bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Random adversarial threshold search enables automated DNA screening

Dana Gretton[1,†], Brian Wang[1,†], Rey Edison[1], Leonard Foner[2], Jens Berlips[2], Theia Vogel[2], Martin Kysel[2], Walther Chen[2], Francesca Sage-Ling[2], Lynn Van Hauwe[2], Stephen Wooster[2], Benjamin Weinstein-Raun[2], Erika A. DeBenedictis[1,3], Andrew B. Liu[4], Emma Chory[1], Hongrui Cui[5], Xiang Li[6], Jiangbin Dong[6], Andres Fabrega[1], Christianne Dennison[1], Otilia Don[1], Cassandra Tong Ye[1], Kaveri Uberoy[1], Ronald L. Rivest[7], Mingyu Gao[6,9], Yu Yu[5,9], Carsten Baum[8,10], Ivan Damgard[8], Andrew C. Yao[2,6,9], and Kevin M. Esvelt[1,2,*]

[1]Media Lab, Massachusetts Institute of Technology, USA
[2]SecureDNA Foundation, Switzerland
[3]Department of Bioengineering, Massachusetts Institute of Technology, USA
[4]Department of Systems Biology, Harvard Medical School, USA
[5]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[6]Institute for Interdisciplinary Information Sciences, Tsinghua University, China
[7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA
[8]Department of Computer Science, Aarhus University, Denmark
[9]Shanghai Qi Zhi Institute, China
[10]DTU Compute, Technical University of Denmark, Denmark
[†]equal contribution
*esvelt@mit.edu

Summary

Searching for exact matches to pre-computed functional variants unique to hazardous genes enables sensitive, secure, and automated DNA synthesis screening.

Abstract

Custom DNA synthesis underpins modern biology, but hazardous genes in the wrong hands could threaten many lives and public trust in science. In 1992, a virology-trained mass murderer tried and failed to obtain physical samples of Ebola; today, viruses can be assembled from synthetic DNA fragments. Screening orders for hazards is unreliable and expensive because similarity search algorithms yield false alarms requiring expert human review. Here we develop "random adversarial threshold" (RAT) search, which looks for exact matches to short nucleic acid and peptide subsequence windows from hazards and predicted functional variants that aren't found in any known innocuous genes. To experimentally assess sensitivity, we used RAT search to protect nine windows from the M13 bacteriophage virus, then invited a "red team" to launch up to 21,000 attacks at each window and measure the fitness of their designed mutants. We identified defensible windows from regulated pathogens, built a curated test database that our M13 experiments indicate will block 99.999% of functional attacks, and verified its sensitivity against orders designed to evade detection. RAT search offers a way to safeguard biotechnology by securely automating DNA synthesis screening.

Introduction

The COVID-19 pandemic demonstrated that society is profoundly vulnerable to new transmissible biological agents, even as virus assembly protocols and inexpensive *de novo* DNA synthesis have made harmful pathogens accessible to a large and growing number of technically skilled individuals[1–4]. Recent publications strongly suggest that future advances will provide genomic blueprints and step-by-step reverse genetics protocols for credible pandemic agents[4–13].

Fortunately, most individuals skilled enough to assemble viruses with reverse genetics cannot synthesize DNA on their own. Members of the International Gene Synthesis Consortium (IGSC), an industry group, are committed to screening DNA synthesis orders above a certain length[14].

The IGSC deserves praise for voluntarily prioritizing safety because doing so is costly: traditional screening methods based on BLAST generate false alarms that require human expert review[15,16]. As the price of synthetic DNA falls, the effective cost of screening grows[15]. Unfortunately, more than two thirds of gene synthesis firms are non-members. A modern equivalent of 1992's virology-trained terrorist[17] can plausibly obtain potential pandemic viruses by ordering the DNA from a company not listed on the IGSC website (Fig. 1a).

Even if all providers did screen requests, adversaries could obtain DNA sufficient to generate a pandemic virus by assembling oligonucleotides shorter than the minimum length for screening; by ordering non-overlapping pieces from multiple suppliers; or by

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

swiftly placing orders for newly identified pandemic viruses to exploit databases that are slow to update, among other attacks (Fig. 1a). Recent red-teaming confirmed many of these weaknesses[18]. Benchtop machines enabling on-site synthesis create another vulnerability[15,19,20]: not only must each device be updated whenever a new threat is identified, but the hazard identification system cannot be stored locally on the device lest it be interrogated and used to build software allowing others to obtain hazards undetected.

Verifiably screening all commercial and benchtop DNA synthesis for current and emerging hazards[21] demands a fully automated, centralized, and privacy-preserving approach that detects short hazardous sequences while triggering negligibly few false alarms. Current algorithms based on sequence alignment, pattern recognition, or exact subsequence matching are insufficiently sensitive, specific, and efficient for universal screening (Fig. 1b-c).

We hypothesized that the unique signature of a hazard can be approximated by compiling all short subsequence windows, predicting functional variants of pseudo-randomly chosen windows to enhance sensitivity, and removing any window that matches a known harmless sequence to ensure specificity. Searching for exact matches to this approximated signature is compatible with privacy-preserving cryptography. Here we describe and experimentally assess the sensitivity of "random adversarial threshold" (RAT) search; our companion paper analyzes specificity and the performance of a cryptographic implementation.

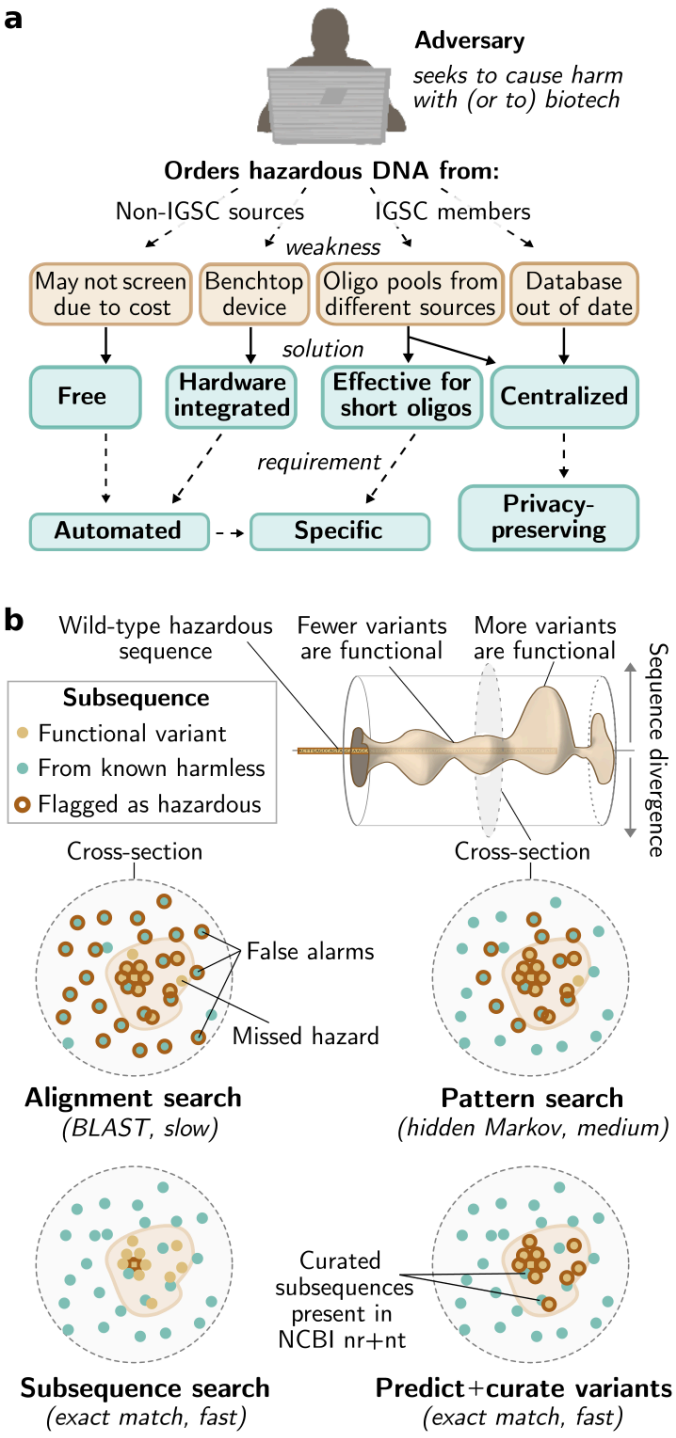


Figure 1 | Achieving robust DNA synthesis screening
a) Adversaries can evade current screening by exploiting weaknesses. Not all providers screen due to the cost, benchtop devices cannot wait for human review, and decentralized sources are vulnerable to split orders and failures to update. Centralized screening can only preserve order privacy with cryptography, which requires a highly specific and efficient search algorithm. **b)** Reliable screening requires sensitively detecting functional variants of hazards without flagging similar subsequences from harmless relatives. Alignment and pattern-based search find similar sequences, but generate false alarms and can miss functional equivalents. Exact match search can only detect wild-type hazards. Predicting functional variants of randomly chosen subsequences and curating to remove harmless matches can detect evasive attempts and avoid false alarms. **c)** In principle, combining exact match search with functional prediction and curation is specific enough to be automated, sensitive enough to thwart adversaries, and efficient enough for cryptographic methods to protect the privacy of both DNA synthesis orders and the entries in the hazard database[22].

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results

Suppose that an adversary seeks to obtain a protected hazard W by incorporating mutations to evade RAT search (Fig. 2b). For each mutated subsequence window w_i, there are three possible outcomes:

1. w_i is present in the hazards database, and the synthesis order is rejected and logged
2. w_i escapes detection, but imposes a fitness cost c_i that reduces functionality
3. w_i escapes detection at negligible cost

Success requires the adversary to achieve the third outcome for most w_i to preserve function. We define the *random adversarial threshold R* as the probability that an adversary with perfect knowledge of the fitness of each variant – but ignorant of which windows and variants are defended – will be detected upon attempting to synthesize functional W.

In theory, defending all variants that do not completely abolish the function of W at an essential window w_i can perfectly thwart the adversary, achieving R=1. In practice, fitness prediction is imperfect, but R can still be maximized by defending windows predicted to be least tolerant of mutations and adding new windows when an attempt is detected (Fig. 2b, Supplementary Fig. 1).

Importantly, an adversary with superior predictive capacity who learns which function-prediction algorithms are used for RAT search can evade screening by choosing the highest-fitness undefended variant known to them for each window (Fig. 2c). By choosing which windows and variants to defend quasi-randomly, we can force the adversary to heavily mutate all windows throughout the hazard in order to evade detection, greatly reducing their odds of obtaining functional W.

Choosing a window size while maintaining specificity

Before experimentally measuring R, we need to know how many variants we can defend without flagging too many innocuous sequences. A key benefit of exact-match screening is the ability to curate the database by removing all peptides and k-mers that match harmless and/or unrelated sequences from sequence repositories (Fig. 3, Supplementary Fig. 2). We distinguish such sequences from hazards and close relatives using taxonomic classification, keywords, and counting the number of windows that match the hazard, among others. As our companion paper demonstrates, RAT search using a large curated database will seldom if ever flag unrelated harmless sequences.

