**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

| | | | |
|---|---|---|---|
| Application No. | : | 17/282,424 | Confirmation No. 7366 |
| Inventor(s) | : | Pardis Sabeti et al. | |
| Filed | : | April 2, 2021 | |
| Art Unit | : | 1634 | |
| Examiner | : | Steven C. Pohnert | |
| Customer No. | : | 198696 | |
| Docket No. | : | BROD-2545US | |
| Title | : | CRISPR EFFECTOR SYSTEM BASED DIAGNOSTICS FOR HEMORRHAGIC FEVER DETECTION | |

**AMENDMENT AND RESPONSE**

**FILED VIA EFS-WEB**

Commissioner:

In response to the Final Office Action mailed November 20, 2024 ("Office Action"), please enter and consider the following amendments and remarks.

Certificate of EFS-WEB Transmission pursuant to 37 C.F.R. § 1.8

I hereby certify that this correspondence is being transmitted via the U.S. Patent and Trademark Office electronic filing system (EFS-Web) to the USPTO, on January 17, 2025.

/Christopher D. Southgate/

Christopher D. Southgate, Ph.D., Esq.

Reg. No. 54,875

# AMENDMENT

## *In the specification*

Please replace TABLE 10A in Example 3 of the specification as filed (on page 116, after paragraph [0354) with the following TABLE 10A where the guide sequences are rearranged to correspond with their guide RNA. No new matter is added.

| | **Ebola Guide Sequences** | |
|---|---|---|
| **Name** | **Full guide sequence** | **Spacer Sequence** |
| EBOV_Guide_2 | TTT AAC CCA AAT AAC TTG CAC AGT TGA TGT TTT AGT CCC CTT CGT TTT TGG GGT AGT CTA AAT CCC CTA TAG TGA GTC GTA TTA ATT TC (SEQ ID NO:159) | ~~CCTTTTCTCCTACTACCAATTTCG GAAG (SEQ ID NO:160)~~ ATCAACTGTGCAAGTTATTTGGGT TAAA (SEQ ID NO:166) |
| EBOV_Guide_9 | AGA ACA CTT GCT GCC ATG CCG GAA GAG GGT TTT AGT CCC CTT CGT TTT TGG GGT AGT CTA AAT CCC CTA TAG TGA GTC GTA TTA ATT TC (SEQ ID NO:161) | ~~CTCCTACTACCAATTTCGGAAGGA ATAG (SEQ ID NO:162)~~ CCTCTTCCGGCATGGCAGCAAGTG TTCT (SEQ ID NO:164) |
| EBOV_Guide_10 | CTA TTC CTT CCG AAA TTG GTA GTA GGA GGT TTT AGT CCC CTT CGT TTT TGG GGT AGT CTA AAT CCC CTA TAG TGA GTC GTA TTA ATT TC (SEQ ID NO:163) | ~~CCTCTTCCGGCATGGCAGCAAGTG TTCT (SEQ ID NO:164)~~ CTCCTACTACCAATTTCGGAAGGA ATAG (SEQ ID NO:162) |
| EBOV_Guide_11 | CTT CCG AAA TTG GTA GTA GGA GAA AAG GGT TTT AGT CCC CTT CGT TTT TGG GGT AGT CTA AAT CCC CTA TAG TGA GTC GTA TTA ATT TC (SEQ ID NO:165) | ~~ATCAACTGTGCAAGTTATTTGGGT TAAA (SEQ ID NO:166)~~ CCTTTTCTCCTACTACCAATTTCG GAAG (SEQ ID NO:160) |
| EBOV_Guide_12 | CAC ACT CCC ATG TAT GAT TGA GCA ATT CGT TTT AGT CCC CTT CGT TTT TGG GGT AGT CTA AAT CCC CTA TAG TGA GTC GTA TTA ATT TC (SEQ ID NO:167) | ~~TGCCCATGAATATTCCCTCAGGAT CTGT (SEQ ID NO:168)~~ GAATTGCTCAATCATACATGGGAG TGTG (SEQ ID NO:172) |
| EBOV_Guide_13 | ATT GGA GCC ACA CTC CCA TGT ATG ATT GGT TTT AGT CCC CTT CGT TTT TGG GGT AGT CTA AAT CCC CTA TAG TGA GTC GTA TTA ATT TC (SEQ ID NO:169) | ~~CAATCATACATGGGAGTGTGGCTC CAAT (SEQ ID NO:170)~~ CAATCATACATGGGAGTGTGGCTC CAAT (SEQ ID NO:170) |
| EBOV_Guide_14 | ACA GAT CCT GAG GGA ATA TTC ATG GGC AGT TTT AGT CCC CTT CGT TTT TGG GGT AGT CTA AAT CCC CTA TAG TGA GTC GTA TTA ATT TC (SEQ ID NO:171) | ~~GAATTGCTCAATCATACATGGGAG TGTG (SEQ ID NO:172)~~ TGCCCATGAATATTCCCTCAGGAT CTGT (SEQ ID NO:168) |

Please replace paragraph [0014] of the specification as filed with the following replacement paragraph [0014]:

> [0014] In some embodiments, the system further comprises nucleic acid amplification reagents. The nucleic acid amplification reagents comprise recombinase polymerase amplification (RPA) reagents, nucleic acid sequence-based amplification (NASBA) reagents, loop-mediated isothermal amplification (LAMP) reagents, strand displacement amplification (SDA) reagents, helicase-dependent amplification (HDA) reagents, nicking enzyme amplification reaction (NEAR) reagents, RT-PCR reagents, multiple displacement amplification (MDA) reagents, rolling circle amplification (RCA) reagents, ligase chain reaction (LCR) reagents, ramification amplification method (RAM) reagents, transposase based amplification reagents; or Programmable CRISPR Nicking Amplification (PCNA)reagents. In some embodiments, the RPA reagents comprise one or more primer pairs selected from the group consisting of SEQ ID NOs: 78, 79, 81-86, 93-108, 127-138, 173-206, 233-248, 285-328, 370-392. In other embodiments, the nucleic acid amplification reagents include transposase-based amplification reagents such as the transposase of prokaryotic transposon Tn5.

Please replace paragraph [0016] of the specification as filed with the following replacement paragraph [0016]:

> [0016] In further embodiments, the CRISPR system effector protein is an RNA-targeting effector protein. In one aspect, the RNA-targeting effector protein comprises one or more Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) domains ~~(HEPN) domains~~, which can, in some embodiments comprise a RxxxxH motif sequence. In an embodiment, the RxxxH motif comprises a R{N/H/K}X1X2X3H sequence (SEQ ID NO:1). In some embodiments, the R{N/H/K}X1X2X3H sequence is defined wherein X1 is R, S, D, E, Q, N, G, or Y, and X2 is independently I, S, T, V, or L, and X3 is independently L, F, N, Y, V, I, S, D, E, or A.

**In the claims**

Please amend the claims as follows. This listing of claims will replace all prior versions and listings of claims in this application.

**1.** (**Currently Amended**) A nucleic acid detection system for detecting the presence of hemorrhagic fever viruses in a sample comprising:

a CRISPR system, comprising a Cas13 effector protein and one or more guide molecules having a direct repeat and a guide sequence designed to bind to one or more corresponding target molecules of one or more hemorrhagic fever viruses ~~selected from the group consisting of~~ chosen from Lassa virus, ~~Hantavirus, Crimean-Congo hemorrhagic fever virus, Lujo virus~~, Ebola virus, and Marburg virus, ~~and Rift Valley fever virus~~, or a combination thereof,

wherein the one or more guide molecules targeting the Lassa virus are guide RNAs comprising a guide sequence having a nucleotide sequence ~~selected from the group consisting of~~ chosen from SEQ ID NOs: [[80,]] 87-92, 109-126, 139-156, [[159-172,]] 207-228, 249-281, and 329-36, [[and 393-416,]]

wherein the one or more guide molecules targeting the Ebola virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 80, 162, 164, 166, 168, 170 and 172,

wherein the one or more guide molecules targeting the Marburg virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 393-416, and

an RNA-based masking construct.

**2.** (**Canceled**)

**3.** **(Previously presented)** The nucleic acid detection system of claim 1, further comprising nucleic acid amplification reagents, wherein the nucleic acid amplification reagents comprise recombinase polymerase amplification (RPA) reagents, nucleic acid sequence-based amplification (NASBA) reagents, loop-mediated isothermal amplification (LAMP) reagents, strand displacement amplification (SDA) reagents, helicase-dependent

amplification (HDA) reagents, nicking enzyme amplification reaction (NEAR) reagents, RT-PCR reagents, multiple displacement amplification (MDA) reagents, rolling circle amplification (RCA) reagents, ligase chain reaction (LCR) reagents, ramification amplification method (RAM) reagents, transposase based amplification reagents, or Programmable CRISPR Nicking Amplification (PCNA) reagents.

**4.** (**Canceled**)

**5.** **(Currently amended)** The nucleic acid detection system of claim 3, wherein the RPA reagents comprise one or more primer pairs selected from the group consisting of SEQ ID NOs: 78, 79, 81-86, 93-108, 127-138, 173-206, 233-248, 285-328, 370-392, or wherein the transposase-based amplification reagents comprise a transposase of prokaryotic transposon 5 (Tn5).

**6.** (**Canceled**)

**7.** **(Previously presented)** The nucleic acid detection system of claim 1, wherein the CRISPR effector protein is an RNA-targeting effector protein, and wherein the RNA-targeting effector protein comprises one or more higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domains, comprising a RxxxxH motif sequence, comprising a R{N/H/K]$X_1X_2X_3$H sequence, wherein $X_1$ is R, S, D, E, Q, N, G, or Y, and $X_2$ is independently I, S, T, V, or L, and $X_3$ is independently L, F, N, Y, V, I, S, D, E, or A.

**8–11.** (**Canceled**)

**12.** **(Previously presented)** The nucleic acid detection system of claim 1, wherein the Cas 13 effector protein is a C2c2 effector protein is within 20 kilobases (kb) of a Cas 1 gene, wherein the C2c2 effector protein is from an organism of a genus selected from the group consisting of: *Leptotrichia*, *Listeria*, *Corynebacter, Sutterella, Legionella, Treponema, Filifactor, Eubacterium, Streptococcus, Lactobacillus, Mycoplasma, Bacteroides, Flaviivola, Flavobacterium, Sphaerochaeta, Azospirillum, Gluconacetobacter, Neisseria, Roseburia, Parvibaculum, Staphylococcus, Nitratifractor, Mycoplasma, Campylobacter,*

and *Lachnospira*; or the C2c2 effector protein is a *L. wadei* F0279 or *L. wadei* F0279 (Lw2) C2c2 effector protein.

**13–16**. (**Canceled**)

**17. (Previously presented)** The nucleic acid detection system of claim 1, wherein the RNA-based masking construct suppresses generation of a detectable positive signal, and wherein:

the RNA-based masking construct suppresses generation of a detectable positive signal by masking the detectable positive signal, or by generating a detectable negative signal instead;

the RNA-based masking construct comprises a silencing RNA that suppresses generation of a gene product encoded by a reporting construct, wherein the gene product generates the detectable positive signal when expressed;

the RNA-based masking construct is a ribozyme that generates the negative detectable signal, and wherein the positive detectable signal is generated when the ribozyme is deactivated, wherein the ribozyme converts a substrate to a first color and wherein the substrate converts to a second color when the ribozyme is deactivated;

the RNA-based masking construct comprises an RNA oligonucleotide to which a detectable ligand and a masking component are attached;

the RNA-based masking construct comprises a nanoparticle held in aggregate by bridge molecules, wherein at least a portion of the bridge molecules comprises RNA, and wherein the solution undergoes a color shift when the nanoparticle is disbursed in solution;

the RNA-based masking construct comprising a quantum dot linked to one or more quencher molecules by a linking molecule, wherein at least a portion of the linking molecule comprises RNA;

the RNA-based masking construct comprises RNA in complex with an intercalating agent, wherein the intercalating agent changes absorbance upon cleavage of the RNA; or

the detectable ligand is a fluorophore and the masking component is a quencher molecule.

**18–21**. (**Canceled**)

**22.** **(Withdrawn - Previously Presented)** The nucleic acid detection system of claim 17, wherein the RNA-based masking agent is an RNA aptamer and/or comprises an RNA-tethered inhibitor, wherein the aptamer or RNA-tethered inhibitor sequesters an enzyme, wherein the enzyme generates a detectable signal upon release from the aptamer or RNA tethered inhibitor by acting upon a substrate; or

wherein the aptamer is an inhibitory aptamer that inhibits an enzyme and prevents the enzyme from catalyzing generation of a detectable signal from a substrate or wherein the RNA-tethered inhibitor inhibits an enzyme and prevents the enzyme from catalyzing generation of a detectable signal from a substrate, and wherein, wherein the enzyme is thrombin, protein C, neutrophil elastase, subtilisin, horseradish peroxidase, beta-galactosidase, or calf alkaline phosphatase, and, wherein, the enzyme is thrombin and the substrate is para-nitroanilide covalently linked to a peptide substrate for thrombin, or 7-amino-4-methylcoumarin covalently linked to a peptide substrate for thrombin;

or wherein the aptamer sequesters a pair of agents that when released from the aptamers combine to generate a detectable signal.

**23–28**. (**Canceled**)

**29.** (**Withdrawn - Previously presented**) The nucleic acid detection system of claim 17, wherein the RNA-based masking construct comprises a nanoparticle held in aggregate by bridge molecules, wherein at least a portion of the bridge molecules comprises RNA, and wherein the solution undergoes a color shift when the nanoparticle is disbursed in solution, wherein the nanoparticle is a colloidal metal, colloidal gold.

**30–32**. (**Canceled**)

**33.** **(Withdrawn - Previously presented)** The nucleic acid detection system of claim 17, wherein the RNA-based masking construct comprises RNA in complex with an intercalating agent, wherein the intercalating agent changes absorbance upon cleavage of the RNA, wherein the intercalating agent is pyronine-Y or methylene blue.

**34–35**. (**Canceled**)

**36.** **(Currently Amended)** The nucleic acid detection system of claim 1, comprising two or more CRISPR systems, each CRISPR system comprising

a~~n~~ Cas13 effector protein and one or more guide molecules designed to bind to one or more corresponding target molecules of one or more hemorrhagic fever viruses; and

a set of RNA-based masking constructs;

wherein each RNA-based masking construct comprises a cutting motif sequence that is preferentially cut by one of the CRISPR effector proteins after the CRISPR effector protein is activated.

**37.** (**Withdrawn**) A method for detecting viral nucleic acid in one or more samples, comprising:

contacting one or more samples with a nucleic acid detection system according to claim 1; and applying said contacted one or more samples to a lateral flow immunochromatographic assay.

**38.** (**Withdrawn - Currently Amended**) A method for detecting viral nucleic acid in a sample comprising:

amplifying the sample nucleic acid;

combining the sample with an Cas13 effector protein, one or more guide molecules designed to bind to corresponding virus specific target molecules ~~according to SEQ ID NOs: 80, 87-92, 109-126, 139-156, 159-172, 207-228, 249-281, 329-366, and 393-416~~, and an RNA-based masking construct,

wherein the one or more guide molecules targeting the Lassa virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 87-92, 109-126, 139-156, 207-228, 249-281, and 329-36,

wherein the one or more guide molecules targeting the Ebola virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 80, 162, 164, 166, 168, 170 and 172,

wherein the one or more guide molecules targeting the Marburg virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 393-416,

wherein each guide RNA further comprises a direct repeat sequence capable of binding to the Cas13 effector protein,

~~wherein the one or more guide molecules are designed to bind to corresponding virus specific target molecules;~~

activating the Cas13 effector protein via binding of the one or more guide molecules to the one or more virus-specific target molecules, wherein activating the Cas13 effector protein results in modification of the RNA-based masking construct such that a detectable positive signal is produced; and

detecting the signal, wherein detection of the signal indicates the presence ~~of a hemorrhagic fever virus selected from the group consisting of~~ Lassa virus, ~~Hantavirus, Crimean-Congo hemorrhagic fever virus~~, ~~Lujo virus~~, Ebola virus, or Marburg virus, ~~and Rift Valley fever virus,~~ or a combination thereof; wherein the method does not include ~~the~~ a step of extracting nucleic acid from the sample, and wherein:

the step of amplifying the sample nucleic acid comprises nucleic acid sequence-based amplification (NASBA), recombinase polymerase amplification (RPA), loop-mediated isothermal amplification (LAMP), strand displacement amplification (SDA), helicase-dependent amplification (HDA), nicking enzyme amplification reaction (NEAR), RT-PCR, multiple displacement amplification (MDA), rolling circle amplification (RCA), ligase chain reaction (LCR), ramification amplification method (RAM), transposase based amplification, or Programmable CRISPR Nicking Amplification (PCNA).[[;]]

wherein the sample is a biological sample comprising blood, plasma, serum, urine, or saliva.

**39.** (**Canceled**)

**40.** (**Withdrawn**) The method of claim 38, wherein amplifying the sample nucleic acid comprises contacting the sample with one or more of the probes according to SEQ ID NOs: 78, 79, 81-86, 93-108, 127-138, 173-206, 233-248, 285-328, 370-392.

**41.** (**Canceled**)

**42.** (**Withdrawn - Previously presented)** The method of claim 38, further comprising the step of applying the sample to one or more lateral flow strips, wherein the lateral flow strip comprises an upstream first antibody directed against a first molecule, and a downstream second antibody directed against a second molecule, and wherein uncleaved RNA-based masking construct is bound by said first antibody if the target nucleic acid is not present in said sample, and wherein cleaved RNA-based masking construct is bound both by said first antibody and said second antibody if the target nucleic acid is present in said sample.

**43.** (**Canceled**)

**44.** (**Previously Presented**) The system of claim 1, wherein the masking construct comprises an RNA oligonucleotide designed to bind a G-quadruplex forming sequence, wherein a G-quadruplex structure is formed by the G-quadruplex forming sequence upon cleavage of the masking construct, and wherein the G-quadruplex structure generates a detectable positive signal.

**45.** (**Withdrawn**) The method of claim 38, further comprising comparing the detectable positive signal with a (synthetic) standard signal.

**46.** **(Currently amended)** The system of claim 1,

wherein the one or more guide molecules comprise guide RNAs for two or more hemorrhagic fever viruses or [[;]]

~~the one or more guide molecules comprise guide RNAs for~~ two or more strains of a hemorrhagic fever virus.[[;]] ~~orwhen the hemorrhagic fever virus of interest is Lassa virus, the one or more guide molecules are guide RNAs selected from the group consisting of SEQ ID NO: 87-92, 109-126, 139-156, 207-228, 249-281, and 329-36;when the hemorrhagic fever virus of interest is Ebola virus, the one or more guide molecules are guide RNAs selected from the group consisting of SEQ ID NO: 80, and 159-172; andwhen the hemorrhagic fever virus of interest is Marburg virus, one or more guide molecules are guide RNAs selected from the group consisting of SEQ ID NO: 393-416.~~

**47.** (**Canceled**)

**48.** (**Withdrawn - Previously Presented**) The method of claim 38, wherein the Lassa virus is SL-IV, N-II, or N-III.

**49.** (**Canceled**)

**50.** **(Withdrawn - Currently amended)** The method of claim ~~47~~38, wherein the direct repeat has a nucleotide sequence of SEQ ID NO: 158.

~~when the hemorrhagic fever virus of interest is Lassa virus, the one or more guide molecules are guide RNAs selected from the group consisting of SEQ ID NO: 87-92, 109-126, 139-156, 207-228, 249-281, 329-36;when the hemorrhagic fever virus of interest is Ebola virus, the one or more guide molecules are guide RNAs selected from the group consisting of SEQ ID NO: 80, 159-172;when the hemorrhagic fever virus of interest is Marburg virus, one or more guide molecules are guide RNAs selected from the group consisting of SEQ ID NO: 393-416.~~

**51.** (**Withdrawn**) The method of claim 38 for distinguishing between two or more hemorrhagic viruses or two or more strains of a hemorrhagic virus, wherein the one or more guide molecules comprise guide RNAs for the two or more hemorrhagic viruses or the two or more strains of a hemorrhagic virus.

**52.** (**Canceled**)

**53.** (**Withdrawn - Currently Amended**) A kit for detecting viral nucleic acids in a sample, comprising

nucleic acid amplification reagents;

a CRISPR system comprising a Cas13 effector protein and one or more of the guide RNAs designed to bind to corresponding target molecules of one or more hemorrhagic fever viruses,

wherein the one or more guide molecules targeting the Lassa virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 87-92, 109-126, 139-156, 207-228, 249-281, and 329-36,

wherein the one or more guide molecules targeting the Ebola virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 80, 162, 164, 166, 168, 170 and 172,

wherein the one or more guide molecules targeting the Marburg virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 393-416,

wherein each guide RNA further comprises a direct repeat sequence capable of binding to the Cas13 effector protein,

~~according to SEQ ID NO: 80, 87-92, 109-126, 139-156, 159-172, 207-228, 249-281, 329-366, and 393-416, wherein the guide RNAs are designed to bind to corresponding target molecules of one or more hemorrhagic fever viruses selected from the group consisting of Lassa virus, Hantavirus, Crimean-Congo hemorrhagic fever virus, Lujo virus, Ebola virus, Marburg virus, and Rift Valley fever virus, or a combination thereof;~~

an RNA-based masking construct;

one or more lateral flow strips; and,

one or more of the probes according to SEQ ID NO: 78, 79, 81-86, 93-108, 127-138, 173-206, 233-248, 285-328, 370-392.

**54.** (**Canceled**)

**55.** (**Withdrawn - Currently Amended**) A diagnostic device comprising one or more individual discrete volumes, each individual discrete volume comprising ~~a~~ the nucleic acid detection ~~CRISPR~~ system of claim 1, wherein:

each individual discrete volume further comprises one or more detection aptamers comprising a masked RNA polymerase promoter binding site or a masked primer binding site;

each individual discrete volume further comprises nucleic acid amplification reagents;

the target molecule is a target RNA and the individual discrete volumes further comprise a primer that binds the target RNA and comprises an RNA polymerase promoter.

**56–58**. (**Canceled**)

**59.** (**Withdrawn - Previously presented)** The device of claim 55, wherein the individual discrete volumes are droplets, or wherein the individual discrete volumes are defined on a solid substrate, wherein the individual discrete volumes are microwells, or are spots define on the substrate; wherein the solid substrate is a flexible materials substrate, and a paper or flexibly polymer-based substrate.

**60–64**. (**Canceled**)

**65.** (**Currently Amended**) The system of claim 1, further comprising a~~n~~ target molecule-specific enrichment CRISPR system, wherein the target molecule-specific enrichment CRISPR system is designed to bind and sequester the corresponding target molecules prior to detection by the detection CRISPR system, and

wherein the target molecule-specific enrichment CRISPR system comprises a catalytically inactive CRISPR effector protein;

wherein the catalytically inactive CRISPR effector protein is a catalytically inactive C2c2; or

wherein the target molecule-specific enrichment CRISPR effector protein comprises a tag, wherein the tag is used to pull down the target molecule-specific enrichment CRISPR effector system, or to bind the target molecule-specific enrichment CRISPR system to a solid substrate.

**66-68**. (**Canceled**)

**69.** **(Previously presented)** The nucleic acid detection system of claim 1, wherein the Cas13 effector protein is a C2c2 effector protein or a Cas13b effector protein from an organism selected from the group consisting of: *Leptotrichia shahii; Leptotrichia wadei* (Lw2); *Listeria seeligeri; Lachnospiraceae bacterium* MA2020; *Lachnospiraceae bacterium* NK4A179; [*Clostridium] aminophilum* DSM 10710; *Carnobacterium gallinarum* DSM 4847; *Carnobacterium gallinarum* DSM 4847 (second CRISPR Loci); *Paludibacter propionicigenes* WB4; *Listeria weihenstephanensis* FSL R9-

0317; *Listeriaceae bacterium* FSL M6-0635; *Leptotrichia wadei* F0279; *Rhodobacter capsulatus* SB 1003; *Rhodobacter capsulatus* R121; *Rhodobacter capsulatus DE442*; *Leptotrichia buccalis* C-1013-b; *Herbinix hemicellulosilytica; [Eubacterium] rectale; Eubacteriaceae bacterium* CHKCI004; *Blautia* sp. Marseille-P2398; *Leptotrichia* sp. oral taxon 879 str. F0557; *Lachnospiraceae bacterium* NK4A144; *Chloroflexus aggregans; Demequina aurantiaca; Thalassospira* sp. TSL5-1; *Pseudobutyrivibrio* sp. OR37; *Butyrivibrio* sp. YAB3001; *Blautia* sp. Marseille-P2398; *Leptotrichia* sp. Marseille-P3007; *Bacteroides ihuae; Porphyromonadaceae bacterium* KH3CP3RA; *Listeria riparia;* and *Insolitispirillum peregrinum.*

**70.** **(New)** The nucleic acid detection system of claim 1, wherein the direct repeat sequence has a nucleotide sequence of SEQ ID NO: 158.

# REMARKS

## Status of the Claims

Claims 1, 3, 5, 7, 12, 17, 36–38, 40, 44–46, 50–51, 53, 55, 59, 65, and 69-70 are pending, with claims 1, 38, and 53 being independent. Claims 2, 4, 6, 8–11, 13–16, 18–21, 23–28, 30–32, 34–35, 39, 41, 43, 47, 49, 52, 54, 56–58, 60–64, and 66–68 are canceled without prejudice or disclaimer as to the subject matter recited therein. Claims 22, 29, 33, 37, 38, 40, 42, 45, 47-48, 50, 51, 53, 55, and 59 are withdrawn pursuant to the Restriction Requirement of February 2, 2024, and the Non-Final Office Action dated May 20, 2024. Claims 1, 5, 36, 38, 46, 50, 53, 55, 59, and 65 are amended. Claim 70 is new. Accordingly, upon entry of this amendment, claims 1, 3, 5, 7, 12, 17, 36, 44, 46, 65, 69 and 70 are under examination on the merits.

Claim 1 is amended to specify the one or more hemorrhagic fever viruses are chosen from Lassa virus, Ebola virus, and Marburg virus, or a combination thereof. Support for this amendment can be found throughout the as-filed application and at least, for example, in paragraphs [0028], [0061], [0158], and [0255] of the corresponding published patent application US20210396756.

Claim 1 is further amended to specify the one or more guide molecules targeting the Lassa virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 87-92, 109-126, 139-156, 207-228, 249-281, and 329-36, the one or more guide molecules targeting the Ebola virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 80, 162, 164, 166, 168, 170 and 172, and the one or more guide molecules targeting the Marburg virus are guide RNAs comprising a guide sequence having a nucleotide sequence chosen from SEQ ID NOs: 393-416. Examples of written description support for this amendment can be found throughout the as-filed application and at least, for example, in paragraph [0061] of the corresponding published patent application US20210396756.

Support for new claim 70 can be found throughout the as-filed application and at least, for example, in paragraph [0325], TABLE 7 of the corresponding published U.S. Patent Application No. 2021/0396756. Appendix I of this response also shows the herein amended Table 10b that is further annotated to show the location of the direct repeat in each of the guide RNAs.

Withdrawn claims 38 and 53 are amended similarly. No new matter has been added.

The foregoing claim amendments are made solely to expedite the prosecution of the present application. Applicants reserve the right to pursue the subject matter of the present claims before being amended in this application or another related application. Entry and consideration of these amendments are respectfully requested.

**Specification**

The specification is objected to as failing to provide a proper antecedent basis for claims 5 and 7. In claim 5, the Examiner alleges "prokaryotic transposon 5" of claim 5 does not have an antecedent basis in the specification. On review, paragraph [0014] of the specification as filed states:

> In other embodiments, the nucleic acid amplification reagents include transposase-based amplification reagents such as Tn5.

A person of ordinary skill in the art at the time of Applicant's filing would understand that Tn5 refers to bacterial transposon 5 that encodes Tn5 transposase.[1] Solely to improve clarity, claim 5 is amended to require "wherein the transposase-based amplification reagents comprise a transposase of prokaryotic transposon 5 (Tn5)." In accordance with 37 CFR 1.75(d)(1), the last sentence of paragraph [0014] is amended as follows:

> In other embodiments, the nucleic acid amplification reagents include transposase-based amplification reagents such as the transposase of prokaryotic transposon Tn5.

No new matter is added.

In claim 7, the Examiner also contends the term "higher eukaryotes and prokaryotes nucleotide-binding" does not have an antecedent basis in the specification. However, a person of ordinary skill in the art at the time of Applicant's filing would understand HEPN to be an acronym for "Higher Eukaryotes and Prokaryotes Nucleotide-binding" [2] In accordance with 37 CFR 1.75(d)(1), paragraph [0016] is amended as follows:

[1] For example, the abstract of the 1993 review by Reznikoff WS (Annu Rev Microbiol. 1993;47:945-63) entitled 'The Tn5 transposon' states *inter alia*: "The bacterial transposon Tn5 encodes two proteins, the transposase and a related protein, the transposition inhibitor, whose relative abundance determines, in part, the frequency of Tn5 transposition."

[2] For example, the abstract of Applicant's 2015 article entitled "C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector" (Abudayyeh et al. Science. 2016 Aug 5;353(6299) states *inter alia*: "Cleavage is mediated by catalytic residues in the two conserved ***Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) domains***, mutations of which generate catalytically inactive RNA-binding proteins."

> In one aspect, the RNA-targeting effector protein comprises one or more Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) domains ~~(HEPN) domains~~, which can, in some embodiments comprise a RxxxxH motif sequence.

No new matter is added. Withdrawal of the objections is respectfully requested.

**Improper Markush Group**

Claims 1, 3, 5, 7, 12, 17, 36, 44, 46, and 65 are rejected for allegedly containing an improper Markush grouping. The Office alleges the claims contain an improper Markush group.

Specifically, the Examiner contends;

> ". . . the nucleic acid guide sequence and primers recited in the instant claims, and the methods which detect them, ***do not share a single structural similarity since each consists of a different nucleic acid guide sequence***. . . The fact that the nucleic acid guide sequence or primer comprise nucleotides *per se* does not support a conclusion that they have a common single structural similarity because the structure of comprising a nucleotide alone is not essential to the common activity of being correlated hemorrhagic fever virus. . . . If the instantly claimed nucleic acid guide sequence are placed in a group with an equal number of nucleic acid guide sequences or primers the skilled artisan could not differentiate those associated with a phenotype from those that are not associated with a phenotype. Thus[,] the one of skill in the art could not identify those nucleic acid guide sequence[s] that are asserted to be associated with the phenotype by their very nature. Thus[,] the instant claims have not met the requirements of a proper Markush group. (Emphasis added)

The Applicant traverses the rejection for the following reasons.

According to the MPEP [3], a Markush claim contains an "improper Markush grouping" if either:

(1) the members of the Markush group do not share a "single structural similarity" or
(2) the members do not share a common use.[4]

MPEP 2117 (II)(A), entitled: "Single Structural Similarity" - Members of a Physical, Chemical, or Art-Recognized Class, further states:

(Emphasis added)
[3] MPEP 2117 (II)
[4] Supplementary Examination Guidelines for Determining Compliance With 35 U.S.C. 112, and for Treatment of Related Issued in Patent Applications, 76 FR 7162, 7166 (2011) (citing In re Harnisch, 631 F.2d 716, 721-22, 206 USPQ 300, 305 (CCPA 1980))

> Members of a Markush group share a "single structural similarity" when they belong to the same recognized physical or chemical class or to the same art-recognized class. A recognized physical class, a recognized chemical class, or an art-recognized class is a class wherein there is an expectation from ***the knowledge in the art that members of the class will behave in the same way in the context of the claimed invention***. . . . Thus, a Markush grouping is ordinarily proper if all the members of the group belong to a recognized class (whether physical, chemical, or art recognized) and are ***disclosed in the specification to possess at least one property in common which is mainly responsible for their function in the claimed invention***, and it is clear from their very nature or from the prior art that all members possess this property. (Emphasis added)

Thus, a recognized physical, chemical class or art-recognized class is not evaluated in a vacuum but ***in the context*** of the claimed invention, the specification and the knowledge in the art.

In the response to the Non-Final Office Action filed on August 20, 2024, the Applicant sought to illustrate this point by citing the PTAB decision in *Ex parte Narva* (Appeal 2018-006168, decided 04-15-2019), as persuasive evidence the Markush groupings found in instant claims 1, 38, and 53 are indeed proper. In the Office's response, the Examiner dismissed consideration of the *Ex parte Narva* PTAB decision merely because the ruling was not precedential. Office Action at pages 4-5.

Regardless of precedential status, PTAB decisions remain informative on how likely the PTAB is to uphold or overturn an Improper Markush rejection under certain circumstances. The Applicant maintains the Examiner would be wise to consider the holding of *Ex parte Narva* which is directly applicable here.

Briefly, Narva's patent application discloses double-stranded RNA molecules for the post-transcriptional RNA interference of genes encoding Ras-opposite proteins ("ROP") in different coleopteran insects.

The rejected independent claim 1 required:

> 1. ***A double-stranded ribonucleic acid (dsRNA) molecule*** comprising a first polyribonucleotide consisting of at least 23 contiguous nucleotides of the *polyribonucleotide encoded by a polynucleotide selected from the group consisting of SEQ ID NO:1, SEQ ID NO:115, SEQ ID NO:120, SEQ ID NO:122, SEQ ID NO:124, SEQ ID NO:126, SEQ ID NO: 131, and SEQ ID NO: 133,*

> wherein *the* ***first polyribonucleotide is hybridized in the dsRNA molecule*** to a second polyribonucleotide that is the complement or reverse complement of the first polyribonucleotide, and
>
> wherein delivery of the dsRNA molecule ***inhibits the expression of a target gene in a coleopteran or hemipteran insect*** selected from the group consisting of Diabrotica virgifera, Euschistus heros, and Meligethes aeneus.

The Board held that the claimed Markush group is not improper, as the claimed species share a single structural similarity and a common use, stating:

> [t]he nucleic acid sequences recited in the rejected claims belong to the same ***recognized chemical class of polyribonucleotides*** that are hybridized in dsRNA molecules and encode ROP proteins. *While the individual sequences differ because they are drawn to ROP sequences of different insects ( e.g., SEQ ID NO: 1 and 115) or different portions of the ROP sequence (e.g., SEQ ID N0:120, 131, and 133),* ***all of the sequences share the common use of silencing ROP proteins.***
> (Emphasis added)

A review of PTAB decisions on the propriety of Improper Markush Grouping rejections of listings of biomarkers, nucleotide, or amino acid sequences reveals the Board has repeatedly held that to the extent an Examiner applie[s] a bright-line rule that focused solely on the sequence differences without consideration of other structural similarities, that approach is **incorrect**.[5,6]

For example, in *Ex parte Richards* (Appeal 2021001219, decided 11-22-21), the PTAB overturned an Improper Markush Group rejection because the claimed nucleotide sequences, while differing in sequence, were described in the specification as capable of being used to capture complementary sequences comprising a causal genetic variant or a sequence within 200 nucleotides of a causal genetic variant and thus shared a unifying structural feature and share a common function. Similarly, the present specification shares the common features of being able to bind target sequences in their corresponding target viruses.

For example, in *Ex parte Ward* (2021003776, decided June 10, 2022), the PTAB overturned an Improper Markush Group rejection, holding that "a Markush grouping is ordinarily proper if all the members of the group belong to a recognized class (whether physical, chemical,

[5] See Harnish, 631 F.2d at 722 ("[l]n determining the propriety of a Markush grouping the compounds must be considered as wholes and not broken down into elements or other components.")
[6] *Ex parte Buyyarapu* (Appeal 2018006665)

or art-recognized) and are disclosed in the specification to possess at least one properly in common which is mainly responsible for their function in the claimed invention, and it is clear from their very nature or from the prior art that all members possess this property." In *Ex parte Ward*, the claimed polypeptides were recognized as belonging to the same chemical class despite comprising different amino acid sequences, and the specification established a shared property that each was correlated with whether a subject had a neurodegenerative disorder or not.

With respect to the instant rejection, claim 1 defines three Markush groups, each comprising guide molecules that specifically bind target sequences in the genome of either the Lassa, Ebola, or Marburg hemorrhagic fever viruses. As stated in paragraph [0064] of the corresponding published patent application US20210396756, each guide RNA or CRISPR RNA (crRNA) comprises a ***direct repeat sequence*** fused or linked to a ***guide or spacer sequence***. The direct repeat sequence is known in the art to be a conserved sequence within the claimed guide RNAs (gRNA) that forms a stem-loop structure that binds tightly to its cognate Cas13 subtype. The ***spacer or guide sequence*** directs the sequence-specific binding of the CRISPR-Cas complex to a target sequence.[7] The application therefore teaches, and claim 1 defines, a set of guide RNA molecules that can bind to Lassa, Ebola, or Marburg target sequences. The claimed guide molecules are part of a recognized class because one of ordinary skill in the art would expect each of these guide molecules to behave the same way ***in the context*** of the claimed invention (i.e., complex with their cognate Cas protein and direct sequence-specific binding of the complex to their corresponding target sequence) based on their knowledge of how guide molecules are structured and function and the explicit teachings of the specification that provide the sequences needed to detect each of the claimed viruses.

The ability to pick an appropriate guide molecule from a line-up proffered by the Examiner ("If the instantly claimed nucleic acid sequence are placed in a group with an equal number of nucleic acid sequence or primers the skilled artisan could not differentiate . . . .." Office Action at 6), is **not** a valid test for assessing the propriety of a Markush group. In taking this approach, the Examiner fails to account for the existing knowledge of guide molecules that one or ordinary skill

[7] Knott, G. J. *et al.* Guide-bound structures of an RNA-targeting A-cleaving CRISPR-Cas13a enzyme. *Nat. Struct. Mol. Biol.* **24**, 825–833 (2017).

in the art would possess and ignores the relevant teachings of the specification establishing which guide molecule sequences will detect which virus.

Thus, not only do the nucleic acid sequences of the claimed guide molecules belong to the same recognized chemical class of polynucleotides, but in the context of the claimed method, all the guide molecules share a unifying structural feature, i.e., a direct repeat and a spacer or guide sequence. The binding of the direct repeat to the Cas13 effector protein is essential for the functionality of the CRISPR-Cas system. All members of the claimed Markush groupings are disclosed in the specification and known to be functionally equivalent in the art.[8] The Markush groupings of the claimed guide RNAs also share the common utility of detecting the presence of hemorrhagic fever viruses in a sample.

In view of the foregoing, the Applicants respectfully submit that the instant claims meet the requirements of a proper Markush group. Reconsideration and withdrawal of the rejection are respectfully requested.

**Claim Rejections under 35 U.S.C. § 112(b)**

Claims 5, 36, and 65 are rejected under 35 U.S.C. § 112(b) as allegedly being indefinite.

***Claim 5***

According to the Examiner, it is unclear if the "prokaryotic transposon 5 (Tn5)" of claim 5 "requires a transposase or a transposition element encompassed by Tn[5]."

As discussed above, a person of ordinary skill in the art would understand Tn5 refers to bacterial transposon 5 that encodes Tn5 transposase.[9] Solely to improve clarity, claim 5 is amended to require "wherein the transposase-based amplification reagents comprise a transposase of prokaryotic transposon 5 (Tn5)." No new matter is added.

[8] *Supplementary Examination Guidelines for Determining Compliance with 35 U.S.C. 112 and for Treatment of Related Issues in Patent Applications* (*"Supplementary Guidelines"*), 76 Fed. Reg. 7162 at 7166 (February 9, 2011) (citing *In re Harnisch,* 631 F.2d 716, 721-22, 206 USPQ 300, 305 (CCPA 1980)) and MPEP § 2117.

[9] For example, the abstract of the 1993 review by Reznikoff WS (Annu Rev Microbiol. 1993;47:945-63) entitled 'The Tn5 transposon' states *inter alia*: "The bacterial transposon Tn5 encodes two proteins, the transposase and a related protein, the transposition inhibitor, whose relative abundance determines, in part, the frequency of Tn5 transposition."

### *Claim 36*

In claim 36, which depends on claim 1, the Examiner alleges the claim is indefinite because of the term "an effector protein." The rejection is rendered moot by the instant amendment of claim 36, which requires "a Cas13 effector protein."

