Virus-free continuous directed evolution in human cells using somatic hypermutation

Stanley Bram[1,2], Jordan Quenneville[1], Graeme Lindsey[1], Sarah Leach[3], Jenna J. Guthmiller[3], Angad P. Mehta*[1,2,4]

[1]Department of Chemistry, University of Illinois at Urbana-Champaign, 600 S Mathews Avenue, Urbana, Illinois 61801, United States.

[2]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1206 W Gregory Dr, Urbana, IL 61801, United States.

[3]Department of Immunology and Microbiology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, United States.

[4]Cancer Center at Illinois, University of Illinois at Urbana-Champaign, 405 N Mathews Ave, Urbana, IL 61801, United States.

*Corresponding author: apm8@illinois.edu

Abstract:

Random mutations followed by natural selection is one of the key mechanisms during natural evolution that results in the generation of new functions. This process is typically slow in nature because of low mutational frequency. In contrast, the B cells of mammalian immune systems have evolved somatic hypermutation (SHM) mechanisms that introduce mutations at the immunoglobulin genomic loci at a significantly higher frequency than the rest of the genome. SHM allows B cells to rapidly evolve new antibody sequences without compromising their fitness as a consequence of genome-wide mutations. In this work, we developed a continuous directed evolution platform in human B cell lines (CODE-HB) that recruits and repurposes the SHM mechanisms to rapidly evolve reporter proteins like the Green Fluorescent Protein. This approach uses a stable, non-immunoglobulin locus within the genomes of human B cell lines. To comprehensively characterize the mutational profile and breadth of this strategy, we performed single-molecule sequencing experiments. We illustrate the utility of the platform by rapidly evolving antibody fragments in a continuous manner by B cell surface display targeting avian subtypes of influenza hemagglutinin (e.g., H5) from emerging strains.

Introduction

Random mutations followed by selection is one of the key mechanisms that results in evolution of new functions in nature. Typically, mutational rates are low so as not to compromise the fitness of the organism as a result of genome-wide mutations. However, this protection against error also results in a relatively slow diversification of biomolecules and the subsequent evolution of new functions. For example, in a typical bacterial or human cell the mutational frequency is around 10^{-10} to 10^{-9} substitution per base.[1] Therefore, for a mutation to occur within a single gene of around 1000 bases, a cell would have to undergo more than a million rounds of cell division. The rate of evolution is further constrained by the number of mutations a genome can tolerate before the organism is not viable. This limit is especially important in the context of directed evolution experiments that rely on biomolecular diversity to evolve biomolecules (proteins, RNAs) with desired functions. Directed evolution has traditionally relied on *in vitro* techniques, such as utilizing randomized oligonucleotide sequences or error-prone polymerase chain reactions (PCR), to diversify genes encoding target biomolecules rather than relying on low mutation frequencies *in vivo*. After *in vitro* diversification, subsequent steps in directed evolution involve screening or selection methods to enrich and identify variants that exhibit a desired phenotype, followed by detailed characterization to establish the link between genotype and phenotype. Given the process's iterative requirements, each round of a directed evolution campaign requires generating biomolecular diversity in vitro before proceeding with screening and selection, rendering it a labor-intensive and time-consuming approach to profile the evolutionary landscape and identify biomolecules with desired properties.[2–7]

To overcome some of these limitations, various continuous directed evolution platforms have been developed[8]. Orthogonal replicons with high mutagenic rates have been established in model bacteria like *E. coli*[9] or model eukaryotes like yeasts[10,11]. Genes encoding proteins of interests are encoded in orthogonal replicons, and are subjected to continuous mutagenesis by expressing an error prone DNA polymerase that specifically replicates the orthogonal replication system and not the genomic DNA. This approach results in the continuous generation of a library of biomolecular variants within the cells

without compromising their viability due to genomic mutations. The cells expressing a library of variants are screened under selection conditions to identify cells expressing variants resulting in a desired phenotype (e.g., antibiotic resistance, protein/protein binding). This step is typically followed by sequencing experiments to identify the mutant gene encoding the selected protein variant. Similar to orthogonal replicons, transcription coupled mutagenesis platforms (e.g., T7 RNA polymerases that have been fused to deaminases) have been engineered to create DNA mismatches, which upon repair result in the generation of biomolecular diversity.[12–14] In addition to these cell-based approaches, virus-based continuous evolution approaches have been developed.[15–17] In these platforms either the viral genome or viral helper plasmids encode genes corresponding to the protein of interest. Biomolecular diversity is generated by either engineering mutagenic polymerases or by expressing mutagenic factors. These approaches are typically linking evolutionary outcomes (binding, enzyme activity) to viral propagation. Each of these approaches requires extensive engineering efforts and all continuous directed evolution efforts in mammalian cells are typically virus mediated approaches.[16,17]

The studies described here were inspired by the mammalian adaptive immune system, which has evolved mechanisms to overcome the evolutionary bottleneck of low error frequencies and mutational tolerance to rapidly evolve proteins in response to antigen exposure. B cells use DNA rearrangement, recombination and somatic hypermutation (SHM) mechanisms to rapidly evolve antibodies.[18–21] The SHM machinery allows B cells to introduce mutations specifically at immunoglobulin genomic loci at a significantly higher frequency than the rest of the genome.[19] This selective diversification allows B cells to rapidly evolve new antibody sequences without compromising genomic fitness due to mutational burden, thus providing an attractive feature for biomolecular evolution. Indeed, methods have been used to evolve antibodies targeting defined epitopes[22] as well as for evolving catalytic antibodies[23]. Moreover, previous studies have attempted evolution of optimized B cell receptor sequences by incorporating the corresponding genetic elements at the immunoglobulin loci of B cells.[24–27] However, such loci are genetically unstable due to recombination and hypermutation events.[28,29] B cells have also been engineered to

express custom antibodies targeting defined epitopes as potential therapeutics.[25,26,30] Of particular interest to this work, a pioneering study by Tsien and coworkers demonstrated that low frequency-virus based genomic integration methods resulted in transient expression and evolution of fluorescent proteins when the corresponding gene was integrated at the immunoglobulin locus.[31] These studies led us to investigate if the SHM mechanisms of B cells can be hijacked and repurposed to engineer proteins of interest, using human B cell lines as chassis for continuous directed evolution.

Here, we describe a novel Continuous Directed Evolution platform in Human B cell lines (CODE-HB) that recruits and repurposes the inherent B cell SHM mechanisms to rapidly, and orthogonally, evolve proteins of interest. We first developed virus free genome editing approaches to incorporate defined genetic elements at a stable, non-immunoglobulin locus in human B cell line genomes. To continuously evolve reporter proteins expressed from this locus, we added defined DNA elements upstream of the gene encoding a reporter protein to facilitate the recruitment of B cell SHM machinery to this non-immunoglobulin stable locus. We demonstrate that these engineered B cell lines can rapidly evolve reporter proteins (e.g., enhanced Green Fluorescent Protein, eGFP). We comprehensively characterize the mutational profile and breadth of this approach by using single-molecule sequencing experiments. To adapt this platform to rapidly evolve antibody fragments in a continuous manner, we developed a B cell surface display platform for displaying Fragment antigen binding domains (F_{ab}) of antibodies. Combining the B cell surface display platform with CODE-HB, we evolved F_{ab} sequences targeting avian subtypes of influenza hemagglutinin, identifying H5 selective antibodies. Such antibodies are important in the context of recent zoonotic spread of H5N1 strains of avian influenza.[32] Given the modularity and simplicity of CODE-HB, we anticipate that this platform can be used for developing biologics as well as for proteins and enzymes with desired properties. Furthermore, the engineered stable system can be used to provide new insights into the SHM mechanisms.

Results:

Precise and stable genome editing of human B cells for reporter protein expression

Our first goal was to develop an approach that would allow incorporation of defined genetic elements at a defined stable locus in B cell lines. Most commonly, lentivirus-based approaches have been used to randomly incorporate genetic elements in the genomes of B cell lines.[31] Recently, CRISPR/Cas9 plasmids along with adeno-associated virus (AAV) containing recombination cassettes have been used in B cells to replace endogenously encoded antibodies with heterologous ones targeting specific antigens[30]. For our studies, we investigated if we could develop a relatively simple, virus-free, plasmid DNA-based approach to edit genomes of B cell lines with a goal to recombinantly express reporter proteins (Figure 1A-C) from a defined and stable genomic locus. We envisioned using a two-plasmid system composed of (i) a CRISPR/Cas9 plasmid DNA containing genetic elements to express Cas9 and guide RNA targeting loci of interest, and (ii) a separate donor plasmid containing homology regions to a specific locus on the B cell genome as well as DNA sequence encoding a defined integration cassette. For these studies, we used a B cell line, RA 1, where the SHM mechanisms have been extensively studied.[29,29,33] We used donor plasmid DNA, p274, and CRISPR/Cas9 plasmid DNA, p276,[34] that were compatible for integration at the safe harbor H11 locus (Chr.22: 31437323 - 31439351) at chromosome 22 in the human genome. Importantly, this locus is a non-immunoglobulin locus in B cells with no reported instability or recombination[18,35]. We then optimized DNA formulations and electroporation conditions to transfect RA 1 B cells with plasmid DNA (see methods section for details). Next, we generated a homology-directed recombination construct with ~700 bp homology arms to the H11 integration site on chromosome 22. Between these homology arms, we placed an EF1a promoter to drive eGFP expression, and a PGK promoter to drive puromycin resistance cassette (puroR) expression. Finally, we developed recovery and puromycin selection conditions for RA 1 cell lines, where selection was tracked phenotypically by eGFP expression (Supplementary Figure 1 and methods section). The cells were then analyzed by microscopy and fluorescence activated cell sorting (FACS), to determine the levels of eGFP expression. As shown in Figure 1D-E, at least 90% of the total cells demonstrated GFP signals. We also confirmed the integration of the desired recombinant DNA at the

targeted integration site in chromosome 22 by isolating the total genomic DNA, followed by PCR analysis (Supplementary Figure 2).

Engineering human B cell lines to continuously evolve a reporter protein

Having established plasmid-based methods for precise genome editing in B cell lines, our next efforts were directed towards repurposing human B cell lines for continuous directed evolution. As mentioned earlier, viral based integration of fluorescent proteins within the IgH locus facilitated their evolution.[31] This prior study provided insights into the genomic locus that preferentially underwent SHM due to its ability to attract transcription factors conducive to this process. Tsien and coworkers observed that a series of variants of fluorescent proteins evolved when the corresponding gene was recombined into the genomic Ig heavy-chain locus in chromosome 14 between gene IgHV7-34-1 and IgHV4-34 genes. We analyzed the sequences upstream of this integration site and identified the presence of a putative SHM enhancer element (henceforth termed proXIV) just upstream of where the integration cassette would have been found. This enhancer element was identified based on two criteria – its proximity to IgHV4-34 and its encoding of a recognition site for the *oct* transcription factor.[36,37] Based on these observations, we added the proXIV SHM enhancer element upstream of our reporter protein sequence, into the stable H11 locus to promote SHM of downstream genes within our genomic integration site in RA 1 cell lines. To test the system (Figure 2A-C), we used an eGFP variant, T66I, that lacked detectable fluorescence as a reporter protein (eGFP*). We thus placed proXIV DNA sequence upstream of the gene encoding eGFP* and the EF1a promoter (plasmid DNA, pSB1). If proXIV was able to successfully recruit endogenous SHM machinery to the non-immunoglobulin locus where the gene encoding eGFP* was integrated, we would expect mutations to accumulate in this gene, with some of the mutations resulting in the reversion of the non-fluorescent eGFP* to fluorescent eGFP, or variants thereof. To convert eGFP T66 to I66, we modified the codon from ACC to ATC. This alteration was designed to leverage the activity of activation-induced deaminase (AID), a key enzyme in SHM that deaminates cytidine residues to uridine[19]. This results in a mismatch and initiates error-prone base excision repair (BER), leading to mutations. A common result of this process is the transition of T to C.[38] Therefore, if this transition

mutation occurs in our experiment, it would revert eGFP* back to its original eGFP form. We also made another control DNA construct that had eGFP* driven by the EF1a promoter but that did not contain proXIV (plasmid DNA, pSB2).