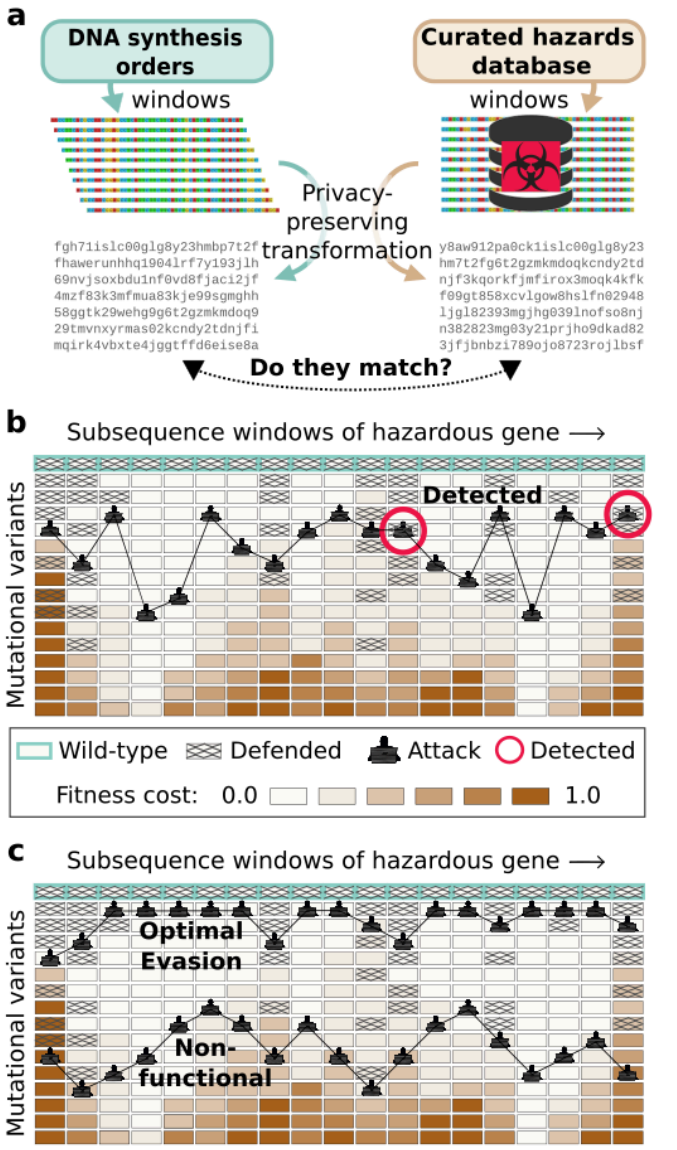


Figure 2 | Random adversarial threshold search

a) Screening relies on detecting matches between subsequences from DNA synthesis orders and from hazards. Efficient exact-match search permits the use of cryptography to preserve the privacy of both orders and hazards[23].
b) To evade screening, the adversary must choose a mutated subsequence for every window across the coding sequence of the hazard without being detected, as shown. This is maximally challenging when the hazards database is populated with predicted functional variants of windows that tolerate few mutations. The "random adversarial threshold" is the probability that an adversary with perfect knowledge of the fitness landscape will be detected. Detection becomes more probable as the defender's predictive capacity increases relative to the attacker's. **c)** An attacker who knows exactly which windows are defended can evade imperfect screening. Choosing windows and variants quasi-randomly forces an imperfect attacker to guess, likely including so many mutations that the resulting hazard is no longer functional.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

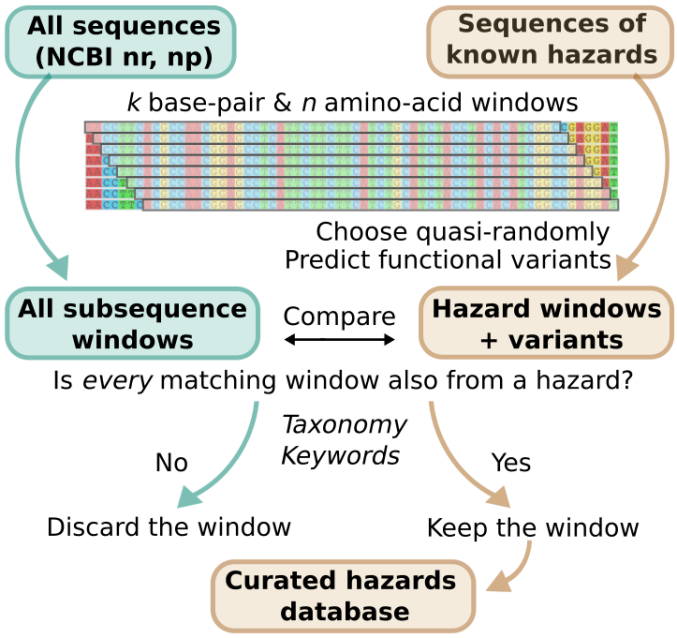


Figure 3 | Building and using the hazards database
To generate and curate a database of hazards for random adversarial threshold screening, subsequences from hazards and predicted functional variants are compared to similarly-sized windows from NCBI repositories. Matches to repository sequences that are not hazardous – as determined by taxonomy, keywords, and fraction of matching windows – are discarded to reduce the false alarm rate.

However, novel sequences absent from natural organisms – which are commonly employed in deep mutational scanning and directed evolution experiments as well as designed proteins – will still randomly trigger false alarms. Random false alarms will occur at a frequency determined by the total amount of DNA synthesized, the number of sequences in the database, and the window length (Appendix A).

Historical efficiency improvements[24] and market projections suggest that global annual synthesis demand may rise to as much as 10^{15} base pairs in a decade. While almost certainly an overestimate, this is counterbalanced by the fact that functional biopolymer sequences are *not* randomly distributed[25]: peptide frequencies are highly biased by amino acid composition and functional constraints[26–28]. For our initial experiments aimed at measuring R, we chose to screen peptides of length 19 because searching for 10^7 functional variants for each of 1,000 hazards would yield one truly random false alarm per 10^{15} base pairs of DNA.

Experimentally testing sensitivity

To test the efficacy of RAT screening against deliberately introduced mutations and learn which types of windows are most easily defended, we selected the harmless M13 virus that infects *E. coli* as a "hazard". Lacking a reliable multiple-mutation variant effect predictor at the time[29], we chose a variety of peptide windows with properties that might be relevant to defense by analyzing all length 19 peptide subsequences using fuNTRp, a computational tool that categorizes residues within proteins as "neutral" if they likely tolerate most mutations, "rheostat" if they suffer reduced fitness from many but not all mutations, or "toggle" if nearly any mutation is deleterious[30]. From four required M13 proteins (Supplementary Fig. 3), we selected nine total windows with fairly low to very low neutral values and a range of rheostat and toggle scores (Supplementary Fig. 4, Supplementary Table 1).

Next, two "blue team" members constructed databases of 10^3 to 10^7 predicted functional variants for each window using a Metropolis-Hastings algorithm[31] that combined the fuNTRp scores of each residue with the BLOSUM62 matrix of observed substitutions across proteins. While recent protein design tools are markedly superior to BLOSUM62[32,33], our method serves as a baseline for defensive efficacy tests that can be substantially improved upon.

"Red team" members experimentally tested the security of RAT search by designing and launching up to 21,000 attacks at each of the nine windows using combinatorial rational design (Fig. 4a). They chose to order oligonucleotide pools with all possible combinations of the four most common substitutions at the six positions with the highest neutral scores, pairwise substitutions of all amino acids at those six positions, and all possible single substitutions. They generated libraries of variants by molecular cloning and measured the effects of each variant on phagemid replicative fitness via complementation.

We defined "functional" variants as those with a measured fitness of at least 0.05 relative to wild-type, which is the level at which the most infectious virus known can no longer spread in an unprotected population[34]. Since approximately 50% of variants chosen by the attackers met this standard, the red team effectively launched ~10^{33} combinatorial, individually functional attacks on the nine windows in the database.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Of the attacks on the most defensible window, 85.2% were detected by a database with 10^6 entries (Fig. 4b). That is, even an adversary with perfect knowledge of subsequence fitness who already possesses the other 99% of the wild-type M13 genome sequence was likely to be detected and thwarted at just this one window. While other windows were less defensible, most still blocked ~30-50% of attacks (Fig. 4c).

These results underscore the extreme difficulty of obtaining a functional hazard by incorporating mutations to evade RAT search. When the nine M13 windows were each defended by 100,000, 1 million, or 10 million variants, 99.3%, 99.7%, and 99.9% of individually functional attacks were detected at one or more windows (Fig. 4d).

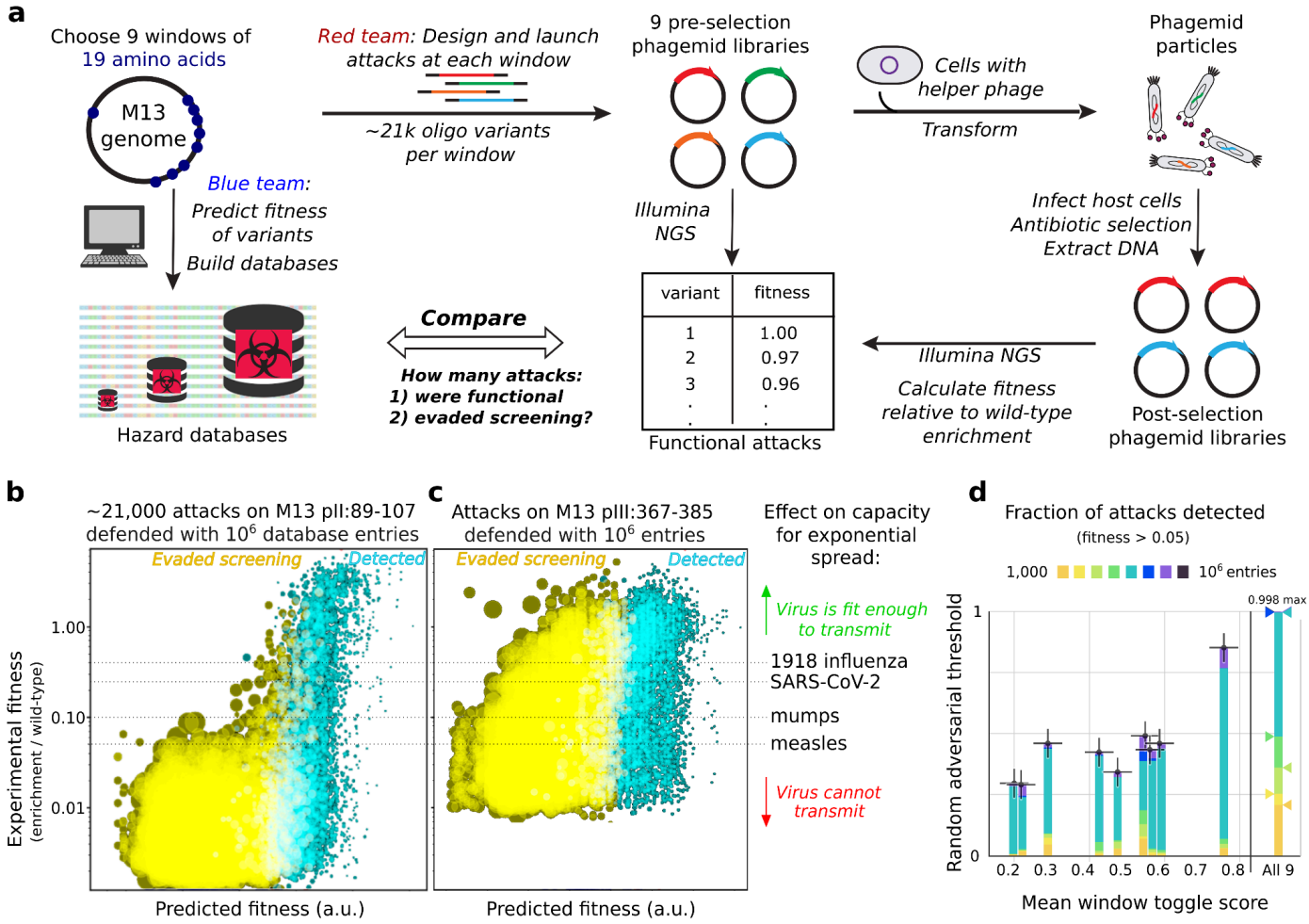


Figure 4 | Incorporating mutations into the genomic blueprint of a virus cannot readily escape screening. a) Team members built defensive databases by predicting functional variants for nine different windows in the genome of M13 bacteriophage. Others launched ~21,000 attacks at each window by synthesizing variants with up to six amino acid changes and using a phagemid assay to measure the fitness of each variant, which we defined as enrichment relative to the wild-type sequence. **b)** At the most defensible window, located within the M13 pII endonuclease, 92% of attacks yielding variants with fitness above 0.05 were thwarted by screening. Smaller dots correspond to sequences with fewer mutations. **c)** At a moderately defensible window located within the M13 pIII receptor-binding protein, 49% of such attacks were thwarted, underscoring the importance of window choice. Potential pandemic pathogens can tolerate only so many mutations impairing fitness before they are no longer capable of sustained transmission. The corresponding fitness lines depict these threshold values for 1918 influenza (R_o~2.5), SARS-CoV-2 (R_o~4), mumps (R_o~10), and measles (R_o~18), which is the most contagious virus known. **d)** The fraction of attacks detected, which corresponds to the random adversarial threshold, as a function of the average fuNTRp toggle score for each of the nine windows for various database sizes (1000, 2000, 5000, 10^4, 5×10^4, 10^5, 5×10^5, 10^6) using a fitness cutoff of 0.05 (sufficient to prevent the sustained spread of measles). Combinatorial screening of all nine windows (right) detected virtually 100% of individually functional attacks. Data

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Since the red team knew exactly which windows to mutate and was able to generate libraries to find functional variants without penalty, these results strongly suggest that R can approach 1.0 unless the attacker has notably superior predictive capability.