### *Claim 65*

According to the Examiner, the recitation of enrichment in claim 65 inherently suggests some are not enriched. The Examiner also alleges the specification and claims do not define how to differentiate enrichment from non-enrichment. Claim 65 is amended herein to require ***a target molecule-specific*** enrichment CRISPR system designed to bind and sequester the corresponding target molecules before detection by the detection CRISPR system. Non-enrichment, therefore, refers to the absence of binding of molecules that are not target molecules.

No new matter is added by these amendments. Withdrawal of the rejections is respectfully requested.

**Claim Rejections Under 35 U.S.C. § 103**

Claims 1, 3, 5, 7, 12, 17, 36, 46, 50, and 69 are rejected under 35 U.S.C § 103 as allegedly being unpatentable over Kim (Current Opinion in Biomedical Engineering 2017, 4:106-115), Rose (Proc. Natl. Acad. Sci. USA Vol. 93, pp. 14998-15000, December 1996), and Genbank KM821910 Lassa virus strain G3283-SLE-2013 segment S glycoprotein precursor (GPC) and nucleoprotein (NP) genes, complete cds (https://www.ncbi.nlm .nih.gov/nuccore/km821910, 1/8/2015).

The Examiner alleges *inter alia* that the cited prior art teaches Cas13a, diagnostic assays, negative-strand RNA viruses, including Lassa, Marburg, and Ebola viruses, and that designing oligonucleotides to hybridize to specific targets is merely routine experimentation. The Examiner then concludes it would have been *prima facie* obvious to one of ordinary skill in the art to design guide strands to detect Lassa virus using sequences comprising SEQ ID NO 86, SEQ ID NO 81, and SEQ ID NO 87 using CRISPR as part of the nucleic acid detection system. According to the Examiner, the artisan would be motivated as detecting Lassa would allow for proper treatment of the hemorrhagic disease caused by the RNA virus. The Examiner further alleges the artisan would be motivated to use a nucleic acid detection system to detect RNA virus as the claims suggest the

detection of hemorrhagic virus, and the artisan would have a reasonable expectation of success as the claims suggest detection of virus by a nucleic acid detection system.

The Applicant respectfully traverses the rejection for the following reasons.

The Examiner assumes that designing guide sequences comprising '**any**' Lassa virus genomic sequence, like primers or probes, is sufficient to detect Lassa virus based solely on the 3271-nucleotide genomic sequence disclosed in Genbank KM821910.

On the contrary, as discussed in paragraph [0008] [10], a person of ordinary skill in the art at the time of Applicant's filing would have known:

> **[0008]** The extreme genetic diversity of the LASV genome hinders development of accurate diagnostic tools. *Efforts to create a single diagnostic test that can detect all viral diversity are complicated by* ***deeply divergent LASV clades, which contain high levels of nucleotide variation and are almost genetically distinct*** (Andersen et al., 2015). LASV's fast rate of viral evolution quickly renders diagnostics obsolete, so diagnostic platforms must be easily adaptable to emerging viral strains in order to maintain clinical relevance (Andersen et al., 2015). (Emphasis added)

Paragraph [320] of the application further states:

> **[0320]** . . . *Given the large amount of genetic divergence between LASV strains, it is* ***impossible*** *to find a region of the genome long enough to bind a crRNA that is perfectly conserved across all strains.* ***crRNAs require high specificity to bind to and cut their target sequence and do not tolerate high numbers of base-pair mismatches in their target sequences*** (Abudayyeh et al., 2016; Gootenberg et al., 2017) . . . .. (Emphasis added)

Thus, a person of ordinary skill in the art at the time of Applicant's filing would have known that the substantial ***genetic divergence between LASV strains*** would preclude incorporating ***"any"*** guide sequence into a crRNA to detect Lassa virus. Contrary to the Examiner's assertion, the artisan would ***not*** have had a reasonable expectation of success because the results of such approach would have been entirely unpredictable ***if not "impossible"*** to achieve based solely on the genomic sequence reported in Genbank KM821910 Lassa virus strain G3283-SLE-2013 because "crRNAs require high specificity to bind to and cut their target sequence and ***do not tolerate high numbers of base-pair mismatches in their target sequences***." Furthermore,

[10] of the corresponding published patent application US20210396756

an artisan reading reports of this genetic diversity in, for example, Andersen et al., 2015,[11] would not have been motivated to attempt to detect the Lassa virus by merely selecting ***"any"*** guide sequence from the Lassa virus genomic sequence disclosed in Genbank KM821910 Lassa virus strain G3283-SLE-2013 as the Examiner repeatedly alleges. Without the requisite expectation of success or motivation to detect the Lassa virus using conventional methods, the claimed invention cannot be obvious in view of the references cited by the Examiner.

Withdrawal of the rejection is respectfully requested.

**Claim Rejections under Double Patenting**

Claims 1, 3, 5, 7, 12, 17, 36, 44, 46, 65 and 69 are provisionally rejected on the grounds of non-statutory obviousness-type double patenting as allegedly being unpatentable over

1) claims 1, 2, 57–59, 61, 64, 66, 69, and 71 of co-pending application no. 18/023,538 ("the '538 Application")
2) the claims of co-pending application no. 16/961,820 ("the '820 Application");
3) the claims of co-pending application no. 16/955,380 ("the '380 Application");
4) the claims of co-pending application no. 18/430,788 ("the '788 Application");
5) the claims of co-pending application no. 16/494,279 ("the '279 Application");
6) the claims of co-pending application no. 16/629,310 ("the '310 Application");
7) claims 1-49 of U.S. Patent No 11,898,142 ("the '142 patent"); and
8) claims 1-35 of U.S. Patent No 11,633,732 ("the '732 patent")

each in combination with Genbank KM821910 Lassa virus strain G3283-SLE-2013 segment S glycoprotein precursor (GPC) and nucleoprotein (NP) genes, complete cds (https://www.ncbi.nlm.nih.govinuccore/km821910, 1/8/2015).

For each rejection, the Examiner reiterates a similar rationale:

> Designing oligonucleotides to hybridize to specific targets, which are equivalents to those taught in the art is routine experimentation. . . . The prior art is replete with guidance and information necessary to permit the ordinary artisan in the field of nucleic acid detection to design primer and probes. As discussed above, the ordinary artisan would be motivated to have designed and tested new primers to

[11] Andersen, K. G. *et al.* Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell* **162**, 738–750 (2015).

> obtain additional oligonucleotides that function to detect Lass[a] virus and identify oligonucleotides with improved properties. Thus, for the reasons provided above, the ordinary artisan would have designed additional oligonucleotides using the teachings in the art at the time the invention was made. The claimed SEQ ID NO[s] are obvious over the cited prior art, absent secondary considerations Therefore it would have been facie obvious to one of ordinary skill in the art to design guide strands to detect Lassa virus using sequences comprising SEQ ID NO 86, SEQ ID NO 81 and SEQ ID NO 87 using CRISPR as part of the nucleic acid detection system. The artisan would be motivated as the detection of Lassa would allow for proper treatment of the disease of the hemorrhagic disease caused by RNA virus. The artisan would be motivated to use nucleic acid detection system to detect RNA virus as the claims suggest the detection of hemorrhagic virus. The artisan would have a reasonable expectation of success as claims of [. . . .] [the co-pending applications] suggest detection of virus were known as was the nucleic acid detection system. Thus[,] it would have been prima facie obvious to one of ordinary skill in the art prior to the effective filing date of the claims[,] the claims of [. . . . .] [the co-pending applications] are a species of the instant claims and thus the instant claims are obvious variants. Dependent claims are obvious as they are coextensive in scope. (Emphasis added)

Applicants respectfully traverse for the following reasons.

The Examiner concedes none of the claims of the ’538 application, the ’820 application, the ’380 application, the ’788 application, the ’279 application, the ’310 application, the ’142 patent, and the ’732 patent specifically teach SEQ ID NO 86, SEQ ID NO 81 or SEQ ID NO 87.

Paragraph [320] of the application states *inter alia*:

> **[0320]** . . . Given the large amount of genetic divergence between LASV strains, it is ***impossible*** to find a region of the genome long enough to bind a crRNA that is perfectly conserved across all strains. crRNAs require high specificity to bind to and cut their target sequence and ***do not tolerate high numbers of base-pair mismatches in their target sequences*** (Abudayyeh et al., 2016; Gootenberg et al., 2017). . .. (Emphasis added)

Thus, a person of ordinary skill in the art at the time of Applicant’s filing would have known that the substantial ***genetic divergence between LASV strains*** would preclude incorporating ***“any”*** guide sequence into a crRNA to detect Lassa virus. Contrary to the Examiner’s assertions, the artisan would ***not*** have had a reasonable expectation of success because the results of such an approach would have been entirely unpredictable ***if not “impossible”*** to

achieve based solely on the genomic sequence reported in Genbank KM821910 Lassa virus strain G3283-SLE-2013 because "crRNAs require high specificity to bind to and cut their target sequence and ***do not tolerate high numbers of base-pair mismatches in their target sequences***." Furthermore, an artisan reading reports of this genetic diversity in, for example, Andersen et al., 2015,[12] would not have been motivated to attempt to detect Lassa virus by merely selecting ***"any"*** guide sequence from the Lassa virus genomic sequence disclosed in Genbank KM821910 Lassa virus strain G3283-SLE-2013 as the Examiner repeatedly alleges. Without the requisite expectation of success or motivation to detect Lassa virus using conventional methods, the claimed invention cannot be obvious over any of the claims of the patents and co-pending patent applications cited by the Examiner. Withdrawal of the rejection is respectfully requested.

**Request for Rejoinder of Withdrawn Claims**

Withdrawn independent claims 38 and 53 include features corresponding to independent claim 1. Accordingly, Applicants respectfully submit that independent claims 38 and 53 are allowable for reasons similar to independent claim 1. Upon allowance of independent claim 1, Applicants request rejoinder and allowance of independent claims 38 and 53.

[12] Andersen, K. G. *et al.* Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell* **162**, 738–750 (2015).

## CONCLUSION

Applicant submits the foregoing as a complete response to the Office Action. Applicant submits that this Amendment and Response places the application in condition for allowance and respectfully requests such action. If any issues can be resolved with an Examiner's Amendment or a telephone conference, please contact Applicant's representative at (478) 292 5119.

No additional fees are believed due for consideration of this paper. However, the Commissioner is hereby authorized to charge any fee deemed necessary for consideration of this paper, including any extension of time fees or additional claim fees, or to credit any overpayment, to Deposit Account Number 50-5193, reference number BROD-2545US.

**JOHNSON, MARCOU, ISAACS & NIX, LLC**
PO Box 691
Hoschton, GA 30548
Tel: 912.257.4864
**198696**
CUSTOMER NUMBER
docketing@jmin.com

Respectfully submitted,

/Christopher D. Southgate/

Attorney for Applicant

Christopher D. Southgate, Ph.D., Esq.
Reg. No. 54,875

| **APPENDIX I: TABLE 10A** | | |
|---|---|---|
| **Name** | **Ebola Guide RNAs*** | **Guide Sequence** |
| EBOV_ Guide_2 | TTTAACCCAAATAACTTGCACAGTTGAT**GTTTTAGTCCCCTTCGTTTTTGGGGTAGTCTAAATC** CCCTATAGTGAGTCGTATTAATTTC (SEQ ID NO:159) | ATCAACTGTGCAAGTTATTTGGGTTAAA (SEQ ID NO:166) |
| EBOV_ Guide_9 | AGAACACTTGCTGCCATGCCGGAAGAGG**GTTTTAGTCCCCTTCGTTTTTGGGGTAGTCTAAATC** CCCTATAGTGAGTCGTATTAATTTC (SEQ ID NO:161) | CCTCTTCCGGCATGGCAGCAAGTGTTCT (SEQ ID NO:164) |
| EBOV_ Guide_10 | CTATTCCTTCCGAAATTGGTAGTAGGAG**GTTTTAGTCCCCTTCGTTTTTGGGGTAGTCTAAATC** CCCTATAGTGAGTCGTATTAATTTC (SEQ ID NO:163) | CTCCTACTACCAATTTCGGAAGGAATAG (SEQ ID NO:162) |
| EBOV_ Guide_11 | CTTCCGAAATTGGTAGTAGGAGAAAAGG**GTTTTAGTCCCCTTCGTTTTTGGGGTAGTCTAAATC** CCCTATAGTGAGTCGTATTAATTTC (SEQ ID NO:165) | CCTTTTCTCCTACTACCAATTTCGGAAG (SEQ ID NO:160) |
| EBOV_ Guide_12 | CACACTCCCATGTATGATTGAGCAATTC**GTTTTAGTCCCCTTCGTTTTTGGGGTAGTCTAAATC** CCCTATAGTGAGTCGTATTAATTTC (SEQ ID NO:167) | GAATTGCTCAATCATACATGGGAGTGTG (SEQ ID NO:172) |
| EBOV_ Guide_13 | ATTGGAGCCACACTCCCATGTATGATTG**GTTTTAGTCCCCTTCGTTTTTGGGGTAGTCTAAATC** CCCTATAGTGAGTCGTATTAATTTC (SEQ ID NO:169) | CAATCATACATGGGAGTGTGGCTCCAAT (SEQ ID NO:170) |
| EBOV_ Guide_14 | ACAGATCCTGAGGGAATATTCATGGGCA**GTTTTAGTCCCCTTCGTTTTTGGGGTAGTCTAAATC** CCCTATAGTGAGTCGTATTAATTTC (SEQ ID NO:171) | TGCCCATGAATATTCCCTCAGGATCTGT (SEQ ID NO:168) |

*: sequence in bold corresponds to the same direct repeat sequence **GTTTTAGTCCCCTTCGTTTTTGGGGTAGTCTAAATC**, i.e., the reverse complement of SEQ ID NO: 158 (GATTTAGACTACCCCAAAAACGAAGGGGACTAAAAC)

cccTATAGTGAGTCGTATTAatttc corresponds to the reverse complement of T7 promoter_25nt: gaaatTAATACGACTCACTATAggg (SEQ ID NO: 157)

Published in final edited form as:

*Science*. 2016 August 5; 353(6299): aaf5573. doi:10.1126/science.aaf5573.

# C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector

**Omar O. Abudayyeh**[1,2,3,4,*], **Jonathan S. Gootenberg**[2,3,4,5,*], **Silvana Konermann**[2,3,4,*], **Julia Joung**[2,3,4], **Ian M. Slaymaker**[2,3,4], **David B.T. Cox**[1,2,3,4,6], **Sergey Shmakov**[7,8], **Kira S. Makarova**[8], **Ekaterina Semenova**[9], **Leonid Minakhin**[9], **Konstantin Severinov**[7,9,10], **Aviv Regev**[2,6], **Eric S. Lander**[2,5,6], **Eugene V. Koonin**[8,†], and **Feng Zhang**[1,2,3,4,†]

[1]Department of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA

[3]McGovern Institute for Brain Research at MIT, Cambridge, Massachusetts 02139, USA

[4]Departments of Brain and Cognitive Science and Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[5]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

[6]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[7]Skolkovo Institute of Science and Technology, Skolkovo, 143025, Russia

[8]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

[9]Waksman Institute for Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[10]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, 123182, Russia

## Abstract

The CRISPR-Cas adaptive immune system defends microbes against foreign genetic elements via DNA or RNA-DNA interference. We characterize the Class 2 type VI-A CRISPR-Cas effector C2c2 and demonstrate its RNA-guided RNase function. C2c2 from the bacterium *Leptotrichia shahii* provides interference against RNA phage. *In vitro* biochemical analysis show that C2c2 is guided by a single crRNA and can be programmed to cleave ssRNA targets carrying complementary protospacers. In bacteria, C2c2 can be programmed to knock down specific

†Correspondence should be addressed to zhang@broadinstitute.org (F.Z.) and koonin@ncbi.nlm.nih.gov (E.V.K.).
*These authors contributed equally to this work.

Supplementary Materials
Supplementary Materials and Methods
Table S1 – S4
Fig S1 – S24
References (51–53)

mRNAs. Cleavage is mediated by catalytic residues in the two conserved HEPN domains, mutations in which generate catalytically inactive RNA-binding proteins. These results broaden our understanding of CRISPR-Cas systems and suggest that C2c2 can be used to develop new RNA-targeting tools.

Almost all archaea and about half of bacteria possess Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated genes (CRISPR-Cas) adaptive immune systems (1, 2), which protect microbes from viruses and other invading DNA through three steps: (i) adaptation, i.e., insertion of foreign nucleic acid segments (spacers) into the CRISPR array in between pairs of direct repeats (DRs), (ii) transcription and processing of the CRISPR array to produce mature CRISPR RNAs (crRNAs), and (iii) interference, whereby Cas enzymes are guided by the crRNAs to target and cleave cognate sequences in the respective invader genomes (3–5). All CRISPR-Cas systems characterized to date follow these three steps, although the mechanistic implementation and proteins involved in these processes display extensive diversity.

The CRISPR-Cas systems are broadly divided into two classes on the basis of the architecture of the interference module: Class 1 systems rely on multi-subunit protein complexes whereas Class 2 systems utilize single effector proteins (1). Within these two classes, types and subtypes are delineated according to the presence of distinct signature genes, protein sequence conservation, and organization of the respective genomic loci. Class 1 systems include type I, where interference is achieved through assembly of multiple Cas proteins into the Cascade complex, and type III systems, which rely on either the Csm (type III-A/D) or Cmr (Type III-B/C) effector complexes which are distantly related to the Cascade (1, 6–11).

Class 2 CRISPR systems comprise type II, characterized by the single-component effector protein Cas9 (12–17), which contains RuvC and HNH nuclease domains, and type V systems, which utilize single RuvC domain-containing effectors such as Cpf1 (18), C2c1, and C2c3 (19). All functionally characterized systems, to date, have been reported to target DNA, and only the multi-component type III-A and III-B systems additionally target RNA (7, 20–25). However, the putative Class 2 type VI system is characterized by the presence of the single effector protein C2c2, which lacks homology to any known DNA nuclease domain but contains two Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) domains (19). Given that all functionally characterized HEPN domains are RNases (26), there is a possibility that C2c2 functions solely as an RNA-guided RNA-targeting CRISPR effector.

HEPN domains are also found in other Cas proteins. Csm6, a component of type III-A systems, and the homologous protein Csx1, in type III-B systems, each contain a single HEPN domain and have been biochemically characterized as ssRNA-specific endoribonucleases (21, 27, 28). In addition, type III systems contain complexes of other Cas enzymes that bind and cleave ssRNA through acidic residues associated with RNA-recognition motif (RRM) domains. These complexes (Cas10-Csm in type III-A and Cmr in type III-B) carry out RNA-guided co-transcriptional cleavage of mRNA in concert with DNA target cleavage (22, 29, 30). In contrast, the roles of Csm6 and Csx1, which cleave their targets with little specificity, are less clear, although in some cases, RNA cleavage by

Csm6 apparently serves as a second line of defense when DNA targeting fails (21). Additionally Csm6 and Csx1 have to dimerize to form a composite active site (27, 28, 31), but C2c2 contains two HEPN domains, suggesting that it functions as a monomeric endoribonuclease.

As is common with Class 2 systems, type VI systems are simply organized. In particular, the type VI locus in *Leptotrichia shahii* contains Cas1, Cas2, C2c2 and a CRISPR array, which is expressed and processed into mature crRNAs (19). In all CRISPR-Cas systems characterized to date, Cas1 and Cas2 are exclusively involved in spacer acquisition (32–37), suggesting that C2c2 is the sole effector protein which utilizes a crRNA guide to achieve interference, likely targeting RNA.

## Reconstitution of the *L. shahii* C2c2 locus in *Escherichia coli* confers RNA-guided immunity

We explored whether LshC2c2 could confer immunity to MS2 (25), a lytic single-stranded (ss) RNA phage, without DNA intermediates in its life cycle, that infects *E. coli*. We constructed a low-copy plasmid carrying the entire LshC2c2 locus (pLshC2c2) to allow for heterologous reconstitution in *E. coli* (fig. S1A). Because expressed mature crRNAs from the LshC2c2 locus have a maximum spacer length of 28nt (fig. S1A) (19), we tiled all possible 28-nt target sites in the MS2 phage genome (Fig. 1A). This resulted in a library of 3,473 spacer sequences (along with 490 non-targeting guides designed to have a Levenshtein distance of ≥8 with respect to the MS2 and *E. coli* genomes) which we inserted between pLshC2c2 direct repeats (DRs). After transformation in of this construct into *E. coli*, we infected cells with varying dilutions of MS2 ($10^{-1}$, $10^{-3}$, and $10^{-5}$) and analyzed surviving cells to determine the spacer sequences carried by cells that survived the infection. Cells carrying spacers that confer robust interference against MS2 are expected to proliferate faster than those that lack such sequences. Following growth for 16 hours, we identified a number of spacers that were consistently enriched across three independent infection replicas in both the $10^{-1}$ and $10^{-3}$ dilution conditions, suggesting that they enabled interference against MS2. Specifically, 147 and 150 spacers showed >1.25 $\log_2$-fold enrichment in all three replicates for the $10^{-1}$ and $10^{-3}$ phage dilutions, respectively; of these two groups of top enriched spacers, 84 are shared (Figs. 1B, S2A–G, table S1). Additionally, no non-targeting guides were found to be consistently enriched among the three $10^{-1}$, $10^{-3}$, or $10^{-5}$ phage replicates (fig. S2D, G). We also analyzed the flanking regions of protospacers on the MS2 genome corresponding to the enriched spacers and found that spacers with a G immediately flanking the 3' end of the protospacer were less fit relative to all other nucleotides at this position (i.e. A, U, or C), suggesting that the 3′ protospacer flanking site (PFS) affects the efficacy of C2c2-mediated targeting (Figs. 1C, S2E–F, S3). Although the PFS is adjacent to the protospacer target, we chose not to use the commonly used protospacer adjacent motif (PAM) nomenclature as it has come to connote a sequence used in self vs. non-self differentiation (38), which is irrelevant in a RNA-targeting system. It is worth noting that the avoidance of G by C2c2 echo the absence of PAMs applicable to other RNA-targeting CRISPR systems and effector proteins (20, 22, 24, 25, 39, 40).

The fact that only ~5% of crRNAs are enriched may reflect other factors influencing interference activity, such as accessibility of the target site that might be affected by RNA binding proteins or secondary structure. In agreement with this hypothesis, the enriched spacers tend to cluster into regions of strong interference where they are closer to each other than one would expect by random chance (fig. S3F–G).

To validate the interference activity of the enriched spacers, we individually cloned four top-enriched spacers into pLshC2c2 CRISPR arrays and observed a 3- to 4-$\log_{10}$ reduction in plaque formation, consistent with the level of enrichment observed in the screen (Figs 1B, S4). We cloned sixteen guides targeting distinct regions of the MS2 *mat* gene (4 guides per possible single-nucleotide PFS). All 16 crRNAs mediated MS2 interference, although higher levels of resistance were observed for the C, A, and U PFS-targeting guides (Figs. 1D, 1E, S5), indicating that C2c2 can be effectively retargeted in a crRNA-dependent fashion to sites within the MS2 genome.

To further validate the observed PFS preference with an alternate approach, we designed a protospacer site in the pUC19 plasmid at the 5' end of the β-lactamase mRNA, which encodes ampicillin resistance in *E. coli,* flanked by five randomized nucleotides at the 3' end. Significant depletion and enrichment was observed for the LshC2c2 locus (****, $p<0.0001$) compared to the pACYC184 controls (Fig. S6A). Analysis of the depleted PFS sequences confirmed the presence of a PFS preference of H (Fig. S6B).

## C2c2 is a single-effector endoRNase mediating ssRNA cleavage with a single crRNA guide

We purified the LshC2c2 protein (fig. S7) and assayed its ability to cleave an *in vitro* transcribed 173-nt ssRNA target (Figs. 2A, S8) containing a C PFS (ssRNA target 1 with protospacer 14). Mature LshC2c2 crRNAs contain a 28-nt direct repeat (DR) and a 28 nt spacer (fig. S1A) (19). We therefore generated an *in-vitro*-transcribed crRNA with a 28-nt spacer complementary to protospacer 14 on ssRNA target 1. LshC2c2 efficiently cleaved ssRNA in a $Mg^{2+}$- and crRNA-dependent manner (Figs. 2B, S9). We then annealed complementary RNA oligos to regions flanking the crRNA target site. This partially double-stranded RNA substrate was not cleaved by LshC2c2, suggesting it is specific for ssRNA (figs. S10A–B).

We tested the sequence constraints of RNA cleavage by LshC2c2 with additional crRNAs complementary to ssRNA target 1 where protospacer 14 is preceded by each PFS variant. The results of this experiment confirmed the preference for C, A, and U PFSs, with little cleavage activity detected for the G PFS target (Fig. 2C). Additionally, we designed 5 crRNAs for each possible PFS (20 total) across the ssRNA target 1 and evaluated cleavage activity for LshC2c2 paired with each of these crRNAs. As expected, we observed less cleavage activity for G PFS-targeting crRNAs compared to other crRNAs tested (Fig. 2D).

We then generated a dsDNA plasmid library with protospacer 14 flanked by 7 random nucleotides to account for any PFS preference. When incubated with LshC2c2 protein and a crRNA complementary to protospacer 14, no cleavage of the dsDNA plasmid library was

observed (fig. S10C). We also did not observe cleavage when targeting a ssDNA version of ssRNA target 1 (fig. S10D). To rule out co-transcriptional DNA cleavage, which has been observed in type III CRISPR-Cas systems (22), we recapitulated the *E. coli* RNA polymerase co-transcriptional cleavage assay (22) (fig. S11A) expressing ssRNA target 1 from a DNA substrate. This assay of purified LshC2c2 and crRNA targeting ssRNA target 1 did not show any DNA cleavage (fig. S11B). Together, these results indicate that C2c2 cleaves specific ssRNA sites directed by the target complementarity encoded in the crRNA, with a H PFS preference.

## C2c2 cleavage depends on local target sequence and secondary structure

Given that C2c2 did not efficiently cleave dsRNA substrates and that ssRNA can form complex secondary structures, we reasoned that cleavage by C2c2 might be affected by secondary structure of the ssRNA target. Indeed, after tiling ssRNA target 1 with different crRNAs (Fig. 2D), we observed the same cleavage pattern regardless of the crRNA position along the target RNA. This observation suggests that the crRNA-dependent cleavage pattern was determined by features of the target sequence rather than the distance from the binding site. We hypothesized that the LshC2c2-crRNA complex binds the target and cleaves exposed regions of ssRNA within the secondary structure elements, with potential preference for certain nucleotides.

In agreement with this hypothesis, cleavage of three ssRNA targets with different sequences flanking identical 28-nt protospacers resulted in three distinct patterns of cleavage (Fig. 3A). RNA-sequencing of the cleavage products for the three targets revealed that cleavage sites mainly localized to uracil-rich regions of ssRNA or ssRNA-dsRNA junctions within the *in silico* predicted co-folds of the target sequence with the crRNA (Figs. 3B–C, S12A–D). To test whether the LshC2c2-crRNA complex prefers cleavage at uracils, we analyzed the cleavage efficiencies of homopolymeric RNA targets (a 28-nt protospacer extended with 120 As or Us regularly interspaced by single bases of G or C to enable oligo synthesis) and found that LshC2c2 preferentially cleaved the uracil target compared to adenine (figs. S12E, S12F). We then tested cleavage of a modified version of ssRNA 4 which had its main site of cleavage, a loop, replaced with each of the four possible homopolymers and found that cleavage only occurred at the uracil homopolymer loop (fig. S12G). To further test whether cleavage was occurring at uracil residues, we mutated single uracil residues in ssRNA 1 that showed cleavage in the RNA-sequencing (Fig. 3B) to adenines. This experiment showed that, by mutating each uracil residue, we could modulate the presence of a single cleavage band, consistent with LshC2c2 cleaving at uracil residues in ssRNA regions (Fig. 3D).

## The HEPN domains of C2c2 mediate RNA-guided ssRNA-cleavage

Bioinformatic analysis of C2c2 has suggested that the HEPN domains are likely to be responsible for the observed catalytic activity (19). Each of the two HEPN domains of C2c2 contains a dyad of conserved arginine and histidine residues (Fig. 4A), in agreement with the catalytic mechanism of the HEPN endoRNAse (26–28). We mutated each of these putative catalytic residues separately to alanine (R597A, H602A, R1278A, H1283A) in the LshC2c2

locus plasmids and assayed for MS2 interference. None of the four mutant plasmids were able to protect *E. coli* from phage infection (Figs. 4B, S13).

We purified the four single-point mutant proteins and assayed their ability to cleave 5'-end-labeled ssRNA target 1 (Fig. 4C). In agreement with our *in vivo* results, all four mutations abolished cleavage activity. The inability of either of the two wild-type HEPN domains to compensate for inactivation of the other implies cooperation between the two domains. These results agree with observations that several bacterial and eukaryotic single-HEPN proteins function as dimers (27, 28, 41).

Catalytically inactive variants of Cas9 retain target DNA binding, allowing for the creation of programmable DNA-binding proteins (12, 13). Electrophoretic mobility shift assays (EMSA) on both the wild-type (Fig. 4D) and R1278A mutant LshC2c2 (Fig. 4E) in complex with crRNA showed the wild-type LshC2c2 complex binding strongly ($K_D \sim 46$ nM, fig. S14A) and specifically to 5'-end-labeled ssRNA target 10 but not to the 5'-end-labeled non-target ssRNA (the reverse complement of ssRNA target 10). The R1278A mutant C2c2 complex showed even stronger ($K_D \sim 7$ nM, fig. S14B) specific binding, indicating that this HEPN mutation results in a catalytically inactive, RNA-programmable RNA-binding protein. The LshC2c2 protein or crRNA alone showed reduced levels of target affinity, as expected (fig. S14C–E). Additionally, no specific binding of LshC2c2-crRNA complex to ssDNA was observed (fig. S15).

These results demonstrate that C2c2 cleaves RNA via a catalytic mechanism distinct from other known CRISPR-associated RNases. In particular, the type III Csm and Cmr multiprotein complexes rely on acidic residues of RRM domains for catalysis, whereas C2c2 achieves RNA cleavage through the conserved basic residues of its two HEPN domains.

## Sequence and structural requirements of C2c2 crRNA

Similar to the type V-B (Cpf1) systems (18), the LshC2c2 crRNA contains a single stem loop in the direct repeat (DR), suggesting that the secondary structure of the crRNA could facilitate interaction with LshC2c2. We thus investigated the length requirements of the spacer sequence for ssRNA cleavage and found that LshC2c2 requires spacers of at least 22 nt length to efficiently cleave ssRNA target 1 (fig. S16A). The stem-loop structure of the crRNA is also critical for ssRNA cleavage, because DR truncations that disturbed the stem loop abrogated target cleavage (fig. S16B). Thus, a DR longer than 24 nt is required to maintain the stem loop necessary for LshC2c2 to mediate ssRNA cleavage.

Single base pair inversions in the stem that preserved the stem structure did not affect the activity of the LshC2c2 complex. In contrast, inverting all four G-C pairs in the stem eliminated the cleavage despite maintaining the duplex structure (fig. S17A). Other perturbations, such as those that introduced kinks and reduced or increased base-pairing in the stem, also eliminated or drastically suppressed cleavage. This suggests that the crRNA stem length is important for complex formation and activity (fig. S17A). We also found that loop deletions eliminated cleavage, whereas insertions and substitutions mostly maintained some level of cleavage activity (fig. S17B). In contrast, nearly all substitutions or deletions

in the region 3' to the DR prevented cleavage by LshC2c2 (fig S18). Together, these results demonstrate that LshC2c2 recognizes structural characteristics of its cognate crRNA but is amenable to loop insertions and most tested base substitutions outside of the 3' DR region. These results have implications for the future application of C2c2-based tools that require guide engineering for recruitment of effectors or modulation of activity (42–44).

## C2c2 cleavage is sensitive to double mismatches in the crRNA-target duplex

We tested the sensitivity of the LshC2c2 system to single mismatches between the crRNA guide and target RNA by mutating single bases across the spacer to the respective complementary bases (e.g., A to U). We then quantified plaque formation with these mismatched spacers in the MS2 infection assay and found that C2c2 was fully tolerant to single mismatches across the spacer as such mismatched spacers interfered with phage propagation with similar efficiency as fully matched spacers (figs. S19A, S20). However, when we introduced consecutive double substitutions in the spacer, we found a ~3 $\log_{10}$-fold reduction in the protection for mismatches in the center, but not at the 5'- or 3'-end, of the crRNA (figs. 19B, S20). This observation suggests the presence of a mismatch-sensitive "seed region" in the center of the crRNA-target duplex.

We generated a set of *in vitro* transcribed crRNAs with mismatches similarly positioned across the spacer region. When incubated with LshC2c2 protein, all single mismatched crRNA supported cleavage (Fig. S19C), in agreement with our *in vivo* findings. When tested with a set of consecutive and non-consecutive double mutant crRNAs, LshC2c2 was unable to cleave the target RNA if the mismatches were positioned in the center, but not at the 5'- or 3'-end of the crRNA (Fig. S19D, S21A), further supporting the existence of a central seed region. Additionally, no cleavage activity was observed with crRNAs containing consecutive triple mismatches in the seed region (fig. S21B).

## C2c2 can be reprogrammed to mediate specific mRNA knockdown *in vivo*

Given the ability of C2c2 to cleave target ssRNA in a crRNA sequence-specific manner, we tested whether LshC2c2 could be reprogrammed to degrade selected non-phage ssRNA targets, and particularly mRNAs, *in vivo*. We co-transformed *E. coli* with a plasmid encoding LshC2c2 and a crRNA targeting the mRNA of red fluorescent protein (RFP) as well as a compatible plasmid expressing RFP (Fig. 5A). For OD-matched samples, we observed an approximately 20% to 92% decrease in RFP positive cells for crRNAs targeting protospacers flanked by C, A, or U PFSs (Fig. 5B, C). As a control, we tested crRNAs containing reverse complements (targeting the dsDNA plasmid) of the top performing RFP mRNA-targeting spacers. As expected, we observed no decrease in RFP fluorescence by these crRNAs (Fig. 5B). We also confirmed that mutation of the catalytic arginine residues in either HEPN domain to alanine precluded RFP knockdown (fig. S22). Thus, C2c2 is capable of general retargeting to arbitrary ssRNA substrates, governed exclusively by predictable nucleic-acid interactions.

When we examined the growth of cells carrying the RFP-targeting spacer with the greatest level of RFP knockdown, we noted that the growth rate of these bacteria was substantially reduced (Fig. 5A, spacer 7). We investigated whether the effect on growth was mediated by the RFP mRNA-targeting activity of LshC2c2 by introducing an inducible-RFP plasmid and an RFP-targeting LshC2c2 locus into *E. coli*. Upon induction of RFP transcription, cells with RFP knockdown showed substantial growth suppression, not observed in non-targeting controls (Fig. 5D, E). This growth restriction was dependent on the level of the RFP mRNA, as controlled by the concentration of the inducer anhydrotetracycline. In contrast, in the absence of RFP transcription, we did not observe any growth restriction nor did we observe any transcription-dependent DNA targeting in our biochemical experiment (fig. S11). These results indicate that RNA-targeting is likely the primary driver of this growth restriction phenotype. We therefore surmised that, in addition to the cleavage of the target RNA, C2c2 CRISPR systems might prevent virus reproduction also via non-specific cleavage of cellular mRNAs, causing programmed cell death (PCD) or dormancy (45, 46).

## C2c2 cleaves collateral RNA in addition to crRNA-targeted ssRNA

Cas9 and Cpf1 cleave DNA within the crRNA-target heteroduplex at defined positions, reverting to an inactive state after cleavage. In contrast, C2c2 cleaves the target RNA outside of the crRNA binding site at varying distances depending on the flanking sequence, presumably within exposed ssRNA loop regions (Figs. 3B, 3C, S12A–D). This observed flexibility with respect to the cleavage distance led us to test whether cleavage of other, non-target ssRNAs also occurs upon C2c2 target binding and activation. Under this model, the C2c2-crRNA complex, once activated by binding to its target RNA, cleaves the target RNA as well as other RNAs non-specifically. We carried out *in vitro* cleavage reactions that included, in addition to LshC2c2 protein, crRNA and its target RNA, one of four unrelated RNA molecules without any complementarity to the crRNA guide (Fig. 6A). These experiments showed that, whereas the LshC2c2-crRNA complex did not mediate cleavage of any of the four collateral RNAs in the absence of the target RNA, all four were efficiently degraded in the presence of the target RNA (Figs. 6B, S23A). Furthermore, R597A and R1278A HEPN mutants were unable to cleave collateral RNA (Fig. S23B).

To further investigate the collateral cleavage and growth restriction *in vivo*, we hypothesized that if a PFS preference screen for LshC2c2 was performed in a transcribed region on the transformed plasmid, then we should be able to detect the PFS preference due to growth restriction induced by RNA targeting. We designed a protospacer site flanked by five randomized nucleotides at the 3’ end in either a non-transcribed region or in a region transcribed from the *lac* promoter (fig. S24A). The analysis of the depleted and enriched PFS sequences identified a H PFS only in the assay with the transcribed sequence but no discernable motif in the non-transcribed sequence (fig. S24B–C).

These results suggest a HEPN-dependent mechanism whereby C2c2 in a complex with crRNA is activated upon binding to target RNA and subsequently cleaves non-specifically other available ssRNA targets. Such promiscuous RNA cleavage could cause cellular toxicity, resulting in the observed growth rate inhibition. These findings imply that, in addition to their likely role in direct suppression of RNA viruses, type VI CRISPR-Cas

systems could function as mediators of a distinct variety of PCD or dormancy induction that is specifically triggered by cognate invader genomes (Fig. 7). Under this scenario, dormancy would slow the infection and supply additional time for adaptive immunity. Such a mechanism agrees with the previously proposed coupling of adaptive immunity and PCD during the CRISPR-Cas defensive response (47).

## Conclusions

In summary, the Class 2 type VI effector protein C2c2 is an RNA-guided RNase that can be efficiently programmed to degrade any ssRNA by specifying a 28-nt sequence on the crRNA (Fig. 10). C2c2 cleaves RNA through conserved basic residues within its two HEPN domains, in contrast to the catalytic mechanisms of other known RNases found in CRISPR-Cas systems (25, 48). Alanine substitution of any of the four predicted HEPN domain catalytic residues converted C2c2 into an inactive programmable RNA-binding protein (dC2c2, analogous to dCas9). Many different spacer sequences work well in our assays although further screening will likely define properties and rules governing optimal function.

These results suggest a broad range of biotechnology applications and research questions (49–51). For example, the ability of dC2c2 to bind to specified sequences could be used to (i) bring effector modules to specific transcripts to modulate their function or translation, which could be used for large-scale screening, construction of synthetic regulatory circuits and other purposes; (ii) fluorescently tag specific RNAs to visualize their trafficking and/or localization; (iii) alter RNA localization through domains with affinity for specific subcellular compartments; and (iv) capture specific transcripts (through direct pull down of dC2c2) to enrich for proximal molecular partners, including RNAs and proteins.

Active C2c2 also has many potential applications such as targeting a specific transcript for destruction, as performed here with RFP. In addition, C2c2, once primed by the cognate target, can cleave other (non-complementary) RNA molecules *in vitro* and inhibit cell growth *in vivo*. Biologically, this promiscuous RNase activity might reflect a PCD/dormancy-based protection mechanism of the type VI CRISPR-Cas systems (Fig. 7). Technologically, it might be used to trigger PCD or dormancy in specific cells such as cancer cells expressing a particular transcript, neurons of a given class, or cells infected by a specific pathogen.

Further experimental study is required to elucidate the mechanisms by which the C2c2 system acquires spacers and the classes of pathogens against which it protects bacteria. The presence of the conserved CRISPR adaptation module consisting of typical Cas1 and Cas2 proteins in the LshC2c2 locus suggests that it is capable of spacer acquisition. Although C2c2 systems lack reverse transcriptases, which mediate acquisition of RNA spacers in some type III systems (52), it is possible that additional host or viral factors could support RNA spacer acquisition. Additionally or alternatively, type VI systems could acquire DNA spacers similar to other CRISPR-Cas variants but then target transcripts of the respective DNA genomes, eliciting PCD and abortive infection (Fig. 7).

