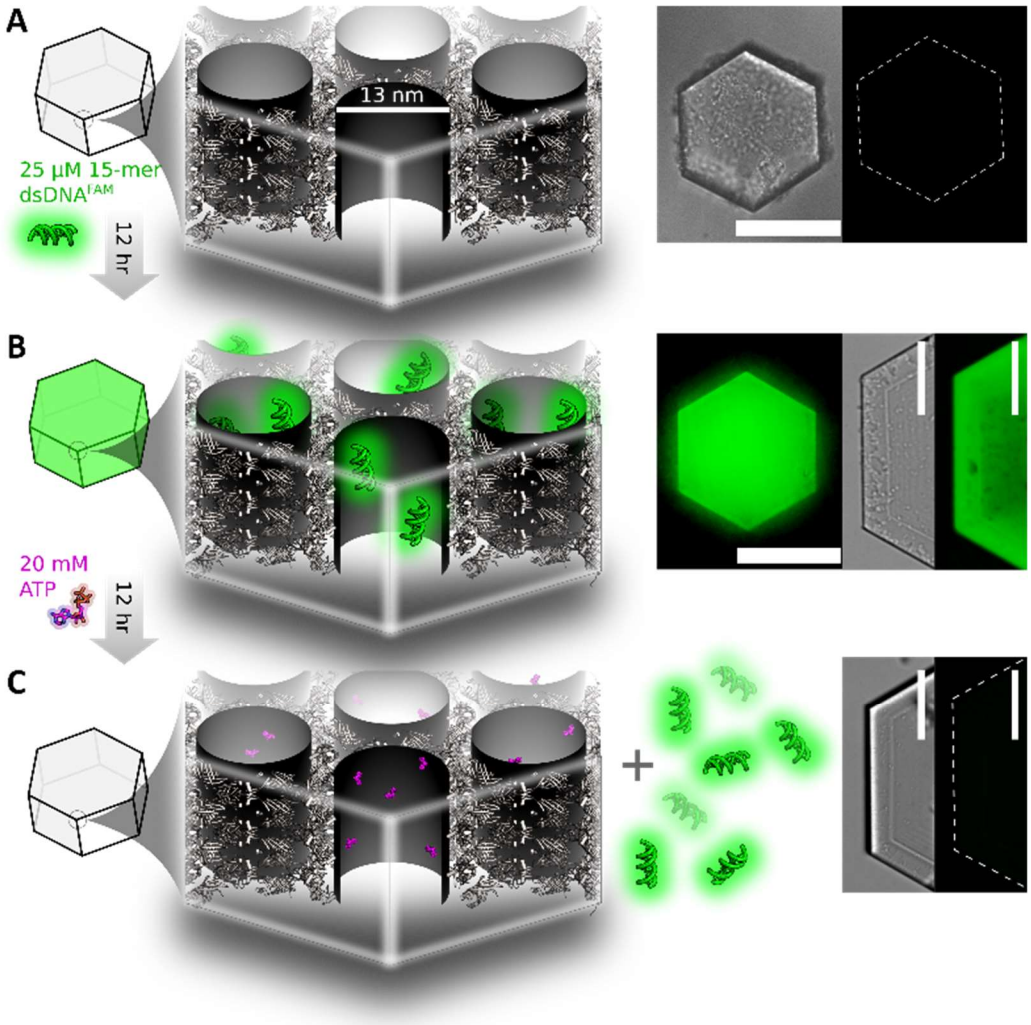



**Figure 3.2. DNA loading and ATP-induced release.**
**(A)** Schematic of a CJ crystal pore, left, and a protein crystal imaged under bright-field and 488 nm excitation (right) revealing low background fluorescence of unloaded microcrystals. **(B)** Microcrystals following 12 hr loading in solution containing fluorophore-labeled (FAM) 15mer dsDNA. **(C)** Imaged microcrystals following 12 hr incubation in 20 mM ATP solution. Previously fluorescent microcrystals exhibited almost no fluorescence due to nucleotide triphosphate-triggered release of microcrystal-adsorbed DNA. Scale bar denotes 100 microns.

fluorescent due to the strong retention of DNA adsorbed to the microcrystal interior despite 2-3 washes in buffer (Fig. 3.2B). The lack of observable DNA release during washing suggested that DNA was adsorbed with very high affinity. Consistent with strong noncovalent binding, it was possible to release guest DNA via incubation in solutions containing adenosine triphosphate (ATP) (Fig. 3.2C). Presumably, nucleotide triphosphate outcompeted microcrystal-adsorbed DNA for binding, resulting in the observed triggered release.

### 3.3.2 Transstadial Persistence of Microcrystals from Larval to Adult Stages

We first confirmed delivery of microcrystals directly to adult mosquitoes via a sugar meal (Fig. S3.6). We then confirmed that larvae ingested microcrystals spiked into their aquatic environment and that ingested microcrystals could be localized within the larval mosquito digestive tract. After being fed non-DNA-loaded protein microcrystals conjugated with the fluorophore Texas Red throughout the second and third instar stages, *Cx. tarsalis* larvae were

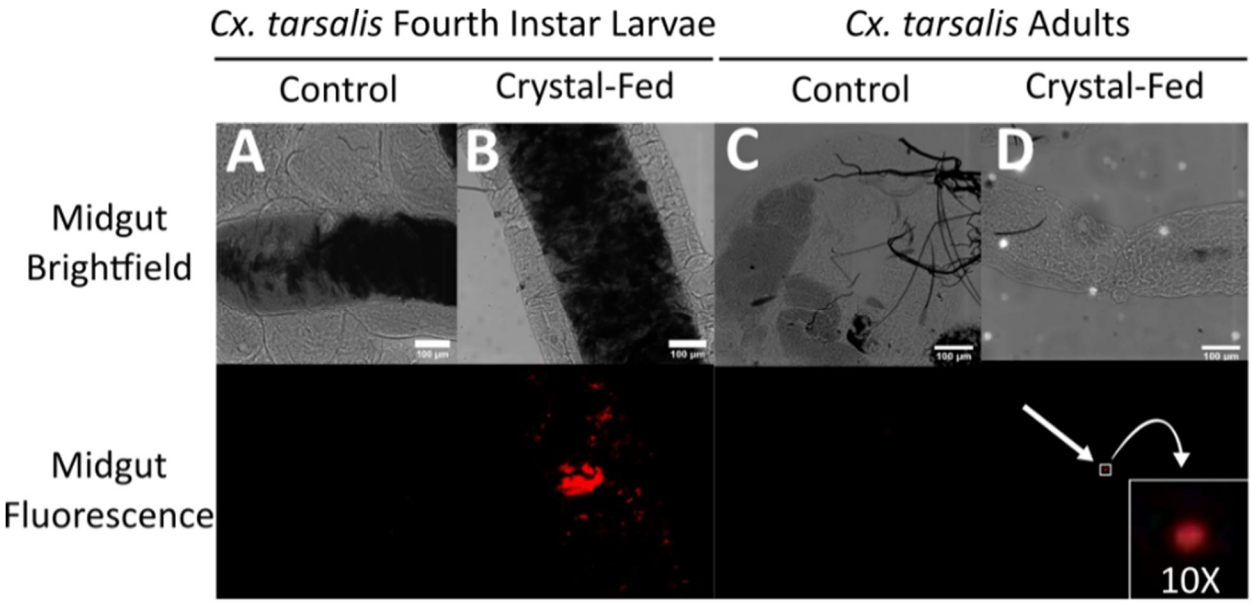



**Figure 3.3. Detection of Texas-Red labeled microcrystals in the midgut of adult and larval mosquitoes.** Composite panels show midgut images from both brightfield and fluorescence microscopy for *Cx. tarsalis* mosquitoes fed either liver powder alone (**A, C**), or liver powder supplemented with Texas-red-labeled protein microcrystals (**B, D**). Scale bar denotes 100 microns.

removed as fourth instar larvae for dissection. Confocal microscopy imaging of larval digestive tracts confirmed the presence of microcrystals (Fig. 3.3B) in 10/10 larval midguts examined whereas microcrystals were absent in 10/10 control larvae fed only liver powder (Fig. 3.3A). Fluorescent-tagged microcrystals were detected in all larvae fed a microcrystal-enhanced diet, although the density of microcrystals within digestive tracts varied to a limited extent by individual (Figs. S3.2, S3.3). Preliminary observations indicated that microcrystals remained within the digestive tract, but additional work to quantify the microcrystal half-life is underway in the Kading laboratory. Microcrystal localization within the digestive tract followed no observable pattern, and microcrystals were roughly evenly dispersed throughout the foregut, midgut, and hindgut (Fig. S3.2). No microcrystals were visualized in the Malpighian tubules supporting the lack of microcrystal excretion via these tubules.

We next sought evidence that microcrystals ingested by mosquito larvae could persist transstadially, or to the next stage along the mosquito life cycle. Remaining fourth instar larvae were allowed to pupate and eclose. Microcrystals were visualized within digestive tracts dissected from newly emerged adult female mosquitoes, demonstrating that microcrystals were retained in the alimentary tract of mosquitoes through development (Fig. 3.3C, D). Microcrystals ingested during larval development were not fully shed in the meconium; instead, a portion remained and could be visualized in the midgut post-eclosion. Microcrystals persisted transstadially in all ten mosquitoes analyzed with this method. Importantly, barcode was detected via qPCR in 82% (59 out of 72) adult mosquitoes reared on barcoded microcrystals as larvae, and microcrystal ingestion did not affect survivorship of adult mosquitoes (Figs. S3.4,

S3.5, Table S3.1). The locations and status of the crystals that are retained for long times are not known at this time.

### 3.3.3 Barcode Recovery and NGS Validation

Mosquitoes that were exposed to microcrystals loaded with a synthetic barcode DNA sequence were subjected to homogenization and DNA extraction. Two primers were selected to amplify an 84-nt segment of synthetic barcode in qPCR experiments with three sample types. The first sample type included three lab colony mosquitoes (*Cx. tarsalis* Kern National Wildlife Refuge (KNWR) strain) that were fed barcode-laden microcrystals as larvae and from which barcode was detected in the emerged adult mosquitoes, designated as Survivorship Replicates 1-3 (SR1-3). The second sample type included three pools of wild-caught *Cx. tarsalis* and *Culiseta inornata* mosquitoes that colonized microcrystal-spiked tubs in the field and were later captured as adults in a CDC light trap, designated as Field Replicates 1-3 (FR1-3). The third sample type included three wild-caught *Cx. pipiens* mosquitoes that colonized microcrystal-spiked tubs placed in the field as larvae and were reared to adults in the laboratory, designated as Larvae Replicates 1-3 (LR1-3). The positive control was naked barcode DNA. The negative control was PCR master mix with no template added. The qPCR melt curves (Fig. 3.4A) had peaks corresponding to the target 84mer barcode peak (~80.6°C) in addition to a slightly higher peak (~82-84°C) observed among mosquitoes from the survivorship experiment as well as field-collected adult and larval specimens (Fig. S3.10).

To verify that the qPCR signal from the primary peak and shoulder peak were both resulting from an authentic barcode, size inspection of the PCR amplicons from samples shown in Figure 3.4A was performed using gel electrophoresis (Fig. 3.4B). All the samples that were

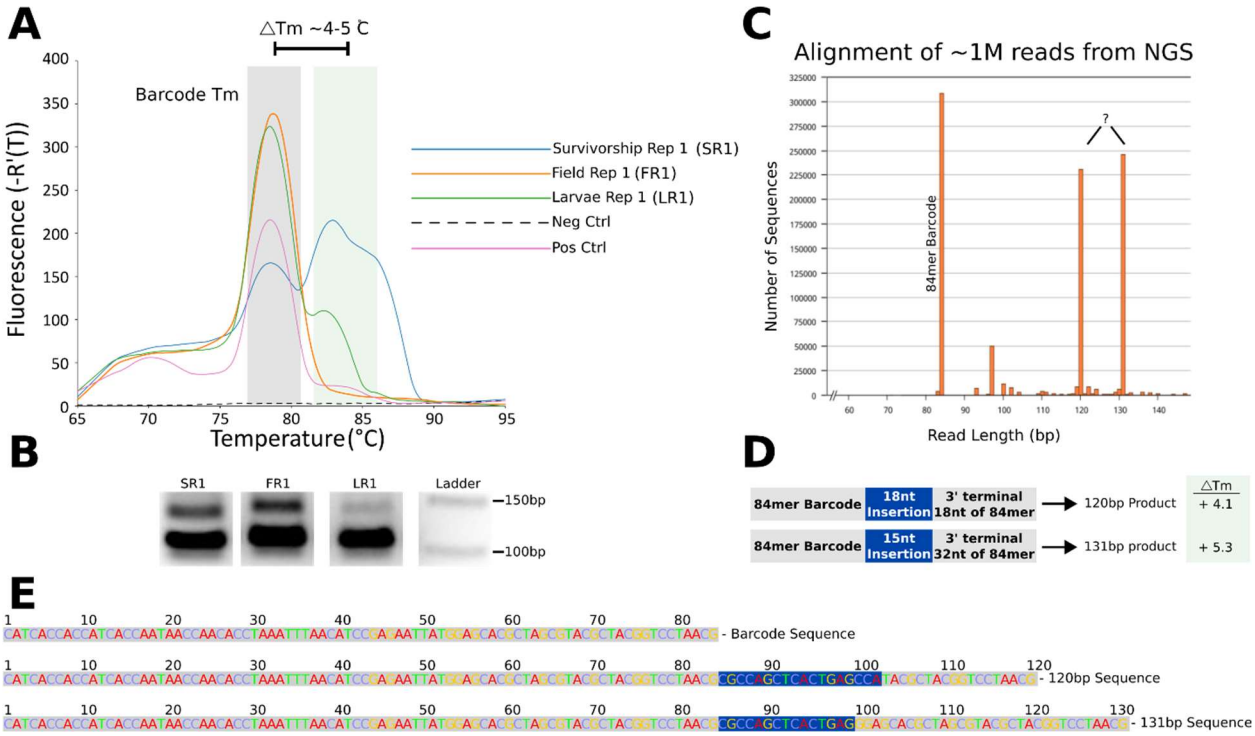



**Figure 3.4. Barcode Detection. (A)** qPCR melt curve displaying target 84-mer barcode peak in addition to a slightly higher peak (~82-84°C) from samples obtained from survivorship, field, and larvae studies. For clarity, a single replicate from survivorship (SR1), field (FR1), and larvae (LR1) are shown. All replicates are shown in fig. S11. **(B)** Gel electrophoresis following PCR amplification of samples shown in (A), have the expected target barcode band in addition to a weaker larger product band consistent with the qPCR melt curves. **(C)** Read length distribution for 1 million reads performed in Geneious Prime® following NGS of the samples shown in (B) reveal two prominent longer amplification products (120 and 131 bp), in addition to the target barcode sequence. **(D)** Inspection of the longer reads shown in (C) reveal a product that is still clearly barcode derived, albeit via an off-target concatenation event. The estimated $T_m$ of the longer products was consistent with the higher melt peaks observed in (A). **(E)** Sequence composition of the intended barcode read and the longer products.

exposed to the synthetic barcode sequence had PCR amplified output products with the expected size for the 84-bp product. There was also a distinct and recurring band for an amplicon larger than expected (Fig. 3.4B). This larger product represented the observed higher melting temp peaks (Fig. 3.4A).

We employed NGS to further verify that the samples recovered from mosquitoes contained authentic barcode sequences, and that the side product that contributed to the qPCR signal was also derived from our synthetic barcode. The larger band in all lanes from the

electrophoresis gel were extracted, prepared for sequencing by adding flanking Illumina adaptors using overhang PCR, and processed with NGS (NovaSeq 6000).

In an analysis of 1 million aligned reads (Fig. 3.4C), performed in Geneious Prime®, the most common read corresponded to the expected 84-bp sequence, *despite extracting the larger band size*. The other two dominant read sequences corresponded to 120 bp and 131 bp sequences (Fig. 3.4D-E) that were almost entirely composed of the original synthetic barcode, with insertions/duplications suggesting an earlier off-target amplification event (e.g. mispriming) during the synthetic barcode amplification. The estimated $T_m$ values (84.7°C and 85.9 °C via Primer3 2.3.7) for these two longer products were consistent with the higher melt temperature region shown in Figure 3.4A.

**3.3.4 Laboratory-Reared Mosquito Barcode Recovery and Microcrystal Protection**

The persistence and recovery of synthetic DNA markers days after ingestion and metamorphosis was a remarkable result. We hypothesized that the host microcrystal would confer protection on guest DNA, thereby allowing guest DNA barcodes to survive in the mosquito midgut environment. In contrast, we expected naked DNA to be degraded (i.e. by nucleases). We further hypothesized that mosquito homogenate would be a similarly harsh environment for unprotected DNA. Microcrystal protection of loaded DNA in harsh solution conditions was directly tested by incubation of DNA-loaded microcrystals with filtered mosquito homogenate, where the guest DNA consisted of a 200-bp barcode sequence. As expected, DNA that was incubated in mosquito homogenate was only recovered if first loaded into microcrystals, as is evident by the presence of the PCR product band in lane 6 and the absence in lane 4 (Fig. S3.9C). DNA barcode was not detected in the negative control samples,

nuclease-free water, and mosquito homogenate. The DNA barcode was recovered via PCR from the positive control samples, DNA in solution, and DNA-loaded microcrystals. Critically, these *in vitro* results were further supported *in vivo* through the elevated detection of barcode from adult mosquitoes fed DNA-loaded microcrystals relative to mosquitoes fed naked barcodes (Fig. S3.7). These results suggest that the primer sequences used do not cross-react with native mosquito homogenate (*Aedes aegypti*) and therefore do not lead to false-positive DNA detection at the estimated PCR product size of 200 bp. Furthermore, these results demonstrate plausible microcrystal protection of guest DNA from degradation under harsh solution conditions, allowing for downstream recovery and analysis via qPCR.

**3.3.5 Field Trial Collections and Barcode Detection**

We made daily observations of larval development in both the 100 uL/week tub and the 1000 uL/week tub. As mosquito larvae presented in containers, pupae were removed, and reared to adulthood in the laboratory for crystal detection. This included 15 male *Culex pipiens* sampled from the 1000 uL tub on August 6, and 9 *Culex pipiens* from the 100 uL tub on August 13 (8 females and 1 male). From the four light traps we established in each cardinal direction of the two bait stations (Fig. S3.10), we collected 34 *Culex pipiens*, 172 *Culex tarsalis*, and 1 *Culiseta inornata* for a total of 207 mosquitoes grouped into 55 pools. Results from this initial pilot study suggest barcode presence in approximately 76% (34 out of 45) of adult mosquito pools and 75% (18 out of 24) of larvae reared from the treated bins (Fig. S3.12). More extensive field evaluations are underway and results will be reported elsewhere.

### 3.4 Discussion

We have developed a new class of lightweight, durable marker particles composed solely of biomolecules (DNA and protein). These particles are suitable for persistently marking mosquitoes for ecological and biosurveillance applications. While these particles may also be suitable for topical application, the current study takes advantage of the size and intrinsic biocompatibility of the particles to assess the possibility of mosquito self-marking via particle ingestion. Microcrystal-adsorbed DNA was consistently trigger-released into the surrounding solution by ATP incubation (Fig. 3.2) and detected by common nucleic-acid amplification approaches (Fig. 3.4). Remarkably, DNA barcode-doped microcrystals ingested by mosquito larvae persisted through metamorphosis to the adult life stage (Fig. 3.3) without significantly affecting detection sensitivity (Fig. S3.8) or adult mosquito survivorship (Fig. S3.4). To our knowledge, this result lacks precedent. Moreover, barcode DNA adsorbed within protein microcrystals possessed an elevated resistance to degradation during *in vitro* mosquito homogenate incubation (Fig. S3.9C) and following ingestion by mosquito larvae (Fig S3.6). We therefore propose that encasing DNA barcodes in the synthesized protein microcrystals provided some protection against barcode degradation by digestive enzymes or the basic pH (~10-11) of the mosquito larval anterior midgut (82).

Our barcoding method is innovative in part due to the use of a designed synthetic DNA barcode strand that was easily recovered from individual mosquitoes or from pools. Specifically, the barcode sequence used in this study exploited nullomer sequences (81) not found in publicly available databases to label mosquitoes (Fig. S3.1). The second major innovation is the protection of potentially vulnerable DNA in the interior of crosslinked porous

protein microcrystals. To this end, we take advantage of our fortuitous empirical discovery that DNA strongly adsorbs to the interior of our porous protein microcrystals (83). The resulting doped microcrystals could be easily integrated into ongoing mosquito surveillance activities, for both research and operational purposes. Larval habitats, particularly artificial containers, can be uniquely marked by the addition of DNA-doped microcrystals (Fig. 3.1 and Fig. S3.10). Barcodes can then be amplified from mosquito pools already undergoing arbovirus testing, thereby linking information on spatial and temporal movement patterns to virus transmission cycles. Extrinsic barcode recovery will directly indicate visited locations for captured mosquitoes, including those that are disease vectors.

While qPCR provides a highly sensitive method to quantify DNA based on the generation of a fluorescence signal, it does have certain limitations when working with environmental field samples. Due to a lack of sequence specificity in SYBR Green-based assays, the presence of non-template DNA and/or template fragments may cause heterogenous melt curves, reducing confidence in positive detection events. While TaqMan qPCR offers sequence specificity, such assays require costly probes and are limited in channel-based multiplexing capability, allowing less than 10 samples to be run in parallel (84), excluding recent fluorescence modulation-based approaches which consume more probe per assay, thus increasing sample processing cost (85). Moreover, multiplexing qPCR requires optimization of primer sequences, requiring additional time in assay design. In contrast, high-throughput NGS offers direct read confirmation of target template with nucleotide resolution. In some application scenarios, the added time and expense of NGS will be justified by eliminating ambiguity in barcode detection from observed

qPCR melt curves, while allowing for multiplexing at a scale (hundreds) presently inaccessible by current probe-based qPCR methods.

An additional limitation of the present study is the barcode design. The barcode was obtained as a singular, full-length sequence (125bp), and PCR was used to synthesize barcode stock for subsequent experiments. Purchasing each barcode sequence is not a particularly scalable approach, considering the potential demand for hundreds of barcodes to represent different field sample sites and times. A more modular and scalable barcode DNA production method is therefore advisable for future studies that differentiate multiple sample sites, and therefore have the potential to provide insight on the degree of connectivity between isolated locations and how that connectivity evolves temporally by taking weekly samples. In particular, the use of barcodes at multiple field sites may allow investigators to understand the degree of habitat sharing and gene flow occurring between mosquitoes originating from different locations. Additional resolution to such data may be introduced by dosing field sites weekly with new, unique barcodes while simultaneously trapping for adult mosquitoes.

In conclusion, DNA-barcoded microcrystals represent an advanced functional biomaterial, and an innovative technology platform for studying mosquito dispersal and subsequently arbovirus circulation by vectors in the field. We have demonstrated proof-of-concept of this technology in the laboratory and in a small pilot field study. Laboratory investigation has shown that 1) *Culex* mosquitoes can be orally marked as larvae by spiking the larval habitat with DNA loaded microcrystals, thereby providing a self-marking strategy for mosquitoes in the field, 2) microcrystals persist transstadially through mosquito development, 3) ingestion of microcrystals by mosquito larvae does not affect adult mosquito survival or

development, and 4) encoded information can be recovered from mosquitoes by qPCR and NGS, with detection sensitivity unaffected by mosquito pool size. This latter result is directly translational to vector surveillance activities in which mosquito pools would also be tested for arbovirus nucleic acid.

One consideration left to be addressed prior to extensive field deployments is the larval feeding behavior of the target mosquito species (86). This approach may be more conducive to mosquito species that dive and filter particles suspended in the water column (i.e. Culex, spp. The focus of these initial studies), as opposed to those species that feed at the surface or are predaceous. Additionally, unit cell calculations from Kowlaski et al. show that microcrystals are approximately 80% solvent due to high porosity, allowing them to readily equilibrate with the surrounding solution and sink to the bottom, resulting in a dynamic availability to mosquitoes (24). While this approach may not function as a direct substitute for research questions requiring knowledge of the marked population size, new doors are opened for gathering more applied information on mosquito populations that may be operationally significant. Collectively, these findings represent the critical first steps to implementation of this technique into mainstream vector research and surveillance activities. Lastly, it is important to consider the possibility of marked mosquitoes transferring DNA barcode to larval and adult mosquitoes that have not previously visited a bait station. Such lateral transfer, as previously described (87), would frustrate data analysis leading to inaccurate conclusions in studies examining population size or dispersal, for example. Future work aims to address this possibility.

## 3.5 Materials and Methods

### 3.5.1 Porous Microcrystal Production and Fluorophore Labeling

Microcrystal-forming protein derived from *Campylobacter jejuni* was expressed, purified and crystallized as described previously (PDB entry 5w17) (24, 26). Microcrystals were crosslinked using 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) or glyoxal during trace-labeling with Texas Red dye (ThermoFisher) as described previously (21). See Supporting Information (Porous Crystal Production, Crystal Fluorophore Labeling) for additional details regarding microcrystal growth, crosslinking and labeling.

### 3.5.2 DNA loading and ATP-induced DNA release

Cross-linked microcrystals were placed in 100 µL TE buffer (10 mM Tris, 1 mM EDTA, pH 7.4) for 1 hour to equilibrate microcrystal pores for DNA loading. Microcrystals were imaged under 488 nm excitation with a fluorescence confocal microscope (Nikon Eclipse Ti-E) equipped with an Andor iXon DU-897 EMCCD camera to establish baseline fluorescence of non-DNA loaded microcrystals. Microcrystals were then immersed in 10 µL of 25 µM 15 base pair double-stranded DNA,15mer (5' - CCGCACGCACGAGGC - 3' ) labeled with 6-FAM (fluorescein) at the 5' terminus (IDT), and sealed in a glass well plate for approximately 12 hours. Following DNA loading, microcrystals were washed with TE buffer to remove unbound 15mer. Microcrystal retention of DNA was confirmed by fluorescence imaging. To trigger release of microcrystal bound 15mer, microcrystals were immersed in 20 mM adenosine triphosphate (ATP) in TE buffer, and sealed in a glass well plate for approximately 12 hours. Microcrystals were then imaged by fluorescence confocal microscopy, confirming ATP-induced DNA release by a reduction in microcrystal fluorescence following ATP incubation.

### 3.5.3 Transstadial persistence of microcrystals from larval to adult stages

Colony-reared, KNWR strain *Cx. tarsalis* were hatched in 9oz clear, plastic cups under standard insectary conditions.(88) Briefly, eggs were hatched in water to produce roughly 100 first instar larvae in each cup. Control larvae were fed 250 µL of 10% liver powder solution each day, while microcrystal-exposed larvae were fed 225 µL liver powder solution thoroughly mixed with 25 µL non-DNA-loaded microcrystals conjugated with the fluorophore Texas Red and suspended in water. At the fourth instar stage, control and microcrystal-fed larvae (n=10 for each) were removed for dissection. A portion of both larvae groups were allowed to pupate and emerge as adults. Immediately after emergence, female *Cx. tarsalis* (n=10) were removed for cold-induced knock down and dissection. Whole digestive tracts were dissected from larvae and adults in PBS and then mounted onto slides with ProLong™ Gold Antifade Mountant with DAPI. Slides were stored in the dark at -20°C for at least 24 hours prior to confocal microscopy. Larvae and adult dissections were imaged with a Nikon Eclipse Ti-E microscope equipped with an Andor iXon DU-897 EMCCD camera at 10X magnification under brightfield and 561nm excitation (10% laser power).

### 3.5.4 *in vivo* qPCR Barcode Recovery

Following extraction, qPCR was performed using 6 µL of each sample as the template for 20 µL reactions with the following primers: fwd 5’ – CATCACCACCATCACCAA – 3’, rev 5’ – CGTTAGGACCGTAGCGTA – 3’. Primers used were designed to amplify an 84bp sub-sequence of the 125bp synthetic barcode (84mer) initially loaded into microcrystals. Primers were designed using Primer3 (89, 90) to enhance template amplification while reducing marked primer-dimer formation observed with an initial primer set (Primer1_114F, Primer1_114R) adopted from

Goswami et al (81). Standards used in qPCR consisted of serial dilutions of the PCR amplified 84mer. The quantitative amplification was performed using the manufacturer's guidelines (Agilent qPCR Brilliant II SYBR Master Mix). Reaction conditions were: 1 cycle of 95 °C for 180 seconds and 45 cycles of 95 °C for 5 seconds and 60 °C for 10 seconds. Melt curve was obtained by 1 cycle of 95 °C for 30 seconds, 65 °C for 30 seconds and 95 °C for 30 seconds. To obtain a quantitative cutoff for the qPCR data we fit each curve as a sum of Gaussian functions using LMFIT (91) and a custom Python script to constrain the center position and width of the component Gaussian functions. On the whole, the resulting functions fit the data remarkably well (Figs. S3.5, S3.12). To provide a conservative cutoff we only classified datasets as "positive" for barcode if the peak height of the ~78 °C Gaussian exceeded the height of the neighboring Gaussian (~74 °C) by a factor of 1.5 and possessed a raw height value greater than 10X (background), a value corresponding to the negative first derivative of the fluorescence (arbitrary units, a.u.) detected during amplification. Raw data and Python code are available on Zenodo (DOI: 10.5281/zenodo.6834837).

**3.5.5 Next-Generation Sequencing (NGS)**

Barcode positive samples from survivorship and field studies detected via qPCR were prepared for sequencing by using 1 µL of each field sample as the template for overhang PCR to append Illumina sequencing primer binding sequences using the following primers: fwd 5' – ACACTCTTTCCCTACACGACGCTC TTCCGATCTCATCACCACCATCACCAA – 3', rev 5' – GTGACTGGAGTTCAGACGTGTGCTC TTCCGATCTCGTTAGGACCGTAGCGTA – 3'. Thermocycling conditions for overhang PCR were: 1 cycle at 98°C for 2:00, 2 cycles at 98°C for 20 sec, 58°C for 30 sec, 72°C for 30 sec and 1 cycle at 72°C for 1:00. An additional overhang PCR was performed

to append Illumina flow cell hybridization sequences using the following primers: fwd 5' – AATGATACGGCGACCACC GAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT – 3', rev 5' – CAAGCAGAAGAC GGCATACGAGATNNNNNNNNNNATATTCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT – 3'. Thermocycling conditions for additional overhang PCR were: 1 cycle at 98°C for 2:00, 2 cycles at 98°C for 15 sec, 63°C for 30 sec, 72°C for 30 sec and 1 cycle at 72°C for 1:00. Lastly, PCR was performed on the full-length template using the following primers: fwd 5' – AATGATA CGGCGACCACCGAGATCT – 3', rev 5' – CAAGCAGAAGACGGCATACGAGAT – 3'. Thermocycling conditions PCR were: 1 cycle at 98°C for 2:00, 30 cycles at 98°C for 15 sec, 58°C for 30 sec, 72°C for 30 sec and 1 cycle at 72°C for 1:00. Following each amplification, PCR cleanup was performed using KAPA Pure Beads (Roche). Size selection for the 218bp barcode library was performed using Monarch DNA Gel Extraction Kit (New England Biolabs). The library was quantified using Qubit 1X dsDNA HS Assay Kit (ThermoFisher) and diluted to 20 nM for sequencing sample prep. Paired end 2x150 cycle sequencing was run on an Illumina NovaSeq 6000 (Genomics and Microarray Core, University of Colorado Anschutz Medical Campus). The ea-utils package was used for initial sample processing including adapter trimming and read joining. FastQC was used to check overall quality of joined reads and to determine total read count of detected barcode (Babraham Bioinformatics). Alignment was performed using Geneious Prime® 2021.0.1.

### 3.5.6 Mosquito processing for barcode detection

Mosquito pools were homogenized by addition of 1 mL mosquito diluent (Dulbecco's Modified Eagle Medium (DMEM) with 20% fetal bovine serum (FBS), 50 µg/mL penicillin/streptomycin,

50 µg/mL gentamicin, 2.5 µL/mL fungizone), two glass Coliroller beads (Novagen), and 75 µL of 100 mM ATP, and homogenized for 3 minutes at 24 Hz using a Retsch Mixer Mill (Retsch). Nucleic acid extractions were performed with MaxMAX Cell-Free DNA extraction kit, using a modified protocol (see Supporting Information – Barcode Recovery from Homogenized Mosquitoes) in a 96-well plate format on a Kingfisher Flex automated extraction platform (Thermo Fisher). qPCR was performed using PowerUp SYBR Green Master Mix (Thermo Fisher), with the primers, cycling conditions, and standard quantification described above for *in vivo* barcode recovery.

**3.5.7 Pilot field trial**

Two black 18 Qt Sterlite wash basins were set outdoors on the CSU Foothills Campus and filled with stagnant organic-rich water taken from the field site to mimic a natural mosquito larval habitat. One tub received 1000 µL of microcrystal stock (weekly) and the other received 100 µL of microcrystal stock weekly for four weeks during July-August 2020. Containers were topped up daily to offset evaporation. When mosquito larvae naturally colonizing these tubs began to pupate, four Centers for Disease Control light traps (John W. Hock) were operated nightly to capture emergent mosquitoes. Additionally, representative fourth instar mosquito larvae or pupae were removed from the tubs and reared to adults in the insectary for the purposes of confirming the presence of barcode DNA. Pooled and individual mosquitoes were processed for barcode detection as described above.

**3.5.8 Barcode Protection by Microcrystals**

To demonstrate the extent to which microcrystals protect PCR-recoverable DNA in a complex matrix, a 50 µL solution containing crosslinked microcrystals in TE buffer was incubated with 50

µL of 10 ng/ µL ~200 bp double-stranded DNA solution (200mer) for 12 hours. Following DNA loading, microcrystals were washed three times in fresh TE buffer using Amicon Ultra-15 centrifugal filter units (Millipore Sigma) at 4700 RPM for 5 minutes. DNA-loaded microcrystals were incubated with an equal volume of filtered mosquito homogenate for 12 hours, followed by addition of 20 mM ATP for another 12-hour incubation. The 200mer/microcrystal/homogenate solution was then used as the template DNA for PCR (forward primer: 5'-AATGATACGGCGACCACCGAGATCT-3', reverse primer: 5'-CAAGCAGAAG ACGGCATACGAGAT-3') using Q5 High-Fidelity DNA Polymerase (NEB) with the following cycling conditions: 1 cycle at 98 °C for 45 sec, 30 cycles at 98 °C for 15 sec, 55 °C for 30 sec and 72 °C for 30 sec, and 1 cycle at 72 °C for 60 sec. The negative control samples consisted of nuclease-free water and, separately, mosquito homogenate. Positive control samples included 200mer in solution and 200mer retrieved from loaded microcrystals, both in the absence of mosquito homogenate. DNA recovery was assessed by 5% agarose gel electrophoresis.

# CHAPTER 4. SCALABLE COMBINATORIAL SYNTHESIS OF SYNTHETIC DNA BARCODE SEQUENCES[3]

## 4.1 Overview

Synthetic DNA barcodes are double-stranded DNA molecules designed to carry recoverable information, information that can be used to represent and track objects and organisms. A material marked with invisible and numerous molecular barcodes could carry a practically indelible identification code that is also amenable to encryption. DNA barcodes offer robust, sensitive detection using standard amplification and sequencing techniques. While numerous research groups have promoted DNA as an information storage medium, less attention has been devoted to the design of economical, scalable DNA barcode libraries. In dynamic tracking schemes, the cost of brute force synthesis of DNA oligos scales poorly since it is necessary to apply additional, unique barcode tags for every object/organism entering circulation. Here, we present an alternative modular approach to sequence design. Barcode sequences were constructed from smaller, interchangeable blocks, allowing for combinatorial assembly of numerous distinct tags. With two generations of modular barcode libraries, we demonstrated design and construction of 256 and, separately, 512 barcode sequences from less than 50 total single-stranded oligonucleotides. Contamination during experimental validation was controlled for by employing a liquid-handling robot for oligonucleotide mixing. Generating barcode sequences in-house reduces dependency upon external entities for unique tag generation, increasing flexibility in barcode generation and deployment. Next generation sequencing (NGS)

[3] The work presented in this chapter is described in the following manuscript: J.D. Stuart, N.R. Wickenkamp, K.A. Davis, C. Meyer, R.C. Kading, C.D. Snow, Scalable Combinatorial Synthesis of Synthetic DNA Barcode Sequences. *Int. J. Mol. Sci.* (In Preparation). **Author contributions**: C.D.S., J.D.S, R.C.K. designed research; J.D.S, C.M., N.R.W., K.A.D. performed research; J.D.S., C.M. analyzed data; J.D.S., C.D.S. wrote the paper; J.D.S., C.D.S., R.C.K. edited the manuscript.

detection of 256 different samples in parallel highlights the multiplexing afforded by the modular barcode design coupled with high-throughput sequencing. Deletion variant analysis of the first-generation library informed sequence design for enhancing barcode assembly specificity in the second-generation library.

**4.2 Introduction**

Inventory and internal stock management are common industrial practices for tracking the flow of raw materials and products from the point of manufacture to the point of sale (92). Material and product tracking is executed, in part, through application of a standard barcode. While barcodes such as one-dimensional Universal Product Code (UPC) barcodes remain an industry standard, these markers possess certain limitations capable of influencing a company's security (93). Specifically, UPC barcodes and two-dimensional Quick Response (QR) codes are ***visible*** markers rendering them susceptible to removal, subversive duplication, or counterfeiting. Also, the products must be large enough to display the barcodes thus limiting the types of products that can be tracked.

In contrast, molecular barcodes possess inherent characteristics surpassing limitations of existing UPC barcodes. Molecular barcodes can be classified into two high-level categories: 1) non-sequence encoding and 2) sequence encoding (94). Examples of non-sequence encoding materials, as given by Paunescu et al., include fluorescent dyes and quantum dots that provide a unique optical signature serving as the barcode (94). However, these materials are limited in the number of barcodes that can be generated due to spectral overlap. Sequence encoding materials include synthetic polymers (95, 96), peptides (97, 98) and DNA (35, 73). These materials boast astronomically high numbers of possible barcodes, making them attractive as

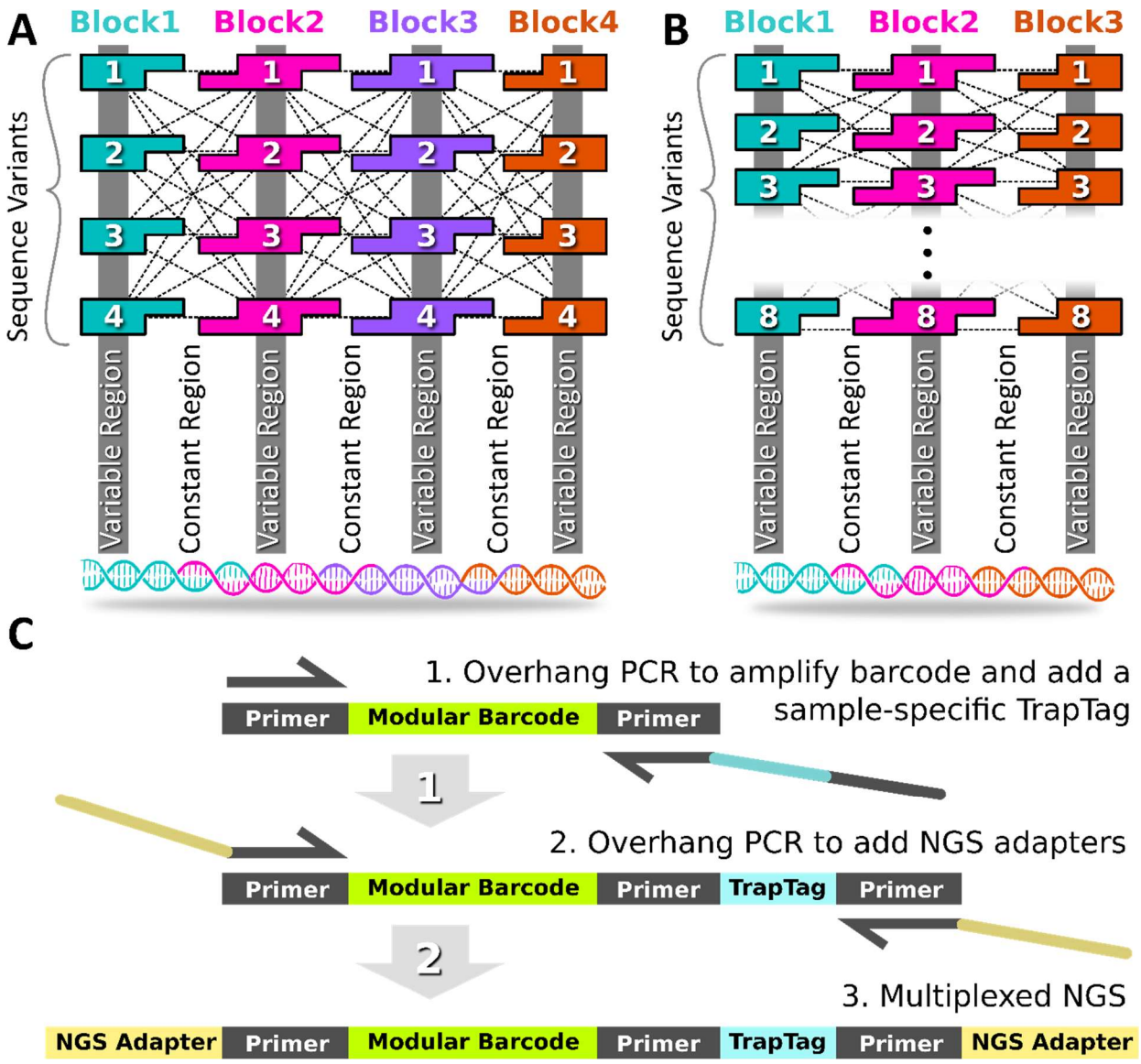



**Figure 4.1. Modular Barcode Design. A**) The first-generation modular barcode library (gen_1) contained 4 blocks with single-stranded overhangs for mixing and annealing. For each block, 4 variants were obtained allowing generation of $4^4$ (256) barcode sequences. **B**) The second-generation modular barcode library (gen_2) contained 3 blocks. The 8 variants per block allowed for generation of $8^3$ (512) sequences. **C)** A UMI Trap Tag sequence was appended using overhang PCR to facilitate multiplex NGS sequencing of multiple samples in parallel.

potential tracking tools. Of these, DNA stands out as the most accessible system due to the ongoing revolution in economical synthesis and sequencing (78).