Moreover, windows do not exist in isolation: an attack featuring mildly deleterious genomic mutations in two different windows that separately reduce fitness to 0.20 typically has a *combined* fitness of 0.04 (or less)[35,36], and consequently will not be functional (Supplementary Fig. 5). We simulated combinatorial attacks by randomly combining functional variants at each window into 10^{10} M13 phage genomes, multiplying the fitness values for all nine windows, and discarding those with a combined fitness below 0.05. Because more high-fitness windows were selected on average, and these are more readily predicted (Supplementary Fig. 6), defending just 50,000 variants per window successfully detected 99.94% of combinatorially functional attacks by the red team.

As expected, some windows were easier to defend than others. The fuNTRp "toggle" score, or predicted sensitivity to mutation at each position[30], appears most predictive of *R* for each of the nine windows (Supplementary Fig. 8). In principle, analyzing the false positive and false negative rates for toggle-based prediction at each window allows us to estimate the optimal number of database entries to include in order to defend still more effectively (Supplementary Fig. 9).

Most importantly, defending 20 windows equivalent to the geometric mean of the nine from M13 with just 50,000 database entries per hazard would block 99.999% of attacks seeking to obtain a pandemic virus as contagious as anything known to science (Fig. 5a, Supplementary Fig. 7).

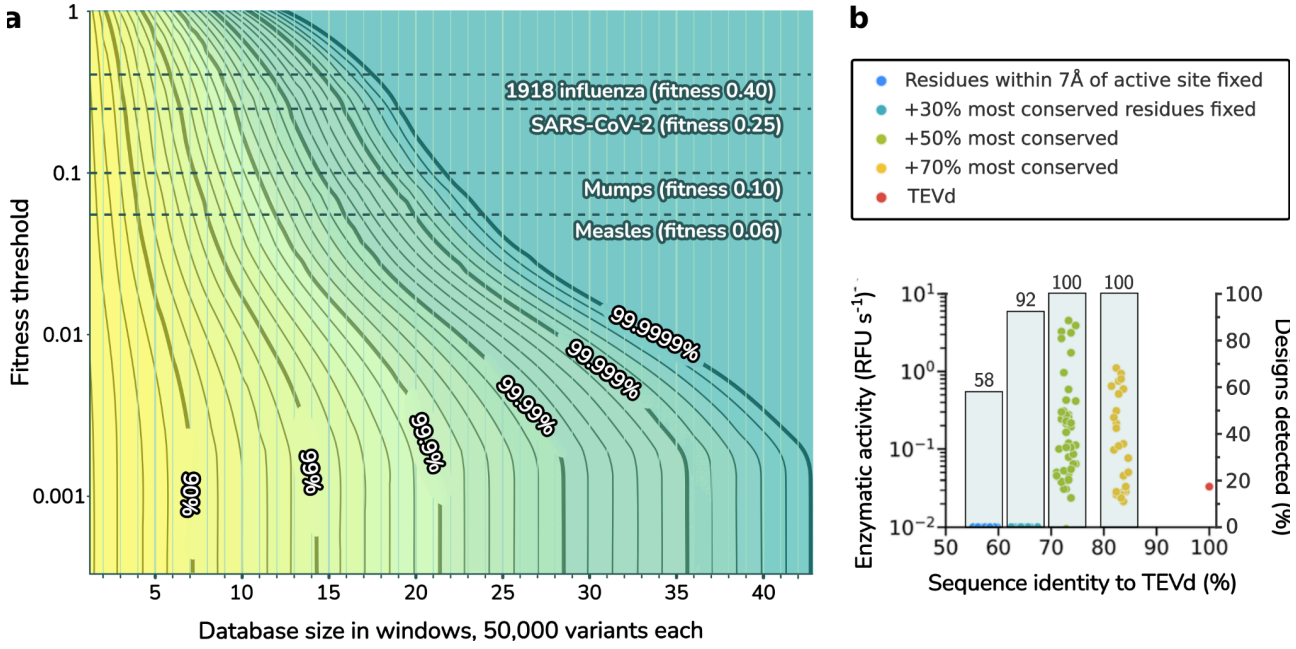


Figure 5 | Adversarial and machine-learning-generated designs are reliably detected. **a)** The cumulative effect of safeguarding more or fewer windows can be extrapolated using powers of the geometric mean of escapee curves (Supplementary Figs. 6-7). Random Adversarial Threshold is plotted as a function of both simulated number of windows protected with 50,000 variants, and fitness at which the hazard is no longer functional. Contours show lines of equal protection as trade-offs between the minimum fitness tolerated and number of windows protected. As an illustration, 16 mean windows would block 99.99% of attempts to synthesize and spread a virus as contagious as measles, or 99.999% of attempts to make a viable pathogen as infectious as SARS-CoV-2. In practice, the database includes 100 or more windows for each hazardous virus. Future prediction tools may influence these results if not incorporated. **b)** Detection of redesigned functional and nonfunctional TEVd protease variants generated with ProteinMPNN, grouped by identity to the wild-type enzyme[37]. All functional redesigns were detected despite the use of inferior variant prediction tools. Future databases will incorporate ProteinMPNN and other tools for prediction.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Defensibility of known hazards

To estimate how many sufficiently defensible windows are available, we analyzed the genomes of all viruses, microbial pathogens, and known genes encoding toxins or pathogenicity islands from the Australia Group controlled pathogen list for mean window toggle scores using fuNTRp. All viruses except swine vesicular disease virus harbored numerous windows with scores above 0.5, each of which blocked over one-third of attacks in our phagemid experiments when defended with 10^6 variants (Supplementary Table 1). Since there are fewer than 100 Australia Group pathogens with publicly available genes or genomes, we can afford to include variants for dozens or even hundreds of windows per public hazard without risking an unacceptable number of false positives. This effectively blocks the synthesis of all close relatives of such hazards even though they were not directly included in the database.

Specificity and sensitivity against adversarial attacks

Having established the robustness of random adversarial threshold search against evasive strategies focused on introducing mutations, we sought to determine the effectiveness of RAT screening against a variety of attacks that defeat both alignment and exact-match search. We built and curated a hazards database from the U.S. Select Agents and Australia Group pathogens (Methods). To prevent adversaries from ordering and assembling short oligonucleotides, we additionally included all 30-mer DNAs as well as 42-mers with all single mutations.

As a preliminary test of specificity, we screened the 2.47-gigabase Clustered Reference Viral DataBase (C-RVDBv24.0) against the bacterial and toxin sequences of our hazard database and manually investigated each hit using BLAST. All matches consisted of toxicity-conferring genes encoded in viral genomes, suggesting that RAT search can be extremely specific. In the companion paper, we report the results of subsequent specificity testing on real-world DNA synthesis orders[23].

To test the sensitivity of RAT search against attempted evasion, we wrote scripts that used several distinct strategies to design gene synthesis orders for hazardous sequences that reliably evade both BLAST and straightforward exact-match screening for 50-mer nucleotides and 16-amino-acid peptides, yet can be assembled into functional hazards in no more than two laboratory steps using standard protocols. Testing revealed that our straightforward implementation of random adversarial threshold search detected every hazardous sequence and passed every harmless order.

Robustness against protein design tools

In principle, machine learning tools may be capable of altering enough residues in hazardous proteins to evade screening while preserving function. To challenge RAT search against the current state of the art for enzyme redesign, we built a hazard database to defend TEV protease, which was recently redesigned with ProteinMPNN[37], and screened all reported designs. Fully 100% of designs with nonzero activity were detected, despite the comparative crudity of fuNTRp+BLOSUM62 prediction (Fig. 5b). Incorporating more advanced design tools into the database generation pipeline may allow screening to detect redesigned threats even more reliably.

Discussion

DNA synthesis screening can help prevent unauthorized individuals from obtaining pandemic-capable agents and other biohazards capable of killing thousands or even millions of people, but the technical limitations inherent to homology search have precluded effective implementation. The success of Kraken and related tools for sequence comparison suggests that exact-match search could better safeguard biotechnology if rendered sufficiently sensitive to deliberate attempts at evasion[38].

Our results demonstrate that predicting functional variants of randomly chosen subsequence windows from hazards, curating them to remove any that match harmless sequences from repositories, and searching DNA synthesis orders for exact matches is more sensitive and efficient than current screening algorithms.

In the companion paper, we describe the design, development, and performance of a free and automated cryptographic screening system based on RAT search that preserves the privacy of orders and database entries[23]. Given suitable legal incentives from clarified liability frameworks and regulatory requirements, it could eventually become universal, including in benchtop devices. U.S. Executive Order 14110, which requires federally funded institutions to purchase DNA from

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

synthesis firms that screen orders in ways that stand up to red-teaming, may provide a market-based impetus for near-universal adoption that could be strengthened through international cooperation. Once the security of the database has been tested via prize competitions, SecureDNA could even be used to screen for emerging hazards without disclosing their nature and highlighting their credible potential for misuse[22].

By providing a way to securely automate DNA synthesis screening, random adversarial threshold search can substantially mitigate the catastrophic risk posed by increasingly widespread access to pandemic-class biological agents.

Methods

Window selection: M13 peptides

The proteins of bacteriophage M13 (Accession NC_003287.2) were analyzed using fuNTRp[30] , which scores each amino acid position by the likelihood that it will accept many (neutral), some (rheostat), or few (toggle) different substitutions without disrupting protein function. We used fuNTRp to identify peptide windows with few predicted neutral positions and varying numbers of toggle and rheostat positions across and within proteins (Supplementary Table 1).

Blue team: Defending the nine windows of M13

All single mutations were included for reasons of caution. For the remaining entries, a Metropolis-Hastings algorithm was used to select combinations of mutations predicted to minimally impair fitness. The mutation tolerance scores estimated by fuNTRp were combined with the BLOSUM62 matrix to generate a probable cost for every possible substitution, with costs multiplied for multiple-mutation combinations.

Red team: Procedurally generating variants for each 19aa peptide window

1. We included the wild-type sequence (1)
2. We included all one-mutants at each position (19*19 = 361)
3. At the six positions predicted to be most neutral, we added all combinations of the four predicted least pathological substitutions according to BLOSUM62 (5^6 = 15625) (overlaps with one-muts and WT at 4*6+1=25)
4. As negative controls (not attacks), we included up-to-six mutants of neutral positions using the two most pathological substitutions according to BLOSUM62 (3^6 = 729) (overlaps with one-muts and WT at 2*6+1=13)
5. We added all pairwise combinations of all possible substitutions at the six most neutral positions (19^2 * 15 pairwise combinations = 5415) (overlaps with 4 most tolerated at 4^2 * 15=240) (overlaps with 2 most pathological at 2^2 * 15=60) (included in case the defenders did not block all of them)

   Total: 1 + 361 - 13 + 15625 - 25 + 5415 - 240 - 60= 21,793 peptide variants at each window

Construction of phagemid libraries

Oligo libraries comprising variants for each 19aa peptide window were synthesized as a pool by Twist Bioscience. Individual libraries were amplified by PCR and ligated into a phagemid backbone—encoding an ampicillin resistance gene, containing an M13 phage origin of replication, and designed for library variant expression upon induction by IPTG—using NEBuilder Hifi DNA Assembly Master Mix (NEB, E2621L). All libraries were then precipitated with isopropanol, transformed into electrocompetent DH5α cells (NEB, C2989K), and plated on 2XYT-carbenicillin-1% glucose; after overnight growth at 37 °C, colonies were counted to ensure >50-fold library coverage. Colonies were scraped with 2XYT and plasmid DNA extracted with the ZymoPURE II Plasmid Maxiprep Kit (Zymo Research, D4203); the extracted plasmid DNA was then precipitated with isopropanol. These plasmid libraries constitute the "pre-selection libraries."

Construction of helper cells

M13cp[39], a plasmid containing all M13 phage genes but with a p15a origin and a chloramphenicol resistance gene replacing the phage origin of replication, was used to construct helper plasmids. Primer pairs were designed for the precise deletion of genes I, II, III, and IV from M13cp following PCR amplification and ligation using the In-Fusion Snap Assembly Master Mix (Takara Bio, 638944). The resulting helper plasmids were transformed into DH5α competent cells (NEB, C2987H), yielding four individual helper cell lines (M13cp-dg1,

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

M13cp-dg2, M13cp-dg3, and M13cp-dg4). The helper cells were made electrocompetent for subsequent same-day transformations. Helper cells are capable of extruding phagemid particles when transformed with a phagemid library variant with a functional gene (complementing the missing phage gene in the helper plasmid) and origin of replication (Supplementary Fig. 3). DNA sequences of helper plasmids and phagemids expressing wild-type proteins are available on Addgene.