The CRISPR-C2c2 system represent a distinct evolutionary path among Class 2 CRISPR-Cas systems. It is likely that other, broadly analogous Class 2 RNA-targeting immune systems exist, and further characterization of the diverse members of Class 2 systems will provide a deeper understanding of bacterial immunity and provide a rich starting point for the development of programmable molecular tools for *in vivo* RNA manipulation.

## Materials and Methods

Expanded materials and methods, including computational analysis, can be found in supplementary materials and methods.

## Bacterial phage interference

The C2c2 CRISPR locus was amplified from DNA from *Leptotrichia shahii DSM 19757* (ATCC, Manassas, VA) and cloned for heterologous expression in *E. coli.* For screens, a library of all possible spacers targeting the MS2 genome were cloned into the spacer array; for individual spacers, single specific spacers were cloned into the array. Interference screens were performed in liquid culture and plated; surviving colonies were harvested for DNA and spacer representation was determined by next-generation sequencing. Individual spacers were tested by spotting on top agar.

## β–lactamase and transcribed/non-transcribed PFS preference screens

Sequences with randomized nucleotides adjacent to protospacer 1 were cloned into pUC19 in corresponding regions. Libraries were screened by co-transformation with LshC2c2 locus plasmid or pACYC184 plasmid control, harvesting of the surviving colonies, and next-generation sequencing of the resulting regions.

## RFP targeting assay

Cells containing an RFP expressing plasmid were transformed with an LshC2c2 locus plasmid with corresponding spacers, grown overnight, and analyzed for RFP fluorescence by flow cytometry. The growth effects of LshC2c2 activity were quantified by titrating inducible RFP levels with dilutions of anhydrotetracycline inducer and then measuring $OD_{600}$.

## *in vitro* nuclease and electrophoretic mobility shift assays

LshC2c2 protein and HEPN mutants were purified for use in *in vitro* reactions; RNA were synthesized via *in vitro* transcription. For nuclease assays, protein was co-incubated with crRNA and either 3' or 5'-labeled targets and analyzed via denaturing gel electrophoresis and imaging or by next-generation sequencing. For electrophoretic mobility shift assays, protein and nucleic acid were co-incubated and then resolved by gel electrophoresis and imaging.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We would like to thank P. Boutz, J. Doench, P. Sharp, and B. Zetsche for helpful discussions and insights; R. Belliveau for overall research support; J. Francis and D. O'Connell for generous MiSeq instrument access; D. Daniels and C. Garvie for providing bacterial incubation space for protein purification; R. Macrae for critical reading of the manuscript; and the entire Zhang laboratory for support and advice. We would like to thank N. Ranu for generously providing pRFP and D. Daniels for providing 6-His-MBP-TEV. O.A.A. is supported by a Paul and Daisy Soros Fellowship, a Friends of the McGovern Institute Fellowship, and the Poitras Center for Affective Disorders.

J.S.G. is supported by a D.O.E. Computational Science Graduate Fellowship. S.S. is supported by the graduate program of Skoltech Data-Intensive Biomedicine and Biotechnology Center for Research, Education, and Innovation. I.S. is supported by the Simons Center for the Social Brain. D.B.T.C. is supported by award number T32GM007753 from the National Institute of General Medical Sciences. K.S.M., E.V.K. and, in part, S.S. are supported by the intramural program of the US Department of Health and Human services (to the National Library of Medicine). K.S. is supported by an NIH grant GM10407, Russian Science Foundation grant 14-14-00988, and Skoltech. F.Z. is a New York Stem Cell Foundation-Robertson Investigator. F.Z. is supported by the NIH through NIMH (5DP1-MH100706 and 1R01-MH110049), NSF, the New York Stem Cell, Simons, Paul G. Allen Family, and Vallee Foundations; and James and Patricia Poitras, Robert Metcalfe, and David Cheng. O.A.A., J.S.G., J.J., E.V.K., S.K., E.S.L., K.S.M., L.M., E.S., K.S., S.S., Y.W., and F.Z. are inventors on provisional patent application 62/181,675 applied for by the Broad Institute, MIT, Harvard, NIH, Skoltech, and Rutgers that covers the C2c2 proteins described in this paper. Deep sequencing data are available at Sequence Read Archive under BioProject accession number PRJNA318890. The authors plan to make the reagents widely available to the academic community through Addgene and to provide software tools via the Zhang lab website (www.genome-engineering.org) and GitHub (github.com/fengzhanglab).

## References and Notes

1. Makarova KS, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015; 13:722–736. [PubMed: 26411297]
2. Makarova KS, et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011; 9:467–477. [PubMed: 21552286]
3. Wright AV, Nunez JK, Doudna JA. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell. 2016; 164:29–44. [PubMed: 26771484]
4. Marraffini LA. CRISPR-Cas immunity in prokaryotes. Nature. 2015; 526:55–61. [PubMed: 26432244]
5. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ. CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci. 2009; 34:401–407. [PubMed: 19646880]
6. Brouns SJ, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008; 321:960–964. [PubMed: 18703739]
7. Hale CR, et al. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell. 2009; 139:945–956. [PubMed: 19945378]
8. Jackson RN, et al. Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science. 2014; 345:1473–1479. [PubMed: 25103409]
9. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. 2008; 322:1843–1845. [PubMed: 19095942]
10. Mulepati S, Heroux A, Bailey S. Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science. 2014; 345:1479–1484. [PubMed: 25123481]
11. Sinkunas T, et al. In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. EMBO J. 2013; 32:385–394. [PubMed: 23334296]

12. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A. 2012; 109:E2579–2586. [PubMed: 22949671]
13. Jinek M, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012; 337:816–821. [PubMed: 22745249]
14. Deltcheva E, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011; 471:602–607. [PubMed: 21455174]
15. Sapranauskas R, et al. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res. 2011; 39:9275–9282. [PubMed: 21813460]
16. Garneau JE, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010; 468:67–71. [PubMed: 21048762]
17. Barrangou R, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007; 315:1709–1712. [PubMed: 17379808]
18. Zetsche B, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015; 163:759–771. [PubMed: 26422227]
19. Shmakov S, et al. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol Cell. 2015; 60:385–397. [PubMed: 26593719]
20. Hale CR, et al. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. Mol Cell. 2012; 45:292–302. [PubMed: 22227116]
21. Jiang W, Samai P, Marraffini LA. Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. Cell. 2016; 164:710–721. [PubMed: 26853474]
22. Samai P, et al. Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell. 2015; 161:1164–1174. [PubMed: 25959775]
23. Staals RH, et al. Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. Mol Cell. 2013; 52:135–145. [PubMed: 24119403]
24. Staals RH, et al. RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. Mol Cell. 2014; 56:518–530. [PubMed: 25457165]
25. Tamulaitis G, et al. Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. Mol Cell. 2014; 56:506–517. [PubMed: 25458845]
26. Anantharaman V, Makarova KS, Burroughs AM, Koonin EV, Aravind L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. Biol Direct. 2013; 8:15. [PubMed: 23768067]
27. Niewoehner O, Jinek M. Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6. RNA. 2016; 22:318–329. [PubMed: 26763118]
28. Sheppard NF, Glover CV 3rd, Terns RM, Terns MP. The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease. RNA. 2016; 22:216–224. [PubMed: 26647461]
29. Goldberg GW, Jiang W, Bikard D, Marraffini LA. Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature. 2014; 514:633–637. [PubMed: 25174707]
30. Deng L, Garrett RA, Shah SA, Peng X, She Q. A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Mol Microbiol. 2013; 87:1088–1099. [PubMed: 23320564]
31. Kim YK, Kim YG, Oh BH. Crystal structure and nucleic acid-binding activity of the CRISPR-associated protein Csx1 of Pyrococcus furiosus. Proteins. 2013; 81:261–270. [PubMed: 22987782]
32. Nunez JK, Lee AS, Engelman A, Doudna JA. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature. 2015
33. Heler R, et al. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature. 2015
34. Nunez JK, et al. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nature structural & molecular biology. 2014; 21:528–534.
35. Diez-Villasenor C, Guzman NM, Almendros C, Garcia-Martinez J, Mojica FJ. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA Biol. 2013; 10:792–802. [PubMed: 23445770]
36. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. 2012; 40:5569–5576. [PubMed: 22402487]

37. Datsenko KA, et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun. 2012; 3:945. [PubMed: 22781758]
38. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature. 2010; 463:568–571. [PubMed: 20072129]
39. Hale CR, Cocozaki A, Li H, Terns RM, Terns MP. Target RNA capture and cleavage by the Cmr type III-B CRISPR-Cas effector complex. Genes Dev. 2014; 28:2432–2443. [PubMed: 25367038]
40. Zhang J, et al. Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol Cell. 2012; 45:303–313. [PubMed: 22227115]
41. Kozlov G, et al. Structural Basis of Defects in the Sacsin HEPN Domain Responsible for Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay (ARSACS). J Biol Chem. 2011; 286:20407–20412. [PubMed: 21507954]
42. Kiani S, et al. Cas9 gRNA engineering for genome editing, activation and repression. Nat Methods. 2015; 12:1051–1054. [PubMed: 26344044]
43. Konermann S, et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature. 2015; 517:583–588. [PubMed: 25494202]
44. Dahlman JE, et al. Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. Nat Biotechnol. 2015; 33:1159–1161. [PubMed: 26436575]
45. Makarova KS, Wolf YI, Koonin EV. Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. Biol Direct. 2009; 4:19. [PubMed: 19493340]
46. Hayes F, Van Melderen L. Toxins-antitoxins: diversity, evolution and function. Crit Rev Biochem Mol Biol. 2011; 46:386–408. [PubMed: 21819231]
47. Makarova KS, Anantharaman V, Aravind L, Koonin EV. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. Biol Direct. 2012; 7:40. [PubMed: 23151069]
48. Benda C, et al. Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. Mol Cell. 2014; 56:43–54. [PubMed: 25280103]
49. Abil Z, Zhao H. Engineering reprogrammable RNA-binding proteins for study and manipulation of the transcriptome. Mol Biosyst. 2015; 11:2658–2665. [PubMed: 26166256]
50. Mackay JP, Font J, Segal DJ. The prospects for designer single-stranded RNA-binding proteins. Nat Struct Mol Biol. 2011; 18:256–261. [PubMed: 21358629]
51. Filipovska A, Rackham O. Designer RNA-binding proteins: New tools for manipulating the transcriptome. RNA Biol. 2011; 8:978–983. [PubMed: 21941129]
52. Silas S, et al. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science. 2016; 351:aad4234. [PubMed: 26917774]
53. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
54. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004; 14:1188–1190. [PubMed: 15173120]
55. Liberzon A, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1:417–425. [PubMed: 26771021]

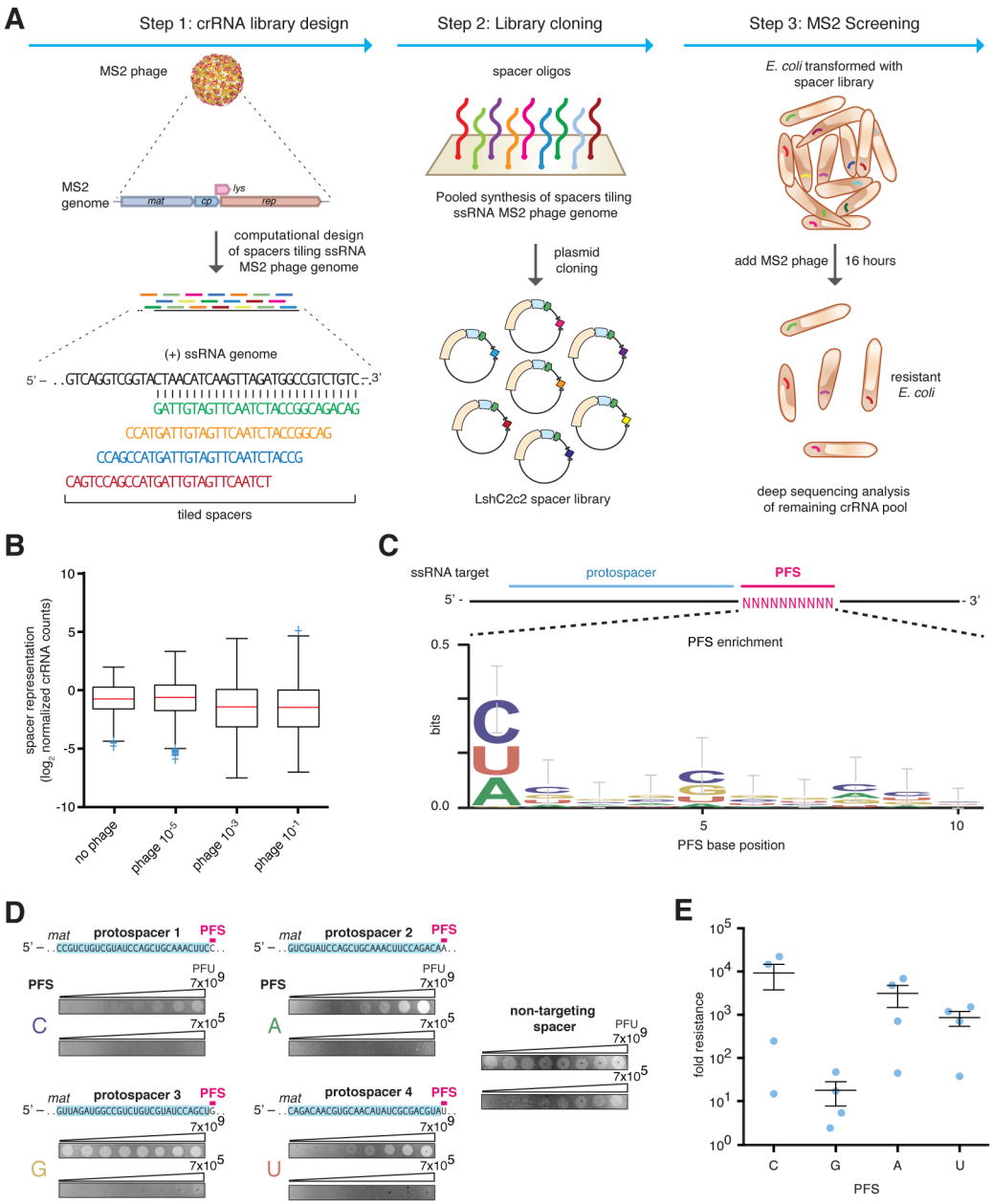



**Figure 1. Heterologous expression of the *Leptotrichia shahii* C2c2 locus mediates robust interference of RNA phage in *Escherichia coli***

A) Schematic for the MS2 bacteriophage interference screen. A library consisting of spacers targeting all possible sequences in the MS2 RNA genome was cloned into the LshC2c2 CRISPR array. Cells transformed with the MS2-targeting spacer library were then treated with phage and plated, and surviving cells were harvested. The frequency of spacers was compared to an untreated control (no phage), and enriched spacers from the phage-treated condition were used for the generation of PFS preference logos.

B) Box plot showing the distribution of normalized crRNA frequencies for the phage-treated conditions and control screen (no phage) biological replicates (n = 3). The box extends from the first to third quartile with whiskers denoting 1.5 times the interquartile range. The mean

is indicated by the red horizontal bar. The $10^{-1}$ and $10^{-3}$ phage dilution distributions are significantly different than each of the control replicates (****, $p < 0.0001$ by ANOVA with multiple hypothesis correction).

C) Sequence logo generated from sequences flanking the 3' end of protospacers corresponding to enriched spacers in the $10^{-1}$ phage dilution condition, revealing the presence of a 3' H PFS (not G).

D) Plaque assay used to validate the functional significance of the H PFS in MS2 interference. All protospacers flanked by non-G PFSs exhibited robust phage interference. Spacer were designed to target the MS2 *mat* gene and their sequences are shown above the plaque images; the spacer used in the non-targeting control is not complementary to any sequence in either the *E. coli* or MS2 genome. Phage spots were applied as series of half-log dilutions.

E) Quantitation of MS2 plaque assay validating the H (non-G) PFS preference. 4 MS2-targeting spacers were designed for each PFS. Each point on the scatter plot represents the average of three biological replicates and corresponds to a single spacer. Bars indicate the mean of 4 spacers for each PFS and standard error (s.e.m).

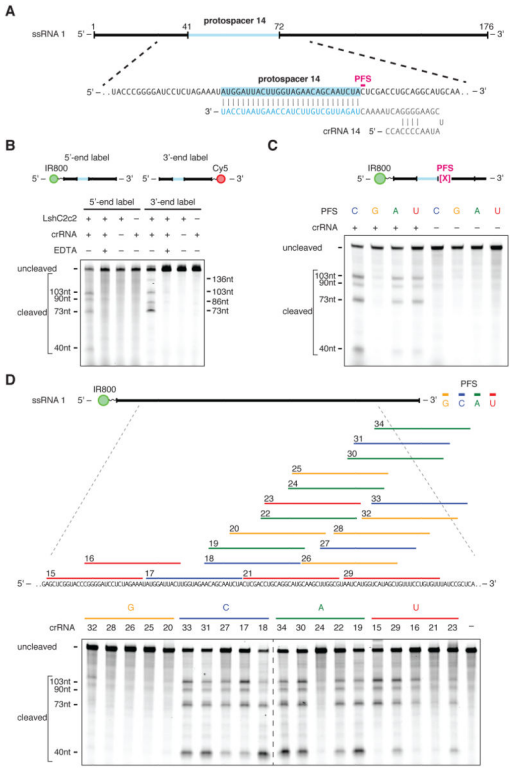



**Figure 2. LshC2c2 and crRNA mediate RNA-guided ssRNA cleavage**

A) Schematic of the ssRNA substrate being targeted by the crRNA. The protospacer region is highlighted in blue and the PFS is indicated by the magenta bar.

B) A denaturing gel demonstrating crRNA-mediated ssRNA cleavage by LshC2c2 after 1 hour of incubation. The ssRNA target is either 5' labeled with IRDye 800 or 3' labeled with Cy5. Cleavage requires the presence of the crRNA and is abolished by addition of EDTA. Four cleavage sites are observed. Reported band lengths are matched from RNA sequencing.

C) A denaturing gel demonstrating the requirement for an H PFS (not G) after 3 hours of incubation. Four ssRNA substrates that are identical except for the PFS (indicated by the magenta X in the schematic) were used for the *in vitro* cleavage reactions. ssRNA cleavage activity is dependent on the nucleotide immediately 3' of the target site. Reported band lengths are matched from RNA sequencing.

D) Schematic showing five protospacers for each PFS on the ssRNA target (top). Denaturing gel showing crRNA-guided ssRNA cleavage activity after 1 hour of incubation. crRNAs correspond to protospacer numbering. Reported band lengths are matched from RNA sequencing.

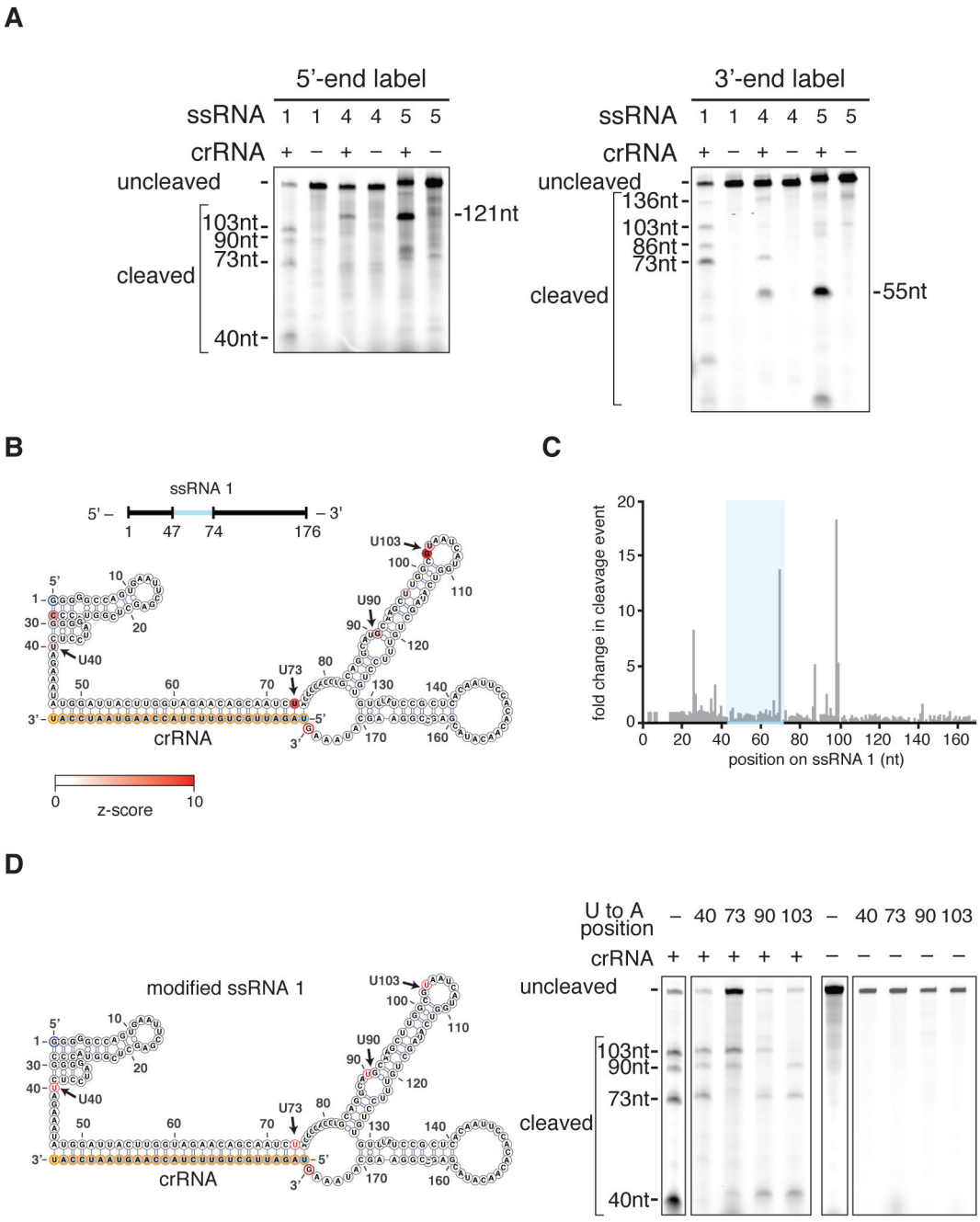



**Figure 3. C2c2 cleavage sites are determined by secondary structure and sequence of the target RNA**

A) Denaturing gel showing C2c2-crRNA-mediated cleavage after 3 hours of incubation of three non-homopolymeric ssRNA targets (1, 4, 5; black, blue and green on figs 3B–C and S12A–D respectively) that share the same protospacer but are flanked by different sequences. Despite identical protospacers, different flanking sequences resulted in different cleavage patterns. Reported band lengths are matched from RNA sequencing.

B) The cleavage sites of non-homopolymer ssRNA target 1 were mapped with RNA-sequencing of the cleavage products. The frequency of cleavage at each base is colored

according to the z-score and shown on the predicted crRNA-ssRNA co-fold secondary structure. Fragments used to generate the frequency analysis contained the complete 5' end. The 5' and 3' end of the ssRNA target are indicated by blue and red outlines, on the ssRNA and secondary structure, respectively. The 5' and 3' end of the spacer (outlined in yellow) is indicated by the blue and orange residues highlighted respectively. The crRNA nucleotides are highlighted in orange.

C) Plot of the frequencies of cleavage sites for each position of ssRNA target 1 for all reads that begin at the 5' end. The protospacer is indicated by the blue highlighted region.

D) Schematic of a modified ssRNA 1 target showing sites (*red)* of single U to A flips *(left).* Denaturing gel showing C2c2-crRNA mediated cleavage of each of these single nucleotide variants after 3 hours of incubation *(right).* Reported band lengths are matched from RNA sequencing.

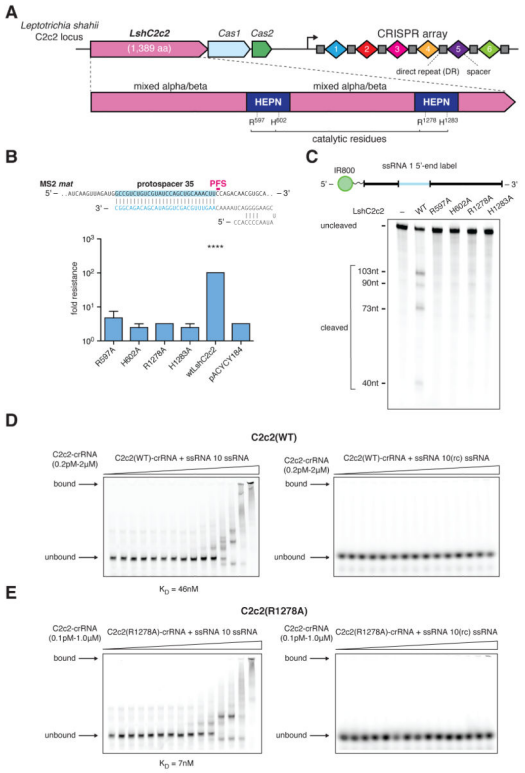



**Figure 4. The two HEPN domains of C2c2 are necessary for crRNA-guided ssRNA cleavage but not for binding**

A) Schematic of the *LshC2c2* locus and the domain organization of the LshC2c2 protein, showing conserved residues in HEPN domains (dark blue).

B) Quantification of MS2 plaque assay with HEPN catalytic residue mutants. For each mutant, the same crRNA targeting protospacer 35 was used. (n=3 biological replicates, ****, $p < 0.0001$ compared to pACYC184 by t-test. Bars represent mean ± s.e.m.)

C) Denaturing gel showing conserved residues of the HEPN motif, indicated as catalytic residues in panel A, are necessary for crRNA-guided ssRNA target 1 cleavage after 3 hours of incubation. Reported band lengths are matched from RNA sequencing.

D) Electrophoretic mobility shift assay (EMSA) evaluating affinity of the wild type LshC2c2-crRNA complex against a targeted (left) and a non-targeted (right) ssRNA substrate. The non-targeted ssRNA substrate is the reverse-complement of the targeted ssRNA 10. EDTA is supplemented to reaction condition to reduce any cleavage activity.

E) Electrophoretic mobility shift assay with LshC2c2(R1278A)-crRNA complex against on-target ssRNA 10 and non-targeting ssRNA (same substrate sequences as in D)

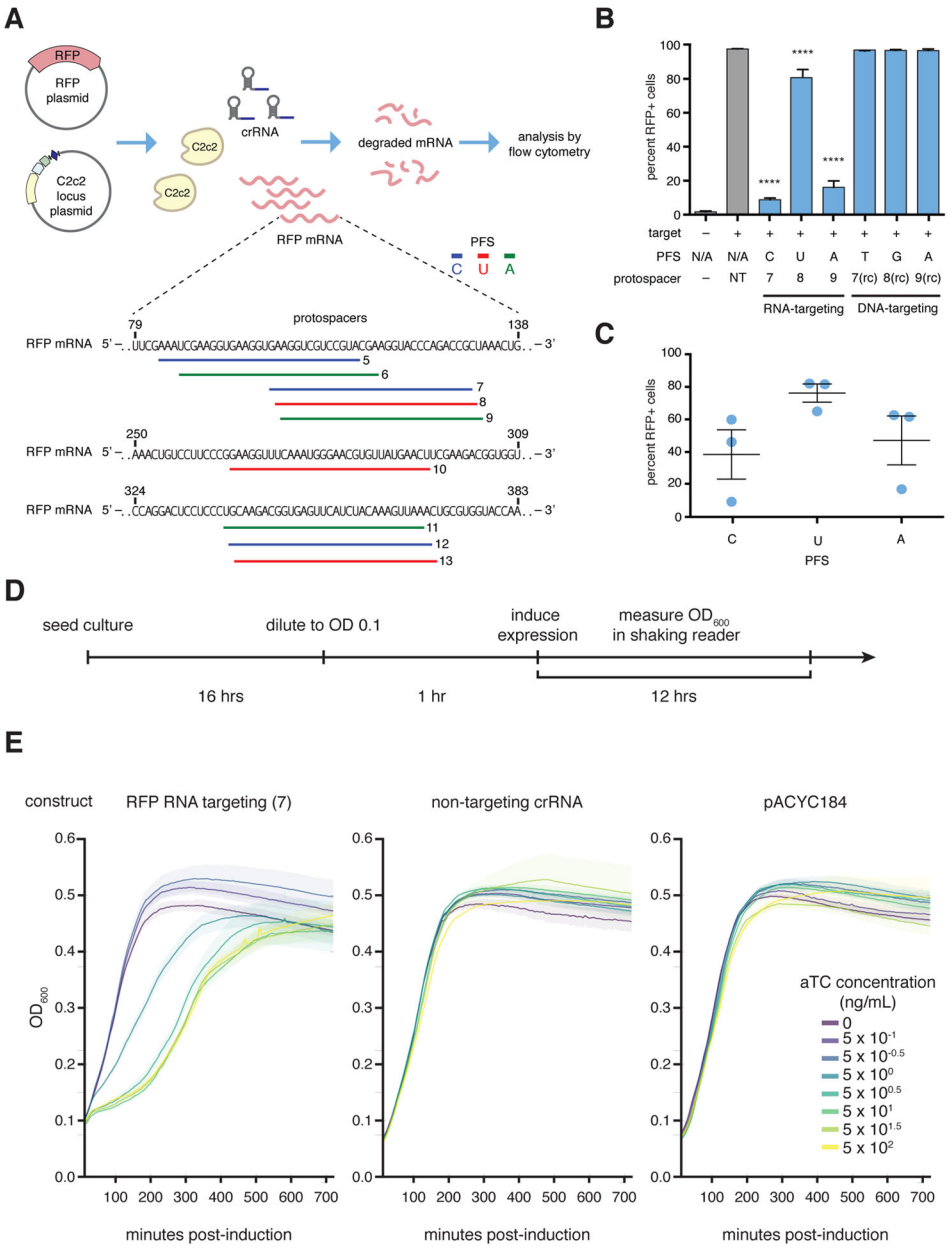



**Figure 5. RFP mRNA knockdown by retargeting LshC2c2**

A) Schematic showing crRNA-guided knockdown of RFP in *E. coli* heterologously expressing the LshC2c2 locus. Three RFP-targeting spacers were selected for each non-G PFS and each protospacer on the RFP mRNA is numbered.

B) RFP mRNA-targeting spacers effected RFP knockdown whereas DNA-targeting spacers (targeting the non-coding strand of the RFP gene on the expression plasmid, indicated as "rc" spacers) did not affect RFP expression. (n=3 biological replicates, ****, $p < 0.0001$ compared to non-targeting guide by ANOVA with multiple hypothesis correction. Bars represent mean ± s.e.m )

C) Quantification of RFP knockdown in *E. coli*. Three spacers each targeting C, U, or A PFS-flanking protospacers (9 spacers, numbered 5–13 as indicated in panel (A)) in the RFP

mRNA were introduced and RFP expression was measured by flow cytometry. Each point on the scatter plot represents the average of three biological replicates and corresponds to a single spacer. Bars indicate the mean of 3 spacers for each PFS and errors bars are shown as the s.e.m.

D) Timeline of *E. coli* growth assay.

E) Effect of RFP mRNA targeting on the growth rate of *E. coli* transformed with an inducible RFP expression plasmid as well as the LshC2c2 locus with non-targeting, RNA targeting (spacer complementary to the RFP mRNA or RFP gene coding strand), and pACYC control plasmid at different anhydrotetracycline (aTc) concentrations.

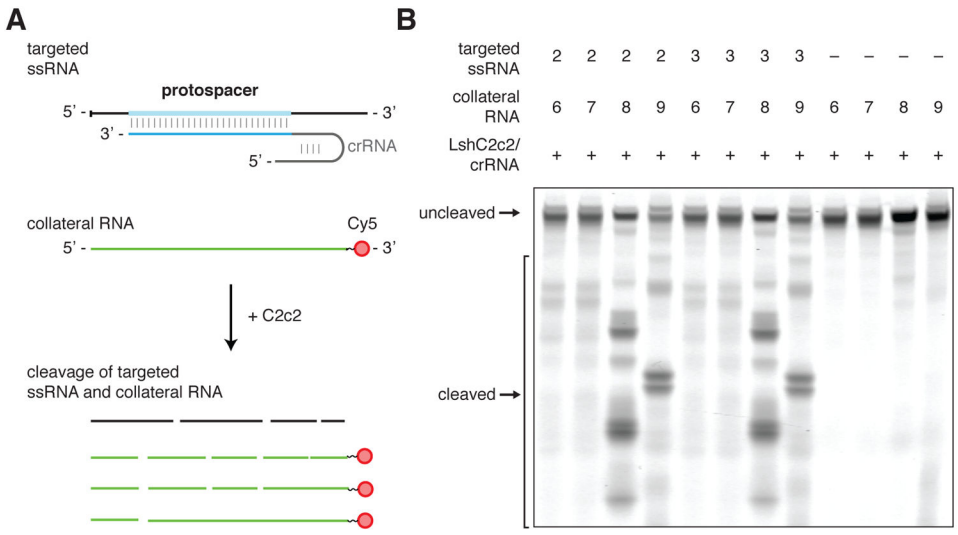



**Figure 6. crRNA-guided ssRNA cleavage activates non-specific RNase activity of LshC2c2**

A) Schematic of the biochemical assay used to detect crRNA-binding-activated non-specific RNase activity on non-crRNA-targeted collateral RNA molecules. The reaction consists of C2c2 protein, unlabeled crRNA, unlabeled target ssRNA, and a second ssRNA with 3' fluorescent labeling and is incubated for 3 hours. C2c2-crRNA mediates cleavage of the unlabeled target ssRNA as well as the 3'-end-labeled collateral RNA which has no complementarity to the crRNA.

B) Denaturing gel showing non-specific RNase activity against non-targeted ssRNA substrates in the presence of target RNA after 3 hours of incubation. The non-targeted ssRNA substrate is not cleaved in the absence of the crRNA-targeted ssRNA substrate.

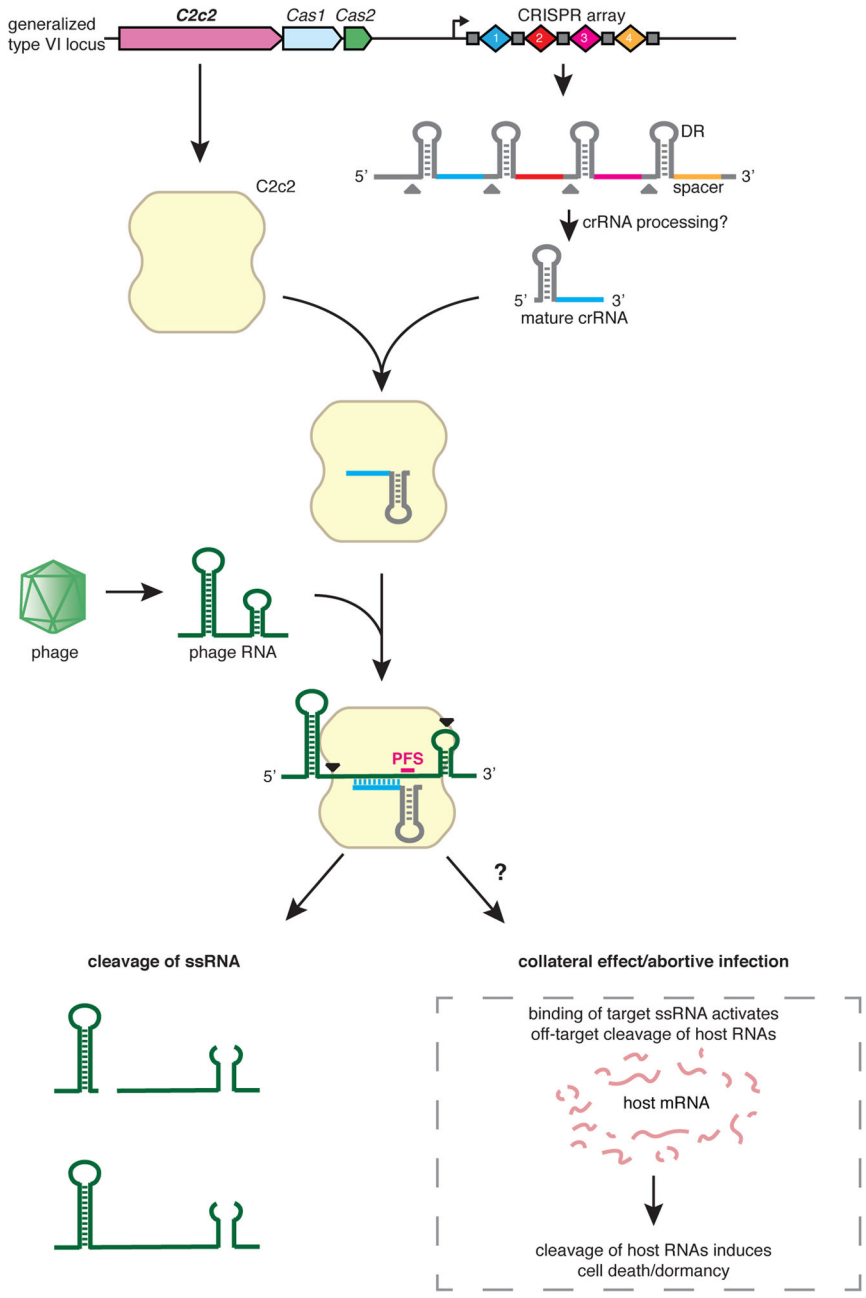



**Figure 7. C2c2 as a putative RNA-targeting prokaryotic immune system**

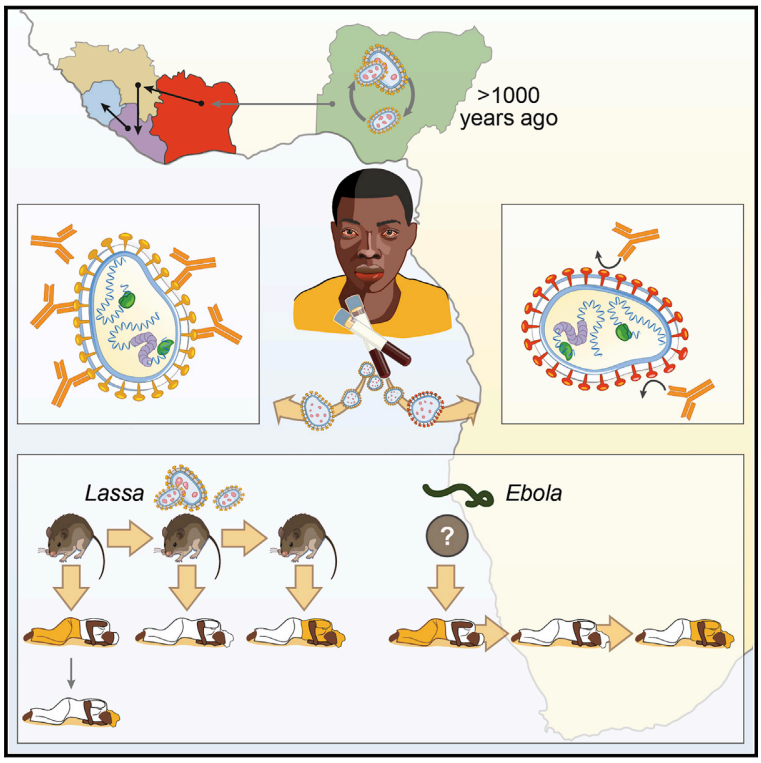
The C2c2-crRNA complex recognizes target RNA via base pairing with the cognate protospacer and cleaves the target RNA. In addition, binding of the target RNA by C2c2-crRNA activates a non-specific RNase activity which may lead to promiscuous cleavage of RNAs without complementarity to the crRNA guide sequence. Through this non-specific RNase activity, C2c2 may also cause abortive infection via programmed cell death or dormancy induction.

Article

Cell

# Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus

## Graphical Abstract




## Highlights

- Lassa virus is a life-threatening pathogen that is endemic in West Africa
- Lassa virus has diverse and ancient origins in Nigeria
- Viral strains from Nigeria and Sierra Leone differ in their translation efficiency
- The virus evolves within hosts to evade immune-determined selection pressures

## Authors

Kristian G. Andersen, B. Jesse Shapiro, Christian B. Matranga, ..., Christian T. Happi, Robert F. Garry, Pardis C. Sabeti

## Correspondence

andersen@scripps.edu (K.G.A.), happic@run.edu.ng (C.T.H.), psabeti@oeb.harvard.edu (P.C.S.)

