IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Application No.	:	17/684,328	Conf. No.: 5469
Inventor(s)	:	Pardis SABETI, *et al.*	
Filed	:	March 1, 2022	
TC/AU	:	1681	
Examiner	:	Angela Marie Bertagna	
Customer No.	:	198696	
Docket No.	:	BROD-5360US	
Title	:	AMPLICON-BASED SEQUENCING USING DNA SPIKE-INS	

AMENDMENT AND RESPONSE

FILED VIA EFS-WEB

Commissioner:

In response to the Non-Final Office Action mailed December 4, 2024 ("Office Action"), please enter and consider the following amendments and remarks.

AMENDMENT

In the Specification:

A substitute specification is submitted herewith.

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 method ~~of~~ for detecting ~~and preventing~~ contamination in ~~amplification-based assays~~ amplicon sequencing, said method ~~comprising~~ consists of:

[[a.]](a) adding a synthetic DNA spike-in (SDSI) to one or more cDNA samples,

wherein each SDSI is capable of amplification simultaneously with a target sequence present in the one or more cDNA samples, ~~and~~

wherein each SDSI comprises

a unique core nucleotide sequence capable of differentiating each SDSI, covalently linked to an upstream 5' primer binding region, and a downstream 3' primer binding region,

wherein the nucleotide sequence of the 5' and 3' primer binding regions is the same in each SDSI;

[[b.]](b) amplifying the one or more target sequence(s) ~~and the SDSI~~ present in the one or more cDNA samples together with the unique core sequence present in each added SDSI;

[[c.]](c) sequencing the amplified target and SDSI ~~sequences~~ sequence(s), and ~~the SDSI(s)SDSI; and~~

[[d.]](d) determining the presence of the amplified SDSI sequence(s) ~~from~~ in the one or more cDNA samples,

wherein detection of a single SDSI unique core sequence in each cDNA ~~per~~ sample indicates contamination-free amplification~~,~~ of each cDNA sample, and

wherein detection of more than one SDSI unique core sequence in a ~~per~~ cDNA sample indicates possible contamination of that cDNA sample with cDNA(s) from a different cDNA sample.

2. (Currently Amended) The method of claim 1~~, wherein the SDSI contains a unique core region and a primer binding region at the 3' end and the 5' end~~, wherein ~~the~~ each SDSI unique core sequence is selected to minimize~~s~~ self-hybridization and cross-hybridization with other nucleic acid molecules present in the one or more cDNA sample~~s~~.

3. (Currently Amended) The method of claim 2, wherein the sequence homology between each SDSI's unique core sequence and the sequence of the nucleic acid molecules in each cDNA sample ~~core sequence homology~~ is less than 65%, or less than 60%, or less than 55%, or less than 50%, or less than 45%, or less than 40%, or less than 35%, or less than 30%, or less than 25%, or less than 20%, or less than 15%, or less than 5%, or less than 1%, and/or

~~to a sample sequence~~ wherein the length of the sequence homology between each SDSI's unique core sequence and the sequence of the nucleic acid molecules in each cDNA sample is less than 15, or less than 20, or less than 25, or less than 30, or less than 35, or less than 40, or less than 45, or less than 50 contiguous bases, and/or

wherein the SDSIs ~~SDSI sequences~~ are from 50[[-]] to 5000 nucleotides in length.

4. (Canceled)

5. (Canceled)

6. (Currently Amended) The method of claim 2, wherein each SDSI's unique ~~the~~ core sequence ~~of the SDSI sequence~~ is derived from a rare organism, optionally wherein the rare organism is a thermophilic archae~~a~~on.

7. (Canceled)

8. (Currently Amended) The method of claim 2, wherein the unique core sequence of the SDSI comprises a nucleotide sequence chosen from any one of ~~as set forth in~~ SEQ ID [[NOS]] NOs: 1 – 96 and 193-291, optionally, wherein the ~~one or more of the~~ unique core sequences comprising nucleotide sequences of SEQ ID[[NOS]] NOs: 16, 57, and 66 are substituted for nucleotide sequences of SEQ ID [[NOS]] NOs: 289, 290, and [[291]]291, respectively.

9. (Currently Amended) The method ~~SDSIs~~ of claim 2, wherein the SDSIs comprise nucleotide sequences of ~~one or more of~~ SEQ ID [[NOS]] NOs: 97-192 and 292-390, optionally, wherein ~~one or more of~~ the unique core sequences comprising nucleotide

sequence of SEQ ID [[NOS]] NOs: 112, 153, and 169 are substituted for nucleotide sequences of SEQ ID [[NOS]] NOs: 388, 389, 390, respectively.

10. (Currently Amended) The method of claim 2, wherein the primer binding ~~sites~~ regions have a Tm between 55-65 °C.

11. (Currently Amended) The SDSIs of claim 2, ~~wherein the primer binding sequences are complementary to the primers having SEQ ID NOS: 391 and 392.~~ wherein a forward primer comprising a nucleotide sequence of SEQ ID NO: 391 and a reverse primer comprising a nucleotide sequence of SEQ ID NO: 392 can anneal to the 5' and 3' primer binding regions of each SDSI, respectively.

12. (Canceled)

13. (Canceled)

14. (Original) The method of claim 1, wherein the concentration of the SDSI ranges from 0.1 femtomolar – 1.0 femtomolar.

15. (Currently amended) The method of claim 1, wherein the one or more cDNA samples [[is]] are for sequencing a pathogen or family of pathogens, optionally, wherein the pathogen is a virus or a bacteria, and the region of the bacterial genome sequenced is associated with antibiotic resistance.

16. (Canceled)

17. (Canceled)

18. (Currently amended) The method of claim 1, wherein ~~each~~ the one or more cDNA samples contain[[s a]] viral nucleic acid sequences.

19. (Currently amended) The method of claim 1, wherein the one or more cDNA samples are for creating one or more sequencing families/clusters.

20. (Currently amended) The method of claim 1, wherein the one or more cDNA samples ~~and SDSI~~ are ~~simultaneously~~ obtained by reverse transcription ~~from~~ of their respective RNA(s).

21. (Withdrawn - Currently amended) A set of synthetic DNA spike-ins (SDSIs), each spike-in [[in]] the set comprising a primer binding sequence at the 3' and 5' end and a unique core sequence between the 3' and 5' primer binding sequences, wherein the unique core sequence of each SDSI is selected to minimize~~s~~ self-hybridization and cross-hybridization with nucleic acids molecules present in each cDNA ~~in the~~ sample.

22. (Withdrawn - Currently amended) The SDSIs of claim 21, wherein the SDSI's sequence is 50-5000 nucleotides in length.

23. (Withdrawn - Currently amended) The SDSIs of claim 21, the sequence homology between each SDSI's unique core sequence and the cDNA sequences in each cDNA sample ~~wherein the core sequence homology~~ is less than 65%, or less than 60%, or less than 55%, or less than 50%, or less than 45%, or less than 40%, or less than 35%, or less than 30%, or less than 25%, or less than 20%, or less than 15%, or less than 5%, or less than 1% ~~to a sample sequence~~.

24. (Withdrawn) The SDSIs of claim 21, wherein the unique core sequence homology is less than 15, or less than 20, or less than 25, or less than 30, or less than 35, or less than 40, or less than 45, or less than 50 contiguous bases in common with the sample sequence.

25. (Withdrawn - Currently amended) The SDSIs of claim 21, wherein the SDSI's unique core sequence is derived from a rare organism, optionally wherein the rare organism is a thermophilic archaea.

26. (Canceled)

27. (Withdrawn - Currently amended) The SDSIs of claim 21, wherein the set of synthetic DNA spike-ins (SDSIs) comprises at least 96 spike-ins.

28. (Withdrawn - Currently amended) The SDSIs of claim 21, wherein the SDSI's unique core sequence ~~are the unique sequences as set forth in~~ is chosen from the nucleotide sequence of SEQ ID [[NOS]] NOs: 1 – 96 and [[195]]193-[[293]]291.

29. (Withdrawn - Currently amended) The SDSIs of claim 21, wherein ~~the~~ each SDSI~~s~~ comprises ~~one or more~~ a nucleotide sequence chosen from ~~of~~ SEQ ID [[NOS]]NOs: 97-192 and [[294]]292-[[392]]390.

30. (Withdrawn - Currently amended) The SDSIs of claim 21, wherein the SDSI's primer binding ~~sites~~ regions have a Tm between 55-65 °C.

31. (Withdrawn - Currently amended) The SDSIs of claim 21, ~~wherein the primer binding sequences are complementary to the primers having SEQ ID NOS: 391 and 392.~~ wherein a forward primer comprising a nucleotide sequence of SEQ ID NO: 391 and a reverse primer comprising a nucleotide sequence of SEQ ID NO: 392 can anneal to the 5' and 3' primer binding regions of each SDSI, respectively.

32. (New) A method for detecting contamination in amplicon sequencing comprising:

(a) adding a synthetic DNA spike-in (SDSI) to one or more cDNA samples,

wherein each SDSI is capable of amplification simultaneously with a target sequence present in the one or more cDNA samples,

wherein each SDSI comprises a unique core nucleotide sequence capable of differentiating each SDSI, covalently linked to an upstream 5' primer binding region and a downstream 3' primer binding region,

wherein the nucleotide sequence of the 5' and 3' primer binding regions is the same in each SDSI, and

wherein the SDSI's unique core sequence is chosen from the nucleotide sequence of SEQ ID NOs: 1-96 and 193-291,

(b) amplifying the one or more target sequence(s) present in the one or more cDNA samples together with the unique core sequence present in each added SDSI;

(c) sequencing the amplified target and SDSI sequence(s), and

(d) determining the presence of the amplified SDSI sequence(s) in the one or more cDNA samples,

wherein detection of a single SDSI unique core sequence in each cDNA sample indicates contamination-free amplification of each cDNA sample, and

wherein detection of more than one SDSI unique core sequence in a cDNA sample indicates possible contamination of that cDNA sample with cDNA(s) from a different cDNA sample.

REMARKS

Status of the Claims

Claims 1-6, 8-11, 14-15, 18-25, and 27-31 are pending in the application, with claims 1 and 21 being independent. Claims 1–6, 8–11, 15, 18–23, 25, and 27–31 are amended solely to improve clarity. Claims 7, 12–13, 16–17, and 26 are canceled without prejudice or disclaimer as to the subject matter recited therein. Claims 21–25 and 27–31 are withdrawn.

Written description for the claim amendments can be found throughout the specification and claims as originally filed. For example, written description of the term "amplicon sequencing" of claim 1 can be found in the Abstract and Title.

No new matter has been added.

Objections to the Specification

In the Office Action, the specification was objected to for the following reasons:

- several sequences in the tables beginning on page 20 and page 93 of the marked-up copy of the substitute specification are not presented in a single box and instead extend over two boxes.
- contains an embedded hyperlink and/or other form of browser-executable code.
- it fails to provide support for all of the subject matter of claim 9. In particular, the substitute specification fails to disclose substituting SEQ ID NO: 169 with SEQ ID NO: 390. Instead, based on para. 12 of the marked-up copy of the substitute specification, it appears that "SEQ ID NO: 162" was intended for "SEQ ID NO: 169" in the claim.

The specification has been amended to address all the objections above. Withdrawal of the objections is respectfully requested.

Applicants have also amended the specification to correct minor clerical or typographical errors that resulted from an error(s) made in good faith by the Applicants. The requested corrections do not introduce new matter nor require substantive examination.

Applicants submit a "Clean Substitute specification" (with all "strikethroughs" and "underlines" removed) and a Marked-Up Copy of the Substitute Specification (with items to be removed marked with a "strikethrough" and items to be inserted marked with an "underline").

The Substitute Specification contains no new matter. Applicants respectfully request that the substitute specification be entered.

Objections to the Drawings

In the Office Action, the Examiner objects to the replacement drawings because they are not labeled "Replacement Sheet." Applicants resubmit the replacement drawings, correctly labeled, to address this issue. Accordingly, withdrawal of the objection is respectfully requested.

Objections to the Claims

In the Office Action, claims 1, 5-6, and 8-11 were objected to for formality issues. Applicant has amended the claims rendering the objection moot. Withdrawal of the objection is respectfully requested.

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

In the Office Action, the Examiner rejected claims 1-6, 8-11, 14-15, and 18-20 under 35 U.S.C. § 112(b) as allegedly lacking clarity.

With respect to claim 11, an analysis of the sequences in the Table on pages 39-47 of the specification as originally filed shows the forward (SEQ ID NO: 391) and reverse primers (SEQ ID NO: 392) can anneal to the 5' and 3' primer binding regions of each SDSI.

For example, the sequence below shows the forward primer (SEQ ID NO: 391) and reverse primer (SEQ ID NO: 392) annealing to the SDSI of SEQ ID NO: 390.

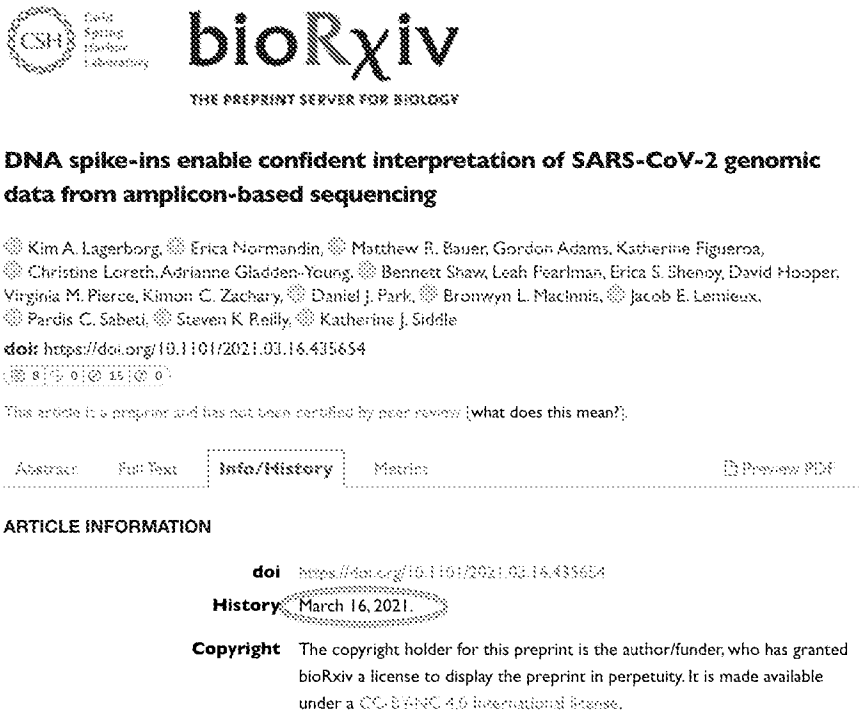
The Applicant submits the rejection under 35 U.S.C. § 112(b) does not apply to the claims as presently presented. Withdrawal of the rejection is respectfully requested.

Claim Rejections Under 35 U.S.C. § 102

1) **In the Office Action, the Examiner rejected claims 1-6, 8-11, 14-15, 18, and 19 under 35 U.S.C. § 102(a)(1) as allegedly anticipated by Lagerborg et al. (bioRxiv [preprint] 2021 Mar 16; doi:10.1101/2021.03.16.435654 ("Lagerborg").**

Applicant respectfully traverses that rejection.

The instant application claims priority to Provisional Patent Applications 63/155,258, filed on **March 1, 2021**, and 63/273,117, filed on October 28, 2021. The bioRxiv preprint cited by the Examiner was made publicly available on **March 16, 2021** (see below).



bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

DNA spike-ins enable confident interpretation of SARS-CoV-2 genomic data from amplicon-based sequencing

Kim A. Lagerborg, Erica Normandin, Matthew R. Bauer, Gordon Adams, Katherine Figueroa, Christine Loreth, Adrianne Gladden-Young, Bennett Shaw, Leah Fearlman, Erica S. Shenoy, David Hooper, Virginia M. Pierce, Kimon C. Zachary, Daniel J. Park, Bronwyn L. MacInnis, Jacob E. Lemieux, Pardis C. Sabeti, Steven K. Reilly, Katherine J. Siddle

doi: https://doi.org/10.1101/2021.03.16.435654

This article is a preprint and has not been certified by peer review (what does this mean?).

Abstract | Full Text | Info/History | Metrics | Preview PDF

ARTICLE INFORMATION

doi https://doi.org/10.1101/2021.03.16.435654

History March 16, 2021.

Copyright The copyright holder for this preprint is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

The instant application's earliest effective filing date, therefore, **predates** the bioRxiv publication date. Consequently, the bioRxiv preprint is not anticipatory prior art under 35 U.S.C. §102(a)(1) with respect to the instant application.

Withdrawal of the rejection is respectfully requested.

2) In the Office Action, the Examiner rejected claims 1, 15, and 18-20 under 35 U.S.C. § 102(a)(1) as allegedly anticipated by US 2019/0249334 A1 to Piehl et al. ("Piehl").

According to the Examiner, Piehl discloses a method that comprises using a spike-in control to detect cross-contamination by (a) adding a spike-in control nucleic acid to a sample containing target nucleic acids, (b) conducting an amplification reaction to generate amplicons from the target nucleic acids and the spike-in control nucleic acid, and (c) sequencing the spike-in control nucleic acids, wherein the presence of a spike-in control nucleic acid other than the spike-in control nucleic acid added in step (a) indicates sample contamination. The Examiner alleges Piehl anticipates the method of claims 1 and 20.

Applicant respectfully traverses that rejection.

In the Summary (paragraph [0004] of the published U.S. Patent Application No. 20190249334), the specification states *inter alia*:

> Provided herein are methods for determining whether a sample has been contaminated, comprising forming a reaction composition in a partition comprising ***at least one spike in control, at least one adapter, and a sample***; ***generating sample fragments***; li***gating the at least one adapter to the sample fragments to generate ligation products***; ***amplifying the reaction composition*** to generate a multiplicity of amplification products comprising a multiplicity of ***ligation product amplicons*** and a multiplicity of ***spike in control amplicons***; and ***sequencing the spike in control amplicons***, wherein the presence of an amplified spike in control not associated with the reaction composition indicates the sample has been contaminated. (Emphasis added)

Claim 1 of the instant application is amended to require:

1. A method for detecting contamination in amplicon sequencing, said method ***consists of***:

(a) ***adding a synthetic DNA spike-in (SDSI) to one or more cDNA samples***,

wherein each SDSI is capable of amplification simultaneously with a target sequence present in the one or more cDNA samples,

wherein each SDSI comprises a unique core nucleotide sequence that differentiates each SDSI, covalently linked to an upstream 5' primer binding region and a downstream 3' primer binding region.

wherein the nucleotide sequence of the 5' and 3' primer binding regions is the same in each SDSI;

(b) ***amplifying the one or more target sequence(s)*** present in the one or more cDNA samples ***together with the unique core sequence present in each added SDSI;***

(c) ***sequencing*** the amplified target and SDSI ~~sequences~~ sequence(s), and
(d) ***determining*** the presence of the amplified SDSI sequence(s) in the one or more cDNA samples,
wherein detection of a single SDSI unique core sequence in each cDNA sample indicates contamination-free amplification of each cDNA sample, and
wherein detection of more than one SDSI unique core sequence in a cDNA sample indicates possible contamination of that cDNA sample with cDNA(s) from a different cDNA sample.
(Emphasis added)

The claimed invention does not require at least one adapter or a ligation step, as disclosed in Piehl. Accordingly, Piehl does not anticipate the claimed invention.

Withdrawal of the rejection is respectfully requested.

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

In the Office Action, the Examiner rejected claims 2-6, 10, and 14 as allegedly being obvious over Piehl.

The Applicant respectfully traverses the rejection.

Piehl teaches ligating at least one adapter to a spike in control and sample fragments, wherein the adapter comprises a primer binding site and/or a bar code (see Summary section in Piehl, paragraphs [0003]-[0006] of the published U.S. Patent Application No. 20190249334 and the claims). A person of ordinary skill in the art at the time of Applicant's filing would have had no motivation to adopt the teaching in Piehl because (1) Piehl fails to teach or suggest each and every one of the elements of the claimed invention, (2) the skilled artisan would not have been even motivated to consider Piehl relevant because the adapter ligation step is not needed. Applicant's SDSIs already contain 5' and 3' primer binding regions. Ligating adapters and multiple primer binding sites to the Applicant's sample fragments would, on the contrary, likely render the claimed invention inoperable.

Accordingly, the claimed invention cannot be *prima facie* obvious in view of Piehl.

Withdrawal of the rejection is respectfully requested. Because

No Waiver

All of Applicant's arguments and amendments are without prejudice or disclaimer. Applicant may not have addressed each specific rejection of the independent and dependent claims because Applicant submits that the independent claims are allowable, as discussed above. Applicant has not acquiesced to any such rejection and reserves the right to address the patentability of any additional claim features in the future.

CONCLUSION

Applicant submits the foregoing as a full and 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 exist that can be resolved with an Examiner's Amendment or a telephone conference, please contact Applicant's representative at 478.292.5119.

In the Non-Final Office Action dated December 4, 2024, a shortened statutory period for reply was set to expire three months from the mailing date of the Office Action, i.e., March 4, 2025. Accordingly, the Applicant is submitting a petition for a one-month extension of time, which will extend the period for reply from March 18, 2025, to April 4, 2025.

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-5360US.

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

Respectfully submitted,

/Christopher Southgate/

Attorney for Applicant

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

PTO/AIA/22 (01-25)
Approved for use through 11/30/2027. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

PETITION FOR EXTENSION OF TIME UNDER 37 CFR 1.136(a) (NOT for Provisional Applications)

Docket Number (Optional): BROD-5360US

Application Number	Filed
17/684,328	March 1, 2022

For: AMPLICON-BASED SEQUENCING USING DNA SPIKE-INS

Art Unit	Examiner
1681	Angela Marie Bertagna

This is a request under the provisions of 37 CFR 1.136(a) to extend the period for filing a reply in the above-identified application.

The requested extension and fee are as follows (check time period desired and enter the appropriate fee below):

	Fee	Small Entity Fee	Micro Entity Fee	
☑ One month (37 CFR 1.17(a)(1))	$235	$94	$47	$ 94
☐ Two months (37 CFR 1.17(a)(2))	$690	$276	$138	$
☐ Three months (37 CFR 1.17(a)(3))	$1,590	$636	$318	$
☐ Four months (37 CFR 1.17(a)(4))	$2,495	$998	$499	$
☐ Five months (37 CFR 1.17(a)(5))	$3,395	$1,358	$679	$

☐ Applicant asserts small entity status. See 37 CFR 1.27.

☐ Applicant certifies micro entity status. See 37 CFR 1.29.
Form PTO/SB/15A or B or equivalent must either be enclosed or have been submitted previously.

☐ A check in the amount of the fee is enclosed.

☐ Payment by credit card. Form PTO-2038 is attached.

☐ The Director has already been authorized to charge fees in this application to a Deposit Account.

☐ The Director is hereby authorized to charge any fees which may be required, or credit any overpayment, to Deposit Account Number _____.

☑ Payment made via USPTO patent electronic filing system.

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

I am the

☐ applicant.

☑ attorney or agent of record. Registration number 79,933.

☐ attorney or agent acting under 37 CFR 1.34. Registration number _____.

/Drew P. Harding/ — Signature

April 4, 2025 — Date

Drew P. Harding — Typed or printed name

(912) 257-4864 — Telephone Number

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. Submit multiple forms if more than one signature is required, see below*.

☐ * Total of _____ forms are submitted.

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. The United States Patent and Trademark Office (USPTO) collects the information in this record under authority of 35 U.S.C. 2. The USPTO's system of records is used to manage all applicant and owner information including name, citizenship, residence, post office address, and other information with respect to inventors and their legal representatives pertaining to the applicant's/owner's activities in connection with the invention for which a patent is sought or has been granted. The applicable Privacy Act System of Records Notice for the information collected in this form is COMMERCE/PAT-TM-7 Patent Application Files, available in the Federal Register at 78 FR 19243 (March 29, 2013). https://www.govinfo.gov/content/pkg/FR-2013-03-29/pdf/2013-07341.pdf

Routine uses of the information in this record may include disclosure to:

1) law enforcement, in the event that the system of records indicates a violation or potential violation of law;

2) a federal, state, local, or international agency, in response to its request;

3) a contractor of the USPTO having need for the information in order to perform a contract;

4) the Department of Justice for determination of whether the Freedom of Information Act (FOIA) requires disclosure of the record;

5) a Member of Congress submitting a request involving an individual to whom the record pertains, when the individual has requested the Member's assistance with respect to the subject matter of the record;

6) a court, magistrate, or administrative tribunal, in the course of presenting evidence, including disclosures to opposing counsel in the course of settlement negotiations;

7) the Administrator, General Services Administration (GSA), or their designee, during an inspection of records conducted by GSA under authority of 44 U.S.C. 2904 and 2906, in accordance with the GSA regulations and any other relevant (i.e., GSA or Commerce) directive, where such disclosure shall not be used to make determinations about individuals;

8) another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c));

9) the Office of Personnel Management (OPM) for personnel research purposes; and

10) the Office of Management and Budget (OMB) for legislative coordination and clearance.

If you do not furnish the information requested on this form, the USPTO may not be able to process and/or examine your submission, which may result in termination of proceedings, abandonment of the application, and/or expiration of the patent.

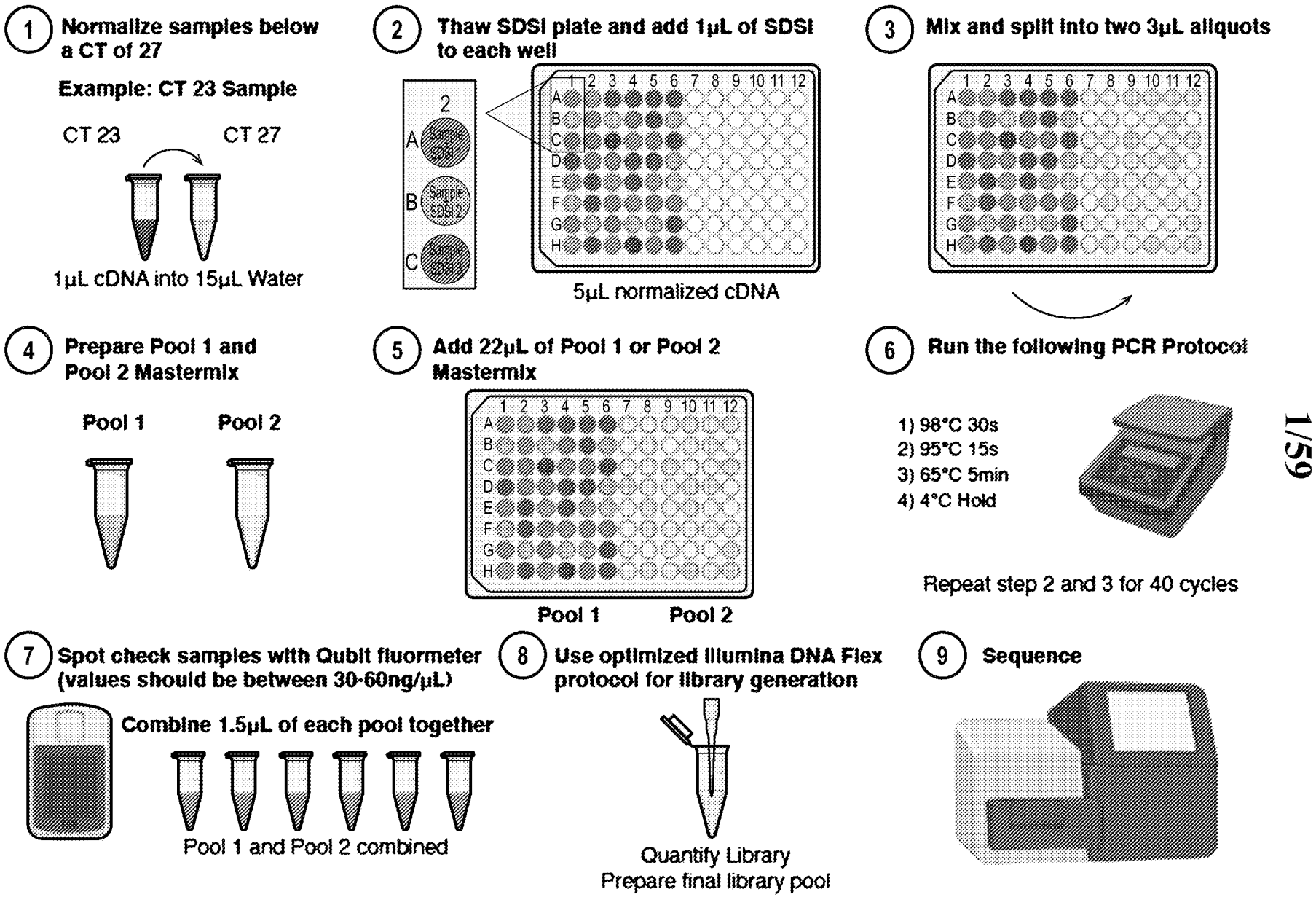
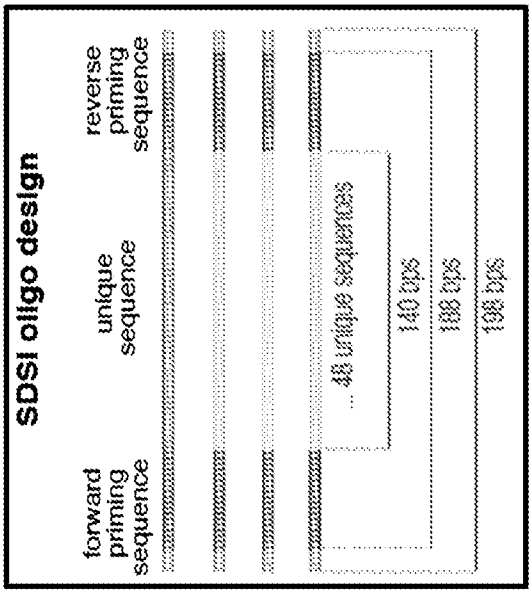

1 Normalize samples below a CT of 27
Example: CT 23 Sample
CT 23
CT 27
1µL cDNA into 15µL Water
2 Thaw SDSI plate and add 1µL of SDSI to each well
5µL normalized cDNA
3 Mix and split into two 3µL aliquots
4 Prepare Pool 1 and Pool 2 Mastermix
Pool 1
Pool 2
5 Add 22µL of Pool 1 or Pool 2 Mastermix
Pool 1
Pool 2
6 Run the following PCR Protocol
1) 98°C 30s
2) 95°C 15s
3) 65°C 5min
4) 4°C Hold
Repeat step 2 and 3 for 40 cycles
7 Spot check samples with Qubit fluormeter (values should be between 30-60ng/µL)
Combine 1.5µL of each pool together
Pool 1 and Pool 2 combined
8 Use optimized Illumina DNA Flex protocol for library generation
Quantify Library
Prepare final library pool
9 Sequence

FIG. 1


SDSI oligo design
forward priming sequence
unique sequence
reverse priming sequence
... 48 unique sequences
140 bps
188 bps
198 bps

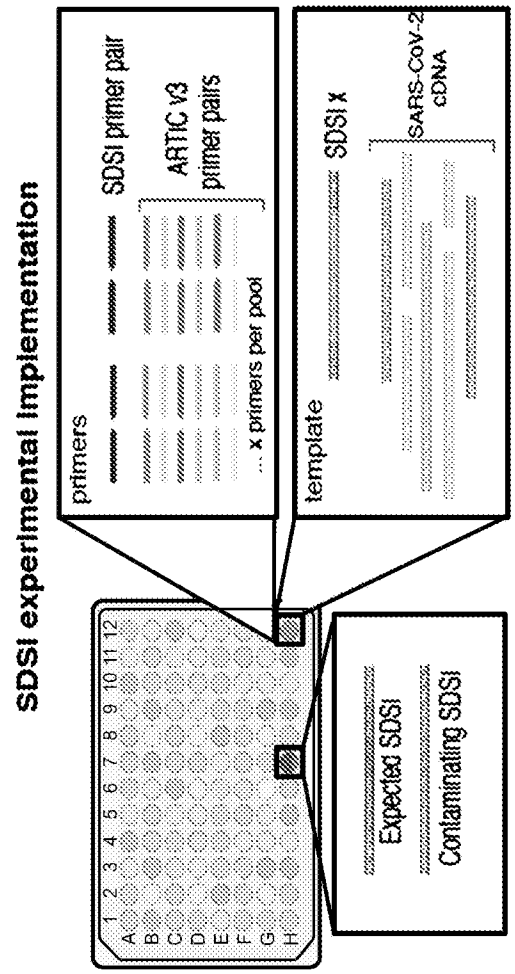

SDSI experimental implementation
primers
SDSI primer pair
ARTIC v3 primer pairs
... x primers per pool
template
SDSI x
SARS-CoV-2 cDNA
Expected SDSI
Contaminating SDSI
1 2 3 4 5 6 7 8 9 10 11 12
A B C D E F G H

FIG. 2A

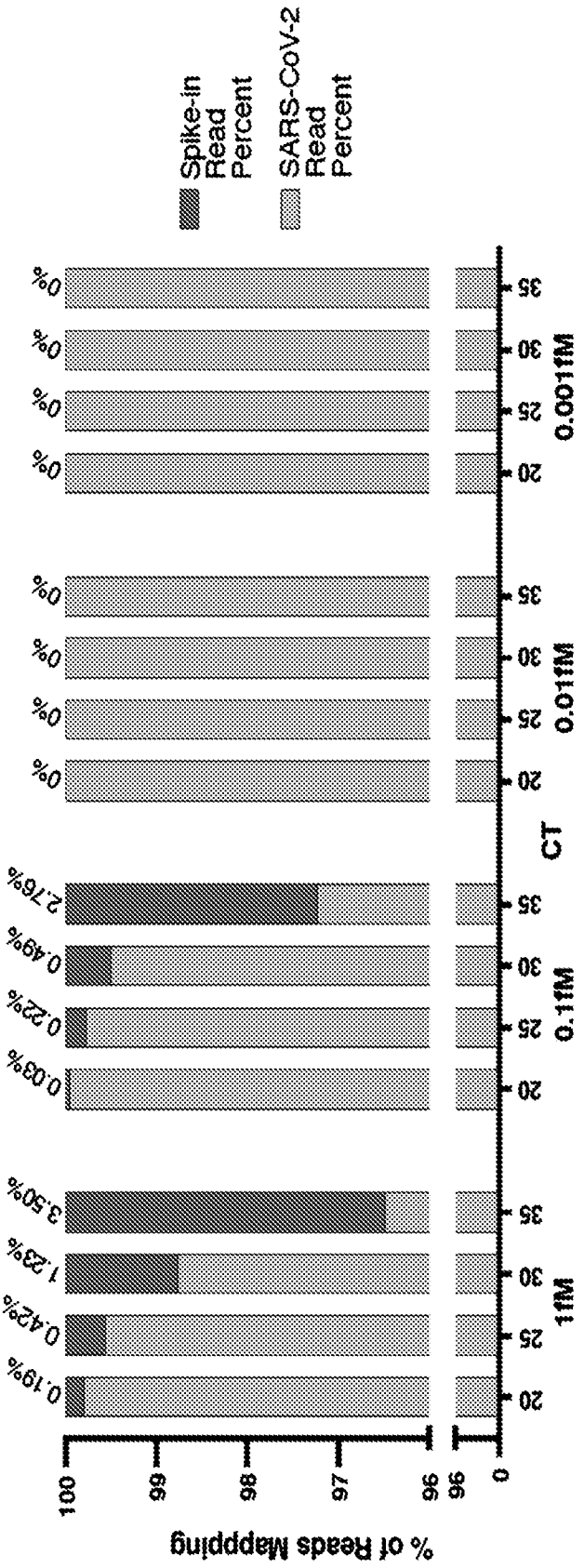

Spike-in Read Percent
SARS-CoV-2 Read Percent
% of Reads Mapping
100
99
98
97
96
0
0.19%
0.42%
1.23%
3.50%
0.03%
0.22%
0.49%
2.76%
0%
CT
1fM
0.1fM
0.01fM
0.001fM
20
25
30
35

FIG. 2B

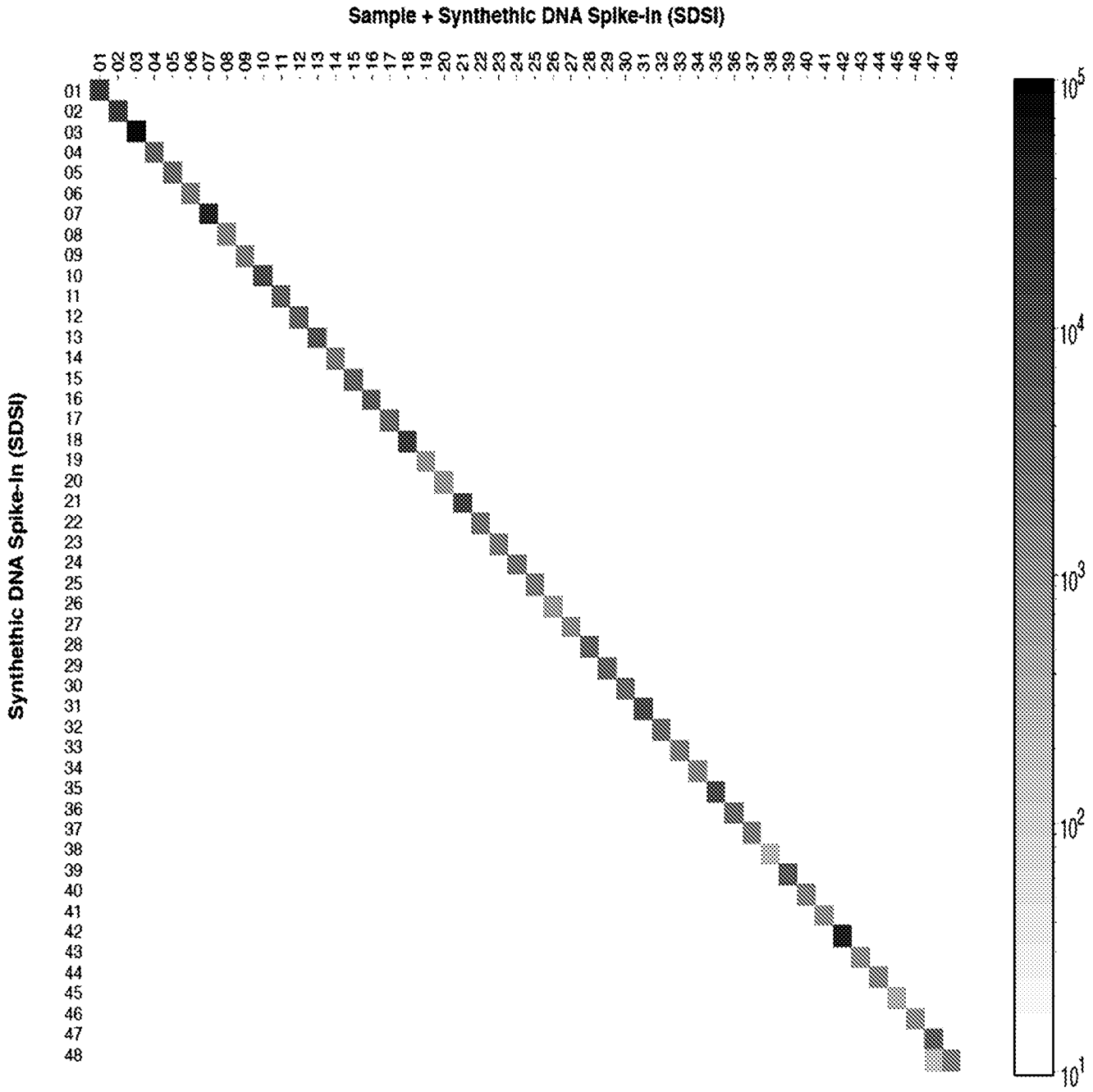

Sample + Synthethic DNA Spike-In (SDSI)
Synthethic DNA Spike-In (SDSI)
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
10^5
10^4
10^3
10^2
10^1

FIG. 2C

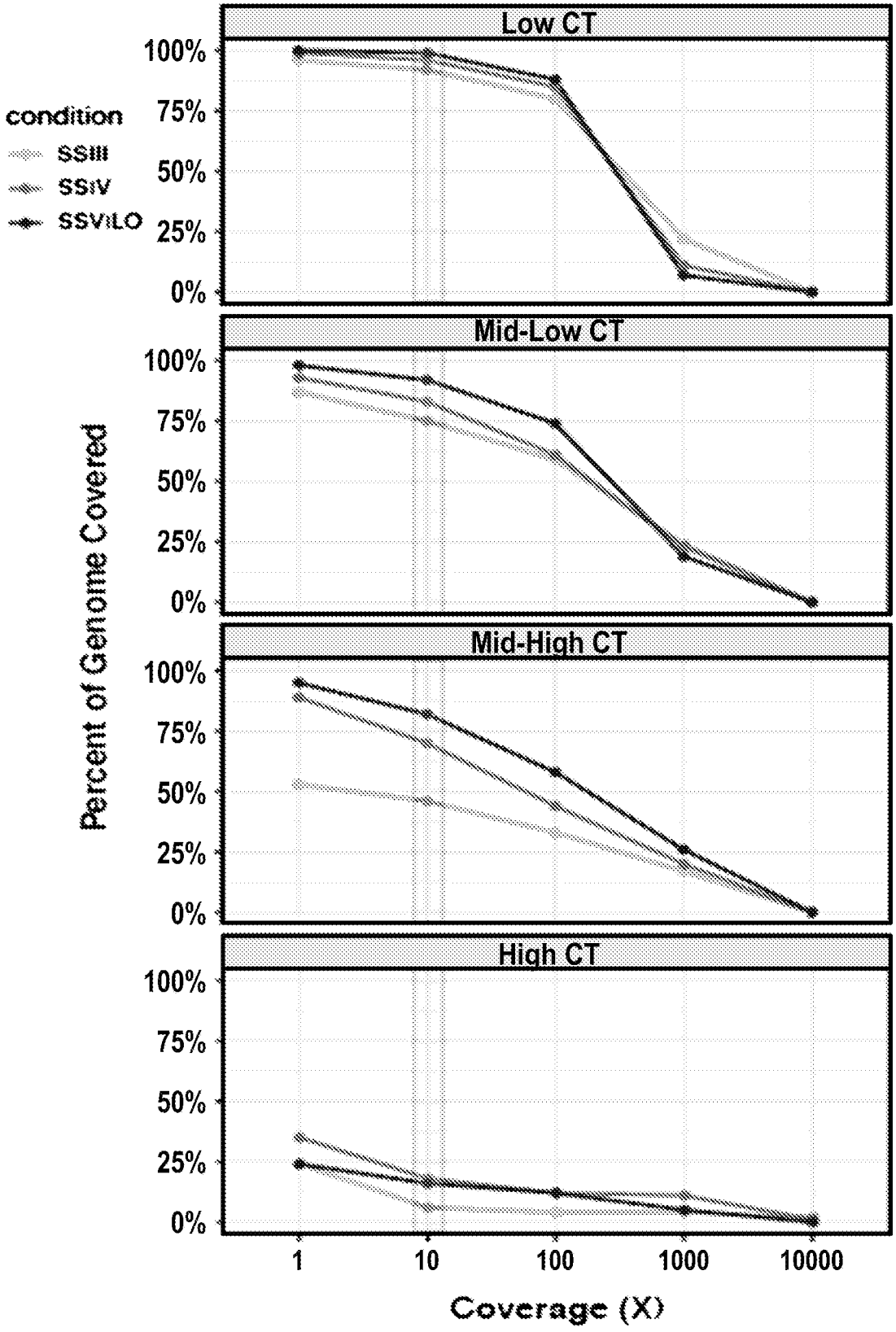

Low CT
Mid-Low CT
Mid-High CT
High CT
condition
SSIII
SSIV
SSVILO
100%
75%
50%
25%
0%
Percent of Genome Covered
1
10
100
1000
10000
Coverage (X)

FIG. 3A

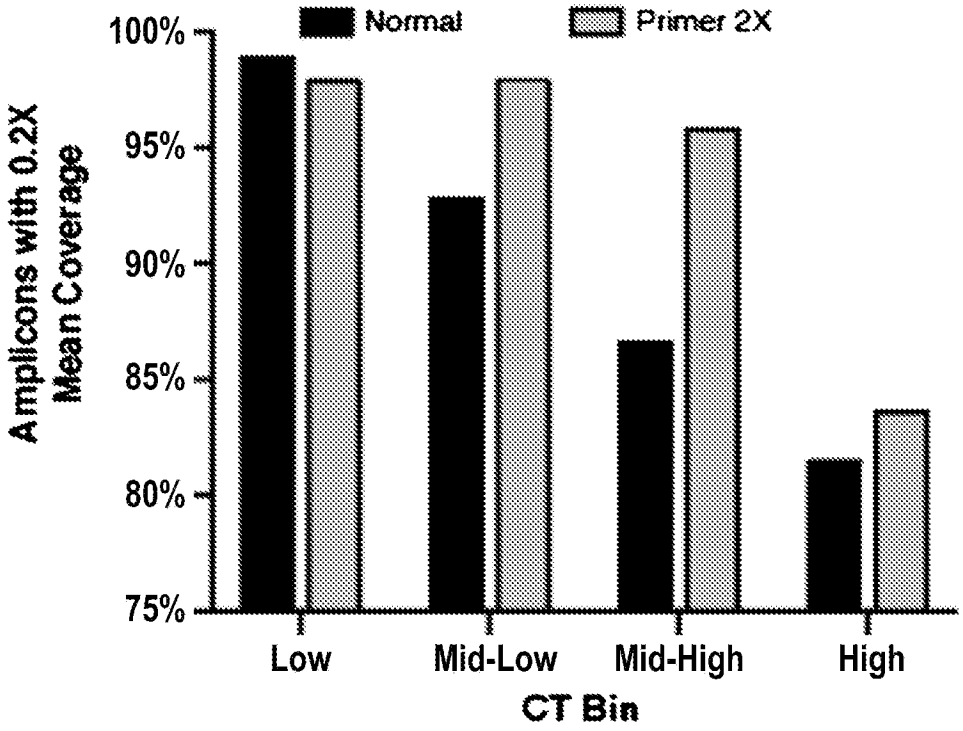

Normal
Primer 2X
100%
95%
90%
85%
80%
75%
Amplicons with 0.2X Mean Coverage
Low
Mid-Low
Mid-High
High
CT Bin

FIG. 3B

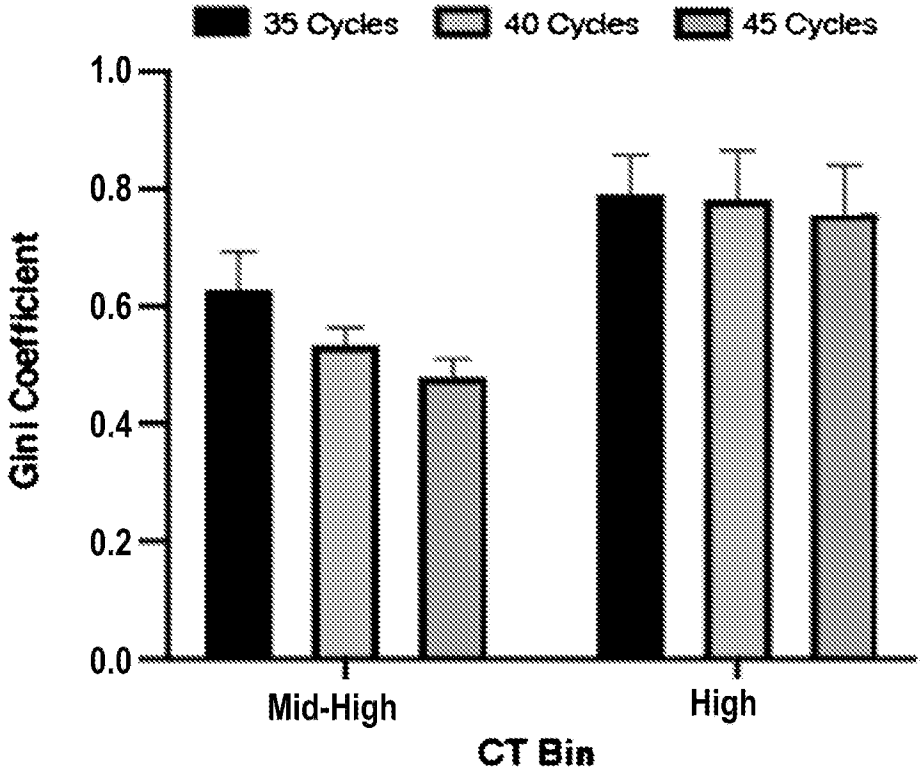

35 Cycles
40 Cycles
45 Cycles
1.0
0.8
0.6
0.4
0.2
0.0
Gini Coefficient
Mid-High
High
CT Bin

FIG. 3C

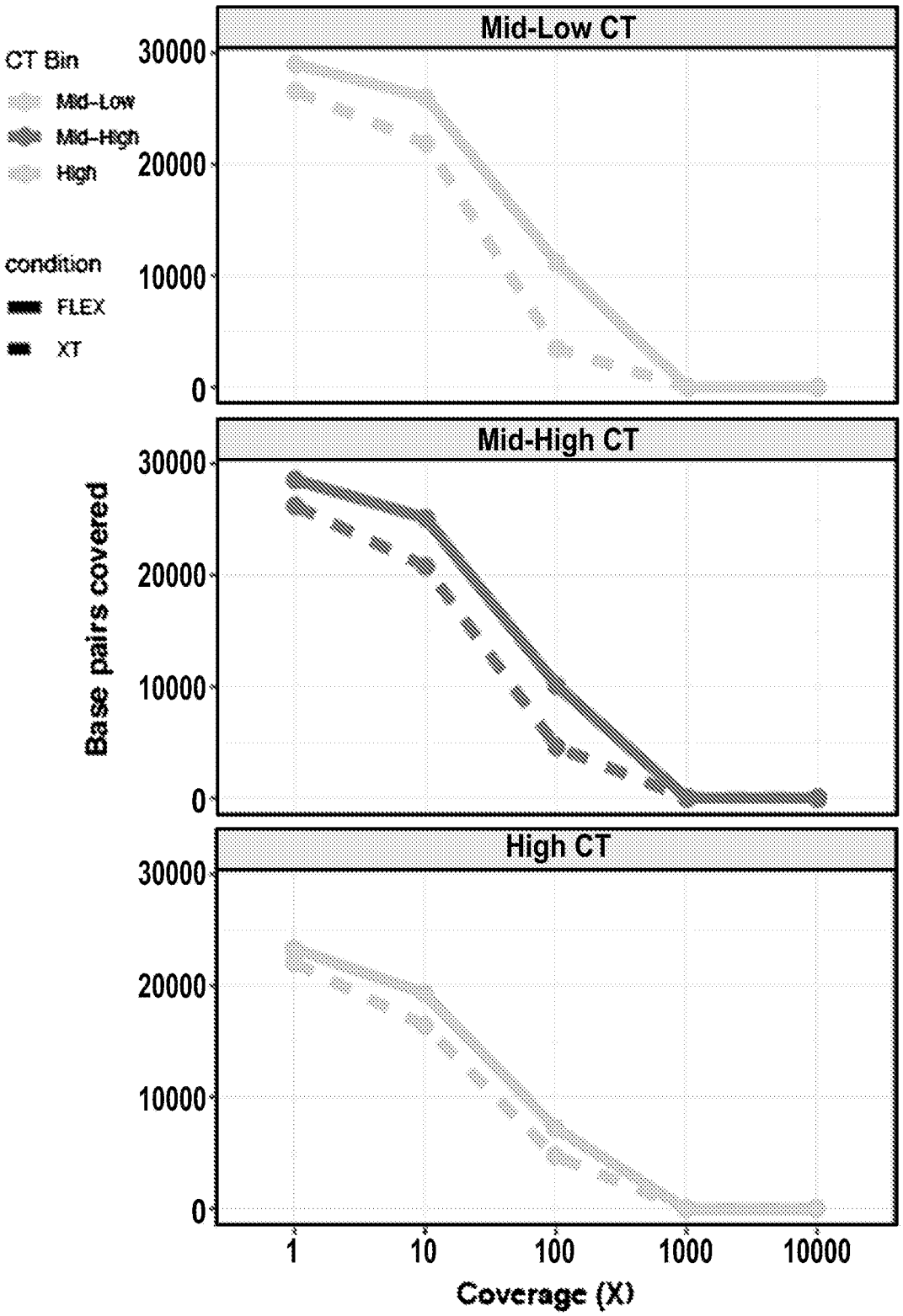

CT Bin
Mid-Low
Mid-High
High
condition
FLEX
XT
Mid-Low CT
Mid-High CT
High CT
Base pairs covered
30000
20000
10000
0
1
10
100
1000
10000
Coverage (X)

FIG. 3D

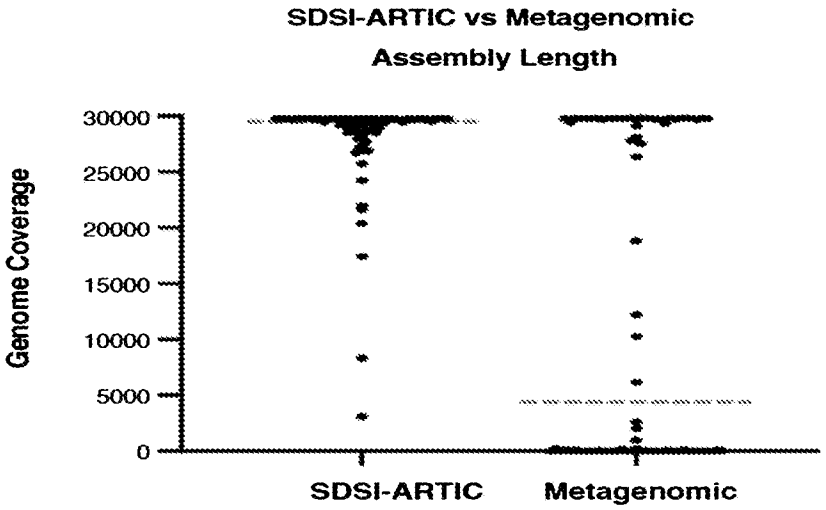

SDSI-ARTIC vs Metagenomic
Assembly Length
Genome Coverage
30000
25000
20000
15000
10000
5000
0
SDSI-ARTIC
Metagenomic

FIG. 4A

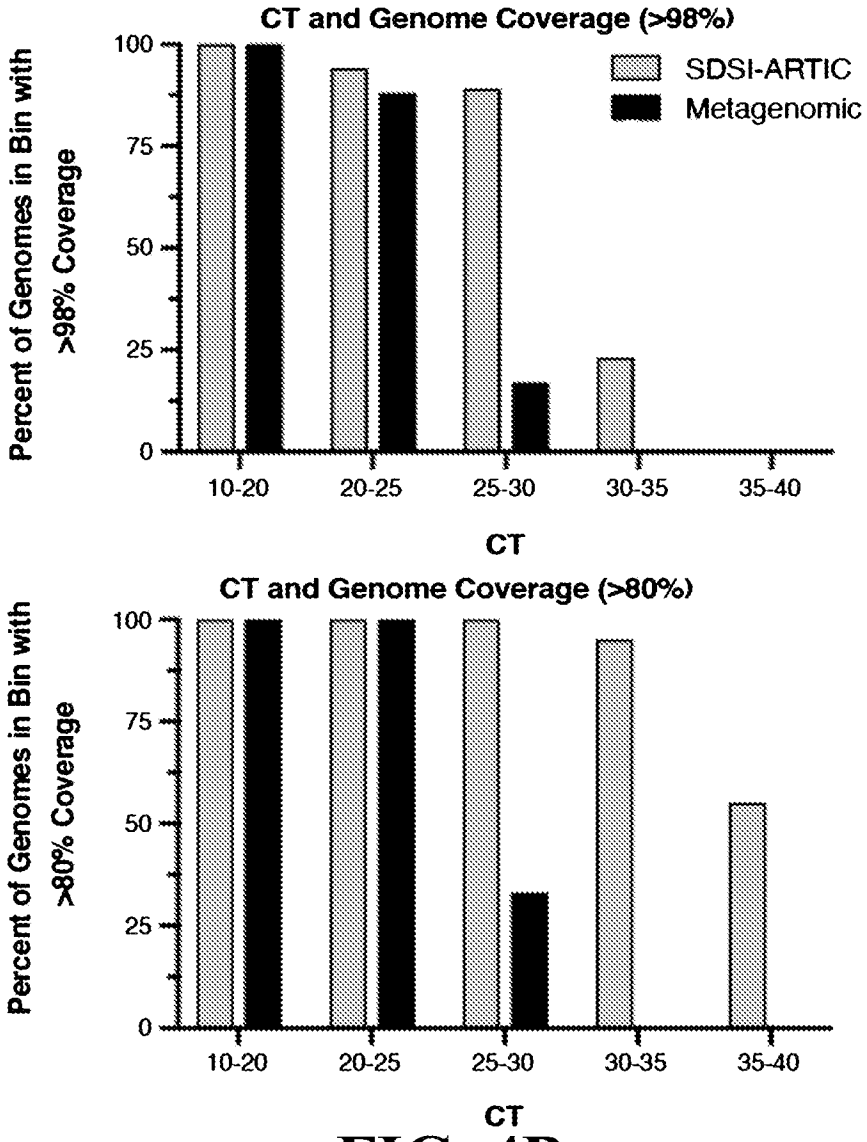

CT and Genome Coverage (>98%)
Percent of Genomes in Bin with >98% Coverage
SDSI-ARTIC
Metagenomic
100
75
50
25
0
10-20
20-25
25-30
30-35
35-40
CT
CT and Genome Coverage (>80%)
Percent of Genomes in Bin with >80% Coverage
100
75
50
25
0
10-20
20-25
25-30
30-35
35-40
CT

FIG. 4B

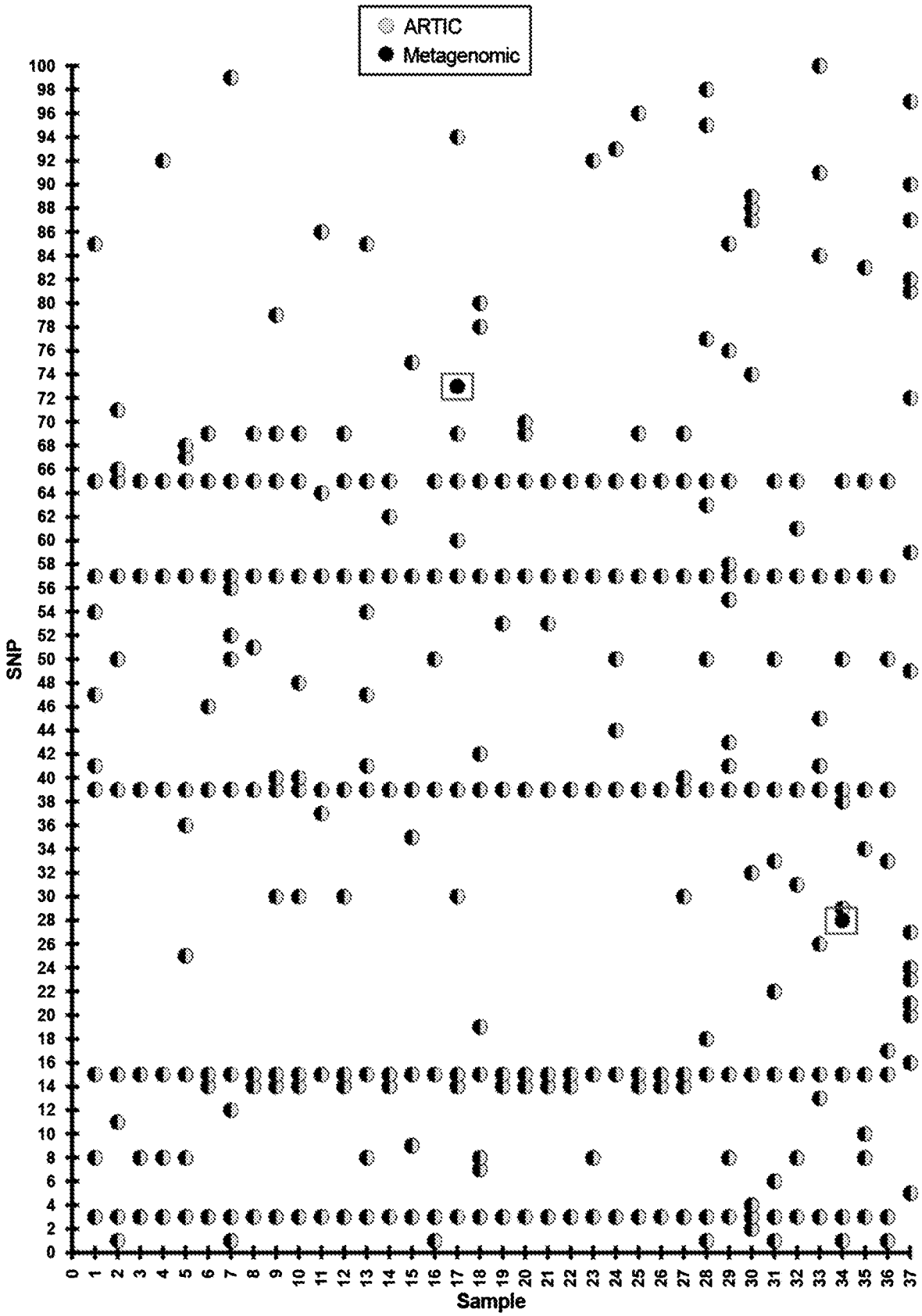


FIG. 4C

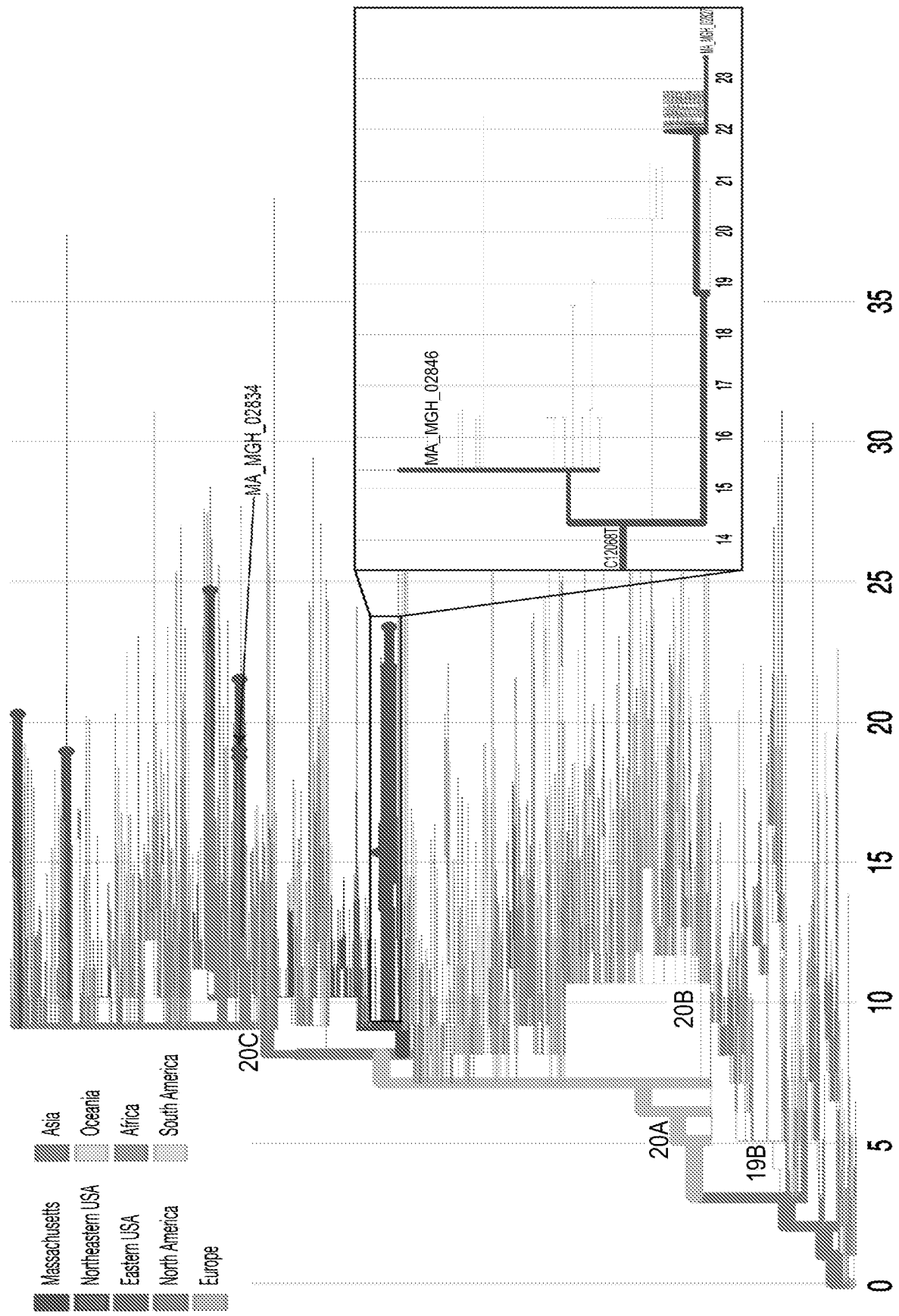

Massachusetts
Northeastern USA
Eastern USA
North America
Europe
Asia
Oceania
Africa
South America
20C
20B
20A
19B
MA_MGH_02834
MA_MGH_02846
C12068T
0
5
10
15
20
25
30
35
Mutations

FIG. 5A

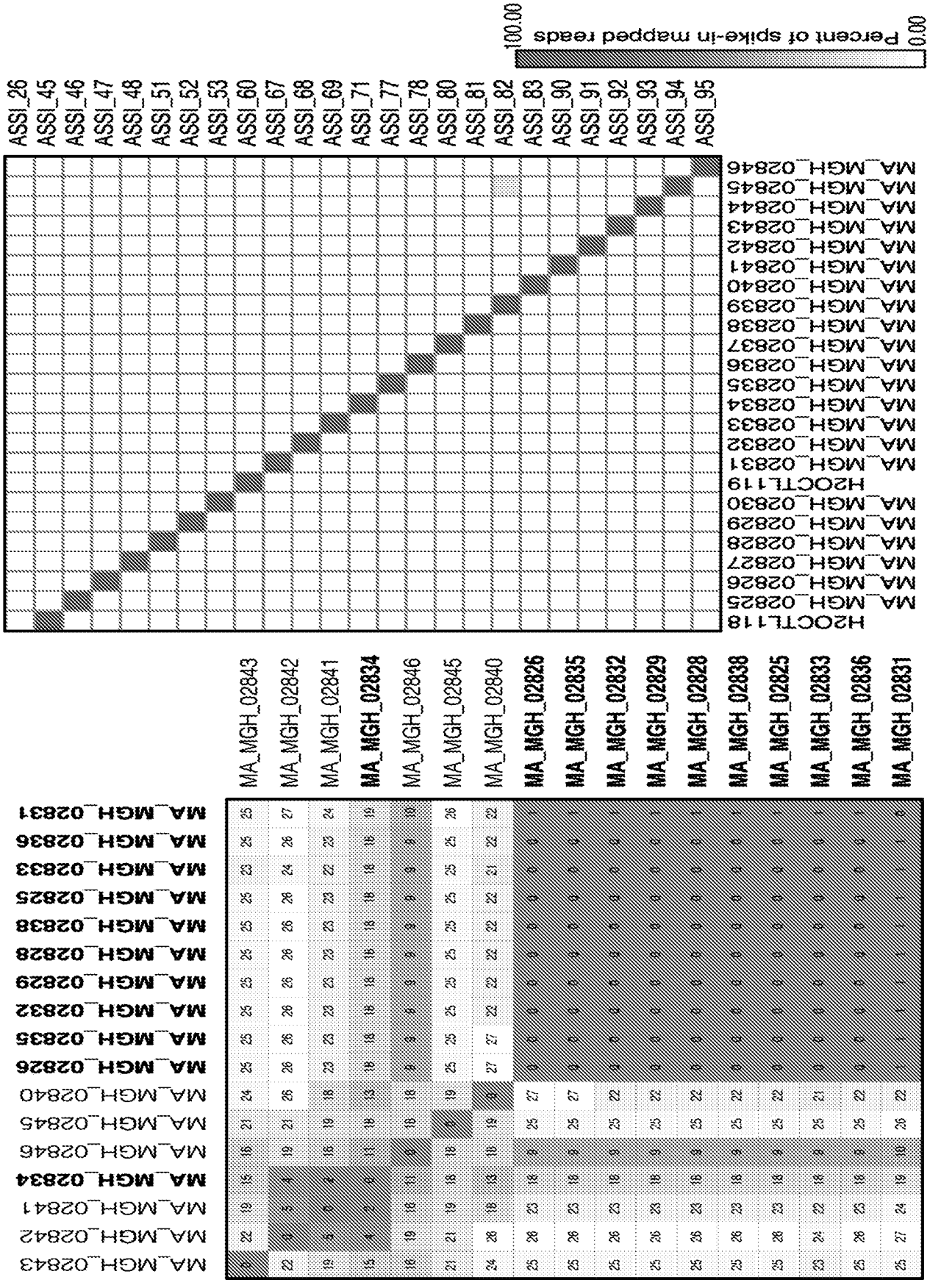

Percent of spike-in mapped reads
100.00
0.00
ASSI_26
ASSI_45
ASSI_46
ASSI_47
ASSI_48
ASSI_51
ASSI_52
ASSI_53
ASSI_60
ASSI_67
ASSI_68
ASSI_69
ASSI_71
ASSI_77
ASSI_78
ASSI_80
ASSI_81
ASSI_82
ASSI_83
ASSI_90
ASSI_91
ASSI_92
ASSI_93
ASSI_94
ASSI_95
H2OCTL118
MA_MGH_02825
MA_MGH_02826
MA_MGH_02827
MA_MGH_02828
MA_MGH_02829
MA_MGH_02830
H2OCTL119
MA_MGH_02831
MA_MGH_02832
MA_MGH_02833
MA_MGH_02834
MA_MGH_02835
MA_MGH_02836
MA_MGH_02837
MA_MGH_02838
MA_MGH_02839
MA_MGH_02840
MA_MGH_02841
MA_MGH_02842
MA_MGH_02843
MA_MGH_02844
MA_MGH_02845
MA_MGH_02846

FIG. 5C

MA_MGH_02843
MA_MGH_02842
MA_MGH_02841
MA_MGH_02834
MA_MGH_02846
MA_MGH_02845
MA_MGH_02840
MA_MGH_02826
MA_MGH_02835
MA_MGH_02832
MA_MGH_02829
MA_MGH_02828
MA_MGH_02838
MA_MGH_02825
MA_MGH_02833
MA_MGH_02836
MA_MGH_02831
MA_MGH_02843
MA_MGH_02842
MA_MGH_02841
MA_MGH_02834
MA_MGH_02846
MA_MGH_02845
MA_MGH_02840
MA_MGH_02826
MA_MGH_02835
MA_MGH_02832
MA_MGH_02829
MA_MGH_02828
MA_MGH_02838
MA_MGH_02825
MA_MGH_02833
MA_MGH_02836
MA_MGH_02831

FIG. 5B

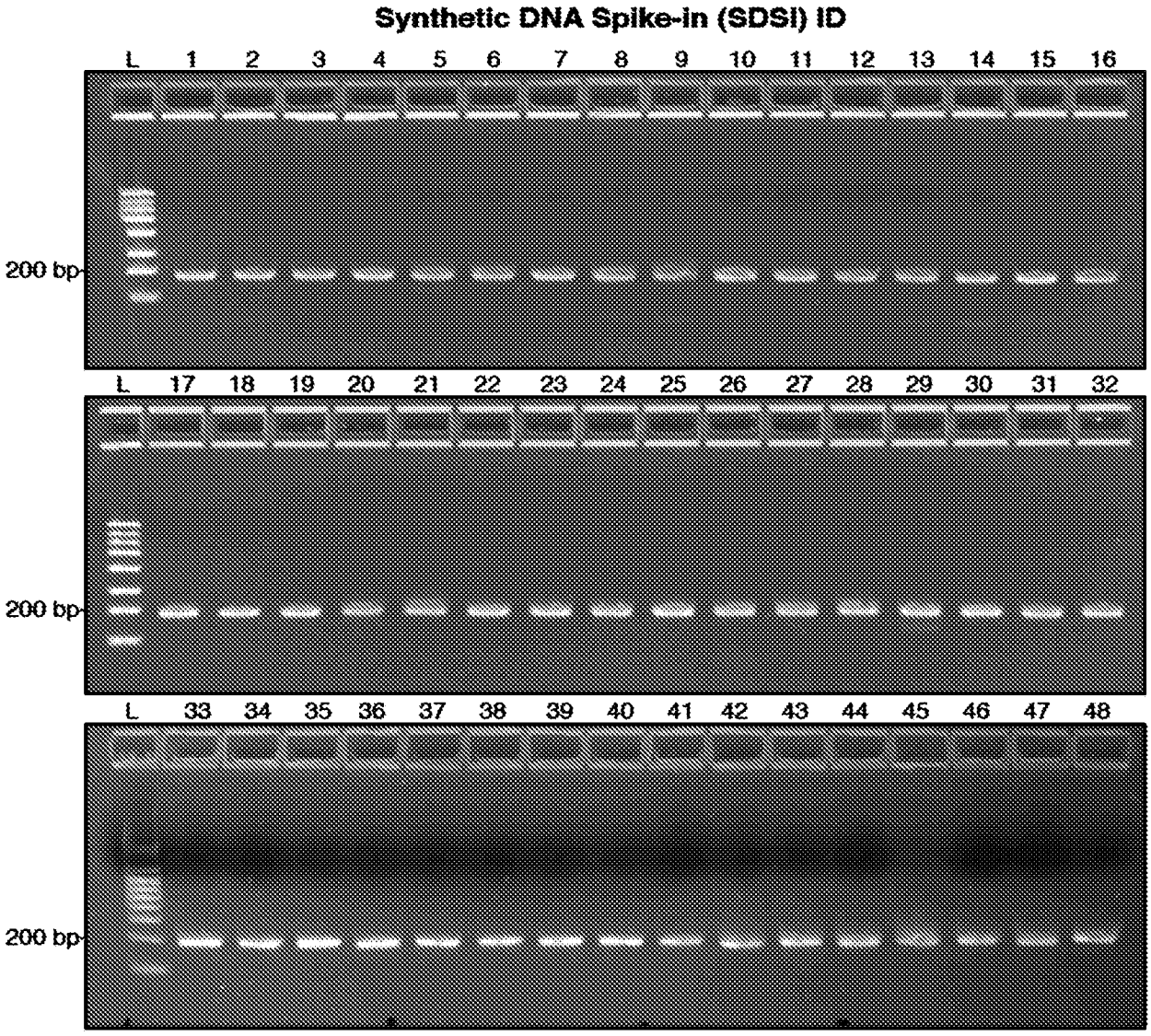

Synthetic DNA Spike-in (SDSI) ID
L 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
200 bp
L 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
200 bp
L 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
200 bp

FIG. 6A

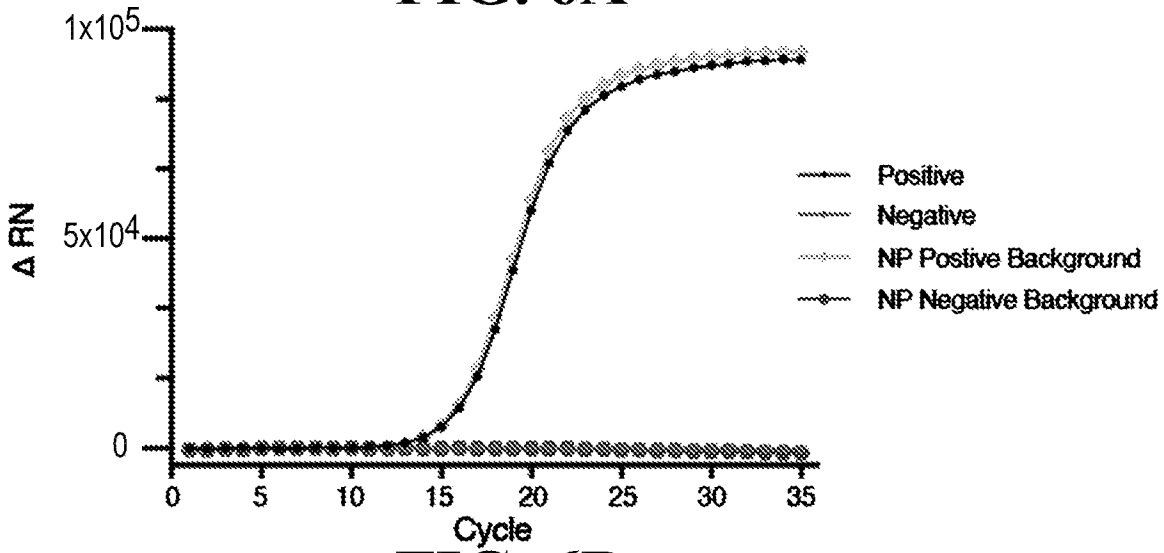

1x10^5
5x10^4
0
Δ RN
0 5 10 15 20 25 30 35
Cycle
Positive
Negative
NP Postive Background
NP Negative Background

FIG. 6B

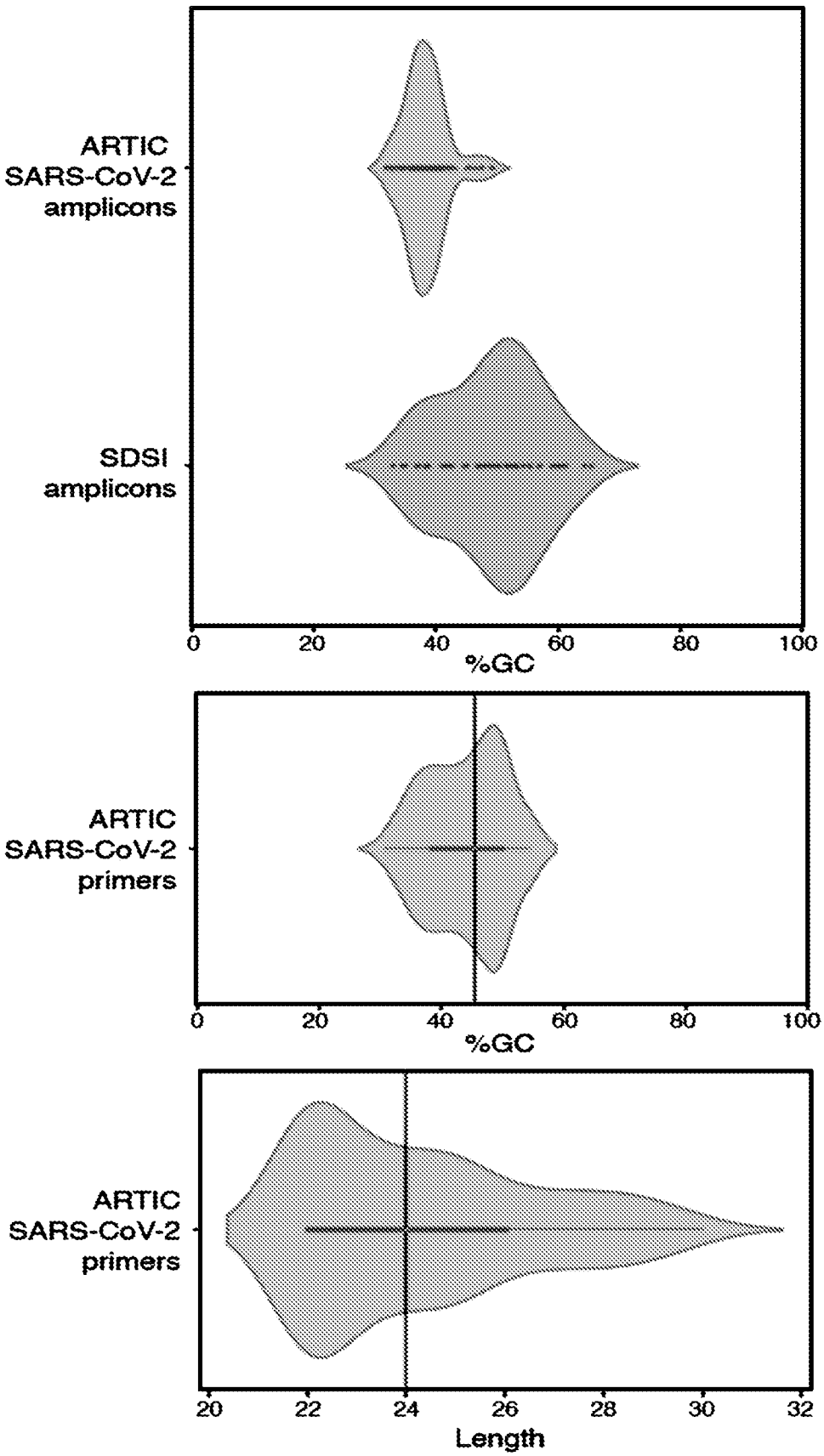

ARTIC SARS-CoV-2 amplicons
SDSI amplicons
0
20
40
60
80
100
%GC
ARTIC SARS-CoV-2 primers
0
20
40
60
80
100
%GC
ARTIC SARS-CoV-2 primers
20
22
24
26
28
30
32
Length

FIG. 6C

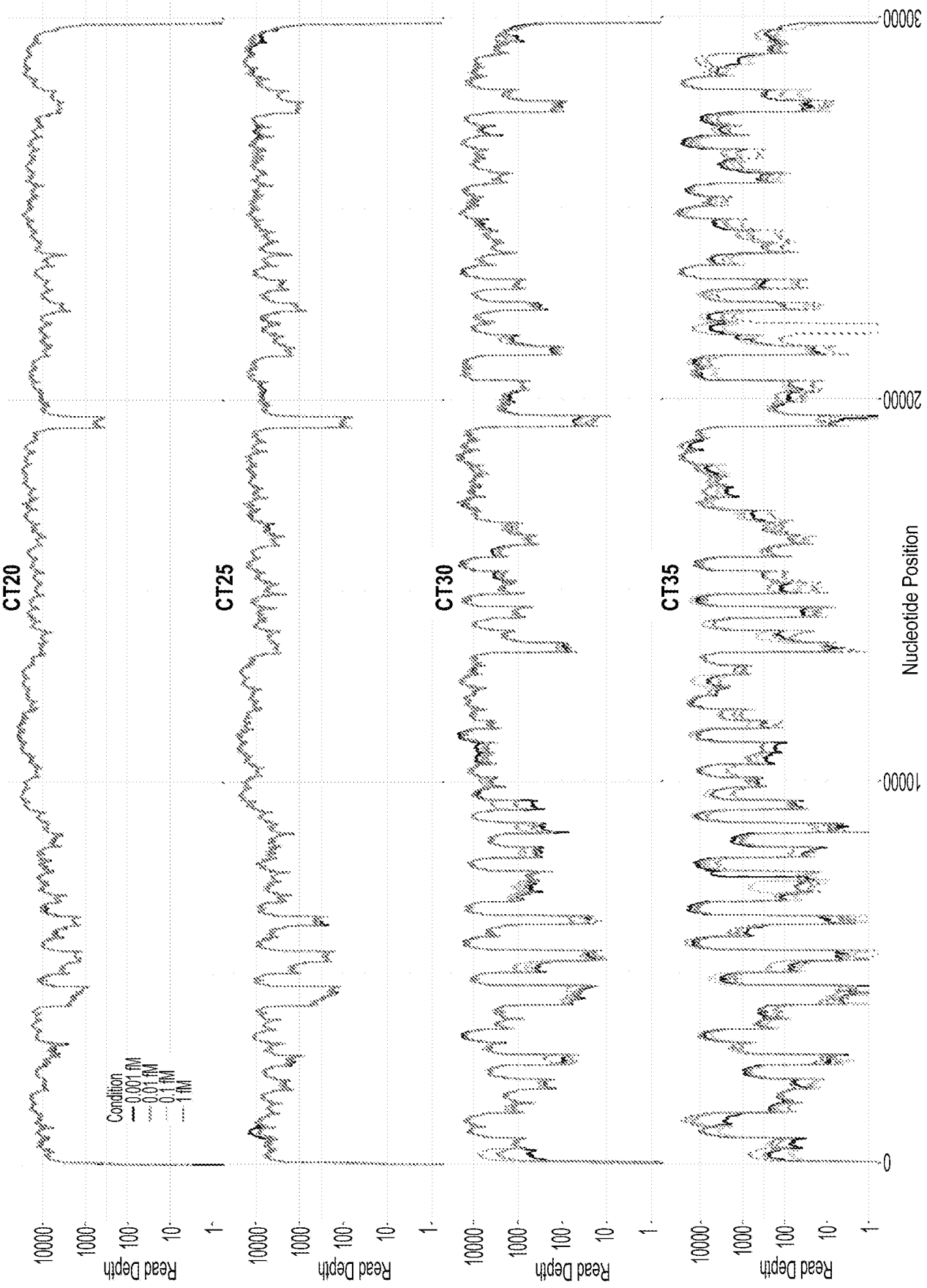

CT20
CT25
CT30
CT35
Read Depth
10000
1000
100
10
1
Condition
0.001 fM
0.01 fM
0.1 fM
1 fM
0
10000
20000
30000
Nucleotide Position