We transfected RA 1 cells with plasmid DNAs p276 and pSB1. We then recovered and selected RA 1 cell lines where the DNA segment containing proXIV, EF1a promoter, eGFP, and the puromycin cassette was integrated into chromosome 22, creating the stable RA 1-proXIV-eGFP* cell line (RA 1-SB1). For control experiments, we transfected RA 1 cells with plasmid DNA p274-eGFP*, generating the stable control RA 1-eGFP* cell line, which lacked the proXIV sequence upstream of the eGFP* sequence (RA 1-SB2). For both cell lines we confirmed the integration of the DNA segment at the chromosome 22 locus using PCR analysis (Supplementary Figure 2). Both cell lines were propagated for multiple passages and at each passage, the cell lines were monitored for the appearance of fluorescence signal corresponding to a reverted eGFP variant. After five passages, we observed the appearance of a fluorescent signal corresponding to eGFP in RA 1-proXIV-eGFP* but not in RA 1-eGFP* cell lines (Figure 2D and Supplementary Figure 3). Subsequent flow cytometry analysis confirmed the presence of significant green fluorescence within a subpopulation of RA 1-proXIV-eGFP* cells (Figure 2E). We then sorted these cells using fluorescence-activated cell sorting (FACS), isolated their RNA, and generated the cDNA corresponding to the eGFP* transcript using reverse transcriptase-PCR (RT-PCR) methods. We then sequenced the amplicon corresponding to eGFP* to detect mutations via Sanger sequencing. In several clones, we observed the reversion of I66 to T66 (Supplementary Figure 4A-B), suggesting that the sequences upstream of eGFP* encoding sequences recruited the SHM machinery to a non-immunoglobulin locus and introduced mutations within this gene. To understand the mutational scope of this approach we performed deep sequencing experiments discussed below.

Single-molecule long read sequencing to elucidate the mutational profile of eGFP*

To comprehensively characterize the mutational profiles within a single eGFP* transcript generated using this approach, we performed single molecule long read PacBio

sequencing. We sorted GFP-positive cells from eGFP* evolution experiments using FACS, isolated RNA, generated cDNA and used the cDNA samples for sequencing. The sequencing reads (Figure 2F and Supplementary Figure 4C-E) demonstrated that the introduced mutation in *eGFP* (C197T, nucleotide change) to produce *eGFP**, was reverted to its original fluorogenic sequence (T197C, nucleotide change). This reversion accounted for 58.52% of the reads. Additionally, the second most prevalent mutation, T197G, converted Ile66 into Ser, effectively recreating the chromophore tripeptide sequence of native GFP, Ser-Tyr-Gly, accounting for 10.19% of mutational reads (Supplementary Table 5).

Other notable mutations included G310A (9.98%) and A596G (7.71%). The G310A substitution changed Asp104 to Asn. The A596G substitution changed Asn199 to Ser. Additional substitutions, although less frequent, also contributed to the diversity of amino acid changes. For instance, C208A changed Gln70 to Lys, introducing a positive charge; C127G changed Leu43 to Val; and T50A changed Val17 to Gly, potentially introducing flexibility. It is worth noting that the T197C substitution frequently occurred both independently and in conjunction with other mutations. The co-occurring mutations that were most enriched in the population alongside T197C include G310A, A596G, C208A, and C127G. Notably, T197A (Ile66Asn) is the only mutation that does not co-occur with T197C, as they are mutually exclusive due to both being substitutions at the same site. This observation suggests that an initial eGFP* integrant may have reverted to T197C, but over the course of passaging, began to accumulate additional mutations such as G310A, A596G, C208A, and C127G. The independent occurrence of T197A implies it might also restore fluorescence, though this requires further experimental validation. The abundance of these additional mutations in the sorted sample may be attributed to their neutral effect on the function of eGFP; these mutations may neither impede nor enhance the protein's fluorescent properties, allowing them to persist in the population without affecting the overall selection for fluorescence. Interestingly, in some cases, T197C was accompanied by as many as three other single point mutations [C127G, C208A, A305G]. We also detected A196G/T197C double mutations, in which Ile66 was mutated to alanine, a known GFP variant.[39,40] The only other A196 mutation was A196C, which was minimally

detected and generates a proline residue not previously described in the GFP literature (based on BLAST-P analysis). A wide range of other mutations (resulting in amino acid changes) were observed in these sequencing experiments (see Supplementary Table 5 and Supplementary Data File 1), demonstrating the remarkable scope of mutations that can be accrued within the gene using this approach. We call this continuous directed evolution platform as CODE-HB.

Evolutionary trajectory of eGFP* evolution using CODE-HB

In order to understand the continuous nature of evolution using CODE-HB, we next constructed phylogenetic trees of the eGFP variants detected in our evolution experiments (Figure 3A). One prominent branch of variants that likely emerged from the initial introduction of the T197C mutation, which reverted the mutated eGFP* back to its original eGFP form. Following this reversion, additional point mutations occurred at sites where substitutions likely did not abolish fluorescence and resulted in lineages such as T197C → C46A and T197C → C46A/T50A/T197C that demonstrated the progressive acquisition of mutations. Interestingly, the T50A mutation also emerged independently without the prior addition of C46A, as observed in lineages like T197C → T50A and T197C → T50A/G58T/T197C. In some cases, the lineage T50A/G58T/T197C acquired the C46A mutation to become a convergent variant C46A/T50A/G58T/T197C. Notably, the T50A mutation had arisen alongside other T197 substitutions such as T197G and T197A, suggesting that T50A may be a beneficial mutation for fluorescence enhancement. Another distinct family introduced mutations at positions C127G and C208A into the T197C background. These mutants, originating as double mutants from T197C, converged to a common variant: C127G/T197C/C208A. This lineage could further acquire additional mutations, resulting in variants like C127G/T197C/C208A/A305G, C127G/T197C/C208A/A596G, and C127G/T197C/C208A/G310A. We also observed evidence of evolutionary convergence with the T50A family, producing variants such as T50A/C127G/T197C/C208A.

Instances were observed where the initial mutation did not occur at position T197 but at different sites, subsequently acquiring a T197 substitution. For example, the mutation

A596G occurred first, followed by the emergence of T197A, T197C, or T197G mutations. Variants like T197C/A596G further gained additional mutations, leading to T197C A596G C654T or T197C G575T A596G. Additionally, mutations such as G8A, G13A, G14A, C127G, and G310A were found to acquire T197G or T197A mutations to restore fluorescence. A family of T197A mutants was also identified, serving as the origin for subsequent descendant lineages.

Mutational bias during eGFP* evolution experiments using CODE-HB

A general mutational bias favoring GC base pairs was observed (Figure 3C). Specifically, GC pairs frequently transitioned to AT pairs. Guanine (G) bases were substituted by either adenine (A) or thymine (T) at relatively equal frequencies, while cytosine (C) bases most often mutated to adenine, followed by thymine, and less frequently to guanine. Conversely, AT pairs displayed a preference for substitution to GC pairs, primarily through transitions. Adenine (A) bases predominantly mutated to guanine, whereas thymine (T) bases commonly transitioned to cytosine. Additionally, cytosine bases showed a higher likelihood of undergoing transversions, whereas guanine bases exhibited a balanced probability of undergoing either transitions or transversions.

A plot of the cumulative occurrences of mutations revealed a modest overall trend line of approximately 0.27 unique instances per base pair (Figure 3B-D). This trend line was interrupted by sections exhibiting acute increases in slope, corresponding to regions acting as hotspots for mutation. These acute increases in slope were quantified at 2.23, 1.349, 1.154, 0.98, and 0.78 occurrences per base pair, respectively, with each subsequent increase being progressively smaller. Notably, the highest increases were observed towards the beginning of the sequence. When these mutational hotspots along the nucleotide sequence were mapped onto the protein structure, they localized to a lateral portion of the enhanced green fluorescent protein (eGFP) barrel, proximal to and including the chromophore (Figure 3E). This spatial correlation suggested that mutations in this area might influence the fluorescence of eGFP, either directly by altering the chromophore or indirectly through interactions that affected the chromophore's environment.

Developing facile F_{ab} surface display platform in human cells

Next, we used CODE-HB to evolve antibody fragments targeting epitopes from emerging strains of avian influenza. We first developed a B cell surface display platform to display fragment antigen binding regions (F_{ab}) of antibodies on the surface of RA 1 cells (Figure 4A-B and Supplementary Figure 5). We selected the F_{ab} region of a well characterized broadly neutralizing antibody, CR9114.[41] Following a method inspired by Moffett et al.[30], a single-chain antibody binding fragment (F_{ab}) was engineered by fusing the light and heavy chains with a 60 amino acid glycine-serine repeat (GlySer) linker. This fusion construct was further modified for surface display by adding a membrane localization sequence at the N-terminus and a major histocompatibility complex (MHC) I transmembrane helix at the C-terminus, facilitating its stable membrane tethering for external presentation.[42] The DNA encoding this chimeric protein sequence (F_{ab}-surface display protein) was subcloned into p274 where the expression of the F_{ab}-surface display protein was driven by the EF1a promoter. Furthermore, we inserted a FLAG-tag between the constant region of the heavy chain and the transmembrane helix to facilitate detection of construct expression independently from antigen binding. The resulting plasmid was called pSB3. To evaluate whether this surface display platform was functional in human cell lines, we transfected the Expi 293 cell line with plasmid pSB3. The expression of the F_{ab} was evaluated by binding of APC-conjugated Anti-FLAG antibodies to these Expi 293 cells. The binding of the surface displayed F_{ab} to the reporter hemagglutinin A, H1, was evaluated by using the fluorescent signal of a C-terminally fused mCherry. Analysis of expression and binding was performed using flow cytometry. To ensure that the emissions of the two fluorophores were distinctly separated, mCherry was excited with a 561 nm laser and detected through a 613/18 nm bandpass filter, while APC was excited with a 640 nm laser and detected using a 660/10 nm bandpass filter. Using this approach, we were able to distinctly delineate expression and binding due to non-overlapping fluorescence signals corresponding to binding (mCherry) and expression (APC) (Supplementary Figure 6A-B). This distinction facilitated accurate and independent evaluations of both expression levels and binding efficacy at the single cell level. Furthermore, the data revealed that the population of CR9114 F_{ab}, when displayed on the

cell surface, maintained its functional integrity, demonstrating effective binding across a spectrum of hemagglutinin subtypes, including H3 and H7 (Supplementary Figure 6C).

Next, we validated that this surface display platform was functional in B cells as well. We transfected plasmid pSB3 and p276 in B cells, recovered them and selected them as before to generate RA-SB3 cell lines. These RA-SB3 cell lines were analyzed for expression of F_{ab}-surface display protein using APC-conjugated Anti-FLAG antibody and the binding of the surface displayed F_{ab} to H1 was evaluated using an H1-mCherry fusion protein. We observed robust expression and H1 binding for these RA 1 mutant cell lines displaying CR9114 F_{ab} (Figure 4C).