## In Brief

Sequencing analysis of ~200 Lassa virus genomes reveals its ancient origins and distinct evolutionary features compared to the Ebola virus.

Andersen et al., 2015, Cell *162*, 738–750
August 13, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.07.020

Cell

Article

# Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus

Kristian G. Andersen,[1,2,3,21,*] B. Jesse Shapiro,[1,2,4,21] Christian B. Matranga,[2,21] Rachel Sealfon,[2,5] Aaron E. Lin,[1,2] Lina M. Moses,[6] Onikepe A. Folarin,[7,8] Augustine Goba,[9] Ikponmwonsa Odia,[7] Philomena E. Ehiane,[7] Mambu Momoh,[9,10] Eleina M. England,[2] Sarah Winnicki,[1,2] Luis M. Branco,[11] Stephen K. Gire,[1,2] Eric Phelan,[2] Ridhi Tariyal,[2] Ryan Tewhey,[1,2] Omowunmi Omoniwa,[7] Mohammed Fullah,[9,10,23] Richard Fonnie,[9,23] Mbalu Fonnie,[9,23] Lansana Kanneh,[9] Simbirie Jalloh,[9] Michael Gbakie,[9] Sidiki Saffa,[9,23] Kandeh Karbo,[9] Adrianne D. Gladden,[2] James Qu,[2] Matthew Stremlau,[1,2] Mahan Nekoui,[1,2] Hilary K. Finucane,[2] Shervin Tabrizi,[1,2] Joseph J. Vitti,[1] Bruce Birren,[2] Michael Fitzgerald,[2] Caryn McCowan,[2] Andrea Ireland,[2] Aaron M. Berlin,[2] James Bochicchio,[2] Barbara Tazon-Vega,[2] Niall J. Lennon,[2] Elizabeth M. Ryan,[2] Zach Bjornson,[12] Danny A. Milner, Jr.,[13] Amanda K. Lukens,[13] Nisha Broodie,[14] Megan Rowland,[11] Megan Heinrich,[11] Marjan Akdag,[11] John S. Schieffelin,[6] Danielle Levy,[6] Henry Akpan,[15] Daniel G. Bausch,[6] Kathleen Rubins,[16] Joseph B. McCormick,[17] Eric S. Lander,[2] Stephan Günther,[18] Lisa Hensley,[19] Sylvanus Okogbenin,[7] Viral Hemorrhagic Fever Consortium,[20] Stephen F. Schaffner,[2] Peter O. Okokhere,[7] S. Humarr Khan,[9,23] Donald S. Grant,[9] George O. Akpede,[7] Danny A. Asogun,[7] Andreas Gnirke,[2] Joshua Z. Levin,[2,22] Christian T. Happi,[7,8,22,*] Robert F. Garry,[6,22] and Pardis C. Sabeti[1,2,13,22,*]

[1]FAS Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[2]Broad Institute, Cambridge, MA 02142, USA
[3]The Scripps Research Institute, Scripps Translational Science Institute, La Jolla, CA 92037, USA
[4]Department of Biological Sciences, University of Montréal, Montréal, QC H2V 2S9, Canada
[5]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[6]Tulane Health Sciences Center, Tulane University, New Orleans, LA 70118, USA
[7]Institute of Lassa Fever Research and Control, Irrua Specialist Teaching Hospital, Irrua, Edo State, Nigeria
[8]Department of Biological Sciences, College of Natural Sciences, Redeemer's University, Redemption City, Osun State, Nigeria
[9]Lassa Fever Laboratory, Kenema Government Hospital, Kenema, Eastern Province, Sierra Leone
[10]Eastern Polytechnic College, Kenema, Eastern Province, Sierra Leone
[11]Zalgen Labs, Germantown, MD 20876, USA
[12]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94304, USA
[13]Department of Immunology and Infectious Disease, Harvard School of Public Health, Boston, MA 02115, USA
[14]College of Medicine, Columbia University, New York, NY 10032, USA
[15]Nigerian Federal Ministry of Health, Abuja, Federal Capital Territory, Nigeria
[16]The National Aeronautics and Space Administration, Johnson Space Center, Houston, TX 77058, USA
[17]The University of Texas School of Public Health, Brownsville, TX 77030, USA
[18]Department of Virology, Bernhard-Nocht-Institute for Tropical Medicine, 20259 Hamburg, Germany
[19]NIAID Integrated Research Facility, Frederick, MD 21702, USA
[20]Tulane University, New Orleans, LA 70118, USA
[21]Co-first author
[22]Co-senior author
[23]Deceased
*Correspondence: andersen@scripps.edu (K.G.A.), happic@run.edu.ng (C.T.H.), psabeti@oeb.harvard.edu (P.C.S.)
http://dx.doi.org/10.1016/j.cell.2015.07.020

## SUMMARY

The 2013–2015 West African epidemic of Ebola virus disease (EVD) reminds us of how little is known about biosafety level 4 viruses. Like Ebola virus, Lassa virus (LASV) can cause hemorrhagic fever with high case fatality rates. We generated a genomic catalog of almost 200 LASV sequences from clinical and rodent reservoir samples. We show that whereas the 2013–2015 EVD epidemic is fueled by human-to-human transmissions, LASV infections mainly result from reservoir-to-human infections. We elucidated the spread of LASV across West Africa and show that this migration was accompanied by changes in LASV genome abundance, fatality rates, codon adaptation, and translational efficiency. By investigating intrahost evolution, we found that mutations accumulate in epitopes of viral surface proteins, suggesting selection for immune escape. This catalog will serve as a foundation for the development of vaccines and diagnostics.

## INTRODUCTION

Viruses that cause human hemorrhagic fevers, such as Ebola, Marburg, and Lassa, are classified as BL-4 agents due to their high fatality rates and lack of effective treatment (Paessler and Walker, 2013). With increasing globalization, changing climatic conditions and an ever-expanding human population, our interactions with these pathogens are likely to increase (Gire et al.,

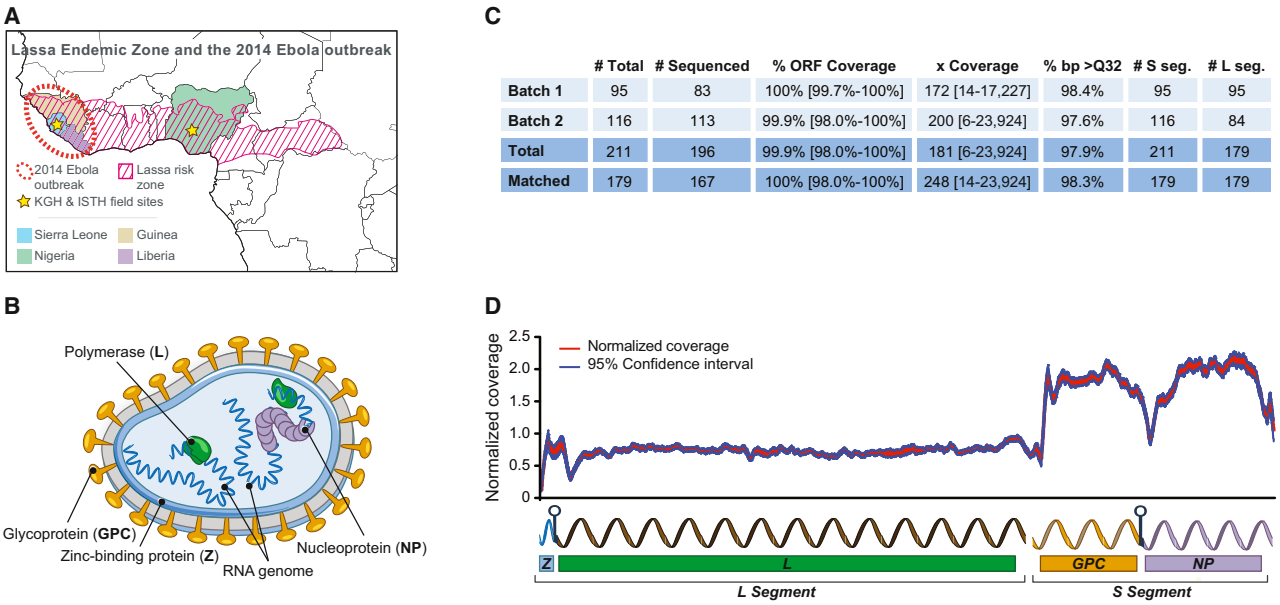



**Figure 1. Lassa Fever Is a Viral Hemorrhagic Fever Endemic in West Africa, where Ebola Virus Disease Broke Out in 2014**

(A) Overview of the LF endemic zone. Study sites are marked. The LF risk zone was defined according to Fichet-Calvet and Rogers (2009).

(B) Schematic of LASV virions.

(C) Summary of LASV sequence data (% ORF coverage = average coverage of open reading frames; x Coverage = median base pair (bp) coverage; % bp > Q32 = fraction of bp with a phred-score > 32.

(D) Plot of the combined normalized (to the sample average) genome coverages (matched dataset, n = 167).

See also Figure S1, Tables S1 and S2, and Data S1.

2012; Lipkin, 2013). The 2013–2015 Ebola virus disease (EVD) epidemic (Baize et al., 2014) is a stark reminder that a better understanding of these viruses is required to develop effective therapeutics and vaccines, as standard containment and isolation can be insufficient to prevent large-scale outbreaks (Pandey et al., 2014).

Lassa virus (LASV) is unique among BL-4 agents in being a common human pathogen, causing endemic disease in much of West Africa, primarily in Sierra Leone, Guinea, Liberia, and Nigeria (Figure 1A). Infection with LASV can lead to acute Lassa fever (LF) with symptoms similar to EVD. LASV is estimated to hospitalize tens of thousands and cause several thousand deaths each year. Case fatality rates (CFRs) among hospitalized LF patients can exceed 50%, although numerous sub-clinical infections are believed to occur (Troup et al., 1970; McCormick and Fisher-Hoch, 2002). Most patients are infected by exposure to excreta from the rodent *Mastomys natalensis*, which functions as a reservoir and maintains persistent infections (Lecompte et al., 2006); human-to-human transmissions have also been reported, however, primarily in hospital settings (McCormick and Fisher-Hoch, 2002; Lo Iacono et al., 2015).

LASV is a single-stranded RNA virus in the family *Arenaviridae*. Its genome consists of two segments, L (7.3 kb) and S (3.4 kb), which encode four proteins: Z (matrix), L (polymerase), NP (nucleoprotein), and GPC, which is post-translationally cleaved into two peptides, GP1 and GP2, that form the transmembrane glycoprotein (Figure 1B). EBOV (*Zaire ebolavirus*) is a single-stranded RNA virus in the family *Filoviridae* with a 19-kb genome encoding seven proteins. While the prevalence of LASV makes it a rare model for studying the evolution of a BL-4 pathogen, only 12 whole-genome LASV sequences were available prior to this study (Djavani et al., 1997; Vieth et al., 2004).

## RESULTS

### Generation of a Large Dataset of Lassa Virus Genomes

We established partnerships with Kenema Government Hospital (KGH), Sierra Leone, and Irrua Specialist Teaching Hospital (ISTH), Nigeria, and collected samples from LF patients between 2008 and 2013. We implemented diagnostics, training, and infrastructure to ensure high-quality and safe sample collection from patients hospitalized with LF (Shaffer et al., 2014).

We sequenced 183 LASV genomes from these clinical samples, 11 LASV genomes from *M. natalensis* field samples, and two genomes from viral laboratory isolates (Figure 1C; Table S1); we deposited all sequence data at NCBI (BioProject PRJNA254017) before publication. Most samples contained >50% human material and yielded <1% LASV reads (Figures S1A and S1B; Table S1). Genome coverage was fairly uniform, with higher coverage of the S than the L segment (Figure 1D), consistent with a greater copy number of S (Southern, 1996).

Since we used an unbiased sequencing approach, we were also able to assemble 7,028 unique open reading frames from the transcriptome of *M. natalensis*, a species not previously sequenced (Figures S1C–S1E; Data S1).

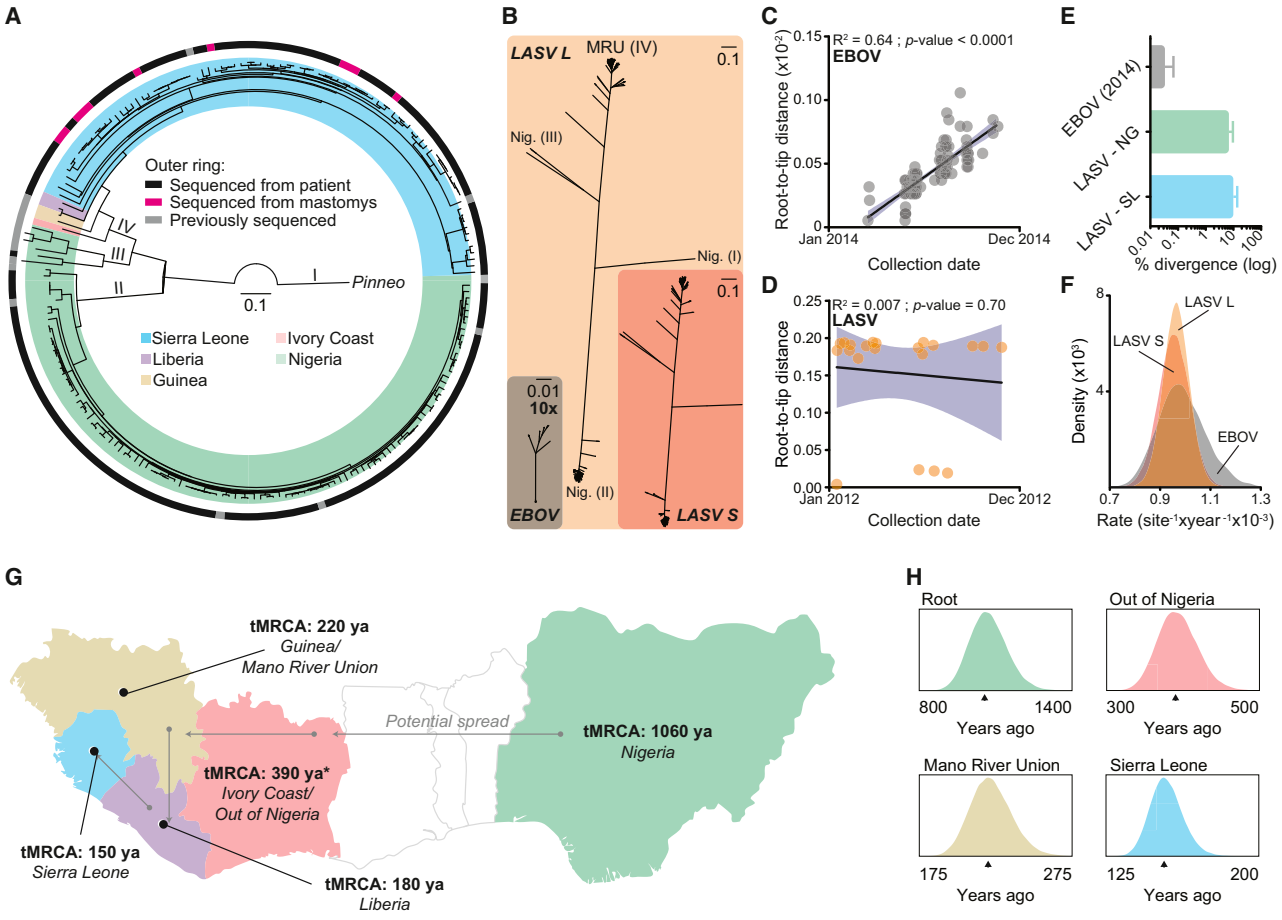



**Figure 2. LASV Is More Diverse than EBOV and Has Ancient Origins in Nigeria**

(A) Phylogenetic tree of LASV S segments (n = 211) (outer ring: gray, previously sequenced; orange, sequenced from *M. natalensis*; scale bar, nucleotide substitutions/site; I–IV, lineages as defined by Bowen et al. (2000).

(B) Scaled trees of LASV L and S segments, as well as EBOV. Trees are shown with the same scale of genetic distance (0.1 nucleotide substitutions/site), except for EBOV, which was magnified 10× (0.01 nucleotide substitutions/site). LASV lineages are shown (Nig., Nigeria; MRU, Mano River Union).

(C and D) Root-to-tip distance versus collection date for (C) EBOV from the West African EVD epidemic (2014; n = 131) or (D) LASV from Sierra Leone (2012; n = 21). Confidence intervals (95%) for linear regression fits are shown in blue.

(E) The % pairwise differences (log scale) in EBOV lineages from the 2014 EVD epidemic (March–October 2014; n = 116) and LASV lineages from Sierra Leone (SL; 2009–2013; n = 60) and Nigeria (NG; 2009–2012; n = 83). The % divergence was calculated within the countries for each year separately and pooled. Error bars represent SD.

(F–H) Bayesian coalescent analysis of LASV samples (matched dataset, n = 179). (F) Substitution rates. (G) LASV L segment tMRCA for each country (median values; ya, years ago). Gray arrows depict the likely spread of LASV. An asterisk indicates a tMRCA that was dependent on only one sequence (AV) from outside Nigeria and MRU. (H) Probability distributions for the estimated tMRCAs with median marked.

See also Figures S1F–S1H, S2, S3, S4, and S5, Table S2, and Data S1, S2, and S3.

## Lassa Virus Strains Are Genetically Diverse and Cluster Based on Geographic Location

We first examined patterns of variation and phylogenetic relationships. We found high levels of LASV nucleotide diversity, with strain variation up to 32% and 25% for the L and S segments (Figures 2A, S1F, and S1G). This is substantially higher than previous findings based on LASV fragments (Bowen et al., 2000), and much higher than EBOV, which is more than 97% conserved across all sequenced strains (Figures 2B and S1H). We confirmed previous findings (Bowen et al., 2000) that LASV clusters into four major clades: three in Nigeria and one from the Mano River Union countries (MRU) of Sierra Leone, Guinea, and Liberia (Figure 2A; Data S1, S2, and S3). We found no evidence for host-specific clades of LASV lineages; rather, samples from humans and *M. natalensis* clustered together (Figure 2A; Data S1, S2, and S3). We did not identify any recombination events within segments, but did find evidence for reassortment between segments in three samples (Figures S2A–S2G). This could be explained by infections of individual hosts with multiple LASV lineages, followed by shuffling of segments, a process previously observed in vitro with LASV (Lukashevich, 1992) and in vivo with other arenaviruses (Stenglein et al., 2015).

## Lassa Virus Infections Are the Result of Multiple Independent Reservoir-to-Human Transmissions

Recent studies suggest that the 2013–2015 EVD epidemic is maintained by sustained human-to-human transmission (Gire et al., 2014) after an initial "spillover" event from a likely animal reservoir (Baize et al., 2014). Similarly, it has been suggested that up to 20% of LF cases also arise from human-to-human transmissions (Lo Iacono et al., 2015). Sustained human-to-human transmission should result in a "ladder-like" structure of the phylogenetic tree along with a strong correlation between a sample's collection date and its genetic distance from the root of the tree over a short time period. Based on data from the 2013–2015 EVD epidemic (Team, 2014), we defined that time period as one year. While collection date is strongly correlated with root-to-tip distance for EBOV from the 2013–2015 EVD epidemic ($R^2$ = 0.64; Figure 2C; Table S2), the same correlation is absent for LASV sampled over a similar time period ($R^2$ = 0.0001; Figure 2D; Table S2).

Human-to-human transmission should also result in clustering of contemporaneous viral sequences on the tree. While this is pervasive across the 2013–2015 EVD epidemic samples (Gire et al., 2014) (Data S1), we found that only 5 out of 169 (3%) LASV sequences from patients resulted in such clusters (Data S1 and S3). As *M. natalensis* serves as the reservoir host for LASV—and presumably maintain LASV diversity via sustained rodent-to-rodent transmission chains—we would expect rodent samples to group into more defined clusters. Indeed, 5 out of 10 (50%) LASV sequences from *M. natalensis* formed clusters consistent with rodent-to-rodent transmissions (Data S1 and S3). Finally, we also found that the average pairwise divergence for EBOV lineages in Sierra Leone from the 2013–2015 EVD epidemic was much lower than that observed for LASV lineages within individual years from Sierra Leone (Figure 2E), despite similar observed substitution rates (Figure 2F). These three lines of evidence suggest that, while EBOV during the 2013–2015 EVD epidemic was transmitted through human-to-human contact, most human LASV infections represent independent transmissions from a genetically diverse reservoir.

## Lassa Virus Has Ancient Origins in Modern-Day Nigeria and Has Recently Spread across West Africa

While EBOV and LASV were both discovered in the latter part of the 20th century—1976 and 1969, respectively—their origins likely vary greatly (Commission, 1978; Frame et al., 1970). Reports suggest that all EVD outbreaks share a common ancestor within the last fifty years (Carroll et al., 2013; Dudas and Rambaut, 2014; Calvignac-Spencer et al., 2014; Gire et al., 2014). In contrast, the widespread persistence of LASV in *M. natalensis*, and evidence in the human genome of natural selection linked to LASV resistance (Andersen et al., 2012), suggest that LASV might be a long-standing human pathogen.

Using molecular dating, we found that extant LASV strains likely originated in modern-day Nigeria more than a thousand years ago and spread into neighboring West African countries within the last several hundred years (Figures 2G and 2H). We first examined evidence for a molecular clock by comparing sample collection dates and root-to-tip distances across the entire LASV tree. In contrast to the shorter timescales analyzed above (Figure 2D), here we found significant evidence for a molecular clock ($R^2$ = 0.38, p value < 0.0001). This allowed us to calculate the time to the most recent common ancestor (tMRCA) using Bayesian coalescent analysis (Drummond et al., 2012). We estimated the tMRCA of sampled extant LASV strains to be a little over one thousand years for the L segment (Figure S2H; Table S2; median = 1,057 years ago [ya]; 915 ya–1,218 ya; 95% highest posterior density [HPD]) and 650 years for the S segment (Figure S2I; Table S2; median = 631 ya; 519 ya–748 ya, 95% HPD). While LASV strains in Nigeria have the same tMRCA as all extant strains, those in Sierra Leone have an estimated tMRCA of only 150 years (Figures 2G and 2H; median = 153 ya; 137 ya–171 ya; 95% HPD).

We tested the sensitivity of our results to key analysis parameters that could severely affect our tMRCA estimates (Wertheim and Kosakovsky Pond, 2011; Wertheim et al., 2013). We found that our estimates were robust to the choice of all tested parameters, including evolutionary model, geographical separations, and exclusion or inclusion of older "anchoring" sequences, e.g., the 1969 Pinneo strain (Figures S3 and S4; Table S2). In linear regression of root-to-tip distance of samples on the date of collection, the sequences from the MRU showed the strongest evidence of temporal structure, suggesting that the dating is most reliable within that region (Figure S4).

## Non-Nigerian Lassa Virus Strains Have Higher Codon Adaptation to Mammalian Hosts

Previous studies have shown that viruses can adapt their codon usage to that of their hosts for translational efficiency (Sharp and Li, 1987; Bahir et al., 2009; Butt et al., 2014; Hershberg and Petrov, 2008). We examined the codon adaptation index (CAI) of LASV and EBOV to different hosts. CAI quantifies how well synonymous codon choice in the viral genome matches that of a potential host genome.

We found that LASV had a higher mean CAI than EBOV, and a similar CAI distribution across different potential mammalian hosts (Figure 3A). There was a strong linear correlation between the CAI of LASV to human and to *M. natalensis*, regardless of which organism LASV was sequenced from (Figure S5A). In agreement with previous studies (Bahir et al., 2009), this suggests that codon adaptation to one mammal also leads to adaptation to another.

We also compared LASV sequences from patients in Sierra Leone to those from Nigeria and found that the former had significantly higher CAI (p value < 0.001, permutation test) (Figure 3B). This apparent "burst" of codon adaptation as LASV spread into Sierra Leone began on the branch leading out of Nigeria and remained high in most non-Nigerian strains (Figures 3C, 3D, and S5B–S5E), with an even distribution across the LASV genome (Figure S5F).

As it has been suggested that dinucleotide usage play a role in determining translational efficiency of RNA viruses (Tulloch et al., 2014), we investigated whether there was a difference between Nigerian and Sierra Leonean strains, but did not observe any significant skew (Figures S5G and S5H).

## Lassa Virus Genome Abundance and Case Fatality Rates Differ between Nigeria and Sierra Leone

Increased codon optimization might lead to increased viral output (Plotkin and Kudla, 2011) and therefore higher viral titers

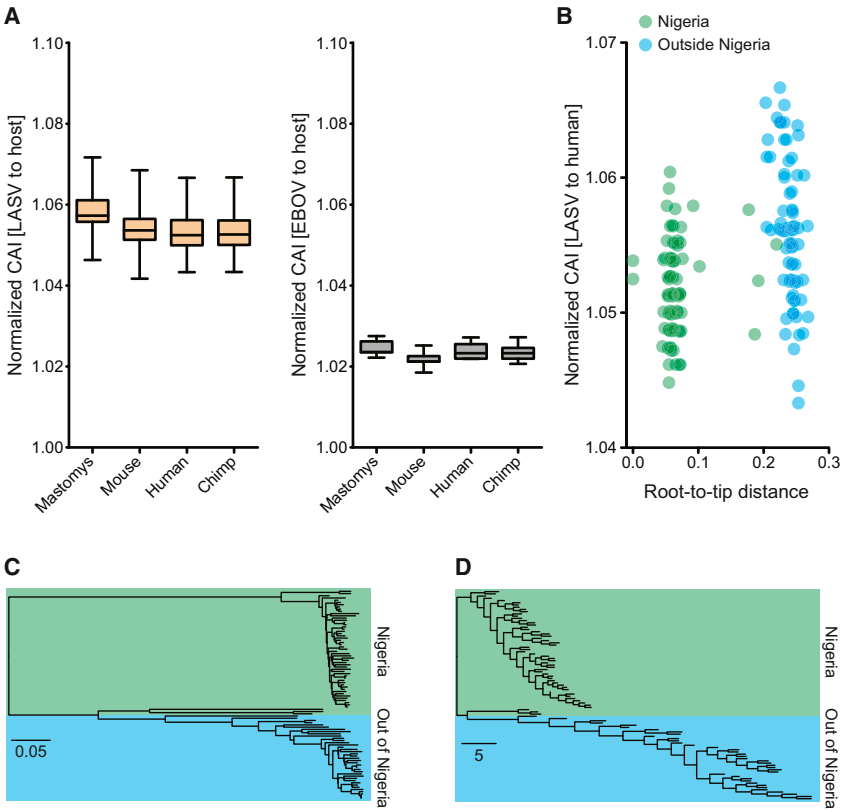



**Figure 3. Increased Codon Adaptation of Non-Nigerian LASV Strains**

(A) Codon adaptation index (CAI) of individual LASV (orange) and EBOV (gray) sequences to four mammalian hosts, normalized by GC and amino acid content.

(B) Normalized CAI (to human) of LASV sequences plotted against their distance (aa substitutions/site) to the root of the tree. (C) Phylogeny of the LASV L genes (scale bar, substitutions/site). (D) A phenogram depicting the phylogeny from (C) with branch lengths representing CAI (scale bar, converted *Z* score).

(C and D) Trees were rooted on Pinneo (data not shown; batch 1 dataset).

See also Table S2.

(Lauring et al., 2012) for non-Nigerian strains. With standardized inclusion criteria at our field sites (Supplemental Experimental Procedures), we tested this hypothesis by using qPCR to quantify LASV genome abundance. We found significantly more LASV genomes in patients from Sierra Leone than in those from Nigeria (Figure 4A). LASV genome abundance in Sierra Leone was similar to that observed in EBOV patients from the same hospital (KGH; Figure 4A) and decreased over the course of the infection (Figure S5I), likely due to treatment with the antiviral drug ribavirin (McCormick et al., 1986). Next, we binned the LASV samples into those in the top or bottom 50% CAI from within each country, and compared LASV genome abundance between bins. In Sierra Leone, individual LASV sequences with high CAI tended to have higher genome copy numbers (p value < 0.05, Mann-Whitney test) but no trend was visible in Nigeria (Figures 4B and 4C). This suggests that CAI may affect LASV replication rate and abundance.

Since increased viremia of LASV in LF patients is correlated with higher fatality rates (McCormick and Fisher-Hoch, 2002), we might also expect CFRs to be higher in patients from Sierra Leone than from Nigeria. Again using strict criteria for inclusion, we found a significantly higher CFR (p value = 0.01; Fisher's exact test) in Sierra Leonean patients than in their Nigerian counterparts (81% versus 60%; Figure 4D). While the treatment options for LF patients are similar in the two countries, other factors could also affect genome abundances and CFRs. In particular, delay in clinical care could bias our estimates; however, self-reported times from onset of symptoms to hospital admission are the same in the two countries (average = 9.3 days; Figure 4E).

## Nigerian Lassa Virus Strains Have Higher Protein Output than Do Sierra Leonean Strains

Although we observed a correlation among CAI, viral genome abundance, and CFR, it remained unclear whether this is driven by differences in protein translation efficiency between Nigerian and Sierra Leonean LASV strains. We designed an experimental system to estimate translational activity for a single LASV gene with different CAI values. We randomly selected 20 LASV sequences from Nigeria and Sierra Leone and fused the first 699 bp of their NP genes ($NP_{1–699}$) to luciferase, before cloning into expression vectors for transfection or in vitro translation experiments (Figure 4F). Readout of luciferase activity allowed us to detect differences in translational activity of the chimeric transcripts. As controls, we codon-optimized one LASV sequence from Nigeria and one from Sierra Leone, for an upper bound on $NP_{1–699}$-luciferase translational efficiency.

For both transfection and in vitro translation experiments, we observed a significant difference in translational output of the tested $NP_{1–699}$-luciferase genes, with Nigerian versions having higher outputs (Figures 4G and 4H). This was the opposite of the expectation based on CAI because the Sierra Leonean sequences had higher CAI (Table S2). Nigerian versions also had higher outputs for the codon-optimized forms of $NP_{1–699}$ (Figure 4H), suggesting that Nigerian sequences are intrinsically more efficient or stable.

To test whether these observations were specific to NP, we repeated the in vitro translation experiment using the first 736 bp of ten LASV GPC genes (Figure 4F). Once again, we found that Nigerian genes had significantly higher translational output (Figure 4I). These results suggest that there is a difference in the translational output between LASV strains from Nigeria and Sierra Leone that is independent of the variation in CAI.

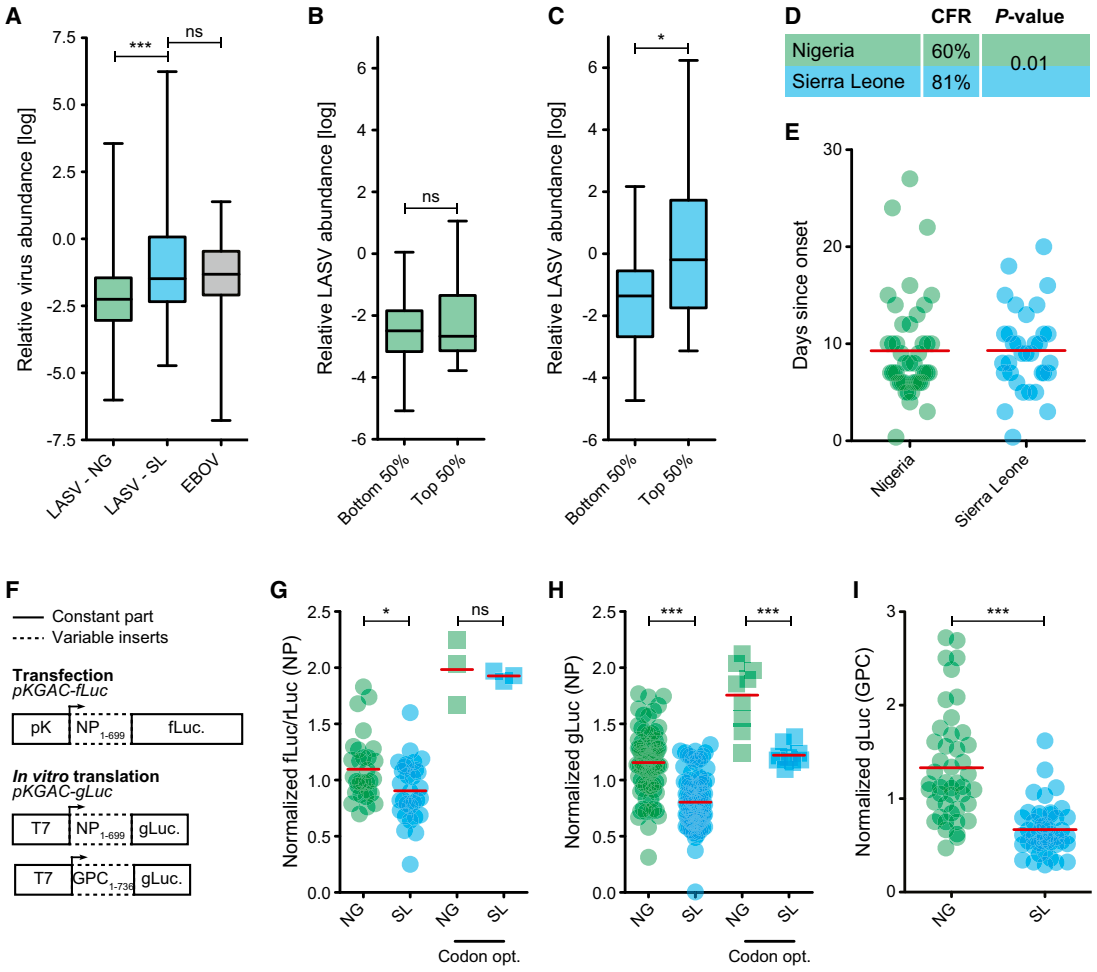



**Figure 4. Difference in Viral Output between Nigerian and Sierra Leonean LASV Strains**

(A) Relative abundance of LASV and EBOV genome copies (log ratio of LASV or EBOV copies/microliter to 18S rRNA copies/microliter; ***p < 0.001, Mann-Whitney test).

(B and C) Relative abundance of LASV genome copies when partitioned into sequences in the top or bottom half of CAI scores. (B) Samples from Nigeria. (C) Samples from Sierra Leone (*p value < 0.05, Mann-Whitney test).

(D) Case fatality rates calculated for patients from Sierra Leone (n = 67) and Nigeria (n = 40). p value from Fisher's exact test.

(E) Patient-reported days from the onset of symptoms until admission to the hospital. The mean values are displayed with red bars.

(F) DNA plasmids encoding the first 699 nucleotides of LASV NP or the first 736 nucleotides of LASV GPC.

(G) NP-reporter expression was measured in HEK293 cells by the ratio of fLuc/rLuc 20 hr post-transfection.

(H and I) In vitro transcription of (H) NP- or (I) GPC-reporter translation measured by gLuc luminescence after 21 hr.

All values were normalized to the average of each biological replicate (n = 3) (G–I). *p < 0.05, ***p < 0.0001, Mann-Whitney test; NG, Nigeria, SL, Sierra Leone. See also Table S2.

## Lassa Virus Is a More Diverse Intrahost than Is Ebola Virus

The long-term evolution of viruses ultimately depends on mutation and selection within individual hosts (Parameswaran et al., 2012). Our deep sequencing allowed us to examine LASV intrahost single-nucleotide variants (iSNVs) within individual human and rodent hosts (Figure 5A). We called variants at a minimum minor allele frequency (minMAF) of 5% and applied stringent filtering (Supplemental Experimental Procedures). We validated subsets of iSNVs using different sequencing technologies and found that our results were consistent across platforms, experimental replicates, library preparations, and variant calling methods (Figures S5J–S5L and S6).

We found that *M. natalensis* generally harbors more LASV iSNVs than humans (median iSNVs/kb = 1.5 versus 0.1; p value < 0.0001; Mann-Whitney test), consistent with longer, more chronic infections (Figures 5B and S7A–S7D). LASV is a significantly more diverse intrahost than is EBOV (accounting for differences in sequence coverage between the two; median bp coverage ~2,000× for EBOV [Gire et al., 2014] and ~250× for LASV; Figure 1C; p value = 0.0005; Mann-Whitney test), with an average number of iSNVs per covered site of $2.1 \times 10^{-3}$ in

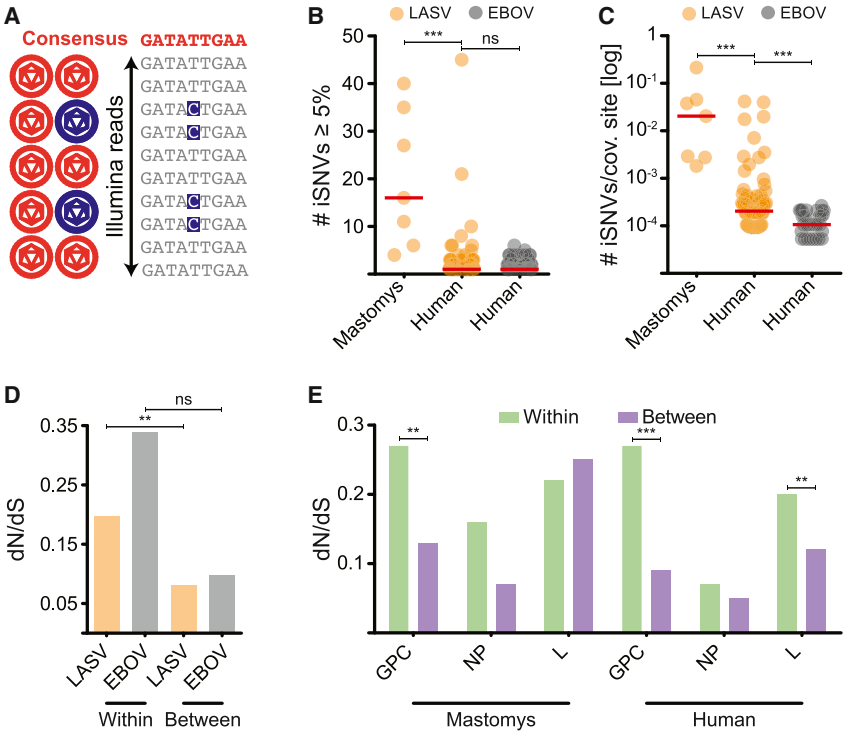



**Figure 5. Genetic Diversity and Selective Pressures within and between Hosts**

(A) iSNVs can be detected in Illumina reads.

(B) The number of iSNVs at 5% MAF or higher, called in LASV (orange) and EBOV (gray).

(C) The normalized number of iSNVs per covered site (80% calling power; 5% MAF). In both (B) and (C), each circle represents one LASV or EBOV sample; red bars denote the median; ***p < 0.001, Mann-Whitney test; ns, not significant. Only samples with >50× coverage are included.

(D) dN/dS ratios for LASV and EBOV based on iSNVs or fixed differences in consensus sequences between hosts.

(E) dN/dS ratios for each LASV gene.

***p < 0.001, **p < 0.01, Permutation test; ns = not significant (D and E). The very short Z gene was excluded.

See also Figures S6 and S7, Table S2, and Data S1.

LF patients, but only 1.3 × $10^{-4}$ in EVD patients (Figure 5C). This difference is primarily driven by a subset of LASV-infected individuals that have >15 iSNVs—diversity similar to that observed in *M. natalensis* (Figure 5C). Such high diversity—with iSNV frequencies that appear stable over the course of infection (Figure S7E)—was never observed in EVD patients (Figures 5B and 5C).

LF is generally considered an acute disease in humans (McCormick and Fisher-Hoch, 2002), but high numbers of iSNVs could be explained by long-term chronic infections and/or adaptive evolution of LASV. An alternative explanation is multiple infections; however, the wide range of allele frequencies (Figure S7F) and general lack of linkage between iSNVs (Table S2) argues against this being the prevailing explanation. In addition, the vast majority of iSNVs (94.4%) are transitions, rather than transversion mutations (Figure S5L), which accumulate over longer evolutionary timescales (Wakeley, 1996). This suggests that most iSNVs are evolutionarily recent and that LASV iSNVs arise mostly via de novo mutation within hosts, and more rarely via transmission and multiple infections by circulating strains.

