**Indels allow antiviral proteins to evolve functional novelty inaccessible by missense mutations**

<https://doi.org/10.5061/dryad.6q573n66b>

We sought to understand the evolutionary paths by which TRIM5α might evolve *de novo* antiviral function against an SIV endemic to sabaeus monkeys (SIVsab), which is a particular difficult evolutionary challenge for TRIM5α. Human TRIM5α has no antiviral activity against this lentivirus, while the rhesus macaque TRIM5α potently inhibits SIVsab. To understand the evolutionary potentials and constraints for this difficult antiviral function, we generated libraries of variants within the lentiviral-specificity-determining v1 loop of full-length TRIM5α. These include:

* Deep mutational scanning (DMS) libraries of all possible single-missense variants within the v1 loop (residues 330-340 of human TRIM5α, “HsDMS” or “HD”; 332-344 of rhesus TRIM5α, “RhDMS” or “RD”);
* Combinatorial libraries sampling either the human or rhesus TRIM5α residue at each of 9 v1 positions where they differ, in either the human (“H2R”) or rhesus (“R2H”) backbone;
* and a deep indel scanning (DIS) library of human TRIM5α (“HsIndel” or “HI”), which simulates DNA slippage errors, sampling in-frame deletions or sequence duplications ranging from 1-9 amino acids at each position in the v1 loop (residues 324-345).

Variant libraries were stably expressed in CRFK cells by low-dose (MOI ~ 0.25) retroviral transduction and selection. Libraries were then infected with SIVsab-GFP and sorted for either GFP-negative cells (human TRIM5α libraries, gain-of-function variants) or GFP-positive cells (rhesus TRIM5α libraries, loss-of-function variants). Finally, we generated Illumina libraries of the v1 loop from each of these experiments. Each experiment included 5 samples:

* the plasmid library, before transduction (“plasmid”)
* a pre-sorted pool (“InA”, “InB”)
* a post-sorted pool (“negA”, “negB” for GFP-negative sorts; “posA”, “posB” for GFP-positive sorts”)

Illumina libraries were barcoded (for each of the 5 samples from each experiment) and appended with Nextera flow cell adapters. Libraries were sequenced with standard Nextera sequencing primers with paired-end 150 base-pair reads.

**Description of the data and file structure**

Uploaded datasets include the raw Illumina reads (forward and reverse reads), pre-processing scripts (.sh files), pre-processed data (.fasta files), R scripts we used to process data, and fully processed data files (.csv files).

* Raw Illumina sequencing reads from each library are included for each experiment as .fq.gz files within the “IlluminaRawData” folder.
* Scripts for pre-processing data (trimming and quality filtering) for each experiment can be found in the “RawDataPreProcessing/Scripts” folder.
* Pre-processed data (i.e. trimmed, quality filtered, and collapsed for unique sequences) can be found as “*Sample\_*collapsed.fasta” files in the relevant experimental folder (“HsDMS”, “RhDMS”,“H2R2Hcombi”, and “HsIndel”), under the subfolder “Data”.
* Pre-processed data were further analyzed using R scripts, as described in the Code/Software section below. All data processing scripts are included in the “Scripts” subfolder for each relevant experiment folder. Output files from processing scripts can be found in the “Analysis” subfolder for each relevant experimental folder.

**Sharing/Access information**

Data can also be found at <https://github.com/jtenthor/T5libraries_SIVsab/tree/main>. Note that the Github repository does not include raw Illumina reads, as these files were too large to upload.

**Code/Software**

Raw data were pre-processed as follows:

* Invariant sequences (TRIM5α homology for PCR amplification; flow cell adapters and barcodes) were trimmed from the 5’ and 3’ end using CutAdapt.
* Paired-end reads were merged to generate a consensus read using NGmerge.
* Low-quality consensus reads were filtered for a minimum score of Q25 at each nucleotide using FASTX-toolkit.
* Remaining reads were collapsed to unique sequences and counts per sequence.