Phagemid growth

Phagemid libraries were transformed into their corresponding helper cells (nucleic acid variant libraries were transformed into M13cp-dg3) by electroporation and plated on 2XYT-carbenicillin-chloramphenicol-1% glucose. After overnight growth at 37 °C, colonies were counted to ensure >15-fold library coverage. Colonies were scraped with 50 mL 2XYT, the bacterial pellet washed sequentially 3x with 50 mL 2XYT, then a 1:1000 dilution used to inoculate a 50 mL phagemid growth culture in 2XYT with maintenance antibiotics and 1% glucose. The culture was grown to OD_{600} = 0.5 with shaking at 37 °C and 250 rpm, at which point the culture was centrifuged and the media replaced with 2XYT containing maintenance antibiotics and 1 mM IPTG. The culture was grown for 16 h at 37 °C and 250 rpm, after which phagemid-containing supernatants were collected by culture centrifugation and filtration through a 0.22 μm filter.

Phagemid infection

Phagemid-containing supernatants were added to 2.5 mL S2060 cells (streptomycin-resistant, Addgene #105064) grown to OD_{600} = 0.5 and allowed to infect at 37 °C and 250 rpm for 1 h. The resulting infected cultures were plated on 2XYT-carbenicillin-streptomycin-1% glucose to select for phagemid-containing cells. After overnight growth at 37 °C, colonies were scraped with 50 mL 2XYT and plasmid DNA extracted with the ZymoPURE II Plasmid Maxiprep Kit (Zymo Research, D4203). These plasmid libraries constitute the "post-selection libraries."

Illumina NGS sequencing

Pre- and post-selection libraries were prepared for illumina NGS sequencing by sequential PCR amplification. PCR amplification was first performed with PrimeSTAR GXL Premix (Takara Bio, R051A) to attach Nextera-style adapter sequences, followed by a second PCR amplification to attach library-specific barcodes and the p5 and p7 indices. Following PCR purification, library concentrations were quantified with qPCR using the NEBNext Library Quant Kit for Illumina (NEB, E7630S), and pre- and post-selection libraries were combined as two pools. Libraries were pooled such that libraries were present in equimolar quantities corrected for library size. Libraries were submitted to the MIT BioMicro Center for MiSeq Illumina sequencing (v3, 2 x 300 bp paired-end). The proportion of attacks detected was analyzed for each window (Supplementary Fig. 10).

Hazard database generation

We began by generating a curated database comprising subsequence windows from U.S. Select Agents, Australia Group hazards, and functional variants that aren't found in innocuous genes and genomes (Fig. 2a). We included all 19-amino-acid peptides and all 42-mers as well as 30-mers from each hazard; reliably assembling hazards from smaller DNA pieces is much more challenging[40,41]. To detect adversarial attempts to obtain functionally equivalent mutants of known hazards, we quasi-randomly selected windows to defend and used variant effect predictors to compute and add millions of variants per window[30,32,42–46].

Next, we curated database entries to remove peptides and k-mers matching harmless sequences found in GenBank nr/nt and protein nr using taxonomic classification, keywords, and exact match quantification, as well as entries with a low Shannon entropy (Fig. 2a). Curation avoids flagging innocuous sequences known to science, eliminating nearly all false alarms. As detailed in our companion paper, the efficiency of exact-match search permits the use of cryptography to protect the privacy of DNA synthesis orders sent to be screened and allows the system to screen for emerging hazards without disclosing what they are[22].

Acknowledgements

We thank S. Golas for assistance with scripting and AWS, K. Sumida for providing data on ProteinMPNN designs, H. Cozzarini for advice on testing, and E. Soice for figure design support. We are grateful for financial support from the Open Philanthropy Project (to MIT, Aarhus University, and SecureBio), an anonymous philanthropist from China (to Tsinghua University), the Aphorism Foundation (to MIT and SecureBio), and Effective Giving (to MIT and SecureBio). The funders

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

had no role in study design, data collection, data analysis, data interpretation, or writing of the report. To preserve international neutrality, no government funds were used to support this project.

Author contributions

K.M.E., D.G., and A.C.Y. conceived the study; L.F., R.R., M.G., Y.Y., C.B., I.D., and A.C.Y. performed the security assessments, D.G. wrote software to implement the search method to defend M13; E.A.D. designed the attacks against M13; B.W. designed and performed all laboratory experiments involving M13; E.C. analyzed sequencing data; D.G. performed the sensitivity analyses; D.G. wrote the prototype database software with assistance from C.D., O.D., T.Y., K.U., and A.F.; the initial screening prototype was developed by H.C., X.L., J.D., M.G., and Y.Y.; professional software development was overseen by L.F. and J.B. and performed by L.F., J.B., T.V., M.K., W.C., F.S-L., L.V.H., S.W., and B.W-R.; and the attacks that defeat alignment and conventional exact-match search were devised and tested by R.E. and K.M.E. All authors contributed to writing the paper.

Competing interests

K.M.E. and D.G. are authors of PCT/US2021/014814 filed by the Massachusetts Institute of Technology. All authors share an interest in preventing future pandemics.

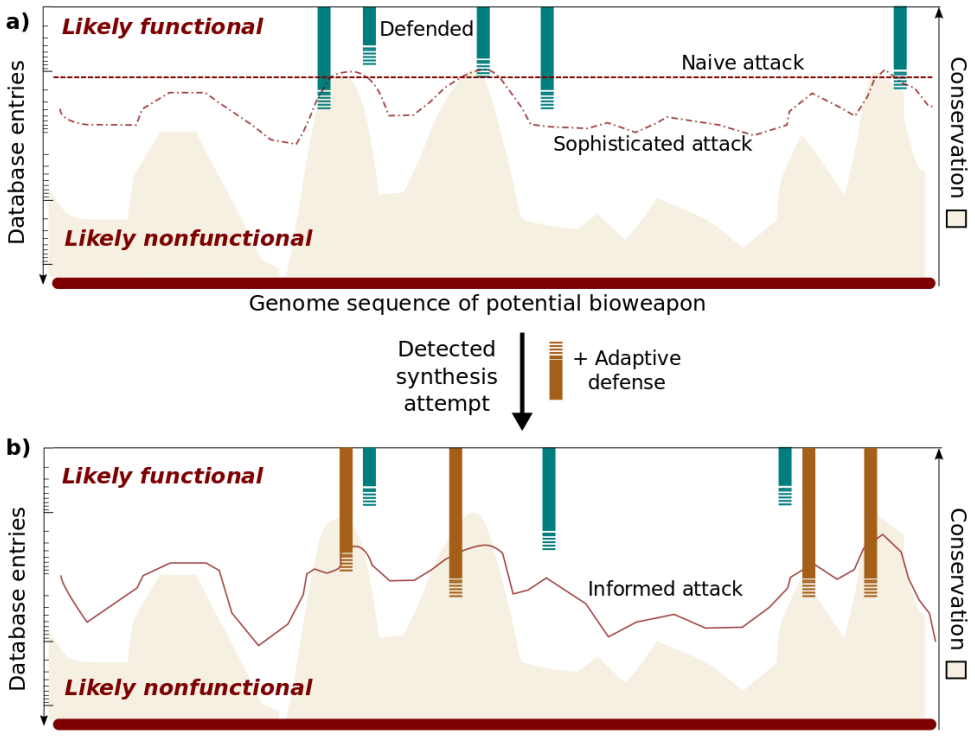
Data availability

Raw experimental data on the fitness of mutant M13 phages and database entries used for defense are available via Figshare.

Code availability

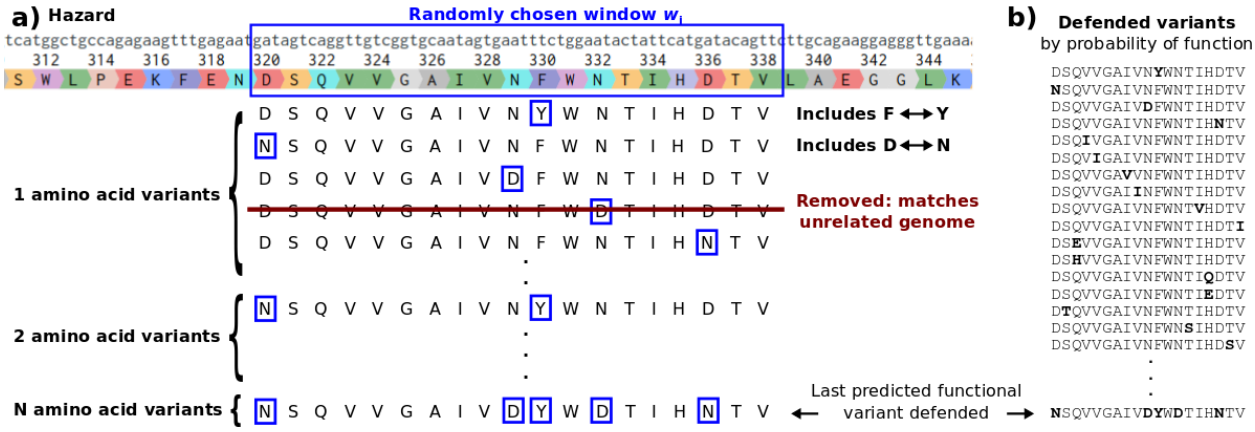
Code is available at https://github.com/dgretton/RAT-DNA-screening. fuNTRp[30] is available on request from Maximilian Miller or at https://bromberglab.org/project/funtrp/.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



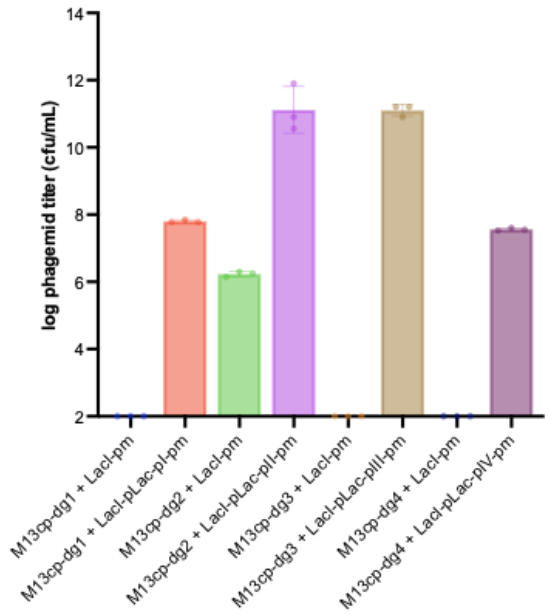
Supplementary Figure 1 | Potential attacks seeking to evade RAT screening and the possibility of adaptive defense. a) Variants from five windows in a region are included in the hazard database, mostly from relatively conserved regions. Most of the variants predicted to be highly functional by a variety of different algorithms at each window are included in the database, so a naive attacker who simply introduces a moderate number of mutations at a constant rate is both highly likely to be detected and risks generating a nonfunctional hazard due to the accumulated fitness cost. A sophisticated attacker may try to tune the number of mutations to the likelihood of obtaining a functional sequence across regions, thereby maximizing the chance of evading screening while preserving function. Their chances improve with the superiority of their variant prediction capabilities relative to the defender, but they still must trade off the risk of generating a nonfunctional hazard against being randomly detected upon picking a database variant at one of the five protected windows. b) If multiple attacks on a particular hazard are detected, the system can adaptively add new windows and defend more variants at each window, precluding informed attacks based on probing or database interrogation. Windows may also be rotated in and out.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