For our purposes, synthetic DNA barcodes are short (~100 - 200bp) pieces of double-stranded DNA of known sequence representing the material's unique signature detected via

sequencing (73). This size range is convenient for signal amplification via PCR as well as reading the stored information via traditional or next-generation DNA sequencing (NGS). The feasibility of DNA serving as a tracking material has been tested in various applications (74). Nanoparticle surface adsorbed DNA, in the form of silica-encapsulated DNA, has been studied as a tracking material for oils (75), trophic pathways (76), reservoir imaging (73) and aquifer characterization (35). Surface adsorption is suggested to afford nucleic acid resistance to nucleases, an advantageous characteristic provided by the surface (77).

While sources in the literature have measured the stability of DNA-based barcodes (73), few sources have described practical scalable methods for DNA barcode sequence design and synthesis in the context of tracking applications. To be clear, many researchers have described novel encoding schemes for storing various information types (e.g., file content, books, digital media) in DNA (99-101). Such encoding, while advantageous for information storage, falls beyond the purview of DNA barcodes as unique tracking tags, where scalable sequence generation and ease-of-detection and differentiation remain the primary focus.

This work describes a combinatorial DNA barcode sequence design strategy that allows the user to construct hundreds of unique sequence tags by combinatorial annealing of less than 50 single-stranded oligonucleotides. To maximize self-assembly fidelity and yield, the sequences were designed computationally using custom Python code and the nucleic acid modeling and design software NUPACK (102). Sequences were designed to ensure proper annealing of single-stranded oligonucleotides and followed criteria common to PCR primer design (e.g., GC content, melting temperature). Experimental validation of top scoring candidate libraries was performed using NGS. The results herein demonstrate that modular

DNA barcodes offer increased autonomy for the user, reducing oligonucleotide cost by taking advantage of combinatorial assembly. Combined with sensitive, amplification-mediated detection, modular barcodes offer an appealing alternative to conventional DNA tagging methodologies that are cost-inefficient and rely heavily upon external nucleic acid synthesis.

## 4.3 Results

### 4.3.1 Modular Barcode Layout

**Table 4.1.** Modular barcode library parameters.

| | Number of Blocks | Variants per Block | Number of variable nucleotides per block | Minimum Hamming distance between variants | Total number of oligos required | Total number of barcodes allowed |
|---|---|---|---|---|---|---|
| Gen_1 | 4 | 4 | 6 | 3 | 32 | 256 |
| Gen_2 | 3 | 8 | 12 | 7 | 48 | 512 |

The modular barcode was composed of multiple double-stranded DNA blocks in series (Fig. 4.1A-B, S4.1). The junction between neighboring blocks contained 10-base pair assembly-encoding single-stranded overhangs, such that all single-stranded oligos annealed in the target order, requiring the overhang sequence domains remain constant across all barcode variants (i.e. "constant regions"). The termini of each assembled barcode remained constant also, allowing PCR amplification of any barcode variant in the library using the same primer set. The unique signature for each barcode resulted from regions internal to each block that are unique to each block variant (i.e. "variable regions"). The shared overhang domains allowed combinatorial mixing of block variants to generate different barcode sequences.

As shown in Table 4.1, the first-generation modular barcode library (gen_1) was comprised of 4 blocks with each block harboring an internal 6-nucleotide (nt) variable sub-domain. Duplicate variable region sequences were permitted for the Gen_1 library. A total of 4 block variants were obtained allowing for construction of 256 ($4^4$) unique sequences. The

second-generation library (gen_2) was comprised of 3 blocks with each block containing an internal 12 nt variable sub-domain. Duplicate variable region sequences were not permitted for the Gen_2 library. A total of 8 block variants were obtained allowing for construction of 512 ($8^3$) unique sequences. The reduced number of blocks in the gen_2 library was motivated by common deletion variants observed in the gen_1 library described below. The expanded variable sub-domain in the gen_2 library (12 nt instead of 7 nt) allowed increased diversity (Hamming distance), thereby reducing likelihood of barcode misidentification in the face of polymerase errors.

**4.3.2 TrapTag**

For multiplexing in NGS, it is important to use unique molecular identifier (UMI) tags.(103, 104) In the context of modular barcode detection, we refer to our UMI as a TrapTag, an 8-nucleotide sequence capable of representing the site and/or time of barcode DNA recovery (Fig. 4.1C). Appending a TrapTag to a barcode amplicon allows one to track barcode DNA with the added dimension of location or time thus increasing the resolution of tracking studies, while simultaneously increasing the information density of sequencing data. Critically, appending TrapTags increased the multiplexing capability of modular barcode libraries by allowing simultaneous reading of multiple distinct collection samples via NGS. Without requiring additional equipment, TrapTags were appended using overhang PCR, a routine technique in modular barcode library preparation, thus preserving overall cost efficiency. The 100 TrapTag sequences purchased for this study were adopted from known Illumina sequencing adapters.(Table S4.1)

### 4.3.3 Sequence Design

Primer3 was used for designing barcode primer pairs (Table 4.2) for both gen_1 and gen_2 libraries (See SI, Primer design in Extended Methods). In contrast to gen_1, the gen_2 TrapTag primer sequence was designed by NUPACK which considered off-target complex formations during sequence design for reducing potential off-target primer binding internal to barcode sequences. Laboratory evaluations of gen_2 TrapTag primer specificity for barcode amplification in the context of insect (mosquito) or contaminating human DNA demonstrated clean amplification of the barcode target of the correct amplicon size (Fig. S4.2). Amplification of barcode DNA was also highly sensitive, with successful amplification down to $10^{-10}$ of the initial barcode concentration (10 ng/µL) of target sequence (Fig. S4.2).

**Table 4.2.** Designed Primer Sequences.

| | Gen_1 Library | | Gen_2 Library | |
|---|---|---|---|---|
| | **Primer Sequence (5' - 3')** | **Design Source** | **Primer Sequence (5' - 3')** | **Design Source** |
| Forward | CCAGTCCTCAACAAGCTG | Primer3 | AGTGCGTGCAGTGAAAGC | Primer3 |
| Reverse | GTTGAAGCCGGTTACCAC | Primer3 | ATGGCGTTGCAAAGTCGG | Primer3 |
| Trap Tag | TTCTGGGTTCCTCATCGC | Primer3 | CGCCTTGATTTCAACTCGGCTCTCCGCTGAACA | NUPACK |

Following overhang sequence design using NUPACK, the sequences for both libraries (256 seqs for gen_1, 512 seqs for gen_2) were constructed *in silico* and analyzed using NUPACK's Tube Analysis function (Fig. 4.2). The NUPACK script is available on the Zenodo repository. The analysis predicted the target complex equilibrium concentration and free energy, while considering off-target complexes (any competing arrangement of up to 4 strands). For both libraries, the target complex was predicted to dominate at equilibrium and the free energy oscillated, minimally, around -300 kcal/mol (gen_1) or -350 kcal/mol (gen_2) suggesting that all combinatorial library members were likely to anneal as intended following

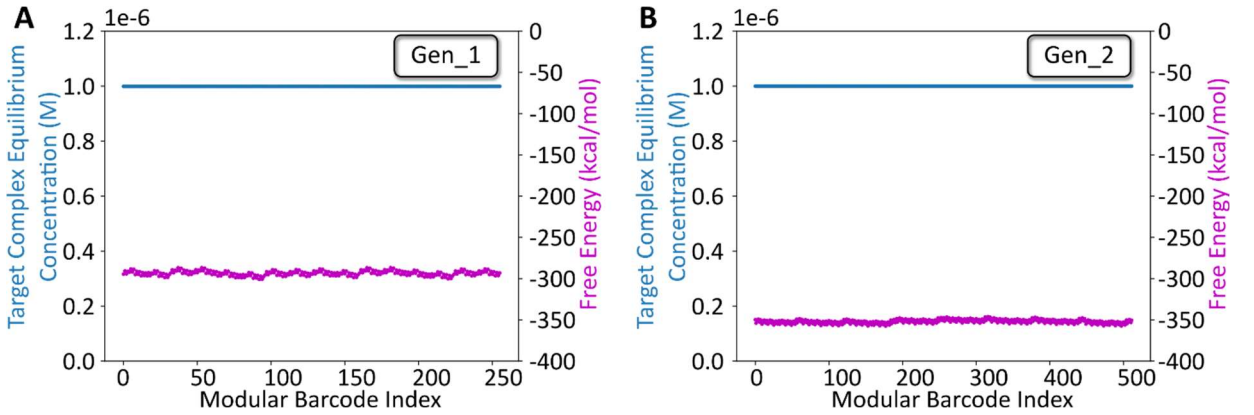



**Figure 4.2. NUPACK Design Analysis**. **A)** Tube analysis results for the gen_1 library. The x-axis is the modular barcode index, a value assigned to each unique sequence within the library. The left y-axis corresponds to the target complex equilibrium concentration predicted by NUPACK after defining a tube containing the 8 oligos at a starting concentration of 1 µM, at room temperature for each barcode. The right y-axis corresponds to the free energy for each barcode complex predicted by NUPACK. **B)** Tube analysis results for the gen_2 library. Notably, the free energy across the entire library is ~50 kcal/mol lower than the gen_1 library (region within dashed lines), suggestive of more favorable binding in the revised barcode design.

mixing. The revised design criteria for the gen_2 library (e.g., reduction in block number from 4 to 3, increased minimum Hamming distance between all variable region sequences from 3nt to 7nt) contributed to the predicted free energy decrease shown in Fig. 4.2B. The same approach was used for designing and analyzing the remaining TrapTag primer sequence for the gen_2 library. Experimental validation was pursued following NUPACK analysis results.

### 4.3.4 NGS barcode recovery

The NGS results for the pooled modular barcode library (gen_1) confirmed detection of all 256 sequences (within several orders of magnitude) within the entire dataset, in a single sequencing run (Fig. 4.3). The wide read distribution likely resulted from inaccurate quantitation of

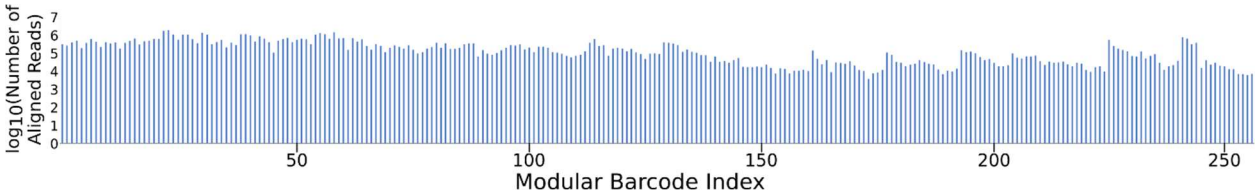



**Figure 4.3. Gen_1 NGS recovery of all 256 modular barcodes**. ~80M joined reads were aligned and assigned using a custom Python script. All barcodes were detected at read quantities ranging from $10^3$ to $10^6$, validating the multiplexing capability provided by the modular barcode design.

individual barcodes prior to attempted equimolar mixing. NGS results for the gen_2 library confirm recovery of all 96 pooled barcodes among ~110M reads (Fig. S4.4). These results highlight the multiplexing capability of the modular barcode design. Given the relatively low complexity of the barcode library, coupled with intent to make data analysis manageable in terms of time and computational cost, the sequencing dataset was downsampled such that only 1M reads were further analyzed out of the 76M paired-end reads received for determining the number and type of deletion variants and 1-nt substitution variants. While downsampling, or subsampling as previously described (105), may not provide complete depiction of the results, any trends observed in the truncated dataset are likely to persist throughout the entire dataset considering the high redundancy of the barcode NGS samples. It is likely that additional 'block' variants would allow parallel modular barcode detection to a greater degree than demonstrated here with 256 barcodes constructed with only 4 variants for each of the 4 blocks. As shown in Table 4.3, out of the 1M read subset for the gen_1 library, 80% were perfectly aligned barcodes, 3% were deletion variants, and 5% were 1-nt substitution variants. For the 1M read subset for the gen_2 library, 26% were perfectly aligned barcodes, 1% were deletion variants, and 3% were 1-nt substitution variants.

**Table 4.3.** Barcode Recovery by library for 1M read subsets.

| | Perfect Barcode Recovered (%) | Deletion Variants (%) | 1-nt Substitution Variants (%) |
|---|---|---|---|
| Gen_1 | 80 | 3 | 5 |
| Gen_2 | 26 | 1 | 3 |

### 4.3.5 Deletion analysis

While the vast majority of the barcode amplicon reads consisted of full length reads (80% of the 1M read data sub-set had length of 128 bp), the ample amount of data returned in the raw NGS

sequencing results was sufficient to reveal relatively rare off-target barcodes. We therefore proceeded to assess the oligonucleotide assembly fidelity during annealing, the first step of modular barcode construction and found a number of amplicons indicative of off-target assembly events. Figure 4.4 displays a heatmap of the 10 most common deletion variants detected out of 1M sequencing reads (gen_1). The number of missing nucleotides within a given barcode segment ranged from 1 to 14, with the majority of variants containing deletions

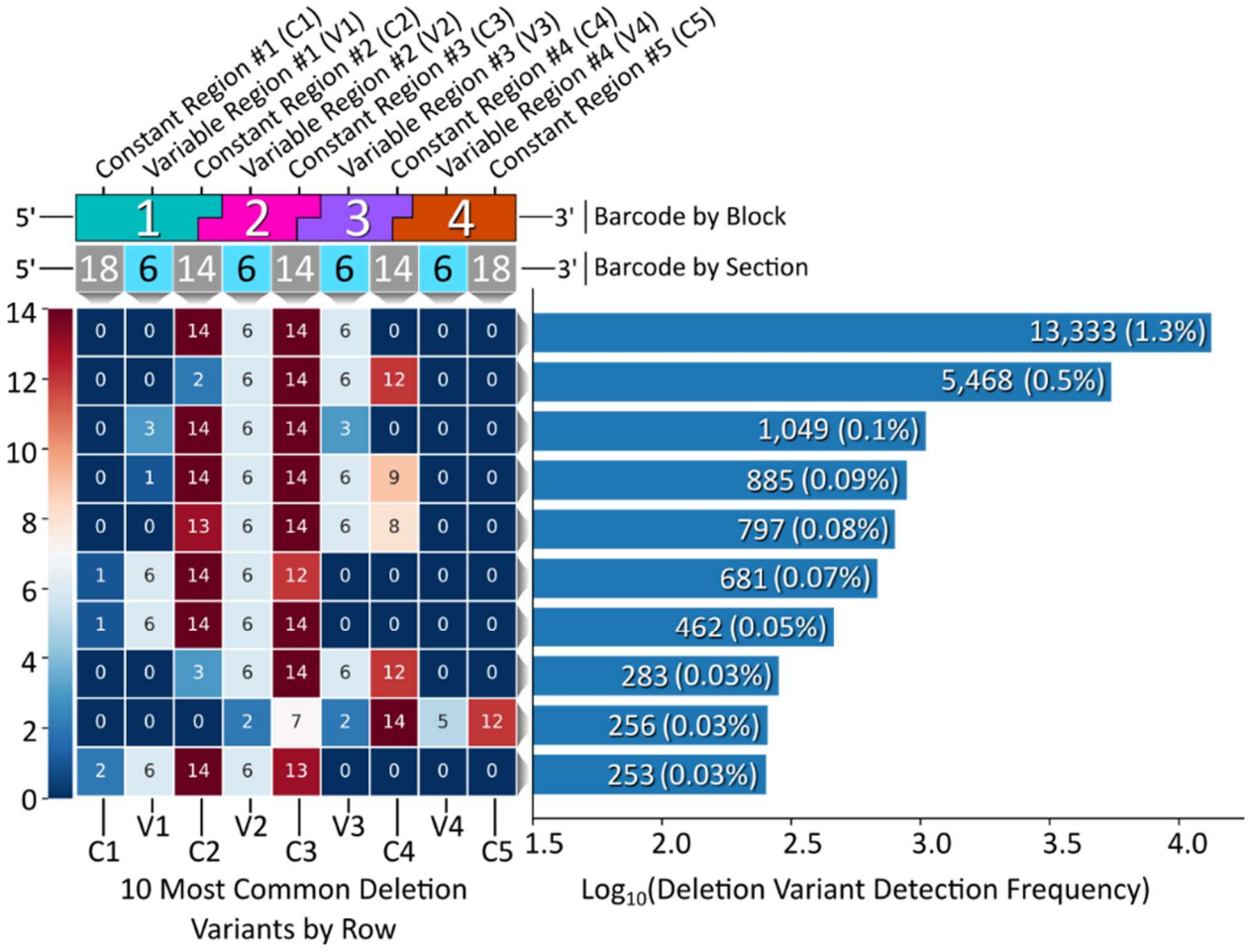



**Figure 4.4. Gen_1 Deletion Variant Analysis.** Left, a heatmap displaying the top 10 deletion variants. Each block corresponds to a certain section within each barcode, top, with constant regions referring to overhang sequence domains between blocks, and variable regions corresponding to the unique sequences internal to each block in the modular design. The number within each block represents the number of deleted nucleotides for that section. Right, a histogram displaying the detection frequency for the top 10 variants out of the 1M read subset.

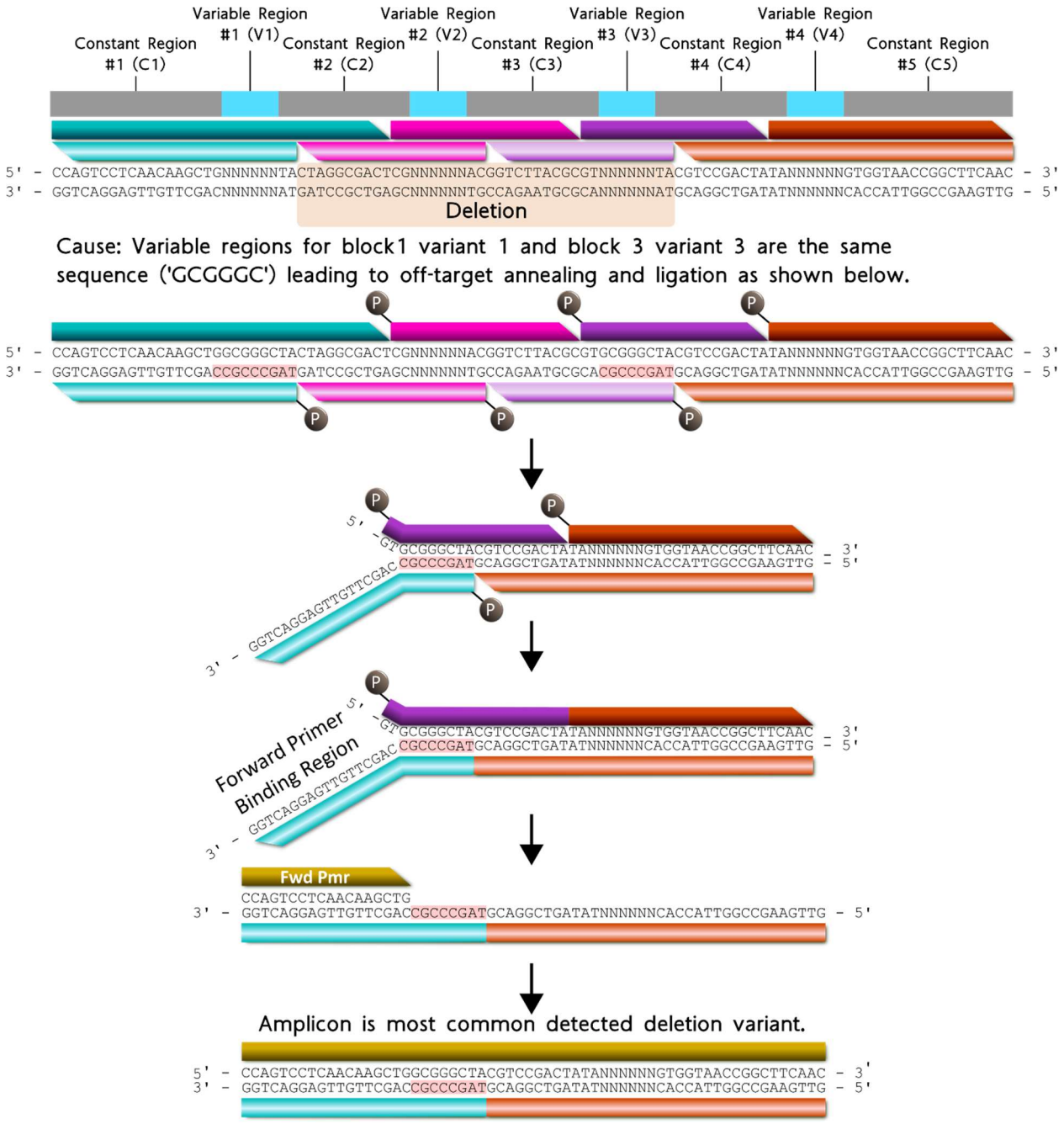



**Figure 4.5.** Proposed formation pathway for the highest deletion variant resulting from identical terminal nucleotides for block 1 (turquoise) and block 3 (purple) bottom strands.

within the first half of the modular barcode encompassing constant regions 2 and 3 and variable regions 2 and 3 internal to blocks 2 and 3.

The most common deletion variant occurred more than twice as often as the second highest variant, suggesting that some aspect of the barcode design favored formation of this

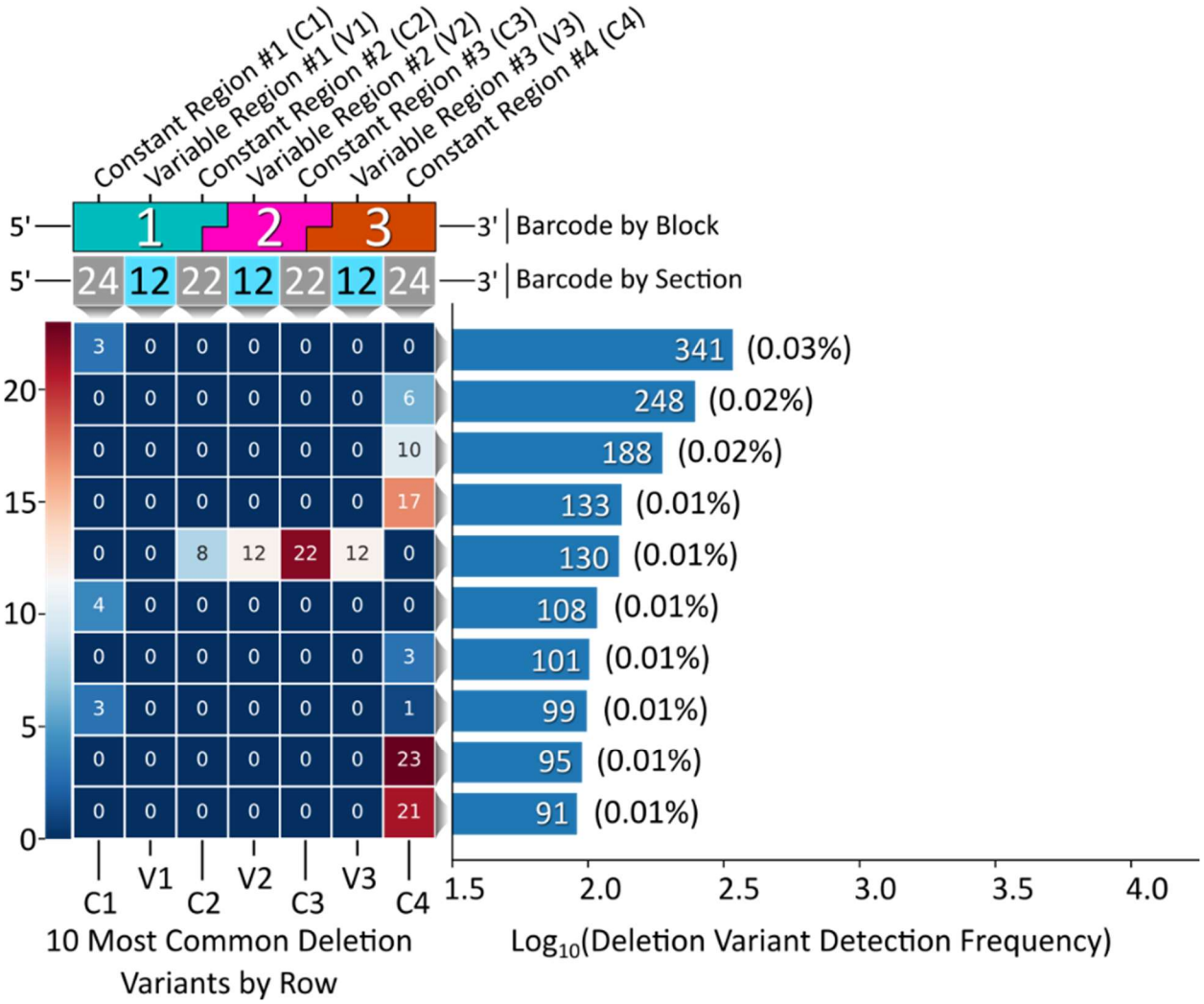



**Figure 4.6. Gen_2 Library Deletion Variant Analysis.** Left, a heatmap corresponding to the 10 most common deletion variants by row. Each block corresponds to a specific barcode domain displayed above the heatmap. Right, a histogram displaying the frequency of detection for each deletion variant out of the 1M read subset.

off-target variant, warranting further inspection. Out of the 13,333 reads for the most common deletion variant, 13,306 contained the variable region sequence 'GCGGGC', the only duplicate variable region sequence used a nd shared by blocks 1 and 3, highlighted in Table 4.4. Additionally, the 5' terminal two nucleotides for the bottom strands of blocks 1 (turquoise) and 3 (purple) were identical. As shown in Figure 4.5, the shared sequence (variable region and terminus) allowed block 1 to anneal adjacent to block 4, becoming ligated. Such an off-target annealing event created a truncated oligo containing the barcode forward primer binding

region in addition to the reverse primer sequence, allowing propagation throughout subsequent library preparation PCR stages.

Deletion variant analysis revealed a sequence motif to control for (i.e., negative design) during NUPACK sequence design, specified as a hard constraint for preventing duplicate nucleotides between block termini, in the subsequent gen_2 modular barcode library (Fig. S4.3). Remarkably, the most common deletion variant from the pooled, 96 barcode Gen_2 library contained no internal deletions, rather a 3-nt terminal deletion (Fig. 4.6). The fifth most common variant containing an internal deletion resulted from partial complementarity between blocks 1 and 3, suggesting an additional design principle for the design of subsequent libraries (Fig. S4.5). Overall, Gen_2 library construction was improved by avoiding duplicate variable region sequences coupled with the NUPACK design constraint incorporation for eliminating duplicate 5' terminal sequences adjacent to variable region sequences.

**4.3.6 Substitution analysis**

The number and type of 1-nucleotide substitutions were analyzed to determine the extent to which a modular barcode library of low complexity relative to conventional sequencing datasets exhibits substitution errors and trends that coincide with previously published data. As shown in Figure 4.7A, the position of 1-nt substitutions were distributed across the entire barcode sequence. The average per-nucleotide substitution frequency, as calculated among all full-length, non-insertion/deletion containing reads, was 0.05 ± 0.04% (Fig. S4.7), slightly lower than a previously reported rate of 0.24 ± 0.06% (106). The frequency of detected substitution types

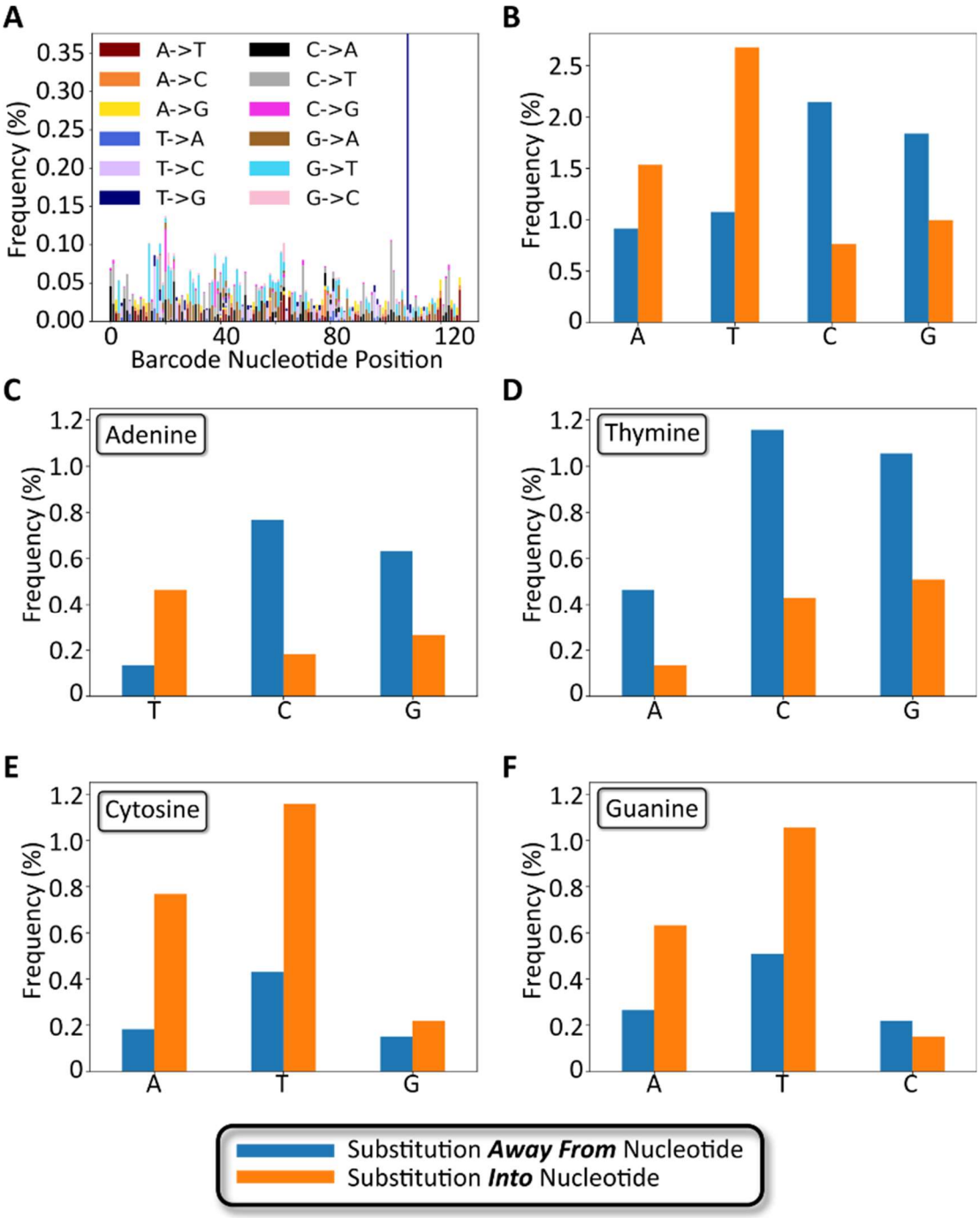



**Figure 4.7. Gen_1 Substitution Variant Analysis. A)** Stacked bar chart showing the number and type of 1-nucleotide substitutions plotted as a function of nucleotide position along the barcode. **B)** Bar chart showing nucleotide substitution frequency broken down by nucleotide. Blue bars represent substitutions away from specified nucleotide. Orange bars represent substitutions into specified nucleotide. **C - F)** Bar charts for substitution frequency for each of the four bases, adenine, thymine, cytosine, and guanine, respectively.

appeared to be random as well, with the exception of T108G that appears with an overall frequency approximating 0.38%. We hypothesize that this particular substitution occurred early

during library preparation PCR amplification of the barcode, rather than occurring as a random substitution during NGS.

The number and type of substitutions are broken down by nucleotide in Figure 4.7B-F. As shown in Figure 4.7B, substitutions away from the wildtype base occurred most frequently for cytosine followed by guanine, with thymine and adenine exhibiting the least substitutions. However, the opposite trend was observed for substitutions into a specific base, with thymine and adenine occurring more frequently than guanine and cytosine. The most frequent substitution types were C -> A, C -> T and G -> T in decreasing order. The least frequent substitution types were A -> C, C -> G, G -> C, and T -> C in decreasing order. Notably, the above trends, with the exception of T -> C as an infrequent substitution, matched trends previously reported by Pfeiffer et al. describing Illumina sequencing error rates (Table S4.2) (107). The comparable substitution errors reported here with previously described data confirm that our low complexity modular barcode libraries did not suffer from elevated substitution errors that would frustrate sequencing data quality and accurate read identification.

**4.4 Conclusions**

Modular DNA barcode libraries represent an efficient combinatorial method of generating numerous, NGS compatible synthetic DNA sequences for tracking applications. Nucleic acid sequences for target complexes were optimized to minimize the average number of incorrectly paired nucleotides (the ensemble defect rate) at equilibrium (108). Structure prediction of designed sequences was performed by NUPACK using dynamic programming to efficiently find the secondary structure hybridization pattern of lowest energy (109). Rigorous computational design and scoring ensured that all single-stranded oligo mixtures favor the target annealed

complex in solution and the primer sequences exhibit minimal propensity toward homo/heterodimer formation due to strict computational design. Experimental validation demonstrated recovery of all 256 gen_1 modular barcode sequences in parallel, emphasizing the degree of sample multiplexing provided by modular barcodes over other technologies, albeit with non-negligible off-target assembly events. The improved second-generation library was larger and had negligible off-target assembly outcomes. Further, amplification of barcode DNA from mosquito-derived samples and in the context of contaminating human DNA was highly sensitive and specific. We used NGS to simultaneously read the barcodes for 256 gen_1 samples, and, separately, 96 gen_2 samples. Analysis of NGS reads in our first-generation library allowed us to quantify polymerase error rates (deletions and substitutions). Observed error rates from the first-generation library motivated design of a second-generation barcode library where amplicon errors would not prevent accurate assignment of the parent barcode. We confirmed that low complexity modular libraries did not increase the likelihood of substitution errors relative to previously reported values.

Looking forward, modular DNA barcodes have the potential to enhance supply chain security and animal tracking by serving as highly unique, microscopic markers suitable for highly sensitive detection. Synthetic DNA in the environment remains vulnerable to degradation (e.g. by nucleases) (75). Therefore, depending on the application, it may be helpful to boost the barcode half-life by embedding otherwise vulnerable DNA inside a protective matrix (31). Barcode particles composed entirely of biomolecules are expected to be edible and biodegradable, resulting in broad application utility.

## 4.5 Methods

### 4.5.1 Sequence Design

Flanking primer sequences for the barcodes were designed by Primer3 fed a random nucleotide sequence generated using an online random DNA sequence generator (110). The block overhang regions and the terminal TrapTag primer sequence (Fig. 1) were designed using NUPACK based on the target secondary structure (102). For the gen_1 library, variable region sequences were designed by first creating all possible 6-nt permutations from the 4 bases: A, T, C, G. The resulting list of 4,096 sequences was filtered to remove homopolymers (e.g., sequences containing 4 or more identical, consecutive nucleotides) resulting in a list of 3,936 sequences. From this list, groups of 4 randomly chosen sequences (performed 4,000X) were used for in silico testing of barcode assembly using NUPACK. The free energy of the target complex was recorded. The top 4 scoring complexes (e.g., complexes with the lowest free energy scores) were chosen for experimental validation. Notably, Hamming distance was not employed as an explicit design parameter for the Gen_1 library, rather calculated *after* sequence design, resulting in a value of 3. In contrast, for the Gen_2 library, we used custom Python code (available upon request) to identify a set of 24 variable region sequences for which all pairs met an explicit and more stringent Hamming dissimilarity cutoff. Specifically, the minimum Hamming distance was 7 between any two barcodes (and also a minimum Hamming distance of 5 even if an indel were to shift a barcode register +/- 1) (Fig. S4.6, Table S4.3). All variant sequences are listed in Table 4.4. Following sequence design, all barcode variants were built *in silico* and predicted assembly fidelity was assessed using NUPACK, ideally ensuring the target barcode remains the dominant complex at equilibrium despite possible off-target

complex formation. Further code was written during gen_2 library design for automating primer specificity checking against certain species or within the library itself (see section titled 'Automated Primer Specificity Check' in extended methods of SI).

### 4.5.2 Primer Specificity and Sensitivity

To further enhance the specificity of the chosen primers and to reduce chances of off-target amplification of potentially contaminating DNA, each candidate primer pair designed by

**Table 4.4.** Modular barcode sequences designating each block variant.

| Gen_1 Variable Region Sequences | | | | Gen_2 Variable Region Sequences | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| TGCGGC | GGCGTC | AGCGGG | CGCTCC | ATCGACTGCGAG | ATGACGAGTGCT | TGTCTCGAGTCT |
| | | | | GCTAGCACTGAG | GAGATCTGCAGT | GCGCTGCTACTG |
| GCGGGC | CCCGGC | GCGGAA | ACTCGT | GTGTGCGCTAGC | CTATCGCGACGT | CACTCAGATGTG |
| | | | | TGCTCTAGTAGC | CATGCTGTCAGC | GATGTGCAGAGA |
| TGGGCG | CCTACC | GCTGCC | ACGCGG | CGATACGAGATC | CGACGTCTATCG | TCTCGTCGTATG |
| | | | | ACTGAGTGTCTC | GAGTCTACGTCG | CAGCAGTCTCGT |
| CGCCGG | GCACAG | GCGGGC | CCTTTG | TGCAGTGACTAG | GTCGCAGTACAG | CAGAGACAGCAG |
| | | | | AGCGTGACGCGT | ACAGTGATCGAC | ATAGCGCACTCA |

Primer3 was used to BLAST against multiple species that could contaminate barcode-positive samples. In addition, we considered downstream applications of this technology which would require amplification of small amounts of barcode DNA from complex sample types. In this case, we sought to design edible barcodes for marking mosquitoes (31). Therefore, the species selected for BLAST search were *Homo sapiens* and *Culicidae*. The BLAST procedure was automated using Python and the Biopython (111) NCBI command line functionality (custom Python scripts within "combinatorial_barcode_scripts.zip" available upon request). Following library construction *in silico*, primer specificity was again checked against the entire barcode library to ensure no off-target mis-priming was likely to occur. Specifically, a FASTA file was created containing all sequences for the barcode library. Candidate gen_2 TrapTag primer pairs

were further evaluated *in vitro* for template specificity by attempting PCR-amplification of DNA extracted from *Culex tarsalis* and *Aedes aegypti* mosquitoes or human saliva, to verify the lack of off-target amplification of insect or contaminating human DNA (see SI, extended methods sections titled 'in vitro Primer Sensitivity' and 'in vitro Primer Specificity'). The modular barcode library FASTA file was then converted to a custom database using Biopython. The BLAST procedure was repeated using the primer pair against the custom database for ensuring primers do not align internal to each barcode sequence.

### 4.5.3 Barcode Construction and Sequencing

Similar experimental validation and analysis methods were employed for both modular barcode libraries. For the gen_1 library, the 32 oligos corresponding to the 4 variants for each of the 4 blocks were purchased from Integrated DNA Technologies (Coralville, Iowa) with 6 oligos containing a 5' phosphate. Each oligo was resuspended to a stock concentration of 100 µM in duplex buffer (100 mM Potassium Acetate, 30 mM HEPES, pH 7.5). A 0.02 pmol/µL working solution was made from each stock solution using duplex buffer. From each of the 8 working solutions corresponding to a single modular barcode sequence, 2 µL was manually transferred to a 0.2 mL PCR tube and mixed. For the gen_2 library, a liquid handling robot (OpenTrons OT-

**Table 4.5.** Primer sequences used in overhang PCR for appending trap tag and sequencing adapters.

| Primer Set # | Forward Primer (5' – 3') | Reverse Primer (5' – 3') |
|---|---|---|
| 1 | ACACTCTTTCCCTACACGACGCTCTTCCGATCT CCAGTCCTCAACAAGCTG | GTTGAAGCCGGTTACCAC |
| 2 | ACACTCTTTCCCTACACGACGCTCTTCCGATCT | TTCTGGGTTCCTCATCGCNNNNNNNN GTTGAAGCCGGTTACCAC |
| 3 | ACACTCTTTCCCTACACGACGCTCTTCCGATCT | GTGACTGGAGTTCAGACGTGTGCTCTT CCGATCTTTCTGGGTTCCTCATCGC |
| 4 | AATGATACGGCGACCACCGAGATCTACACTCT TTCCCTACACGACGCTCTTCCGATCT | CAAGCAGAAGACGGCATACGAGATNN NNNNNNNNATATTCACGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCT |

2) was employed for initial oligo mixing to reduce chances of contamination. Oligo mixtures were heated to 94 °C for 4 minutes using a heat block, followed by gradual cooling for 1 hr by turning off the heat block. Following annealing, 2 µL T4 DNA Ligase Buffer (NEB) and 1µL T4 DNA Ligase (NEB) were added to the annealed mixture followed by incubation at room temperature for 10 minutes. The ligation reaction was heat inactivated by 10-minute incubation at 65 °C. Our OpenTrons Python pipetting script can be found within "combinatorial_barcode_scripts.zip" available upon request.