FIG. 7

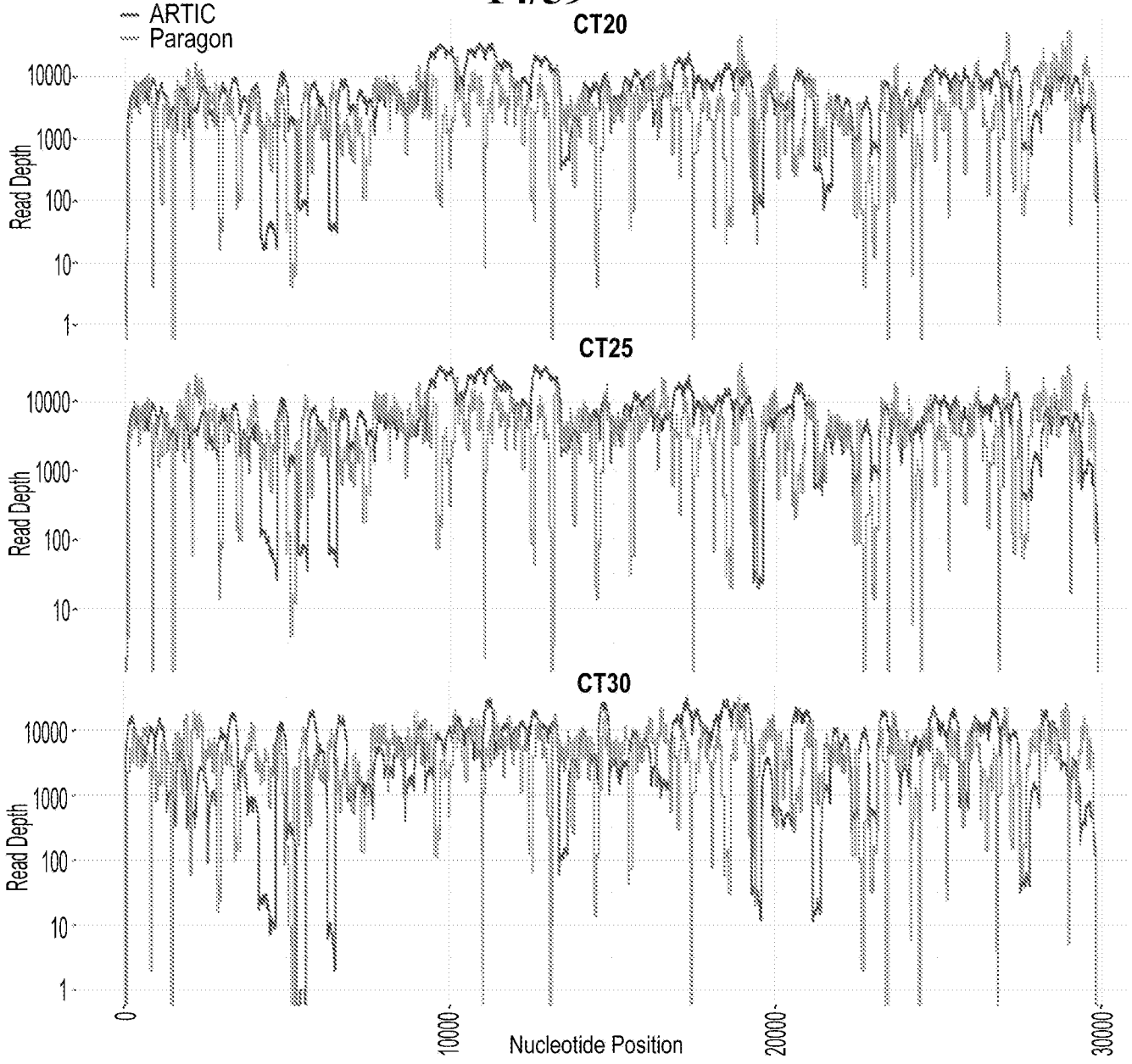

ARTIC
Paragon
CT20
CT25
CT30
Read Depth
10000
1000
100
10
1
0
10000
20000
30000
Nucleotide Position

FIG. 8A

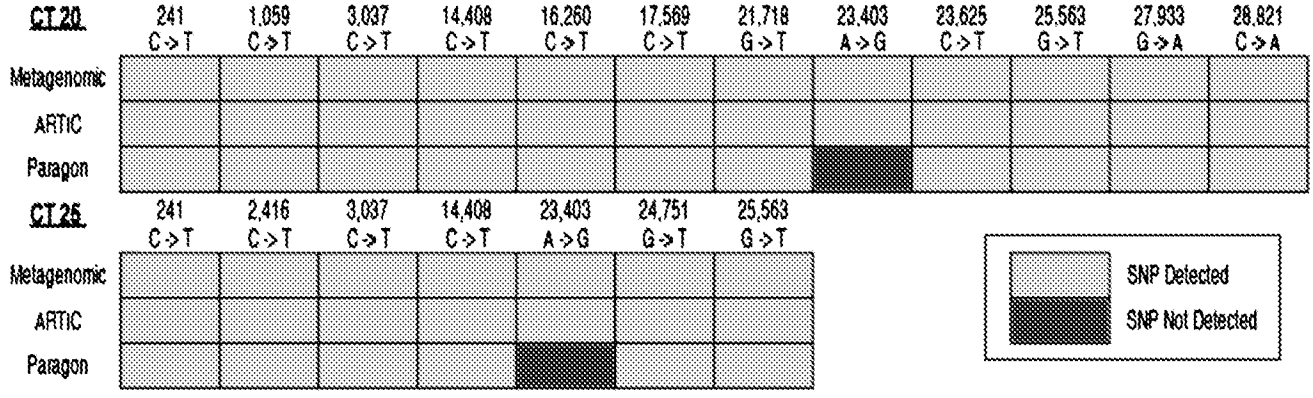

CT20
241 C->T
1,059 C->T
3,037 C->T
14,408 C->T
16,260 C->T
17,569 C->T
21,718 G->T
23,403 A->G
23,625 C->T
25,563 G->T
27,933 G->A
28,821 C->A
Metagenomic
ARTIC
Paragon
CT25
241 C->T
2,416 C->T
3,037 C->T
14,408 C->T
23,403 A->G
24,751 G->T
25,563 G->T
SNP Detected
SNP Not Detected

FIG. 8B

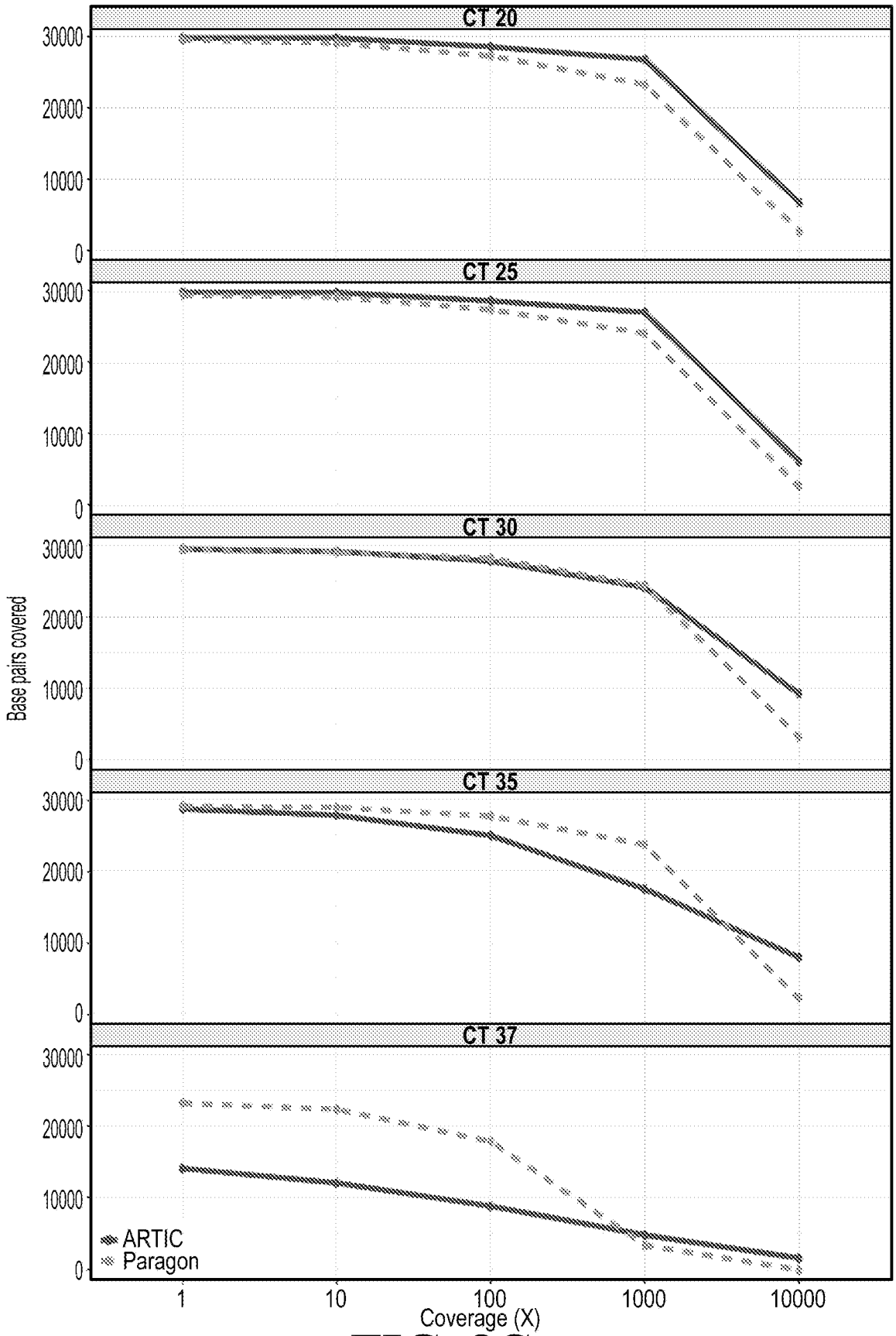

CT 20
CT 25
CT 30
CT 35
CT 37
30000
20000
10000
0
Base pairs covered
ARTIC
Paragon
1
10
100
1000
10000
Coverage (X)

FIG. 8C

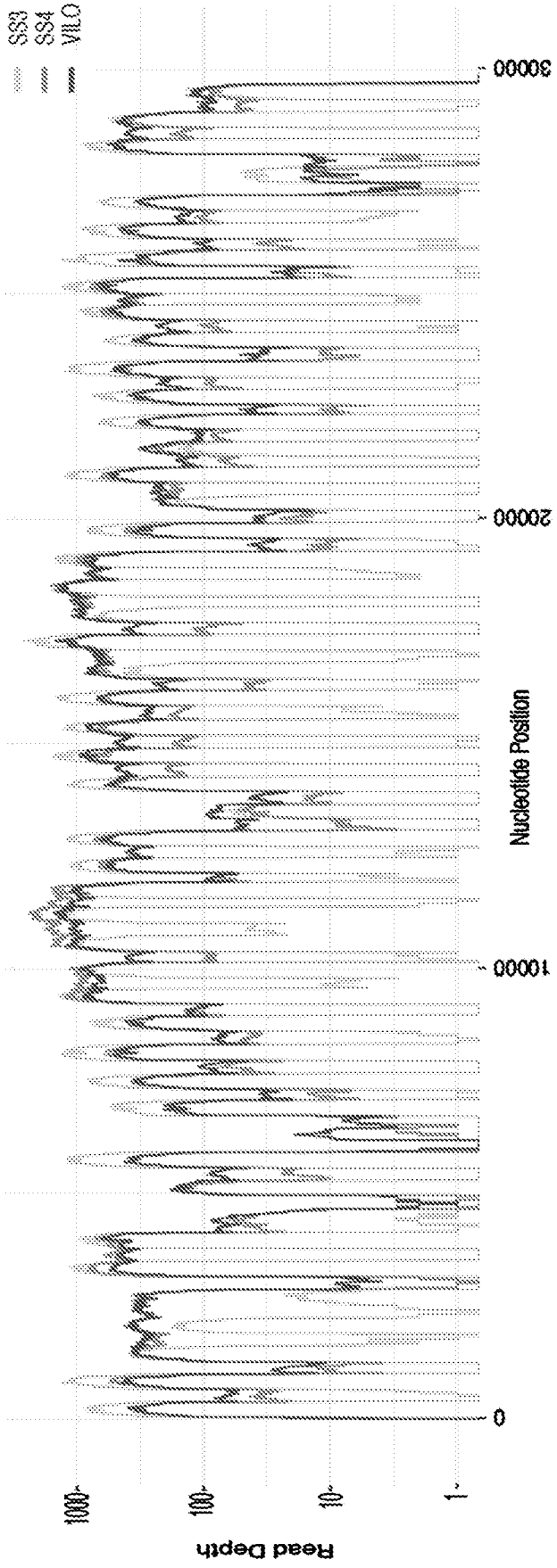

SS3
SS4
VILO
30000
20000
10000
0
Nucleotide Position
1000
100
10
1
Read Depth

FIG. 9A

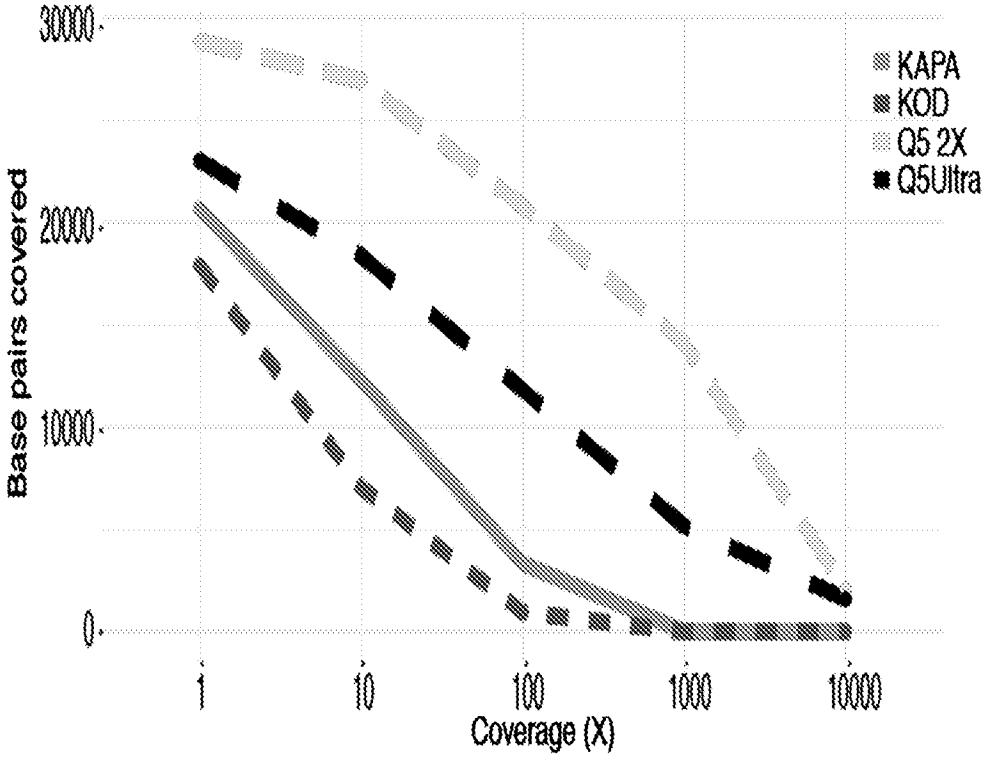

30000
20000
10000
0
Base pairs covered
1
10
100
1000
10000
Coverage (X)
KAPA
KOD
Q5 2X
Q5Ultra

FIG. 9B

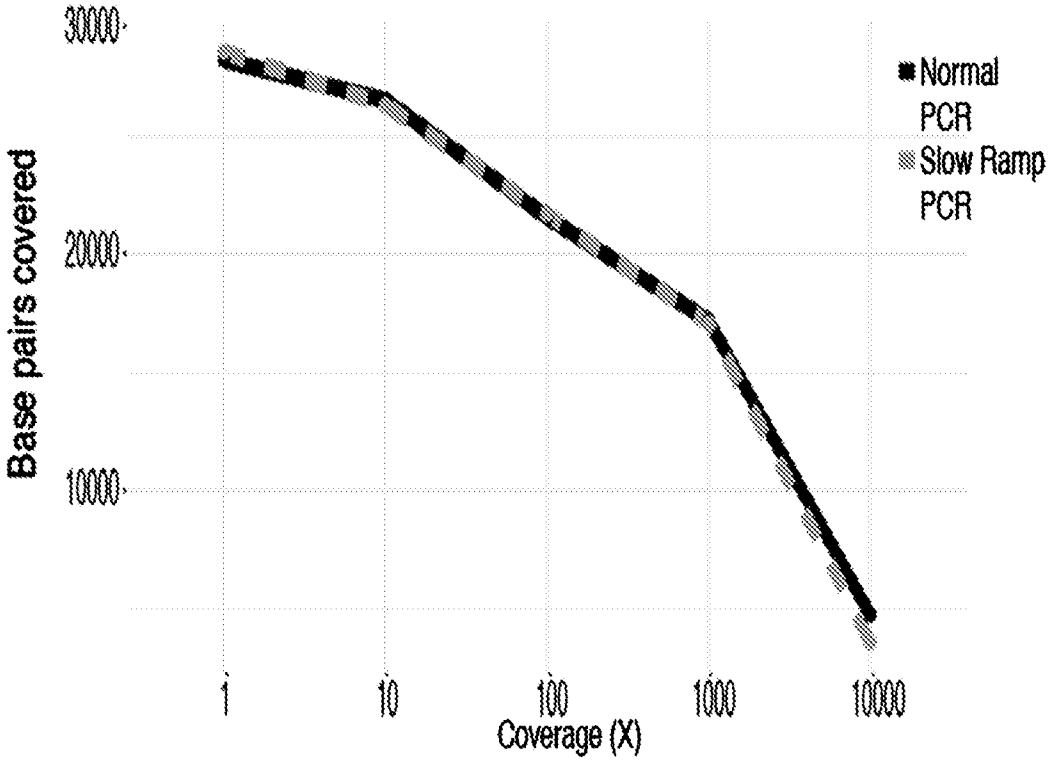

30000
20000
10000
Base pairs covered
1
10
100
1000
10000
Coverage (X)
Normal PCR
Slow Ramp PCR

FIG. 9C

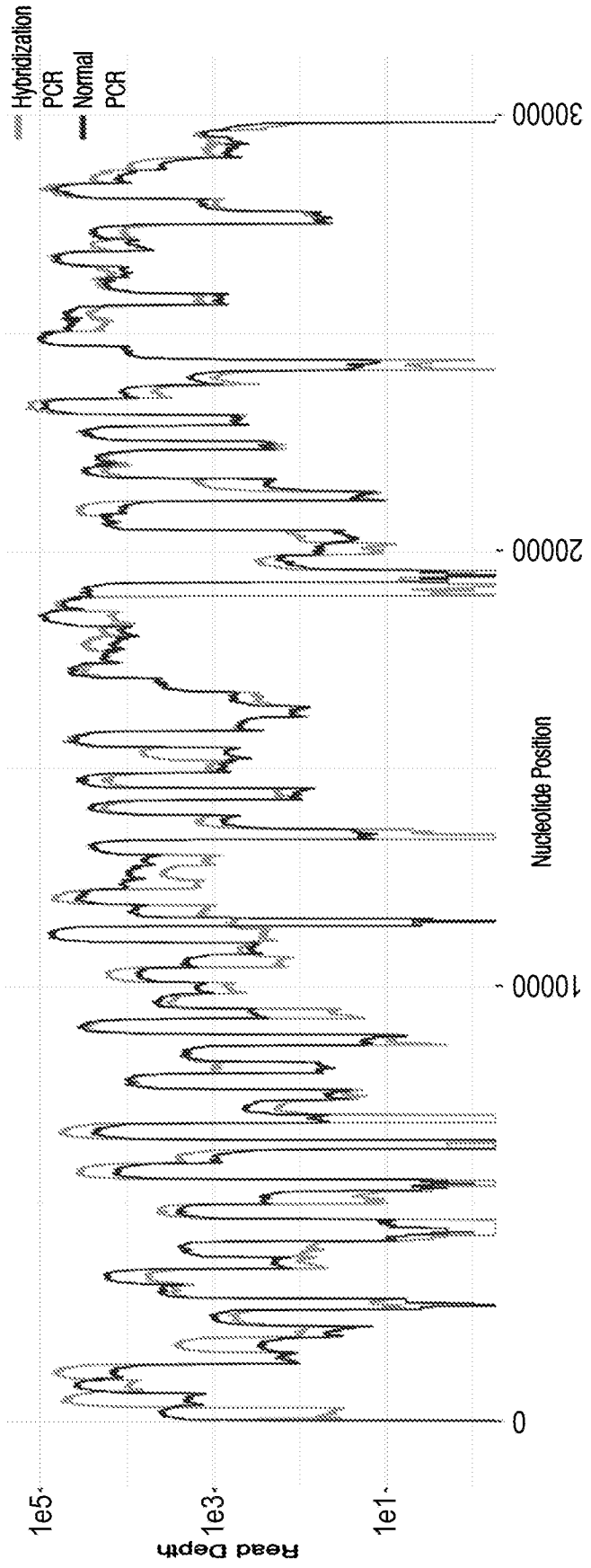

Hybridization
PCR
Normal
PCR
30000
20000
10000
0
Nucleotide Position
1e5
1e3
1e1
Read Depth

FIG. 9D

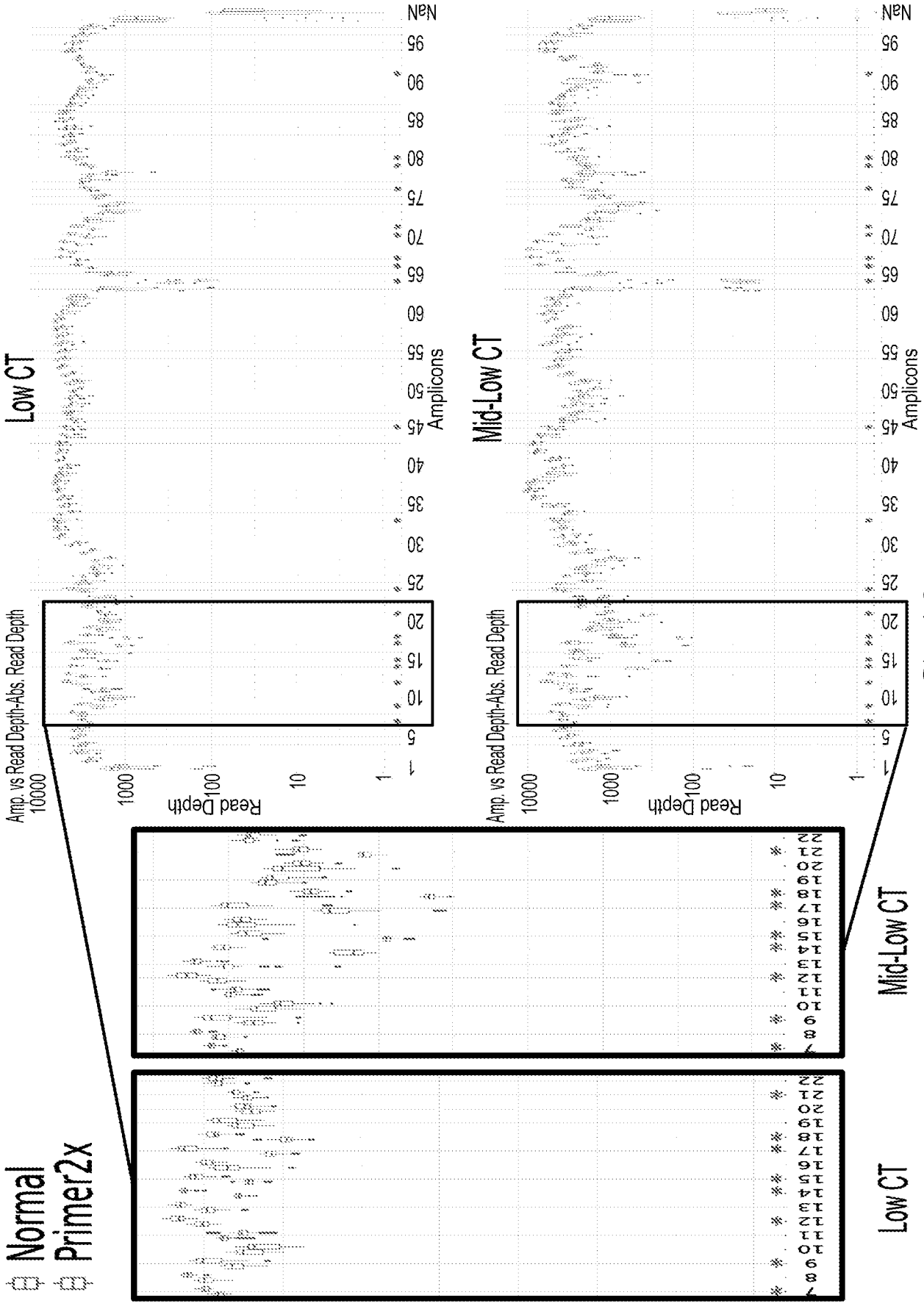

Normal
Primer2x
Low CT
Mid-Low CT
Low CT
Mid-Low CT
Amplicons
Read Depth
Amp. vs Read Depth-Abs. Read Depth
NaN

FIG. 10

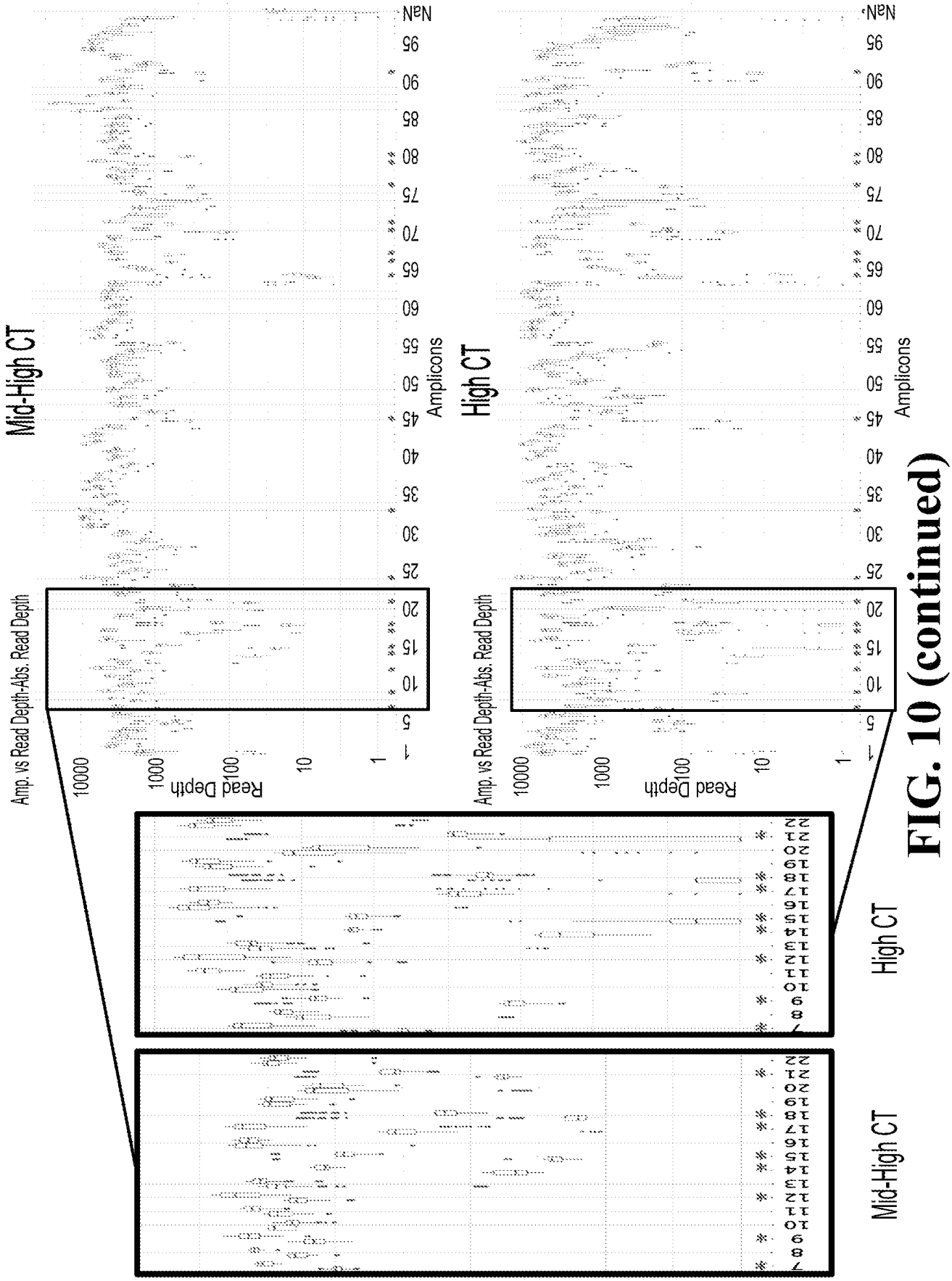


FIG. 10 (continued)

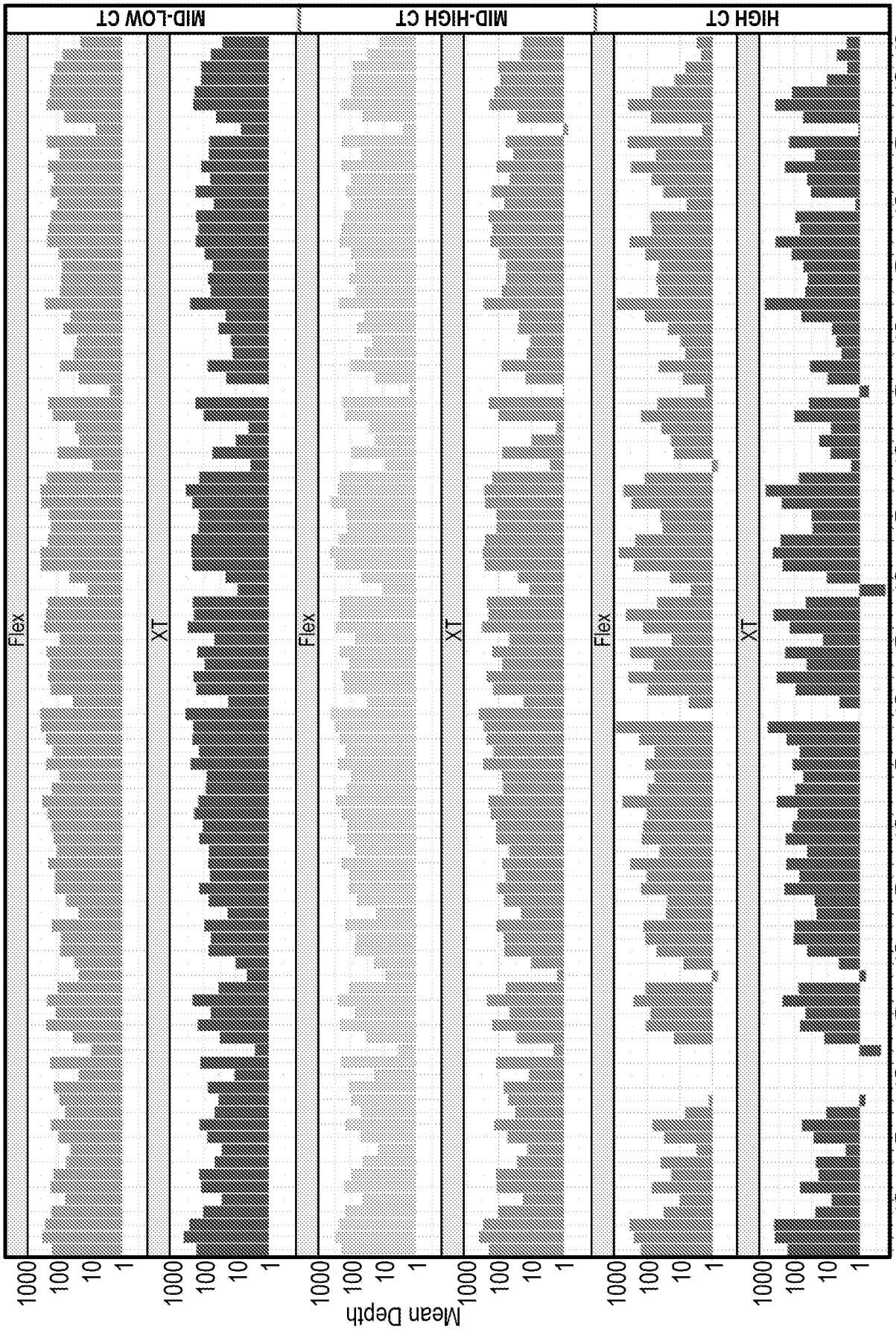

HIGH CT
MID-HIGH CT
MID-LOW CT
Flex
XT
Mean Depth
1000
100
10
1
Amplicons
01
05
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
98

FIG. 11

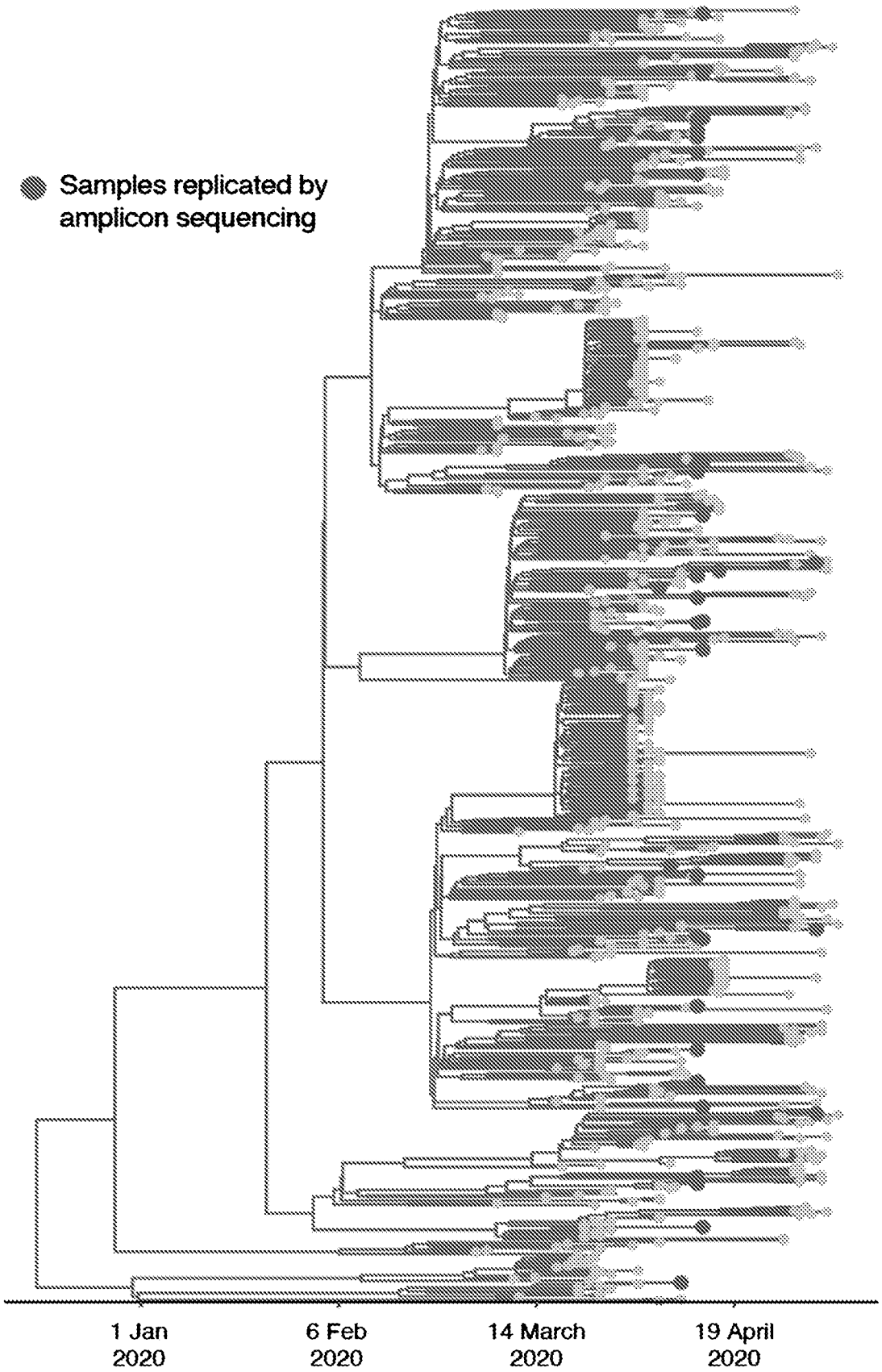

Samples replicated by amplicon sequencing
1 Jan 2020
6 Feb 2020
14 March 2020
19 April 2020

FIG. 12A

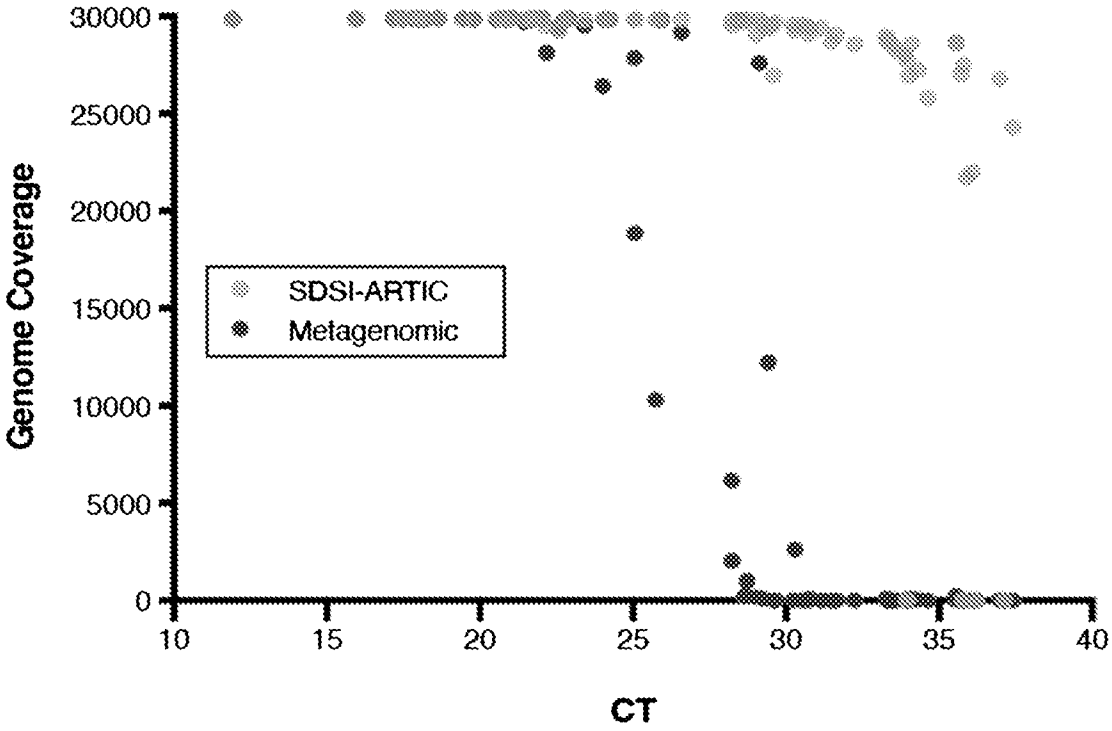


FIG. 12B

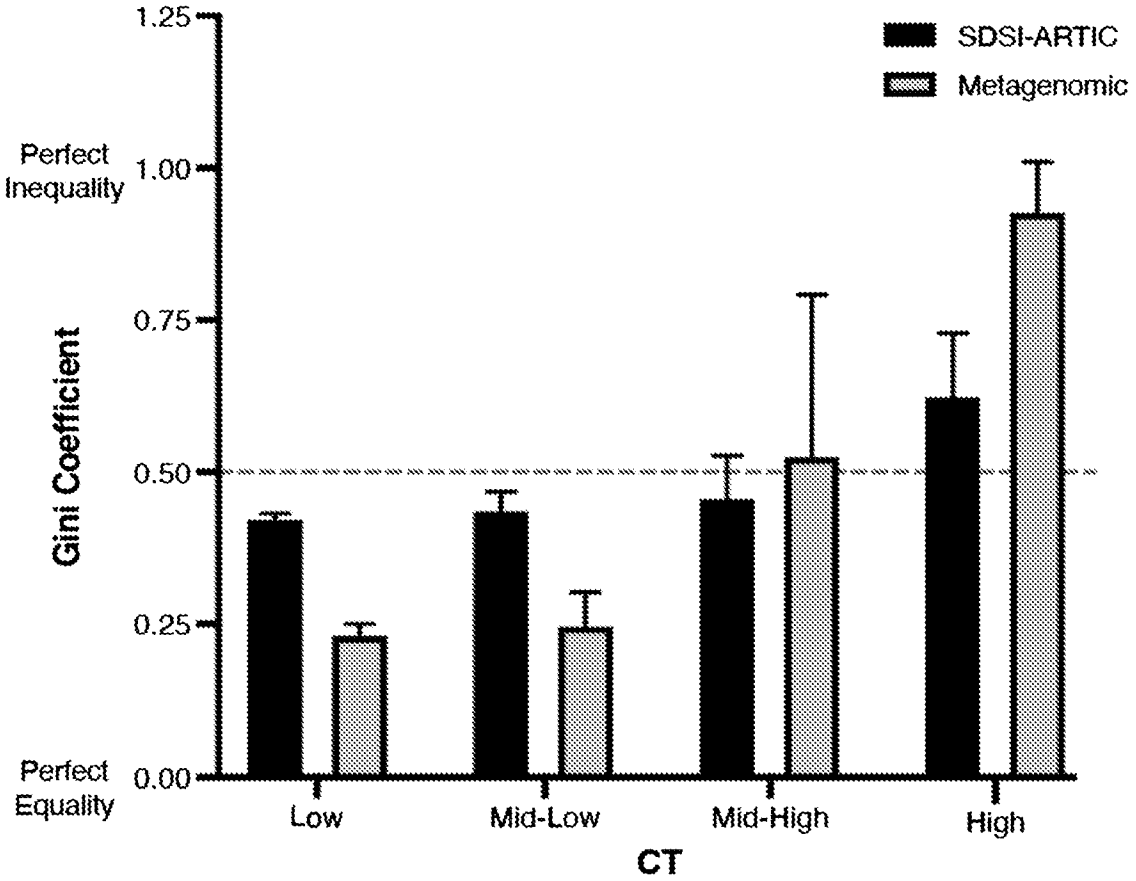


FIG. 12C

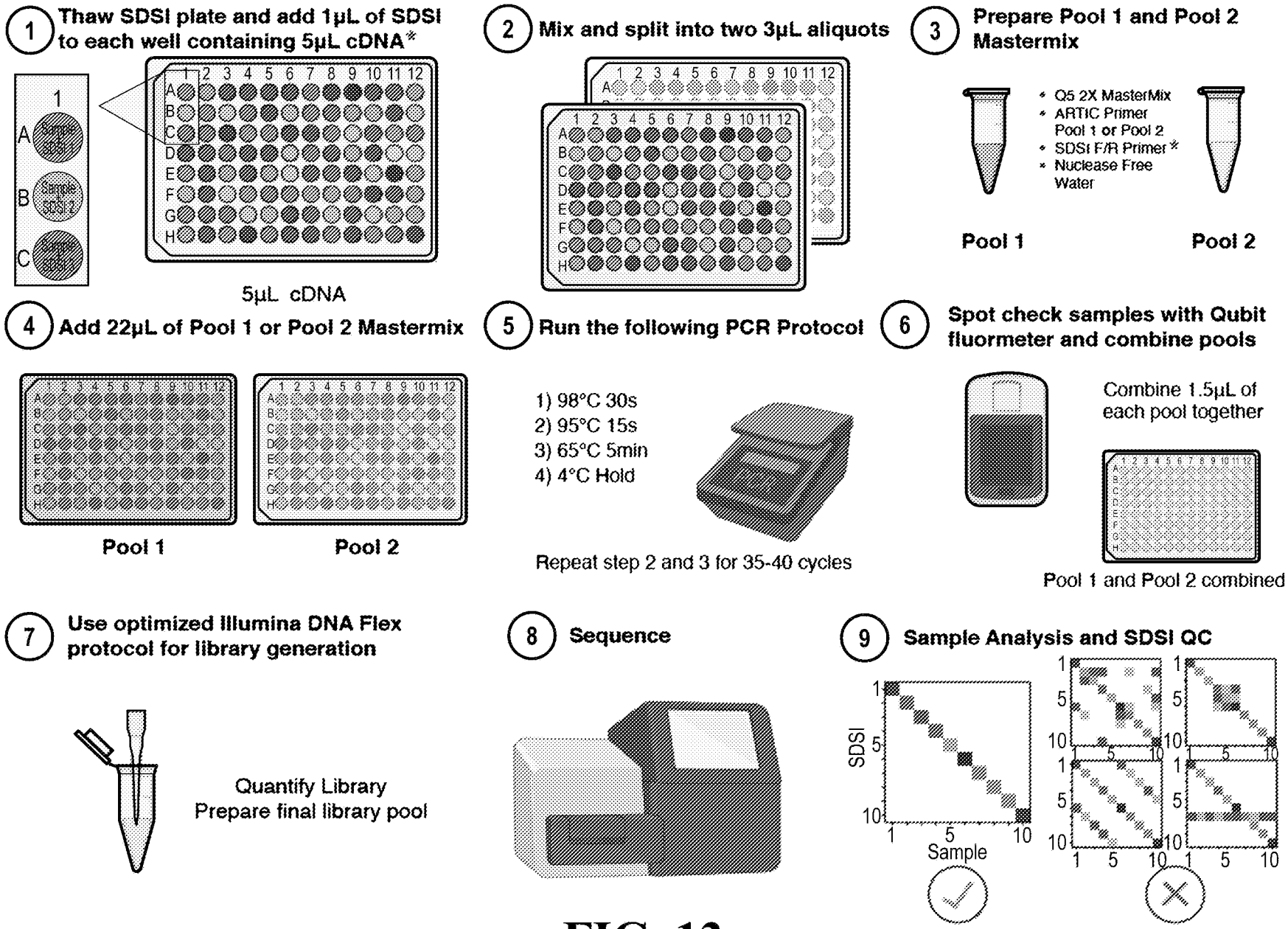

1 Thaw SDSI plate and add 1µL of SDSI to each well containing 5µL cDNA*
5µL cDNA
2 Mix and split into two 3µL aliquots
3 Prepare Pool 1 and Pool 2 Mastermix
Q5 2X MasterMix
ARTIC Primer Pool 1 or Pool 2
SDSI F/R Primer*
Nuclease Free Water
Pool 1
Pool 2
4 Add 22µL of Pool 1 or Pool 2 Mastermix
Pool 1
Pool 2
5 Run the following PCR Protocol
1) 98°C 30s
2) 95°C 15s
3) 65°C 5min
4) 4°C Hold
Repeat step 2 and 3 for 35-40 cycles
6 Spot check samples with Qubit fluormeter and combine pools
Combine 1.5µL of each pool together
Pool 1 and Pool 2 combined
7 Use optimized Illumina DNA Flex protocol for library generation
Quantify Library
Prepare final library pool
8 Sequence
9 Sample Analysis and SDSI QC
SDSI
Sample

FIG. 13

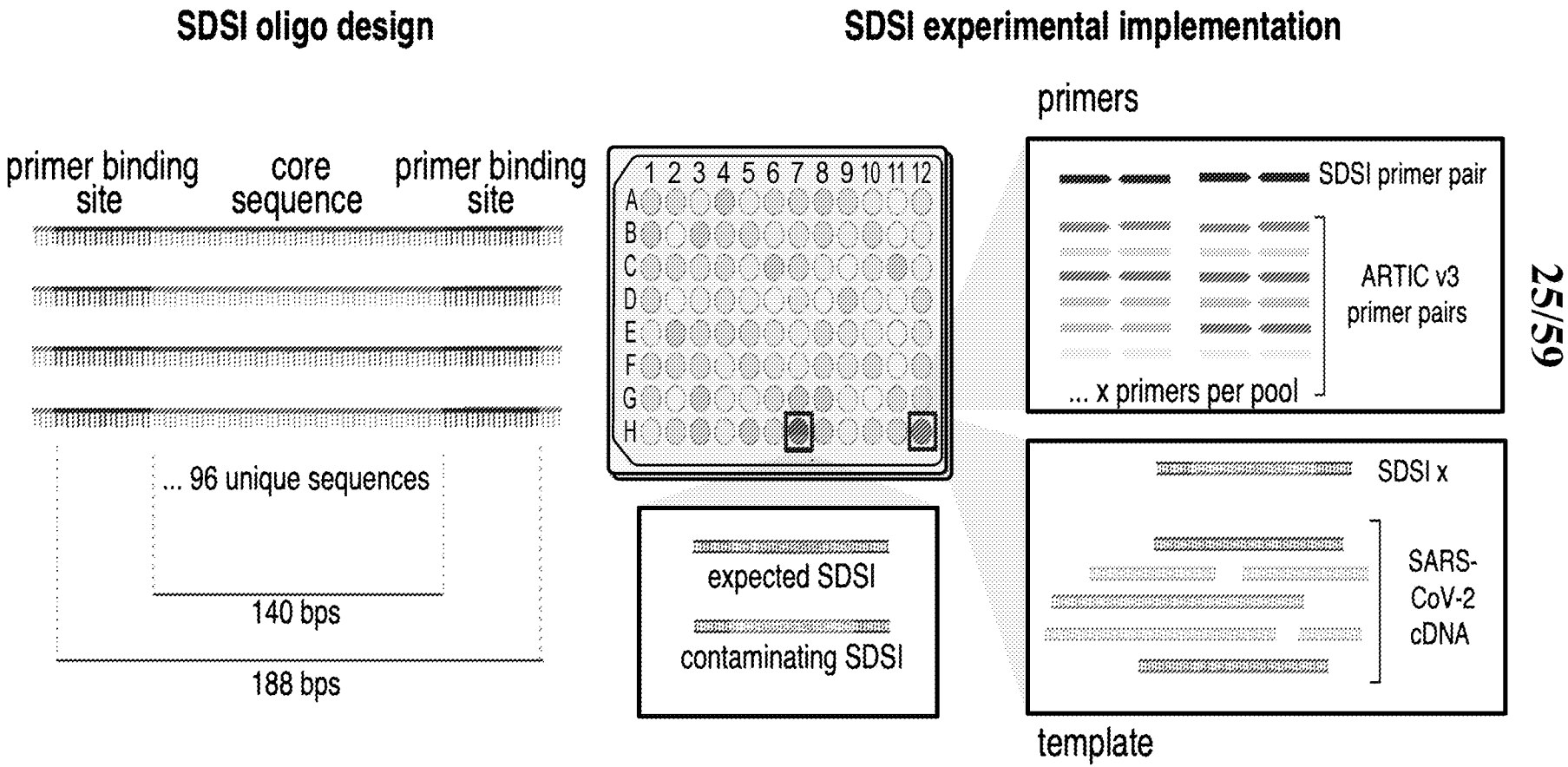


FIG. 14A

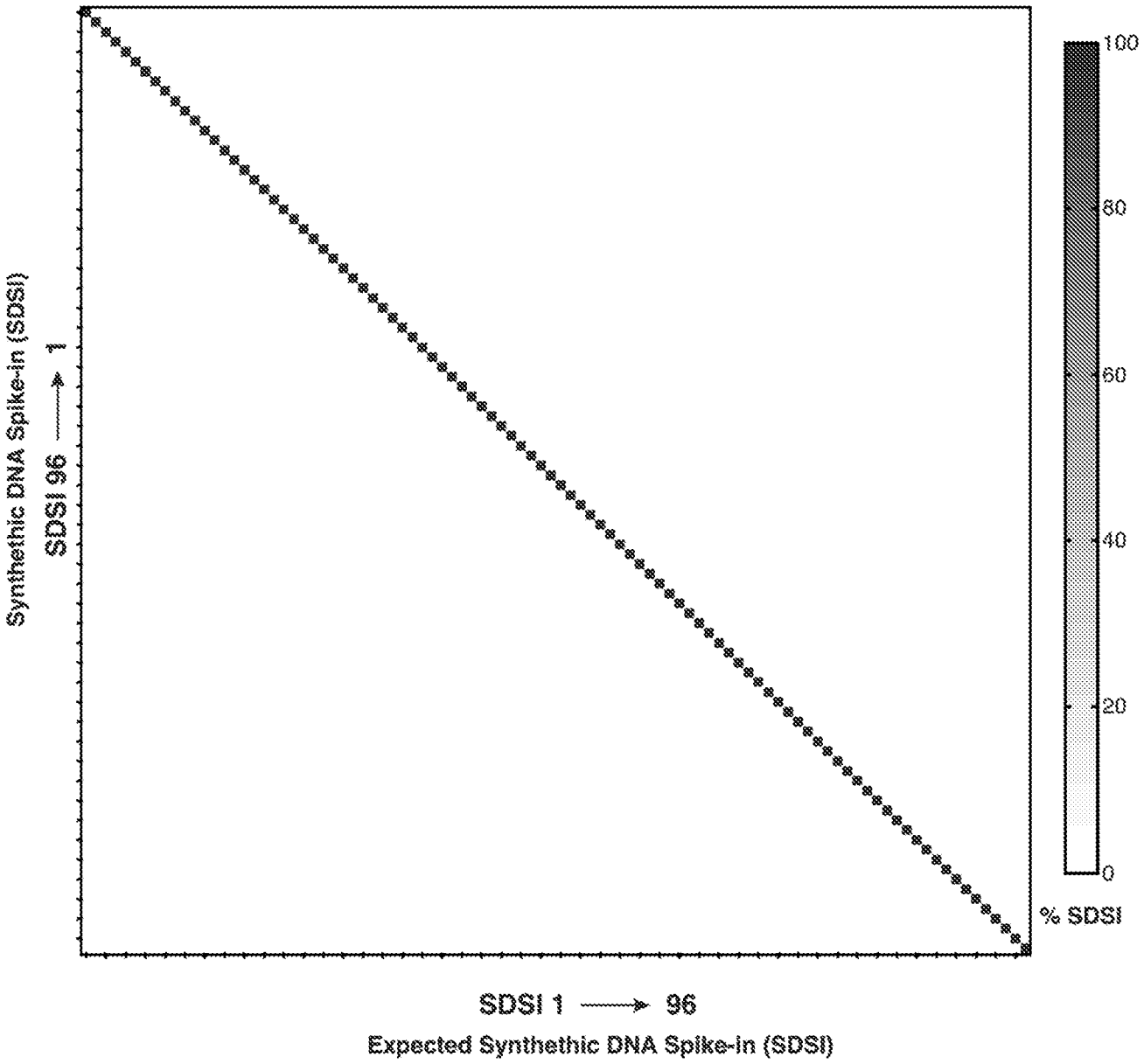

Synthetic DNA Spike-in (SDSI)
SDSI 96 ⟶ 1
SDSI 1 ⟶ 96
Expected Synthethic DNA Spike-in (SDSI)
100
80
60
40
20
0
% SDSI

FIG. 14B

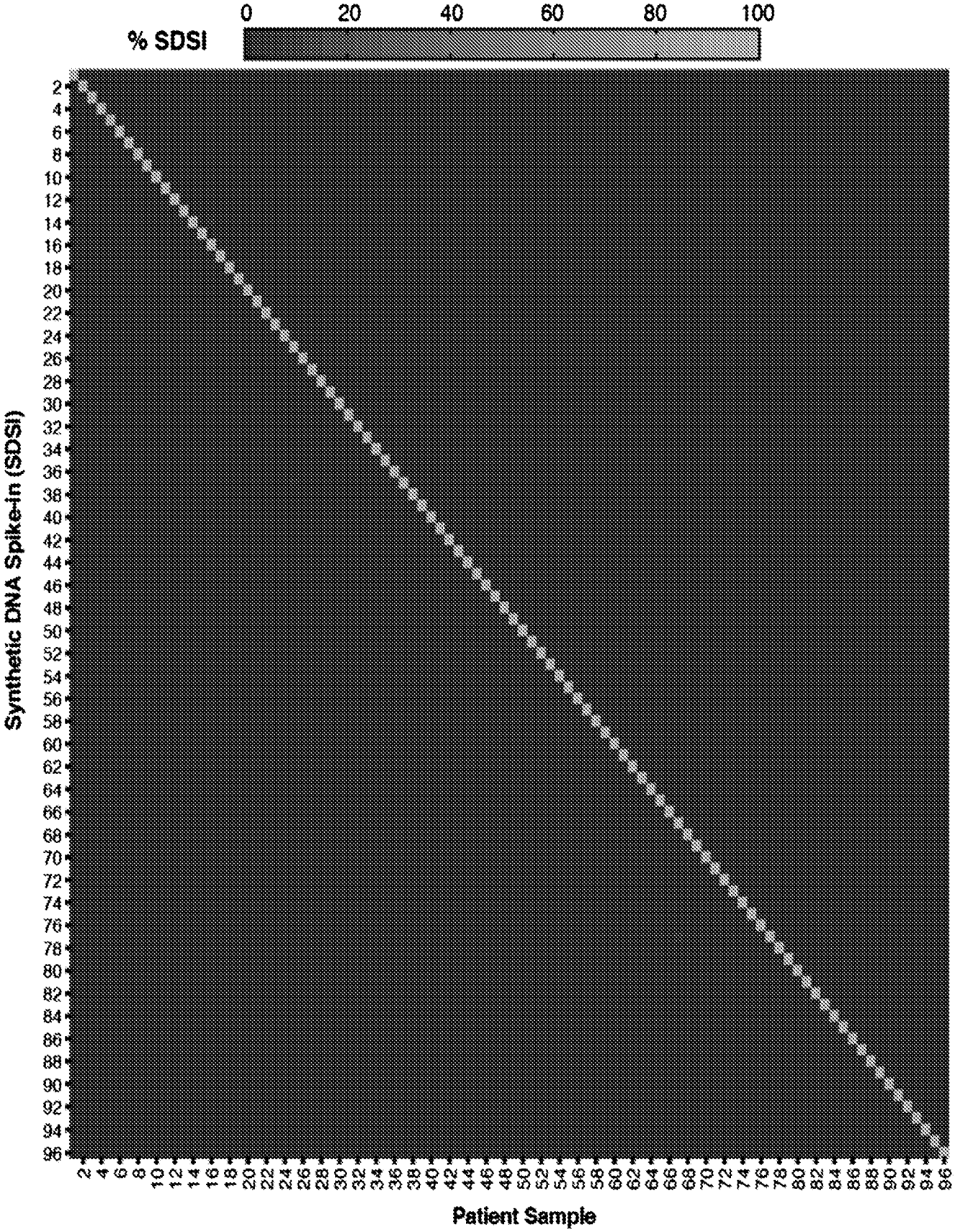

% SDSI
0
20
40
60
80
100
Synthetic DNA Spike-in (SDSI)
Patient Sample

FIG. 15A

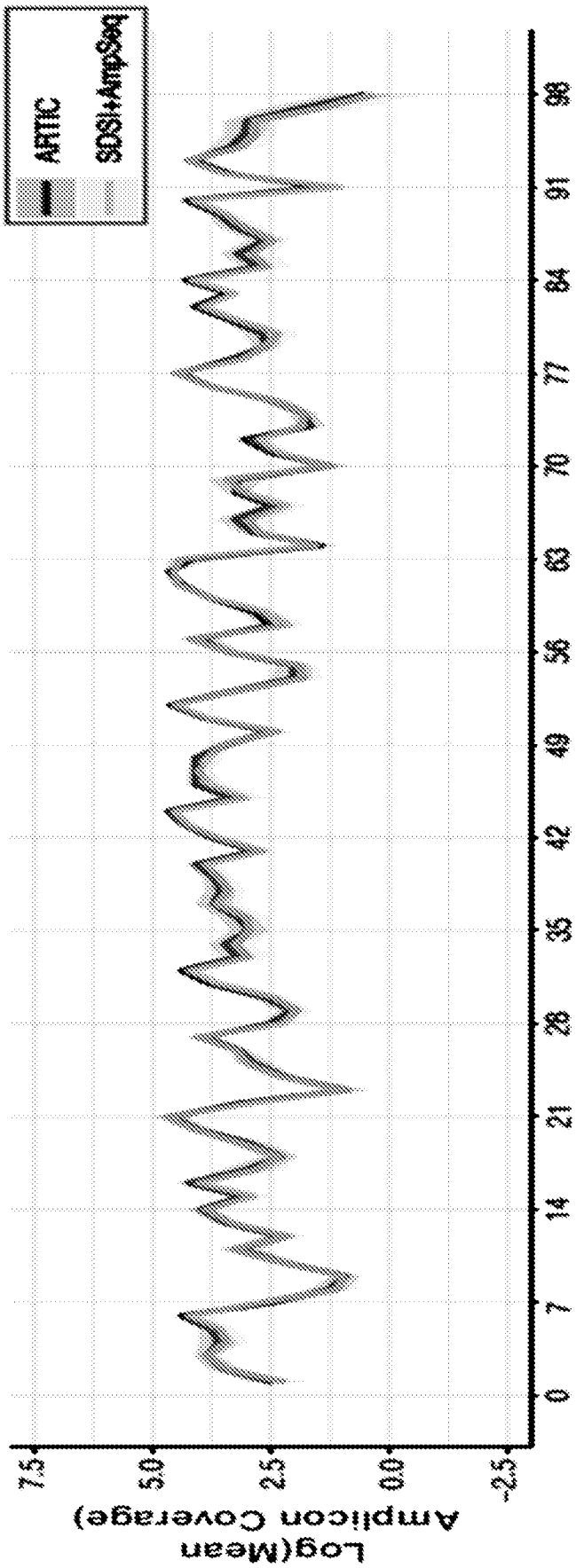

AIRTIC
SDSI+AmpSeq
Log(Mean Amplicon Coverage)
7.5
5.0
2.5
0.0
-2.5
0
7
14
21
28
35
42
49
56
63
70
77
84
91
98

FIG. 15B

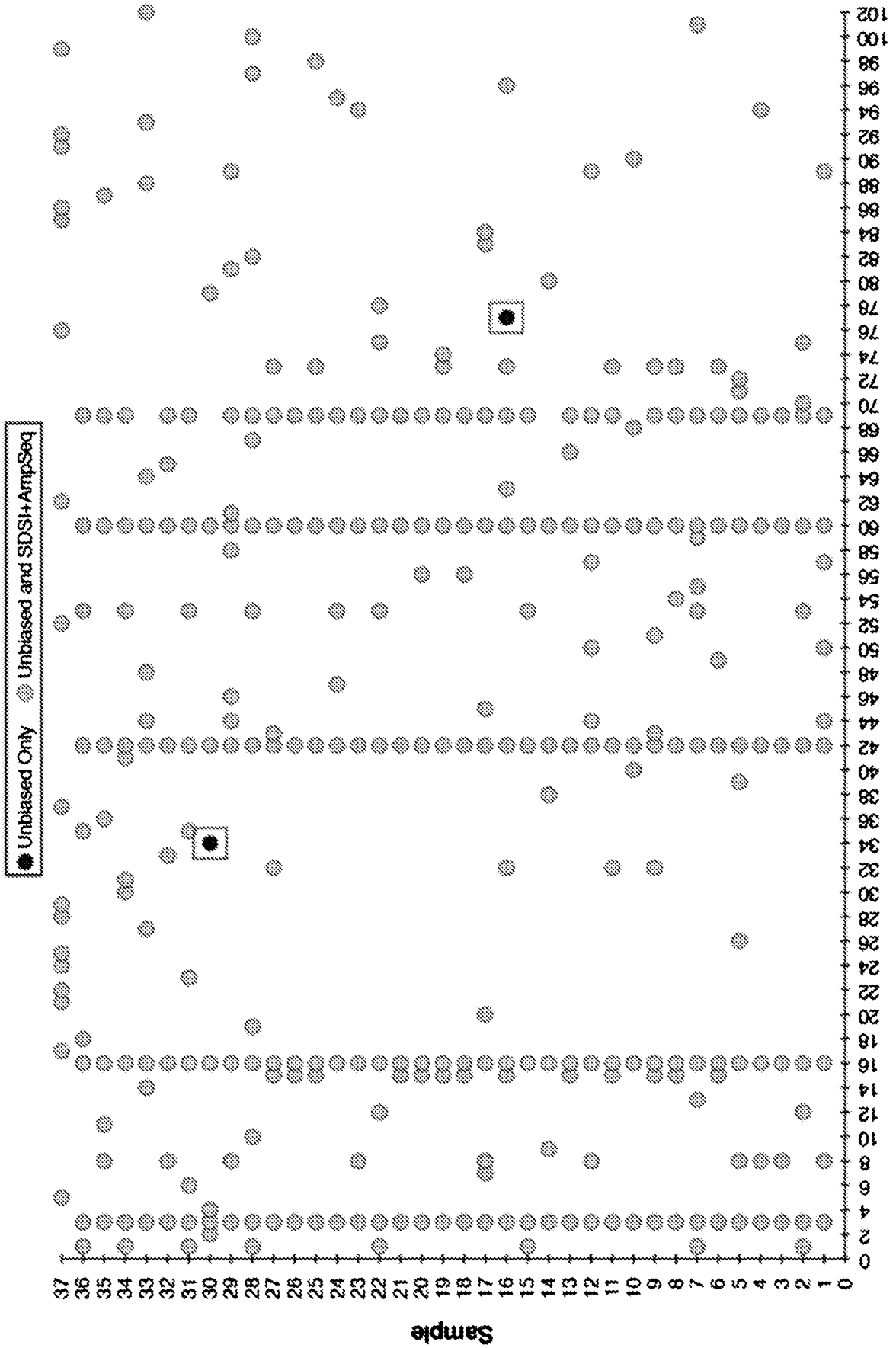


FIG. 15C

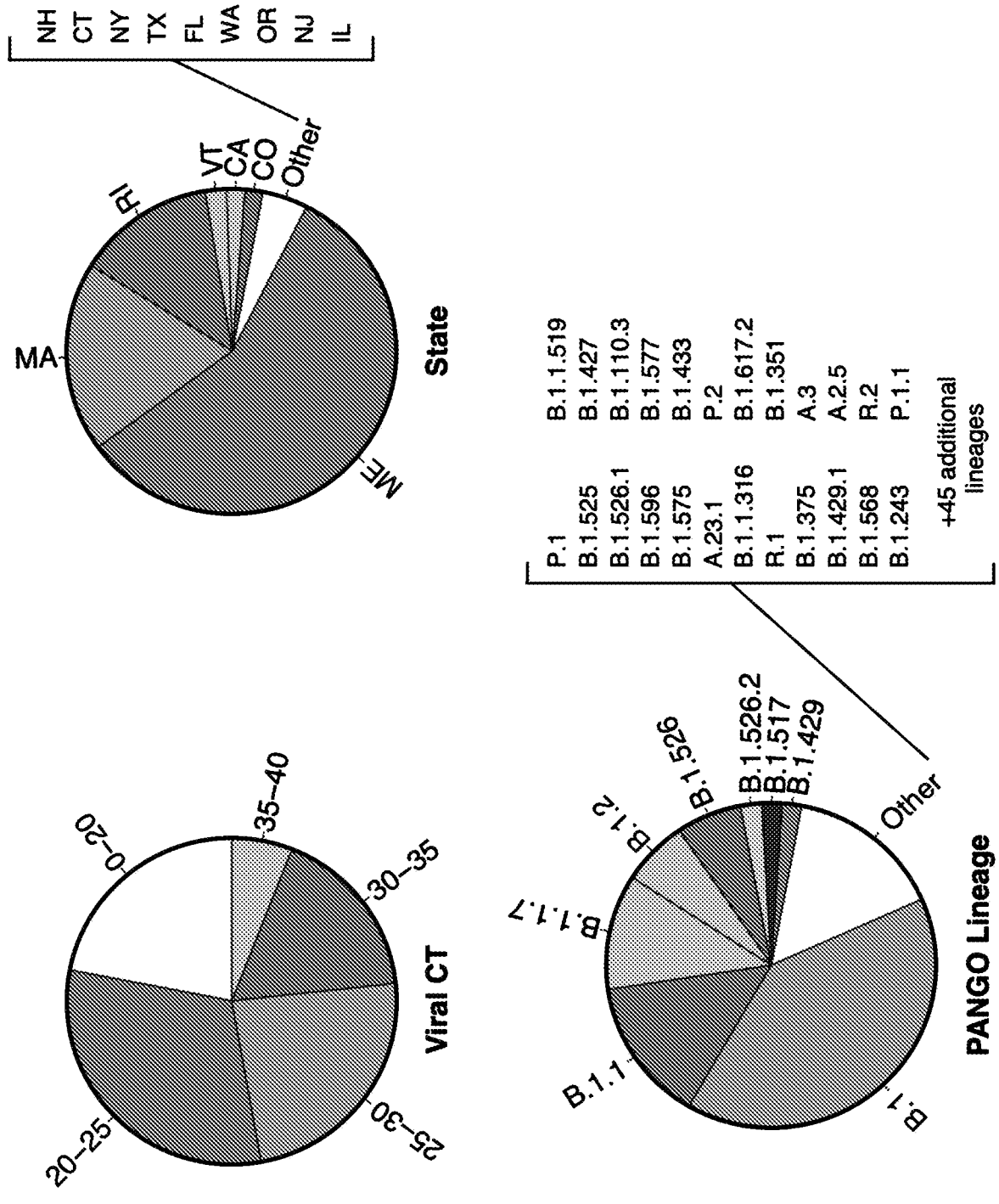

NH
CT
NY
TX
FL
WA
OR
NJ
IL
VT
CA
CO
Other
RI
MA
ME
State
0–20
35–40
30–35
20–25
25–30
Viral CT
P.1
B.1.525
B.1.526.1
B.1.596
B.1.575
A.23.1
B.1.1.316
R.1
B.1.375
B.1.429.1
B.1.568
B.1.243
B.1.1.519
B.1.427
B.1.110.3
B.1.577
B.1.433
P.2
B.1.617.2
B.1.351
A.3
A.2.5
R.2
P.1.1
+45 additional lineages
B.1.526.2
B.1.517
B.1.429
Other
B.1.526
B.1.2
B.1.1.7
B.1.1
B.1
PANGO Lineage

FIG. 16A

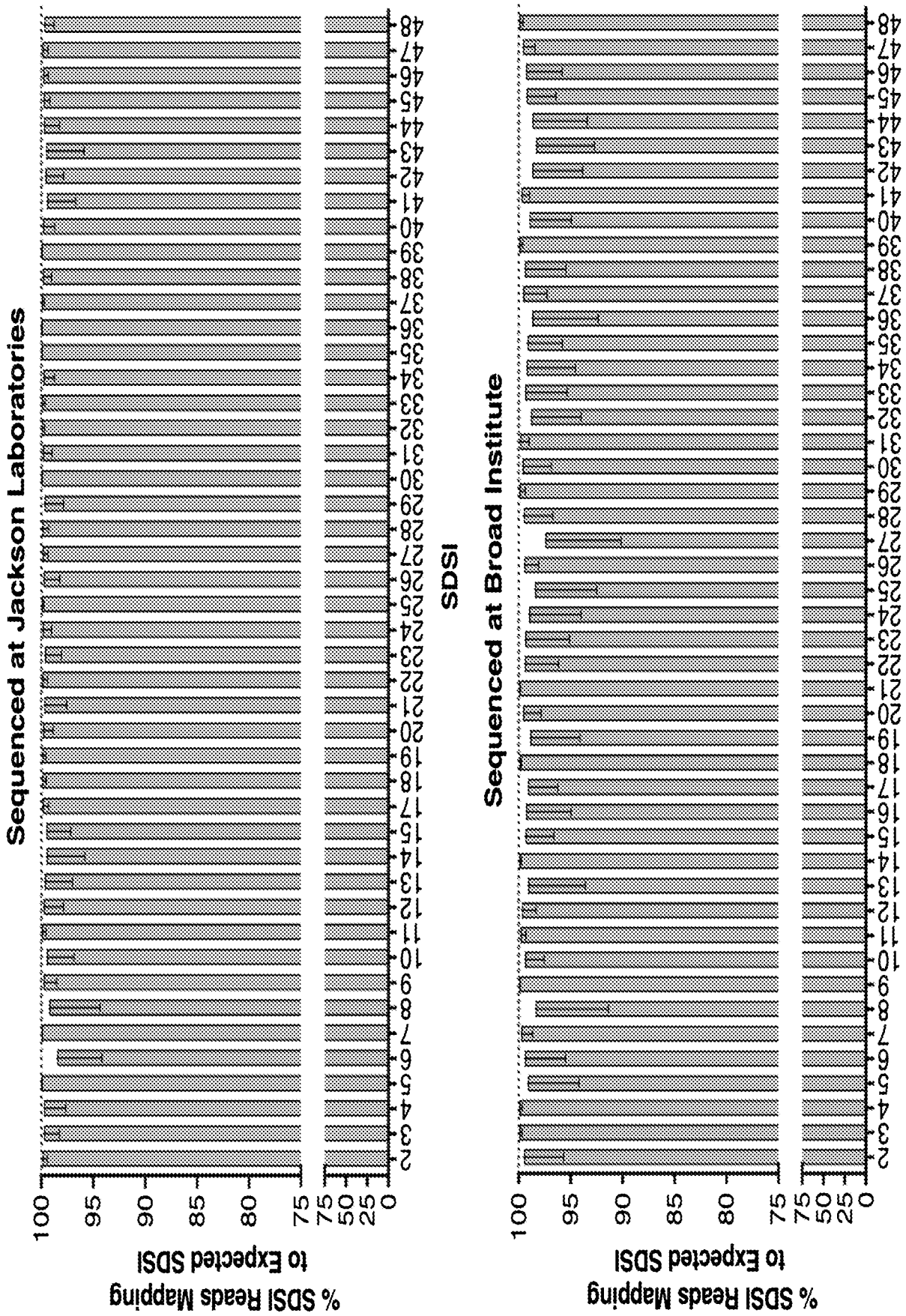

Sequenced at Jackson Laboratories
% SDSI Reads Mapping to Expected SDSI
100
95
90
85
80
75
75
50
25
0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
SDSI
Sequenced at Broad Institute
% SDSI Reads Mapping to Expected SDSI
100
95
90
85
80
75
75
50
25
0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
SDSI

FIG. 16B

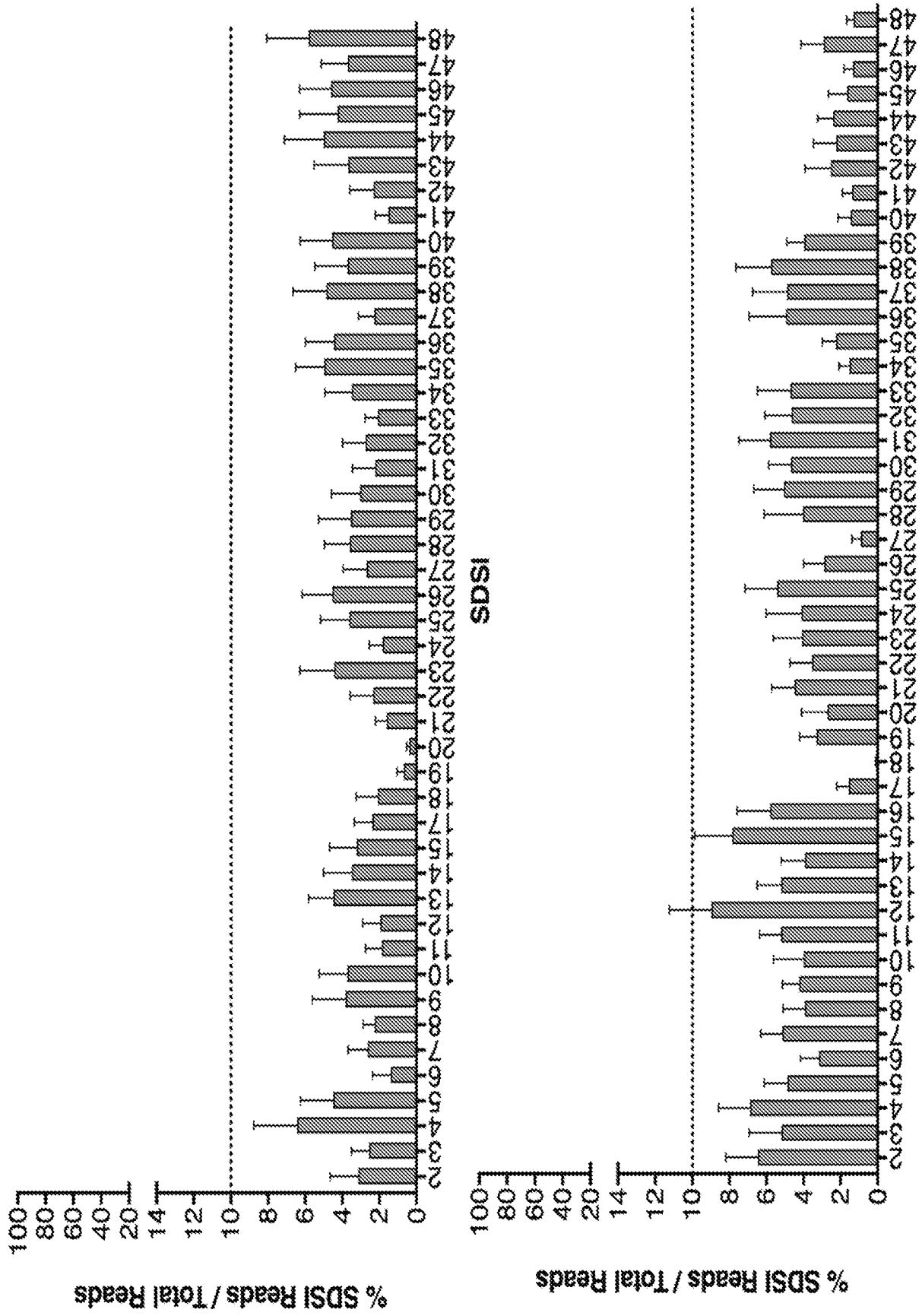


FIG. 16C

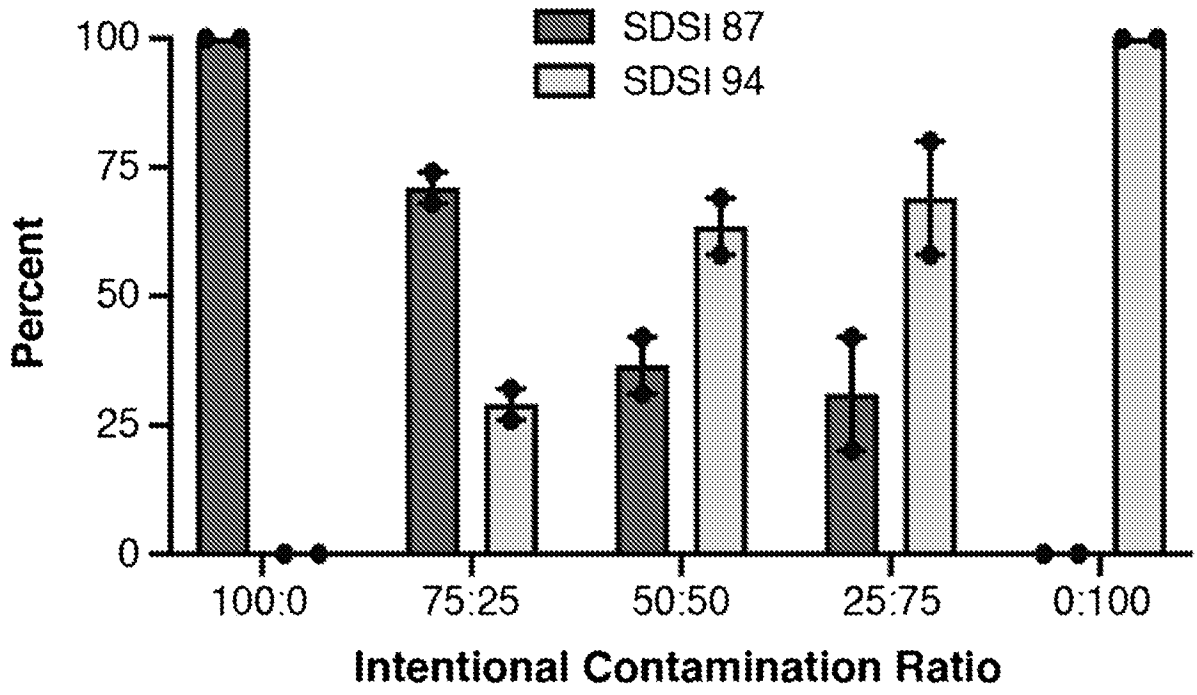

SDSI 87
SDSI 94
100
75
50
25
0
Percent
100:0
75:25
50:50
25:75
0:100
Intentional Contamination Ratio

FIG. 17A

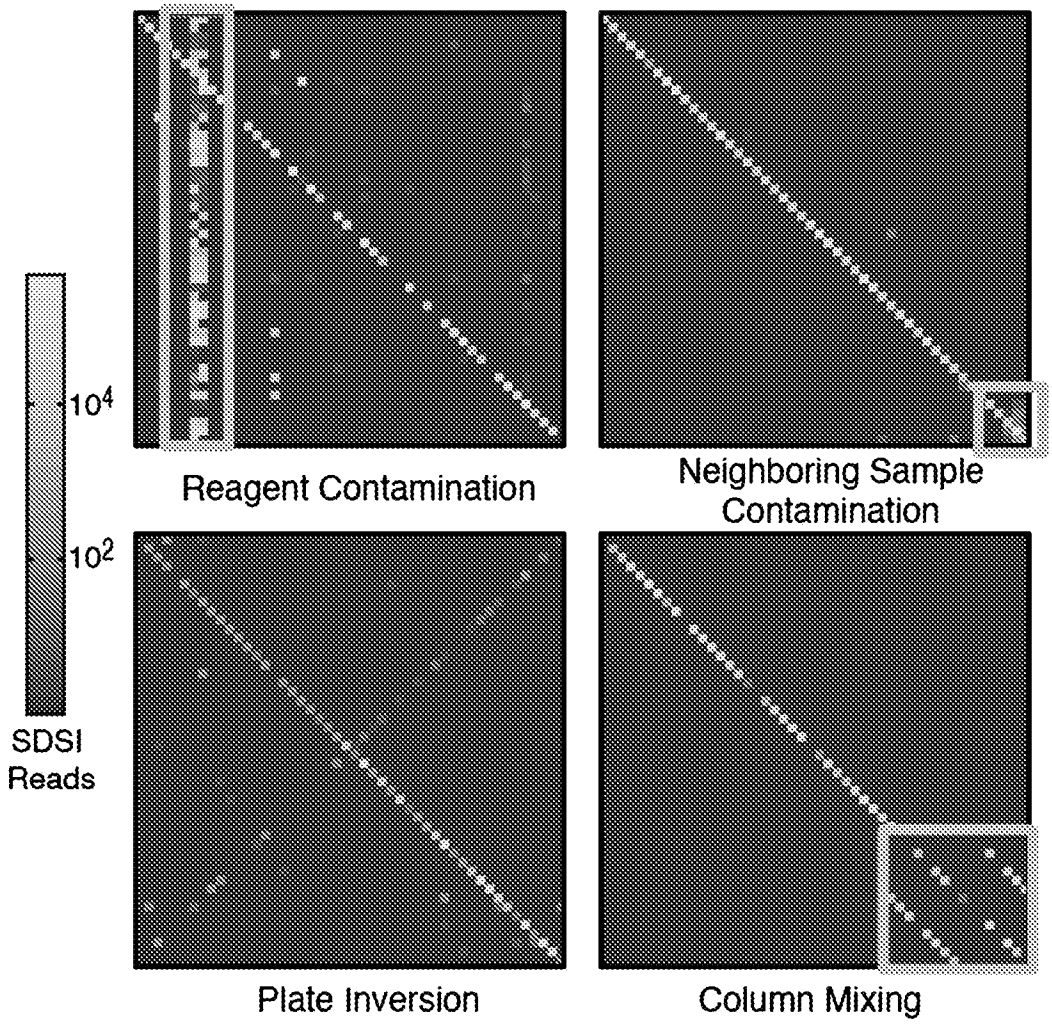

10^4
10^2
SDSI
Reads
Reagent Contamination
Neighboring Sample
Contamination
Plate Inversion
Column Mixing

