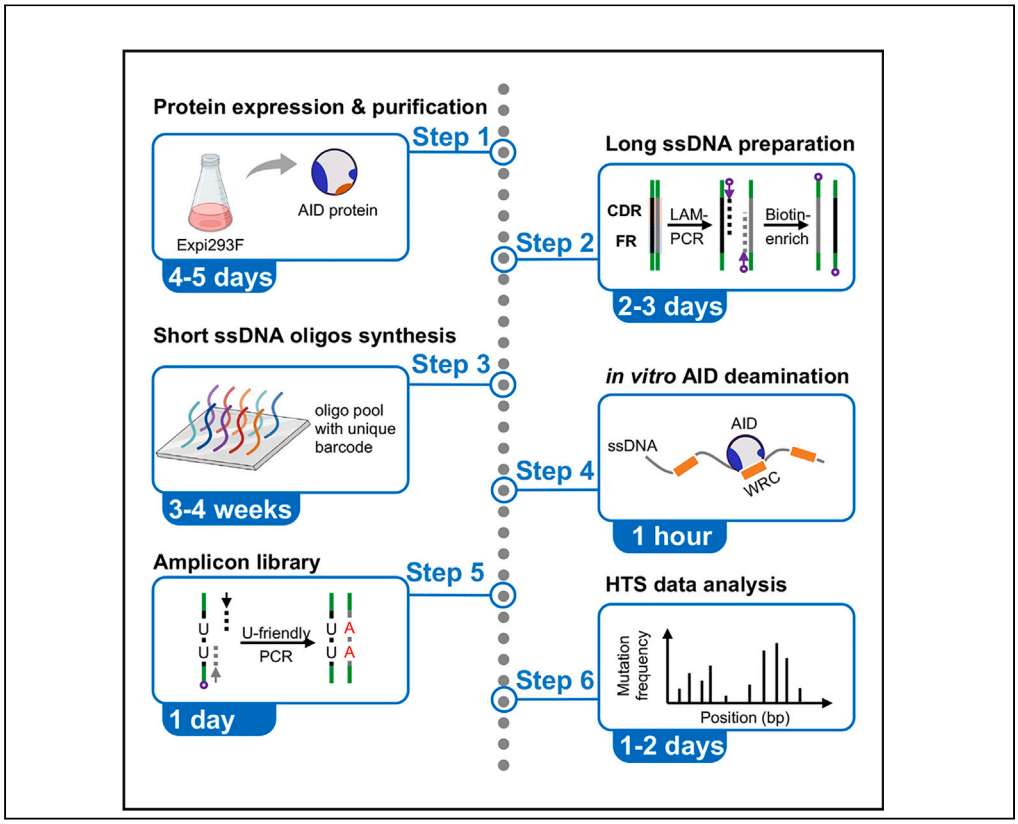
STAR Protocols

CellPress
OPEN ACCESS

Protocol

A high-throughput protocol for deamination of long single-stranded DNA and oligo pools containing complex sequences



Cytidine deaminases as DNA mutators play important roles in immunity and genome stability. Here, we present a high-throughput protocol for deamination of long single-stranded (ss) DNA or oligo pools containing complex sequences. We describe steps for the preparation of both enzyme (activation-induced deaminase, AID) and ssDNA substrates, the deamination reaction, uracil-friendly amplification, and data analysis. This assay can be used to determine the intrinsic mutation profile of a single antibody gene or a pool of selected regions on genomic DNA.

Yanyan Wang, Senxin Zhang, Xiaoqi Zheng, Leng-Siew Yeap, Fei-Long Meng

xqzheng@shsmu.edu.cn (X.Z.)
yeaplengsiew@shsmu.edu.cn (L.-S.Y.)
feilong.meng@sibcb.ac.cn (F.-L.M.)

Highlights

Stepwise high-throughput assay for AID/APOBEC cytidine deaminase activity

Reveal antibody variable exon mutation profiles *in vitro*

Intrinsic mutation profiling of up to 600K nt DNA in one reaction

Simplified optimization and evaluation of the antibody sequence

Wang et al., STAR Protocols 4, 102602
December 15, 2023
https://doi.org/10.1016/j.xpro.2023.102602

STAR Protocols

CellPress
OPEN ACCESS

Protocol

A high-throughput protocol for deamination of long single-stranded DNA and oligo pools containing complex sequences

Yanyan Wang,[1,6,7] Senxin Zhang,[2,6] Xiaoqi Zheng,[3,*] Leng-Siew Yeap,[1,4,*] and Fei-Long Meng[5,8,*]

[1]Shanghai Institute of Immunology, State Key Laboratory of Oncogenes and Related Genes, Department of Immunology and Microbiology, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

[2]Department of Mathematics, Shanghai Normal University, Shanghai 200234, China

[3]Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[4]Center for Immune-Related Diseases at Shanghai Institute of Immunology, Department of Endocrinology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

[5]State Key Laboratory of Molecular Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200031, China

[6]These authors contributed equally

[7]Technical contact

[8]Lead contact

*Correspondence: wangyanyan2019@sibcb.ac.cn (Y.W.), xqzheng@shsmu.edu.cn (X.Z.), yeaplengsiew@shsmu.edu.cn (L.-S.Y.), feilong.meng@sibcb.ac.cn (F.-L.M.)
https://doi.org/10.1016/j.xpro.2023.102602

SUMMARY

Cytidine deaminases as DNA mutators play important roles in immunity and genome stability. Here, we present a high-throughput protocol for deamination of long single-stranded (ss) DNA or oligo pools containing complex sequences. We describe steps for the preparation of both enzyme (activation-induced deaminase, AID) and ssDNA substrates, the deamination reaction, uracil-friendly amplification, and data analysis. This assay can be used to determine the intrinsic mutation profile of a single antibody gene or a pool of selected regions on genomic DNA.

For complete details on the use and execution of this protocol, please refer to Wang et al. (2023).[1]

BEFORE YOU BEGIN

AID is a cytidine deaminase that acts on the ssDNA substrate.[2–7] Following the discovery of AID, *in vitro* biochemical experiments have greatly advanced the mechanistic study of antibody diversification. Although previous studies have described the *in vitro* deamination assay of AID/APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide) enzymes,[3,5,6,8] none have determined the high-throughput cytidine deaminase activity on long ssDNA (~200 nt) or oligo pools (50–100 nt) substrates in an unbiased manner. Therefore, we first verified that the *in vitro* deamination profile of AID on a single long ssDNA substrate strongly correlates with the intrinsic SHM profile *in vivo*.[1] The expansion of substrates into oligo pools containing complex sequences allows the high-throughput assay of AID-initiated mutations in numerous antibody sequences or selected regions on genomic DNA. This assay allows rapid and sensitive evaluation of antibody sequence mutation profiles *in vitro*, which is useful for antibody sequence optimization, antibody sequence evolution studies, and validation of potential AID-initiated mutations in cancer genomes.

Prepare buffers and plasmids

Timing: ~2 days

1. Prepare sufficient buffers as indicated.
2. Prepare the indicated plasmid using a standard Maxiprep protocol and filter through a 0.22 μm filter prior to use.

Prepare Expi293F cells

Timing: ~5 days

3. Thaw Expi293F cells and culture the cells in serum-free medium.
   a. Quickly thaw the Expi293F cells in a 37°C water bath.
   b. Add 9 mL of serum-free medium to 1 mL of cells (3×10^7 cells/mL).
   c. Centrifuge the cells at 400 × *g* for 5 min at 4°C to pellet the cell debris.
   d. Discard the supernatant and resuspend the cells in 10 mL of fresh serum-free medium.
   e. Add the 10 mL of cells to a 125 mL sterile flask.
   f. Add a further 20 mL of fresh serum-free medium to the flask.
   g. Mix well and culture the cells on an orbital shaker incubator at 37°C, 120 rpm, 5% CO_2.