The inactivated ligation mixture product (102bp for the gen_1 library, 128 bp for the gen_2 library) was used as the template for overhang PCR with the following reaction conditions: 1 cycle of 98 °C for 45 seconds, 30 cycles of 98 °C for 30 seconds, 61 °C for 30 seconds, 72 °C for 30 seconds, and 1 cycle of 72 °C for 1 minute. Overhang PCR was performed using the Table 4.5 primer sets for amplifying barcode DNA, attaching the TrapTag and terminal sequencing adapters (Fig. 4.1). All PCR reactions, including those for validating primer specificity, were performed using the same thermocycling conditions described immediately above for overhang PCR. Following amplification, PCR cleanup was performed using KAPA Pure Beads (Roche). Size selection for the 262bp barcode library was performed using Monarch DNA Gel Extraction Kit (New England Biolabs). The library was quantified using Qubit 1X dsDNA HS Assay Kit (ThermoFisher) and each library member was diluted to 20 nM for sequencing sample prep. Paired end 2x150 cycle sequencing was run on an illumina NovaSEQ 6000 (Genomics and Microarray Core, University of Colorado Anschutz Medical Campus).

For NGS read analysis, the ea-utils package was used for initial sample processing including adapter trimming and read joining. FastQC was used to check overall quality of joined

reads and to determine total read count of detected barcode (Babraham Bioinformatics). Using the Biopython package and custom Python code, reads were aligned to a generic barcode template containing 'N's for the four variable sequence regions within each barcode (script can be found within "combinatorial_barcode_scripts.zip" available upon request). Following alignment, demultiplexing was performed based on the TrapTag UMI appended to each individual positive sample. NGS data were analyzed by gathering various statistics on recovered reads including, but not limited to, the number of reads containing no substitutions/deletions, the number/type of reads containing 1-nucleotide substitutions, and the number/type of reads containing variable nucleotide deletions. The average substitution rate per nucleotide was calculated for all non-indel reads by taking the average of the substitution rates determined for all nucleotide positions along the barcode, as described previously (107).

# CHAPTER 5. SUMMARY AND FUTURE DIRECTIONS

The goal of this work was to characterize a novel composite nanomaterial comprised of porous protein crystals and double-stranded DNA. We showed that fluorescently labeled guest DNA loads into host crystals predominantly along the axial nanopores. Modeling of guest transport and equilibrium diffusion, along with Fluorescence Recovery After Photobleaching (FRAP) experiments, result in attenuated pore diffusion coefficient value relative to bulk diffusion. We also showed that porous protein crystals loaded with barcode DNA can be used for tagging and tracking mosquitoes. Crystals can be ingested by both mosquito larvae and adult mosquitoes. DNA barcode recovery from mosquito specimens was performed using PCR, qPCR and NGS. Lastly, we showed that modular DNA barcodes present an economic, scalable DNA construction method for generating a library of numerous, distinct DNA sequences. Candidate sequences were designed and scored computationally using Python and a nucleic acid secondary structure prediction program for experimental validation of top candidates. Increased sequence diversity resulted in increased barcode annealing specificity, thus yielding the target barcode sequence.

While a biophysical characterization of guest DNA transport into host crystals was described, the models explored for simulating guest loading suffered inaccuracies regarding predicted binding and unbinding rate constants, and thus limited model utility. Future work includes optimizing the models employed for enhancing accurate computational depiction of experimental guest loading datasets. Additionally, the method employed for performing crystal adsorption isotherm experiments was meticulous (e.g., individual crystal measuring under a stereomicroscope) which hinders scalability in performing subsequent isotherms with various guest types (e.g., RNA, protein, fluorescent dyes). The method also relied upon using numerous,

large crystals for increasing overall incubation volume capable of measuring solution guest concentration using existing technology (e.g., Qubit 2 Fluorometer). A future alternative approach to explore would be to use a suspension of protein microcrystals of known mass which would reduce crystal-handling steps and increase throughput of experimental trials.

Remarkably, we showed that DNA barcode-loaded crystals can be ingested by mosquito larvae and adults followed by barcode recovery using standard molecular biology techniques. While barcode loaded crystals were applied to insects orally, such a route suggests the marking material may persist along the alimentary canal for a limited time followed by elimination in excreta. Future work includes measuring how long barcodes remain recoverable following ingestion. The barcode-crystal dose, or concentration, exposed to insect would also influence lifetime of barcode recovery. Additional future work includes performing dosing studies for assessing how barcode exposure amount influences detection over time. Such information (ingested barcode half-life, optimal barcode exposure dosage) will contribute to performing enhanced field trials with optimal barcode recovery likelihood.

Modular DNA barcodes represent a novel, effective design strategy for economical and scalable generation of unique DNA sequences. Given the unique context of using NGS for boasting the multiplexing capabilities of barcode detection, the analysis approach almost entirely relied upon custom Python scripts for separately exploring detectable trends in sequencing datasets (e.g., deletion variants, 1-nt substitution types). Future work includes compiling all analysis steps into a single pipeline capable of receiving a single-input (NGS dataset) and generating publication-quality output figures visually displaying results for subsequent interpretation. Also, the large file size of NGS datasets, coupled with the relatively

low complexity of modular DNA barcode libraries, motivated downsampling of datasets for analyzing a smaller portion ( ~ 1M reads) on a personal laptop. Additional future work includes altering existing Python analysis for compatibility for being executed on a High-Performance Computer (HPC) cluster, allowing timely analysis of entire sequencing datasets for enhancing accuracy of results obtained from analysis.

BIBLIOGRAPHY


1. B. D. Malhotra, M. A. Ali, "Chapter 5 - Nanocomposite Materials: Biomolecular Devices" in Nanomaterials for Biosensors, B. D. Malhotra, M. A. Ali, Eds. (William Andrew Publishing, 2018), https://doi.org/10.1016/B978-0-323-44923-6.00005-4, pp. 145-159.
2. S. Akgöl, F. Ulucan-Karnak, C. İ. kuru, K. Kuşat, The usage of composite nanomaterials in biomedical engineering applications. *Biotechnol Bioeng* **118**, 2906-2922 (2021).
3. N. F. Attia, S. E. A. Elashery, H. Oh, "Chapter 9 - Nanomaterials-based antibacterial textiles" in Nanosensors and Nanodevices for Smart Multifunctional Textiles, A. Ehrmann, T. A. Nguyen, P. Nguyen Tri, Eds. (Elsevier, 2021), https://doi.org/10.1016/B978-0-12-820777-2.00009-1, pp. 135-147.
4. A. El.Shafei, A. Abou-Okeil, ZnO/carboxymethyl chitosan bionano-composite to impart antibacterial and UV protection for cotton fabric. *Carbohydrate Polymers* **83**, 920-925 (2011).
5. M. M. Harussani, S. M. Sapuan, G. Nadeem, T. Rafin, W. Kirubaanand, Recent applications of carbon-based composites in defence industry: A review. *Defence Technology* **18**, 1281-1300 (2022).
6. V. Suvarna, A. Nair, R. Mallya, T. Khan, A. Omri, Antimicrobial Nanomaterials for Food Packaging. *Antibiotics (Basel, Switzerland)* **11** (2022).
7. S. Jafarzadeh, A. Salehabadi, S. M. Jafari, "10 - Metal nanoparticles as antimicrobial agents in food packaging" in Handbook of Food Nanotechnology, S. M. Jafari, Ed. (Academic Press, 2020), https://doi.org/10.1016/B978-0-12-815866-1.00010-8, pp. 379-414.

8. Q. Zhang, J.-Q. Huang, W.-Z. Qian, Y.-Y. Zhang, F. Wei, The Road for Nanomaterials Industry: A Review of Carbon Nanotube Production, Post-Treatment, and Bulk Applications for Composites and Energy Storage. *Small* **9**, 1237-1265 (2013).

9. K. Jiang *et al.*, Superaligned Carbon Nanotube Arrays, Films, and Yarns: A Road to Applications. *Adv Mater* **23**, 1154-1161 (2011).

10. A. Kubacka *et al.*, Understanding the antimicrobial mechanism of TiO2-based nanocomposite films in a pathogenic bacterium. *Sci Rep-Uk* **4**, 4134 (2014).

11. A. Samanta, I. L. Medintz, Nanoparticles and DNA – a powerful and growing functional combination in bionanotechnology. *Nanoscale* **8**, 9037-9095 (2016).

12. X. Liu *et al.*, Complex silica composite nanomaterials templated with DNA origami. *Nature* **559**, 593-598 (2018).

13. H. Wang, Y. Wang, J. Jin, R. Yang, Gold Nanoparticle-Based Colorimetric and "Turn-On" Fluorescent Probe for Mercury(II) Ions in Aqueous Solution. *Analytical Chemistry* **80**, 9021-9028 (2008).

14. Z. Zhou, Y. Du, S. Dong, DNA-Ag nanoclusters as fluorescence probe for turn-on aptamer sensor of small molecules. *Biosensors & bioelectronics* **28**, 33-37 (2011).

15. Z. Zhu *et al.*, Regulation of Singlet Oxygen Generation Using Single-Walled Carbon Nanotubes. *J Am Chem Soc* **130**, 10856-10857 (2008).

16. V. V. Sokolova, I. Radtke, R. Heumann, M. Epple, Effective transfection of cells with multi-shell calcium phosphate-DNA nanoparticles. *Biomaterials* **27**, 3147-3153 (2006).

17. M. Kojima, S. Abe, T. Ueno, Engineering of protein crystals for use as solid biomaterials. *Biomaterials Science* **10**, 354-367 (2022).

18. M. A. Dessau, Y. Modis, Protein Crystallization for X-ray Crystallography. *Jove-J Vis Exp* UNSP e2285

10.3791/2285 (2011).

19. A. McPherson, J. A. Gavira, Introduction to protein crystallization. *Acta Crystallogr F* **70**, 2-20 (2014).

20. F. A. Quiocho, F. M. Richards, Intermolecular Cross Linking of Protein in Crystalline State - Carboxypeptidase-A. *P Natl Acad Sci USA* **52**, 833-& (1964).

21. L. F. Hartje *et al.*, Characterizing the Cytocompatibility of Various Cross-Linking Chemistries for the Production of Biostable Large-Pore Protein Crystal Materials. *Acs Biomater Sci Eng* **4**, 826-831 (2018).

22. C. J. Lusty, A gentle vapor-diffusion technique for cross-linking of protein crystals for cryocrystallography. *J Appl Crystallogr* **32**, 106-112 (1999).

23. T. R. Huber, E. C. McPherson, C. E. Keating, C. D. Snow, Installing Guest Molecules at Specific Sites within Scaffold Protein Crystals. *Bioconjugate Chem* **29**, 17-22 (2018).

24. A. E. Kowalski *et al.*, Gold nanoparticle capture within protein crystal scaffolds. *Nanoscale* **8**, 12693-12696 (2016).

25. J. Sprenger *et al.*, Guest-protein incorporation into solvent channels of a protein host crystal (hostal). *Acta crystallographica. Section D, Structural biology* **77**, 471-485 (2021).

26. A. E. Kowalski *et al.*, Porous protein crystals as scaffolds for enzyme immobilization. *Biomater Sci* 10.1039/c8bm01378k (2019).

27. S. Lopez *et al.*, Cross-Linked Artificial Enzyme Crystals as Heterogeneous Catalysts for Oxidation Reactions. *J Am Chem Soc* **139**, 17994-18002 (2017).

28. C. G. Sonwane, S. K. Bhatia, Characterization of Pore Size Distributions of Mesoporous Materials from Adsorption Isotherms. *The Journal of Physical Chemistry B* **104**, 9099-9110 (2000).

29. T. Hashimoto *et al.*, Encapsulation of biomacromolecules by soaking and co-crystallization into porous protein crystals of hemocyanin. *Biochemical and Biophysical Research Communications* **509**, 577-584 (2019).

30. Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K. F. Au, Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348-1365 (2021).

31. J. D. Stuart *et al.*, Mosquito Tagging Using DNA-Barcoded Nanoporous Protein Microcrystals. *PNAS Nexus* 10.1093/pnasnexus/pgac190, pgac190 (2022).

32. D. H. Lin, A. Hoelz, The Structure of the Nuclear Pore Complex (An Update). *Annu Rev Biochem* **88**, 725-783 (2019).

33. M. Stewart, Nuclear export of mRNA. *Trends Biochem Sci* **35**, 609-617 (2010).

34. B. C. Durney, C. L. Crihfield, L. A. Holland, Capillary electrophoresis applied to DNA: determining and harnessing sequence and structure to advance bioanalyses (2009-2014). *Analytical and bioanalytical chemistry* **407**, 6923-6938 (2015).

35. G. Mikutis *et al.*, Silica-Encapsulated DNA-Based Tracers for Aquifer Characterization. *Environmental Science & Technology* **52**, 12142-12152 (2018).

36. J. Yun *et al.*, Role of the DksA-like protein in the pathogenesis and diverse metabolic activity of Campylobacter jejuni. *Journal of bacteriology* **190**, 4512-4520 (2008).

37. C. The UniProt, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489 (2021).

38. A. Cvetkovic *et al.*, Quantifying anisotropic solute transport in protein crystals using 3-D laser scanning confocal microscopy visualization. *Biotechnol Bioeng* **86**, 389-398 (2004).

39. A. Cvetkovic, C. Picioreanu, A. J. J. Straathof, R. Krishna, L. A. M. van der Wielen, Quantification of binary diffusion in protein crystals. *J Phys Chem B* **109**, 10561-10566 (2005).

40. A. Cvetkovic, C. Picioreanu, A. J. J. Straathof, R. Krishna, L. A. M. van der Wielent, Relation between pore sizes of protein crystals and anisotropic solute diffusivities. *J Am Chem Soc* **127**, 875-879 (2005).

41. J. Sprenger, C. L. Lawson, C. von Wachenfeldt, L. Lo Leggio, J. Carey, Crystal structures of Val58Ile tryptophan repressor in a domain-swapped array in the presence and absence of L-tryptophan. *Acta crystallographica. Section F, Structural biology communications* **77**, 215-225 (2021).

42. M. Schmidt, Reaction Initiation in Enzyme Crystals by Diffusion of Substrate. *Crystals* **10** (2020).

43. S. Pandey *et al.*, Observation of substrate diffusion and ligand binding in enzyme crystals using high-repetition-rate mix-and-inject serial crystallography. *IUCrJ* **8**, 878-895 (2021).

44. M. W. Martynowycz, T. Gonen, Ligand Incorporation into Protein Microcrystals for MicroED by On-Grid Soaking. *Structure* **29**, 88-95.e82 (2021).

45. S. Geremia, M. Campagnolo, N. Demitri, L. N. Johnson, Simulation of Diffusion Time of Small Molecules in Protein Crystals. *Structure* **14**, 393-400 (2006).

46. K. Mori, B. Kuhn, Imaging Ca2+ Concentration and pH in Nanopores/Channels of Protein Crystals. *The Journal of Physical Chemistry B* **122**, 9646-9653 (2018).

47. T. Sato, K. Hata, K. Nakatani, Mass Transfer in Mesoporous Microparticles Studied by Confocal Fluorescence Recovery after Photobleaching. *Analytical Sciences* **33**, 647-650 (2017).

48. J. Gutenwik, B. Nilsson, A. Axelsson, Coupled diffusion and adsorption effects for multiple proteins in agarose gel. *AIChE Journal* **50**, 3006-3018 (2004).

49. J. Gutenwik, B. Nilsson, A. Axelsson, Effect of hindered diffusion on the adsorption of proteins in agarose gel using a pore model. *Journal of Chromatography A* **1048**, 161-172 (2004).

50. H. Ishikawa-Ankerhold, R. Ankerhold, G. Drummen, Fluorescence Recovery After Photobleaching (FRAP). *eLS* doi:10.1002/9780470015902.a0003114

10.1002/9780470015902.a0003114 (2014).

51. L. F. Hartje, B. Munsky, T. W. Ni, C. J. Ackerson, C. D. Snow, Adsorption-Coupled Diffusion of Gold Nanoclusters within a Large-Pore Protein Crystal Scaffold. *J Phys Chem B* **121**, 7652-7659 (2017).

52. R. Eymard, T. Gallouët, R. Herbin, "Finite volume methods" in Handbook of Numerical Analysis. (Elsevier, 2000), vol. 7, pp. 713-1018.

53. A. Shukla, A. K. Singh, P. Singh, A Comparative Study of Finite Volume Method and Finite Difference Method for Convection-Diffusion Problem. *American Journal of Computational and Applied Mathematics* **1**, 67-73 (2012).

54. P. Dechadilok, W. M. Deen, Hindrance Factors for Diffusion and Convection in Pores. *Industrial & Engineering Chemistry Research* **45**, 6953-6959 (2006).

55. A. Cvetkovic, A. J. Straathof, R. Krishna, L. A. van der Wielen, Adsorption of xanthene dyes by lysozyme crystals. *Langmuir* **21**, 1475-1480 (2005).

56. G. Koulouras *et al.*, EasyFRAP-web: a web-based tool for the analysis of fluorescence recovery after photobleaching data. *Nucleic Acids Res* **46**, W467-W472 (2018).

57. M. Kang, C. A. Day, A. K. Kenworthy, E. DiBenedetto, Simplified Equation to Extract Diffusion Coefficients from Confocal FRAP Data. *Traffic* **13**, 1589-1600 (2012).

58. J. Schindelin *et al.*, Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012).

59. L. D. Kramer, A. T. Ciota, Dissecting vectorial capacity for mosquito-borne viruses. *Curr Opin Virol* **15**, 112-118 (2015).

60. R. B. Matlock, S. R. Skoda, Mark-recapture estimates of recruitment, survivorship and population growth rate for the screwworm fly, Cochliomyia hominivorax. *Med Vet Entomol* **23**, 111-125 (2009).

61. D. Cianci *et al.*, Estimating Mosquito Population Size From Mark-Release-Recapture Data. *J Med Entomol* **50**, 533-542 (2013).

62. C. A. Guerra *et al.*, A global assembly of adult female mosquito mark-release-recapture data to inform the control of mosquito-borne pathogens. *Parasites & Vectors* **7** (2014).

63. N. O. Verhulst, J. A. C. M. Loonen, W. Takken, Advances in methods for colour marking of mosquitoes. *Parasites & Vectors* **6**, 200 (2013).

64. B. J. Johnson *et al.*, Use of rhodamine B to mark the body and seminal fluid of male Aedes aegypti for mark-release-recapture experiments and estimating efficacy of sterile male releases. *Plos Neglect Trop D* **11** (2017).

65. E. E. Wilkins, S. C. Smith, J. M. Roberts, M. Benedict, Rubidium marking of Anopheles mosquitoes detectable by field-capable X-ray spectrometry. *Med Vet Entomol* **21**, 196-203 (2007).

66. G. L. Hamer *et al.*, Evaluation of a Stable Isotope Method to Mark Naturally-Breeding Larval Mosquitoes for Adult Dispersal Studies. *J Med Entomol* **49**, 61-70 (2012).

67. R. Faiman *et al.*, Marking mosquitoes in their natural larval sites using (2)H-enriched water: a promising approach for tracking over extended temporal and spatial scales. *Methods in ecology and evolution* **10**, 1274-1285 (2019).

68. B. L. Dickens, H. L. Brant, Effects of marking methods and fluorescent dusts on Aedes aegypti survival. *Parasit Vectors* **7**, 65 (2014).

69. D. Rojas-Araya, B. W. Alto, N. Burkett-Cadena, D. A. Cummings, Detection of Fluorescent Powders and Their Effect on Survival and Recapture of Aedes aegypti (Diptera: Culicidae). *J Med Entomol* **57**, 266-272 (2020).

70. J. G. Juarez *et al.*, Dispersal of female and male Aedes aegypti from discarded container habitats using a stable isotope mark-capture study design in South Texas. *Sci Rep-Uk* **10** (2020).

71. I. Filipović *et al.*, Using spatial genetics to quantify mosquito dispersal for control programs. *BMC Biology* **18**, 104 (2020).

72. H. Schmidt *et al.*, Transcontinental dispersal of Anopheles gambiae occurred from West African origin via serial founder events. *Communications Biology* **2**, 473 (2019).

73. X.-Z. Kong *et al.*, Tomographic Reservoir Imaging with DNA-Labeled Silica Nanotracers: The First Field Validation. *Environmental Science & Technology* **52**, 13681-13689 (2018).

74. A. Glover, N. Aziz, J. Pillmoor, D. W. J. McCallien, V. B. Croud, Evaluation of DNA as a taggant for fuels. *Fuel* **90**, 2142-2146 (2011).

75. M. Puddu, D. Paunescu, W. J. Stark, R. N. Grass, Magnetically Recoverable, Thermostable, Hydrophobic DNA/Silica Encapsulates and Their Application as Invisible Oil Tags. *ACS Nano* **8**, 2677-2685 (2014).

76. C. A. Mora, D. Paunescu, R. N. Grass, W. J. Stark, Silica particles with encapsulated DNA as trophic tracers. *Molecular Ecology Resources* **15**, 231-241 (2015).

77. I. H. Sabir, J. Torgersen, S. Haldorsen, P. Aleström, DNA tracers with information capacity and high detection sensitivity tested in groundwater studies. *Hydrogeology Journal* **7**, 264-272 (1999).

78. P. A. Carr, G. M. Church, Genome engineering. *Nat Biotechnol* **27**, 1151-1162 (2009).

79. R. Faiman *et al.*, A novel fluorescence and DNA combination for versatile, long-term marking of mosquitoes. *Methods in ecology and evolution* **12**, 1008-1016 (2021).

80. G. Hampikian, T. Andersen, Absent sequences: nullomers and primes. *Pac Symp Biocomput*, 355-366 (2007).

81. J. Goswami, M. C. Davis, T. Andersen, A. Alileche, G. Hampikian, Safeguarding forensic DNA reference samples with nullomer barcodes. *J Forensic Leg Med* **20**, 513-519 (2013).

82. P. J. Linser, K. E. Smith, T. J. Seron, M. Neira Oviedo, Carbonic anhydrases and anion transport in mosquito midgut pH regulation. *The Journal of experimental biology* **212**, 1662-1671 (2009).

83. D. Wang, J. D. Stuart, A. A. Jones, C. D. Snow, M. J. Kipper, Measuring interactions of DNA with nanoporous protein crystals by atomic force microscopy. *Nanoscale* **13**, 10871-10881 (2021).

84. C. T. Wittwer, M. G. Herrmann, C. N. Gundry, K. S. J. Elenitoba-Johnson, Real-Time Multiplex PCR Assays. *Methods* **25**, 430-442 (2001).

85. A. Rajagopal *et al.*, Significant Expansion of Real-Time PCR Multiplexing with Traditional Chemistries using Amplitude Modulation. *Sci Rep-Uk* **9**, 1053 (2019).

86. R. W. Merritt, R. H. Dadd, E. D. Walker, Feeding behavior, natural food, and nutritional relationships of larval mosquitoes. *Annu Rev Entomol* **37**, 349-376 (1992).

87. J. R. Hagler, C. G. Jackson, METHODS FOR MARKING INSECTS: Current Techniques and Future Prospects. *Annual Review of Entomology* **46**, 511-543 (2001).

88. E. Kauffman *et al.*, Rearing of Culex spp. and Aedes spp. Mosquitoes. *Bio Protoc* **7**, e2542 (2017).

89. A. Untergasser *et al.*, Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**, e115-e115 (2012).

90. T. Koressaar, M. Remm, Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289-1291 (2007).

91. M. Newville, Stensitzki, Till, Allen, Daniel B., Ingargiola, Antonio, LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python. *Zenodo* 10.5281/zenodo.11813 (2014).

92. D. Singh, A. Verma, Inventory Management in Supply Chain. *Materials Today: Proceedings* **5**, 3867-3872 (2018).

93. H. Ringsberg, "Bar Coding for Product Traceability" in Reference Module in Food Science. (Elsevier, 2016), https://doi.org/10.1016/B978-0-08-100596-5.03165-6.

94. D. Paunescu, W. J. Stark, R. N. Grass, Particles with an identity: Tracking and tracing in commodity products. *Powder Technology* **291**, 344-350 (2016).

95. D. Karamessini *et al.*, Abiotic Sequence-Coded Oligomers as Efficient In Vivo Taggants for the Identification of Implanted Materials. *Angewandte Chemie* **130**, 10734-10738 (2018).

96. M. J. Austin, A. M. Rosales, Tunable biomaterials from synthetic, sequence-controlled polymers. *Biomaterials Science* **7**, 490-505 (2019).

97. J. Gooch, C. Koh, B. Daniel, V. Abbate, N. Frascione, Establishing evidence of contact transfer in criminal investigation by a novel 'peptide coding' reagent. *Talanta* **144**, 1065-1069 (2015).

98. J. Gooch, H. Goh, B. Daniel, V. Abbate, N. Frascione, Monitoring Criminal Activity through Invisible Fluorescent "Peptide Coding" Taggants. *Analytical Chemistry* **88**, 4456-4460 (2016).

99. Y. Erlich, D. Zielinski, DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950-954 (2017).

100. G. M. Church, Y. Gao, S. Kosuri, Next-Generation Digital Information Storage in DNA. *Science* **337**, 1628-1628 (2012).

101. N. Goldman *et al.*, Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77 (2013).

102. J. N. Zadeh *et al.*, NUPACK: Analysis and Design of Nucleic Acid Systems. *J Comput Chem* **32**, 170-173 (2011).

103. T. Kivioja *et al.*, Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**, 72-74 (2012).

104. T. Smith, A. Heger, I. Sudbery, UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499 (2017).

105. D. G. Robinson, J. D. Storey, subSeq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics* **30**, 3424-3426 (2014).

106. E. J. Fox, K. S. Reid-Bayliss, M. J. Emond, L. A. Loeb, Accuracy of Next Generation Sequencing Platforms. *Next generation, sequencing & applications* **1** (2014).

107. F. Pfeiffer *et al.*, Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep-Uk* **8**, 10950 (2018).

108. B. R. Wolfe, N. A. Pierce, Sequence Design for a Test Tube of Interacting Nucleic Acid Strands. *ACS Synthetic Biology* **4**, 1086-1100 (2015).

109. M. E. Fornace, N. J. Porubsky, N. A. Pierce, A Unified Dynamic Programming Framework for the Analysis of Interacting Nucleic Acid Strands: Enhanced Models, Scalability, and Speed. *ACS Synthetic Biology* **9**, 2665-2678 (2020).

110. P. Stothard, The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**, 1102-1104 (2000).

111. P. J. A. Cock *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).

APPENDIX I. Supplemental Information for Chapter 2 |
Characterization of Guest DNA Transport and Adsorption within Host Porous Protein Crystals

**Authors**
Julius D. Stuart[a,1], Szu-Hsuan (Ashlyn) Chen[b,1], Christopher D. Snow[b]

**Author Affiliations**
[a]Department of Chemistry, Colorado State University, Fort Collins, CO 80523; [b]Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO 80523

**Corresponding Author**
Christopher D. Snow
(970) 491-5276
Christopher.snow@colostate.edu
356 Scott Bioengineering
Colorado State University
Fort Collins, CO 80523

**Author contributions**: C.D.S., J.D.S., A.C. designed research; J.D.S., A.C. performed research; J.D.S., A.C. analyzed data; J.D.S., A.C. wrote the paper; C.D.S., J.D.S., A.C. edited the manuscript.

[1]J.D.S. and A.C. contributed equally to this work.

**This section includes:**
Figure S2.1-2.5
Extended Methods

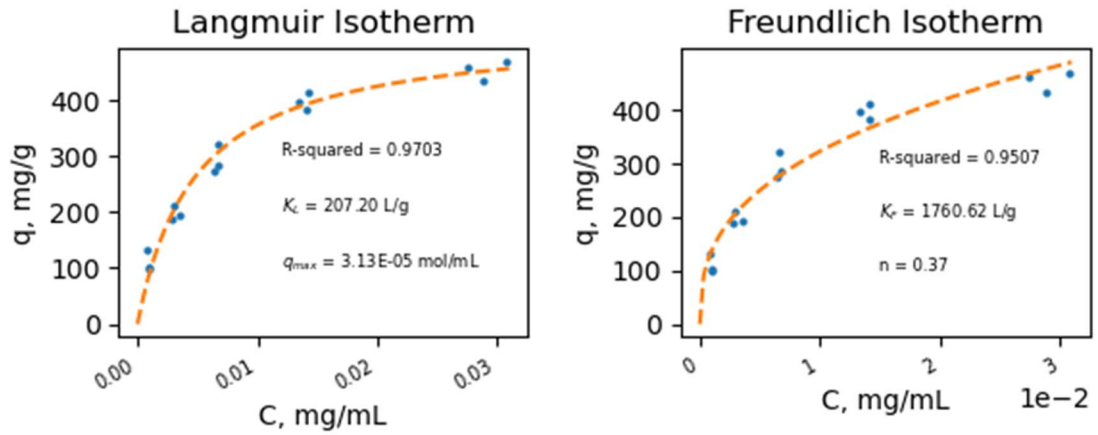



**Figure S2.1.** Isotherm Results. The adsorption isotherm data was fit with the Langmuir model (left) and the Freundlich model (right).

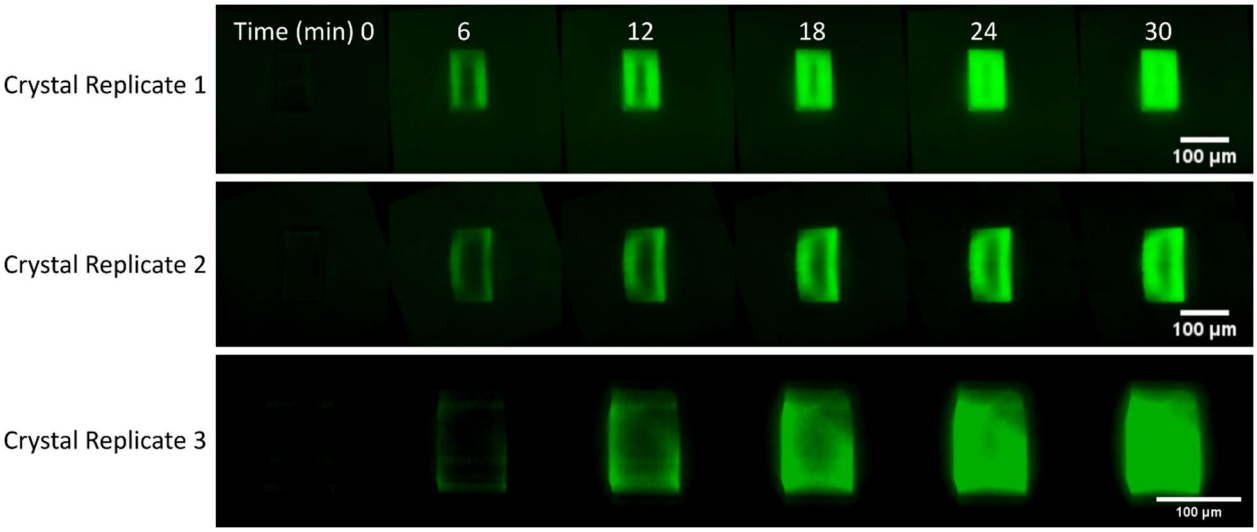



**Figure S2.2.** Timelapse confocal imaging of FAM-labeled 15mer loading into all crystal replicates over the course of 30 minutes.

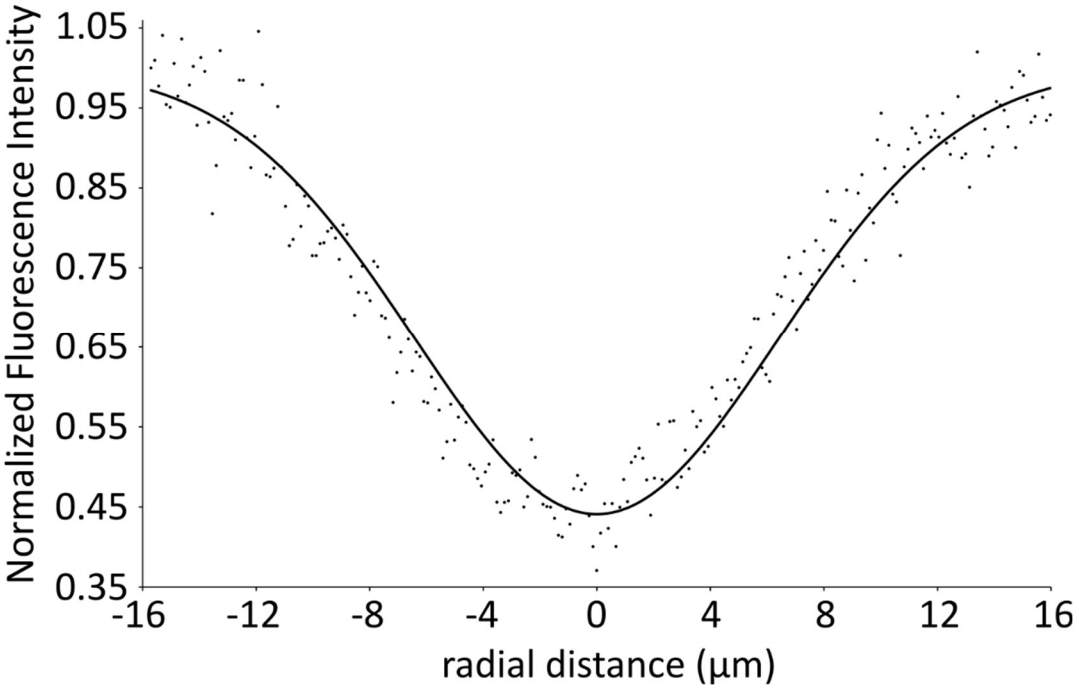



**Figure S2.3.** Normalized Mean Postbleach Profile. The normalized fluorescence intensity (circles) immediately after photobleaching plotted as a function of radial distance from the center of the bleached region. Experimental data were fit (solid line) with the following equation to determine the effective bleaching radius ($r_e$): $1 - K \exp(-2x^2/r_e^2)$. The resultant fitting allowed calculation of $D_{confocal}$ resulting in a value of 1.3e-10 $cm^2$/sec, falling within one order of magnitude of the modeled FRAP $D_{pore}$ value.

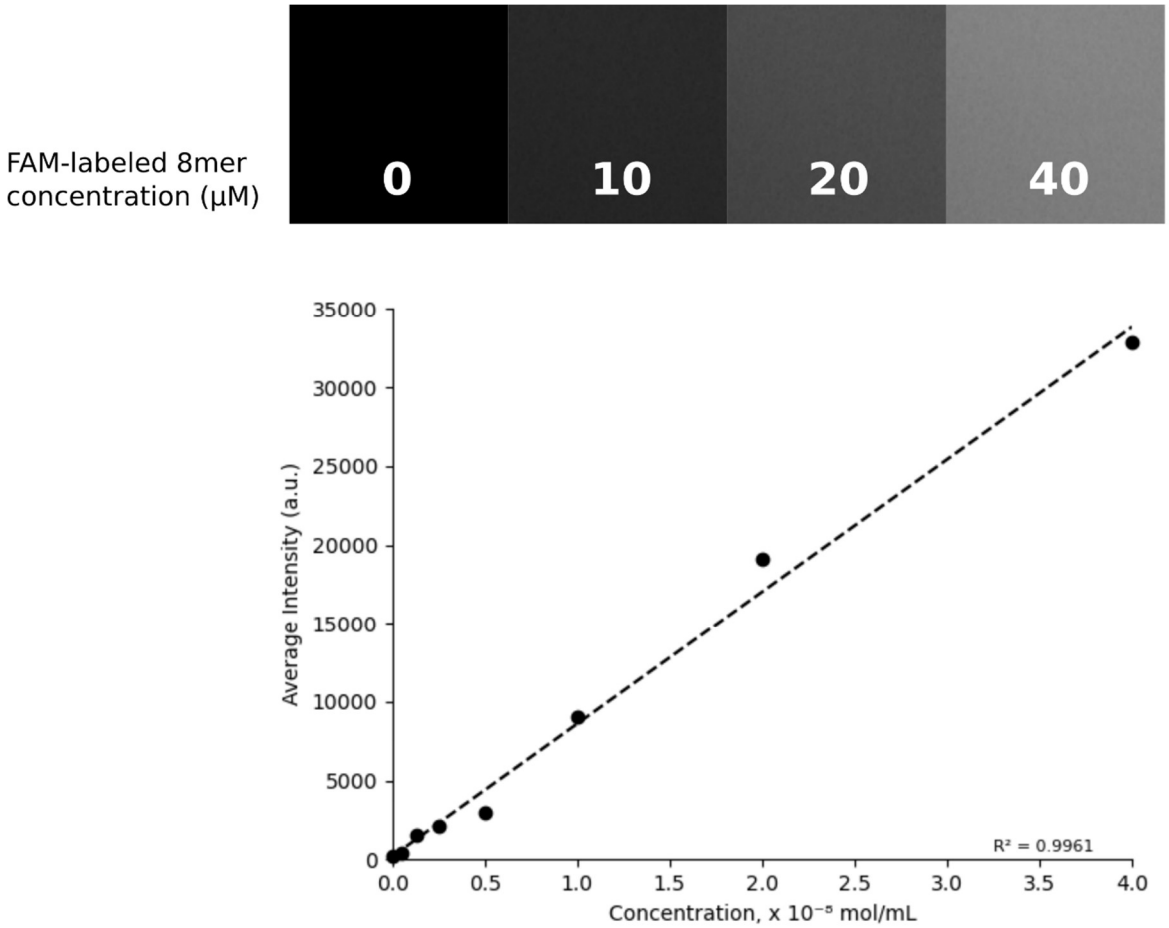



**Figure S4.** Standard curve of confocal images (grayscale) of increasing concentrations of FAM-labeled 8mer DNA.

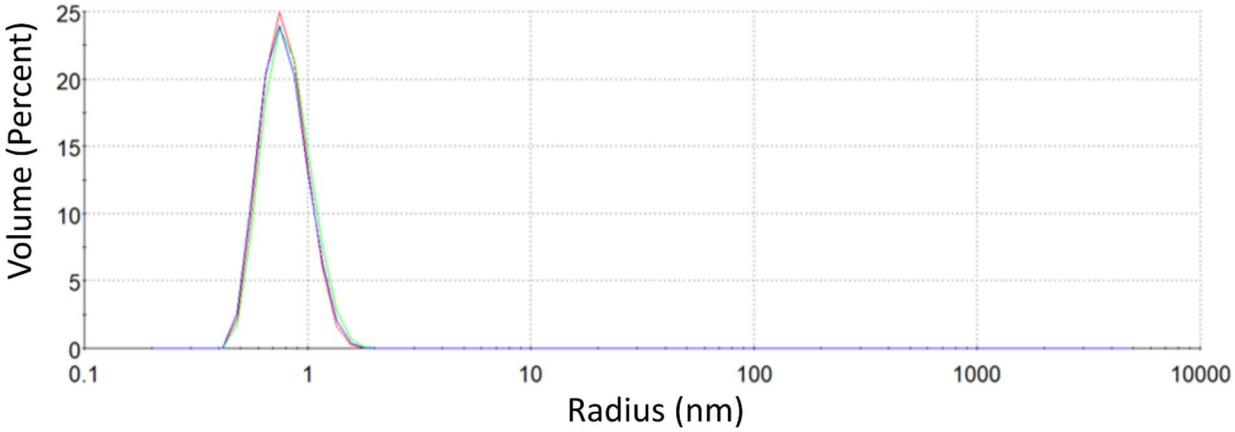



**Figure S5.** Dynamic Light Scattering (DLS) for 3 replicates of FAM-labeled 8mer in TE buffer yielding an average particle radius of 0.8 ± 0.2 nm.

**Extended Methods**

**Alternative $D_{confocal}$ Calculation from Confocal Datasets**

The normalized mean postbleach profile was plotted for calculating the effective postbleach radius ($r_e$) as previously described(57) by fitting the profile with the following formula using Microsoft Excel Solver tool:

$$f(x) = 1 - K \exp\left(-\frac{2x^2}{r_e^2}\right)$$

where *x* is the radial distance from the bleach spot center and *K* refers to the bleaching depth. Concurrently, FRAP analysis was performed using the online tool easyFRAP-web(56) for calculating the half-time for recovery ($\tau_{1/2}$) and mobile fraction. The diffusion coefficient from confocal FRAP datasets ($D_{confocal}$) was calculated as previously described(57) using the following formula:

$$D_{confocal} = \frac{r_n^2 + r_e^2}{8\tau_{1/2}}$$

where $r_n$ is the user defined nominal bleaching radius.

APPENDIX II. Supplemental Information for Chapter 3 |
Mosquito Tagging Using DNA-Barcoded Nanoporous Protein Microcrystals

**Authors**
Julius D. Stuart[a,1], Daniel A. Hartman[b,c,1], Lyndsey I. Gray[b], Alec A. Jones[d], Natalie R. Wickenkamp[b], Christine Hirt[b,e], Aya Safira[d,f], April R. Regas[g], Therese M. Kondash[h,i], Margaret L. Yates[j], Sergei Driga[k], Christopher D. Snow[a,d,j,k], Rebekah C. Kading[b]

**Author Affiliations**
[a]Department of Chemistry, Colorado State University, Fort Collins, CO 80523; [b]Department of Microbiology, Immunology, and Pathology, Colorado State University, Fort Collins, CO 80523; [c]Department of Entomology, Cornell University, Ithaca NY 14853 (current);[d]School of Biomedical Engineering, Colorado State University, Fort Collins, CO 80523;[e]Invitae, Longmont, CO 80503 (current);[f]Just-Evotec Biologics, Seattle WA 98109 (current);[g]College of Veterinary Medicine and Biological Sciences, Colorado State University, Fort Collins, CO 80523;[h]Department of Environmental Health and Radiological Sciences, Colorado State University, Fort Collins, CO 80523; [i]H3 Environmental, Albuquerque, NM 87109 (current);[j]Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, CO 80523; [k]Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO 80523

**Corresponding Author**
Rebekah C. Kading
(970) 491-7833
Rebekah.Kading@colostate.edu
176 CVID
Colorado State University
Fort Collins, CO 80523

[1]J.D.S. and D.A.H. contributed equally to this work.