FIG. 17B

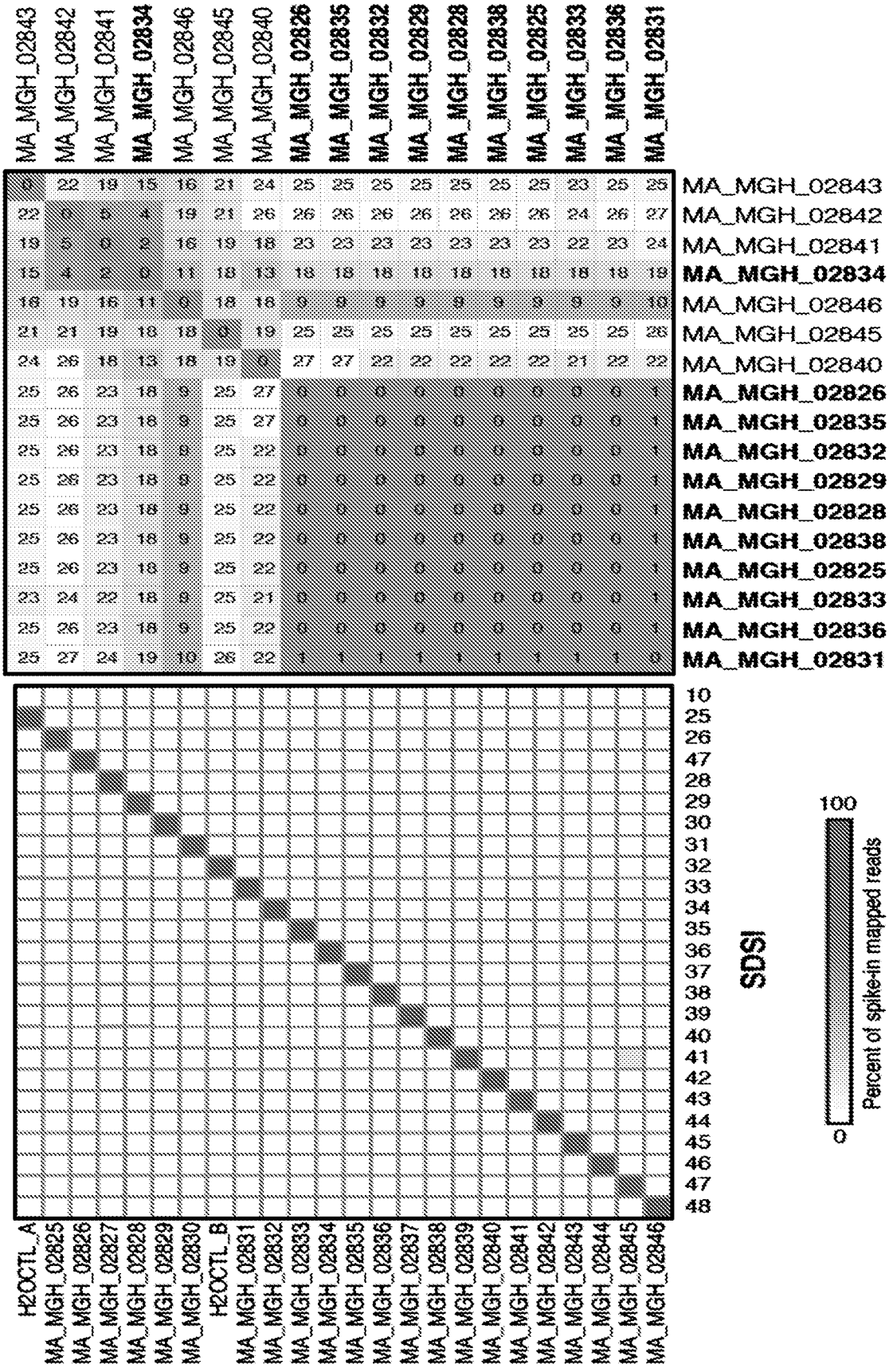

MA_MGH_02843
MA_MGH_02842
MA_MGH_02841
MA_MGH_02834
MA_MGH_02846
MA_MGH_02845
MA_MGH_02840
MA_MGH_02826
MA_MGH_02835
MA_MGH_02832
MA_MGH_02829
MA_MGH_02828
MA_MGH_02838
MA_MGH_02825
MA_MGH_02833
MA_MGH_02836
MA_MGH_02831
SDSI
100
Percent of spike-in mapped reads
0
H2OCTL_A
H2OCTL_B

FIG. 17C

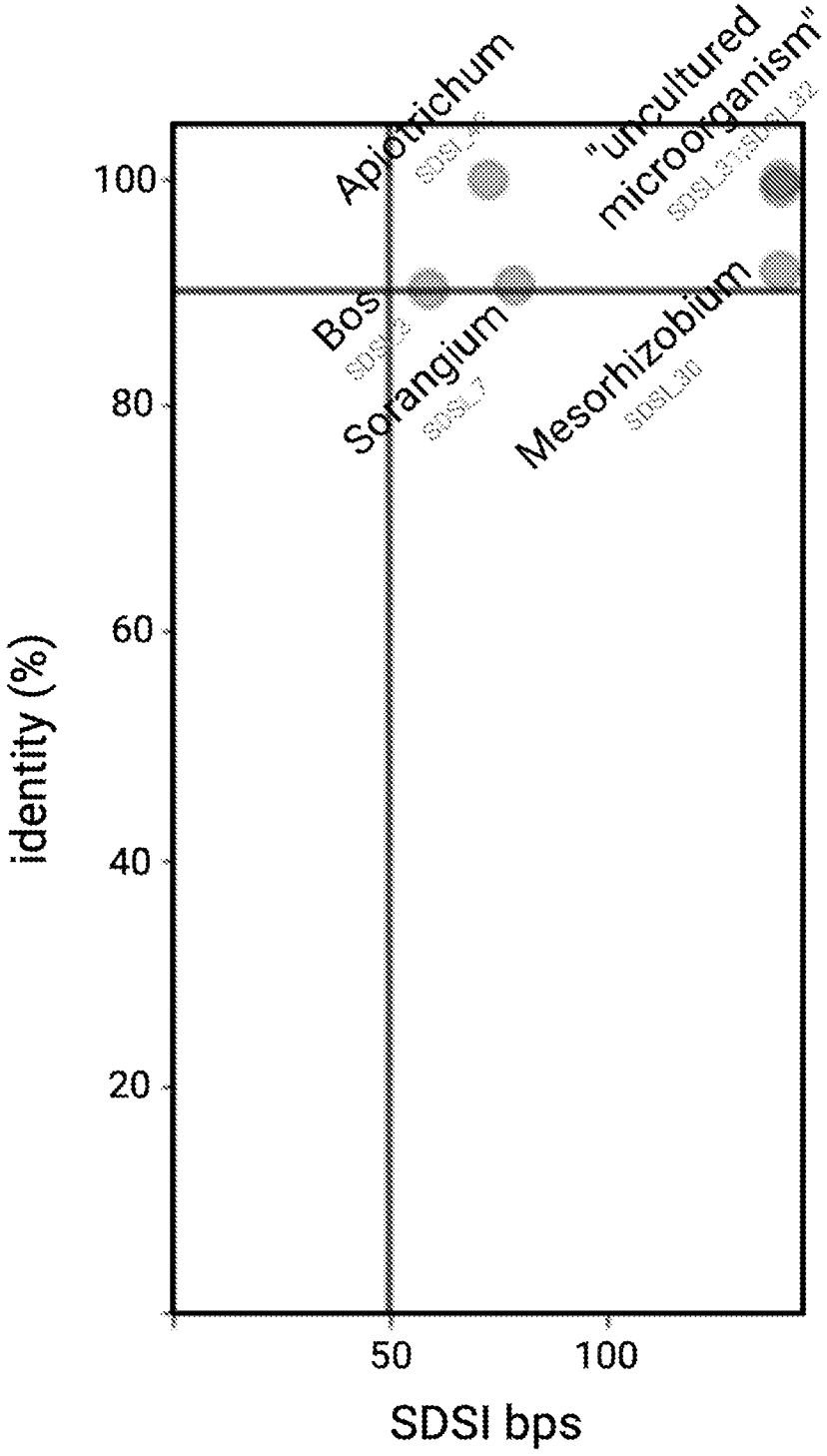

Apiotrichum
SDSL_6
"uncultured microorganism"
SDSL32-SDSL32
Bos
SDSL2
Sorangium
SDSL7
Mesorhizobium
SDSL30
100
80
60
40
20
identity (%)
50
100
SDSI bps

FIG. 18A

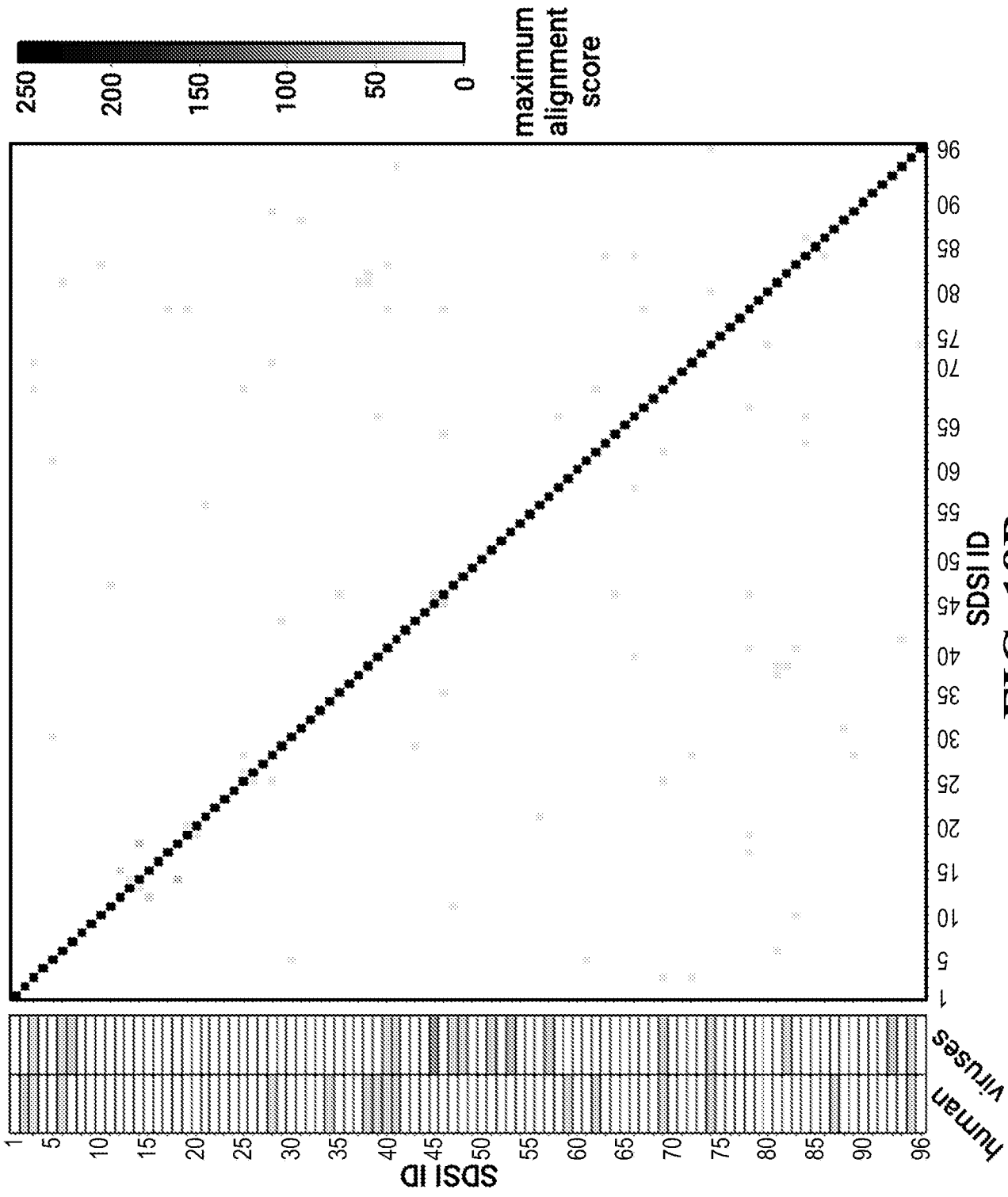

maximum alignment score
250
200
150
100
50
0
SDSI ID
1
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
96
human viruses
SDSI ID

FIG. 18B

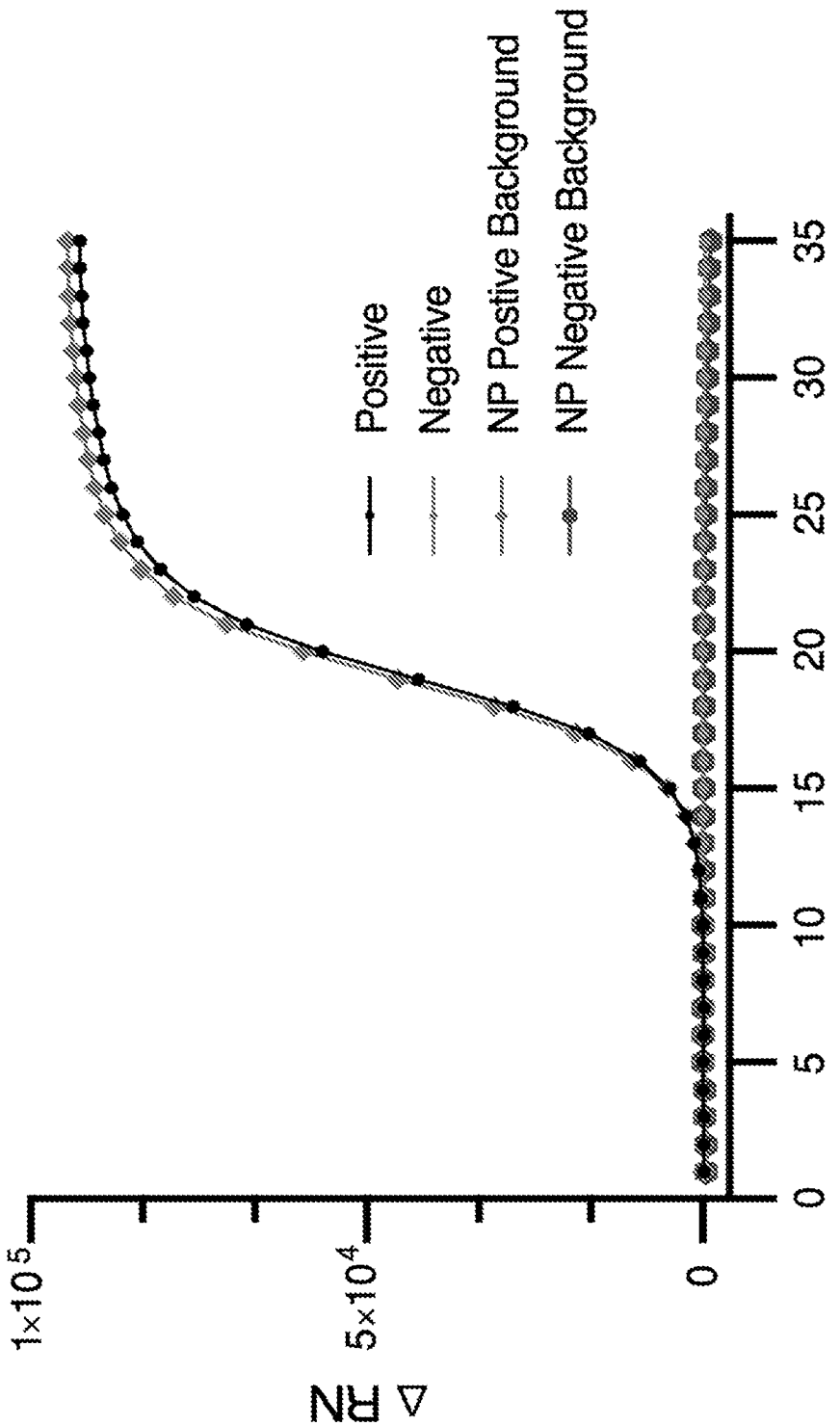

Positive
Negative
NP Postive Background
NP Negative Background
ΔRN
1×10^5
5×10^4
0
0
5
10
15
20
25
30
35
Cycle

FIG. 19A

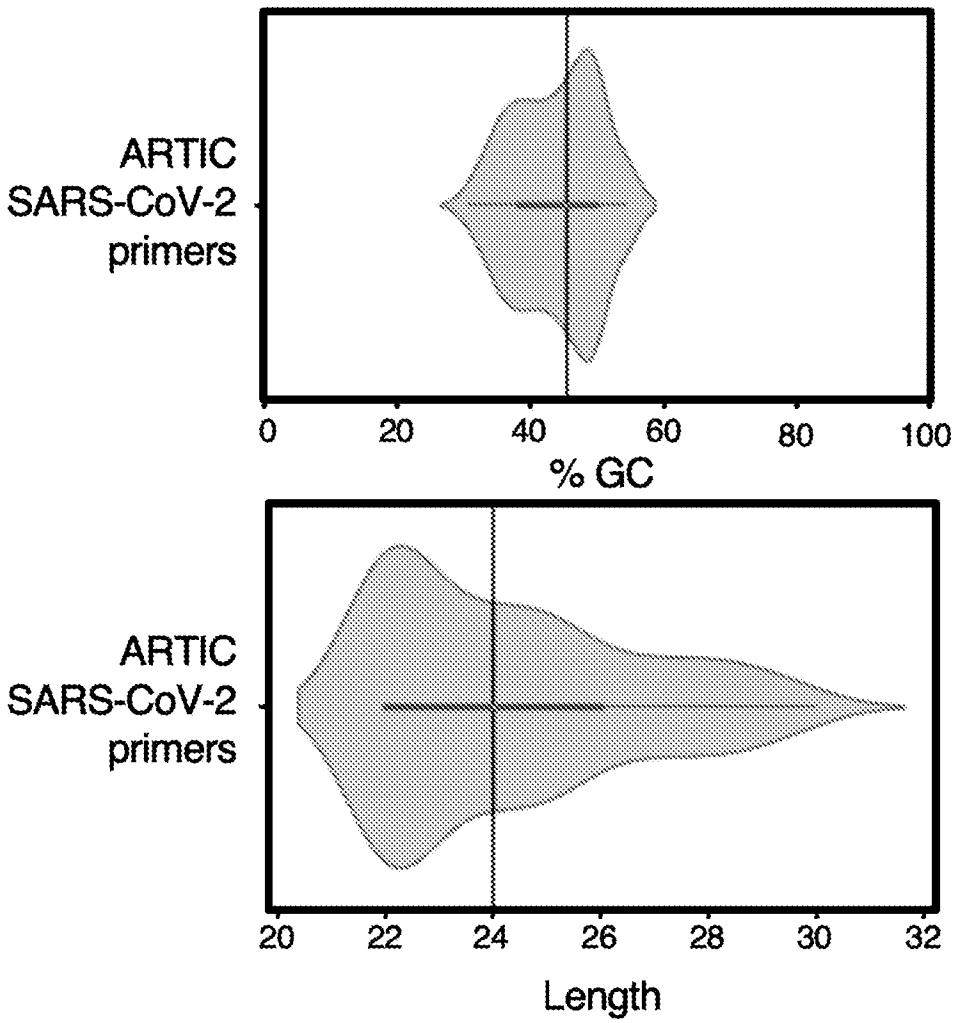


FIG. 19B

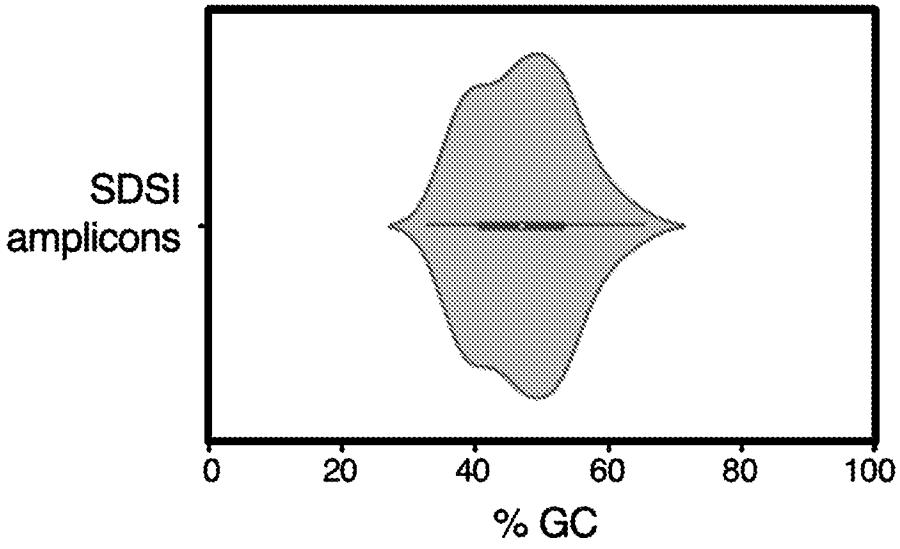


FIG. 19C

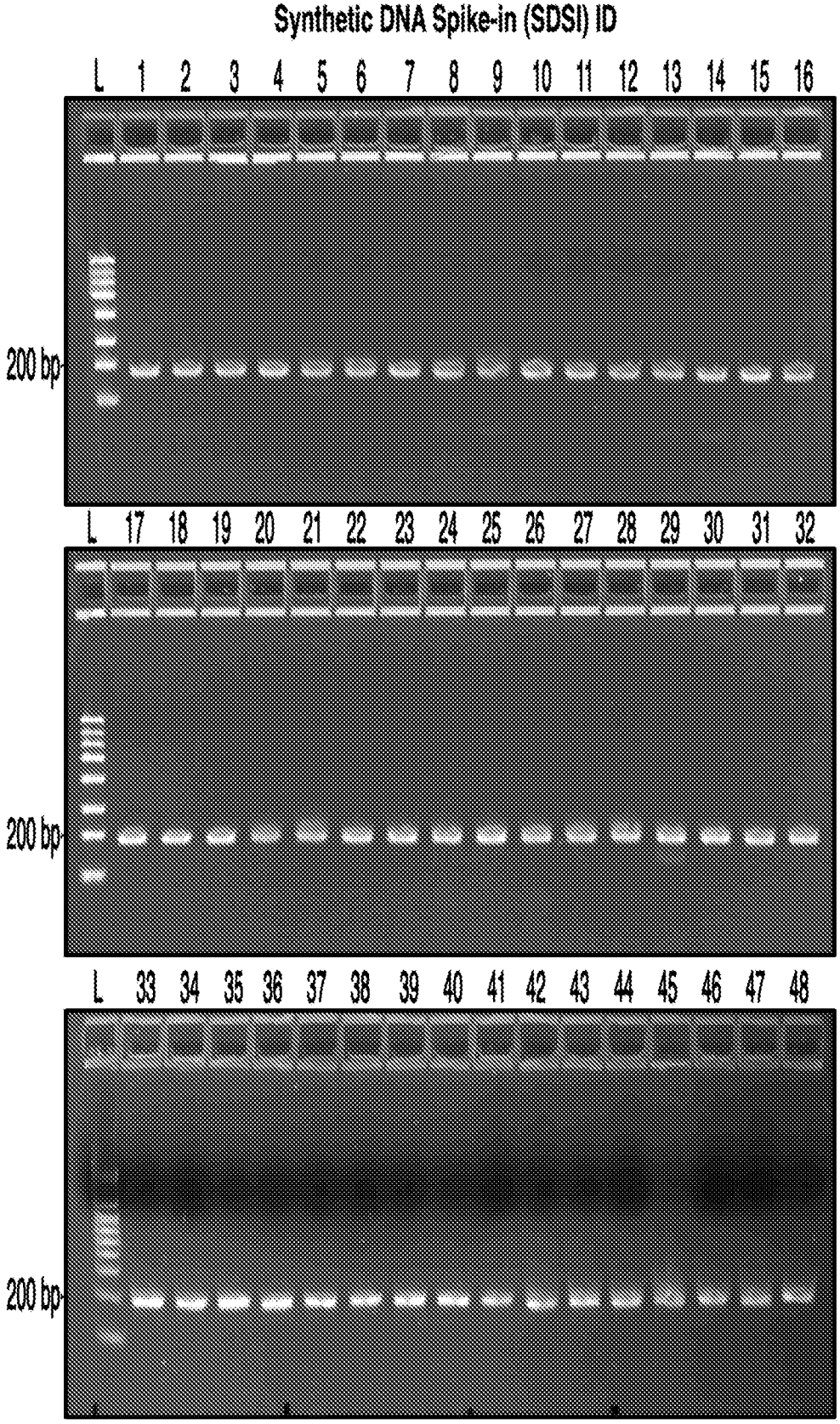

Synthetic DNA Spike-in (SDSI) ID
L 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
200 bp
L 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
200 bp
L 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
200 bp

FIG. 19D

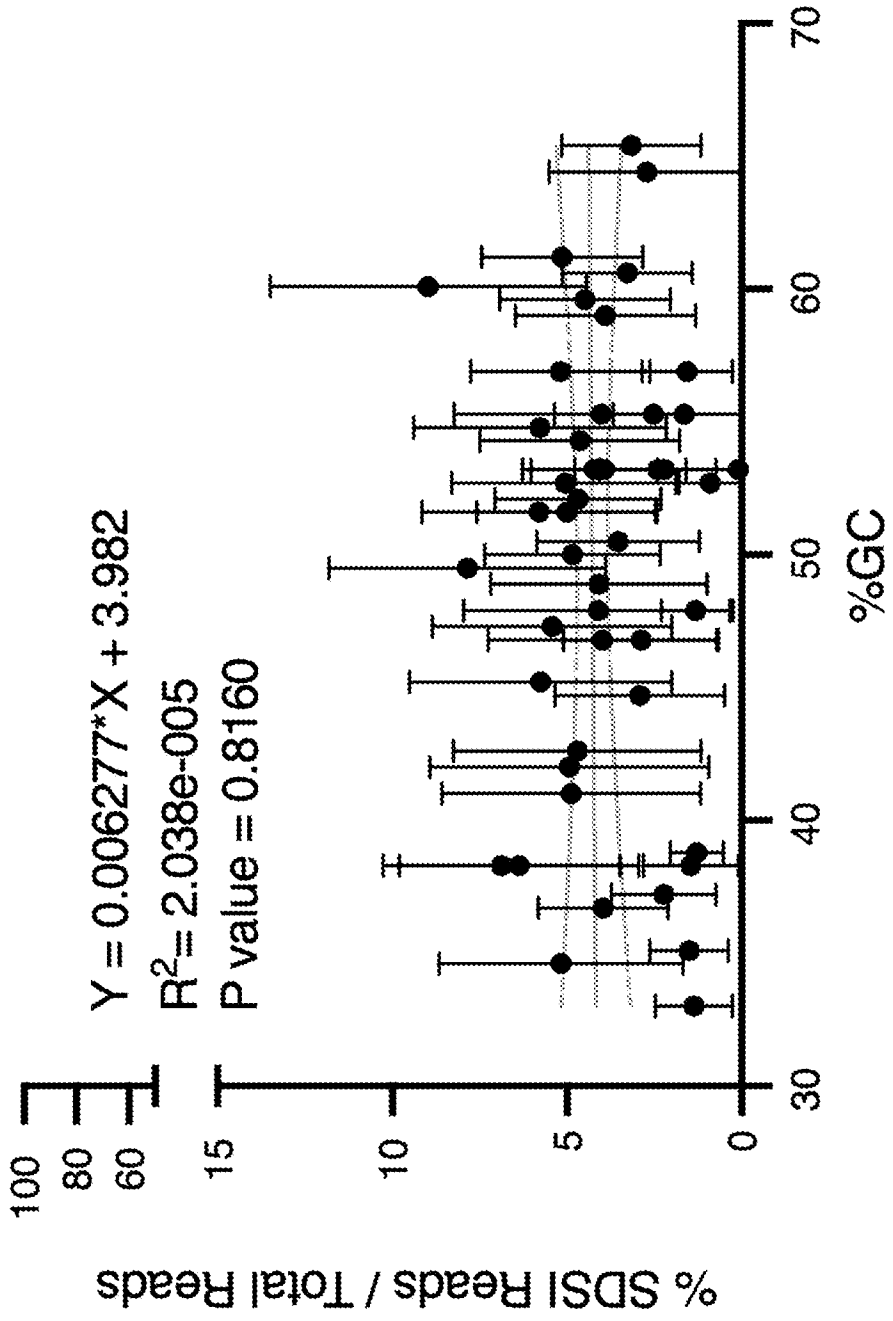

% SDSI Reads / Total Reads
100
80
60
15
10
5
0
30
40
50
60
70
%GC
Y = 0.006277*X + 3.982
R2= 2.038e-005
P value = 0.8160

FIG. 19E

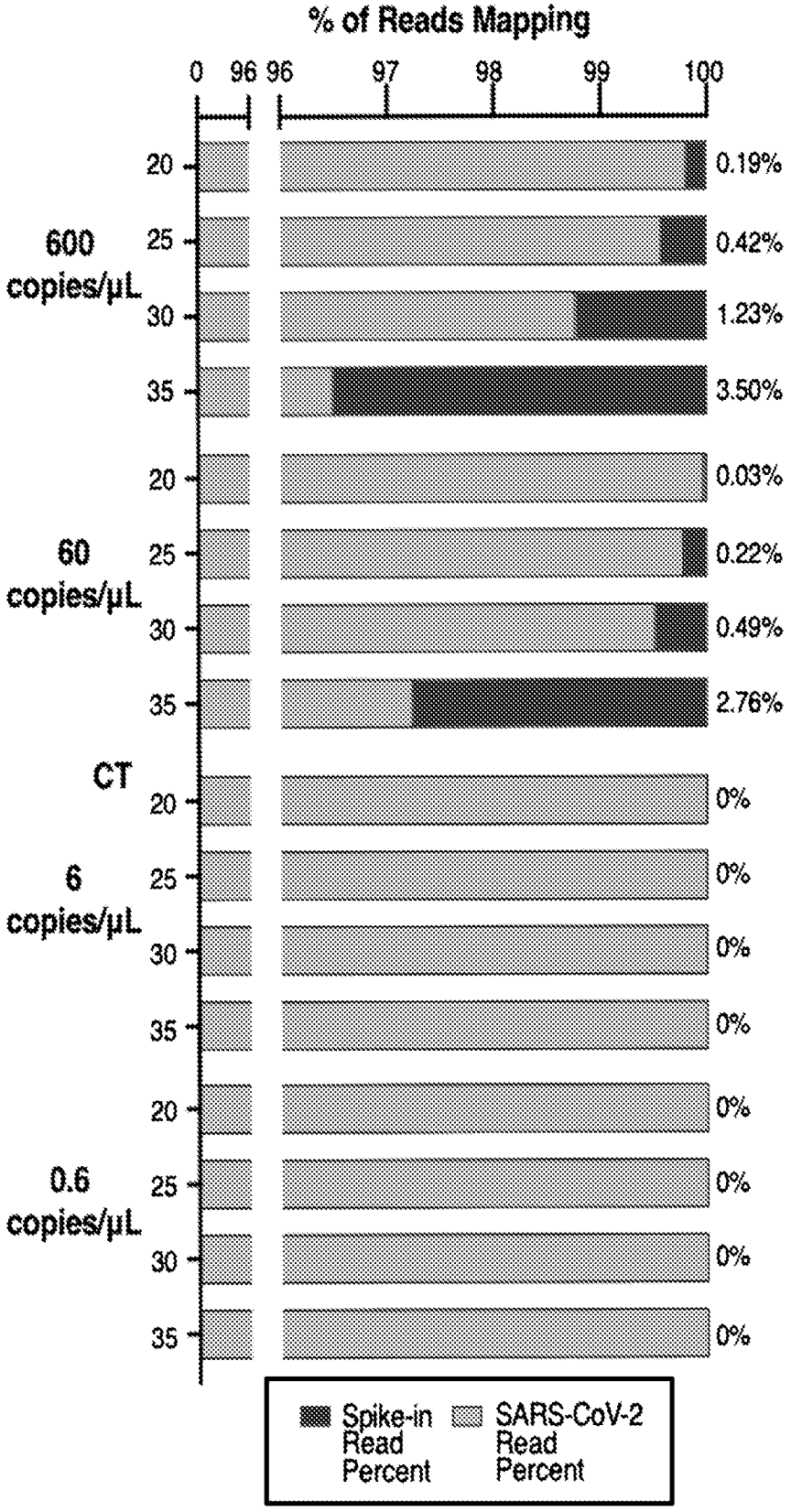

% of Reads Mapping
0
96
96
97
98
99
100
600 copies/µL
60 copies/µL
CT
6 copies/µL
0.6 copies/µL
20
25
30
35
0.19%
0.42%
1.23%
3.50%
0.03%
0.22%
0.49%
2.76%
0%
Spike-in Read Percent
SARS-CoV-2 Read Percent

FIG. 20A

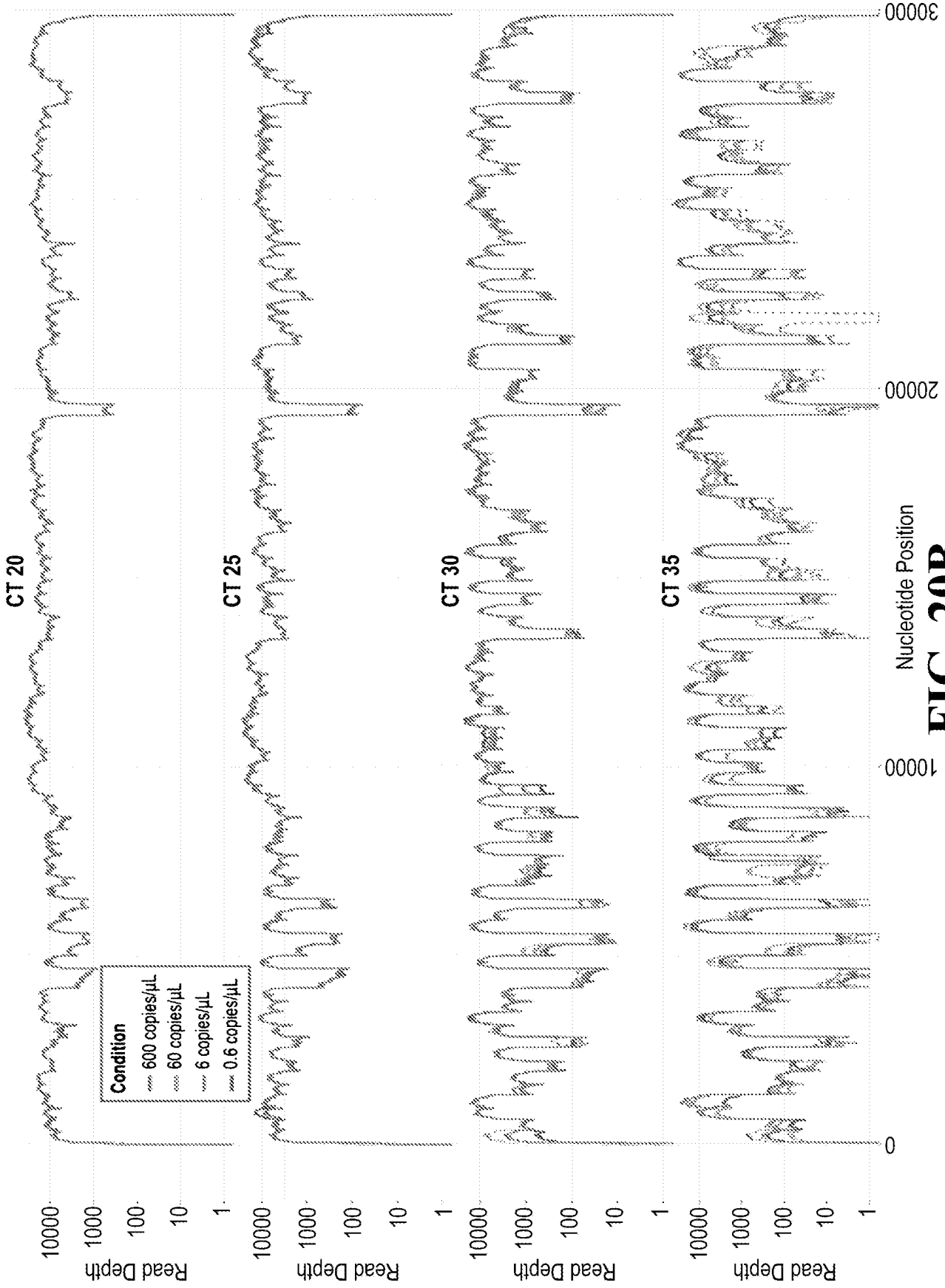

CT 20
CT 25
CT 30
CT 35
Read Depth
10000
1000
100
10
1
Condition
600 copies/µL
60 copies/µL
6 copies/µL
0.6 copies/µL
0
10000
20000
30000
Nucleotide Position

FIG. 20B

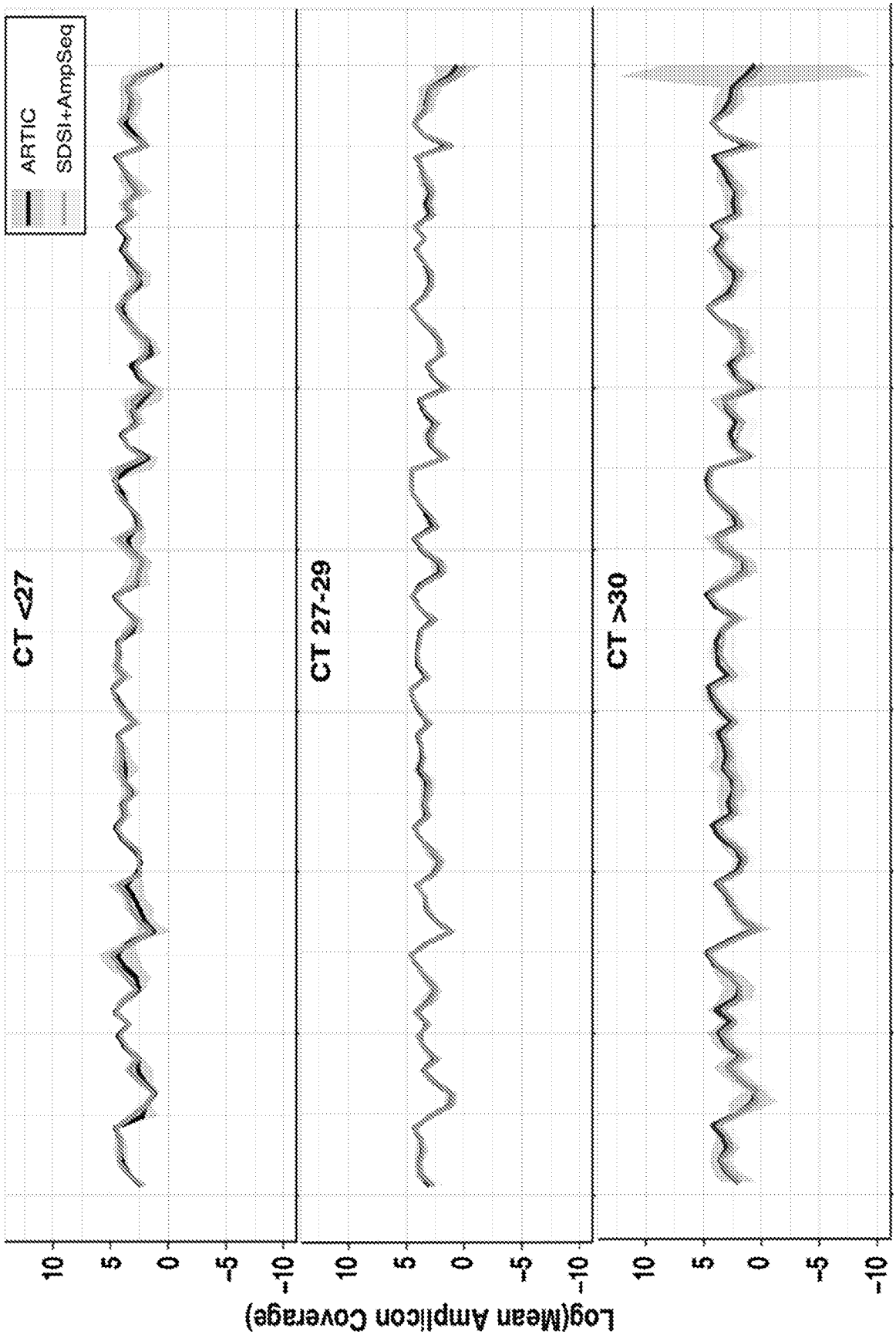

ARTIC
SDSI+AmpSeq
CT <27
CT 27-29
CT >30
Log(Mean Amplicon Coverage)
10
5
0
−5
−10
0
14
28
42
56
70
84
98
Amplicon

FIG. 21A

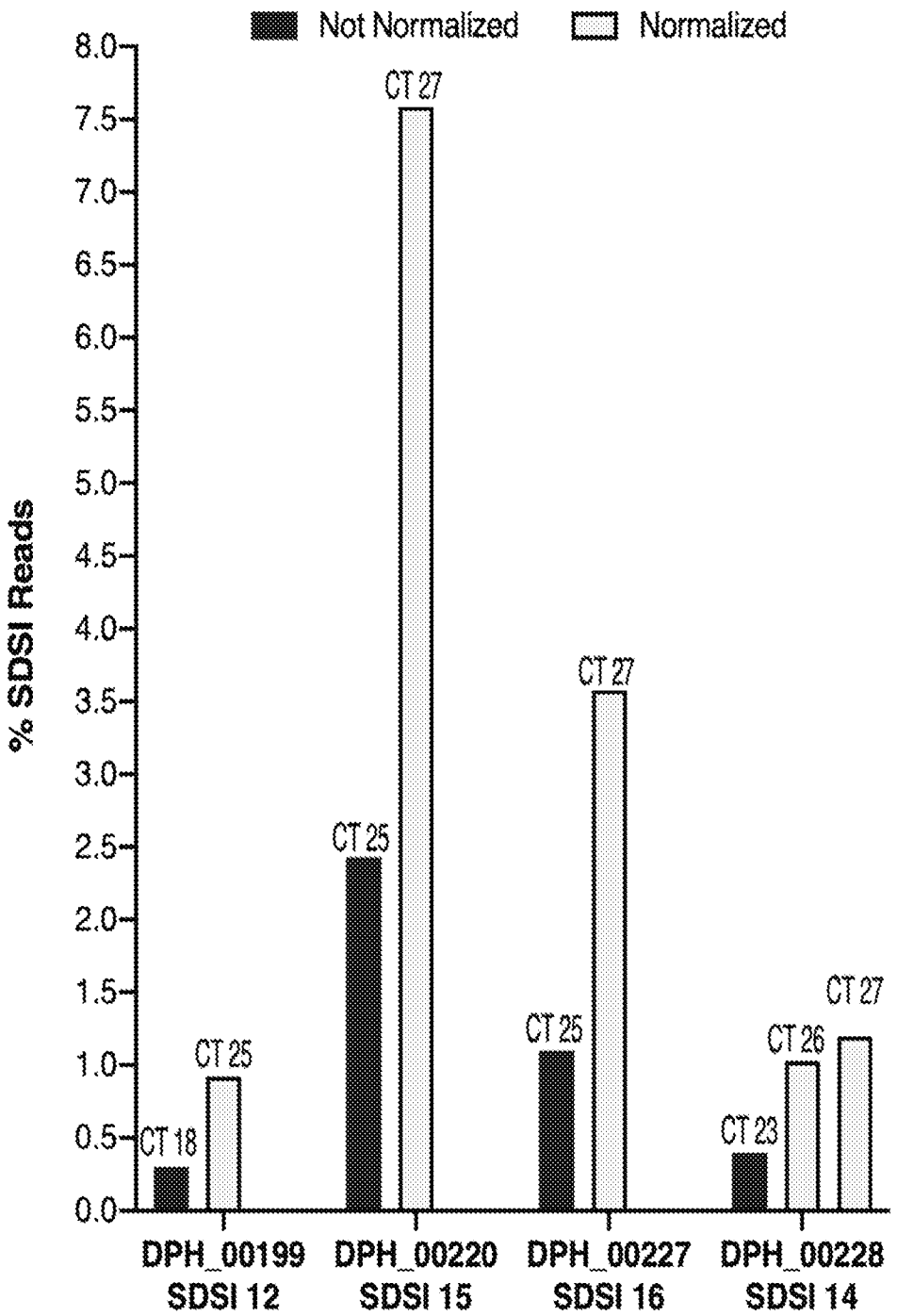

Not Normalized
Normalized
% SDSI Reads
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
CT 18
CT 25
CT 25
CT 27
CT 25
CT 27
CT 23
CT 26
CT 27
DPH_00199
SDSI 12
DPH_00220
SDSI 15
DPH_00227
SDSI 16
DPH_00228
SDSI 14

FIG. 21B

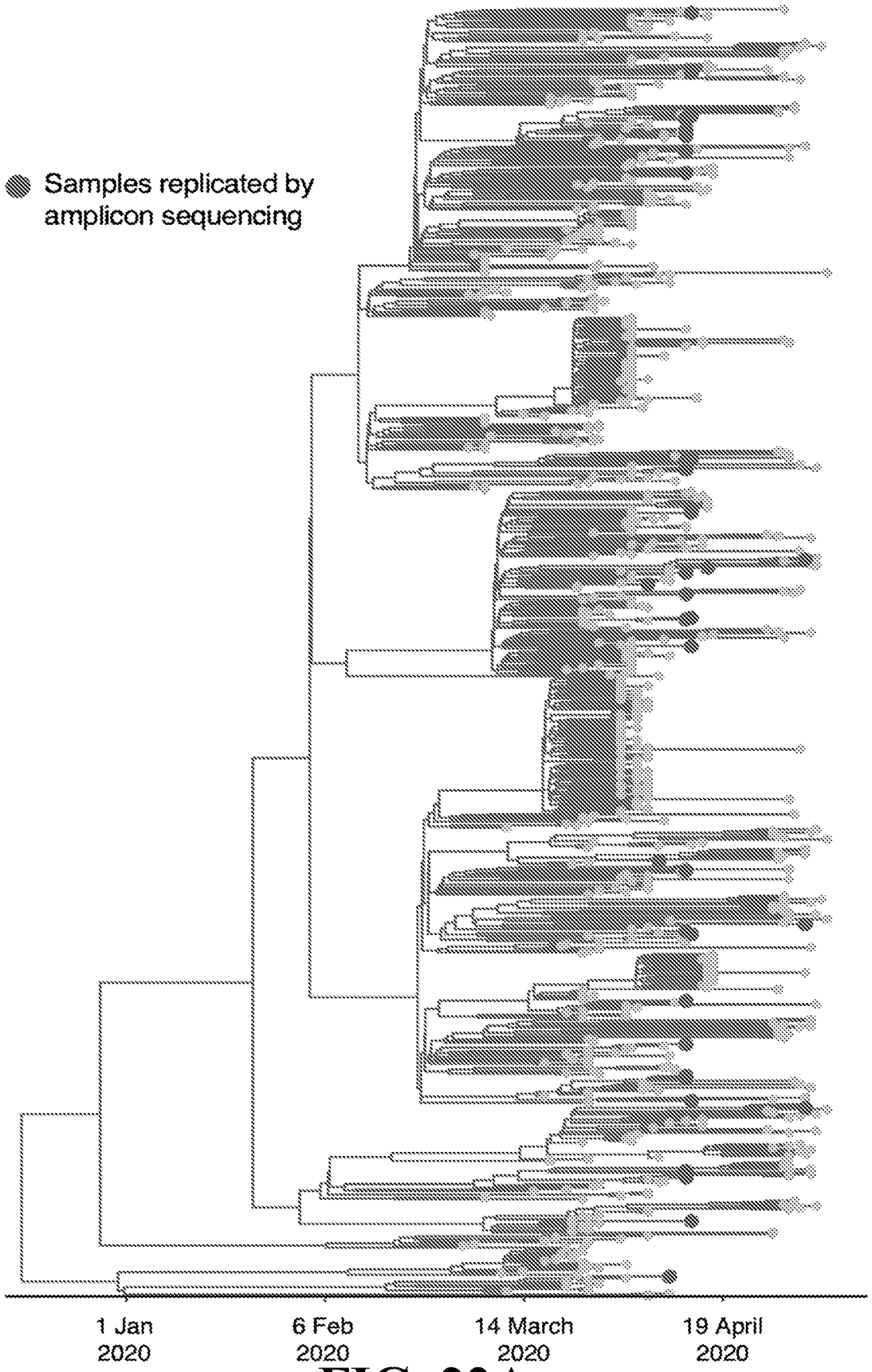

Samples replicated by amplicon sequencing
1 Jan 2020
6 Feb 2020
14 March 2020
19 April 2020

FIG. 22A

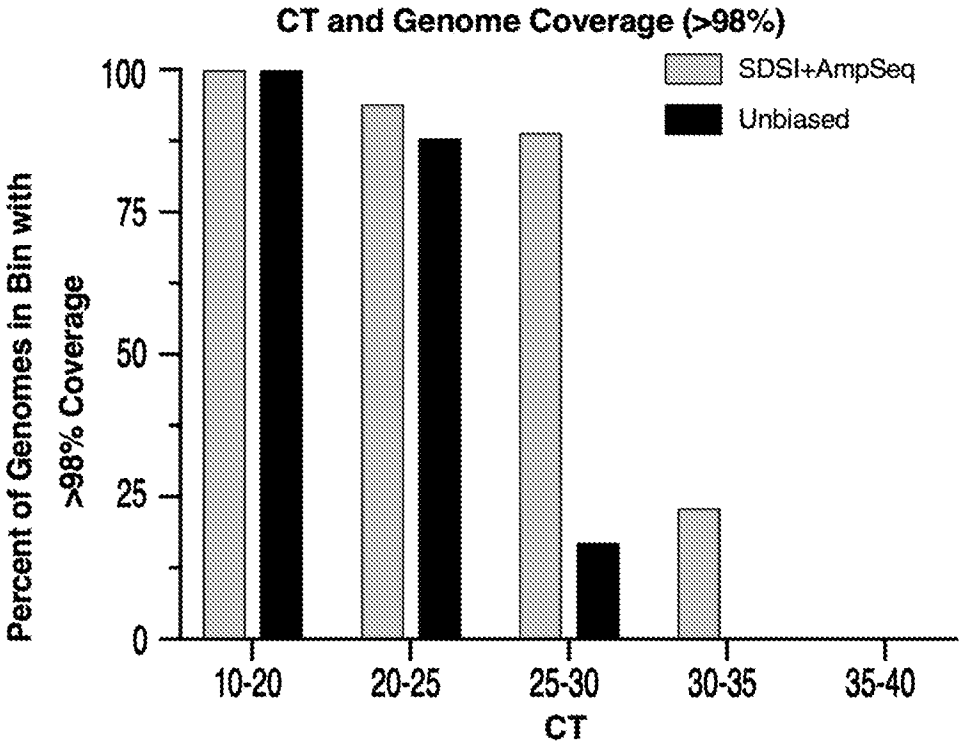


FIG. 22B

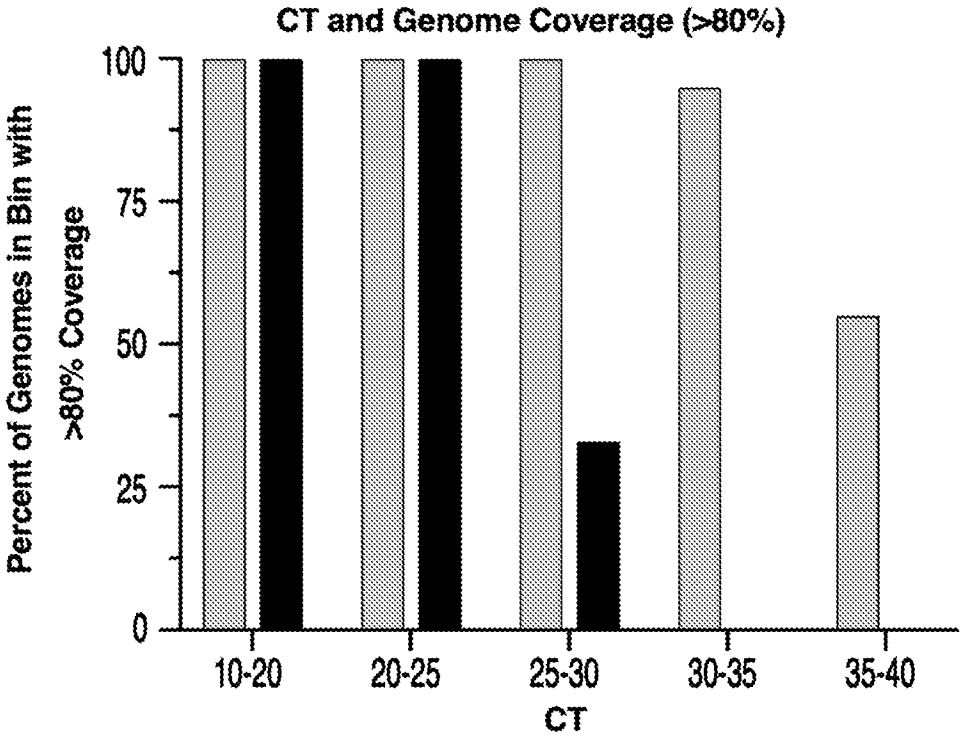


FIG. 22C

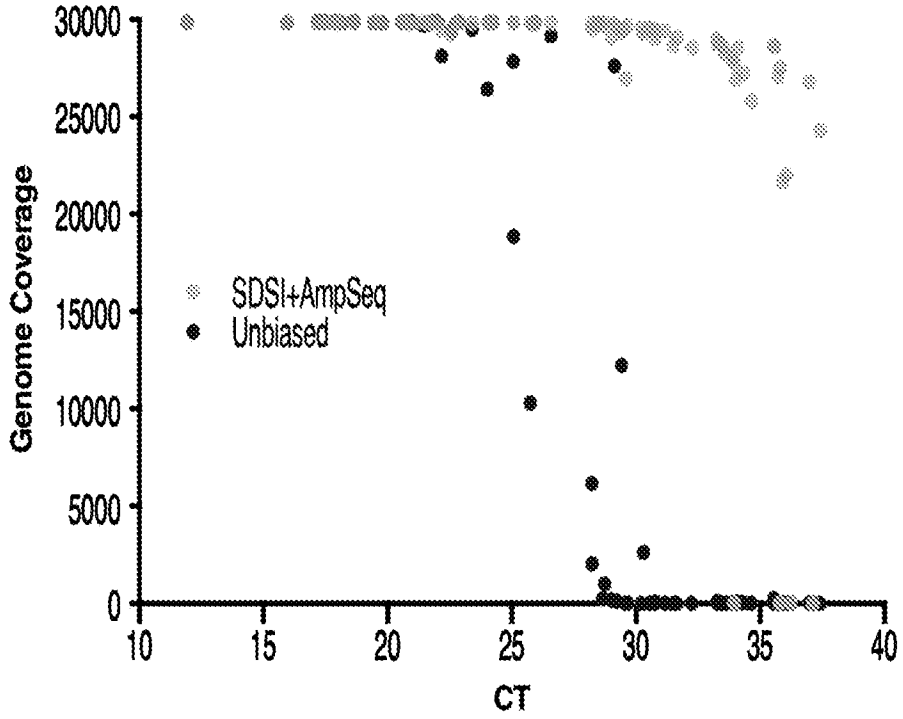


FIG. 22D

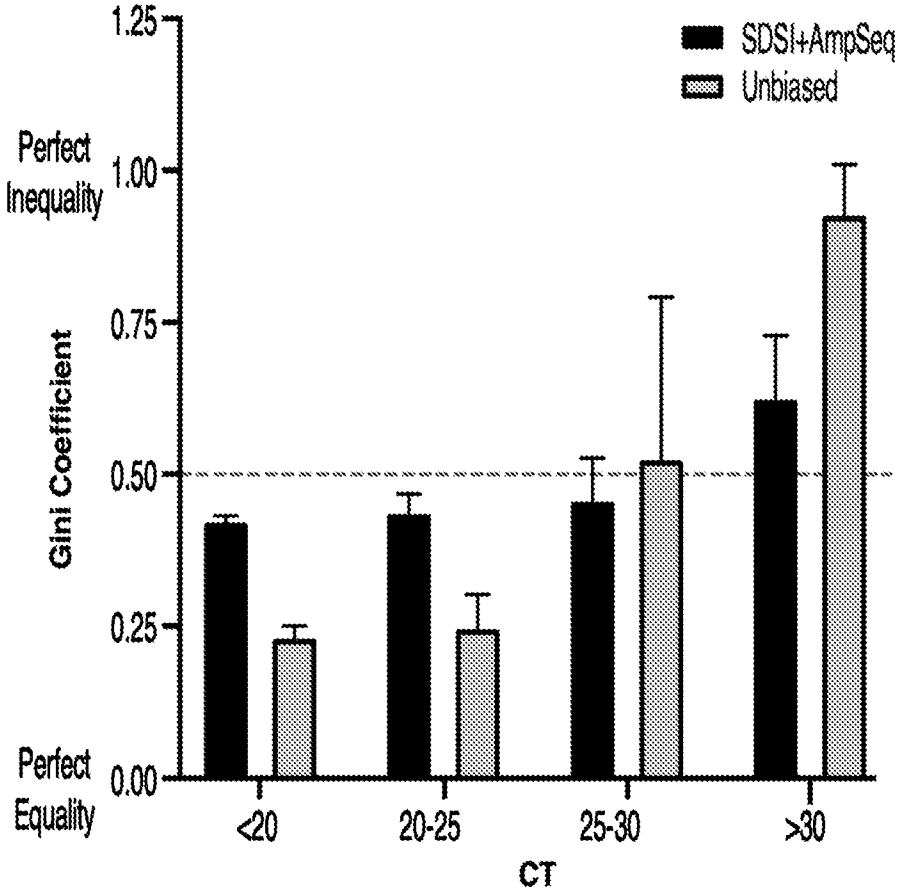


FIG. 22E

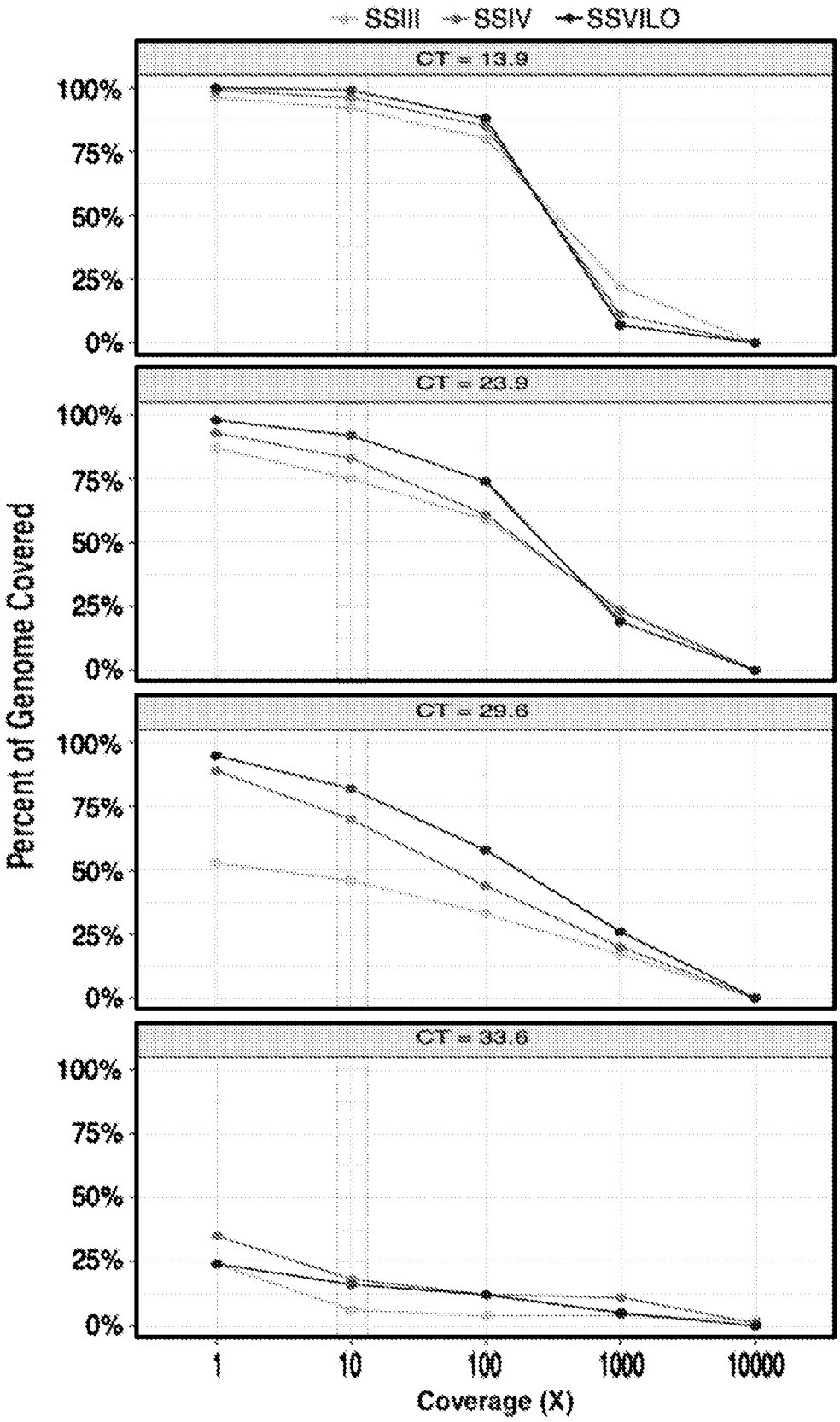

SSIII
SSIV
SSVILO
CT = 13.9
CT = 23.9
CT = 29.6
CT = 33.6
100%
75%
50%
25%
0%
Percent of Genome Covered
1
10
100
1000
10000
Coverage (X)

FIG. 23A

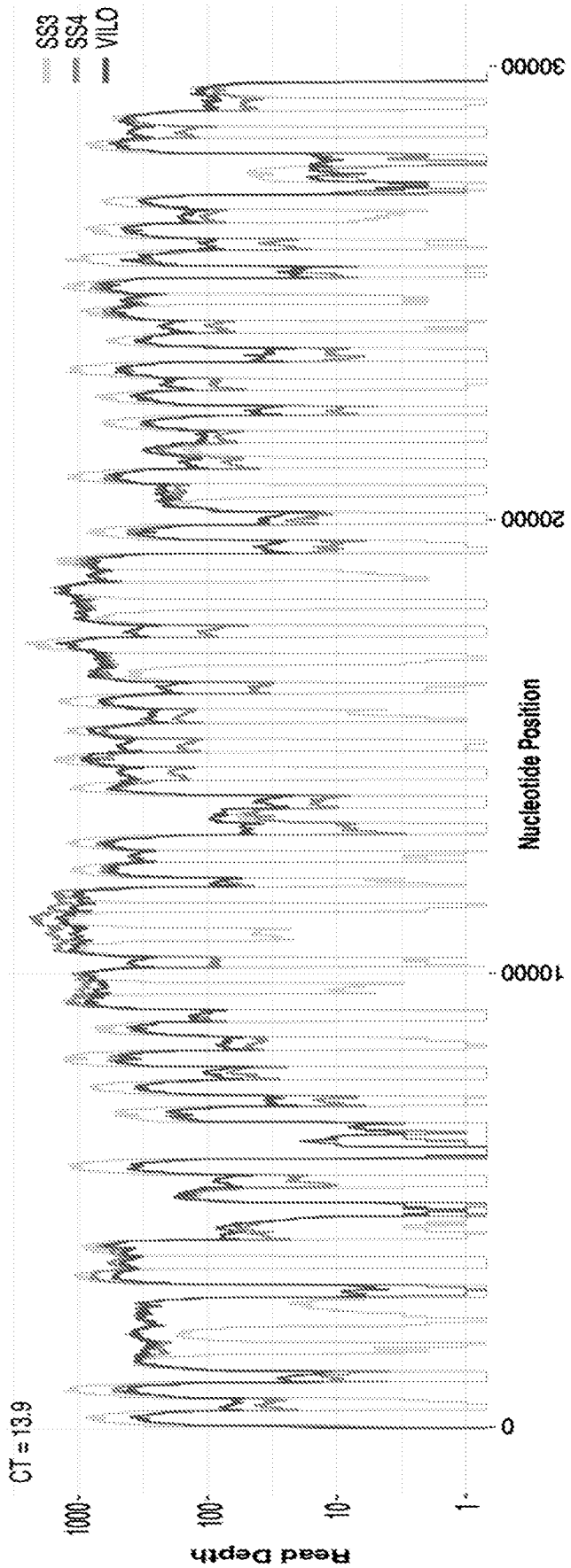

SS3
SS4
VILO
30000
20000
10000
0
Nucleotide Position
CT = 13.9
1000
100
10
1
Read Depth

FIG. 23B

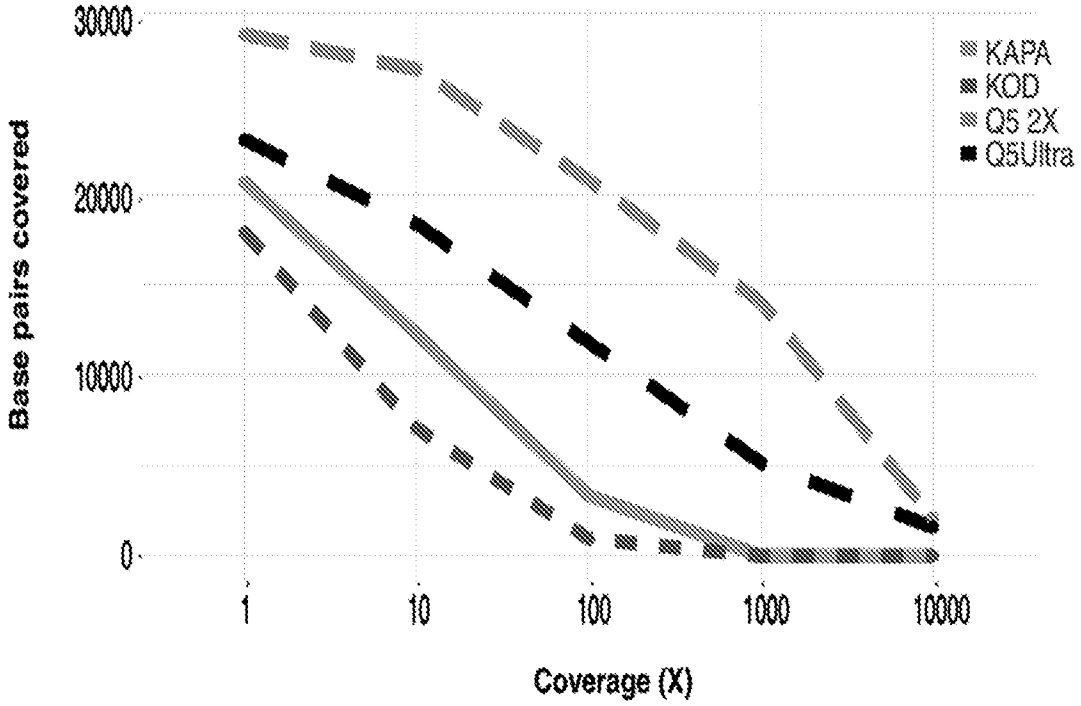


FIG. 23C

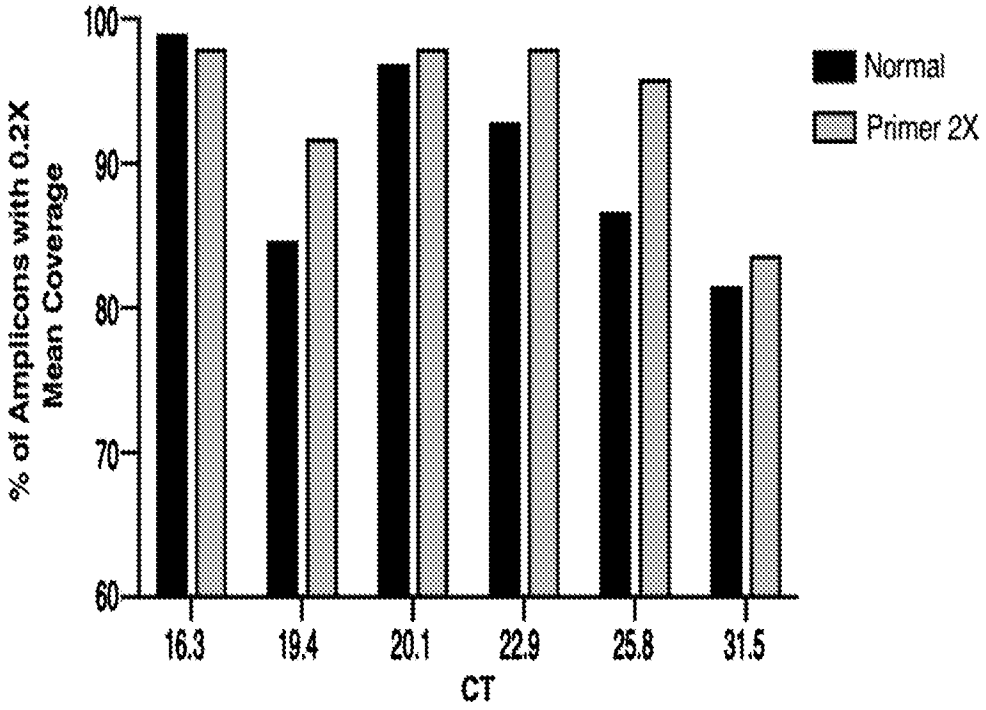


FIG. 23D

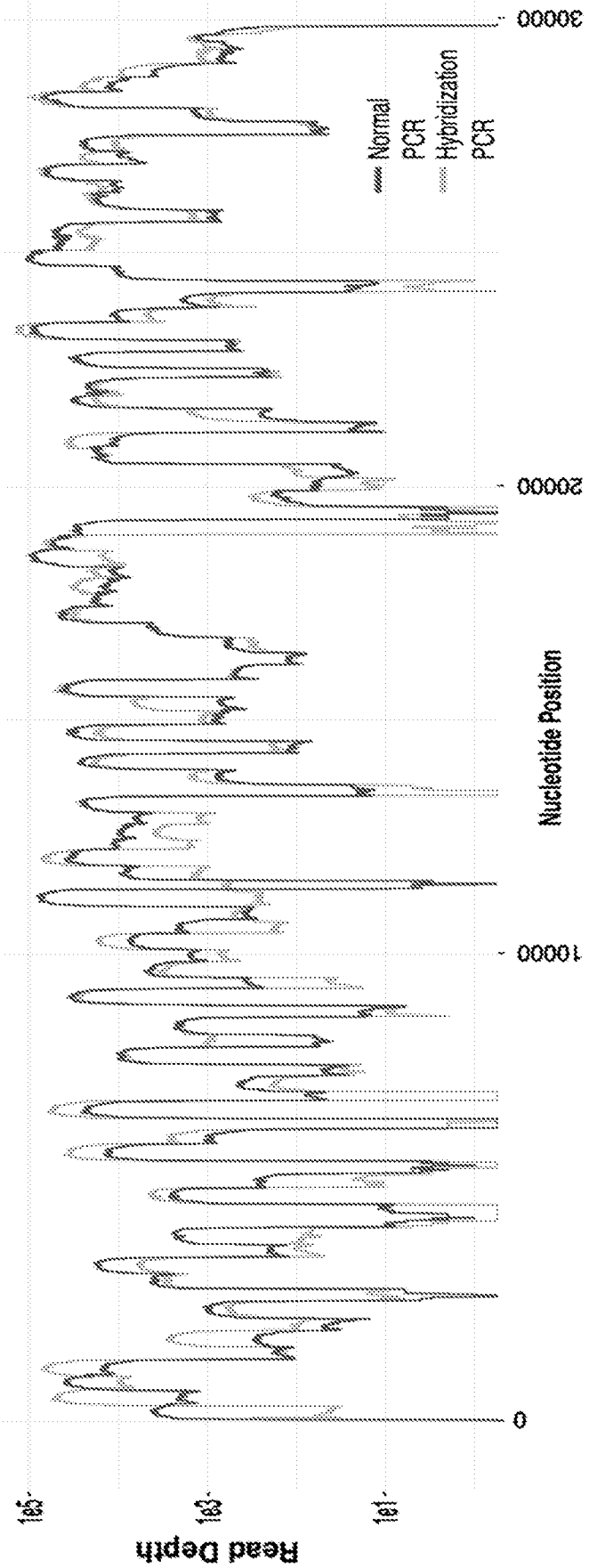

Normal
PCR
Hybridization
PCR
30000
20000
10000
0
Nucleotide Position
1e5
1e3
1e1
Read Depth

FIG. 23E

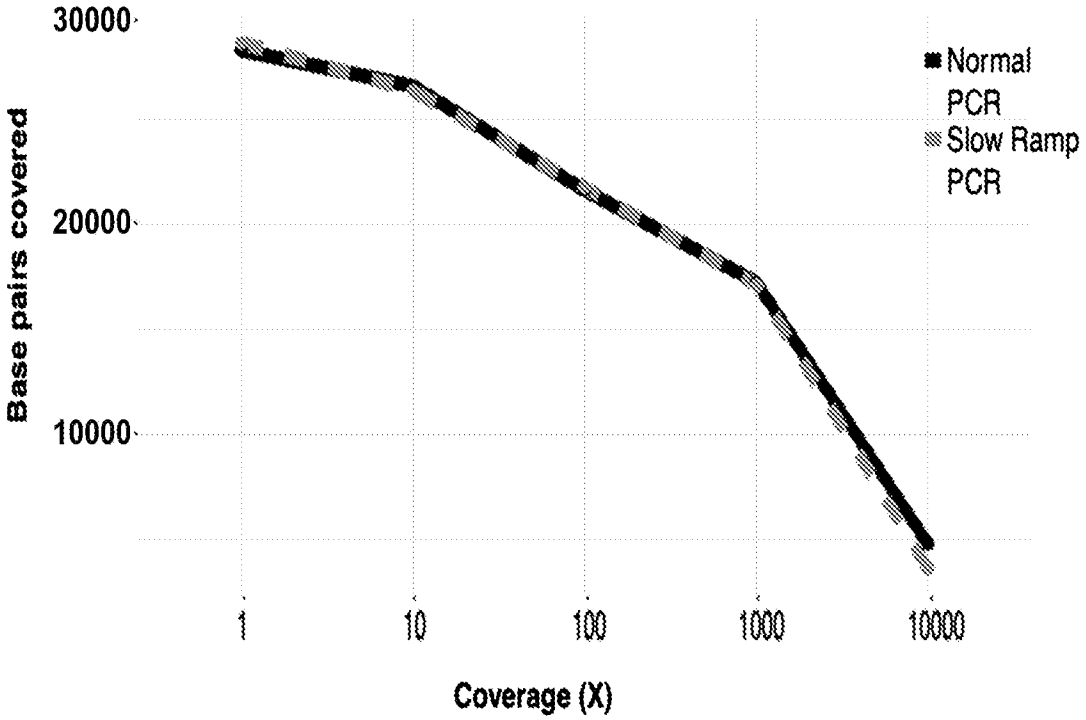


FIG. 23F

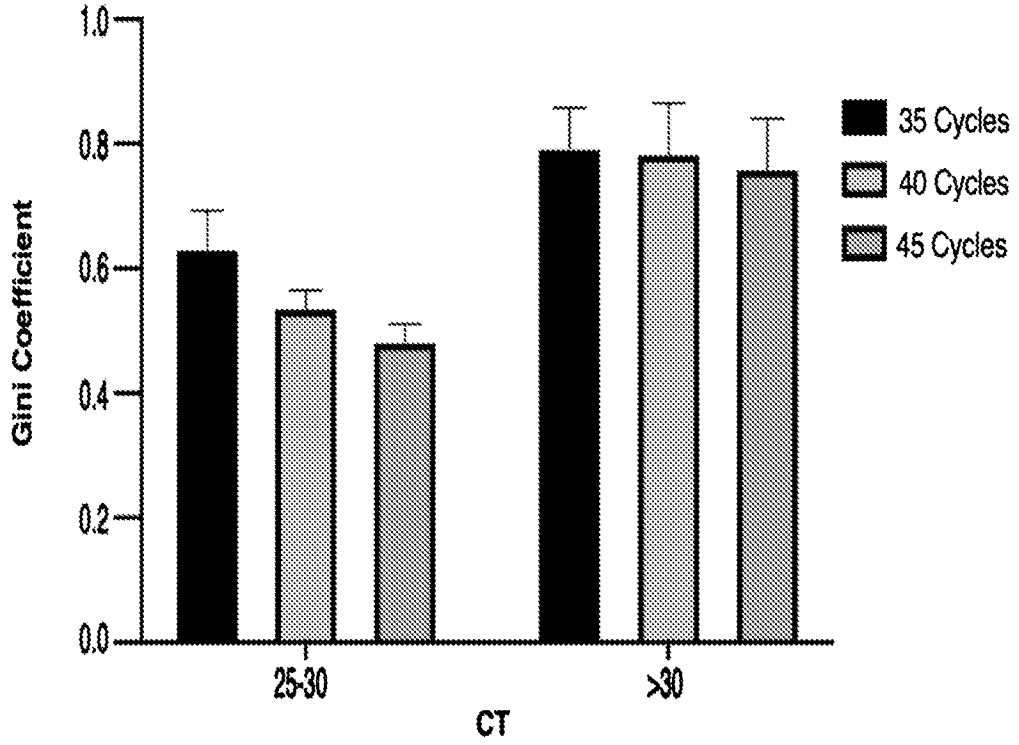


FIG. 23G

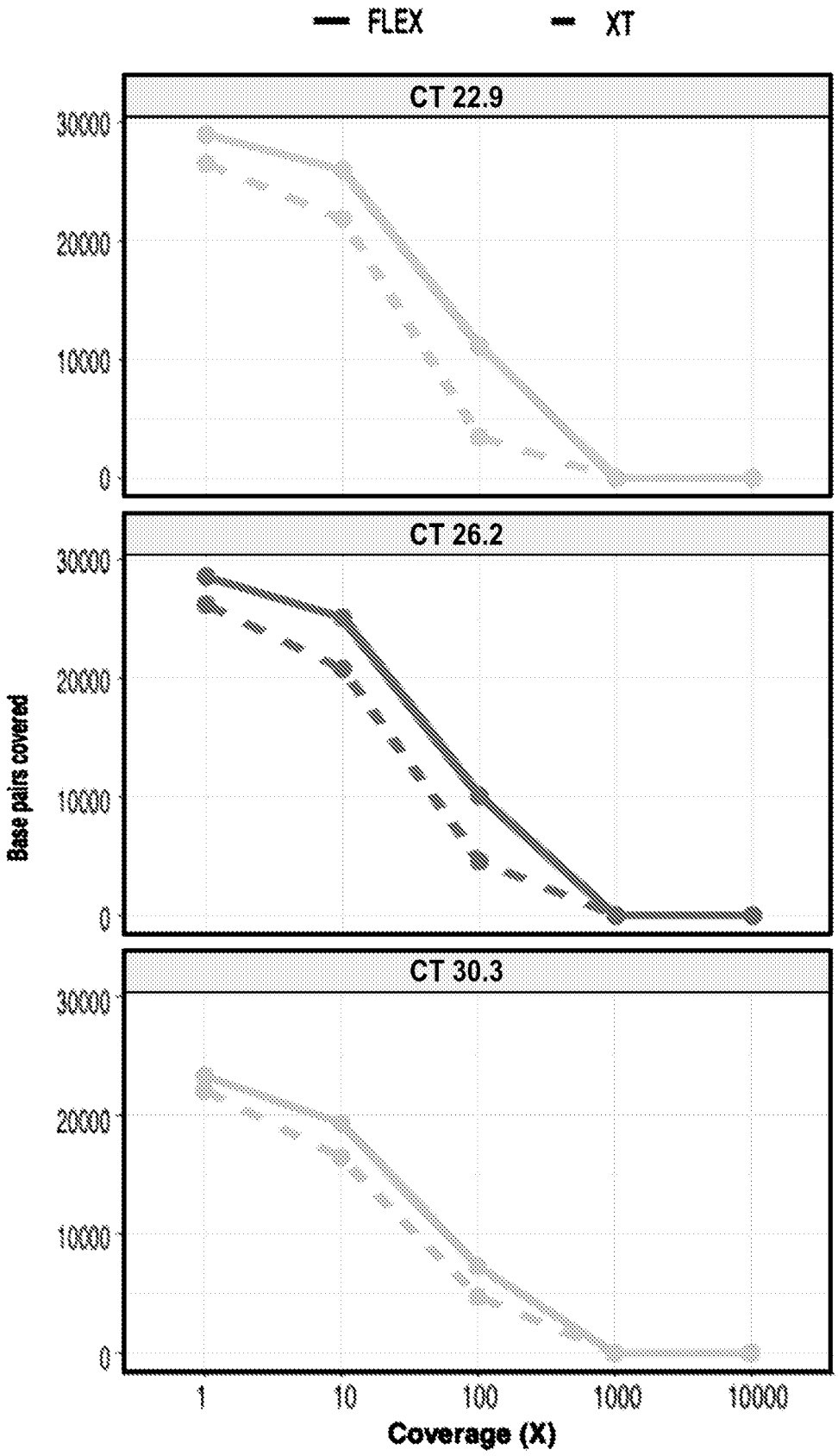

FLEX
XT
CT 22.9
CT 26.2
CT 30.3
30000
20000
10000
0
Base pairs covered
1
10
100
1000
10000
Coverage (X)

FIG. 23H

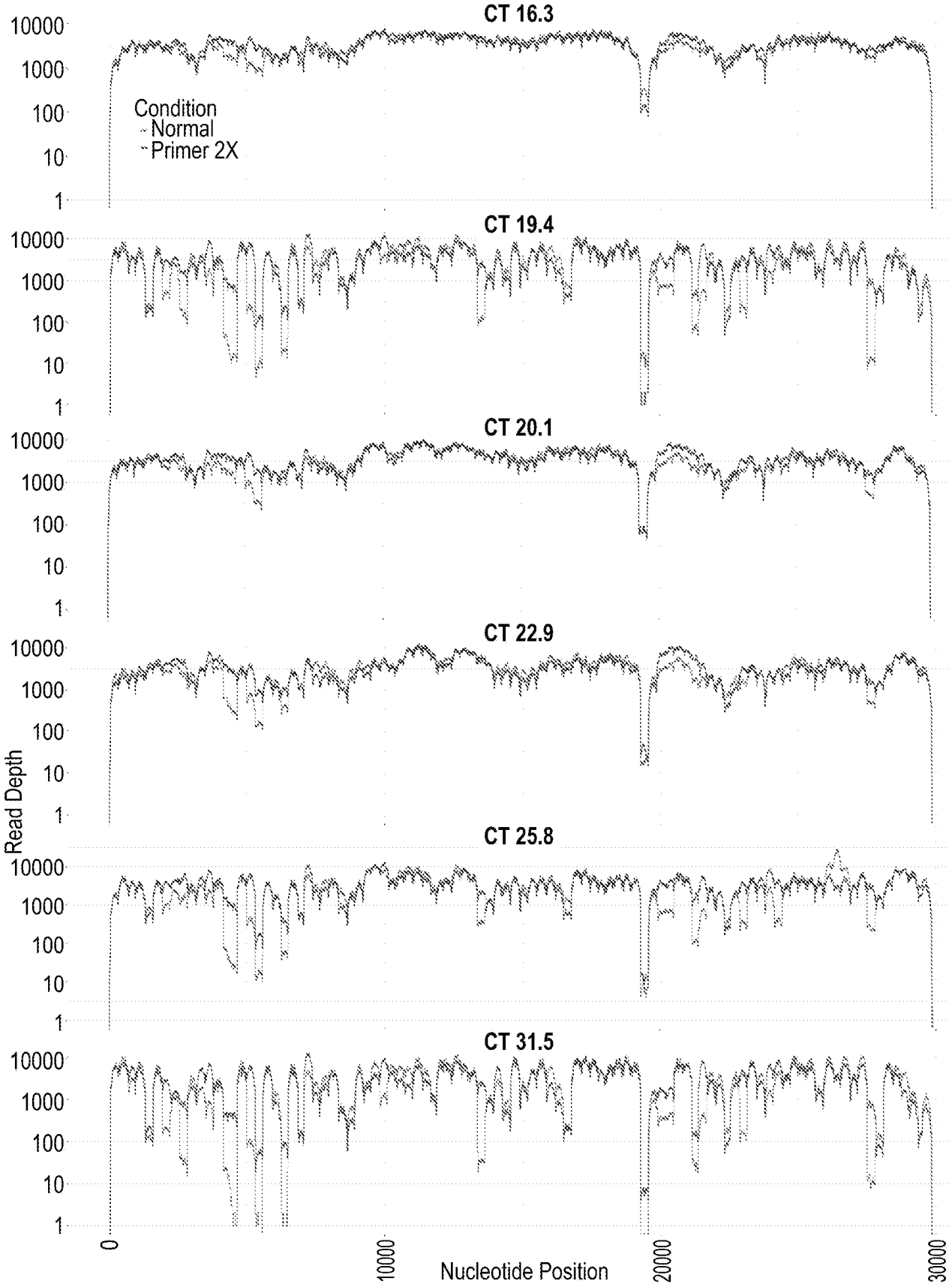

CT 16.3
CT 19.4
CT 20.1
CT 22.9
CT 25.8
CT 31.5
Condition
Normal
Primer 2X
Read Depth
10000
1000
100
10
1
0
10000
20000
30000
Nucleotide Position