⚠ CRITICAL: Cells are counted daily and should be maintained at 0.5–2.5 × 10^6 cells/mL. Although regular HEK293 cells can be used, we recommend the use of Expi293F cells because they provide consistently high protein yields. The culture volume should not exceed 25% of the maximum flask volume.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Escherichia coli DH5α Competent Cells	TIANGEN	Cat# CB101
Escherichia coli Stbl3 Competent Cells	Exonbio	Cat# CC104-01
Chemicals, peptides, and recombinant proteins		
MBP-hAID	Wang et al.[1]	https://doi.org/10.1016/j.cell.2023.03.030
HEPES	Sangon Biotech	Cat# A600264
NaOH	Sangon Biotech	Cat# A100583
NaCl	Sangon Biotech	Cat# A610476
$MgCl_2 \cdot 6H_2O$	Sangon Biotech	Cat# A610328
$EDTA \cdot Na_2 \cdot 2H_2O$	Sangon Biotech	Cat# A610185
Tris base	Sangon Biotech	Cat# A600194
HCl	HUSHI	Cat# 10011018
PMSF	Sangon Biotech	Cat# A610425
Isopropanol	Sangon Biotech	Cat# A600918
DTT	Sangon Biotech	Cat# A620058
Glycerol	Sangon Biotech	Cat# A600232
Maltose monohydrate	Amresco	Cat# 1B1184
Xylene Cyanol FF	Sangon Biotech	Cat# A630005
Bromophenol blue	Sangon Biotech	Cat# A602230
Acetic acid	Sangon Biotech	Cat# A501931
Union-293 medium	Union-Biotech	Cat# UP0050
Conical cell culture flask-125 mL	Thermo Fisher Scientific	Cat# 4115-0125
Conical cell culture flask-250 mL	Thermo Fisher Scientific	Cat# 4115-0250

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Conical cell culture flask-1000 mL	Thermo Fisher Scientific	Cat# 4115-1000
PEI-MAX-Transfection Grade Linear Polysciences Polyethylenimine Hydrochloride (MW 40,000)	Polysciences	Cat# 24765-1
Amylose resin	NEB	Cat# E8021S
Bradford reagent	Sigma-Aldrich	Cat# B6916
DNase I, RNase-free	Sangon Biotech	Cat# A610099
Agarose	BIOWEST	Cat# BY-R0100
dNTPs	TIANGEN	Cat# CD111-13
100 bp DNA Ladder	TIANGEN	Cat# MD109-02
1 kb DNA Ladder	TIANGEN	Cat# MD111-02
6× Protein Loading Buffer	TransGen	Cat# DL101-02
PageRuler Prestained Protein Ladder	Thermo Fisher Scientific	Cat# 26616
Critical commercial assays		
NucleoBond Xtra Midi	MACHEREY-NAGEL	Cat# 740410.50
SDS-PAGE Preparation kit	Sangon Biotech	Cat# C631100
TransStart FastPfu DNA Polymerase	TransGen Biotech	Cat# AP221-01
PfuTurbo Cx Hot Start high-fidelity DNA Polymerase	Agilent	Cat# 600410
Q5U Hot Start High-Fidelity DNA Polymerase	NEB	Cat# M0515
Phusion High-Fidelity DNA polymerase	Thermo Fisher Scientific	Cat# F530N
Universal DNA Purification Kit	TIANGEN	Cat# DP214-03
Qubit dsDNA HS Assay Kit	Life Technologies	Cat# Q32851
ClonExpress II One Step Cloning Kit	Vazyme	Cat# C112
Dynabead MyOne Streptavidin C1	Invitrogen	Cat# 65001
Dynabeads MyOne Silane	Invitrogen	Cat# 37002D
Deposited data		
SHM-amplicon-seq	Wang et al.[1]	PRJNA918596
Experimental models: Cell lines		
Expi293F cells	Thermo Fisher Scientific	Cat# A14527
Oligonucleotides		
Synthesized ssDNA oligos pool	Wang et al.[1]	https://doi.org/10.1016/j.cell.2023.03.030
ssDNA generated by LAM-PCR	Wang et al.[1]	https://doi.org/10.1016/j.cell.2023.03.030
Recombinant DNA		
Plasmid: pcDNA3.4-MBP-hAID	Xie et al.[9]	https://doi.org/10.15252/embj.2021109324
Software and algorithms		
Cutadapt	Martin	https://github.com/marcelm/cutadapt
Bowtie2	Langmead and Salzberg	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Deeptools	Ramirez et al.[10]	https://github.com/deeptools/deepTools
Samtools	Li et al.[11]	http://samtools.sourceforge.net/
Fastq-multx	Aronesty	https://github.com/brwnj/fastq-multx
Fastp	Chen et al.[12]	https://github.com/OpenGene/fastp
R Studio – Open source edition	Rstudio	https://www.rstudio.com/
Prism version 9.0	GraphPad Software	https://www.graphpad.com/scientific-software/prism/
SHM pipeline	Yeap et al.[13]	https://doi.org/10.1016/j.cell.2015.10.042
Other		
CO_2 resistant shaker	Thermo Fisher Scientific	Cat# 88881102
High-pressure crusher	Union- Biotech	Cat# UH-03
TGem Plus Spectrophotometer	TIANGEN	Cat# OSE-260-01
Qubit 3.0 Fluorometer	Invitrogen	Cat#Q33216
DynaMag™-PCR Magnet	Thermo Fisher Scientific	Cat# 492025
PCR thermal cycler	Eppendorf	Cat# EP6331000025
Syringe-driven filters (0.22 μm)	BIOFIL	Cat# FPE204030
Gravity chromatography columns	YEASEN	Cat# 20523ES08
Centrifuge 5424	Eppendorf	N/A

(*Continued on next page*)

Continued		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
Centrifuge 5424R	Eppendorf	N/A
Sorvall LYNX 6000 Superspeed Centrifuge	Thermo Fisher Scientific	Cat# 75006590

MATERIALS AND EQUIPMENT

0.5 M HEPES pH 7.4

Reagent	Final concentration	Amount
HEPES	0.5 M	119.15 g
NaOH (2.5 M)	N/A	Until pH 7.4
ddH_2O	-	To 1 L
Total	**N/A**	**1 L**

Autoclave and store at 20°C–25°C for up to 1 year.

4 M NaCl

Reagent	Final concentration	Amount
NaCl	4 M	234 g
ddH_2O	-	To 1 L
Total	**N/A**	**1 L**

Autoclave and store at 20°C–25°C for up to 1 year.

2 M $MgCl_2$

Reagent	Final concentration	Amount
$MgCl_2 \cdot 6H_2O$	2 M	406.6 g
ddH_2O	-	To 1 L
Total	**N/A**	**1 L**

Autoclave and store at 20°C–25°C for up to 1 year.

0.5 M Tris-HCl pH 7.4

Reagent	Final concentration	Amount
Tris base	0.5 M	60.57 g
HCl	N/A	Until pH 7.4
ddH_2O	-	To 1 L
Total	**N/A**	**1 L**

Autoclave and store at 20°C–25°C for up to 1 year.

2.5 M NaOH

Reagent	Final concentration	Amount
NaOH	2.5 M	10 g
ddH_2O	-	To 100 mL
Total	**N/A**	**100 m L**

Store at 20°C–25°C in plastic containers instead of glass containers for up to 1 month.

0.5 M EDTA pH 8.0		
Reagent	Final concentration	Amount
EDTA·Na_2·$2H_2O$	0.5 M	186.12 g
NaOH (2.5 M)	N/A	Until pH 8.0
ddH_2O	-	To 1 L
Total	**N/A**	**1 L**
Autoclave and store at 20°C–25°C for up to 1 year.		

100 mM PMSF (Phenylmethylsulfonyl fluoride)		
Reagent	Final concentration	Amount
PMSF	100 mM	174 mg
Isopropanol	N/A	To 10 mL
Total	**N/A**	**10 mL**
Aliquot to 1 mL per tube and store at −20°C for up to 1 year.		

1 M DTT (Dithiothreitol)		
Reagent	Final concentration	Amount
DTT	1 M	1.55 g
ddH_2O	N/A	To 10 mL
Total	**N/A**	**10 mL**
Aliquot into 1 mL per tube and store at −20°C for up to 1 year.		

Lysis buffer		
Reagent	Final concentration	Amount
HEPES pH 7.4 (0.5 M)	20 mM	20 mL
NaCl (4 M)	150 mM	18.75 mL
Glycerol	10%	50 mL
$MgCl_2$ (2 M)	5 mM	1.25 mL
ddH_2O	-	To 500 mL
Total	**N/A**	**500 mL**
Autoclave or filter through a 0.22 μm filter, store at 4°C for up to 2 weeks. Prior to use, add 10 μL PMSF (100 mM), 1 μL DNase I (150 KU) and 1 μL DTT (1 M) per 1 mL buffer.		

Protein purification wash buffer		
Reagent	Final concentration	Amount
HEPES pH 7.4 (0.5 M)	20 mM	20 mL
NaCl (4 M)	150 mM	18.75 mL
Glycerol	10%	50 mL
ddH_2O	-	To 500 mL
Total	**N/A**	**500 mL**
Autoclave or filter through a 0.22 μm filter, store at 4°C for up to 2 weeks. Prior to use, add 1 μL DTT (1 M) per 1 mL buffer.		

Protein elute buffer		
Reagent	Final concentration	Amount
HEPES pH 7.4 (0.5 M)	20 mM	2 mL
NaCl (4 M)	150 mM	1.875 mL
Glycerol	10%	5 mL
Maltose monohydrate	10 mM	180 mg
ddH_2O	-	To 50 mL
Total	**N/A**	**50 mL**
Autoclave or filter through a 0.22 μm filter, store at 4°C for up to 2 weeks. Prior to use, add 1 μL DTT (1 M) per 1 mL buffer.		

6× DNA loading dye		
Reagent	Final concentration	Amount
EDTA pH 8.0 (0.5 M)	30 mM	6 mL
Glycerol	36%	36 mL
Xylene Cyanol FF	0.05%	50 mg
Bromophenol Blue	0.05%	50 mg
ddH_2O	-	To 100 mL
Total	**N/A**	**100 mL**
Store at 20°C–25°C for up to 1 year.		

50× TAE buffer		
Reagent	Final concentration	Amount
Tris base	2 M	242.28 g
$EDTA{\cdot}Na_2{\cdot}2H_2O$	50 mM	18.61 g
Acetic acid	5.71%	57.1 mL
ddH_2O	-	To 1 L
Total	**N/A**	**1 L**
Store at 20°C–25°C for up to 1 year.		