**This section includes:**
DNA and protein sequence content for crystal loading and recovery
Figures S3.1 – S3.12
Table S3.1
Extended Materials and Methods

**DNA Sequence (5’ – 3’) for Microcrystal Protein Monomer**

TTAAGAAGGAGATATACATATGAAAAAAGTTCTGCTGAGCAGCCTGGTTGCAGTTAGCCTGCTGAGTACCGGTCT
GTTTGCAAAAGAATATACCCTGGATAAAGCCCATACCGATGTTGGCTTTAAAATCAAACATCTGCAGATTAGCAAT
GTGAAAGGCAACTTTAAAGATTATAGCGCAGTGATCGATTTTGATCCGGCAAGTGCAGAATTCAAAAAACTGGAT
GTGACCATTAAAATCGCCAGCGTGAATACCGAAAATCAGACCCGTGATAATCATCTGCAGCAGGATGACTTCTTCA
AAGCCAAAAAATACCCGGATATGACCTTTACCATGAAAAAATACGAGAAAATCGATAACGAAAAAGGCAAAATGA
CCGGCACCCTGACCATTGCCGGTGTTAGCAAAGATATTGTTCTGGATGCAGAAATTGGTGGTGTTGCCAAAGGTA
AAGATGGCAAAGAAAAAATTGGCTTTAGCCTGAACGGCAAAATCAAACGTAGCGATTTCAAATTTGCAACCAGCA
CCAGCACCATTACCCTGAGTGATGACATTAATCTGAACATTGAAGTGAAAGCCAACGAGAAAGAAGGTGGTAGTC
ATCACCACCACCATCACTAATAACTCGAGCACCACCACCACCACCACTGAGATCCGGCTG

**Protein Sequence for Microcrystal Protein Monomer**

MKEYTLDKAHTDVGFKIKHLQISNVKGNFKDYSAVIDFDPASAEFKKLDVTIKIASVNTENQTRDNHLQQDDFFKAKKYP
DMTFTMKKYEKIDNEKGKMTGTLTIAGVSKDIVLDAEIGGVAKGKDGKEKIGFSLNGKIKRSDFKFATSTSTITLSDDINL
NIEVEANEKEGGSHHHHHH

**200mer sequences (5’ – 3’)**

Nuclease-free water sample:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTACTAGGCGACTCGACGGT
CTTACGCGTTACGTCCGACTATAGAGCTTAGATTAGCGACGTTAAGATCGGAAGAGCACACGTCTGAACTCCAGTC
ACACAGGCGCNNNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG

Mosquito homogenate sample:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATACTAGACCGCTCGATCC
GACCTAGCGTACCTAGTACGTTACGACGACTAAGCATACCGCTAAGATCGGAAGAGCACACGTCTGAACTCCAGT
CACCATAGAGTNNNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG

Loaded microcrystals in nuclease-free water sample:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTCTCGTCCGACGGTCTTA
CGCGTTACGCCAAGTCTGCTAGCGTACGCTACGGTCTTGGACTCAGATCGGAAGAGCACACGTCTGAACTCCAGTC
ACTGCGAGACNNNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG

Loaded microcrystals in mosquito homogenate sample:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCAGAATTCGACGGTCTT
ACGCGTTACGATGAGGCCGCTAGCGTACGCTACGGTCACTAAGATAGATCGGAAGAGCACACGTCTGAACTCCAG
TCACTCTCTACTNNNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG

**200mer forward primer DNA sequence (5’ – 3’)**

AATGATACGGCGACCACCGAGATCT

**200mer reverse primer DNA sequence (5’ – 3’)**

CAAGCAGAAGACGGCATACGAGAT

**125mer DNA Sequence (5’ – 3’)**

TAGGCGACTCGACGGTCTTACGCGTTACGTATGATATGCATCACCACCATCACCAATAACCAACACCTAAATTTAAC
ATCCGAGAATTATGGAGCACGCTAGCGTACGCTACGGTCCTAACGCGC

**125mer forward primer DNA sequence (5’ – 3’)**

TAGGCGACTCGACGGTCTTACGCGTTACGT

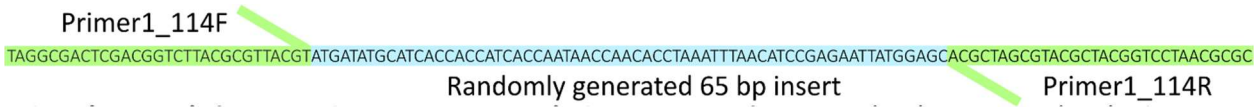
**125mer reverse primer DNA sequence (5’ – 3’)**

GCGCGTTAGGACCGTAGCGTACGCTAGCGT

**125mer revised forward primer DNA sequence (5’ – 3’)**
CATCACCACCATCACCAA

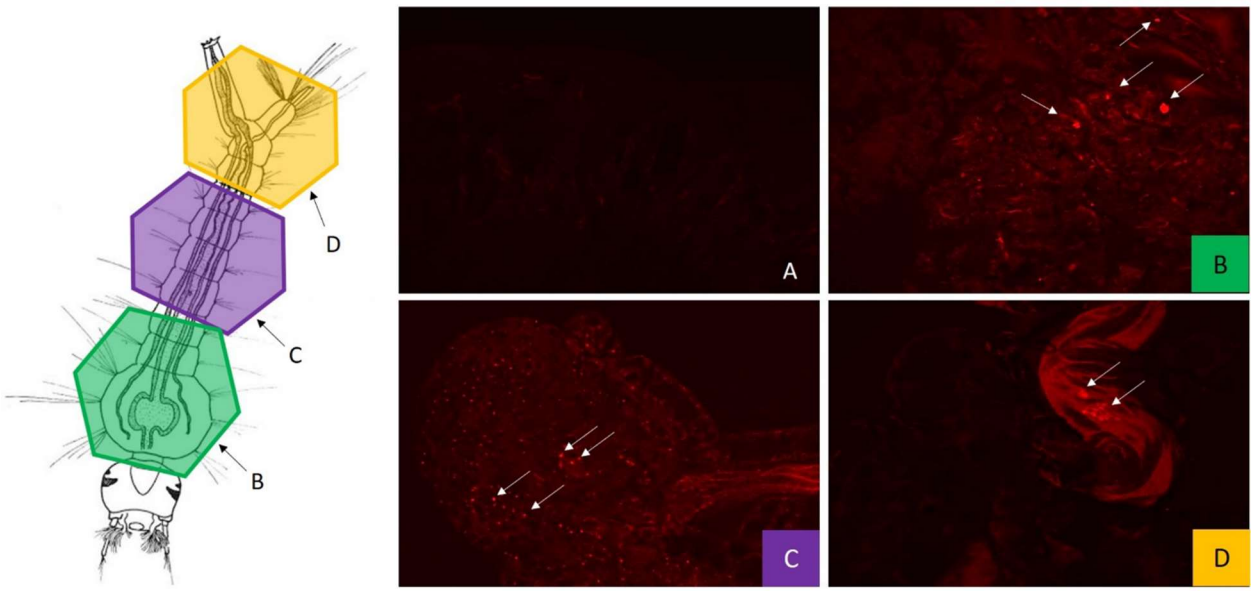
**15mer TAMRA DNA Sequence (5’ – 3’)**
TAMRA - CGGAGCACGCACGCC

**15mer Fluorescein DNA Sequence (5’ – 3’)**
FAM - CCGCACGCACGAGGC

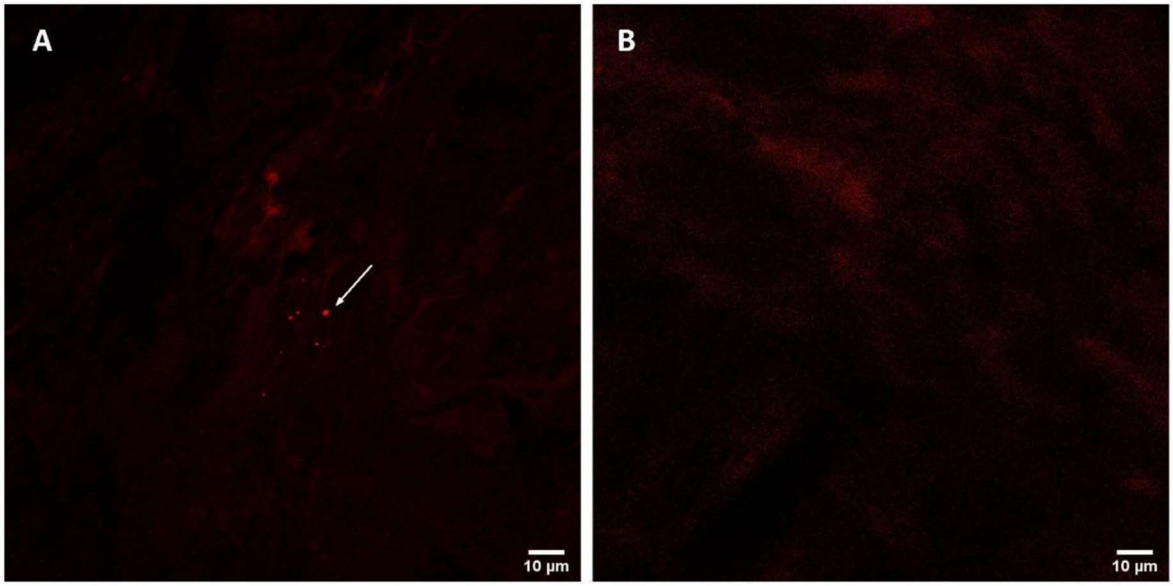



**Supplemental Figure 3.1. DNA Barcode Sequence Design**

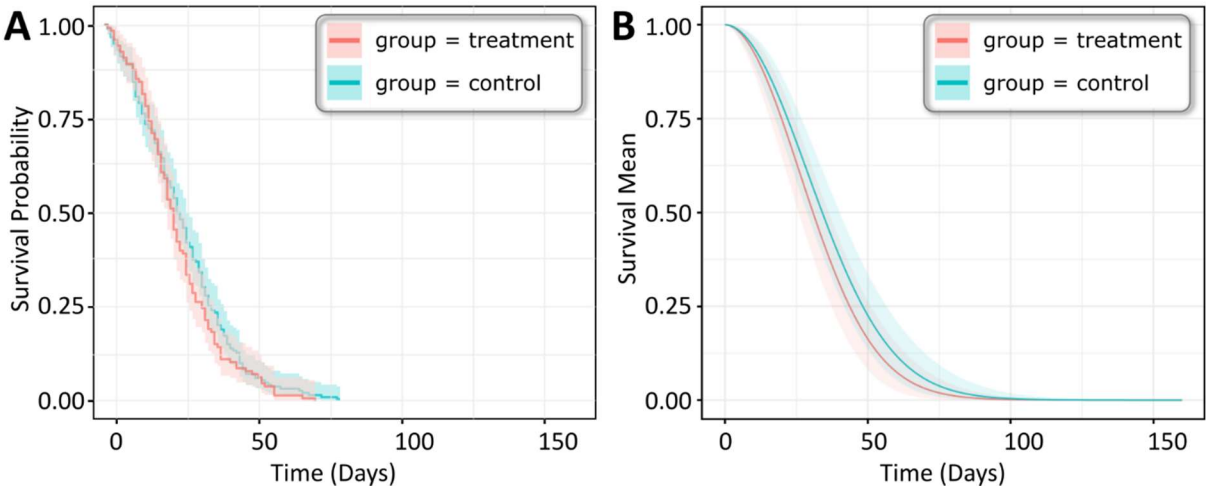
A randomly generated 65 bp insert using a publicly available online Sequence Manipulation Suite(110) is flanked by nullomer barcode primers provided in the Supplemental Information for Goswami et al. and follow the same naming convention(81).




**Supplemental Figure 3.2.** Detection of Texas Red labeled crystals in multiple regions (panels **B – D**) of the larval midgut of *Culex tarsalis* mosquitoes corresponding to labeled regions in diagram (left). No fluorescence was detected from non-crystal fed larvae (panel **A**). Scale bar not shown.




**Supplemental Figure 3.3. A)** Microcrystals loaded with Texas Red visualized inside the alimentary tract of larval *Culex tarsalis* mosquitoes. **B)** Negative control: *Culex tarsalis* larval alimentary tract after consuming liver powder alone. Images taken at 40X magnification. Scale bar denotes 10 µm.



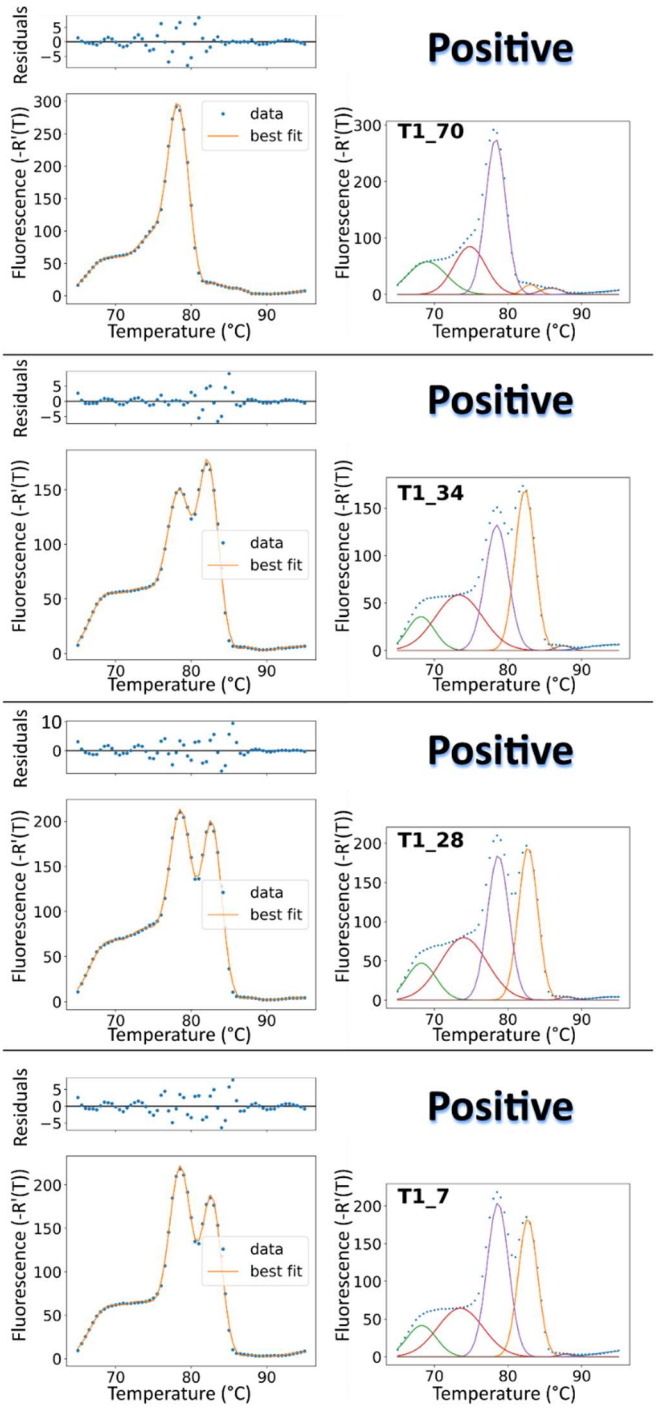

**Supplemental Figure 3.4. A)** Kaplan-Meier curves for two combined replicates of adult mosquito survivorship. Lines indicate survival probabilities for treatment (larvae fed microcrystals mixed with liver powder) and control (larvae fed liver powder alone) mosquito groups. Shaded areas indicate 95% confidence intervals around survival probabilities. **B)** Posterior predictions for adult mosquito survival probabilities from the Weibull survival model are shown as lines for treatment and control groups. Plotting the posterior predictions show overlap between 95% credible intervals over the entire time course.

| Adult Mosquito Age | % barcode positive | n |
|---|---|---|
| 1-10 | 100 | 14 |
| 11-20 | 73 | 26 |
| 21-30 | 81 | 21 |
| 31-40 | 83 | 6 |

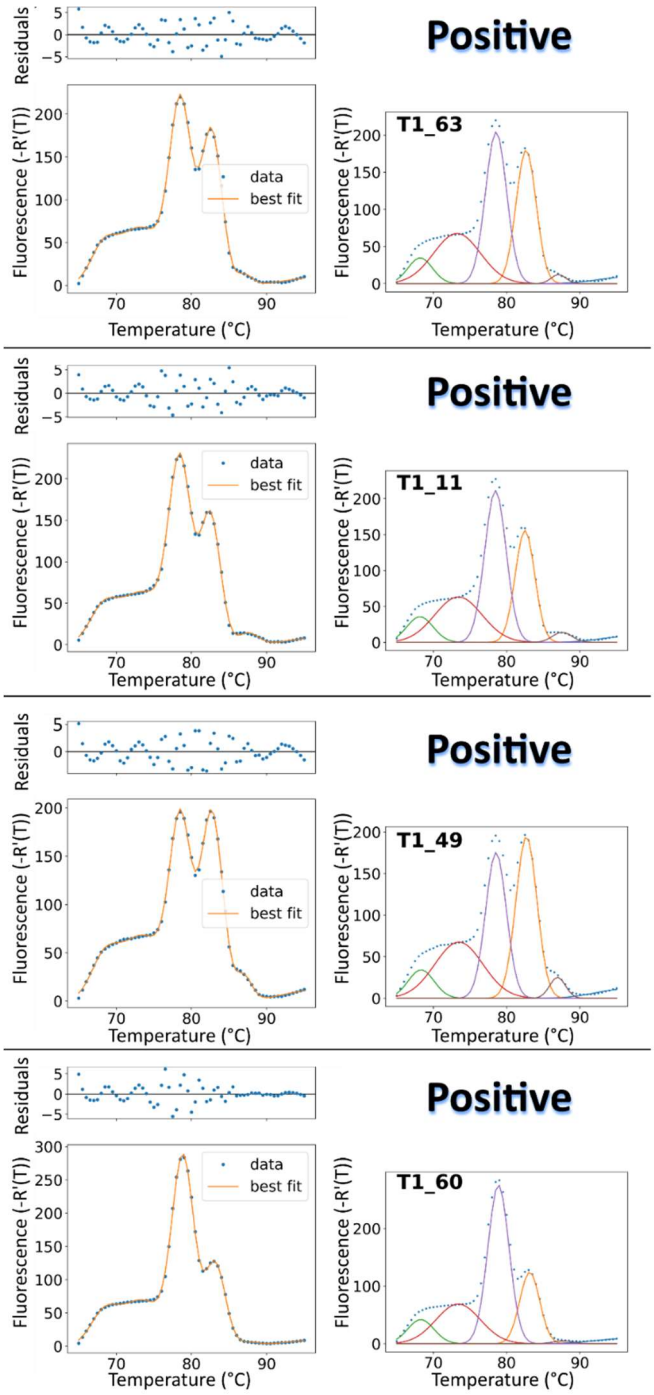
**Supplemental Table 3.1. Temporal Analysis of Survivorship Barcode Detection.** Overall, barcode detection exhibits a decreasing trend with age. However, the majority of mosquitoes remained barcode-positive out to 40 days post-emergence, suggesting an elevated likelihood of barcode persistence for the mosquito lifetime and subsequent recovery.

**Microcrystals do not affect adult mosquito survival or development**

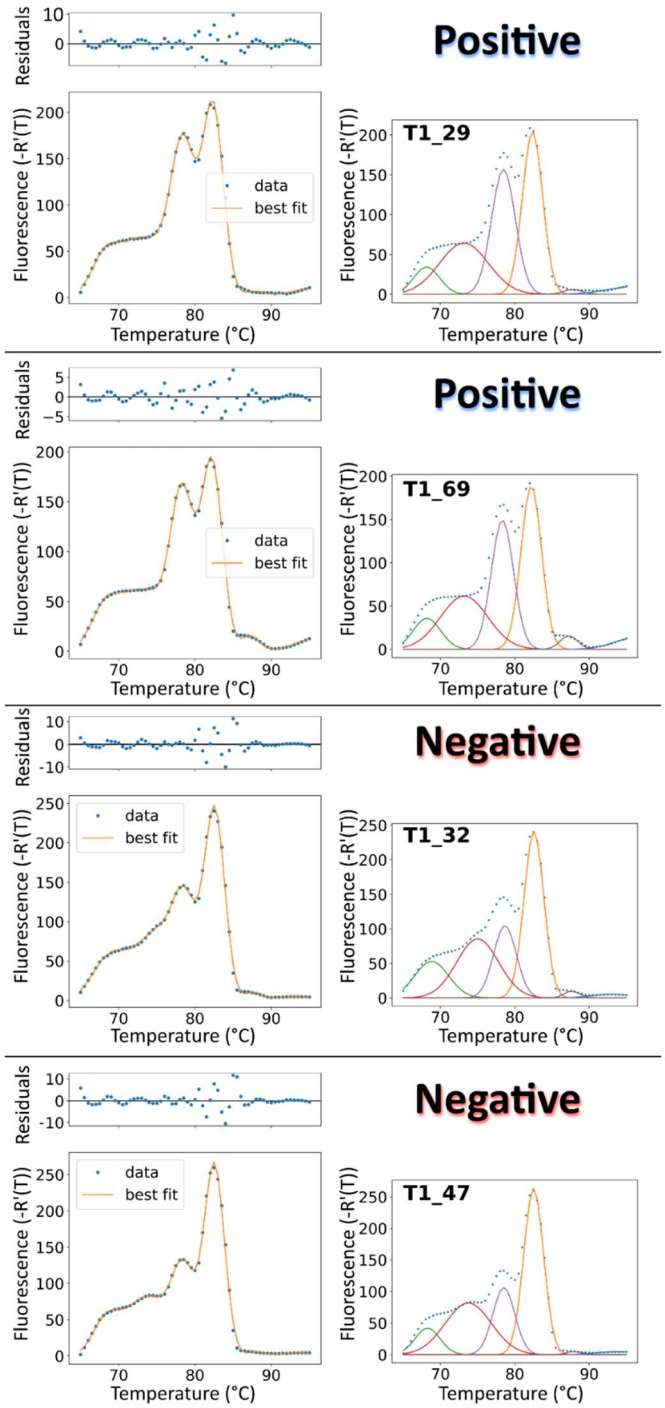
Larval *Culex tarsalis* mosquitoes reared on either liver powder alone (control) or liver powder with the addition of DNA barcoded microcrystals (treatment) were reared to adulthood and tracked individually for length of survivorship in days. Of the two replicates completed, the adult survivorship of mosquitoes was not significantly different between control and treatment groups. Crystal-fed mosquitoes across all replicates lived an average of 24 ± 13 days as adults (n=125), with the longest-lived mosquito reared on microcrystals surviving 67 days. By comparison, control mosquitoes survived an average of 26 ± 19 days as adults (n=177). One mosquito in the control group survived 196 days, a clear outlier; the next longest-lived control mosquito survived 69 days. qPCR of individual mosquitoes from the treatment group suggests an approximate 82 % barcode recovery rate, based on analysis (fig. S5) of melt curve data (n = 72). Preliminarily, a significant effect on larval survivorship was observed (p = 2.747e-05, Fisher's Exact Test) between larvae reared on microcrystals plus liver powder as opposed to those fed liver powder alone. However, this warrants further evaluation as only two replicates were available for this analysis and the effect of larval density was not considered but may have also affected larval survivorship.



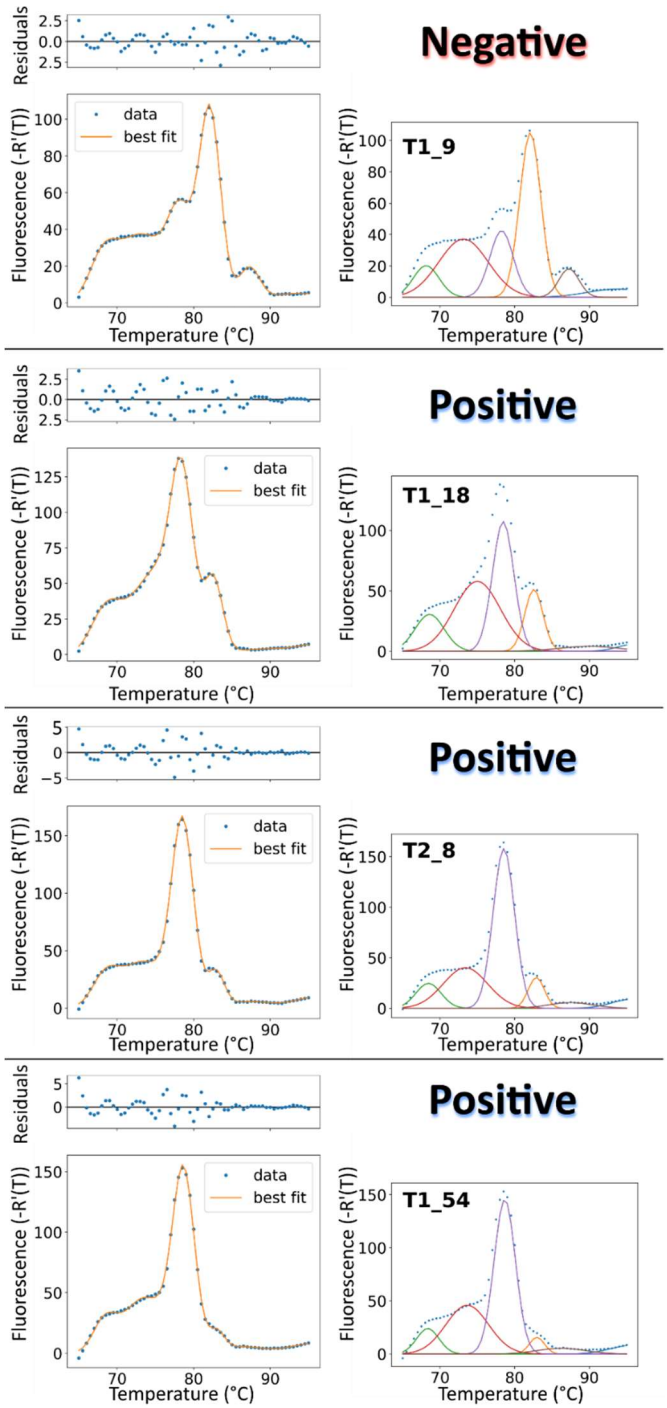

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (1/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



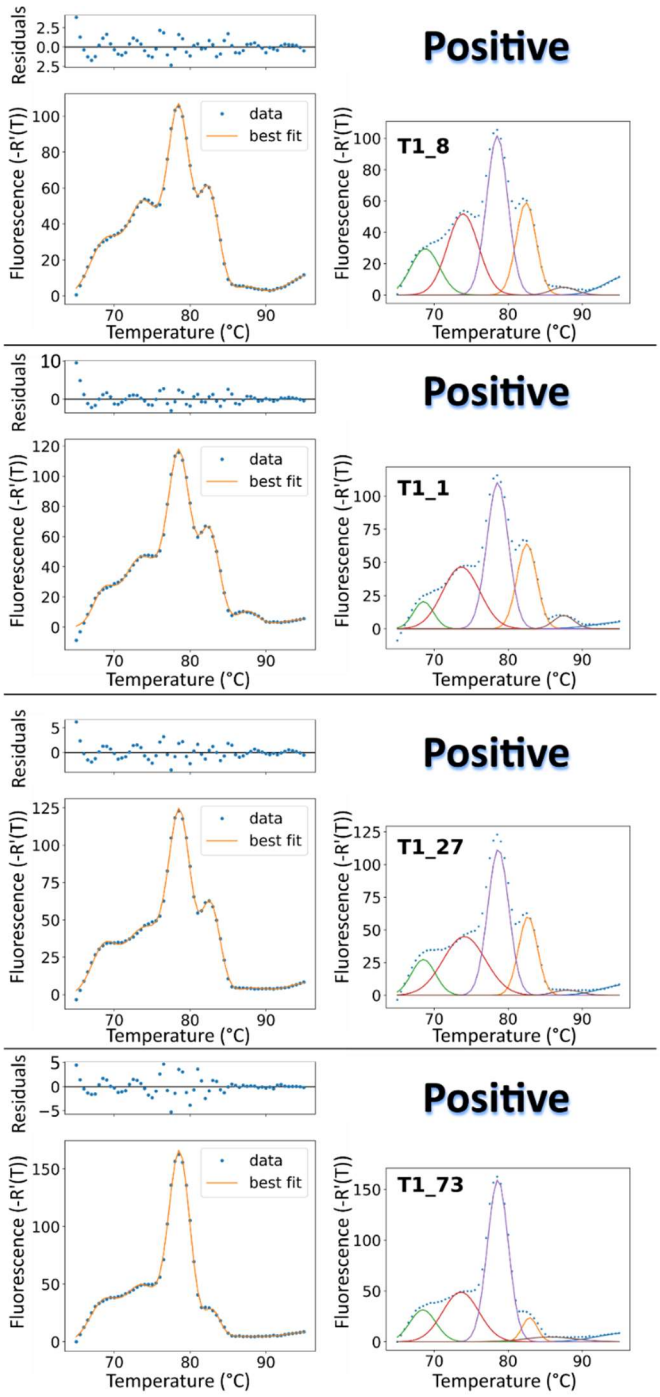

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (2/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



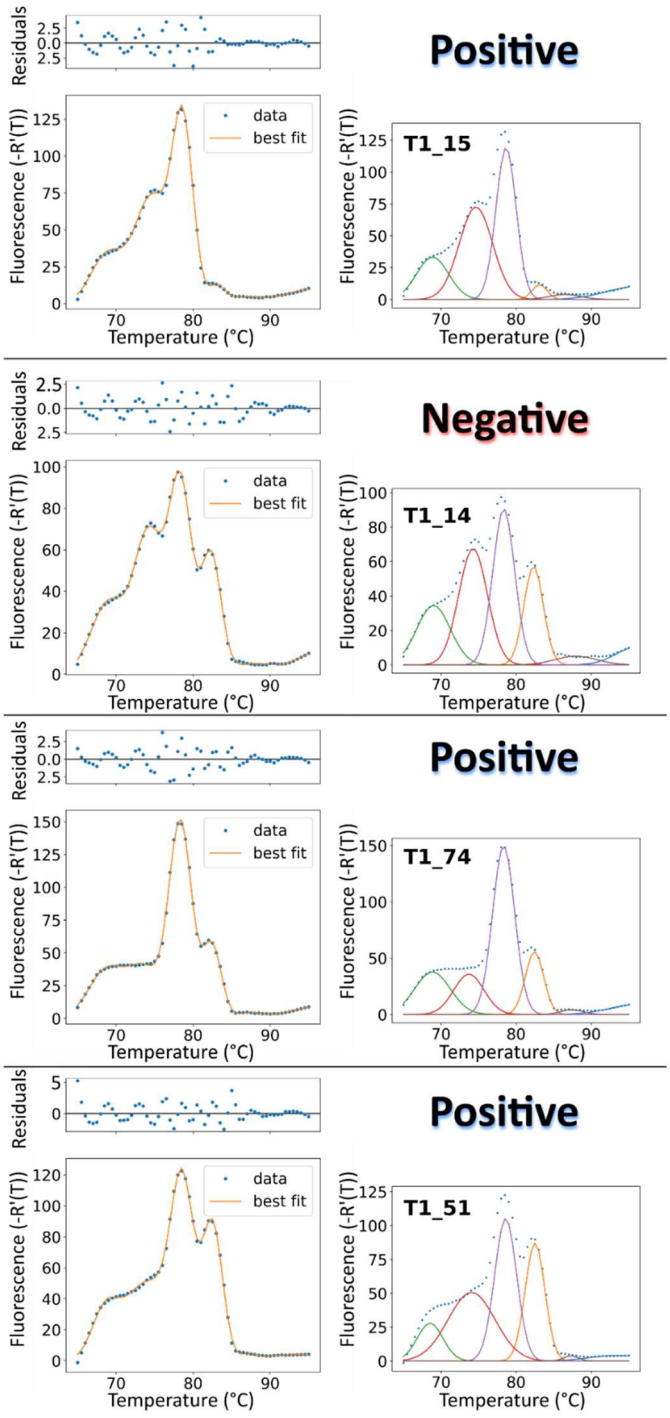

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (3/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



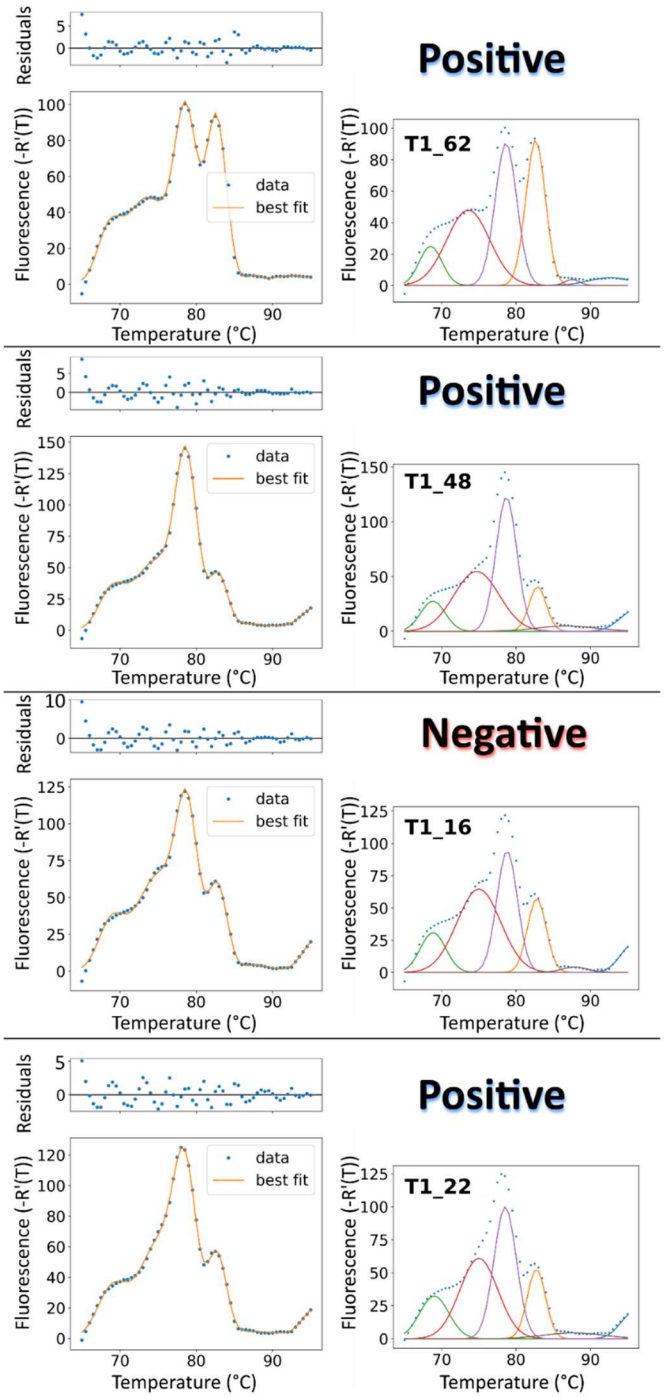

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (4/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



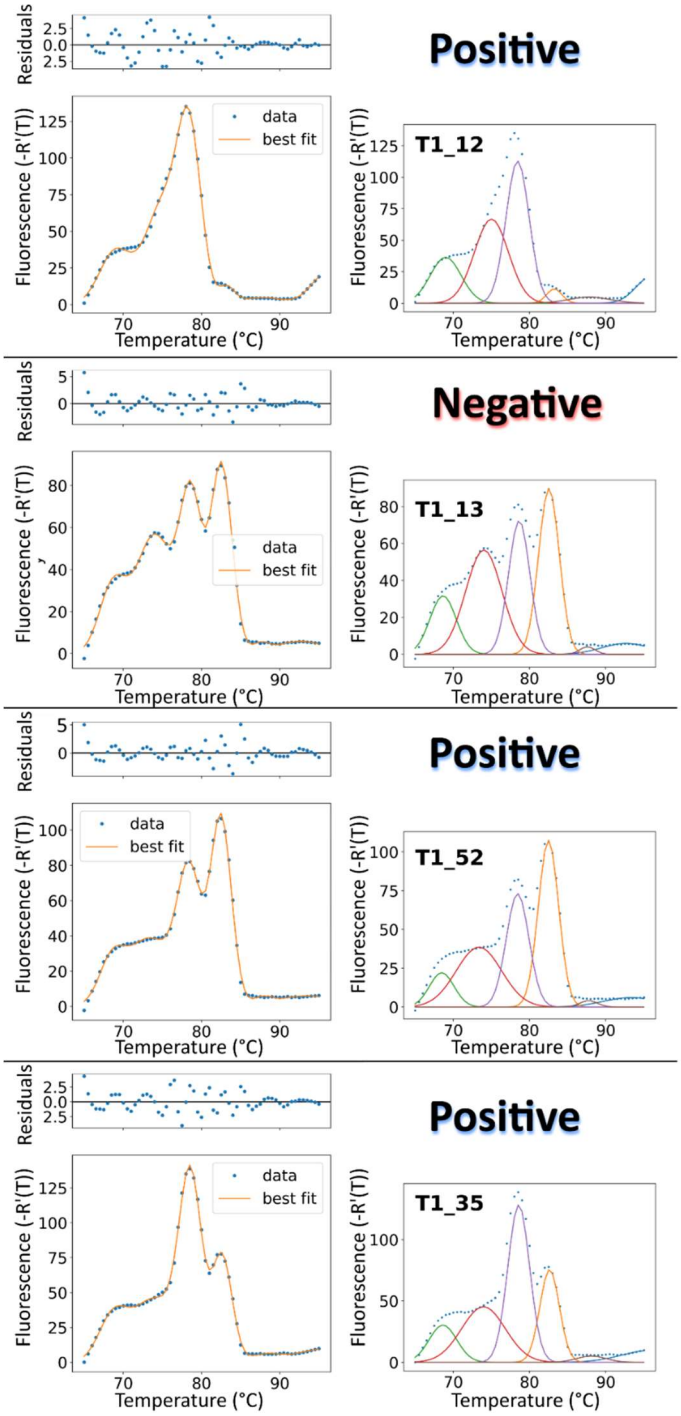

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (5/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



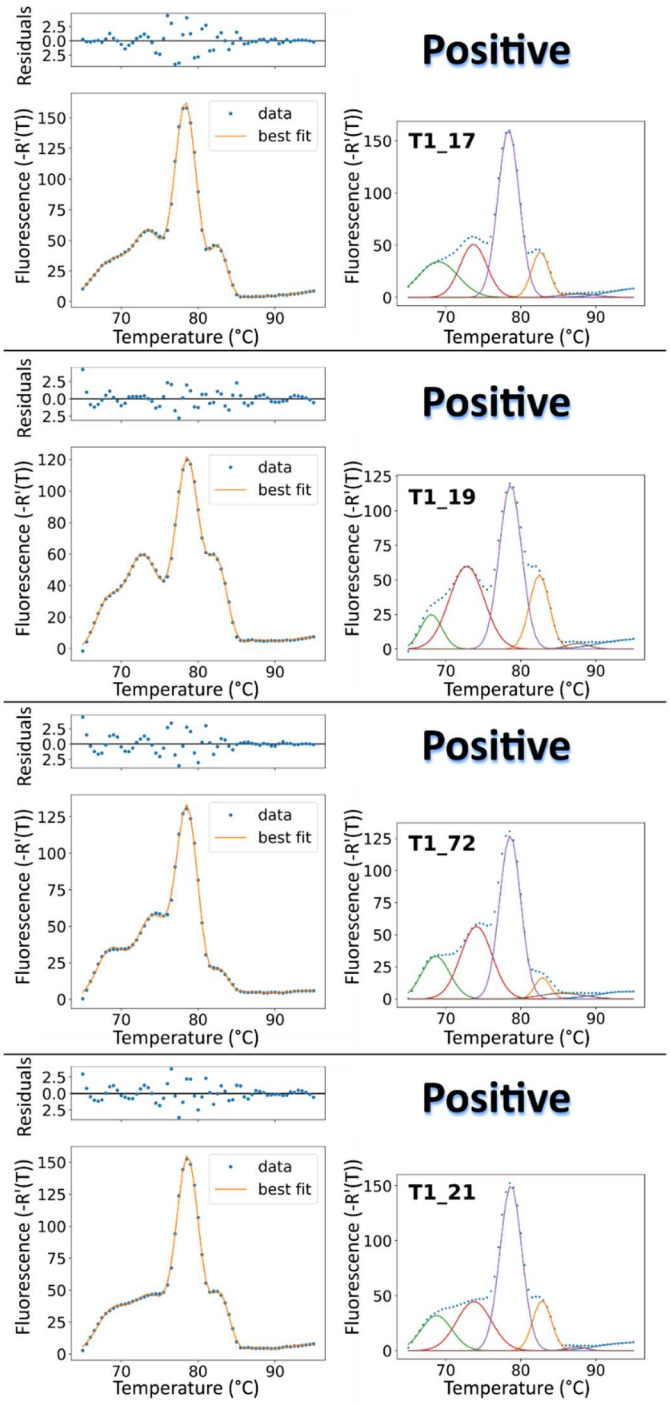

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (6/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