FIG. 24

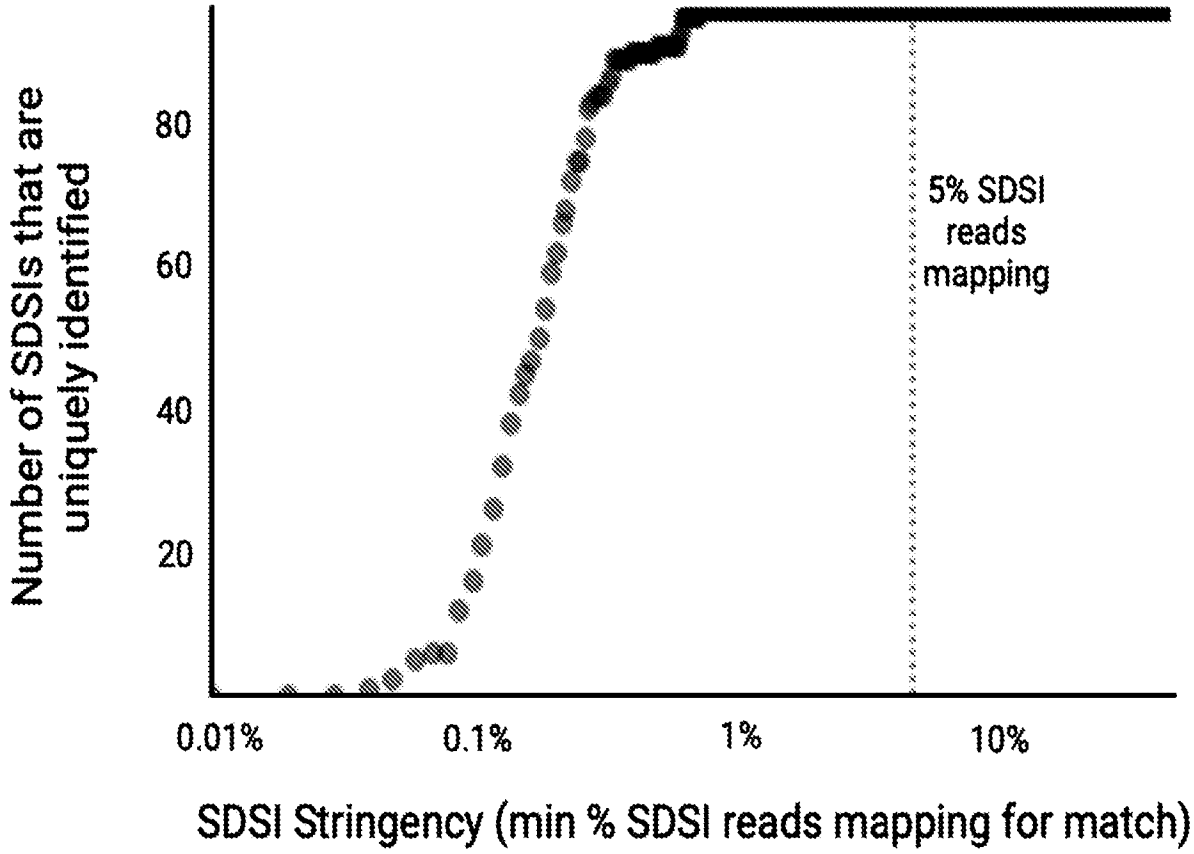

Number of SDSIs that are uniquely identified
80
60
40
20
5% SDSI reads mapping
0.01%
0.1%
1%
10%
SDSI Stringency (min % SDSI reads mapping for match)

FIG. 25

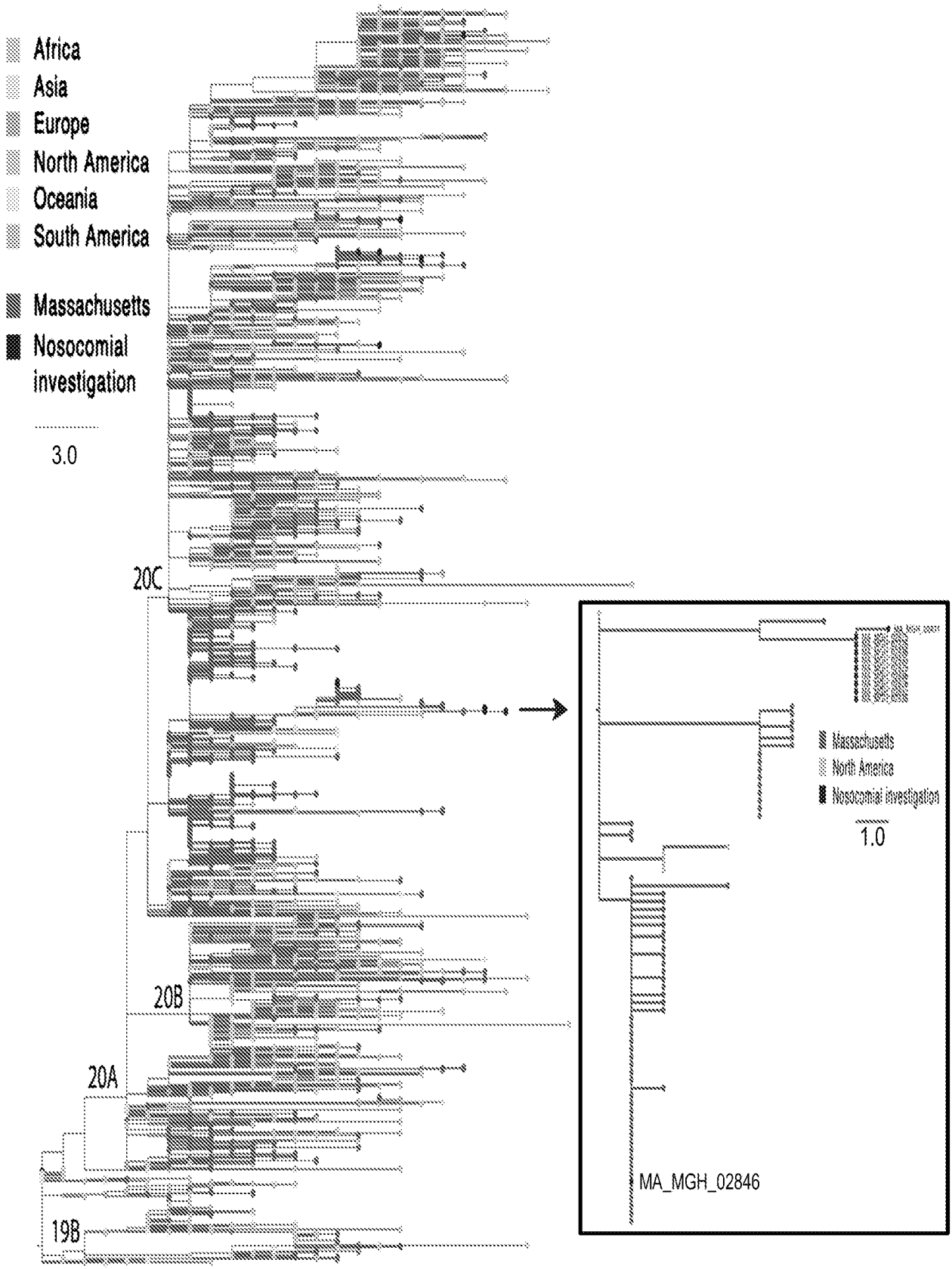

Africa
Asia
Europe
North America
Oceania
South America
Massachusetts
Nosocomial investigation
3.0
20C
20B
20A
19B
Massachusetts
North America
Nosocomial investigation
1.0
MA_MGH_02846

FIG. 26

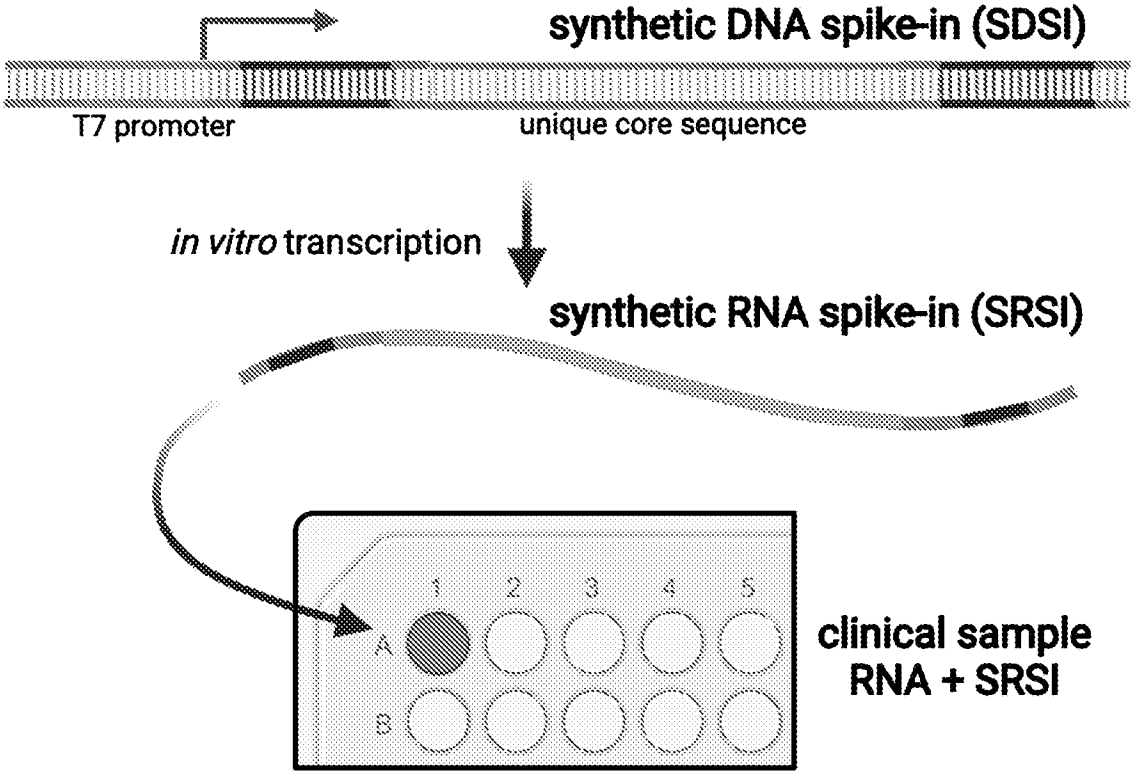

synthetic DNA spike-in (SDSI)
T7 promoter
unique core sequence
in vitro transcription
synthetic RNA spike-in (SRSI)
1
2
3
4
5
A
B
clinical sample
RNA + SRSI

FIG. 27A

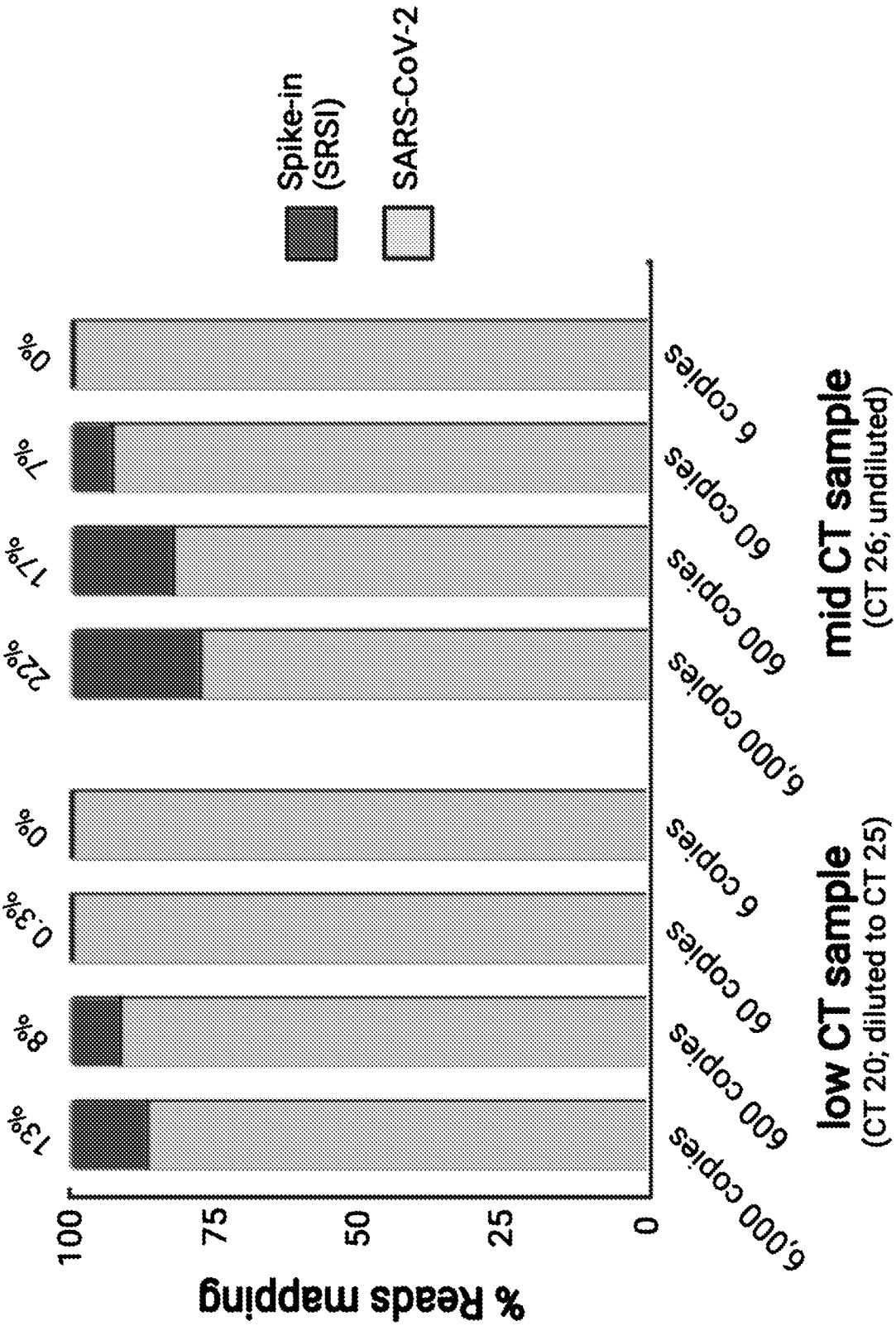

Spike-in (SRSI)
SARS-CoV-2
% Reads mapping
100
75
50
25
0
13%
8%
0.3%
0%
22%
17%
7%
0%
6,000 copies
600 copies
60 copies
6 copies
6,000 copies
600 copies
60 copies
6 copies
low CT sample
(CT 20; diluted to CT 25)
mid CT sample
(CT 26; undiluted)

FIG. 27B

AMPLICON-BASED SEQUENCING USING DNA SPIKE-INS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application Nos. 63/155,258, filed March 1, 2021, and 63/273,117, filed October 10, 2021. The entire contents of the above-identified applications are hereby fully incorporated herein by reference.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

[0002] This invention was made with government support under Grant Nos. AI110818, AI147868, HG010669, and CK000490 awarded by the National Institutes of Health, Grant No. 223-101-8101 awarded by the United States Food and Drug Administration, and Grant No. 75D30120C09605 awarded by the Centers for Diseases Control. The government has certain rights in the invention.

REFERENCE TO AN ELECTRONIC SEQUENCE LISTING

[0003] The contents of the electronic sequence listing ("BROD-5360US_ST25.txt"; Size is 205,235 bytes and it was created on February 17, 2022) is herein incorporated by reference in its entirety.

TECHNICAL FIELD

[0004] The subject matter disclosed herein is generally directed to synthetic DNA spike-ins and their use for detecting, quantifying, and preventing amplification contamination in genome profiling analysis.

BACKGROUND

[0005] The COVID-19 pandemic has demonstrated, once again, the crucial role of genomic sequencing in combatting infectious disease outbreaks globally. Monitoring the emergence of pathogens and the spread of variants of concern has become commonplace in government, academic, and private laboratories[1,2]. Genomics data provides insights into the diversity, evolution and transmission of a virus, a critical guide for public health interventions ranging from contact tracing, identifying cases of reinfection, or documenting resistance to clinical interventions[3–6]. In

the year since, genomic data have provided new insights into the diversity, evolution and transmission of the virus, which has increasingly been used to guide impactful public health interventions. In particular, scientists have employed viral genome sequencing to characterize the fine-scale epidemiology of clusters and superspreading events (Lemieux et al., 2021, Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events, *Science*, 371(6529); Popa et al., 2020, Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2, *Science Translational Medicine*, 12(573); Volz et al., 2021, Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data, *bioRxiv*, *medRxiv*). More recently, genome sequencing to monitor the emergence of new lineages and the spread of variants of concern (VoC) has become paramount (Washington et al., 2021, Genomic epidemiology identifies emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States, *medRxiv*). As laboratories are now performing viral genomic sequencing on SARS-CoV-2 at an unprecedented scale[7,8], it highlights the need for stringent requirements to ensure the integrity of genomes being produced.

[0006] Multiplexed amplicon-based genome sequencing methods have accelerated the massive scale of SARS-CoV-2 genomic surveillance due to their improved sensitivity, cost, and speed over other, lower-amplification RNA sequencing approaches, such as unbiased metagenomic sequencing[9].Unsurprisingly, amplicon-based approaches that target the SARS-CoV-2 genome for amplification and subsequent sequencing have become the genomic surveillance method of choice during the ongoing pandemic (over 90% of Short Read Archive submissions). In just a year since the first genome sequence enabled the identification of SARS-CoV-2, hundreds of thousands of complete genomes have been sequenced and released by a relatively small group of several hundred laboratories. An open-access tiled primer set developed by the ARTIC network (artic.network/) is the most widely used method for SARS-CoV-2 specific genome amplification followed by sequencing on either Illumina or nanopore instruments (Quick et al., 2017; Tyson et al., 2020). A wide array of protocols and publications are now available that integrate these ARTIC primers with different amplification and library construction indexing strategies (Baker et al., 2020; Gohl et al., 2020). Approaches such as batching samples by viral load to increase sensitivity are impractical to scale to current needs, resulting in incomplete recovery of viral genomes, especially from low titer samples.

[0007] However, the risk for contamination during the amplification stage is especially high as the 35 or more cycles of virus-specific PCR produces trillions of SARS-CoV-2 amplicons in a single reaction. Other high-risk modes of contamination, including sample swaps, cross-contamination of samples, or aerosolization, can occur throughout the sample processing pipeline. With many laboratories performing viral sequencing by processing multiple large batches in parallel, the potential for contamination increases[10]. Even small amounts of sample mixing or contaminating amplicons could potentially confound studies where viral detection is sensitive to only tens of molecules[10,11]. Moreover, as SARS-CoV-2 has relatively low genetic diversity and often spreads in local outbreaks or clusters[11,12], many genomes are expected to be identical at the consensus level[11,15–17], a pattern that could also be observed due to contamination. The risk of contamination, and the challenges in detecting it, can confound a wide array of genomic analyses including estimates of the frequencies of variants, lineage dynamics, and transmission events. Additionally, methods to address the critical risk of sample processing errors in clinical sequencing could enable its use more widely in clinical decision making.

[0008] To meet the genomic surveillance goals laid out by local and world governments, sequencing efforts will need to be scaled to thousands of centers, many performing viral genomics for the first time. Additional laboratories will enter the SARS-CoV-2 sequencing space with an emphasis to rapidly surveil VoCs for clinical significance, with even higher requirements to ensure the integrity of SARS-CoV-2 genomes being produced. While inclusion of internal standards is commonplace in many experimental approaches[13–15] and some technical assay controls exist for DNA sequencing[16–18], the use of internal controls is currently rare in amplicon-based genomic surveillance. Here Applicants developed and extensively tested a sample identification method using 96 synthetic DNA spike-ins (SDSIs) for amplicon-based sequencing approaches. Using the widely used open-access ARTIC tiled primer design (artic.network/), Applicants implemented these SDSIs for SARS-CoV-2 genomic sequencing from thousands of residual diagnostic (clinical) samples. The resulting user-friendly and highly versatile SDSI+AmpSeq protocol can be easily implemented to improve the quality of genomic data generated for epidemiological and clinical investigations of human pathogens (**FIG. 1 and FIG. 13, Table 6**).

[0009] Citation or identification of any document in this application is not an admission that such a document is available as prior art to the present invention.

SUMMARY

[0010] In one aspect, the present invention provides for a method of detecting and preventing contamination in one or more cDNA samples comprising adding a synthetic DNA spike-in (SDSI) to each cDNA sample, wherein each SDSI is capable of amplification simultaneously with the cDNA, and wherein each SDSI comprises a unique sequence capable of differentiating each SDSI; amplifying one or more of the cDNA samples and SDSI; sequencing the amplified sample; and determining the number of reads of the spike-in from the one or more samples. In certain example embodiments, the sample is associated with drug resistance. In certain example embodiments, the sample is for sequencing a pathogen or family of pathogens. In certain example embodiments, the pathogen is a virus. In certain example embodiments, the pathogen is a bacteria and the region sequenced is associated with antibiotic resistance. In certain example embodiments, each sample contains a viral nucleic acid sequence. In certain example embodiments, the samples are for creating one or more sequencing families/clusters.

[0011] In certain example embodiments, the SDSI contains a core region and a primer binding region at the 3' end and the 5' end. In certain example embodiments, the core sequence of the SDSI is derived from a rare organism. In certain example embodiments, the rare organism is a thermophilic archaea. In certain example embodiments, the core sequence homology is less than 65%, or less than 60%, or less than 55%, or less than 50%, or less than 45%, or less than 40%, or less than 35%, or less than 30%, or less than 25%, or less than 20%, or less than 15%, or less than 5%, or less than 1% to a sample sequence. In certain example embodiments, the core sequence homology is less than 15, or less than 20, or less than 25, or less than 30, or less than 35, or less than 40, or less than 45, or less than 50 contiguous bases in common with the sample sequence.

[0012] In certain example embodiments, the synthetic DNA spike-in sequences are 50 – 5000 nucleotides in length. In certain example embodiments, the SDSI minimizes self-hybridization and cross-hybridization with nucleic acids in the sample. In certain example embodiments, the primer binding sites of the SDSI have a Tm between 55-65 ^{o}C. In certain example embodiments, the method further comprises a plurality of SDSIs. In certain example embodiments, the core sequence of the synthetic DNA comprises a sequence as set forth in SEQ ID NOS: 1 – 96 and 193-291. In certain example embodiments, the primer binding sequences are complementary to the primers having SEQ ID NOS: 391 and 392. In certain example embodiments the SDSIs comprise one or

more of SEQ ID NOS: 97-192 and 292-390. In example embodiments, sequences can be used in the alternative. In one example embodiment, sequence SEQ ID NO: 289 can substitute for sequence SEQ ID NO: 16. In one example embodiment, sequence SEQ ID NO: 290 can substitute for sequence SEQ ID NO: 57. In one example embodiment, sequence SEQ ID NO: 291 can substitute for sequence SEQ ID NO: 66. In one example embodiment, sequence SEQ ID NO: 388 can substitute for sequence SEQ ID NO: 112. In one example embodiment, sequence SEQ ID NO: 389 can substitute for sequence SEQ ID NO: 153. In one example embodiment, sequence SEQ ID NO: 390 can substitute for sequence SEQ ID NO: 162. In one example embodiment, one or more of SEQ ID NOS: 16, 57, 66, 112, 153, and 162 can be substituted with their alternative sequence SEQ ID NOS: 289, 290, 291, 388, 389, and 390, respectively.

[0013] In certain example embodiments, the concentration of synthetic DNA spike-ins range from 0.1 femtomolar – 1.0 femtomolar. In certain example embodiments, the presence of an amplified spike-in corresponding to the spike-in added to a sample indicates a decreased risk of contamination. In certain example embodiments, the presence of an amplified spike-in corresponding to the spike-in not added to a sample indicates an increased risk of contamination.

[0014] In another aspect, the present invention is a set of synthetic DNA spike-ins (SDSIs), each SDSI in the set comprising a primer binding sequence at the 3' and 5' end and a unique core sequence between the 3' and 5' primer binding sequences. In certain example embodiments, the set comprises at least 96 spike-ins. In certain example embodiments, the unique core sequence is derived from a rare organism. In certain example embodiments, the rare organism is a thermophilic archaea. In certain example embodiments, the core sequence homology is less than 65%, or less than 60%, or less than 55%, or less than 50%, or less than 45%, or less than 40%, or less than 35%, or less than 30%, or less than 25%, or less than 20%, or less than 15%, or less than 5%, or less than 1% to a sample sequence. In certain example embodiments, the core sequence homology is less than 15, or less than 20, or less than 25, or less than 30, or less than 35, or less than 40, or less than 45, or less than 50 contiguous bases in common with the sample sequence.

[0015] In certain example embodiments, the sequence is 50-5000 nucleotides in length. In certain example embodiments, the SDSIs minimizes self-hybridization and cross-hybridization with nucleic acids in the sample. In certain example embodiments, the primer binding sites have a Tm between 55-65 ^{o}C. In certain example embodiments, the core sequence are the unique

sequences as set forth SEQ ID NOS: 1 – 96 and 193-291. In certain example embodiments, the primer binding sequences are complementary to the primers having SEQ ID NOS: 391 and 392. In certain example embodiments, the SDSIs comprise one or more of SEQ ID NOS: 97-192 and 292-390. In example embodiments, sequences can be used in the alternative. In one example embodiment, sequence SEQ ID NO: 289 can substitute for sequence SEQ ID NO: 16. In one example embodiment, sequence SEQ ID NO: 290 can substitute for sequence SEQ ID NO: 57. In one example embodiment, sequence SEQ ID NO: 291 can substitute for sequence SEQ ID NO: 66. In one example embodiment, sequence SEQ ID NO: 388 can substitute for sequence SEQ ID NO: 112. In one example embodiment, sequence SEQ ID NO: 389 can substitute for sequence SEQ ID NO: 153. In one example embodiment, sequence SEQ ID NO: 390 can substitute for sequence SEQ ID NO: 162. In one example embodiment, one or more of SEQ ID NOS: 16, 57, 66, 112, 153, and 162 can be substituted with their alternative sequence SEQ ID NOS: 289, 290, 291, 388, 389, and 390, respectively.

[0016] These and other aspects, objects, features, and advantages of the example embodiments will become apparent to those having ordinary skill in the art upon consideration of the following detailed description of example embodiments.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] An understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention may be utilized, and the accompanying drawings of which:

[0018] **FIG. 1 – SDSI-ARTIC Amplicon-Sequencing Protocol** – Illustrative workflow for 48 samples through the SDSI+ARTIC amplicon-sequencing pipeline. A synthetic DNA spike-ins (SDSI) will be added to each sample to allow for contamination tracking and accurate sample identification.

[0019] **FIG. 2A-2C – Synthetic DNA oligos spiked into amp-seq reactions flag contamination and sample swaps** – **A.** Schematic detailing SDSI design. Each oligo contains 140 bp of sui generis sequence flanked by unique primer binding sites. Primers designed to amplify SDSIs are added to ARTIC primer pools, and a unique SDSI is added to each clinical sample. Identification of multiple SDSIs in the same sample indicates contamination. **B.** In a titration of

SDSIs across clinical samples with variable CTs, the number of reads mapping to both SARS-CoV-2 and the SDSI were quantified, and the percentage of each was calculated. **C.** For each of 48 unique clinical samples (on the horizontal axis), reads mapping to each of 48 unique SDSIs (on the vertical axis) were quantified; the log of this read count is represented by the intensity of color displayed. Samples and SDSIs were ordered such that the intended match is on the diagonal of this matrix, thus any off-diagonal signal would reveal non-specific identification of SDSIs or contamination of SDSIs across samples.

[0020] FIG. 3A-3D – Maximizing Genome Recovery and Coverage with SDSI-ARTIC – **A.** The percent of the target genome covered at various depths of coverage when three reverse transcriptases, Superscript III, Superscript IV, and Superscript VILO were used for cDNA synthesis. Data represents four individual samples. **B.** Amplicons with at least 0.2X of the mean amplicon coverage with the normal ARTIC v3 primer pools or with a modified primer pool with a 2X concentration of 20 different ARTIC primer pairs. Four samples with low, mid-low, mid-high, and high CTs were used. **C.** Gini coefficients for two mid-high CT samples and four high CT when using either 35, 40, or 45 cycles for the ARTIC PCR. Error bars represent standard deviation. **D.** Comparison of Nextera DNA Flex and Nextera XT on the number of SARS-CoV-2 base pairs covered at various depths of coverage for three samples at different CTs.

[0021] FIG. 4A-4C – Improved amp-seq assembles more complete genomes than metagenomic sequencing with few errors across a wide range of samples – **A.** SDSI+ARTIC (N=81) and metagenomic (N=81) assembly lengths. All samples were downsampled to 975,000 reads. Dotted line indicates median assembly length (SDSI+ARTIC = 29,577; Metagenomic = 4,389) **B.** Percent of assemblies with greater than 98% or 80% coverage in different CT bins (SDSI+ARTIC N=81, Metagenomic N=81) (downsampled to 975,000 reads). **C.** SNP concordance plot between SDSI+ARTIC and metagenomic consensus sequences. Two discordant SNPs, outlined in a red box, were found.

[0022] FIG. 5A-5C – Rapid deployment of optimized amp-seq to determine a nosocomial transmission cluster – **A.** Phylogenetic tree showing the location of the putative cluster sequences in the context of a global subset of circulating SARS-CoV-2 diversity. Zoom box shows the 10 highly similar cluster genomes. Sample named on the main tree is the one putative cluster sample that was excluded from the cluster based on genome sequence. **B.** Distance matrix showing

pairwise differences between the 17 complete genomes assembled from this sample set. Putative cluster samples are bolded. **C.** Spike-in counts for each of the 24 samples and water controls in this sequencing batch.

[0023] FIG. 6A-6C – Spike-in validation – **A.** 100fmol DNA spike-in amplified under standard ARTIC PCR conditions for 40 cycles run on 2.2% agarose gel image with 188bp amplified spike-in (SDSI 1-48) **B.** RT-PCR for Spike-in and spike-in specific primers, Spike-in specific primers water control, Spike-in with COVID positive cDNA and spike-in specific primers, COVID positive cDNA and spike-in specific primers. **C.** Both SDSIs and ARTIC amplicons avoid extremes of GC content, and the two have generally overlapping distributions. SDSI primers also have a length and GC content similar to the average ARTIC v3 primer, resulting in a compatible TM.

[0024] FIG. 7 – SDSI Titration – Coverage plots for four different SDSI concentrations (1fM, 0.1fM, 0.01fM, 0.001fM) at four different CT dilutions (CT=20,25,30,35).

[0025] FIG. 8A-8C – Comparison to alternate amp-seq strategies – **A.** Three representative coverage plots for CT 20, CT 25, and CT 30 samples. **B.** SNP detection for the CT 20 and CT 25 sample. ARTIC and Paragon consensus sequences were compared to the metagenomic consensus sequences. The SNP that was not called in Paragon was due to low coverage at that position. Analysis was performed with assemblies generated with a minimum coverage of both 3 and 20, yielding identical results. **C.** Base pairs of the SARS-CoV-2 genome covered for the modified ARTIC pipeline versus Paragon CleanPlex Panel at different depths of coverage. Five samples at varying CTs were compared.

[0026] FIG. 9A-9D – RT comparisons for cDNA length – **A.** Read depth across each nucleotides position for the same sample (CT=13.89) when using three different reverse transcriptases (SSIII, SSIV, or SSVILO) for cDNA synthesis. **B.** Base pairs of the SARS-CoV-2 genome covered at various depths when using different enzymes for the ARTIC PCR. **C.** Base pairs of the SARS-CoV-2 genome covered at various depths when using either normal ramping speed (3°C/s) for the ARTIC PCR or reduce the ramping (1.5°C/s). **D.** Read depth across each nucleotides position for normal ARTIC PCR vs an alternate hybridization PCR.

[0027] FIG. 10 – Increasing primer concentration 2-fold in regions of low amplicon coverage – Red asterisk indicates amplicons in which the primer pairs were spiked in at 2X the

concentration of the others in the pool. Box plots showing the distribution of absolute sequencing coverage (log10) per amplicon for ARTIC PCR conditions (Normal) and Primer 2x concentrations for 4 representative samples. The boxes are plotted by the Q1, median, and Q3, the whiskers by Q1/Q4, and the outliers by the dots.

[0028] FIG. 11 – Modified Flex outperforms XT in coverage depth and evenness at lower cost – Illumina Nextera XT and modified Illumina Nextera Flex library construction on three samples with varying CTs. Asterisks indicate amplicons with large levels of drop out that were improved with the Nextera Flex. Plotted is the mean sequencing depth (log10) per amplicon.

[0029] FIG. 12A-12C – SDSI+ARTIC over a diverse set of samples is advantageous when compared to metagenomics – **A.** Time-measured maximum clade credibility tree of 772 genomes from Massachusetts, reported in Lemieux et al., 2020. The 89 samples compared for metagenomic and amplicon sequencing are shown with red dots. **B.** Genome coverage for metagenomics versus SDSI+ARTIC amplicon sequencing pipeline (N=81, excluded samples had no detectable CT). All samples downsampled to 975,000 reads. **C.** Gini coefficients grouped by CT (N=70, excluded samples that did not generate assemblies in either one or both methods). Dashed red line represents the median.

[0030] FIG. 13 – SDSI+AmpSeq Protocol. Illustrative workflow for 96 samples through the SDSI+AmpSeq amplicon-sequencing pipeline. A unique, synthetic DNA spike-in (SDSI) will be added to each cDNA sample to allow for contamination tracking and accurate sample identification in analysis. Asterisks indicate additional steps to the standard ARTIC pipeline.

[0031] FIG. 14A-14B – Synthetic DNA oligos spiked into amp-seq reactions designed to flag contamination and sample swaps. A. Schematic of SDSI design. Each oligo contains 140 bp of unique sequence flanked by common primer binding sites. Primers designed to amplify all SDSIs are added to ARTIC primer pools, and a unique SDSI is added to each clinical sample. Identification of multiple SDSIs in the same sample indicates contamination. **B.** Percent of SDSI reads mapping for each of the 96 SDSIs (horizontal axis) were quantified for each of the 96 SDSIs (vertical axis). Any off-diagonal signal would indicate non-specific identification of SDSIs.

[0032] FIG. 15A-15C – SDSI+AmpSeq amplicon coverage and genome concordance. A. Percent of SDSI for SDSI 1-96 in patient samples. **B.** Log of the mean amplicon coverage for the same clinical samples run with and without an SDSI (n=14). A unique SDSI was used in each

sample. The solid blue line represents SDSI+AmpSeq and the solid black line is ARTIC only with no SDSI. Blue and black shading around the solid lines represents the confidence interval. There were no statistical differences (p-value > 0.05) in the mean amplicon coverage for each amplicon between the groups (two-tailed Mann Whitney t-test and multiple comparison two-stage step-up Benjamini, Krieger, and Yekutieli test with FDR set to 5%). **C.** SNV concordance plot between SDSI+AmpSeq and unbiased consensus sequences. Two discordant SNVs, outlined in a red box, were found. Blue dots represent SNVs found in both the unbiased and SDSI+AmpSeq method, whereas black dots indicate the SNV was only present in unbiased.

[0033] FIG. 16A-16C – SDSI+AmpSeq performs well across thousands of samples. A. Sample diversity from two different institutions representing a range of CTs, viral lineages, and states of sample collection from samples where the data was available. **B.** The percent of SDSI reads out of the sum of all SDSI reads that map to the correct spike-in (Left: JAX, N=3,838, Right: Broad, N=2,903). Error bars represent SEM. **C.** The percent of SDSI reads over the total of all sequenced reads for all SARS-CoV-2 positive samples (Left: JAX, N=3,093, Right: Broad, N=2,670). Error bars represent SEM.

[0034] FIG. 17A-17C – SDSI+AmpSeq is used to identify sample swaps and contamination. A. Intentional SDSI contamination experiment (run in duplicate) assessing if different ratios of contamination between SDSI 87 and SDSI 94 (SDSI 87:SDSI 94) were detectable with the SDSI+AmpSeq method. **B.** Examples of experimental errors that were caught using the SDSI+AmpSeq method. **C.** Top: Distance matrix showing pairwise differences between the 17 complete genomes assembled from this sample set. Putative cluster samples are bolded. Bottom: Spike-in counts for each of the 24 samples and water controls in this sequencing batch.

[0035] FIG. 18A-18B – SDSI core sequence in silico validation. Applicants surveyed the core SDSI sequences by BLASTn to identify significant homology. **A.** Significant homology between SDSIs and anything in the NCBI database outside the domain archaea was identified and the SDSI and genus were plotted if identity (y-axis) was greater than 90% and query cover (x-axis) was greater than 50 bps. **B.** For each SDSI, Applicants identified and plotted (see color scale) the maximum alignment score for a significant homology to human (taxid:9606) and viral (taxid:10239) sequences in the NCBI database. Applicants also identified and plotted the alignment score for each pairwise combination of SDSIs.

[0036] FIG. 19A-19E – Spike-in validation. A. RT-PCR for an SDSI in water and a SARS-CoV-2 positive clinical sample background. Mastermix and SDSI specific primers were added to all samples. SARS-CoV-2 positive clinical sample is cDNA generated from a nasopharyngeal (NP) swab. **B.** The distribution of GC content and length for ARTIC v3 primers. **C.** The distribution of GC content of SDSI amplicons. **D.** 100fmol DNA spike-in amplified under standard ARTIC PCR conditions for 40 cycles run on 2.2% agarose gel image with 188bp amplified spike-in (SDSI 1-48). **E.** % SDSI reads over total reads for SDSI (2-48) over a range of SDSI GC% (33%-65.4%) showed no significant read depth bias. Error bars represent 95% CI. Linear regression p-value=0.8160 (Broad, N=2,903).

[0037] FIG. 20A-20B – SDSI Titration. A. In a titration of SDSI 49 across one clinical sample (CT=16) mock diluted to various CTs (CT=20,25,30,35), the number of reads mapping to both SARS-CoV-2 and the SDSI were quantified, and the percentage of each was calculated. SDSI 49 was tested at 600,60,6, and 0.6 copies/uL in each mock diluted sample. **B.** Coverage plots for the SDSI 49 titration experiment.

[0038] FIG. 21A-21B – ARTIC SARS-CoV-2 amplicon sequencing with and without SDSI and normalization. A. In three different CT bins, Applicants showed coverage plots with confidence intervals for multiple samples sequenced with and without SDSIs (CT<27, n=4; CT 27-29, n=6; CT>30, n=4). The solid blue line represents SDSI+AmpSeq and the solid black line is ARTIC only with no SDSI. Blue and black shading around the solid lines represents the confidence interval. There were no significant differences (p-value > 0.05) between the with and without SDSI group for the mean coverage at any of the amplicons (two-tailed Mann Whitney t-test and multiple comparison two-stage step-up Benjamini, Krieger, and Yekutieli test with FDR set to 5%). **B.** The percentage of SDSI reads for 4 different SDSIs was assessed within 4 clinical samples that were run with and without CT normalization of the cDNA prior to the ARTIC PCR.

[0039] FIG. 22A-22E – SDSI+AmpSeq over a diverse set of samples has superior genome recovery and more coverage uniformity at higher CTs. A. Time-measured maximum clade credibility tree of 772 genomes from Massachusetts, reported in Lemieux et al., 2021. The 89 samples compared for metagenomic and amplicon sequencing are shown with red dots. **B.** Percent of assemblies with greater than 98% or **C.** 80% coverage in different CT bins (SDSI+AmpSeq N=81; Unbiased N=81) (downsampled to 975,000 reads). **D.** Genome coverage for unbiased

metagenomic sequencing versus SDSI+AmpSeq amplicon sequencing pipeline (N=81, excluded samples had no detectable CT). All samples downsampled to 975,000 reads. **E.** Gini coefficients grouped by CT (N=70, excluded samples that did not generate assemblies in either one or both methods). Dashed red line represents the median. Error bars represent standard deviation.

[0040] FIG. 23A-23H – Maximizing Genome Recovery and Coverage with SDSI+AmpSeq. A. The percent of the target genome covered at various depths of coverage for four individual samples (CT=13.9, 23.9, 29.6, 33.6), with each undergoing cDNA with three different reverse transcriptases (SSIII, SSIV, or SSVILO). Yellow bar highlights comparison between the reverse transcriptases at a coverage depth of 10X. **B.** Read depth across each nucleotide position for the same sample (CT=13.9) when using these reverse transcriptases. **C.** Base pairs of the SARS-CoV-2 genome covered at various depths when using different enzymes for the ARTIC PCR (n=1). **D.** Amplicons with at least 0.2X of the mean amplicon coverage with the normal ARTIC v3 primer pools or with a modified primer pool with a 2X concentration of 20 poor-performing ARTIC primer pairs. Six individual samples with different CTs were used. **E.** Read depth across each nucleotide position for normal ARTIC PCR vs an alternate hybridization PCR (n=1). **F.** Base pairs of the SARS-CoV-2 genome covered at various depths when using either normal ramping (3°C/s) or reduced ramping (1.5°C/s) speed for the ARTIC PCR (n=1). **G.** Gini coefficients for two mid-high CT samples and four high CT samples when using either 35, 40, or 45 cycles for the ARTIC PCR. Error bars represent standard deviation. **H.** Comparison of Nextera DNA Flex and Nextera XT on the number of SARS-CoV-2 base pairs covered at various depths of coverage for three samples with different CTs.

[0041] FIG. 24 – Increasing primer concentration 2-fold in regions of low amplicon coverage. Data represents 6 individual samples at different CTs.

[0042] FIG. 25 – Unique identification of SDSIs given varying thresholds of SDSI mapping stringency. Applicants considered a range of cutoffs of the percentage of all SDSI-mapped reads mapping to a given SDSI (0.01%-50%, with a step size of 0.01). For an experiment where Applicants sequenced SDSIs without any clinical sample, Applicants calculated, at each cutoff, the number of SDSIs (y-axis) in the set Applicants present (96 total) for which only the expected SDSI had a proportion of mapped reads that exceeded the cutoff (x-axis). Assuming no contamination, all 96 SDSIs should be identified uniquely, i.e. no other SDSI should have a

proportion of mapped reads that exceeds the cutoff. The dotted line at x=5% represents the stringency cutoff that Applicants recommend in practice to detect contamination events.

[0043] FIG. 26 – Deployment of SDSI+AmpSeq to assess for possible nosocomial transmission. Phylogenetic tree showing the location of the putative cluster sequences in the context of a global subset of circulating SARS-CoV-2 diversity. Zoom box shows the 10 highly similar cluster genomes. Sample named on the main tree is the one putative cluster sample that was excluded from the cluster based on genome sequence.

[0044] FIG. 27A-27B – Modification enables addition of spike-ins to RNA. A. A schematic of how to design, produce, and apply synthetic RNA spike-ins (SRSIs). **B.** A limited titration experiment where SRSIs of varying concentrations were added to two clinical samples with low and intermediate SARS-CoV-2 Cts. SRSIs were added to the sample at the RNA stage; the sample with a low CT (20) was then normalized to CT 25 at the cDNA stage, whereas the sample with mid CT (26) was not normalized.

[0045] The figures herein are for illustrative purposes only and are not necessarily drawn to scale.

DETAILED DESCRIPTION OF THE EXAMPLE EMBODIMENTS

General Definitions

[0046] Unless defined otherwise, technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this disclosure pertains. Definitions of common terms and techniques in molecular biology may be found in Molecular Cloning: A Laboratory Manual, 2nd edition (1989) (Sambrook, Fritsch, and Maniatis); Molecular Cloning: A Laboratory Manual, 4th edition (2012) (Green and Sambrook); Current Protocols in Molecular Biology (1987) (F.M. Ausubel et al. eds.); the series Methods in Enzymology (Academic Press, Inc.): PCR 2: A Practical Approach (1995) (M.J. MacPherson, B.D. Hames, and G.R. Taylor eds.): Antibodies, A Laboratory Manual (1988) (Harlow and Lane, eds.): Antibodies A Laboratory Manual, 2nd edition 2013 (E.A. Greenfield ed.); Animal Cell Culture (1987) (R.I. Freshney, ed.); Benjamin Lewin, Genes IX, published by Jones and Bartlet, 2008 (ISBN 0763752223); Kendrew *et al.* (eds.), The Encyclopedia of Molecular Biology, published by Blackwell Science Ltd., 1994 (ISBN 0632021829); Robert A. Meyers (ed.), Molecular Biology

and Biotechnology: a Comprehensive Desk Reference, published by VCH Publishers, Inc., 1995 (ISBN 9780471185710); Singleton *et al.*, Dictionary of Microbiology and Molecular Biology 2nd ed., J. Wiley & Sons (New York, N.Y. 1994), March, Advanced Organic Chemistry Reactions, Mechanisms and Structure 4th ed., John Wiley & Sons (New York, N.Y. 1992); and Marten H. Hofker and Jan van Deursen, Transgenic Mouse Methods and Protocols, 2nd edition (2011).

[0047] As used herein, the singular forms "a", "an", and "the" include both singular and plural referents unless the context clearly dictates otherwise.

[0048] The term "optional" or "optionally" means that the subsequent described event, circumstance or substituent may or may not occur, and that the description includes instances where the event or circumstance occurs and instances where it does not.

[0049] The recitation of numerical ranges by endpoints includes all numbers and fractions subsumed within the respective ranges, as well as the recited endpoints.

[0050] The terms "about" or "approximately" as used herein when referring to a measurable value such as a parameter, an amount, a temporal duration, and the like, are meant to encompass variations of and from the specified value, such as variations of +/-10% or less, +/-5% or less, +/-1% or less, and +/-0.1% or less of and from the specified value, insofar such variations are appropriate to perform in the disclosed invention. It is to be understood that the value to which the modifier "about" or "approximately" refers is itself also specifically, and preferably, disclosed.

[0051] As used herein, a "biological sample" may contain whole cells and/or live cells and/or cell debris. The biological sample may contain (or be derived from) a "bodily fluid". The present invention encompasses embodiments wherein the bodily fluid is selected from amniotic fluid, aqueous humour, vitreous humour, bile, blood serum, breast milk, cerebrospinal fluid, cerumen (earwax), chyle, chyme, endolymph, perilymph, exudates, feces, female ejaculate, gastric acid, gastric juice, lymph, mucus (including nasal drainage and phlegm), pericardial fluid, peritoneal fluid, pleural fluid, pus, rheum, saliva, sebum (skin oil), semen, sputum, synovial fluid, sweat, tears, urine, vaginal secretion, vomit and mixtures of one or more thereof. Biological samples include cell cultures, bodily fluids, cell cultures from bodily fluids. Bodily fluids may be obtained from a mammal organism, for example by puncture, or other collecting or sampling procedures.

[0052] The terms "subject," "individual," and "patient" are used interchangeably herein to refer to a vertebrate, preferably a mammal, more preferably a human. Mammals include, but are

not limited to, murines, simians, humans, farm animals, sport animals, and pets. Tissues, cells and their progeny of a biological entity obtained in vivo or cultured in vitro are also encompassed.

[0053] Various embodiments are described hereinafter. It should be noted that the specific embodiments are not intended as an exhaustive description or as a limitation to the broader aspects discussed herein. One aspect described in conjunction with a particular embodiment is not necessarily limited to that embodiment and can be practiced with any other embodiment(s). Reference throughout this specification to "one embodiment", "an embodiment," "an example embodiment," means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, appearances of the phrases "in one embodiment," "in an embodiment," or "an example embodiment" in various places throughout this specification are not necessarily all referring to the same embodiment, but may. Furthermore, the particular features, structures or characteristics may be combined in any suitable manner, as would be apparent to a person skilled in the art from this disclosure, in one or more embodiments. Furthermore, while some embodiments described herein include some but not other features included in other embodiments, combinations of features of different embodiments are meant to be within the scope of the invention. For example, in the appended claims, any of the claimed embodiments can be used in any combination.

[0054] All publications, published patent documents, and patent applications cited herein are hereby incorporated by reference to the same extent as though each individual publication, published patent document, or patent application was specifically and individually indicated as being incorporated by reference.

OVERVIEW

[0055] Embodiments disclosed herein provide a method of detecting and preventing contamination during genome profiling using synthetic DNA spike-ins (SDSIs). Embodiments disclosed herein also provide methods to track sample contamination by implementing synthetic DNA spike-ins (SDSIs) for sample verification. Embodiments disclosed herein also provide synthetic DNA spike-ins (SDSIs) and methods for producing synthetic DNA spike-ins (SDSIs). The global spread and continued evolution of SARS-CoV-2 has driven an unprecedented surge in viral genomic surveillance. Amplicon-based sequencing methods provide a sensitive, low-cost and rapid approach but suffer a high potential for contamination, which can undermine laboratory

processes and results. This challenge will only increase with expanding global production of sequences by diverse laboratories for epidemiological and clinical interpretation, as well in genomic surveillance in future outbreaks. Applicants present SDSI+AmpSeq, an approach which uses synthetic DNA spike-ins (SDSIs) to track samples and detect inter-sample contamination through the sequencing workflow. Applying SDSIs to the ARTIC Consortium's amplicon design, Applicants demonstrated their utility and efficiency in a real-time investigation of a suspected hospital cluster of SARS-CoV-2 cases and across thousands of diagnostic samples at multiple laboratories. Applicants established that SDSI+AmpSeq provides increased confidence in genomic data by detecting and in some cases correcting for relatively common, yet previously unobserved modes of error without impacting genome recovery.

[0056] The methods described herein add a unique SDSI to each sample (e.g., cDNA) before performing a sequence amplification process during which the samples and SDSIs are amplified in the same reactions. This procedure can be repeated in parallel for each sample undergoing analysis. After the samples have been amplified, the presence of the SDSI is measured. If the SDSI introduced before amplification is the only SDSI present, then the sample is determined to be uncontaminated. However, the presence of any other SDSI immediately reveals contamination of the sample. This method provides a reliable safety measure for pathogen-genome studies and the resulting therapeutic and preventative medicine.

Synthetic DNA Spike-In (SDSI)

[0057] In one aspect, the present invention is directed to SDSI's and uses thereof. An example SDSI comprises, in a 5' to 3' direction, a 5' primer binding sequence, a core sequence, and a 3' primer binding sequence. In one example embodiment, spike-ins comprise sequences derived from a rare organism. A rare organism is a species that is limited in number or geographic occurrence relative to the distribution and abundance of other species making up the pool of interest.(Raphael, M. *et al.*, Conservation of Rare or Little-Known Species: Biological, Social, and Economic Considerations. Bibliovault OAI Repository (2007) the University of Chicago Press) In some embodiments, the rare organism is an archaea. In some embodiments, the archaea is thermophilic. A thermophilic archaea may exist in environments with temperatures greater than 50°C. In certain embodiments, the present invention includes spike-ins. In certain embodiments, a spike-in comprises a DNA sequence that is not from the target organism. In certain embodiments, a spike

in is an RNA molecule that can be added to a sample comprising pathogen RNA. In certain embodiments, the RNA is converted to cDNA concurrently with pathogen RNA. The RNA spike in cDNA can then be amplified with pathogen cDNA using pathogen specific primers and spike-in specific primers.

[0058] In certain embodiments, a spike-in sequence is compared to the target organism and the host for the target organism to limit homology. Limited homology can be determined using a BLAST search of all SDSIs. In one example embodiment, a permissive BLAST search is used (*e.g.*, blastn; 5000 max targets; E=10; ws=11; no mask for low-complexity). Results may be filtered by species of interest, *e.g. homo sapiens*. In one example embodiment, results can be filtered for a pathogen of interest (e.g., SARS-CoV-2). The query coverage and sequence identity may each be set for 35-100%, preferably, 50-100%, and sequences having no significant hits can be selected for use as a spike-in. In certain embodiments, a spike-in set comprises different DNA sequences that can be easily distinguished using sequencing.

[0059] In certain embodiments, the GC content of the spike-ins promote similar amplification rates across pathogen targets and the different SDSIs in our set. In one example embodiment, a spike-in comprises a similar GC content as the target organism. In another example embodiment, the GC content of the primer may range from 30% – 80%. (Buck, G.A. *et al.*, Design Strategies and Performance of Custom DNA Sequencing Primers, BioTechniques (1999) 27:3, 528-536). In another example embodiment, the GC content of the primer may range from or between 30%-40% nucleotides, or between 40%-50% nucleotides, or between 50%-60% nucleotides, or between 60%-70% nucleotides, or between 70%-80% nucleotides. In general GC content extremes are avoided. For example, sequences may have a median of 50% GC content, preferably, between 35-65%. In another example embodiment, the GC content of the primer may range from or between 40%-70%, or between 30%-50% nucleotides, or between 30%-60% nucleotides, or between 30%-70% nucleotides.

Core Sequence

[0060] Each SDSI in the set is differentiated by its core sequences. The SDSI cores are designed to minimize self-hybridization and cross-hybridization with others nucleic acids in a given sample. Accordingly, core sequences are selected based on the type of target sequence to be amplified and the type of sample the target sequence is to be derived from. For example, in the

context of detecting a pathogen in a human sample, core sequence should be selected with minimal homology to the target pathogen, other common microbes and non-target pathogens that might be present in the sample, and human sequences as well. In certain example embodiments, the core sequence has a homology of less than about 65%, or less than 64%, or less than 63%, or less than 62%, or less than 61%, or less than 60%, or less than 59%, or less than 58%, or less than 57%, or less than 56%, or less than 55%, or less than 54%, or less than 53%, or less than 52%, or less than 51%, or less than 50%, or less than 49%, or less than 48%, or less than 47%, or less than 46%, or less than 45%, or less than 44%, or less than 43%, or less than 42%, or less than 41%, or less than 40%, or less than 35%, or less than 30%, or less than 25%, or less than 20%, or less than 15%, or less than 10%, or less than 5%, or less than 1%.

[0061] The core sequence may vary in length between 50-5,000 nucleotides, or between 50-nucleotides, or between 50-4,500 nucleotides, or between 50-4,000 nucleotides, or between 50-4,000 nucleotides, or between 50-3,500 nucleotides, or between 50-3,000 nucleotides, or between 50-2,500 nucleotides, or between 50-2,000 nucleotides, or between 50-1,500 nucleotides, or between 50-1,000 nucleotides, or between 50-500 nucleotides.

[0062] The core sequence may vary in length between 50-60 nucleotides, or between 50-70 nucleotides, or between 50-80 nucleotides, or between 50-90 nucleotides, or between 50-100 nucleotides, or between 50-110 nucleotides, or between 50-120 nucleotides, or between 50-130 nucleotides, or between 50-140 nucleotides, or between 50-150 nucleotides, or between 50-160 nucleotides, or between 50-170 nucleotides, or between 50-180 nucleotides, or between 50-190 nucleotides, or between 50-200 nucleotides, or between 50-210 nucleotides, or between 50-220 nucleotides, or between 50-230 nucleotides, or between 50-240 nucleotides, or between 50-250 nucleotides, or between 50-260 nucleotides, or between 50-270 nucleotides, or between 50-280 nucleotides, or between 50-290 nucleotides, or between 50-300 nucleotides, or between 50-310 nucleotides, or between 50-320 nucleotides, or between 50-330 nucleotides, or between 50-340 nucleotides, or between 50-350 nucleotides, or between 50-360 nucleotides, or between 50-370 nucleotides, or between 50-380 nucleotides, or between 50-390 nucleotides, or between 50-400 nucleotides, or between 50-410 nucleotides, or between 50-420 nucleotides, or between 50-430 nucleotides, or between 50-440 nucleotides, or between 50-450 nucleotides, or between 50-460 nucleotides, or between 50-470 nucleotides, or between 50-480 nucleotides, or between 50-490

nucleotides, or between 50-500 nucleotides, or between 50-510 nucleotides, or between 50-520 nucleotides, or between 50-530 nucleotides, or between 50-540 nucleotides, or between 50-550 nucleotides, or between 50-560 nucleotides, or between 50-570 nucleotides, or between 50-580 nucleotides, or between 50-590 nucleotides, or between 50-600 nucleotides, or between 50-610 nucleotides, or between 50-620 nucleotides, or between 50-630 nucleotides, or between 50-640 nucleotides, or between 50-650 nucleotides, or between 50-660 nucleotides, or between 50-670 nucleotides, or between 50-680 nucleotides, or between 50-690 nucleotides, or between 50-700 nucleotides, or between 50-710 nucleotides, or between 50-720 nucleotides, or between 50-730 nucleotides, or between 50-740 nucleotides, or between 50-750 nucleotides, or between 50-760 nucleotides, or between 50-770 nucleotides, or between 50-780 nucleotides, or between 50-790 nucleotides, or between 50-800 nucleotides, or between 50-810 nucleotides, or between 50-820 nucleotides, or between 50-830 nucleotides, or between 50-840 nucleotides, or between 50-850 nucleotides, or between 50-860 nucleotides, or between 50-870 nucleotides, or between 50-880 nucleotides, or between 50-890 nucleotides, or between 50-900 nucleotides, or between 50-910 nucleotides, or between 50-920 nucleotides, or between 50-930 nucleotides, or between 50-940 nucleotides, or between 50-950 nucleotides, or between 50-960 nucleotides, or between 50-970 nucleotides, or between 50-980 nucleotides, or between 50-990 nucleotides, or between 50-1000 nucleotides, or between 50-1010 nucleotides.

[0063] The core sequence may vary in length between 100-5,000 nucleotides, or between 1,000-5,000 nucleotides, or between 2,000-5,000 nucleotides, or between 3,000-5,000 nucleotides, or between 4,000-5,000 nucleotides.

[0064] The core sequence may vary in length between 75-150 nucleotides, or between 100-150 nucleotides, or between 100-200 nucleotides, or between 100-300 nucleotides, or between 150-200, or between 150-250 nucleotides.

[0065] The homology to a target sequence or non-target sequence in the sample across the size of a given core sequence may be less than 1 nucleotide, or may be less than 2 nucleotides, or may be less than 3 nucleotides, or may be less than 4 nucleotides, or may be less than 5 nucleotides, or may be less than 6 nucleotides, or may be less than 7 nucleotides, or may be less than 8 nucleotides, or may be less than 9 nucleotides, or may be less than 10 nucleotides, or may be less than 11 nucleotides, or may be less than 12 nucleotides, or may be less than 13 nucleotides, or may be less

than 14 nucleotides, or may be less than 15 nucleotides, or may be less than 16 nucleotides, or may be less than 17 nucleotides, or may be less than 18 nucleotides, or may be less than 19 nucleotides, or may be less than 20 nucleotides, or may be less than 21 nucleotides, or may be less than 22 nucleotides, or may be less than 23 nucleotides, or may be less than 24 nucleotides, or may be less than 25 nucleotides~~,~~.

[0066] The homology to a target sequence or non-target sequence in the sample across the size of a given core sequence may vary in length between 1-5 nucleotides, or between 1-10 nucleotides, or between 1-15 nucleotides, or between 1-20 nucleotides, or between 1-25 nucleotides, or between 1-5 nucleotides, or between 5-10 nucleotides, or between 10-15 nucleotides, or between 15-20 nucleotides, or between 20-25 nucleotides, or between 1-10 nucleotides, or between 10-20 nucleotides, or between 20-30 nucleotides.

[0067] These SDSIs can be implemented in a wide range of genome profiling applications including, but not limited to, investigations of SARS-CoV-2 epidemiology and emerging viral variants. Exemplary SDSIs are provided in Table 1.

[0068] **Table 1.** Sequences of 96 unique SDSIs. The unique core of each SDSIs is 140 bps long (SEQ ID NOS: 1-96 and 193-291). The unique SDSIs including the priming regions (SEQ ID NOS: 97-192 and 292-390). Alternative sequences are also included. SEQ ID NOS: 16, 57, 66, 112, 153, and 162 can be, in the alternative, substituted with 289, 290, 291, 388, 389, and 390 respectively. Sequences for forward and reverse primers for amplifying the SDISs (SEQ ID NOS: 391 and 392 respectively).

Table 1

SEQ ID	Core Sequences
1	CAATTGCTCCCTCGTATCCCTTGTACATTATCTCAGCTCCGCTTAATGATATTAATTTTACCTTGA GTGTTTTTGCTAAAGCCTTTGCCATCATCGTTTTACCTACTCCAGGTGGCCCGTAAAGCAACACA GCTTTGGCA
2	TTCTCCAAAACCTACCCAGTTCTCCGAGGAACCTCTTAGCATCTGTTAAATCGTTATTAGTATTA GCTTCCACCATCTCAAGTTCCTTTAAGGCGTTACTCACACTCTTCTTACCTATCTTTTAGAGAACC ACTCGTCAG
3	GTTATCAAAGCCCTTAAAGAGTGGTAGGGGCAAAAGTCTGAAGCGTCCTTACTTAACTGGAGTA TCTGAGATGGCCTTAATCCGCTTAGGTCTTTAATTTTATCCCTTAATGAACATTCCCTGCACTCTA TGTCTTCGGG
4	GAGATGTAGCAGACGGGCTAAGAGTTTCAAACCCTCTAAGGATCACTACAAACAAGAGAGAGA GACAATCCTCTCTTTTGTCTTGTCATTGTGTTTCAAACCCTCTAAGGATCACTACAAACATCTTTA ACATAGATACC

5	GACCGGACGTTGTGATCACGGGTACCTTGATCTGGTACTCAAAGGTTTGCCCCCGTGAAGTCTG GTACATGGCTAGACACGTCACTCCATTCGAGGGACATTCGAAGTTAGAGAAGGGCAGAGCGAT ACATCAGATATAT
6	GTCTTTTCTCTACTAATTCTCCTCACGAGATCTCTAAACATTCTTGCTGAAAGAGGATCCAAACC TAATGTAGGTTCGTCAAGCAATAAAATTGGAGGATCAGTTATTAATGCTCTTGCTAAGGCTAGT TTCCTCTGCAT
7	GATTTTGCCATCATTAAAAACAACAATTTGATCACCCATAGTCATAGCTTCTAATTGATCGTGAG TTACATAAATACTTGTGGTGTTTAACATACGGTGAATATTTACAATTTCTCTTCGCATGTTTTCTC TTAGTTTAG
8	GTATCTTTCAATTCTCGAAAGAAAAGGTTACAAGTCTCATAGATTTATTCCTCTTCACTGTTGTA CGTTGGCAGCTAGAGAGAGTTTAGATTATGAGAAAATTAAGAGAATATATGAGGATTCGTTTTC TTGGTTTAAGT
9	CTAATTGATTTTCCTGTACCATGTGGTAAAACAACGCTACCTCTTAATTGTTGATCTGCTTTTCTA GTATCAAGATTTAATCTAAAAGCTAAATCAACTGAAGCATCAAATTTTGTATAAGAAGTTTTTTT CACTAATTC
10	TCGGTTTTCCCGTGAACTAATAAACACCTACTGGAGCCAAGAACGGGTCAGAATTGATGGAATA AACGTTGCGGAGAATGAAATTAATTTGTACATCAGAGACATTGATGACAACGGTGACCCTATAC AGTCAACTATAC
11	CTTAATGGAAAGTATGCTTTAGATACCTTCTGGAACGCTATCTCACTTGGCGGGAATTCAGATAT GGAGAGTAAATTAAGGGATCTGGAAGTAAAGTTAATGTCGTTAATCTATTTAAATGAGTCACCA TTAAAATCACC
12	CATAATATGTTAGAGGTAGAATTTCTTTGTGATAGAATATTATTGATGAATGATGGAAGAGAAT TAGCATTAGGAAAACCTAAGGAACTGGTAAAGGATACAGAATCTAAGAATCTTGAAGAGGTTT TCCTTAAACTTGT
13	CCTTACTTCATCTCTCAAGATAAGGGTAATAAGTTCACTTCAAATATCTGGTCTTATCGCAAGTT GATTGAGGCTATAGTGTATAAGCTCTATGAGTATGGTATAAACGTGTTCCTCGTTGTAGAGTAT AACACTTCACG
14	AGTCTAGGTTTTAATTCTTCAACTGCTTCAAATACTAGCTTACTGTAGTTATCTGCCCTCATGTTA GGATATATATCTGGAATATAAGGAGGTTGATGAGTTATAAGAAGTGGATGAAATTGTTGTCACA CACTCCCCTA
15	CTACCTCTTCGGCCTTGTACCAACGTACCCCTGATACAAGTTCCAAGCAGAGATGGAAAACTCG AAGATGGTATCACCCAAGATGAGATACGATATCAATGAAGGCGAGCCTAGGTACAAGTAAAGG GATACCACGAGAG
16	CTCGTAAGCGTTTCCTACCCTCGAGAGGGCCATCCTGGTGGTGAGGAAGTCGTCGAAGTGGGCT AAGTAAAAAGCGAAGATCTCGACCCACAATTACCTCCTCCTGTACACCAGGAATACCCCTATCA GGATAGAGATAC
17	GCGCGTCCGGGTCGCGGCCGGGGACGACCGTCTTGACGAAGTCGGTCGACCCCTCGTCGGTCGA GATGGTCGTCACCTCGGTGTCGAGGCCGTACGTTTCGAGCGCGTCGCGTACCAGTTCGCCGTCC GCGTCGGGACGG
18	CATGTACTCGTTCCAGAAGGTGAGTTCGCTCCCCTCGATTTCGACCTCGCCCACGTCGAAGCCGC CGGTCGTTTCGAGCGCGAACGACTCGACGGGACCGACGAGCGAAACTTCGCCGCCGAGCACGT CGGCGACGCGTT
19	CTCGATGCGCTCGGGCTTGTAGGACTCCCCGAGGGCGTCCTTGTTGGTGAAGACGTTTTGTTTTC GCTCGAACCGGCGCATTAGCGTCGGTCCGTTGTAGCGTCCCCTTATTTAAAACCCCGATTTCATC TGATTCATGT
20	TCACGGTCCGCGACGTGAATCGGGCGTTCCAGTCGGCGTTCGGCTACGACGCCGACGACGTGGT CGGAAGCGACCTCCTCGGGCGAATCGTGCCCCCGGTGCCGGACCCGGACCCGGTGCCGGAACC GGGGGACGACGAG
21	GCGTCCGCGAGTTCATCCTGAACGTCGTCCCGCTGTCGCCCGGCGAGGAGCGCGGGGCGGGCTA CGCCATCTACACCGACATCACGGAGCGGAAGACCCGCGAAAGCGAGCTAGAGCGACAGAACGA GCGATTGGAGGAG

22	GCGAGACCGGCGACGAGGTGCGCTTCGACACCGCCGAGCGGGCGCTCGAACAGATGGAGGAAC TCATCGACGACCTGCTGTCGCTCGCCCGTCGCGGCCAACTGGTCGACGAGACGGAGCGCGTCGA CCTCGGGGCGGTC
23	ACGAACTCGTCGGTGAACATCTCGTCTTCCGGGGAGCCCGCCGCTCATGGCCTGCCCCCGCCGT AAGCTGCTGCATAAACCCGCTCCAAAATATACGGATCATTCACCCCTTGGAATCGCTCAATCAG ATCAATGTACAC
24	TGCGTACATTCCCCCTAAGCGGCTCCCAATATACAGACGCCGGTTAACGACAGCTGGCGACCCT GTGATCTCAGTACCGGTGTCGAATGACCACATCAGCTTGCCTGTCCGTGCATGGAGTTCGTATA CGTACCCGTCGT
25	AGATAGATGAGCCGATCAGAGATCGCTGGTGAGTTGGTAATTGTCCCGACATAGACACGCCAA CGTTCTGTTCCATCTGCTGCGTCGTAGGTCGCGAGATACGGCCAGCCACCAACATACACAATCC CATCGACGAGGAC
26	ATACACCACCCCATCAGCAACAACTGAATCATGATTAAGTATCGCACCAGCATCGTAGCGCCAG CGTTCACTGCCAGTGGTGCTATCGAATGCATAGAAGATATGCTCCTAATCGCCAATATCAGTAC TTCACAAAGCCG
27	TCGACGAGGAGAGGGGCGAGTACATCTGCACGCTTACGGGAGAGGTAGTTGAGGAGACGGTTA TAGATACAGGGCCCGAATGGAGGGCTTACACACCTGAGGAGAGGACCCGCAGAAGCCGCGTGG GCAGCCCGCTTACC
28	AGTCGATGGCTGCGGCAGCTGTCTATGCTGCCTGCCGTATACGCGGCATACCCAGGAGTATAGA CGACATAGCGGAGGTCGTGAAGGGTGGCCGTAAGGAGGTTGCCCGCTGCTACCGCCTCATAGTC CGCGAGCTGAAG
29	GTGGAGTCTTTTGTCACACCGCAGAGGCGTAGCGCTGCAGAGCAGGAGCCCAAGCCTACTGCCA ACATAGAGAACATAGTGGCTACAGTATCCCTCGACCAGACTCTAGACCTGAACCTCATAGAGAG GAGCATACTGAC
30	CGTCGCCTGGGTTAAGAGGATGTTCGGCCTCTCCAAGGCGGGTCACGGAGGCACGCTGGACCCG AAGGTCACCGGCGTCCTCCCCGTAGCCCTGGAGGAAGCAACCAAGGTCATAGGCCTGGTGGTG CACACGAGCAAGG
31	CGTGGGCGAGATCTACCAGAGGCCGCCGCTCCGCAGCAGTGTTAAGAGAAGCCTCCGCGTCAA GAGGATATACGAGATAGAGCTGCTGGAGTACAACGGCAGGTACGCGCTCATGAGGGTGCTCTG CGAGGCCGGCACAT
32	CGCTGGAAGAACGAGGGCAAGGAGGACCTGCTGCGGAGCTACATCAAGCCCGTCGAGTACGCC GTGAGCCACCTGCCCAAGATAGTTATACGCGATACCGCGGTGGACGCCATAGCCCATGGCGCG AACCTCGCGGTGCC
33	GGGAGACCCCAAGGTGACCGGCGTCCTACCAGTGGGGCTCGCCAACAGCACCAAGGTCATTGG TAATGTTATACATAGTGTTAAAGAATACGTGATGGTTATACAGCTCCACGGCGATGTAGCCGAG CAGGATTTAAGAA
34	TAGAGGGAAAGACTGTAGCTTTCATTCCTAGGCACGGAAAGAGACACAGAATACCTCCACATA AGATAAATTATAGAGCTAATATATGGGCATTAAAAGAACTAGGAGTGAAATGGGTCATCTCAG TTTCTGCCGTAGGA
35	TGAGGGAGCTCAGGAGGACTCGCACGGGGCCCTACAGGGAGGATGAGACACTTGTAAGGCTCC AGGACGTCAGCGAGGCCCTGCTCCTGTGGAGGAGCAACGGGGATGAGAGGTATCTTAGACGCA TCGTGCTACCCGTT
36	GAAACATCTATCGCCCACCTCCCGAAGATAATGATCTTGGATACAGCTGTCGACGCCATAGCAC ATGGTGCCAACCTGGCTGCCCCAGGCGTCGCCAGGTTAACCAGGAACATCGCGAAGGGTAGTA CCGTAGCGATCCT
37	TCGCTATCCCCGTGTACAGCATGGTGGGGGTGCCGATGCCCGGGTAGAACTTGGTGACGCTCTC CAGCTTCTCGAGGACGGTTTCCTTGGGGAGGCTCGCGGTGTCCACGAGGGTTATCGCGTCCTCG GCGCCGTCGCCG
38	CGAGGACGCGAAGAGCGCGGTGGATGTGGACGCGCCGCCGCACACGTAGCCGTCGAGGTAGCG CGGAACCATCGGCGACATCAGCCCCACGACGCGACCCGAGGCGTTGCCGAGGATCACGTCGAG CGTCACGCGCGGCA

39	CTCGACACCGTGCCGTTGCCCTCCTCTAAGTAGTCGGAAAGCCTCATCCGCGACTCCAGCTTCGC CACCGGCTCCTCGAGCAGGAGGAGGACGCGGTTGATGCGGTAGGACGCACTGCCCGCCTCCAG CACCGCGCCGTC
40	TCTATGGTGTAGAACGGGTCGTTGCGGAGCCAGCCTGGCGGCACGTACCGGTCGTCCGCTATCG CCAGCGATCTCTCGAAGAGGTCGAGGTAGGCGGACGCGTTGGCGAACGCCCCGTGTATCACGA CGTCTATCCCGCC
41	GTATAGGTTTCAGGTATTGATAATGCATAGGAGGTTTTTAAAACCTTGAGCCGCATAGTCTTCTG GATGGGCGAGAGACATGGTTAAGTATAAGTGCGGCAGGTGCGGATACGTCTTCGACGACGAGG AGATGAAGAGGA
42	CCTACGCCGGGTGCGTAGGAGGGCTCGAGTACATCCATGTCTATACTGATGTATGTTTTACCCA GGTCGCCTAGTGCCAGGGGTCCCTTTAACGCTTCCAGGATAGAGTACACGGTGACGTCTCTAGT CTTCTTCAAGAA
43	CTACTAGCGTGTCAACGGAGCTCTTCAACGCCTTTACTATTGGATAGGTTATAAGGTGCTCGCCT CCGAGGAATCCCAGGAGCATGCCGGGATACTCGTCTACAACGCCTTTCACCACGTCACCTATGA TTCTTAAAGAG
44	CATAGGTGACATGGGGTTTCCCATTGACTCTATAAAGCCGTATCCTTTAAGCGGAGTGCAATTG GTCTACGCTTTGCTTAACAACAGGTATTTCCTACCGGGTAGAGAGGGCTCGCTCATAGCTTTAG GTAGCGTGACGG
45	GGTATCTCACCGCTTGTCACCATAGTATCCCTCAGGTACTCCAGTATTCTTGAGAGAAACGCACC TAAGCCGGATCTCAGGTTTGAATCCATAAGAACTATGAGTGAAGCGGGATTGAAGCCCCTGCTG TTTCTAAGACC
46	TAAGGGAGATAGAGAAACGCATCAAAATACCCTTGGGGAAACTGCGTGCAGGGGTTCAATATG GAGTAGAGGTCTCAGACATAAAGGAGAAGATAGCTGCTTACGCTAGGAGGAAGGGGCTTAAAT ACTTCCCATCGGCA
47	TGTGAACCTCGTGCCCGGCTCTAAGTCGTGAGGGCTTGCAACATAGGTGGGGAGGAACCCGAG CAACGGGTAAGAAGACAGGATAAGCGGTATCGCTATGAAGAGGGCTGAGAAAAGGACATATA CTCCTGAGCCCGTCC
48	CGAACATGCCTTCCCCGTCTATATAGACCCAGTAGAGTTTAAAAACTTAACCAGAGACGGCTTG TGAGCCGGATCTCTCCCCCGCTAGGCCCTGGATTGGGCTCGCTCCTCCTGGGACCCCGGCCTCCA CATGCTCGGGA
49	CCTGAAGGGCTCGGCTACCCTGAAGACGGGCTTCTGCGCGACCGCCGCGTACTCCGCCGTGGAG CGGTAGAAGAGCGAGGCTGTCTCCGTGAGCCTGACCATTCCGTACAGGGCGACTGCGACGAGC ACTATGACTGCGA
50	GTCAAGGTGCTGATGCCGAAGGCGACTTTCGACACCGACGATGCCGCCGACGCCCTGGCCATTG CCATCTGCCACGCGCATCACCGGCACAGTGTTGCCTATAGGATGGCGCTGGCCGGATAAGTTTG TTCTTGACCTGT
51	TCTCGGTTCGGCAATAAGTAATACCAACGAGGTATTACCATGCGCGTGACCAGCAAAGGCCAA GTGACGATCCCAAAGGAGATACGGGATCATTTGGGGATTGGGCCGGGCTCCGAGGTGGAGTTC GTGCCCACAGACGA
52	CTCGATCATATGGCCGGCACGTTGGACTTGGGAGGCATGACAACGGACGAGTATATGGAGTGG CTGAGGGGTCCACGTGAAGATCTCGACATTGATTGACACAAATGTCCTGATCGATGTTTGGGGT CCTGCCGGACAGG
53	CAGGTGTATTTTACACACCTGGACAGCCAGCATATGATGCTAGCACTCGGTGTCCCCTTATCAC GGTTTCCCGCATTGTAAAGTTTTCGCGCCTGCTGCGCCCCGTAGGGCCTGGATTCATGTCTCAGA ATCCATCTCCG
54	CTGGAGCCTGTTAGTTGTTACAGGTTCACCGGTTGTCGGAGTATTCAGATCATTGAGCCAGCAG TTGATGGCTGCCTGTAGTTCACTGGTTGTGATGTAAGCTGCTCCATCGGAATCAACATCGTTCCA TGGGTTCCAGT
55	ACGGTCTTGCTTTCTCCTGAATCCATTTCACCTGTCCAGACCCATTCATAGCGGTTAGCTTCACT GAGGTTCTGCTTGAAGACACCGTCATCATTGTTAGATGAGGTTATTGTCCAGCCGGCAGGAATG ACTTCTTCGAA

56	GTCAGCAGCTCTTCATAGAAGTTCTGGTTTGCAATATCCCTCTGGGCAATGACAGGGTAGTCGA CTTCGTTTGCAGTCAGGTGGACTGCATACAGGGACTTGCTGATGTCCGGGGTATATCCACTGTG AGGAGCATAGTA
57	ACCCGTCAGTCGTGACGTCCTCCGCTCCTCCTATGCTATCTCCACACACCCACTCACGTTCTTGC TTCTTTACTACACCCTCTTTATTCAGCTCTTCGAGAACATTATTAATGTGACCCTTAGAGATATAT TCATTATAC
58	GTGCCTCCTCAAGCGACTGCTTAAACCCAATTACATCTGATTTATCCTTTATTTTAGGGCCTATA GAATCTATGAATAATTCGGCGATTCTTATTATTTCTAAAACCAATTCGTCTGTTTTGAGTGGTGT GCCTTCTTCA
59	CATCCCATGCATTTTCATAATAATCGGAATTCAAATCCTCTATATTGAATTTTATCTTAACATTTG ACATAATCATTTTCTCCTTACAGAAGAGATCCAGCTAAGCTTACTCATAAATGGTAGTACCATG CCAATATTGG
60	CGTAGCCCGCACCTTCCTCTGGTTTAGCACCAGCGGTCCCCACAGAGTACCCATCATCCCGAAG GATATGCTGGCAACAGTGGGCACGGGTCTCGCTCGTTGCCTGACTTAACAGGATGCTTCACAGT ACGAACTGACGA
61	CCTGATAGGCCGCAGATTCATCCTAAGGCGCCGGAGCTTTTGACCACAGAACATTCCAGTATCT ATGGTATATCTGGAATTATCACCAGTTTCCCGGTGTTATGCCAGACCTTAGGGCAGATTATCCAC GTGTTACTGAG
62	TGTTTGGCTTGATACTAATAAAAGCACAGCTAAAATGAAAATAAGCCGATATTTGTGATTCATG CAACTCACCCTTTTCTACATAAACAAAATACTAACCCGAAAACCGAAATTGAAATTAATGCAGA GAAACCAGGTGA
63	TTAACGGCACCAACAGTTATTATATTTTTAGCAGTCCCGGGTGAAGTAATTATGGAATAGTTGTT AGAATTACTGTTCTTATTACCAGCTGATTTGAAAGCAATTATACCTGCATCACGAATTGCAGCAT CATAATATTC
64	GGCTCAGACGACTGAAAAAGCAACGATTGGAATAATAGGGGGTTCTGGGCTCTATGATCCTGGT ATTTTGACTAACAGCAGAGAAATAAAAGTATATACACCCTATGGGGAACCTAGCGATTTGATAA CGATAGGTAACA
65	CGCAGAACAGGTTCCTTCTATTGGATATTCATCTTCGGCTGCAGTTGCAGGAAGAGTAAGGATA TATACTACGGTCTTGCTTTCTCCTGAATCCATTTCACCTGTCCAGACCCATTCATAGCGGTTAGC TTCACTGAGGT
66	CTTCCTCCACGCATTTGTTGTGGTGCTGATGGCGTATTCTCTGGAATTTGGGATGATTCTGGAAA TCCATCCTCAGACACTTCAGATATTTTAGTCTTACTTCCAGCGTTTAATTGAACCTTACCTTTAAA AGCAGTAGT
67	GAAACTTACCTTATCAGTGTCATTAAGCATATTGCTTCCAAGACCCATTGAAGCACTTACATCGT TGATACACAGGTGCCAGGAATAGTATTCCTCAGTCTCACTATAATCCTCGTTGGTGTAGCCTTCA AGAGAGTCAA
68	GTTTAAGCAATTCTTCGGATGAAAGATGGCGCTCTATAGGAATTTGTTCTGGTCTAGCCATAAG GCATTATTTGTACTTAATTAGTAATAAATGTTTAGTTAATGACTATAAATCTGCAATTGGAGTCT CAAATTTTCAA
69	AACATGAAGGATGTGTGTAAGAGGAAACGTTATTAACAGACGTAATCAGGAGGATAGTTATGC CCTAAAAACAGCAGAGTTAAGGTTTAAAAATAAGATAAGAACTCAGTTGAGGTTTATCCATTAA TCCCATTAATCCT
70	ACTTTCTAAAAGCGCTTGGAGCACGTATCAGGTCAAGTCTTTCAACCTTAAATGCTGCCAGTGC CGTAAGTAGTGCAGTTATGTTGCTTATTGAAACAAACAACTTAGCCCACTTATTACCTCTTGTCA GTGTTTTTGAT
71	GTATCCGCTGATATATCCTGGGGATATAGATCGCTCTGAAATGGTTACATCTATCGGTTTTAAGG ACAGTTCCAACACTATTGGACCTTGCAGCTATGACAGGAATAATCTGTTTATCGAGCACAGTTG AATTTGACCTA
72	TCAATACCTAATTCTTTCCTTAGAGTGCTATTTTGATTGAATTCCCTCAGGAAAGATTCAAAATT TAAGTAGCCGAGCTTACATCTTGAAATTTCCATCTTTATTATGTTGCTCAGGCTTAATGCTTCTA AGTATGGGTT