## Natural Selection Is Acting on the Lassa Virus Glycoprotein

Next, we investigated the role of natural selection in shaping intrahost variation. In LASV, we observed a significantly higher dN/dS (p value = 0.0013; permuted McDonald-Kreitman test)—a measure of selective constraint at the protein level—within hosts than between hosts (Figure 5D). For EBOV, the trend was in the same direction but was not statistically significant (Figure 5D). Assuming that dN represents mostly deleterious mutations (Shapiro et al., 2009), this is consistent with these mutations being purged by purifying selection over evolutionary time (Rocha et al., 2006). Because purifying selection has less time to act within a single host, dN/dS is higher within hosts than between. However, the dN/dS of ~0.2 for LASV iSNVs is still much less than ~1 expected in the absence of any selection (Anisimova and Liberles, 2007). In contrast, iSNVs in certain other viruses, such as dengue, approach the neutral expectation of dN/dS ~1 (Holmes, 2003). Also, reflecting purifying selection on LASV within hosts, the dN/dS ratio appears to decrease at higher iSNV frequencies (Figure S7G). EBOV intrahost dN/dS is higher than LASV (Figure 5D), consistent with LASV intrahost populations being subject to stronger (or a longer duration of) purifying selection.

Intrahost dN/dS varied widely across LASV genes, suggesting different selective pressures on individual genes. Most notably, GPC genes sequenced from both human and *M. natalensis*, had a significantly higher dN/dS ratio within hosts than between hosts (Figure 5E). GPC encodes the only protein partially exposed on the outside of the LASV particle (Figure 1B). It has a significantly higher within-host dN/dS than does NP (p value < 0.05; Fisher's exact test), the neighboring gene on the S segment (Figure 5E), but similar between-host dN/dS. These results suggest either a GPC-specific relaxation of within-host purifying selection or within-host diversifying (positive) selection (Baum et al., 2003).

## Nonsynonymous Lassa Virus Intrahost Variants Accumulate in Predicted Epitopes in the Glycoprotein

We hypothesized that immune pressures on LASV GPC could drive within-host diversifying selection, favoring nonsynonymous iSNVs that impair immune detection by disrupting epitopes. This phenomenon has been reported for other viruses; e.g., at the population-wide level in pandemic influenza A virus (Bhatt et al., 2011). To evaluate whether iSNVs disrupt epitopes,

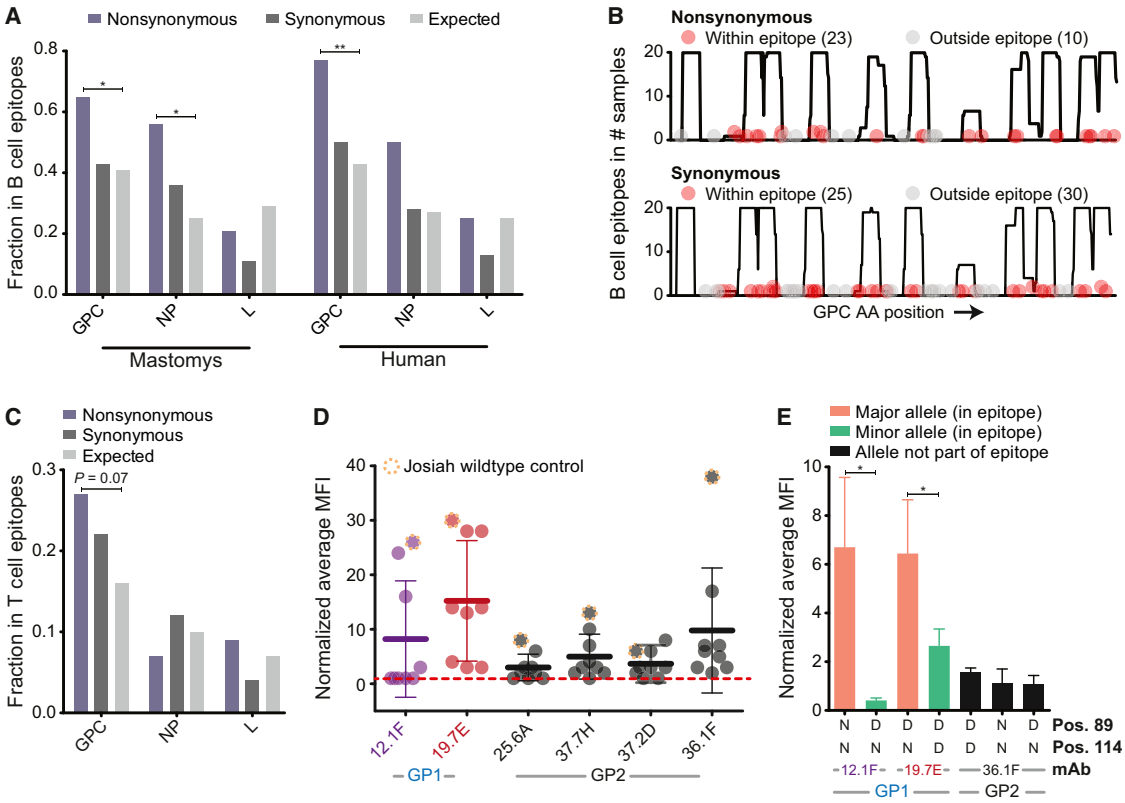



**Figure 6. Nonsynonymous iSNVs Are Overrepresented within Predicted B Cell Epitopes in LASV GPC**

(A) The fraction of iSNVs within predicted B cell epitopes. The observed fraction is compared to the expected fraction (**p < 0.01, *p < 0.05, binomial test).

(B) Overlap between GPC epitopes and iSNVs. Epitopes were predicted separately in each sample (y axis) and overlaid with iSNVs from that sample.

(C) Fraction of iSNVs falling within predicted T cell epitopes (p value = binomial test).

(D and E) Binding of monoclonal antibodies (mAbs) to iSNV mutants in predicted B cell epitopes was tested in HEK293 cells. (D) Each circle corresponds to the normalized average mean fluorescence intensity (MFI) measured by flow cytometry of each LASV GPC construct carrying either wild-type or iSNV mutations (Supplemental Experimental Procedures). Each tested mAb is shown on the x axis. The MFI was normalized to the MFI of the empty vector control for each experiment. (E) Binding to the GP1-specific mAbs 12.1F and 19.7E using constructs carrying either the major or minor population-wide allele at positions 89 and 114. For comparison, binding to mAb 36.1F, which requires GP2, is also shown. All MFI values are normalized to the MFI of binding to the GP2-specific mAb 37.2D. Error bars show the SD from four independent experiments; *p < 0.05, Mann-Whitney test.

See also Table S2 and Data S1.

we used a machine-learning method (El-Manzalawy et al., 2008) to predict linear B cell epitopes in each LASV protein.

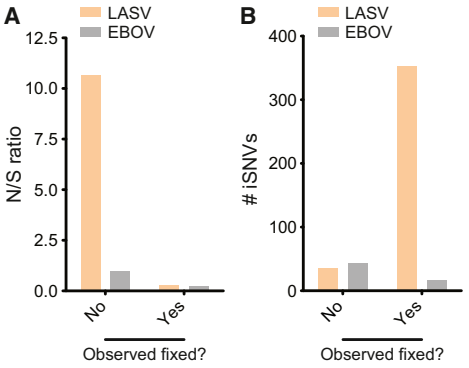
Nonsynonymous iSNVs in LASV GPC occurred in predicted B cell epitopes significantly more than expected by chance (Figures 6A and 6B; p value < 0.01; binomial test). This was true for LASV samples from patients and *M. natalensis* independently, although the signal was stronger in patients (Figure 6A). In contrast, synonymous iSNVs were randomly distributed across GPC, consistent with their lack of impact on epitope structure (Figures 6A and 6B). We observed a similar but weaker trend for NP, although this difference only reached statistical significance in *M. natalensis* (Figure 6A).

To test if nonsynonymous iSNVs interfere with B cell epitope recognition, we reran the B cell epitope predictions, changing single amino acids within the epitopes from the consensus call to the iSNV variant. For 14 of the 18 predicted B cell epitopes, changing the iSNV from the consensus to the variant allele significantly reduced the epitope score (Table S2; p value = 0.015, Sign test).

To test if nonsynonymous iSNVs also appear to fall within T cell epitopes, we predicted T cell epitopes in each LASV protein (Supplemental Experimental Procedures). We found that nonsynonymous iSNVs accumulated to some extent in LASV GPC, although the results did not reach statistical significance (Figure 6C; p value = 0.07; binomial test).

## Intrahost Variants Interfere with Antibody Binding

To investigate the functional effects of a subset of LASV iSNVs, we created iSNV mutations in predicted B cell epitopes in GP1 (Supplemental Experimental Procedures), expressed them in HEK293 cells, and tested their binding to a panel of GPC-specific monoclonal antibodies (mAbs) using flow cytometry. These mutations led to a significant drop in the average mean fluorescence intensity (MFI) for GP1-specific mAbs (Figure 6D), consistent




C

| All LASV iSNVs[1] | | Fixed? | | Fisher test | |
|---|---|---|---|---|---|
| | Class | No | Yes | Odds ratio | *P*-value |
| *Minor allele count* | NS | 32 | 81 | 35.3 | 6.92E-16 |
| | S | 3 | 271 | | |

| All EBOV iSNVs[1] | | Fixed? | | Fisher test | |
|---|---|---|---|---|---|
| | Class | No | Yes | Odds ratio | *P*-value |
| *Minor allele count* | NS | 21 | 3 | 4.04 | 4.27E-02 |
| | S | 22 | 13 | | |

| Z0947 LASV iSNVs[2] | | DAF | | Fisher test | |
|---|---|---|---|---|---|
| | Class | >50% | <50% | Odds ratio | *P*-value |
| *Z0948 allele count* | NS | 7 | 0 | >16 | 1.65E-03 |
| | S | 5 | 13 | | |

**Figure 7. Biased Fixation of Nonsynonymous iSNVs**

(A) iSNVs that are never observed as fixed differences between consensus sequences have a higher N/S ratio in both LASV and EBOV.

(B) LASV iSNVs are more commonly seen as fixed differences than are EBOV iSNVs. The data displayed in (A) and (B) are tabulated in the top two panels of (C).

(C) Biased fixation of iSNVs at the population-wide level ("All iSNVs") and in a pair of *M. natalensis* ("Z0947 iSNVs"). At the population level (top and middle tables), the "Fixed?" column indicates whether or not the minor iSNV allele is observed in any other LASV (top) or EBOV (middle) consensus sequences.

The "DAF" columns indicate the derived allele frequency in Z0947, with derived/ancestral allele states inferred from Z0948. [1]Fixation criterion: the minor iSNV is fixed (100%) in one or more other consensus sequences. [2]The DAF is defined as the frequency in Z0947 of the allele not fixed in the Z0948 consensus sequence. See also Figure S7H and Table S2.

with diminished mAb binding. Similarly, when we investigated the effects of single-point mutations within GP1 epitopes, we found that minor alleles in the LASV population displayed significantly reduced binding to GP1-specific mAbs (Figure 6E).

These observations suggest that the host adaptive immune system imposes selective pressures on the intrahost viral population, driving an accumulation of nonsynonymous iSNVs in LASV GPC.

### Nonsynonymous Lassa Virus Intrahost Variants Tend Not to Become Fixed in Other Hosts

To further explore the evolution of LASV within and between hosts, we investigated how often iSNVs become fixed in other consensus sequences. We defined an iSNV as "fixed" if its minor allelic variant was observed in one or more LASV consensus sequences. We observed a significantly higher nonsynonymous-to-synonymous ratio (N/S) for unfixed compared to fixed iSNVs (Figure 7A), suggesting a selective bias against the fixation of nonsynonymous iSNVs. LASV and EBOV both have similar numbers of unfixed iSNVs, but LASV has many more fixed iSNVs, likely due to higher rates of iSNV fixation (or transmission) in LASV than in EBOV (Figure 7B). However, the putative transmitted (fixed) iSNVs tend to be biased toward synonymous mutations. This bias is much stronger in LASV (Figure 7C, top) but still detectable in EBOV (Figure 7C, middle). The bias cannot be attributed to differences in minor allele frequencies between nonsynonymous and synonymous iSNVs (p value > 0.1; Kolmogorov-Smirnov test) or to a correlation between MAF and prevalence in consensus sequences (p value > 0.1 for both N and S; Pearson's correlation); therefore, it is best attributed to selection against transmission and/or fixation of nonsynonymous iSNVs.

A single suspected transmission event, between a pair of *M. natalensis* captured on the same day from the same household, provided an opportunity to observe iSNV fixation dynamics on short timescales. The two samples, Z0947 and Z0948, are nearest-neighbors on the LASV phylogeny (Data S1, S2, and S3), suggesting recent (but not necessarily direct) transmission (Supplemental Experimental Procedures). Assuming that transmission occurred from Z0948 to Z0947, we observed that derived alleles reaching high frequency (DAF > 0.5) in Z0947 tended to be nonsynonymous, while derived alleles remaining at lower frequency (DAF < 0.5) were always synonymous (Figure 7C, bottom). Other transmission scenarios (Figure S7H; Supplemental Experimental Procedures) also confirm that nonsynonymous iSNVs reach high frequency within a host, but fail to be transmitted to the next host. Along with the dN/dS and epitope analyses, this supports a model in which nonsynonymous iSNVs rise to high frequency within an individual due to positive selection, but are less likely to become fixed in other hosts due to purifying selection.

## DISCUSSION

### Comparing Ebola Virus and Lassa Virus Evolutionary Dynamics

EBOV and LASV are RNA viruses that can lead to illnesses with similar clinical symptoms, yet they differ markedly in their epidemiology and evolutionary dynamics. LASV is more than an order of magnitude more diverse than EBOV (Figure 2B), and molecular dating suggests that it has been circulating in Nigeria for over a thousand years, followed by a more recent spread across West Africa (Figure 2G). In contrast, it has been suggested that the Makona variant of EBOV responsible for the EVD epidemic in West Africa was introduced over the last decade (Dudas and Rambaut, 2014; Calvignac-Spencer et al., 2014; Gire et al., 2014).

These analyses, however, provide lower-bound tMRCA estimates of sampled (extant) viral lineages; the true ages of all LASV and EBOV lineages are likely much older (Taylor et al., 2010). Because of limited sampling from Guinea and Liberia,

our LASV dating analysis is likely most accurate within Sierra Leone, although we achieved comparable results when considering the entire dataset or the individual regions alone (Figure S3E). Furthermore, our 400-year-old "out of Nigeria" estimate relies on a single sequence from the Ivory Coast; additional sampling outside the MRU and Nigeria could push back this date.

Because of the high heterogeneity among LASV lineages, continuous monitoring of its mutational spectrum and evolutionary change will be critical for the development of effective vaccines and diagnostics. Since LASV strains cluster by geography, it is more conserved within individual countries. For example, average sequence identity among lineages from Sierra Leone is 90% at the nucleotide level and 95% at the amino acid level (Figure S1F). A useful strategy might therefore be to develop diagnostics, vaccines, and sequence-specific therapeutics that are country specific or that target the most conserved features of the viral genome.

The 2013–2015 West African EVD epidemic likely originated from a single zoonotic transmission event (Baize et al., 2014), followed by sustained human-to-human transmission and clock-like, linear accumulation of mutations (Figure 2C). In contrast, LASV has a clock-like signature on the timescale of decades (Figure S4B), but not on shorter timescales (Figure 2D). Combined with the intermingling of human and *M. natalensis* samples on the phylogenetic tree (Figure 2A), this is consistent with a genetically diverse pool of LASV being maintained in its rodent reservoir, with most human infections caused by genetically distinct viruses. A recent study suggested that human-to-human transmission of LASV may account for up to 20% of all cases (Lo Iacono et al., 2015), but we found little support for this in our dataset. This does not rule out human-to-human transmission entirely, but it suggests that human transmission chains are the exception rather than the rule.

LASV is more polymorphic within hosts than EBOV, and *M. natalensis* hosts harbor more polymorphic LASV populations than humans (Figures 5B and 5C). Since most LF and EVD patients have 0 or 1 iSNVs, the difference difference between LASV and EBOV is mostly driven by a subset of LF patients with many LASV iSNVs (Figure 5B and 5C). LASV iSNV frequencies tend to remain stable over the relatively short period of hospitalization (Figure S7E), suggesting that intrahost de novo mutations and frequency changes may take time to develop or may occur early in the infection. These observations suggest that—at least in some patients—LASV infections could last longer than EBOV infections, allowing more time for the generation of polymorphism within hosts.

Longer infections also provide more time for natural selection to eliminate deleterious mutations from the viral population. Consistent with longer infection periods in LASV, dN/dS ratios are lower within LF patients than in EVD patients (Figure 5D).

While these findings are consistent with the existence of chronic LASV infections in humans, they do not constitute proof. Further studies are needed to verify the causes of high-diversity LASV infection and the prevalence of nonacute human infections. Compelling evidence could come from longitudinal sampling asymptomatic carriers of LASV, for example.

## Codon Adaptation, Translational Efficiency, and Genome Abundance

We uncovered significant differences in genome abundance, CFR, CAI, and translational efficiency of LASV strains from Sierra Leone and from Nigeria. The increase in CAI of non-Nigerian strains (Figures 3B–3D), along with higher viral copy numbers and CFRs in human patients (Figures 4A–4D), would seem to suggest that the virus evolved toward greater human virulence. There are indeed many examples suggesting that pathogens with natural reservoirs may evolve toward greater human virulence, so long as they remain avirulent in the reservoir host (Ewald, 1994). We have not, however, been able to establish causality among these observations. While clinical care for LF patients is similar at ISTH and KGH (Supplemental Experimental Procedures)—and time from symptom onset to hospitalization appears to be the same (Figure 4E)—several additional parameters are beyond our control, including variations in socioeconomic, clinical, and human genetic factors between Sierra Leone and Nigeria. These prevent us from determining whether the difference in CFRs has a viral genetic basis and whether variations in CAI, viral genome abundance, and CFRs are causally linked. In addition, while we observed significantly higher LASV genome abundance in LF patients from Sierra Leone, we could not determine whether this difference translates into higher infectious titers. Controlled animal studies in BL-4 laboratories comparing strains from the two countries would help to resolve this important question.

We expected that the increased CAI of Sierra Leonean LASV lineages would lead to increased translational output in an experimental system. Surprisingly, the opposite was true. Nigerian LASV lineages had significantly higher in vitro protein outputs than their Sierra Leonean counterparts (Figures 4G–4I). Translation-independent mechanisms, such as post-translational modifications and protein stability, could explain these observations, regardless of CAI differences.

If the progenitor to Sierra Leone LASV strains indeed had lower translational output, the emergence of LASV strains with increased CAI could have been driven by positive selection to compensate. The higher CAI could then have led to higher viral titers in human patients. Alternatively, the increase in CAI could have been caused by genetic drift. Under this scenario, mutational biases (Supplemental Experimental Procedures) and drift—combined with insufficient time for mutations to be exposed to purifying selection in Sierra Leone—led to an increase in CAI, independently of positive selection. The Sierra Leonean tMRCA of the LASV population is indeed relatively recent (Figure 2G), having possibly undergone a recent population bottleneck (Lalis et al., 2012). Further work will be required to disentangle the adaptive and neutral contributions to the evolution of CAI and determine whether changes in CAI affect viral fitness and CFRs.

## Immune Selection on Lassa Virus within Hosts

The observation that LASV GPC has the highest intrahost dN/dS (Figure 5E) and the most nonsynonymous iSNVs in predicted epitopes (Figure 6) suggests that it is a target of immune-driven diversifying selection within hosts, both in humans and *M. natalensis*. This effect was most strongly supported when

we investigated the involvement of B cells (Figure 6A), although we also found weaker evidence for the role of T cells (Figure 6C).

Given that human LASV infections are typically thought to be of short duration, it is perhaps surprising that there would be enough time for the host to mount an antibody response, and for viral diversity to be measurably shaped by this selective pressure. However, this is not entirely implausible, since it is known that LASV-specific antibodies appear relatively quickly, with experimentally infected nonhuman primates developing detectable IgM titers after 9 days and IgG titers after 12 days (Baize et al., 2009). Furthermore, it has also been shown that 25% of LF patients in Sierra Leone are IgM-positive upon admission and 10% are IgG-positive (Shaffer et al., 2014). Combined with the observation that several LF patients have LASV iSNV numbers similar to those observed in *M. natalensis* (Figures 5B and 5C), sufficient time for antibody-mediated responses has likely elapsed in a number of LF patients when they present at the hospital. We were unable to test seropositivity in our cohort, but future studies could assess whether there is a correlation between the presence or absence of LASV-specific antibodies and the number of iSNVs. Our prediction would be that LF patients with high numbers of LASV iSNVs should also tend to be IgM- and/or IgG-positive upon admission.

## Conclusions

Our dataset of full-length LASV genomes yields insights into the biogeography, evolution, and spread of a hemorrhagic fever-causing virus. Further improvements in collecting, sequencing, and analyzing LASV and EBOV samples combined with concurrent experiments in BL-4 laboratories will allow us to pursue open areas of inquiry. The catalog of LASV genetic diversity presented here is thus an essential foundation for the development of vaccines and diagnostics for a highly diverse, continuously evolving, and unusually prevalent BL-4 agent.

## EXPERIMENTAL PROCEDURES

### Ethics Statement

Subjects were recruited for this study using protocols approved by relevant human subjects committees. All patients were treated with a similar standard of care.

### Samples

All samples were acquired on the day of admission before any treatment regimens had begun. 10 ml of whole blood was collected, and plasma or serum was prepared. Diagnostic tests for the presence of LASV were performed on-site. Rodents (all from Sierra Leone) were trapped in case households and humanely sacrificed, and samples were collected from serum and/or spleen. All samples were inactivated in either Buffer AVL (QIAGEN) or TRIzol (Life Technologies), following the manufacturer's standard protocols and stored in solar-driven −20°C freezers.

### Case Fatality Rates

CFRs were calculated from patients that had all of the following characteristics: (1) known outcome, (2) positive for LASV in the field, (3) confirmed positive upon retesting in the United States, and (4) sequencing confirmed the presence of LASV.

### Sequencing

cDNA synthesis and Illumina library construction were performed, libraries were pooled, and sequenced on the Illumina platform. All the data were deposited at NCBI (BioProject PRJNA254017). EBOV data are available under PRJNA257197.

### Assembly of LASV Genomes

LASV genomes were de novo assembled, followed by read mapping and duplicate removal.

### Trees

Maximum likelihood phylogenies were made with RAxML (v.7.3.0) (Stamatakis et al., 2005) and rooted using the 1969 Pinneo strain.

### Molecular Dating

BEAST (v.1.7.4) (Drummond and Rambaut, 2007) was used with a model incorporating a log normal relaxed clock, exponential growth, HKY$\gamma$i with four categories, and codon partitioning (srd06).

### Intrahost Variants

iSNVs were called and filtered taking into consideration minimum coverage, minimum frequency (5%), strand-bias, and base quality.

### Intrahost Selection

A modified version of the McDonald-Kreitman test (McDonald and Kreitman, 1991) was applied to assess evidence for natural selection.

### Epitope Prediction

B cell epitopes were predicted using BCPRED (El-Manzalawy et al., 2008). T cell epitopes were predicted using NetCTL (Larsen et al., 2007).

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, two tables, and three data files and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.07.020.

## AUTHOR CONTRIBUTIONS

All authors contributed in the development of this work. K.G.A., E.P., O.A.F., S.H.K., C.T.H., R.F.G., and P.C.S. initiated the study; K.G.A., B.J.S., and P.C.S. conceptualized the study; K.G.A., C.B.M., A.E.L., S.K.G., E.M.E., S.W., A.B.G., A.L., C.M., A.I., B.T.-V., E.M.R., I.O., P.E.E., A.G., Z.B., L.M.B., M.R., and M.H. performed experimental work; K.G.A., B.J.S., R.S., A.E.L., H.K.F., and M.N. analyzed the data; K.G.A., S.K.G., L.M.M., M.S., R.T., E.P., L.M.B., S.J., D.L., D.S.G., A.G., R.F., K.K., M.M., M.F., G.OA., D.A.A., P.O.O., W.O., I.O., P.E.E., O.F., C.T.H., R.F.G., and P.C.S. performed field work; L.M.M. led the rodent collections; and K.G.A., B.J.S., C.B.M., R.S., S.F.S., and P.C.S. wrote the paper.

## ACKNOWLEDGMENTS

The authors thank C. Edwards, X. Yang, M. Zody, B. Han, A. Andersen, I. Shlyakhter, D. Park, M. Henn, M. Busby, M. Kellis, C. Bishop, A. Haislip, L. Burchfield, A. Matthews, and F. Viloria for their help and advice. We also thank the Ministry of Health and Sanitation Government of Sierra Leone, Oyo State Ministry of Health, and the Federal Ministry of Health, Nigeria. A full list of members of the Viral Hemorrhagic Fever Consortium can be found at www.vhfc.org. This project was funded with federal funds from the NIH, Department of Health and Human Services (1DP2OD006514-01) and NIAID Contracts (HHSN272200900049C, HHSN272201000022C, HHSN272200900018C, and U19AI110818). The authors received additional support from USAMRAA (W81XWH-10-1-0098), a Packard Foundation Fellowship for Science and Engineering, a Broad Institute SPARC award, and the German Research Foundation (GU 883/1-1). K.G.A. was supported by a fellowship from the Carlsberg Foundation; B.J.S. was supported by a Harvard MIDAS CCDD postdoctoral fellowship and Canada Research Chair; and R.S. and A.E.L. received NSF GRFP fellowships (DGE 1122374 and 1144152). Tulane University and industry

partners have filed patent applications for several Lassa-related technologies and may receive royalties or profits if commercial products are developed.

Received: February 12, 2015
Revised: April 26, 2015
Accepted: June 12, 2015
Published: August 13, 2015

## REFERENCES

Andersen, K.G., Shylakhter, I., Tabrizi, S., Grossman, S.R., Happi, C.T., and Sabeti, P.C. (2012). Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 868–877.

Anisimova, M., and Liberles, D.A. (2007). The quest for natural selection in the age of comparative genomics. Heredity *99*, 567–579.

Bahir, I., Fromer, M., Prat, Y., and Linial, M. (2009). Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol. Syst. Biol. *5*, 311.

Baize, S., Marianneau, P., Loth, P., Reynard, S., Journeaux, A., Chevallier, M., Tordo, N., Deubel, V., and Contamin, H. (2009). Early and strong immune responses are associated with control of viral replication and recovery in lassa virus-infected cynomolgus monkeys. J. Virol. *83*, 5890–5903.

Baize, S., Pannetier, D., Oestereich, L., Rieger, T., Koivogui, L., Magassouba, N., Soropogui, B., Sow, M.S., Keïta, S., De, C., et al. (2014). Emergence of Zaire Ebola virus disease in Guinea — preliminary report. N. Engl. J. Med. *371*, 1418–1425.

Baum, J., Thomas, A.W., and Conway, D.J. (2003). Evidence for diversifying selection on erythrocyte-binding antigens of Plasmodium falciparum and P. vivax. Genetics *163*, 1327–1336.

Bhatt, S., Holmes, E.C., and Pybus, O.G. (2011). The genomic rate of molecular adaptation of the human influenza A virus. Mol. Biol. Evol. *28*, 2443–2451.

Bowen, M.D., Rollin, P.E., Ksiazek, T.G., Hustad, H.L., Bausch, D.G., Demby, A.H., Bajani, M.D., Peters, C.J., and Nichol, S.T. (2000). Genetic diversity among Lassa virus strains. J. Virol. *74*, 6992–7004.

Butt, A.M., Nasrullah, I., and Tong, Y. (2014). Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. PLoS ONE *9*, e90905.

Calvignac-Spencer, S., Schulze, J.M., Zickmann, F., and Renard, B.Y. (2014). Clock rooting further demonstrates that Guinea 2014 EBOV is a member of the Zaïre lineage. PLoS Curr. *6*, 6.

Carroll, S.A., Towner, J.S., Sealy, T.K., McMullan, L.K., Khristova, M.L., Burt, F.J., Swanepoel, R., Rollin, P.E., and Nichol, S.T. (2013). Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences. J. Virol. *87*, 2608–2616.

Commission, R.O.A.I. (1978). Ebola haemorrhagic fever in Zaire, 1976. Bull. World Health Organ. *56*, 271–293.

Djavani, M., Lukashevich, I.S., Sanchez, A., Nichol, S.T., and Salvato, M.S. (1997). Completion of the Lassa fever virus sequence and identification of a RING finger open reading frame at the L RNA 5′ end. Virology *235*, 414–418.

Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. *7*, 214.

Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. *29*, 1969–1973.

Dudas, G., and Rambaut, A. (2014). Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak. PLoS Curr. *6*, 6.

El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008). Predicting linear B-cell epitopes using string kernels. J. Mol. Recognit. *21*, 243–255.

Ewald, P.W. (1994). Evolution of Infectious Disease (New York: Oxford University Press).

Fichet-Calvet, E., and Rogers, D.J. (2009). Risk maps of Lassa fever in West Africa. PLoS Negl. Trop. Dis. *3*, e388.

Frame, J.D., Baldwin, J.M.J., Jr., Gocke, D.J., and Troup, J.M. (1970). Lassa fever, a new virus disease of man from West Africa. I. Clinical description and pathological findings. Am. J. Trop. Med. Hyg. *19*, 670–676.

Gire, S.K., Stremlau, M., Andersen, K.G., Schaffner, S.F., Bjornson, Z., Rubins, K., Hensley, L., McCormick, J.B., Lander, E.S., Garry, R.F., et al. (2012). Epidemiology. Emerging disease or diagnosis? Science *338*, 750–752.

Gire, S.K., Goba, A., Andersen, K.G., Sealfon, R.S., Park, D.J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., et al. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science *345*, 1369–1372.

Hershberg, R., and Petrov, D.A. (2008). Selection on codon bias. Annu. Rev. Genet. *42*, 287–299.

Holmes, E.C. (2003). Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J. Virol. *77*, 11296–11298.

Lalis, A., Leblois, R., Lecompte, E., Denys, C., Ter Meulen, J., and Wirth, T. (2012). The impact of human conflict on the genetics of Mastomys natalensis and Lassa virus in West Africa. PLoS ONE *7*, e37068.

Larsen, M.V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O., and Nielsen, M. (2007). Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinformatics *8*, 424.

Lauring, A.S., Acevedo, A., Cooper, S.B., and Andino, R. (2012). Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. Cell Host Microbe *12*, 623–632.

Lecompte, E., Fichet-Calvet, E., Daffis, S., Koulémou, K., Sylla, O., Kourouma, F., Doré, A., Soropogui, B., Aniskin, V., Allali, B., et al. (2006). Mastomys natalensis and Lassa fever, West Africa. Emerg. Infect. Dis. *12*, 1971–1974.

Lipkin, W.I. (2013). The changing face of pathogen discovery and surveillance. Nat. Rev. Microbiol. *11*, 133–141.

Lo Iacono, G., Cunningham, A.A., Fichet-Calvet, E., Garry, R.F., Grant, D.S., Khan, S.H., Leach, M., Moses, L.M., Schieffelin, J.S., Shaffer, J.G., et al. (2015). Using modelling to disentangle the relative contributions of zoonotic and anthroponotic transmission: the case of lassa fever. PLoS Negl. Trop. Dis. *9*, e3398.

Lukashevich, I.S. (1992). Generation of reassortants between African arenaviruses. Virology *188*, 600–605.

McCormick, J.B., and Fisher-Hoch, S.P. (2002). Lassa fever. Curr. Top. Microbiol. Immunol. *262*, 75–109.

McCormick, J.B., King, I.J., Webb, P.A., Scribner, C.L., Craven, R.B., Johnson, K.M., Elliott, L.H., and Belmont-Williams, R. (1986). Lassa fever. Effective therapy with ribavirin. N. Engl. J. Med. *314*, 20–26.

McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652–654.

Paessler, S., and Walker, D.H. (2013). Pathogenesis of the viral hemorrhagic fevers. Annu. Rev. Pathol. *8*, 411–440.

Pandey, A., Atkins, K.E., Medlock, J., Wenzel, N., Townsend, J.P., Childs, J.E., Nyenswah, T.G., Ndeffo-Mbah, M.L., and Galvani, A.P. (2014). Strategies for containing Ebola in West Africa. Science *346*, 991–995.

Parameswaran, P., Charlebois, P., Tellez, Y., Nunez, A., Ryan, E.M., Malboeuf, C.M., Levin, J.Z., Lennon, N.J., Balmaseda, A., Harris, E., and Henn, M.R. (2012). Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity. J. Virol. *86*, 8546–8558.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. *12*, 32–42.

Rocha, E.P., Smith, J.M., Hurst, L.D., Holden, M.T., Cooper, J.E., Smith, N.H., and Feil, E.J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. *239*, 226–235.

Shaffer, J.G., Grant, D.S., Schieffelin, J.S., Boisen, M.L., Goba, A., Hartnett, J.N., Levy, D.C., Yenni, R.E., Moses, L.M., Fullah, M., et al.; Viral Hemorrhagic Fever Consortium (2014). Lassa fever in post-conflict sierra leone. PLoS Negl. Trop. Dis. *8*, e2748.

Shapiro, B.J., David, L.A., Friedman, J., and Alm, E.J. (2009). Looking for Darwin's footprints in the microbial world. Trends Microbiol. *17*, 196–204.

Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281–1295.

Southern, P. (1996). Arenaviridae: the viruses and their replication. In Fundamental Virology, B.N. Fields, D.M. Knipe, and P.M. Howley, eds. (Philadelphia: Lippincott-Raven), pp. 675–690.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics *21*, 456–463.

Stenglein, M.D., Jacobson, E.R., Chang, L.W., Sanders, C., Hawkins, M.G., Guzman, D.S., Drazenovich, T., Dunker, F., Kamaka, E.K., Fisher, D., et al. (2015). Widespread recombination, reassortment, and transmission of unbalanced compound viral genotypes in natural arenavirus infections. PLoS Pathog. *11*, e1004900.

Taylor, D.J., Leach, R.W., and Bruenn, J. (2010). Filoviruses are ancient and integrated into mammalian genomes. BMC Evol. Biol. *10*, 193.

Team, W.H.O.E.R. (2014). Ebola virus disease in West Africa - the first 9 months of the epidemic and forward projections. N. Engl. J. Med. *371*, 1481–1495.

Troup, J.M., White, H.A., Fom, A.L., and Carey, D.E. (1970). An outbreak of Lassa fever on the Jos plateau, Nigeria, in January-February 1970. A preliminary report. Am. J. Trop. Med. Hyg. *19*, 695–696.

Tulloch, F., Atkinson, N.J., Evans, D.J., Ryan, M.D., and Simmonds, P. (2014). RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. eLife *3*, e04531.

Vieth, S., Torda, A.E., Asper, M., Schmitz, H., and Günther, S. (2004). Sequence analysis of L RNA of Lassa virus. Virology *318*, 153–168.

Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends Ecol. Evol. *11*, 158–162.

Wertheim, J.O., and Kosakovsky Pond, S.L. (2011). Purifying selection can obscure the ancient age of viral lineages. Mol. Biol. Evol. *28*, 3355–3365.

Wertheim, J.O., Chu, D.K., Peiris, J.S., Kosakovsky Pond, S.L., and Poon, L.L. (2013). A case for the ancient origin of coronaviruses. J. Virol. *87*, 7039–7045.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

| APPLICATION NO. | FILING DATE | FIRST NAMED INVENTOR | ATTORNEY DOCKET NO. | CONFIRMATION NO. |
|---|---|---|---|---|
| 14/212,469 | 03/14/2014 | Ramesh Buyyarapu | P11485.2US(72022-US-NP2) | 1039 |

25212 7590 03/19/2019

DOW AGROSCIENCES LLC
9330 ZIONSVILLE RD
INDIANAPOLIS, IN 46268

| EXAMINER |
|---|
| STANKOVIC, BRATISLAV |

| ART UNIT | PAPER NUMBER |
|---|---|
| 1663 | |

| NOTIFICATION DATE | DELIVERY MODE |
|---|---|
| 03/19/2019 | ELECTRONIC |

**Please find below and/or attached an Office communication concerning this application or proceeding.**

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

PatentsUS@pioneer.com

UNITED STATES PATENT AND TRADEMARK OFFICE

---

BEFORE THE PATENT TRIAL AND APPEAL BOARD

---

*Ex parte* RAMESH BUYYARAPU, RUIHUA REN, MUSTAFA MCPHERSON, SIVA P. KUMPATLA, CHANDRA CHANNABASAVARAHYA, JOE SPINKS, and KELLY PARLIAMENT

---

Appeal 2018-006665
Application 14/212,469
Technology Center 1600

---

Before ULRIKE W. JENKS, TIMOTHY G. MAJORS, and MICHAEL A. VALEK, *Administrative Patent Judges*.

VALEK, *Administrative Patent Judge.*

DECISION ON APPEAL

Appellants[1] submit this appeal under 35 U.S.C. § 134(a) involving claims to methods for producing a cotton plant comprising reniform nematode (RN) resistance. The Examiner rejected the claims on the basis

---

[1] Appellants identify the real party in interest as Dow AgroSciences, LLC. App. Br. 2. Herein we refer to the Final Office Action mailed April 12, 2017 ("Final Act."), Response After Final Action filed June 12, 2107 ("Response After Final"), Appeal Brief filed November 14, 2017 ("App. Br."), Examiner's Answer mailed April 12, 2018 ("Ans."), and Reply Brief filed June 12, 2018 ("Reply Br.").

that they contain an improper Markush grouping. We have jurisdiction under 35 U.S.C. § 6(b).

We REVERSE.

STATEMENT OF THE CASE

Claims 12, 16–19, 23, 33, 34, 36 and 37 are on appeal and can be found in the Claims Appendix of the Appeal Brief. Claims 16–19, 23, 33, 34, 36 and 37 all depend from claim 12. Claim 12 contains the Markush group noted in italics below:

> 12. A method for producing a cotton plant comprising reniform nematode (RN) resistance, the method comprising:
>
> crossing a first parental cotton plant comprising the trait of RN resistance with a second parental cotton plant that is sensitive to RN infection as compared to the first parental cotton plant, to produce progeny cotton plants, wherein the first parental cotton plant comprises at least one marker that is linked to the RN resistance trait, the marker being *selected from the group consisting of SEQ ID NOs:58-62*, and wherein the second parental cotton plant does not comprise the marker that is linked to the RN resistance trait;
>
> isolating genomic DNA from the progeny cotton plants;
>
> screening the genomic DNA of the progeny cotton plants for the presence of the marker that is linked to the RN resistance trait; and
>
> selecting a progeny cotton plant having genomic DNA comprising the marker that is linked to the RN resistance trait, thereby producing a cotton plant comprising RN resistance.

App. Br. 16 (emphasis added).

The claimed Markush group consists of five distinct nucleotide sequences listed as SEQ ID NOs:58-62. These sequences range from about 300–600 nucleotides in length and are referred to in the Specification as DASCTP_28910_164, DC7_56523319, DCTE1_240981_97, DCTE1_317966_63 and DASCTP_1656_527. Ans. 6–7; App. Br. 3. Like

all nucleotide sequences, the markers corresponding to SEQ ID NOs:58-62 share a common sugar–phosphate backbone. But, as Appellants acknowledge, their primary structure, *i.e.*, the "series of nucleobases" joined by that common sugar-phosphate backbone, is "different." App. Br. 11.

Examiner determined that the Markush grouping in claim 12 is improper and finally–rejected Appellants' claims on that basis. There are no other rejections before us on appeal.[2]

Accordingly, the issue is: Does the evidence of record support Examiner's determination that claim 12, and claims 16–19, 23, 33, 34, 36 and 37 by dependency, contain an improper Markush grouping?

*Analysis*

According to the MPEP, "[a] Markush claim contains an 'improper Markush grouping' if either: (1) the members of the Markush group do not share a 'single structural similarity' or (2) the members do not share a common use." MPEP § 706.03(y)(II). Moreover, where the members do not belong to a recognized physical, chemical or art-recognized class, "the common use must flow from the substantial structural feature." MPEP § 706.03(y)(II)(B). Whether a Markush group is proper depends on the particular facts at issue and "must be decided on a case-by-case basis." MPEP § 706.03(y)(IV).