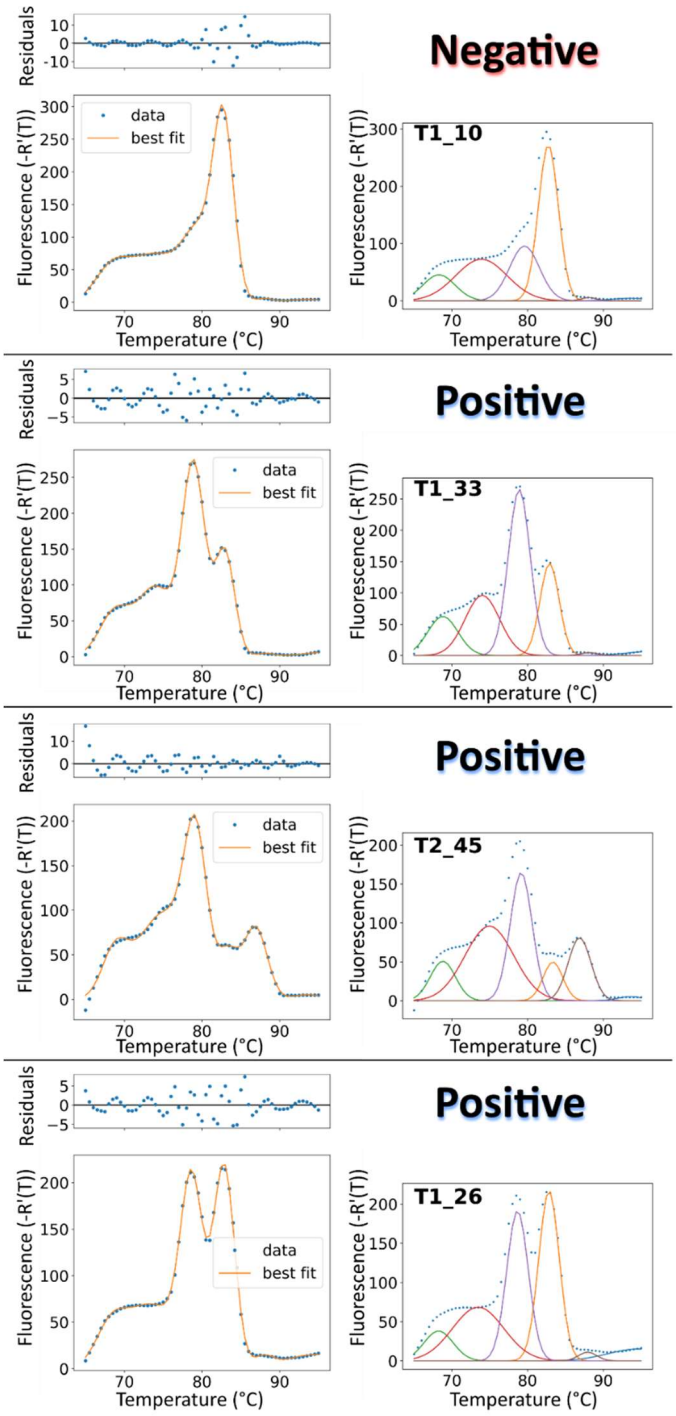

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (7/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



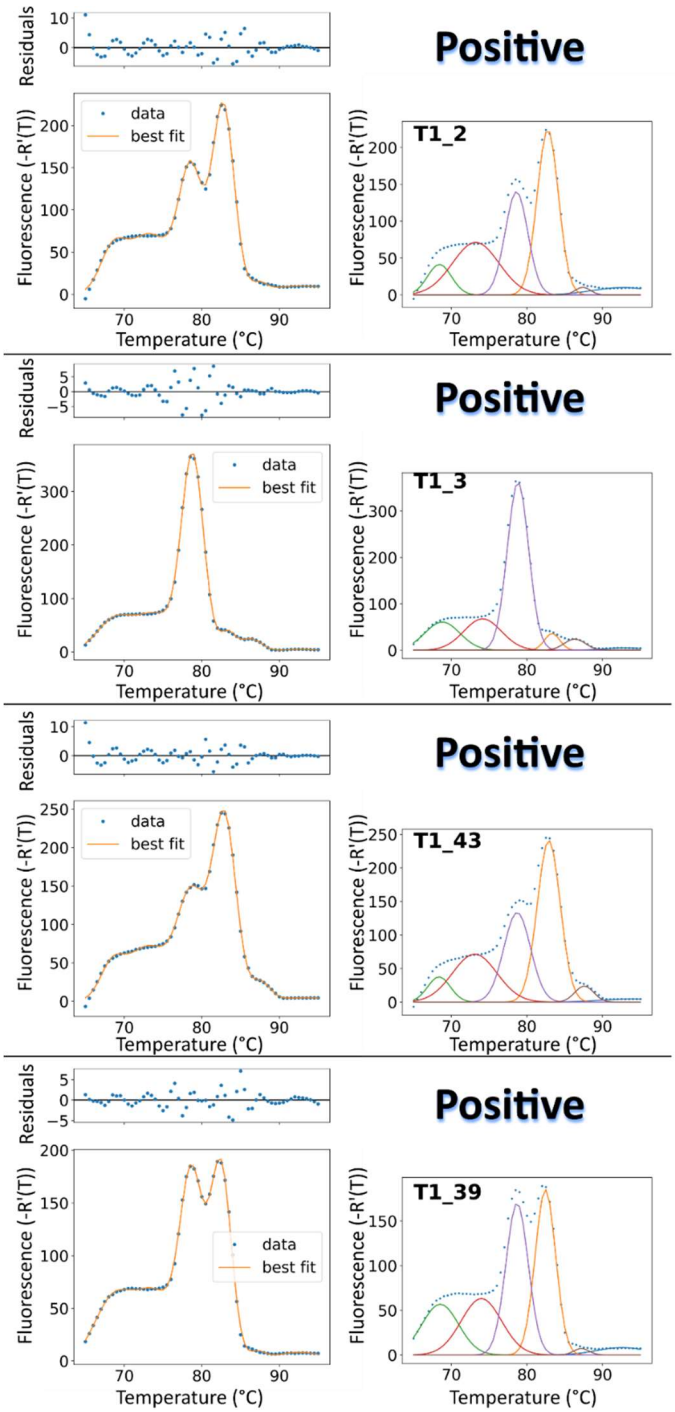

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (8/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



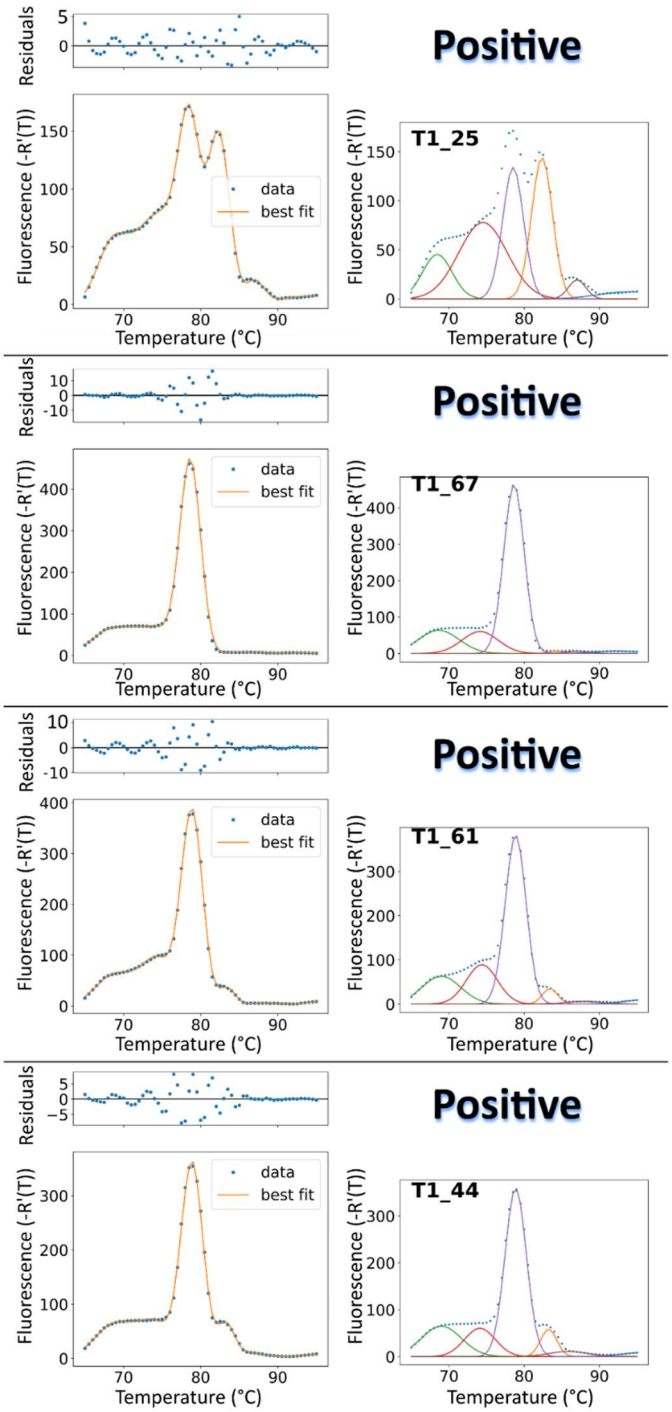

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (9/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



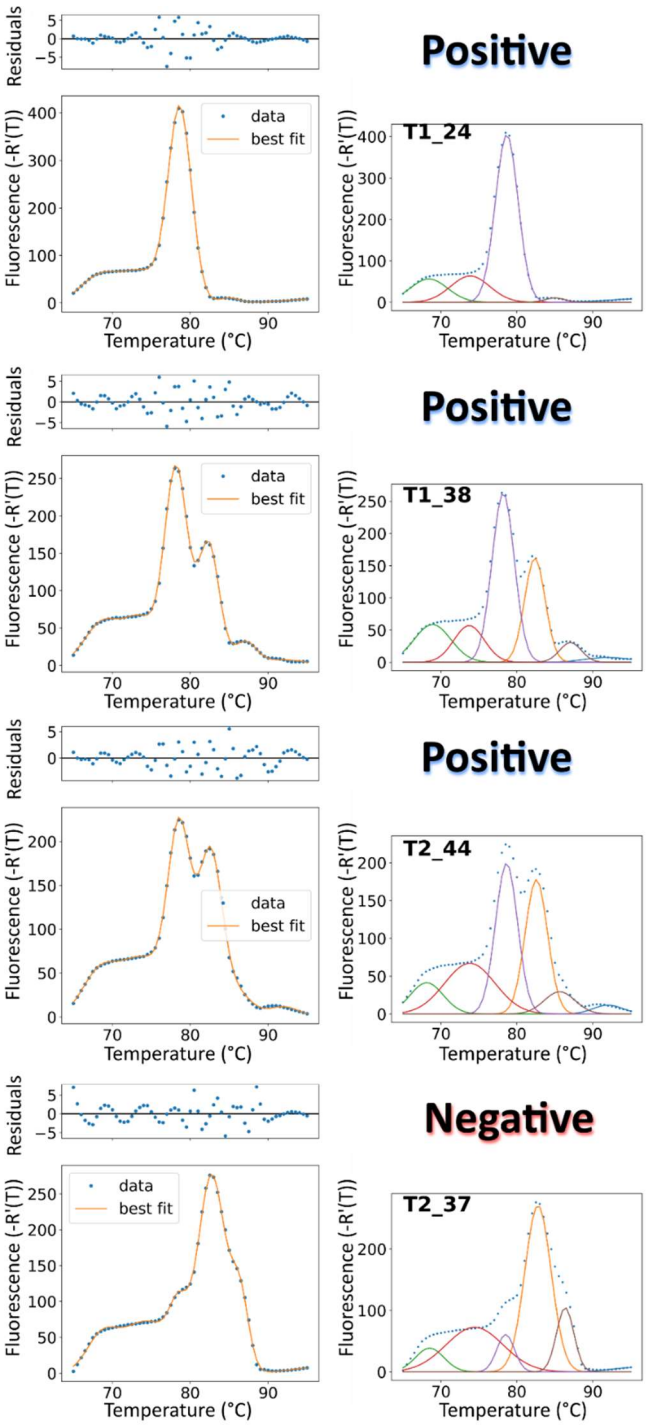

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (10/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



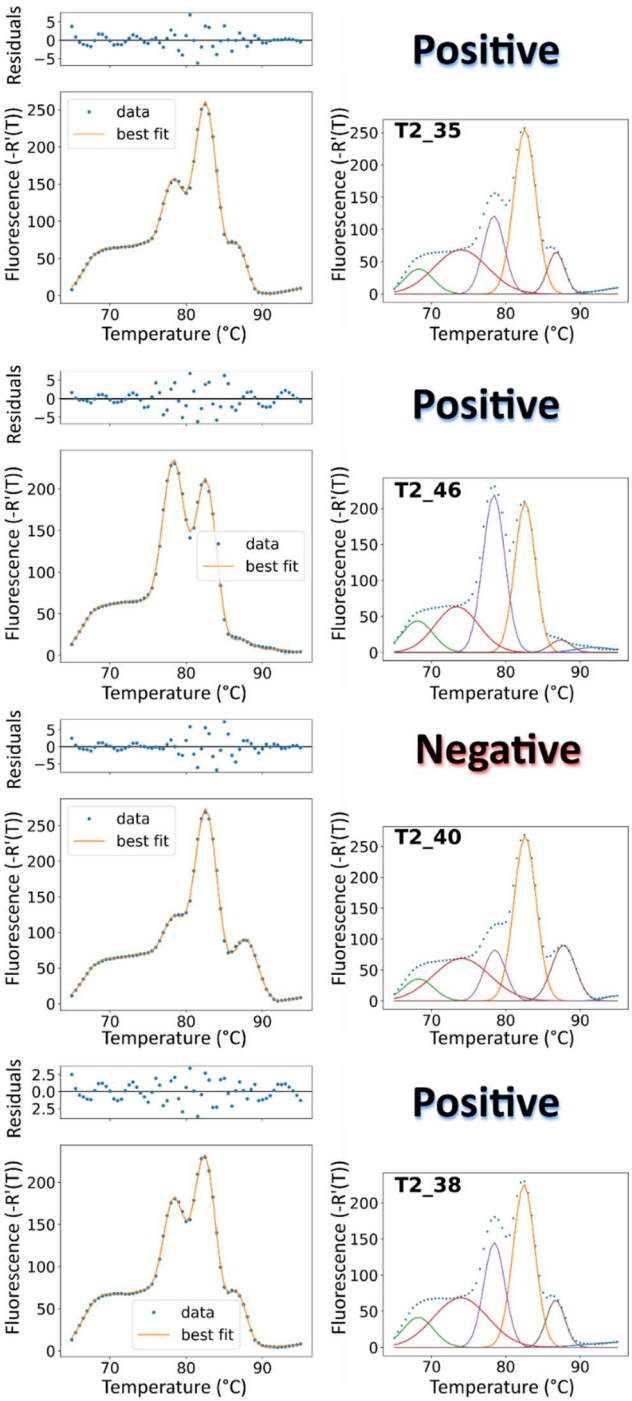

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (11/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



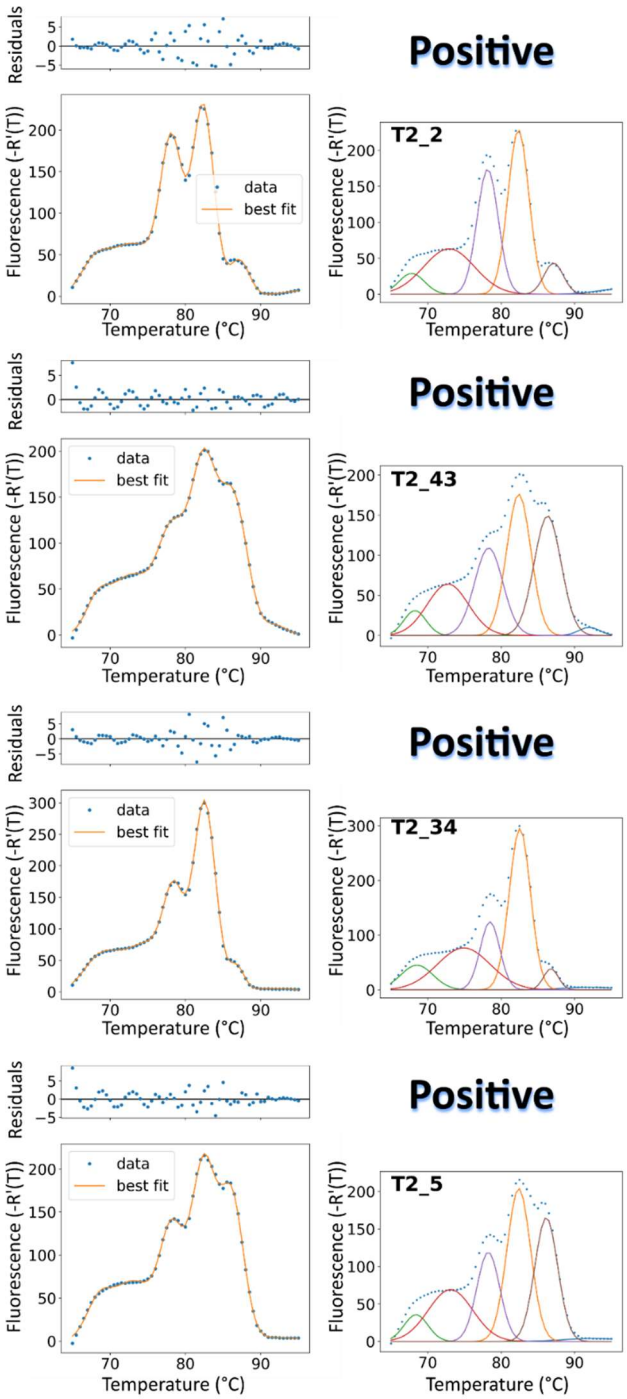

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (12/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



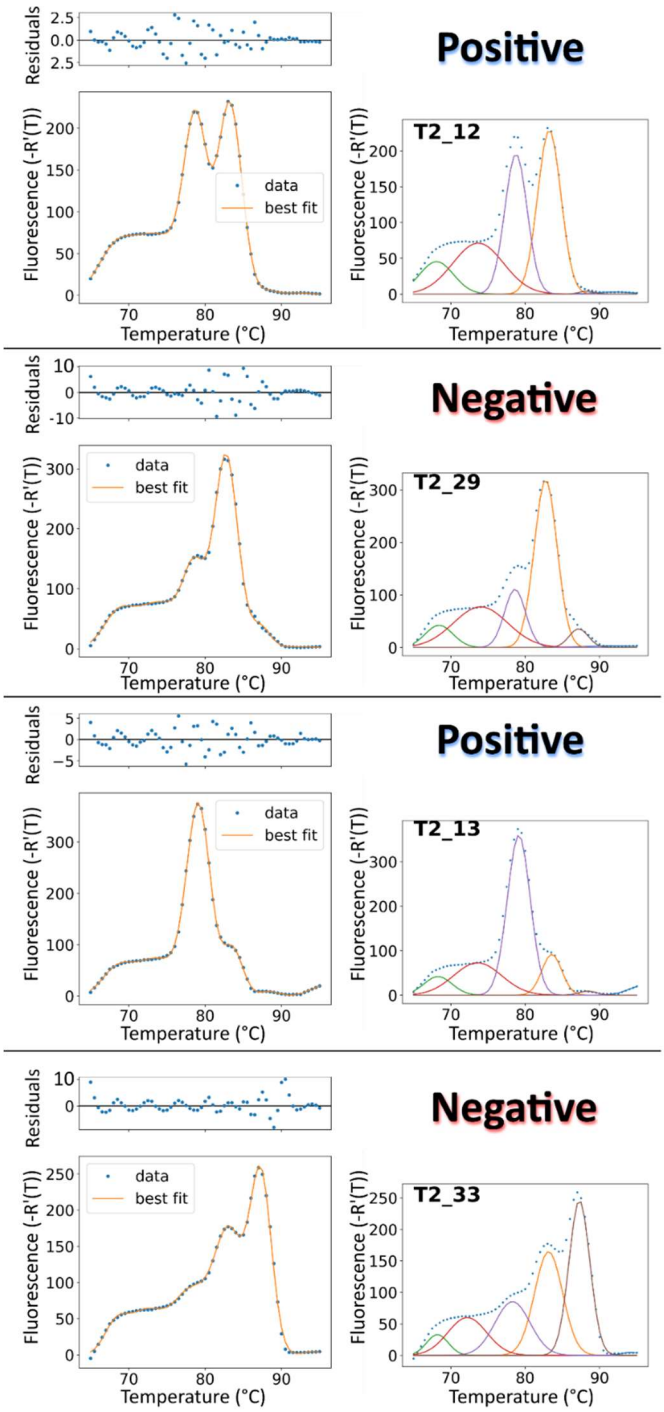

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (13/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



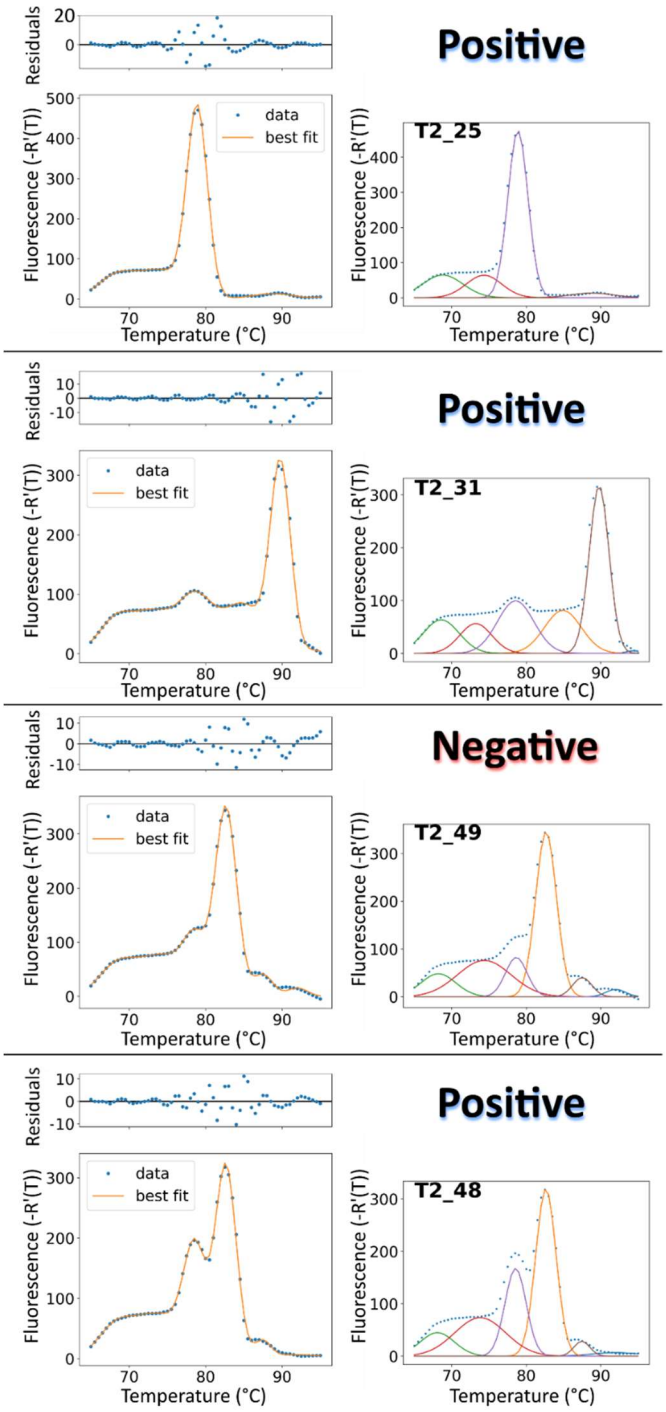

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (14/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



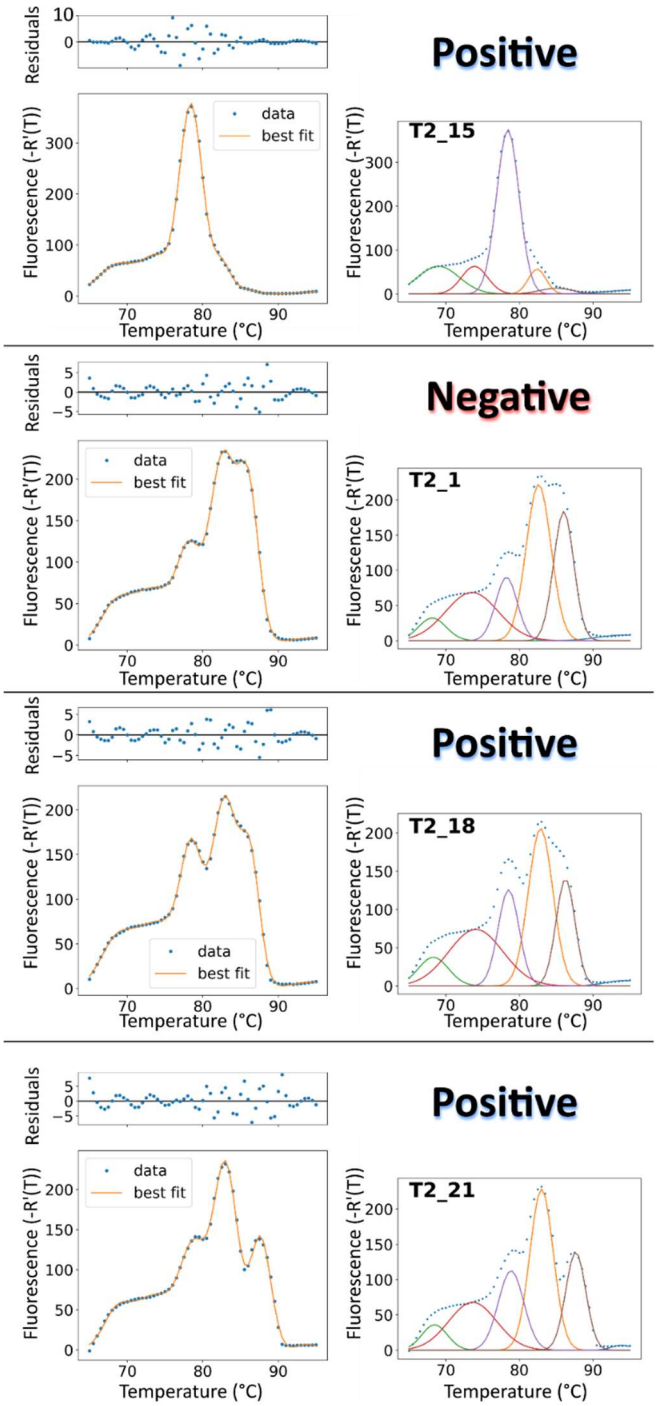

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (15/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



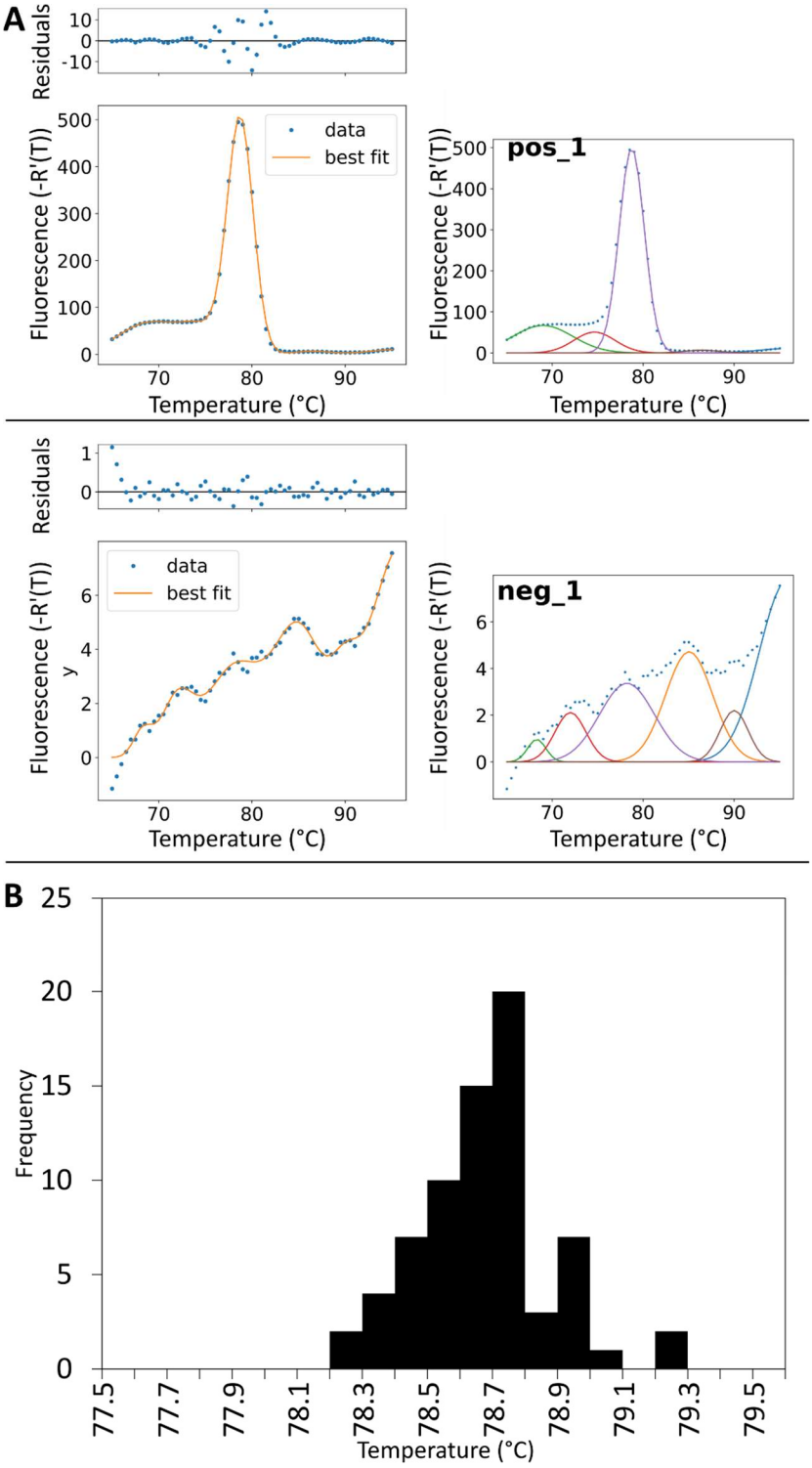

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (16/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



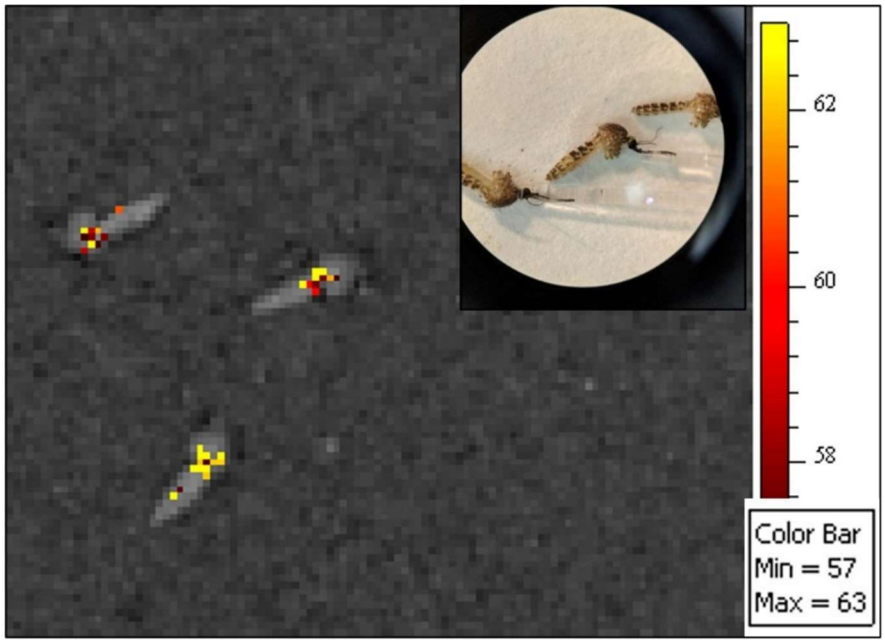

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (17/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



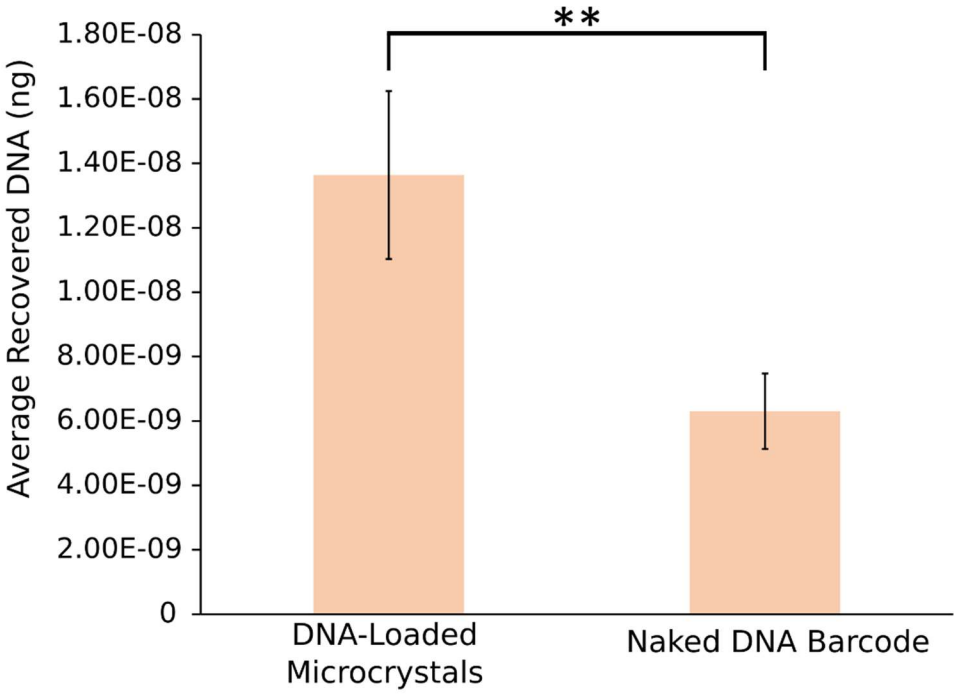

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (18/19).** Complete melt curves for individual mosquitoes from survivorship experiments. T1 and T2 denote replicate group 2 and 3, respectively. Melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



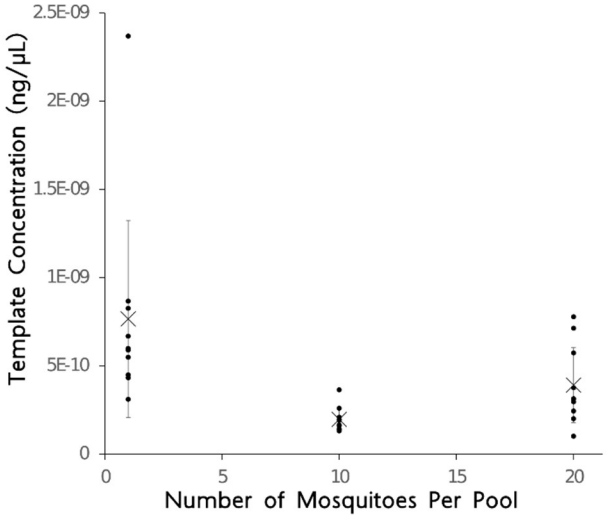

**Supplemental Figure 3.5. Melt curves and scoring results for qPCR of mosquito extractions from survivorship experiments (19/19). (A)** Complete melt curves for a positive control (top) and negative control (bottom) overlaid with LMFIT(91) results.
**(B)** Histogram of fitted barcode peak centers for all 72 samples revealing a narrow distribution for identified barcode peaks.




**Supplemental Figure 3.6. Ingestion of fluorescein-labeled crystals by adult *Culex tarsalis* mosquitoes.** Mosquitoes were immobilized by removal of legs and wings. The proboscis of each mosquito was inserted into a capillary tube containing microcrystals in sugar solution. After approximately 30 minutes of feeding, mosquitoes were live-imaged using an In Vivo Imaging System (IVIS). A fluorescent signal is present inside the mosquito body indicating that crystals were ingested.

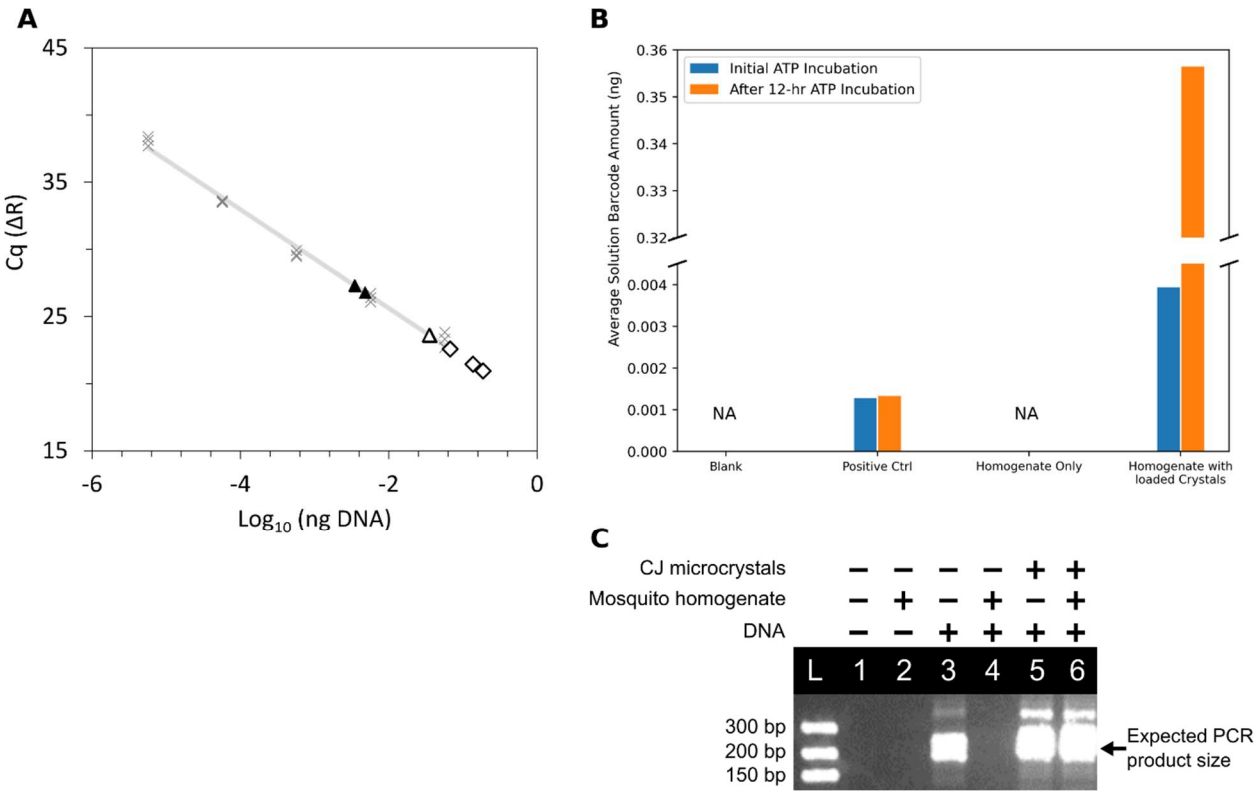



**Supplemental Figure 3.7. Environmental Persistence.** Bar chart showing elevated DNA barcode recovery from mosquitoes fed DNA-loaded microcrystals mixed with liver powder (n = 36) relative to mosquitoes fed naked DNA barcode in liver powder solution (n = 45). Mosquito larvae were fed 6.5 ng of either naked barcode or barcoded microcrystals daily during the 2nd, 3rd, and 4th instar life stages. Adult mosquitoes were harvested upon emergence for DNA extraction and barcode detection as described in the Materials and Methods. Error bars represent one standard deviation (p = 0.0075).

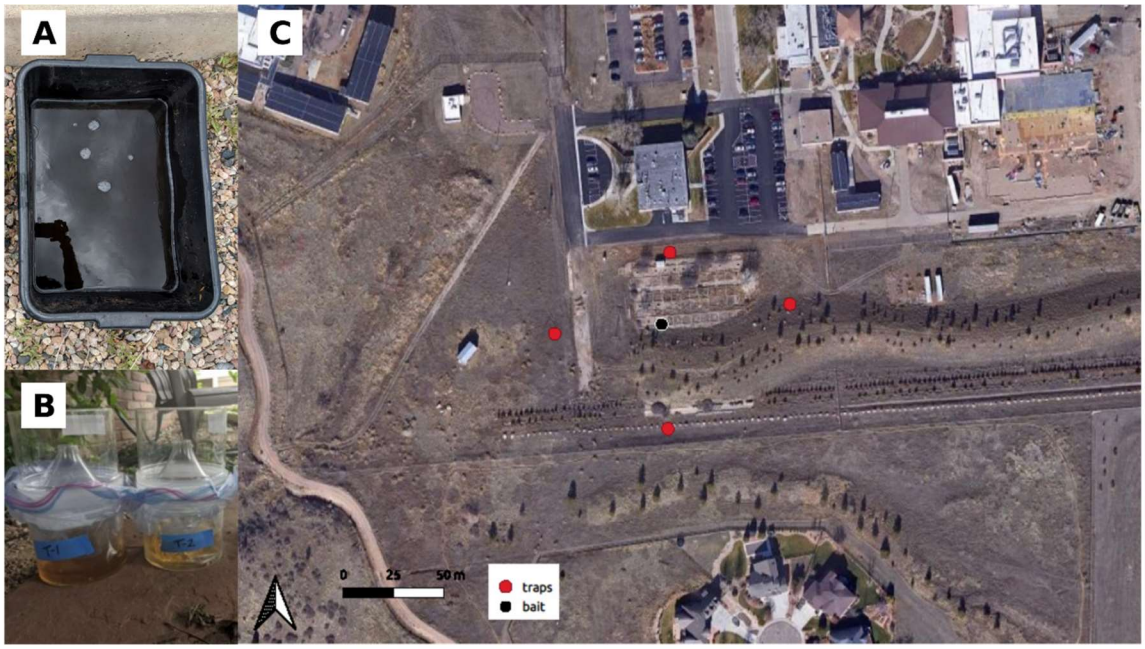



**Supplemental Figure 3.8. Barcode Detection Sensitivity.** The recovered barcode amount (y-axis) plotted as a function of mosquito pool size (x-axis) demonstrates barcode DNA remains quantifiable at pools up to 20, proving detection sensitivity remains preserved despite greater pool sizes.