2× Binding and Washing (B&W) buffer		
Reagent	Final concentration	Amount
NaCl (4 M)	2 M	25 mL
Tris-HCl pH 7.4 (0.5 M)	10 mM	1 mL
EDTA pH 8.0 (0.5 M)	1 mM	100 μL
ddH_2O	-	To 50 mL
Total	**N/A**	**50 mL**
Store at 20°C–25°C for up to 1 year.		

1× deamination reaction buffer		
Reagent	Final concentration	Amount
HEPES pH 7.4 (0.5 M)	20 mM	400 μL
NaCl (4 M)	150 mM	375 μL
DTT (1 M)	1 mM	10 μL
ddH_2O	-	To 10 mL
Total	**N/A**	**10 mL**
Filter through a 0.22 μm filter, store at 4°C for up to 2 weeks.		

PEI-MAX solution		
Reagent	Final concentration	Amount
PEI MAX	150 μM	240 mg
NaOH (2.5 M)	N/A	Until pH 7.0
ddH_2O	-	To 40 mL
Total	**N/A**	**40 mL**
Filter through a 0.22 μm filter, store at 4°C for up to 6 months.		

STEP-BY-STEP METHOD DETAILS

Recombinant AID protein expression and purification

Timing: 4–5 days

In this section, recombinant AID is expressed in Expi293F cells and purified using amylose resin. The total yield of purified AID is about 1.25 mg from 500 mL culture.

1. Expi293F cell culture and transfection. Troubleshooting 1.
   a. Day 1. Grow starter cultures in 125–250 mL conical cell culture flasks with vented caps, in volumes between 30 and 60 mL.
      i. Maintain cells at 0.5–5 × 10^6 cells/mL.
      ii. Passenger cells the day before transfection.
      iii. Incubate the cells in an orbital shaker incubator at 37°C, 120 rpm, 5% CO2.
   b. Day 2. Transfect Expi293F cells with pcDNA3.4-MBP-hAID plasmid using PEI-MAX and culture the cells in serum-free medium at 37°C with 5% CO2 for 60 h.
      i. Before transfection, change the cell culture medium to fresh medium and keep the final density at 2.0 × 10^6 cells/mL.
      ii. For a 500-mL culture, add 500 μg of filter-sterilized pcDNA3.4-MBP-hAID plasmid into 12.5 mL fresh medium and gently mix by inverting 5–6 times.
      iii. Pipette 250 μL (PEI-MAX:plasmid = 3:1, mass ratio) of the filter-sterilized PEI-MAX solution to the other 12.5 mL fresh medium and gently mix by inverting 5–6 times.
      iv. Mix the plasmid and PEI-MAX master mix in the above two steps, then incubate for 20 min at 20°C–25°C.
      v. Add the 25 mL of plasmid:PEI-MAX:medium mixture to the 500 mL of cell culture and shake the bottle for 5 s.
      vi. Incubate the transfected cells in an orbital shaker incubator at 37°C, 120 rpm, 5% CO2 for 60 h.

⚠ CRITICAL: Serum-free Expi293F medium should be pre-warmed in a 37°C water bath prior to any cell culture work. The Expi293F serum-free medium is a ready-to-use complete medium without any further additives and can be repaced by other media of the same class, e.g., Expi293 expression medium (A1435101; ThermoFisher Scientific).

Note: Cells are harvested at 2–3 days post-transfection without changing or adding new medium. The pCDNA3.4-MBP-AID plasmid will be deposited to Addgene. The plasmid was generated by inserting an N-terminal His-MBP tag and a PreScission protease cleavage site following the human AID coding sequence downstream of the CMV promoter based on the pCDNA3.4 vector (Thermo Fisher Scientific).

Note: The amount of plasmid and PEI-MAX can be adjusted proportionally according to the cell number.

2. AID protein purification. Troubleshooting 2.

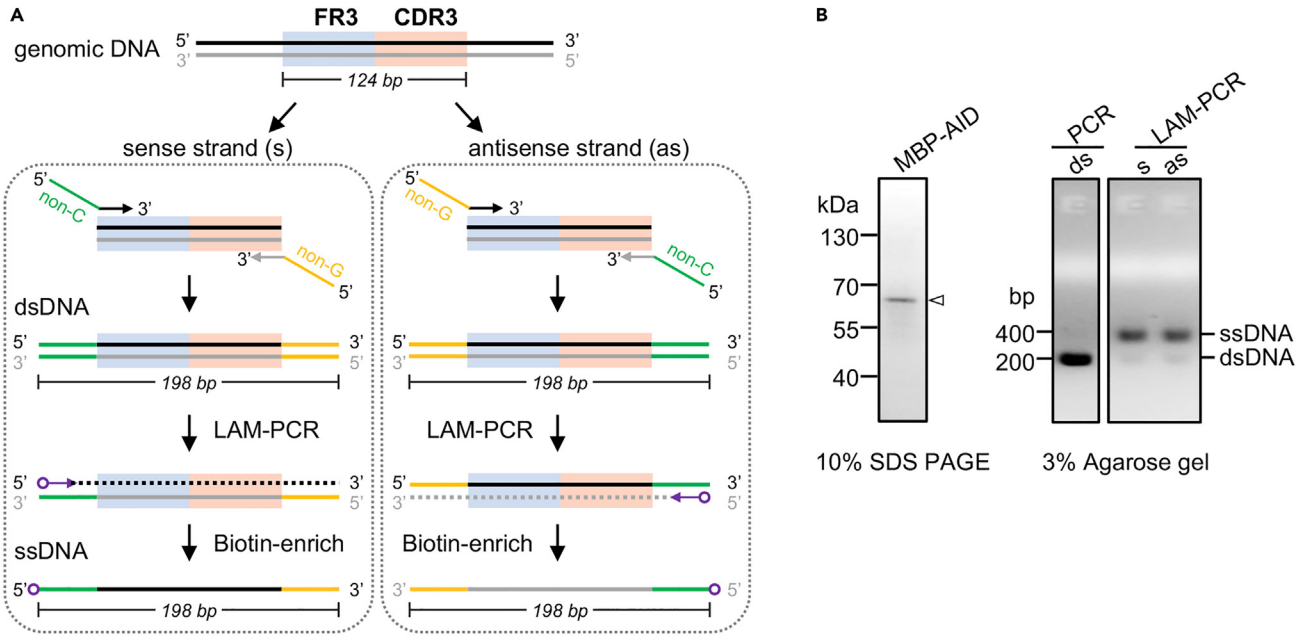


Figure 1. Preparation of long ssDNA substrates and recombinant AID protein

(A) Overview of the preparation of long ssDNA substrates. The ssDNA was prepared with a biotin-conjugated non-C primer via linear-amplification mediated (LAM)-PCR.

(B) Recombinant purified AID protein separated by the SDS-PAGE followed by Coomassie blue staining are shown at left. Prepared dsDNA and ssDNA substrates were separated on the agarose gel followed by Ethidium bromide staining are shown at right.

a. Day 5. Harvest cells by centrifugation at 2,000 × *g* for 20 min at 4°C.

⚠ CRITICAL: All the buffer for purification should be autoclaved or filtered with a 0.22 μm filter before use. Keep the whole operation on ice or at 4°C.

b. Discard supernatant and thoroughly resuspend cells in 30–40 mL of pre-chilled lysis buffer.

Note: Add PMSF, DTT and DNase I to the lysis buffer prior to use.

c. Lyse the cells using a high-pressure crusher (manufacture model) or similar instrument at 100 Pa for 1 min followed by 300 Pa for 5 min.
d. Centrifuge the lysate at 45,000 × *g* for 40 min at 4°C to pellet the cell debris.
e. Equilibrate 1 mL of amylose resin with 40 mL of pre-chilled ddH_2O and then with 20 mL of pre-chilled protein purification wash buffer in a gravity chromatography column.

Note: Add DTT to the wash buffer prior to use.

f. Load the supernatant onto the column, and allow it to flow slowly under gravity (1 mL per minute). Repeat this loading step two more times to increase the final protein yield.
g. Wash the resin with 10–20 column volumes of protein purification wash buffer.
h. Elute the protein from the column using 10 × 500 μL protein elute buffer.

Note: Add DTT to the protein elute buffer prior to use. Add 500 μL protein elute buffer each time. In general, the protein concentration is higher from elute 2 to elute 4.

i. Measure the concentration and purity of the recombinant AID protein by Bradford assay and gel electrophoresis (Figure 1B).

j. Aliquot the protein and snap freeze with liquid nitrogen, then store at −80°C.

Long ssDNA substrates preparation

Timing: 2–3 days

In this section, long ssDNA substrates (~200 nt) are generated using a linear-amplification mediated PCR (LAM-PCR) approach followed by biotin-streptavidin purification. With this protocol, 1,500 ng of ssDNA can be prepared.

3. dsDNA template preparation.
   a. Day 1. Design primers to amplify the specific DNA sequence.