Examiner finds that the markers in the five member Markush group here "do not share a substantial [structural] feature and/or common use that flows from the substantial structural feature," because "they have no conserved structure throughout the genus other than a phosphodiester

[2] Claim 35 was rejected for indefiniteness, but has since been cancelled. Response after Final 5.

backbone." Final Act. 3–4. Examiner also determines that they share no common use because they "have different effects on trait expression, as evidenced by the different LOD numbers and different % explanation." Ans. 11 (referring to Spec. ¶ 253, Table 2).

Appellants argue that, in addition to their common sugar-phosphate backbone, the claimed markers share a structural feature because they are all located in "close physical proximity to the RN resistance locus on chromosome 21." App. Br. 12. According to Appellants, this feature is "directly responsible for their shared function of being linked to RN resistance, because their proximity results in the co-segregation of the markers with the RN resistance." *Id.* Moreover, according to Appellants, the fact that the markers have different nucleotide sequences is, in fact, integral to their common use because it allows them to be used collectively in "common assays" to "provide[] a better result than their independent use." *Id.* at 13. Thus, Appellants urge that Examiner erred by applying "a hard rule that different nucleotide sequences cannot be presented" in a Markush group, rather than considering the structure of the claimed markers as a whole in light of the particular facts of this case. Reply 6–8.

On the record before us, we determine that Appellants have the better position. The MPEP makes clear that the propriety of a particular Markush grouping is a fact-specific inquiry that "must" be decided on a case-by-case basis. MPEP § 706.03(y)(IV); *see also In re Harnisch*, 631 F.2d 716, 722 (CCPA 1980) ("[W]e decide this and like cases on their facts on a case-by-case basis."). To the extent Examiner applied a bright-line rule that focused solely on the sequence differences without consideration of other structural similarities, that approach is incorrect. *See Harnish*, 631 F.2d at 722 ("[I]n

determining the propriety of a Markush grouping the compounds must be considered as wholes and not broken down into elements or other components."); *see also* MPEP § 706.03 (y)(IV). Accordingly, we look to whole of the structure in the context of the particular facts of the claimed invention to determine whether the nucleotide sequences in SEQ ID NOs:58-62 are properly grouped together.

Here, the record supports, and Examiner does not dispute, that the claimed markers were selected out of a larger set of identified markers and grouped together because these five sequences are particularly tightly linked to the RN resistance trait locus in cotton. *See* Spec. ¶ 251 ("The major QTL[3] region was concentrated at the distal end of the chromosome, between 0 and 32.63 cM, with a maximum LOD score of 13.93 observed between DASCTP_28910_164 and DASCTP_1656_527. This major QTL explained 29.8% of the variation in the resistance phenotype."). The Specification further teaches that all five of these markers are physically located proximate to each other in a particular region, *i.e.*, the distal end between about 29–32 cM in chromosome 21. *Id.* at ¶¶ 251; 253, Table 2. Accordingly, in addition to sharing a sugar-phosphate backbone, each member of the claimed Markush group shares an additional structural feature, *i.e.*, their specific physical proximity to each other and overall location within cotton genome.

The record further demonstrates that each of the five group members shares a common use that directly flows from this shared structural similarity. The Specification teaches that "[i]n general, the closer one marker is to another marker or gene . . . the more tightly they are linked."

---

[3] According to the Specification, "QTL" stands for "quantitative trait locus." Spec. ¶ 44.

Spec. ¶ 89. Thus, the physical location of the claimed sequences is at least partially responsible for their utility as markers to screen for RN resistance in cross-bred cotton plants. In addition, the Specification reports that the five members of this Markush group have the highest LOD and % Explanation values out of a much larger set of markers identified though Appellants' experiments. *Id.* ¶ 154, Table 2. That data supports Appellants' argument that these sequences are particularly suited for use as markers to be used, alone or in combination, to screen for RN resistance in the cross-bred progeny plants of the claimed method. *See* Reply 7.

We disagree with Examiner's argument that there is no common use because the Specification reports slightly different LOD and % Explanation values for each of the five markers in the claimed group. *See* Ans. 11. At most, the data evidences that some members of the Markush group may be slightly more indicative of RN resistance than others. That is a difference in degree, not kind. There is no evidence, nor does Examiner find, that the members of the claimed Markush group are not alternatively useful as markers for RN resistance. For all of these reasons, we determine that Examiner's rejection is not supported by the evidence of record and therefore reverse.

## SUMMARY

We reverse the rejection of claims 12, 16–19, 23, 33, 34, 36 and 37 on the basis that they contain an improper Markush group.

## REVERSED

**In re Harnisch 631 F.2d 716 (C.C.P.A. 1980)**
(United States Court of Customs and Patent Appeals)

Prepared by UNCTAD's Intellectual Property Unit

## Case summary

The United States Court of Customs and Patent Appeals (CCPA), which was the predecessor of the United States Court of Appeals for the Federal Circuit reversed a decision by the United States Patent Office (USPTO) to reject a patent application on the basis of alleged improper use of the "Markush claims" format. The CCPA confirmed unity of the claimed invention at issue and therefore considered the Markush claims format as being properly used.

## The facts

The appellant, Harnisch, applied for a product patent on a group of coumarin compounds used as dyestuffs for synthetic or natural fibers as well as plastics and liquids. Instead of claiming each chemical entity contained in the group by its structure, Harnisch limited his claims to a number of general structural characteristics shared by all coumarin compounds, leaving a large number of compounds unspecified ("Markush claims", in reference to the first user of this claims format). The USPTO rejected the application for failing to meet the USPTO's own requirements for properly using the Markush claims format. Harnisch appealed to the CCPA. The CCPA reversed the decision of the USPTO.

## The legal issues

The main issue in this case was the appropriate application of the requirements to admit Markush claims format as laid down in the USPTO's patent examination guidelines. Markush claims are a sub-category of structural claims. This claims format enables the inventor to claim a large number of unspecified compounds along with a specified general structure and some specified alternatives available under that structure. Unspecified elements may have certain structural differences and untested properties as compared to the specified elements. Their properties are only theoretically inferred from the equivalence with specified compounds within the claim. Depending on the extent to which they are admitted, Markush claims may result in product patents with a very broad scope of protection. The admissibility of Markush claims is not addressed under TRIPS. Before the EPO and USPTO, such claims are generally admissible if all claimed alternatives share a common structure and have a common property or activity.[1]

[1] For more details on this claims format, see WHO-ICTSD-UNCTAD, "Guidelines for the examination of pharmaceutical patents: Developing a public health perspective." Working Paper, by Carlos M. Correa, Geneva, January 2007, pp. 12-14 (http://www.iprsonline.org/resources/docs/Correa_Patentability%20Guidelines.pdf) .

In the present case, the Court specifically examined whether the recourse by the appellant to the use of the Markush claims format was in disregard of the administrative requirement of unity of invention. Many national patent laws provide that one patent application can only cover one invention or a group of closely related inventions. The rationale behind this rule is to facilitate the classification of the patent by the patent office and, importantly, to prevent applicants from avoiding multiple patent fees by filing only one application for a multitude of unrelated inventions. The CCPA in this context emphasized the fact that all claimed alternatives of the appellant's invention were coumarin compounds of similar structure and all may be used as dyes. Accordingly, the Court decided that Harnisch's application was in line with the unity of invention requirement and the use of the Markush claims format was thus admissible. The CCPA affirmed precedent according to which the examination of similarity in structure and properties should not be carried out too rigorously, as there would always be certain differences among the compounds of a claimed group. Rather, an overall assessment should suffice, and the characterization of several compounds as belonging to the same genus should not be "repugnant to scientific classification."[2]

The Court stated that in the past, other courts had raised doubts about the conformity of the Markush claims format with other patent law requirements, in particular the requirements of definiteness and enablement disclosure. The definiteness requirement in national patent laws, which is not expressly referred to in TRIPS, obliges a patent applicant to distinctly claim the subject matter of a patent in a way that defines its boundaries. Markush claims could be problematic in that respect, as they do not specify every claimed compound, but define the boundary of an overall group of compounds.

In addition, the disclosure requirement, which is stated in Article 29.1 of the TRIPS Agreement, requires applicants to disclose the invention in a manner sufficiently clear and complete for the invention to be carried out by a person skilled in the art. The Markush claims format leaves a substantial number of compounds undisclosed. The claims only refer to an overall common structure and a few examples of sub-elements. In the areas of chemistry and biotechnology, the development of sub-compounds with certain properties from an overall common structure may be difficult to predict, thus making it difficult, if not impossible for a skilled expert to make the unspecified compounds based on the patent document. The patent applicant is likely to counter such alleged lack of enabling disclosure by arguing that the undisclosed alternatives will be predictable for a skilled expert to develop. However, such approach could call into question the non-obvious nature of the undisclosed elements, arguably making them fail the non-obviousness test. The patent applicant may thus be caught between the enablement/disclosure requirement on the one hand and the inventive step test on the other. Despite these conceptual problems related to the use of Markush claims, the CCPA limited its analysis to the unity of invention requirement and reversed the decision of the USPTO to refuse the patent grant.

[2] See the decision of the CCPA at p. 5.

**Points of significance**

- The extent to which Markush claims are admitted lies entirely within national discretion. This format is particularly used in countries interested in promoting the R&D-based pharmaceutical industry. In developing countries, whose pharmaceutical industries depend on generic production, broad patent claims as promoted by the use of the Markush formula could result in inappropriate blocking effects for generic competitors. The boundaries of the claimed groups would not be entirely clear for competitors, possibly generating a chilling effect for fear of patent infringement. In addition, the patent, once fallen into the public domain, would be difficult to carry out for a competitor, due to the lack of precise disclosure of the claimed elements.
- The judiciary may play an important role in limiting the use of Markush claims. Based on the enablement disclosure requirement (Article 29.1, TRIPS), the courts may reject claims that extend to elements of the invention that are not specified and supported by the examples given in the patent document. Where unspecified elements are obvious to carry out, the court may in certain cases reject the inventive / non-obvious character of the compound at issue.
- If Markush claims are generally admitted under national law or patent examination guidelines, the courts have to proceed their analysis on a case-by-case basis.

**Key words**
Markush claims; unity of invention; disclosure; enablement; definiteness of claims.

**Available at :** https://casetext.com/case/in-re-harnisch

HHS Public Access

Author manuscript

*Nat Struct Mol Biol*. Author manuscript; available in PMC 2018 May 21.

Published in final edited form as:

*Nat Struct Mol Biol*. 2017 October ; 24(10): 825–833. doi:10.1038/nsmb.3466.

# Guide-bound structures of an RNA-targeting A-cleaving CRISPR-Cas13a enzyme

**Gavin J. Knott**[1,9], **Alexandra East-Seletsky**[1,9], **Joshua C. Cofsky**[1], **James M. Holton**[2,3,4], **Emeric Charles**[1], **Mitchell R. O'Connell**[1,5], and **Jennifer A. Doudna**[1,2,6,7,8,*]

[1]Department of Molecular and Cell Biology, University of California, Berkeley, California, 94720, USA

[2]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, USA

[3]Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, USA

[4]Department of Biochemistry and Biophysics, University of California, San Francisco, USA

[6]Department of Chemistry, University of California, Berkeley, California, 94720, USA

[7]Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA

[8]Innovative Genomics Institute, University of California, Berkeley, California 94720, USA

## Abstract

CRISPR adaptive immune systems protect bacteria from infections by deploying CRISPR RNA (crRNA)-guided enzymes to recognize and cut foreign nucleic acids. Type VI-A CRISPR-Cas systems include the Cas13a enzyme, an RNA-activated ribonuclease (RNase) capable of crRNA processing and single-stranded RNA degradation upon target transcript binding. Here we present the 2.0 Å resolution crystal structure of a crRNA-bound *L. bacterium* Cas13a (LbaCas13a), representing a recently discovered Cas13a enzyme subtype. This structure and accompanying biochemical experiments define for the first time the Cas13a catalytic residues that are directly responsible for crRNA maturation. In addition, the orientation of the foreign-derived target RNA-specifying sequence in the protein interior explains the conformational gating of Cas13a nuclease activation. These results describe how Cas13a enzymes generate functional crRNAs and how

*Correspondence should be addressed to J.A.D (doudna@berkeley.edu).

[5]Present address: Center for RNA Biology and Department of Biochemistry and Biophysics, School of Medicine and Dentistry, University of Rochester, Rochester, NY 14642, USA.

[9]These two authors contributed equally.

**ONLINE CONTENT**

The atomic coordinates and structure factors for the *Lba*Cas13a-crRNA complexes have been deposited to the Protein Data Bank (PDB) under the accession codes 5W1H for LbaCas13a:crRNA complex (24-nt spacer), 5W1I for LbaCas13a:crRNA complex (20-nt spacer) and 5WLH for LbaCas13a(H328A):pre-crRNA complex (24-nt spacer).

**AUTHOR CONTRIBUTIONS**

G.J.K, A.E-S, M.R.O and J.A.D. designed the study. G.J.K., A.E-S., J.C.C. and E.C. expressed and purified proteins. G.J.K prepared and crystallized the Cas13a complexes with assistance from J.C.C, collected X-ray data, and determined the crystal structure with support from J.H. A.E.S carried out biochemical assays with assistance from G.J.K and J.C.C. G.J.K drafted the manuscript, and all authors reviewed and edited the manuscript.

catalytic activity is blocked prior to target RNA recognition, with implications for both bacterial immunity and diagnostic applications.

## INTRODUCTION

A repertoire of defensive mechanisms evolved to protect bacteria against phage and other mobile genetic elements [1]. Unique among these mechanisms are CRISPR-Cas (clustered regularly interspaced short palindromic repeats, CRISPR-associated) adaptive immune systems, which use RNA-guided nucleases to direct cleavage of invading DNA or RNA [2,3]. The programmability of DNA-targeting CRISPR-Cas enzymes has driven a revolution in technologies for gene editing, transcriptional control, and genomic imaging [4–6]. Similarly, RNA-targeting CRISPR-Cas systems offer the potential for RNA-programmed RNA detection and development as scaffolds for designer RNA binding proteins [7–14].

Type VI-A RNA-targeting CRISPR-Cas systems are defined by the single effector protein Cas13a (formerly C2c2), which forms a target surveillance complex with a CRISPR RNA (crRNA). The crRNA, encoded by the genomic CRISPR locus, contains a defined 5′ hairpin structure preceding a 3′ foreign-derived variable region (or 3′ spacer) [8,15]. In Type VI-A CRISPR-Cas systems, Cas13a catalyzes maturation of the transcribed pre-crRNA using a metal ion-independent active site that has not been well defined [10,16]. Once assembled with its mature crRNA, Cas13a is competent for foreign single-stranded RNA (ssRNA) target surveillance. Recognition of a spacer-complementary ssRNA transcript (ssRNA-activator) activates the HEPN (higher eukaryotes and prokaryotes nucleotide-binding) domains of the enzyme to create a composite general RNase active site [8,16,17], distinct from that responsible for crRNA maturation [10]. Across the Cas13a family, this composite HEPN active-site is functionally diverse, in terms of both nucleotide cleavage preference and turnover efficiency [14]. Two distinct subfamilies, named for their cleavage preference (adenosine (A)-cleaving or uridine (U)-cleaving), exist within the Cas13a protein family and can be harnessed in parallel for RNA detection [14].

The molecular cues of Cas13a HEPN nuclease activation are unknown, limiting the development of Cas13a-based tools. Recent structural studies on the apo and binary complex of *Leptotrichia shahii* Cas13a established the HEPN domains form a surface exposed composite active site[16]. While an attractive possibility, multiplexed RNA detection from the two Cas13a subfamilies is limited by an incomplete mechanistic understanding of the different crRNA binding specificities and catalytic efficiencies, especially in light of substantial sequence divergence within the family[14].

To determine the structural basis for the divergent activities of the A-cleaving Cas13a subfamily, we solved a 2.0 Å resolution crystal structure of *Lachnospiraceae bacterium* Cas13a (LbaCas13a). This structure and accompanying biochemical experiments revealed the active-site for pre-crRNA processing in the A-cleaving Cas13a subfamily. The structure also showed that LbaCas13a shields the spacer region of the crRNA from target hybridization, providing clues to the conformational rearrangement required to activate the HEPN nuclease for general RNase activity. Finally, structural comparisons to the previously modeled U-cleaving subfamily member, LshCas13a[16], rationalize how these enzyme

subfamilies recognize unique sets of crRNAs that enable orthogonal RNA detection in complex mixtures. Together, these findings provide new insights into the mechanisms controlling crRNA generation by, and target RNA activation of, Cas13a.

## RESULTS

### Global architecture of LbaCas13a:crRNA complex

The Cas13a enzyme family depends on two distinct nuclease activities for biological function, yet neither nuclease active-site has been well described in terms of mechanism or conformational regulation. To further investigate the mechanisms of pre-crRNA processing and activator driven HEPN nuclease activity, we studied a divergent A-cleaving homolog, LbaCas13a, enabling comparisons to previously investigated U-cleaving subfamily members LshCas13a and *Leptotrichia buccalis* (Lbu) Cas13a (Fig. 1a). We determined the crystal structure of LbaCas13a (residues 1-1437) in complex with a mature crRNA to 2.0 Å resolution by native single-wavelength anomalous dispersion (SAD) (Fig. 1b–d; Table 1; Supplementary Fig. S1a). To obtain a mature crRNA for binary complex crystallization, LbaCas13a was incubated with a 60-nt pre-crRNA, allowing the enzyme to process the crRNA, before purifying the resulting complex (Supplementary Fig. S1b). The mature crRNA within the structure was resolved as 41 well-ordered nucleotides that includes the 28-nt 5′ direct repeat (henceforth referred to as the 5′ handle) (A(-28)-C(-1)) and 13 nucleotides of the 3′ spacer. Only the first 13 nucleotides (C(1)-G(13)) of the 3′ spacer are well-ordered within the crystal structure with the remaining 11 nucleotides G(14)-C(24) present, but disordered within the structure (Fig. 1b; Supplementary Fig. S1c).

LbaCas13a adopts a predominantly α-helical fold comprising six discrete domains that encase the crRNA to form an active surveillance complex (Fig. 1c, d). Consistent with other single-protein CRISPR effectors, its overall architecture is bi-lobed; the N-terminal domain (NTD) and the Helical-1 domain form the crRNA recognition (REC) lobe, and the HEPN1, Helical-2, Helical-3, and HEPN2 domains form the nuclease (NUC) lobe. The crRNA 5′ handle is bound by the REC lobe within a positively charged channel formed between the NTD and Helical-1 domains, while the spacer is channeled into a cavity formed within the NUC lobe (Fig. 1d; Supplementary Fig. 1c). The NUC lobe in turn encompasses two distinct structural domains, NUC1 and NUC2, that "sandwich" the targeting region of the crRNA into a planar orientation (Fig. 1d; Supplementary Fig. 1c). The NUC1 lobe contains the N-terminal half of HEPN1 (HEPN1-I), connected by a large α-helical domain, Helical-2. The NUC2 lobe comprises the C-terminal half of HEPN1 (HEPN1-II), Helical-3 and HEPN2 domains. The overall architecture of this A-cleaving Cas13a is related to the U-cleaving *Leptotrichia shahii* Cas13a (LshCas13a) [16], with notable differences in the NTD and Helical-3 domain disposition (Supplementary Fig. S2).

### Recognition of a compact and divergent crRNA handle

A hallmark of CRISPR interference complexes is highly specific and tight binding between effector proteins and their cognate crRNA [18–22]. Multiple sequence alignment of the crRNA repeat across both Cas13a subfamilies suggests that the 5′ handle of Type VI-A CRISPR-Cas systems forms a conserved hairpin structure comprising a 5-nt base-paired stem, an 8–9

nt loop, and a 2-nt bulge within the stem (Fig. 2a) [14]. Previous work established that LbaCas13a can process and utilize crRNA handles from all A-cleaving subfamily members but cannot utilize crRNA handles from the related U-cleaving subfamily despite the similarity of their predicted structures (Fig. 2a) [14]. We wondered if our structure could inform us on the molecular basis for the restrictions on Cas13a subfamily crRNA exchangeability.

This structure of LbaCas13a revealed that the crRNA preference is determined by a combination of loop conformation and a divergent set of contacts between the 5′ handle and Cas13a (Fig. 2b–f; Supplementary Fig. S3a). LbaCas13a binds to its cognate crRNA with high affinity (Supplementary Fig. S4a) through an extensive network of conserved sugar-phosphate and nucleobase interactions within a cavity formed by the NTD and Helical-1 domains (Fig. 1d; Fig. 2b; Supplementary Fig. S3a). The 5′ handle of LbaCas13a's cognate crRNA is highly compact, as the loop (A(-18)-A(-11)) is condensed into a helical stack (Fig. 2b–d). Only A(-14) is extruded from the cavity and disordered in the solvent (Fig. 2b; Supplementary Fig. S1c). This conformation is in stark contrast to the structure of U-cleaving LshCas13a's crRNA, in which A(-13), A(-15), U(-17) and A(-18) are flipped out of the loop (Fig. 2b).

The 5′ handle makes both structure- and sequence-specific contacts with LbaCas13a (Fig. 2c, d, f; Supplementary Fig. S3a). In LbaCas13a, the loop is stabilized by the arrangement of the purines into a quadruple and triple base stack on either side of the single extruded nucleotide, A(-14). The quadruple π-stack (A(-11)-A(-15)) is extended to include Y376 of Helical-1 and is stabilized by base-specific interactions with R369, E410, and E406 (Fig. 2c). Notably, the nucleobases involved in these interactions, G(-12) and A(-15), are strictly conserved within the Type VI-A family (Fig. 2a). The equivalent nucleotides in the Lsh-crRNA (G(-15) and A(-18)) make no base-specific contacts with LshCas13a, consistent with our previous observation that LbaCas13a and LshCas13a cannot functionally exchange crRNA [14]. Parallel to the quadruple base stack, nucleotides G(-16)-A(-18) stack between the NTD and Helical-1 domain, forming a triple base stack that is stabilized by R270 hydrogen bonding to backbone phosphates and G(-16) (Fig. 2d). The close proximity of the two sets of stacking nucleotides is further stabilized by a hydrogen bond between the ribose 2′-hydroxyl of G(-12) and nucleobase N3 of A(-17), another contact unique to the compact loop of LbaCas13a.

Beyond stabilizing the crRNA loop, Helical-1 also positions the 5′ flank of the stem-loop in a shallow, surface-exposed groove at the interface with HEPN2 (Fig. 1d; Fig. 2e). crRNAs from the A-cleaving subfamily maintain a 5-nt 5′ flank, which wraps around the exterior of LbaCa13a, with A(-25) and G(-26) stacking with F1300 and F422, respectively (Fig. 2e). Upstream of the 3′ flank, Lba crRNA contains a two-nucleotide bulge, neighboring an A-U base pair that is invariant across the direct repeats of the Type VI-A family. In contrast, the identity of the bulged bases varies as a function of subfamily (Fig. 2a), suggesting that this portion of the sequence may be a crucial feature for crRNA recognition among the two subfamilies. Consistent with this hypothesis, the AA 3′-bulge observed in LbaCas13a (A(-6)-A(-5)) is read sequence-specifically (Fig. 2f; Supplementary Fig. S3a). A(-6) engages in a conserved pocket formed by HEPN1-II and Helical-2, where it stacks with Y928

ordering the nucleobase N7 and amine to hydrogen bond with E921 and Q927, respectively. A(-5) is buried in a pocket formed at the interface of the NTD, Helical-1, and HEPN1-I domains, where it interacts with the backbone carbonyl and side-chain hydroxyl of T21 (Fig. 2f). The set of contacts that recognize A(-6) in LbaCas13a closely resemble those reading out the equivalent A(-8) in LshCas13a [16]. In contrast, the second nucleotide of the AC-bulge in LshCas13a (C(-7)) engages in a nucleobase-specific pocket, suggesting the A-cleaving subfamily has diverged to allow for adenosine at this position.

While the REC lobe confers crRNA 5′ handle specificity in both U- and A-cleaving Cas13a subfamilies, the landscape of protein-RNA interactions is markedly different. Our structure has revealed differences in crRNA conformation that could not be predicted through simple RNA-folding algorithms. The structure- and sequence-specific contacts observed in LbaCas13a serve to rationalize the previous finding that LbaCas13a cannot effectively utilize non-cognate crRNA from the U-cleaving subfamily [14].

## Spacer recognition and activation of LbaCas13a

Cas13a-crRNA complexes show no general RNase activity in the absence of a complementary ssRNA-activator, suggesting that the enzyme is able to achieve complete conformational inhibition until the activator is bound. The binary LbaCas13-crRNA structure represents the target search complex, allowing us to understand the structural constraints on both target identification and HEPN nuclease activation.

We observed unambiguous density for a contiguous stretch of 13 spacer nucleotides immediately proximal to the repeat (C(1)-G(13)) (Fig. 3a). The remaining 11 spacer nucleotides, while present in the crystal (Supplementary Fig. S1b), were disordered in the structure (Supplementary Fig. S1c). The repeat-proximal spacer adopts a distorted U-turn conformation in a pocket between NUC1 and NUC2 (Fig. 3b). Several sugar-phosphate interactions stabilize the highly-distorted RNA conformation within the NUC lobe (Supplementary Fig. S3a). While the discontinuous spacer nucleotides within LshCas13a appear to follow a similar trajectory, our structure provides a more holistic view of spacer orientation within the NUC lobe. The crystal structure of LshCas13a also revealed pre-ordered A-form spacer nucleotides bound at the surface of the NTD [16], which we did not observe in LbaCas13a.

The conformation of the LbaCas13a spacer sequence led us to wonder if spacer length would dramatically impact the HEPN nuclease activity of LbaCas13a. Other homologs have demonstrated HEPN nuclease activation with a variety of spacer-lengths, ranging from 20–28 nucleotides [10,17]. We did not observe a significant effect of spacer length on rates of LbaCas13a *trans*-ssRNA cleavage across this range (Supplementary Fig. S4b–c). We solved an additional structure of LbaCa13a complexed with a crRNA containing a 20-nt spacer (Table 1; Supplementary Fig. S1b). While the overall protein structure was identical to the 24-nt spacer containing complex, we observed disorder and partial occupancy of the crRNA spacer within the NUC lobe of the 20-nt spacer containing complex (Supplementary Fig. S1d). These observations suggest that LbaCas13a accommodates a range of spacer lengths, sequence identities, and conformations within the spacer cavity, consistent with its role as an adaptive RNA-guided immune system.

We examined the entirety of the bound spacer within LbaCa13a to identify interactions between the NUC lobe and crRNA spacer. The HEPN domains, Helical-2 and Helical-3 domains make a number of conserved contacts with the sugar-phosphate backbone of the spacer (Fig. 3c; Supplementary Fig. S3a–c). Notably, the 3′ end of the spacer (C(11)-G(13)) emerges from the NUC lobe, where it adopts an A-form conformation as C(11) stacks with F1293 and F1338 of HEPN2, effectively gating access to the proximal spacer sequence that is sandwiched within the NUC lobe (Fig. 3c). The nucleotides emerging from the NUC lobe adopt a similar conformation in LshCas13a and in our second structure (20nt-spacer) of LbaCas13a (Fig. 3c; Supplementary Fig. S1d–e). Additionally, there are two highly conserved pockets that extensively contact the sugar-phosphate backbone around spacer nucleotides A(6) and A(8), rendering the nucleobases solvent inaccessible (Supplementary Fig. S3b–c). These structural observations suggest that the Cas13a-crRNA complex orders the repeat-proximal spacer into a distorted U-turn while conformationally gating HEPN nuclease activation through an extensive network of interactions with the sugar-phosphate backbone. Further biochemical and structural studies will be required to determine how the spacer rearranges upon activator-binding and how this conformational change triggers Cas13a HEPN domain activation.

We also sought to understand how crRNA-bound LbaCas13a is competent for target search but not active for *trans*-ssRNA cleavage. Evidence from both biochemical and structural experiments has suggested that U-cleaving Cas13a enzymes are dependent on the formation of a composite active-site comprised of two $RX_4H$ motifs, one contributed by each HEPN domain [8,10,16,17]. In agreement with the catalytic requirements of the U-cleaving subfamily, double mutation of the conserved arginine and histidine within either HEPN1 (R600, H605) or HEPN2 (R1243, H1248) abolished general RNase activity of LbaCas13a (Fig. 3d; Supplementary Fig. 4e, f). In the LbaCas13a structure, the HEPN1 motif adopts a conformation in which α2 of HEPN1-I is distorted, burying H605 of the $RX_4H$ motif distant from the putative active-site beneath a loop of HEPN2 (Fig. 3e). Given this remote and solvent inaccessible positioning, we questioned whether H605 participates in HEPN nuclease activity. Similar to double mutations in HEPN1 or HEPN2, the single H605A substitution was sufficient to render LbaCas13a catalytically inactive (Fig. 3d), suggesting that activation of the enzyme requires H605 to join the surface-exposed HEPN motif. Overall, these data demonstrate that the HEPN nuclease remains inactive prior to ssRNA-activator binding due to the crRNA-assisted separation of HEPN active site residues. These data extend the previous mechanistic understanding gained from the LshCas13a structure, where it was observed that the HEPN domains were surface exposed to form a composite active site [16]. Activator binding must induce a synergistic conformational change involving the release of the spacer from the NUC lobe cavity and accommodation of the activator-spacer RNA duplex, similar to the target RNA binding mechanism observed in human Ago2 [23].

## Active site for Cas13a crRNA maturation

Pre-crRNA processing is broadly conserved within the Cas13a family, yet despite extensive mutagenesis studies, the precise active-site residues and chemical mechanism are unknown [10,14,16]. Our binary LbaCas13a-crRNA structure represents the product-bound state of the

pre-crRNA processing reaction, prompting us to examine the mature crRNA 5′-terminus for insights into the catalytic mechanism. The structure revealed that the 5′-flank (C(-24)-A(-28)) is directed away from the core of the enzyme and into a surface-exposed groove formed by the Helical-1 and HEPN2 domains (Fig. 4a). Both Helical-1 and HEPN2 extensively interact with the sugar-phosphate backbone of the 5′-flank, stabilizing the crRNA in the product-bound state (Supplementary Fig. S3a). The residues within LbuCas13a whose mutation results in processing defects are predicted by conservation to line this processing groove between Helical-1 and HEPN2 (Supplementary Fig. S5). The 5′ terminal nucleotide within the LbaCas13a structure, A(-28), has a well-resolved 5′-hydroxyl that is surrounded by H328, K435, and K1320, amino acids that are strictly conserved within the Cas13a A-cleaving subfamily (Fig. 4a; Supplementary Fig. S3a, Supplementary Fig. S5).

To test the involvement of H328, K435, and K1320 in crRNA maturation, we substituted alanine at these positions and assayed each point mutant's capacity for pre-crRNA cleavage. All three single point mutants were completely inactive for pre-crRNA processing, confirming the role of H328, K435 and K1320 in catalysis of crRNA maturation (Fig. 4b; Supplementary Fig. S4d). Further biochemical testing of additional residues near A(-28) revealed that alanine substitutions at K432, D1268, or K1305 completely ablate activity, while mutations of W325 or N1232 modestly impact processing (Fig 4b; Supplementary Fig. S4d). To confirm that the observed mutant phenotypes are catalytic deficiencies and not substrate binding defects, we assayed mutant HEPN-nuclease activity and binding affinity of a pre-crRNA mimic substrate. With the exception of D1268A, the processing mutants maintained near wild-type levels of ssRNA *trans*-cleavage activity (Supplementary Fig. S4e–f). Furthermore, all of the catalytically implicated LbaCas13a mutants retained wild-type binding affinity for both the pre-crRNA mimic and the mature crRNA (Supplementary Fig. S4g–i).

These structural and biochemical observations are consistent with crRNA processing occurring via an acid-base mechanism, where the reaction proceeds through a cation metal-independent pathway in which the 2′-hydroxyl nucleophilically attacks the scissile phosphate in the 3′ position. To assess the role of divalent cations in pre-crRNA processing, we performed processing reactions in the presence of increasing concentrations of the metal chelator EDTA and observed minimal impacts on LbaCas13a processing activity (Fig. 4c; Supplementary Fig. S4j). To confirm the role of the 2′-hydroxyl of G(-29), we assayed the ability of LbaCas13a to process a pre-crRNA substrate in which the catalytic ribonucleotide is substituted for a 2′-deoxynucleotide (dG(-29) substrate). LbaCas13a was unable to cleave this pre-crRNA mimic substrate (Fig. 4f), providing biochemical evidence for the metal-independent acid-base mechanism of pre-crRNA processing.

To provide support for this model of crRNA processing, we crystallized the crRNA-processing deficient LbaCas13a H328A with a pre-crRNA substrate bearing two nucleotides 5′ to the scissile phosphate (Table1; Fig. 4d) Wild-type LbaCas13a can process this truncated pre-crRNA (Supplementary Fig. S4k), suggesting that this structure represents the pre-catalytic, substrate-bound state of crRNA processing. The structure revealed that K435 and K1320 are positioned around the scissile phosphate, while W325 and N1232 stabilize a distorted C2′-endo conformation of the G(-29) ribose. Furthermore, modeling of H328 into

the mutant structure suggests that it would be positioned to contact the scissile phosphate. Interestingly, we did not observe any of the biochemically implicated residues within hydrogen-bonding distance of the ribose 2′-hydroxyl. The pre-crRNA-bound structure revealed that the 2′-hydroxyl of G(-29) contacts N1232 and a single water molecule that is coordinated by HEPN2 residues K1305 and K1320. These structural observations led us to hypothesize that the G(-29) 2′-hydroxyl is deprotonated by an activated water. Consistent with this hypothesis, K1305 is flanked by two carboxylate moieties (D1268 and E1235) that may enable it to act as a general base to activate the coordinated water.

Together these data suggest a model for LbaCa13a-mediated crRNA processing in which conserved residues within the pre-crRNA processing groove between the Helical-1 and HEPN2 domains form an acid-base catalytic center (Fig. 4e). LbaCas13a binds to its cognate pre-crRNA, positioning the 5′-flank within the processing groove where W325 and N1232 stabilize the ribose of G(-29) in a distorted C2′-endo conformation. This coordination sets up the ribose 2′-hydroxyl for nucleophilic attack on the scissile phosphate, at which point K1305, an activated water, K1320, K435 and H328 participate in proton transfer to facilitate the formation of a 2′-3′-cyclic phosphate on G(-29), leaving behind a mature crRNA with a ribose 5′-hydroxyl on A(-28).

## DISCUSSION

The signature proteins of Type VI-A CRISPR-Cas systems, Cas13a, are a remarkably diverse family of RNA-guided enzymes that degrade ssRNA non-specifically upon binding a spacer-complementary ssRNA-activator. In this work, we present high-resolution structures of the LbaCas13a:crRNA complex, revealing the active site for pre-crRNA processing and suggesting a structural gating mechanism of RNA-guided HEPN nuclease activity. We also uncovered the key conformational differences between two divergent Cas13a subfamilies, highlighting the molecular basis for orthogonal crRNA preferences. Finally, this study reveals that Cas13a ssRNA targeting activity is constrained during target search by both crRNA spacer conformation and HEPN domain arrangement.

Initial studies determined that residues important for pre-crRNA processing by two Cas13a homologs, LbuCas13a[10] and LshCas13a[16], are located in the groove between the Helical-1 and HEPN2 domains. Our study extends this work by establishing that the active site for pre-crRNA processing by all A-cleaving Cas13a homologs is in the Helical-1/HEPN2 interface based on structural, biochemical and phylogenetic data. Protein sequence divergence and differing sites of cleavage within pre-crRNA preclude direct comparison to all of the U-cleaving Cas13a subfamily members[14]. Nonetheless, biochemical evidence suggests that LbuCas13a pre-crRNA processing is also metal ion-independent and likely occurs in an active site located between the Helical-1 and HEPN2 domains[10,14]. These results are consistent with our identification of a conserved RNA processing groove that likely serves to stabilize the crRNA backbone.

After associating with its crRNA, the Cas13a:crRNA complex is competent for target surveillance. The high-resolution structures presented in this paper reveal conserved contacts between the spacer and the two halves of the NUC lobe, suggesting that these regions may

function in conformational sensing of ssRNA-activator binding. The repeat-proximal portion of the spacer is highly distorted and sequestered within the core of the enzyme, making it unlikely that this pocket could accommodate a complementary activator RNA without significant conformational rearrangements. This contrasts with other RNA-guided nuclease complexes, including Cas9 and Argonaute, which pre-order their guides within an extended but accessible cavity [24,25]. Due to the anchoring of the crRNA 5′ handle within the REC lobe and the U-turn conformation of the spacer, we expect that the activator-bound enzyme may contain distortions from a canonical double helix, similar to the spacer:target duplex observed in Type III Cmr RNA-targeting systems [26]. We propose that both Cas13a and the distorted crRNA spacer must undergo a concerted conformational change upon ssRNA-activator association that drives HEPN domain dimerization, forming a composite active site and activating Cas13a for ssRNA cleavage.

Cas13a represents a novel paradigm within RNA-guided nucleases where the ordered region of the target-complementary sequence is sequestered within the enzyme, with disordered regions of the guide RNA exposed to solvent. Other RNA-guided nucleases, including the Argonaute family and Cas9, utilize a seed region within their guides for initial nucleic-acid recognition to direct target-binding into a pre-ordered A- or B-form guide [27,28]. The pre-organization of the seed sequence accelerates target search and mismatch discrimination by decreasing the entropic cost of target binding [29]. Cas13a is not following this model for target search, as the solvent-accessible crRNA regions are disordered and the ordered crRNA regions are incompatible with duplex formation. Due to the kinked and inaccessible spacer nucleotides, complementary ssRNA-activator binding can only occur once the system overcomes this additional energetic barrier, raising the question of what energetic parameter drives ternary complex formation.

The Cas13a enzyme family continues to be developed as a platform for RNA detection and programmable RNA binding [13,14]. Bioengineering of Cas13a will require a complete understanding of how the enzyme carries out target search and is subsequently activated to become a general RNase. Our high-resolution target surveillance structure provides key details about crRNA recognition, processing and spacer ordering that will advance the development of LbaCas13a as valuable tool for RNA detection.

## METHODS

### LbaCas13a protein production and purification

The expression and purification of *Lachnospiraceae bacterium* NK4A179 Cas13a (LbaCas13a) was carried out as described elsewhere [14] with modifications, summarized here: $His_6$-MBP-TEV-tagged Cas13a was prepared in *E. coli* Rosetta2 (DE3) grown in TB at 37°C. At an $OD_{600}$ of 0.4–0.5, cultures were cooled on ice for 15 minutes prior to induction with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and expression overnight at 16°C. Soluble $His_6$-MBP-TEV Cas13a was loaded onto a 5 mL HiTrap NiNTA column (GE Healthcare) and eluted over a linear imidazole (0.01–0.3 M) via FPLC (ÄKTA Pure). After overnight dialysis and TEV digestion, the eluted LbaCas13a was further purified by ion exchange (5 mL HiTrap SP, GE Healthcare) and size-exclusion chromatography (S200 16/60, GE Healthcare) with 2 mM TCEP supplemented into the gel filtration buffer.

### *In vitro* transcription and purification of RNA

For biochemical experiments, RNA was transcribed and radiolabeled as described previously [14]. For crystallization, pre-crRNA was transcribed *in vitro* from a PCR-amplified double-stranded DNA template encoding a hepatitis delta virus (HDV) ribozyme 3′ to the pre-crRNA sequence. Transcriptions were carried out as described previously [10] with the following modifications: pre-crRNAs were transcribed at 37°C overnight, followed by treatment with DNase I and 5 mM $MgCl_2$ (37°C, 1 hr) to quench the reaction and drive ribozyme cleavage. Before complexing with LbaCas13a, pre-crRNA was annealed in 1X annealing buffer (10 mM Na-HEPES pH 7.4, 30 mM KCl, 1.5 mM $MgCl_2$) by heating at 75°C for 5 min followed by rapid cooling on ice.