Evolving antibody fragments by combining F_{ab} surface display with CODE-HB

Our next goal was to use CODE-HB to evolve the F_{ab} sequences. Particularly, we investigated if the CR9114 F_{ab} can be evolved to exhibit a greater degree of binding towards hemagglutinin A (HA) from emerging strains of avian influenza (H5); identifying H5 selective antibodies is especially important in the context of recent zoonotic spread of H5N1 strains of avian influenza.[32] We started by inserting the CR9114 F_{ab} surface display protein into a B cell integration vector containing the proXIV enhancer sequence to generate plasmid pSB4 (Figure 5A). We transfected plasmid pSB4 and p276 into RA 1, recovered cells and selected them as before to generate RA 1-SB4. We also utilized the control plasmid pSB3 that was identical to pSB4 but lacked proXIV. Using this integration plasmid, we generated a control cell line RA 1-SB3. To refine the selection process for cells exhibiting increased affinity towards H5, a critical examination was undertaken to distinguish between levels of construct expression and binding efficiency (Figure 5B). Utilizing a control cohort transfected with a vector devoid of the proXIV sequence facilitated the establishment of a normative regression line for binding relative to expression. This baseline enabled the definition of a "High Binding" gate for events surpassing expected binding levels based on expression alone (Figure 5C). Cells that exceeded this benchmark within the proXIV-inclusive sample, yet not present in the control, were posited as having undergone favorable somatic hypermutations within their recombinant CR9114 construct sequences.

Upon utilizing the "High Binding" gate for selection, an enrichment of approximately 40,000 cells was achieved that were cultivated to saturation. Subsequent iterations of FACS sorting and analysis of this population highlighted an intriguing divergence from the control group's fluorescence profile, characterized by a decreasing mean fluorescence intensity in the expression channel, contrasted with a consistent intensity in the binding channel (Supplementary Figure 7A). This unexpected trend prompted a deeper molecular investigation, leading to the isolation of RNA from the sorted population and Sanger sequencing of the CR9114 amplicon. The sequencing results revealed a prevalent mutation that specifically resulted in the conversion of Tyr532 to Phe532 (TAC to TTC), adjacent to the FLAG-tag region, suggesting that a targeted selection pressure in this area resulted in a phenotype that could be detected (Supplementary Figure 7B). These observations demonstrated that a phenotype can be linked to the genotype and can be evolved continuously by using CODE-HB.

In addition to this population, we also observed a dynamic population of high binders and low binders as compared to the control group (Figure 5D and Supplementary Figure 7C). The enrichment process showed a marked improvement in target population isolation across successive sorting steps. Initially, the percentage of events that escaped the control wild-type binding gating was 2.97%. This number increased substantially to 94.48% after the second sort, and further to 97.07% following the third sort, demonstrating progressive refinement of the target cell population (Supplementary Figure 8). Correspondingly, the histogram data illustrated a significant rise in the mean fluorescence intensity (F.I.) of the H5-binding channel (FITC+), with values of 361.8, 716.87, and 1155.62 in the initial, second, and third sorts, respectively (Figure 5E-F). For this verification experiment His-tagged H5 was used and FITC-conjugated anti-His antibody was used to detect binding of cells to His-tagged H5. Post-sorting, the cells were expanded, and their RNA was isolated for subsequent conversion to cDNA.

Single-molecule long read sequencing to elucidate the mutational profiles of evolved F_{ab}

Once the cDNA of the transcript corresponding to the F_{ab}-surface display protein was generated, long read single molecule sequencing was performed (Figure 6A). Third generation sequencing generated a total of 642,858 reads, of which 471,562 met the expression threshold based on published PacBio error rates (15 observations).[43]

Further analysis focused on heavy chain variants encompassing both the variable (Hv) and constant (Hc) regions, due to the heavy chain-only binding mode of CR9114 to HA as indicated by the crystal structure (PDB: 4CQI).[44] Unique reads from the top 200 variants were shortlisted, resulting in a final set of 16 variants that included both variable and constant regions of the heavy chain. Notably, these variants represented only 0.25% of the total reads.

Analysis of these variants revealed a variety of nucleotide substitutions, including transitions, transversions, insertions, and deletions (Supplementary Figure 9A-F and Supplementary Data File 2). Approximately 22% of the unique read sequences were discarded because they introduced insertion or deletion mutations when aligned to the reference sequence. Of the remaining sequences, 82% corresponded to a single mutational profile comprising at least two mutations: A1149G and A1601T. The first mutation, A1149G, was a silent mutation that altered the codon of a glutamic acid residue without changing the amino acid sequence. The second mutation, A1601T, resulted in an amino acid change from tyrosine to phenylalanine within the FLAG tag sequence (tAc→tTC, Tyr→Phe; see Supplementary Figure 9B). The subsequent multiple sequence alignment (MSA) included 208 unique sequences, from which the average mutation rate was calculated to be 1.73 mutations per kilobase.

We performed bioinformatics analysis to elucidate the evolutionary trajectory of the diversified F_{ab} variants. The initial variant identified was A1601T, detected by a decrease in binding to APC-conjugated anti-FLAG antibodies. The mutation of a tyrosine residue within the FLAG tag to phenylalanine may potentially reduce antibody recognition by eliminating a hydroxyl-moiety that participated in binding to the secondary antibody. The A1601T variant subsequently gave rise to the double mutant A1149G A1601T, from which

all the tested variants in the study emerged (Supplementary Figure 10). Interestingly, the maximum count for sequential descendent branches was 4, evidenced by the lineage A1601T → A1149G, A1601T → C107A/A1149G, and A1601T → C107A/G119T/A1149G/A1601T – a count similar to that achieved from the eGFP* experiments. Following the nucleotide-level analysis, remaining reads were translated, and protein level variations were analyzed. Many amino acids were not altered (Supplementary Figure 11). Across the F_{ab} sequence, two mutational hot-spots appeared. The first, while not part of the shortlisted set, sat within the light chain region of CR9114 specifically between amino acid positions Thr128-Leu144. Within this region, a marked increase in mutational frequency was observed, with the exception of Pro142, which was never mutated. Mutations within this region were primarily to another single amino acid, with few residues showing mutation to two different amino acids. The second mutational hot-spot was found within the heavy chain of the CR9114 F_{ab} region. In contrast to the light chain region mutational hot-spot, within this region the specific residues that showed alterations were: Gly343Arg, Met347Val, Ile373Phe, Phe394Leu, and Phe394Ser. Beyond these variants, a variety of other mutations were present across the F_{ab} sequence with varying frequency (Supplementary Data File 2).

Validating CR9114-F_{ab} variants for increased binding towards hemagglutinin H5

Next, we identified and validated F_{ab} sequences that demonstrated enhanced binding. Functional characterization of these variants (Figure 6A-B and Supplementary Figure 12) involved determining their EC50 values through hemagglutinin H5 titration against the surface-displayed form. Overall, the variants exhibited lower EC50 values compared to the CR9114 wild-type control, reflecting enhanced binding affinities. Among these, two variants, Ser427Pro and Trp461Arg, demonstrated the lowest EC50 values of around 4.2 nM (Figure 6C-D). These values were approximately three times lower than that of the CR9114 wild-type control, highlighting the superior, yet modestly improved, binding affinity of these selected variants. Interestingly, the two variants with the lowest EC50 values, Ser427Pro and Trp461Arg, both contained mutations in the heavy chain *constant* region. This finding is particularly noteworthy as it suggested that alterations in the constant region, typically associated with structural and stability roles rather than direct

antigen binding, can significantly influence the binding affinity of the antibody by optimizing an already-evolved binding interface. This observation highlights the potential for targeted mutations in this constant region to improve therapeutic antibody efficacy. The mutant W461R was tested for retention of binding towards other hemagglutinin variants (H1, H3, H5) and was found to not only retain, but also exhibit more favorable EC50 values than the wild-type CR9114 (Supplementary Figure 13).

In addition to these observations, variants in the heavy *variable* region that demonstrated enhanced binding relative to the controls were also detected (Figure 6B-C). The Ile373Phe variant of CR9114 introduced a point mutation within a critical complementarity-determining region (CDR) loop on the heavy chain, which interfaces directly with the hemagglutinin H5 antigen (crystal structure PDB: 4CQI).The Met347Val variant of the CR9114 antibody is another point mutation within the beta sheet of the heavy chain variable region, a critical area for maintaining the antibody's structural integrity. This region, situated between a beta strand and an alpha helix, displayed a degree of torsion established by this potential axis of interaction that was flanked on all sides by aromatic residues. The Gln412Arg variant of the CR9114 antibody involved a point mutation within a flexible loop of the heavy chain variable region, situated near a hydrogen bond network comprising residues Gln6, Thr107, Ser7, and Ser21. According to the crystal structure (PDB: 4CQI), the amide nitrogen of Gln412 resided approximately 5.5 angstroms from the oxygen of Thr107, though the flexibility of the loop might allow it to migrate closer. Q6 was positioned 3 angstroms from Thr107, forming part of this network. With the substitution of glutamine to arginine at position 412, an extended carbon chain and an additional interacting nitrogen were introduced, which may have enhanced the stability of the hydrogen bond network as a whole. This stabilization could have contributed to the upstream integrity of the CDR loops, essential for antigen binding (Supplementary Figure 14).

The results herein underscore the potent application of CODE-HB for the refinement of antibody engineering, highlighting a methodology to not only uncover new variants, but also uncover unique functional insights related to the evolved molecules.

Discussion

Directed evolution platforms have significantly advanced the ability to rapidly evolve biomolecules with defined functions. Advances in molecular biology and synthetic biology further resulted in the development of continuous directed evolution platforms where cell or virus-based systems are used (see introduction section). One of the key advantages of these platforms is the ability of engineered cells to introduce biomolecular diversity *in situ* without requiring *in vitro* library generation. Another attractive feature of some of these systems is the design where the evolutionary outcomes are linked to the propagation of cells or viruses with selective growth advantage. Several other efforts are ongoing to build, optimize and utilize continuous directed evolution platforms in model bacteria, yeasts, mammalian cells and viruses. Our efforts were inspired by the SHM mediated rapid antibody evolution in B cells. Several studies have extensively characterized the biochemistry and intrinsic mutation rates of SHM, and the timing of SHM in relation to phenotype evolution in B cells and B cell lines, including RA 1 cell lines.[20,29,45]

We focused our efforts on investigating if the SHM mechanisms in B cells can be repurposed for engineered continuous directed evolution. Through a series of human B cell engineering efforts, we developed a continuous directed evolution platform in human B cell lines that recruits and repurposes the SHM mechanisms, to rapidly and orthogonally evolve proteins of interest from a stable, safe harbor locus within human B cells. We demonstrated that this platform can be used to rapidly evolve eGFP variants. We also developed a B cell surface display platform for displaying Fragment antigen binding domains, F_{ab}, of antibodies. Combining the F_{ab} surface display platform with CODE-HB, we evolved F_{ab} sequences targeting avian subtypes of influenza hemagglutinin (e.g., H5). We performed extensive single-molecule sequencing and bioinformatics studies to comprehensively characterize the mutational profile and scope of this approach. Our phylogenetic analysis demonstrated a continuous evolution of various lineages of variants in each evolutionary experiment. Further, based on our observations, the estimated mutation rate achieved by the system was around 5×10^{-5} to 9×10^{-5} substitutions per base per generation, translating to 5×10^{-2} to 9×10^{-2} substitutions per kilobase per generation (0.005% to 0.009%, based on both eGFP* and F_{ab} evolution experiments,

Supplementary Table 6). This rate can be contextualized by comparing it with other mutagenesis systems. Phage-Assisted Continuous Evolution (PACE), demonstrates a higher mutation rate of 7.2×10^{-3} substitutions per base per generation, which equates to 2.3 substitutions per kilobase (0.23%)[46]. Mutazyme, an error-prone DNA polymerase from Agilent Technologies, operates within an even higher mutation rate range of 0.3% to 1.6%. In contrast, the OrthoRep system reports a mutation rate of 1×10^{-5} per base per generation, translating to 1×10^{-2} substitutions per kilobase per generation (0.001%).[10] These comparisons illustrate that this system's mutation rate is intermediate among commonly used mutagenesis platforms - it is significantly lower than that of PACE and Mutazyme, yet higher than that of OrthoRep, potentially allowing for more rapid generation of genetic diversity. We believe that it might be possible to further modulate the mutation rates of CODE-HB by incorporating various SHM recruiting sequences that have been previously identified.[47]

The evolutionary outcomes for the F_{ab} fragment primarily resulted in solutions containing at least three mutations, with EC_{50} improvements being modest—approximately threefold over the wild type—compared to the 100- to 1000-fold enhancements often reported in other studies. However, this observation should be contextualized within the framework of the experimental design and the characteristics of the starting sequence. These experiments were initiated with CR9114, a mature antibody that had already undergone extensive optimization through somatic hypermutation.[41] This prior optimization suggests that CR9114 was likely approaching its affinity and functional ceilings within the given sequence space. In highly optimized systems like CR9114, additional enhancements tend to be incremental rather than drastic. Moreover, the limited timeframe of the experiments (~20 passages) may not have allowed for the accumulation of multiple beneficial mutations required for substantial functional changes. The observed solutions are consistent with the concept of diminishing returns in protein evolution, where each successive mutation offers smaller gains. Therefore, while the improvements achieved were modest, they are meaningful within the constraints of starting with a highly optimized, globular antibody and conducting the experiments over a limited time period.