Note: As an example, we chose a 124-bp substrate that covers the FR3-CDR3 sequence of the mouse V_HB1-8 exon, which has biased WRC mutations in the CDR3 *in vivo*.[13,14] To facilitate downstream PCR reactions, an additional non-C sequence is added to both ends of the dsDNA template, resulting in a final length of 198 bp (Figure 1A).

   b. Set up eight 50-μL PCR reactions for each sequence as below:
   PCR reaction master mix

Reagent	Final concentration	Amount
Genomic DNA	0.2 ng/μL	10 ng
Phusion High-Fidelity DNA Polymerase	0.02 U/μL	0.5 μL
FW-non-C-sB18 or FW-non-G-asB18 (10 μM)	0.2 μM	1 μL
RV-non-G-sB18 or RV-non-C-asB18 (10 μM)	0.2 μM	1 μL
dNTPs mix (10 mM each)	0.2 mM each	1 μL
5× HF buffer	1×	10 μL
ddH_2O	N/A	to 50 μL

Note: non-C or non-G sequences are underlined, gene-specific sequences are shown in blue. To amplify the sense strand substrate:

FW-non-C-sB18: 5'-AGATGTGGATGAGGAAGGTTAGAGTGAGTGTGGATGTAGACAAACCCTCCAGCA-3'
RV-non-G-sB18: 5'-TACAACACACACCTTCTAACCACACTCTCACACTGGTGCCTTGGCCCCAGTAG-3'
To amplify the anti-sense strand substrate:
FW-non-G-asB18: 5'- TACAACACACACCTTCTAACCACACTCTCACACTCAAACCCTCCAGCA-3'
RV-non-C-asB18: 5'- AGATGTGGATGAGGAAGGTTAGAGTGAGTGTGGATGTAGAGGTGCCTTGGCCC-3'

   c. Set a PCR program to amplify the DNA fragments as below:
   PCR cycling conditions

Steps	Temperature	Time	Cycles
Initial Denaturation	98°C	30 s	1
Denaturation	98°C	10 s	30–35 cycles
Annealing	60°C	20 s	
Extension	72°C	15–30 s/kb	

(*Continued on next page*)

Continued

Steps	Temperature	Time	Cycles
Final extension	72°C	5–10 min	1
Hold	4°C	forever	

Note: For different templates, the annealing temperature, extension time and number of cycles can be adjusted.

Pause point: Amplified DNA products can be stored at −20°C for months.

d. Separate the PCR products by agarose gel electrophoresis.
   i. Pool the eight tubes of PCR products.
   ii. Add 80 μL 6× DNA loading dye directly to the pool.
   iii. Run the sample by a 2% agarose gel in 1× TAE buffer.
e. Cut the 198-bp DNA fragment from the gel.
f. Add 2–3 mL of PC buffer included in the Universal DNA Purification Kit to the excised gels and allow to completely dissolve at 20°C–25°C.
g. Add 500 μL BL buffer to a Universal DNA Purification column and spin at 13,523 × *g* for 1 min at 20°C–25°C, discard the flow-through.
h. Add the solubilized gel mixture to the balanced column and spin at 13,523 × *g* for 1 min at 20°C–25°C, discard the flowthrough.
i. Add 600 μL PW buffer to the column, spin at 13,523 × *g* for 1 min at 20°C–25°C, discard the flowthrough. Repeat this step one more time.
j. Spin the column at 13,523 × *g* for 2 min at 20°C–25°C, transfer the column to a new 1.5 mL microtube.
k. Add 50–100 μL ddH_2O to the column, incubate for 2 min at 20°C–25°C, spin at 13,523 × *g* for 2 min at 20°C–25°C.
l. Check the DNA concentration using a NanoDrop or similar instrument.

Pause point: Purified DNA products can be stored at −20°C for months.

4. LAM-PCR and size selection. Troubleshooting 3.
   a. Day 2. Prepare four 50-μL PCR reactions for sense or antisense strand as follows:
   Primers for LAM-PCR:
   Biotinylated primer: 5′ Biotin-AGATGTGGATGAGGAAGGTT-3′.
   PCR reaction master mix

Reagent	Final concentration	Amount
dsDNA template	10 ng/μL	500 ng
TransStart FastPfu DNA Polymerase	0.025 U/μL	0.5 μL
Biotinylated primer (1 μM)	10 nM	0.5 μL
dNTPs mix (2.5 mM each)	75 μM each	1.5 μL
5× TransStart FastPfu buffer	1×	10 μL
ddH_2O	N/A	to 50 μL

   b. Set the PCR program to linear-amplify ssDNA as below:
   PCR cycling conditions

Steps	Temperature	Time	Cycles
Initial Denaturation	98°C	2 min	1
Denaturation	95°C	30 s	80 cycles
Annealing	58°C	30 s	
Extension	72°C	90 s	
Final extension	72°C	2 min	1
Hold	4°C	<4 h	

Note: Do not leave PCR products in the PCR machine for too long at 4°C after the PCR amplification, as the TransStart FastPfu DNA polymerase may resect the 3′ ends of the ssDNA products.

c. Size selection.
   i. Add 60 μL Dynabeads MyOne Silane (2.4 mg) to each 50 μL PCR products.
   ii. Mix by gentle pipetting 5–6 times.
   iii. Incubate at 20°C–25°C for 10 min to remove excess biotinylated primer.
d. Retrieve the beads on a magnetic stand for 2 min, and discard the supernatant.
e. Wash the beads three times with 200 μL 70% ethanol.
f. Remove the 70% ethanol and dry the beads for 3–5 min at 20°C–25°C.
g. Resuspend the beads in 40 μL Tris-HCl (10 mM, pH7.4).
h. Retrieve the beads on a magnetic stand for 2 min.
i. Combine the supernatants of 4 reactions in a new tube, final volume 160 μL.

Pause point: The ssDNA fragments can be stored at −20°C for up to one week. Longer storage is not recommended as ssDNA products are not stable.

Note: Completely resuspend the streptavidin beads by vortexing for at least 30 s prior to pipetting. Different concentrations of Dynabeads can be used to capture DNA of different sizes. The final concentration can be adjusted according to specific conditions.

5. Streptavidin affinity purification.
   a. Day 3. Add 40 μL 5 M NaCl (1 M final) and 2 μL 0.5 M EDTA (pH 8.0, 5 mM final) to the 160 μL ssDNA products.
   b. Transfer 20 μL Dynabead MyONE C1 Streptavidin Beads (200 μg) to a new 1.5-mL microtube, add 600 μL 1× Binding and Washing (B&W) buffer and mix by pipetting.

Note: Before pipetting streptavidin beads, fully resuspend the beads by vortexing for at least 30 s.

   c. Retrieve the beads on a magnetic stand for 2 min, and discard the supernatant. Repeat this step three times.
   d. Resuspend the beads with pooled ssDNA products from step 5a, and incubate the mixture on a rotary shaker at 20°C–25°C for at least 2 h.

Note: Although 2 h is sufficient for the beads to capture most of the biotinylated PCR products, longer incubation times (up to 4 h) can be used.

   e. Retrieve the DNA-beads complex on the magnetic stand and wash the DNA-beads complex with 200 μL 1× Binding and Washing (B&W) buffer, 200 μL 10 mM NaOH, 200 μL 1× Binding and Washing (B&W) buffer and 200 μL Tris-HCl (10 mM, pH7.4), respectively.
   f. Retrieve the DNA-beads on a magnetic stand for 2 min, and discard the supernatant.

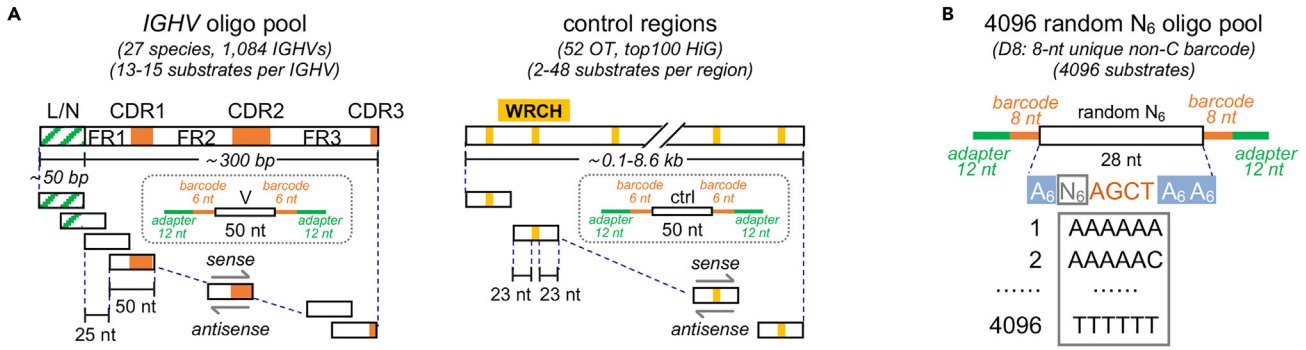


Figure 2. Design of short ssDNA oligo pools containing complex sequences

(A) Illustration of short ssDNA oligo pools. Two oligo pools, corresponding to the sense and antisense strands, were synthesized to cover the *IGHVs* and control regions. Each sequence was tiled into a 50-nt oligo pool at 25-nt resolution. L/N notes the leader sequence or 50 N added to the 5′ side of V-REGION without leader. The final substrates are 5′ - '12 nt adapter' - '6 nt barcode' - 50-nt - '6 nt barcode' - '12 nt adapter' - 3′.