### LbaCas13a:pre-crRNA complex formation

For crystallization, the LbaCas13a:crRNA or LbaCas13a (H328A) pre-crRNA complex was reconstituted *in vitro* by incubating purified LbaCas13a protein with pre-crRNA at a molar ratio of 1:1.3 at 37°C for 1 hr in complexing buffer (20 mM Na-HEPES pH 7.0, 200 mM KCl, 1 mM TCEP). The complex was purified by size-exclusion chromatography (S200 10/300 Increase, GE Healthcare) in complexing buffer and concentrated using a 10 kDa Centricon (Thermo Fisher) before immediate use in crystallization experiments. The purity and integrity of protein and RNA were assessed using SDS-PAGE (Coomassie blue staining) and 15% (v/v) TBE-urea-polyacrylamide gel electrophoresis (SYBR Gold straining), respectively (Supplementary Fig. S1b).

### LbaCas13a crystallization and data collection

Initial screening was carried out with JCSG Core Suites (Qiagen) in 96-well sitting drop vapor diffusion format (Intelli-plate, Hampton Research) at 293 ± 1 K. Diffraction-quality crystals of the LbaCas13a:crRNA complex (20-nt and 24-nt containing spacer) ($A_{280}$=4.0) and LbaCas13a:pre-crRNA complex (24-nt containing spacer) ($A_{280}$=5.7) grew in 0.2 M ammonium iodide and 20% (w/v) PEG 3,350. Crystals were cryoprotected in 10–15% (v/v) glycerol, and high-resolution diffraction data were collected at the Advanced Light Source (ALS) beam line 8.3.1 on a Pilatus3 S 6M (Dectris) detector under cryogenic conditions. To produce isomorphous crystals for phasing by native single-anomalous dispersion (SAD), sitting drop vapor diffusion experiments containing 2 μL purified LbaCas13a:crRNA complex (24-nt containing spacer) complex ($A_{280}$=4.0) and 2 μL reservoir solution (0.2 M ammonium iodide, 20–25% (w/v) PEG 3,350) were prepared in 24-well Cryschem plates (Hampton Research). A total of 52 datasets were collected from 26 crystals (cryoprotected in 10% (v/v) glycerol) under cryogenic conditions at 6.0 keV (ALS beam line 8.3.1, Pilatus3 S 6M). Each dataset covered 360° of crystal rotation divided into 0.2° images collected in 45° inverse-beam wedges at 5 Hz. To minimize pixel readout error, the detector was moved by 10 mm before collecting a second 360°-pass from the same position of each crystal.

### Structure determination by native SAD and model refinement

High-resolution data collected from LbaCas13a crystals at 11.1 keV were processed and scaled in *XDS* [30] before merging in *AIMLESS* [31,32]. After processing with *XDS*, each SAD dataset was compared to every other dataset after applying each of the 27 possible re-

indexing operations found by *OTHERCELL* [31], and then computing an isomorphous R-factor with a 4 Å resolution cut-off using *SCALEIT* [31]. Pairs of datasets agreeing to better than 20% were grouped and merged using *XSCALE* [30] producing a final cluster containing 20 wedges in the space group C2 (data statistics summarized in Table 1). The merged anomalous data were prepared for substructure search in *SHELXC* [33], and input into *SHELXD* set to search for 65 sites using a high-resolution cutoff of 3.5 Å. Nine solutions were found after 25,000 trials and each solution was input separately into *SHELXE*. The correct substructure solution was found to contain a total of 115 sites, 32 of which were correct as determined using *PHENIX.emma* [34]. A readily interpretable map (CC 25%) was obtained after 30 cycles of density modification and model building in *SHELXE. PHENIX.AutoBuild* [35] was used to extend the initial α-helical model, complemented with manual model-building in *COOT* [36] and RNA-modelling with *RCrane* [37]. The atomic models were completed by iterative building and refinement in *COOT* and *PHENIX.refine* [38] against the high-resolution data using the $2mF_o$-$DF_c$, $mF_o$-$DF_c$, anomalous Fourier difference and $2mF_o$-$DF_c$ composite omit [39] maps. The final models for LbaCas13a:crRNA (24-nt spacer), LbaCas13a:crRNA (20-nt spacer), and LbaCas13a(H328A):pre-crRNA (24-nt spacer) were validated in *MOLPROBITY* [40] and submitted to the Protein Data Bank (PDB) under the accession codes 5W1H, 5W1I, and 5WLH respectively. All figures were generated in *PyMOL* (PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC).

### Radiolabeled crRNA maturation assays

Pre-crRNA processing assays were performed as previously described [14]. Briefly, 100 nM LbaCas13a was incubated with <1 nM 5′-radiolabeled pre-crRNA substrates at 37°C in processing buffer (20 mM HEPES-Na pH 6.8, 50 mM KCl, 5 mM $MgCl_2$, 10 μg/mL BSA, 100 μg/mL tRNA, 0.01% Igepal CA-630, and 5% glycerol). Reactions were quenched (100% formamide, 0.025% bromophenol blue, and 200 μg/mL heparin), denatured at 95°C for 5 min and resolved by 15% denaturing PAGE (0.5x TBE buffer). All bands were visualized by phosphorimaging (Typhoon, GE Healthcare) and quantified with ImageQuant (GE Healthcare). The percent cleavage was determined as the ratio of the intensity of the product band to the total intensity of both the product and uncleaved pre-crRNA bands, and normalized for background within each measured substrate. For clarity, in some figure panels the percentage of uncleaved substrates is presented (calculated as 100-%cleaved).

### Fluorescent ssRNA nuclease assays

Target cleavage assays were performed as previously described with minimal modification [14]. Briefly, 100 nM LbaCas13a was complexed with 50 nM pre-crRNAs at 37°C in processing buffer for 30–60 mins. Alternatively, cleavage buffer (20 mM HEPES-Na pH 6.8, 50 mM KCl, 5 mM $MgCl_2$) was used for reactions instead of processing buffer as noted in the figure legends, resulting in less fluorescent signal generated. Generally, 150 nM RNase Alert reporter (IDT) and 10 nM of ssRNA-activator were added to initiate the reaction. Reactions were incubated in a fluorescence plate reader (Tecan Infinite Pro F2000) for 60 min at 37°C, with fluorescence measurements taken every 5 min ($\lambda_{ex}$: 485 nm; $\lambda_{em:}$ 535 nm). Background corrected values were calculated by subtracting signal in the negative control containing no activating RNA. For endpoint analyzes, the background corrected florescence value from the 60 min time point is presented (mean ± st. dev., n= 3). Apparent

rates were calculated according to a single-exponential decay using Prism7 (GraphPad) fitted to the entire curve, and calculated rates were plotted with their associated standard deviations (n=3).

### Filter-binding assays

Filter-binding assays were carried out as previously described [10]. Briefly, Cas13a and radiolabeled crRNA were incubated for 1 hr at 37°C in RNA-processing buffer (20 mM HEPES-Na pH 6.8, 50 mM KCl, 5 mM $MgCl_2$, 10 $\mu$g/mL BSA, 100 $\mu$g/mL yeast tRNA, 0.01% Igepal CA-630 and 5% glycerol). Tufryn, Protran and Hybond-N+ membranes were assembled onto a dot-blot apparatus in the order listed. The membranes were washed twice with 50 $\mu$L cleavage buffer (20 mM HEPES-Na pH 6.8, 50 mM KCl, 5 mM $MgCl_2$, and 5% glycerol) before the sample was applied to the membranes. Membranes were then washed with 50 $\mu$L cleavage buffer, dried, and visualized by phosphorimaging. Data were quantified with ImageQaunt TL Software (GE Healthcare) and fit to a binding isotherm using Prism7 (GraphPad). All experiments were carried out in triplicate.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank Nan Ma and Kaihong Zhou for technical assistance, Christine L. Gee for assistance with refinement, and members of the Doudna laboratory for helpful discussions. G.J.K. acknowledges support from the Howard Hughes Medical Institute. This work was supported in part by a Frontiers Science award from the Paul Allen Institute to J.A.D, and the National Science Foundation (MCB-1244557 to J.A.D.). J.A.D is an Investigator of the Howard Hughes Medical Institute. J.A.D. is a co-founder of Caribou Biosciences, Inc., Editas Medicine, and Intellia Therapeutics.

## References

1. Dy RL, Richter C, Salmond GP, Fineran PC. Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu Rev Virol. 2014; 1:307–31. [PubMed: 26958724]
2. Barrangou R, Marraffini LA. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. Mol Cell. 2014; 54:234–44. [PubMed: 24766887]
3. Marraffini LA. CRISPR-Cas immunity in prokaryotes. Nature. 2015; 526:55–61. [PubMed: 26432244]
4. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014; 157:1262–78. [PubMed: 24906146]
5. Wright AV, Nunez JK, Doudna JA. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell. 2016; 164:29–44. [PubMed: 26771484]
6. Barrangou R, Doudna JA. Applications of CRISPR technologies in research and beyond. Nat Biotechnol. 2016; 34:933–941. [PubMed: 27606440]
7. O'Connell MR, et al. Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature. 2014; 516:263–6. [PubMed: 25274302]
8. Shmakov S, et al. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Molecular Cell. 2015; 60:385–397. [PubMed: 26593719]
9. Nelles DA, et al. Programmable RNA Tracking in Live Cells with CRISPR/Cas9. Cell. 2016; 165:488–96. [PubMed: 26997482]
10. East-Seletsky A, et al. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. Nature. 2016

11. Shmakov S, et al. Diversity and evolution of class 2 CRISPR-Cas systems. Nat Rev Microbiol. 2017
12. Smargon AA, et al. Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. Mol Cell. 2017
13. Gootenberg JS, et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science. 2017; 356:438–442. [PubMed: 28408723]
14. East-Seletsky A, O'Connell MR, Burstein D, Knott GJ, Doudna JA. RNA Targeting by Functionally Orthogonal Type VI-A CRISPR-Cas Enzymes. Mol Cell. 2017; 66:373–383 e3. [PubMed: 28475872]
15. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology. 2009; 155:733–40. [PubMed: 19246744]
16. Liu L, et al. Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities. Cell. 2017; 168:121–134 e12. [PubMed: 28086085]
17. Abudayyeh OO, et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science. 2016; 353:557.
18. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science. 2010; 329:1355–8. [PubMed: 20829488]
19. Sternberg SH, Haurwitz RE, Doudna JA. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. RNA. 2012; 18:661–72. [PubMed: 22345129]
20. Li H. Structural Principles of CRISPR RNA Processing. Structure. 2015; 23:13–20. [PubMed: 25435327]
21. Hochstrasser ML, Doudna JA. Cutting it close: CRISPR-associated endoribonuclease structure and function. Trends Biochem Sci. 2015; 40:58–66. [PubMed: 25468820]
22. Charpentier E, Richter H, van der Oost J, White MF. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol Rev. 2015; 39:428–41. [PubMed: 25994611]
23. Schirle NT, Sheu-Gruttadauria J, MacRae IJ. Structural basis for microRNA targeting. Science. 2014; 346:608–13. [PubMed: 25359968]
24. Jiang F, Doudna JA. The structural biology of CRISPR-Cas systems. Curr Opin Struct Biol. 2015; 30:100–11. [PubMed: 25723899]
25. Nakanishi K. Anatomy of RISC: how do small RNAs and chaperones activate Argonaute proteins? Wiley Interdiscip Rev RNA. 2016; 7:637–60. [PubMed: 27184117]
26. Osawa T, Inanaga H, Sato C, Numata T. Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog. Mol Cell. 2015; 58:418–30. [PubMed: 25921071]
27. Parker JS, Parizotto EA, Wang M, Roe SM, Barford D. Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein. Mol Cell. 2009; 33:204–14. [PubMed: 19187762]
28. Jiang F, Zhou K, Ma L, Gressel S, Doudna JA. STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. Science. 2015; 348:1477–81. [PubMed: 26113724]
29. Salomon WE, Jolly SM, Moore MJ, Zamore PD, Serebrov V. Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides. Cell. 2015; 162:84–95. [PubMed: 26140592]
30. Kabsch W. Xds. Acta Crystallogr D Biol Crystallogr. 2010; 66:125–32. [PubMed: 20124692]
31. Winn MD, et al. Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr. 2011; 67:235–42. [PubMed: 21460441]
32. Evans PR, Murshudov GN. How good are my data and what is the resolution? Acta Crystallogr D Biol Crystallogr. 2013; 69:1204–14. [PubMed: 23793146]
33. Powell HR. The Rossmann Fourier autoindexing algorithm in MOSFLM. Acta Crystallogr D Biol Crystallogr. 1999; 55:1690–5. [PubMed: 10531518]
34. Adams PD, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010; 66:213–21. [PubMed: 20124702]

35. Terwilliger TC, et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallogr D Biol Crystallogr. 2008; 64:61–9. [PubMed: 18094468]

36. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010; 66:486–501. [PubMed: 20383002]

37. Keating KS, Pyle AM. Semiautomated model building for RNA crystallography using a directed rotameric approach. Proc Natl Acad Sci U S A. 2010; 107:8177–82. [PubMed: 20404211]

38. Afonine PV, et al. Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr D Biol Crystallogr. 2012; 68:352–67. [PubMed: 22505256]

39. Terwilliger TC, et al. Iterative-build OMIT maps: map improvement by iterative model building and refinement without model bias. Acta Crystallogr D Biol Crystallogr. 2008; 64:515–24. [PubMed: 18453687]

40. Chen VB, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010; 66:12–21. [PubMed: 20057044]

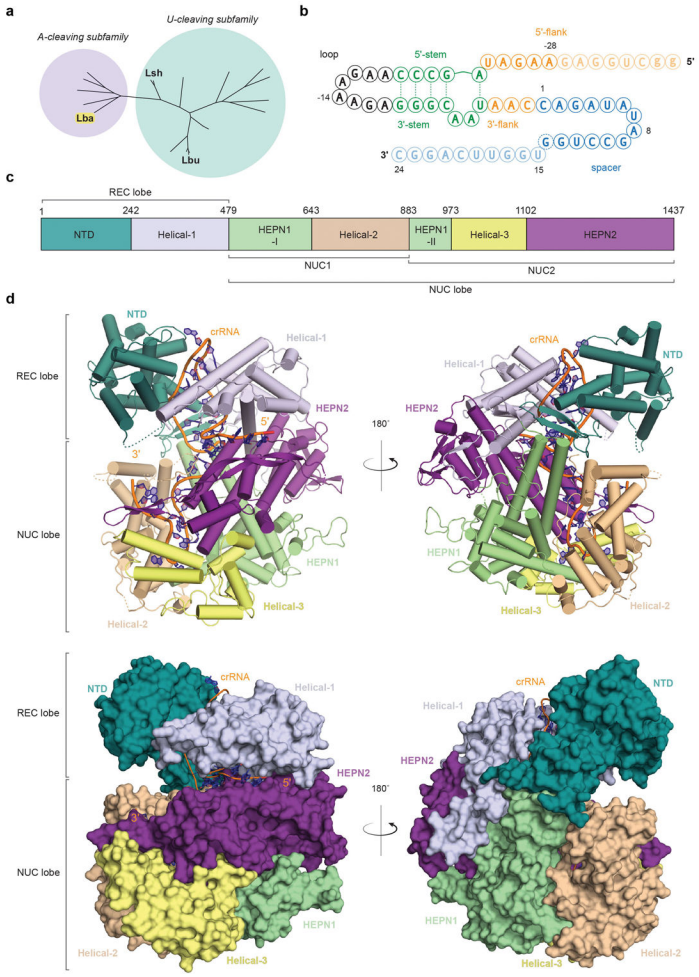



**Figure 1. Overall structure of the LbaCas13a:crRNA complex**

**a**, Phylogenetic tree of all Cas13a family members with colored circles highlighting the A- and U-cleaving subfamilies (adapted from [14]). The homolog described in this study, LbaCas13a, is highlighted in yellow. **b**, Schematic representation of the crRNA repeat scaffold crystallized in this study. Nucleotides at the 5′ and 3′ end that were disordered in the structure are shown in semitransparent text. The ribose and nucleobase of G(14) were disordered. Only the phosphate of G(14) was modeled in the structure and is represented here as a dashed circle. **c**, Domain organization schematic of LbaCas13a shown with the REC, NUC, NUC1, and NUC2 lobes annotated. The color code used is consistent throughout the manuscript. **d,** Two views of the LbaCas13a:crRNA complex are shown related by a 180° rotation as cartoon (top) or surface (bottom) representations. The crRNA backbone is shown in orange.

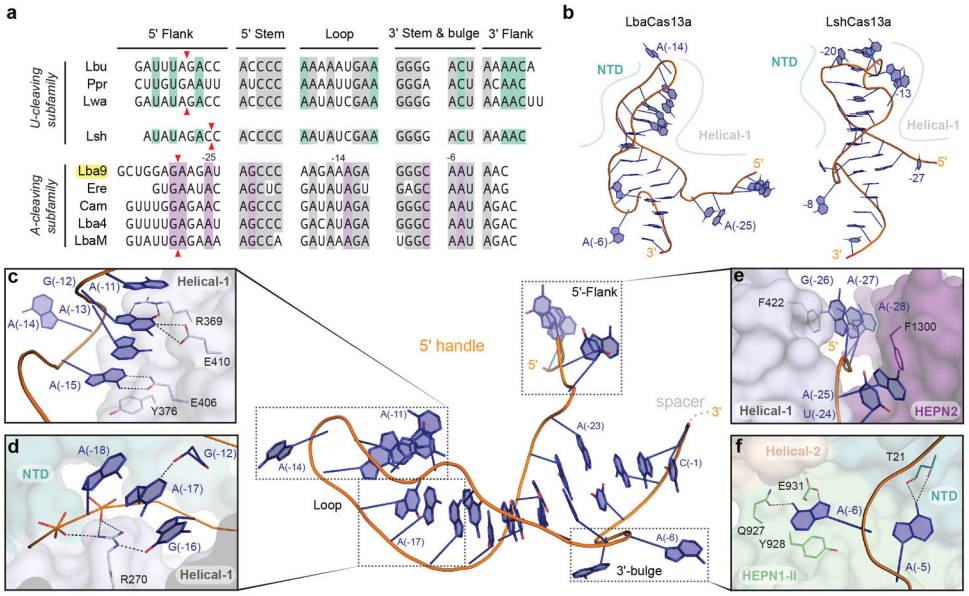



**Figure 2. Structure and recognition of the crRNA 5′ handle**

**a**, Multiple sequence alignment of select type VI-A Cas13a crRNA 5′ handles representing both subfamilies adapted from [14]. Nucleotides conserved across all Cas13a homologs, amongst just the U-cleaving, or A-cleaving subfamily are shown in grey, teal or purple respectively. The sites of pre-crRNA processing for each Cas13a homolog are noted by red arrows. LbaCas13a, is highlighted in yellow with the repeat nucleotides numbered negatively from the start of the spacer. **b**, Side by side comparison of the crRNA crystallized in this study (LbaCas13a, left) and previously [16] (LshCas13a, right). **c–f,** Highlights of specific interactions between the crRNA 5′ handle and LbaCas13a. Purine stacks within the loop region of the 5′ handle interacting with Helical-1 and the NTD (**c–d**). Residues involved in conserved contacts between LbaCa13a and the crRNA loop are shown as sticks, with hydrogen-bonding interactions denoted by dashed lines. Binding of the 5′ flank (A(-28)-U(-24)) of LbaCas13a within the shallow Helical-1/HEPN2 groove via F422 and F1300 stacking with G(-26) and A(-25) respectively (**e**), and sequence-specific recognition of the dinucleotide 3′ AA-bulge (**f**).

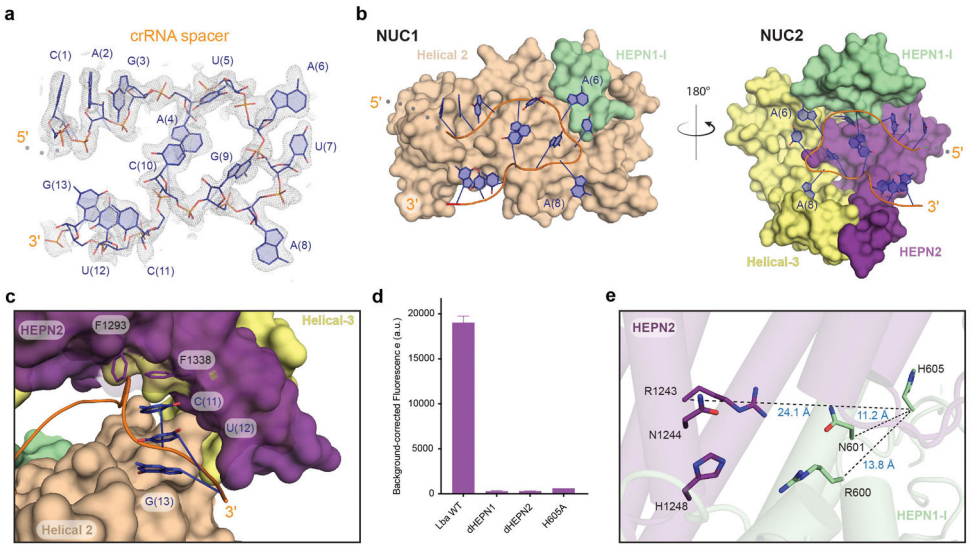



**Figure 3. Structure and recognition of the crRNA spacer**

**a**, Simulated annealing m$F_o$-D$F_c$ omit electron density map of the crRNA spacer contoured at 2σ. **b**, Opposing perspectives of the two halves of the NUC lobe that distort the crRNA spacer into a stabilized U-turn conformation. The half of the NUC lobe projecting out of the page is omitted in each image for clarity. **c**, The 3′ spacer emerging from the NUC lobe where it interacts with HEPN2. The spacer is shown in cartoon representation (orange) with the A-form nucleobases C(11)-G(13) shown as sticks. The amino acids F1338 and F1293 are shown as sticks stacking with C(11) gating access to the upstream spacer. **d**, Endpoint total fluorescence values for ssRNA-targeting assay by LbaCas13a HEPN mutants (background-corrected mean ± st. dev, n=3). Fitted apparent rates for these reactions are presented in Supplementary Fig. S4e. **e**, View of the composite HEPN nuclease active-site with the catalytic residues ($RX_4H$) shown in stick representation. Notably, H605 of HEPN1-I is removed from the surface exposed composite site with distances shown.

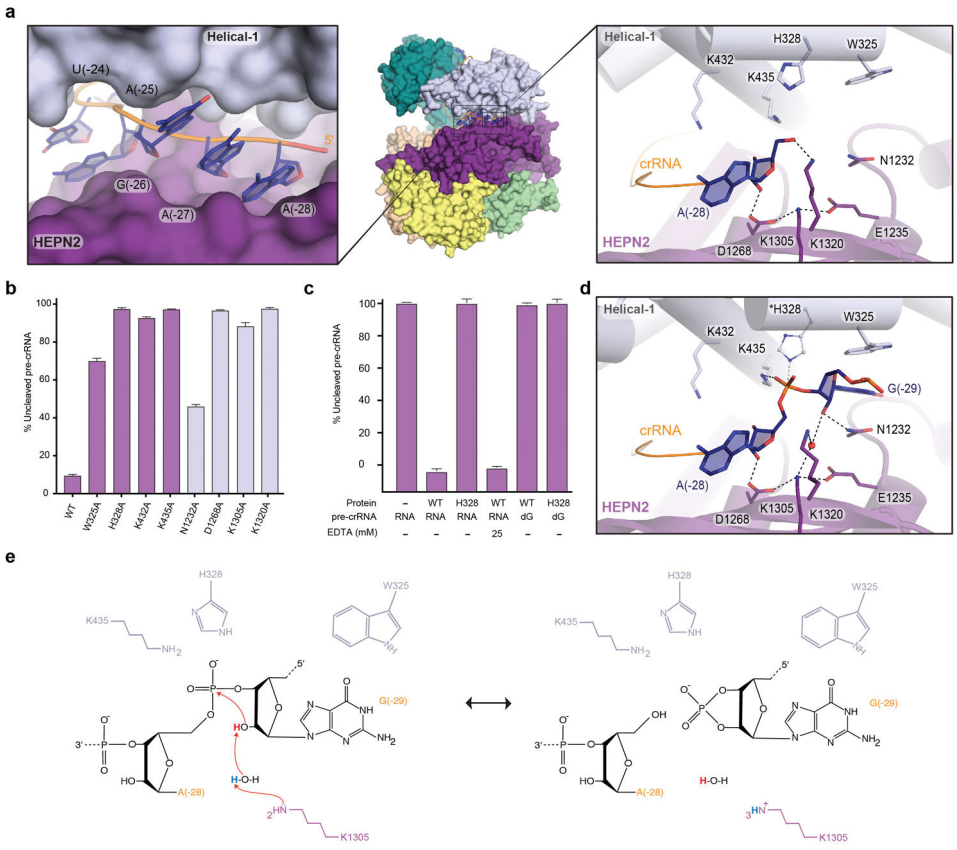



**Figure 4. Structural basis for crRNA maturation by LbaCas13a**

**a**, The pre-crRNA processing groove formed between Helical-1 and HEPN2 domains of wild-type LbaCas13a (left). Close-up view of the pre-crRNA processing site in LbaCas13a (right). The 5′-hydroxyl and amino acids implicated in processing are shown as sticks with hydrogen-bonding interactions denoted by dashed lines. **b**, LbaCas13a-mediated pre-crRNA processing measured under single-turnover conditions for alanine substitutions within the pre-crRNA processing groove. For clarity, 60 min endpoints (mean ± st. dev, n=3) are graphed as the percentage of substrate remaining (uncleaved pre-crRNA). Time course data are plotted and rates calculated in Supplementary Fig. S4. **c,** Bar graph of LbaCas13a-mediated pre-crRNA processing under single-turnover conditions in the presence of EDTA or a modified pre-crRNA dG(-29). Percentage of uncleaved substrate was measured after 60 min (background correct mean ± st. dev, n=3). **d,** Close-up view of the pre-catalytic state of LbaCas13a (H328A) showing the pre-crRNA processing groove formed between Helical-1 and HEPN2 domains. The amino acids surrounding the scissile phosphate between A(-28) and G(-29) are shown as sticks with hydrogen-bonding interactions denoted by dashed lines. H328, marked with an asterisk, is shown in a ball-and-stick representation modeled in the conformation observed in the structure of wild-type LbaCas13a. **e,** Proposed model for acid-base-catalyzed maturation of pre-crRNA by LbaCas13a.

## Table 1

Data collection and refinement statistics

| | LbaCas13a:crRNA (24-nt spacer)[+] | LbaCas13a:crRNA (20-nt spacer)[+] | LbaCas13a:pre-crRNA (24-nt spacer)[+] | LbaCas13a:crRNA (24-nt spacer) (SAD)[‡] |
|---|---|---|---|---|
| **Data collection** | | | | |
| Space group | C 1 2 1 | P 1 | C 1 2 1 | C 1 2 1 |
| Cell dimensions | | | | |
| *a*, *b*, *c* (Å) | 150.84, 92.21, 115.86 | 88.78, 89.21, 116.31 | 150.84, 92.21, 115.86 | 150.69, 91.45, 115.37 |
| α β γ (°) | 90.00, 92.42, 90.00 | 86.14, 86.97, 66.16 | 90.00, 92.42, 90.00 | 90.00, 92.76, 90.00 |
| Resolution (Å) | 47.12-1.99 (2.02-1.99) [*] | 48.57-2.20 (2.24-2.20) [*] | 47.11-1.80 (1.86-1.80) [*] | 46.94-3.50 (3.83-3.50) [*] |
| $R_{merge}$ | 0.079 (0.841) | 0.110 (1.060) | 0.048 (0.666) | 0.406 (0.534)[†] |
| $I / \sigma I$ | 15.2 (2.0) | 8.6 (1.3) | 18.1 (2.2) | 36.0 (30.0) |
| CC (1/2) | 0.99 (0.70) | 0.996 (0.466) | 0.99 (0.77) | 0.99 (0.99) |
| Completeness (%) | 98.8 (89.7) | 97.5 (96.6) | 97.3 (95.3) | 100.0 (100.0) |
| Redundancy | 6.9 (6.0) | 3.6 (3.7) | 5.0 (5.2) | 118.6 (120.2) |
| **Refinement** | | | | |
| Resolution (Å) | 47.12-1.99 (2.02-1.99) | 48.57-2.20 (2.24-2.20) | 47.11-1.80 (1.86-1.80) [*] | |
| No. reflections | 107,256 | 161,267 | 142,815 | |
| $R_{work}$ / $R_{free}$ | 18.50 / 21.63 | 19.65 / 23.30 | 18.59/ 21.85 | |
| No. atoms | 12,499 | 24, 245 | 12,394 | |
| Protein | 11,646 | 23, 086 | 11,601 | |
| Ligand/ion | 41 | 16 | 26 | |
| Water | 812 | 1,143 | 767 | |
| *B*-factors | | | | |
| Protein | 36.35 | 44.31 | 37.08 | |
| Ligand/ion | 56.91 | 83.97 | 64.51 | |
| Water | 39.68 | 41.30 | 37.95 | |
| R.m.s. deviations | | | | |
| Bond lengths (Å) | 0.003 | 0.002 | 0.006 | |
| Bond angles (°) | 0.482 | 0.415 | 0.830 | |

[+] A single dataset was scaled and merged for each crystal structure.

[‡] A total of 20 datasets were scaled and merged.

[*] Values in parentheses are for highest-resolution shell.

[†] The high $R_{merge}$ for the SAD data is due to non-isomorphism between the 20 merged crystals.

*Annu. Rev. Microbiol. 1993. 47:945–63*


# THE TN5 TRANSPOSON

*W. S. Reznikoff*
Department of Biochemistry, College of Agricultural and Life Sciences, University of Wisconsin-Madison, Madison, Wisconsin 53706

KEY WORDS: transposase, end sequences, gene regulation, *cis*-activity



## CONTENTS



## ABSTRACT

The bacterial transposon Tn*5* encodes two proteins, the transposase and a related protein, the transposition inhibitor, whose relative abundance determines, in part, the frequency of Tn*5* transposition. The synthesis of these proteins is programmed by a complex set of genetic regulatory elements. The host DNA methylation function, *dam,* inhibits transposase promoter recognition and indirectly enhances the transposition inhibitor promoter. The inhibitor lacks the N-terminal 55 amino acids of the transposase, suggesting that this sequence plays a key role in the transposition process. An intact N-terminal sequence is required for the transposase's recognition of the 19-bp end DNA sequences. This is the first critical step in the transposition process. Transposase–end DNA interaction is itself regulated by an intricate series of reactions involving several host proteins: DnaA, Dam, and Fis. The transposase is a unique protein in that it acts primarily *in cis* and inhibits its own activity *in trans*. Models to explain these properties are described. Finally circumstantial evidence suggests that transposition occurs preferentially from newly replicated DNA that has yet to be partitioned to progeny cells. This



0066-4227/93/1001-0945$02.00

timing of transposition is likely to have a selective advantage for the host and the transposable element.

## *Introduction*

Transposition is a recombination process in which DNA sequences termed transposable elements move from an original site on a DNA molecule to a new site on the same or on a different DNA molecule. In addition, transposable elements can cause, and are associated with, other types of genetic rearrangements such as deletions, inversions, and chromosome fusions. The genomes of prokaryotic and eukaryotic organisms contain these elements. One could consider them as an ancient genetic machinery for causing genomic rearrangements and, therefore, for facilitating genome evolution. In addition, their associated biochemical reactions are likely to be similar to other interesting events involving the interaction of proteins and DNA. For these reasons transposable elements are of considerable interest.

The transposable elements found in *Escherichia coli* fall into three general classes determined by their mechanisms of transposition. Transposons such as Tn*3* and γδ transpose through a two-step replicative mechanism in which a cointegrate (fused replicon) structure is an intermediate. Transposons Tn*10* and Tn*903* transpose through a conservative cut-and-paste mechanism. Bacteriophage Mu and related viruses represent the third class of transposable element. In these cases the transposition can occur through either of the above two mechanisms, depending upon the proteins involved and the precise nature of the DNA strand cutting after an intermediate is formed between the transposable element and the target DNA sequence. Tn*5* is generally assumed to transpose via a conservative mechanism (2); however, this presumption has not been critically tested. The reader is encouraged to examine the models and evidence for these transposition mechanisms in the recent monograph by Berg & Howe (3).

The conservative and replicative mechanisms of transposition share many basic characteristics. The transposable element encodes two critical functions required for the process—the end sequences and a protein termed the transposase. The element is defined by the specific sequences at its end. Transposition and related events remove the transposable element from its original sequence context precisely at the ends of these sequences. Changes in any base pair of these sequences typically reduces the frequency of or abolishes transposition. The transposable element also encodes a protein called the transposase. The transposase is a critical participant in many transposition functions including: specifically binding to the end sequences, bringing the two ends together through a protein oligomerization process, cutting or nicking the DNA adjacent to the end sequences, and inserting the transposable element DNA into a DNA target site.

Host proteins also play critical roles in the transposition process such as facilitating the end-sequence binding of the transposase, nucleating the higher-order structure in which the ends are brought together, performing the necessary repair or replication functions, and regulating several steps in the transposition process.

Transposition is, in general, a quite rare, highly regulated process. Such tight regulation might enable the host cell to strike a balance between insuring proliferation of the transposable element and insuring the cell's own genetic survival—the very process of transposition causes chromosome breakage and rearrangements. The mechanisms of the regulation vary remarkably among the many cases studied, although some aspects are common to all, and the strategies, perhaps not surprisingly, are similar to ones found for other bacterial genetic systems. In addition to the role of host functions, transposable element encoded–functions often play a critical role in this regulation.

Implicit in the above general description is the fact that several very interesting molecular events are involved in and regulate the transposition process. Studying these events will help us understand other genetic processes. For instance, elucidating how transposition is regulated in a given system will tell us how complex DNA metabolizing processes might be controlled and give yet more examples of how gene expression can be modulated. The transposase is a protein that performs multiple complex functions. Protein structure/function studies would seek answers to the questions about the organization of the peptide domains that perform these functions and how they interact. The terminal DNA sequences are at first glance simply an example of a target for protein binding. However, their functionality is considerably more complex because they are often the target for more than one protein, and the protein–end sequence interaction is associated with at least two events: protein binding and DNA strand scission. Work done on other systems suggests that short DNA sequences can dictate several extremely intricate reactions, and transposable element end sequences may be a perfect example of such compact complexity. Transposable elements use host functions like biochemical parasites. Understanding how they use these functions will tell us much about the transposition process and about the functionality of the host functions themselves. Because many of these host functions are ones involved in host DNA metabolism, understanding their role in the transposition process may give us insights as to how host DNA metabolism is organized and regulated.

Finally, there are always surprising connections in biology. By studying transposition we embark on a largely unknown path into the cell's metabolism. For instance, if the host functions that Tn*5* uses are organized in the cell in a unique spatial fashion, studying Tn*5* transposition may reveal that arrangement.

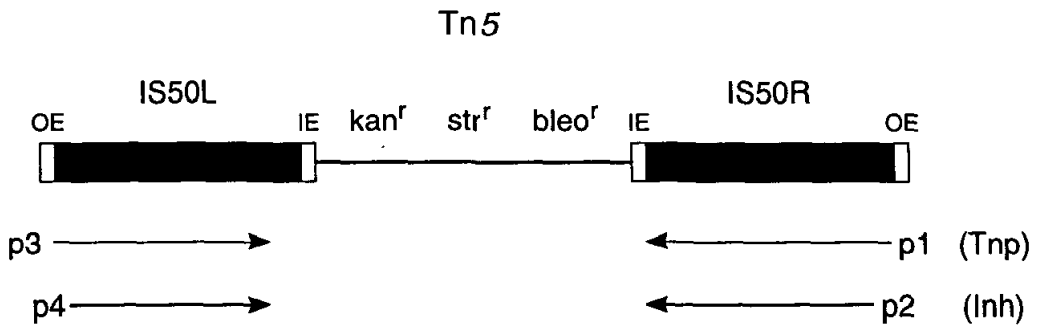



*Figure 1* Transposon Tn*5*. Tn*5* is a composite transposon in which genes encoding three antibiotic resistance proteins are bracketed by two IS*50* elements, IS*50*L and IS*50*R. Both IS*50* elements are delineated by 19-bp sequences, the inside end (IE) and the outside end (OE). IS*50*R encodes the transposase (Tnp or p1) and a second protein that inhibits transposition (Inh or p2). Tnp and Inh are translated in the same reading frame, but the Inh AUG is 55 codons downstream from the Tnp AUG. IS*50*L contains an ochre codon that results in the synthesis of p3 and p4, nonfunctional analogues of Tnp and Inh.



My laboratory and several other investigators have been studying the bacterial transposable element Tn*5* as a model system. Tn*5* is an example of a composite transposon in which antibiotic resistance genes are flanked by two nearly identical insertion sequences, IS*50*R and IS*50*L (see Figure 1 for a schematic). IS*50*R is a fully functional transposable element, while IS*50*L contains an ochre codon that results in the synthesis of inactive proteins (33).

IS*50*R encodes the transposase (p1 or Tnp) (15, 18, 19, 33), and the Tn*5* Tnp has two opposing activities. *In cis* it acts as a transposase, catalyzing the transposition of the Tn*5* or IS*50* sequences from which it was encoded (15, 19). However, *in trans* it primarily acts as an inhibitor of transposition (42). The paradox of Tnp inhibiting the activity of other Tnp molecules is discussed in greater detail below. The Tnp inhibitory activity is one means by which Tn*5* transposition is down-regulated.

A second protein [the inhibitor (Inh or p2)] is also encoded by IS*50*R (15, 18, 19, 45). Inh is translated in the same reading frame as Tnp but lacks the N-terminal 55 amino acids. Inh's only known function is to inhibit Tn*5* transposition, which is thought to be the major means of down-regulating this process. The properties of Inh relative to the Tnp suggest that the N-terminal 55 amino acids of Tnp play an important role in its activity. The possible function of this sequence is an important topic of investigation, and this review discusses our current understanding of this function.

The relative abundance of Tnp and Inh plays a major role in determining the Tn*5* transposition frequency. Interestingly, separate, apparently competing, promoters program Tnp and Inh syntheses (21, 43). Thus the regulation of these promoter activities becomes a crucial question. As discussed below, it is the host that regulates the relative activity of these two promoters, and

the regulatory mechanism that has evolved appears to link the occurrence of transposition to the DNA replication process.

Tn5 has also evolved a mechanism for preventing the spurious synthesis of Tnp by virtue of accidental placement within an active transcription unit (21). The translation initiation signals were designed such that they will only function as part of a correctly initiated Tnp mRNA (21, 37). The strategy to accomplish this translation control may be widespread and is discussed below.

Each IS*50* is bounded by two unique 19-bp end sequences [termed outside (OE) and inside (IE) ends] that are critical for transposition (17, 34). The distinguishing feature between IS*50* and Tn5 transposition is the choice of the transposable element ends. Tn5 transposition utilizes two OEs, whereas IS*50* transposition uses an OE and an IE sequence. Tnp binds to both of these sequences during the transposition process, and they likely perform other functions vital for Tnp activity. In addition, they are the sequences recognized by host functions. As we shall see, a host function binds to the OE to enhance transposition while the host functions affecting the IE down-regulate transposition. Two of the relevant host functions (DnaA and Dam) also link the transposition process to DNA replication.

Finally, we are just now beginning to obtain clues as to the possible relationship of Tn5 transposition with overall cell processes. Some of these clues, which point to DNA replication, have been suggested above. Others will become obvious during the discussion of the lethal effect of overproducing Tnp.

Douglas Berg (whose laboratory has contributed much of what we know about Tn5) recently published an excellent review of Tn5 (2). Therefore, this review is not a comprehensive treatment of the subject. Rather, this chapter concentrates on areas of current and future research interest raised by the questions implied above. The specific topics covered are:

1. How are the syntheses of Tnp and Inh regulated?
2. What are the possible functions for the Tnp N-terminal sequence?
3. What sort of protein recognition reactions occur at the 19-bp terminal sequences?
4. What is the molecular basis for the *cis*-active nature of the transposase?
5. How might transposition be linked to chromosome replication and/or cell division?