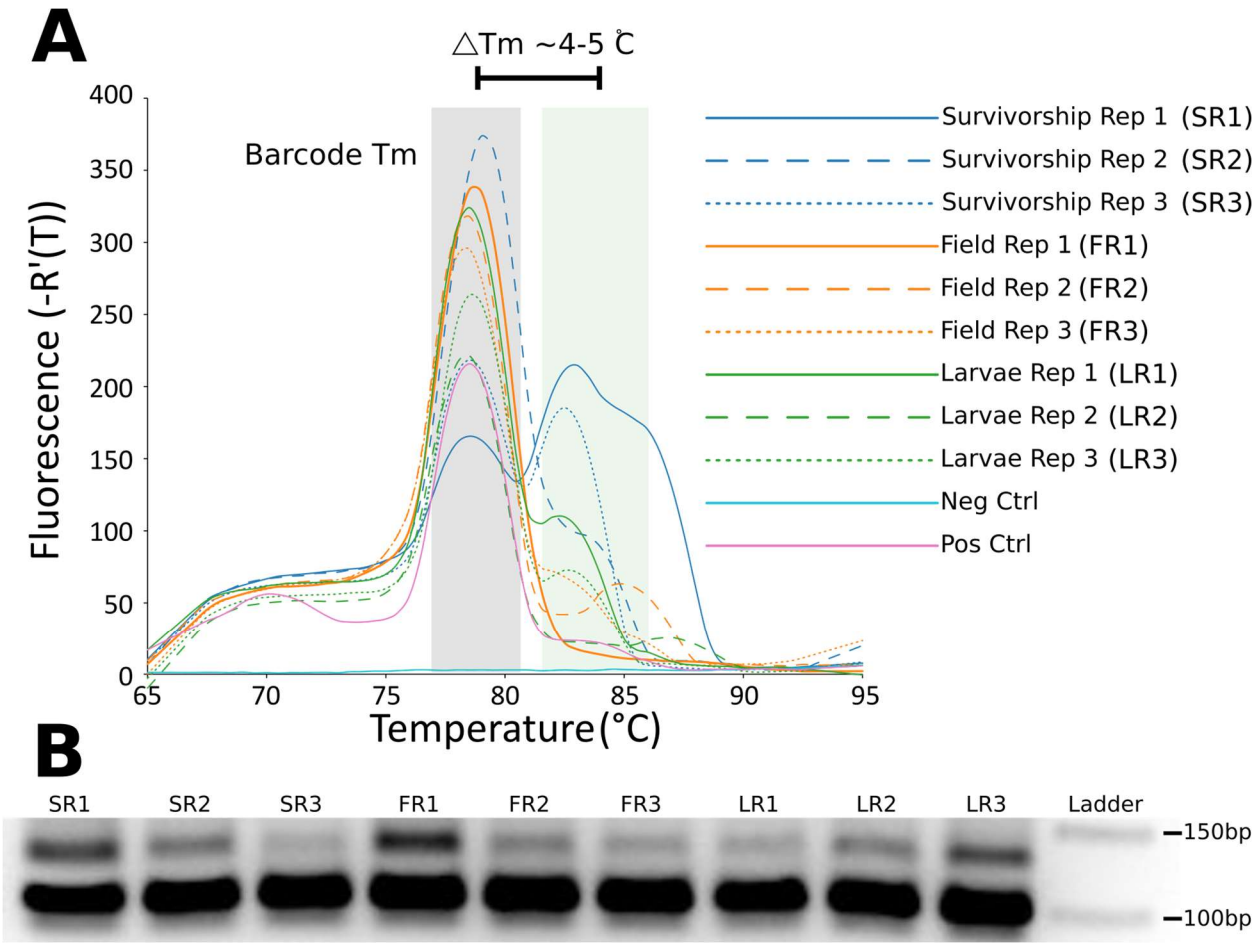
To determine the sensitivity of DNA detection as it relates to mosquito pool size, one crystal-fed mosquito was placed into pools of non-crystal-fed mosquitoes at increasing ratios. For the initial experiment, pool sizes of 1, 10 and 20 mosquitoes were assessed using 10 replicates per treatment (fig. S8). Observed melt curve variations, corresponding to samples SR1-3 in fig. 4, may result from variation in sample homogenization, although larger quantities of mosquito genomic DNA may be a contributing factor. Regardless, these data demonstrate that detection sensitivity of a single barcoded mosquito would not be compromised by pool size, for pools of up to 20 mosquitoes.



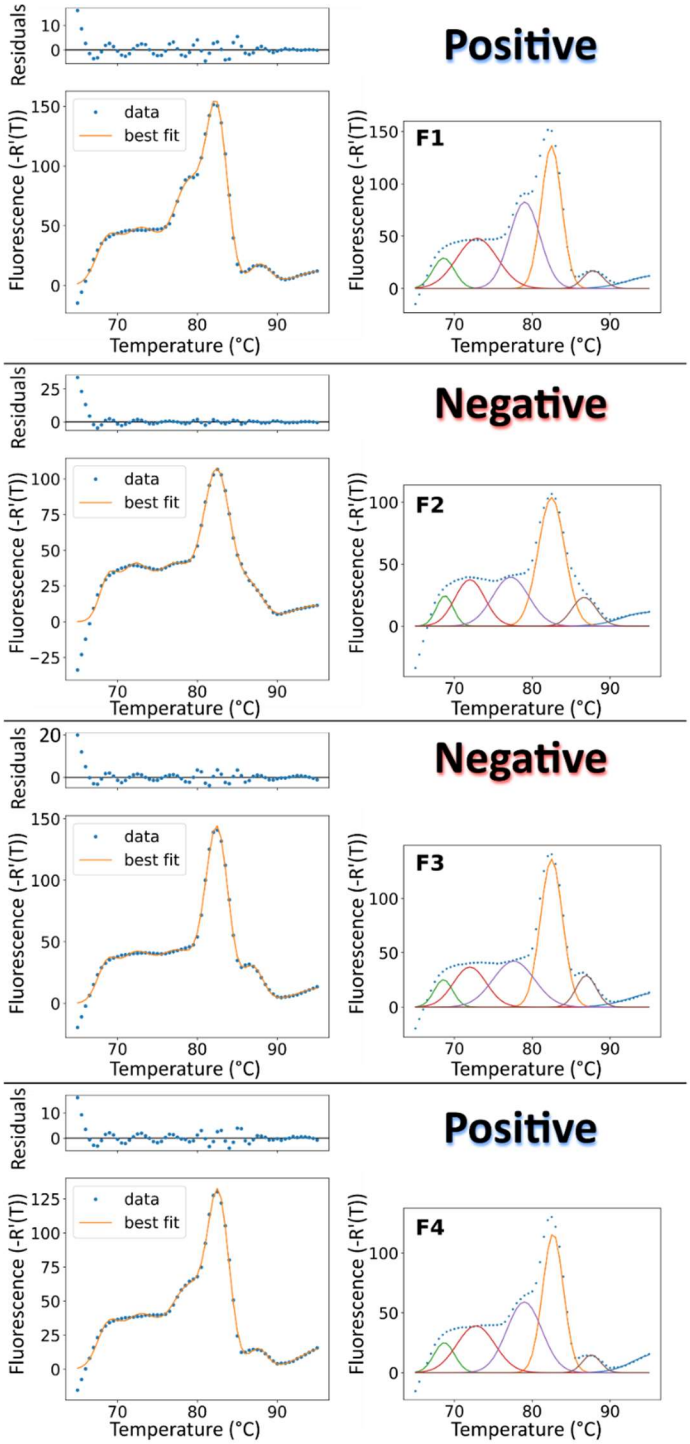

**Supplemental Figure 3.9. qPCR DNA Barcode Recovery Following Mosquito Homogenate Incubation and Crystal Protection. A)** Standards plotted (x) with linear fit (gray line, R2 = 0.99). Triplicate measurements of the initial mosquito homogenate and crystal mixture demonstrated detectable barcode DNA in solution (black triangles). Incubation (12hr) with 20 mM ATP resulted in elevated barcode recovery (empty diamonds), exceeding the standard curve. A 10-fold dilution of the crystal/DNA/homogenate/ATP replicates brought the solution DNA concentration within a quantifiable range (empty triangles). **B)** Comparison of solution DNA Barcode amount upon initial ATP addition and after 12-hr incubation. No Cq values were detected for either the TE buffer blank or the homogenate only negative control. The solution DNA amount for the positive control (125mer in solution) remains virtually unchanged throughout the incubation period. Notably, the solution DNA amount for the crystal/homogenate mixture increases by over 2-orders of magnitude following ATP incubation. No Cq value was detected when using the 20 mM ATP as an additional negative control template (Data not shown). **C)** Barcode in solution was not detected following incubation with mosquito homogenate (lane 4) but was present in the sample that contained DNA-loaded microcrystals (lane 6) suggesting microcrystals protect DNA barcodes from degradation. Barcode detection was absent in the negative control samples, nuclease-free water (lane 1) and mosquito homogenate (lane 2). The expected PCR product was observed for both positive control samples, 200mer in solution (lane 3) and 200mer loaded in microcrystals (lane 5).



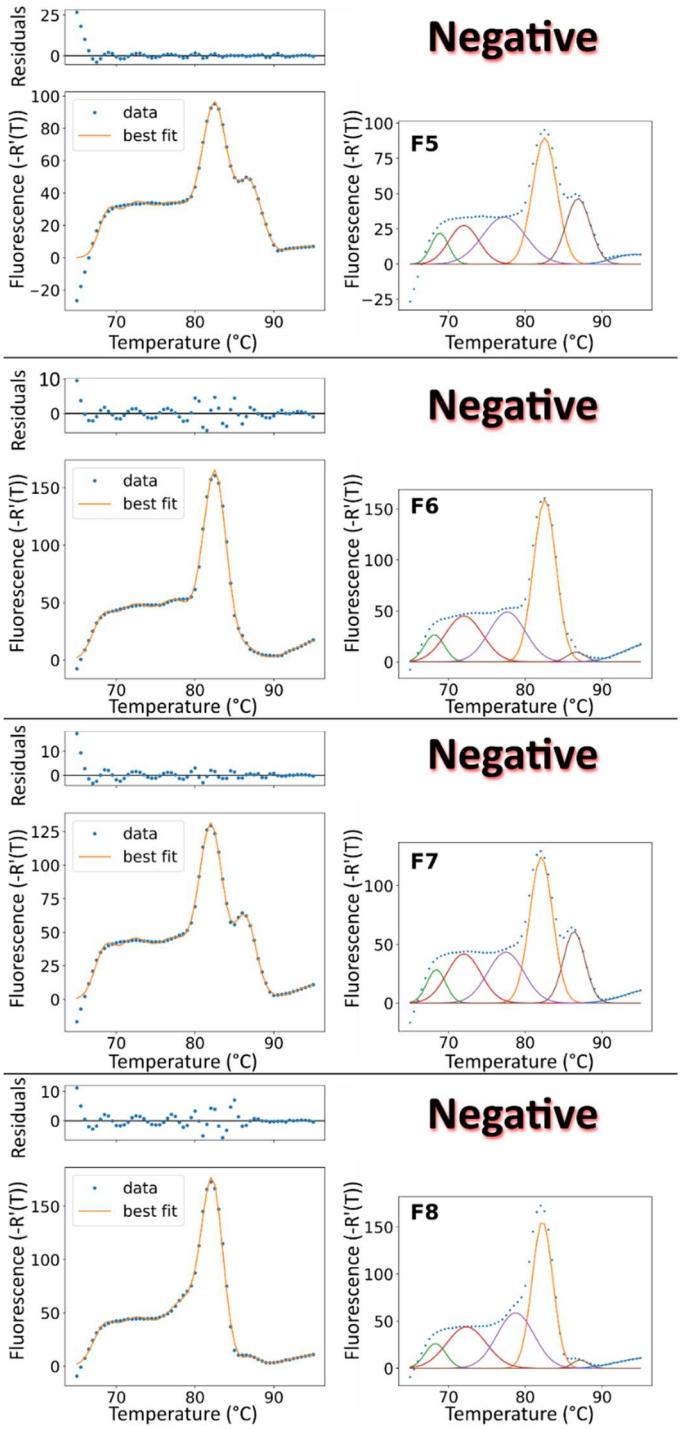

**Supplemental Figure 3.10.** Pilot field test on the CSU Foothills Campus during the summer of 2020. **A)** Standardized concentrations of microcrystals were deployed into 18Q Sterilite wash basins, which also served to contain microcrystals within the environment. **B)** Mosquito rearing containers into which representative fourth instar larvae and pupae were picked from the wash basins and reared to adult mosquitoes in the insectary and assayed for the presence of barcode. **C)** Location of the pilot study on the CSU Foothills Campus. Red circles denote the locations of CDC light traps. Black circle denotes the location of the wash basin spiked with barcoded microcrystals. Scale bar denotes 50 meters. Microcrystal field deployments have been approved by the CSU Institutional Biosafety Committee (19-032B) which includes a local environmental risk assessment.



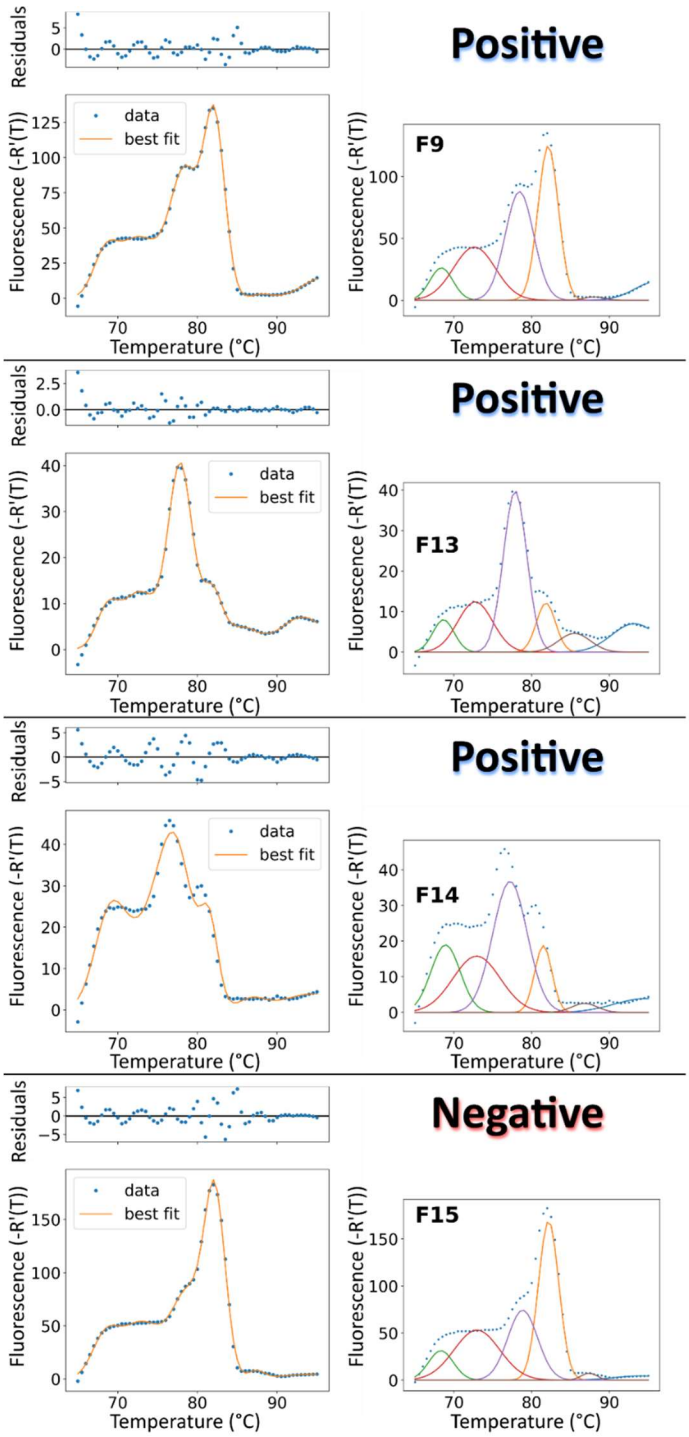

**Supplemental Figure 3.11. Barcode Detection for All Replicates. (A)** qPCR melt curves for the three replicates from survivorship (SR1-3), field (FR1-3), and larvae (LR1-3) studies. **(B)** Gel electrophoresis results of all samples shown in (A) displaying the target 84-mer barcode band in addition to a fainter slightly higher band.



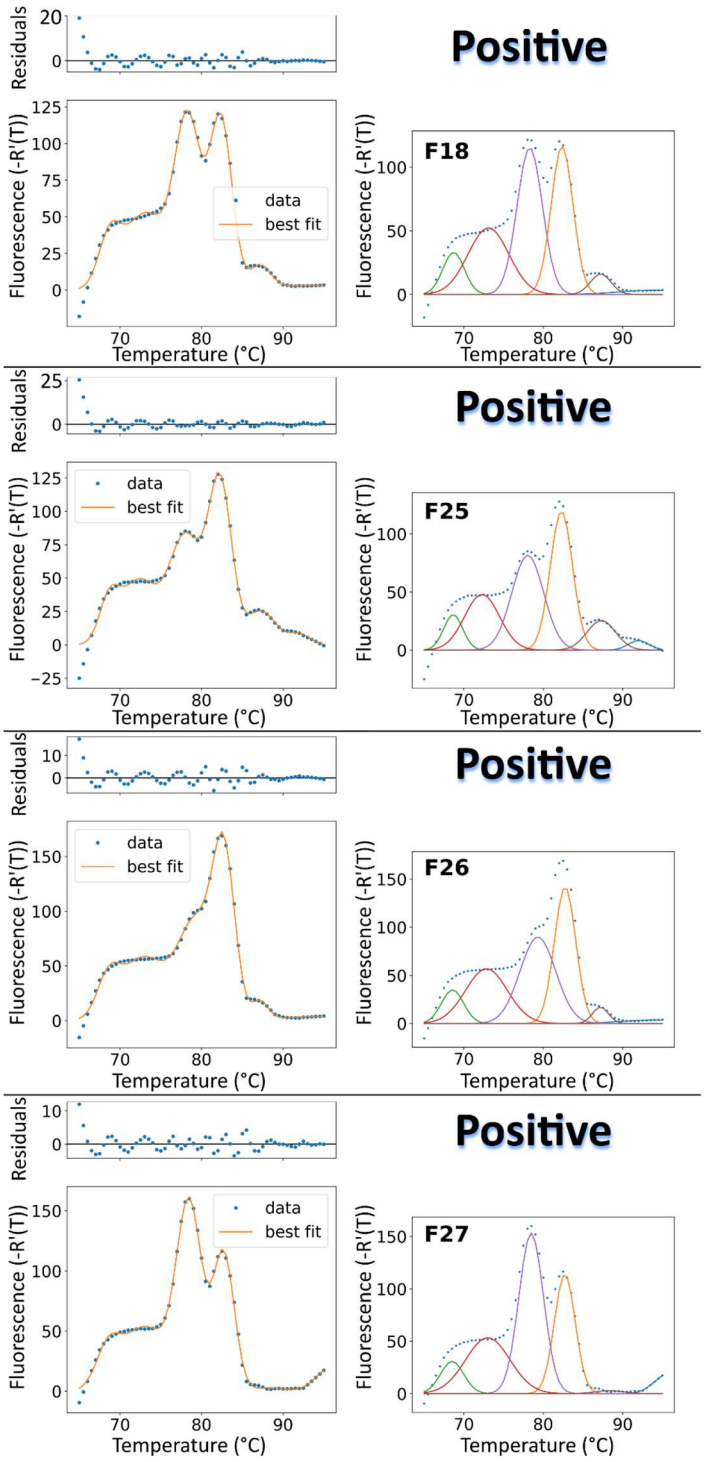

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (1/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



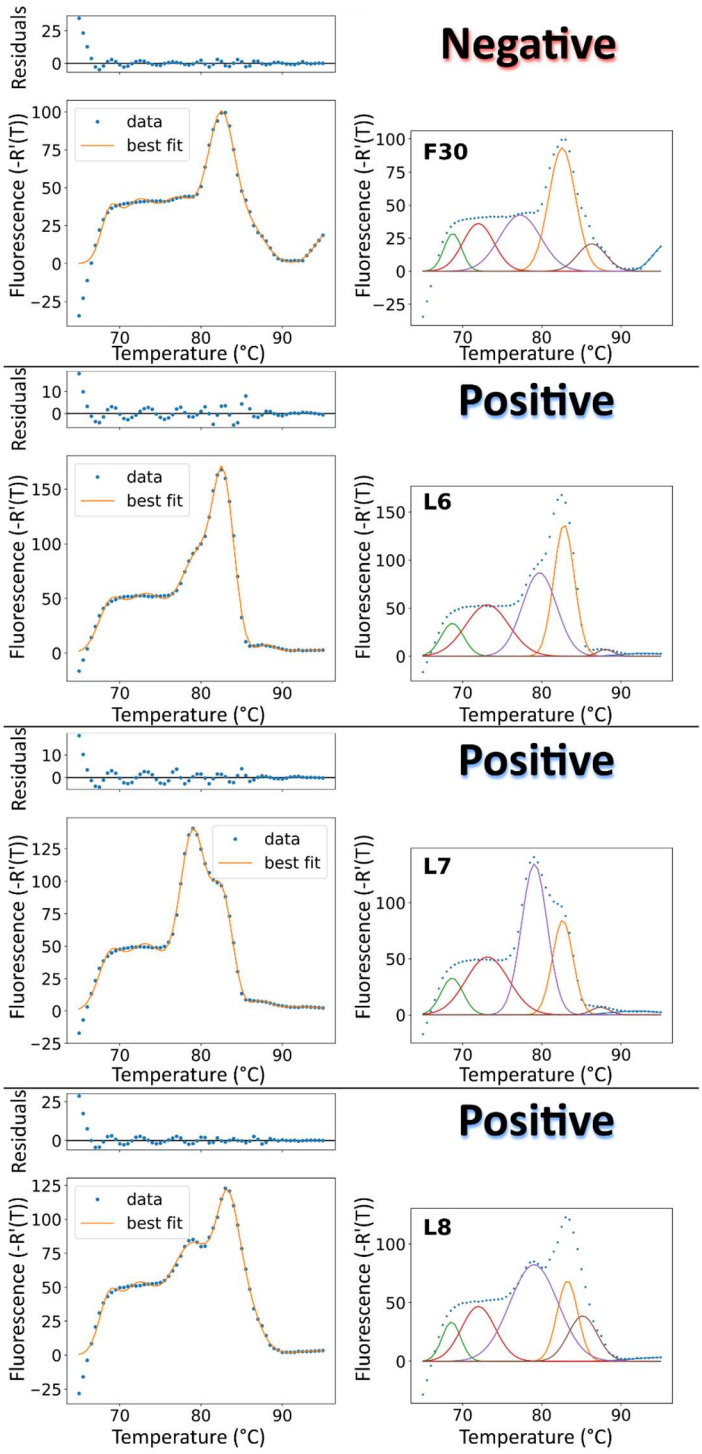

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (2/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



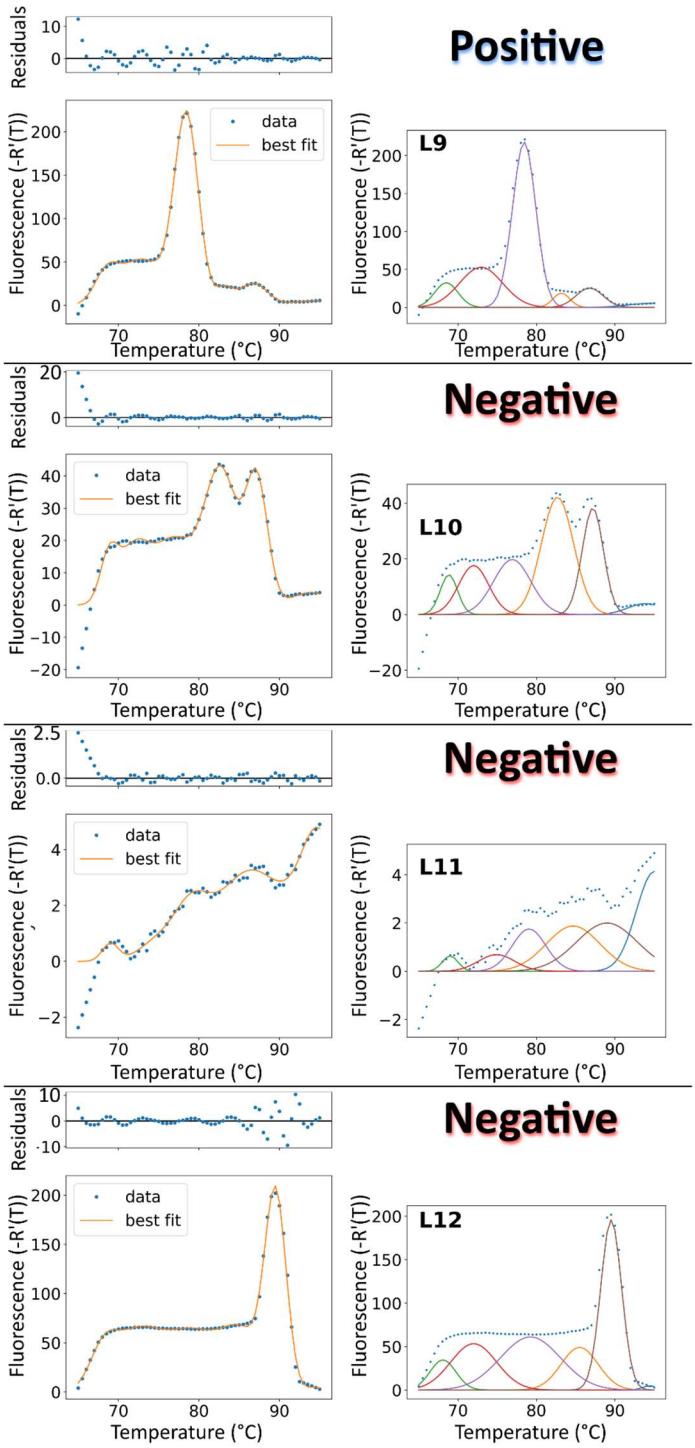

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (3/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



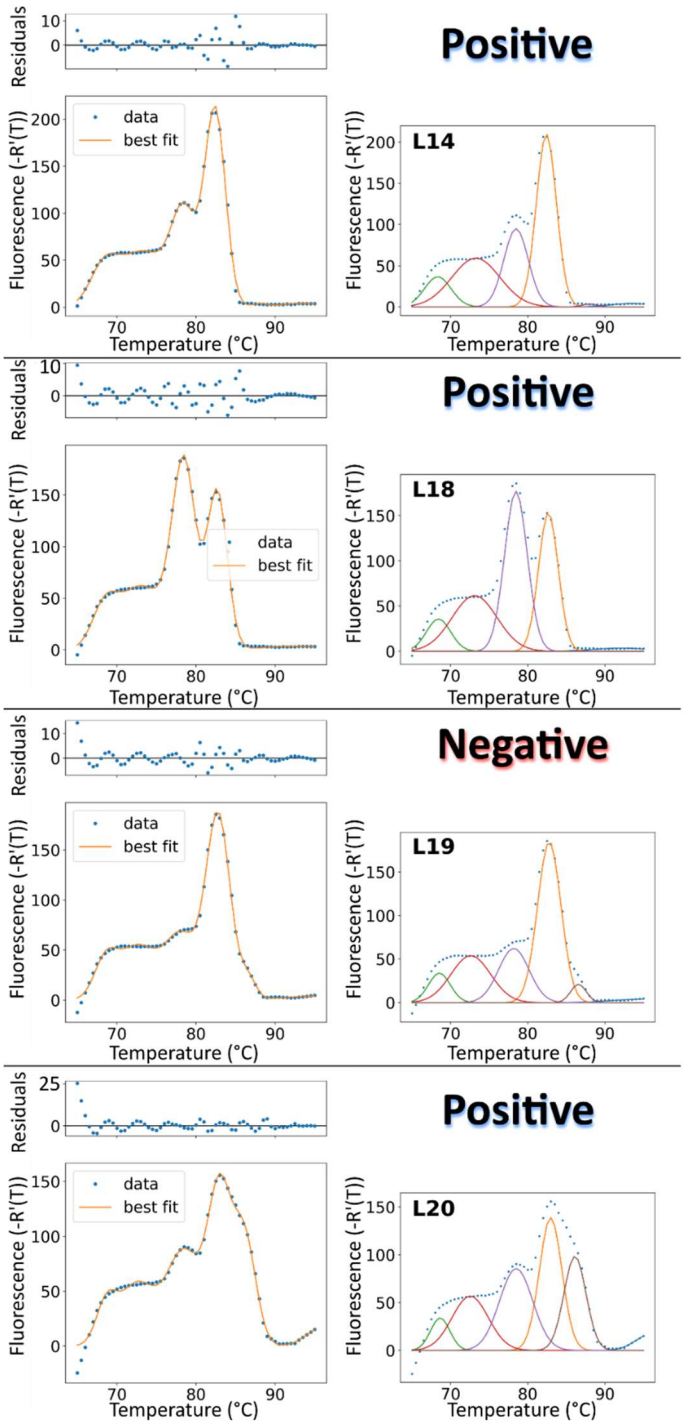

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (4/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



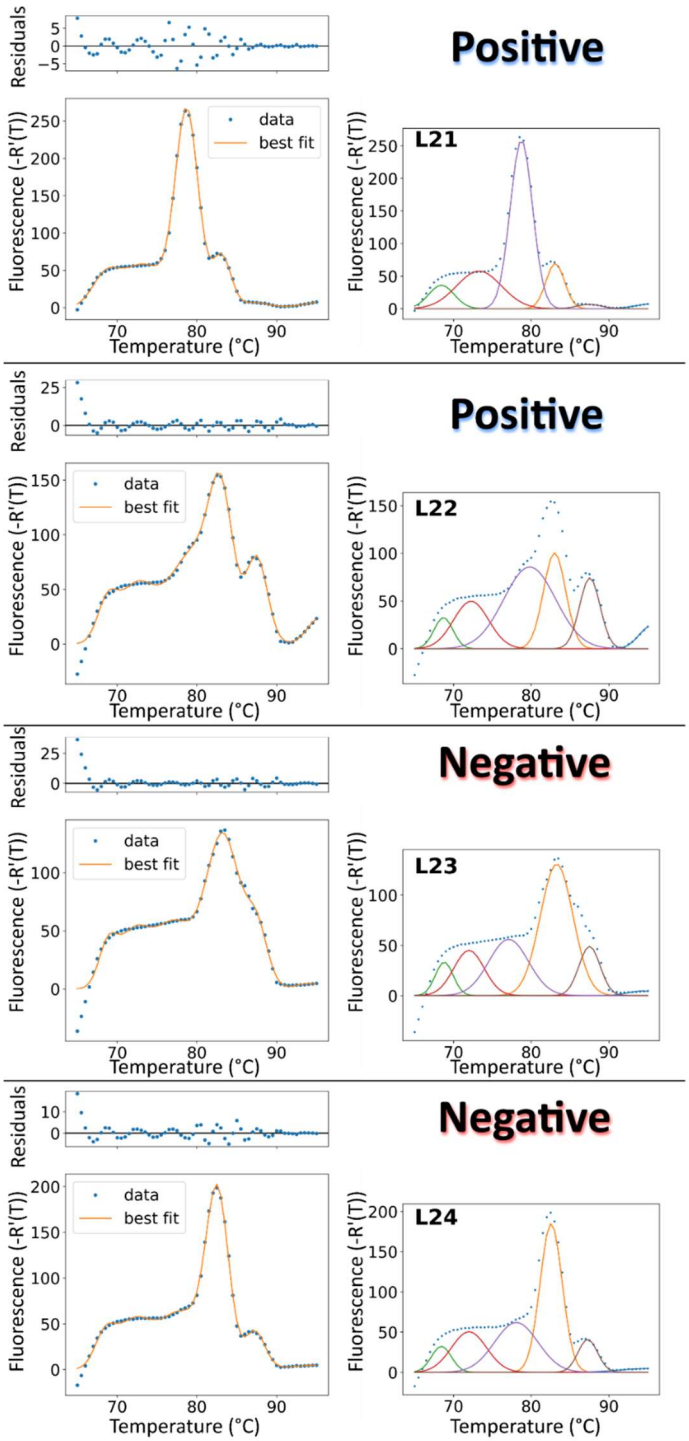

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (5/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



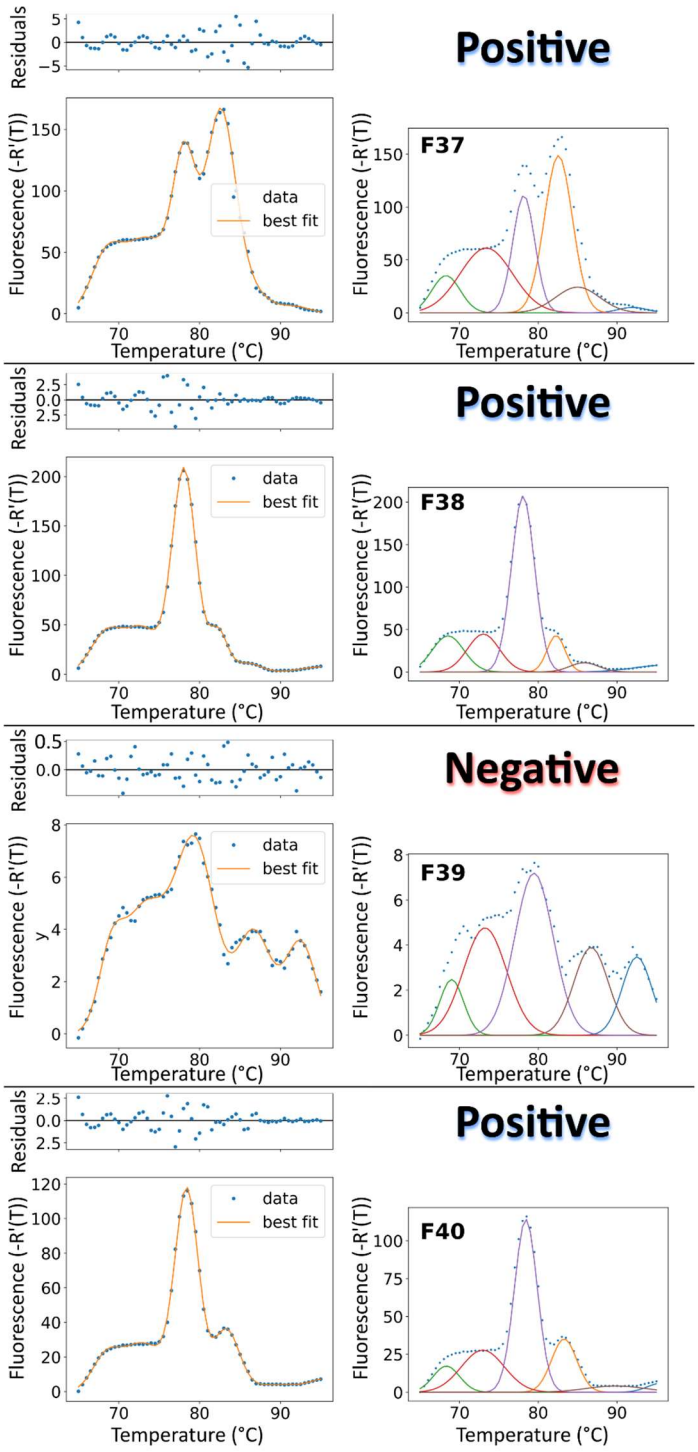

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (6/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



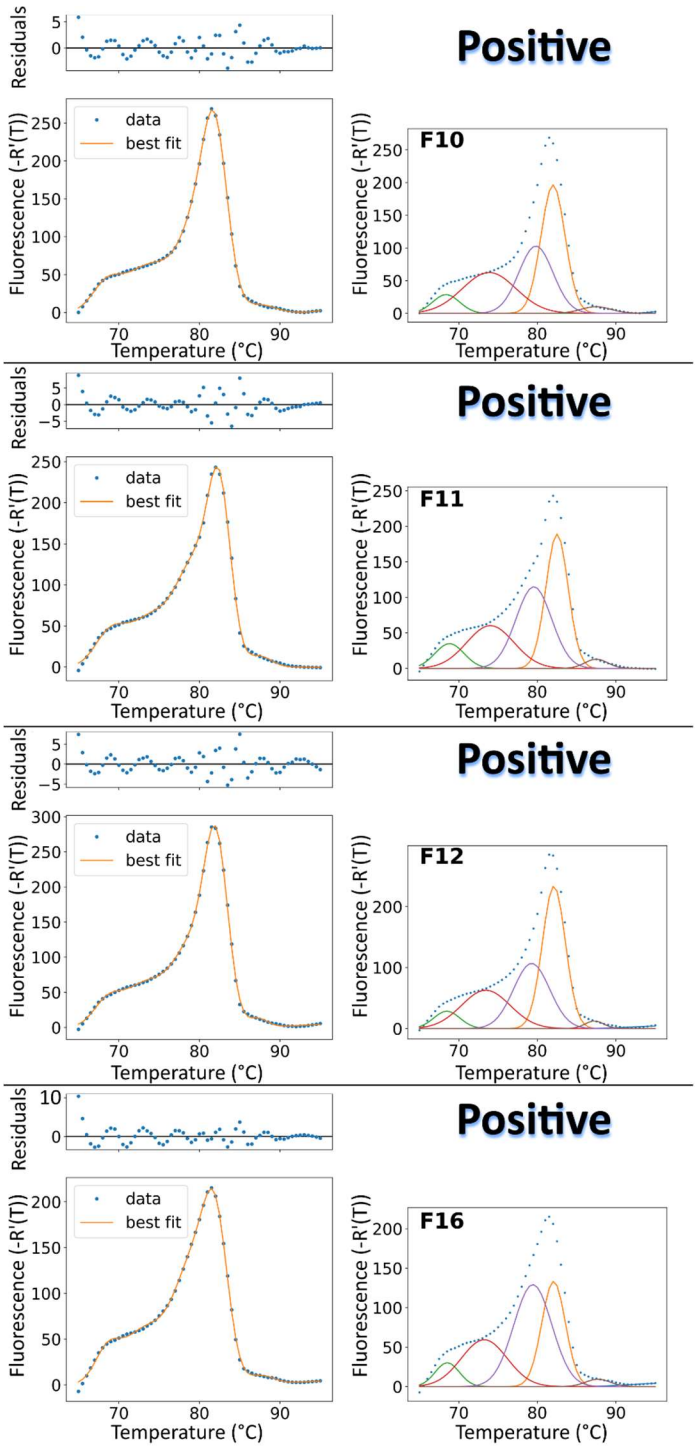

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (7/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



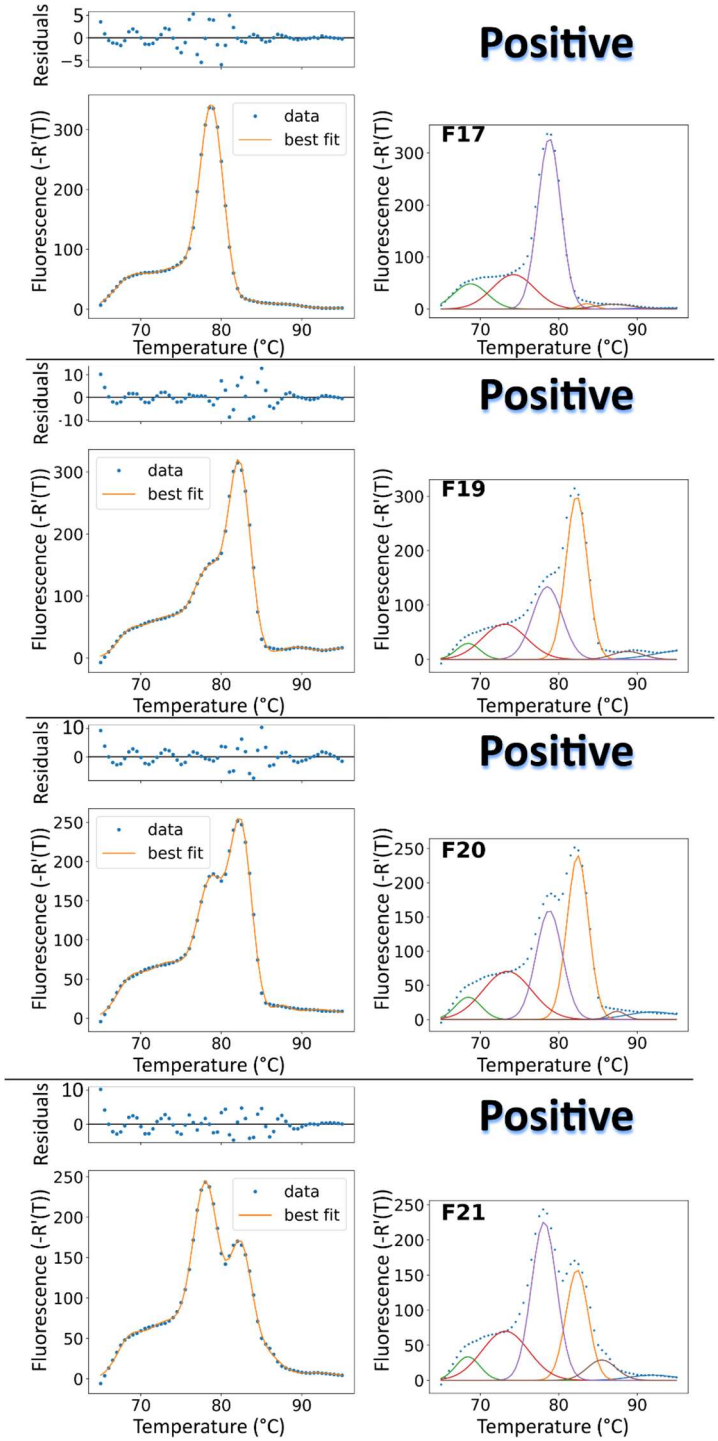

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (8/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