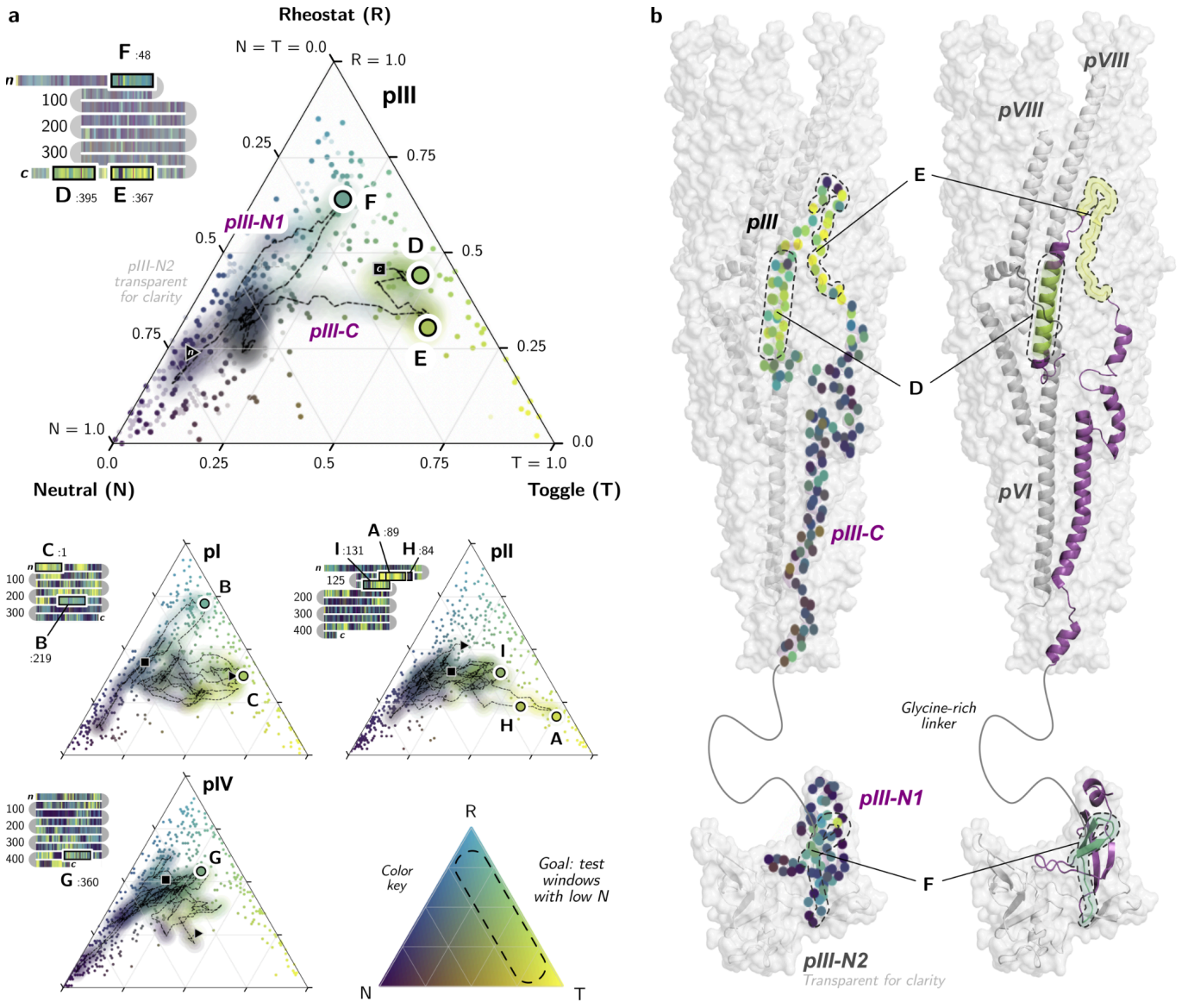
Supplementary Figure 2 | Functional variant prediction and curation. The hazard database includes predicted functional variants of quasi-randomly chosen windows comprising 20-amino-acid peptides from hazardous proteins or 42-base pair DNA/RNA sequences from the noncoding regions of hazardous genomes. a) Examples of variant peptide sequences. Variants matching unrelated sequences in GenBank are removed. b) The random adversarial threshold increases as variants are added to the database. The exact method used to predict the function of variants in order to generate the list will be randomized across several prediction methods to prevent adversaries from predicting the contents of the hazard database.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 3 | Fitness costs of loss-of-function mutations in M13 genes bearing windows. The fitness of phagemids encoding filamentous phage genes I-IV (LacI-pLac-pI-pm, LacI-pLac-pII-pm, LacI-pLac-pIII-pm, and LacI-pLac-pIV-pm, respectively) was quantified in cells carrying helper plasmids deleted for the gene in question (M13cp-dg1, M13cp-dg2, M13cp-dg3, or M13cp-dg4). In all cases, cells extruded phagemid particles when induced with 1 mM IPTG, as measured by infection of recipient cells with 3 independent biological replicates. Data from helper cells transformed with phagemids lacking the wild-type phage genes are provided for comparison. Phagemid titers below the limit of detection (100 cfu/mL) are plotted at the limit of detection. For all genes, loss of function reduced fitness by over 100-fold, which is greater than the reduction to fitness 0.05 of wild-type designated as too costly for the most contagious human virus to spread. The difference in maximum titers suggests that the optimal level of each protein differs from the level produced upon induction, which may affect the relative fitness of variants. Notably, phagemid overproduction may artificially increase the measured fitness of variants with reduced activity when excess activity is costly, resulting in an apparent mutant fitness greater than wild-type. Data plotted as mean ± standard deviation.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



A	VEIKASPAKVLQGHNVFGT	**D**	FRGVFAFLLYVATFMYVFS	**G**	IATTVNLRDGQTLLLGGLT
B	YSYLTPYLSHGRYFKPLNL	**E**	PQSVECRPFVFGAGKPYEF	**H**	NFYPCVEIKASPAKVLQGH
C	MAVYFVTGKLGSGKTLVSV	**F**	YANYEGCLWNATGVVVCTG	**I**	LLDVNATTISRIDAAFSAR

Supplementary Figure 4 | fuNTRp window analyses for M13 bacteriophage proteins. a) Ternary plots and color-bar snake plots show the probabilities that each residue of M13 phage proteins pI-pIV are neutral (N, purple), toggle (T, yellow), or rheostat (R, cyan), N + T + R = 1 (color key at bottom). The nine defended windows are highlighted (A-I, legend at bottom). Ternary plot and snake plot of protein pIII enlarged to show detail. On ternary plots, scatter points show NTR scores for all residues in each protein, while 19-residue moving average (dotted trace) from n-terminus (triangle) to c-terminus (square) shows local average NTRs. 19-residue windows were optimized on the basis of average NTR scores, meaning that all possible window choices fall on the dotted trace. Windows chosen for testing minimized neutral scores (N≈0, right diagonal edge), but varied in proportion of toggle *vs*. neutral. Colors of dots for defended windows represent their average NTRs. **b)** 3D structure of pIII in context of assembled phage virion tip (surfaces)[47,48]. Dots colored by NTR (left) indicate alpha carbon atoms. Ribbon representation (right) shows the structure of pIII (purple), with defended windows in pIII (D-F) highlighted, colored by average NTR. Flexible glycine-rich N1-N2 linker not shown in structure. Using a low neutral score as a proxy for functional importance, minimum-neutral windows appear to correspond to regions with many contacts with nearby proteins, pVI (gray) and pVIII (light gray) in this case (D, E), and binding sites (F).

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Impact of toggle positions given few neutral residues

Sequence (**protein**:amino acid start)	Toggle	Rheostat	Neutral	Fraction of attacks blocked
VEIKASPAKVLQGHNVFGT (**pII**:89)	14.4	3.48	1.12	0.852
MAVYFVTGKLGSGKTLVSV (**pI**:1)	10.43	7.11	1.46	0.492
FRGVFAFLLYVATFMYVFS (**pIII**: 395)	9.03	8.37	1.6	0.345

Comparison of toggle-dominant and rheostat-dominant windows given low neutrality

MAVYFVTGKLGSGKTLVSV (**pI**:1)	10.43	7.11	1.46	0.492
YSYLTPYLSHGRYFKPLNL (**pI**:219)	4.15	13.62	1.23	0.293
PQSVECRPFVFGAGKPYEF (**pIII**:367)	10.64	5.76	2.6	0.437
YANYEGCLWNATGVVVCTG (**pIII**:48)	3.8	12.17	3.03	0.297

Comparing toggle vs rheostat positions given more neutral residues

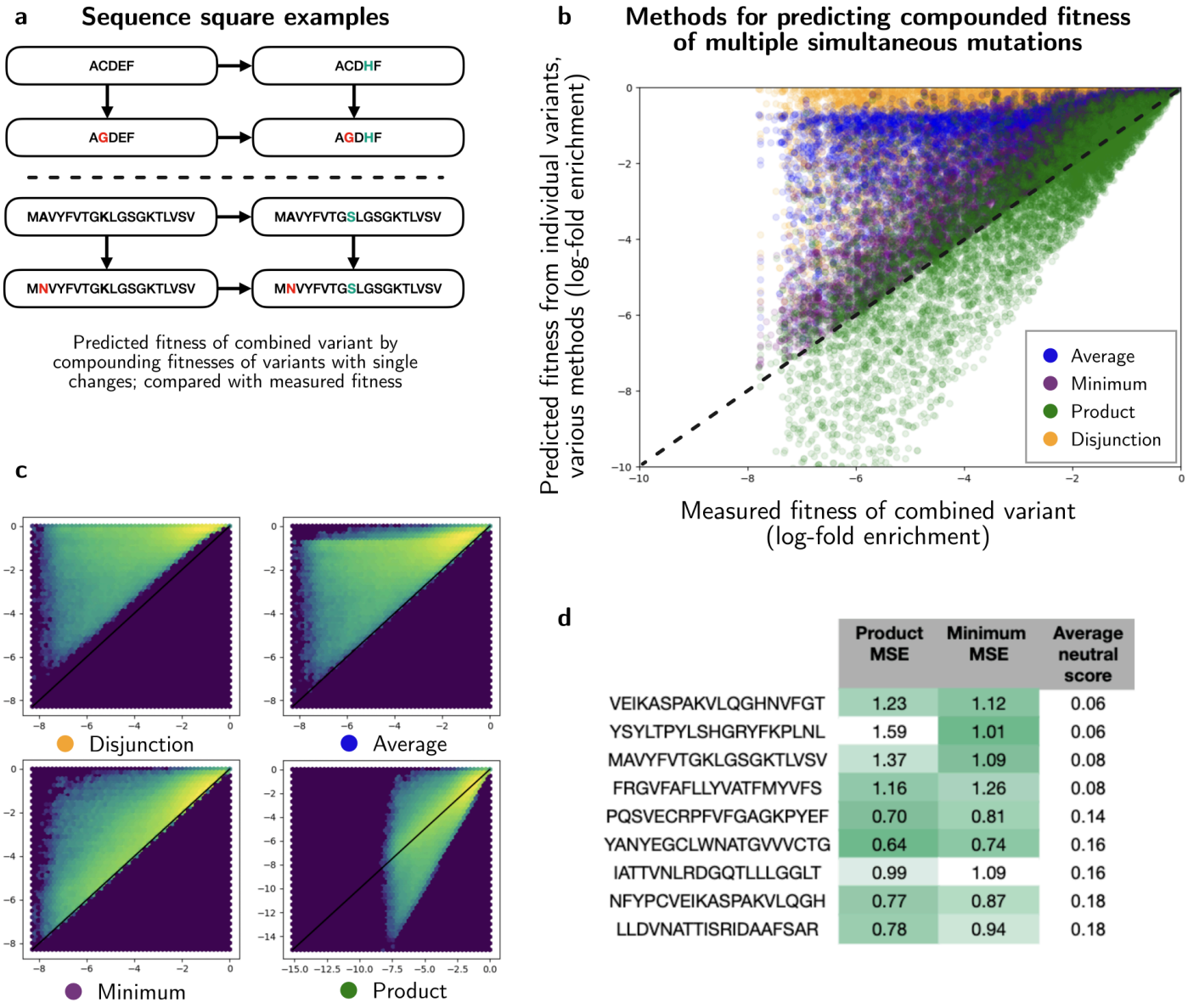
NFYPCVEIKASPAKVLQGH (**pII**:84)	11.14	4.39	3.47	0.477
YANYEGCLWNATGVVVCTG (**pIII**:48)	3.8	12.17	3.03	0.297
IATTVNLRDGQTLLLGGLT (**pIV**:360)	5.5	10.42	3.08	0.462

Assessing effect of spread of neutral residues for comparable mean neutral scores

NFYPCVEIKASPAKVLQGH (**pII**:84)	11.14	4.39	3.47	0.477
LLDVNATTISRIDAAFSAR (**pII**:131)	8.09	7.43	3.48	0.427

Supplementary Table 1 | Strategic selection of windows to assess impact of funNTRp attributes on defensibility. Windows from M13 phage proteins were chosen to compare how different fuNTRp scores affect the fraction of combinatorial attacks blocked when screening 10^6 predicted variants at each window. Colors highlight values of interest. The top section shows that with few neutral residues, more toggles (purple) increase defensibility over more rheostats (cyan). The middle section compares toggle-dominant to rheostat-dominant windows. The bottom section compares windows with comparable mean neutral scores that are either spread out or concentrated. Some windows are included multiple times to enable comparisons. This strategic selection of windows provided insights into optimizing the use of funNTRp outputs for identifying highly defensible sequences when generalizing this approach to populate the hazard database used by the full SecureDNA system for screening DNA synthesis orders. Note that two of the windows partly overlap.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