All pre-processing was done using Unix shell executable files (.sh) on Mac Terminal. Scripts for pre-processing data can be found in the “RawDataPreProcessing/Scripts” folder. Published and freely available code necessary to execute these files includes:

* CutAdapt (doi:<https://doi.org/10.14806/ej.17.1.200>)
* NGmerge (DOI: [10.1186/s12859-018-2579-2](https://doi.org/10.1186/s12859-018-2579-2) ). We note that NGmerge, as available, did not accept our current Illumina data quality scores – we needed to provide an updated “qual\_profile\_41.txt” file
* FASTX-toolkit (<http://hannonlab.cshl.edu/fastx_toolkit/>)

Pre-processed data were further processed in R Studio (version 2023.12.1+402). All data processing scripts are included in the “Scripts” subfolder for each relevant experiment folder. Loaded packages necessary to run these scripts include:

* Biostrings, v2.72.0 (DOI: [10.18129/B9.bioc.Biostrings](https://doi.org/doi:10.18129/B9.bioc.Biostrings) )
* Tidyverse, v2.0.0 (<https://www.tidyverse.org/packages/>)
* SeqinR, v4.2-36 (<https://seqinr.r-forge.r-project.org/>)

DMS data were analyzed as follows:

1. Fasta files were merged based on sequence into a single table containing the number of read counts for each unique sequence detected in all 4 samples.
2. Sequences were translated and identified for mutated codon(s) relative to wild-type TRIM5α.
3. Sequences were filtered for:
   1. expected length (33 nucleotides for human TRIM5α; 39 nucleotides for rhesus TRIM5α)
   2. ≤ 1 mutated codon
   3. codon ending in a C or G at the 3rd position (as libraries were created using NNS codons, where N = any nucleotide and S = C or G)
4. A pseudocount of 1 read was added to each sequence in each sample to avoid dividing by (or returning) 0.
5. Read counts for each read in each sample were normalized to the total number of reads in each sample (e.g. “InA”), expressed as counts per million (cpm).
6. Sequences with < 10 cpm in either input library were removed for having too much noise in subsequent analysis.
7. Enrichment was calculated for each sequence by dividing sorted cpm by input cpm for each replicate.
8. All synonymously-coding variants were averaged to yield an enrichment score for each protein sequence for each replicate. The standard deviation of enrichment was calculated across synonymous variants and replicates. The only exception was for wild-type-synonymous variants, which were plotted separately to give a visual representation of the standard deviation in each replicate. Fully wild-type nucleotide sequence was removed from further analysis because (1) it was sometimes a strong outlier relative to all other wild-type variants, and (2) this effect could have been due to plasmid contamination of samples with wild-type TRIM5α.

All other libraries were processed in a similar manner, except that instead of filtering as in step 3 above, fasta files were mapped to a table of expected sequences based on library construction, and only sequences matching these exact nucleotide sequences were retained. The expected sequence files can be found as csv files in the “Data” subfolder for each relevant experimental folder. Typically, >90% of reads matched these expected sequences. However, for the combinatorial library converting human into rhesus TRIM5α, we noticed a large fraction (~25%) of reads being lost at this step. Subsequent analysis identified two frameshift mutations, which were then included (as combinatorial variants) in the expected sequence list “CombiSeqListActual”. After this correction, >90% of reads matched expected sequences. Fully wild-type sequence was retained for this library, as we failed to include wild-type sequences in our library design; however, it was removed from analysis of the indel library, which was designed with 7 wild-type-synonymous variants.

Gain-of-function cut-offs were set based on either premature stop codons (for DMS or indel scanning libraries; > [mean + 2 SD] in each replicate) or frameshift variants (for combinatorial human-to-rhesus library; > max score in each replicate). This choice was both more conservative for detecting any antiviral function, and it also frequently included more controls to give a better estimate of assay noise. Loss-of-function cut-offs were based on wild- type rhesus TRIM5α variants (> [mean + 2 SD] in each replicate).