(B) Illustration of 4096 random N_6 ssDNA oligo pools. The final substrates are 5′ - '12 nt adapter' - '8 nt barcode' - 28-nt - '8 nt barcode' - '12 nt adapter' - 3′.

g. Resuspend the DNA-beads complex with 30–50 μL ddH_2O. Heat the final ssDNA at 95°C for 10 min and flash-cool the sample on ice.
h. Retrieve the ssDNA supernatant on the magnetic stand.
i. Determine the ssDNA concentration using a NanoDrop or a similar machine.
j. Aliquot the ssDNA and store at −20°C.
k. Evaluate the purity of the ssDNA by running 200 ng of ssDNA on a 3% agarose gel in 1× TAE buffer (Figure 1B).

ssDNA oligo pools preparation

Timing: 3–4 weeks

In this section, we design oligo pools as deamination substrates. As an example, IGHV oligo pools containing 1,084 IGHVs from 27 species, and 4,096 random N_6 oligo pool are designed.

⚠ CRITICAL: Oligo pool design takes about a few days and oligo synthesis takes 2–3 weeks. The non-C primers and non-C barcodes should be added to both sides of each oligo. For the IGHV oligo pool, the sense and antisense strand oligo pools are designed separately. We use the sense strand as an example. The antisense strand oligo pool contains the same set of primers and barcodes as the sense strand, but only reverse-complement tested sequences. All custom scripts used in this section are available at https://github.com/ZhangSenxin/Deamination-HTS-Pipeline (https://doi.org/10.5281/zenodo.8271540).

6. IGHV oligo pool design.
   a. Week1. Download the functional L-PART1-IGHV-EXON segments from 16 species and the functional V-REGION without leader sequences from 11 species from the IMGT database.[15]

   Note: Chicken has only 1 functional V gene, so we include other pseudogenes as well. For the V-REGION without leader information, we arbitrarily add 50 of N on the 5′ side.

   b. Select the *01 allele of each V gene, resulting in a total of 1,084 Vs.

Table V gene segment . Species and V gene segment information

Classification	Scientific name	Common name	V num.	V Type
Mammal	*Homo sapiens*	Human	52	VH
	Macaca fascicularis	Crab-eating macaque	61	VH
	Macaca mulatta	Rhesus monkey	83	VH
	Mus musculus	Mouse	129	VH
	Rattus norvegicus	Rat	114	VH
	Canis lupus familiaris	Dog	35	VH
	Vicugna pacos	Alpaca	74	VH, VHH
	Ornithorhynchus anatinus	Platypus	35	VH
	Equus caballus	Horse	19	VH
	Oryctolagus cuniculus	Rabbit	38	VH
	Sus scrofa	Pig	13	VH
	Bos taurus	Cattle	13	VH
	Ovis aries	Sheep	7	VH
Bird	*Gallus gallus*	Chicken	81	VH
Bony fish	*Ictalurus punctatus*	Catfish	38	VH
	Gadus morhua	Cod	39	VH
	Danio rerio	Zebrafish	35	VH
	Salmo salar	Salmon	67	VH
	Oncorhynchus mykiss	Trout	43	VH
Cartilaginous fish	*Hydrolagus colliei*	Ratfish	12	VH
	Heterodontus francisci	Horn shark	17	VH
	Orectolobus maculatus	Wobbegong	8	VH, VNAR
	Carcharhinus leucas	Bull shark	4	VH
	Ginglymostoma cirratum	Nurse shark	53	VH, VNAR
	Carcharhinus plumbeus	Sandbar shark	8	VH
	Leucoraja erinacea	Little skate	5	VH
	Raja eglanteria	Clearnose skate	1	VH

c. Divide each V gene into a 50-nt oligonucleotide pool at 25-nt resolution.

Note: Since the average length of each V is about 350 bp, each V is tiled into 13–15 oligos (Figure 2A).

d. Pool all of the 50-nt sequences together and remove duplicates, resulting in a total of 9,753 unique sequences.
e. Select 52 AID off-targets (OT) in mouse CSR-activated B cells and the first exon of highly-transcribed genes (HiG)[16] in mouse GC B cells as controls.
f. For OT or HiG regions, select 50-nt sequences with a WRCH motif in the middle (23-nt-WRCH-23-nt), resulting in a total of 1,355 unique OT sequences and 892 unique HiG sequences, respectively. (g-i) Unique barcode design.
g. Extract and cluster the 9-nt sequences on both sides of the IGHVs and control sequences using custom scripts as follows: Troubleshooting 9.

```
>python3 cluster_inbarcode.py -i SEQ_FILE -o CLUSTER_PATH

>python3 barcode_assignment.py -i CLUSTER_PATH -o SAVE_FILE -m
```

h. Determine the largest sequence number in the cluster of identical 9-nt sequences.

Note: For example, in the current IGVH pool, the largest cluster contains ~500 sequences because many IGVHs from different species share some degree of similarity.

i. Select 30 of the 6-nt non-C barcodes with at least 2 nt difference, and add the barcodes to both sides of the 50-nt sequence, resulting in ssDNA 62 nt long.
6-nt non-C barcode

Table barcodes. 30 of the 6-nt non-C barcodes

AAAATG	ATTAGT	GGAATT	GTGGAT	TGTTGA	TTGGTG
AAGTTA	ATTATA	GGATGG	TAGTTG	TTAAGT	TTTGAG
AGGAAT	ATTGTG	GTTTAA	TATAAG	TTATTT	TGAGTG
AGTTAG	GAATGA	GTAGTT	TATTGT	TTGATT	TGATTA
ATATTG	GATTTG	GTATAG	TGAAAT	TTGGGA	TGTGAT

Note: There are 900 different combinations of 30 × 30 barcodes, which is enough to cover the largest cluster in the IGVH pool (~500 sequences). Different clusters use the same set of barcodes. In the IGVH pool, the 15-nt in-barcodes on both side of ssDNA sequence are used to demultiplex the reads in the following steps. The length of the non-C barcode can be extended accordingly.

j. Run a simulation test using a custom script to demultiplex each sequence according to the in-barcodes as follows:

```
>python3 fastq_simulation.py -I SEQ_FILE -o FASTQ_PATH
```

i. Generate raw data files with 10 paired-reads for each sequence, named as R1_10.fq.gz and R2_10.fq.gz
ii. Prepare the metafiles.

Note: Each metafile contains 500 sequences. The format is as follows: (ID, 'barcode1+9 bp sequence'-'the reverse complementary sequence of barcode2+9 bp sequence').

Table meta.test1.txt. Example of metafile1 for simulation test

WYY1	AAAATGAAAGGGGCC-CATTTTATGTGCTAC
…	…
WYY500	AAAATGCGCTTTGGA-CTATACCTTTCACTT

Table meta.test2.txt. Example of metafile2 for simulation test

WYY501	AAAATGCGGTGGTCA-CATTTTGGTCACGTG
…	…
WYY1000	AAAATGGAAAGCCTC-CATTTTATAAGTAAT

…

iii. Perform the simulation test using this code:

```
>fastq-multx -x -B meta.test1.txt -m 1 -d 1 -b R1_10.fq.gz R2_10.fq.gz -o test1/%_R1.fq.gz
test1/%_R2.fq.gz

> out.test1_m1.txt
```

iv. If each sequence is accurately split into 10 reads, the simulation is successful, indicating that the in-barcodes can be used to distinguish each sequence.

k. Add the 12-nt non-C PCR primers to both sides of the 62-nt sequences, resulting in a final length of 86 nt.

Note: The final oligo sequence is 5′- GTAAGGGTGAGG-barcode1-'50-nt test sequence'--barcode2-GATAGGGTGGTG-3′ (Figure 2A).

7. 4,096 N_6 oligo pool design.
   a. Week1. Generate 4096 sequences of $A_6N_6AGCTA_{12}$ *in silico*. N represents A, G, C or T.
   b. Add a unique 8-nt non-C barcode pair on each side of the 28-nt sequences, for a total length of 44 nt.
   c. Add 12-nt non-C PCR-primers to each side of the above sequence, resulting in a final length of 68 nt.

Note: The oligo sequence is 5′- GTAAGGGTGAGG-barcode1- $A_6N_6AGCTA_{12}$-barcode2-GATAGGGTGGTG-3′ (Figure 2B).

   d. As in Step 6j, run a simulation test to demultiplex each sequence according to the in-barcodes.
8. Week2-Week4. Synthesize the oligo pool using a commercial vendor (Twist Bioscience or similar).

In vitro AID deamination assay

Timing: 1–1.5 h

This section describes how to perform the *in vitro* AID deamination reaction.

9. Reaction setup. Troubleshooting 4, 5.
   a. Thaw AID protein on ice from the −80°C refrigerator.
      i. Centrifuge the enzyme at 9,391 × *g* for 10 min at 4°C prior to use.
      ii. Retrieve the supernatant carefully. Discard the ~5 μL of liquid at the bottom to remove any aggregated proteins.
   b. Heat the ssDNA at 95°C for 10 min, and quickly place the tube into ice to completely denature the substrates.
   c. Perform the reaction in a 10-μL format.
      i. Add the pre-cooled reaction buffer.
      ii. Add the ssDNA and AID.