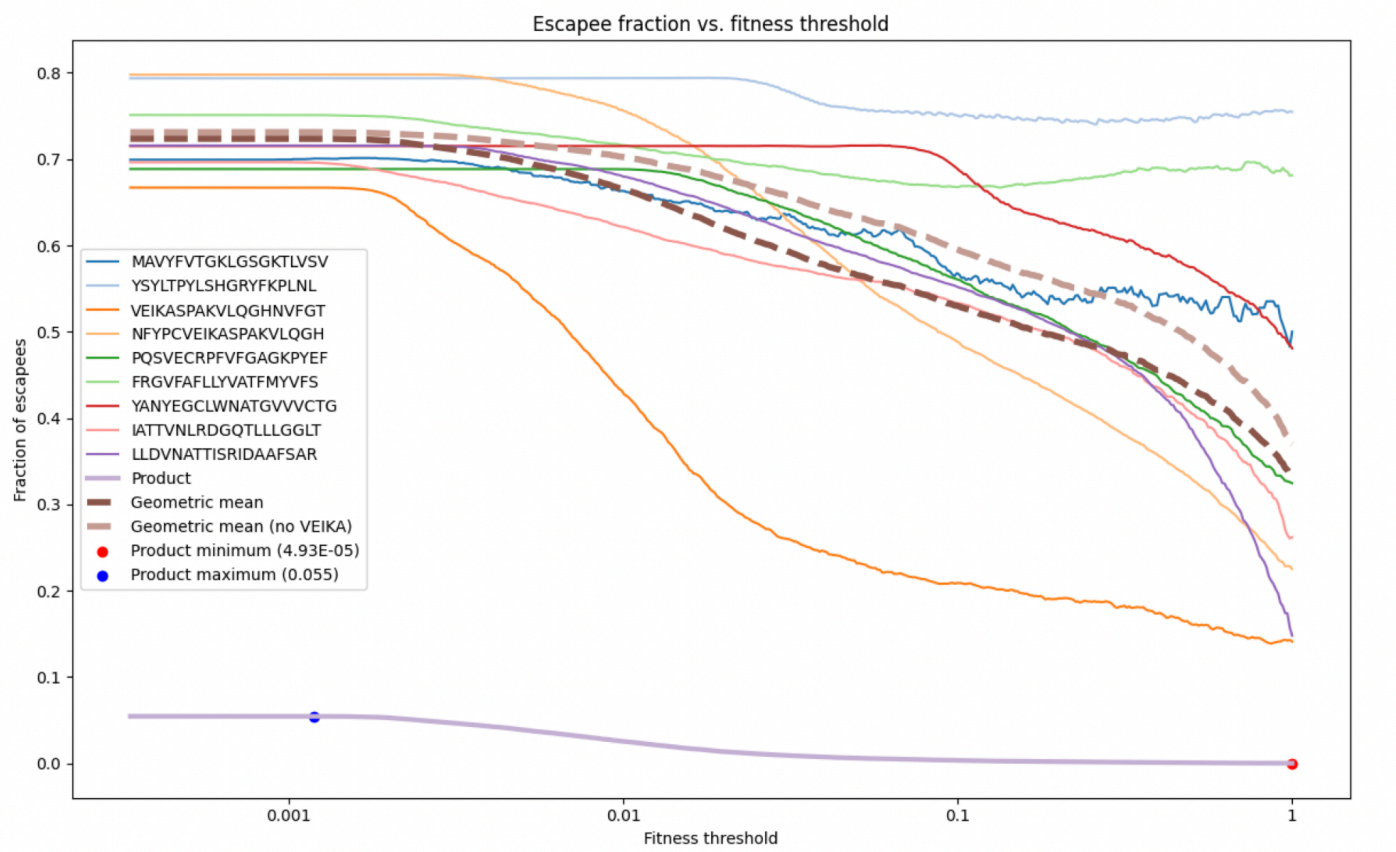


d

	Product MSE	Minimum MSE	Average neutral score
VEIKASPAKVLQGHNVFGT	1.23	1.12	0.06
YSYLTPYLSHGRYFKPLNL	1.59	1.01	0.06
MAVYFVTGKLGSGKTLVSV	1.37	1.09	0.08
FRGVFAFLLYVATFMYVFS	1.16	1.26	0.08
PQSVECRPFVFGAGKPYEF	0.70	0.81	0.14
YANYEGCLWNATGVVVCTG	0.64	0.74	0.16
IATTVNLRDGQTLLLGGLT	0.99	1.09	0.16
NFYPCVEIKASPAKVLQGH	0.77	0.87	0.18
LLDVNATTISRIDAAFSAR	0.78	0.94	0.18

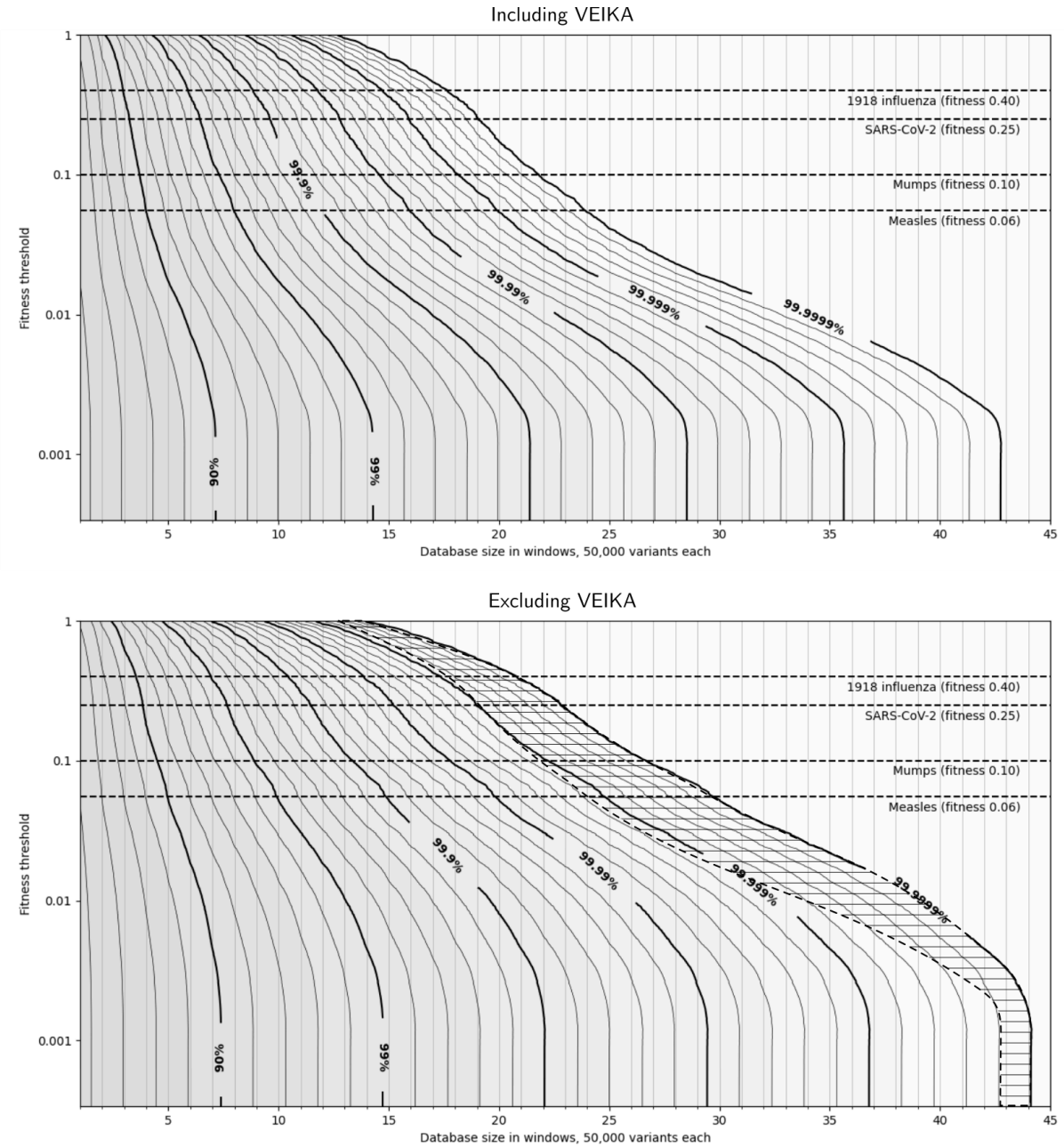
Supplementary Figure 5 | Effects of combinatorial fitness reductions. For its security claims, SecureDNA assumes that fitness reductions due to individual mutations will tend to compound when such mutations are present simultaneously. Fitness of combined mutations are estimated by various methods. a) Sequence "squares" are shown, representing a variant, its orthogonal single mutants, and their combined double mutant. Top: A conceptual instance. Bottom: A sequence square from the dataset with variants of the fragment MAVYFVTGKLGSGKTLVSV. b) Fitness models for the fragment VEIKASPAKVLQGHNVFGT are assessed, comparing the average, minimum, product, and disjunction ($1-(1-f_1)(1-f_2)$) of 1,000 single-mutant variants against their double-mutants' fitnesses. Dotted line denotes perfect correlation. Heuristically, product models statistically independent changes and minimum models breaking changes. c) Hex heat map of the fitness methods for 10^6 sequence squares from the fragment VEIKASPAKVLQGHNVFGT shows highest concentration near the diagonal for the minimum and product methods. Across all fragments, minimum and product consistently outperformed others by mean square error (MSE). d) Fragments sorted by neutral score. For fragments with low neutral scores, expected to be regions where breaking changes are likely, the minimum method is superior by MSE. Conversely, the product method is superior for fragments with high neutral scores, suggesting that small independent fitness impacts can be multiplicatively combined to estimate the total fitness reduction.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



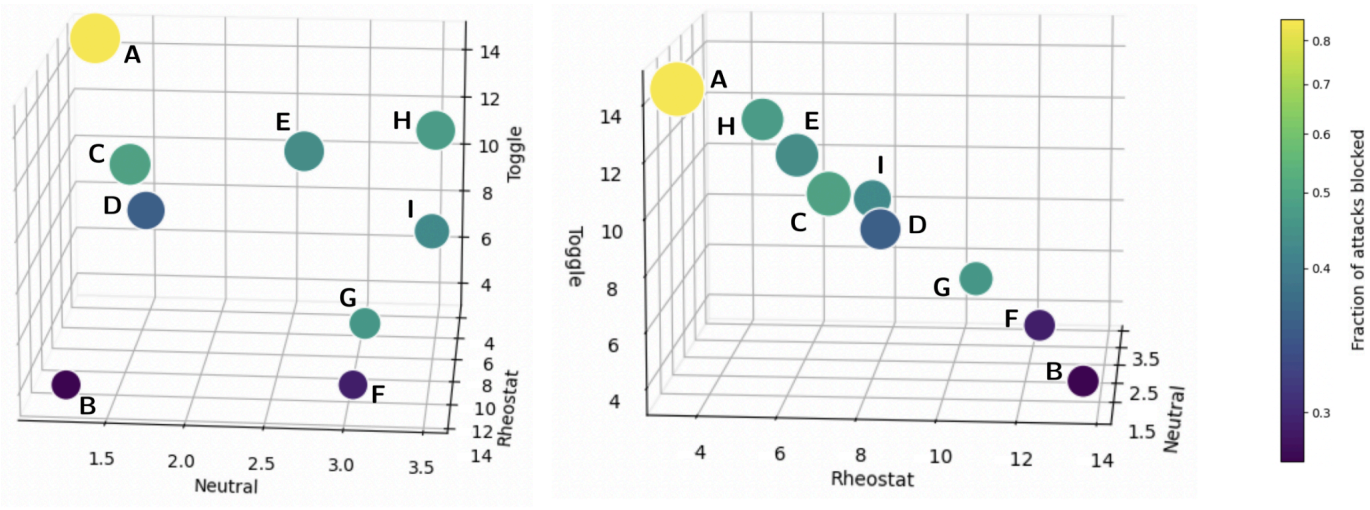
Supplementary Figure 6 | Fraction of undetected attacks against minimum fitness thresholds. The adversary is assumed to possess perfect knowledge of variant fitness out to six amino acid mutations relative to the 19-amino acid wild-type subsequence. The horizontal axis depicts the tolerance level of the attacker to fitness hits: more damaging hits are towards the left, while higher-performing but correspondingly restrictive fitness cutoffs are on the right. Vertical axis is the fraction of escapee variants that had fitness higher than the cutoff that were not in the 50,000-variant database. A decreasing trend indicates that an attacker attempting to synthesize an agent with high fitness is more likely to match a functional variant in the hazard database: that is, the fuNTRp+BLOSUM62 classifier accurately predicts fitness at the window. Some sequence windows (VEIKA) offer robust protection even at low fitness cutoffs, whereas others (YSYLT, FRGVF) protect a roughly fixed fraction of variants regardless of the fitness threshold, indicating that our classifier has limited power to distinguish more fit variants at these windows. Classifiers based on other variant effect predictors may differ in their predictive ability across distinct windows. When all nine of these windows are combined multiplicatively (bold purple), simulating a defense based on all of them, adversaries rarely predict functional variant genomes at any fitness threshold. The geometric mean (dark brown, dotted) represents the average window such that the effect of combining more than 9 windows may be extrapolated multiplicatively. Geometric mean excluding VEIKA, which may be an outlier, also shown (light brown, dotted).