We anticipate that this platform can be used for developing biologics as well as for peptides, proteins and enzymes with desired properties. This platform may allow the exploration of human cells for continuous directed evolution of medicinally relevant proteins like antibodies, antibody fragments, fluorescent proteins, metabolic enzymes, reporters and other membrane proteins that often fail to function properly *in vivo* when evolved in bacterial or yeast cells.[14] We anticipate that these synthetic systems also have the potential to be repurposed to elucidate finer molecular details of the SHM mechanisms.

Materials and Methods:

Bacterial strains, mammalian cell lines, growth media, DNA constructs.

E. coli DH5α was used for all transformations and bacterial cultures and grown in either LB or 2xYT medium. Expi 293 were used for transient transfection and initial construct testing and were grown in Expi 293 Expression Medium. RA 1 (ATCC CRL-1596) were used for SHM experiments and grown in RPMI 1640 Medium supplemented with 10% FBS; both cell lines were validated by STR analysis. The oligonucleotides used in this study are listed in Supplementary Table 1, the synthesized DNA fragments are listed in Supplementary Table 2, the plasmid DNAs used in this study are listed in Supplementary Table 3 and the B cell lines generated in this study are listed in Supplementary Table 4.

Preparation of plasmid DNA samples for transfection of RA 1 cells.

Using the QIAprep® Spin Miniprep Kit (Qaigen: 27104), plasmids for transfection were prepared from E. coli cultures. Initially, E. coli containing the desired plasmids were grown in 2 mL of 2x YT media, inoculated from glycerol stock. The next day, this culture was diluted 1% into 30 mL of LB media in a 125 mL Erlenmeyer flask and incubated at 37°C with 220 rpm shaking. After 24 hours, the culture was collected in 50 mL conical tube and centrifuged at 10,000 xg for 5 minutes, the supernatant was decanted, and the pellet was resuspended in 800 µL of P1 buffer with RNase. Following vortexing, 1 mL of P2 buffer was added and the mixture was lysed for 5-7 minutes with shaking at 150 rpm. Then, 1.4 mL of N3 buffer was added, the tube was inverted for mixing, and left with shaking for 3

minutes to form a precipitate. This was centrifuged at 12,000 xg for 5 minutes, and the clear supernatant was distributed among two spin columns for purification, following the kit's protocol. The DNA was then serially eluted from the two columns using 45 uL of Dnase-free water, using the pass-through of the first column as the elution media for the next column.

Transfection of RA 1 Cells.

RA 1 cells were cultured to a density of 1.5-2.0 million cells/mL using growth media RPMI 1640 + 10% FBS. For transfection, the Mirus Ingenio electroporation solution (Mirus Bio: MIR 50114) was pre-warmed by shaking at 37°C for 10 minutes. Meanwhile, 7 million cells were centrifuged at 100xg for 10 minutes at room temperature, resuspended in 100 µL Mirus electroporation buffer, and combined with 20 µL DNA solution, containing 10 µg integration vector and 10 µg Cas9/sgRNA plasmid. A final volume of 110 µL was then transferred to a 0.2 cm electroporation cuvette.

Electroporation was performed using the BioRad XCell Total Electroporation system (BioRad: 1652660) at 133V, 950 µF, infinite resistance, in a 0.2 cm cuvette, at room temperature. Post-electroporation, cells rested at room temperature for 5-7 minutes, then were resuspended in 1 mL RPMI + 10% FBS and transferred to a 6-well plate in a total volume of 3 mL RPMI + 10% FBS. Cells were then grown in this vessel for 24 hours, washed, and resuspended in 10 mL RPMI + 10% FBS media in a T025 flask, and grown for 48 hours. For determining the concentration of puromycin required for the selection of cell lines with genomic integrant, we titrated puromycin levels in the growth medium. Cell viability was evaluated using trypan blue staining and microscopic examination, as well as assessing the impact on growth rates via cell counts, all aiming to find an optimal puromycin concentration that allowed for selection without irreversibly compromising cell health or inducing irreversible morphological alterations. Using this approach, we determined that the ideal starting concentration of puromycin for effective selection of our RA 1 mutant cell lines was 0.3 µg/mL. The RA 1 cells that were transfected with p274 and p276 were recovered and selected in growth medium containing 0.3 µg/ml of puromycin

for 84 hours. The volume of growth medium was increased in a stepwise manner from 10 mL to 30 mL over the course of one week.

Selection of RA 1 Cells

Transfected RA 1 cells, post-cultivation in T025 flasks for 48 hours, were passaged into fresh T025 flasks at a density of 500,000 cells/mL in 10 mL volume of growth media, supplemented with 0.3 µg/mL puromycin (Sigma Aldrich: P8833). This selection medium was utilized to grow the cells for 84 hours. Subsequently, cell cultures were diluted to 15 mL by adding 5 mL of growth media and incubated for an additional 48 hours. Following this, cells were further diluted to 20 mL with 5 mL of growth media and again grown for 24 hours. The dilution process continued, increasing the volume to 30 mL with an addition of 10 mL growth media for another 24 hours. At the conclusion of the selection regimen, cells exhibiting puromycin resistance were deemed "enriched" for the integration event. These enriched cells were cryopreserved at a density of 10,000,000 cells/mL. The cryopreservation protocol involved resuspending cells in growth media amended with 5% DMSO, followed by a 30-minute chill on dry ice before storage in the vapor phase of liquid nitrogen. Our genomic DNA analysis confirms the integration of the recombinant cassette at the desired locus and suggests the generation of heterozygous mutants (Supplementary Figure 2). Typical transfection efficiencies, as detected by eGFP fluorescence 48 hours post-transfection, range from 2% to 7%. This suggests that, theoretically, less than 10% of the variants in any produced library are incorporated into the final selected population of integrant (based on eGFP integration). The low incorporation rate may certainly limit the diversity of the starting mutant library, but an average transfection would produce up to 350,000 variants as novel starting points from the wild-type sequence to subsequently diversify.

RNA isolation and cDNA preparation

Total RNA from sorted cells was isolated using the PureLink RNA Mini Kit (ThermoFisher: 12183018A), following the manufacturer's instructions. Cells (5,000,000) were resuspended in 600 µL Lysis Buffer with 1% BME, vortexed for cell lysis, and homogenized via a 21-gauge needle. After adding 1.5 volumes of ethanol (96-100%) and

vortexing, the mixture was column-purified and eluted in 18 µL RNase-free water. The isolated RNA underwent reverse transcription with the Superscript III First Strand Synthesis Kit (ThermoFisher: 18080051), per the manufacturers' protocol. For this, 16 µL of the isolated RNA solution was combined with the provided set of 2 µL poly-dT primer and 2 µL dNTPs, heated at 65°C for 5 minutes, then cooled on ice for 1 minute. A mix of 2 µL reverse transcriptase, 8 µL MgCl2, 4 µL DTT, 4 µL 10X RT buffer, and 2 µL RNase Out (total 20 µL) was prepared. Both mixes were combined, vortexed, and incubated for 50 minutes at 50°C, then denatured at 85°C for 5 minutes. Adding 2 µL RNase H and incubating for 20 minutes at 37°C finalized the cDNA mixture that was ready for use in PCR.

PCR Optimization and Amplification from cDNA

PCR optimization for individual cDNA samples involved initial dilution of the cDNA to 10%, which then served as a template in PCR mixtures. PCRs were performed using PrimeStar MAX (Takara: R047A) to maximize replication fidelity. In these reactions, each primer was serially diluted 2-fold from a starting concentration of 1.6 µM. For the amplification of CR9114-based amplicons, the optimization process utilized primers SB 274A and SB 212B. Similarly, eGFP-based amplicons were obtained through optimization with primers SB 193A and SB 193B.

Microscopy

Images were produced by the Bio-Rad ZOE Fluorescent Cell Imager.

FACS Analysis and Sorting

To identify FLAG-tagged F_{ab} displays, APC-conjugated Anti-FLAG antibodies (Biolegend 637308) and FITC-conjugated Anti-His antibodies (ICLLab CHIS-45A-Z) were utilized, targeting the 6xHis tag on hemagglutinins H3 and H5 (SinoBiological 40859-V08H and 11700-V08H, respectively). H1-mCherry was produced by chimeric protein fusion of mCherryXL to the C-terminus of the hemagglutinin H1 ectodomain. FACS analysis and sorting were performed on a BD FACSMelody or BD Accuri C6 system, using a combination of FCS Express 6 and FlowJo11 for FACS plot generation and data analysis.

Binding versus expression was quantified by establishing a linear regression between the FITC and APC signals. Cells displaying APC fluorescence two logarithms higher than untransfected controls were classified as "Expressing". High-binding cells were isolated using triangular gating, adjusted so the gate's hypotenuse was parallel to the regression line and elevated to include no more than 2% of parent gate events. For sorting, 6-8 million cells were processed at the maximum flow rate and collected into a collection tube with 3 mL of growth media, all maintained at 5°C.

Third generation sequencing data analysis:

Multi-FASTA files corresponding to PacBio sequencing data were analysed using Python (3.10.12). Briefly, multi-FASTA files were imported using BioPython (1.59) and collapsed to unique reads and their occurrence. Unique reads with an occurrence threshold below 15 were discarded. Surviving unique reads were then aligned to the reference sequence using BioPython with the following alignment parameters: match_score = 2, mismatch_score = -3, open_gap_score = -5, extend_gap_score = -2, query_right_open_gap_score = 0, query_left_open_gap_score = 0, query_right_extend_gap_score = -1, query_left_extend_gap_score = -1. Alignments which introduced gaps in the interior of the alignment were discarded. Surviving unique reads were then used as a basis for mutational profile calculations. Data handling was done using Pandas (2.2.2). Histogram generation was done using matplotlib (3.9.0). Sunburst plots and pie charts were generated using Plotly (5.22.0)

Structural Analysis and Modeling

Crystal structures of hemagglutinin H5 and CR9114 (PDB: 4FQI) were visualized using PyMOL. The predicted crystal structure of the chimeric CR9114 surface-displayed protein was calculated using AlphaFold3 and visualized using PyMOL.