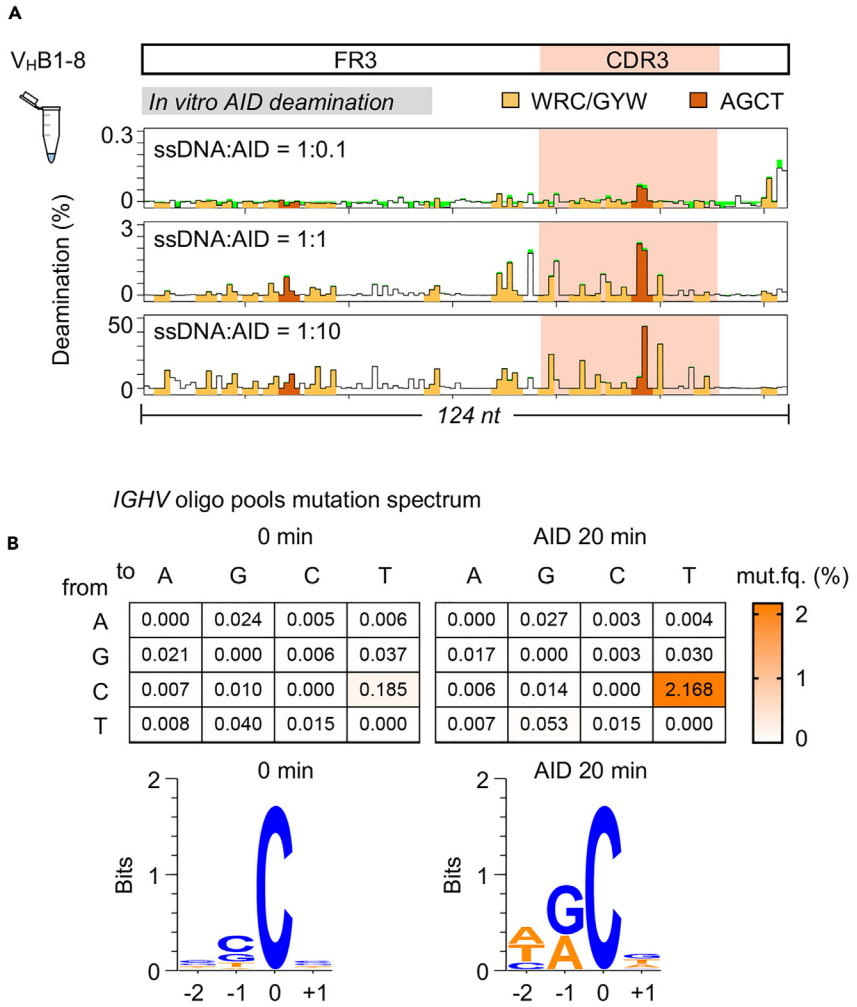
Note: The deamination reaction buffer contains 20 mM HEPES pH7.4, 10/150 mM NaCl and 1 mM DTT.

      iii. Pipette the mixture up and down several times.
      iv. Quick-spined on a desktop centrifuge for 5 s.
      v. Incubate the tube at 37°C for 5–60 min.

Note: In our hands, the reaction reaches a plateau after 60 min.

10. After incubation, stop the reaction by heating (95°C, 10 min).

Note: The salt concentration in the AID protein buffer should be counted in the final reaction salt concentration. Reaction time, substrate/enzyme ratio should be tested for each batch of purified enzyme, as exampled in Figures 3A and 3B.



from \ to	0 min A	0 min G	0 min C	0 min T	AID 20 min A	AID 20 min G	AID 20 min C	AID 20 min T
A	0.000	0.024	0.005	0.006	0.000	0.027	0.003	0.004
G	0.021	0.000	0.006	0.037	0.017	0.000	0.003	0.030
C	0.007	0.010	0.000	0.185	0.006	0.014	0.000	2.168
T	0.008	0.040	0.015	0.000	0.007	0.053	0.015	0.000

Figure 3. HTS-data analysis of long ssDNA substrates and short ssDNA oligo pools

(A) Mutation profiles of a FR3-CDR3 region from the *V_HB1-8* allele in *in vitro* deamination assay with different molar ratio of ssDNA:AID. Deamination frequency, i.e., substitution in total sequencing reads, at each nucleotide along the 124-nt region is plotted as a bar graph with green error bars, representing mean with SEM from 3 repeats. WRC/GYW motifs are labeled in yellow, and AGCT is labeled in orange.

(B) Deamination products of *IGHV* oligo pools were subjected to mutation spectrum (upper) and seqlogo analyses (lower). The seqlogo was obtained from the top16 tetramer with high mutation frequency in all 64 NNCN combinations.

Pause point: The final reaction product can be stored for up to one week at −20°C. Longer storage is not recommended because ssDNA is not stable.

Amplicon-seq library preparation

Timing: 1 day

This step describes how to prepare the amplicon sequencing library from deamination products for high-throughput sequencing (HTS).

11. U-friendly PCR. Troubleshooting 6.
    a. Set up a 10-μL PCR reaction for each sample as below:

PCR reaction master mix		
Reagent	Final concentration	Amount
Deaminated products	N/A	1 μL
PfuTurbo Cx Hot Start High-Fidelity DNA Polymerase	0.025–0.05 U/μL	0.1–0.2 μL
non-C-VB18-F or SubP5 (10 μM)	200 nM	0.2 μL
non-C-VB18-R or SubP7 (10 μM)	200 nM	0.2 μL
dNTPs mix (10 mM each)	200 nM each	0.2 μL
10× PfuTurbo Cx reaction buffer	1×	1 μL
ddH_2O	N/A	to 10 μL

Note: Primers for long ssDNA substrates (underlined sequences are used as the primer sites for the next step PCR, in-barcodes are in red, non-C or non-G sequences are in blue). A pair of primers used for *VB1-8* FR3-CDR3 amplicon is shown.

non-C-VB18-F-27: 5′- TTCCCTACACGACGCTCTTCCGATCTATTCCTAGAGTGAGTGTGGATGTAGA-3′
non-G-VB18-R-46: 5′- AGTTCAGACGTGTGCTCTTCCGATCTTCCCGATTCTAACCACACTCTCACACT-3′

Primers for short ssDNA oligo pools: (underlined sequences are used as the primer sites for the next step PCR, in-barcodes are in red, 12-nt non-C or non-G sequences are in blue).

SubP5F25: 5′-TTCCCTACACGACGCTCTTCCGATCTACTGATTAGGTGAGTAAGGGTGAGG-3′
SubP7F38: 5′- AGTTCAGACGTGTGCTCTTCCGATCTCTAGCTTCATCCACTTCACCACCCTATC-3′

    b. Set up the following PCR program:

PCR cycling conditions			
Steps	Temperature	Time	Cycles
Initial Denaturation	95°C	2 min	1
Denaturation	95°C	30 s	20 cycles
Annealing	60°C	30 s	
Extension	72°C	1 min for targets ≤ 1 kb	
Final extension	72°C	10 min	1
Hold	4°C	Forever	

Note: This 1st step of PCR can also use other U-friendly enzymes, such as Q5U Hot Start High-Fidelity DNA Polymerase, or Phusion U DNA polymerase. The in-barcode is used to distinguish the difference batch of experiments, which is different from the in-barcode designed in step 6 or 7.

12. Amplicon-seq HTS library preparation.
    a. Set up a 25-μL PCR reaction for each sample as below:

PCR reaction master mix		
Reagent	Final concentration	Amount
1st PCR products	N/A	1 μL

(*Continued on next page*)

Continued		
Reagent	Final concentration	Amount
Phusion High-Fidelity DNA Polymerase	0.02 U/μL	0.25 μL
PE-P5-short (10 μM)	200 nM	0.5 μL
PE-P7-index5 (10 μM)	200 nM	0.5 μL
dNTPs mix (10 mM each)	200 nM each	0.5 μL
5× HF buffer	1×	5 μL
ddH_2O	N/A	to 25 μL

Note: Primers for 2nd step PCR (underlined sequences overlap with 1st step PCR primers, barcode is in red):

5'- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC-3'
PE-P7-index5:
5'- CAAGCAGAAGACGGCATACGAGATACAGTGGTGACTGGAGTTCAGACGTGT-3'

b. Set the following PCR program to amplify the DNA fragments:

PCR cycling conditions			
Steps	Temperature	Time	Cycles
Initial Denaturation	98°C	30 s	1
Denaturation	98°C	10 s	20 cycles
Annealing	60°C	30 s	
Extension	72°C	15–30 s/kb	
Final extension	72°C	10 min	1
Hold	4°C	Forever	

Pause point: Amplified DNA products can be stored at −20°C for months.