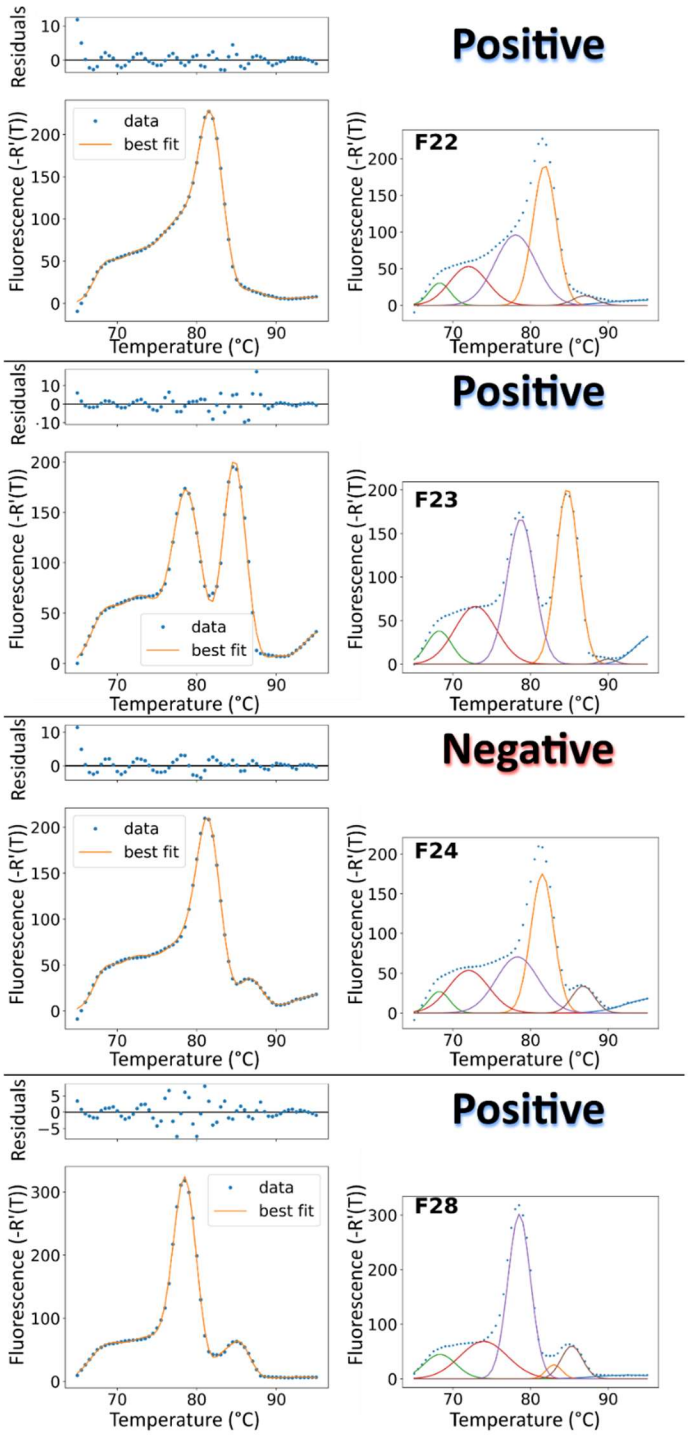

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (9/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



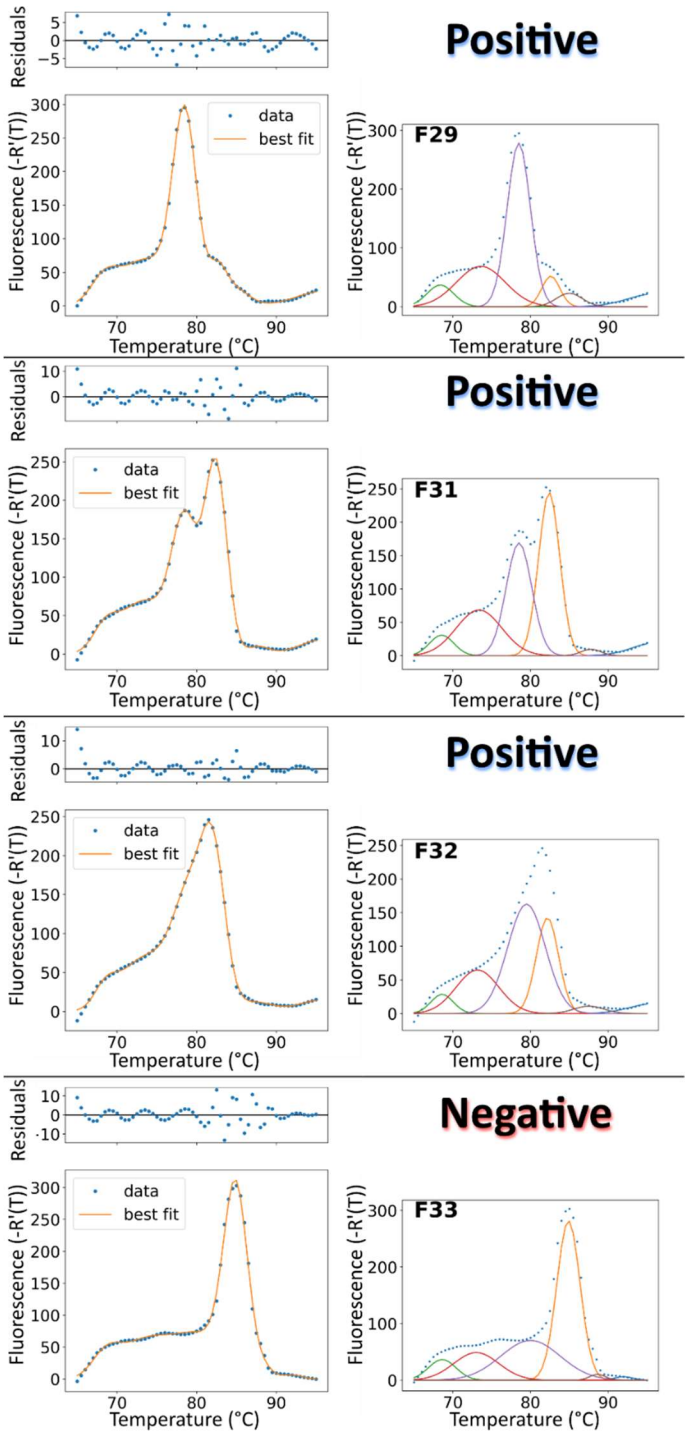

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (10/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



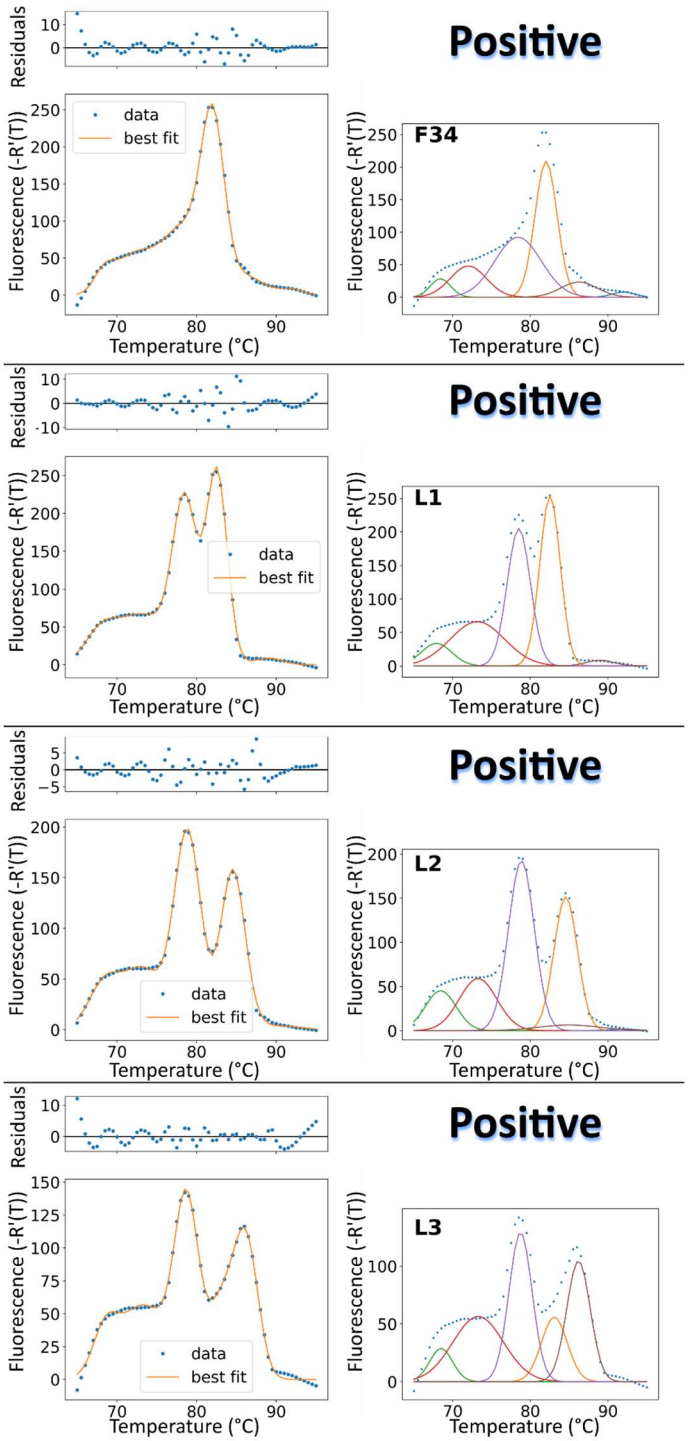

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (11/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



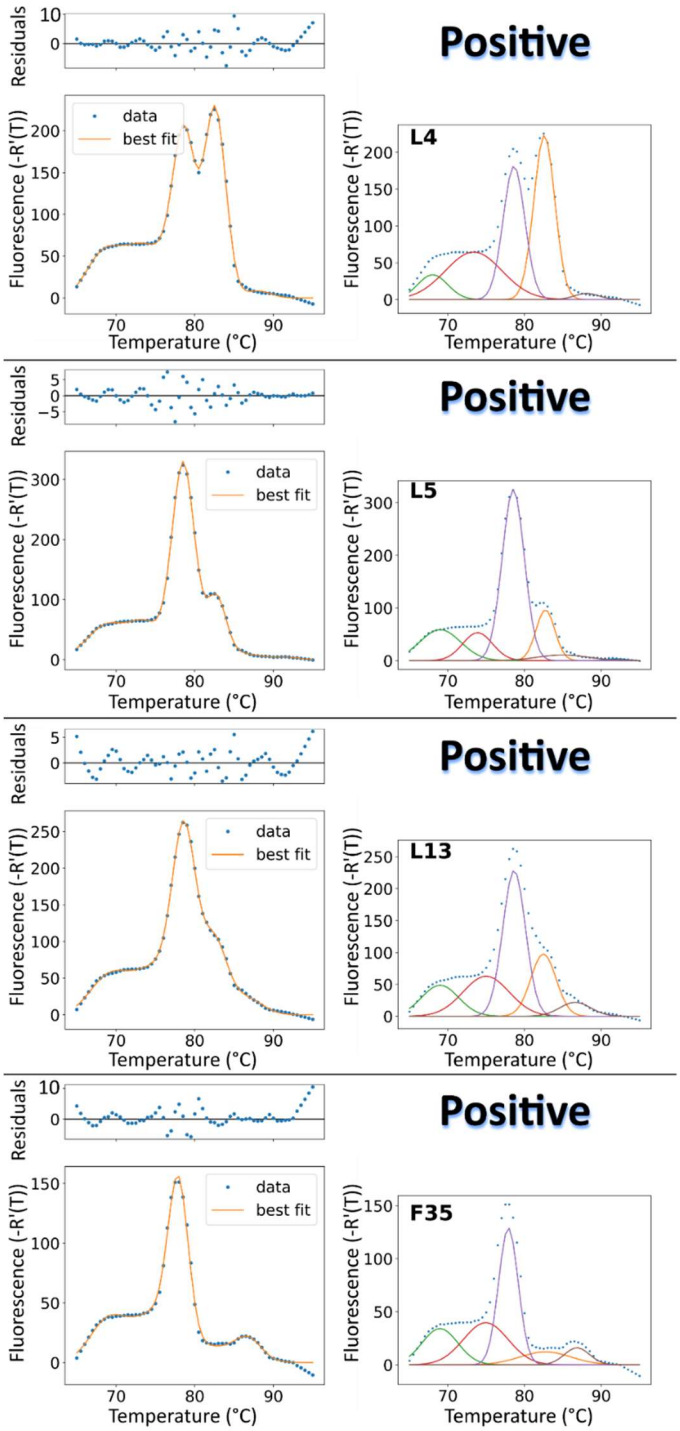

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (12/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



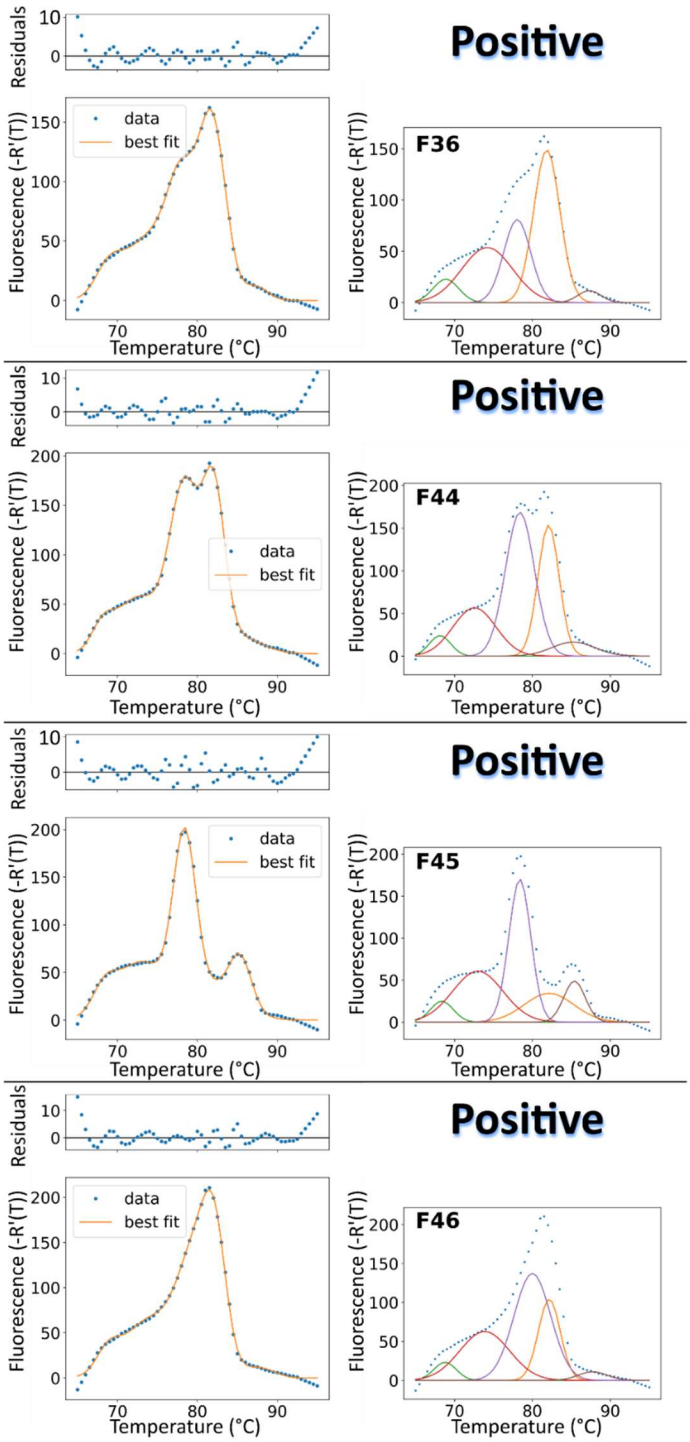

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (13/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



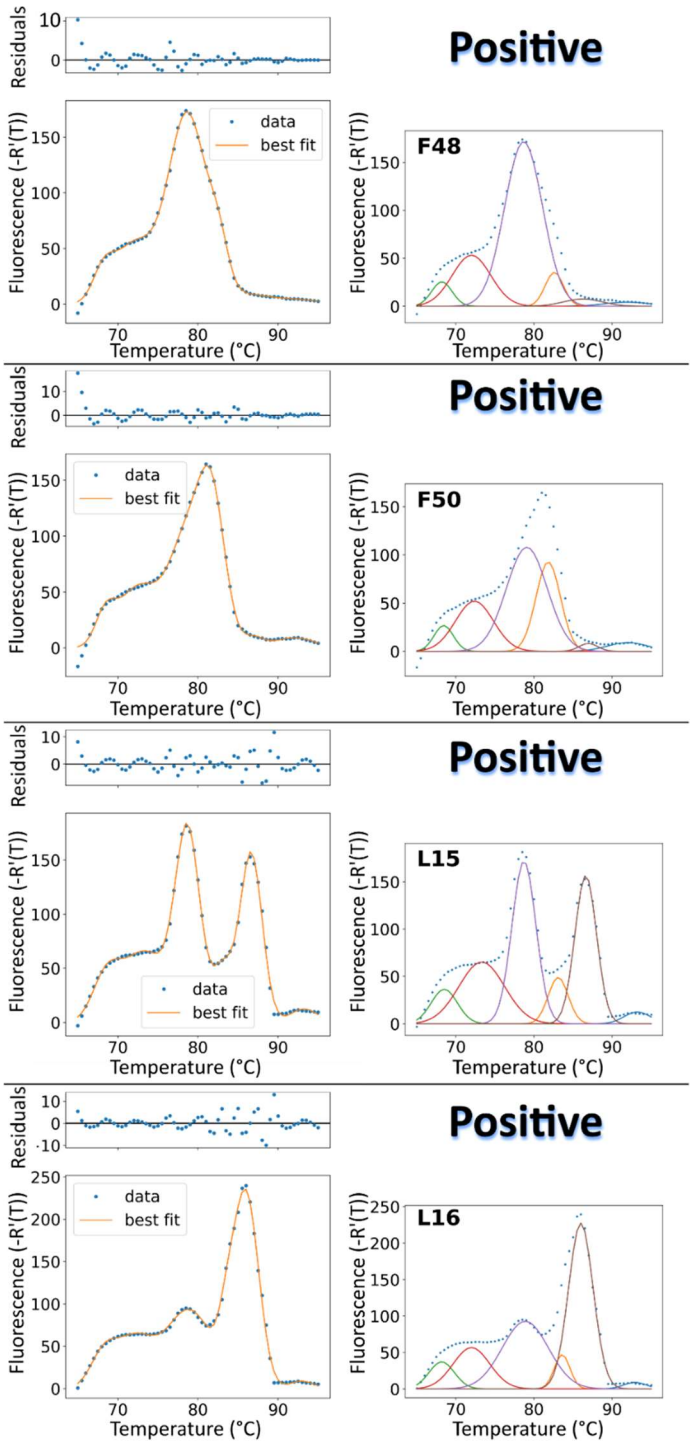

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (14/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.



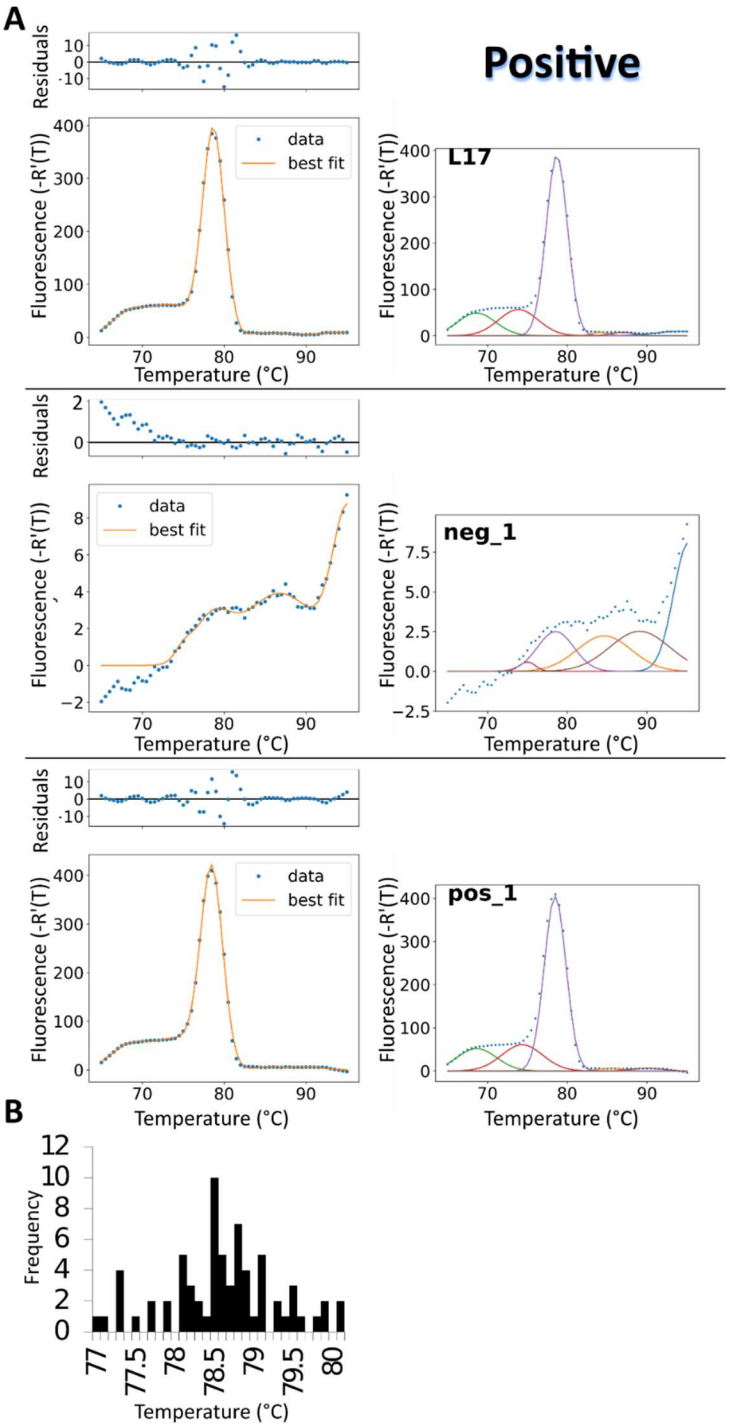

**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (15/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.




**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (16/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.




**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (17/18).** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode.




**Supplemental Figure 3.12. Field Trial qPCR Melt Curve Analysis (18/18). (A)** Complete melt curves from field collected mosquito pools (F) or laboratory-reared field collected mosquito larvae (L). As with survivorship samples, melt curves containing a peak at ~78.5 °C with a height at least 50% greater than the neighboring peak at ~ 74 °C using the python package LMFIT(91) were scored as positive. Left plots display raw qPCR melt curve data overlaid by the obtained fit along with the corresponding residuals. Right plots display raw qPCR data overlaid with the set of obtained gaussian peaks from computational analysis. The purple peak represents the detected DNA barcode. **(B)** Histogram of fitted barcode peak centers displaying a wider distribution of centers compared to survivorship samples.

## Extended Materials and Methods

### Barcode Recovery from Homogenized Mosquitoes

Mosquitoes were homogenized in mosquito diluent (77.9% DMEM, 20% FBS, 1% penicillin/streptomycin, 1% amphotericin B, 0.1% gentamycin) using a bead beater (Retsch MM400) set at 24hz for 1 minute. ATP was added for a final concentration of 2mM prior to incubation. Sample homogenate was set on an orbital shaker at room temperature overnight for a minimum of 8h prior to extraction. Extraction was performed with a MaxMAX Cell-Free DNA extraction kit, using a modified protocol in a 96-well plate format on a Kingfisher Flex extraction platform (Thermo Fisher). qPCR was performed using PowerUP SYBR Green Master Mix (Thermo Fisher), with the primers, cycling conditions, and standard quantification described for in vivo qPCR barcode recovery.

### Survival Analysis of Mosquitoes Reared on Crystals

Second instar *Culex tarsalis* mosquito larvae were reared in mosquito breeders (Bioquip), with 200 larvae in each container. The non-crystal fed control container received 200 µL of liver powder solution daily, and the crystal-reared treatment group received 200 µL of liver powder plus 25 µL of barcode-loaded crystal solution. Pupae were picked and placed individually in emergence containers. Upon emergence, mosquitoes received sugar cubes and water *ad libitum* and survivorship was followed until death. The following parameters were monitored: time to pupation, time to emergence, and time to death. Survival time was defined as the number of days from emergence to death. Mosquitoes found dead were frozen for barcode detection and determination of crystal persistence over time. Survival analysis was performed using a Bayesian survival model, with survival time following a Weibull distribution, and the mean of this distribution varying by group (crystal+liver powder vs. liver powder only) according to a linear regression function. Parameters were estimated by Markov Chain Monte Carlo using STAN. The regression coefficient for the 'group' factor did not differ significantly from zero, indicating no significant effect of crystal ingestion on adult mosquito survival. Plotting the posterior predictions show overlap between 95% credible intervals over the entire time course (fig. S4).

### Barcode detection sensitivity

*Culex tarsalis* mosquito larvae were raised on a diet of CJ crystals loaded with DNA mixed into liver powder as previously described. Adult female mosquitoes were separated and frozen at -80°C. Non-crystal fed adult female mosquitoes were also separated and frozen. One crystal-fed mosquito was placed in samples of increasing numbers of non-crystal fed mosquitoes for totals of 1, 10, and 20 mosquitoes per pool, to represent the common pool sizes used for pathogen surveillance programs. Negative control groups consisted of the sample total number of mosquitoes (1, 10, 20) but without a crystal-fed mosquito. Mosquito pools were homogenized by adding 1mL of mosquito diluent (DMEM with 20% FBS, 50 ug/mL penicillin/streptomycin, 50 ug/ml gentamicin, 2.5 µl/ml fungizone), two glass Coliroller beads (Novagen), and 75µl of 100 mM ATP, and homogenizing for 3 minutes at 24 Hz using a Retsch Mixer Mill (Retsch). Barcodes were recovered from mosquitoes as described above.

***in vitro* qPCR validation**

To prepare crystals for qPCR, 4 CJ crystals (~200 µm diameter) per replicate were immersed in 10 µL of approximately 50 ng/µL of a 125 base pair double-stranded DNA oligonucleotide (125mer, 5'- TAGGCGACTCGACGGTCTTACGCGTTACGTATGATATGCATCACCACCATC ACCAATAACCAACACCTAAATTTAACATCCGAGAATTATGGAGCACGCTAGCGTACGCTACGGTCCTAAC GCGC-3') and sealed in a glass well plate for approximately 12 hours, followed by washing with TE buffer as described for DNA loading to remove unbound 125mer. Following washing, 5 µL of solution was removed from the DNA-loaded crystal mixture, followed by addition of 5 µL filtered mosquito homogenate and incubated for approximately 12 hours. Following incubation, 10 µL of 40 mM ATP were added to the crystal/DNA/homogenate mixture and an aliquot of the solution is stored. The 125mer loaded crystal/homogenate/ATP solution was sealed in a glass well plate for approximately 12 hours. Following the 12hr incubation, an additional aliquot of the solution is stored to compare solution 125mer concentration pre and post ATP incubation via qPCR with the following primers [31]: fwd 5'-TAGGCGACTCGACGGTCT TACGCGTTACGT-3', rev 5'- GCGCGTTAGGACCGTA GCGTACGCTAGCGT-3'.

Standards used in qPCR consisted of serial dilutions of the pcr amplified 125mer template in nuclease free water. Starting with an initial 125mer template concentration of approximately 57 ng/µL, the following serial dilutions were made: $10^{-3}$, $10^{-4}$, $10^{-5}$, $10^{-6}$, and $10^{-7}$. The quantitative amplification was performed per the manufacturer's instructions (Luna® Universal qPCR Master Mix). Reaction conditions were: 1 cycle of 95 °C for 3:00 min and 40 cycles of 95 °C for 15 seconds followed by 72 °C for 30 seconds. Melt curve was obtained by 1 cycle of 95 °C for 30 seconds, 1 cycle of 65 °C for 30 seconds and 1 cycle of 95 °C for 30 seconds. TE buffer and filtered mosquito homogenate were separately used as the templates for negative controls. The positive control was 125mer in solution.

***in vivo* Environmental Persistence qPCR**

A master mix was prepared using 2X qPCR mix (Agilent #600882), 125mer revised forward primer (10 µM), 125mer reverse primer (10 µM) and nuclease water. Master mix (14 µL) was combined with 6 µL of unknown sample, template DNA or water. Standard curves were prepared using 4 to 5 100-fold serial dilutions of known concentration of barcode DNA. All reactions were performed in duplicate under the following cycling conditions: 1 cycle of 95 °C for 3:00 min and 50 cycles of 95 °C for 5 s, 60 °C for 10 s. Melt curve analysis was performed by spanning 65 °C to 95 °C, + 0.5 °C/cycle with 5 s/cycle. Initial concentrations were extrapolated from standard curves for each qPCR run. These quantities were combined into a single Excel spreadsheet. Single-factor ANOVA was used to assess statistical significance, before and after omitting samples with a Cq standard deviation ≥ 1. With inclusion of the omitted data, this result remains statistically significant ($p = 0.029$, $n = 42$ for loaded crystals and $n = 45$ for naked DNA).

**Porous Crystal Production**

Briefly, protein possessing a C-terminal hexahistidine tag was cloned into pSB3 expression vector. Protein expression was performed in BL21(DE3) *Escherichia coli* cells using Terrific broth with 0.4 mM IPTG induction at 25 °C for 16 hours. Cells were spun down, resuspended in lysis

buffer (50 mM HEPES, 500 mM NaCl, 25 mM Imidazole, 10 % glycerol, pH 7.4), and sonicated. Cell lysate was purified using immobilized metal affinity chromatography (IMAC) containing HisPur Ni-NTA resin (Thermo Fisher Scientific) and dialyzed into ammonium sulfate storage buffer (10 mM HEPES, 500 mM $NH_4(SO_4)_2$, 10% glycerol, pH 7.4) overnight at 4 °C. Purified protein was concentrated to 15 mg/mL by using Amicon Ultra-15 Centrifugal Filters (MWCO 10K, Millipore Sigma), aliquoted and stored at -30 °C. Crystals were grown overnight with sitting drop vapor diffusion (for single-crystal confocal imaging) or batch crystallization (for mosquito feeding) in 3.3 – 3.55M ammonium sulfate, 0.1 M Bis-Tris, pH 6.5 at 20 °C.
Crystal cross-linking was performed as described previously(21). Crystals were transferred to 4.2 M trimethylamine N-oxide (TMAO), pH 7.4 for 1 hour to remove excess protein in solution. Crystals were then placed in fresh 4.2M TMAO containing 50 mM imidazole and 40 mg/mL 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) for 2 hours. Crystals were then placed in quench solution (50 mM borate, pH 10) for 1 hour, followed by washing and storage in 4.2 M TMAO, pH 7.4.

**Crystal Fluorophore Labeling**

Following crystallization, crystals were loop-transferred into a wash solution comprised of 90% mTacsimate (1.83 M malonic acid, 0.25 M sodium citrate, 0.12 M succinic acid, 0.3 M D-L malic acid, 0.4 M acetic acid, 0.5 M sodium formate, 0.16 M sodium tartrate, pH 6.5) and 10% glycerol for 1 hour. This is an altered recipe based on Tacsimate (Hampton Research) that does not contain primary amines that could interfere with crosslinking. Crystals were crosslinked in fresh wash solution containing 1% glyoxal and 250 mM borane dimethylamine complex (DMAB) for 2 hours. Crystals were quenched and labeled with Texas Red dye by transferring to quench solution containing 0.25 M carbohydrazide, 0.25 mM Texas Red-X (ThermoFisher), and 100 mM DMAB in phosphate buffered saline (137 mM sodium chloride, 2.7 mM potassium chloride, 10 mM sodium phosphate dibasic, 1.8 mM potassium phosphate dibasic), pH 7.5 for 1 hour. Following crosslinking and labeling, crystals were washed and stored in 4.2 M TMAO.

APPENDIX III. Supplemental Information for Chapter 4 |
Scalable Combinatorial Synthesis of Synthetic DNA Barcode Sequences

**Authors**
Julius D. Stuart[a], Natalie R. Wickenkamp[b], Kaleb A. Davis[b], Camden Meyer[c], Rebekah C. Kading[b], Christopher D. Snow[c]

**Author Affiliations**
[a]Department of Chemistry, Colorado State University, Fort Collins, CO 80523; [b]Department of Microbiology, Immunology, and Pathology, Colorado State University, Fort Collins, CO 80523; [c]Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO 80523;

**Corresponding Author**
Christopher D. Snow
(970) 491-5276
Christopher.snow@colostate.edu
356 Scott Bioengineering
Colorado State University
Fort Collins, CO 80523

**Author contributions:** C.D.S., J.D.S, R.C.K. designed research; J.D.S, C.M., N.R.W., K.A.D. performed research; J.D.S., C.M. analyzed data; J.D.S., C.D.S. wrote the paper; J.D.S., C.D.S., R.C.K. edited the manuscript.

**This section includes:**
Tables S4.1- 4.3
Figure S4.1-4.8
Extended Methods

**Table S4.1. TrapTag Sequence Information**

| TrapTag_ oligo_ID | Illumina Index ID | TrapTag Index Sequence 5'-3' (8 nt) | TrapTag Oligo Sequence to order 5'-3' (51 nt) (contains *reverse complement* of TrapTag Index Sequence) |
|---|---|---|---|
| tt_001 | UDI0001 | CCGCGGTT | TTCTGGGTTCCTCATCGCAACCGCGGGTTGAA GCCGGTTACCAC |
| tt_002 | UDI0002 | TTATAACC | TTCTGGGTTCCTCATCGCGGTTATAAGTTGAA GCCGGTTACCAC |
| tt_003 | UDI0003 | GGACTTGG | TTCTGGGTTCCTCATCGCCCAAGTCCGTTGAA GCCGGTTACCAC |
| tt_004 | UDI0004 | AAGTCCAA | TTCTGGGTTCCTCATCGCTTGGACTTGTTGAA GCCGGTTACCAC |
| tt_005 | UDI0005 | ATCCACTG | TTCTGGGTTCCTCATCGCCAGTGGATGTTGAA GCCGGTTACCAC |
| tt_006 | UDI0006 | GCTTGTCA | TTCTGGGTTCCTCATCGCTGACAAGCGTTGAA GCCGGTTACCAC |
| tt_007 | UDI0007 | CAAGCTAG | TTCTGGGTTCCTCATCGCCTAGCTTGGTTGAA GCCGGTTACCAC |
| tt_008 | UDI0008 | TGGATCGA | TTCTGGGTTCCTCATCGCTCGATCCAGTTGAA GCCGGTTACCAC |
| tt_009 | UDI0009 | AGTTCAGG | TTCTGGGTTCCTCATCGCCCTGAACTGTTGAA GCCGGTTACCAC |
| tt_010 | UDI0010 | GACCTGAA | TTCTGGGTTCCTCATCGCTTCAGGTCGTTGAA GCCGGTTACCAC |
| tt_011 | UDI0011 | TCTCTACT | TTCTGGGTTCCTCATCGCAGTAGAGAGTTGAA GCCGGTTACCAC |
| tt_012 | UDI0012 | CTCTCGTC | TTCTGGGTTCCTCATCGCGACGAGAGGTTGA AGCCGGTTACCAC |
| tt_013 | UDI0013 | CCAAGTCT | TTCTGGGTTCCTCATCGCAGACTTGGGTTGAA GCCGGTTACCAC |
| tt_014 | UDI0014 | TTGGACTC | TTCTGGGTTCCTCATCGCGAGTCCAAGTTGAA GCCGGTTACCAC |
| tt_015 | UDI0015V2 | CAGTAGGC | TTCTGGGTTCCTCATCGCGCCTACTGGTTGAA GCCGGTTACCAC |
| tt_016 | UDI0015 | GGCTTAAG | TTCTGGGTTCCTCATCGCCTTAAGCCGTTGAA GCCGGTTACCAC |
| tt_017 | UDI0016V2 | TGACGAAT | TTCTGGGTTCCTCATCGCATTCGTCAGTTGAA GCCGGTTACCAC |
| tt_018 | UDI0016 | AATCCGGA | TTCTGGGTTCCTCATCGCTCCGGATTGTTGAA GCCGGTTACCAC |
| tt_019 | UDI0017 | TAATACAG | TTCTGGGTTCCTCATCGCCTGTATTAGTTGAA GCCGGTTACCAC |
| tt_020 | UDI0018 | CGGCGTGA | TTCTGGGTTCCTCATCGCTCACGCCGGTTGAA GCCGGTTACCAC |
| tt_021 | UDI0019 | ATGTAAGT | TTCTGGGTTCCTCATCGCACTTACATGTTGAA GCCGGTTACCAC |

**Table S4.1. TrapTag Sequence Information (Continued)**

| TrapTag_ oligo_ID | Illumina Index ID | TrapTag Index Sequence 5'-3' (8 nt) | TrapTag Oligo Sequence to order 5'-3' (51 nt) (contains *reverse complement* of TrapTag Index Sequence) |
|---|---|---|---|
| tt_022 | UDI0020 | GCACGGAC | TTCTGGGTTCCTCATCGCGTCCGTGCGTTGAA GCCGGTTACCAC |
| tt_023 | UDI0021 | GGTACCTT | TTCTGGGTTCCTCATCGCAAGGTACCGTTGAA GCCGGTTACCAC |
| tt_024 | UDI0022 | AACGTTCC | TTCTGGGTTCCTCATCGCGGAACGTTGTTGAA GCCGGTTACCAC |
| tt_025 | UDI0023 | GCAGAATT | TTCTGGGTTCCTCATCGCAATTCTGCGTTGAA GCCGGTTACCAC |
| tt_026 | UDI0024 | ATGAGGCC | TTCTGGGTTCCTCATCGCGGCCTCATGTTGAA GCCGGTTACCAC |
| tt_027 | UDI0025 | ACTAAGAT | TTCTGGGTTCCTCATCGCATCTTAGTGTTGAA GCCGGTTACCAC |
| tt_028 | UDI0026 | GTCGGAGC | TTCTGGGTTCCTCATCGCGCTCCGACGTTGAA GCCGGTTACCAC |
| tt_029 | UDI0027 | CTTGGTAT | TTCTGGGTTCCTCATCGCATACCAAGGTTGAA GCCGGTTACCAC |
| tt_030 | UDI0028 | TCCAACGC | TTCTGGGTTCCTCATCGCGCGTTGGAGTTGAA GCCGGTTACCAC |
| tt_031 | UDI0029 | CCGTGAAG | TTCTGGGTTCCTCATCGCCTTCACGGGTTGAA GCCGGTTACCAC |
| tt_032 | UDI0030 | TTACAGGA | TTCTGGGTTCCTCATCGCTCCTGTAAGTTGAA GCCGGTTACCAC |
| tt_033 | UDI0031 | GGCATTCT | TTCTGGGTTCCTCATCGCAGAATGCCGTTGAA GCCGGTTACCAC |
| tt_034 | UDI0032 | AATGCCTC | TTCTGGGTTCCTCATCGCGAGGCATTGTTGAA GCCGGTTACCAC |
| tt_035 | UDI0033 | TACCGAGG | TTCTGGGTTCCTCATCGCCCTCGGTAGTTGAA GCCGGTTACCAC |
| tt_036 | UDI0034 | CGTTAGAA | TTCTGGGTTCCTCATCGCTTCTAACGGTTGAA GCCGGTTACCAC |
| tt_037 | UDI0035 | AGCCTCAT | TTCTGGGTTCCTCATCGCATGAGGCTGTTGAA GCCGGTTACCAC |
| tt_038 | UDI0036 | GATTCTGC | TTCTGGGTTCCTCATCGCGCAGAATCGTTGAA GCCGGTTACCAC |
| tt_039 | UDI0037 | TCGTAGTG | TTCTGGGTTCCTCATCGCCACTACGAGTTGAA GCCGGTTACCAC |
| tt_040 | UDI0038 | CTACGACA | TTCTGGGTTCCTCATCGCTGTCGTAGGTTGAA GCCGGTTACCAC |
| tt_041 | UDI0039 | TAAGTGGT | TTCTGGGTTCCTCATCGCACCACTTAGTTGAA GCCGGTTACCAC |
| tt_042 | UDI0040 | CGGACAAC | TTCTGGGTTCCTCATCGCGTTGTCCGGTTGAA GCCGGTTACCAC |