Calculation of EC50

The EC50 of surface-displayed CR9114 and its variants was determined by fitting the sigmoidal data to a dose-response curve using the Hill equation protocol in OriginPro 2024, and subsequently extracting the EC50 parameter.

Construction of plasmid DNAs.

All cloning related PCRs were performed using PrimeStar HS polymerase (Takara: R040A). The integration plasmid, p274, was sourced from Addgene (ID 164851), and the Cas9/sgRNA plasmid, p276, was also sourced from Addgene (ID 164850). Cloning on p274 was performed by extracting the proXIV sequence from RA 1 genomic DNA using primers SB 168A and SB 168B, which was then cloned directly upstream of the EF1a promoter via Gibson Assembly. Single-point mutagenesis to create an eGFP* variant was conducted using the KLD method with primers SB 171A and SB 171B. To substitute the eGFP component of the native p274 sequence with novel genes, p274 was linearized using primers SB 209A and SB 209B. F_{ab} constructs, synthesized as gBlocks by IDT, were amplified with primers SB 274A and SB 212B and cloned into p274 using Gibson Assembly. For the purpose of F_{ab} evolution, p274 was linearized with SB 209A and SB 275B, introducing an optional BsiWI cut site for restriction/ligation cloning. PCR bands were purified using a combination of the QIAprep® Spin Miniprep Kit (Qaigen: 27104) and GeneJET Gel Extraction Kit (ThermoFisher: K0691). PCR products, in 30 µL reactions, were electrophoresed on a 1% agarose gel at 130V for 14 minutes. The fluorescent bands were then excised and dissolved in 400 µL of GeneJET binding buffer, incubated at 55°C for 10 minutes, and vortexed. To this, 300 µL of 100% ethanol was added and the mixture was vortexed again before being transferred to a spin column. The purification process involved sequential washing with 500 µL of PB Buffer and 750 µL of PE Buffer from the QIAprep® kit, and then 750 µL of Wash Buffer from the GeneJET kit. After centrifugation to dry the spin column, DNA was eluted in DNase-free water.

pSB1: pSB2 was linearized by PCR using the oligonucleotides SB 169 A/B. The proXIV element was amplified by PCR from the genome of RA 1 using the oligonucleotides SB 168 A/B. The amplified proXIV fragment was inserted into linearized pSB2 by Gibson assembly to afford pSB1.

pSB2: p274 was linearized by PCR using the oligonucleotides SB 171 A/B, which encode the eGFP* knockout mutation. The linearized p274 was subject to blunt-end ligation by Kinase-Ligase-DpnI (KLD), to afford the plasmid pSB2.

pSB3: p274 was linearized by PCR using the oligonucleotides SB 209A / SB 275B. A gBlock of the light/heavy fusion protein F_{ab}-CR9114 (PDB: 4FQI) was codon optimized for H. sapiens and amplified in three steps to incorporate the secretion signal, using the oligonucleotides SB 246A / SB 274B, followed by GL 7B / SB 274B, lastly by SB 274 A/B. A gBlock of the linker-FLAG-MHCI helix domain of the chimera was codon optimized for H. sapiens and amplified using the oligonucleotides SB 275A / SB 252B. The two fragments were inserted into the linearized p274 by Gibson assembly to afford the plasmid pSB3.

pSB4: pSB1 was linearized using the oligonucleotides SB 209A / SB 275B. The complete chimeric F_{ab}-CR9114-Linker-FLAG-MHCI helix gene was amplified by PCR of pSB3 using oligonucleotides SB 274A / SB 252B. The gene was inserted into linearized pSB1 by Gibson assembly to afford the plasmid pSB4.

Acknowledgments

The research reported in this publication was supported by the HHMI EPI grant (UI Award #111112). This funding was received by A.P.M. The authors thank Alvaro Hernandez, Chris Wright and Christopher Fields (all at Roy J. Carver Biotechnology Center, UIUC) for helpful discussions on sequencing studies. A.P.M. thanks Prof. Wilfred van der Donk, Prof. Nicholas Wu and Prof. Beth Stadtmueller for helpful discussions on biochemical experiments. A.P.M. thanks Prof. Wilfred van der Donk for reading the manuscript and providing critical feedback.

Author contributions

A.P.M. and S.B. conceived the project. A.P.M., S.B. and G.L. designed experiments. S.B. performed biochemical experiments, microscopy experiments and obtained all data reported in this manuscript. Sequencing was performed at the Roy J. Carver

Biotechnology Center at University of Illinois Urbana-Champaign. J.Q. analyzed the sequencing data. S.L. and J.J.G. provided recombinant hemagglutinin. A.P.M., S.B. and J.Q. analyzed the data. All authors helped with manuscript preparation.

Competing interests

Authors declare that they have no competing interests.

Data Availability

All the data except the raw sequencing files is listed in the manuscript and the supporting information. The raw sequencing datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

References:

1. Drake, J.W., Charlesworth, B., Charlesworth, D., and Crow, J.F. (1998). Rates of spontaneous mutation. Genetics *148*, 1667–1686.

2. Hermes, J.D., Blacklow, S.C., and Knowles, J.R. (1990). Searching sequence space by definably random mutagenesis: improving the catalytic potency of an enzyme. Proc. Natl. Acad. Sci. *87*, 696–700. https://doi.org/10.1073/pnas.87.2.696.

3. Chen, K., and Arnold, F.H. (2020). Engineering new catalytic activities in enzymes. Nat. Catal. *3*, 203–213. https://doi.org/10.1038/s41929-019-0385-5.

4. Winter, G., Griffiths, A.D., Hawkins, R.E., and Hoogenboom, H.R. (1994). Making antibodies by phage display technology. Annu. Rev. Immunol. *12*, 433–455.

5. Xiao, H., and Schultz, P.G. (2016). At the interface of chemical and biological synthesis: an expanded genetic code. Cold Spring Harb. Perspect. Biol. *8*, a023945.

6. Peisajovich, S.G., and Tawfik, D.S. (2007). Protein engineers turned evolutionists. Nat. Methods *4*, 991–994.

7. Zeymer, C., and Hilvert, D. (2018). Directed evolution of protein catalysts. Annu. Rev. Biochem. *87*, 131–157.

8. Morrison, M.S., Podracky, C.J., and Liu, D.R. (2020). The developing toolkit of continuous directed evolution. Nat. Chem. Biol. *16*, 610–619.

9. Tian, R., Rehm, F.B., Czernecki, D., Gu, Y., Zürcher, J.F., Liu, K.C., and Chin, J.W. (2024). Establishing a synthetic orthogonal replication system enables accelerated evolution in E. coli. Science *383*, 421–426.

10. Ravikumar, A., Arzumanyan, G.A., Obadi, M.K., Javanpour, A.A., and Liu, C.C. (2018). Scalable, continuous evolution of genes at mutation rates above genomic error thresholds. Cell *175*, 1946–1957.

11. Arzumanyan, G.A., Gabriel, K.N., Ravikumar, A., Javanpour, A.A., and Liu, C.C. (2018). Mutually orthogonal DNA replication systems in vivo. ACS Synth. Biol. *7*, 1722–1729.

12. Moore, C.L., Papa III, L.J., and Shoulders, M.D. (2018). A processive protein chimera introduces mutations across defined DNA regions in vivo. J. Am. Chem. Soc. *140*, 11560–11564.

13. Chen, H., Liu, S., Padula, S., Lesman, D., Griswold, K., Lin, A., Zhao, T., Marshall, J.L., and Chen, F. (2020). Efficient, continuous mutagenesis in human cells using a pseudo-random DNA editor. Nat. Biotechnol. *38*, 165–168. https://doi.org/10.1038/s41587-019-0331-8.

14. Hendel, S.J., and Shoulders, M.D. (2021). Directed evolution in mammalian cells. Nat. Methods *18*, 346–357.

15. Esvelt, K.M., Carlson, J.C., and Liu, D.R. (2011). A system for the continuous directed evolution of biomolecules. Nature *472*, 499–503.

16. Berman, C.M., Papa III, L.J., Hendel, S.J., Moore, C.L., Suen, P.H., Weickhardt, A.F., Doan, N.-D., Kumar, C.M., Uil, T.G., and Butty, V.L. (2018). An adaptable platform for directed evolution in human cells. J. Am. Chem. Soc. *140*, 18093–18103.

17. Jewel, D., Kelemen, R.E., Huang, R.L., Zhu, Z., Sundaresh, B., Cao, X., Malley, K., Huang, Z., Pasha, M., and Anthony, J. (2023). Virus-assisted directed evolution of enhanced suppressor tRNAs in mammalian cells. Nat. Methods *20*, 95–103.

18. Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

19. Di Noia, J.M., and Neuberger, M.S. (2007). Molecular mechanisms of antibody somatic hypermutation. Annu Rev Biochem *76*, 1–22.

20. Neuberger, M.S., and Rada, C. (2007). Somatic hypermutation: activation-induced deaminase for C/G followed by polymerase η for A/T. J. Exp. Med. *204*, 7–10.

21. Sale, J.E., Bemark, M., Williams, G.T., Jolly, C.J., Ehrenstein, M.R., Rada, C., Milstein, C., and Neuberger, M.S. (2001). In vivo and in vitro studies of immunoglobulin gene somatic hypermutation. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *356*, 21–28.

22. Milstein, C., and Cuello, A. (1983). Hybrid hybridomas and their use in immunohistochemistry. Nature *305*, 537–540.

23. Schultz, P.G., and Lerner, R.A. (1995). From molecular diversity to catalysis: lessons from the immune system. Science *269*, 1835–1842.

24. Seo, H., Masuda, H., Asagoshi, K., Uchiki, T., Kawata, S., Sasaki, G., Yabuki, T., Miyai, S., Takahashi, N., Hashimoto, S., et al. (2021). Streamlined human antibody generation and optimization by exploiting designed immunoglobulin loci in a B cell line. Cell. Mol. Immunol. *18*, 1545–1561. https://doi.org/10.1038/s41423-020-0440-9.

25. Nahmad, A.D., Raviv, Y., Horovitz-Fried, M., Sofer, I., Akriv, T., Nataf, D., Dotan, I., Carmi, Y., Burstein, D., Wine, Y., et al. (2020). Engineered B cells expressing an anti-HIV antibody enable memory retention, isotype switching and clonal expansion. Nat. Commun. *11*, 5851. https://doi.org/10.1038/s41467-020-19649-1.

26. Nahmad, A.D., Lazzarotto, C.R., Zelikson, N., Kustin, T., Tenuta, M., Huang, D., Reuveni, I., Nataf, D., Raviv, Y., Horovitz-Fried, M., et al. (2022). In vivo engineered B cells secrete high titers of broadly neutralizing anti-HIV antibodies in mice. Nat. Biotechnol. *40*, 1241–1249. https://doi.org/10.1038/s41587-022-01328-9.

27. Yin, Y., Quinlan, B.D., Ou, T., Guo, Y., He, W., and Farzan, M. (2021). In vitro affinity maturation of broader and more-potent variants of the HIV-1–neutralizing antibody CAP256-VRC26. 25. Proc. Natl. Acad. Sci. *118*, e2106203118.

28. Goossens, T., Klein, U., and Küppers, R. (1998). Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease. Proc. Natl. Acad. Sci. *95*, 2463–2468.

29. Sale, J.E., and Neuberger, M.S. (1998). TdT-accessible breaks are scattered over the immunoglobulin V domain in a constitutively hypermutating B cell line. Immunity *9*, 859–869.

30. Moffett, H.F., Harms, C.K., Fitzpatrick, K.S., Tooley, M.R., Boonyaratanakornkit, J., and Taylor, J.J. (2019). B cells engineered to express pathogen-specific antibodies protect against infection. Sci. Immunol. *4*, eaax0644.