13. HTS-library purification.
    a. Add 5 μL 6× DNA loading dye directly into the PCR products and run the sample on a 2% agarose gel in 1× TAE buffer.
    b. Excise the target DNA fragments from the gel.
    c. Add 2–3 mL of PC buffer provided in the Universal DNA Purification Kit into the excised gels and allow to fully dissolve at 20°C–25°C.
    d. Add 500 μL BL buffer to a Universal DNA Purification column and spin at 13,523 × *g* for 1 min at 20°C–25°C, discard the flow through.
    e. Load the dissolved gel mixture onto the balanced column and spin at 13,523 × *g* for 1 min at 20°C–25°C, discard the flow-through.
    f. Add 600 μL buffer PW to the column, spin at 13,523 × *g* for 1 min at 20°C–25°C, discard the flow-through. Repeat this step one more time.
    g. Spin the column at 13,523 × *g* for 2 min at 20°C–25°C, transfer the column to a new 1.5 mL tube.
    h. Add 50–100 μL ddH_2O to the column, incubate for 2 min at 20°C–25°C, spin at 13,523 × *g* for 2 min at 20°C–25°C.
    i. Measure the concentration using a Qubit Quantification Kit.

HTS data analysis

Timing: 1–2 days

This section describes how to perform the HTS data analysis. Data from long-ssDNA and oligo-pool substrates are processed differently.

⚠ CRITICAL: All custom scripts used in this section are available at https://github.com/ZhangSenxin/Deamination-HTS-Pipeline (https://doi.org/10.5281/zenodo.8271540).

14. Create a metafile.txt file for Hiseq run. Troubleshooting 7, 8.

Note: The metafile contains the configuration information needed to process sequence reads for a given sample. Incorrect information can lead to errors at various stages of the pipeline or produce incorrect results. The metadata file is a tab-delimited plain text file containing the following header line with each subsequent line describing the design of a single sample.

Table meta.wyy.txt. Example of the metafile for Hiseq run

Experiment	Genotype	Allele	Mouse/cell	Tissue/clone	PE-P7-index	Barcode	Forward_primer	Reverse_primer	In-barcode-F	In-barcode-R	Reference	Start	End
WYY15599	sVB18	LAM	AID	r1	R706	GCCAAT	AGAGTGAGTGTGGATGTAGA	TTCTAACCACACTCTCACACT	CACGAT	CGGAAT	VB18.fa	41	164

"experiment"- the unique name of the sample, this ID will be used to name most files generated by the pipeline.

"genotype", "allele", "mouse/cell", "tissue/clone"- information to describe the sample.

"PE-P7-index"- the unique PE-P7 index ID.

"barcode"- sequence of the PE-P7-index.

"forward_primer"- sequence of the gene-specific primer on P5 side.

"reverse_primer"- sequence of the gene-specific primer on P7 side.

"in-barcode-F"- barcode on the P5 side that you will add to your forward_primer.

"in-barcode-R"- barcode on the P7 side that you will add to your reverse_primer.

"reference"- The pipeline uses this name to find the reference sequence and bowtie2 index on the filesystem.

"start"- position of the next nucleotide at the end of the forward_primer.

"end"- position of the next nucleotide at the end of the reverse_primer.

15. For long ssDNA substrates, run the SHM pipeline[13] to generate the result files. Multiple samples can be processed in the same time.

```
>perl SHMPipeline.pl --meta meta.wyy.txt --in raw --out result_wyy

--ref ./ref --threads 8--ow &> wyy.nohup.txt
```

Note: The SHM pipeline reads the metafile and calls bowtie2 to align the forward and reverse reads to the reference sequence.

16. The WYY*_filt_profile.txt can be used directly to plot the mutation profile using the custom scripts as follows:

```
>For WYY in *txt

>do RScript ./SHMPlot2_zhou1_WRC.R ./$WYY/$WYY.pdf tstart=xx tend=xxx plotrows=1 ymax=0.35
figureheight=2 showsequence=T or F; done
```

Note: Where tstart and tend are the start and end filled in the metafile, plotrows defines the number of rows displayed in the plot, ymax means the maximum value of the mutation frequency, figureheight defines the height of the plot, showsequence = T or F means to show or hide the sequence information, as illustrated on Figure 3A.

17. For short ssDNA oligo pools, demultiplex the raw reads and remove the 12-bp adapters at the 5′ and 3′ ends of the oligos using cutadapt v1.9 as follows:

```
>bash cutadapter.sh
```

18. Filter out reads with Illumina Sequencing Quality Scores less than 30.

```
>source activate fastp

>fastp -i WYY*_recut_R1.fq.gz -I WYY*_recut_R2.fq.gz -o WYY*_recut_R1_q30.fq.gz -O WYY*_r-
ecut_R2_q30.fq.gz -q 30 -u 10 -A
```

19. Divide each substrate according to the in-barcodes.
    a. Create multiple metafile_cut.txt files with no header to split the reads.

Note: Since the fastq-multx program has an upper limit for each run, each metafile contains 500 rows and 2 columns, the first column is the sample ID, the second column is the internal barcode information, the format is 5′barcode-3′barcode (reverse complement).

    b. Divide each sequence into different files using the custom script as follows:

```
>bash multx.sh
```

20. Annotate and generate mutation profiles using the custom script as follows:

```
>python3 read_fq.gz.py -m ./array1_ref.txt -f ./array1/
-s ./array1/out2/ -c 1 > array1.txt
```

Note: Where -m: oligo pool reference file, -f: the data path generated in step 19b, -s: save path, -c defines the sense strand (-c 1, default) or the antisense strand (-c 0).

21. Mutation spectrum analysis.
    a. Mutation spectrum statistics for A, G, C, T using the custom script as follows (Figure 3B):

```
>python3 mut_summary.py -m ./IGHV_ref_kabat.txt
-f ./array1/out2/ -s ./array1/out3_kabat/ > mutation_summary01_out.txt
```

Note: Where -m: oligo pool reference file containing region information (one sequence spans up to 3 regions), -f: the data path generated in step 20 -s: save path.

    b. Mutation spectrum statistics for C in WRC, SYC, WGCW, AGCT using the custom script as follows:

```
>python3 mut_summary2.py -m ./IGHV_ref_kabat.txt
-f ./array1/out2/ -s ./array1/out3_kabat/ > mutation_summary02_out.txt
```

Note: Where -m: oligo pool reference file containing region information (one sequence spans up to 3 regions), -f: the data path generated in step 20, -s: save path.

22. Combine the tiling assay into a full-length V gene. Troubleshooting 10.

```
>python3 mapping2full_length_sequence.py
-m ./full_length_reference.csv
-r ./IGHV_ref_kabat.txt -f ./array1/out2/
-s ./array1/full_length_out/
-i ./array1/full_length_mismatch.txt > full_length_out.txt
```

Note: Where -m: sequence reference file containing full length V, -r: oligo pool reference file containing region information (one sequence spans up to 3 regions), -f: the data path generated in step 20, -s: save path, -i file path to save mismatch information (if exist).

Note: This step combines each 50-nt segment from a V gene and generate the full-length V mutation information.

23. Plot mutation frequency for V genes. Troubleshooting 11.
    a. Pre-process the data:

```
>python3 data_produce.py -m ./full_length_reference.csv
-r ./V_region_kabat.csv -f ./array1/full_length_out/
-s ./array1/plot_data/ -p ./species.txt > data_produce_out.txt
```

Note: Where -m: sequence reference file containing full length sequence, -r: region reference file containing V gene region info, -f: the data path generated in step 22, -s: save path, -p species information.

b. Plot mutation frequency based on specie pre-processed data (as shown in Wang et al.[1]):

```
>python3 figure_plot.py -f ./array1/plot_data/
-s ./array1/plot_figure/ -p ./species.txt
```

Note: Where -f: the data path generated in step 23a, -s: save path, -p species reference file.

EXPECTED OUTCOMES

Following this protocol, we first obtained the full-length recombinant human AID protein and used this protein to perform the *in vitro* deamination assay on long ssDNA substrates or short ssDNA oligo pools, as shown in Figures 1 and 2. Finally, we prepared the amplicon library for HTS and analyzed the data, as shown in Figure 3.

LIMITATIONS

First, the *in vitro* AID deamination profile correlated strongly with the *in vivo* mutation pattern at C/G sites but not at A/T sites, as mutations in the latter depend on cellular DNA repair machineries,[17] which is absent in the *in vitro* experiments. Second, Although the LAM-PCR method generates majority ssDNA, a small amount of dsDNA was also observed. Other methods to generate long-ssDNA substrates could be applied.

TROUBLESHOOTING

Problem 1

AID protein expression with low yield (related to step 1: Expi293F culture and transfection).

Potential solution

Check the condition of the cells under the microscope. If most of the cells are irregular in shape, it means that the cells are not in good condition and are not suitable for transfection. During the culture process, it is necessary to add the fresh medium in time to ensure that the cells are in the best condition before transfection.

Problem 2

AID protein purification with low yield (related to step 2: AID purification).

Potential solution

Using fresh prepared buffer at each step, and process the samples as fast as possible.

Problem 3

Very low long ssDNA concentration (<3 ng/μL) (related to step 4: LAM-PCR).