73	AGATATCCTTTGAAATTCTCGTAATTGCTGAAGGCCACTACTTCATCAGGTCTGATGCAATCTTT AATCTGAACATTGCTTTCTGAGGTCTTAGGAATAATCCTGTAAGGGAGTCGGATATTGTTCGTTA AGATGCTCTT
74	ATAGAGGGACCTAGATTTTCAACGAGGGCAGAAAGTAGAATTTGGAGGGAAGTTTATAAAGCC GATATCATAGGGATGACTTTAGTTCCAGAAGTAAATTTAGCTTGCGAAATGCAAATGTGCTATG CAACAATTGCGAT
75	GTCTTCAGCATAGTACCAGCTTATGTTGTCACCATCGTTCAGTACGTTACCACCAAGTCCACTGC CTGCAGCTACATCATTAATGTACAGGAACCAGGCATAGTAACCGCCAGTTGATATGTAGTCTTC ACCTTCGATTC
76	ACTCTCCATCATGACAGCCAGATCGGTCATAGCATCGATTGTGTACTCTTCGTCGGGATTGTTGT ATGGAATGAACTTATAGTTCTCACCTGCTACCTGATCCACTGTCATTTCTGCAAGAGTCTGCACT GTGGTAATTC
77	ATATTCCGTATTTCTTATCAAACCGATCGTGAAGATTTGACAAAGGCTTAACTTTAGGGCTCCAC TTCTCATTATTAGCCTTAGAATATAAAGCGTAACCGTAAGCCTGAGGAACGTAAAGCTTAGGAG ATTCAATCCCG
78	TAAAATTAGCCGAAGGCTTCCCATTACCGAAAAAGTCGTTTATTAGCTCTTCATCCTTCTTCTCC ACGTCCGCCCATTCCTCTCCTTCCCTTGGAATTTTAAGCTCGTCCCAGCTGACTCTTATGGGCAA TTCAATATCC
79	GTATAAACTTTTGATATAACCTTGCCTAATTTGATATCATAGCTTATGTTTGGCGCTATCCCCCA CTTGTAGAGGGTCGCGTTATATTCTCTAATAGCAAGAGAGATACAAGATTCGTTAACGTTATTT ATATCACTCTC
80	TCCGGAGGAATCTATCATATTAAACCTCCTCAAAATCGCCTCCTCTTGATTGCTTAAAGGCTGTG AATTACAAAGCTTATTTAATGCGTCCCAAAGCGTTAAGTAATAATTATTTATATTAAACACTACT ATTTCAGTAG
81	GTTCCTCCTCAATTCAATTGGACTGAAGGAGGGTACGTTCTGGAAAACAGAGCGTAAAAGAGAT ATAGAACGTAGTATACACATAGCTGGAAAAAGAACAATCATTAAGACAATAAAGAACTTTATG GAAAAGAGTAGAA
82	TCGTGTAAAGGTTGTATAATTCAAGCCTCAGAACATTTCGAACTCCTTACAAAATCGTTTAAACT TTCTAAGGCATAAATTTACTAGAAATTGTCATTTATGAGAATGTAACTATATAGATGGTAAAAT TATTAATCCTC
83	GGCTGAAAAATAGGTTCGATCCGCCTCCTCACTTCTTCTCCTTCTTGCCCTCGGCCTCGGAGGAG GCCTCTATTCCCAGCTTCTTGGCCTCCTCCTCGGTCGTCATGAACAGGCTAGTCCTCTGCCTTCC GCCCATGCTC
84	GACCTAGCCTTACGCACAGCCCTCTCCACAACCTCCTCAAGCTTATCCCAGTCAATAGAGCTCAT TACAAGTTAACCACGCCCACCTTTAATATAAACCTTTACCCCTCGTGGCAATTAACTTTAACCGC TACTCCGGTG
85	TGGCCCTTAGACCTCTGCCCATGCTTAGGCGCTTACCCACACCTATTAGTACGGCGCCAATGCCC ACGGCCATGAAGTACATTAAGGCACCCATGGTTGCACCGTAGAGTGCCGTGAATGTTCCGTAGA ATACACCGGCC
86	TCGGCGAATCTGTCGAGCTCCATGACGTCCACAGAGCCGCCGAACTTGGCCGAGAATCTATCGG CCTGGGCGGTGCGCCTCCCTATCAGCAAAACCCTGGGCGCCGTCAGTAGCGCGACGGCCCTGGC GATTCCCCTGGC
87	TCCAGGTAGGATCTGGCCGAGAGGGAGGACGCCGCGCTGTTGTGCTCCGGGAACCCTAGAGTC ACGACCGCCTTGACGCCTATACGTTCGGCGTATTCAGCGACGGCGGCGCCGGTGCCGCCCGTCA GCGTGACGGCAAG
88	GCAAGAGAATACATTTTTGATGATAAGAGAAGCTTGTGGCATACTTTCTTAGGCTTTATTTCAGC ATTCACTTTAGCGTATTCTATCGTTATTTTGCTATTGTTCACATTGTATCAAGTGAGAGAAAGAG AGAAGCCAAC
89	AGAATCAAAGGAGTGGTGTAAAGATGGAGAGAAAAAAAGGTTGGCATCCTATTTATGTGAGTG AAGCGGTTTTAAGTAAGTTAGATAAAGAGAGAGAAGAAATTAAAGAAGAATTAGGTATTCCAA AGGAAGAGAATTTG

90	GTTCAGCATAAAAGACGGTTTCACGGGCCAAAGCCTAAGCGGCGTAACGGTGAAAGAAGGAGA TACGGTTTTGGGCACGATTGACGACGGCGGGACGCTGGAGCTCACGAGGGGCACTCACACCTTG ACTTTCGAGAAGC
91	CTGATGTTATAGAAGTCCGCAAGGACGGCTCTGTCATCTCGCCCGAGGGTGGGAAATACTATCT CGGCGACATAAGCGGCCCGACACAAATTAGCATCAAGTTCAAGGCCGGCGCGGTGGGAACCCA CGGCTTCACTATC
92	TCTCCCTCAACCTTCGCGGGGAGAACGGCGCGGAGTACTGGACGGGCTACGCGGACGCGCTGG AAGACCTGTTGAAGAAAATCCAGAGGCGGGAGGTGAGGGCATGAGAAGGTATTGTTACATCAC GTGGGGATGGATCA
93	GAGCGCCGGGAGGTGAGGGCATGAGTGAGGAATTGATGTTTGGTCGTGTCGTGGAGTATGTTCA GCATAGTTTCTACAAGAAACCGTTTCCTCTTGGCAGTGAGCTCAAGAATGCAGTAGAGAAGGTT ATGGAAACAGGA
94	AGGTCAGAGCCCACGTGGCAACTTTTGAGGTTCTGACAAAAGACTATGTTCGTGAGAAATACAA AGACATCATAGAGTTCATGAGGGAGAAAGGGACAGTATCGAGAAAGGAACTGCGGAAGAAGT TCTTCTTGCTTGCT
95	GTACCTCAAAATACAGAATCATATTTTACAATCGCTTGGAAATATTAATATCAACAATACGCAA GTCCAAATTAACGTCCCTGGCAAACAGGTGACAATTTATACCCACGAAATACTAGATAACGCCA AAAAGGCACTCG
96	CTTTGTATACTTAGATCAGGAAATGGAGCTAAAAGGCACTATCAAGAAGACAAAAGATTCCTG GAGAGAAACATTTAAAGAGTACTCCAAGACAGACAGCGAATATCTAATAAATTACAGACTGTT TTCAATACTCCCTC
	Primer and Core Sequence
97	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAATTGCTCCCTCGTATCCCTTGTACATTATCTCAGC TCCGCTTAATGATATTAATTTTACCTTGAGTGTTTTTGCTAAAGCCTTTGCCATCATCGTTTTACC TACTCCAGGTGGCCCGTAAAGCAACACAGCTTTGGCACACATCATGTAGTAGACGACCAAGAC AGT
98	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTCTCCAAAACCTACCCAGTTCTCCGAGGAACCTCT TAGCATCTGTTAAATCGTTATTAGTATTAGCTTCCACCATCTCAAGTTCCTTTAAGGCGTTACTC ACACTCTTCTTACCTATCTTTTAGAGAACCACTCGTCAGCACATCATGTAGTAGACGACCAAGA CAGT
99	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTATCAAAGCCCTTAAAGAGTGGTAGGGGCAAAA GTCTGAAGCGTCCTTACTTAACTGGAGTATCTGAGATGGCCTTAATCCGCTTAGGTCTTTAATTT TATCCCTTAATGAACATTCCCTGCACTCTATGTCTTCGGGCACATCATGTAGTAGACGACCAAGA CAGT
100	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGATGTAGCAGACGGGCTAAGAGTTTCAAACCCT CTAAGGATCACTACAAACAAGAGAGAGAGACAATCCTCTCTTTTGTCTTGTCATTGTGTTTCAA ACCCTCTAAGGATCACTACAAACATCTTTAACATAGATACCCACATCATGTAGTAGACGACCAA GACAGT
101	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACCGGACGTTGTGATCACGGGTACCTTGATCTGGT ACTCAAAGGTTTGCCCCCGTGAAGTCTGGTACATGGCTAGACACGTCACTCCATTCGAGGGACA TTCGAAGTTAGAGAAGGGCAGAGCGATACATCAGATATATCACATCATGTAGTAGACGACCAA GACAGT
102	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTCTTTTCTCTACTAATTCTCCTCACGAGATCTCTAA ACATTCTTGCTGAAAGAGGATCCAAACCTAATGTAGGTTCGTCAAGCAATAAAATTGGAGGATC AGTTATTAATGCTCTTGCTAAGGCTAGTTTCCTCTGCATCACATCATGTAGTAGACGACCAAGAC AGT
103	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGATTTTGCCATCATTAAAAACAACAATTTGATCACC CATAGTCATAGCTTCTAATTGATCGTGAGTTACATAAATACTTGTGGTGTTTAACATACGGTGAA TATTTACAATTTCTCTTCGCATGTTTTCTCTTAGTTTAGCACATCATGTAGTAGACGACCAAGAC AGT

104	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATCTTTCAATTCTCGAAAGAAAAGGTTACAAGTC TCATAGATTTATTCCTCTTCACTGTTGTACGTTGGCAGCTAGAGAGAGTTTAGATTATGAGAAAA TTAAGAGAATATATGAGGATTCGTTTTCTTGGTTTAAGTCACATCATGTAGTAGACGACCAAGA CAGT
105	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTAATTGATTTTCCTGTACCATGTGGTAAAACAACG CTACCTCTTAATTGTTGATCTGCTTTTCTAGTATCAAGATTTAATCTAAAAGCTAAATCAACTGA AGCATCAAATTTTGTATAAGAAGTTTTTTTCACTAATTCCACATCATGTAGTAGACGACCAAGAC AGT
106	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGGTTTTCCCGTGAACTAATAAACACCTACTGGAG CCAAGAACGGGTCAGAATTGATGGAATAAACGTTGCGGAGAATGAAATTAATTTGTACATCAG AGACATTGATGACAACGGTGACCCTATACAGTCAACTATACCACATCATGTAGTAGACGACCAA GACAGT
107	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTAATGGAAAGTATGCTTTAGATACCTTCTGGAAC GCTATCTCACTTGGCGGGAATTCAGATATGGAGAGTAAATTAAGGGATCTGGAAGTAAAGTTAA TGTCGTTAATCTATTTAAATGAGTCACCATTAAAATCACCCACATCATGTAGTAGACGACCAAG ACAGT
108	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATAATATGTTAGAGGTAGAATTTCTTTGTGATAGA ATATTATTGATGAATGATGGAAGAGAATTAGCATTAGGAAAACCTAAGGAACTGGTAAAGGAT ACAGAATCTAAGAATCTTGAAGAGGTTTTCCTTAAACTTGTCACATCATGTAGTAGACGACCAA GACAGT
109	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTTACTTCATCTCTCAAGATAAGGGTAATAAGTTC ACTTCAAATATCTGGTCTTATCGCAAGTTGATTGAGGCTATAGTGTATAAGCTCTATGAGTATGG TATAAACGTGTTCCTCGTTGTAGAGTATAACACTTCACGCACATCATGTAGTAGACGACCAAGA CAGT
110	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTCTAGGTTTTAATTCTTCAACTGCTTCAAATACT AGCTTACTGTAGTTATCTGCCCTCATGTTAGGATATATATCTGGAATATAAGGAGGTTGATGAGT TATAAGAAGTGGATGAAATTGTTGTCACACACTCCCCTACACATCATGTAGTAGACGACCAAGA CAGT
111	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTACCTCTTCGGCCTTGTACCAACGTACCCCTGATA CAAGTTCCAAGCAGAGATGGAAAACTCGAAGATGGTATCACCCAAGATGAGATACGATATCAA TGAAGGCGAGCCTAGGTACAAGTAAAGGGATACCACGAGAGCACATCATGTAGTAGACGACCA AGACAGT
112	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGTAAGCGTTTCCTACCCTCGAGAGGGCCATCCT GGTGGTGAGGAAGTCGTCGAAGTGGGCTAAGTAAAAAGCGAAGATCTCGACCCACAATTACCT CCTCCTGTACACCAGGAATACCCCTATCAGGATAGAGATACCACATCATGTAGTAGACGACCAA GACAGT
113	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGCGTCCGGGTCGCGGCCGGGGACGACCGTCTTG ACGAAGTCGGTCGACCCCTCGTCGGTCGAGATGGTCGTCACCTCGGTGTCGAGGCCGTACGTTT CGAGCGCGTCGCGTACCAGTTCGCCGTCCGCGTCGGGACGGCACATCATGTAGTAGACGACCAA GACAGT
114	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATGTACTCGTTCCAGAAGGTGAGTTCGCTCCCCTC GATTTCGACCTCGCCCACGTCGAAGCCGCCGGTCGTTTCGAGCGCGAACGACTCGACGGGACCG ACGAGCGAAACTTCGCCGCCGAGCACGTCGGCGACGCGTTCACATCATGTAGTAGACGACCAA GACAGT
115	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGATGCGCTCGGGCTTGTAGGACTCCCCGAGGGC GTCCTTGTTGGTGAAGACGTTTTGTTTTCGCTCGAACCGGCGCATTAGCGTCGGTCCGTTGTAGC GTCCCCTTATTTAAAACCCCGATTTCATCTGATTCATGTCACATCATGTAGTAGACGACCAAGAC AGT
116	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCACGGTCCGCGACGTGAATCGGGCGTTCCAGTCG GCGTTCGGCTACGACGCCGACGACGTGGTCGGAAGCGACCTCCTCGGGCGAATCGTGCCCCCGG TGCCGGACCCGGACCCGGTGCCGGAACCGGGGGACGACGAGCACATCATGTAGTAGACGACCA AGACAGT

117	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGTCCGCGAGTTCATCCTGAACGTCGTCCCGCTGT CGCCCGGCGAGGAGCGCGGGGCGGGCTACGCCATCTACACCGACATCACGGAGCGGAAGACCC GCGAAAGCGAGCTAGAGCGACAGAACGAGCGATTGGAGGAGCACATCATGTAGTAGACGACC AAGACAGT
118	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGAGACCGGCGACGAGGTGCGCTTCGACACCGCC GAGCGGGCGCTCGAACAGATGGAGGAACTCATCGACGACCTGCTGTCGCTCGCCCGTCGCGGC CAACTGGTCGACGAGACGGAGCGCGTCGACCTCGGGGCGGTCCACATCATGTAGTAGACGACC AAGACAGT
119	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACGAACTCGTCGGTGAACATCTCGTCTTCCGGGGAG CCCGCCGCTCATGGCCTGCCCCCGCCGTAAGCTGCTGCATAAACCCGCTCCAAAATATACGGAT CATTCACCCCTTGGAATCGCTCAATCAGATCAATGTACACCACATCATGTAGTAGACGACCAAG ACAGT
120	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGCGTACATTCCCCCTAAGCGGCTCCCAATATACAG ACGCCGGTTAACGACAGCTGGCGACCCTGTGATCTCAGTACCGGTGTCGAATGACCACATCAGC TTGCCTGTCCGTGCATGGAGTTCGTATACGTACCCGTCGTCACATCATGTAGTAGACGACCAAG ACAGT
121	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGATAGATGAGCCGATCAGAGATCGCTGGTGAGTT GGTAATTGTCCCGACATAGACACGCCAACGTTCTGTTCCATCTGCTGCGTCGTAGGTCGCGAGA TACGGCCAGCCACCAACATACACAATCCCATCGACGAGGACCACATCATGTAGTAGACGACCA AGACAGT
122	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATACACCACCCCATCAGCAACAACTGAATCATGATT AAGTATCGCACCAGCATCGTAGCGCCAGCGTTCACTGCCAGTGGTGCTATCGAATGCATAGAAG ATATGCTCCTAATCGCCAATATCAGTACTTCACAAAGCCGCACATCATGTAGTAGACGACCAAG ACAGT
123	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGACGAGGAGAGGGGCGAGTACATCTGCACGCTT ACGGGAGAGGTAGTTGAGGAGACGGTTATAGATACAGGGCCCGAATGGAGGGCTTACACACCT GAGGAGAGGACCCGCAGAAGCCGCGTGGGCAGCCCGCTTACCCACATCATGTAGTAGACGACC AAGACAGT
124	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTCGATGGCTGCGGCAGCTGTCTATGCTGCCTGCC GTATACGCGGCATACCCAGGAGTATAGACGACATAGCGGAGGTCGTGAAGGGTGGCCGTAAGG AGGTTGCCCGCTGCTACCGCCTCATAGTCCGCGAGCTGAAGCACATCATGTAGTAGACGACCAA GACAGT
125	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGAGTCTTTTGTCACACCGCAGAGGCGTAGCGCT GCAGAGCAGGAGCCCAAGCCTACTGCCAACATAGAGAACATAGTGGCTACAGTATCCCTCGAC CAGACTCTAGACCTGAACCTCATAGAGAGGAGCATACTGACCACATCATGTAGTAGACGACCA AGACAGT
126	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTCGCCTGGGTTAAGAGGATGTTCGGCCTCTCCAA GGCGGGTCACGGAGGCACGCTGGACCCGAAGGTCACCGGCGTCCTCCCCGTAGCCCTGGAGGA AGCAACCAAGGTCATAGGCCTGGTGGTGCACACGAGCAAGGCACATCATGTAGTAGACGACCA AGACAGT
127	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTGGGCGAGATCTACCAGAGGCCGCCGCTCCGCA GCAGTGTTAAGAGAAGCCTCCGCGTCAAGAGGATATACGAGATAGAGCTGCTGGAGTACAACG GCAGGTACGCGCTCATGAGGGTGCTCTGCGAGGCCGGCACATCACATCATGTAGTAGACGACC AAGACAGT
128	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCTGGAAGAACGAGGGCAAGGAGGACCTGCTGCG GAGCTACATCAAGCCCGTCGAGTACGCCGTGAGCCACCTGCCCAAGATAGTTATACGCGATACC GCGGTGGACGCCATAGCCCATGGCGCGAACCTCGCGGTGCCCACATCATGTAGTAGACGACCA AGACAGT
129	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGGAGACCCCAAGGTGACCGGCGTCCTACCAGTGG GGCTCGCCAACAGCACCAAGGTCATTGGTAATGTTATACATAGTGTTAAAGAATACGTGATGGT TATACAGCTCCACGGCGATGTAGCCGAGCAGGATTTAAGAACACATCATGTAGTAGACGACCA AGACAGT

130	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAGAGGGAAAGACTGTAGCTTTCATTCCTAGGCAC GGAAAGAGACACAGAATACCTCCACATAAGATAAATTATAGAGCTAATATATGGGCATTAAAA GAACTAGGAGTGAAATGGGTCATCTCAGTTTCTGCCGTAGGACACATCATGTAGTAGACGACCA AGACAGT
131	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGAGGGAGCTCAGGAGGACTCGCACGGGGCCCTAC AGGGAGGATGAGACACTTGTAAGGCTCCAGGACGTCAGCGAGGCCCTGCTCCTGTGGAGGAGC AACGGGGATGAGAGGTATCTTAGACGCATCGTGCTACCCGTTCACATCATGTAGTAGACGACCA AGACAGT
132	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAACATCTATCGCCCACCTCCCGAAGATAATGATC TTGGATACAGCTGTCGACGCCATAGCACATGGTGCCAACCTGGCTGCCCCAGGCGTCGCCAGGT TAACCAGGAACATCGCGAAGGGTAGTACCGTAGCGATCCTCACATCATGTAGTAGACGACCAA GACAGT
133	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGCTATCCCCGTGTACAGCATGGTGGGGGTGCCGA TGCCCGGGTAGAACTTGGTGACGCTCTCCAGCTTCTCGAGGACGGTTTCCTTGGGGAGGCTCGC GGTGTCCACGAGGGTTATCGCGTCCTCGGCGCCGTCGCCGCACATCATGTAGTAGACGACCAAG ACAGT
134	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAGGACGCGAAGAGCGCGGTGGATGTGGACGCGC CGCCGCACACGTAGCCGTCGAGGTAGCGCGGAACCATCGGCGACATCAGCCCCACGACGCGAC CCGAGGCGTTGCCGAGGATCACGTCGAGCGTCACGCGCGGCACACATCATGTAGTAGACGACC AAGACAGT
135	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGACACCGTGCCGTTGCCCTCCTCTAAGTAGTCG GAAAGCCTCATCCGCGACTCCAGCTTCGCCACCGGCTCCTCGAGCAGGAGGAGGACGCGGTTG ATGCGGTAGGACGCACTGCCCGCCTCCAGCACCGCGCCGTCCACATCATGTAGTAGACGACCAA GACAGT
136	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTATGGTGTAGAACGGGTCGTTGCGGAGCCAGCCT GGCGGCACGTACCGGTCGTCCGCTATCGCCAGCGATCTCTCGAAGAGGTCGAGGTAGGCGGAC GCGTTGGCGAACGCCCCGTGTATCACGACGTCTATCCCGCCCACATCATGTAGTAGACGACCAA GACAGT
137	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATAGGTTTCAGGTATTGATAATGCATAGGAGGTT TTTAAAACCTTGAGCCGCATAGTCTTCTGGATGGGCGAGAGACATGGTTAAGTATAAGTGCGGC AGGTGCGGATACGTCTTCGACGACGAGGAGATGAAGAGGACACATCATGTAGTAGACGACCAA GACAGT
138	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTACGCCGGGTGCGTAGGAGGGCTCGAGTACATC CATGTCTATACTGATGTATGTTTTACCCAGGTCGCCTAGTGCCAGGGGTCCCTTTAACGCTTCCA GGATAGAGTACACGGTGACGTCTCTAGTCTTCTTCAAGAACACATCATGTAGTAGACGACCAAG ACAGT
139	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTACTAGCGTGTCAACGGAGCTCTTCAACGCCTTTA CTATTGGATAGGTTATAAGGTGCTCGCCTCCGAGGAATCCCAGGAGCATGCCGGGATACTCGTC TACAACGCCTTTCACCACGTCACCTATGATTCTTAAAGAGCACATCATGTAGTAGACGACCAAG ACAGT
140	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATAGGTGACATGGGGTTTCCCATTGACTCTATAAA GCCGTATCCTTTAAGCGGAGTGCAATTGGTCTACGCTTTGCTTAACAACAGGTATTTCCTACCGG GTAGAGAGGGCTCGCTCATAGCTTTAGGTAGCGTGACGGCACATCATGTAGTAGACGACCAAG ACAGT
141	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGTATCTCACCGCTTGTCACCATAGTATCCCTCAGG TACTCCAGTATTCTTGAGAGAAACGCACCTAAGCCGGATCTCAGGTTTGAATCCATAAGAACTA TGAGTGAAGCGGGATTGAAGCCCCTGCTGTTTCTAAGACCCACATCATGTAGTAGACGACCAAG ACAGT
142	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAAGGGAGATAGAGAAACGCATCAAAATACCCTTG GGGAAACTGCGTGCAGGGGTTCAATATGGAGTAGAGGTCTCAGACATAAAGGAGAAGATAGCT GCTTACGCTAGGAGGAAGGGGCTTAAATACTTCCCATCGGCACACATCATGTAGTAGACGACCA AGACAGT

143	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTGAACCTCGTGCCCGGCTCTAAGTCGTGAGGGCT TGCAACATAGGTGGGGAGGAACCCGAGCAACGGGTAAGAAGACAGGATAAGCGGTATCGCTAT GAAGAGGGCTGAGAAAAGGACATATACTCCTGAGCCCGTCCCACATCATGTAGTAGACGACCA AGACAGT
144	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAACATGCCTTCCCCGTCTATATAGACCCAGTAGA GTTTAAAAACTTAACCAGAGACGGCTTGTGAGCCGGATCTCTCCCCCGCTAGGCCCTGGATTGG GCTCGCTCCTCCTGGGACCCCGGCCTCCACATGCTCGGGACACATCATGTAGTAGACGACCAAG ACAGT
145	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTGAAGGGCTCGGCTACCCTGAAGACGGGCTTCT GCGCGACCGCCGCGTACTCCGCCGTGGAGCGGTAGAAGAGCGAGGCTGTCTCCGTGAGCCTGA CCATTCCGTACAGGGCGACTGCGACGAGCACTATGACTGCGACACATCATGTAGTAGACGACCA AGACAGT
146	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTCAAGGTGCTGATGCCGAAGGCGACTTTCGACAC CGACGATGCCGCCGACGCCCTGGCCATTGCCATCTGCCACGCGCATCACCGGCACAGTGTTGCC TATAGGATGGCGCTGGCCGGATAAGTTTGTTCTTGACCTGTCACATCATGTAGTAGACGACCAA GACAGT
147	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTCGGTTCGGCAATAAGTAATACCAACGAGGTATT ACCATGCGCGTGACCAGCAAAGGCCAAGTGACGATCCCAAAGGAGATACGGGATCATTTGGGG ATTGGGCCGGGCTCCGAGGTGGAGTTCGTGCCCACAGACGACACATCATGTAGTAGACGACCA AGACAGT
148	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGATCATATGGCCGGCACGTTGGACTTGGGAGG CATGACAACGGACGAGTATATGGAGTGGCTGAGGGGTCCACGTGAAGATCTCGACATTGATTG ACACAAATGTCCTGATCGATGTTTGGGGTCCTGCCGGACAGGCACATCATGTAGTAGACGACCA AGACAGT
149	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGGTGTATTTTACACACCTGGACAGCCAGCATATG ATGCTAGCACTCGGTGTCCCCTTATCACGGTTTCCCGCATTGTAAAGTTTTCGCGCCTGCTGCGC CCCGTAGGGCCTGGATTCATGTCTCAGAATCCATCTCCGCACATCATGTAGTAGACGACCAAGA CAGT
150	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTGGAGCCTGTTAGTTGTTACAGGTTCACCGGTTGT CGGAGTATTCAGATCATTGAGCCAGCAGTTGATGGCTGCCTGTAGTTCACTGGTTGTGATGTAA GCTGCTCCATCGGAATCAACATCGTTCCATGGGTTCCAGTCACATCATGTAGTAGACGACCAAG ACAGT
151	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACGGTCTTGCTTTCTCCTGAATCCATTTCACCTGTCC AGACCCATTCATAGCGGTTAGCTTCACTGAGGTTCTGCTTGAAGACACCGTCATCATTGTTAGAT GAGGTTATTGTCCAGCCGGCAGGAATGACTTCTTCGAACACATCATGTAGTAGACGACCAAGAC AGT
152	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTCAGCAGCTCTTCATAGAAGTTCTGGTTTGCAATA TCCCTCTGGGCAATGACAGGGTAGTCGACTTCGTTTGCAGTCAGGTGGACTGCATACAGGGACT TGCTGATGTCCGGGGTATATCCACTGTGAGGAGCATAGTACACATCATGTAGTAGACGACCAAG ACAGT
153	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACCCGTCAGTCGTGACGTCCTCCGCTCCTCCTATGC TATCTCCACACACCCACTCACGTTCTTGCTTCTTTACTACACCCTCTTTATTCAGCTCTTCGAGAA CATTATTAATGTGACCCTTAGAGATATATTCATTATACCACATCATGTAGTAGACGACCAAGAC AGT
154	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGCCTCCTCAAGCGACTGCTTAAACCCAATTACAT CTGATTTATCCTTTATTTTAGGGCCTATAGAATCTATGAATAATTCGGCGATTCTTATTATTTCTA AAACCAATTCGTCTGTTTTGAGTGGTGTGCCTTCTTCACACATCATGTAGTAGACGACCAAGAC AGT
155	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATCCCATGCATTTTCATAATAATCGGAATTCAAAT CCTCTATATTGAATTTTATCTTAACATTTGACATAATCATTTTCTCCTTACAGAAGAGATCCAGC TAAGCTTACTCATAAATGGTAGTACCATGCCAATATTGGCACATCATGTAGTAGACGACCAAGA CAGT

156	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTAGCCCGCACCTTCCTCTGGTTTAGCACCAGCGG TCCCCACAGAGTACCCATCATCCCGAAGGATATGCTGGCAACAGTGGGCACGGGTCTCGCTCGT TGCCTGACTTAACAGGATGCTTCACAGTACGAACTGACGACACATCATGTAGTAGACGACCAAG ACAGT
157	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTGATAGGCCGCAGATTCATCCTAAGGCGCCGGA GCTTTTGACCACAGAACATTCCAGTATCTATGGTATATCTGGAATTATCACCAGTTTCCCGGTGT TATGCCAGACCTTAGGGCAGATTATCCACGTGTTACTGAGCACATCATGTAGTAGACGACCAAG ACAGT
158	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTTTGGCTTGATACTAATAAAAGCACAGCTAAAAT GAAAATAAGCCGATATTTGTGATTCATGCAACTCACCCTTTTCTACATAAACAAAATACTAACC CGAAAACCGAAATTGAAATTAATGCAGAGAAACCAGGTGACACATCATGTAGTAGACGACCAA GACAGT
159	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTAACGGCACCAACAGTTATTATATTTTTAGCAGTC CCGGGTGAAGTAATTATGGAATAGTTGTTAGAATTACTGTTCTTATTACCAGCTGATTTGAAAGC AATTATACCTGCATCACGAATTGCAGCATCATAATATTCCACATCATGTAGTAGACGACCAAGA CAGT
160	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCTCAGACGACTGAAAAAGCAACGATTGGAATAA TAGGGGGTTCTGGGCTCTATGATCCTGGTATTTTGACTAACAGCAGAGAAATAAAAGTATATAC ACCCTATGGGGAACCTAGCGATTTGATAACGATAGGTAACACACATCATGTAGTAGACGACCA AGACAGT
161	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCAGAACAGGTTCCTTCTATTGGATATTCATCTTC GGCTGCAGTTGCAGGAAGAGTAAGGATATATACTACGGTCTTGCTTTCTCCTGAATCCATTTCAC CTGTCCAGACCCATTCATAGCGGTTAGCTTCACTGAGGTCACATCATGTAGTAGACGACCAAGA CAGT
162	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTCCTCCACGCATTTGTTGTGGTGCTGATGGCGTA TTCTCTGGAATTTGGGATGATTCTGGAAATCCATCCTCAGACACTTCAGATATTTTAGTCTTACT TCCAGCGTTTAATTGAACCTTACCTTTAAAAGCAGTAGTCACATCATGTAGTAGACGACCAAGA CAGT
163	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAACTTACCTTATCAGTGTCATTAAGCATATTGCT TCCAAGACCCATTGAAGCACTTACATCGTTGATACACAGGTGCCAGGAATAGTATTCCTCAGTC TCACTATAATCCTCGTTGGTGTAGCCTTCAAGAGAGTCAACACATCATGTAGTAGACGACCAAG ACAGT
164	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTTAAGCAATTCTTCGGATGAAAGATGGCGCTCTA TAGGAATTTGTTCTGGTCTAGCCATAAGGCATTATTTGTACTTAATTAGTAATAAATGTTTAGTT AATGACTATAAATCTGCAATTGGAGTCTCAAATTTTCAACACATCATGTAGTAGACGACCAAGA CAGT
165	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAACATGAAGGATGTGTGTAAGAGGAAACGTTATTA ACAGACGTAATCAGGAGGATAGTTATGCCCTAAAAACAGCAGAGTTAAGGTTTAAAAATAAGA TAAGAACTCAGTTGAGGTTTATCCATTAATCCCATTAATCCTCACATCATGTAGTAGACGACCA AGACAGT
166	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACTTTCTAAAAGCGCTTGGAGCACGTATCAGGTCAA GTCTTTCAACCTTAAATGCTGCCAGTGCCGTAAGTAGTGCAGTTATGTTGCTTATTGAAACAAAC AACTTAGCCCACTTATTACCTCTTGTCAGTGTTTTTGATCACATCATGTAGTAGACGACCAAGAC AGT
167	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATCCGCTGATATATCCTGGGGATATAGATCGCTC TGAAATGGTTACATCTATCGGTTTTAAGGACAGTTCCAACACTATTGGACCTTGCAGCTATGAC AGGAATAATCTGTTTATCGAGCACAGTTGAATTTGACCTACACATCATGTAGTAGACGACCAAG ACAGT
168	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCAATACCTAATTCTTTCCTTAGAGTGCTATTTTGAT TGAATTCCCTCAGGAAAGATTCAAAATTTAAGTAGCCGAGCTTACATCTTGAAATTTCCATCTTT ATTATGTTGCTCAGGCTTAATGCTTCTAAGTATGGGTTCACATCATGTAGTAGACGACCAAGAC AGT

169	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGATATCCTTTGAAATTCTCGTAATTGCTGAAGGCC ACTACTTCATCAGGTCTGATGCAATCTTTAATCTGAACATTGCTTTCTGAGGTCTTAGGAATAAT CCTGTAAGGGAGTCGGATATTGTTCGTTAAGATGCTCTTCACATCATGTAGTAGACGACCAAGA CAGT
170	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATAGAGGGACCTAGATTTTCAACGAGGGCAGAAAG TAGAATTTGGAGGGAAGTTTATAAAGCCGATATCATAGGGATGACTTTAGTTCCAGAAGTAAAT TTAGCTTGCGAAATGCAAATGTGCTATGCAACAATTGCGATCACATCATGTAGTAGACGACCAA GACAGT
171	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTCTTCAGCATAGTACCAGCTTATGTTGTCACCATC GTTCAGTACGTTACCACCAAGTCCACTGCCTGCAGCTACATCATTAATGTACAGGAACCAGGCA TAGTAACCGCCAGTTGATATGTAGTCTTCACCTTCGATTCCACATCATGTAGTAGACGACCAAG ACAGT
172	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACTCTCCATCATGACAGCCAGATCGGTCATAGCATC GATTGTGTACTCTTCGTCGGGATTGTTGTATGGAATGAACTTATAGTTCTCACCTGCTACCTGAT CCACTGTCATTTCTGCAAGAGTCTGCACTGTGGTAATTCCACATCATGTAGTAGACGACCAAGA CAGT
173	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATATTCCGTATTTCTTATCAAACCGATCGTGAAGAT TTGACAAAGGCTTAACTTTAGGGCTCCACTTCTCATTATTAGCCTTAGAATATAAAGCGTAACCG TAAGCCTGAGGAACGTAAAGCTTAGGAGATTCAATCCCGCACATCATGTAGTAGACGACCAAG ACAGT
174	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAAAATTAGCCGAAGGCTTCCCATTACCGAAAAAG TCGTTTATTAGCTCTTCATCCTTCTTCTCCACGTCCGCCCATTCCTCTCCTTCCCTTGGAATTTTAA GCTCGTCCCAGCTGACTCTTATGGGCAATTCAATATCCCACATCATGTAGTAGACGACCAAGAC AGT
175	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATAAACTTTTGATATAACCTTGCCTAATTTGATA TCATAGCTTATGTTTGGCGCTATCCCCCACTTGTAGAGGGTCGCGTTATATTCTCTAATAGCAAG AGAGATACAAGATTCGTTAACGTTATTTATATCACTCTCCACATCATGTAGTAGACGACCAAGA CAGT
176	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCGGAGGAATCTATCATATTAAACCTCCTCAAAAT CGCCTCCTCTTGATTGCTTAAAGGCTGTGAATTACAAAGCTTATTTAATGCGTCCCAAAGCGTTA AGTAATAATTATTTATATTAAACACTACTATTTCAGTAGCACATCATGTAGTAGACGACCAAGA CAGT
177	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCCTCCTCAATTCAATTGGACTGAAGGAGGGTAC GTTCTGGAAAACAGAGCGTAAAAGAGATATAGAACGTAGTATACACATAGCTGGAAAAAGAAC AATCATTAAGACAATAAAGAACTTTATGGAAAAGAGTAGAACACATCATGTAGTAGACGACCA AGACAGT
178	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGTGTAAAGGTTGTATAATTCAAGCCTCAGAACAT TTCGAACTCCTTACAAAATCGTTTAAACTTTCTAAGGCATAAATTTACTAGAAATTGTCATTTAT GAGAATGTAACTATATAGATGGTAAAATTATTAATCCTCCACATCATGTAGTAGACGACCAAGA CAGT
179	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCTGAAAAATAGGTTCGATCCGCCTCCTCACTTCT TCTCCTTCTTGCCCTCGGCCTCGGAGGAGGCCTCTATTCCCAGCTTCTTGGCCTCCTCCTCGGTCG TCATGAACAGGCTAGTCCTCTGCCTTCCGCCCATGCTCCACATCATGTAGTAGACGACCAAGAC AGT
180	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACCTAGCCTTACGCACAGCCCTCTCCACAACCTCC TCAAGCTTATCCCAGTCAATAGAGCTCATTACAAGTTAACCACGCCCACCTTTAATATAAACCTT TACCCCTCGTGGCAATTAACTTTAACCGCTACTCCGGTGCACATCATGTAGTAGACGACCAAGA CAGT
181	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGGCCCTTAGACCTCTGCCCATGCTTAGGCGCTTAC CCACACCTATTAGTACGGCGCCAATGCCCACGGCCATGAAGTACATTAAGGCACCCATGGTTGC ACCGTAGAGTGCCGTGAATGTTCCGTAGAATACACCGGCCCACATCATGTAGTAGACGACCAAG ACAGT

182	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGGCGAATCTGTCGAGCTCCATGACGTCCACAGA GCCGCCGAACTTGGCCGAGAATCTATCGGCCTGGGCGGTGCGCCTCCCTATCAGCAAAACCCTG GGCGCCGTCAGTAGCGCGACGGCCCTGGCGATTCCCCTGGCCACATCATGTAGTAGACGACCAA GACAGT
183	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCAGGTAGGATCTGGCCGAGAGGGAGGACGCCGC GCTGTTGTGCTCCGGGAACCCTAGAGTCACGACCGCCTTGACGCCTATACGTTCGGCGTATTCA GCGACGGCGGCGCCGGTGCCGCCCGTCAGCGTGACGGCAAGCACATCATGTAGTAGACGACCA AGACAGT
184	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCAAGAGAATACATTTTTGATGATAAGAGAAGCTT GTGGCATACTTTCTTAGGCTTTATTTCAGCATTCACTTTAGCGTATTCTATCGTTATTTTGCTATT GTTCACATTGTATCAAGTGAGAGAAAGAGAGAAGCCAACCACATCATGTAGTAGACGACCAAG ACAGT
185	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGAATCAAAGGAGTGGTGTAAAGATGGAGAGAAA AAAAGGTTGGCATCCTATTTATGTGAGTGAAGCGGTTTTAAGTAAGTTAGATAAAGAGAGAGA AGAAATTAAAGAAGAATTAGGTATTCCAAAGGAAGAGAATTTGCACATCATGTAGTAGACGAC CAAGACAGT
186	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCAGCATAAAAGACGGTTTCACGGGCCAAAGCC TAAGCGGCGTAACGGTGAAAGAAGGAGATACGGTTTTGGGCACGATTGACGACGGCGGGACGC TGGAGCTCACGAGGGGCACTCACACCTTGACTTTCGAGAAGCCACATCATGTAGTAGACGACCA AGACAGT
187	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTGATGTTATAGAAGTCCGCAAGGACGGCTCTGTCA TCTCGCCCGAGGGTGGGAAATACTATCTCGGCGACATAAGCGGCCCGACACAAATTAGCATCA AGTTCAAGGCCGGCGCGGTGGGAACCCACGGCTTCACTATCCACATCATGTAGTAGACGACCAA GACAGT
188	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTCCCTCAACCTTCGCGGGGAGAACGGCGCGGAG TACTGGACGGGCTACGCGGACGCGCTGGAAGACCTGTTGAAGAAAATCCAGAGGCGGGAGGTG AGGGCATGAGAAGGTATTGTTACATCACGTGGGGATGGATCACACATCATGTAGTAGACGACC AAGACAGT
189	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGCGCCGGGAGGTGAGGGCATGAGTGAGGAATTG ATGTTTGGTCGTGTCGTGGAGTATGTTCAGCATAGTTTCTACAAGAAACCGTTTCCTCTTGGCAG TGAGCTCAAGAATGCAGTAGAGAAGGTTATGGAAACAGGACACATCATGTAGTAGACGACCAA GACAGT
190	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGTCAGAGCCCACGTGGCAACTTTTGAGGTTCTGA CAAAAGACTATGTTCGTGAGAAATACAAAGACATCATAGAGTTCATGAGGGAGAAAGGGACAG TATCGAGAAAGGAACTGCGGAAGAAGTTCTTCTTGCTTGCTCACATCATGTAGTAGACGACCAA GACAGT
191	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTACCTCAAAATACAGAATCATATTTTACAATCGCT TGGAAATATTAATATCAACAATACGCAAGTCCAAATTAACGTCCCTGGCAAACAGGTGACAATT TATACCCACGAAATACTAGATAACGCCAAAAAGGCACTCGCACATCATGTAGTAGACGACCAA GACAGT
192	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTTGTATACTTAGATCAGGAAATGGAGCTAAAAG GCACTATCAAGAAGACAAAAGATTCCTGGAGAGAAACATTTAAAGAGTACTCCAAGACAGACA GCGAATATCTAATAAATTACAGACTGTTTTCAATACTCCCTCCACATCATGTAGTAGACGACCA AGACAGT
	Core Sequence
193	CAAATGCTTTCAGTGGTTTCCAATGCCCTGACCAGCCTGTACTCAACTAAGGGATATGACGTAA CATTCAGTGACCTGATCGCCGCCATTCAGGCAATGAAGGGCTACGATGACAGCGCAAACGCTA AACTCGTCGTGGA
194	AGTCGGAGCTTATAACCACAGATCCTGAAGTATTGAGAAAGAGAAGGGGATGGTGAGATAAAC ATGAGGTGTGNNNAAAGCTGACAATATGTACGCGTGCTTGAAGGACACTGTTGTGAAGGAAAG GTATCCTGTCGCTA

195	ATGCAGTATATGGCAGGTGCGAGCAACTGCTTTACCATGGGTGCCATGGTGCAGAACGGGAGA AGCTCCCTCTACTGGAAGGTTAAGGACGCAAATTTCTGGTCCAAGACTTTCGAGAGCAAGTTGC GCGTTCTGGGGCT
196	GGCTGCGGAAGCCAGCAAGGAACCTTGTCTCTTCAGGAGAGGCAGGATATACGTCACATTCAAT GTAAGGAGGGTGGCACTGAGCGGTGCGGGACAGTTGACTGCCGCCAGGACTTATGCCATAGGC AACACCGATGCGA
197	GATGAGGTTAAGGGGATATGCGAGAGATTCGGCAAGGTCGTGGATGCCGTGAACTCTCCTGTGC TTACCGAGAGTAATGCCTCGTACAGGAATGCGGGGCTGGTGCGTGCCAGGTTCAACTGGGACTA CATCAGGCCCGA
198	TGGTCTCCACGGAAGGCTACATGGACAGGGCAATAGGCGTCCAGGATATCGGCTACCTGTTCTG GCAGGCAGGTCCCACCGCAATGAAGGATATGAGAATTTACAACGGTCCCGGTGGTCTGATCGTT CTGCCTTTCTAT
199	CTTTCTGCGCGCAACAGGCTGGCCAAGGGACTTCCGAAGAGCCTGGACATGTTTGCCAGCGTGG AAGGTCGTGACCTTGGGTACGATCCGAGGTACATAACAGAGGAAGATTACAAGACCATTATGA CCAAGGCCCGTCT
200	GGCATGATAGATGAGGATGGCTACGAGGTACCGAAGGGTGAAGACCCCAACGACCCCAGAAGT GCACACACCTTTGGTTGGGTCGACCAATCAGATGGAGGCACATCCAATGGTGGCATGCAGTCCG GTGGGAGCTCTCA
201	CTTAAGGACGGGGACGTGACAAAGCTGTATTCTCAGGACGATTACCTCAGGGTCAGCAGGCTCA AGTTCAGCGAGAATCCGATGCTTGGCATCGTCAAGAATACGGATGGCACAGGGGAGGTTATAG GTCCGTCCTTTGC
202	AGGGACCTTCTGTGGAACATAATCTCGGGTGCCCTGAATGCGGGAAGGGAACAGCTCTACGGG GATGCATTCGGCGGTCCTAAGATAGAGCAGTACGTGAAAGCACTCACGCAGGTGCTGTATGACC TGTCTGTCAACAG
203	TGAAGGAACGCGTCATCGTCCACAAGAGCACTACGAGGAACGAAGTCCTTGAAGAGTTCAAGG CGTCTAAGGAGCCGAAGGTCCTATTTGCGATAAAGATGGAAGAGGGTACGGATTTCAGGGATG ACCAGGCAAGGTGG
204	CAGATATTGGTCAAGACTCCTTACCAGGATCTGGGAGACGAGTGGGTCCGCCTCCATAGGGAGA AGATGGGACGGAGATGGTACGAGATATCCGCCCTCCAGCAGGTCATCCAGGCGAGCGGCAGGA TAATGAGGAACGA
205	CAGGGACTGGGGAGACACCTACGTCCTTGACATGAACGCCATGAAGCTCATCCGCATGTACGA AAAGGAATGCCCCCGCTGGTTTTTGAAGAGGTTGAAACTATGACGCATCACAATATCACCTTCC CCGTCCCTCCCGA
206	GCAATGTCCGCAACAGTTGACAACGTTGCACGGGTCGCGGGATGGCTCAGGGCGACTGCAGTCT CCAGCGATTTCAGGCCCGTCACCCTGAAGAAGTACGTCCTTACCCCCCGCCATATCATGGATGA GAAAGGGGATAC
207	CGAGATACAGCAGATGCTGGGGCGCGCCGGGAGGGCGAAATACGATTCCATGGGCTACGGCTA CATCTGCTCATCCGACGTTCACCTCCAGGACGTGTATAAGACGTACGTTCATGGCCGTCTGGAG AGCGTAAAATCAA
208	GGCACTGGACGAGTTCTTCTCCACCACGCTGGCACGCCACGAGGGTGCCCGTCTGGAAGAATGG ATAGACAACAGCCTTGTCTTCCTGCAGGACAACGACATGATAGTCGGGGGACGTTCCTTCACGG CTACCCCCTTCG
209	CCATCCTTGCGGACTGGATAGACGAGAAGCCCGAAAGCGACATCGTCAATAAATACAACATCT GGCCTGCCGACCTGAGGAGCAGGGTTGAGTTGGCCGAATGGCTCTCGCATTCCCTTTACGAGAT CTCGAGGGTCCTG
210	GGCGACTATTCCTACGTCAGCGTTGCGGAATATTTCTCCAGCTCAAGGATAATAGCCACTACCG CTTCCCCCGGTGGCGACAGGGAGAAGATAAACGAGATCATGCGCCACCTGAGAATAGAGAACC TTGAGGTGAGGGA
211	ATCGACGCTCCAGCTGTTCAGGGACGGTGCGGTCAGGATACTCGTAGCAACGCAGGTTGGGGA GGAAGGACTGGACGTACCGGCTGCAGATACCGTCATATTCTACGAGCCGGTGGCAAGCGAGGT CCGCTCAATCCAGA

212	CGCGGTCCTCTTTCCCAGCTTCAGTCCTTTGGCTTTCCTATGTTCTGCTGGTACTCTTCCCATTGC TCTCTCTGTTGTTTCTCCTGGCTTTTCCTGAAGTTTTCGAGGGCTTCGTCCATGCTGTCATAAGCG TTATGCGA
213	CGAACTCCTTGACGCTCCTGGAGATGACCAGTTTCTCTATCTCCACCCTCCCGCTCTTCATGTCC GATATTATTTTCCTCGCCCTTCTCAGTGCCTCGTCCACGTTCCGGTCGAGCACGAGATTGAACAT TTCCATGAGT
214	CCCTTTCCTCGAGGTCTTTGTCTATTCCCTTCACCGTTTCCCTGGCCCAAGCAGTGATGCTGGAC CCGATACTGGGATCGGTAAACCTGTAGAAGCTTGATGCAAACACTCCGTAGAACGAATTCATAA GCACCTTCACC
215	CAGTTCTCCTTCCACCCCGTCCGCCTCTTCTATGACGGTCGCACTTTCCGTCGAGCAGACGCCCT ACGGCTGGATGGATCAGTATCCCTCGTCGGTTGTTGCCCATGTCACGGGAGGCATCCCTCCCTA CGCCTATCACT
216	CAATCTGGAGGCTTTCGCCGTATCCGTCGAAGTCACGGACTCATCGGGGCATTCAGTCTCAGGT GCAATCATGATCAACTACGGTTCCATAGACCTCTCCCCGTTTGGATACATGGTAACTTTGATTTT TCCGGTGATCA
217	GAACATCCATTGCTGACCATCACGATCGATGGCTCAAAGGACACGTTCAAAACTGGCGATGTCC TGGAATGGTTGACCGAAAGTGACATCTCAAACATGCACAATGTTGCGTCCTTCACAAAATCTCT CCTGAGGATAGT
218	GCATGCTGCCCGATGGCCAATGGTTCGGGGAGGTCATTGGGAAGGACGTGCAGGGAAATCCCT ACGGCATTGATTATACAATGTGGTTGCCGTTTAACACCTACGTTAGGGATAAGCTCAGTTACAA TAGTTGGGGGAAG
219	CCAGCATTCCTCCTCTGAGGGAGTTCGGAAGGTAAAATCTCCTGATGACATCTGAGTCCCTGGC GCCCATTGTCTTGGCTGAGTAGACTTCCATTTTACTTGTCGTGCTGCCAGTTGAAAATGCAAGTA CTGCTATCATC
220	GTTCCCTGCAGGAATGATTGTTCAACTGCACTCGTCAGTACATGATAGAACAATCTTGAAGCTG ACAGATGGAGCGCATAGAAGATAAGGAATGCAAGAGGTACCAGTACTAGTACCACTGCATATA TTCCTGCGTATAG
221	CGATTGCAGGAACGTGGAAAGTGTGCGGTTAATGTATGACTCATTGCTCACATCGAATCGATAC AGAAGACCGCTGTTTGCTGCCAGGTAGGAATCTATGTTATTCAGGCCAACAACCACATCGTATC CCGCCCCCTTTG
222	AGGATTGAACCGGTTGGTGTCAACGTAGAATCCTCATTGACCCGCCACATAAAACTGAACATTC CAATTGTGTCGTCTCCTATGGATACGGTCTCTGAGGCAGATATGGCAATTGCACTAGCAAGACT CGGTGGTATTGG
223	CTTATTATACGCGACCTTTACACTGTAAGCCCGGAAACACCTGTTGACGATGCAATCCGTACTAT GAGGGAGAAGCGAATCGCTGGGCTCCCAGTGATATTGAACGGCAAACTTGTCGGAATACTTAC GAACAGGGACAT
224	GGTACGGCAAGATAGGCTCAGGGAAATTTGTACCAGAGGGAGTTGAAGGAGCAGTTCCGTACA AAGGTAAAGTTGCAGATGCAGTCTTTCAATTGATCGGGGGCCTGAAGTCGGGGATGGGGTATAC TGGCTCGCCCACA
225	GGTGGAAGCGTTGAGGAGTTTGTCACTCTATCGAGGAGAGTGGAGGCAGCGGGATTCGACAAG GTCGAGCTCAATTTGTCCTGCCCACACGTTCAGGGAGTTGGATCCGAGGTAGGACAGGATGTAG GTCTTGTAGAAGA
226	GACACTTATAGACAGGCTAGACAAGAAGACGAAGACAAGGATATTCTTCTCACTTGAGCGATT GATGAAGTGCGGCATAGGGATTTGTGACAGTTGCAGCATCAACGGCATCCGGGTATGCAAGGA CGGAACAATTTTCG
227	CTTCGCAACTGCAAAGAGGTAGCTTCTGGATGCTTCCCTGGAACTATCCCTACATTGCTGTTATC TTACTAGTGGTACTGATTTATGCAGCAATAGAGGACCTTAGGAAGAGGAAAATAACAACTATA ACCTTCCTTGCA
228	GTGACAGTTGGAACTGGTCTATCTCCCCGGTATTTTAATAAGTTTATAGGCGTAGCAAAGGCAT ATACGACAAGAGTAGGGGAGGGGATATTTCCTACTGAGATGTTTGGGGAAGAGGCAGATAGAC TTAGAACCCTAGG

229	GAAGAAGACTTAAAGGATTTAGGTAGAGAGCTTAAGGTACCAAGAAGACCGTTCAAAAAGTTA ACGCATAGAGAAGCTGTTNATATATTGAGATCTCATGGCATAAAAGCAAGTTATGAACATGAG ATACCTTGGGAAGC
230	ACGGGGAGGCTGTCTCAGGAGCTGAAAGAGAATATAGAGCGGAGAAGGTTATTGAGAGGATGA GAGCTACTGGTGAGAACCCTGCAAAATACGGTTGGTACATTGAAATGTTGAAATATGGTATTCC GCCGAGTGCAGGG
231	ATATGCAGATTTAGATGAGATTATAGGGGTTGCATCTAAGGCAGGAATAGATTGCATAACTATA GATGGGTCAGAAGGTGGAACAGGTATGAGCCCTATAGCTGCGATGAGAGAACTAGGATATCCA ACGCTAGTATGTC
232	GGACACGAAATTGCTGAAGCAGCTGGCTCAACATGGTATATCGACAATTTCTGGGATAAACTCA AAGAGGGCTGTGTAGCATATCTAAACATAGATTCACCTGGATTAAAAGATGCAACAAGATATAT CGCTTACGCGTC
233	GTAACTTCTGGAAACGCCCAATCAAAACAGATCATGACACCAAAGCTAAAATTATCTTCCCTAA TAGCTTCTATAGGTGTATCTCCAGGTTGAAATATTAGCTTCTCTTTGGCAAATAAGTGAAGTTTC CTATACTTTCC
234	CCAGATAGCCCAATAGCATCAATTTCCGTTGCAATAATAGGTACAGTACACAAAGAACACGTAA TTTTCAGCGACACTGCAAATACAGGCGACTTAATAATTTTTGCCATAGATCTCGATGGAACATTT CACCCTAAGTT
235	GTTCTAATTCCTCTCTTACAGCTTTAAAAGCAATCACAGCAGATTCCAAAATATCATCCATATCA TCCAGAGCTATAATAACACCTCTTGAAGTTTTCCCAATCTTATGCCCACTTCTTCCAACTCTTTG AACCAAACGA
236	GTAACTTGTCTGGGAGACATATATTGGACAACTAAATCAACGGTTCCTACATCAATCCCTAACT CCATAGATGATGTACAAATAAGACCTTTCAACTCACCGTCTTTAAATAACCTTTCAACTTCTATA CGAACATCTCT
237	ACTCAATGAACCATGATGCACATCAATACTTAGATTAGGATCGTATAAGTGAAGCCTAGAAGCT AGTATCTCAGCTATTTCACGAGTGTTTACAAAAGTAAGCATAGAGCGGCTCTTTTCTAATAACTC AACCAATACCC
238	CAGTTAAATCATCTTAACTCACAAATATTAAGGCTTTAATTTCTGAGGGAGTGCAAAATGAAAA CTGACGTAGTAATAGTAGGTGCAGGGCCCGCAGGCATGTTTGCTGCACATGAATTGGCAACTAA ATCTAATCTGAA
239	AAAAATAGCCAAGGATCCAAAATTCCGTGTATATACAAAAACCTTCGATGACCTTACACGTGTA TTTTGCGTTAATTATCGAGGCTTCGTCGTCCAAGAAGTCTACGGAGATATCGTTGGTGTTAACGG CCACACTCTAA
240	TCAAACAAAAATCTGAAAATGCCAATTTTGCATTTCTAGTTCGAGTTGAACTCACCGAACCGCT TGAAGACACAACCGCCTACGGATTCTCAATAGCCAAATTAGCAACTACCATAGGTGGAGGAAA ACCAATTCTTCAA
241	CGAGATACTGAATTTCCAAAACTCAAAGGATATAGAATTGTTAGAATCGCAACACATCCGCAAG TTATGAGCATGGGACTAGGAAGTGAAGGGTTGTCAAAACTTTGCCAAGAAGCCGAAAAGAGAG GACTAGATTGGGT
242	CGAAGTTTTTATCCTCCTCGGTCCAAGTCACACTGGTTACCCAGGCGTTGGAATAATGACAGAA GGCATCTGGAAAACTTCTTTAGGAGAAATATCAATAGATGAAACTCTCTCGAATACTATTTTAA ATAATTGTGACC
243	TGACACACTACGGCACCTACTATGGATACACACCAGCTGGTGTTGAACCATTAACCAAAGTTTT AGAATGGATATACCAGACGGACAAACAAGTTATTGAGAGAATTAAAAGATTAGATGGAGCAGG AGTAATAGAATAT
244	CTGAAAAGTTCATTCCAATTGTTAAATCGCCATCTTGGAAACACGGCACAAGAAAAGGGAAAG GATTTAGCATCGGTGAGATTAAAGCAGCCGAGATAGATATTAGTATGGCAGTTAAACTCGGTAT ACCCATTGATAAA
245	GGGAATAATAATTAAAATAATGTGGCACACCTTTTAGCTTCTTTTCATCTCATATTTTCAAAGAA GCCTTCCAGGTGTGCCTCATCGGTGTCCCCCGCTGCGGAGACACGGTATCATCGTATCCGCCGA AGGAAACTCAA

246	GACATTGCCTATCAATTACTTCAAGCCGGAATGCAAGTTCCCGGTTTCAGAAGGTCGCCAAAGA TAATAGAAAGAATTTTAGAAAGATATATTCCAACAGTCACCGTACTAGGCGGCATTATTGTAGG ATTAATAGCTGC
247	TGTCGTTCAGGGAGGTATAAAAATGCCAGAACCACGCTACCGGTCAAGGTCTTTAAGAAGACG ATACGTACACACACCTGGAGGAAAAACCGTCATCCATTACAGGAGAAAAAAACCTGACGTTGC AAAATGCGCATTAT
248	GTGGTCAACCTCTCAGAGGAATTCCCAGACTAAGGCCAGGAGAATTCAGAAAGTTGACAAAAA GTCAACGAAGACCAGAGAGACCTTTCGGTGGATATCTATGCCACAAATGCTTAGCAATGGAAAT CAAGAAAGCTGTT
249	ATAGGATGAATCTAACTGGGGCGACCCGGTAGATAACTGAGAGTGTAGGAGGTGAAATAATTG AGCGCAATAGAAGTAGGTAGAATATGTGTTAAAACTAGTGGAAGAGAAGCAGGAAGAAAGTG CGTTATTGTTGAAAT
250	ACACCATTTCCTAATATTTTAGTAACTAGATATGTTTGTTATAGTATTAGGGTGAAGTATTTGTA TGAAAGAAAGTTGCCATCAGACATTAAAAGAGAGATTCTAGTAAAAAGTGAAGCAGAAACTGA CCCTGCTTATGG
251	CACATGAGAGAACTTAGAAGAACACGTACAGGACCCTTTAAAGAAGATGAAACCCTAGTAACT CTTCACGATGTAGTTGATGCTTACTATTTTTGGAAGGAAGATGGAGAAGAAGAATTTCTACGAA AAGTCATACAACC
252	AATGGAAAAGGGTTTAGAACACCTACCTCACATTTGGATTAGAGATTCTGCTGTAGATGCAATA TGCCATGGGGCAAACTTAGCAGCTCCTGGTGTTGTAAAACTTCATGACGGTATATCACCTGGAG ACTTAATAGTAA
253	CGCTGATCATACATGTGCATTGTCTTTAAATACACTAGTAACGTTAATAATATCTAGCAATTTTA GATAAAAATAACTAGCAGTGCCGGGGTAGCCAAGTGGACTACAGGCCTTATACCGGTTAGGGC GCGGGCCTGGAG
254	CATGCCTTAACGAGAGGCATGGGATGGGGGAGCTGTGAGCCCCCCGAACCGGCAGATGAGGGG AAGGGTGCAAAGCATCCCTTAACGCCGGAAGCTCCCGACTTCAGTCGTGGAGCAGCTCACTGCT TTGACGAAAGGTT
255	GAACTTGCAAGGAAGGCCGGTGTTGATTATGAGACAAAGCTGTTGGTCAGGGGCAAGGAACCG GCTGAGGACATAATAGAATTTGCTGACGAGATCAGGGCAAGTCTCATTGTAATAGGGGTTAGG AAGAGGAGACCCGC
256	TCCAGAAGAGATTCAAAGCTCTCGTATTCAATGTCCCCACCAAATTTCTGGTCGCGCTCAATTTT GACTTTACCAAAAGCGGGGAAAACGTAGTGCTTTGCTAGGTCTATTATCGGATTTCCTTCTACA ACCTTTGGCGG
257	GATTTGCTCATTTTCTCCCCGTCGAGTCCTGAGATTATCGGCGTATGGATGCAGATCGGTGCCTT GTAACCGAGGGCCGGCAGATTCTCCCTTGCGAGCATGTGGATCTTTCTCTGATCTATTCCACCAA CCGCCACATC
258	TCCGGGAGTTGCAGAACCAAGCATGGAAATTGCTAGAGATCCCGAAAAGGTTTACGAGTACAC GAATAAGTGGAACACGGTTGCAATTATCACTGATGGCTCGAGGGTCTTGGGACTGGGCAACATC GGTGCGATGGCTT
259	GTGGTGTTATCAAGAGGGAATATATTGCTCAGATGGCAGAGGATCCGATAGTCTTTGCCTTATC AAACCCGGTGCCTGAGATCTATCCGCAGGAGGCAAAGGAAGCCGGAGCCAGGATCGTAGGAAC TGGTAGGAGCGAC
260	GGGATCTGTTAGTATGGCATTCAGAGCCTTTATGTCCTCATCGGTAAGCTTGTCCGATGGCAGAT CGTATTTCACGATGTCTGAAGGAGTAACTCCGAGAAACTTCGCTTCTGGTGTCGCAAGATACTC CGAGAGATGCG
261	TGAGTGCGGCTTACTCTGCACTGTGCGAGATCGATGAGGTCGTTGTTGTTGCCCCCATAACGCA GATGAGCGGAGTGGGGAGGAGCATATCCATAATGCGGCCGGTTCGTTTTTTCGAGCTCGAAATA GATGGCATGAGG
262	AGGGGAAGGGAGTACTACTGGATTCATGGGGTGGAAGTCGAAAGCGCTGAGCCTGGAACGGAC ATACACGCACTCAGAAACGGGTATGTCTCCATTACACCGATATCCTTAAATGCAACTTCGGACT GCGAAGCTTTAAG

263	ATAGTTTTATGGAGGGTGGTTGGACATGAATGAAAGGGCAAAGAAGGTCATTCTTATTGTGGAT GACGATTTGGCTCTGCTTGAAGCTCTTGAACTGATGCTTCGAGGCAAGTATGAGGTTGTGAAGG TGACAAATGGGA
264	ATGTCGATTCCGAAATAGCAGGGAGCAATTATCGGTGGGCTTCCGACCCTTAAATGGATTTCCT TCGCTCCCGCCTTTCTTATCATGTCGACTATTCTTTTGGATGTTGTTGCCCGCACAATGCTGTCGT CAACCAGCAC
265	ACTTTCTGAGGGAAAAACATTGTTGCTTATCCTAAAGAGTTTACAAGCAAGAAGCTGGAAACAA ACTCTGGATGTTATTAATTTAGAGCCTGCAGCAGCATATACAATGTTTAGAGCGGCAATAAAGA AACTATACAAAG
266	GTGGTTGAGAGGCTGCTTGAAGGCATTGCAAAGAATGAAAGGGTAGCTTACGGATTGGAGGAG GTTAGGAGGGCAAAAGAGTATGGAGCAATTGAGGTTCTGTTGGTTTCAGATGACTTCCTGCTCA CCGAGCGTGAGAA
267	TCGCTTCGAGATTCCTGATAGGAGTGGGAGTTGCCGGGGTTTACGTGCCTACGATAAAAATAAT ATCCGTCTGGTTCAGGCAGAATGAGTTTGCAACTGCTACTGGGATTCTTTTCGCGATTGGAAATC TAGGAGCGATT
268	GAGGTATCGCCTACTTAGAGAGTTCGTAAAGTCGGAGATATTGGAGGAAGTTAAATTTGAAAAC GTTGTGGACGAGTACTGGGTTGCGGAACCATTCATAAAGATCATAATTTTTGAGGATCTCGAAA ACCAGAAATTGA
269	CTAATCCGATTATCGATTCTACGCTTCCTGATGGTAGCAGGCTTCAGGCTACCCTAGGAACAGA AATTACACCTAGAGGCTCGAGCTTCACGGTGAGAAAATTTACAACCCAGCCACTGACCCCGTTA GATCTAGTGAGG
270	CAAAATTATATCGATAGAGGATACCAGAGAGATAAAGCTCCATCATGAGAACTGGCTGGCTCA GGTGACGAGAACGGGGATAGGAGAGCAGGAAATTGACATGTATGACCTTCTCAAAGCCGCCTT GAGACAGAGACCGG
271	GAATCAGTTTGTTAAATGGGATGCGAAGAAAAATTCGCATGTTGAGGTAGGGATTCCGAAAAA GCTAGAGAAAATCGCGATGTCGAGAGTGGACGATGCTTACGCGGAGCTGGAAAGAAGAAGGA GGTATTTGGAGTGGA
272	TCAGTGAAGTTAGCACGGAATTCGAAAGGATAGTGGTTCTCGTTGAAATGGGAGAGGATTTGG AAAGCGCAATGAGGTTTGTTGCAGAAACAACTCCCTCAGAGAGGCTCAGGGTTTTTCTGGAGAA CTTTATTGATGTG
273	GCTGGAGCGGGAGGCGTATCAACGCTTGCCCTCAATCCGTTACCCGAAGTTCCAGAATACTTTG AGTATTTCCAGTCCGAATAGAAGCAGAGCACCTCTCGATCGACTAGAGTCTTTCTGCTAGCTCTT GCACCCTCATC
274	GCGGAAATCTCTGCTGAAAACACCTTGACTTTTTCTTCGTATATCTCCCATTCCATCAGGCACCA CCAACTTTGGTCCTGCAAAGAGTCATCGGTGCCCCATCTGCTACGGGAACGATCTGAAAGGCTT TACCACAGAAT
275	TCCGGTTGCAGGATTGGTCTCCCCACCTCTCGAGCCTATGAGGAATACCCCATTCCTGCAGAGCT CGAGAAGCTCTTCGAATTCAAGATCCCCCTTCTGCAGGAATGTGTTGCTCATTCTGACAATCGG AAAAGCAACTC
276	AACTCCTCGATTGTTGGGTCATCGATTATTGTCACGTTCTCTCCTGCAATTCTCTCTCCAATCTTT CCAGCAAGAACGCTGTTTTCCTGCAGAACGTGATCTGCCTCGACCGCATGCCCGAAAGCTTCGT GAATAAAAAC
277	CTTTCTTCAAGAATGCTTTCTGCGGCGATAAGCCCAGTAACAGCCGCTCCAACTATTCCCCTGCT TATTCCGGCTCCATCGCCAATTGCATAGATGTACGGTATGCTTGTCCTCATCTTCTCGTCAACCT TAAGCTTCAA
278	TGCTGGATTTTTCTTTGGCCTTGCTGTGGCCGTTGACTAGACAAAAGTCGCCGTACTCCTCTCTT ATAACCCAGCCCCTCGGGCAGGTGCAGAACGTGCGCATGTAGTCGTCATGCCTCTGTGTGATTA TTCTCAGCTTT
279	TTGCCTTGGAATTTTCCGCCACTTCGATCTTATACTTTTTTACCCATTTTTCCAGCCAGTCGGCAC CGCTCCTCCCAACTGCAATTATGAGTTTGTCGTAGCCGAACTTGTCCCCATCGTTCGTCTTCACG ATCTTTTCT

280	AATCACCACCGACTGTGAAGCTCGAAGGATAGTTGGGGTTGGCATAATTCAGCTTTCCATCCGA AAGTCCTCCAGCACCACCCACACCAGAAGTAATGTTGCAGGGATCGCATTTCTTGCAATAGCTT TGCGAAAGGTCA
281	TTAACCAACCTCTTTCGCATCAAAATCCCAACTGCGGCATCCGTTATCAGCGTTACATCGATTCC ATCTTTCATAAGCTCGTAGCAGGTGAGCCTAGAGCCTTGGTTCAGCGGCCTCGTTTCGCAGGCG AAAACCTTTAC
282	TTATCGAGTTAATAGCTATCAGTGTTGCTATTACGATCGTTGCGATCCCATCAAAGATGTTATGA CCGAAGGAGATAGCAATAATGCCAAATATCGCTGCAAGCGTCGATAAGGAGTCGTTAAAACTC TCAAACATCACT
283	CTCCCTTCTAAGCTTCGTGATATCTGCATTGCCAATATCAACTAGAAATTCGATTGAGATAAGCT TGTCTCTTGCGGTTAAGCTTGTTCTCTCGATATTTATACCGAAATTTAGCAATACACCCGTGATA TCTCTCACGA
284	AAGCGGGGCTTTTGCCTTTCCAATTCCGCCGCAACCAACCGTTGGACTTATCAAACCGGAACCT TTCAACTCCGAGATTAAAGAGCCTGGCTCCTTATCGTGCTTAATAGCAATTTCTACAATGTCTTC CCCGCATACAA
285	CTAGTTCTTGGTTTTCGTCGACGTTGACCTTGTAGAACTCTACATCTGGAAACTCCTTTGAAAGC TTTTCGAGCACTGGGCTGAGATACCTGCACGGCATGCACCAGTCGGCGTAGAAGTCAACAACAA CAAGCTTATCC
286	CTCCGATCGTCTTTAAAGCTTGCAAGTCTAAATCCTCGCCCCAGGGAATTTCCTGGGATTTTCTC GCAATCTCTATCGCCGAAGTATAGGTTATCCTCGGGAATGGTATCTCGGGGACTTCGAGCTTTA GTTCGAGAATA
287	GAGGTTTTGTCCCTATTGGGTTTCTCATTGCCTGCAGCATTTCTTCTCTGCTCAGAGCTCTGCAGC CATCGCCTTTCATTCTTAAAATGCTAACCTCCCAATCATCCGGAAAATCGAGCTCTATTTCCCTA TCCTGCCAG
288	TGTTTACAGGCTGGTGGGTGGGGAAAGGAGTGTTAAGGGCAAAAGGAGTGTAAGCAAGTTCAG GGTTGCGATTGCGATTCTTCTGGCATTCATTCTGATATATCCTACATACCGCATAGCCGAGATTC AAAGCAGTGGGG
289	CAGGGTCAGGAGGATTCACGAGATAGAAGTCCTCGAGGTGAGAGGCAGGTTCGCGCTTATAAG GGTTCTCAGCGACCCCGGCACGTACATGAGGAAGCTGGCCCACGACATCGGGCTATTGCTCGGA GTAGGTGCACACA
290	GAAGTCAGTCATAGAATCAATGGTGTATTCTTCATCAGGGTTATTATACGGAATGAACTTATAG TTCTCACCTGCTACCTGATCCACTGTCATTTCTGCAAGAGTCTGCACTGTGGTAATTCCACCTTCT TCCATCCGGG
291	AGTAAGGGAATCAATGTCTTCCATTGCTGTAAGGGTTACTGTTACCTTTGTAGAAGTCAGACCG TAATTGGTCAGCAGCTCTTCATAGAAGTTCTGGTTTGCAATATCCCTCTGGGCAATGACAGGGT AGTCGACTTCGT
	Primer and Core Sequence
292	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAAATGCTTTCAGTGGTTTCCAATGCCCTGACCAGC CTGTACTCAACTAAGGGATATGACGTAACATTCAGTGACCTGATCGCCGCCATTCAGGCAATGA AGGGCTACGATGACAGCGCAAACGCTAAACTCGTCGTGGACACATCATGTAGTAGACGACCAA GACAGT
293	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTCGGAGCTTATAACCACAGATCCTGAAGTATTG AGAAAGAGAAGGGGATGGTGAGATAAACATGAGGTGTGNNNAAAGCTGACAATATGTACGCG TGCTTGAAGGACACTGTTGTGAAGGAAAGGTATCCTGTCGCTACACATCATGTAGTAGACGACC AAGACAGT
294	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATGCAGTATATGGCAGGTGCGAGCAACTGCTTTACC ATGGGTGCCATGGTGCAGAACGGGAGAAGCTCCCTCTACTGGAAGGTTAAGGACGCAAATTTCT GGTCCAAGACTTTCGAGAGCAAGTTGCGCGTTCTGGGGCTCACATCATGTAGTAGACGACCAAG ACAGT

295	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCTGCGGAAGCCAGCAAGGAACCTTGTCTCTTCA GGAGAGGCAGGATATACGTCACATTCAATGTAAGGAGGGTGGCACTGAGCGGTGCGGGACAGT TGACTGCCGCCAGGACTTATGCCATAGGCAACACCGATGCGACACATCATGTAGTAGACGACCA AGACAGT
296	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGATGAGGTTAAGGGGATATGCGAGAGATTCGGCAA GGTCGTGGATGCCGTGAACTCTCCTGTGCTTACCGAGAGTAATGCCTCGTACAGGAATGCGGGG CTGGTGCGTGCCAGGTTCAACTGGGACTACATCAGGCCCGACACATCATGTAGTAGACGACCAA GACAGT
297	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGGTCTCCACGGAAGGCTACATGGACAGGGCAATA GGCGTCCAGGATATCGGCTACCTGTTCTGGCAGGCAGGTCCCACCGCAATGAAGGATATGAGA ATTTACAACGGTCCCGGTGGTCTGATCGTTCTGCCTTTCTATCACATCATGTAGTAGACGACCAA GACAGT
298	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTTCTGCGCGCAACAGGCTGGCCAAGGGACTTCCG AAGAGCCTGGACATGTTTGCCAGCGTGGAAGGTCGTGACCTTGGGTACGATCCGAGGTACATAA CAGAGGAAGATTACAAGACCATTATGACCAAGGCCCGTCTCACATCATGTAGTAGACGACCAA GACAGT
299	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCATGATAGATGAGGATGGCTACGAGGTACCGAA GGGTGAAGACCCCAACGACCCCAGAAGTGCACACACCTTTGGTTGGGTCGACCAATCAGATGG AGGCACATCCAATGGTGGCATGCAGTCCGGTGGGAGCTCTCACACATCATGTAGTAGACGACCA AGACAGT
300	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTAAGGACGGGGACGTGACAAAGCTGTATTCTCA GGACGATTACCTCAGGGTCAGCAGGCTCAAGTTCAGCGAGAATCCGATGCTTGGCATCGTCAAG AATACGGATGGCACAGGGGAGGTTATAGGTCCGTCCTTTGCCACATCATGTAGTAGACGACCAA GACAGT
301	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGGACCTTCTGTGGAACATAATCTCGGGTGCCCTG AATGCGGGAAGGGAACAGCTCTACGGGGATGCATTCGGCGGTCCTAAGATAGAGCAGTACGTG AAAGCACTCACGCAGGTGCTGTATGACCTGTCTGTCAACAGCACATCATGTAGTAGACGACCAA GACAGT
302	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGAAGGAACGCGTCATCGTCCACAAGAGCACTACG AGGAACGAAGTCCTTGAAGAGTTCAAGGCGTCTAAGGAGCCGAAGGTCCTATTTGCGATAAAG ATGGAAGAGGGTACGGATTTCAGGGATGACCAGGCAAGGTGGCACATCATGTAGTAGACGACC AAGACAGT
303	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGATATTGGTCAAGACTCCTTACCAGGATCTGGGA GACGAGTGGGTCCGCCTCCATAGGGAGAAGATGGGACGGAGATGGTACGAGATATCCGCCCTC CAGCAGGTCATCCAGGCGAGCGGCAGGATAATGAGGAACGACACATCATGTAGTAGACGACCA AGACAGT
304	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGGGACTGGGGAGACACCTACGTCCTTGACATGA ACGCCATGAAGCTCATCCGCATGTACGAAAAGGAATGCCCCCGCTGGTTTTTGAAGAGGTTGAA ACTATGACGCATCACAATATCACCTTCCCCGTCCCTCCCGACACATCATGTAGTAGACGACCAA GACAGT
305	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCAATGTCCGCAACAGTTGACAACGTTGCACGGGT CGCGGGATGGCTCAGGGCGACTGCAGTCTCCAGCGATTTCAGGCCCGTCACCCTGAAGAAGTAC GTCCTTACCCCCCGCCATATCATGGATGAGAAAGGGGATACCACATCATGTAGTAGACGACCAA GACAGT
306	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAGATACAGCAGATGCTGGGGCGCGCCGGGAGGG CGAAATACGATTCCATGGGCTACGGCTACATCTGCTCATCCGACGTTCACCTCCAGGACGTGTA TAAGACGTACGTTCATGGCCGTCTGGAGAGCGTAAAATCAACACATCATGTAGTAGACGACCA AGACAGT
307	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCACTGGACGAGTTCTTCTCCACCACGCTGGCACG CCACGAGGGTGCCCGTCTGGAAGAATGGATAGACAACAGCCTTGTCTTCCTGCAGGACAACGA CATGATAGTCGGGGGACGTTCCTTCACGGCTACCCCCTTCGCACATCATGTAGTAGACGACCAA GACAGT

308	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCATCCTTGCGGACTGGATAGACGAGAAGCCCGAA AGCGACATCGTCAATAAATACAACATCTGGCCTGCCGACCTGAGGAGCAGGGTTGAGTTGGCC GAATGGCTCTCGCATTCCCTTTACGAGATCTCGAGGGTCCTGCACATCATGTAGTAGACGACCA AGACAGT
309	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCGACTATTCCTACGTCAGCGTTGCGGAATATTTC TCCAGCTCAAGGATAATAGCCACTACCGCTTCCCCCGGTGGCGACAGGGAGAAGATAAACGAG ATCATGCGCCACCTGAGAATAGAGAACCTTGAGGTGAGGGACACATCATGTAGTAGACGACCA AGACAGT
310	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATCGACGCTCCAGCTGTTCAGGGACGGTGCGGTCA GGATACTCGTAGCAACGCAGGTTGGGGAGGAAGGACTGGACGTACCGGCTGCAGATACCGTCA TATTCTACGAGCCGGTGGCAAGCGAGGTCCGCTCAATCCAGACACATCATGTAGTAGACGACCA AGACAGT
311	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCGGTCCTCTTTCCCAGCTTCAGTCCTTTGGCTTTC CTATGTTCTGCTGGTACTCTTCCCATTGCTCTCTCTGTTGTTTCTCCTGGCTTTTCCTGAAGTTTTC GAGGGCTTCGTCCATGCTGTCATAAGCGTTATGCGACACATCATGTAGTAGACGACCAAGACAG T
312	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAACTCCTTGACGCTCCTGGAGATGACCAGTTTCT CTATCTCCACCCTCCCGCTCTTCATGTCCGATATTATTTTCCTCGCCCTTCTCAGTGCCTCGTCCA CGTTCCGGTCGAGCACGAGATTGAACATTTCCATGAGTCACATCATGTAGTAGACGACCAAGAC AGT
313	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCCTTTCCTCGAGGTCTTTGTCTATTCCCTTCACCGT TTCCCTGGCCCAAGCAGTGATGCTGGACCCGATACTGGGATCGGTAAACCTGTAGAAGCTTGAT GCAAACACTCCGTAGAACGAATTCATAAGCACCTTCACCCACATCATGTAGTAGACGACCAAGA CAGT
314	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGTTCTCCTTCCACCCCGTCCGCCTCTTCTATGACG GTCGCACTTTCCGTCGAGCAGACGCCCTACGGCTGGATGGATCAGTATCCCTCGTCGGTTGTTGC CCATGTCACGGGAGGCATCCCTCCCTACGCCTATCACTCACATCATGTAGTAGACGACCAAGAC AGT
315	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAATCTGGAGGCTTTCGCCGTATCCGTCGAAGTCAC GGACTCATCGGGGCATTCAGTCTCAGGTGCAATCATGATCAACTACGGTTCCATAGACCTCTCC CCGTTTGGATACATGGTAACTTTGATTTTTCCGGTGATCACACATCATGTAGTAGACGACCAAG ACAGT
316	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAACATCCATTGCTGACCATCACGATCGATGGCTCA AAGGACACGTTCAAAACTGGCGATGTCCTGGAATGGTTGACCGAAAGTGACATCTCAAACATG CACAATGTTGCGTCCTTCACAAAATCTCTCCTGAGGATAGTCACATCATGTAGTAGACGACCAA GACAGT
317	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCATGCTGCCCGATGGCCAATGGTTCGGGGAGGTC ATTGGGAAGGACGTGCAGGGAAATCCCTACGGCATTGATTATACAATGTGGTTGCCGTTTAACA CCTACGTTAGGGATAAGCTCAGTTACAATAGTTGGGGGAAGCACATCATGTAGTAGACGACCA AGACAGT
318	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCAGCATTCCTCCTCTGAGGGAGTTCGGAAGGTAA AATCTCCTGATGACATCTGAGTCCCTGGCGCCCATTGTCTTGGCTGAGTAGACTTCCATTTTACT TGTCGTGCTGCCAGTTGAAAATGCAAGTACTGCTATCATCCACATCATGTAGTAGACGACCAAG ACAGT
319	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCCCTGCAGGAATGATTGTTCAACTGCACTCGTC AGTACATGATAGAACAATCTTGAAGCTGACAGATGGAGCGCATAGAAGATAAGGAATGCAAGA GGTACCAGTACTAGTACCACTGCATATATTCCTGCGTATAGCACATCATGTAGTAGACGACCAA GACAGT
320	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGATTGCAGGAACGTGGAAAGTGTGCGGTTAATGT ATGACTCATTGCTCACATCGAATCGATACAGAAGACCGCTGTTTGCTGCCAGGTAGGAATCTAT GTTATTCAGGCCAACAACCACATCGTATCCCGCCCCCTTTGCACATCATGTAGTAGACGACCAA GACAGT

321	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGATTGAACCGGTTGGTGTCAACGTAGAATCCTC ATTGACCCGCCACATAAAACTGAACATTCCAATTGTGTCGTCTCCTATGGATACGGTCTCTGAG GCAGATATGGCAATTGCACTAGCAAGACTCGGTGGTATTGGCACATCATGTAGTAGACGACCAA GACAGT
322	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTATTATACGCGACCTTTACACTGTAAGCCCGGAA ACACCTGTTGACGATGCAATCCGTACTATGAGGGAGAAGCGAATCGCTGGGCTCCCAGTGATAT TGAACGGCAAACTTGTCGGAATACTTACGAACAGGGACATCACATCATGTAGTAGACGACCAA GACAGT
323	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGTACGGCAAGATAGGCTCAGGGAAATTTGTACCA GAGGGAGTTGAAGGAGCAGTTCCGTACAAAGGTAAAGTTGCAGATGCAGTCTTTCAATTGATCG GGGGCCTGAAGTCGGGGATGGGGTATACTGGCTCGCCCACACACATCATGTAGTAGACGACCA AGACAGT
324	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGTGGAAGCGTTGAGGAGTTTGTCACTCTATCGAG GAGAGTGGAGGCAGCGGGATTCGACAAGGTCGAGCTCAATTTGTCCTGCCCACACGTTCAGGG AGTTGGATCCGAGGTAGGACAGGATGTAGGTCTTGTAGAAGACACATCATGTAGTAGACGACC AAGACAGT
325	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACACTTATAGACAGGCTAGACAAGAAGACGAAGA CAAGGATATTCTTCTCACTTGAGCGATTGATGAAGTGCGGCATAGGGATTTGTGACAGTTGCAG CATCAACGGCATCCGGGTATGCAAGGACGGAACAATTTTCGCACATCATGTAGTAGACGACCA AGACAGT
326	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTCGCAACTGCAAAGAGGTAGCTTCTGGATGCTTC CCTGGAACTATCCCTACATTGCTGTTATCTTACTAGTGGTACTGATTTATGCAGCAATAGAGGAC CTTAGGAAGAGGAAAATAACAACTATAACCTTCCTTGCACACATCATGTAGTAGACGACCAAG ACAGT
327	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGACAGTTGGAACTGGTCTATCTCCCCGGTATTTT AATAAGTTTATAGGCGTAGCAAAGGCATATACGACAAGAGTAGGGGAGGGGATATTTCCTACT GAGATGTTTGGGGAAGAGGCAGATAGACTTAGAACCCTAGGCACATCATGTAGTAGACGACCA AGACAGT
328	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAGAAGACTTAAAGGATTTAGGTAGAGAGCTTAA GGTACCAAGAAGACCGTTCAAAAAGTTAACGCATAGAGAAGCTGTTNATATATTGAGATCTCAT GGCATAAAAGCAAGTTATGAACATGAGATACCTTGGGAAGCCACATCATGTAGTAGACGACCA AGACAGT
329	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACGGGGAGGCTGTCTCAGGAGCTGAAAGAGAATAT AGAGCGGAGAAGGTTATTGAGAGGATGAGAGCTACTGGTGAGAACCCTGCAAAATACGGTTGG TACATTGAAATGTTGAAATATGGTATTCCGCCGAGTGCAGGGCACATCATGTAGTAGACGACCA AGACAGT
330	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATATGCAGATTTAGATGAGATTATAGGGGTTGCATC TAAGGCAGGAATAGATTGCATAACTATAGATGGGTCAGAAGGTGGAACAGGTATGAGCCCTAT AGCTGCGATGAGAGAACTAGGATATCCAACGCTAGTATGTCCACATCATGTAGTAGACGACCA AGACAGT
331	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGACACGAAATTGCTGAAGCAGCTGGCTCAACATG GTATATCGACAATTTCTGGGATAAACTCAAAGAGGGCTGTGTAGCATATCTAAACATAGATTCA CCTGGATTAAAAGATGCAACAAGATATATCGCTTACGCGTCCACATCATGTAGTAGACGACCAA GACAGT
332	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTAACTTCTGGAAACGCCCAATCAAAACAGATCAT GACACCAAAGCTAAAATTATCTTCCCTAATAGCTTCTATAGGTGTATCTCCAGGTTGAAATATTA GCTTCTCTTTGGCAAATAAGTGAAGTTTCCTATACTTTCCCACATCATGTAGTAGACGACCAAGA CAGT
333	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCAGATAGCCCAATAGCATCAATTTCCGTTGCAATA ATAGGTACAGTACACAAAGAACACGTAATTTTCAGCGACACTGCAAATACAGGCGACTTAATA ATTTTTGCCATAGATCTCGATGGAACATTTCACCCTAAGTTCACATCATGTAGTAGACGACCAA GACAGT