31. Wang, L., Jackson, W.C., Steinbach, P.A., and Tsien, R.Y. (2004). Evolution of new nonantibody proteins via iterative somatic hypermutation. Proc. Natl. Acad. Sci. *101*, 16745–16749.

32. Eisfeld, A.J., Biswas, A., Guan, L., Gu, C., Maemura, T., Trifkovic, S., Wang, T., Babujee, L., Dahn, R., and Halfmann, P.J. (2024). Pathogenicity and transmissibility of bovine H5N1 influenza virus. Nature, 1–3.

33. Klein, G., Giovanella, B., Westman, A., Stehlin, J.S., and Mumford, D. (1975). An EBV-genome-negative cell line established from an American Burkitt lymphoma; receptor characteristics. EBV infectibility and permanent conversion into EBV-positive sublines by in vitro infection. Intervirology *5*, 319–334.

34. Pryzhkova, M.V., and Jordan, P.W. (2020). Adaptation of human testicular niche cells for pluripotent stem cell and testis development research. Tissue Eng. Regen. Med. *17*, 223–235.

35. Zhu, F., Gamboa, M., Farruggio, A.P., Hippenmeyer, S., Tasic, B., Schüle, B., Chen-Tsai, Y., and Calos, M.P. (2014). DICE, an efficient system for iterative genomic editing in human pluripotent stem cells. Nucleic Acids Res. *42*, e34–e34.

36. Song, S., Cao, C., Choukrallah, M.-A., Tang, F., Christofori, G., Kohler, H., Wu, F., Fodor, B.D., Frederiksen, M., Willis, S.N., et al. (2021). OBF1 and Oct factors control the germinal center transcriptional program. Blood *137*, 2920–2934. https://doi.org/10.1182/blood.2020010175.

37. Wang, V.E., Tantin, D., Chen, J., and Sharp, P.A. (2004). B cell development and immunoglobulin transcription in Oct-1-deficient mice. Proc. Natl. Acad. Sci. *101*, 2005–2010.

38. Zheng, N.-Y., Wilson, K., Jared, M., and Wilson, P.C. (2005). Intricate targeting of immunoglobulin somatic hypermutation maximizes the efficiency of affinity maturation. J. Exp. Med. *201*, 1467–1478.

39. Grotjohann, T., Testa, I., Reuss, M., Brakemann, T., Eggeling, C., Hell, S.W., and Jakobs, S. (2012). rsEGFP2 enables fast RESOLFT nanoscopy of living cells. Elife *1*, e00248.

40. El Khatib, M., Martins, A., Bourgeois, D., Colletier, J.-P., and Adam, V. (2016). Rational design of ultrastable and reversibly photoswitchable fluorescent proteins for super-resolution imaging of the bacterial periplasm. Sci. Rep. *6*, 18459.

41. Dreyfus, C., Laursen, N.S., Kwaks, T., Zuijdgeest, D., Khayat, R., Ekiert, D.C., Lee, J.H., Metlagel, Z., Bujny, M.V., and Jongeneelen, M. (2012). Highly conserved protective epitopes on influenza B viruses. Science *337*, 1343–1348.

42. Wieczorek, M., Abualrous, E.T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., and Freund, C. (2017). Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. Front. Immunol. *8*, 292.

43. Lang, D., Zhang, S., Ren, P., Liang, F., Sun, Z., Meng, G., Tan, Y., Li, X., Lai, Q., and Han, L. (2020). Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. Gigascience *9*, giaa123.

44. Barrack, K.L., Fyfe, P.K., and Hunter, W.N. (2015). The structure of tubulin-binding cofactor A from Leishmania major infers a mode of association during the early stages of microtubule assembly. Acta Crystallogr. Sect. F Struct. Biol. Commun. *71*, 539–546.

45. Senigl, F., Maman, Y., Dinesh, R.K., Alinikula, J., Seth, R.B., Pecnova, L., Omer, A.D., Rao, S.S., Weisz, D., and Buerstedde, J.-M. (2019). Topologically associated domains delineate susceptibility to somatic hypermutation. Cell Rep. *29*, 3902–3915.

46. Badran, A.H., and Liu, D.R. (2015). Development of potent in vivo mutagenesis plasmids with broad mutational spectra. Nat. Commun. *6*, 8425.

47. Kohler, K.M., McDonald, J.J., Duke, J.L., Arakawa, H., Tan, S., Kleinstein, S.H., Buerstedde, J.-M., and Schatz, D.G. (2012). Identification of core DNA elements that target somatic hypermutation. J. Immunol. *189*, 5314–5326.

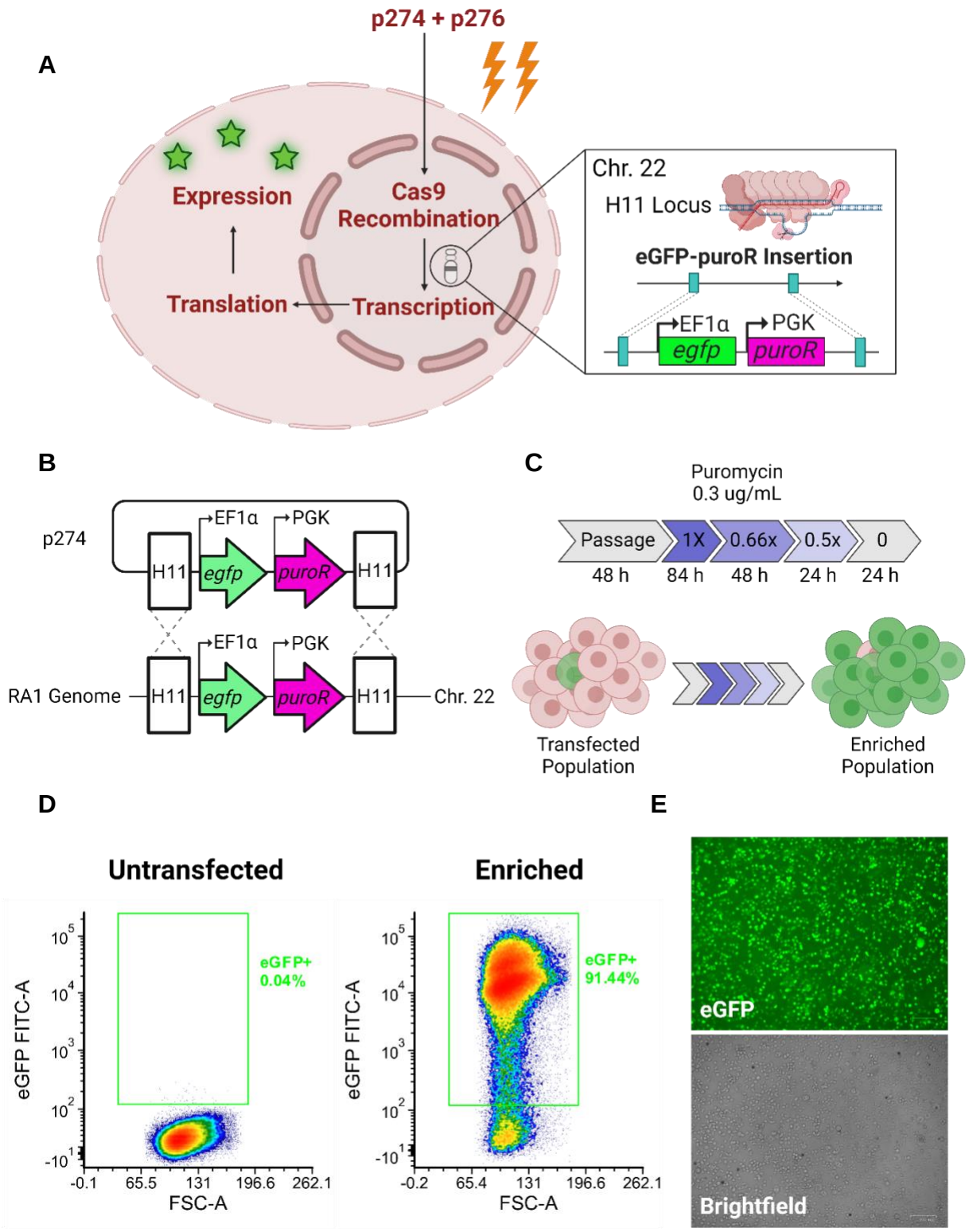


Figure 1: **Streamlining RA 1 transfection and enrichment workflow**. (**A**) Transfection scheme for RA 1 cells using a two-plasmid system. (**B**) Integration scheme of homology

cassette, following DSB formation by RNA guided Cas9. (**C**) Selection scheme for transfected RA 1 cells containing the puromycin resistance cassette, puroR. (**D**) FACS density plot of gated eGFP+ cells in untransfected RA 1 cells and those enriched through puromycin selection; FL1 represents the FITC fluorescence channel. (**E**) Imaging performed on enriched RA 1-274-eGFP cells following puromycin selection. Scale bar indicates 100 µM.