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 7 | Effect of excluding outlier window. Comparison of analyses including VEIKA (top) and excluding VEIKA (bottom), which may be an outlier. Top, same data as Figure 5: original extrapolation of levels of protection, or RAT (Random Adversarial Threshold), by powers of the geometric mean of all 9 escapee curves (Supplementary Fig. 6, dark brown dotted line). The number of simulated windows, each protected with 50,000 variants, is plotted on the horizontal axis. The fitness at which the hazard is no longer functional is the vertical axis. Contours show lines of constant RAT, or equal protection, as a trade-off between choices of minimum fitness tolerated and number of windows protected. Bottom: identical plot where the effect of simulated windows is extrapolated using the geometric mean of 8 escapee curves, excluding VEIKA (Supplementary Fig. 6, light brown dotted line). Shaded region shows displacement of 99.9999% contour, which moved the most. Greatest displacement was for measles. Excluding VEIKA from the estimate for the mean window, a virus as infectious as measles may require up to 6 extra windows to achieve the same RAT. Even if VEIKA is an outlier, 100 or more windows should suffice to block combinatorial attacks with high confidence.

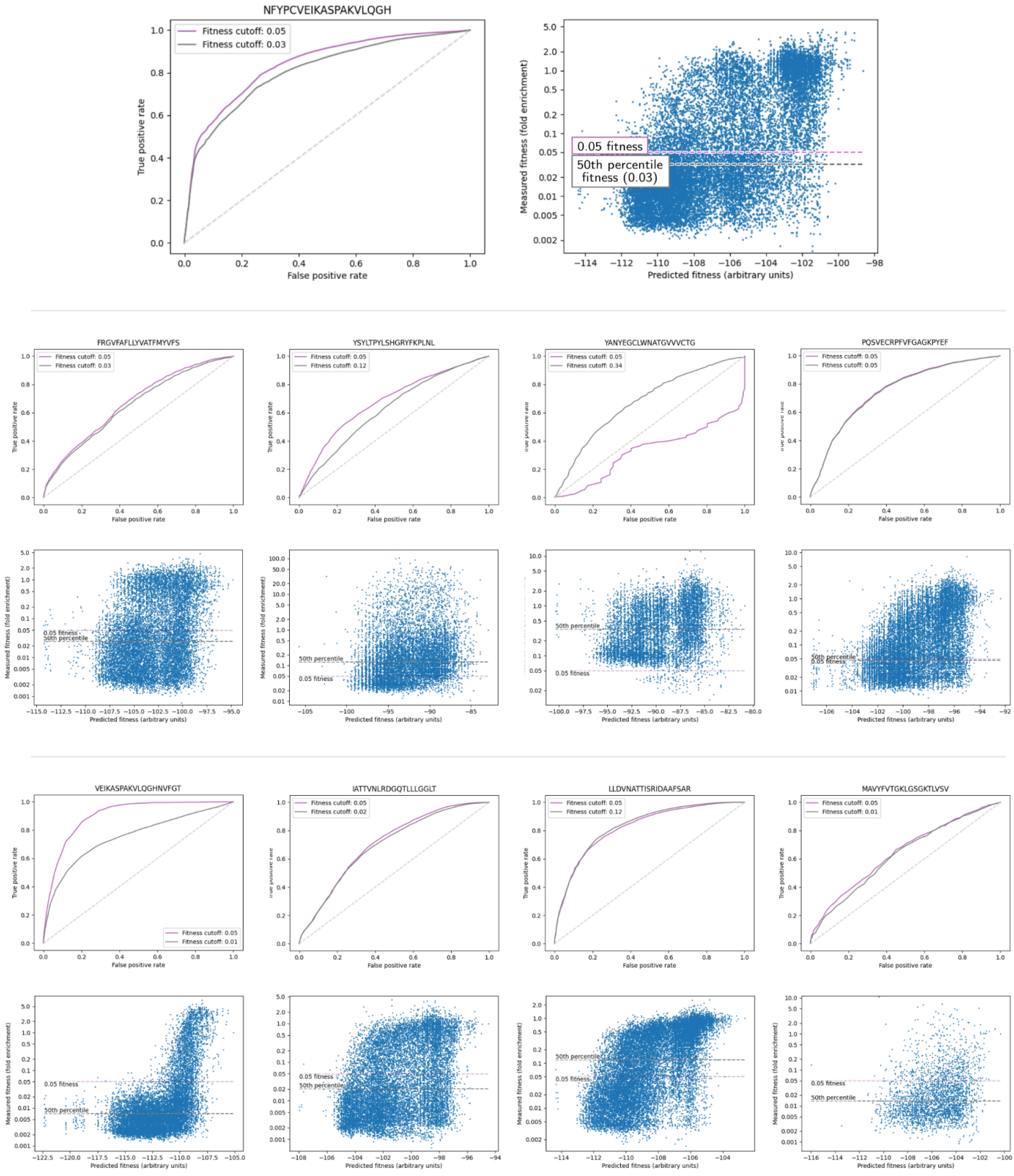
bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



A	VEIKASPAKVLQGHNVFGT	**D**	FRGVFAFLLYVATFMYVFS	**G**	IATTVNLRDGQTLLLGGLT
B	YSYLTPYLSHGRYFKPLNL	**E**	PQSVECRPFVFGAGKPYEF	**H**	NFYPCVEIKASPAKVLQGH
C	MAVYFVTGKLGSGKTLVSV	**F**	YANYEGCLWNATGVVVCTG	**I**	LLDVNATTISRIDAAFSAR

Supplementary Figure 8 | Predicted residue mutability correlates with defensibility. Points represent 19-codon windows from the M13 phage genome, plotted in 3D feature space by their average fuNTRp neutral, toggle, and rheostat scores per residue. Lighter color indicates a higher fraction of 21,000 combinatorial attacks blocked by screening 10^6 predicted functional variants of that window. The fraction of attacks blocked serves as an experimental measurement of the window's robustness to mutations. Windows with higher toggle scores generally blocked a higher fraction of attacks, as toggle residues are predicted to be less tolerant of mutations. funNTRp classifies residues as neutral (tolerating most mutations), toggle (highly sensitive), or rheostat (intermediate mutability). The summed scores per residue equal a constant, confining points to a plane. This analysis informed subsequent window selection to maximize screening efficacy.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

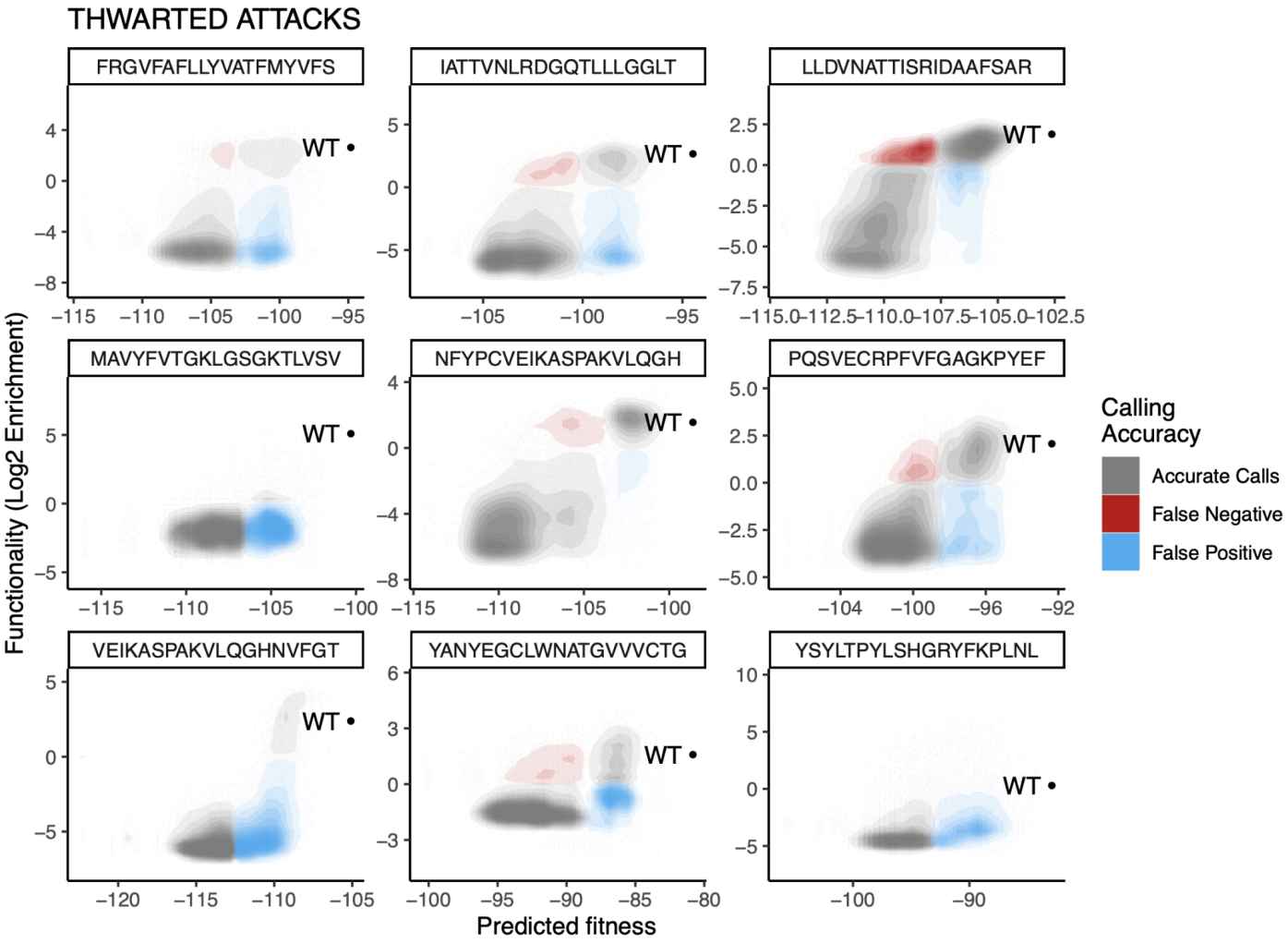


Supplementary Figure 9 | Receiver-operating-characteristic curves. Left: Receiver-operating-characteristic (ROC) curves for fuNTRp+BLOSUM62 prediction for nine 19-amino acid windows across the genome of M13 bacteriophage. ROC assesses performance of computational prediction of variant

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

functionality for design of the hazard database. Curves show true positive rate vs false positive rate in the range 0 to 1 across a range of classification thresholds. ROC curves higher toward (0, 1) indicate better performance. Dotted line from (0, 0) to (1, 1) corresponds to random guessing. Metric combines fuNTRp scores and BLOSUM62 substitution scores to predict fitness effect of amino acid substitutions. Functional variants defined as empirical fitness >0.05 wild-type (purple). ROC for median empirical fitness shown in gray. Right: fuNTRp+BLOSUM62 prediction metric vs measured relative fitness. Bottom: similar curves and point clouds for all fragments. Best-performing classifier VEIKASPAKVLQGHNVFGT demonstrates high discrimination ability, capturing >90% of functional variants at <30% false positive rates, preventing database overfill while retaining impactful variants. Poorer-performing MAVYFVTGKLGSGKTLVSV classifier still contributes protective value by correctly predicting some functional variants, while trading off less favorably with database size. YANYEGCLWNATGVVVCTG produced no meaningful ROC curve at 0.05 fitness threshold due to data sparsity; median curve indicates roughly average metric performance where data exists.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 10 | Analyses of red-team attacks against defended windows of M13 phage. Predicted fitness using fuNTRp and BLOSUM62 to estimate fitness of sequence variants is shown on the horizontal axis. Vertical axis shows variant fitness, in log fold enrichment after one round of phage replication and selection. Headmap shows density of points in each of three categories: true positives and true true negatives, grouped as "accurate calls;" false positives, which represent an instance when the predictor rated a variant's fitness too highly and incorrectly screened it, representing an inefficiency; and false negatives, which represent successful attacks that were not thwarted at the window shown. YSYLT had negligible true positives but also no false negatives. VEIKA had almost exclusively accurate calls and false positives, representing excellent performance. LLDVNA was the worst performer by this metric in that it introduced the greatest density of false negatives; to restore security, other windows with better performance must be included in combination with this window. As this cannot be known in advance, this result highlights the importance of including many windows, which assist in driving RAT to 1.0.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix A: Theoretical specificity analysis

Curation ensures that RAT search will never flag known harmless sequences, but random matches to novel sequences will occur at a frequency determined by the total amount of novel DNA synthesized and the number of sequences in the database. The probability that any translation of a random 60-mer will match a specific 20 amino acid peptide is $8.0*10^{-25}$. If 10^{15} unique 60-mers of oligonucleotides will be synthesized in 2035[50], we expect approximately one random false positive for the entire world's DNA synthesis in that year. Similar numbers are obtained for DNA screening. While biological sequences are far from isotropic[51], almost all orders comprise or encode known sequences. Therefore, random false alarms will overwhelmingly occur in oligonucleotide libraries for experiments such as deep mutational scanning and directed evolution; removing a random oligo from a library of many thousands or millions is not expected to impact the results of such an experiment. Empirical measurements of specificity on datasets of real-world orders are detailed in the companion paper.