334	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCTAATTCCTCTCTTACAGCTTTAAAAGCAATCA CAGCAGATTCCAAAATATCATCCATATCATCCAGAGCTATAATAACACCTCTTGAAGTTTTCCCA ATCTTATGCCCACTTCTTCCAACTCTTTGAACCAAACGACACATCATGTAGTAGACGACCAAGA CAGT
335	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTAACTTGTCTGGGAGACATATATTGGACAACTAA ATCAACGGTTCCTACATCAATCCCTAACTCCATAGATGATGTACAAATAAGACCTTTCAACTCA CCGTCTTTAAATAACCTTTCAACTTCTATACGAACATCTCTCACATCATGTAGTAGACGACCAAG ACAGT
336	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACTCAATGAACCATGATGCACATCAATACTTAGATT AGGATCGTATAAGTGAAGCCTAGAAGCTAGTATCTCAGCTATTTCACGAGTGTTTACAAAAGTA AGCATAGAGCGGCTCTTTTCTAATAACTCAACCAATACCCCACATCATGTAGTAGACGACCAAG ACAGT
337	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGTTAAATCATCTTAACTCACAAATATTAAGGCTT TAATTTCTGAGGGAGTGCAAAATGAAAACTGACGTAGTAATAGTAGGTGCAGGGCCCGCAGGC ATGTTTGCTGCACATGAATTGGCAACTAAATCTAATCTGAACACATCATGTAGTAGACGACCAA GACAGT
338	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAAAAATAGCCAAGGATCCAAAATTCCGTGTATATA CAAAAACCTTCGATGACCTTACACGTGTATTTTGCGTTAATTATCGAGGCTTCGTCGTCCAAGAA GTCTACGGAGATATCGTTGGTGTTAACGGCCACACTCTAACACATCATGTAGTAGACGACCAAG ACAGT
339	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCAAACAAAAATCTGAAAATGCCAATTTTGCATTTC TAGTTCGAGTTGAACTCACCGAACCGCTTGAAGACACAACCGCCTACGGATTCTCAATAGCCAA ATTAGCAACTACCATAGGTGGAGGAAAACCAATTCTTCAACACATCATGTAGTAGACGACCAA GACAGT
340	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAGATACTGAATTTCCAAAACTCAAAGGATATAG AATTGTTAGAATCGCAACACATCCGCAAGTTATGAGCATGGGACTAGGAAGTGAAGGGTTGTC AAAACTTTGCCAAGAAGCCGAAAAGAGAGGACTAGATTGGGTCACATCATGTAGTAGACGACC AAGACAGT
341	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAAGTTTTTATCCTCCTCGGTCCAAGTCACACTGG TTACCCAGGCGTTGGAATAATGACAGAAGGCATCTGGAAAACTTCTTTAGGAGAAATATCAATA GATGAAACTCTCTCGAATACTATTTTAAATAATTGTGACCCACATCATGTAGTAGACGACCAAG ACAGT
342	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGACACACTACGGCACCTACTATGGATACACACCA GCTGGTGTTGAACCATTAACCAAAGTTTTAGAATGGATATACCAGACGGACAAACAAGTTATTG AGAGAATTAAAAGATTAGATGGAGCAGGAGTAATAGAATATCACATCATGTAGTAGACGACCA AGACAGT
343	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTGAAAAGTTCATTCCAATTGTTAAATCGCCATCTT GGAAACACGGCACAAGAAAAGGGAAAGGATTTAGCATCGGTGAGATTAAAGCAGCCGAGATA GATATTAGTATGGCAGTTAAACTCGGTATACCCATTGATAAACACATCATGTAGTAGACGACCA AGACAGT
344	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGGAATAATAATTAAAATAATGTGGCACACCTTTTA GCTTCTTTTCATCTCATATTTTCAAAGAAGCCTTCCAGGTGTGCCTCATCGGTGTCCCCCGCTGC GGAGACACGGTATCATCGTATCCGCCGAAGGAAACTCAACACATCATGTAGTAGACGACCAAG ACAGT
345	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACATTGCCTATCAATTACTTCAAGCCGGAATGCAA GTTCCCGGTTTCAGAAGGTCGCCAAAGATAATAGAAAGAATTTTAGAAAGATATATTCCAACAG TCACCGTACTAGGCGGCATTATTGTAGGATTAATAGCTGCCACATCATGTAGTAGACGACCAAG ACAGT
346	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTCGTTCAGGGAGGTATAAAAATGCCAGAACCAC GCTACCGGTCAAGGTCTTTAAGAAGACGATACGTACACACACCTGGAGGAAAAACCGTCATCC ATTACAGGAGAAAAAAACCTGACGTTGCAAAATGCGCATTATCACATCATGTAGTAGACGACC AAGACAGT

347	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGTCAACCTCTCAGAGGAATTCCCAGACTAAGG CCAGGAGAATTCAGAAAGTTGACAAAAAGTCAACGAAGACCAGAGAGACCTTTCGGTGGATAT CTATGCCACAAATGCTTAGCAATGGAAATCAAGAAAGCTGTTCACATCATGTAGTAGACGACCA AGACAGT
348	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATAGGATGAATCTAACTGGGGCGACCCGGTAGATA ACTGAGAGTGTAGGAGGTGAAATAATTGAGCGCAATAGAAGTAGGTAGAATATGTGTTAAAAC TAGTGGAAGAGAAGCAGGAAGAAAGTGCGTTATTGTTGAAATCACATCATGTAGTAGACGACC AAGACAGT
349	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACACCATTTCCTAATATTTTAGTAACTAGATATGTT TGTTATAGTATTAGGGTGAAGTATTTGTATGAAAGAAAGTTGCCATCAGACATTAAAAGAGAGA TTCTAGTAAAAAGTGAAGCAGAAACTGACCCTGCTTATGGCACATCATGTAGTAGACGACCAAG ACAGT
350	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCACATGAGAGAACTTAGAAGAACACGTACAGGACC CTTTAAAGAAGATGAAACCCTAGTAACTCTTCACGATGTAGTTGATGCTTACTATTTTTGGAAGG AAGATGGAGAAGAAGAATTTCTACGAAAAGTCATACAACCCACATCATGTAGTAGACGACCAA GACAGT
351	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAATGGAAAAGGGTTTAGAACACCTACCTCACATTT GGATTAGAGATTCTGCTGTAGATGCAATATGCCATGGGGCAAACTTAGCAGCTCCTGGTGTTGT AAAACTTCATGACGGTATATCACCTGGAGACTTAATAGTAACACATCATGTAGTAGACGACCAA GACAGT
352	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCTGATCATACATGTGCATTGTCTTTAAATACACT AGTAACGTTAATAATATCTAGCAATTTTAGATAAAAATAACTAGCAGTGCCGGGGTAGCCAAGT GGACTACAGGCCTTATACCGGTTAGGGCGCGGGCCTGGAGCACATCATGTAGTAGACGACCAA GACAGT
353	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATGCCTTAACGAGAGGCATGGGATGGGGGAGCTG TGAGCCCCCCGAACCGGCAGATGAGGGGAAGGGTGCAAAGCATCCCTTAACGCCGGAAGCTCC CGACTTCAGTCGTGGAGCAGCTCACTGCTTTGACGAAAGGTTCACATCATGTAGTAGACGACCA AGACAGT
354	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAACTTGCAAGGAAGGCCGGTGTTGATTATGAGAC AAAGCTGTTGGTCAGGGGCAAGGAACCGGCTGAGGACATAATAGAATTTGCTGACGAGATCAG GGCAAGTCTCATTGTAATAGGGGTTAGGAAGAGGAGACCCGCCACATCATGTAGTAGACGACC AAGACAGT
355	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCAGAAGAGATTCAAAGCTCTCGTATTCAATGTCC CCACCAAATTTCTGGTCGCGCTCAATTTTGACTTTACCAAAAGCGGGGAAAACGTAGTGCTTTG CTAGGTCTATTATCGGATTTCCTTCTACAACCTTTGGCGGCACATCATGTAGTAGACGACCAAGA CAGT
356	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGATTTGCTCATTTTCTCCCCGTCGAGTCCTGAGATT ATCGGCGTATGGATGCAGATCGGTGCCTTGTAACCGAGGGCCGGCAGATTCTCCCTTGCGAGCA TGTGGATCTTTCTCTGATCTATTCCACCAACCGCCACATCCACATCATGTAGTAGACGACCAAGA CAGT
357	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCGGGAGTTGCAGAACCAAGCATGGAAATTGCTA GAGATCCCGAAAAGGTTTACGAGTACACGAATAAGTGGAACACGGTTGCAATTATCACTGATG GCTCGAGGGTCTTGGGACTGGGCAACATCGGTGCGATGGCTTCACATCATGTAGTAGACGACCA AGACAGT
358	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGTGTTATCAAGAGGGAATATATTGCTCAGATG GCAGAGGATCCGATAGTCTTTGCCTTATCAAACCCGGTGCCTGAGATCTATCCGCAGGAGGCAA AGGAAGCCGGAGCCAGGATCGTAGGAACTGGTAGGAGCGACCACATCATGTAGTAGACGACCA AGACAGT
359	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGGATCTGTTAGTATGGCATTCAGAGCCTTTATGTC CTCATCGGTAAGCTTGTCCGATGGCAGATCGTATTTCACGATGTCTGAAGGAGTAACTCCGAGA AACTTCGCTTCTGGTGTCGCAAGATACTCCGAGAGATGCGCACATCATGTAGTAGACGACCAAG ACAGT

360	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGAGTGCGGCTTACTCTGCACTGTGCGAGATCGATG AGGTCGTTGTTGTTGCCCCCATAACGCAGATGAGCGGAGTGGGGAGGAGCATATCCATAATGCG GCCGGTTCGTTTTTTCGAGCTCGAAATAGATGGCATGAGGCACATCATGTAGTAGACGACCAAG ACAGT
361	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGGGAAGGGAGTACTACTGGATTCATGGGGTGGA AGTCGAAAGCGCTGAGCCTGGAACGGACATACACGCACTCAGAAACGGGTATGTCTCCATTAC ACCGATATCCTTAAATGCAACTTCGGACTGCGAAGCTTTAAGCACATCATGTAGTAGACGACCA AGACAGT
362	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATAGTTTTATGGAGGGTGGTTGGACATGAATGAAA GGGCAAAGAAGGTCATTCTTATTGTGGATGACGATTTGGCTCTGCTTGAAGCTCTTGAACTGAT GCTTCGAGGCAAGTATGAGGTTGTGAAGGTGACAAATGGGACACATCATGTAGTAGACGACCA AGACAGT
363	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATGTCGATTCCGAAATAGCAGGGAGCAATTATCGG TGGGCTTCCGACCCTTAAATGGATTTCCTTCGCTCCCGCCTTTCTTATCATGTCGACTATTCTTTT GGATGTTGTTGCCCGCACAATGCTGTCGTCAACCAGCACCACATCATGTAGTAGACGACCAAGA CAGT
364	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACTTTCTGAGGGAAAAACATTGTTGCTTATCCTAAA GAGTTTACAAGCAAGAAGCTGGAAACAAACTCTGGATGTTATTAATTTAGAGCCTGCAGCAGCA TATACAATGTTTAGAGCGGCAATAAAGAAACTATACAAAGCACATCATGTAGTAGACGACCAA GACAGT
365	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGTTGAGAGGCTGCTTGAAGGCATTGCAAAGAA TGAAAGGGTAGCTTACGGATTGGAGGAGGTTAGGAGGGCAAAAGAGTATGGAGCAATTGAGGT TCTGTTGGTTTCAGATGACTTCCTGCTCACCGAGCGTGAGAACACATCATGTAGTAGACGACCA AGACAGT
366	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGCTTCGAGATTCCTGATAGGAGTGGGAGTTGCCG GGGTTTACGTGCCTACGATAAAAATAATATCCGTCTGGTTCAGGCAGAATGAGTTTGCAACTGC TACTGGGATTCTTTTCGCGATTGGAAATCTAGGAGCGATTCACATCATGTAGTAGACGACCAAG ACAGT
367	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGGTATCGCCTACTTAGAGAGTTCGTAAAGTCGG AGATATTGGAGGAAGTTAAATTTGAAAACGTTGTGGACGAGTACTGGGTTGCGGAACCATTCAT AAAGATCATAATTTTTGAGGATCTCGAAAACCAGAAATTGACACATCATGTAGTAGACGACCAA GACAGT
368	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTAATCCGATTATCGATTCTACGCTTCCTGATGGTA GCAGGCTTCAGGCTACCCTAGGAACAGAAATTACACCTAGAGGCTCGAGCTTCACGGTGAGAA AATTTACAACCCAGCCACTGACCCCGTTAGATCTAGTGAGGCACATCATGTAGTAGACGACCAA GACAGT
369	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAAAATTATATCGATAGAGGATACCAGAGAGATAA AGCTCCATCATGAGAACTGGCTGGCTCAGGTGACGAGAACGGGGATAGGAGAGCAGGAAATTG ACATGTATGACCTTCTCAAAGCCGCCTTGAGACAGAGACCGGCACATCATGTAGTAGACGACCA AGACAGT
370	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAATCAGTTTGTTAAATGGGATGCGAAGAAAAATT CGCATGTTGAGGTAGGGATTCCGAAAAAGCTAGAGAAAATCGCGATGTCGAGAGTGGACGATG CTTACGCGGAGCTGGAAAGAAGAAGGAGGTATTTGGAGTGGACACATCATGTAGTAGACGACC AAGACAGT
371	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCAGTGAAGTTAGCACGGAATTCGAAAGGATAGTG GTTCTCGTTGAAATGGGAGAGGATTTGGAAAGCGCAATGAGGTTTGTTGCAGAAACAACTCCCT CAGAGAGGCTCAGGGTTTTTCTGGAGAACTTTATTGATGTGCACATCATGTAGTAGACGACCAA GACAGT
372	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCTGGAGCGGGAGGCGTATCAACGCTTGCCCTCAA TCCGTTACCCGAAGTTCCAGAATACTTTGAGTATTTCCAGTCCGAATAGAAGCAGAGCACCTCT CGATCGACTAGAGTCTTTCTGCTAGCTCTTGCACCCTCATCCACATCATGTAGTAGACGACCAAG ACAGT

373	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGGAAATCTCTGCTGAAAACACCTTGACTTTTTCT TCGTATATCTCCCATTCCATCAGGCACCACCAACTTTGGTCCTGCAAAGAGTCATCGGTGCCCCA TCTGCTACGGGAACGATCTGAAAGGCTTTACCACAGAATCACATCATGTAGTAGACGACCAAGA CAGT
374	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCGGTTGCAGGATTGGTCTCCCCACCTCTCGAGCC TATGAGGAATACCCCATTCCTGCAGAGCTCGAGAAGCTCTTCGAATTCAAGATCCCCCTTCTGC AGGAATGTGTTGCTCATTCTGACAATCGGAAAAGCAACTCCACATCATGTAGTAGACGACCAAG ACAGT
375	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAACTCCTCGATTGTTGGGTCATCGATTATTGTCACG TTCTCTCCTGCAATTCTCTCTCCAATCTTTCCAGCAAGAACGCTGTTTTCCTGCAGAACGTGATCT GCCTCGACCGCATGCCCGAAAGCTTCGTGAATAAAAACCACATCATGTAGTAGACGACCAAGA CAGT
376	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTTCTTCAAGAATGCTTTCTGCGGCGATAAGCCCA GTAACAGCCGCTCCAACTATTCCCCTGCTTATTCCGGCTCCATCGCCAATTGCATAGATGTACGG TATGCTTGTCCTCATCTTCTCGTCAACCTTAAGCTTCAACACATCATGTAGTAGACGACCAAGAC AGT
377	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGCTGGATTTTTCTTTGGCCTTGCTGTGGCCGTTGAC TAGACAAAAGTCGCCGTACTCCTCTCTTATAACCCAGCCCCTCGGGCAGGTGCAGAACGTGCGC ATGTAGTCGTCATGCCTCTGTGTGATTATTCTCAGCTTTCACATCATGTAGTAGACGACCAAGAC AGT
378	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTGCCTTGGAATTTTCCGCCACTTCGATCTTATACTT TTTTACCCATTTTTCCAGCCAGTCGGCACCGCTCCTCCCAACTGCAATTATGAGTTTGTCGTAGC CGAACTTGTCCCCATCGTTCGTCTTCACGATCTTTTCTCACATCATGTAGTAGACGACCAAGACA GT
379	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAATCACCACCGACTGTGAAGCTCGAAGGATAGTTG GGGTTGGCATAATTCAGCTTTCCATCCGAAAGTCCTCCAGCACCACCCACACCAGAAGTAATGT TGCAGGGATCGCATTTCTTGCAATAGCTTTGCGAAAGGTCACACATCATGTAGTAGACGACCAA GACAGT
380	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTAACCAACCTCTTTCGCATCAAAATCCCAACTGCG GCATCCGTTATCAGCGTTACATCGATTCCATCTTTCATAAGCTCGTAGCAGGTGAGCCTAGAGCC TTGGTTCAGCGGCCTCGTTTCGCAGGCGAAAACCTTTACCACATCATGTAGTAGACGACCAAGA CAGT
381	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTATCGAGTTAATAGCTATCAGTGTTGCTATTACGA TCGTTGCGATCCCATCAAAGATGTTATGACCGAAGGAGATAGCAATAATGCCAAATATCGCTGC AAGCGTCGATAAGGAGTCGTTAAAACTCTCAAACATCACTCACATCATGTAGTAGACGACCAAG ACAGT
382	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCCCTTCTAAGCTTCGTGATATCTGCATTGCCAAT ATCAACTAGAAATTCGATTGAGATAAGCTTGTCTCTTGCGGTTAAGCTTGTTCTCTCGATATTTA TACCGAAATTTAGCAATACACCCGTGATATCTCTCACGACACATCATGTAGTAGACGACCAAGA CAGT
383	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAAGCGGGGCTTTTGCCTTTCCAATTCCGCCGCAACC AACCGTTGGACTTATCAAACCGGAACCTTTCAACTCCGAGATTAAAGAGCCTGGCTCCTTATCG TGCTTAATAGCAATTTCTACAATGTCTTCCCCGCATACAACACATCATGTAGTAGACGACCAAG ACAGT
384	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTAGTTCTTGGTTTTCGTCGACGTTGACCTTGTAGA ACTCTACATCTGGAAACTCCTTTGAAAGCTTTTCGAGCACTGGGCTGAGATACCTGCACGGCAT GCACCAGTCGGCGTAGAAGTCAACAACAACAAGCTTATCCCACATCATGTAGTAGACGACCAA GACAGT
385	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCCGATCGTCTTTAAAGCTTGCAAGTCTAAATCCT CGCCCCAGGGAATTTCCTGGGATTTTCTCGCAATCTCTATCGCCGAAGTATAGGTTATCCTCGGG AATGGTATCTCGGGGACTTCGAGCTTTAGTTCGAGAATACACATCATGTAGTAGACGACCAAGA CAGT

386	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGGTTTTGTCCCTATTGGGTTTCTCATTGCCTGCA GCATTTCTTCTCTGCTCAGAGCTCTGCAGCCATCGCCTTTCATTCTTAAAATGCTAACCTCCCAA TCATCCGGAAAATCGAGCTCTATTTCCCTATCCTGCCAGCACATCATGTAGTAGACGACCAAGA CAGT
387	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTTTACAGGCTGGTGGGTGGGGAAAGGAGTGTTA AGGGCAAAAGGAGTGTAAGCAAGTTCAGGGTTGCGATTGCGATTCTTCTGGCATTCATTCTGAT ATATCCTACATACCGCATAGCCGAGATTCAAAGCAGTGGGGCACATCATGTAGTAGACGACCA AGACAGT
388	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGGGTCAGGAGGATTCACGAGATAGAAGTCCTCG AGGTGAGAGGCAGGTTCGCGCTTATAAGGGTTCTCAGCGACCCCGGCACGTACATGAGGAAGC TGGCCCACGACATCGGGCTATTGCTCGGAGTAGGTGCACACACACATCATGTAGTAGACGACCA AGACAGT
389	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAGTCAGTCATAGAATCAATGGTGTATTCTTCATC AGGGTTATTATACGGAATGAACTTATAGTTCTCACCTGCTACCTGATCCACTGTCATTTCTGCAA GAGTCTGCACTGTGGTAATTCCACCTTCTTCCATCCGGGCACATCATGTAGTAGACGACCAAGA CAGT
390	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTAAGGGAATCAATGTCTTCCATTGCTGTAAGGGT TACTGTTACCTTTGTAGAAGTCAGACCGTAATTGGTCAGCAGCTCTTCATAGAAGTTCTGGTTTG CAATATCCCTCTGGGCAATGACAGGGTAGTCGACTTCGTCACATCATGTAGTAGACGACCAAGA CAGT
	Forward and Reverse Primers
391	TCTCCTTCTTAGCTTCGTGAGAAC
392	CTTGGTCGTCTACTACATGATGTG

Primer Binding Sequences

[0069] The 5' and 3' primer binding sequences are selected to be complementary to a SDSI 5' and 3' primer which is included in an amplification reaction and used to amplify SDSIs present in a given sample. The primer binding sites may be optimized for multiplex amplification with a set of primers used to amplify a genome for sequencing. In one example embodiments, the 5' and 3' primer binding sites have a Tm of between 55-65 ^{o}C. In one example embodiment, the 5' and 3' primer binding site are complementary to primers having SEQ ID NOS: 391 and 392.

Methods of Detecting and Preventing Sample Contamination

[0070] In one example embodiment, a method of detecting and preventing contamination in one or more amplification reactions comprises adding a SDSI according to the example embodiments disclosed above to a one or more samples to be assayed. An amplification reaction is then used to amplify a target sequence in the samples. The amplification reaction will include probes and primers needed to amplify the target sequence and to amplify the SDSI. The amplicons generated from the amplification step are then used the one or more samples, sequencing the amplified samples and determining the number of reads of the SDSI from the one or more samples,

wherein detection of only a single SDSI in the sample indicates contamination free amplification of the same, and wherein detection of multiple SDSI's indicates possible contamination of the sample. Samples identified as potentially contaminated may then be discarded or marked for repeat to confirm accuracy of results.

Amplification

[0071] The present invention solves this problem by providing for the sequencing of spike-DNA sequences at concentrations that can be amplified concurrently with the nucleic acids of interest. In one example embodiment, sequencing includes extracting total RNA or DNA from a biological sample, such as a sample collected with a swab (e.g., nasal, rectal, vaginal). Methods of extracting total RNA or DNA are known in the art and commercial kits are available. The presence of a pathogen may be confirmed in a sample. Exemplary methods for confirming include PCR, RT-PCR and RT-qPCR. In certain embodiments, sequencing includes DNase treatment to remove residual DNA. In certain example embodiments, sequencing may include depletion of ribosomal RNA (rRNA). In certain example embodiments, cDNA may be prepared from total RNA using RT-PCR. In certain example embodiments, RT-PCR may be performed using random hexamer priming. In one example embodiments, a SDSI is added to each cDNA sample. The SDSI can be added to the total cDNA sample. In certain example embodiments, cDNA samples may be normalized to a constant amplification level. In certain example embodiments, real time PCR may be performed on the cDNA using one or more standard primers and a Ct value is used to normalize cDNA samples. As used herein, standard primers refer to a primer set that is used for every sample. In certain embodiments, the standard primers are directed to a region of the pathogen to be sequenced. The samples can be diluted such that all of the samples for amplification have the same Ct value in the amplification reaction. In certain embodiments, each sample is normalized to a Ct value less than 35, 34, 33, 32, 30, 29, 28, 27, 26, 25, or 24. In preferred embodiments, the samples are normalized to a Ct value of 26 to 28, preferably 27. In one example embodiment, a SDSI is added to the normalized sample used for PCR amplification of the pathogen. The cDNA may be amplified in the same reaction with pathogen specific primers and primers specific to the SDSI. Amplification may be performed in a multi-well plate (e.g., a standard PCR plate).

[0072] In certain example embodiments, the primer concentration is 100μM. In certain example embodiments, the primer concentration is between 50μM-150 μM, or between 50μM-

200μM, or between 50μM-250μM, or between 50μm-250μM or between 50μm-300μM or between 50μm-350μM or between 50μm-400μM or between 50μm-450μM or between 50μm-500μM. In certain example embodiments, the primer concentrations is between 50μm-70μM or between 70μm-90μM or between 90μm-110μM or between 110μm-130μM or between 130μm-150μM or between 150μm-170μM or between 170μm-190μM or between 190μm-210μM or between 210μm-230μM or between 230μm-250μM or between 250μm-270μM or between 270μm-290μM or between 290μm-310μM or between 310μm-330μM or between 330μm-350μM or between 350μm-370μM or between 370μm-390μM or between 390μm-410μM or between 410μm-430μM or between 430μm-450μM or between 450μm-470μM or between 470μm-490μM. In certain example embodiments, the primer concentration is between 50μm-100μM or between 100μm-150μM or between 150μm-200μM or between 200μm-250μM or between 250μm-300μM or between 300μm-350μM or between 350μm-400μM or between 400μm-450μM or between 450μm-500μM.

[0073] In certain example embodiments, a spike-in may be relatively the same length as the amplicons generated for the target organism. In one example embodiment, spike-ins are the same size and share the same priming region to ensure similar amplification performance. In certain embodiments, a spike-in for MNase-seq, ChIP-seq, and genomic DNA are around 150 nucleotides in length. In one example embodiment, a spike-in accounts for 0.1% - 3.5% reads. A spike-in to total sample ratio may be from 1,000:1 to 50:1. In one example embodiment, a spike-in includes primer binding sites on the 3' end and/or the 5' end. (Chen K., *et al.*, The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. MolCell Biol (2016) 36:662–667) The primers and primer binding sites on the SDSI may range between 15 – 40 nucleotides in length. The primer's melting temperature (T_m) may range from 40°C – 95 °C, preferably between 55-65 °C.

Sequencing

[0074] After amplification of cDNA, standard sequence library generation can be performed. In certain embodiments, sequencing comprises high-throughput (formerly "next-generation") technologies to generate sequencing reads. In DNA sequencing, a read is an inferred sequence of base pairs (or base pair probabilities) corresponding to all or part of a single DNA fragment. A typical sequencing experiment involves fragmentation of the genome into millions of molecules

or generating complementary DNA (cDNA) fragments, which are size-selected and ligated to adapters. The set of fragments is referred to as a sequencing library, which is sequenced to produce a set of reads. Methods for constructing sequencing libraries are known in the art (see, *e.g.*, Head et al., Library construction for next-generation sequencing: Overviews and challenges. Biotechniques. 2014; 56(2): 61–77). A "library" or "fragment library" may be a collection of nucleic acid molecules derived from one or more nucleic acid samples, in which fragments of nucleic acid have been modified, generally by incorporating terminal adapter sequences comprising one or more primer binding sites and identifiable sequence tags. In certain embodiments, the library members (e.g., genomic DNA, cDNA) may include sequencing adaptors that are compatible with use in, e.g., Illumina's reversible terminator method, long read nanopore sequencing, Roche's pyrosequencing method (454), Life Technologies' sequencing by ligation (the SOLiD platform) or Life Technologies' Ion Torrent platform. Examples of such methods are described in the following references: Margulies *et al* (Nature 2005 437: 376-80); Schneider and Dekker (Nat Biotechnol. 2012 Apr 10;30(4):326-8); Ronaghi *et al.* (Analytical Biochemistry 1996 242: 84-9); Shendure *et al.* (Science 2005 309: 1728-32); Imelfort *et al.* (Brief Bioinform. 2009 10:609-18); Fox *et al.* (Methods Mol. Biol. 2009; 553:79-108); Appleby *et al.* (Methods Mol. Biol. 2009; 513:19-39); and Morozova *et al.* (Genomics. 2008 92:255-64), which are incorporated by reference for the general descriptions of the methods and the particular steps of the methods, including all starting products, reagents, and final products for each of the steps.

[0075] In one example embodiment, any suitable RNA or DNA amplification technique may be used to amplify a sample and SDSI. In one example embodiment, the RNA or DNA amplification is an isothermal amplification. The isothermal amplification may be nucleic-acid sequenced-based amplification (NASBA), recombinase polymerase amplification (RPA), loop-mediated isothermal amplification (LAMP), strand displacement amplification (SDA), helicase-dependent amplification (HDA), or nicking enzyme amplification reaction (NEAR). In certain example embodiments, non-isothermal amplification methods may be used which include, but are not limited to, PCR, multiple displacement amplification (MDA), rolling circle amplification (RCA), ligase chain reaction (LCR), or ramification amplification method (RAM).

Example Applications

[0076] In one example embodiment, the present invention is used to improve any method of sequencing wherein the nucleic acids to be sequenced are amplified (i.e., amplicon-based methods). In certain example embodiments, the amplification method preferentially amplifies a contaminant nucleic acid if it is present in a sample. In preferred embodiments, samples comprising a pathogen of interest are sequenced. In more preferred embodiments, the pathogen of interest includes variants that can be clustered into families or a lineage. As used herein, the term "variant" refers to any virus having one or more mutations as compared to a known virus. A strain is a genetic variant or subtype of a virus. The terms 'strain', 'variant', and 'isolate' may be used interchangeably. In certain embodiments, a variant has developed a "specific group of mutations" that causes the variant to behave differently than that of the strain it originated from. In certain example embodiments, the families of variants are important for tracking and responding to epidemics and pandemics. For example, sequencing can be used to determine variants that are emerging as the dominant variants causing disease or are spreading more quickly. In another example, sequencing variants can be used to track community transmission and superspreading events (see e.g., Lemieux *et al.*, 2020). Variants may also include those that are resistant to a specific treatment, such as drug resistance. In certain embodiments, variants are associated with more severe disease. As used herein, the term "epidemic" refers to the rapid spread of disease to a large number of people in a given population within a short period of time or the occurrence of more cases of disease, injury, or other health condition than expected in a given area or among a specific group of persons during a particular period. For example, in meningococcal infections, an attack rate in excess of 15 cases per 100,000 people for two consecutive weeks is considered an epidemic. Epidemics of infectious disease are generally caused by several factors including a change in the ecology of the host population (*e.g.*, increased stress or increase in the density of a vector species), a genetic change in the pathogen reservoir or the introduction of an emerging pathogen to a host population (by movement of pathogen or host). Generally, an epidemic occurs when host immunity to either an established pathogen or newly emerging novel pathogen is suddenly reduced below that found in the endemic equilibrium and the transmission threshold is exceeded. An epidemic may be restricted to one location; however, if it spreads to other countries or continents and affects a substantial number of people, it may be termed a pandemic. Effective

preparations for a response to a pandemic are multi-layered. The first layer is a disease surveillance system, which includes sequencing of all variants in a population. In certain embodiments, sequencing contaminants that were amplified from a sample would provide an incorrect identification and clustering of the variants.

[0077] Any method of sequencing variants in pathogens, such as viral pathogens, is applicable to the present invention (see e.g., Lemieux *et al.*, 2020). Current sequencing methods all suffer from the risk of contamination and the user would be blind to whether the results were accurate.

[0078] In certain example embodiments, a pathogen with a DNA genome is sequenced. Sequencing may include whole genome sequencing. Whole genome sequencing (also known as WGS, full genome sequencing, complete genome sequencing, or entire genome sequencing) is the process of determining the complete DNA sequence of an organism's genome at a single time. This entails sequencing all of an organism's chromosomal DNA as well as DNA contained in the mitochondria and, for plants, in the chloroplast. "Whole genome amplification" ("WGA") refers to any amplification method that aims to produce an amplification product that is representative of the genome from which it was amplified. In certain embodiments, the SDSIs of the present invention are added at the amplification step. Non-limiting WGA methods include Primer extension PCR (PEP) and improved PEP (I-PEP), Degenerated oligonucleotide primed PCR (DOP-PCR), Ligation-mediated PCR (LMP), T7-based linear amplification of DNA (TLAD), and Multiple displacement amplification (MDA).

[0079] In certain example embodiments, the present invention includes whole exome sequencing. Exome sequencing, also known as whole exome sequencing (WES), is a genomic technique for sequencing all of the protein-coding genes in a genome (known as the exome) (see, e.g., Ng et al., 2009, Nature volume 461, pages 272–276). It consists of two steps: the first step is to select only the subset of DNA that encodes proteins. These regions are known as exons – humans have about 180,000 exons, constituting about 1% of the human genome, or approximately 30 million base pairs. The second step is to sequence the exonic DNA using any high-throughput DNA sequencing technology. In certain embodiments, whole exome sequencing is used to determine germline mutations in genes associated with disease.

[0080] In certain example embodiments, targeted sequencing is used in the present invention (see, e.g., Mantere *et al.*, PLoS Genet 12 e1005816 2016; and Carneiro *et al.* BMC Genomics,

2012 13:375). Targeted gene sequencing panels are useful tools for analyzing specific mutations in a given sample. Focused panels contain a select set of genes or gene regions that have known or suspected associations with the disease or phenotype under study. In certain embodiments, targeted sequencing is used to detect mutations associated with a disease in a subject in need thereof. Targeted sequencing can increase the cost-effectiveness of variant discovery and detection. In certain embodiments, targeted sequencing includes amplification and the SDSIs of the present invention are added at the amplification step.

[0081] In one example embodiment, the mitochondrial genome from more than one sample is sequenced. In certain embodiments, mitochondrial genome sequencing includes amplification and the SDSIs of the present invention are added at or before the amplification step. An exemplary method includes MitoRCA-seq (see e.g., Ni *et al.*, MitoRCA-seq reveals unbalanced cytosine~~cytocine~~ to thymine transition in Polg mutant mice. Sci Rep. 2015 Jul 27;5:12049. doi: 10.1038/srep12049). The method employs rolling circle amplification, which enriches the full-length circular mtDNA by either custom mtDNA-specific primers or a commercial kit and minimizes the contamination of nuclear encoded mitochondrial DNA (Numts). In certain embodiments, RCA-seq is used to detect low-frequency mtDNA point mutations starting with as little as 1 ng of total DNA.

[0082] In another example embodiment, multiple displacement amplification (MDA) is used to generate a sequencing library. Multiple displacement amplification (MDA, is a non-PCR-based isothermal method based on the annealing of random hexamers to denatured DNA, followed by strand-displacement synthesis at constant temperature (Blanco et al. J. Biol. Chem. 1989, 264, 8935-8940). It has been applied to samples with small quantities of genomic DNA, leading to the synthesis of high molecular weight DNA with limited sequence representation bias (Lizardi et al. Nature Genetics 1998, 19, 225-232; Dean et al., Proc. Natl. Acad. Sci. U. S. A. 2002, 99, 5261-5266). As DNA is synthesized by strand displacement, a gradually increasing number of priming events occur, forming a network of hyper- branched DNA structures. The reaction can be catalyzed by enzymes such as the Phi29 DNA polymerase or the large fragment of the Bst DNA polymerase. The Phi29 DNA polymerase possesses a proofreading activity resulting in error rates 100 times lower than Taq polymerase (Lasken et al. Trends Biotech. 2003, 21, 531-535). In certain

embodiments, the SDSIs of the present invention are added to samples and amplified during MDA or in a subsequent amplification step.

[0083] In one example embodiment, is sequencing comprises sequencing of SARS-CoV-2 variants. The scale of the SARS-CoV-2 pandemic has led to a particular focus on reducing the cost and time of amplicon-based methods, often at the cost of slightly reduced sensitivity. However, viral loads of SARS-CoV-2 can vary widely between individuals, in particular when samples are caught early in infection or follow-up sampling is needed. An open-access tiled primer set developed by the ARTIC network is the most widely used method for SARS-CoV-2 specific genome amplification followed by sequencing on either Illumina or nanopore instruments (Quick *et al.*, 2017; Tyson e*t al.*, 2020). A wide array of protocols and publications are now available that integrate these ARTIC primers with different amplification and library construction indexing strategies (Baker *et al.*, 2020; Gohl *et al.*, 2020). Approaches such as batching samples by viral load to increase sensitivity are impractical to scale to current needs, resulting in incomplete recovery of viral genomes, especially from low titer samples.

[0084] In certain embodiments, the methods described herein can be used to sequence viral samples with low viral loads. A viral load may also be interchangeably referred to as viral burden or viral titer. A viral load may be expressed in viral particles per mL, infectious particles per mL, copies per mL, or virus per mL. A low viral load may be a cycle threshold (CT) > 30 or copies per mL < 10^4. A high viral load may be a CT < 30 or par or copies per mL > 10^5. For example, viral loads lower than 10,000, 1,000, 500, 400, 300, 200, 100, 50, 40, 30, 20, 10 viral particles. In certain embodiments, a single viral particle is sequenced.

[0085] In certain embodiments, the SDSI is used to detect and prevent contamination in genomic analysis samples of pathogens. A pathogen may include viruses, bacteria, fungi, and protozoa. In certain embodiments, a virus may belong to any morphological category including helical, envelope, or icosahedral. In certain embodiments, a virus me comprise of DNA or RNA, may be single stranded or double stranded, and may be linear or circular. In certain embodiments, the genome of the virus may be one nucleic acid molecule or several nucleic acid segments. In certain embodiments a virus may belong to the family: Adenoviridae, Papovaviridae, Parvoviridae, Herpesviridae, Poxviridae, Anelloviridae, Pleolipoviridae, Reoviridae, Picornaviridae, Caliciviridae, Togaviridae, Arenaviridae, Flaviviridae, Orthomyxoviridae, Paramyxoviridae,

Bunyaviridae, Rhabdoviridae, Filoviridae, Astroviridae, Bornaviridae, Arteriviridae, Hepeviridae, Retroviridae, Caulimoviridae, Hepadnaviridae, Coronaviridae. In certain embodiment, the virus is SARS-CoV-2. (Gelderblom HR. Structure and Classification of Viruses. In: Baron S, editor. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston; 1996. Chapter 41).

~~[0001]~~**[0086]** In an exemplary embodiment, the pathogen sequenced is a coronavirus. As used herein, "coronavirus" refers to enveloped viruses with a positive-sense single-stranded RNA genome and a nucleocapsid of helical symmetry that constitute the subfamily *Orthocoronavirinae*, in the family *Coronaviridae* (see, e.g., Woo PC, Huang Y, Lau SK, Yuen KY. Coronavirus genomics and bioinformatics analysis. Viruses. 2010;2(8):1804-1820). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the virus causing the ongoing Coronavirus Disease 19 (COVID19) pandemic (see, e.g., Zhou, et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273). In preferred embodiments, the virus is SARS-CoV-2 or variants thereof. In preferred embodiments, the disease treated is COVID-19. SARS-CoV-2 is the third zoonotic betacoronavirus to cause a human outbreak after SARS-CoV in 2002 and Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012 (de Wit et al., 2016, SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol *14*, 523-534). While there are many thousands of variants of SARS-CoV-2, (Koyama, Takahiko Koyama; Platt, Daniela; Parida, Laxmi (June 2020). "Variant analysis of SARS-CoV-2 genomes". Bulletin of the World Health Organization. 98: 495–504) there are also much larger groupings called clades. Several different clade nomenclatures for SARS-CoV-2 have been proposed. As of December 2020, GISAID, referring to SARS-CoV-2 as hCoV-19 identified seven clades (O, S, L, V, G, GH, and GR) (Alm E, Broberg EK, Connor T, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020 [published correction appears in Euro Surveill. 2020 Aug;25(33):]. Euro Surveill. 2020;25(32):2001410). Also as of December 2020, Nextstrain identified five (19A, 19B, 20A, 20B, and 20C) (Cited in Alm et al. 2020). Guan et al. identified five global clades (G614, S84, V251, I378 and D392) (Guan Q, Sadykov M, Mfarrej S, et al. A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic. Int J Infect Dis. 2020;100:216-223). Rambaut et al. proposed the term "lineage" in a 2020 article in Nature Microbiology; as of December 2020, there

have been five major lineages (A, B, B.1, B.1.1, and B.1.777) identified (Rambaut, A.; Holmes, E.C.; O'Toole, Á.; et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology". 5: 1403–1407).

~~[0002]~~[0087] Exemplary, non-limiting variants applicable to the present invention are described below. Genetic variants of SARS-CoV-2 have been emerging and circulating around the world throughout the COVID-19 pandemic (see, e.g., The US Centers for Disease Control and Prevention; www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html). Exemplary, non-limiting variants applicable to the present disclosure include variants of SARS-CoV-2, particularly those having substitutions of therapeutic concern. **Table A** shows exemplary, non-limiting genetic substitutions in SARS-CoV-2 variants.

Table A.

Spike Protein Substitution	**Common Pango Lineages with Spike Protein Substitutions**
L452R	A.2.5, B.1, B.1.429, B.1.427, B.1.617.1, B.1.526.1, B.1.617.2, C.36.3
E484K	B.1.1.318, B.1.1.7, B.1.351, B.1.525, B.1.526, B.1.621, B.1.623, P.1, P.1.1, P.1.2, R.1
K417N, E484K, N501Y	B.1.351, B.1.351.3
K417T, E484K, N501Y	P.1, P.1.1, P.1.2
A67V, del69-70, T95I, del142-144, Y145D, del211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F	B.1.1.529 and BA lineages

Phylogenetic Assignment of Named Global Outbreak (PANGO) Lineages is software tool developed by members of the Rambaut Lab. The associated web application was developed by the Centre for Genomic Pathogen Surveillance in South Cambridgeshire and is intended to implement the dynamic nomenclature of SARS-CoV-2 lineages, known as the PANGO nomenclature. It is available at cov-lineages.org.

~~[0003]~~[0088] In some embodiments, the SARS-CoV-2 variant is and/or includes: B.1.1.7, also known as Alpha (WHO) or UK variant, having the following spike protein substitutions: 69del,

70del, 144del, (E484K*), (S494P*), N501Y, A570D, D614G, P681H, T716I, S982A, and D1118H (K1191N*); B.1.351, also known as Beta (WHO) or South Africa variant, having the following spike protein substitutions: D80A, D215G, 241del, 242del, 243del, K417N, E484K, N501Y, D614G, and A701V; B.1.427, also known as Epsilon (WHO) or US California variant, having the following spike protein substitutions: L452R, and D614G; B.1.429, also known as Epsilon (WHO) or US California variant, having the following spike protein substitutions: S13I, W152C, L452R, and D614G; B.1.617.2, also known as Delta (WHO) or India variant, having the following spike protein substitutions: T19R, (G142D), 156del, 157del, R158G, L452R, T478K, D614G, P681R, and D950N; P.1, also known as Gamma (WHO) or Japan/Brazil variant, having the following spike protein substitutions: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, and T1027I; and B.1.1.529 also known as Omicron (WHO), having the following spike protein substitutions: A67V, del69-70, T95I, del142-144, Y145D, del211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F, or any combination thereof.

~~[0004]~~[0089] In some embodiments, the SARS-CoV-2 variant is classified and/or otherwise identified as a Variant of Concern (VOC) by the World Health Organization and/or the U.S. Centers for Disease Control. A VOC is a variant for which there is evidence of an increase in transmissibility, more severe disease (e.g., increased hospitalizations or deaths), significant reduction in neutralization by antibodies generated during previous infection or vaccination, reduced effectiveness of treatments or vaccines, or diagnostic detection failures.

~~[0005]~~[0090] In some embodiments, the SARS-Cov-2 variant is classified and/or otherwise identified as a Variant of High Consequence (VHC) by the World Health Organization and/or the U.S. Centers for Disease Control. A variant of high consequence has clear evidence that prevention measures or medical countermeasures (MCMs) have significantly reduced effectiveness relative to previously circulating variants.

~~[0006]~~[0091] In some embodiments, the SARS-Cov-2 variant is classified and/or otherwise identified as a Variant of Interest (VOI) by the World Health Organization and/or the U.S. Centers for Disease Control. A VOI is a variant with specific genetic markers that have been associated with changes to receptor binding, reduced neutralization by antibodies generated against previous

infection or vaccination, reduced efficacy of treatments, potential diagnostic impact, or predicted increase in transmissibility or disease severity.

~~[0007]~~[0092] In some embodiments, the SARS-Cov-2 variant is classified and/or is otherwise identified as a Variant of Note (VON). As used herein, VON refers to both "variants of concern" and "variants of note" as the two phrases are used and defined by Pangolin (cov-lineages.org) and provided in their available "VOC reports" available at cov-lineages.org.

~~[0008]~~[0093] In some embodiments the SARS-Cov-2 variant is a VOC. In some embodiments, the SARS-CoV-2 variant is or includes an Alpha variant (e.g., Pango lineage B.1.1.7), a Beta variant (e.g., Pango lineage B.1.351, B.1.351.1, B.1.351.2, and/or B.1.351.3), a Delta variant (e.g., Pango lineage B.1.617.2, AY.1, AY.2, AY.3 and/or AY.3.1); a Gamma variant (e.g., Pango lineage P.1, P.1.1, P.1.2, P.1.4, P.1.6, and/or P.1.7), a Omicon variant (B.1.1.529) or any combination thereof.

~~[0009]~~[0094] In some embodiments the SARS-Cov-2 variant is a VOI. In some embodiments, the SARS-CoV-2 variant is or includes an Eta variant (e.g., Pango lineage B.1.525 (Spike protein substitutions A67V, 69del, 70del, 144del, E484K, D614G, Q677H, F888L)); an Iota variant (e.g., Pango lineage B.1.526 (Spike protein substitutions L5F, (D80G*), T95I, (Y144-*), (F157S*), D253G, (L452R*), (S477N*), E484K, D614G, A701V, (T859N*), (D950H*), (Q957R*))); a Kappa variant (e.g., Pango lineage B.1.617.1 (Spike protein substitutions (T95I), G142D, E154K, L452R, E484Q, D614G, P681R, Q1071H)); Pango lineage variant B.1.617.2 (Spike protein substitutions T19R, G142D, L452R, E484Q, D614G, P681R, D950N)), Lambda (e.g., Pango lineage C.37); or any combination thereof.

~~[0086]~~[0095] In some embodiments SARS-Cov-2 variant is a VON. In some embodiments, the SARS-Cov-2 variant is or includes Pango lineage variant P.1 (alias, B.1.1.28.1.) as described in Rambaut et al. 2020. Nat. Microbiol. 5:1403-1407) (spike protein substitutions: T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, TI027I)); an Alpha variant (e.g., Pango lineage B.1.1.7); a Beta variant (e.g., Pango lineage B.1.351, B.1.351.1, B.1.351.2, and/or B.1.351.3); Pango lineage variant B.1.617.2 (Spike protein substitutions T19R, G142D, L452R, E484Q, D614G, P681R, D950N)); an Eta variant (e.g., Pango lineage B.1.525); Pango lineage variant A.23.1 (as described in Bugembe et al. medRxiv. 2021. doi:

~~https://~~doi.org/10.1101/2021.02.08.21251393) (spike protein substitutions: F157L, V367F, Q613H, P681R); or any combination thereof.

~~[0087]~~[0096] In certain embodiments, the pathogen sequenced is a pathogenic bacteria and may include: spirochetes; spirilla; vibrios; gram-negative aerobic rods and cocci; enterics; pyogenic cocci; and endospore-forming bacteria; actinomycetes and related bacteria; rickettsias and chlamydiae; mycoplasmas, which are groups defined by some bacteriological criteria. A pathogenic bacteria may include: *Escherichia coli, Salmonella enterica, Salmonella typhi, Shigella dysenteriae, Yersinia~~Yersina~~ pestis, Pseudomonas aeruginosa, Vibrio cholerae, Bordetella pertussis, Haemophilus influenza, Helicobacter pylori, Campylobacter jejuni, Neisseria gonorrhoeae, Neisseria meningitidis, Brucella abortus, Bacteroides fragilis, Staphylococcus aureus, Streptococcus pyogenes, Streptococcus pneumoniae, Bacillus anthracis, Bacillus cereus, Clostridium tetani, Clostridium perfringens, Clostridium botulinum, Clostridium difficile, Corynebacterium diphtheriae, Listeria monocytogenes, Mycobacterium tuberculosis, Mycobacterium leprae, Chlamydia trachomatis, Chlamydia pneumoniae, Mycoplasma pneumoniae, Rickettsias~~Rickettisas~~, Treponema pallidum, Borrelia burgdorferi*, or a variant thereof. (Todar, K. Textbook of Bacteriology (2020) Online).

~~[0088]~~[0097] In an exemplary embodiment, the pathogen sequenced is a pathogenic fungi and may include: Aspergillus; Blastomyces; Candida; Coccidioides; Cryptococcus; Fusarium; Microsporum; Epidermophyton; Trichophyton; Histoplasma; Rhizopus; Mucor; Rhizomucor; Syncephalastrum; Cunninghamella; Apophysomyces; Lichtheimia (formerly Absidia); Eumycetoma; Pneumocystis; Trichophyton; Microsporum; Epidermophyton; Sporothrix; Paracoccidioides; Talaromyces or a variant or species thereof. (CDC)

~~[0089]~~[0098] In an exemplary embodiment, the pathogen sequenced is a pathogenic protozoa belonging to the group: Sarcodina; Mastigophora; Ciliophora; or Sporozoa defined by their mode of movement. (CDC) In certain embodiments, the pathogenic protozoa may include: Entamoeba; Trichomonas; Leishmania; Chilomonas; Giardia; Isopora; Sarcocystis; Nosema; Balantidium; Eimeria; Histomonas; Trypanosoma; Plasmodium; Babesia; or Haemoproteus or a variant or species thereof.

~~[0090]~~[0099] Further embodiments are illustrated in the following Examples which are given for illustrative purposes only and are not intended to limit the scope of the invention.

EXAMPLES

~~[0091]~~[0100] Here Applicants designed, optimized, and implemented a novel sample identification method using synthetic DNA spike-ins (SDSIs) that is broadly compatible with SARS-CoV-2 sequencing approaches and settings. Applicants implemented these SDSIs for Illumina sequencing with SARS-CoV-2 specific amplification using the ARTIC consortium's amplicon designs. To maximize epidemiological utility by increasing the number of genomes recovered from samples with low viral loads, Applicants benchmarked key amplification and library construction steps. Applicants propose a modified protocol, hereafter termed SDSI+ARTIC, that provides increased confidence in the veracity of genomes with minimal extra cost and time that can be applied to investigations of SARS-CoV-2 epidemiology and emerging viral variants (**FIG. 1**).

Example 1 – Design and in silico validation of novel amplicon spike-ins

~~[0092]~~[0101] Applicants sought to design a robust system for contamination tracing and sample tracking applicable to a wide-variety of viral sequencing strategies via known synthetic DNA sequences. Applicants envisioned that these novel synthetic DNA spike-ins (SDSIs) would consist of a uniquely identifiable sequence such that each sample in a sequencing batch could be paired with a different SDSI, enabling in-sample labeling. SDSIs should be sufficiently distinct from one another as well as common laboratory or human pathogens to ensure reliable identification. Each unique sequence is then flanked by constant priming regions so that a single additional primer set can be integrated into a multiplexed PCR to co-amplify the SDSI with the sample **(FIG 2A)**. In labeling all amplified viral genomic material in a laboratory setting, Applicants could track sample swaps and viral contamination with exquisite resolution and accuracy.

~~[0093]~~[0102] Excerpting DNA sequences from diverse, exotic archaea genomes to serve as the unique portion of the SDSI precludes false detection and cross-identification. To balance common sequencing library construction constraints, DNA synthesis costs, and providing enough sequence to be uniquely identifiable, Applicants generated SDSIs with a 140 bp stretch of variable sequence. Applicants confirmed that the various SDSIs were significantly different from each other to mitigate cross-identification; among all SDSIs, the minimum pairwise Hamming distances of the 140 bp stretch of unique sequence was 84 (mean=105; max=121). Since false detection of SDSI would occur if its sequence shared significant homology with other genetic material in a sample,

Applicants based these sequences on archaea, which are divergent from organisms found in typical laboratory or clinical settings **(Table 2)**. A permissive search performed against the entire NCBI database confirmed that 44/48 SDSI sequences had significant homology (>75% sequence identity over >75% query cover) exclusively within the domain archaea; the remaining SDSIs had homology to a handful of bacterial genuses unlikely to be found in laboratories (**Table 2**). In considering the application of these SDSIs to ARTIC SARS-CoV-2 amplicon sequencing, Applicants also specifically verified that each unique SDSI sequences were unlikely to be confused with expected COVID-19 clinical sample content, confirming that each sequence had very limited homology (nothing >50% sequence identity over >50% query cover) to both *Homo sapiens* and SARS-CoV-2. In designing these amplicon sequences Applicants also avoided extremes of GC content (range: 35-65%) in order to promote similar amplification rates across different SDSIs, as well as other potential targets of the multiplexed reaction, such as viral amplicons. Applicants specifically ensured that the SDSIs had similar GC content to ARTIC SARS-CoV-2 amplicons (**FIG. 6**).

~~[0094]~~**[0103]** Similarly, the design of common primers for SDSI amplicons enabled compatibility with a broad spectrum of amplicon-based sequencing reactions, including in clinical settings. To preclude off-target priming in the PCR reaction that could outcompete amplification of a primary target, Applicants limited SDSI primer homology to common organisms, particularly on the 3' end of the primer. Applicants specifically confirmed that primers were unlikely to amplify human or SARS-CoV-2 to promote SDSI primer integration into the ARTIC SARS-CoV-2 amplicon sequencing PCR reaction. Primers were compatible with ARTIC v3 primer sets, with a similar length (24 bps each) and GC content (45.8% each) (**FIG. 6**).

Example 2 – Application of spike-ins to ARTIC SARS-CoV-2 sequencing

~~[0095]~~**[0104]** Applicants demonstrated that the addition of SDSIs into the ARTIC multiplexed PCR provided a sample-specific internal control and did not detrimentally affect the amplification of SARS-CoV-2 RNA. SDSI primers did not produce any nonspecific amplification, including in the presence of NP swab RNA, supporting the expectation that primers shared limited homology with genomic material from clinical samples **(FIG. 6)**. All SDSIs amplified in an ARTIC SARS-CoV-2 PCR reaction with SDSI primers included, in each case yielding a single clean product of the expected size **(FIG. 6)**. Applicants next sought to ensure that inclusion of the SDSI oligo and

SDSI primers did not limit amplification of SARS-CoV-2 RNA. To prevent SDSIs overtaking the amplification and sequencing of SARS-CoV-2 amplicons, Applicants optimized the amount of SDSI added to each reaction through limited titration (**FIG. 7**). Applicants found that 1µl of a 1fM SDSI resulted in the reliable detection of the SDSI across a range of CT values (CT 20, 25, 30, 35) while the majority of reads (>96%) still mapped to SARS-CoV-2 **(Table 3; FIG. 2B)**.

~~**[0096]**~~**[0105]** Applicants performed SDSI+ARTIC sequencing on a batch of 48 SARS-CoV-2+ clinical samples to demonstrate its feasibility and utility in tracking samples and identifying contamination. After adding a different SDSI to each sample, Applicants found that 47/48 SDSIs were identified exclusively in the anticipated sample, validating the use of SDSIs as an internal control for sample tracking. One SDSI (SDSI_48) was detected in the sample that it was added to as well as a neighboring sample in the batch **(FIG. 2C)**. Applicants suspect that this represents unintentional within-batch contamination that was likely a consequence of spillover between neighboring wells. This case reveals the insidious nature of commonplace contamination and underscores the importance of this novel method for identifying it.

~~**[0097]**~~**[0106]** As shorter amplicons have been purported to yield superior recovery for low viral load samples (Antonov et al., 2005; No et al., 2019)), Applicants explored extending SDSIs to the Paragon Genomics' CleanPlex SARS-CoV-2 panel, but identified fatal shortcomings. Paragon amplicons are on average half the size of ARTIC (149bp vs 343bp), and compatible with the SDSI length 140bp. (Antonov et al., 2005; No et al., 2019) (*SARS-CoV-2 COVID-19 Coronavirus Research and Surveillance*, n.d.)(Antonov et al., 2005; No et al., 2019). However, the Paragon panel had dropout regions even in low CT samples which resulted in missed SNP calls compared to ARTIC across 5 samples (CTs = 20-37), consistent with other reports (**FIG 8A, 8B**) (Klempt et al., 2020). Although this panel did recover more of the genome in very high CT samples (>35), Applicants did not proceed with SDSI integration as the uneven and unreliable genome coverage across most clinical CTs limited Paragon's epidemiological utility (**FIG. 8C**).

Example 3 – Improving genome recovery and coverage for Illumina-based ARTIC SARS-CoV-2 sequencing

~~**[0098]**~~**[0107]** Applicants benchmarked various alterations to Illumina-based SDSI+ARTIC sequencing in order to maximize the number of complete, high-quality genomes recovered from clinically diverse samples. Higher CT samples prove especially challenging to sequence but their

recovery is still of critical importance to epidemiological and clinical applications of viral genomics. Applicants found that substituting a more processive reverse transcriptase provided the single biggest benefit. Comparing cDNA produced with Superscripts III, IV, or IV-VILO across a range of clinical CTs (low CT: <20, mid-low CT: 20-25, mid-high CT: 25-30, and high CT: >30), SSIV-VILO and SSIV produced the highest number of amplicons with at least 10X coverage across 13 samples (SSIII: 72.64%, SSIV: 81.93%, SSIV-VILO: 86.97%) **(FIG. 3A)**. These processive reverse transcriptases also displayed lower variability as measured by the percent of amplicons with <20% mean coverage (SSIII: 36.89% SSIV: 31.24% SSIV-VILO: 22.45%) **(FIG. 9A)**. Applicants also tested five DNA polymerase and conditions in the SDSI+ARTIC PCR reaction **(Methods)** and found that Q5 Hot Start High-Fidelity 2X Master Mix and KAPA reactions yielded the highest amplification (average 85.3nM and 56nM respectively) **(FIG. 9B)**.

~~[0099]~~[0108] Applicants also attempted protocol modifications to increase sequence depth uniformity in SDSI+ARTIC, which is crucial for recovering complete genomes in the fewest number of reads. When Applicants increased (2x) primer concentrations (20.8nM final) for low efficiency amplicons, Applicants observed increased coverage in these amplicons that enabled whole genome recovery for multiple samples, especially those with higher CTs **(FIG. 3B; FIG. 10; Table 4).** Other groups have also noted that alternative primers or changes in annealing temperature can reduce the formation of certain primer interactions, and Applicants suspect exploration of these avenues would further optimize SDSI+ARTIC (Itokawa et al., 2020). Applicants also attempted to recover high CT samples by increasing the number of PCR cycles and observed greater coverage uniformity with increasing cycles (**FIG. 3C**). However, at 45 cycles Applicants observed 3 SNPs that were not present in lower-amplified samples. To avoid erroneous SNP calls, Applicants decided to implement and optimize the SDSI's for a 40 cycle PCR. Additional modifications such as DNA-rehybridization steps (Mathieu-Daudé et al., 1996) or slower temperature ramp speeds had no significant effects **(FIG. 9C, 9D)**.

~~[0100]~~[0109] Applicants reduced the potential for highly amplified library contamination within the laboratory or clinical setting by scaling down (.5X) the Illumina DNA Flex library construction kit, which also reduced per sample cost without impacting performance (**Table 5; Table 6**). In benchmarking library construction methods, Applicants confirmed Nextera DNA Flex generated greater coverage depth than DNA XT (**FIG. 10**). In combination, the final suggested modifications

to Illumina ARTIC sequencing include using more processive reverse transcriptases, 40 cycles of PCR, and 2X primer concentration to recover higher CT samples, as well as a .5X scale down of Illumina DNA Flex to produce less concentrated, and thus less likely to contaminate libraries at a halved cost. Integrating these modifications into the SDSI approach may enable greater genomic surveillance in a limited number of samples.

Example 4 – SDSI-ARTIC sequencing benchmarks well against metagenomic sequencing.

~~[0101]~~[0110] Highlighting the reliability and robustness of this approach, Applicants observed high sequence correlation and superior genome recovery with SDSI+ARTIC compared to an unbiased metagenomics approach, the gold standard in generating error-free viral genomes. Applicants sequenced a small batch of six samples (CTs = 16-31) using ARTIC without SDSIs, and generated full length genomes with 100% concordance to those generated with metagenomic sequencing, indicating the accuracy of ARTIC-based sequencing methods (Lemieux et al., 2021). Applicants then resequenced 89 unique patient samples with SDSI+ARTIC that were previously sequenced using the same standard metagenomics approach (Lemieux et al., 2021) to serve as a direct comparison. The 89 samples in the validation batch consisted of diverse viral lineages and a broad range of CTs (range = 11.9-37.4; mean = 27.4) (**FIG. 12A**). SDSI+ARTIC outperformed metagenomic sequencing in terms of genome recovery, with increased median assembly lengths (29,577 bp and 4,389 bp respectively) (**FIG. 4A**), and a higher number of complete (>98%) genomes assembled (50 and 31 respectively). Applicants recovered even more partial (>80%) genomes with SDSI+ARTIC when compared to metagenomic sequencing (75 vs 36 respectively). Notably, 5 complete genomes recovered for SDSI+ARTIC had a CT above 30 (**FIG. 4B; FIG. 12B).** Applicants also assessed coverage uniformity in both methods, as increasing uniformity reduces the sequencing depth required to generate reliable genomes, thus improving throughput and efficiency. (So et al., 2018). As measured by a gini coefficient for each sample that generated an assembly, uniformity decreased in both methods above a CT of 25 but was markedly worse for metagenomics (**FIG. 12C**).

~~[0102]~~[0111] SDSI+ARTIC displayed high concordance in sequence variant identification to metagenomics, producing only two divergent SNP calls out of 331 total SNPs across 38 genomes (**FIG. 4C**). Notably, this discordance was present with both relaxed (n=3) and conservative (n=20) minimum coverage thresholds. The discordant SNPs, observed in two samples, were present in

different regions. However, both were located in ARTIC primer regions and matched the primer sequence even though primer trimming was performed and confirmed by manual inspection. Additionally, the coverage depth in the regions of the SNPs was greater than 1000X for both platforms in both samples. Applicants believe these errors likely arose during the ARTIC PCR, suggesting a discordance rate of 0.6% between amplicon-based and metagenomic sequencing. Notably these few mismatches did not result in lineage misassignment for either sample. To evaluate concordance, Applicants compared consensus sequences without down sampling using only samples that produced a full genome to make the most equivalent comparison (Methods).

Example 5 – Rapid deployment of SDSI-ARTIC sequencing confirms a suspected nosocomial transmission cluster

~~**[0103]**~~**[0112]** SDSI+ARTIC is a powerful method for public health interventions, especially as superspreading events - and clusters of cases linked to close contact settings more broadly - have become a defining feature of the SARS-CoV-2 pandemic ((Adam et al., 2020; Dearlove et al., 2020; Lemieux et al., 2021; Wong & Collins, 2020)). Viral genomes can reveal whether these clusters are linked through transmission, based on shared viral sequences, providing useful information for public health interventions. Such outbreak investigations of single cases leading to many are distinguishable due to low viral sequence variation but requires higher levels of confidence to ensure such a pattern has not occurred due to laboratory contamination. To demonstrate the utility of the novel SDSIs and modified protocol, Applicants applied the method to investigate a putative cluster of 14 SARS-CoV-2 cases from Massachusetts General Hospital (MGH), for which the infection control unit had suspicion of a nosocomial outbreak. Applicants sequenced 24 samples; 14 samples believed to be part of the cluster based on traditional contact-tracing, 8 unlinked samples and 2 negative controls.

~~**[0104]**~~**[0113]** The SDSI+ARTIC method enabled fast and confident identification of a nosocomial cluster, with samples processed within 24 hours and final genomes assembled within 52 hours of bio-sample receipt. Applicants assembled 14 complete genomes (>98% complete) of which 9 were from cluster-associated samples. Those samples that did not yield a full genome were those with lower viral loads (CT > 30). Phylogenetic analysis showed that samples from the cluster were genetically highly similar and clustered together (**FIG. 5A, 5B**) to the exclusion of other samples from Boston around the same time, strongly suggesting that this cluster did reflect

transmission within the hospital. One sample, MA-MGH-02834, differed from other cluster-associated samples by 18-19 consensus-level variants suggesting that this infection was likely acquired separately and not as part of the same nosocomial transmission. Analysis of the SDSIs confirmed that genome sequence similarity was not the result of cross-contamination from highly amplified final libraries (**FIG. 5C**).

Example 6 – Discussion on novel amplicon spike-ins

~~[0105]~~[0114] As the SARS-CoV-2 pandemic intensifies and new genomic variants continue to emerge, it is imperative to build robust experimental confidence into genomic surveillance data interpretation. Here, Applicants report a novel design and implementation of Synthetic DNA Spike-ins (SDSI) as an essential component for tracking and tracing contamination, a potential confounder in amplicon-based sequencing methods of SARS-CoV-2. The in-silico design generated robust synthetic targets at low costs while mitigating inter-spike-in sequence homology as well as homology with human, SARS-CoV-2, and common laboratory reagents. While broadly applicable to most amplicon-based approaches, as a proof-of-principle Applicants coupled the SDSIs to an improved ARTIC amplicon sequencing protocol yielding faster throughput with an overall reduced cost compared to existing Illumina DNA Flex-based protocols.

~~[0106]~~[0115] SDSIs can readily be adopted by laboratories and platforms of all sizes with only minor changes to existing methodologies, little additional cost per sample ($0.006), and no interruption to standard workflow methodologies. Additional synthetic targets could be designed using the same principles to expand into 384 well formats and beyond. Primer sites could also be modulated for integration with new advancements in amplicon sequencing, like tailed primer approaches (Gohl et al., 2020). More broadly, standardizing controls across the viral surveillance community will increase accuracy and integrity of SARS-CoV-2 genomic data worldwide. These SDSIs not only enable profiling of in-batch contamination, but also laboratory-wide detection as their presence in other data (amplicon, metagenomic, qPCR, or otherwise) would indicate a tagged amplification and thus contamination. Moreover the approach is applicable to both Illumina and Nanopore sequencing platforms as well as any other existing or future tiled amplicon panel, such as those previously used for Zika, Ebola, and other recent outbreaks (Quick et al., 2016) (Metsky et al., 2017). SDSIs could serve as a broad tool for tracing potential contamination across a plethora

of fields that employ amplicon based genomic sequencing, such as food safety, species identification or environmental sampling.

~~[0107]~~**[0116]** In optimizing the SDSI+ARTIC protocol Applicants tested and incorporated a number of cost and time saving adjustments. Modifications that can be used include implementing liquid handlers in high volume settings such as public health laboratories. Additional methodological improvements could allow for direct PCR amplification of SARS-CoV-2 using primers with indexing adapter compatible ends (Baker et al., 2020; Gohl et al., 2020) or the inclusion of unique molecular identifiers to understand intra-host variation. The SDSIs were designed to be compatible with such potential future approaches. Applicants note that there is still considerable non-uniformity in per-amplicon coverage for samples with low viral loads highlighting the need for methods that can confidently capture this information. A recent update to the ARTIC protocol for nanopore suggests that a change in the annealing temperature from 65°C to 63°C can reduce dropout of amplicon 64 (Tyson et al., 2020), a particularly poorly performing amplicon. The results show that 2x primer concentration for a subset of underperforming amplicons improved performance, and matching primer concentrations with amplicon efficiency would likely yield more uniform coverage (**Table 4**). Alternative approaches for the recovery of genomes from samples with low viral load include the use of targeted enrichment approaches (Houldcroft et al., 2017; Metsky et al., 2019) are more costly and time-consuming.

~~[0108]~~**[0117]** Amplicon based sequencing methods fill a critical need for rapid turn around and full genome recovery for epidemiological surveillance where SNP identification is crucial. While benchmarking the modified protocol against the gold standard approach of metagenomics Applicants observed discordant SNPs were rare (2/331). This emphasizes the need for caution and replication of libraries for highly important samples. Other commercial amplicon-based designs such as those by Paragon Genomics are significantly faster workflows and use smaller size amplicons, but the ARTIC primer set results in better overall coverage for the majority of samples (up to CT=35) and genome accuracy. Applicants believe subsequent generations of amplicon-based sequencing will address this pressing need pushing cost down while increasing genomic surveillance accuracy, which is sorely needed in the public health setting. The rapid deployment of SDSI+ARTIC confirming a nosocomial infection cluster further emphasizes the utility of the SDSIs to confidently identify samples of high genetic similarity.

The potential emergence of SARS-CoV-2 immune and vaccine escape variants underscores the ongoing necessity of accurate, reliable, and accessible genome sequencing. The modifications and suggestions build upon a remarkable global genomic surveillance response that has developed new tools for the rapid sequencing of viral genomes at an unprecedented rate. In light of the latest surges in SARS-CoV-2 cases globally and the emergence of more transmissible lineages and variants of concern that are rising in frequency in multiple continents, continual innovation in these protocols to improve their efficiency, cost-effectiveness and reliability are essential to meet the growing need for genomic surveillance of SARS-CoV-2. Moreover, stringent sample tracking and contamination detection strategies must become a standard practice, maximizing the utility of genomic data and its increasing importance for shaping public health interventions.

Example 7– Design and characterization of synthetic DNA spike-ins for AmpSeq

~~**[0109]**~~**[0118]** Applicants designed a simple and flexible system for sample tracking and contamination tracing using a core uniquely identifiable DNA sequence flanked by constant priming regions that satisfy several design requirements. This design allows in-sample tracking through the addition of a different SDSI to each sample during sample processing. Following sequencing, the data can be analyzed for both the presence of the expected SDSI and any other SDSI, illuminating both sample misassignment and contamination with high resolution and accuracy (**FIG 13**). Applicants focused the initial design on highly stable DNA oligos that would be added to sample cDNA and could capture contamination at or after the critical viral amplification step, including contamination generated during amplification and in handling amplified material. By using a longer unique core sequence, as compared to a short barcode system, these SDSIs are compatible with both tagmentation- and ligation-based sequencing approaches. The constant priming regions mean that only a single primer pair needs to be added into the existing multiplexed PCR step to co-amplify all SDSIs with the primary reaction target(s) **(FIG 14A)**. In particular, Applicants sought to design a system that could be integrated into diverse amplicon-based viral sequencing approaches. 96 distinct DNA sequences from the genomes of diverse, uncommon archaea serve as the core portion of each SDSI, precluding false detection and cross-identification **(Table 1, Methods)**. By using extremophilic archaea, the designs maximized evolutionary distance from common human pathogens. To avoid false positive results the core SDSI sequences should be sufficiently distinct from one another, as well as sequences commonly

found in laboratories and clinical samples. A permissive BLASTn search performed against the entire NCBI database confirmed that the unique SDSI core sequences had limited homology outside the domain archaea, specifically to genera unlikely to be found in laboratories (**FIG 18A**). While this limited homology outside of the domain archaea maximized the potential for broad applications, Applicants also confirmed that none of the core sequences shared significant homology with *Homo sapiens* or known viral genomes (**Methods**). Applicants considered significant homology as >90% sequence identity over 50 bps, as library construction can result in the generation of small fragments. Similarly, Applicants confirmed that all SDSIs were significantly different from each other to prevent misidentification; among all pairwise combinations of SDSIs, the greatest homology occurred between SDSI 14 and 18, which had 15 mismatches over 66 bps (**FIG 18B**). Sequencing of the SDSIs confirmed that each of the 96 constructs resulted in a robust and specific signal of mapped reads (**FIG 14B**).

~~**[0110]**~~**[0119]** Applicants selected a pair of primers and corresponding priming regions on each SDSI that are highly specific and show reliable amplification across SDSIs and under standard PCR conditions. Using Primer-BLAST, Applicants predicted that these sequences had limited homology to common organisms and thus were unlikely to amplify nonspecific templates that could outcompete amplification of a primary target. Experimentally Applicants confirmed that the SDSI primers did not produce any nonspecific amplification, including in the presence of cDNA from a nasopharyngeal (NP) swab sample **(FIG 19A)**. The primer pair also had a common length (24 bps), GC content (45.8%), and melting temperature (62℃ and 63℃, respectively in the SDSI+AmpSeq protocol), ensuring their compatibility with many multiplexed PCR reactions, including the most widely used SARS-CoV-2 amplicon sequencing strategy (artic.network/) (**FIG 19B**). Since each SDSI was identically sized and shared a priming region, a similar amplification rate was expected across all SDSIs. Applicants avoided extremes of GC content in SDSI amplicons (range: 33-65%) in order to promote similar amplification rates across different SDSIs and to viral amplicons (e.g., the GC content of the SARS-CoV-2 genome is roughly 37±5%) [19] (**FIG 19C**). Applicants confirmed experimentally that all SDSIs amplified in an ARTIC SARS-CoV-2 PCR reaction with SDSI primers included, in each case yielding a single clean product of the

expected size (**FIG 19D**). Furthermore, Applicants observed that GC content did not significantly bias the number of SDSI reads detected in clinical samples (**FIG 19E**).

Example 8 – Validation of an SDSI+AmpSeq SARS-CoV-2 sequencing approach

~~[0111]~~[0120] Applicants determined that the addition of SDSIs into the ARTIC multiplexed PCR did not detrimentally affect or otherwise alter the amplification of SARS-CoV-2 cDNA from clinical samples. First, to prevent SDSIs from overtaking the amplification and sequencing of SARS-CoV-2 amplicons, Applicants optimized the amount of SDSI added to each reaction through limited titration. Using a randomly selected SDSI (SDSI 49), Applicants found that the highest concentration tested, 600 copies/μL, resulted in reliable SDSI detection with >96% of reads still mapping to SARS-CoV-2 and no apparent alteration in coverage across the genome **(FIG 20A,B)**. Applicants then validated the specificity of the 96 selected SDSIs in a batch of clinical samples to confirm that there was no unpredicted cross-mapping, misidentification, or significant differences in amplification rate (**FIG 15A**). To assess more precisely how the addition of SDSIs would affect SARS-CoV-2 genome sequencing in clinical samples, Applicants processed 14 samples, spanning a range of CT values (CT range = 25-33), with both the standard ARTIC and SDSI+AmpSeq methods. For each amplicon, across all samples, there was no significant difference in coverage between the ARTIC and SDSI+AmpSeq conditions (**FIG 15B**). Even in samples with low viral loads (CT>30), Applicants found that there were no significant differences in amplicon coverage (**FIG 21A**). Additionally, within the 14 samples processed with and without an SDSI, Applicants see a 100% genome concordance rate illustrating the addition of the SDSIs does not impact recovery of accurate genomes.

~~[0112]~~[0121] As extensive PCR can result in the propagation of numerous types of errors, such as DNA polymerase base substitution errors, PCR recombination events, template switching, and thermocycling induced DNA damage, Applicants further compared SARS-CoV-2 genome concordance between the SDSI+AmpSeq method and unbiased, metagenomic sequencing[9,10,20]. Applicants performed SDSI+AmpSeq on a batch of 89 unique patient samples previously sequenced with unbiased metagenomics[21]. The samples consisted of diverse viral lineages and a broad range of viral loads (CT range = 11.9-37.4; mean = 27.4) with the more sensitive amplicon sequencing method generating more complete genomes at higher CTs (**FIG 22A-D**). Applicants assessed the coverage uniformity between the methods, as increasing

uniformity reduces the sequencing depth required to generate reliable genomes[22]. Applicants found that unbiased sequencing had more uniform coverage up to a CT of 25 (N=31, Gini Coefficient = 0.240 ± 0.046 (unbiased) vs 0.428 ± 0.026 (SDSI+AmpSeq)), while SDSI+AmpSeq generated more uniform coverage for samples above a CT of 25 (N=39, Gini Coeff = 0.766 ± 0.265 (unbiased) vs 0.554 ± 0.124 (SDSI+AmpSeq)) **(FIG 22E)**. For the 37 samples that assembled a full genome in both methods, only two out of 332 total single nucleotide variants (SNVs) identified compared to the reference (Wuhan-Hu-1) were divergently identified by SDSI+AmpSeq **(FIG 15C**). Each SNVs was observed in only one sample, and both fell within an ARTIC primer region, despite primer trimming during analysis; this suggests that PCR error from the ARTIC protocol may have contributed to the discrepancy[23]. Manual inspection of one SNV, (a C9565T mutation in unbiased sequencing) indicated the presence of intra-host variation in both methods with a variant allele frequency of 39.4% (SDSI+AmpSeq) and 59.2% (unbiased sequencing). Overall, the discordance rate between SNV calling for SDSI+AmpSeq and unbiased sequencing was 0.6%, a percentage that is reasonable with SNV rates and sequencing based errors. Consistent with previous reports from other groups, ARTIC amplicon sequencing maintains a high level of concordance at the consensus genome level[10], even with the addition of SDSIs.

~~**[0113]**~~**[0122]** Applicants explored a number of other technical modifications to the ARTIC amplicon sequencing protocol in order to improve genome recovery, limit contamination points, and enhance reproducibility of the SDSI approach. Foremost, increasing cDNA length by use of more processive reverse transcriptases improves amplicon coverage **(FIG 23A,B)**. Amplification of ARTIC amplicons and SDSIs by Q5 Hot Start High-Fidelity 2x Master Mix results in higher amplification **(FIG 23C, Table 7**). Applicants found that increasing (2X) primer concentrations (20.8nM final) for poor performing amplicons increased coverage in these amplicons, even enabling whole genome recovery for multiple samples supporting that primer rebalancing can enable greater coverage[24,25] **(FIG 23D, FIG 24, Table 4).** Applicants then explored the effects of different numbers of PCR cycles, DNA-hybridization steps, and temperature ramp speeds. Both DNA-hybridization steps and temperature ramping provided no significant changes in amplification **(FIG 23E,F)**. Although it may lead to a potential increase in erroneous SNV calls[23], additional cycles of PCR can be beneficial for low viral load samples by increasing genome coverage uniformity **(FIG 23G).** Using a standardized cDNA input, Applicants found that the

DNA Flex library preparation kit resulted in an increased depth of coverage for the SARS-CoV-2 genome across all CT values tested, compared to Nextera XT (**FIG 23H**). To further mitigate the risk of contamination from such highly amplified libraries, Applicants took advantage of the self-normalizing feature of the DNA flex kit and found that limiting the tagmentation beads by scaling down (.5X) all components of the DNA Flex library construction reagents restricted library over-amplification. Notably, this limitation did not impact final library size distributions or SDSI amplification, while having the desired effect of generating final sequencing libraries at half their original concentrations (**Table 8**). This approach also had the added benefit of nearly halving the library construction cost per sample (**Methods**). Applicants have summarized the results of the optimizations within the full SDSI+AmpSeq protocol (~~https://~~benchling.com/s/prt-R95g0tCxKOeCAqn8lAk3); additionally, Applicants have found that the SDSIs can be easily integrated with numerous protocol alterations.

Example 9 – Implementation of SDSIs to SARS-CoV-2 clinical samples at scale

~~[0114]~~[0123] The SDSI+AmpSeq method is compatible with a range of viral CTs, SARS-CoV-2 lineages, origin of the patient sample, and laboratory in which the pipeline is implemented demonstrating that this is a robust and flexible approach that can be readily implemented for surveillance. A half plate of SDSIs were used at two large-scale sequencing facilities, the Broad Institute and Jackson Laboratories (JAX), for SDSI+AmpSeq SARS-CoV-2 surveillance across a total of 6,741 clinical samples and controls (JAX: N=3,838; Broad: N=2,903). Individual batches typically consisted of 92 clinical samples with 4 designated water controls. Clinical samples were largely from Maine, Massachusetts, and Rhode Island from December 2020 to July 2021 and covered a wide range of viral CT values (CT 8.4-39.9) and pango lineages (77 total lineages) (**FIG 16A**). The SDSI+AmpSeq method worked robustly despite minor implementation differences in protocols between the two laboratories including alterations in the cDNA synthesis enzymes (SSIV vs Lunascript), CT normalization implementation, and library construction approaches (0.5X Illumina DNA Flex vs Illumina COVID-Seq) (**Methods**).

~~[0115]~~[0124] The SDSI+AmpSeq is a tractable and easily-implemented method for genome quality control when applied to high-throughput processing of clinical samples. Across thousands of clinical samples, the SDSIs performed consistently and reliably (**FIG 16B,C**). The mean percentage of SDSI reads that mapped to the expected SDSI was above 95% for all SDSIs in both

laboratories (**FIG 16B**). This demonstrated that across a large set of highly variable clinical samples, there were no systemic issues of misidentification for specific SDSIs. Additionally, across all samples from both institutions, the percentage of all SDSI reads in SARS-CoV-2 positive samples averaged 3.71% (90% of samples fell between 0.002-9.989%) (**FIG 16C**). Each SDSI consumed roughly the same read percentage, with no SDSI consistently absent or regularly taking up more than 10% of the sample reads, supporting the prediction that the unique constructs amplified at similar rates. Importantly, this low, but consistent percentage of reads mapping to SDSIs allows for their implementation without needing to greatly increase sequencing depth. Across batches, SDSIs also take up roughly similar shares of the reads, indicating that the SDSI+AmpSeq method is consistent over time. Notably, the SDSIs performed well with and without prior normalization of cDNA based on CT, however normalizing did increase the percentage of SDSI reads **(FIG 21B, FIG 16B left, Methods)**. Normalization of viral CT may provide an additional level of quality control that is especially important for labs with limited sequencing capacities.

Example 10 – SDSI+AmpSeq provides highly confident genome sequencing and analysis

~~[0116]~~[0125] SDSIs enable detection of sample swaps and contamination events that occur in large scale batch processing which may otherwise go undetected. In a controlled experiment, Applicants demonstrated that the SDSI+AmpSeq approach provides a feasible method to accurately detect contamination. Applicants mixed two SDSIs at various ratios prior to the ARTIC PCR and found that those SDSI ratios were reflected in the sequencing output (**FIG 17A**). With evidence of SDSI's robust detection, uniqueness, and ability to detect intentional contamination, Applicants proceeded to use them to identify sample swaps and contamination in large batch processing. Across thousands of SARS-CoV-2 samples processed, SDSIs detected in samples to which they were not intentionally added allowed for identification of multiple key modes of error (**FIG 17B**). As plotted, a plate without contaminating events or sample swaps should display a simple diagonal pattern with 1-1 matching of expected and observed SDSIs. In some cases, off-diagonal events occur in clear patterns, enabling speculation on the nature of the contamination, clearly demonstrating the utility of SDSIs as an internal control and in-sample label. Applicants observed cases where a plate was likely inverted when SDSI+AmpSeq pool 1 was mixed with pool 2 (**FIG 15B**). The SDSI+AmpSeq approach allows researchers to detect entire flawed batches that

may not have been flagged with standard controls (as in the case with the plate inversion where water controls in plate corners would not have been affected). In another example, SDSIs were detected unexpectedly throughout a batch, indicating that SDSI (and possibly SARS-CoV-2 and other genetic material) contaminated a common reagent.

~~[0117]~~[0126] SDSI+AmpSeq also enables fine-resolution insight into sample processing errors with high specificity. In one example, SDSI counts indicated columns were unintentionally mixed together (**FIG 17B**). Here, in-sample labeling in all wells allowed researchers to confidently move forward with analyses on unaffected samples. In other cases, samples are associated with both the expected SDSI and SDSIs that were expected in neighboring samples. This indicates a potential spillover event or pipetting errors. Again, genomes generated from samples with suspicious SDSI profiles can be investigated further, and potentially removed from analyses and/or reprocessed. Applicants recommend manual curation of genomes assembled from any samples with <95% of SDSI reads mapping to the expected SDSI. This level of impurity is likely attributable to sample processing contamination, given minimal baseline crosstalk from sources like indexing primer or oligo synthesis observed (Methods, **FIG 25**). Moreover, these patterns of contamination events identified via use of SDSI+AmpSeq illuminated key sources of error in processing pipelines and provided an opportunity to improve processing fidelity in subsequent batches.