## *The Regulation of Transposase and Inhibitor Synthesis*

Studies by Biek & Roth (4) first indicated that the frequency of Tn5 transposition was regulated by Tn5-encoded functions. Tn5 sequences that were newly introduced into a cell transposed at dramatically lower frequencies in a cell already containing Tn5 as opposed to a cell lacking Tn5. In addition,

we (18) found that the Tn*5* transposition frequency was constant regardless of the copy number of Tn*5* (hence the transposition frequency of an individual Tn*5* decreased in the presence of additional Tn*5*s). We now know that this down-regulation is primarily (but not entirely; see below) a consequence of the Tn*5*-encoded Inh protein. Inh functions *in trans* to block transposition. Moreover, the frequency of transposition is in part set by the abundance of Tnp and the ratio of Tnp to Inh. Thus the incoming Tn*5* in the Biek & Roth experiment (4) encountered a preexisting pool of inhibiting Inh; as the copy number of Tn*5* increases, the concentration of *trans*-acting Inh increases, but the amount of *cis*-acting Tnp per Tn*5* remains constant.

What then determines the abundance of Tnp and Inh? The answers to this question are not only interesting for Tn*5*, they also reveal genetic regulatory motifs found in other systems.

Tnp and Inh are expressed from overlapping, probably competitive, promoters. This conclusion was deduced by a deletion analysis of in vivo promoter activities (21). As shown in Figure 2, the T2 transcript can only encode the Inh while the T1 transcript encodes both proteins [but the Inh is translated inefficiently from the T1 message (37)]. Such overlapping promoters are found in other systems (e.g. see 30) and can program contradictory functions. Thus by studying this Tn*5* arrangement we will elucidate a more general regulation strategy.

DNA (*dam*) methylation down-regulates the synthesis of the T1 (Tnp) transcript and appears to up-regulate the synthesis of the T2 transcript (43). The frequency of Tn*5* transposition is 10-fold higher in *dam* hosts, and this change in transposition seems to mirror (and presumably is the consequence of) a four- to fivefold increase in T1 mRNA and a twofold decrease in T2 mRNA. The opposite effect on promoter activity suggests that the promoters compete for RNA polymerase. An inspection of the DNA sequence indicates, and site-specific mutation studies prove, that the relevant GATC *dam*

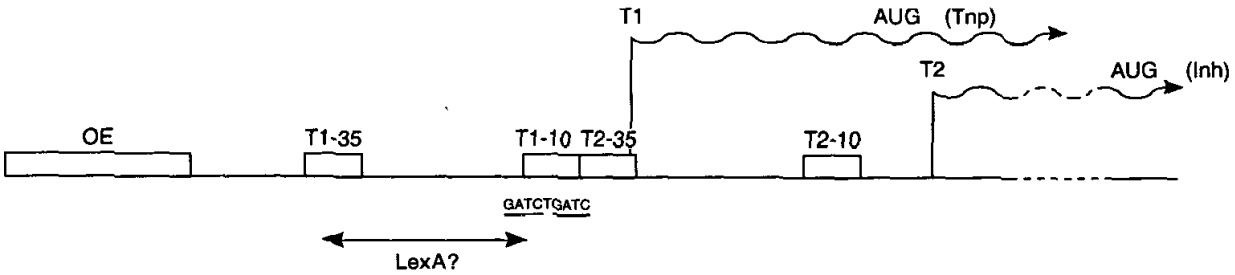



*Figure 2* Controlling elements for Tn*5*/IS*50* Tnp and Inh synthesis. The 5′ end of IS*50* is defined by the 19-bp OE sequence (see Figure 4, below). The promoter for transposase synthesis (T1) initiates transcription 66 bp from the end of IS*50* (21). Between the −35 and −10 regions of T1 is a weak LexA binding site (23). Overlapping the T1 −10 region are two Dam methylation sites that when methylated down regulate Tnp mRNA synthesis (43). The T2 (Inh) promoter overlaps T1 (21). The AUGs for Tnp and Inh are indicated (21).

methylation sites overlap with the −10 region of the T1 promoter. The specific relevance of *dam* regulation of Tnp (and Inh) synthesis is that it should couple transposition to DNA replication (about which more is discussed below) because newly replicated DNA is hemimethylated. Also, this regulation should stimulate transposition off of DNA that is newly introduced into the cell, which is exactly the observed result (27, 32).

Other transposon systems [in particular Tn*10* (31)] also display Dam methylation control of Tnp synthesis and of transposition. Moreover, these studies demonstrate in a quite precise manner a type of gene regulation displayed in many systems—chemical modification of specific DNA sequences to modulate protein binding. This mechanism reappears in the discussion of protein recognition of the IS*50* IE sequence.

Tn*5* could transpose into highly expressed genes, thereby giving rise to read-through transcripts into the transposase gene. Production of these transcripts might result in spurious transposase synthesis. However, a set of experiments (21) has demonstrated that read-through expression of transposase does not occur because such messages do not program the translation of Tnp.

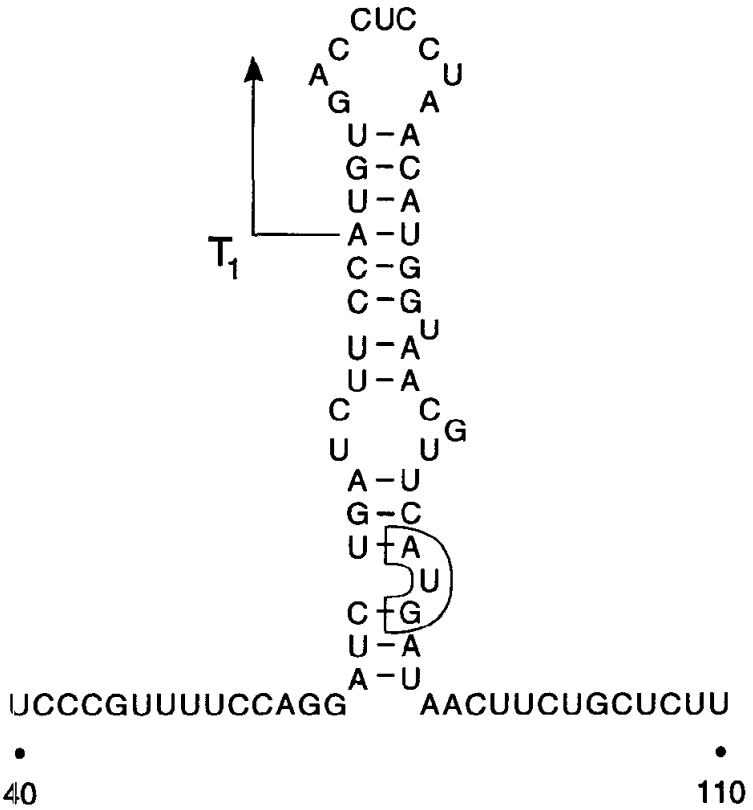



*Figure 3* Secondary structure in read-through transcripts blocks Tnp translation. Transcripts that read through the end of IS*50* do not express Tnp because of a secondary structure that occludes the Shine-Dalgarno sequence end and the AUG (21, 37). The start of the T1 transcript (which will not form a secondary structure) and the Tnp AUG are indicated. This figure is similar to one shown in Ref. 21.

An RNA secondary structure in these read-through RNAs occludes the Tnp translation initiation signals (37). As shown in Figure 3, RNA sequences upstream of the T1 transcript start site are necessary for the formation of this secondary structure. This mechanism prevents read-through expression of Tn*5* Tnp. The same mechanism appears to be shared by several other transposable elements [IS*2* (13), IS*3* (39), IS*4* (20), IS*5* (22), IS*150* (38), and Tn*10* (5)] and may be a general motif for *Escherichia coli* gene systems [for instance, the same pattern is seen for *lac* (36, 46) and *gal* (26)].

Do other mechanisms exist that regulate transposase synthesis? Recent reports present conflicting evidence regarding the possible role of LexA in regulating transposase synthesis. Our laboratory has found that LexA has no significant control over transposase levels (40), while Kuan & Tessman postulate that LexA represses transposase synthesis approximately fivefold, presumably by binding to a weak LexA binding site upstream of the T1 promoter (23) (see Figure 2). These studies differ in two respects: (*a*) Weinreich et al performed direct tests of a LexA effect (40), but the Kuan & Tessman studies were indirect (23), and (*b*) the sequence contexts of the two Tn*5* systems were different. The former difference is a matter of technique, while the latter may suggest an interesting molecular phenomenon. Many genetic regulatory proteins are enhanced in their DNA-binding activities through cooperative binding to a secondary site (1). For instance, binding of the *lac* repressor to its operator is thought to be stabilized through a bond to a secondary binding site, thereby forming a looped DNA structure (10, 11). Future studies might examine whether such a nearby LexA binding site is present in the Kuan-Tessman system, but not in the one which we studied, and if so, whether such a fortuitous nearby LexA binding site could facilitate LexA binding to its weak Tn*5* site.

## *The Transposase N-Terminal Sequence Is Critical for Sequence-Specific DNA Binding and Other Transposase Activities*

The transposase and its inhibitor are identical in sequence except that the Inh is missing 55 N-terminal amino acids (20). Because Inh does not promote transposition, it is missing one or more critical Tnp activities (15–17, 33, 45), suggesting that the N-terminal 55 amino acids encode a domain of importance. In vitro analyses of purified Tnp and purified Inh have shown that at least one property missing in the Inh preparations is a DNA sequence–specific binding activity. The Tnp, on the other hand, binds to OE and IE end sequences specifically and shows the expected sensitivity to OE sequence mutations and Dam methylation of the IE (N. de la Cruz, M. Weinreich, T. Wiegand, M. Krebs & W. Reznikoff, submitted; R. A. Jilk & W. Reznikoff, unpublished results). Recent studies have demonstrated that deletion of 11

N-terminal amino acids or introduction of 4 different N-terminal missense mutations destroys the DNA-binding activity of Tnp (M. D. Weinreich & W. Reznikoff, unpublished results). These experiments show that the N-terminal 55 amino acids are necessary for the sequence-specific DNA-binding activity and imply that the DNA-binding domain is in this region. Determining the precise DNA-binding domain will be instructive because the Tn5 transposase does not contain a previously classified DNA-binding motif, and yet the activity is clearly there.

Another surprising property requires the N-terminal three amino acids of Tnp. Overproduction of the transposase (in the absence of transposition) kills host cells. Deletion of as little as three amino acids prevents this killing (M. D. Weinreich & W. Reznikoff, unpublished results). A possible biological



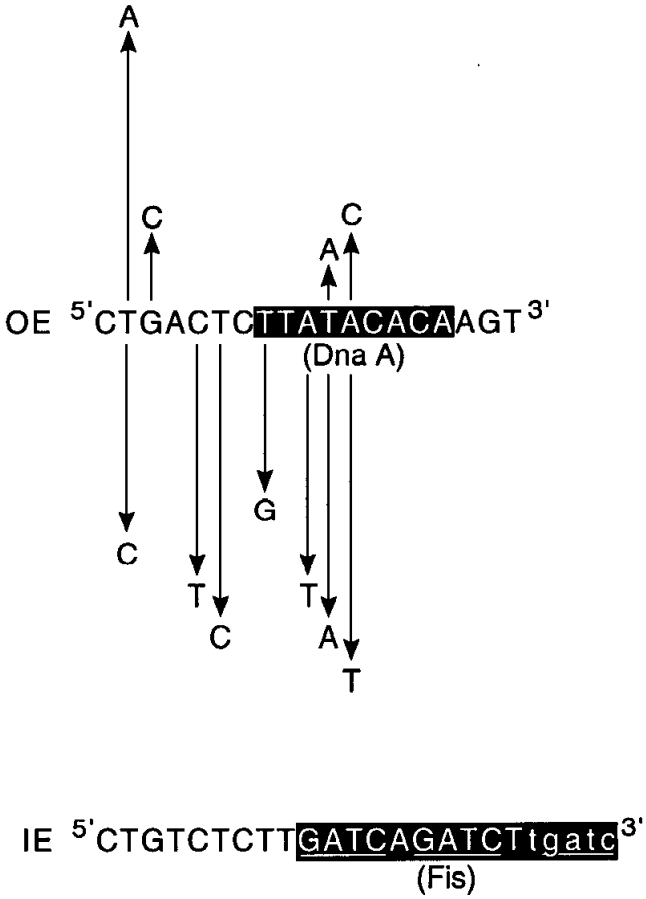



*Figure 4* (*Top*) Outside and inside end sequences. The DnaA box is indicated for the outside end sequence (12, 17, 34, 44). Downward-pointing arrows denote mutations that act like a deletion in the adjacent deletion experiment. Upward-pointing arrows denote mutations that constitute the second class in the adjacent deletion experiment (16). The 11A change gave rise to one deletion in the second class. (*Bottom*) For the inside end sequence, the Dam sites and the overlapping Fis binding site are indicated (34, 41). Dam site methylation regulates both Tnp binding (43) and Fis binding (41). The lower-case letters indicate base pairs outside of the required 19-bp sequence. The Tnp/Inh termination codon is the complement to positions 12–14.

significance for the cell killing will be described in the section on chromosome replication below.

## *Recognition of Terminal 19–Base Pair Sequences*

The Tn*5* and IS*50* transposable elements are defined by the terminal 19-bp sequences. Tn*5* is delineated by two inverted OE sequences and IS*50* by an OE and an IE sequence (see Figure 4) (17, 34). All transposase-mediated genetic events (save one mentioned below) occur at the precise boundary of the relevant 19-bp sequences, and almost all changes in these sequences greatly reduce or abolish the frequency of transposition (24, 29). Presumably the importance of these sequences lies in the fact that they are recognized by proteins during the transposition process.

Although we are not certain of all of the salient facts, the protein-DNA interactions occurring at the OE and IE sequences clearly display a remarkable degree of complexity; at least two proteins recognize the OE and three proteins recognize the IE. In addition, different portions of the OE sequence may play roles in different steps in the transposition process.

The OE end is recognized by both the transposase and the host protein DnaA (17, 29, 44). Tnp recognition of OE is fundamental to the transposition process. Tnp binds specifically to OE sequences in vitro as judged by gel retardation assays (42; N. de la Cruz, M. Weinreich, T. Wiegand, M. Krebs & W. Reznikoff, submitted). However, transposase recognition of the OE may be much more complex than a simple binding reaction. For instance, some base pairs could be recognized during the strand-cleavage reaction.

The first evidence that various OE sequences dictate different steps in the transposition process came from a complicated set of in vivo observations (16). Tnp can catalyze deletions adjacent to its site of insertion starting at the end of the OE sequence. These are most easily detected in situations in which the OE far removed (distal) from the location of the adjacent deletion is defective, greatly reducing or blocking transposition. Two classes of adjacent deletions have been found, and the nature of the class depends upon the precise change in the distal OE. If the distal OE is deleted, then the adjacent deletions start immediately adjacent to the functional OE. Some OE point mutations (which totally block transposition) also fall into this class. Mutant OE sequences of this class that have been tested for transposase binding in vitro bind transposase with a lower affinity than the wild-type OE (R. A. Jilk & W. Reznikoff, unpublished results). The second class of adjacent deletion is unusual in that this type starts one bp removed from the OE. These adjacent deletions are found in association with another subset of distal OE mutations, some of which are only partially defective in transposition and do not seem to impair transposase binding in vitro (16; R. A. Jilk & W. Reznikoff, unpublished results). Figure 4 displays the distribution of bp defects. Because

different subsets of OE point mutations give rise to different classes of adjacent deletions, this implies a functional fine structure to the OE. This fine structure is also seen in the binding of Tnp to the OE. Some OE mutations disrupt Tnp binding while others do not. A possible explanation for some of the mutations that do not impair transposase binding is that they are recognized by a different protein, DnaA. However, this explanation does not fit some cases (2A, 3C) because the mutations are located outside of the DnaA box as defined by sequence homology (12, 44). On the other hand, some mutations that presumably alter transposase binding (8G, 10T, 11A) are within the DnaA box, suggesting that these base pairs may be recognized by both proteins.

A closer examination of the results summarized in Figure 4 suggests an additional level of complexity. At positions 2 and 12 different mutations are associated with different classes of adjacent deletion formation—perhaps in these two cases the same base pair is recognized in different ways at different steps in the transposition process. Possibly both DnaA and transposase recognize position 12, and different base pair changes may discriminate differentially between these two protein-recognition processes. For position 2, Tnp might make different molecular contacts with the same base pair at different steps in the overall reaction.

DnaA also recognizes the OE sequence. This observation was first made through footprinting studies of the OE (12). Later experiments tested the relevance of this binding, showing that Tn*5* and IS*50* transposition occurs at a reduced efficiency in *dna* hosts, and that a Tn*5* derivative with a mutation in the OE DnaA box did not appear to manifest sensitivity to the host *dna* genotype (29, 44). Current experiments are examining the effect of DnaA on transposase binding and are attempting to determine the in vitro properties of DnaA binding to various OE mutant sequences.

Thus, some end-sequence base pairs are recognized in very complex ways, at different steps in the transposition reaction and/or by different proteins. A similar complexity for end sequences has been suggested for other transposable elements (7, 14).

The IE sequence is even more complicated. It contains recognition sites for two proteins (transposase and Dam DNA methylase); it overlaps a binding site for a third protein, Fis, and also encodes the termination codon for the transposase (see Figure 4). The IE sequence and the OE sequence differ at 7 out of 19 positions, yet the transposase binds in vitro to an unmethylated IE with an affinity resembling its binding to the OE (R. A. Jilk & W. Reznikoff, unpublished results). This Tnp-IE binding result is consistent with the observed transposition preferences (9, 25). This result is quite surprising because it suggests considerable flexibility in the DNA recognition domain of Tnp. Alternatively, we could hypothesize that the positions that differ are not recognized by Tnp either because they do not play a critical protein-rec-

ognition role [position 4 may be an example of this (24)] or because they are recognized by some other protein (e.g. DnaA). However, the in vivo adjacent deletion studies suggest that mutations at some of these dissimilar positions do reduce Tnp binding directly (16). Thus we conclude that Tnp specifically recognizes two quite diverse sequences—another reflection of the complex substructures of IE and OE.

Dam DNA methylation inhibits transposase binding to the IE. Transposition of IS*50* is sensitive to Dam DNA methylation because of the inhibition of two separate sequence-specific protein-DNA interactions. As described above, transcription initiation programmed by the T1 (transposase) promoter is directly inhibited by Dam DNA methylation of sites overlapping the −10 region (43). However, Dam DNA methylation also inhibits IS*50* (but not Tn*5*) transposition even when the transposase synthesis is programmed by a Dam-insensitive promoter (*tac*), indicating a direct effect of Dam methylation on IE recognition during transposition (24, 29, 43). This effect is presumably mediated through the Dam sites within the IE (see Figure 4). We recently showed that Dam methylation of IE DNA fragments reduces transposase binding in vitro (R. A. Jilk & W. Reznikoff, unpublished experiments). This observation is consistent with direct Dam DNA methylation control over transposition reactions using the IE. Very similar effects of DNA methylation that regulate Tn*10*/IS*10* transposition have also been observed (31). As noted below, DNA methylation regulation of transposition may suggest a coupling between transposition and DNA replication.

The role of Dam DNA methylation regulation of IS*50* transposition (utilizing the IE) is complicated by the fact that a second protein (Fis) binds to a sequence overlapping the IE, and this binding (which is also blocked by Dam DNA methylation) is thought to compete with transposase-IE binding. Some in vivo experiments designed to study the frequency of IS*50* transposition actually used various recombinant DNA constructs that have the same general structure as IS*50* (OE–transposase gene–reporter gene–IE) but differ in the size of the IE sequence. These studies gave quite different results depending upon whether a 19- or a 24-bp sequence was used for IE. The 24-bp IE construct demonstrates much lower transposition frequencies in a *dam* host (9, 25). We now know that the 24-bp sequence (but not the 19-bp sequence) contains the overlapping Fis-binding site, and that Fis acts to inhibit transposition in a Dam-dependent manner (41). In vitro gel retardation and footprinting studies confirm that Fis binds to this sequence and that it binds efficiently only to the unmethylated site (41). The physiological significance of this Fis effect is unclear, but because Fis abundance varies with the stage of bacterial growth, it may act to dampen IS*50* transposition from newly replicated DNA during the exponential phase, when it is most abundant.

The translation termination codon for the transposase (and its inhibitor) is located within the IE at positions 14–12. We have no information as to whether this influences transposition utilizing an IE or whether Tnp bound to the IE can influence the synthesis of the transposase.

## *p1 Is a* Cis-*Active Transposase (and a* Trans-*Active Transposition Inhibitor)*

The Tn*5* transposase functions primarily *in cis,* acting preferentially on OE-OE or OE-IE sequences located close to the site of transposase synthesis on the same relicon (15, 19). This observation has also been made for the transposase proteins encoded by other transposons such as Tn*10* and Tn*903* (8, 28). This preferential *cis* activity could result from either a chemical or functional instability in the transposase, and/or from a sequestration of the transposase during or shortly after synthesis. The Tn*903* transposase is known to be chemically unstable (8), but various studies have indicated that Tn*5* Tnp is, by and large, chemically stable in vivo (see 33).

My current hypothesis (pictured in Figure 5) is that the Tn*5* Tnp is *cis*-active owing to protein conformation changes that are regulated by two factors; by the release of the Tnp protein from the translation apparatus and by the oligomerization of Tnp with monomers of Inh or Tnp. We propose that the nearly complete, tethered Tnp has a high affinity end-sequence binding activity and can initiate the transposition process prior to completion of its translation. Such a property would obviously lead to *cis* but not *trans* activity. The released completed Tnp is proposed to have two conformations in equilibrium, a low-abundance active conformation similar to the tethered protein and a high-abundance inactive conformation. Furthermore, the formation of Tnp-Tnp or Tnp-Inh oligomers probably occurs with the inactive conformation (or oligomerization inactivates the active conformation). The

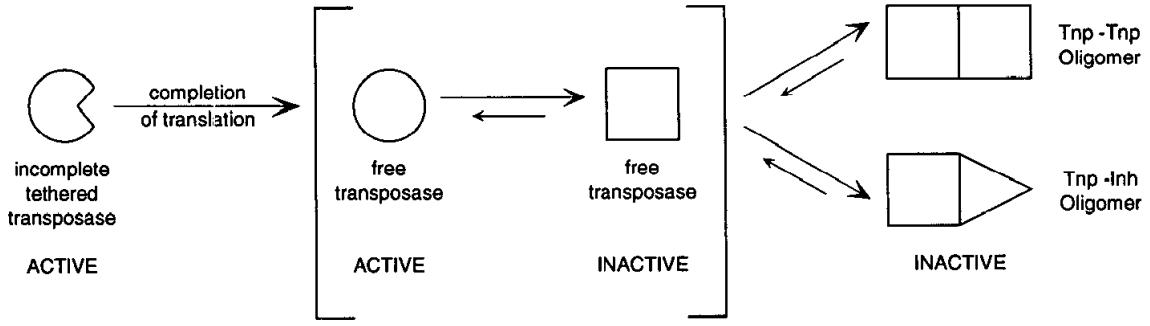



*Figure 5* Transposase acts *in cis*. The Tn*5* transposase functions primarily *in cis* (14, 18). Presumably, Tnp tethered to the translation apparatus can initiate the transposition reaction. Once Tnp synthesis is complete, the molecule favors an inactive conformation, and Tnp-Tnp or Tnp-Inh oligomerization prior to end-sequence binding will also inactivate Tnp.

longer a Tnp monomer exists without end-sequence binding the higher is the probability of oligomerization. Thus this property would also lead to *cis* activity.

The proposed model includes aspects of sequestration (the incomplete tethered Tnp is held in close proximity to its gene) and functional instability (a conformational change facilitated by oligomerization). The evidence for this model is quite circumstantial, but several of its predictions are clear.

The primary evidence supporting the notion that the tethered nascent protein may bind end sequences with an enhanced affinity came from gel retardation experiments examining the binding of Tnp subfragments to OE DNA. An in vitro synthesized Tnp subfragment lacking 100 carboxyl terminal amino acids binds specifically to OE DNA and does so with an affinity ~10-fold higher than that of full-length Tnp (L. Mahnke & W. Reznikoff, unpublished results). This observation is consistent with the OE-binding domain residing at the N terminus (see section 2 of this review) and suggests that this OE binding domain is occluded in most of the intact Tnp.

However, the completed Tnp is not completely inactive as Tnp does have some *trans* activity. We have recently described Tnp mutants that increase the frequency of transposition both *in cis* and *in trans* (42). These mutant proteins may have altered the equilibrium between active and inactive conformations of Tnp, and the mutations are believed to have a conformational effect because the mutant Tnp proteins are more sensitive to proteolytic degradation (T. Wiegand & W. Reznikoff, unpublished results).

In vivo studies suggest that protein oligomerization regulates Tnp activity. The Inh protein is known to inhibit transposase activity *in trans* (15, 18, 19, 45). Two simple models explain this inhibitory activity. Inh might bind to the end sequences forming inactive Inh–end sequence complexes, thereby blocking Tnp access to these sequences. However, Inh and other N-terminal deletions of Tnp cannot bind end-sequence DNA efficiently (N. de la Cruz, M. Weinreich, T. Wiegand, M. Krebs & W. Reznikoff, submitted); thus this model is not correct. Alternatively, Inh could form mixed oligomers with Tnp and alter Tnp activity. Evidence suggests that Inh does oligomerize with Tnp and that the oligomers do have altered DNA-binding activity (N. de la Cruz, M. Weinreich, T. Wiegand, M. Krebs & W. Reznikoff, submitted).

Also suggesting that Tnp activity is regulated by oligomerization is the observation that Tnp itself inhibits transposition *in trans*. This property of Tnp was discovered during an in vivo analysis of a Tnp mutant that fails to make the Inh protein (42). The MA56 mutant destroys the Inh start codon and introduces an alanine in place of methionine at codon 56 of Tnp. This mutant is fully functional for transposition *in cis* but inhibits transposition *in trans*.

Although the model described above and in Figure 5 is consistent with all

the observations, direct experiments are needed to test it. For instance, do the hypertransposing mutants alter the conformation of Tnp? Will tethered, incomplete Tnp bind to OE DNA with a high affinity? Does oligomerization favor the inactive Tnp conformation?

## *Transposition May Be Linked to Chromosome Replication*

Some circumstantial evidence might link Tn*5* transposition to the process of chromosome replication and partition. In particular the following host functions are known to influence the frequency of Tn*5* transposition:

Dam: a protein catalyzing postreplicative DNA methylation; *dam* cells have elevated transposition frequencies (43).

DnaA: an *oriC*-binding protein required for the initiation of DNA replication (29, 44); *dnaA* cells have reduced transposition frequencies.

SulA: a protein that inhibits cell division; *sulA* cells have elevated transposition frequencies (35).

In addition, DeLong & Syvanen (6) reported that a small fraction of transposase is associated with the inner membrane fraction upon cell lysis. These correlations are intriguing but do not show a mechanistic link.

Observations that suggest a functional link between transposition and chromosome partition came from experiments analyzing the cellular consequences of transposase over expression. Cell death occurs in the absence of transposition as well as any other Tn*5*-encoded function (such as Inh or the end sequences) (M. D. Weinreich, unpublished results). The dying cells form long multinuclear filaments, indicating a failure in partition-dependent septation. *E. coli* mutants resistant to transposase overproduction killing continue to divide when transposase is overproduced. The mutations mapped using Hfr crosses reside at more than one locus. The one that has been most closely examined maps near the *fts* locus at 76 min (M. D. Weinreich & W. Reznikoff, unpublished results). *fts* genes encode functions required for septation.

These results point to a possible link between transposition and chromosome-partition function. This link is in addition to the preferential transposition of Tn*5* or IS*50* sequences off of newly replicated DNA (which results from *dam* regulatory effects) and to the sharing of DnaA in both transposition and replication processes. However, at least the link to chromosome partition can be bypassed; transposition does occur in the mutants resistant to transposase overproduction (M. D. Weinreich & W. Reznikoff, unpublished results).

How is this coupling accomplished and what is the advantage of this arrangement? We hope an analysis of the mutants mentioned above will suggest an answer to the first question. As for the second, many of the host functions participating in Tn*5* transposition are assumed to be components of a sequestered organelle involved in host-chromosome replication. One possi-

ble advantage to Tn*5* accruing from such an association is that a relationship between Tn*5* and the same organelle might have evolved in order to insure that the organelle's transposition apparatus would have efficient access to these functions.

Alternatively, the proposed coupling of Tn*5* transposition to chromosome replication and partition might decrease the risk of genomic "suicide" occurring as a result of the conservative cut-and-paste transposition process. Transposition is most likely to happen soon after DNA replication (when there are two copies of the donor DNA sequence) due to the influence of *dam* DNA methylation on transposase synthesis and inside-end usage. If, however, transposase inhibits chromosome partitioning, then this might delay cell division in the very cells in which transposition is most likely to have occurred, e.g. cells that contain transposase, assuring that the ongoing chromosome replication would be completed prior to cell division. Thus transposition would be biased towards cells in which a copy of the parental genome is maintained and can be inherited.

## *Conclusion*

The Tn*5*/IS*50* system has been an amazingly productive object of experimental inquiry. The work to date has elucidated (and future work will continue to elucidate) the mechanisms by which transposition occurs and is regulated and will also continue to provide insights into important aspects of molecular biology.

One of the interesting aspects of Tn*5* molecular biology is the density of information encoded within a short sequence. The IS*50* sequence is 1533 bp in length. This element encodes all of the Tn*5* functions required for transposition and all of the Tn*5* regulatory mechanisms. Thus in the first 259 bp, starting at the OE, we find a complex Tnp-specific sequence, a DnaA-binding site, an imperfect symmetry element that blocks the expression of read-through mRNA, a promoter for Tnp mRNA synthesis, a possible LexA binding site, Dam methylation sites regulating the Tnp promoter activity, a promoter for Inh mRNA synthesis, translation initiation signals for Tnp and Inh, and sequences encoding the important N terminus of Tnp. Most of the rest of the sequence up to the Fis site, which overlaps the IE, is involved with encoding Tnp and Inh, although *cis*-active sites may also be in this region [e.g. a Fis binding site of unknown function (41)].

I have not discussed the bulk of the protein-coding sequences in this review because the required Tnp (and Inh) structure-function studies have not been performed. For instance, the Tnp presumably executes the following operations: binding of OE and IE, bringing the ends together (or inactivating the unbound Tnp) through oligomerization, cutting the DNA next to the end sequences, cleaving target DNA to give 9-bp 5′ overhangs, and inserting the

transposable element into target DNA. Each of the functions requires particular Tnp domains of critical importance. Where are they? How are they organized in a three-dimensional structure? Understanding the molecular basis of Tnp *cis* activity should also reveal fascinating insights into this protein's structure-function relationships. These and other questions will be the object of future experiments, which will require detailed genetic, biochemical, and structural analyses.

Even among the transposition operations for which we have substantial information, much work is left to be done. For instance, how does DnaA binding to the OE influence the Tnp-OE interaction? How do the different base pairs of the OE interact with Tnp and at what steps of the transposition reaction does this occur? How do the different base pairs of the IE function with regard to Tnp activity? How does Fis perturb the Tnp-IE interaction? What is the active form of the Tnp, and how many molecules are involved in synaptic complex formation?

Finally, research efforts should be directed at determining if transposition is coupled to the cell-division cycle and how that linkage is accomplished.



## ACKNOWLEDGMENTS

The research in the author's laboratory has been supported by grants from the NIH (GM19670), the NSF (DMB-9020517), and the American Cancer Society (MV-554). The author is indebted to the many members of his laboratory who have contributed their imagination and work, and the editor of this volume for very helpful suggestions.

## *Literature Cited*

1. Bellomy, G. R., Record, M. T. Jr. 1990. DNA loop formation: role in gene regulation and implications for DNA structure. In *Progress in Nucleic Acid Research and Molecular Biology*, ed. W. Cohn, K. Moldave, 39:81–128. New York: Academic
2. Berg, D. E. 1989. Transposon Tn*5*. See Ref. 3, pp. 185–210
3. Berg, D. E., Howe, M. M. 1989. *Mobile DNA*. Washington, DC: Am. Soc. Microbiol.
4. Biek, D., Roth, J. R. 1990. Regulation of Tn*5* transposition in *Salmonella typhimurium*. *Proc. Natl. Acad. Sci. USA* 77:6047–51
5. Davis, M. A., Simons, R. W., Kleckner, N. 1985. Tn10 protects itself at two levels from fortuitous activation by external promoters. *Cell* 43:379–97
6. DeLong, A., Syvanen, M. 1990. Membrane association of the Tnp and Inh proteins of IS*50*R. *J. Bacteriol.* 172: 5516–19
7. Derbyshire, K. M., Hwang, L., Grindley, N. D. F. 1987. Genetic analysis of the interaction of the insertion sequence IS*903* transposase with its terminal inverted repeats. *Proc. Natl. Acad. Sci. USA* 84:8049–53
8. Derbyshire, K. M., Kramer, M., Grindley, N. D. F. 1990. Role of instability in the *cis* action of the insertion sequence IS*903* transposase. *Proc. Natl. Acad. Sci. USA* 87:4048–52
9. Dodson, K. W., Berg, D. E. 1989. Factors affecting transposition activity of IS*50* and Tn*5* ends. *Gene* 76:207–13
10. Eismann, E., von Wilcken-Bergmann, B., Muller-Hill, B. 1987. Specific destruction of the second *lac* operator decreases repression of the *lac* operon in *Escherichia coli* fivefold. *J. Mol. Biol.* 195:949–52

11. Flashner, Y., Gralla, J. D. 1988. Dual mechanism of repression at a distance in the *lac* operon. *Proc. Natl. Acad. Sci. USA* 85:8968–72
12. Fuller, R. S., Funnell, B. E., Kornberg, A. 1984. The dnaA protein complex with the E. coli chromosomal replicative origin (oriC) and other DNA sites. *Cell* 38:889–900
13. Ghosal, D., Sommer, H., Saedler, H. 1979. Nucleotide sequence of the transposable DNA-element IS2. *Nucleic Acids Res.* 6:1111–22
14. Huisman, O., Errada, P. R., Signon, L., Kleckner, N. 1989. Mutational analysis of IS10's outside end. *EMBO J.* 8:2101–9
15. Isberg, R. R., Lazaar, A. L., Syvanen, M. 1982. Regulation of Tn5 by the right repeat proteins: control at the level of the transposition reactions? *Cell* 30:883–92
16. Jilk, R. A., Makris, J. C., Borchardt, L., Reznikoff, W. S. 1993. Implications of Tn*5* associated adjacent deletions. *J. Bacteriol.* 175:1264–71
17. Johnson, R. C., Reznikoff, W. S. 1983. DNA sequences at the ends of transposon Tn*5* required for transposition. *Nature* 304:280–82
18. Johnson, R. C., Reznikoff, W. S. 1984. The role of the IS*50*R proteins in the promotion and control of Tn*5* transposition. *J. Mol. Biol.* 177:645–61
19. Johnson, R. C., Yin, J. C.-P., Reznikoff, W. S. 1982. The control of Tn5 transposition in Escherichia coli is mediated by protein from the right repeat. *Cell* 30:873–82
20. Klaer, R., Kuhn, S., Tillmann, E., Fritz, H.-J., Starlinger, P. 1981. The sequence of IS*4*. *Mol. Gen. Genet.* 181:169–75
21. Krebs, M. P., Reznikoff, W. S. 1986. Transcriptional and translational sites of IS*50*. Control of transposase and inhibitor expression. *J. Mol. Biol.* 192: 781–91
22. Kroger, M., Hobom, G. 1982. Structural analysis of insertion sequence IS*5*. *Nature* 297:159–62
23. Kuan, C.-T., Tessman, I. 1991. LexA protein of *Escherichia coli* represses expression of the Tn*5* transposase gene. *J. Bacteriol.* 173:6406–10
24. Makris, J. C., Nordmann, P. L., Reznikoff, W. S. 1988. Mutational analysis of IS*50* and Tn*5* ends. *Proc. Natl. Acad. Sci. USA* 85:2224–28
25. Makris, J. C., Reznikoff, W. S. 1989. Orientation of IS*50* transposase gene and transposition. *J. Bacteriol.* 171: 5212–14
26. Merril, C. R., Gottesman, M. E., Adhya, S. L. 1981. *Escherichia coli gal* operon proteins made after prophage lambda induction. *J. Bacteriol.* 147: 875–87
27. McCommas, S. A., Syvanen, M. 198. Temporal control of Tn*5* transposition. *J. Bacteriol.* 170:889–94
28. Morisato, D., Way, J. C., Kim, H.-J., Kleckner, N. 1983. Tn10 transposase acts preferentially on nearby transposon ends in vivo. *Cell* 32:799–807
29. Phadnis, S. H., Berg, D. E. 1987. Identification of base pairs in the outside end of insertion sequence IS*50* that are needed for IS*50* and Tn*5* transposition. *Proc. Natl. Acad. Sci. USA* 84:9118–22
30. Reznikoff, W. S., Bertrand, K., Donnelly, C., Krebs, M., Maquat, L. E., et al. 1987. Complex promoters. In *RNA Polymerase and the Regulation of Transcription,* ed. W. S. Reznikoff, R. R. Burgess, J. E. Dahlberg, C. A. Gross, M. T. Record, Jr., M. P. Wickens, pp. 105–13. New York: Elsevier
31. Roberts, D. E., Hoopes, B. C., McClure, W. R., Kleckner, N. 1985. IS10 transposition is regulated by DNA adenine methylation. *Cell* 43:117–30
32. Rosetti, O. L., Altman, R., Young, R. 1984. Kinetics of Tn*5* transposition. *Gene* 32:91–8
33. Rothstein, S. J., Reznikoff, W. S. 1981. The functional differences in the inverted repeats of the Tn5 are caused by a single base pair nonhomology. *Cell* 23:191–99
34. Sasakawa, C., Carle, G. F., Berg, D. E. 1983. Sequences essential for transposition at the termini of IS*50*. *Proc. Natl. Acad. Sci. USA* 80:7293–97
35. Sasakawa, C., Uno, Y., Yoshikawa, M. 1987. *lon-sulA* regulatory function affects the efficiency of transposition of Tn*5* from λ b221 cI857 Pam Oam to the chromosome. *Biochem. Biophys. Res. Commun.* 142:879–84
36. Schulz, V. P., Reznikoff, W. S. 1990. In vitro secondary structure analysis of mRNA from *lacZ* translation initiation mutants. *J. Mol. Biol.* 211:427–45
37. Schulz, V. P., Reznikoff, W. S. 1991. Translation initiation of IS*50*R readthrough transcripts. *J. Mol. Biol.* 221: 65–80
38. Schwartz, E., Kroger, M., Rak, B. 1988. IS*150:* distribution, nucleotide sequence and phylogenetic relationships of a new *E. coli* insertion element. *Nucleic Acids Res.* 16:6789–99
39. Timmerman, K. P., Tu, C.-P. D.

1985. Complete sequence of IS*3*. *Nucleic Acids Res.* 13:2127–39
40. Weinreich, M. D., Makris, J. C., Reznikoff, W. S. 1991. Induction of the SOS response in *Escherichia coli* inhibits Tn*5* and IS*50* transposition. *J. Bacteriol.* 173:6910–18
41. Weinreich, M. D., Reznikoff, W. S. 1992. Fis plays a role in Tn*5* and IS*50* transposition. *J. Bacteriol.* 174:4530–37
42. Wiegand, T. W., Reznikoff, W. S. 1992. Characterization of two hyper-transposing Tn*5* mutants. *J. Bacteriol.* 174:1229–39
43. Yin, J. C.-P., Krebs, M. P., Reznikoff, W. S. 1988. The effect of *dam* methylation on Tn*5* transposition. *J. Mol. Biol.* 199:35–46
44. Yin, J. C.-P., Reznikoff, W. S. 1987. *dnaA*, an essential host gene, and Tn*5* mutants. *J. Bacteriol.* 169:4637–45
45. Yin, J. C.-P., Reznikoff, W. S. 1988. p2 and the inhibition of Tn*5* transposition. *J. Bacteriol.* 170:3008–15
46. Yu, X. M., Munson, L. M., Reznikoff, W. S. 1984. Molecular cloning and sequence analysis of *trp-lac* fusion deletions. *J. Mol. Biol.* 172:355–62

*Annual Review of Microbiology*
*Volume 47, 1993*

# CONTENTS



vi (*continued*)