Potential solution

Double-check the biotin-conjugated primer with your vendor. Use correct primer and/or dNTPs concentrations for step 4a. Test the primer by gradient PCR and select an optimal primer melting temperature (Tm) for step 4b. Do not let the beads dry out, use a pipette tip to remove any residual ethanol on the side wall and add Tris-HCl (10 mM, pH7.4) when the edges of the beads start to dry for step 4f.

Problem 4

AID has poor enzymatic activity, or the overall mutation level is very low (related to step 9: *in vitro* deamination assay).

Potential solution

Use freshly purified AID, long storage time affects its activity. Do not refreeze the protein. For step 9c, test the molar ratio gradient of AID to substrates and the incubation time course.

Problem 5

The overall mutation level is too high (related to step 9: *in vitro* deamination assay).

Potential solution

Reduce the amount of AID and shorten the incubation time.

Problem 6

No or weak PCR band (related to step 11–13: Amplicon library preparation).

Potential solution

This is usually caused by the failure of the U-friendly PCR. The dense WRCs on the substrate will produce clustered Us after deamination by AID, which will affect the efficiency of the U-friendly PCR. You can halve the input template, double the PfuTurbo Cx Hot Start High-Fidelity DNA Polymerase, or extend the PCR extension time for step 11. However, it is not recommended to increase the number of cycles.

Problem 7:

Sequence contamination with many other irrelevant template sequences (related to step 14–22: HTS data analysis).

Potential solution

Clean your bench, pipette and etc. Use filtered tips.

Problem 8

For ssDNA oligo pools, less than 70% of the sequencing results can be mapped to the reference sequences. ~30% of the sequences have no reads (related to step 14–22: HTS data analysis).

Potential solution

The substrates may be partially degraded as ssDNA is not stable. When first dissolving the substrate, use nuclease-free TE buffer to prepare the stock solution and store the aliquots at −20°C or −80°C.

Problem 9

The cluster_inbarcode.py script returns a MemoryError (related to step 6g).

Potential solution

It is likely that there are too many sequences. Since the complexity of the distance matrix calculation is O(n2), a large amount of memory will be occupied during the matrix calculation. About 10 GB of memory will be used for 9753 sequences.

Problem 10

Pipeline does not run (related to step 22).

Potential solution

- Incorrect metafile.

Once we have created the full_length_reference.csv metafile and used it as input for mapping2full_length_sequence.py, the order of each row cannot be changed. The outputs of this step are order dependent and are needed in the following steps, such as data processing and plotting. Therefore, if you encounter an error message or unexpected results in the following steps, it would be a good idea to check the full_length_reference.csv meta file.

- Incorrect input path.

When using this pipeline, there are several input options to consider. Some require a direct path to a file, while others require a folder path, which can be confusing. To address this issue, we have included a readme file (available on GitHub) that provides examples for each script and indicates the type of each input option. If an input option ends with "_FILE", it requires a file path, and if it doesn't, it requires a folder path. Here is an example:

```
>python3 cluster_inbarcode.py -i SEQ_FILE -o CLUSTER_PATH
```

- Incorrect mapping to full-length reference.

After performing basic mutation statistics or motif mutation statistics, scripts will generate a mismatch_file.txt. This file records any files that were unsuccessfully demultiplexed. We need this file to run mapping2full_length_sequence.py, in case we encounter a FileNotFoundError.

Problem 11

Failure to pre-process data and plot figures (related to step 23).

Potential solution

In this step, we assume that there are sequencing data for both sense and antisense strands, since the data pre-processing using the data_produce.py script needs to be run twice, once for the sense strand and once for the antisense strand. This will allow us to provide two correct paths for the complete visualization in the "Plot" step.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Fei-Long Meng (Feilong.Meng@sibcb.ac.cn).

Materials availability

Plasmids, primers, recombinant proteins, experimental strains, and any other research reagents generated by the authors will be distributed upon request to other research investigators under a material transfer agreement.

Data and code availability

- Data used in this protocol has been generated and analyzed in a previous work, see Wang et al.[1] The dataset for analyses included in this protocol has been deposited at Gene Expression Omnibus (GEO) database with an accession number BioProject: PRJNA918596.
- Code used in this study is publicly available at https://github.com/ZhangSenxin/Deamination-HTS-Pipeline (https://doi.org/10.5281/zenodo.8271540).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was financially supported by the National Key R&D Program of China (2021YFA1301400 to L.-S.Y.), NSFC (32090040, 31970880 to F.-L.M.; 31722020, 81671634, and 81861138014 to L.-S.Y.; 61972257 to X.Z.), Natural Science Foundation of Shanghai (20490760200 to F.-L.M.), and Chinese Academy of Sciences (JCTD-2020-17 to F.-L.M. and X.Z. and 318GJHZ2022010MI to F.-L.M.). Y.W. is a Yuhe postdoctoral fellow.

AUTHOR CONTRIBUTIONS

Y.W. and S.Z. wrote the protocol and drew the graph. F.-L.M., L.-S.Y., and X.Z. edited and revised the protocol.

DECLARATION OF INTERESTS

Shanghai Institute of Biochemistry and Cell Biology and Shanghai Jiao Tong University School of Medicine have filed a patent application based on the protocol in this article.

REFERENCES

1. Wang, Y., Zhang, S., Yang, X., Hwang, J.K., Zhan, C., Lian, C., Wang, C., Gui, T., Wang, B., Xie, X., et al. (2023). Mesoscale DNA feature in antibody-coding sequence facilitates somatic hypermutation. Cell *186*, 2193–2207.e19. https://doi.org/10.1016/j.cell.2023.03.030.

2. Bransteitter, R., Pham, P., Scharff, M.D., and Goodman, M.F. (2003). Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc. Natl. Acad. Sci. USA *100*, 4102–4107. https://doi.org/10.1073/pnas.0730835100.

3. Chaudhuri, J., Tian, M., Khuong, C., Chua, K., Pinaud, E., and Alt, F.W. (2003). Transcription-targeted DNA deamination by the AID antibody diversification enzyme. Nature *422*, 726–730. https://doi.org/10.1038/nature01574.

4. Dickerson, S.K., Market, E., Besmer, E., and Papavasiliou, F.N. (2003). AID mediates hypermutation by deaminating single stranded DNA. J. Exp. Med. *197*, 1291–1296. https://doi.org/10.1084/jem.20030481.

5. Pham, P., Bransteitter, R., Petruska, J., and Goodman, M.F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature *424*, 103–107. https://doi.org/10.1038/nature01760.

6. Yu, K., Huang, F.T., and Lieber, M.R. (2004). DNA substrate length and surrounding sequence affect the activation-induced deaminase activity at cytidine. J. Biol. Chem. *279*, 6496–6500. https://doi.org/10.1074/jbc.M311616200.

7. Larijani, M., Petrov, A.P., Kolenchenko, O., Berru, M., Krylov, S.N., and Martin, A. (2007). AID associates with single-stranded DNA with high affinity and a long complex half-life in a sequence-independent manner. Mol. Cell Biol. *27*, 20–30. https://doi.org/10.1128/MCB.00824-06.

8. Shi, K., Carpenter, M.A., Banerjee, S., Shaban, N.M., Kurahashi, K., Salamango, D.J., McCann, J.L., Starrett, G.J., Duffy, J.V., Demir, Ö., et al. (2017). Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. Nat. Struct. Mol. Biol. *24*, 131–139. https://doi.org/10.1038/nsmb.3344.

9. Xie, X., Gan, T., Rao, B., Zhang, W., Panchakshari, R.A., Yang, D., Ji, X., Cao, Y., Alt, F.W., Meng, F.L., and Hu, J. (2022). C-terminal deletion-induced condensation sequesters AID from IgH targets in immunodeficiency. EMBO J. *41*, e109324. https://doi.org/10.15252/embj.2021109324.

10. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191. https://doi.org/10.1093/nar/gku365.

11. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

12. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

13. Yeap, L.S., Hwang, J.K., Du, Z., Meyers, R.M., Meng, F.L., Jakubauskaitė, A., Liu, M., Mani, V., Neuberg, D., Kepler, T.B., et al. (2015). Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. Cell *163*, 1124–1137. https://doi.org/10.1016/j.cell.2015.10.042.

14. Bross, L., Fukita, Y., McBlane, F., Démollière, C., Rajewsky, K., and Jacobs, H. (2000). DNA double-strand breaks in immunoglobulin genes undergoing somatic hypermutation. Immunity *13*, 589–597. https://doi.org/10.1016/s1074-7613(00)00059-5.

15. Lefranc, M.P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., et al. (2009). IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. *37*, D1006–D1012. https://doi.org/10.1093/nar/gkn838.

16. Meng, F.L., Du, Z., Federation, A., Hu, J., Wang, Q., Kieffer-Kwon, K.R., Meyers, R.M., Amor, C., Wasserman, C.R., Neuberg, D., et al. (2014). Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. Cell *159*, 1538–1548. https://doi.org/10.1016/j.cell.2014.11.014.

17. Zeng, X., Winter, D.B., Kasmer, C., Kraemer, K.H., Lehmann, A.R., and Gearhart, P.J. (2001). DNA polymerase eta is an A-T mutator in somatic hypermutation of immunoglobulin variable genes. Nat. Immunol. *2*, 537–541. https://doi.org/10.1038/88740.