For DNA of length N and peptides of length N/3,

$FA_{DNA}(N)$ = (novel N-mers/yr) * (N-mers in *D*) / (2 frames * 4^N)

$FA_{AA}(N)$ = (novel DNA windows/yr) * (windows in D) / [2 frames * (20 amino acids * 61 / 64 codons)$^{(N/3)}$]

Appendix B: Measures for Customer and Provider Data Privacy

This study utilizes unfiltered customer sequence data from specified time intervals provided by multiple DNA synthesis companies. Acknowledging legitimate privacy concerns, this appendix aims to assure readers as well as the data providers and their customers that stringent measures were taken to fully protect sensitive information in the analysis. The identities of customers were anonymized and not known to the researchers, preventing any possibility of directly revealing them accidentally or intentionally. Only the rates of detected hazards and approximate dataset sizes are reported, with no other specific sequence information disclosed, such that the only sequence content revealed pertains to hazards. We utilize data from three or more customers per provider and three or more providers total. If there were only two customers per provider, one customer could attribute hazards exclusively to the other. However, with data from at least three customers included, no single customer can definitively attribute a detected hazard to a specific other customer. Similarly, if there were only two customers, one could roughly estimate the order volume of the other by subtracting their own volume from the approximate total given. But with three or more customers aggregated, no single customer can determine the exact order volumes of other specific customers, only an upper bound on their own fraction of their provider's business, a quantity that could be derived from public sources like earnings reports. Regarding the privacy of contributing providers, we can reasonably assume that each contributing provider knows the contents of the dataset contributed to the analysis, allowing them to potentially deduce which data originated from other providers. However, with data incorporated from at least three different providers, no single provider can definitively attribute a particular dataset to another specific provider. To further obscure which hazards may be attributable to which provider, and to confirm that screening is working as intended even in low-hazard-content sets, we intentionally inject known hazard sequences into test data. Since only lower bounds on dataset sizes are provided, reporting hazard detection rates with low precision ensures that any small rates are indistinguishable from zero. Thus, a reported rate of <0.1% could reflect an arbitrary number of actual detections in the provider data, including zero, preventing positive identification of which provider's data may have contained specific hazards. In summary, by utilizing data from three or more customers per provider and at least three providers, along with intentionally limiting the precision of the reported hazard rates and dataset quantifications, this methodology aims to effectively evaluate the screening system while stringently protecting the privacy of all data providers and their customers.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix C: Screening for emerging hazards

RAT search is compatible with cryptographic approaches capable of obscuring the identities of entries in an emerging hazard database. Such a system would enable a researcher concerned about an emerging potential biological weapon to safely take action to restrict global access without creating information hazards[52,53] by securely conveying their concern to one of the biologist curators responsible for emerging hazards. If a curator concurs that the threat is serious, they could use their unique key to add sequences from the hazard to the encrypted emerging hazard database without requiring further disclosure. Several times as many genes or genomes would be chosen as "decoys": related genes or agents that might seem to pose a threat, but are not actually of concern. The use of decoys can ensure that anyone who finds a match to the database will learn only that it corresponds to a plausible-seeming threat, not that it is a credible weapon. This would ensure that adversaries cannot learn whether a specific hazard can be used to build a weapon of mass destruction by attempting to synthesize it. A detailed explanation of the cryptographic approaches required is available elsewhere[23]. Emerging hazards would only be added to the database after multiple years of testing to verify that the implementation is secure.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. OECD. OECD: Graduates by field. https://stats.oecd.org/Index.aspx?DataSetCode=EDU_GRAD_FIELD.
2. Kang, K., Falkenheim, J. & Kam, L. Doctorate recipients from U.S. universities, 2022. https://ncses.nsf.gov/pubs/nsf24300/data-tables.
3. Neumann, G., Ozawa, M. & Kawaoka, Y. Reverse genetics of influenza viruses. *Methods Mol. Biol.* **865**, 193–206 (2012).
4. Xie, X. *et al.* Engineering SARS-CoV-2 using a reverse genetic system. *Nat. Protoc.* **16**, 1761–1784 (2021).
5. Tumpey, T. M. *et al.* Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* **310**, 77–80 (2005).
6. Kennedy, D. Better never than late. *Science* vol. 310 195 (2005).
7. Herfst, S. *et al.* Airborne transmission of influenza A/H5N1 virus between ferrets. *Science* **336**, 1534–1541 (2012).
8. Imai, M. *et al.* Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* **486**, 420–428 (2012).
9. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**, e1006698 (2017).
10. Grange, Z. L. *et al.* Ranking the risk of animal-to-human spillover for newly discovered viruses. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
11. Warren, C. J. *et al.* Primate hemorrhagic fever-causing arteriviruses are poised for spillover to humans. *Cell* **185**, 3980–3991.e18 (2022).
12. Sun, H. *et al.* Airborne transmission of human-isolated avian H3N8 influenza virus between ferrets. *Cell* **186**, 4074–4084.e11 (2023).
13. Hou, Y. J. *et al.* Host range, transmissibility and antigenicity of a pangolin coronavirus. *Nat Microbiol* **8**, 1820–1833 (2023).
14. International Gene Synthesis Consortium. Harmonized Screening Protocol V2. https://genesynthesisconsortium.org/wp-content/uploads/IGSCHarmonizedProtocol11-21-17.pdf (2017).
15. Diggans, J. & Leproust, E. Next Steps for Access to Safe, Secure DNA Synthesis. *Front Bioeng Biotechnol* **7**, 86 (2019).
16. Beal, J., Clore, A. & Manthey, J. Studying pathogens degrades BLAST-based pathogen identification. *Sci. Rep.* **13**, 5390 (2023).
17. Danzig, R. *et al. Aum Shinrikyo: Insights into How Terrorists Develop Biological and Chemical Weapons*. http://www.jstor.org/stable/resrep06323 (2012).
18. Edison R, Toner S, Esvelt KM. Evaluating the adequacy of DNA synthesis screening. *To be made available in May 2024 when superior screening software based on random adversarial threshold search has been available to providers for three months, in keeping with cybersecurity norms*. (2024).
19. Nouri, A. & Chyba, C. F. DNA synthesis security. *Methods Mol. Biol.* **852**, 285–296 (2012).
20. Seydel, C. DNA writing technologies moving toward synthetic genomes. *Nat. Biotechnol.* **41**, 1504–1509 (2023).
21. Esvelt, K. M. Inoculating science against potential pandemics and information hazards. *PLoS Pathog.* **14**, e1007286 (2018).
22. The SecureDNA cryptography team. Hiding dangerous DNA in plain sight. *submitted*.
23. Baum C, Berlips J, Chen W, Cui H, Damgard I, Dong J, Esvelt Km, Gao M, Gretton D, Foner L, Kysel M, Li K, Li L, Li X, Rivest R, Sage-Ling F, Shamir A, Vaikuntanathan V, Van Hauwe L, Vogel T, Weinstein-Raun B, Wichs D, Wooster S, Yao AC, Yu Y, Zhang H. A system capable of verifiably and privately screening global DNA synthesis. *SecureDNA Project* (2024). https://www.securedna.org/download/System_Screening_Global_DNA_Synthesis.pdf
24. Carlson, R. Carlson Curves - synthesis. *synthesis* http://www.synthesis.cc/synthesis/category/Carlson+Curves (2016).
25. Bussi, Y., Kapon, R. & Reich, Z. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLoS One* **16**, e0258693 (2021).
26. Mouratidis, I. *et al.* Quasi-prime peptides: identification of the shortest peptide sequences unique to a species. *NAR Genom Bioinform* **5**, lqad039 (2023).
27. Poznański, J. *et al.* Global pentapeptide statistics are far away from expected distributions. *Sci. Rep.* **8**, 15178 (2018).
28. Weidmann, L., Dijkstra, T., Kohlbacher, O. & Lupas, A. N. Minor deviations from randomness have huge repercussions on the functional structuring of sequence space. *bioRxiv* 706119 (2021) doi:10.1101/706119.
29. Pan, Q., Nguyen, T. B., Ascher, D. B. & Pires, D. E. V. Systematic evaluation of computational tools to predict the effects of mutations on protein stability in the absence of experimental structures. *Brief. Bioinform.* **23**, (2022).
30. Miller, M., Vitale, D., Kahn, P. C., Rost, B. & Bromberg, Y. funtrp: identifying protein positions for variation driven functional tuning. *Nucleic Acids Res.* **47**, e142

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.20.585782; this version posted April 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(2019).
31. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
32. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020).
33. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
34. Guerra, F. M. *et al.* The basic reproduction number (R0) of measles: a systematic review. *Lancet Infect. Dis.* **17**, e420–e428 (2017).
35. Burns, C. C. *et al.* Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. *J. Virol.* **80**, 3259–3272 (2006).
36. Clarke, D. K. *et al.* Synergistic attenuation of vesicular stomatitis virus by combination of specific G gene truncations and N gene translocations. *J. Virol.* **81**, 2056–2064 (2007).
37. Sumida, K. H. *et al.* Improving Protein Expression, Stability, and Function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).
38. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
39. Chasteen, L., Ayriss, J., Pavlik, P. & Bradbury, A. R. M. Eliminating helper phage from phage display. *Nucleic Acids Res.* **34**, e145 (2006).
40. Gibson, D. G. Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res.* **37**, 6984–6990 (2009).
41. Plesa, C., Sidore, A. M., Lubock, N. B., Zhang, D. & Kosuri, S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* **359**, 343–347 (2018).
42. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35**, 3823–3835 (2007).
43. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
44. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
45. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst* **6**, 116–124.e3 (2018).
46. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
47. Jia, Q. & Xiang, Y. Cryo-EM structure of a bacteriophage M13 mini variant. *Nat. Commun.* **14**, 5421 (2023).
48. Lubkowski, J., Hennecke, F., Plückthun, A. & Wlodawer, A. The structural basis of phage display elucidated by the crystal structure of the N-terminal domains of g3p. *Nat. Struct. Biol.* **5**, 140–147 (1998).
49. Titus, A. J. *et al.* SIG-DB: Leveraging homomorphic encryption to securely interrogate privately held genomic databases. *PLoS Comput. Biol.* **14**, e1006454 (2018).
50. Guesstimating the Size of the Global Array Synthesis Market. https://synbiobeta.com/guesstimating-size-global-array-synthesis-market/ (2017).
51. Liu, Z., Venkatesh, S. S. & Maley, C. C. Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples. *BMC Genomics* **9**, 509 (2008).
52. Bostrom, N. Information Hazards: A Typology of Potential Harms from Knowledge. *Review of Contemporary Philosophy* **10**, 44–79 (2012).
53. Lewis, G., Millett, P., Sandberg, A., Snyder-Beattie, A. & Gronvall, G. Information Hazards in Biotechnology. *Risk Anal.* **39**, 975–981 (2019).