~~[0118]~~[0127] To demonstrate the application of the SDSIs for confident interpretation of sequencing data Applicants used SDSI+AmpSeq to investigate a putative SARS-CoV-2 cluster from Massachusetts General Hospital (MGH) for which the Infection Control Unit suspected nosocomial transmission, a context in which both sample swaps and contamination could easily undermine findings. Applicants sequenced 22 samples with SDSI+AmpSeq (14 samples suspected to be part of the cluster based on epidemiological contact-tracing and 8 unlinked samples as controls), within 24 hours and final genomes were assembled within 52 hours of biosample receipt. Of the 11 samples that Applicants assembled genomes from that were suspected to be part of the cluster, 10 were genetically highly similar (0-1 consensus nucleotide difference) **(FIG 17C)** and distinct from other samples from Massachusetts around the same time **(FIG 26)**, strongly suggesting that this cluster did arise from nosocomial transmission. Analysis of the SDSIs confirmed that genome sequence similarity among cluster-associated samples was not the result of cross-contamination **(FIG 17C)**. Indeed, 23/24 libraries (22 patient samples and 2 water

controls) contained >95% SDSI-mapped reads corresponding to the expected SDSI. One sample that was not part of the cluster (MA_MGH_02845) showed 18% of reads from a second SDSI, which was added to a different sample in the batch (MA_MGH_02839). Applicants re-sequenced both samples implicated in the contamination event. Applicants confirmed that the two genome sequences for MA_MGH_02845 were 100% concordant, and no genome was assembled for MA_MGH_02839 in either attempt, likely due to its very low viral load (CT = 37). This example illustrates how SDSIs can be used to isolate and validate only those samples implicated in contamination events and altogether increase confidence in cluster investigations.

~~[0119]~~[0128] To further increase the confidence in AmpSeq methods for viral genomics, Applicants sought to capture contamination and sample swaps that might occur before the cDNA stage. Applicants explored the feasibility of modifying the SDSI approach to enable synthetic RNA spike-ins (SRSI) from the same constructs, which could be added to clinical sample RNA to provide end-to-end quality control. For a subset of SDSIs, Applicants included a T7 promoter site to enable in-vitro production of these constructs as RNAs. For two clinical samples representing low (20) and mid (26) CTs, Applicants detected reads from the RNA spike-ins added directly to extracted viral RNA as a proof of principle (**FIG 27**). Notably, this approach did not require any additional protocol modifications, and Applicants therefore expect it to be a highly versatile and user-friendly method when deployed at scale for complete end-to-end sample tracking.

Example 11– Discussion on SDSI+AmpSeq

~~[0120]~~[0129] Amplicon-based sequencing methods crucially empower rapid, full genome recovery for emerging SARS-CoV-2 variant surveillance; however, robust tools are needed to ensure accuracy in genomic data. SDSI+AmpSeq is a reliable technique for detecting key modes of contamination, addressing this critical gap in standard controls and practices. SDSIs do not compromise genome quality, have been successfully deployed in thousands of clinical samples, and are in use across multiple laboratories with differing protocols. These SDSIs revealed numerous instances of sample swaps and contamination, many of which would go unnoticed with standard batch-level controls. SDSIs further provide critical confidence in the interpretation of clusters of identical genomes, a renewed challenge in the surveillance of more transmissible variants. The common primer design of the SDSI approach enables them to be readily applied to

multiple short amplicon designs and sequencing strategies, adding only minor changes to existing protocols and minimal additional cost.

~~[0121]~~**[0130]** SDSIs overcome multiple modes of error in the production of amplicon-based genomic sequencing data and are a critical component of quality control measures. The approach is most effective when adopted fully within a laboratory setting and thus Applicants propose routine use of the SDSI+AmpSeq method to flag laboratory-wide contamination. Applicants have implemented SDSI's across diverse approaches and provide an extensively tested protocol with ARTIC v3 and Illumina-based tagmentation. It can also be applied to other sequencing pipelines, though this potentially requires further optimization. The pathogen-exclusion design criteria allows the 96 validated SDSIs to be immediately incorporated into other tiled amplicon panels, such as existing ones for Zika, Ebola, and other viruses of epidemic potential[26,27].

~~[0122]~~**[0131]** The SDSI-labeling paradigm is broadly applicable to many amplicon-based needs: amenable to a variety of technical enhancements, flexible to remaining error modes, and expandable to additional targets. One could apply the same design parameters to expand the set of SDSIs, such as to 384 well formats. Additionally, uniquely permuted sets of any size could be created for specific sample batches. To design larger panels of SDSIs, Applicants could use artificial core sequences, rather than excerpting from archaea. Primer sites could also be easily adapted for integration with new advancements in amplicon sequencing, like tailed primer approaches or new primer schemes[28–32]. In its current implementation, the SDSIs detect contamination or workflow errors that occur during and after amplification, but not issues arising at the RNA or cDNA generation stage, and act qualitatively, rather than quantitatively. Further refinement of the RNA spike-in approach could address other modes of contamination, enabling end-to-end sample tracking at scale. Future work improving quantification and SDSI analysis pipelines may enable them to serve as within sample controls, since samples or batches with outlier SDSI read counts may reveal missing or defective PCR components, incomplete mixing, thermocycling issues, or other types of experimental error.

The integration of SDSIs can mitigate a critical vulnerability of amplicon-based sequencing while preserving the many advantages, increasing the robustness of its use across laboratory and clinical settings. Adoption of controls across the viral surveillance community would increase accuracy and integrity of genomic data worldwide. Looking forward, SDSIs could serve as a crucial

component in improving data integrity in amplicon based genomic sequencing beyond infectious disease surveillance, such as food safety, species identification and environmental sampling.

Example 12 – Methods

SDSI design and in silico validation

~~[0123]~~[0132] Applicants designed synthetic DNA fragments that each contained a 140 bp unique sequence and constant priming regions. Core SDSI sequence homology to sequences from various organisms was predicted by a permissive BLAST search (blastn; 5000 max targets; E=10; word size=11; no mask for low complexity). Applicants considered homologies identified with this BLASTn search described above that were additionally >50 bps (>35% query cover) and >90% sequence identity to be significant homologies. For all 96 selected SDSIs, there were no such significant homologies when results were filtered to all *Homo sapiens* (taxid:9606) or viral (taxid:10239) sequences in the NCBI database. For significant homologies to bacterial or eukaryotic sequences in the NCBI database (excluding archaea: taxid:2157), Applicants report both the SDSI and the genus it mapped to in each case (**FIG 18A**). Using the same BLASTn parameters, Applicants also mapped SDSIs against a custom database including SDSI core sequences and found no significant homologies between SDSIs. As there were no significant homologies between SDSIs and human, virus, or other SDSI sequences, Applicants noted the maximum alignment scores for any non-significant homology identified and reported these (**FIG 18B)**.

~~[0124]~~[0133] Applicants confirmed that SDSI primers and amplicons were predicted to amplify specifically and consistently with ARTIC v3 amplicons. Applicants used Primer-BLAST to predict 50-5000 bp amplicons produced on templates in the entire nr database; no amplicons were identified. Applicants calculated the length and GC content of SDSI primers and full SDSI amplicon sequences and ARTIC v3 primers and amplicons using Geneious Prime (2019.2.1) and compared their distributions (**FIG 19B-C**). ARTIC and SDSI primer melting temperatures were matched and calculated using the New England Biolabs online calculator (tmcalculator.neb.com).

SDSI experimental validation

~~[0125]~~[0134] Applicants sought to validate in silico predictions for the performance of the SDSI primers and amplicons. Applicants ordered primers (IDT) (oligo sequences in **Supplementary Data File 1**) and performed qPCR using the Q5 Hotstart 2x Mastermix, with 500nM SDSI primers

and 0.17X SYBR Gold (ThermoFisher #S11494), and without ARTIC primer pools. Applicants performed this assay in triplicate in 10µL reactions on a QuantStudio 6 with the following cycling conditions: 95°C for 30 seconds, followed by 35 cycles of 95°C for 15 seconds and 65°C for 5 minutes. Applicants tested 4 conditions: (1) 0.5µL of an SDSI gene block (IDT) (1pM), (2) 0.5µL of an SDSI gene block + 0.5µL of cDNA from an NP swab, (3) 0.5µL of cDNA from an NP swab, and (4) no template to detect any nonspecific amplification of the primers (**FIG 19A**). Applicants performed PCR on each SDSI oligo, using the standard SDSI+AmpSeq PCR conditions (benchling.com/s/prt-R95g0tCxKOeCAqn8lAk3), then ran the PCR products on a 2.2% agarose gel to confirm that these primers amplified the SDSIs and that the product was clean and of the expected size (**FIG 19D**).

~~[0126]~~[0135] Applicants ordered unique oligos as TruGrade ultramers (IDT), then resuspended and stored them at 10µM in water (oligo sequences in **Table 1**). Further characterization for identification of 96 SDSIs was achieved by direct PCR amplification with primers containing the constant SDSI handle and an Illumina P5/P7 adapter followed by sequencing with a MiSeq Nano 2x150bp kit (Illumina #MS-102-2002). SDSI reads were quantified by mapping each SDSI against other SDSIs with the align_and_count_multiple_report wdl implemented in Terra, as described below, and purity and sequence fidelity of SDSIs was achieved by calculating the percentage of reads mapping to each SDSI out of total SDSI reads (**FIG 14B**). Given these same data, Applicants explored the SDSI mapping stringency threshold. Applicants determined whether each SDSI was uniquely identified over a range of SDSI stringency thresholds (0.01%-50% of SDSI reads mapping, with a step size of 0.01%) (**FIG 25**). Applicants tested 142 total unique SDSIs; all SDSIs amplified successfully with high sequence fidelity and purity (>95% of reads mapped to the expected SDSI in the experiment described above). The final set of 96 SDSIs were chosen after first pass validation in a combination of clinical sample amplification tests, GC cutoffs, and sequence homology cutoffs. SDSIs excluded because of poor amplification or impurity in clinical sample processing were not retested to determine whether error was technical or biological.

Sample collection and study design

~~[0127]~~[0136] Research was conducted at the Broad Institute with an exempt determination from the Broad Office of Research Subjects Protections and with approval from the MIT Institutional Review Board under protocol #1612793224. Samples were obtained from Massachusetts General

Hospital (MGH), Massachusetts Department of Public Health, the Rhode Island Department of Public Health and the Broad Institute Clinical Research Sequencing Platform. Samples from Massachusetts General Hospital (MGH) fall under Partners Institutional Review Board under protocol #2019P003305. Samples were secondary-use or residual clinical and diagnostic specimens (referred to collectively throughout as clinical samples), obtained by researchers under a waiver of consent. All samples were nasopharyngeal or anterior nares swabs in a stabilizing medium (e.g., MTM or VTM). These unique biological materials are not available to other researchers as they are human patient samples from clinical excess material and thus are of limited volume. Samples sequenced at Jackson Laboratories (JAX) were approved under protocol 2020-NHSR-019-BH.

Viral CT Determination

~~[0128]~~**[0137]** Viral cycle threshold (CT) for all samples sequenced at the Broad Institute were obtained using the CDC RT-qPCR assay with the N1 probe as previously described[21]. Viral CTs for samples sequenced at JAX were obtained from various providers and thus the RT-qPCR assays used are variable.

CT Normalization

~~[0129]~~**[0138]** CT normalization was performed by first setting a desired mock viral CT and calculating the difference between this desired mock viral CT and the measured viral CT of a given sample, rounding to the nearest whole number. Applicants next calculated the number of doublings required for the mock viral CT (assuming 100% PCR efficiency) and multiplied this by the volume of cDNA input to be used for the normalization. The final volume of water used to dilute the cDNA was the doubling factor minus the volume of cDNA input. An example calculation is illustrated below:

Example of CT Normalization:

N= Difference between actual and mock
X= Volume (μL) of cDNA to use for normalization
DF=Doubling factor is $X(2^N)$
Volume water for dilution (μL)=DF-X

Actual viral CT=23
Desired mock viral CT=27
N=27-23=4

X=1μL
DF= $1(2^4)$
Volume water for dilution (μL) = 16-1 = 15μL

Add 1μL of cDNA to 15μL nuclease free water.

~~[0130]~~[0139] This CT normalization was done for certain method development samples which are described throughout the manuscript as being "mock diluted" or "normalized to CT X". The nosocomial cluster was normalized to CT 27. The majority of batch data generated at the Broad Institute underwent CT normalization to CT 25. Batch data from JAX did not undergo CT normalization. CT normalization of the cDNA prior to the ARTIC PCR should reduce the potential for generating excessively large libraries from very high viral load samples, keep the percentage of SDSI reads in a detectable range (**FIG 21B**), and further reduce the need for additional normalization steps later in the pipeline.

cDNA generation and ARTIC amplification optimization

Reverse transcriptase

~~[0131]~~[0140] Applicants tested reverse transcriptase enzymes using extracted RNA from four SARS-CoV-2 positive clinical samples (CTs = 13.9, 23.9, 29.6, 33.6) **(FIG 23A,B)**. Applicants added 2μL of purified DNase treated RNA as input into SuperScript III (Thermo #18080093), SuperScript IV (Thermo #18091050), or SuperScript IV VILO (Thermo #11756500). Superscript IV (SSIV) reactions incubated at room temperature for 10 minutes, followed by 50°C for 60 minutes and an inactivation step at 80°C for 10min. Superscript IV VILO shared the same protocol, but with a temperature of 85°C for the inactivation step. Applicants input 2.5μL of cDNA for ARTIC pool #1 PCR under standard conditions for 40 cycles. Applicants then tested the resulting pool #1 using the scaled down Illumina DNA Flex library construction (as described in Methods below) and sequenced on Illumina Miseq (V2 reagent kit) with 2 x 150 bp paired end sequencing.

ARTIC PCR enzyme

~~[0132]~~[0141] Applicants tested PCR enzyme efficiency using extracted RNA from SARS-CoV-2 positive clinical samples followed by cDNA generation using SuperScript IV and diluted the resulting cDNA to a mock CT value of 35 for standardization across all PCR enzyme tests. Applicants set up the standard ARTIC PCR pool #1 and pool #2 using an input of 2.5μL, altering only the PCR enzyme and corresponding buffer. Applicants tested NEB Q5 Hot Start High-fidelity

2x Master Mix (Q5 2X MM) (NEB #M0494L), NEB Q5 Hot Start High-fidelity 2x Master Mix plus .01% SDS, NEB Q5 Ultra II Master Mix (NEB #M0544L), KAPA HiFi HotStart (Roche #KK2601), and KOD Hot Start DNA polymerase (Sigma-Aldrich #71842) **(FIG 23C)**. Applicants quantified the resulting ARTIC PCR amplicons using a High Sensitivity DNA Qubit kit, then input 25ng from each pool (50ng total) into scaled down Illumina DNA Flex library construction. The resulting libraries (except Q5 plus .01% SDS, which had no visible product using the Tapestation D1000 High Sensitivity Kit) were quantified and pooled on Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

Rehybridization PCR

~~[0133]~~**[0142]** Applicants optimized PCR cycling conditions on mock CT 35 cDNA (generated as described above) using standard ARTIC PCR primer conditions. Applicants performed a catch-up/rehybridization PCR under the following conditions: 98°C for 30s, 95°C for 15s then 65°C for 5 min (10 cycles), 95°C for 15s then 80°C for 30s then 65°C for 5 min (2 cycles), 95°C for 15s then 65°C for 5 min (8 cycles), 4°C hold **(FIG 23E)**. Applicants quantified the resulting ARTIC PCR amplicons using a High Sensitivity DNA Qubit kit, then input 25ng from each pool (50ng total) into scaled down Illumina DNA Flex library construction. Applicants then quantified these libraries and pooled on Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

Cycle Test

~~[0134]~~**[0143]** Applicants further optimized ARTIC PCR by modifying PCR cycle numbers. Extracted RNA from six SARS-CoV-2 positive clinical samples ranging from CT 27-37 were converted to cDNA with Superscript IV and amplified under standard ARTIC PCR reaction components (with Q5 2X MM) modifying the final number of cycles of PCR from 35, 40 and 45 **(FIG 23G)**. Applicants quantified cDNA and used at a standard 50ng of input for scaled down Illumina DNA Flex Library Construction, then quantified the resulting libraries and pooled on Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

Ramp Test

~~[0135]~~**[0144]** Applicants used mock CT 35 cDNA to test the effect of decreased ramp speed on genome recovery and coverage. ARTIC PCR conditions for this experiment were 98°C for 30 seconds, followed by 40 cycles of 95°C for 15 seconds and 65°C for 5 minutes with a cooling and heating ramping speed of 3°C/s. Applicants tested a slow ramp PCR protocol with the ramp speed

reduced to 1.5°C/s **(FIG 23F)**. Libraries were constructed with Illumina DNA Flex and were sequenced on Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

Primer Concentration Optimization

~~[0136]~~[0145] Under standard ARTIC protocol conditions, Applicants ordered lyophilized ARTIC v3 primers from IDT and resuspended in water at 100μM each. Pool #1 primers consisted of all odd numbered amplicons whereas pool #2 primers consisted of all even numbered amplicons. To generate the 100μM pool #1 primer stock, Applicants combined 5μL of each 100μM pool #1 primer, and repeated this protocol for the even numbered primers to give a 100μM pool #2 primer stock. Applicants selected a total of 20 amplicons as regions of low coverage from previous sequencing data (**Table 4**). Low coverage amplicons were present in both pools, with 11 coming from pool #1 and 9 coming from pool #2. For the primer 2X pools, Applicants spiked in primers for the corresponding amplicons at 2X the concentration (20.8nM final) of the other primers in the pool. For these low coverage primers, Applicants used 10μL of the 100μM stock rather than 5μL. Applicants diluted both the original and 2X primer pools 1:10 in nuclease free water to generate a 10μM working stock. Applicants then selected 8 samples with varying CT values to determine if selectively increasing primer concentrations reduced amplicon dropout **(FIG 23D)**. Applicants used the SDSI+AmpSeq protocol (without the SDSI or SDSI primers) and processed each sample with both the original primer pool, as well as the 2X primer pool, then sequenced these 16 samples on an Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing. Only 6 of the 8 samples generated complete genomes (>98%) in both conditions and were used for further analysis.

CT normalization experiment

~~[0137]~~[0146] The CT normalization experiment was performed by taking four individual clinical samples (CT = 18-25) with four randomly selected SDSIs and either not normalizing the cDNA or normalizing to CT 25, 26, or 27 prior to the ARTIC PCR **(FIG 21B)**. Samples were processed with the standard SDSI+AmpSeq protocol described below and were sequenced on a NextSeq 500 Mid Output Kit v2.5 (300 Cycles).

Illumina DNA Flex

~~[0138]~~[0147] Applicants performed a head-to-head comparison of standard Illumina Nextera DNA Flex and Nextera XT (Illumina #FC-131-1096) library construction kits **(FIG 23H)**. The Nextera XT protocol was performed as previously described[21,33]. Both library construction

methods were compared on post ARTIC v1 PCR amplicons from clinical samples. In short, applicants amplified samples with a range of SARS-CoV-2 viral CT values (CTs = 22.9, 26.2, 30.3) with ARTIC v1 primers, producing 400 bp size fragments. Applicants then quantified amplicons from each ARTIC primer pool and pooled in equal molar concentrations. Standard Nextera DNA Flex input was 100ng (50ng from each pool) and 1ng (.5ng from each pool) for Nextera XT. Applicants quantified and pooled the resulting libraries before sequencing on an Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

~~[0139]~~**[0148]** Applicants optimized Illumina DNA Flex library construction (Illumina #20018705) construction with the goal of reducing normalization steps, cost and increasing throughput. Applicants scaled down (.5X) Illumina DNA Flex throughout the standard Illumina sequencing protocol, also scaling down sample input for a total of 50ng (25ng from each primer pool). Due to the CT normalization step, applicants removed the pre-DNA Flex DNA concentration and pooling step. Applicants used 1-2µL of post ARTIC PCR amplicon as input into the scaled down DNA Flex library construction and performed post library construction quantification and pooling with more uniform library size and concentration, further reducing time and cost of pooling libraries for sequencing. This protocol was used for all method development experiments, the cluster investigation, and a portion of the batch data generated from both the Broad Institute and JAX.

SDSI+AmpSeq SDSI titration in ARTIC SARS-CoV-2 sequencing

~~[0140]~~**[0149]** To determine an optimal concentration for SDSIs in ARTIC SARS-CoV-2 sequencing, applicants diluted SDSI 49 to 0.6, 6, 60, and 600 copies/µL (1, 0.1, 0.01, and 0.001fM); 1µL of SDSI 49 was added to 5µL of cDNA, to be split to 2x3µL for each ARTIC pool (**FIG 20, Table 1**). SDSI primers were added to each ARTIC pool with a final concentration of 40nM. The cDNA from one clinical sample (MA_MGH_00195; CT =16) was mock diluted to CT 20,25,30, and 35 for this experiment using the protocol described within the CT normalization section. Based on the results of this experiment, SDSIs were used at 6e2 copies/µL (1fM) for all method development data. Batch processing modifications to this approach from the Broad Institute and Jackson Laboratories are detailed below.

SDSI+AmpSeq Protocol

~~[0141]~~[0150] Full protocol details can be found here: benchling.com/s/prt-R95g0tCxKOeCAqn8lAk3 **(FIG 13)**. In short, cDNA synthesis is performed on 2.5µL of DNAse-treated viral RNA with SSIV following the manufacturer's protocol with an extension of the 50°C incubation from 10 minutes to 60 minutes. An additional cDNA normalization step can be performed (see above) or one can move directly into the ARTIC PCR by taking 5µL of cDNA and mixing this with 1uL of a 1fM SDSI (equal to 600 copies/µL). After mixing, split into 2 x 3µL aliquots and add ARTIC primer pool 1 or pool 2, as well as 1µM of the spike-in forward and reverse primers (40nM final concentration in the ARTIC pool). The ARTIC PCR conditions were 98°C for 30 seconds, followed by 40 cycles of 95°C for 15 seconds and 65°C for 5 minutes. Pool 1 and pool 2 PCR reactions were combined and taken through library construction with scaled down Illumina DNA Flex.

Broad Institute Sample Processing

~~[0142]~~[0151] The batch data from the Broad Institute was generated using SDSI+AmpSeq with minor modifications **(FIG 16)**. In short, SSIV was used for cDNA synthesis. Q5 2X MM was used for the ARTIC PCR which was run for 35 cycles. The SDSIs were spiked in at 6e3 copies/µL and the SDSI specific primers were added to each ARTIC pool at a final concentration of 40nM. Library construction was performed either with the scaled down Illumina DNA Flex (previously described) or COVID-seq (Illumina #20043675). Samples were sequenced on a NovaSeq 6000 SP Reagent Kit v1 (300 cycles) or v1.5 kits (300 cycles), or NextSeq 500 v2 kit (300 cycles).

~~[0143]~~[0152] The GC percent for each SDSIs and percent SDSI reads over total reads correlation for SDSI (2-48) was performed with the samples sequenced at the Broad Institute (N=2,903) **(FIG 19E)**. A linear regression was used to evaluate significance (p-value = 0.8160).

Jackson Laboratory Sample Processing

~~[0144]~~[0153] Data generated at Jackson Laboratory (JAX) used two different protocols publicly available here: github.com/tewhey-lab/SARS-CoV-2-Consensus **(FIG 16)**. All samples included 6e2 copies/µL of SDSIs and the SDSI specific primers were added to each ARTIC pool at a final concentration of 4nM. Samples processed from December 2020 to April 2021 used Lunascript (NEB #E3010) for cDNA synthesis and Q5 2X MM for the ARTIC PCR which was run for 35 cycles. These samples used scaled down Illumina DNA Flex for library construction. Samples

sequenced after April 2021 used the standard COVID-seq protocol. All samples were sequenced on a NextSeq500 using paired 75 bp reads by the Genome Technology group on Jackson Laboratory's Bar Harbor campus. The validation of all SDSIs in clinical samples (**FIG 15A**) was performed with this protocol and is presented as the percent of SDSI reads over the total of all reads for each sample.

~~[0145]~~[0154] Of note, the SDSIs (used at the lowest recommended concentration of 6e2 copies/uL) were reliably detected in the samples sequenced at JAX. This reliable detection however is also dependent on the sequencing depth used by the institution.

SDSI impact on genome recovery

~~[0146]~~[0155] For +/- SDSI experiments testing impact on recovery of viral genomes, fourteen clinical samples spanning a range of CTs (CT = 17.6-30) were selected **(FIG 15B, FIG 21A)**. Samples were CT normalized and split after cDNA synthesis into 2 x 5µL aliquots. Samples below CT 20 were normalized to CT 25 and samples between CT 20-25 were normalized to CT 26. Fourteen randomly selected SDSIs were used with each sample receiving either an SDSI (600 copies/µL) and the SDSI specific primers (40nM final concentration in the ARTIC pool) or just the ARTIC pool 1 and pool 2 mastermix with additional nuclease free water and no SDSI primers. Samples were processed according to the SDSI+AmpSeq protocol using scaled down Illumina DNA Flex for library construction, sequenced on a NextSeq 500 Mid Output Kit v2.5 (300 Cycles) and analyzed as described below.

~~[0147]~~[0156] Statistical analysis for the plus/minus SDSI experiment involved analysis of the mean coverage for all 98 amplicons for the full sample set with a two-tailed Mann Whitney t-test and multiple comparison two-stage step-up Benjamini, Krieger, and Yekutieli test with FDR set to 5%. All 98 amplicons were found to be not significantly different (p-value > 0.05) between the plus and minus SDSI group. Samples were also separated into three CT bins (CT <27 (n=4), 27-29 (n=6), CT>30 (n=4)) and this test repeated for each CT bin. This analysis also revealed that there was no significant difference (p-value > 0.05) in the mean coverage across any amplicon for any CT bin.

Intentional SDSI contamination experiment

~~[0148]~~[0157] The intentional contamination experiment used SDSI 87 and SDSI 94 (SDSI 87: SDSI 94). The SDSIs were mixed at five different proportions (100:0, 75:25, 50:50, 25:75, and

0:100) **(FIG 17A)**. Each condition was performed in duplicate. All validation experiment samples were processed according to the SDSI+AmpSeq protocol using scaled down Illumina DNA Flex for library construction. Samples were processed with the standard SDSI+AmpSeq protocol and sequenced on a NextSeq 500 Mid Output Kit v2.5 (300 Cycles).

Production and application of synthetic RNA spike-ins (SRSI)

~~[0149]~~**[0158]** Applicants ordered SDSI oligos with minor modifications to enable in-vitro transcription of RNAs (including a T7 promoter upstream of the SDSI amplicon, as well as 17 bps of constant sequence within the primer region) (Twist Bioscience) (sequences in attached **Sup Data File 1**). For two SDSIs (SDSI 1 and SDSI 4) applicants in-vitro transcribed RNA using a T7 transcription kit (NEB E2050), quantified by RNA screen tape (Agilent 5067-5579 and 5067-5580), then diluted in water to 10fM (6,000 copies/μL), 1fM (600 copies/μL), 100aM (60 copies/μL), and 10aM (6 copies/μL). Applicants added 1μL of SRSI at each concentration directly to 5 μL of RNA from two patient samples with high and intermediate viral loads, respectively, and prepared sequencing libraries using the SDSI+AmpSeq protocol (without the SDSI addition step at the cDNA stage). For the sample with a high viral load, applicants performed a dilution at the cDNA stage (diluting 32-fold for a mock Ct of 25 rather than 20). Reads mapping to unique SDSI sequences and SARS-CoV-2 were quantified using the align_and_count_multiple_report and assemble_refbased wdls respectively, and % SDSI/total reads was reported (**FIG 27**).

Computational analysis workflow

~~[0150]~~**[0159]** Applicants analyzed sequencing data on the Terra platform (app.terra.bio) using viral-ngs 2.1.28 with workflows that are publicly available on the Dockstore Tool Repository Service (dockstore.org/organizations/BroadInstitute/collections/pgs). Samples were demultiplexed using the demux_plus workflow with a spike in database file for the SDSIs. Applicants performed any separate analyses to quantify read counts, including those for SDSIs, with the align_and_count_multiple_report workflow with the relevant database. For most analyses involving direct comparisons between samples, applicants performed downsampling to the lowest number of reads passing filter with the downsample workflow. Applicants performed assembly using the assemble_refbased workflow to the following reference fasta: *www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta*. Applicants used iVar version 1.2.1 for primer trimming on all samples followed by assembly with minimap2 set to a minimum

coverage of either 3, 10, or 20, skipping deduplication procedures. The computational pipeline for all samples sequenced at JAX is publicly available at the following: github.com/tewhey-lab/SARS-CoV-2-Consensus.

~~[0151]~~[0160] Samples from the batch data were subset in the following way for analysis. All samples with a present SDSI were used for the percent of SDSI reads out of the sum of all SDSI reads analysis (JAX: N=3,838, Broad: N=2,903). Samples with known experimental contamination errors or where the dominant (>50%) SDSI was not the correct SDSI were removed. For the percent of SDSI reads over the total of all sequenced reads analysis (JAX: N=3,093, Broad: N=2,670), non-template controls (waters) and clinical samples with no detectable viral load (CT>40 or not detected via qPCR as described above) were removed from analysis.

Metagenomic sequencing and comparison

~~[0152]~~[0161] Metagenomic sequencing data and genome assemblies used for the comparison of amplicon-based sequencing were prepared, sequenced, analyzed as described previously,[21] and the data are publicly available at NCBI's GenBank and SRA databases under BioProject PRJNA622837. Applicants prepared amplicon sequencing libraries from the sample RNA extract following the SDSI+AmpSeq protocol (**FIG 13**). In order to increase sample throughput and bypass an additional more laborious quantification step post the ARTIC PCR, applicants normalized cDNA samples that had a high viral load (CT<27) to a CT of 27. To prepare for the ARTIC PCR, applicants transferred 5μL of the normalized cDNA to a new plate and added 1μL of a SDSI (600 copies/μL). After mixing, applicants transferred 3μL to a new plate, added ARTIC PCR pool #1 mastermix and pool #2 mastermix to the respective plates, and on a thermal cycler incubated at 98°C for 30s, followed by 40 cycles of 95°C for 15s and 65°C for 5min. Applicants then combined in equal molar amounts of amplified samples for a total of 50ng and processed through .5X Illumina Flex library construction pipeline. Applicants sequenced the concordance data set on a NovaSeq 6000 SP Reagent Kit v1 (300 cycles) and analyzed as detailed in the methods below. For SNV analysis, the coverage depth over each divergent SNV was greater than 1000X for both platforms, and both SNV calls persisted at relaxed (n=3) and conservative (n=20) minimum coverage thresholds. Primer trimming using iVar version 1.2.1 was manually confirmed.

Suspected nosocomial cluster investigation

~~[0153]~~[0162] Applicants received NP swab samples in UTM and extracted RNA from 200µL of biosample as previously described[21]. Applicants prepared amplicon sequencing libraries as described above and analyzed them as detailed in the methods below. A pairwise distance was calculated between all partial genomes (>80% complete), excluding gaps, to determine whether samples were likely to be the result of nosocomial transmission **(FIG 17C)**. Applicants calculated the proportion of reads that mapped to a given SDSI out of all reads that mapped to any SDSI. Data has been made available in both the Short Read Archive and NCBI GenBank under Bioproject PRJNA622837. GenBank accessions for SARS-CoV-2 genomes from this set of samples are MW454553 - MW454562.

~~[0154]~~[0163] For phylogenetic tree reconstruction applicants placed the suspected nosocomial cluster in a broader genomic context by performing a subsampling of the genome sequences available in GISAID (as of January 26 2021) **(FIG 26)**. Applicants used the sarscov2_nextstrain workflow to perform a Massachusetts-weighted subsampling of samples from 1 January 2020 - 1 November 2020. Applicants' subsampled dataset included 3146 sequences; 1449 samples from Massachusetts, 1425 samples from elsewhere in the United States and 283 from other countries. Applicants constructed a maximum likelihood tree using iqtree with a GTR substitution model and edited and interpreted the tree in Figtree v1.4.4.

Data presentation

~~[0155]~~[0164] Data analysis and graphing was performed using R Statistical Software (version 1.3.959; R Foundation for Statistical Computing, Vienna, Austria), GraphPad PRISM (version 9.0.2; GraphPad Software, La Jolla California USA, www.graphpad.com) and Python (version 3.7). Applicants created original figures using BioRender (BioRender.com).

Code availability

~~[0156]~~[0165] Viral genomes were processed using the Terra platform (app.terra.bio) using viral-ngs 2.1.1 with workflows that are publicly available on the Dockstore Tool Repository Service (dockstore.org/organizations/BroadInstitute/collections/pgs). Downstream analyses were performed using Geneious or standard R packages. Custom scripts used to generate figures are available upon request.

Methods

Data availability

~~[0157]~~**[0166]** Sequences and genome assembly data are publicly available on NCBI's Genbank and SRA databases under BioProject PRJNA622837. GenBank accessions for SARS-CoV-2 genomes newly reported in this study are MW454553 - MW454562.

References

1. Washington, N. L. et al. Genomic epidemiology identifies emergence and rapid transmission of SARS

CoV-2 B.1.1.7 in the United States. medRxiv (2021) doi:10.1101/2021.02.06.21251159.

2. Walensky, R. P., Walke, H. T. & Fauci, A. S. SARS-CoV-2 Variants of Concern in the United States—Challenges and Opportunities. JAMA vol. 325 1037 (2021).

3. Wang, P. et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature 593, 130–135 (2021).

4. Focosi, D., Tuccori, M., Baj, A. & Maggi, F. SARS-CoV-2 Variants: A Synopsis of In Vitro Efficacy Data of Convalescent Plasma, Currently Marketed Vaccines, and Monoclonal Antibodies. Viruses 13, (2021).

5. Wang, P. et al. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. Cell Host Microbe 29, 747–751.e4 (2021).

6. Naveca, F. et al. SARS-CoV-2 reinfection by the new Variant of Concern (VOC) P. 1 in Amazonas, Brazil. virological. org (2021).

7. Organization, W. H. & Others. Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health, 8 January 2021. (2021).

8. COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk. An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet Microbe 1, e99–e100 (2020).

9. Chiara, M. et al. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Brief. Bioinform. (2020) doi:10.1093/bib/bbaa297.

10. Charre, C. et al. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. Virus Evol 6, veaa075 (2020).

11. Rausch, J. W., Capoferri, A. A., Katusiime, M. G., Patro, S. C. & Kearney, M. F. Low genetic diversity may be an Achilles heel of SARS-CoV-2. Proceedings of the National Academy

of Sciences of the United States of America vol. 117 24614–24616 (2020).

12. Endo, A., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott, S., Kucharski, A. J. & Funk, S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Res 5, 67 (2020).

13. Lagerborg, K. A., Watrous, J. D., Cheng, S. & Jain, M. High-Throughput Measure of Bioactive Lipids Using Non-targeted Mass Spectrometry. Methods Mol. Biol. 1862, 17–35 (2019).

14. Boja, E. S. & Rodriguez, H. Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins. Proteomics 12, 1093–1110 (2012).

15. Chen, K. et al. The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. Molecular and Cellular Biology vol. 36 662–667 (2016).

16. Illumina COVIDSeq Test. emea.illumina.com/products/by-type/ivd-products/covidseq.html.

17. Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 21, 1543–1551 (2011).

18. Quail, M. A. et al. SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. BMC Genomics 15, 110 (2014).

19. Dilucca, M., Forcelloni, S., Pavlopoulou, A., Georgakilas, A. G. & Giansanti, A. Codon usage and evolutionary rates of the 2019-nCoV genes. Cold Spring Harbor Laboratory 2020.03.25.006569 (2020) doi:10.1101/2020.03.25.006569.

20. Potapov, V. & Ong, J. L. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLOS ONE vol. 12 e0169774 (2017).

21. Lemieux, J. E. et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science 371, (2021).

22. So, A. P. et al. A robust targeted sequencing approach for low input and variable quality DNA from clinical samples. NPJ Genom Med 3, 2 (2018).

23. Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 20, 8 (2019).

24. Pipelines R&D, D. N. A. et al. COVID-19 ARTIC v3 Illumina library construction and sequencing protocol v5. protocols.io (2020) doi:10.17504/protocols.io.bibtkann.

25. Lam, C. et al. Sars-CoV-2 Genome Sequencing Methods Differ In Their Ability To Detect

Variants From Low Viral Load Samples. J. Clin. Microbiol. JCM0104621 (2021).

26. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228–232 (2016).

27. Metsky, H. C. et al. Zika virus evolution and spread in the Americas. Nature 546, 411–415 (2017).

28. Gohl, D. M. et al. A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. BMC Genomics 21, 863 (2020).

29. Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R. & Kuroda, M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. PLoS One 15, e0239403 (2020).

30. Tyson, J. R. et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv (2020) doi:10.1101/2020.09.04.283077.

31. VarSkip: VarSkip multiplex PCR designs for SARS-CoV-2 sequencing. (Github).

32. primer_schemes/nCoV-2019 at master · artic-network/artic-ncov2019. github.com/artic-network/artic-ncov2019.

33. Wong, F., & Collins, J. J. (2020). Evidence that coronavirus superspreading is fat-tailed. In *Proceedings of the National Academy of Sciences* (Vol. 117, Issue 47, pp. 29416–29418). doi.org/10.1073/pnas.2018490117.

34. Adam, D. C., Wu, P., Wong, J. Y., Lau, E. H. Y., Tsang, T. K., Cauchemez, S., Leung, G. M., & Cowling, B. J. (2020). Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. In *Nature Medicine* (Vol. 26, Issue 11, pp. 1714–1719). doi.org/10.1038/s41591-020-1092-0.

35. Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18.

36. Antonov, J., Goldstein, D. R., Oberli, A., Baltzer, A., Pirotta, M., Fleischmann, A., Altermatt, H. J., & Jaggi, R. (2005). Reliable gene expression measurements from degraded RNA by quantitative real-time PCR depend on short amplicons and a proper normalization. *Laboratory Investigation; a Journal of Technical Methods and Pathology*, *85*(8), 1040–1050.

37. *Artic Network*. (n.d.). Retrieved February 2, 2021, from artic.network/

38. Baker, D. J., Kay, G. L., Aydin, A., Le-Viet, T., Rudder, S., Tedim, A. P., Kolyva, A., Diaz, M., de Oliveira Martins, L., Alikhan, N.-F., Meadows, L., Bell, A., Gutierrez, A. V., Trotter, A. J., Thomson, N. M., Gilroy, R., Griffith, L., Adriaenssens, E. M., Stanley, R., … O'Grady, J. (2020). CoronaHiT: large scale multiplexing of SARS-CoV-2 genomes using Nanopore sequencing. In *Cold Spring Harbor Laboratory* (p. 2020.06.24.162156). doi.org/10.1101/2020.06.24.162156.

39. Dearlove, B., Lewitus, E., Bai, H., Li, Y., Reeves, D. B., Gordon Joyce, M., Scott, P. T., Amare, M. F., Vasan, S., Michael, N. L., Modjarrad, K., & Rolland, M. (2020). A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. In *Proceedings of the National Academy of Sciences* (Vol. 117, Issue 38, pp. 23652–23662). doi.org/10.1073/pnas.2008281117.

40. Houldcroft, C. J., Beale, M. A., & Breuer, J. (2017). Clinical and biological insights from viral genome sequencing. *Nature Reviews. Microbiology*, *15*(3), 183–192.

41. Klempt, P., Brož, P., Kašný, M., Novotný, A., Kvapilová, K., & Kvapil, P. (2020). Performance of Targeted Library Preparation Solutions for SARS-CoV-2 Whole Genome Analysis. *Diagnostics (Basel, Switzerland)*, *10*(10). doi.org/10.3390/diagnostics10100769.

42. Mathieu-Daudé, F., Welsh, J., Vogt, T., & McClelland, M. (1996). DNA Rehybridization During PCR: The "C o t Effect" and Its Consequences. *Nucleic Acids Research*, *24*(11), 2080–2086.

43. Metsky, H. C., Siddle, K. J., Gladden-Young, A., Qu, J., Yang, D. K., Brehio, P., Goldfarb, A., Piantadosi, A., Wohl, S., Carter, A., Lin, A. E., Barnes, K. G., Tully, D. C., Corleis, B., Hennigan, S., Barbosa-Lima, G., Vieira, Y. R., Paul, L. M., Tan, A. L., … Matranga, C. B. (2019). Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nature Biotechnology*, *37*(2), 160–168.

44. No, J. S., Kim, W.-K., Cho, S., Lee, S.-H., Kim, J.-A., Lee, D., Song, D. H., Gu, S. H., Jeong, S. T., Wiley, M. R., Palacios, G., & Song, J.-W. (2019). Comparison of targeted next-generation sequencing for whole-genome sequencing of Hantaan orthohantavirus in Apodemus agrarius lung tissues. *Scientific Reports*, *9*(1), 16631.

45. Popa, A., Genger, J.-W., Nicholson, M. D., Penz, T., Schmid, D., Aberle, S. W., Agerer, B., Lercher, A., Endler, L., Colaço, H., Smyth, M., Schuster, M., Grau, M. L., Martínez-Jiménez,

F., Pich, O., Borena, W., Pawelka, E., Keszei, Z., Senekowitsch, M., … Bergthaler, A. (2020). Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Science Translational Medicine*, *12*(573). doi.org/10.1126/scitranslmed.abe2555.

46. Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T. F., Beutler, N. A., Burton, D. R., Lewis-Ximenez, L. L., de Jesus, J. G., Giovanetti, M., Hill, S. C., Black, A., Bedford, T., Carroll, M. W., Nunes, M., … Loman, N. J. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols*, *12*(6), 1261–1276.

47. *SARS-CoV-2 COVID-19 Coronavirus research and surveillance*. (n.d.). Retrieved February 12, 2021, from www.paragongenomics.com/product/cleanplex-sars-cov-2-panel/

Sethuraman, N., Jeremiah, S. S., & Ryo, A. (2020). Interpreting Diagnostic Tests for SARS-CoV-2. *JAMA: The Journal of the American Medical Association*, *323*(22), 2249–2251.

48. Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C. V., Boyd, O., Loman, N. J., McCrone, J. T., Gonçalves, S., … The COVID-19 Genomics UK (COG-UK) consortium. (2021). Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. In *bioRxiv*. medRxiv. doi.org/10.1101/2020.12.30.20249034.

Tables

Table 2. Non-limiting set of designed spike-ins.

Oligo ID	Sequence to order (5' to 3')	Nonarchaeal genuses with significant homology*
SDSI forward primer	TCTCCTTCTTAGCTTCGTGAGAAC (SEQ ID NO: 391)	n/a
SDSI reverse primer	CTTGGTCGTCTACTACATGATGTG (SEQ ID NO: 392)	n/a

SDSI 1	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACCGGACGTTGTGATCACGGG TACCTTGATCTGGTACTCAAAGGTTTGCCCCCGTGAAGTCTGGTACATGGCT AGACACGTCACTCCATTCGAGGGACATTCGAAGTTAGAGAAGGGCAGAGC GATACATCAGATATATCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 393)	none
SDSI 2	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTAATGGAAAGTATGCTTTAGA TACCTTCTGGAACGCTATCTCACTTGGCGGGAATTCAGATATGGAGAGTAA ATTAAGGGATCTGGAAGTAAAGTTAATGTCGTTAATCTATTTAAATGAGTC ACCATTAAAATCACCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 394)	none
SDSI 3	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATAATATGTTAGAGGTAGAATT TCTTTGTGATAGAATATTATTGATGAATGATGGAAGAGAATTAGCATTAGG AAAACCTAAGGAACTGGTAAAGGATACAGAATCTAAGAATCTTGAAGAGG TTTTCCTTAAACTTGTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 395)	none
SDSI 4	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTCTAGGTTTTAATTCTTCAAC TGCTTCAAATACTAGCTTACTGTAGTTATCTGCCCTCATGTTAGGATATATA TCTGGAATATAAGGAGGTTGATGAGTTATAAGAAGTGGATGAAATTGTTGT CACACACTCCCCTACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 396)	none
SDSI 5	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGTAAGCGTTTCCTACCCTCG AGAGGGCCATCCTGGTGGTGAGGAAGTCGTCGAAGTGGGCTAAGTAAAAA GCGAAGATCTCGACCCACAATTACCTCCTCCTGTACACCAGGAATACCCCT ATCAGGATAGAGATACCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 397)	none
SDSI 6	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCACGGTCCGCGACGTGAATCG GGCGTTCCAGTCGGCGTTCGGCTACGACGCCGACGACGTGGTCGGAAGCG ACCTCCTCGGGCGAATCGTGCCCCCGGTGCCGGACCCGGACCCGGTGCCGG AACCGGGGGACGACGAGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 398)	none
SDSI 7	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGTCCGCGAGTTCATCCTGAAC GTCGTCCCGCTGTCGCCCGGCGAGGAGCGCGGGGCGGGCTACGCCATCTAC ACCGACATCACGGAGCGGAAGACCCGCGAAAGCGAGCTAGAGCGACAGA ACGAGCGATTGGAGGAGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 399)	none
SDSI 8	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACGAACTCGTCGGTGAACATCTC GTCTTCCGGGGAGCCCGCCGCTCATGGCCTGCCCCCGCCGTAAGCTGCTGC ATAAACCCGCTCCAAAATATACGGATCATTCACCCCTTGGAATCGCTCAAT CAGATCAATGTACACCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 400)	none
SDSI 9	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGCGTACATTCCCCCTAAGCGGC TCCCAATATACAGACGCCGGTTAACGACAGCTGGCGACCCTGTGATCTCAG TACCGGTGTCGAATGACCACATCAGCTTGCCTGTCCGTGCATGGAGTTCGT ATACGTACCCGTCGTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 401)	none
SDSI 10	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATACACCACCCCATCAGCAACA ACTGAATCATGATTAAGTATCGCACCAGCATCGTAGCGCCAGCGTTCACTG CCAGTGGTGCTATCGAATGCATAGAAGATATGCTCCTAATCGCCAATATCA GTACTTCACAAAGCCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 402)	none

SDSI 11	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGAGTCTTTTGTCACACCGCA GAGGCGTAGCGCTGCAGAGCAGGAGCCCAAGCCTACTGCCAACATAGAGA ACATAGTGGCTACAGTATCCCTCGACCAGACTCTAGACCTGAACCTCATAG AGAGGAGCATACTGACCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 403)	none
SDSI 12	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTCGCCTGGGTTAAGAGGATG TTCGGCCTCTCCAAGGCGGGTCACGGAGGCACGCTGGACCCGAAGGTCAC CGGCGTCCTCCCCGTAGCCCTGGAGGAAGCAACCAAGGTCATAGGCCTGGT GGTGCACACGAGCAAGGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 404)	none
SDSI 13	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTGGGCGAGATCTACCAGAGG CCGCCGCTCCGCAGCAGTGTTAAGAGAAGCCTCCGCGTCAAGAGGATATA CGAGATAGAGCTGCTGGAGTACAACGGCAGGTACGCGCTCATGAGGGTGC TCTGCGAGGCCGGCACATCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 405)	none
SDSI 14	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCTGGAAGAACGAGGGCAAGG AGGACCTGCTGCGGAGCTACATCAAGCCCGTCGAGTACGCCGTGAGCCAC CTGCCCAAGATAGTTATACGCGATACCGCGGTGGACGCCATAGCCCATGGC GCGAACCTCGCGGTGCCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 406)	none
SDSI 15	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGGAGACCCCAAGGTGACCGGC GTCCTACCAGTGGGGCTCGCCAACAGCACCAAGGTCATTGGTAATGTTATA CATAGTGTTAAAGAATACGTGATGGTTATACAGCTCCACGGCGATGTAGCC GAGCAGGATTTAAGAACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 407)	none
SDSI 16	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAGAGGGAAAGACTGTAGCTTT CATTCCTAGGCACGGAAAGAGACACAGAATACCTCCACATAAGATAAATT ATAGAGCTAATATATGGGCATTAAAAGAACTAGGAGTGAAATGGGTCATC TCAGTTTCTGCCGTAGGACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 408)	none
SDSI 17	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGAGGGAGCTCAGGAGGACTCG CACGGGGCCCTACAGGGAGGATGAGACACTTGTAAGGCTCCAGGACGTCA GCGAGGCCCTGCTCCTGTGGAGGAGCAACGGGGATGAGAGGTATCTTAGA CGCATCGTGCTACCCGTTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 409)	none
SDSI 18	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAACATCTATCGCCCACCTCCC GAAGATAATGATCTTGGATACAGCTGTCGACGCCATAGCACATGGTGCCAA CCTGGCTGCCCCAGGCGTCGCCAGGTTAACCAGGAACATCGCGAAGGGTA GTACCGTAGCGATCCTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 410)	none
SDSI 19	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGCTATCCCCGTGTACAGCATG GTGGGGGTGCCGATGCCCGGGTAGAACTTGGTGACGCTCTCCAGCTTCTCG AGGACGGTTTCCTTGGGGAGGCTCGCGGTGTCCACGAGGGTTATCGCGTCC TCGGCGCCGTCGCCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 411)	none
SDSI 20	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAGGACGCGAAGAGCGCGGTG GATGTGGACGCGCCGCCGCACACGTAGCCGTCGAGGTAGCGCGGAACCAT CGGCGACATCAGCCCCACGACGCGACCCGAGGCGTTGCCGAGGATCACGT CGAGCGTCACGCGCGGCACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 412)	none

SDSI 21	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTATGGTGTAGAACGGGTCGTT GCGGAGCCAGCCTGGCGGCACGTACCGGTCGTCCGCTATCGCCAGCGATCT CTCGAAGAGGTCGAGGTAGGCGGACGCGTTGGCGAACGCCCCGTGTATCA CGACGTCTATCCCGCCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 413)	none
SDSI 22	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTACGCCGGGTGCGTAGGAGG GCTCGAGTACATCCATGTCTATACTGATGTATGTTTTACCCAGGTCGCCTAG TGCCAGGGGTCCCTTTAACGCTTCCAGGATAGAGTACACGGTGACGTCTCT AGTCTTCTTCAAGAACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 414)	none
SDSI 23	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTACTAGCGTGTCAACGGAGCTC TTCAACGCCTTTACTATTGGATAGGTTATAAGGTGCTCGCCTCCGAGGAAT CCCAGGAGCATGCCGGGATACTCGTCTACAACGCCTTTCACCACGTCACCT ATGATTCTTAAAGAGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 415)	none
SDSI 24	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATAGGTGACATGGGGTTTCCCA TTGACTCTATAAAGCCGTATCCTTTAAGCGGAGTGCAATTGGTCTACGCTTT GCTTAACAACAGGTATTTCCTACCGGGTAGAGAGGGCTCGCTCATAGCTTT AGGTAGCGTGACGGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 416)	none
SDSI 25	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGTATCTCACCGCTTGTCACCAT AGTATCCCTCAGGTACTCCAGTATTCTTGAGAGAAACGCACCTAAGCCGGA TCTCAGGTTTGAATCCATAAGAACTATGAGTGAAGCGGGATTGAAGCCCCT GCTGTTTCTAAGACCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 417)	none
SDSI 26	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAAGGGAGATAGAGAAACGCAT CAAAATACCCTTGGGGAAACTGCGTGCAGGGGTTCAATATGGAGTAGAGG TCTCAGACATAAAGGAGAAGATAGCTGCTTACGCTAGGAGGAAGGGGCTT AAATACTTCCCATCGGCACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 418)	none
SDSI 27	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTGAACCTCGTGCCCGGCTCTA AGTCGTGAGGGCTTGCAACATAGGTGGGGAGGAACCCGAGCAACGGGTAA GAAGACAGGATAAGCGGTATCGCTATGAAGAGGGCTGAGAAAAGGACATA TACTCCTGAGCCCGTCCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 419)	none
SDSI 28	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAACATGCCTTCCCCGTCTATA TAGACCCAGTAGAGTTTAAAAACTTAACCAGAGACGGCTTGTGAGCCGGAT CTCTCCCCCGCTAGGCCCTGGATTGGGCTCGCTCCTCCTGGGACCCCGGCCT CCACATGCTCGGGACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 420)	none
SDSI 29	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTCGGTTCGGCAATAAGTAATA CCAACGAGGTATTACCATGCGCGTGACCAGCAAAGGCCAAGTGACGATCC CAAAGGAGATACGGGATCATTTGGGGATTGGGCCGGGCTCCGAGGTGGAG TTCGTGCCCACAGACGACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 421)	mesorhizobium; neorhizobium
SDSI 30	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGATCATATGGCCGGCACGTT GGACTTGGGAGGCATGACAACGGACGAGTATATGGAGTGGCTGAGGGGTC CACGTGAAGATCTCGACATTGATTGACACAAATGTCCTGATCGATGTTTGG GGTCCTGCCGGACAGGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 422)	mesorhizobium; neorhizobium; rhizobium; neorhizobium; aminobacter; sinorhizobium; shinella;

SDSI 31	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGGTGTATTTTACACACCTGGACAGCCAGCATATGATGCTAGCACTCGGTGTCCCCTTATCACGGTTTCCCGCATTGTAAAGTTTTCGCGCCTGCTGCGCCCCGTAGGGCCTGGATTCATGTCTCAGAATCCATCTCCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 423)	'uncultured bacteria'
SDSI 32	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTAGCCCGCACCTTCCTCTGGTTTAGCACCAGCGGTCCCCACAGAGTACCCATCATCCCGAAGGATATGCTGGCAACAGTGGGCACGGGTCTCGCTCGTTGCCTGACTTAACAGGATGCTTCACAGTACGAACTGACGACACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 424)	'uncultured bacteria'
SDSI 33	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAACTTACCTTATCAGTGTCATTAAGCATATTGCTTCCAAGACCCATTGAAGCACTTACATCGTTGATACACAGGTGCCAGGAATAGTATTCCTCAGTCTCACTATAATCCTCGTTGGTGTAGCCTTCAAGAGAGTCAACACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 425)	none
SDSI 34	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTTAAGCAATTCTTCGGATGAAAGATGGCGCTCTATAGGAATTTGTTCTGGTCTAGCCATAAGGCATTATTTGTACTTAATTAGTAATAAATGTTTAGTTAATGACTATAAATCTGCAATTGGAGTCTCAAATTTTCAACACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 426)	none
SDSI 35	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAACATGAAGGATGTGTGTAAGAGGAAACGTTATTAACAGACGTAATCAGGAGGATAGTTATGCCCTAAAAACAGCAGAGTTAAGGTTTAAAAATAAGATAAGAACTCAGTTGAGGTTTATCCATTAATCCCATTAATCCTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 427)	none
SDSI 36	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATCCGCTGATATATCCTGGGGATATAGATCGCTCTGAAATGGTTACATCTATCGGTTTTAAGGACAGTTCCAACACTATTGGACCTTGCAGCTATGACAGGAATAATCTGTTTATCGAGCACAGTTGAATTTGACCTACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 428)	none
SDSI 37	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATATTCCGTATTTCTTATCAAACCGATCGTGAAGATTTGACAAAGGCTTAACTTTAGGGCTCCACTTCTCATTATTAGCCTTAGAATATAAAGCGTAACCGTAAGCCTGAGGAACGTAAAGCTTAGGAGATTCAATCCCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 429)	none
SDSI 38	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAAAATTAGCCGAAGGCTTCCCATTACCGAAAAAGTCGTTTATTAGCTCTTCATCCTTCTTCTCCACGTCCGCCCATTCCTCTCCTTCCCTTGGAATTTTAAGCTCGTCCCAGCTGACTCTTATGGGCAATTCAATATCCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 430)	none
SDSI 39	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCGGAGGAATCTATCATATTAAACCTCCTCAAAATCGCCTCCTCTTGATTGCTTAAAGGCTGTGAATTACAAAGCTTATTTAATGCGTCCCAAAGCGTTAAGTAATAATTATTTATATTAAACACTACTATTTCAGTAGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 431)	none
SDSI 40	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCCTCCTCAATTCAATTGGACTGAAGGAGGGTACGTTCTGGAAAACAGAGCGTAAAAGAGATATAGAACGTAGTATACACATAGCTGGAAAAAGAACAATCATTAAGACAATAAAGAACTTTATGGAAAAGAGTAGAACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 432)	none

SDSI 41	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGTGTAAAGGTTGTATAATTCA AGCCTCAGAACATTTCGAACTCCTTACAAAATCGTTTAAACTTTCTAAGGC ATAAATTTACTAGAAATTGTCATTTATGAGAATGTAACTATATAGATGGTA AAATTATTAATCCTCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 433)	none
SDSI 42	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCTGAAAAATAGGTTCGATCC GCCTCCTCACTTCTTCTCCTTCTTGCCCTCGGCCTCGGAGGAGGCCTCTATT CCCAGCTTCTTGGCCTCCTCCTCGGTCGTCATGAACAGGCTAGTCCTCTGCC TTCCGCCCATGCTCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 434)	none
SDSI 43	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCAGCATAAAAGACGGTTTC ACGGGCCAAAGCCTAAGCGGCGTAACGGTGAAAGAAGGAGATACGGTTTT GGGCACGATTGACGACGGCGGGACGCTGGAGCTCACGAGGGGCACTCACA CCTTGACTTTCGAGAAGCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 435)	none
SDSI 44	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTGATGTTATAGAAGTCCGCAA GGACGGCTCTGTCATCTCGCCCGAGGGTGGGAAATACTATCTCGGCGACAT AAGCGGCCCGACACAAATTAGCATCAAGTTCAAGGCCGGCGCGGTGGGAA CCCACGGCTTCACTATCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 436)	none
SDSI 45	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTCCCTCAACCTTCGCGGGGAG AACGGCGCGGAGTACTGGACGGGCTACGCGGACGCGCTGGAAGACCTGTT GAAGAAAATCCAGAGGCGGGAGGTGAGGGCATGAGAAGGTATTGTTACAT CACGTGGGGATGGATCACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 437)	none
SDSI 46	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGCGCCGGGAGGTGAGGGCAT GAGTGAGGAATTGATGTTTGGTCGTGTCGTGGAGTATGTTCAGCATAGTTT CTACAAGAAACCGTTTCCTCTTGGCAGTGAGCTCAAGAATGCAGTAGAGAA GGTTATGGAAACAGGACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 438)	none
SDSI 47	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGTCAGAGCCCACGTGGCAAC TTTTGAGGTTCTGACAAAAGACTATGTTCGTGAGAAATACAAAGACATCAT AGAGTTCATGAGGGAGAAAGGGACAGTATCGAGAAAGGAACTGCGGAAG AAGTTCTTCTTGCTTGCTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 439)	none
SDSI 48	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTACCTCAAAATACAGAATCAT ATTTTACAATCGCTTGGAAATATTAATATCAACAATACGCAAGTCCAAATT AACGTCCCTGGCAAACAGGTGACAATTTATACCCACGAAATACTAGATAAC GCCAAAAAGGCACTCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 440)	none

Table 3. SDSI and Viral Read Percentages

Mock CT	Viral reads %	Spike-in reads %
20	99.56	0.18
25	99.19	0.38
30	98.47	1.11
35	99.65	3.17

Table 4. Exemplary ARTIC v3 Primers and Primers Spiked in at 2X

Name	Pool	Sequence	Length	%GC	Spiked in at 2X
nCoV-2019_1_LEFT	nCoV-2019_1	ACCAACCAACTTTCGATCTCTTGT (SEQ ID NO: 441)	24	41.7	
nCoV-2019_1_RIGHT	nCoV-2019_1	CATCTTTAAGATGTTGACGTGCCTC (SEQ ID NO: 442)	25	44.0	
nCoV-2019_2_LEFT	nCoV-2019_2	CTGTTTTACAGGTTCGCGACGT (SEQ ID NO: 443)	22	50.0	
nCoV-2019_2_RIGHT	nCoV-2019_2	TAAGGATCAGTGCCAAGCTCGT (SEQ ID NO: 444)	22	50.0	
nCoV-2019_3_LEFT	nCoV-2019_1	CGGTAATAAAGGAGCTGGTGGC (SEQ ID NO: 445)	22	54.6	
nCoV-2019_3_RIGHT	nCoV-2019_1	AAGGTGTCTGCAATTCATAGCTCT (SEQ ID NO: 446)	24	41.7	
nCoV-2019_4_LEFT	nCoV-2019_2	GGTGTATACTGCTGCCGTGAAC (SEQ ID NO: 447)	22	54.6	
nCoV-2019_4_RIGHT	nCoV-2019_2	CACAAGTAGTGGCACCTTCTTTAGT (SEQ ID NO: 448)	25	44.0	
nCoV-2019_5_LEFT	nCoV-2019_1	TGGTGAAACTTCATGGCAGACG (SEQ ID NO: 449)	22	50.0	
nCoV-2019_5_RIGHT	nCoV-2019_1	ATTGATGTTGACTTTCTCTTTTTGGAGT (SEQ ID NO: 450)	28	32.1	
nCoV-2019_6_LEFT	nCoV-2019_2	GGTGTTGTTGGAGAAGGTTCCG (SEQ ID NO: 451)	22	54.6	
nCoV-2019_6_RIGHT	nCoV-2019_2	TAGCGGCCTTCTGTAAAACACG (SEQ ID NO: 452)	22	50.0	
nCoV-2019_7_LEFT_alt0	nCoV-2019_1	CATTTGCATCAGAGGCTGCTCG (SEQ ID NO: 453)	22	54.6	X
nCoV-2019_7_RIGHT_alt5	nCoV-2019_1	AGGTGACAATTTGTCCACCGAC (SEQ ID NO: 454)	22	50.0	X
nCoV-2019_8_LEFT	nCoV-2019_2	AGAGTTTCTTAGAGACGGTTGGGA (SEQ ID NO: 455)	24	45.8	
nCoV-2019_8_RIGHT	nCoV-2019_2	GCTTCAACAGCTTCACTAGTAGGT (SEQ ID NO: 456)	24	45.8	
nCoV-2019_9_LEFT_alt4	nCoV-2019_1	TTCCCACAGAAGTGTTAACAGAGG (SEQ ID NO: 457)	24	45.8	X
nCoV-2019_9_RIGHT_alt2	nCoV-2019_1	GACAGCATCTGCCACAACACAG (SEQ ID NO: 458)	22	54.6	X
nCoV-2019_10_LEFT	nCoV-2019_2	TGAGAAGTGCTCTGCCTATACAGT (SEQ ID NO: 459)	24	45.8	
nCoV-2019_10_RIGHT	nCoV-2019_2	TCATCTAACCAATCTTCTTCTTGCTCT (SEQ ID NO: 460	27	37.0	

nCoV-2019_11_LEFT	nCoV-2019_1	GGAATTTGGTGCCACTTCTGCT (SEQ ID NO: 461)	22	50.0	
nCoV-2019_11_RIGHT	nCoV-2019_1	TCATCAGATTCAACTTGCATGGCA (SEQ ID NO: 462)	24	41.7	
nCoV-2019_12_LEFT	nCoV-2019_2	AAACATGGAGGAGGTGTTGCAG (SEQ ID NO: 463)	22	50.0	X
nCoV-2019_12_RIGHT	nCoV-2019_2	TTCACTCTTCATTTCCAAAAAGCTTGA (SEQ ID NO: 464)	27	33.3	X
nCoV-2019_13_LEFT	nCoV-2019_1	TCGCACAAATGTCTACTTAGCTGT (SEQ ID NO: 465)	24	41.7	
nCoV-2019_13_RIGHT	nCoV-2019_1	ACCACAGCAGTTAAAACACCCT (SEQ ID NO: 466)	22	45.5	
nCoV-2019_14_LEFT_alt4	nCoV-2019_2	TGGCAATCTTCATCCAGATTCTGC (SEQ ID NO: 467)	24	45.8	X
nCoV-2019_14_RIGHT_alt2	nCoV-2019_2	TGCGTGTTTCTTCTGCATGTGC (SEQ ID NO: 468)	22	50.0	X
nCoV-2019_15_LEFT_alt1	nCoV-2019_1	AGTGCTTAAAAAGTGTAAAAGTGCCT (SEQ ID NO: 469)	26	34.6	X
nCoV-2019_15_RIGHT_alt3	nCoV-2019_1	ACTGTAGCTGGCACTTTGAGAGA (SEQ ID NO: 470)	23	47.8	X
nCoV-2019_16_LEFT	nCoV-2019_2	AATTTGGAAGAAGCTGCTCGGT (SEQ ID NO: 471)	22	45.5	
nCoV-2019_16_RIGHT	nCoV-2019_2	CACAACTTGCGTGTGGAGGTTA (SEQ ID NO: 472)	22	50.0	
nCoV-2019_17_LEFT	nCoV-2019_1	CTTCTTTCTTTGAGAGAAGTGAGGACT (SEQ ID NO: 473)	27	40.7	X
nCoV-2019_17_RIGHT	nCoV-2019_1	TTTGTTGGAGTGTTAACAATGCAGT (SEQ ID NO: 474)	25	36.0	X
nCoV-2019_18_LEFT_alt2	nCoV-2019_2	ACTTCTATTAAATGGGCAGATAACAACT GT (SEQ ID NO: 475)	30	33.3	X
nCoV-2019_18_RIGHT_alt1	nCoV-2019_2	GCTTGTTTACCACACGTACAAGG (SEQ ID NO: 476)	23	47.8	X
nCoV-2019_19_LEFT	nCoV-2019_1	GCTGTTATGTACATGGGCACACT (SEQ ID NO: 477)	23	47.8	
nCoV-2019_19_RIGHT	nCoV-2019_1	TGTCCAACTTAGGGTCAATTTCTGT (SEQ ID NO: 478)	25	40.0	
nCoV-2019_20_LEFT	nCoV-2019_2	ACAAAGAAAACAGTTACACAACAACCA (SEQ ID NO: 479)	27	33.3	
nCoV-2019_20_RIGHT	nCoV-2019_2	ACGTGGCTTTATTAGTTGCATTGTT (SEQ ID NO: 480)	25	36.0	
nCoV-2019_21_LEFT_alt2	nCoV-2019_1	GGCTATTGATTATAAACACTACACACCC T (SEQ ID NO: 481	29	37.9	X
nCoV-2019_21_RIGHT_alt0	nCoV-2019_1	GATCTGTGTGGCCAACCTCTTC (SEQ ID NO: 482)	22	54.6	X

nCoV-2019_22_LEFT	nCoV-2019_2	ACTACCGAAGTTGTAGGAGACATTATAC T (SEQ ID NO: 483)	29	37.9	
nCoV-2019_22_RIGHT	nCoV-2019_2	ACAGTATTCTTTGCTATAGTAGTCGGC (SEQ ID NO: 484)	27	40.7	
nCoV-2019_23_LEFT	nCoV-2019_1	ACAACTACTAACATAGTTACACGGTGT (SEQ ID NO: 485)	27	37.0	
nCoV-2019_23_RIGHT	nCoV-2019_1	ACCAGTACAGTAGGTTGCAATAGTG (SEQ ID NO: 486)	25	44.0	
nCoV-2019_24_LEFT	nCoV-2019_2	AGGCATGCCTTCTTACTGTACTG (SEQ ID NO: 487)	23	47.8	X
nCoV-2019_24_RIGHT	nCoV-2019_2	ACATTCTAACCATAGCTGAAATCGGG (SEQ ID NO: 488)	26	42.3	X
nCoV-2019_25_LEFT	nCoV-2019_1	GCAATTGTTTTTCAGCTATTTTGCAGT (SEQ ID NO: 489)	27	33.3	
nCoV-2019_25_RIGHT	nCoV-2019_1	ACTGTAGTGACAAGTCTCTCGCA (SEQ ID NO: 490)	23	47.8	
nCoV-2019_26_LEFT	nCoV-2019_2	TTGTGATACATTCTGTGCTGGTAGT (SEQ ID NO: 491)	25	40.0	
nCoV-2019_26_RIGHT	nCoV-2019_2	TCCGCACTATCACCAACATCAG (SEQ ID NO: 492)	22	50.0	
nCoV-2019_27_LEFT	nCoV-2019_1	ACTACAGTCAGCTTATGTGTCAACC (SEQ ID NO: 493)	25	44.0	
nCoV-2019_27_RIGHT	nCoV-2019_1	AATACAAGCACCAAGGTCACGG (SEQ ID NO: 494)	22	50.0	
nCoV-2019_28_LEFT	nCoV-2019_2	ACATAGAAGTTACTGGCGATAGTTGT (SEQ ID NO: 495)	26	38.5	
nCoV-2019_28_RIGHT	nCoV-2019_2	TGTTTAGACATGACATGAACAGGTGT (SEQ ID NO: 496)	26	38.5	
nCoV-2019_29_LEFT	nCoV-2019_1	ACTTGTGTTCCTTTTTGTTGCTGC (SEQ ID NO: 497)	24	41.7	
nCoV-2019_29_RIGHT	nCoV-2019_1	AGTGTACTCTATAAGTTTTGATGGTGTGT (SEQ ID NO: 498)	29	34.5	
nCoV-2019_30_LEFT	nCoV-2019_2	GCACAACTAATGGTGACTTTTTGCA (SEQ ID NO: 499)	25	40.0	
nCoV-2019_30_RIGHT	nCoV-2019_2	ACCACTAGTAGATACACAAACACCAG (SEQ ID NO: 500)	26	42.3	
nCoV-2019_31_LEFT	nCoV-2019_1	TTCTGAGTACTGTAGGCACGGC (SEQ ID NO: 501)	22	54.6	
nCoV-2019_31_RIGHT	nCoV-2019_1	ACAGAATAAACACCAGGTAAGAATGAG T (SEQ ID NO: 502)	28	35.7	
nCoV-2019_32_LEFT	nCoV-2019_2	TGGTGAATACAGTCATGTAGTTGCC (SEQ ID NO: 503)	25	44.0	
nCoV-2019_32_RIGHT	nCoV-2019_2	AGCACATCACTACGCAACTTTAGA (SEQ ID NO: 504)	24	41.7	

nCoV-2019_33_LEFT	nCoV-2019_1	ACTTTTGAAGAAGCTGCGCTGT (SEQ ID NO: 505)	22	45.5	X
nCoV-2019_33_RIGHT	nCoV-2019_1	TGGACAGTAAACTACGTCATCAAGC (SEQ ID NO: 506)	25	44.0	X
nCoV-2019_34_LEFT	nCoV-2019_2	TCCCATCTGGTAAAGTTGAGGGT (SEQ ID NO: 507)	23	47.8	
nCoV-2019_34_RIGHT	nCoV-2019_2	AGTGAAATTGGGCCTCATAGCA (SEQ ID NO: 508)	22	45.5	
nCoV-2019_35_LEFT	nCoV-2019_1	TGTTCGCATTCAACCAGGACAG (SEQ ID NO: 509)	22	50.0	
nCoV-2019_35_RIGHT	nCoV-2019_1	ACTTCATAGCCACAAGGTTAAAGTCA (SEQ ID NO: 510)	26	38.5	
nCoV-2019_36_LEFT	nCoV-2019_2	TTAGCTTGGTTGTACGCTGCTG (SEQ ID NO: 511)	22	50.0	
nCoV-2019_36_RIGHT	nCoV-2019_2	GAACAAAGACCATTGAGTACTCTGGA (SEQ ID NO: 512)	26	42.3	
nCoV-2019_37_LEFT	nCoV-2019_1	ACACACCACTGGTTGTTACTCAC (SEQ ID NO: 513)	23	47.8	
nCoV-2019_37_RIGHT	nCoV-2019_1	GTCCACACTCTCCTAGCACCAT (SEQ ID NO: 514)	22	54.6	
nCoV-2019_38_LEFT	nCoV-2019_2	ACTGTGTTATGTATGCATCAGCTGT (SEQ ID NO: 515)	25	40.0	
nCoV-2019_38_RIGHT	nCoV-2019_2	CACCAAGAGTCAGTCTAAAGTAGCG (SEQ ID NO: 516)	25	48.0	
nCoV-2019_39_LEFT	nCoV-2019_1	AGTATTGCCCTATTTTCTTCATAACTGGT (SEQ ID NO: 517)	29	34.5	
nCoV-2019_39_RIGHT	nCoV-2019_1	TGTAACTGGACACATTGAGCCC (SEQ ID NO: 518)	22	50.0	
nCoV-2019_40_LEFT	nCoV-2019_2	TGCACATCAGTAGTCTTACTCTCAGT (SEQ ID NO: 519)	26	42.3	
nCoV-2019_40_RIGHT	nCoV-2019_2	CATGGCTGCATCACGGTCAAAT (SEQ ID NO: 520)	22	50.0	
nCoV-2019_41_LEFT	nCoV-2019_1	GTTCCCTTCCATCATATGCAGCT (SEQ ID NO: 521)	23	47.8	
nCoV-2019_41_RIGHT	nCoV-2019_1	TGGTATGACAACCATTAGTTTGGCT (SEQ ID NO: 522)	25	40.0	
nCoV-2019_42_LEFT	nCoV-2019_2	TGCAAGAGATGGTTGTGTTCCC (SEQ ID NO: 523)	22	50.0	
nCoV-2019_42_RIGHT	nCoV-2019_2	CCTACCTCCCTTTGTTGTGTTGT (SEQ ID NO: 524)	23	47.8	
nCoV-2019_43_LEFT	nCoV-2019_1	TACGACAGATGTCTTGTGCTGC (SEQ ID NO: 525)	22	50.0	
nCoV-2019_43_RIGHT	nCoV-2019_1	AGCAGCATCTACAGCAAAAGCA (SEQ ID NO: 526)	22	45.5	

nCoV-2019_44_LEFT_alt3	nCoV-2019_2	CCACAGTACGTCTACAAGCTGG (SEQ ID NO: 527)	22	54.6	
nCoV-2019_44_RIGHT_alt0	nCoV-2019_2	CGCAGACGGTACAGACTGTGTT (SEQ ID NO: 528)	22	54.6	
nCoV-2019_45_LEFT_alt2	nCoV-2019_1	AGTATGTACAAATACCTACAACTTGTGCT (SEQ ID NO: 529)	29	34.5	X
nCoV-2019_45_RIGHT_alt7	nCoV-2019_1	TTCATGTTGGTAGTTAGAGAAAGTGTGTC (SEQ ID NO: 530)	29	37.9	X
nCoV-2019_46_LEFT_alt1	nCoV-2019_2	CGCTTCCAAGAAAAGGACGAAGA (SEQ ID NO: 531)	23	47.8	
nCoV-2019_46_RIGHT_alt2	nCoV-2019_2	CACGTTCACCTAAGTTGGCGTAT (SEQ ID NO: 532)	23	47.8	
nCoV-2019_47_LEFT	nCoV-2019_1	AGGACTGGTATGATTTTGTAGAAAACCC (SEQ ID NO: 533)	28	39.3	
nCoV-2019_47_RIGHT	nCoV-2019_1	AATAACGGTCAAAGAGTTTTAACCTCTC (SEQ ID NO: 534)	28	35.7	
nCoV-2019_48_LEFT	nCoV-2019_2	TGTTGACACTGACTTAACAAAGCCT (SEQ ID NO: 535)	25	40.0	
nCoV-2019_48_RIGHT	nCoV-2019_2	TAGATTACCAGAAGCAGCGTGC (SEQ ID NO: 536)	22	50.0	
nCoV-2019_49_LEFT	nCoV-2019_1	AGGAATTACTTGTGTATGCTGCTGA (SEQ ID NO: 537)	25	40.0	
nCoV-2019_49_RIGHT	nCoV-2019_1	TGACGATGACTTGGTTAGCATTAATACA (SEQ ID NO: 538)	28	35.7	
nCoV-2019_50_LEFT	nCoV-2019_2	GTTGATAAGTACTTTGATTGTTACGATGGT (SEQ ID NO: 539)	30	33.3	
nCoV-2019_50_RIGHT	nCoV-2019_2	TAACATGTTGTGCCAACCACCA (SEQ ID NO: 540)	22	45.5	
nCoV-2019_51_LEFT	nCoV-2019_1	TCAATAGCCGCCACTAGAGGAG (SEQ ID NO: 541)	22	54.6	
nCoV-2019_51_RIGHT	nCoV-2019_1	AGTGCATTAACATTGGCCGTGA (SEQ ID NO: 542)	22	45.5	
nCoV-2019_52_LEFT	nCoV-2019_2	CATCAGGAGATGCCACAACTGC (SEQ ID NO: 543)	22	54.6	
nCoV-2019_52_RIGHT	nCoV-2019_2	GTTGAGAGCAAAATTCATGAGGTCC (SEQ ID NO: 544)	25	44.0	
nCoV-2019_53_LEFT	nCoV-2019_1	AGCAAAATGTTGGACTGAGACTGA (SEQ ID NO: 545)	24	41.7	
nCoV-2019_53_RIGHT	nCoV-2019_1	AGCCTCATAAAACTCAGGTTCCC (SEQ ID NO: 546)	23	47.8	
nCoV-2019_54_LEFT	nCoV-2019_2	TGAGTTAACAGGACACATGTTAGACA (SEQ ID NO: 547)	26	38.5	
nCoV-2019_54_RIGHT	nCoV-2019_2	AACCAAAAACTTGTCCATTAGCACA (SEQ ID NO: 548)	25	36.0	

nCoV-2019_55_LEFT	nCoV-2019_1	ACTCAACTTTACTTAGGAGGTATGAGCT (SEQ ID NO: 549)	28	39.3	
nCoV-2019_55_RIGHT	nCoV-2019_1	GGTGTACTCTCCTATTTGTACTTTACTGT (SEQ ID NO: 550)	29	37.9	
nCoV-2019_56_LEFT	nCoV-2019_2	ACCTAGACCACCACTTAACCGA (SEQ ID NO: 551)	22	50.0	
nCoV-2019_56_RIGHT	nCoV-2019_2	ACACTATGCGAGCAGAAGGGTA (SEQ ID NO: 552)	22	50.0	
nCoV-2019_57_LEFT	nCoV-2019_1	ATTCTACACTCCAGGGACCACC (SEQ ID NO: 553)	22	54.6	
nCoV-2019_57_RIGHT	nCoV-2019_1	GTAATTGAGCAGGGTCGCCAAT (SEQ ID NO: 554)	22	50.0	
nCoV-2019_58_LEFT	nCoV-2019_2	TGATTTGAGTGTTGTCAATGCCAGA (SEQ ID NO: 555)	25	40.0	
nCoV-2019_58_RIGHT	nCoV-2019_2	CTTTTCTCCAAGCAGGGTTACGT (SEQ ID NO: 556)	23	47.8	
nCoV-2019_59_LEFT	nCoV-2019_1	TCACGCATGATGTTTCATCTGCA (SEQ ID NO: 557)	23	43.5	
nCoV-2019_59_RIGHT	nCoV-2019_1	AAGAGTCCTGTTACATTTTCAGCTTG (SEQ ID NO: 558)	26	38.5	
nCoV-2019_60_LEFT	nCoV-2019_2	TGATAGAGACCTTTATGACAAGTTGCA (SEQ ID NO: 559)	27	37.0	
nCoV-2019_60_RIGHT	nCoV-2019_2	GGTACCAACAGCTTCTCTAGTAGC (SEQ ID NO: 560)	24	50.0	
nCoV-2019_61_LEFT	nCoV-2019_1	TGTTTATCACCCGCGAAGAAGC (SEQ ID NO: 561)	22	50.0	
nCoV-2019_61_RIGHT	nCoV-2019_1	ATCACATAGACAACAGGTGCGC (SEQ ID NO: 562)	22	50.0	
nCoV-2019_62_LEFT	nCoV-2019_2	GGCACATGGCTTTGAGTTGACA (SEQ ID NO: 563)	22	50.0	
nCoV-2019_62_RIGHT	nCoV-2019_2	GTTGAACCTTTCTACAAGCCGC (SEQ ID NO: 564)	22	50.0	
nCoV-2019_63_LEFT	nCoV-2019_1	TGTTAAGCGTGTTGACTGGACT (SEQ ID NO: 565)	22	45.5	
nCoV-2019_63_RIGHT	nCoV-2019_1	ACAAACTGCCACCATCACAACC (SEQ ID NO: 566)	22	50.0	
nCoV-2019_64_LEFT	nCoV-2019_2	TCGATAGATATCCTGCTAATTCCATTGT (SEQ ID NO: 567)	28	35.7	X
nCoV-2019_64_RIGHT	nCoV-2019_2	AGTCTTGTAAAAGTGTTCCAGAGGT (SEQ ID NO: 568)	25	40.0	X
nCoV-2019_65_LEFT	nCoV-2019_1	GCTGGCTTTAGCTTGTGGGTTT (SEQ ID NO: 569)	22	50.0	
nCoV-2019_65_RIGHT	nCoV-2019_1	TGTCAGTCATAGAACAAACACCAATAGT (SEQ ID NO: 570)	28	35.7	

nCoV-2019_66_LEFT	nCoV-2019_2	GGGTGTGGACATTGCTGCTAAT (SEQ ID NO: 571)	22	50.0	X
nCoV-2019_66_RIGHT	nCoV-2019_2	TCAATTTCCATTTGACTCCTGGGT (SEQ ID NO: 572)	24	41.7	X
nCoV-2019_67_LEFT	nCoV-2019_1	GTTGTCCAACAATTACCTGAAACTTACT (SEQ ID NO: 573)	28	35.7	X
nCoV-2019_67_RIGHT	nCoV-2019_1	CAACCTTAGAAACTACAGATAAATCTTGGG (SEQ ID NO: 574)	30	36.7	X
nCoV-2019_68_LEFT	nCoV-2019_2	ACAGGTTCATCTAAGTGTGTGTGT (SEQ ID NO: 575)	24	41.7	
nCoV-2019_68_RIGHT	nCoV-2019_2	CTCCTTTATCAGAACCAGCACCA (SEQ ID NO: 576)	23	47.8	
nCoV-2019_69_LEFT	nCoV-2019_1	TGTCGCAAAATATACTCAACTGTGTCA (SEQ ID NO: 577)	27	37.0	
nCoV-2019_69_RIGHT	nCoV-2019_1	TCTTTATAGCCACGGAACCTCCA (SEQ ID NO: 578)	23	47.8	
nCoV-2019_70_LEFT	nCoV-2019_2	ACAAAAGAAAATGACTCTAAAGAGGGTTT (SEQ ID NO: 579)	29	31.0	X
nCoV-2019_70_RIGHT	nCoV-2019_2	TGACCTTCTTTTAAAGACATAACAGCAG (SEQ ID NO: 580)	28	35.7	X
nCoV-2019_71_LEFT	nCoV-2019_1	ACAAATCCAATTCAGTTGTCTTCCTATTC (SEQ ID NO: 581)	29	34.5	X
nCoV-2019_71_RIGHT	nCoV-2019_1	TGGAAAAGAAAGGTAAGAACAAGTCCT (SEQ ID NO: 582)	27	37.0	X
nCoV-2019_72_LEFT	nCoV-2019_2	ACACGTGGTGTTTATTACCCTGAC (SEQ ID NO: 583)	24	45.8	
nCoV-2019_72_RIGHT	nCoV-2019_2	ACTCTGAACTCACTTTCCATCCAAC (SEQ ID NO: 584)	25	44.0	
nCoV-2019_73_LEFT	nCoV-2019_1	CAATTTTGTAATGATCCATTTTTGGGTGT (SEQ ID NO: 585)	29	31.0	
nCoV-2019_73_RIGHT	nCoV-2019_1	CACCAGCTGTCCAACCTGAAGA (SEQ ID NO: 586)	22	54.6	
nCoV-2019_74_LEFT	nCoV-2019_2	ACATCACTAGGTTTCAAACTTTACTTGC (SEQ ID NO: 587)	28	35.7	
nCoV-2019_74_RIGHT	nCoV-2019_2	GCAACACAGTTGCTGATTCTCTTC (SEQ ID NO: 588)	24	45.8	
nCoV-2019_75_LEFT	nCoV-2019_1	AGAGTCCAACCAACAGAATCTATTGT (SEQ ID NO: 589)	26	38.5	
nCoV-2019_75_RIGHT	nCoV-2019_1	ACCACCAACCTTAGAATCAAGATTGT (SEQ ID NO: 590)	26	38.5	
nCoV-2019_76_LEFT_alt3	nCoV-2019_2	GGGCAAACTGGAAAGATTGCTGA (SEQ ID NO: 591)	23	47.8	X
nCoV-2019_76_RIGHT_alt0	nCoV-2019_2	ACCTGTGCCTGTTAAACCATTGA (SEQ ID NO: 592)	23	43.5	X

nCoV-2019_77_LEFT	nCoV-2019_1	CCAGCAACTGTTTGTGGACCTA (SEQ ID NO: 593)	22	50.0	
nCoV-2019_77_RIGHT	nCoV-2019_1	CAGCCCCTATTAAACAGCCTGC (SEQ ID NO: 594)	22	54.6	
nCoV-2019_78_LEFT	nCoV-2019_2	CAACTTACTCCTACTTGGCGTGT (SEQ ID NO: 595)	23	47.8	
nCoV-2019_78_RIGHT	nCoV-2019_2	TGTGTACAAAAACTGCCATATTGCA (SEQ ID NO: 596)	25	36.0	
nCoV-2019_79_LEFT	nCoV-2019_1	GTGGTGATTCAACTGAATGCAGC (SEQ ID NO: 597)	23	47.8	X
nCoV-2019_79_RIGHT	nCoV-2019_1	CATTTCATCTGTGAGCAAAGGTGG (SEQ ID NO: 598)	24	45.8	X
nCoV-2019_80_LEFT	nCoV-2019_2	TTGCCTTGGTGATATTGCTGCT (SEQ ID NO: 599)	22	45.5	X
nCoV-2019_80_RIGHT	nCoV-2019_2	TGGAGCTAAGTTGTTTAACAAGCG (SEQ ID NO: 600)	24	41.7	X
nCoV-2019_81_LEFT	nCoV-2019_1	GCACTTGGAAAACTTCAAGATGTGG (SEQ ID NO: 601)	25	44.0	
nCoV-2019_81_RIGHT	nCoV-2019_1	GTGAAGTTCTTTTCTTGTGCAGGG (SEQ ID NO: 602)	24	45.8	
nCoV-2019_82_LEFT	nCoV-2019_2	GGGCTATCATCTTATGTCCTTCCCT (SEQ ID NO: 603)	25	48.0	
nCoV-2019_82_RIGHT	nCoV-2019_2	TGCCAGAGATGTCACCTAAATCAA (SEQ ID NO: 604)	24	41.7	
nCoV-2019_83_LEFT	nCoV-2019_1	TCCTTTGCAACCTGAATTAGACTCA (SEQ ID NO: 605)	25	40.0	
nCoV-2019_83_RIGHT	nCoV-2019_1	TTTGACTCCTTTGAGCACTGGC (SEQ ID NO: 606)	22	50.0	
nCoV-2019_84_LEFT	nCoV-2019_2	TGCTGTAGTTGTCTCAAGGGCT (SEQ ID NO: 607)	22	50.0	
nCoV-2019_84_RIGHT	nCoV-2019_2	AGGTGTGAGTAAACTGTTACAAACAAC (SEQ ID NO: 608)	27	37.0	
nCoV-2019_85_LEFT	nCoV-2019_1	ACTAGCACTCTCCAAGGGTGTT (SEQ ID NO: 609)	22	50.0	
nCoV-2019_85_RIGHT	nCoV-2019_1	ACACAGTCTTTTACTCCAGATTCCC (SEQ ID NO: 610)	25	44.0	
nCoV-2019_86_LEFT	nCoV-2019_2	TCAGGTGATGGCACAACAAGTC (SEQ ID NO: 611)	22	50.0	
nCoV-2019_86_RIGHT	nCoV-2019_2	ACGAAAGCAAGAAAAAGAAGTACGC (SEQ ID NO: 612)	25	40.0	
nCoV-2019_87_LEFT	nCoV-2019_1	CGACTACTAGCGTGCCTTTGTA (SEQ ID NO: 613)	22	50.0	
nCoV-2019_87_RIGHT	nCoV-2019_1	ACTAGGTTCCATTGTTCAAGGAGC (SEQ ID NO: 614)	24	45.8	

nCoV-2019_88_LEFT	nCoV-2019_2	CCATGGCAGATTCCAACGGTAC (SEQ ID NO: 615)	22	54.6	
nCoV-2019_88_RIGHT	nCoV-2019_2	TGGTCAGAATAGTGCCATGGAGT (SEQ ID NO: 616)	23	47.8	
nCoV-2019_89_LEFT_alt2	nCoV-2019_1	CGCGTTCCATGTGGTCATTCAA (SEQ ID NO: 617)	22	50.0	
nCoV-2019_89_RIGHT_alt4	nCoV-2019_1	ACGAGATGAAACATCTGTTGTCACT (SEQ ID NO: 618)	25	40.0	
nCoV-2019_90_LEFT	nCoV-2019_2	ACACAGACCATTCCAGTAGCAGT (SEQ ID NO: 619)	23	47.8	
nCoV-2019_90_RIGHT	nCoV-2019_2	TGAAATGGTGAATTGCCCTCGT (SEQ ID NO: 620)	22	45.5	
nCoV-2019_91_LEFT	nCoV-2019_1	TCACTACCAAGAGTGTGTTAGAGGT (SEQ ID NO: 621)	25	44.0	X
nCoV-2019_91_RIGHT	nCoV-2019_1	TTCAAGTGAGAACCAAAAGATAATAAGCA (SEQ ID NO: 622)	29	31.0	X
nCoV-2019_92_LEFT	nCoV-2019_2	TTTGTGCTTTTTAGCCTTTCTGCT (SEQ ID NO: 623)	24	37.5	
nCoV-2019_92_RIGHT	nCoV-2019_2	AGGTTCCTGGCAATTAATTGTAAAAGG (SEQ ID NO: 624)	27	37.0	
nCoV-2019_93_LEFT	nCoV-2019_1	TGAGGCTGGTTCTAAATCACCCA (SEQ ID NO: 625)	23	47.8	
nCoV-2019_93_RIGHT	nCoV-2019_1	AGGTCTTCCTTGCCATGTTGAG (SEQ ID NO: 626)	22	50.0	
nCoV-2019_94_LEFT	nCoV-2019_2	GGCCCCAAGGTTTACCCAATAA (SEQ ID NO: 627)	22	50.0	
nCoV-2019_94_RIGHT	nCoV-2019_2	TTTGGCAATGTTGTTCCTTGAGG (SEQ ID NO: 628)	23	43.5	
nCoV-2019_95_LEFT	nCoV-2019_1	TGAGGGAGCCTTGAATACACCA (SEQ ID NO: 629)	22	50.0	
nCoV-2019_95_RIGHT	nCoV-2019_1	CAGTACGTTTTTGCCGAGGCTT (SEQ ID NO: 630)	22	50.0	
nCoV-2019_96_LEFT	nCoV-2019_2	GCCAACAACAACAAGGCCAAAC (SEQ ID NO: 631)	22	50.0	
nCoV-2019_96_RIGHT	nCoV-2019_2	TAGGCTCTGTTGGTGGGAATGT (SEQ ID NO: 632)	22	50.0	
nCoV-2019_97_LEFT	nCoV-2019_1	TGGATGACAAAGATCCAAATTTCAAAGA (SEQ ID NO: 633)	28	32.1	
nCoV-2019_97_RIGHT	nCoV-2019_1	ACACACTGATTAAAGATTGCTATGTGAG (SEQ ID NO: 634)	28	35.7	
nCoV-2019_98_LEFT	nCoV-2019_2	AACAATTGCAACAATCCATGAGCA (SEQ ID NO: 635)	24	37.5	
nCoV-2019_98_RIGHT	nCoV-2019_2	TTCTCCTAAGAAGCTATTAAAATCACATGG (SEQ ID NO: 636)	30	33.3	

Table 5. Time and Cost Comparison of FLEX vs XT

Library Prep Kit	Cost Per Sample ($)	Time (hrs)
Illumina DNA Flex	45.96	10
Illumina Nextera XT	64.43	13.5

Table 6. Cost of SDSI+AmpSeq

Processing Step	Reagent	Vendor	Item Number	Cost (dollars)	Number of Reactions	Cost per Reaction
Biosample Extraction	MagMAX™ mirVana™Total RNA Isolation Kit	Thermo Fisher Scientific	A27828	495	96	5.16
	SSIV RT master mix	Thermo Fisher Scientific	18090050	383	50	7.66
cDNA Synthesis	Random hexamers (50ng/ul)	Thermo Fisher Scientific	N808127	91	100	0.91
	dNTPs (10nM)	Thermo Fisher Scientific	18427-013	99	100	0.99
	5x RT buffer	Thermo Fisher Scientific	18090050	x	x	x
	DTT (100mM)	Thermo Fisher Scientific	18090050	x	x	x
	Superase rnase inhibitor	Thermo Fisher Scientific	10777-019	188	125	1.50
ARTIC PCR	Q5 Hot Start High-Fidelity 2X Master Mix	New England BioLabs	M0494L	845	500	1.69
	Artic Primers Pool#1 and Pool#2	IDT		30	500	0.06
Spike-ins	Spike in Primers (Forward/Reverse)	IDT		500	1000000	0.00
	Spike-in targets n=96	IDT		5821	1000000	0.01
Post Artic Pooling Quantification	Qubit™ dsDNA HS Assay Kit	Thermo Fisher Scientific	Q32854	308	500	0.62

Library Construction	Nextera DNA flex Library Prep (n=96)	Illumina	20018705	4153	190	21.86
	Nextera index UD Set A (n=96)	Illumina	20027213	672	384	1.75
Library Quantification	High Sensitivity D1000 ScreenTape	Agilent	5067-5584	362	112	3.23
	High Sensitivity D1000 Sample Buffer	Agilent	5067-5603	59.14	112	0.53
TOTAL:						45.96

Table 7. Final Library Quantification for Enzyme Optimization Experiment

PCR Enzyme	KOD
Q5 2X MM	**Quant (nM)** 85.3
Q5 2X MM + 0.01% S	0.258
Q5 Ultra	34.5
KAPA	56.0
	8.1

Table 8. Library Size DNA Flex

Sample	CT Dilution	Standard DNA Flex Library Size (bp)	Standard DNA Flex .5X DNA Flex Library Concentration (nM)	.5X DNA Flex Library Size (bp)	.5X DNA Flex Library Concentration (nM)
MA_MGH_00109	15.39	340	332	92	54.3
MA_MGH_00110	26.39	293	271	13.4	6.84
MA_MGH_00113	31.93	211	207	3.05	1.84

***

~~[0158]~~[0167] Various modifications and variations of the described methods, pharmaceutical compositions, and kits of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific embodiments, it will be understood that it is capable of further modifications and that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are obvious to those skilled in the art are intended to be within the scope of the invention. This application is intended to cover any variations, uses, or adaptations of the invention following, in general, the principles of the invention and including such departures from the present disclosure come within known customary practice within the art to which the invention pertains and may be applied to the essential features herein before set forth.