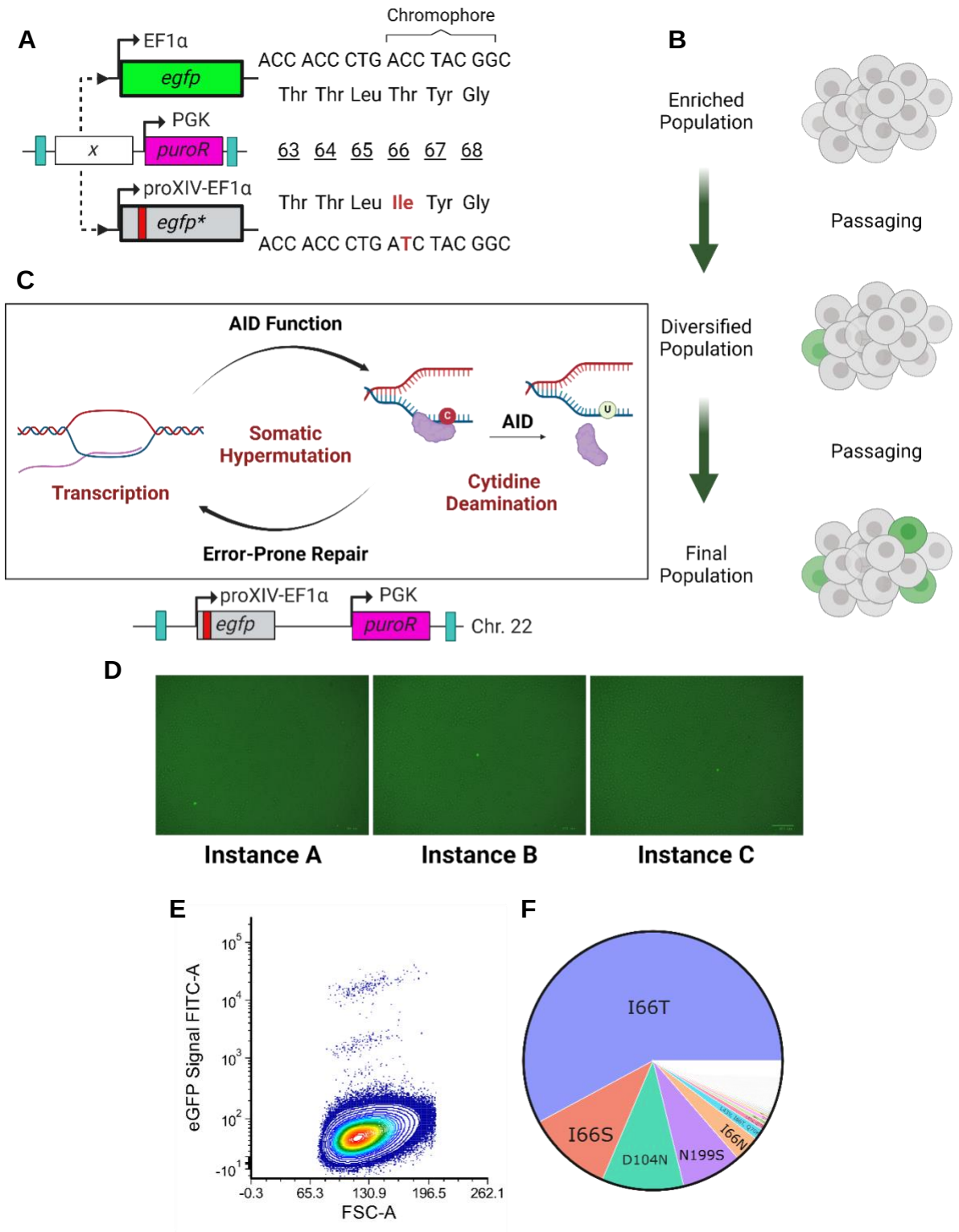

A
EF1α
egfp
Chromophore
ACC ACC CTG ACC TAC GGC
Thr Thr Leu Thr Tyr Gly
PGK
x
puroR
63 64 65 66 67 68
proXIV-EF1α
egfp*
Thr Thr Leu Ile Tyr Gly
ACC ACC CTG ATC TAC GGC
B
Enriched Population
Passaging
Diversified Population
Passaging
Final Population
C
AID Function
Somatic Hypermutation
Transcription
AID
Cytidine Deamination
Error-Prone Repair
proXIV-EF1α
egfp
PGK
puroR
Chr. 22
D
Instance A
Instance B
Instance C
E
eGFP Signal FITC-A
FSC-A
-0.3
65.3
130.9
196.5
262.1
F
I66T
I66S
D104N
N199S
I66N

Figure 2: Creation of eGFP knockout, eGFP*, and subsequent restoration of fluorescence through SHM. (**A**) Scheme for introduction of a C to T mutation resulting in change of the Thr residue in the GFP chromophore responsible for fluorescence generation to Ile. (**B**) Scheme for serial passaging and observation of emerging eGFP+ fluorescence. (**C**) Scheme for SHM function in RA 1-proXIV-eGFP* cells. (**D**) Imaging of eGFP+ cells within a single sample. (**E**) FACS density plot for eGFP+ cells, as sorted from the bulk population of enriched RA 1-274-proXIV-eGFP* sample. (**F**) Sunburst plot from NGS analysis showing the emergence of high frequency variants.

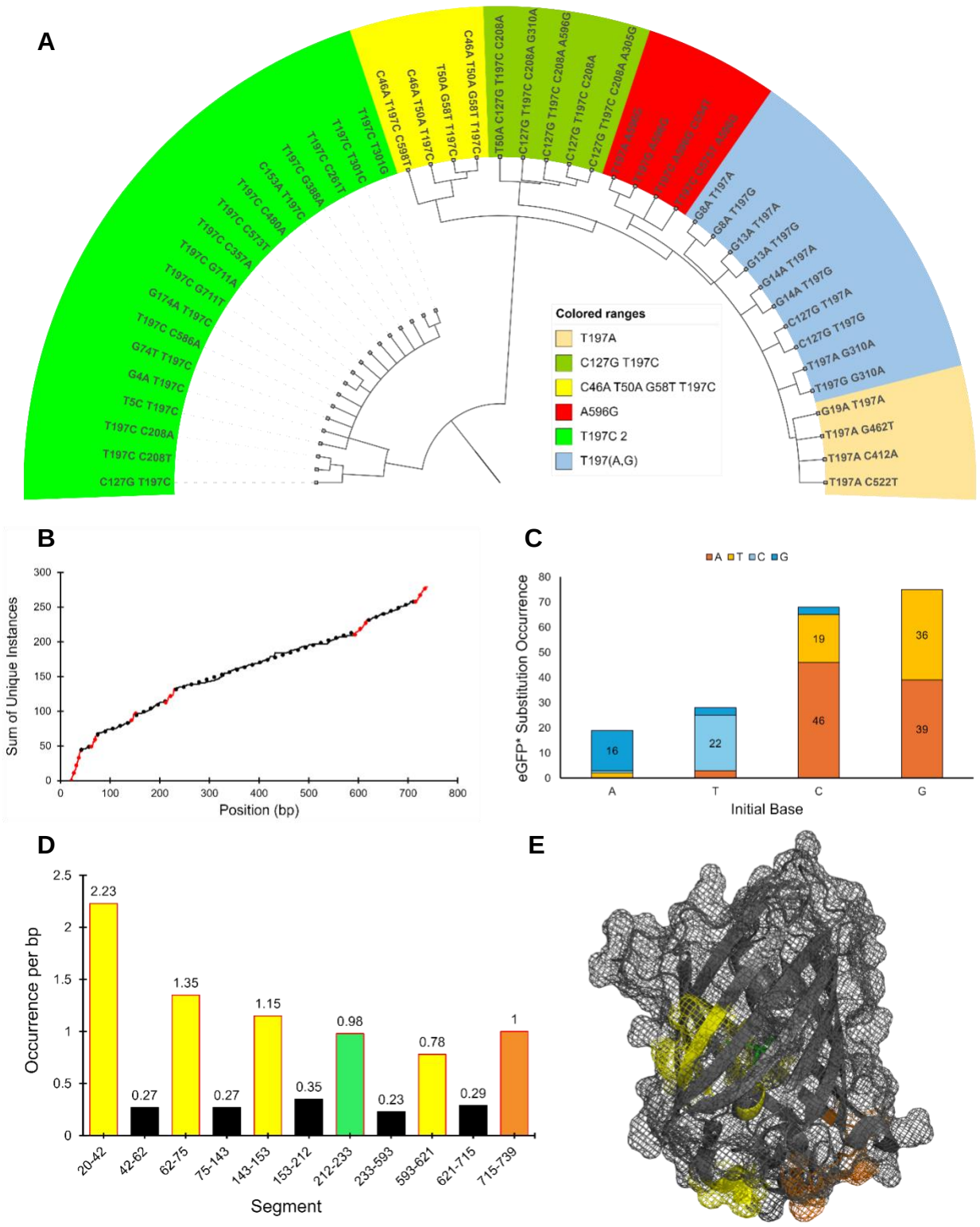

A
Colored ranges
T197A
C127G T197C
C46A T50A G58T T197C
A596G
T197C 2
T197(A,G)
B
Sum of Unique Instances
Position (bp)
C
A T C G
eGFP* Substitution Occurrence
Initial Base
D
Occurrence per bp
Segment
E

Figure 3. The mutational landscape of the eGFP* library. (A) The phylogenetic tree of various eGFP* lineages. (**B**) Integral plot of occurrences over the sequence length of eGFP*. (**C**) Bar chart depicting the mutational bias within eGFP*. (**D**) Bar chart of slopes retrieved from each segment of the integral plot, representing mutational hotspots within eGFP*. (**E**) Crystal structure of eGFP (PDB: 2Y0G), overlayed with color codes corresponding to the hotspot bar chart.

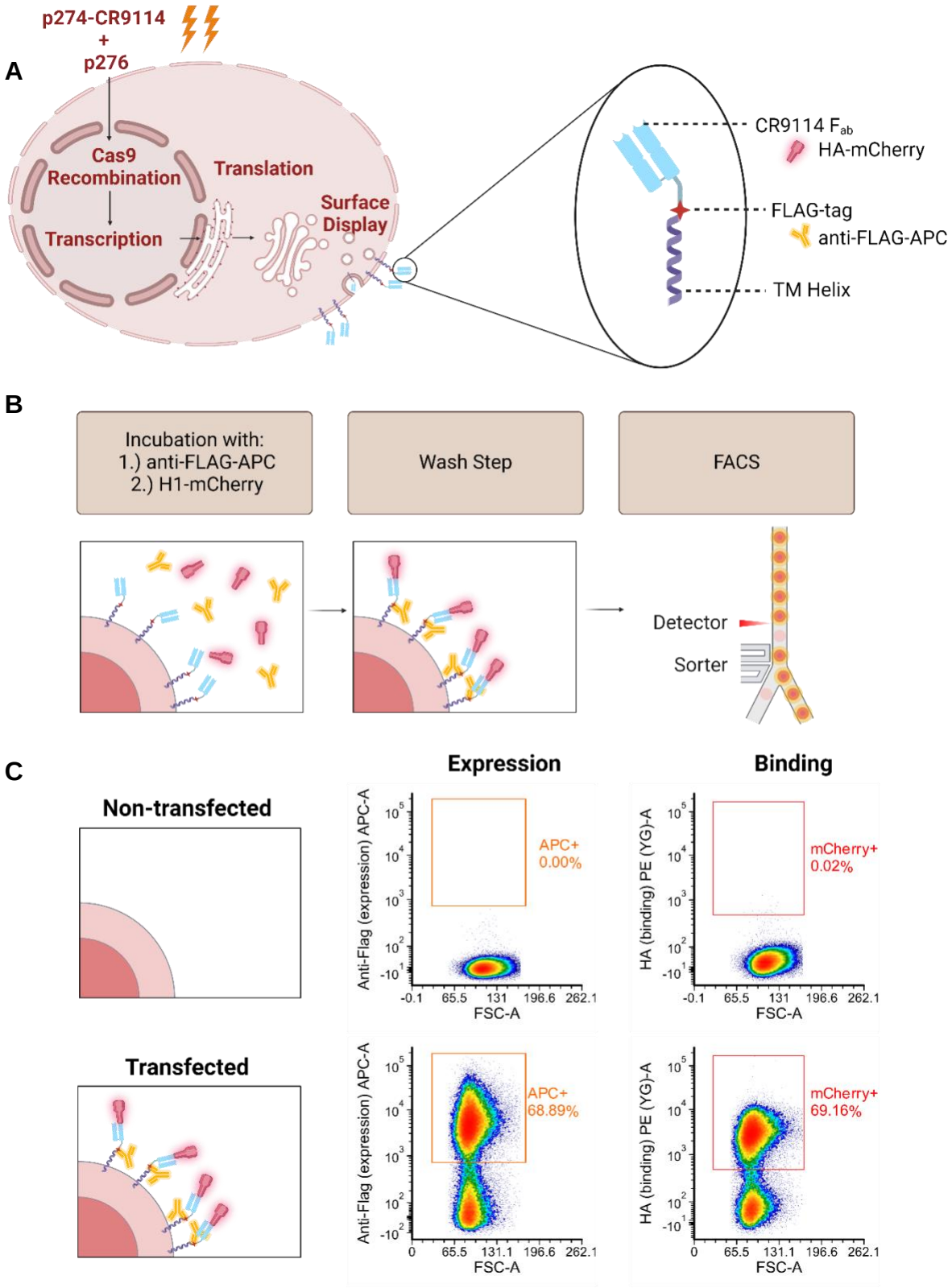

A
p274-CR9114
+
p276
Cas9
Recombination
Translation
Transcription
Surface
Display
CR9114 Fab
HA-mCherry
FLAG-tag
anti-FLAG-APC
TM Helix
B
Incubation with:
1.) anti-FLAG-APC
2.) H1-mCherry
Wash Step
FACS
Detector
Sorter
C
Expression
Binding
Non-transfected
Transfected
Anti-Flag (expression) APC-A
HA (binding) PE (YG)-A
FSC-A
APC+
0.00%
mCherry+
0.02%
APC+
68.89%
mCherry+
69.16%

Figure 4. Chimeric protein design of F_{ab} display construct and function in vivo. (**A**) Scheme of F_{ab} display on the surface of RA 1 cells, utilizing fluorescent HA-mCherry and fluorescent anti-FLAG-APC mAbs to parameterize the chimeric protein into separate binding and expression variables. (**B**) Scheme of FACS experiment. (**C**) Density FACS plots of enriched RA 1-CR9114 cells stained with anti-FLAG-APC and H1-mCherry.