**Table S4.1. TrapTag Sequence Information (Continued)**

| TrapTag_ oligo_ID | Illumina Index ID | TrapTag Index Sequence 5'-3' (8 nt) | TrapTag Oligo Sequence to order 5'-3' (51 nt) (contains *reverse complement* of TrapTag Index Sequence) |
|---|---|---|---|
| tt_043 | UDI0041 | ATATGGAT | TTCTGGGTTCCTCATCGCATCCATATGTTGAA GCCGGTTACCAC |
| tt_044 | UDI0042 | GCGCAAGC | TTCTGGGTTCCTCATCGCGCTTGCGCGTTGAA GCCGGTTACCAC |
| tt_045 | UDI0043 | AAGATACT | TTCTGGGTTCCTCATCGCAGTATCTTGTTGAA GCCGGTTACCAC |
| tt_046 | UDI0044 | GGAGCGTC | TTCTGGGTTCCTCATCGCGACGCTCCGTTGAA GCCGGTTACCAC |
| tt_047 | UDI0045 | ATGGCATG | TTCTGGGTTCCTCATCGCCATGCCATGTTGAA GCCGGTTACCAC |
| tt_048 | UDI0046 | GCAATGCA | TTCTGGGTTCCTCATCGCTGCATTGCGTTGAA GCCGGTTACCAC |
| tt_049 | UDI0047 | GTTCCAAT | TTCTGGGTTCCTCATCGCATTGGAACGTTGAA GCCGGTTACCAC |
| tt_050 | UDI0048 | ACCTTGGC | TTCTGGGTTCCTCATCGCGCCAAGGTGTTGAA GCCGGTTACCAC |
| tt_051 | UDI0049 | ATATCTCG | TTCTGGGTTCCTCATCGCCGAGATATGTTGAA GCCGGTTACCAC |
| tt_052 | UDI0050 | GCGCTCTA | TTCTGGGTTCCTCATCGCTAGAGCGCGTTGAA GCCGGTTACCAC |
| tt_053 | UDI0051 | AACAGGTT | TTCTGGGTTCCTCATCGCAACCTGTTGTTGAA GCCGGTTACCAC |
| tt_054 | UDI0052 | GGTGAACC | TTCTGGGTTCCTCATCGCGGTTCACCGTTGAA GCCGGTTACCAC |
| tt_055 | UDI0053 | CAACAATG | TTCTGGGTTCCTCATCGCCATTGTTGGTTGAA GCCGGTTACCAC |
| tt_056 | UDI0054 | TGGTGGCA | TTCTGGGTTCCTCATCGCTGCCACCAGTTGAA GCCGGTTACCAC |
| tt_057 | UDI0055V2 | GTTCGCCG | TTCTGGGTTCCTCATCGCCGGCGAACGTTGAA GCCGGTTACCAC |
| tt_058 | UDI0055 | AGGCAGAG | TTCTGGGTTCCTCATCGCCTCTGCCTGTTGAA GCCGGTTACCAC |
| tt_059 | UDI0056V2 | CACGAGCG | TTCTGGGTTCCTCATCGCCGCTCGTGGTTGAA GCCGGTTACCAC |
| tt_060 | UDI0056 | GAATGAGA | TTCTGGGTTCCTCATCGCTCTCATTCGTTGAAG CCGGTTACCAC |
| tt_061 | UDI0057 | TGCGGCGT | TTCTGGGTTCCTCATCGCACGCCGCAGTTGAA GCCGGTTACCAC |
| tt_062 | UDI0058 | CATAATAC | TTCTGGGTTCCTCATCGCGTATTATGGTTGAA GCCGGTTACCAC |
| tt_063 | UDI0059 | GATCTATC | TTCTGGGTTCCTCATCGCGATAGATCGTTGAA GCCGGTTACCAC |
| tt_064 | UDI0060 | AGCTCGCT | TTCTGGGTTCCTCATCGCAGCGAGCTGTTGAA GCCGGTTACCAC |

**Table S4.1. TrapTag Sequence Information (Continued)**

| TrapTag_ oligo_ID | Illumina Index ID | TrapTag Index Sequence 5'-3' (8 nt) | TrapTag Oligo Sequence to order 5'-3' (51 nt) (contains *reverse complement* of TrapTag Index Sequence) |
|---|---|---|---|
| tt_065 | UDI0061 | CGGAACTG | TTCTGGGTTCCTCATCGCCAGTTCCGGTTGAA GCCGGTTACCAC |
| tt_066 | UDI0062 | TAAGGTCA | TTCTGGGTTCCTCATCGCTGACCTTAGTTGAA GCCGGTTACCAC |
| tt_067 | UDI0063 | TTGCCTAG | TTCTGGGTTCCTCATCGCCTAGGCAAGTTGAA GCCGGTTACCAC |
| tt_068 | UDI0064 | CCATTCGA | TTCTGGGTTCCTCATCGCTCGAATGGGTTGAA GCCGGTTACCAC |
| tt_069 | UDI0065 | ACACTAAG | TTCTGGGTTCCTCATCGCCTTAGTGTGTTGAA GCCGGTTACCAC |
| tt_070 | UDI0066 | GTGTCGGA | TTCTGGGTTCCTCATCGCTCCGACACGTTGAA GCCGGTTACCAC |
| tt_071 | UDI0067 | TTCCTGTT | TTCTGGGTTCCTCATCGCAACAGGAAGTTGAA GCCGGTTACCAC |
| tt_072 | UDI0068 | CCTTCACC | TTCTGGGTTCCTCATCGCGGTGAAGGGTTGA AGCCGGTTACCAC |
| tt_073 | UDI0069 | GCCACAGG | TTCTGGGTTCCTCATCGCCCTGTGGCGTTGAA GCCGGTTACCAC |
| tt_074 | UDI0070 | ATTGTGAA | TTCTGGGTTCCTCATCGCTTCACAATGTTGAA GCCGGTTACCAC |
| tt_075 | UDI0071 | ACTCGTGT | TTCTGGGTTCCTCATCGCACACGAGTGTTGAA GCCGGTTACCAC |
| tt_076 | UDI0072 | GTCTACAC | TTCTGGGTTCCTCATCGCGTGTAGACGTTGAA GCCGGTTACCAC |
| tt_077 | UDI0073 | CAATTAAC | TTCTGGGTTCCTCATCGCGTTAATTGGTTGAA GCCGGTTACCAC |
| tt_078 | UDI0074 | TGGCCGGT | TTCTGGGTTCCTCATCGCACCGGCCAGTTGAA GCCGGTTACCAC |
| tt_079 | UDI0075 | AGTACTCC | TTCTGGGTTCCTCATCGCGGAGTACTGTTGAA GCCGGTTACCAC |
| tt_080 | UDI0076 | GACGTCTT | TTCTGGGTTCCTCATCGCAAGACGTCGTTGAA GCCGGTTACCAC |
| tt_081 | UDI0077 | TGCGAGAC | TTCTGGGTTCCTCATCGCGTCTCGCAGTTGAA GCCGGTTACCAC |
| tt_082 | UDI0078 | CATAGAGT | TTCTGGGTTCCTCATCGCACTCTATGGTTGAA GCCGGTTACCAC |
| tt_083 | UDI0079 | ACAGGCGC | TTCTGGGTTCCTCATCGCGCGCCTGTGTTGAA GCCGGTTACCAC |
| tt_084 | UDI0080 | GTGAATAT | TTCTGGGTTCCTCATCGCATATTCACGTTGAA GCCGGTTACCAC |
| tt_085 | UDI0081 | AACTGTAG | TTCTGGGTTCCTCATCGCCTACAGTTGTTGAA GCCGGTTACCAC |
| tt_086 | UDI0082 | GGTCACGA | TTCTGGGTTCCTCATCGCTCGTGACCGTTGAA GCCGGTTACCAC |

**Table S4.1. TrapTag Sequence Information (Continued)**

| TrapTag_ oligo_ID | Illumina Index ID | TrapTag Index Sequence 5'-3' (8 nt) | TrapTag Oligo Sequence to order 5'-3' (51 nt) (contains *reverse complement* of TrapTag Index Sequence) |
|---|---|---|---|
| tt_087 | UDI0083 | CTGCTTCC | TTCTGGGTTCCTCATCGCGGAAGCAGGTTGAAGCCGGTTACCAC |
| tt_088 | UDI0084 | TCATCCTT | TTCTGGGTTCCTCATCGCAAGGATGAGTTGAAGCCGGTTACCAC |
| tt_089 | UDI0085 | AGGTTATA | TTCTGGGTTCCTCATCGCTATAACCTGTTGAAGCCGGTTACCAC |
| tt_090 | UDI0086 | GAACCGCG | TTCTGGGTTCCTCATCGCCGCGGTTCGTTGAAGCCGGTTACCAC |
| tt_091 | UDI0087 | CTCACCAA | TTCTGGGTTCCTCATCGCTTGGTGAGGTTGAAGCCGGTTACCAC |
| tt_092 | UDI0088 | TCTGTTGG | TTCTGGGTTCCTCATCGCCCAACAGAGTTGAAGCCGGTTACCAC |
| tt_093 | UDI0089 | TATCGCAC | TTCTGGGTTCCTCATCGCGTGCGATAGTTGAAGCCGGTTACCAC |
| tt_094 | UDI0090 | CGCTATGT | TTCTGGGTTCCTCATCGCACATAGCGGTTGAAGCCGGTTACCAC |
| tt_095 | UDI0091 | GTATGTTC | TTCTGGGTTCCTCATCGCGAACATACGTTGAAGCCGGTTACCAC |
| tt_096 | UDI0092 | ACGCACCT | TTCTGGGTTCCTCATCGCAGGTGCGTGTTGAAGCCGGTTACCAC |
| tt_097 | UDI0093 | TACTCATA | TTCTGGGTTCCTCATCGCTATGAGTAGTTGAAGCCGGTTACCAC |
| tt_098 | UDI0094 | CGTCTGCG | TTCTGGGTTCCTCATCGCCGCAGACGGTTGAAGCCGGTTACCAC |
| tt_099 | UDI0095 | TCGATATC | TTCTGGGTTCCTCATCGCGATATCGAGTTGAAGCCGGTTACCAC |
| tt_100 | UDI0096 | CTAGCGCT | TTCTGGGTTCCTCATCGCAGCGCTAGGTTGAAGCCGGTTACCAC |

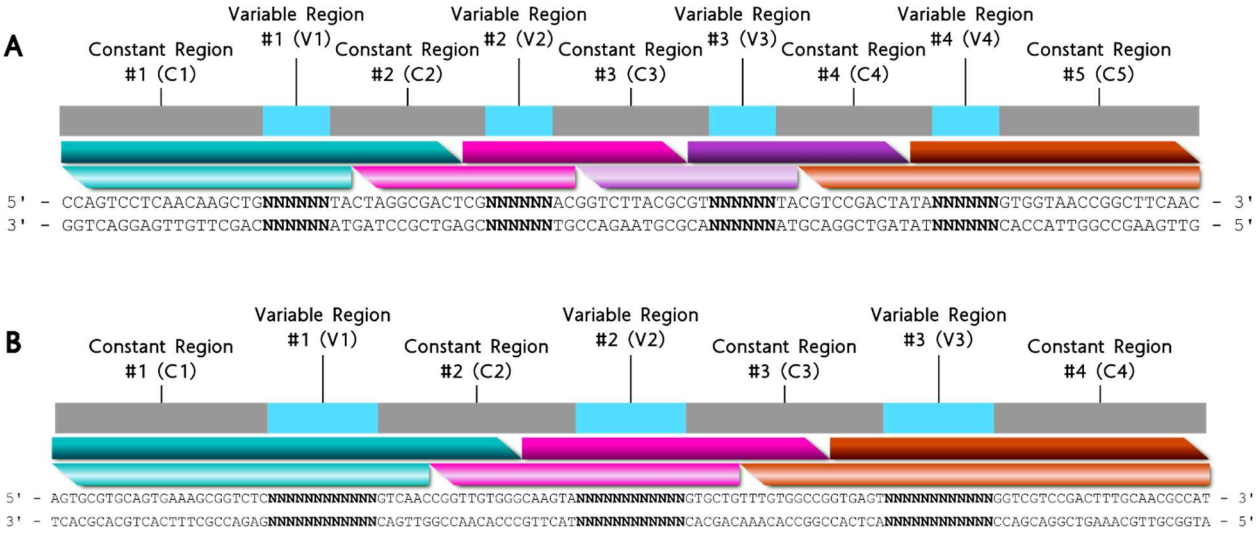



**Figure S4.1. Modular Barcode Layout. A)** Modular barcode layout for the Gen_1 library displaying 4 blocks containing a total of 4 variable region sequences. **B)** Modular barcode layout for the Gen_2 library displaying 3 blocks containing a total of 3 variable region sequences.

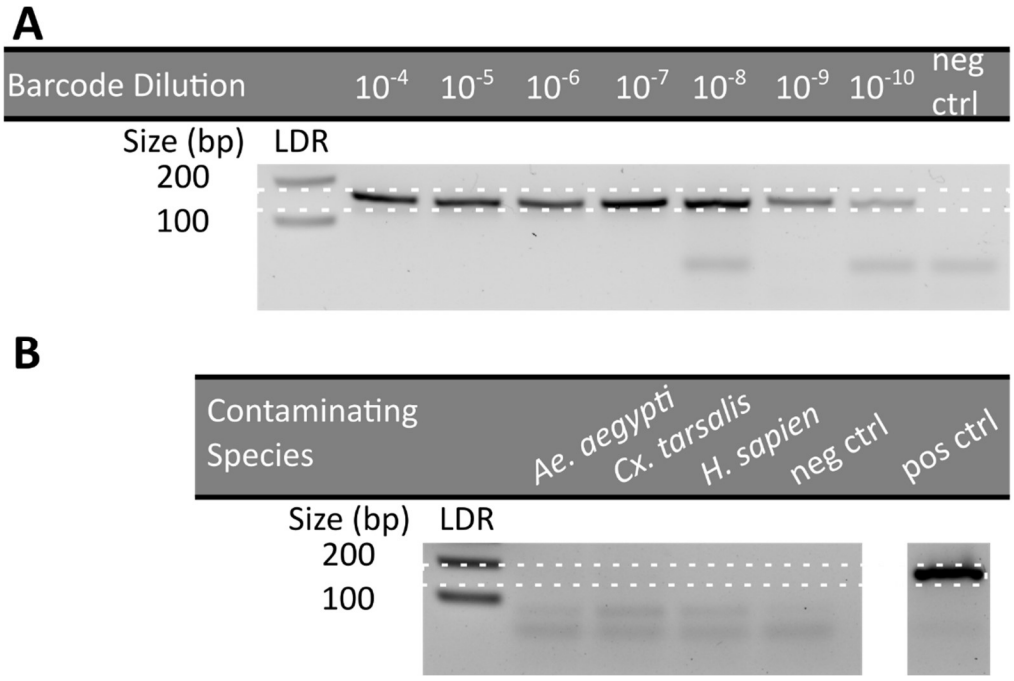
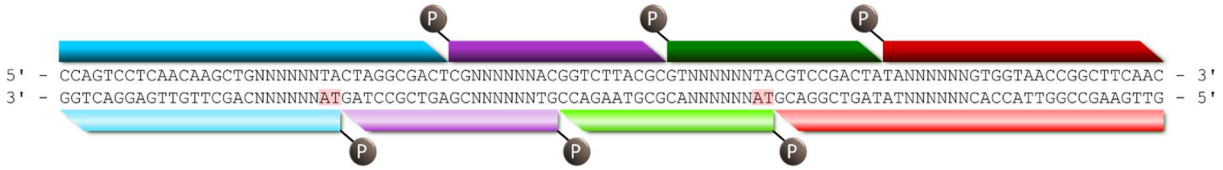
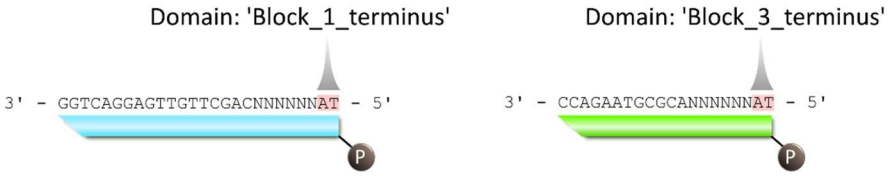



**Figure S4.2. Primer Sensitivity and Specificity. A)** Gel electrophoresis results following PCR with barcode template ranging from $10^{-4}$ – $10^{-10}$ dilutions of the initial barcode concentration, highlighting the sensitivity of the designed primer set for amplifying the target 161bp barcode amplicon. **B)** Gel electrophoresis results following PCR of barcode in the presence of additional contaminating species. *Ae. aegypti* refers to a pool of 15 *Aedes aegypti* mosquitos. *Cx. tarsalis* refers to a pool of 15 *Culex tarsalis* mosquitos. *H. sapien* refers to human saliva. Remarkably, the designed primer pair exhibits strict specificity for amplifying barcode only with no observed off-target amplification of contaminating species.

**Problem**: Identical 5' terminal nucleotides (highlighted below) allow off-target annealing




**Solution**:
1) Assign ***domain*** names to the 5' terminal end of the light blue and light green strands corresponding to the highlighted nucleotides..




2) Use Nupack's ***Diversity*** constraint to force the specified domains to contain non-identical nucleotides during the design run.

```
my_hard_constraints = [ . . .
                       Diversity(word = 4, types = 4, scope = [Block_1_terminus, Block_3_terminus])
                       . . . ]
```

*word*: Length in nucletides of the 'window' probed by the constaint
*types*: The number of nucleotide types to include that must occur in the word
*scope*: accepts a list of ***concatenated*** domains for the constaint to act upon

**Result:** Following a design run, a unique nucleotide occupies each of the 4 positions in the 'Block_1_terminus' and 'Block_3_terminus' domains.

**Figure S4.3. Negative Design with NUPACK.** The 5’ terminal sequence regions of blocks 1 and 3, light blue and light green, respectively, are assigned domain names used as input for the Diversity constraint from NUPACK which prevents identical nucleotides from appearing in the specified domains following a design run.

| Substitution Type | Frequency (%) | Frequency as reported by Pfeiffer et al. (%) |
|---|---|---|
| C -> T | 0.010 | 0.11 |
| G -> T | 0.008 | 0.11 |
| C -> A | 0.006 | 0.13 |
| G -> A | 0.005 | - |
| T -> G | 0.004 | - |
| A -> T | 0.004 | - |
| T -> C | 0.003 | 0.04 |
| A -> G | 0.002 | - |
| C -> G | 0.002 | 0.04 |
| A -> C | 0.001 | 0.04 |
| G -> C | 0.001 | 0.04 |
| T -> A | 0.001 | - |

**Table S4.2.** Comparison of detected substitution frequencies with values reported by Pfeiffer *et al*.(107)

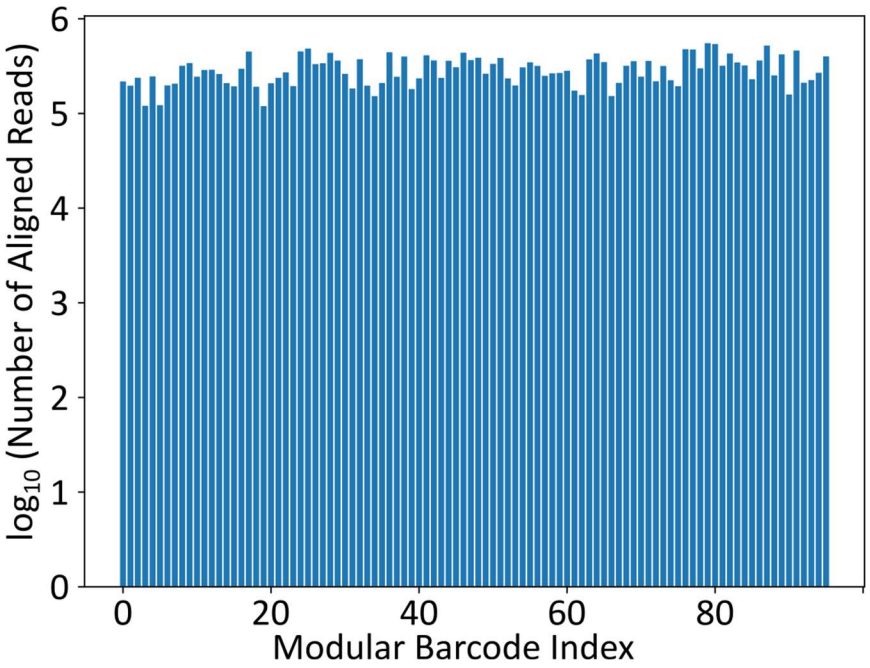



**Figure S4.4. Gen_2 Barcode Recovery.** Following alignment of ~109M joined reads using a custom Python script, all 96 pooled barcodes were detected at read quantities above $10^5$.

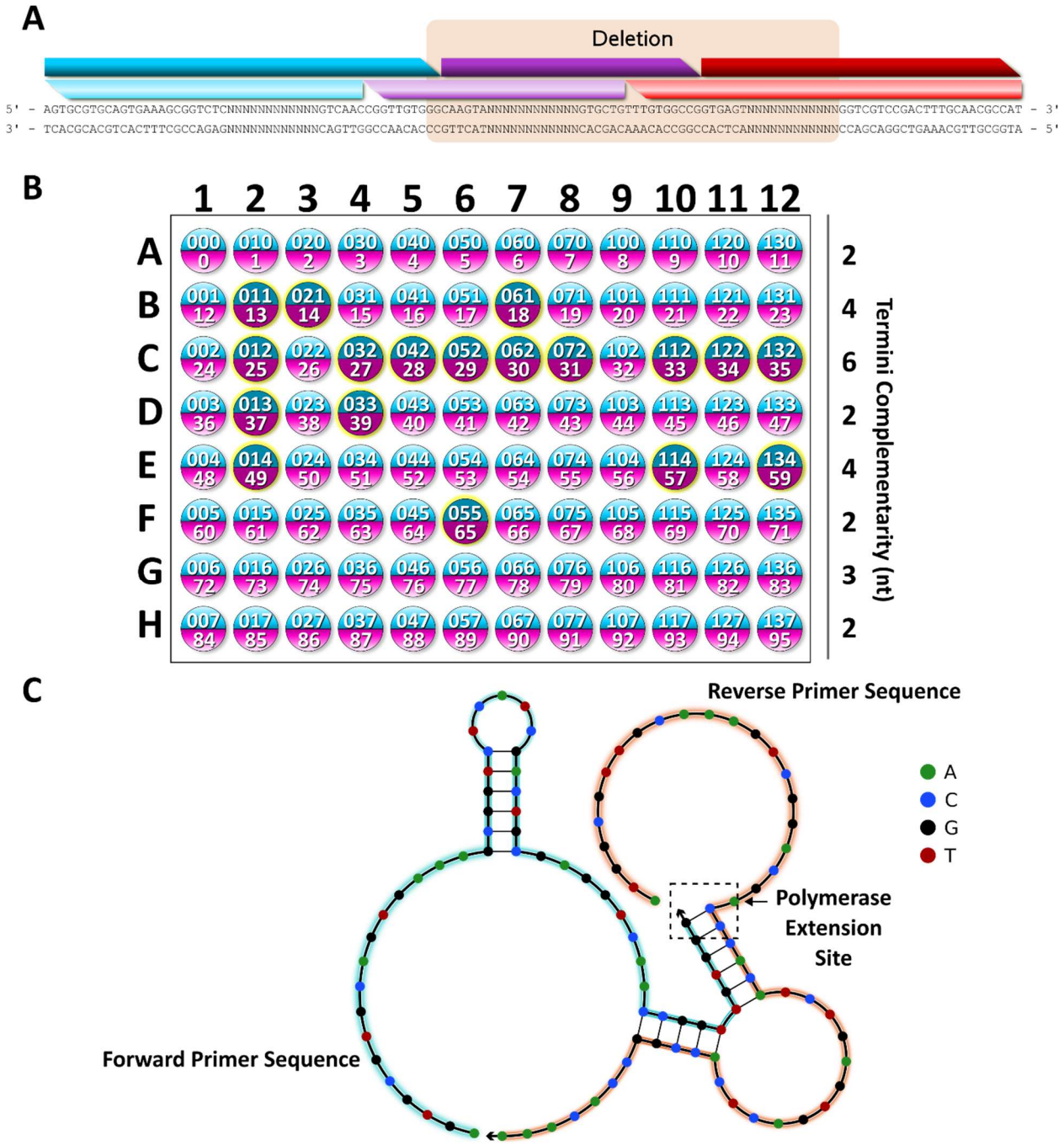



**Figure S4.5. Gen_2 Deletion Variant Analysis. A)** Flawed reads were very rare in the pooled 96 barcodes from the Gen_2 library (main text Fig. 6). Here we sought to better understand the origin of the fifth most common deletion variant (Fig. 6 row 5). The deletion spanned blocks 2 and 3 such that the only remaining variable region sequence originated in block 1. Of the 130 reads detected, 125 reads had 'intact' TrapTag sequences (e.g., no indels or substitutions), which were cross-referenced to the 96-well plate used for oligo mixing for discerning the origin of the off-target assembly. **B)** Layout of the 96 well plate configuration used for appending unique TrapTags to each of the 96 barcodes. For each well, the top 3 digits correspond to the "barcode", the variable region indices for each of the 3 blocks. The bottom digit corresponds to the TrapTag index UMI. Darker shaded wells indicate originating locations

for the off-target assembly shown in A. Thus, this rare off-target assembly nonetheless occurred in at least 18 distinct wells. Approximately 25% of all deletion variant reads resulted from well C4, containing block 1 variant index 0, block 2 variant index 3, and block 3 variant index 2 (i.e. barcode “032”). **C)** The predicted secondary structure of the block 1 variant index 0 top strand (cyan strand background) with the block 3 variant index 2 bottom strand (orange strand background) by NUPACK at the annealing temperature (58 °C) employed for barcode amplification. Notably, the six 3’ terminal nucleotides for the block 1 strand (TGTGGG) have the potential to anneal to the last two bases in constant region 4 (CC) followed by the first four bases in the variable region sequence of the block 3 strand (CACA), allowing polymerase extension during PCR, resulting in formation of the detected deletion variant. This off-target 6-bp complementarity can occur between the third variant (index 2) for block 3 (which contains CACATCTGAGTG in the reverse strand) and any block 1 variant (which all end with TGTGGG) , which explains why the third row of the barcode assembly plate is the dominant source. In aggregate, 116 out of the 125 traceable reads (93%) originated in row C, corresponding to barcodes **2. The other two rows with the most unwanted block1 to block3 pairing correspond to block3 variant indices 1 and 4, which both feature a 4-bp complementary region since their variable regions start with CA in the reverse direction. The total size of the Watson-Crick complementarity region between the 3’ terminus of forward block 1 and the reverse block 3 variants (that would lead to the observed deletion variant) is listed to the right of the corresponding row in the panel B table. Elimination of this off-target assembly may be a useful negative design principle for future combinatorial barcode library design.

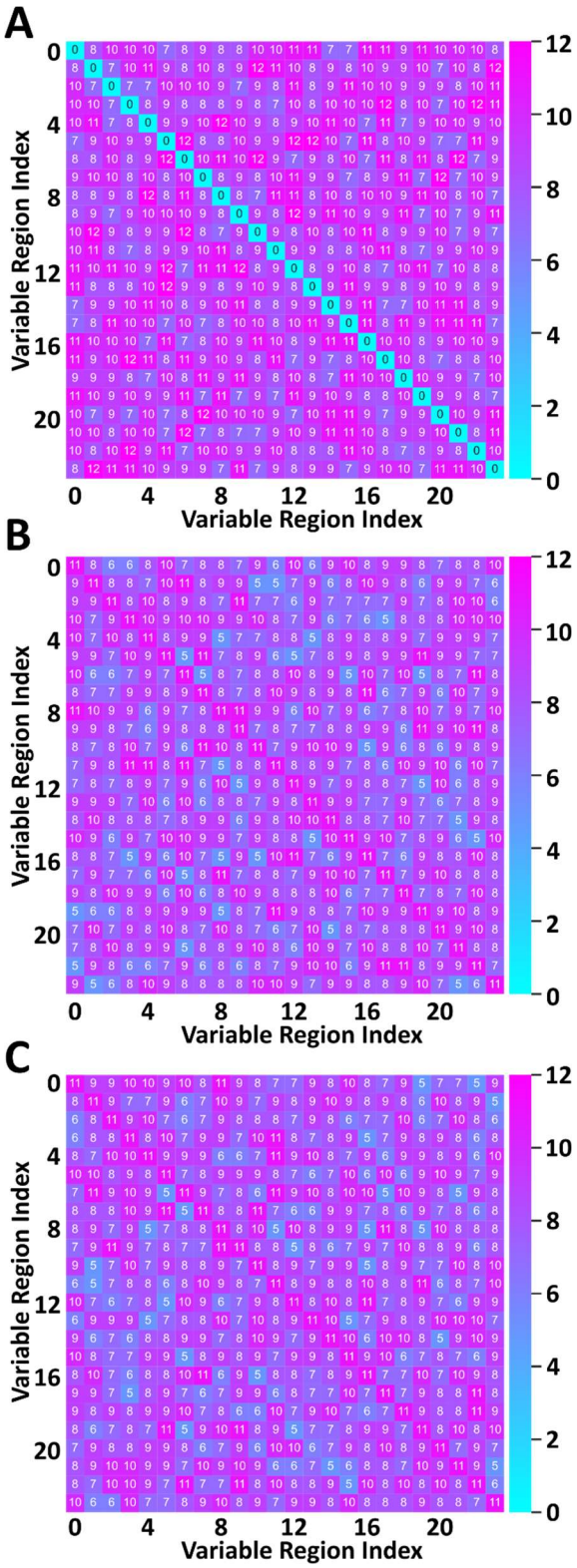



**Figure S4.6. Variable Region Hamming Distance. (A)** Heat map corresponding to the Hamming Distance results between all 24 designed variable region sequences. **(B, C)** Hamming distance results between all variable region sequences containing 1-nt insertions and 1-nt deletions, respectively.

| Block | Block Variant | Variable Region Index | Variable Region Sequence (5' - 3'), top strand | Variable Region Sequence (5' - 3'), bottom strand |
|---|---|---|---|---|
| 1 | 0 | 0 | ATCGACTGCGAG | CTCGCAGTCGAT |
| | 1 | 1 | GCTAGCACTGAG | CTCAGTGCTAGC |
| | 2 | 2 | GTGTGCGCTAGC | GCTAGCGCACAC |
| | 3 | 3 | TGCTCTAGTAGC | GCTACTAGAGCA |
| | 4 | 4 | CGATACGAGATC | GATCTCGTATCG |
| | 5 | 5 | ACTGAGTGTCTC | GAGACACTCAGT |
| | 6 | 6 | TGCAGTGACTAG | CTAGTCACTGCA |
| | 7 | 7 | AGCGTGACGCGT | ACGCGTCACGCT |
| 2 | 0 | 8 | ATGACGAGTGCT | AGCACTCGTCAT |
| | 1 | 9 | GAGATCTGCAGT | ACTGCAGATCTC |
| | 2 | 10 | CTATCGCGACGT | ACGTCGCGATAG |
| | 3 | 11 | CATGCTGTCAGC | GCTGACAGCATG |
| | 4 | 12 | CGACGTCTATCG | CGATAGACGTCG |
| | 5 | 13 | GAGTCTACGTCG | CGACGTAGACTC |
| | 6 | 14 | GTCGCAGTACAG | CTGTACTGCGAC |
| | 7 | 15 | ACAGTGATCGAC | GTCGATCACTGT |
| 3 | 0 | 16 | TGTCTCGAGTCT | AGACTCGAGACA |
| | 1 | 17 | GCGCTGCTACTG | CAGTAGCAGCGC |
| | 2 | 18 | CACTCAGATGTG | CACATCTGAGTG |
| | 3 | 19 | GATGTGCAGAGA | TCTCTGCACATC |
| | 4 | 20 | TCTCGTCGTATG | CATACGACGAGA |
| | 5 | 21 | CAGCAGTCTCGT | ACGAGACTGCTG |
| | 6 | 22 | CAGAGACAGCAG | CTGCTGTCTCTG |
| | 7 | 23 | ATAGCGCACTCA | TGAGTGCGCTAT |

**Table S4.3.** Assigned index values representing individual variable region sequences for Hamming distance analysis.

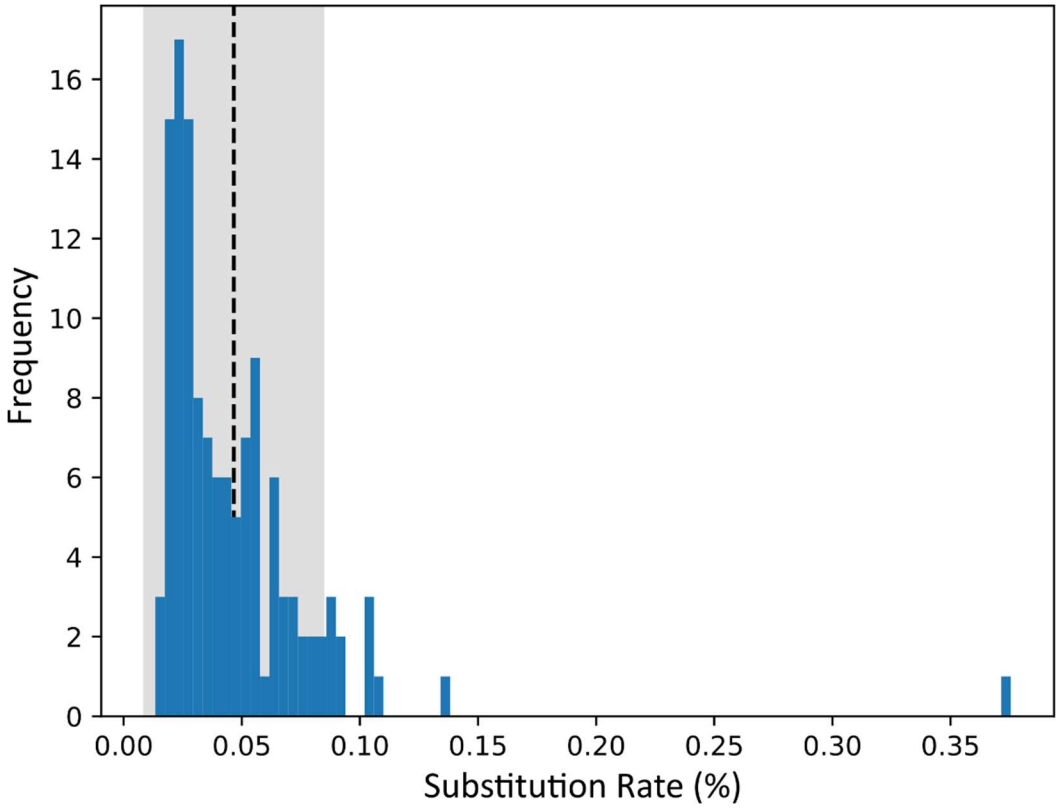



**Figure S4.7. Histogram of 1-nt Substitutions.** The dashed line represents the average substitution rate (0.05%). The gray shaded region represents ± the standard deviation (0.04%).

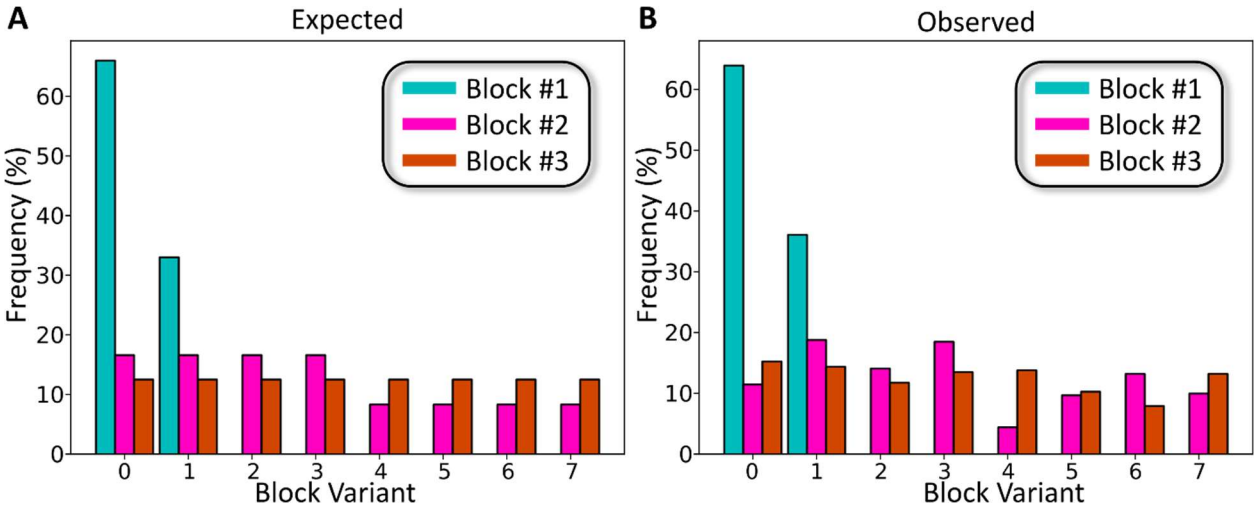



**Figure S4.8.** Here we assess the 341 pooled reads that constitute the most common deletion variant for the Gen_2 library (row 1 in main text Figure 6). **A)** Given the 96 barcode synthesis targets in play (Fig. S3B) we can compute the expected distribution of block variants for the pooled 96 barcodes assuming that an equimolar population of all barcodes was subjected to NGS and that the chance of a truncated NGS read does not depend on the barcode sequence. **B)** The observed distribution of block variants detected from the most common deletion variant for the Gen_2 library. The x-axis denotes the block variant and the y-axis denotes detection frequency as percent. The expected and observed frequencies are reasonably similar, consistent with the idea that the most common barcode read imperfection is simply a random endemic NGS artifact rather than a flaw in the barcode design and synthesis.

**EXTENDED METHODS**

**Primer Design**

Primer sequences for both Gen_1 and Gen_2 modular barcode libraries were designed using Primer3 (89). Specifically, a randomly generated 100,000nt sequence using an online sequence generator (110) was the input for the primer design code. Design parameters included primer sets with length of 18nt, melting temperature of 60 °C, GC content of 40 – 60% and a GC clamp of 2nt. The output designed 20 primer pairs were further sorted based on the sum of the following penalties assigned by Primer3: primer left self any TH, primer right self any TH, primer left self end TH, primer right self end TH, primer left hairpin TH, primer right hairpin TH, primer left end stability, primer right end stability, primer pair complementarity any TH, primer pair complementarity end TH. The primer pair with the lowest sum of penalties was chosen as barcode primer set.

**Automated Primer Specificity Check**

Custom Python code was written for checking for potential off-target amplification of candidate contaminating species (Culicidae, Homo sapiens). A spreadsheet of candidate primer pairs designed using Primer3 served as the input. After loading in the spreadsheet, the Biopython package is used for performing a blast search against the specified species for each primer pair. The program then parses the blast output files (xml) searching for instances where a primer pair align on opposing strands (requisite for exponential amplification) of an organism's genome and calculates the distance between that possible amplicon. Amplicon lengths much greater than the target barcode length ( > 1kbp) were treated as non-threatening from a barcode detection/identification perspective. The results for each primer pair were written out to separate sub-directories for further analysis.

***in vitro* Primer Sensitivity**

The barcode G-Block (5' – AGTGCGTGCAGTGAAAGCGGTCTCATCGACTGCGAGGTCAACCGGTTGTGGGC AAGTAATGACGAGTGCTGTGCTGTTTGTGGCCGGTGAGTTGTCTCGAGTCTGGTCGTCCGACTTTGCAACGCCAT – 3') was rehydrated to a 10ng/µL solution. 25µL of water was added to 250ng dried barcode. This solution was diluted $10^{-2}$ by adding 1µL of barcode solution to 99µL of water. This $10^{-2}$ solution was further diluted by adding 10µL of solution to 90µL of water and this serial dilution was repeated until the $10^{-10}$ dilution was achieved for each barcode. PCR was set up with primer set #1 (Table 5 of the main text) using the following master mix for each reaction. 25µL of GoTaq Green 2xMM, 2.5µL of 10µM forward primer, 2.5µL of 10µM reverse primer, 17µL of water, and 3µL of DNA template. 8 Reactions were prepared with the first 7 reactions using the previously prepared $10^{-4}$ through the $10^{-10}$ barcode as the template. The last remaining reaction used 3µL of water as the template for a negative control. Thermocycling of the two sets took place in tandem with the following conditions:

1. 96.0C° for 2:00min
2. 96.0C° for 20sec
3. 60.0C° for 20sec
4. 72.0C° for 30sec
5. Go to step 2 39 times
6. 72.0C° for 5:00min
7. 4.0C° for ∞

PCR product was run via gel electrophoresis on a 2% agarose gel in 1X TAE buffer at 90V for 60 minutes.

***in vitro* Primer Specificity**

The designed primers and barcode template were tested for specificity by running them against a panel of samples known to be negative for barcode. PCR was set up for primer set #1 (Table 5 of the main text) using the following master mix for each reaction: 25µL of GoTaq Green 2xMM, 2.5µL of 10µM NDX3-iseqFWD forward primer, 2.5µL of 10µM NDX3-Rev reverse primer, 17µL of water, and 3µL of DNA template. 9 reactions were prepared for each reaction set and the template samples for the first 8 were the same for each. Sample 1 a pool of 15 lab reared *Aedes aegypti* mosquitoes processed via the mosquito processing and extraction protocol. Sample 2 a pool of 15 lab reared *Culex tarsalis* mosquitoes processed via the mosquito processing and extraction protocol. Sample 3 was an aliquot of human saliva processed using the extraction protocol. Sample 4 water processed using the extraction protocol. Sample 5 a pool of 15 lab reared *Culex tarsalis* mosquitoes processed via the mosquito processing and extraction protocol. Sample 6 a pool of 15 lab reared *Aedes aegypti* mosquitoes processed via the mosquito processing and extraction protocol. Sample 7 was an aliquot of human saliva processed using the extraction protocol. Sample 8 water processed using the extraction protocol. Sample 9 was $10^{-9}$ diluted barcode.

Thermocycling of the two sets took place in tandem with the following conditions:

8. 96.0C° for 2:00min
9. 96.0C° for 20sec
10. 60.0C° for 20sec
11. 72.0C° for 30sec
12. Go to step 2 39 times
13. 72.0C° for 5:00min
14. 4.0C° for ∞

PCR product was run via gel electrophoresis on a 2% agarose gel in 1X TAE buffer at 90V for 60 minutes.