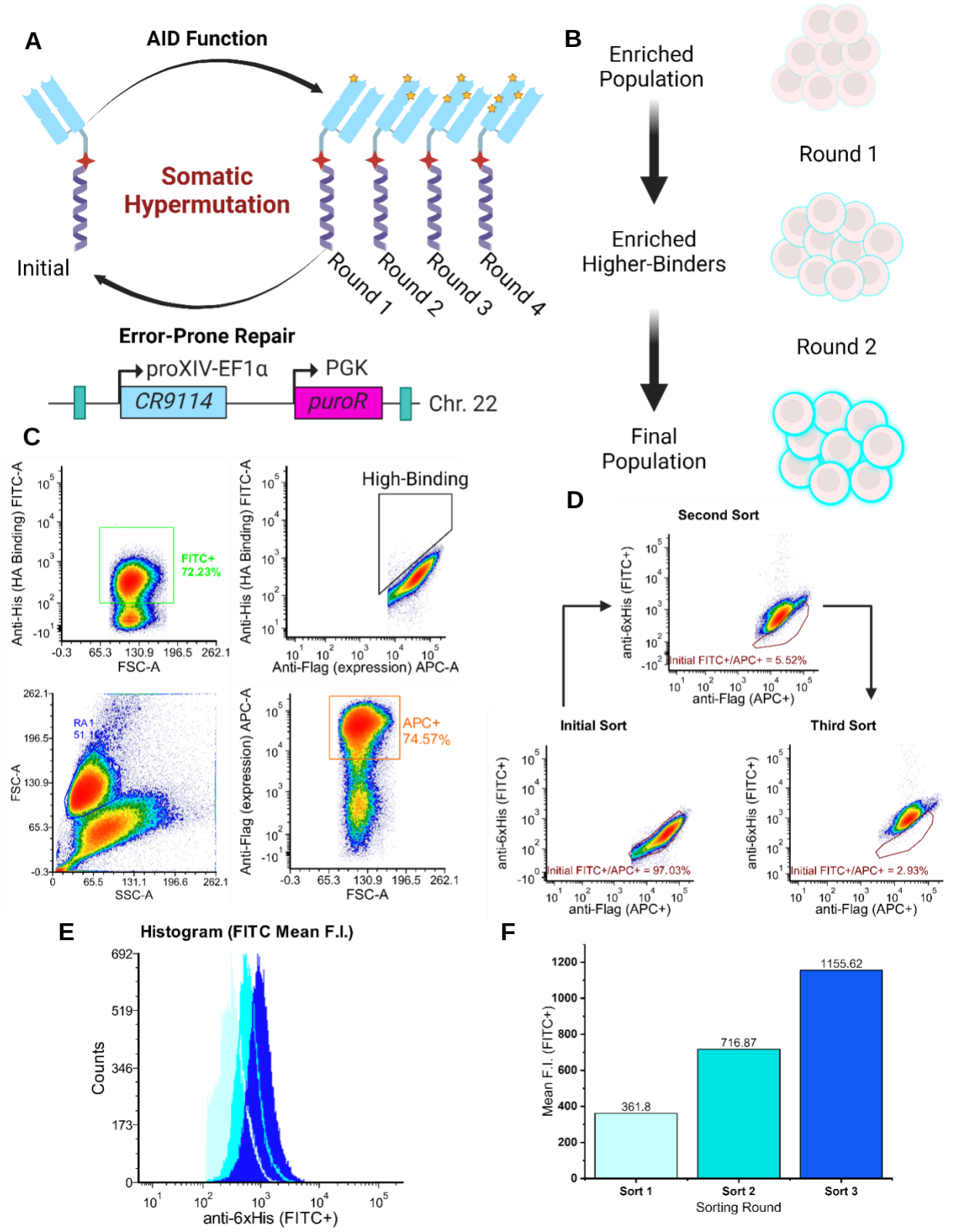
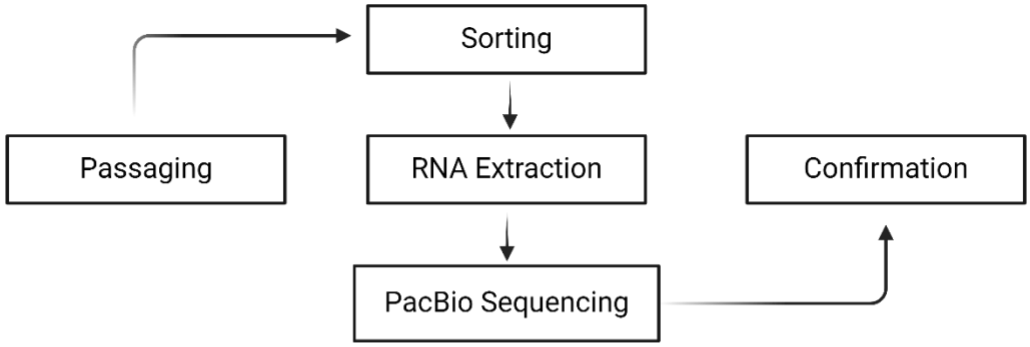

A
AID Function
Somatic Hypermutation
Initial
Round 1
Round 2
Round 3
Round 4
Error-Prone Repair
proXIV-EF1α
PGK
CR9114
puroR
Chr. 22
B
Enriched Population
Round 1
Enriched Higher-Binders
Round 2
Final Population
C
Anti-His (HA Binding) FITC-A
FSC-A
FITC+ 72.23%
High-Binding
Anti-Flag (expression) APC-A
SSC-A
APC+ 74.57%
D
Second Sort
anti-6xHis (FITC+)
anti-Flag (APC+)
Initial FITC+/APC+ = 5.52%
Initial Sort
Initial FITC+/APC+ = 97.03%
Third Sort
Initial FITC+/APC+ = 2.93%
E
Histogram (FITC Mean F.I.)
Counts
anti-6xHis (FITC+)
F
Mean F.I. (FITC+)
361.8
716.87
1155.62
Sort 1
Sort 2
Sort 3
Sorting Round

Figure 5. Workflow for the continuous directed evolution of higher-binding CR9114 variants. (**A**) Schematic for the incorporation of mutations into the gene encoding CR9114 F_{ab} by iterative cycles of SHM**.** (**B**) Scheme for the iterative cycle of enriching cells binding to H5-mCherry, followed by diversification of the CR9114 gene through SHM that occurs during transcription, followed by selection of higher-binding variants and subsequent enrichment. (**C**) FACS density plots of the RA 1 gating strategy (bottom left) for expression (APC+, bottom right) and binding to H5 (FITC+, top left) channels, yielding a linear relationship from which higher-binding variants may be gated and sorted for growth (top right). (**D**) FACS density plots for iterative sorting and passaging (expression vs binding). For this verification experiment His-tagged H5 was used and FITC-conjugated anti-His antibody was used to detect binding of cells to His-tagged H5. (**E**) Histogram representing each iterative sort in terms of H5-binding. (**F**) Bar chart representing mean F.I. of the histogram.

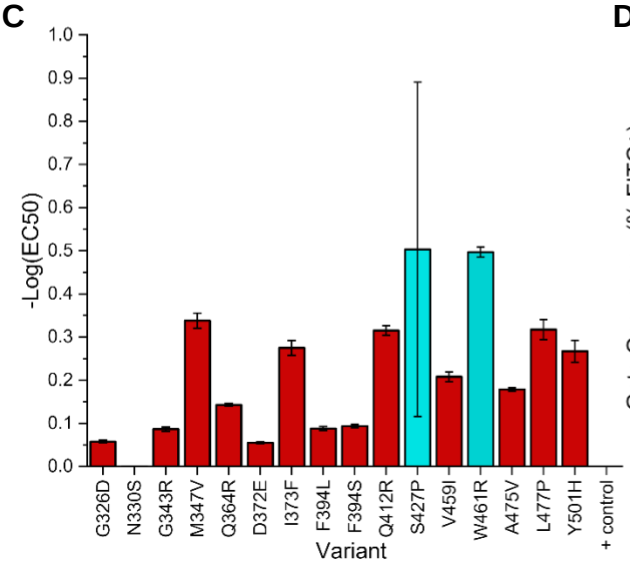
A



B

A.A. Substitution	Read Freq. (%)	Initial Codon	Subst. Codon	EC50 (nM)	Region	Score
G326D	0.0036	GGC	GAC	11.74	Hv	mut202
N330S	0.0053	AAC	AGC	13.38	Hv	mut121
G343R	0.0265	GGT	CGT	10.98	Hv	mut21
M347V	0.0271	ATG	GTG	6.16	Hv	mut20
Q364R	0.0115	CAG	CGG	9.65	Hv	mut42
D372E	0.0042	GAT	GAG	11.82	Hv	mut170
I373F	0.0235	ATT	TTT	7.12	Hv	mut26
F394L	0.0575	TTT	CTT	10.95	Hv	mut10
F394S	0.0174	TTT	TCT	10.8	Hv	mut29
Q412R	0.0102	CAG	CGG	6.49	Hv	mut52
S427P	0.0070	TCC	CCC	4.21	Hc	mut78
V459I	0.0098	GTA	ATA	8.31	Hc	mut54
W461R	0.0125	TGG	CGG	4.27	Hc	mut36
A475V	0.0134	GCG	GTG	8.89	Hc	mut32
L477P	0.0074	CTC	CCC	6.46	Hc	mut74
Y501H	0.0089	TAT	CAT	7.25	Hc	mut61
wild-type				13.41		

C



D

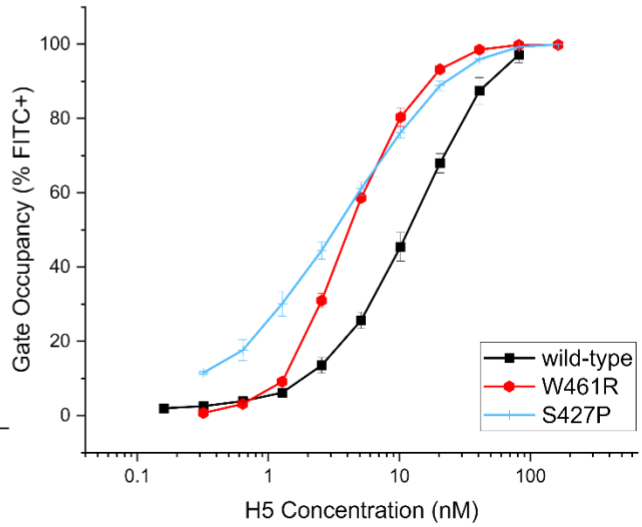


Figure 6. Isolation of enriched CR9114 variants obtained by SHM and their experimental validation. (**A**) Experimental scheme of the characterizing higher-binding variants. (**B**) Table of read frequencies for the tested variants alongside their substituted codons, encoded amino acids, and EC50 values; variants were named based on the amino acid substitution found in the heavy chain of CR9114 and numbered based on the read frequency of their unique nucleotide substitution in the sequencing data - lower numbers indicate a higher degree enrichment. Point mutations listed for the F_{ab} fragment also include at least one additional mutation, specifically A1601T and/or A1149G. (**C**) Bar chart representing the EC50 values of the 16 CR9114 heavy chain variants. (**D**) Titration curve for the two high-affinity CR9114 variants, W461R and S427P, relative to the CR9114 wild-type control.