AMPLICON-BASED SEQUENCING USING DNA SPIKE-INS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application Nos. 63/155,258, filed March 1, 2021, and 63/273,117, filed October 10, 2021. The entire contents of the above-identified applications are hereby fully incorporated herein by reference.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

[0002] This invention was made with government support under Grant Nos. AI110818, AI147868, HG010669, and CK000490 awarded by the National Institutes of Health, Grant No. 223-101-8101 awarded by the United States Food and Drug Administration, and Grant No. 75D30120C09605 awarded by the Centers for Diseases Control. The government has certain rights in the invention.

REFERENCE TO AN ELECTRONIC SEQUENCE LISTING

[0003] The contents of the electronic sequence listing ("BROD-5360US_ST25.txt"; Size is 205,235 bytes and it was created on February 17, 2022) is herein incorporated by reference in its entirety.

TECHNICAL FIELD

[0004] The subject matter disclosed herein is generally directed to synthetic DNA spike-ins and their use for detecting, quantifying, and preventing amplification contamination in genome profiling analysis.

BACKGROUND

[0005] The COVID-19 pandemic has demonstrated, once again, the crucial role of genomic sequencing in combatting infectious disease outbreaks globally. Monitoring the emergence of pathogens and the spread of variants of concern has become commonplace in government, academic, and private laboratories[1,2]. Genomics data provides insights into the diversity, evolution and transmission of a virus, a critical guide for public health interventions ranging from contact tracing, identifying cases of reinfection, or documenting resistance to clinical interventions[3–6]. In

the year since, genomic data have provided new insights into the diversity, evolution and transmission of the virus, which has increasingly been used to guide impactful public health interventions. In particular, scientists have employed viral genome sequencing to characterize the fine-scale epidemiology of clusters and superspreading events (Lemieux et al., 2021, Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events, *Science*, 371(6529); Popa et al., 2020, Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2, *Science Translational Medicine*, 12(573); Volz et al., 2021, Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data, *bioRxiv*, *medRxiv*). More recently, genome sequencing to monitor the emergence of new lineages and the spread of variants of concern (VoC) has become paramount (Washington et al., 2021, Genomic epidemiology identifies emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States, *medRxiv*). As laboratories are now performing viral genomic sequencing on SARS-CoV-2 at an unprecedented scale[7,8], it highlights the need for stringent requirements to ensure the integrity of genomes being produced.

[0006] Multiplexed amplicon-based genome sequencing methods have accelerated the massive scale of SARS-CoV-2 genomic surveillance due to their improved sensitivity, cost, and speed over other, lower-amplification RNA sequencing approaches, such as unbiased metagenomic sequencing[9].Unsurprisingly, amplicon-based approaches that target the SARS-CoV-2 genome for amplification and subsequent sequencing have become the genomic surveillance method of choice during the ongoing pandemic (over 90% of Short Read Archive submissions). In just a year since the first genome sequence enabled the identification of SARS-CoV-2, hundreds of thousands of complete genomes have been sequenced and released by a relatively small group of several hundred laboratories. An open-access tiled primer set developed by the ARTIC network (artic.network/) is the most widely used method for SARS-CoV-2 specific genome amplification followed by sequencing on either Illumina or nanopore instruments (Quick et al., 2017; Tyson et al., 2020). A wide array of protocols and publications are now available that integrate these ARTIC primers with different amplification and library construction indexing strategies (Baker et al., 2020; Gohl et al., 2020). Approaches such as batching samples by viral load to increase sensitivity are impractical to scale to current needs, resulting in incomplete recovery of viral genomes, especially from low titer samples.

[0007] However, the risk for contamination during the amplification stage is especially high as the 35 or more cycles of virus-specific PCR produces trillions of SARS-CoV-2 amplicons in a single reaction. Other high-risk modes of contamination, including sample swaps, cross-contamination of samples, or aerosolization, can occur throughout the sample processing pipeline. With many laboratories performing viral sequencing by processing multiple large batches in parallel, the potential for contamination increases[10]. Even small amounts of sample mixing or contaminating amplicons could potentially confound studies where viral detection is sensitive to only tens of molecules[10,11]. Moreover, as SARS-CoV-2 has relatively low genetic diversity and often spreads in local outbreaks or clusters[11,12], many genomes are expected to be identical at the consensus level[11,15–17], a pattern that could also be observed due to contamination. The risk of contamination, and the challenges in detecting it, can confound a wide array of genomic analyses including estimates of the frequencies of variants, lineage dynamics, and transmission events. Additionally, methods to address the critical risk of sample processing errors in clinical sequencing could enable its use more widely in clinical decision making.

[0008] To meet the genomic surveillance goals laid out by local and world governments, sequencing efforts will need to be scaled to thousands of centers, many performing viral genomics for the first time. Additional laboratories will enter the SARS-CoV-2 sequencing space with an emphasis to rapidly surveil VoCs for clinical significance, with even higher requirements to ensure the integrity of SARS-CoV-2 genomes being produced. While inclusion of internal standards is commonplace in many experimental approaches[13–15] and some technical assay controls exist for DNA sequencing[16–18], the use of internal controls is currently rare in amplicon-based genomic surveillance. Here Applicants developed and extensively tested a sample identification method using 96 synthetic DNA spike-ins (SDSIs) for amplicon-based sequencing approaches. Using the widely used open-access ARTIC tiled primer design (artic.network/), Applicants implemented these SDSIs for SARS-CoV-2 genomic sequencing from thousands of residual diagnostic (clinical) samples. The resulting user-friendly and highly versatile SDSI+AmpSeq protocol can be easily implemented to improve the quality of genomic data generated for epidemiological and clinical investigations of human pathogens (**FIG. 1 and FIG. 13, Table 6**).

[0009] Citation or identification of any document in this application is not an admission that such a document is available as prior art to the present invention.

SUMMARY

[0010] In one aspect, the present invention provides for a method of detecting and preventing contamination in one or more cDNA samples comprising adding a synthetic DNA spike-in (SDSI) to each cDNA sample, wherein each SDSI is capable of amplification simultaneously with the cDNA, and wherein each SDSI comprises a unique sequence capable of differentiating each SDSI; amplifying one or more of the cDNA samples and SDSI; sequencing the amplified sample; and determining the number of reads of the spike-in from the one or more samples. In certain example embodiments, the sample is associated with drug resistance. In certain example embodiments, the sample is for sequencing a pathogen or family of pathogens. In certain example embodiments, the pathogen is a virus. In certain example embodiments, the pathogen is a bacteria and the region sequenced is associated with antibiotic resistance. In certain example embodiments, each sample contains a viral nucleic acid sequence. In certain example embodiments, the samples are for creating one or more sequencing families/clusters.

[0011] In certain example embodiments, the SDSI contains a core region and a primer binding region at the 3' end and the 5' end. In certain example embodiments, the core sequence of the SDSI is derived from a rare organism. In certain example embodiments, the rare organism is a thermophilic archaea. In certain example embodiments, the core sequence homology is less than 65%, or less than 60%, or less than 55%, or less than 50%, or less than 45%, or less than 40%, or less than 35%, or less than 30%, or less than 25%, or less than 20%, or less than 15%, or less than 5%, or less than 1% to a sample sequence. In certain example embodiments, the core sequence homology is less than 15, or less than 20, or less than 25, or less than 30, or less than 35, or less than 40, or less than 45, or less than 50 contiguous bases in common with the sample sequence.

[0012] In certain example embodiments, the synthetic DNA spike-in sequences are 50 – 5000 nucleotides in length. In certain example embodiments, the SDSI minimizes self-hybridization and cross-hybridization with nucleic acids in the sample. In certain example embodiments, the primer binding sites of the SDSI have a Tm between 55-65 ^{o}C. In certain example embodiments, the method further comprises a plurality of SDSIs. In certain example embodiments, the core sequence of the synthetic DNA comprises a sequence as set forth in SEQ ID NOS: 1 – 96 and 193-291. In certain example embodiments, the primer binding sequences are complementary to the primers having SEQ ID NOS: 391 and 392. In certain example embodiments the SDSIs comprise one or

more of SEQ ID NOS: 97-192 and 292-390. In example embodiments, sequences can be used in the alternative. In one example embodiment, sequence SEQ ID NO: 289 can substitute for sequence SEQ ID NO: 16. In one example embodiment, sequence SEQ ID NO: 290 can substitute for sequence SEQ ID NO: 57. In one example embodiment, sequence SEQ ID NO: 291 can substitute for sequence SEQ ID NO: 66. In one example embodiment, sequence SEQ ID NO: 388 can substitute for sequence SEQ ID NO: 112. In one example embodiment, sequence SEQ ID NO: 389 can substitute for sequence SEQ ID NO: 153. In one example embodiment, sequence SEQ ID NO: 390 can substitute for sequence SEQ ID NO: 162. In one example embodiment, one or more of SEQ ID NOS: 16, 57, 66, 112, 153, and 162 can be substituted with their alternative sequence SEQ ID NOS: 289, 290, 291, 388, 389, and 390, respectively.

[0013] In certain example embodiments, the concentration of synthetic DNA spike-ins range from 0.1 femtomolar – 1.0 femtomolar. In certain example embodiments, the presence of an amplified spike-in corresponding to the spike-in added to a sample indicates a decreased risk of contamination. In certain example embodiments, the presence of an amplified spike-in corresponding to the spike-in not added to a sample indicates an increased risk of contamination.

[0014] In another aspect, the present invention is a set of synthetic DNA spike-ins (SDSIs), each SDSI in the set comprising a primer binding sequence at the 3' and 5' end and a unique core sequence between the 3' and 5' primer binding sequences. In certain example embodiments, the set comprises at least 96 spike-ins. In certain example embodiments, the unique core sequence is derived from a rare organism. In certain example embodiments, the rare organism is a thermophilic archaea. In certain example embodiments, the core sequence homology is less than 65%, or less than 60%, or less than 55%, or less than 50%, or less than 45%, or less than 40%, or less than 35%, or less than 30%, or less than 25%, or less than 20%, or less than 15%, or less than 5%, or less than 1% to a sample sequence. In certain example embodiments, the core sequence homology is less than 15, or less than 20, or less than 25, or less than 30, or less than 35, or less than 40, or less than 45, or less than 50 contiguous bases in common with the sample sequence.

[0015] In certain example embodiments, the sequence is 50-5000 nucleotides in length. In certain example embodiments, the SDSIs minimizes self-hybridization and cross-hybridization with nucleic acids in the sample. In certain example embodiments, the primer binding sites have a Tm between 55-65 ^{o}C. In certain example embodiments, the core sequence are the unique

sequences as set forth SEQ ID NOS: 1 – 96 and 193-291. In certain example embodiments, the primer binding sequences are complementary to the primers having SEQ ID NOS: 391 and 392. In certain example embodiments, the SDSIs comprise one or more of SEQ ID NOS: 97-192 and 292-390. In example embodiments, sequences can be used in the alternative. In one example embodiment, sequence SEQ ID NO: 289 can substitute for sequence SEQ ID NO: 16. In one example embodiment, sequence SEQ ID NO: 290 can substitute for sequence SEQ ID NO: 57. In one example embodiment, sequence SEQ ID NO: 291 can substitute for sequence SEQ ID NO: 66. In one example embodiment, sequence SEQ ID NO: 388 can substitute for sequence SEQ ID NO: 112. In one example embodiment, sequence SEQ ID NO: 389 can substitute for sequence SEQ ID NO: 153. In one example embodiment, sequence SEQ ID NO: 390 can substitute for sequence SEQ ID NO: 162. In one example embodiment, one or more of SEQ ID NOS: 16, 57, 66, 112, 153, and 162 can be substituted with their alternative sequence SEQ ID NOS: 289, 290, 291, 388, 389, and 390, respectively.

[0016] These and other aspects, objects, features, and advantages of the example embodiments will become apparent to those having ordinary skill in the art upon consideration of the following detailed description of example embodiments.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] An understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention may be utilized, and the accompanying drawings of which:

[0018] **FIG. 1 – SDSI-ARTIC Amplicon-Sequencing Protocol** – Illustrative workflow for 48 samples through the SDSI+ARTIC amplicon-sequencing pipeline. A synthetic DNA spike-ins (SDSI) will be added to each sample to allow for contamination tracking and accurate sample identification.

[0019] **FIG. 2A-2C – Synthetic DNA oligos spiked into amp-seq reactions flag contamination and sample swaps** – **A.** Schematic detailing SDSI design. Each oligo contains 140 bp of sui generis sequence flanked by unique primer binding sites. Primers designed to amplify SDSIs are added to ARTIC primer pools, and a unique SDSI is added to each clinical sample. Identification of multiple SDSIs in the same sample indicates contamination. **B.** In a titration of

SDSIs across clinical samples with variable CTs, the number of reads mapping to both SARS-CoV-2 and the SDSI were quantified, and the percentage of each was calculated. **C.** For each of 48 unique clinical samples (on the horizontal axis), reads mapping to each of 48 unique SDSIs (on the vertical axis) were quantified; the log of this read count is represented by the intensity of color displayed. Samples and SDSIs were ordered such that the intended match is on the diagonal of this matrix, thus any off-diagonal signal would reveal non-specific identification of SDSIs or contamination of SDSIs across samples.

[0020] FIG. 3A-3D – Maximizing Genome Recovery and Coverage with SDSI-ARTIC – **A.** The percent of the target genome covered at various depths of coverage when three reverse transcriptases, Superscript III, Superscript IV, and Superscript VILO were used for cDNA synthesis. Data represents four individual samples. **B.** Amplicons with at least 0.2X of the mean amplicon coverage with the normal ARTIC v3 primer pools or with a modified primer pool with a 2X concentration of 20 different ARTIC primer pairs. Four samples with low, mid-low, mid-high, and high CTs were used. **C.** Gini coefficients for two mid-high CT samples and four high CT when using either 35, 40, or 45 cycles for the ARTIC PCR. Error bars represent standard deviation. **D.** Comparison of Nextera DNA Flex and Nextera XT on the number of SARS-CoV-2 base pairs covered at various depths of coverage for three samples at different CTs.

[0021] FIG. 4A-4C – Improved amp-seq assembles more complete genomes than metagenomic sequencing with few errors across a wide range of samples – **A.** SDSI+ARTIC (N=81) and metagenomic (N=81) assembly lengths. All samples were downsampled to 975,000 reads. Dotted line indicates median assembly length (SDSI+ARTIC = 29,577; Metagenomic = 4,389) **B.** Percent of assemblies with greater than 98% or 80% coverage in different CT bins (SDSI+ARTIC N=81, Metagenomic N=81) (downsampled to 975,000 reads). **C.** SNP concordance plot between SDSI+ARTIC and metagenomic consensus sequences. Two discordant SNPs, outlined in a red box, were found.

[0022] FIG. 5A-5C – Rapid deployment of optimized amp-seq to determine a nosocomial transmission cluster – **A.** Phylogenetic tree showing the location of the putative cluster sequences in the context of a global subset of circulating SARS-CoV-2 diversity. Zoom box shows the 10 highly similar cluster genomes. Sample named on the main tree is the one putative cluster sample that was excluded from the cluster based on genome sequence. **B.** Distance matrix showing

pairwise differences between the 17 complete genomes assembled from this sample set. Putative cluster samples are bolded. **C.** Spike-in counts for each of the 24 samples and water controls in this sequencing batch.

[0023] FIG. 6A-6C – Spike-in validation – **A.** 100fmol DNA spike-in amplified under standard ARTIC PCR conditions for 40 cycles run on 2.2% agarose gel image with 188bp amplified spike-in (SDSI 1-48) **B.** RT-PCR for Spike-in and spike-in specific primers, Spike-in specific primers water control, Spike-in with COVID positive cDNA and spike-in specific primers, COVID positive cDNA and spike-in specific primers. **C.** Both SDSIs and ARTIC amplicons avoid extremes of GC content, and the two have generally overlapping distributions. SDSI primers also have a length and GC content similar to the average ARTIC v3 primer, resulting in a compatible TM.

[0024] FIG. 7 – SDSI Titration – Coverage plots for four different SDSI concentrations (1fM, 0.1fM, 0.01fM, 0.001fM) at four different CT dilutions (CT=20,25,30,35).

[0025] FIG. 8A-8C – Comparison to alternate amp-seq strategies – **A.** Three representative coverage plots for CT 20, CT 25, and CT 30 samples. **B.** SNP detection for the CT 20 and CT 25 sample. ARTIC and Paragon consensus sequences were compared to the metagenomic consensus sequences. The SNP that was not called in Paragon was due to low coverage at that position. Analysis was performed with assemblies generated with a minimum coverage of both 3 and 20, yielding identical results. **C.** Base pairs of the SARS-CoV-2 genome covered for the modified ARTIC pipeline versus Paragon CleanPlex Panel at different depths of coverage. Five samples at varying CTs were compared.

[0026] FIG. 9A-9D – RT comparisons for cDNA length – **A.** Read depth across each nucleotides position for the same sample (CT=13.89) when using three different reverse transcriptases (SSIII, SSIV, or SSVILO) for cDNA synthesis. **B.** Base pairs of the SARS-CoV-2 genome covered at various depths when using different enzymes for the ARTIC PCR. **C.** Base pairs of the SARS-CoV-2 genome covered at various depths when using either normal ramping speed (3°C/s) for the ARTIC PCR or reduce the ramping (1.5°C/s). **D.** Read depth across each nucleotides position for normal ARTIC PCR vs an alternate hybridization PCR.

[0027] FIG. 10 – Increasing primer concentration 2-fold in regions of low amplicon coverage – Red asterisk indicates amplicons in which the primer pairs were spiked in at 2X the

concentration of the others in the pool. Box plots showing the distribution of absolute sequencing coverage (log10) per amplicon for ARTIC PCR conditions (Normal) and Primer 2x concentrations for 4 representative samples. The boxes are plotted by the Q1, median, and Q3, the whiskers by Q1/Q4, and the outliers by the dots.

[0028] FIG. 11 – Modified Flex outperforms XT in coverage depth and evenness at lower cost – Illumina Nextera XT and modified Illumina Nextera Flex library construction on three samples with varying CTs. Asterisks indicate amplicons with large levels of drop out that were improved with the Nextera Flex. Plotted is the mean sequencing depth (log10) per amplicon.

[0029] FIG. 12A-12C – SDSI+ARTIC over a diverse set of samples is advantageous when compared to metagenomics – **A.** Time-measured maximum clade credibility tree of 772 genomes from Massachusetts, reported in Lemieux et al., 2020. The 89 samples compared for metagenomic and amplicon sequencing are shown with red dots. **B.** Genome coverage for metagenomics versus SDSI+ARTIC amplicon sequencing pipeline (N=81, excluded samples had no detectable CT). All samples downsampled to 975,000 reads. **C.** Gini coefficients grouped by CT (N=70, excluded samples that did not generate assemblies in either one or both methods). Dashed red line represents the median.

[0030] FIG. 13 – SDSI+AmpSeq Protocol. Illustrative workflow for 96 samples through the SDSI+AmpSeq amplicon-sequencing pipeline. A unique, synthetic DNA spike-in (SDSI) will be added to each cDNA sample to allow for contamination tracking and accurate sample identification in analysis. Asterisks indicate additional steps to the standard ARTIC pipeline.

[0031] FIG. 14A-14B – Synthetic DNA oligos spiked into amp-seq reactions designed to flag contamination and sample swaps. A. Schematic of SDSI design. Each oligo contains 140 bp of unique sequence flanked by common primer binding sites. Primers designed to amplify all SDSIs are added to ARTIC primer pools, and a unique SDSI is added to each clinical sample. Identification of multiple SDSIs in the same sample indicates contamination. **B.** Percent of SDSI reads mapping for each of the 96 SDSIs (horizontal axis) were quantified for each of the 96 SDSIs (vertical axis). Any off-diagonal signal would indicate non-specific identification of SDSIs.

[0032] FIG. 15A-15C – SDSI+AmpSeq amplicon coverage and genome concordance. A. Percent of SDSI for SDSI 1-96 in patient samples. **B.** Log of the mean amplicon coverage for the same clinical samples run with and without an SDSI (n=14). A unique SDSI was used in each

sample. The solid blue line represents SDSI+AmpSeq and the solid black line is ARTIC only with no SDSI. Blue and black shading around the solid lines represents the confidence interval. There were no statistical differences (p-value > 0.05) in the mean amplicon coverage for each amplicon between the groups (two-tailed Mann Whitney t-test and multiple comparison two-stage step-up Benjamini, Krieger, and Yekutieli test with FDR set to 5%). **C.** SNV concordance plot between SDSI+AmpSeq and unbiased consensus sequences. Two discordant SNVs, outlined in a red box, were found. Blue dots represent SNVs found in both the unbiased and SDSI+AmpSeq method, whereas black dots indicate the SNV was only present in unbiased.

[0033] FIG. 16A-16C – SDSI+AmpSeq performs well across thousands of samples. A. Sample diversity from two different institutions representing a range of CTs, viral lineages, and states of sample collection from samples where the data was available. **B.** The percent of SDSI reads out of the sum of all SDSI reads that map to the correct spike-in (Left: JAX, N=3,838, Right: Broad, N=2,903). Error bars represent SEM. **C.** The percent of SDSI reads over the total of all sequenced reads for all SARS-CoV-2 positive samples (Left: JAX, N=3,093, Right: Broad, N=2,670). Error bars represent SEM.

[0034] FIG. 17A-17C – SDSI+AmpSeq is used to identify sample swaps and contamination. A. Intentional SDSI contamination experiment (run in duplicate) assessing if different ratios of contamination between SDSI 87 and SDSI 94 (SDSI 87:SDSI 94) were detectable with the SDSI+AmpSeq method. **B.** Examples of experimental errors that were caught using the SDSI+AmpSeq method. **C.** Top: Distance matrix showing pairwise differences between the 17 complete genomes assembled from this sample set. Putative cluster samples are bolded. Bottom: Spike-in counts for each of the 24 samples and water controls in this sequencing batch.

[0035] FIG. 18A-18B – SDSI core sequence in silico validation. Applicants surveyed the core SDSI sequences by BLASTn to identify significant homology. **A.** Significant homology between SDSIs and anything in the NCBI database outside the domain archaea was identified and the SDSI and genus were plotted if identity (y-axis) was greater than 90% and query cover (x-axis) was greater than 50 bps. **B.** For each SDSI, Applicants identified and plotted (see color scale) the maximum alignment score for a significant homology to human (taxid:9606) and viral (taxid:10239) sequences in the NCBI database. Applicants also identified and plotted the alignment score for each pairwise combination of SDSIs.

[0036] FIG. 19A-19E – Spike-in validation. A. RT-PCR for an SDSI in water and a SARS-CoV-2 positive clinical sample background. Mastermix and SDSI specific primers were added to all samples. SARS-CoV-2 positive clinical sample is cDNA generated from a nasopharyngeal (NP) swab. **B.** The distribution of GC content and length for ARTIC v3 primers. **C.** The distribution of GC content of SDSI amplicons. **D.** 100fmol DNA spike-in amplified under standard ARTIC PCR conditions for 40 cycles run on 2.2% agarose gel image with 188bp amplified spike-in (SDSI 1-48). **E.** % SDSI reads over total reads for SDSI (2-48) over a range of SDSI GC% (33%-65.4%) showed no significant read depth bias. Error bars represent 95% CI. Linear regression p-value=0.8160 (Broad, N=2,903).

[0037] FIG. 20A-20B – SDSI Titration. A. In a titration of SDSI 49 across one clinical sample (CT=16) mock diluted to various CTs (CT=20,25,30,35), the number of reads mapping to both SARS-CoV-2 and the SDSI were quantified, and the percentage of each was calculated. SDSI 49 was tested at 600,60,6, and 0.6 copies/uL in each mock diluted sample. **B.** Coverage plots for the SDSI 49 titration experiment.

[0038] FIG. 21A-21B – ARTIC SARS-CoV-2 amplicon sequencing with and without SDSI and normalization. A. In three different CT bins, Applicants showed coverage plots with confidence intervals for multiple samples sequenced with and without SDSIs (CT<27, n=4; CT 27-29, n=6; CT>30, n=4). The solid blue line represents SDSI+AmpSeq and the solid black line is ARTIC only with no SDSI. Blue and black shading around the solid lines represents the confidence interval. There were no significant differences (p-value > 0.05) between the with and without SDSI group for the mean coverage at any of the amplicons (two-tailed Mann Whitney t-test and multiple comparison two-stage step-up Benjamini, Krieger, and Yekutieli test with FDR set to 5%). **B.** The percentage of SDSI reads for 4 different SDSIs was assessed within 4 clinical samples that were run with and without CT normalization of the cDNA prior to the ARTIC PCR.

[0039] FIG. 22A-22E – SDSI+AmpSeq over a diverse set of samples has superior genome recovery and more coverage uniformity at higher CTs. A. Time-measured maximum clade credibility tree of 772 genomes from Massachusetts, reported in Lemieux et al., 2021. The 89 samples compared for metagenomic and amplicon sequencing are shown with red dots. **B.** Percent of assemblies with greater than 98% or **C.** 80% coverage in different CT bins (SDSI+AmpSeq N=81; Unbiased N=81) (downsampled to 975,000 reads). **D.** Genome coverage for unbiased

metagenomic sequencing versus SDSI+AmpSeq amplicon sequencing pipeline (N=81, excluded samples had no detectable CT). All samples downsampled to 975,000 reads. **E.** Gini coefficients grouped by CT (N=70, excluded samples that did not generate assemblies in either one or both methods). Dashed red line represents the median. Error bars represent standard deviation.

[0040] FIG. 23A-23H – Maximizing Genome Recovery and Coverage with SDSI+AmpSeq. A. The percent of the target genome covered at various depths of coverage for four individual samples (CT=13.9, 23.9, 29.6, 33.6), with each undergoing cDNA with three different reverse transcriptases (SSIII, SSIV, or SSVILO). Yellow bar highlights comparison between the reverse transcriptases at a coverage depth of 10X. **B.** Read depth across each nucleotide position for the same sample (CT=13.9) when using these reverse transcriptases. **C.** Base pairs of the SARS-CoV-2 genome covered at various depths when using different enzymes for the ARTIC PCR (n=1). **D.** Amplicons with at least 0.2X of the mean amplicon coverage with the normal ARTIC v3 primer pools or with a modified primer pool with a 2X concentration of 20 poor-performing ARTIC primer pairs. Six individual samples with different CTs were used. **E.** Read depth across each nucleotide position for normal ARTIC PCR vs an alternate hybridization PCR (n=1). **F.** Base pairs of the SARS-CoV-2 genome covered at various depths when using either normal ramping (3°C/s) or reduced ramping (1.5°C/s) speed for the ARTIC PCR (n=1). **G.** Gini coefficients for two mid-high CT samples and four high CT samples when using either 35, 40, or 45 cycles for the ARTIC PCR. Error bars represent standard deviation. **H.** Comparison of Nextera DNA Flex and Nextera XT on the number of SARS-CoV-2 base pairs covered at various depths of coverage for three samples with different CTs.

[0041] FIG. 24 – Increasing primer concentration 2-fold in regions of low amplicon coverage. Data represents 6 individual samples at different CTs.

[0042] FIG. 25 – Unique identification of SDSIs given varying thresholds of SDSI mapping stringency. Applicants considered a range of cutoffs of the percentage of all SDSI-mapped reads mapping to a given SDSI (0.01%-50%, with a step size of 0.01). For an experiment where Applicants sequenced SDSIs without any clinical sample, Applicants calculated, at each cutoff, the number of SDSIs (y-axis) in the set Applicants present (96 total) for which only the expected SDSI had a proportion of mapped reads that exceeded the cutoff (x-axis). Assuming no contamination, all 96 SDSIs should be identified uniquely, i.e. no other SDSI should have a

proportion of mapped reads that exceeds the cutoff. The dotted line at x=5% represents the stringency cutoff that Applicants recommend in practice to detect contamination events.

[0043] FIG. 26 – Deployment of SDSI+AmpSeq to assess for possible nosocomial transmission. Phylogenetic tree showing the location of the putative cluster sequences in the context of a global subset of circulating SARS-CoV-2 diversity. Zoom box shows the 10 highly similar cluster genomes. Sample named on the main tree is the one putative cluster sample that was excluded from the cluster based on genome sequence.

[0044] FIG. 27A-27B – Modification enables addition of spike-ins to RNA. A. A schematic of how to design, produce, and apply synthetic RNA spike-ins (SRSIs). **B.** A limited titration experiment where SRSIs of varying concentrations were added to two clinical samples with low and intermediate SARS-CoV-2 Cts. SRSIs were added to the sample at the RNA stage; the sample with a low CT (20) was then normalized to CT 25 at the cDNA stage, whereas the sample with mid CT (26) was not normalized.

[0045] The figures herein are for illustrative purposes only and are not necessarily drawn to scale.

DETAILED DESCRIPTION OF THE EXAMPLE EMBODIMENTS

General Definitions

[0046] Unless defined otherwise, technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this disclosure pertains. Definitions of common terms and techniques in molecular biology may be found in Molecular Cloning: A Laboratory Manual, 2nd edition (1989) (Sambrook, Fritsch, and Maniatis); Molecular Cloning: A Laboratory Manual, 4th edition (2012) (Green and Sambrook); Current Protocols in Molecular Biology (1987) (F.M. Ausubel et al. eds.); the series Methods in Enzymology (Academic Press, Inc.): PCR 2: A Practical Approach (1995) (M.J. MacPherson, B.D. Hames, and G.R. Taylor eds.): Antibodies, A Laboratory Manual (1988) (Harlow and Lane, eds.): Antibodies A Laboratory Manual, 2nd edition 2013 (E.A. Greenfield ed.); Animal Cell Culture (1987) (R.I. Freshney, ed.); Benjamin Lewin, Genes IX, published by Jones and Bartlet, 2008 (ISBN 0763752223); Kendrew *et al.* (eds.), The Encyclopedia of Molecular Biology, published by Blackwell Science Ltd., 1994 (ISBN 0632021829); Robert A. Meyers (ed.), Molecular Biology

and Biotechnology: a Comprehensive Desk Reference, published by VCH Publishers, Inc., 1995 (ISBN 9780471185710); Singleton *et al.*, Dictionary of Microbiology and Molecular Biology 2nd ed., J. Wiley & Sons (New York, N.Y. 1994), March, Advanced Organic Chemistry Reactions, Mechanisms and Structure 4th ed., John Wiley & Sons (New York, N.Y. 1992); and Marten H. Hofker and Jan van Deursen, Transgenic Mouse Methods and Protocols, 2nd edition (2011).

[0047] As used herein, the singular forms "a", "an", and "the" include both singular and plural referents unless the context clearly dictates otherwise.

[0048] The term "optional" or "optionally" means that the subsequent described event, circumstance or substituent may or may not occur, and that the description includes instances where the event or circumstance occurs and instances where it does not.

[0049] The recitation of numerical ranges by endpoints includes all numbers and fractions subsumed within the respective ranges, as well as the recited endpoints.

[0050] The terms "about" or "approximately" as used herein when referring to a measurable value such as a parameter, an amount, a temporal duration, and the like, are meant to encompass variations of and from the specified value, such as variations of +/-10% or less, +/-5% or less, +/-1% or less, and +/-0.1% or less of and from the specified value, insofar such variations are appropriate to perform in the disclosed invention. It is to be understood that the value to which the modifier "about" or "approximately" refers is itself also specifically, and preferably, disclosed.

[0051] As used herein, a "biological sample" may contain whole cells and/or live cells and/or cell debris. The biological sample may contain (or be derived from) a "bodily fluid". The present invention encompasses embodiments wherein the bodily fluid is selected from amniotic fluid, aqueous humour, vitreous humour, bile, blood serum, breast milk, cerebrospinal fluid, cerumen (earwax), chyle, chyme, endolymph, perilymph, exudates, feces, female ejaculate, gastric acid, gastric juice, lymph, mucus (including nasal drainage and phlegm), pericardial fluid, peritoneal fluid, pleural fluid, pus, rheum, saliva, sebum (skin oil), semen, sputum, synovial fluid, sweat, tears, urine, vaginal secretion, vomit and mixtures of one or more thereof. Biological samples include cell cultures, bodily fluids, cell cultures from bodily fluids. Bodily fluids may be obtained from a mammal organism, for example by puncture, or other collecting or sampling procedures.

[0052] The terms "subject," "individual," and "patient" are used interchangeably herein to refer to a vertebrate, preferably a mammal, more preferably a human. Mammals include, but are

not limited to, murines, simians, humans, farm animals, sport animals, and pets. Tissues, cells and their progeny of a biological entity obtained in vivo or cultured in vitro are also encompassed.

[0053] Various embodiments are described hereinafter. It should be noted that the specific embodiments are not intended as an exhaustive description or as a limitation to the broader aspects discussed herein. One aspect described in conjunction with a particular embodiment is not necessarily limited to that embodiment and can be practiced with any other embodiment(s). Reference throughout this specification to "one embodiment", "an embodiment," "an example embodiment," means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, appearances of the phrases "in one embodiment," "in an embodiment," or "an example embodiment" in various places throughout this specification are not necessarily all referring to the same embodiment, but may. Furthermore, the particular features, structures or characteristics may be combined in any suitable manner, as would be apparent to a person skilled in the art from this disclosure, in one or more embodiments. Furthermore, while some embodiments described herein include some but not other features included in other embodiments, combinations of features of different embodiments are meant to be within the scope of the invention. For example, in the appended claims, any of the claimed embodiments can be used in any combination.

[0054] All publications, published patent documents, and patent applications cited herein are hereby incorporated by reference to the same extent as though each individual publication, published patent document, or patent application was specifically and individually indicated as being incorporated by reference.

OVERVIEW

[0055] Embodiments disclosed herein provide a method of detecting and preventing contamination during genome profiling using synthetic DNA spike-ins (SDSIs). Embodiments disclosed herein also provide methods to track sample contamination by implementing synthetic DNA spike-ins (SDSIs) for sample verification. Embodiments disclosed herein also provide synthetic DNA spike-ins (SDSIs) and methods for producing synthetic DNA spike-ins (SDSIs). The global spread and continued evolution of SARS-CoV-2 has driven an unprecedented surge in viral genomic surveillance. Amplicon-based sequencing methods provide a sensitive, low-cost and rapid approach but suffer a high potential for contamination, which can undermine laboratory

processes and results. This challenge will only increase with expanding global production of sequences by diverse laboratories for epidemiological and clinical interpretation, as well in genomic surveillance in future outbreaks. Applicants present SDSI+AmpSeq, an approach which uses synthetic DNA spike-ins (SDSIs) to track samples and detect inter-sample contamination through the sequencing workflow. Applying SDSIs to the ARTIC Consortium's amplicon design, Applicants demonstrated their utility and efficiency in a real-time investigation of a suspected hospital cluster of SARS-CoV-2 cases and across thousands of diagnostic samples at multiple laboratories. Applicants established that SDSI+AmpSeq provides increased confidence in genomic data by detecting and in some cases correcting for relatively common, yet previously unobserved modes of error without impacting genome recovery.

[0056] The methods described herein add a unique SDSI to each sample (e.g., cDNA) before performing a sequence amplification process during which the samples and SDSIs are amplified in the same reactions. This procedure can be repeated in parallel for each sample undergoing analysis. After the samples have been amplified, the presence of the SDSI is measured. If the SDSI introduced before amplification is the only SDSI present, then the sample is determined to be uncontaminated. However, the presence of any other SDSI immediately reveals contamination of the sample. This method provides a reliable safety measure for pathogen-genome studies and the resulting therapeutic and preventative medicine.

Synthetic DNA Spike-In (SDSI)

[0057] In one aspect, the present invention is directed to SDSI's and uses thereof. An example SDSI comprises, in a 5' to 3' direction, a 5' primer binding sequence, a core sequence, and a 3' primer binding sequence. In one example embodiment, spike-ins comprise sequences derived from a rare organism. A rare organism is a species that is limited in number or geographic occurrence relative to the distribution and abundance of other species making up the pool of interest.(Raphael, M. *et al.*, Conservation of Rare or Little-Known Species: Biological, Social, and Economic Considerations. Bibliovault OAI Repository (2007) the University of Chicago Press) In some embodiments, the rare organism is an archaea. In some embodiments, the archaea is thermophilic. A thermophilic archaea may exist in environments with temperatures greater than 50°C. In certain embodiments, the present invention includes spike-ins. In certain embodiments, a spike-in comprises a DNA sequence that is not from the target organism. In certain embodiments, a spike

in is an RNA molecule that can be added to a sample comprising pathogen RNA. In certain embodiments, the RNA is converted to cDNA concurrently with pathogen RNA. The RNA spike in cDNA can then be amplified with pathogen cDNA using pathogen specific primers and spike-in specific primers.

[0058] In certain embodiments, a spike-in sequence is compared to the target organism and the host for the target organism to limit homology. Limited homology can be determined using a BLAST search of all SDSIs. In one example embodiment, a permissive BLAST search is used (*e.g.*, blastn; 5000 max targets; E=10; ws=11; no mask for low-complexity). Results may be filtered by species of interest, *e.g. homo sapiens*. In one example embodiment, results can be filtered for a pathogen of interest (e.g., SARS-CoV-2). The query coverage and sequence identity may each be set for 35-100%, preferably, 50-100%, and sequences having no significant hits can be selected for use as a spike-in. In certain embodiments, a spike-in set comprises different DNA sequences that can be easily distinguished using sequencing.

[0059] In certain embodiments, the GC content of the spike-ins promote similar amplification rates across pathogen targets and the different SDSIs in our set. In one example embodiment, a spike-in comprises a similar GC content as the target organism. In another example embodiment, the GC content of the primer may range from 30% – 80%. (Buck, G.A. *et al.*, Design Strategies and Performance of Custom DNA Sequencing Primers, BioTechniques (1999) 27:3, 528-536). In another example embodiment, the GC content of the primer may range from or between 30%-40% nucleotides, or between 40%-50% nucleotides, or between 50%-60% nucleotides, or between 60%-70% nucleotides, or between 70%-80% nucleotides. In general GC content extremes are avoided. For example, sequences may have a median of 50% GC content, preferably, between 35-65%. In another example embodiment, the GC content of the primer may range from or between 40%-70%, or between 30%-50% nucleotides, or between 30%-60% nucleotides, or between 30%-70% nucleotides.

Core Sequence

[0060] Each SDSI in the set is differentiated by its core sequences. The SDSI cores are designed to minimize self-hybridization and cross-hybridization with others nucleic acids in a given sample. Accordingly, core sequences are selected based on the type of target sequence to be amplified and the type of sample the target sequence is to be derived from. For example, in the

context of detecting a pathogen in a human sample, core sequence should be selected with minimal homology to the target pathogen, other common microbes and non-target pathogens that might be present in the sample, and human sequences as well. In certain example embodiments, the core sequence has a homology of less than about 65%, or less than 64%, or less than 63%, or less than 62%, or less than 61%, or less than 60%, or less than 59%, or less than 58%, or less than 57%, or less than 56%, or less than 55%, or less than 54%, or less than 53%, or less than 52%, or less than 51%, or less than 50%, or less than 49%, or less than 48%, or less than 47%, or less than 46%, or less than 45%, or less than 44%, or less than 43%, or less than 42%, or less than 41%, or less than 40%, or less than 35%, or less than 30%, or less than 25%, or less than 20%, or less than 15%, or less than 10%, or less than 5%, or less than 1%.

[0061] The core sequence may vary in length between 50-5,000 nucleotides, or between 50-nucleotides, or between 50-4,500 nucleotides, or between 50-4,000 nucleotides, or between 50-4,000 nucleotides, or between 50-3,500 nucleotides, or between 50-3,000 nucleotides, or between 50-2,500 nucleotides, or between 50-2,000 nucleotides, or between 50-1,500 nucleotides, or between 50-1,000 nucleotides, or between 50-500 nucleotides.

[0062] The core sequence may vary in length between 50-60 nucleotides, or between 50-70 nucleotides, or between 50-80 nucleotides, or between 50-90 nucleotides, or between 50-100 nucleotides, or between 50-110 nucleotides, or between 50-120 nucleotides, or between 50-130 nucleotides, or between 50-140 nucleotides, or between 50-150 nucleotides, or between 50-160 nucleotides, or between 50-170 nucleotides, or between 50-180 nucleotides, or between 50-190 nucleotides, or between 50-200 nucleotides, or between 50-210 nucleotides, or between 50-220 nucleotides, or between 50-230 nucleotides, or between 50-240 nucleotides, or between 50-250 nucleotides, or between 50-260 nucleotides, or between 50-270 nucleotides, or between 50-280 nucleotides, or between 50-290 nucleotides, or between 50-300 nucleotides, or between 50-310 nucleotides, or between 50-320 nucleotides, or between 50-330 nucleotides, or between 50-340 nucleotides, or between 50-350 nucleotides, or between 50-360 nucleotides, or between 50-370 nucleotides, or between 50-380 nucleotides, or between 50-390 nucleotides, or between 50-400 nucleotides, or between 50-410 nucleotides, or between 50-420 nucleotides, or between 50-430 nucleotides, or between 50-440 nucleotides, or between 50-450 nucleotides, or between 50-460 nucleotides, or between 50-470 nucleotides, or between 50-480 nucleotides, or between 50-490

nucleotides, or between 50-500 nucleotides, or between 50-510 nucleotides, or between 50-520 nucleotides, or between 50-530 nucleotides, or between 50-540 nucleotides, or between 50-550 nucleotides, or between 50-560 nucleotides, or between 50-570 nucleotides, or between 50-580 nucleotides, or between 50-590 nucleotides, or between 50-600 nucleotides, or between 50-610 nucleotides, or between 50-620 nucleotides, or between 50-630 nucleotides, or between 50-640 nucleotides, or between 50-650 nucleotides, or between 50-660 nucleotides, or between 50-670 nucleotides, or between 50-680 nucleotides, or between 50-690 nucleotides, or between 50-700 nucleotides, or between 50-710 nucleotides, or between 50-720 nucleotides, or between 50-730 nucleotides, or between 50-740 nucleotides, or between 50-750 nucleotides, or between 50-760 nucleotides, or between 50-770 nucleotides, or between 50-780 nucleotides, or between 50-790 nucleotides, or between 50-800 nucleotides, or between 50-810 nucleotides, or between 50-820 nucleotides, or between 50-830 nucleotides, or between 50-840 nucleotides, or between 50-850 nucleotides, or between 50-860 nucleotides, or between 50-870 nucleotides, or between 50-880 nucleotides, or between 50-890 nucleotides, or between 50-900 nucleotides, or between 50-910 nucleotides, or between 50-920 nucleotides, or between 50-930 nucleotides, or between 50-940 nucleotides, or between 50-950 nucleotides, or between 50-960 nucleotides, or between 50-970 nucleotides, or between 50-980 nucleotides, or between 50-990 nucleotides, or between 50-1000 nucleotides, or between 50-1010 nucleotides.

[0063] The core sequence may vary in length between 100-5,000 nucleotides, or between 1,000-5,000 nucleotides, or between 2,000-5,000 nucleotides, or between 3,000-5,000 nucleotides, or between 4,000-5,000 nucleotides.

[0064] The core sequence may vary in length between 75-150 nucleotides, or between 100-150 nucleotides, or between 100-200 nucleotides, or between 100-300 nucleotides, or between 150-200, or between 150-250 nucleotides.

[0065] The homology to a target sequence or non-target sequence in the sample across the size of a given core sequence may be less than 1 nucleotide, or may be less than 2 nucleotides, or may be less than 3 nucleotides, or may be less than 4 nucleotides, or may be less than 5 nucleotides, or may be less than 6 nucleotides, or may be less than 7 nucleotides, or may be less than 8 nucleotides, or may be less than 9 nucleotides, or may be less than 10 nucleotides, or may be less than 11 nucleotides, or may be less than 12 nucleotides, or may be less than 13 nucleotides, or may be less

than 14 nucleotides, or may be less than 15 nucleotides, or may be less than 16 nucleotides, or may be less than 17 nucleotides, or may be less than 18 nucleotides, or may be less than 19 nucleotides, or may be less than 20 nucleotides, or may be less than 21 nucleotides, or may be less than 22 nucleotides, or may be less than 23 nucleotides, or may be less than 24 nucleotides, or may be less than 25 nucleotides.

[0066] The homology to a target sequence or non-target sequence in the sample across the size of a given core sequence may vary in length between 1-5 nucleotides, or between 1-10 nucleotides, or between 1-15 nucleotides, or between 1-20 nucleotides, or between 1-25 nucleotides, or between 1-5 nucleotides, or between 5-10 nucleotides, or between 10-15 nucleotides, or between 15-20 nucleotides, or between 20-25 nucleotides, or between 1-10 nucleotides, or between 10-20 nucleotides, or between 20-30 nucleotides.

[0067] These SDSIs can be implemented in a wide range of genome profiling applications including, but not limited to, investigations of SARS-CoV-2 epidemiology and emerging viral variants. Exemplary SDSIs are provided in Table 1.

[0068] **Table 1.** Sequences of 96 unique SDSIs. The unique core of each SDSIs is 140 bps long (SEQ ID NOS: 1-96 and 193-291). The unique SDSIs including the priming regions (SEQ ID NOS: 97-192 and 292-390). Alternative sequences are also included. SEQ ID NOS: 16, 57, 66, 112, 153, and 162 can be, in the alternative, substituted with 289, 290, 291, 388, 389, and 390 respectively. Sequences for forward and reverse primers for amplifying the SDISs (SEQ ID NOS: 391 and 392 respectively).

Table 1

SEQ ID	Core Sequences
1	CAATTGCTCCCTCGTATCCCTTGTACATTATCTCAGCTCCGCTTAATGATATTAATTTTACCTTGA GTGTTTTTGCTAAAGCCTTTGCCATCATCGTTTTACCTACTCCAGGTGGCCCGTAAAGCAACACA GCTTTGGCA
2	TTCTCCAAAACCTACCCAGTTCTCCGAGGAACCTCTTAGCATCTGTTAAATCGTTATTAGTATTA GCTTCCACCATCTCAAGTTCCTTTAAGGCGTTACTCACACTCTTCTTACCTATCTTTTAGAGAACC ACTCGTCAG
3	GTTATCAAAGCCCTTAAAGAGTGGTAGGGGCAAAAGTCTGAAGCGTCCTTACTTAACTGGAGTA TCTGAGATGGCCTTAATCCGCTTAGGTCTTTAATTTTATCCCTTAATGAACATTCCCTGCACTCTA TGTCTTCGGG
4	GAGATGTAGCAGACGGGCTAAGAGTTTCAAACCCTCTAAGGATCACTACAAACAAGAGAGAGA GACAATCCTCTCTTTTGTCTTGTCATTGTGTTTCAAACCCTCTAAGGATCACTACAAACATCTTTA ACATAGATACC

5	GACCGGACGTTGTGATCACGGGTACCTTGATCTGGTACTCAAAGGTTTGCCCCCGTGAAGTCTG GTACATGGCTAGACACGTCACTCCATTCGAGGGACATTCGAAGTTAGAGAAGGGCAGAGCGAT ACATCAGATATAT
6	GTCTTTTCTCTACTAATTCTCCTCACGAGATCTCTAAACATTCTTGCTGAAAGAGGATCCAAACC TAATGTAGGTTCGTCAAGCAATAAAATTGGAGGATCAGTTATTAATGCTCTTGCTAAGGCTAGT TTCCTCTGCAT
7	GATTTTGCCATCATTAAAAACAACAATTTGATCACCCATAGTCATAGCTTCTAATTGATCGTGAG TTACATAAATACTTGTGGTGTTTAACATACGGTGAATATTTACAATTTCTCTTCGCATGTTTTCTC TTAGTTTAG
8	GTATCTTTCAATTCTCGAAAGAAAAGGTTACAAGTCTCATAGATTTATTCCTCTTCACTGTTGTA CGTTGGCAGCTAGAGAGAGTTTAGATTATGAGAAAATTAAGAGAATATATGAGGATTCGTTTTC TTGGTTTAAGT
9	CTAATTGATTTTCCTGTACCATGTGGTAAAACAACGCTACCTCTTAATTGTTGATCTGCTTTTCTA GTATCAAGATTTAATCTAAAAGCTAAATCAACTGAAGCATCAAATTTTGTATAAGAAGTTTTTTT CACTAATTC
10	TCGGTTTTCCCGTGAACTAATAAACACCTACTGGAGCCAAGAACGGGTCAGAATTGATGGAATA AACGTTGCGGAGAATGAAATTAATTTGTACATCAGAGACATTGATGACAACGGTGACCCTATAC AGTCAACTATAC
11	CTTAATGGAAAGTATGCTTTAGATACCTTCTGGAACGCTATCTCACTTGGCGGGAATTCAGATAT GGAGAGTAAATTAAGGGATCTGGAAGTAAAGTTAATGTCGTTAATCTATTTAAATGAGTCACCA TTAAAATCACC
12	CATAATATGTTAGAGGTAGAATTTCTTTGTGATAGAATATTATTGATGAATGATGGAAGAGAAT TAGCATTAGGAAAACCTAAGGAACTGGTAAAGGATACAGAATCTAAGAATCTTGAAGAGGTTT TCCTTAAACTTGT
13	CCTTACTTCATCTCTCAAGATAAGGGTAATAAGTTCACTTCAAATATCTGGTCTTATCGCAAGTT GATTGAGGCTATAGTGTATAAGCTCTATGAGTATGGTATAAACGTGTTCCTCGTTGTAGAGTAT AACACTTCACG
14	AGTCTAGGTTTTAATTCTTCAACTGCTTCAAATACTAGCTTACTGTAGTTATCTGCCCTCATGTTA GGATATATATCTGGAATATAAGGAGGTTGATGAGTTATAAGAAGTGGATGAAATTGTTGTCACA CACTCCCCTA
15	CTACCTCTTCGGCCTTGTACCAACGTACCCCTGATACAAGTTCCAAGCAGAGATGGAAAACTCG AAGATGGTATCACCCAAGATGAGATACGATATCAATGAAGGCGAGCCTAGGTACAAGTAAAGG GATACCACGAGAG
16	CTCGTAAGCGTTTCCTACCCTCGAGAGGGCCATCCTGGTGGTGAGGAAGTCGTCGAAGTGGGCT AAGTAAAAAGCGAAGATCTCGACCCACAATTACCTCCTCCTGTACACCAGGAATACCCCTATCA GGATAGAGATAC
17	GCGCGTCCGGGTCGCGGCCGGGGACGACCGTCTTGACGAAGTCGGTCGACCCCTCGTCGGTCGA GATGGTCGTCACCTCGGTGTCGAGGCCGTACGTTTCGAGCGCGTCGCGTACCAGTTCGCCGTCC GCGTCGGGACGG
18	CATGTACTCGTTCCAGAAGGTGAGTTCGCTCCCCTCGATTTCGACCTCGCCCACGTCGAAGCCGC CGGTCGTTTCGAGCGCGAACGACTCGACGGGACCGACGAGCGAAACTTCGCCGCCGAGCACGT CGGCGACGCGTT
19	CTCGATGCGCTCGGGCTTGTAGGACTCCCCGAGGGCGTCCTTGTTGGTGAAGACGTTTTGTTTTC GCTCGAACCGGCGCATTAGCGTCGGTCCGTTGTAGCGTCCCCTTATTTAAAACCCCGATTTCATC TGATTCATGT
20	TCACGGTCCGCGACGTGAATCGGGCGTTCCAGTCGGCGTTCGGCTACGACGCCGACGACGTGGT CGGAAGCGACCTCCTCGGGCGAATCGTGCCCCCGGTGCCGGACCCGGACCCGGTGCCGGAACC GGGGGACGACGAG
21	GCGTCCGCGAGTTCATCCTGAACGTCGTCCCGCTGTCGCCCGGCGAGGAGCGCGGGGCGGGCTA CGCCATCTACACCGACATCACGGAGCGGAAGACCCGCGAAAGCGAGCTAGAGCGACAGAACGA GCGATTGGAGGAG

22	GCGAGACCGGCGACGAGGTGCGCTTCGACACCGCCGAGCGGGCGCTCGAACAGATGGAGGAAC TCATCGACGACCTGCTGTCGCTCGCCCGTCGCGGCCAACTGGTCGACGAGACGGAGCGCGTCGA CCTCGGGGCGGTC
23	ACGAACTCGTCGGTGAACATCTCGTCTTCCGGGGAGCCCGCCGCTCATGGCCTGCCCCCGCCGT AAGCTGCTGCATAAACCCGCTCCAAAATATACGGATCATTCACCCCTTGGAATCGCTCAATCAG ATCAATGTACAC
24	TGCGTACATTCCCCCTAAGCGGCTCCCAATATACAGACGCCGGTTAACGACAGCTGGCGACCCT GTGATCTCAGTACCGGTGTCGAATGACCACATCAGCTTGCCTGTCCGTGCATGGAGTTCGTATA CGTACCCGTCGT
25	AGATAGATGAGCCGATCAGAGATCGCTGGTGAGTTGGTAATTGTCCCGACATAGACACGCCAA CGTTCTGTTCCATCTGCTGCGTCGTAGGTCGCGAGATACGGCCAGCCACCAACATACACAATCC CATCGACGAGGAC
26	ATACACCACCCCATCAGCAACAACTGAATCATGATTAAGTATCGCACCAGCATCGTAGCGCCAG CGTTCACTGCCAGTGGTGCTATCGAATGCATAGAAGATATGCTCCTAATCGCCAATATCAGTAC TTCACAAAGCCG
27	TCGACGAGGAGAGGGGCGAGTACATCTGCACGCTTACGGGAGAGGTAGTTGAGGAGACGGTTA TAGATACAGGGCCCGAATGGAGGGCTTACACACCTGAGGAGAGGACCCGCAGAAGCCGCGTGG GCAGCCCGCTTACC
28	AGTCGATGGCTGCGGCAGCTGTCTATGCTGCCTGCCGTATACGCGGCATACCCAGGAGTATAGA CGACATAGCGGAGGTCGTGAAGGGTGGCCGTAAGGAGGTTGCCCGCTGCTACCGCCTCATAGTC CGCGAGCTGAAG
29	GTGGAGTCTTTTGTCACACCGCAGAGGCGTAGCGCTGCAGAGCAGGAGCCCAAGCCTACTGCCA ACATAGAGAACATAGTGGCTACAGTATCCCTCGACCAGACTCTAGACCTGAACCTCATAGAGAG GAGCATACTGAC
30	CGTCGCCTGGGTTAAGAGGATGTTCGGCCTCTCCAAGGCGGGTCACGGAGGCACGCTGGACCCG AAGGTCACCGGCGTCCTCCCCGTAGCCCTGGAGGAAGCAACCAAGGTCATAGGCCTGGTGGTG CACACGAGCAAGG
31	CGTGGGCGAGATCTACCAGAGGCCGCCGCTCCGCAGCAGTGTTAAGAGAAGCCTCCGCGTCAA GAGGATATACGAGATAGAGCTGCTGGAGTACAACGGCAGGTACGCGCTCATGAGGGTGCTCTG CGAGGCCGGCACAT
32	CGCTGGAAGAACGAGGGCAAGGAGGACCTGCTGCGGAGCTACATCAAGCCCGTCGAGTACGCC GTGAGCCACCTGCCCAAGATAGTTATACGCGATACCGCGGTGGACGCCATAGCCCATGGCGCG AACCTCGCGGTGCC
33	GGGAGACCCCAAGGTGACCGGCGTCCTACCAGTGGGGCTCGCCAACAGCACCAAGGTCATTGG TAATGTTATACATAGTGTTAAAGAATACGTGATGGTTATACAGCTCCACGGCGATGTAGCCGAG CAGGATTTAAGAA
34	TAGAGGGAAAGACTGTAGCTTTCATTCCTAGGCACGGAAAGAGACACAGAATACCTCCACATA AGATAAATTATAGAGCTAATATATGGGCATTAAAAGAACTAGGAGTGAAATGGGTCATCTCAG TTTCTGCCGTAGGA
35	TGAGGGAGCTCAGGAGGACTCGCACGGGGCCCTACAGGGAGGATGAGACACTTGTAAGGCTCC AGGACGTCAGCGAGGCCCTGCTCCTGTGGAGGAGCAACGGGGATGAGAGGTATCTTAGACGCA TCGTGCTACCCGTT
36	GAAACATCTATCGCCCACCTCCCGAAGATAATGATCTTGGATACAGCTGTCGACGCCATAGCAC ATGGTGCCAACCTGGCTGCCCCAGGCGTCGCCAGGTTAACCAGGAACATCGCGAAGGGTAGTA CCGTAGCGATCCT
37	TCGCTATCCCCGTGTACAGCATGGTGGGGGTGCCGATGCCCGGGTAGAACTTGGTGACGCTCTC CAGCTTCTCGAGGACGGTTTCCTTGGGGAGGCTCGCGGTGTCCACGAGGGTTATCGCGTCCTCG GCGCCGTCGCCG
38	CGAGGACGCGAAGAGCGCGGTGGATGTGGACGCGCCGCCGCACACGTAGCCGTCGAGGTAGCG CGGAACCATCGGCGACATCAGCCCCACGACGCGACCCGAGGCGTTGCCGAGGATCACGTCGAG CGTCACGCGCGGCA

39	CTCGACACCGTGCCGTTGCCCTCCTCTAAGTAGTCGGAAAGCCTCATCCGCGACTCCAGCTTCGC CACCGGCTCCTCGAGCAGGAGGAGGACGCGGTTGATGCGGTAGGACGCACTGCCCGCCTCCAG CACCGCGCCGTC
40	TCTATGGTGTAGAACGGGTCGTTGCGGAGCCAGCCTGGCGGCACGTACCGGTCGTCCGCTATCG CCAGCGATCTCTCGAAGAGGTCGAGGTAGGCGGACGCGTTGGCGAACGCCCCGTGTATCACGA CGTCTATCCCGCC
41	GTATAGGTTTCAGGTATTGATAATGCATAGGAGGTTTTTAAAACCTTGAGCCGCATAGTCTTCTG GATGGGCGAGAGACATGGTTAAGTATAAGTGCGGCAGGTGCGGATACGTCTTCGACGACGAGG AGATGAAGAGGA
42	CCTACGCCGGGTGCGTAGGAGGGCTCGAGTACATCCATGTCTATACTGATGTATGTTTTACCCA GGTCGCCTAGTGCCAGGGGTCCCTTTAACGCTTCCAGGATAGAGTACACGGTGACGTCTCTAGT CTTCTTCAAGAA
43	CTACTAGCGTGTCAACGGAGCTCTTCAACGCCTTTACTATTGGATAGGTTATAAGGTGCTCGCCT CCGAGGAATCCCAGGAGCATGCCGGGATACTCGTCTACAACGCCTTTCACCACGTCACCTATGA TTCTTAAAGAG
44	CATAGGTGACATGGGGTTTCCCATTGACTCTATAAAGCCGTATCCTTTAAGCGGAGTGCAATTG GTCTACGCTTTGCTTAACAACAGGTATTTCCTACCGGGTAGAGAGGGCTCGCTCATAGCTTTAG GTAGCGTGACGG
45	GGTATCTCACCGCTTGTCACCATAGTATCCCTCAGGTACTCCAGTATTCTTGAGAGAAACGCACC TAAGCCGGATCTCAGGTTTGAATCCATAAGAACTATGAGTGAAGCGGGATTGAAGCCCCTGCTG TTTCTAAGACC
46	TAAGGGAGATAGAGAAACGCATCAAAATACCCTTGGGGAAACTGCGTGCAGGGGTTCAATATG GAGTAGAGGTCTCAGACATAAAGGAGAAGATAGCTGCTTACGCTAGGAGGAAGGGGCTTAAAT ACTTCCCATCGGCA
47	TGTGAACCTCGTGCCCGGCTCTAAGTCGTGAGGGCTTGCAACATAGGTGGGGAGGAACCCGAG CAACGGGTAAGAAGACAGGATAAGCGGTATCGCTATGAAGAGGGCTGAGAAAAGGACATATA CTCCTGAGCCCGTCC
48	CGAACATGCCTTCCCCGTCTATATAGACCCAGTAGAGTTTAAAAACTTAACCAGAGACGGCTTG TGAGCCGGATCTCTCCCCCGCTAGGCCCTGGATTGGGCTCGCTCCTCCTGGGACCCCGGCCTCCA CATGCTCGGGA
49	CCTGAAGGGCTCGGCTACCCTGAAGACGGGCTTCTGCGCGACCGCCGCGTACTCCGCCGTGGAG CGGTAGAAGAGCGAGGCTGTCTCCGTGAGCCTGACCATTCCGTACAGGGCGACTGCGACGAGC ACTATGACTGCGA
50	GTCAAGGTGCTGATGCCGAAGGCGACTTTCGACACCGACGATGCCGCCGACGCCCTGGCCATTG CCATCTGCCACGCGCATCACCGGCACAGTGTTGCCTATAGGATGGCGCTGGCCGGATAAGTTTG TTCTTGACCTGT
51	TCTCGGTTCGGCAATAAGTAATACCAACGAGGTATTACCATGCGCGTGACCAGCAAAGGCCAA GTGACGATCCCAAAGGAGATACGGGATCATTTGGGGATTGGGCCGGGCTCCGAGGTGGAGTTC GTGCCCACAGACGA
52	CTCGATCATATGGCCGGCACGTTGGACTTGGGAGGCATGACAACGGACGAGTATATGGAGTGG CTGAGGGGTCCACGTGAAGATCTCGACATTGATTGACACAAATGTCCTGATCGATGTTTGGGGT CCTGCCGGACAGG
53	CAGGTGTATTTTACACACCTGGACAGCCAGCATATGATGCTAGCACTCGGTGTCCCCTTATCAC GGTTTCCCGCATTGTAAAGTTTTCGCGCCTGCTGCGCCCCGTAGGGCCTGGATTCATGTCTCAGA ATCCATCTCCG
54	CTGGAGCCTGTTAGTTGTTACAGGTTCACCGGTTGTCGGAGTATTCAGATCATTGAGCCAGCAG TTGATGGCTGCCTGTAGTTCACTGGTTGTGATGTAAGCTGCTCCATCGGAATCAACATCGTTCCA TGGGTTCCAGT
55	ACGGTCTTGCTTTCTCCTGAATCCATTTCACCTGTCCAGACCCATTCATAGCGGTTAGCTTCACT GAGGTTCTGCTTGAAGACACCGTCATCATTGTTAGATGAGGTTATTGTCCAGCCGGCAGGAATG ACTTCTTCGAA

56	GTCAGCAGCTCTTCATAGAAGTTCTGGTTTGCAATATCCCTCTGGGCAATGACAGGGTAGTCGA CTTCGTTTGCAGTCAGGTGGACTGCATACAGGGACTTGCTGATGTCCGGGGTATATCCACTGTG AGGAGCATAGTA
57	ACCCGTCAGTCGTGACGTCCTCCGCTCCTCCTATGCTATCTCCACACACCCACTCACGTTCTTGC TTCTTTACTACACCCTCTTTATTCAGCTCTTCGAGAACATTATTAATGTGACCCTTAGAGATATAT TCATTATAC
58	GTGCCTCCTCAAGCGACTGCTTAAACCCAATTACATCTGATTTATCCTTTATTTTAGGGCCTATA GAATCTATGAATAATTCGGCGATTCTTATTATTTCTAAAACCAATTCGTCTGTTTTGAGTGGTGT GCCTTCTTCA
59	CATCCCATGCATTTTCATAATAATCGGAATTCAAATCCTCTATATTGAATTTTATCTTAACATTTG ACATAATCATTTTCTCCTTACAGAAGAGATCCAGCTAAGCTTACTCATAAATGGTAGTACCATG CCAATATTGG
60	CGTAGCCCGCACCTTCCTCTGGTTTAGCACCAGCGGTCCCCACAGAGTACCCATCATCCCGAAG GATATGCTGGCAACAGTGGGCACGGGTCTCGCTCGTTGCCTGACTTAACAGGATGCTTCACAGT ACGAACTGACGA
61	CCTGATAGGCCGCAGATTCATCCTAAGGCGCCGGAGCTTTTGACCACAGAACATTCCAGTATCT ATGGTATATCTGGAATTATCACCAGTTTCCCGGTGTTATGCCAGACCTTAGGGCAGATTATCCAC GTGTTACTGAG
62	TGTTTGGCTTGATACTAATAAAAGCACAGCTAAAATGAAAATAAGCCGATATTTGTGATTCATG CAACTCACCCTTTTCTACATAAACAAAATACTAACCCGAAAACCGAAATTGAAATTAATGCAGA GAAACCAGGTGA
63	TTAACGGCACCAACAGTTATTATATTTTTAGCAGTCCCGGGTGAAGTAATTATGGAATAGTTGTT AGAATTACTGTTCTTATTACCAGCTGATTTGAAAGCAATTATACCTGCATCACGAATTGCAGCAT CATAATATTC
64	GGCTCAGACGACTGAAAAAGCAACGATTGGAATAATAGGGGGTTCTGGGCTCTATGATCCTGGT ATTTTGACTAACAGCAGAGAAATAAAAGTATATACACCCTATGGGGAACCTAGCGATTTGATAA CGATAGGTAACA
65	CGCAGAACAGGTTCCTTCTATTGGATATTCATCTTCGGCTGCAGTTGCAGGAAGAGTAAGGATA TATACTACGGTCTTGCTTTCTCCTGAATCCATTTCACCTGTCCAGACCCATTCATAGCGGTTAGC TTCACTGAGGT
66	CTTCCTCCACGCATTTGTTGTGGTGCTGATGGCGTATTCTCTGGAATTTGGGATGATTCTGGAAA TCCATCCTCAGACACTTCAGATATTTTAGTCTTACTTCCAGCGTTTAATTGAACCTTACCTTTAAA AGCAGTAGT
67	GAAACTTACCTTATCAGTGTCATTAAGCATATTGCTTCCAAGACCCATTGAAGCACTTACATCGT TGATACACAGGTGCCAGGAATAGTATTCCTCAGTCTCACTATAATCCTCGTTGGTGTAGCCTTCA AGAGAGTCAA
68	GTTTAAGCAATTCTTCGGATGAAAGATGGCGCTCTATAGGAATTTGTTCTGGTCTAGCCATAAG GCATTATTTGTACTTAATTAGTAATAAATGTTTAGTTAATGACTATAAATCTGCAATTGGAGTCT CAAATTTTCAA
69	AACATGAAGGATGTGTGTAAGAGGAAACGTTATTAACAGACGTAATCAGGAGGATAGTTATGC CCTAAAAACAGCAGAGTTAAGGTTTAAAAATAAGATAAGAACTCAGTTGAGGTTTATCCATTAA TCCCATTAATCCT
70	ACTTTCTAAAAGCGCTTGGAGCACGTATCAGGTCAAGTCTTTCAACCTTAAATGCTGCCAGTGC CGTAAGTAGTGCAGTTATGTTGCTTATTGAAACAAACAACTTAGCCCACTTATTACCTCTTGTCA GTGTTTTTGAT
71	GTATCCGCTGATATATCCTGGGGATATAGATCGCTCTGAAATGGTTACATCTATCGGTTTTAAGG ACAGTTCCAACACTATTGGACCTTGCAGCTATGACAGGAATAATCTGTTTATCGAGCACAGTTG AATTTGACCTA
72	TCAATACCTAATTCTTTCCTTAGAGTGCTATTTTGATTGAATTCCCTCAGGAAAGATTCAAAATT TAAGTAGCCGAGCTTACATCTTGAAATTTCCATCTTTATTATGTTGCTCAGGCTTAATGCTTCTA AGTATGGGTT

73	AGATATCCTTTGAAATTCTCGTAATTGCTGAAGGCCACTACTTCATCAGGTCTGATGCAATCTTT AATCTGAACATTGCTTTCTGAGGTCTTAGGAATAATCCTGTAAGGGAGTCGGATATTGTTCGTTA AGATGCTCTT
74	ATAGAGGGACCTAGATTTTCAACGAGGGCAGAAAGTAGAATTTGGAGGGAAGTTTATAAAGCC GATATCATAGGGATGACTTTAGTTCCAGAAGTAAATTTAGCTTGCGAAATGCAAATGTGCTATG CAACAATTGCGAT
75	GTCTTCAGCATAGTACCAGCTTATGTTGTCACCATCGTTCAGTACGTTACCACCAAGTCCACTGC CTGCAGCTACATCATTAATGTACAGGAACCAGGCATAGTAACCGCCAGTTGATATGTAGTCTTC ACCTTCGATTC
76	ACTCTCCATCATGACAGCCAGATCGGTCATAGCATCGATTGTGTACTCTTCGTCGGGATTGTTGT ATGGAATGAACTTATAGTTCTCACCTGCTACCTGATCCACTGTCATTTCTGCAAGAGTCTGCACT GTGGTAATTC
77	ATATTCCGTATTTCTTATCAAACCGATCGTGAAGATTTGACAAAGGCTTAACTTTAGGGCTCCAC TTCTCATTATTAGCCTTAGAATATAAAGCGTAACCGTAAGCCTGAGGAACGTAAAGCTTAGGAG ATTCAATCCCG
78	TAAAATTAGCCGAAGGCTTCCCATTACCGAAAAAGTCGTTTATTAGCTCTTCATCCTTCTTCTCC ACGTCCGCCCATTCCTCTCCTTCCCTTGGAATTTTAAGCTCGTCCCAGCTGACTCTTATGGGCAA TTCAATATCC
79	GTATAAACTTTTGATATAACCTTGCCTAATTTGATATCATAGCTTATGTTTGGCGCTATCCCCCA CTTGTAGAGGGTCGCGTTATATTCTCTAATAGCAAGAGAGATACAAGATTCGTTAACGTTATTT ATATCACTCTC
80	TCCGGAGGAATCTATCATATTAAACCTCCTCAAAATCGCCTCCTCTTGATTGCTTAAAGGCTGTG AATTACAAAGCTTATTTAATGCGTCCCAAAGCGTTAAGTAATAATTATTTATATTAAACACTACT ATTTCAGTAG
81	GTTCCTCCTCAATTCAATTGGACTGAAGGAGGGTACGTTCTGGAAAACAGAGCGTAAAAGAGAT ATAGAACGTAGTATACACATAGCTGGAAAAAGAACAATCATTAAGACAATAAAGAACTTTATG GAAAAGAGTAGAA
82	TCGTGTAAAGGTTGTATAATTCAAGCCTCAGAACATTTCGAACTCCTTACAAAATCGTTTAAACT TTCTAAGGCATAAATTTACTAGAAATTGTCATTTATGAGAATGTAACTATATAGATGGTAAAAT TATTAATCCTC
83	GGCTGAAAAATAGGTTCGATCCGCCTCCTCACTTCTTCTCCTTCTTGCCCTCGGCCTCGGAGGAG GCCTCTATTCCCAGCTTCTTGGCCTCCTCCTCGGTCGTCATGAACAGGCTAGTCCTCTGCCTTCC GCCCATGCTC
84	GACCTAGCCTTACGCACAGCCCTCTCCACAACCTCCTCAAGCTTATCCCAGTCAATAGAGCTCAT TACAAGTTAACCACGCCCACCTTTAATATAAACCTTTACCCCTCGTGGCAATTAACTTTAACCGC TACTCCGGTG
85	TGGCCCTTAGACCTCTGCCCATGCTTAGGCGCTTACCCACACCTATTAGTACGGCGCCAATGCCC ACGGCCATGAAGTACATTAAGGCACCCATGGTTGCACCGTAGAGTGCCGTGAATGTTCCGTAGA ATACACCGGCC
86	TCGGCGAATCTGTCGAGCTCCATGACGTCCACAGAGCCGCCGAACTTGGCCGAGAATCTATCGG CCTGGGCGGTGCGCCTCCCTATCAGCAAAACCCTGGGCGCCGTCAGTAGCGCGACGGCCCTGGC GATTCCCCTGGC
87	TCCAGGTAGGATCTGGCCGAGAGGGAGGACGCCGCGCTGTTGTGCTCCGGGAACCCTAGAGTC ACGACCGCCTTGACGCCTATACGTTCGGCGTATTCAGCGACGGCGGCGCCGGTGCCGCCCGTCA GCGTGACGGCAAG
88	GCAAGAGAATACATTTTTGATGATAAGAGAAGCTTGTGGCATACTTTCTTAGGCTTTATTTCAGC ATTCACTTTAGCGTATTCTATCGTTATTTTGCTATTGTTCACATTGTATCAAGTGAGAGAAAGAG AGAAGCCAAC
89	AGAATCAAAGGAGTGGTGTAAAGATGGAGAGAAAAAAAGGTTGGCATCCTATTTATGTGAGTG AAGCGGTTTTAAGTAAGTTAGATAAAGAGAGAGAAGAAATTAAAGAAGAATTAGGTATTCCAA AGGAAGAGAATTTG

90	GTTCAGCATAAAAGACGGTTTCACGGGCCAAAGCCTAAGCGGCGTAACGGTGAAAGAAGGAGA TACGGTTTTGGGCACGATTGACGACGGCGGGACGCTGGAGCTCACGAGGGGCACTCACACCTTG ACTTTCGAGAAGC
91	CTGATGTTATAGAAGTCCGCAAGGACGGCTCTGTCATCTCGCCCGAGGGTGGGAAATACTATCT CGGCGACATAAGCGGCCCGACACAAATTAGCATCAAGTTCAAGGCCGGCGCGGTGGGAACCCA CGGCTTCACTATC
92	TCTCCCTCAACCTTCGCGGGGAGAACGGCGCGGAGTACTGGACGGGCTACGCGGACGCGCTGG AAGACCTGTTGAAGAAAATCCAGAGGCGGGAGGTGAGGGCATGAGAAGGTATTGTTACATCAC GTGGGGATGGATCA
93	GAGCGCCGGGAGGTGAGGGCATGAGTGAGGAATTGATGTTTGGTCGTGTCGTGGAGTATGTTCA GCATAGTTTCTACAAGAAACCGTTTCCTCTTGGCAGTGAGCTCAAGAATGCAGTAGAGAAGGTT ATGGAAACAGGA
94	AGGTCAGAGCCCACGTGGCAACTTTTGAGGTTCTGACAAAAGACTATGTTCGTGAGAAATACAA AGACATCATAGAGTTCATGAGGGAGAAAGGGACAGTATCGAGAAAGGAACTGCGGAAGAAGT TCTTCTTGCTTGCT
95	GTACCTCAAAATACAGAATCATATTTTACAATCGCTTGGAAATATTAATATCAACAATACGCAA GTCCAAATTAACGTCCCTGGCAAACAGGTGACAATTTATACCCACGAAATACTAGATAACGCCA AAAAGGCACTCG
96	CTTTGTATACTTAGATCAGGAAATGGAGCTAAAAGGCACTATCAAGAAGACAAAAGATTCCTG GAGAGAAACATTTAAAGAGTACTCCAAGACAGACAGCGAATATCTAATAAATTACAGACTGTT TTCAATACTCCCTC
	Primer and Core Sequence
97	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAATTGCTCCCTCGTATCCCTTGTACATTATCTCAGC TCCGCTTAATGATATTAATTTTACCTTGAGTGTTTTTGCTAAAGCCTTTGCCATCATCGTTTTACC TACTCCAGGTGGCCCGTAAAGCAACACAGCTTTGGCACACATCATGTAGTAGACGACCAAGAC AGT
98	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTCTCCAAAACCTACCCAGTTCTCCGAGGAACCTCT TAGCATCTGTTAAATCGTTATTAGTATTAGCTTCCACCATCTCAAGTTCCTTTAAGGCGTTACTC ACACTCTTCTTACCTATCTTTTAGAGAACCACTCGTCAGCACATCATGTAGTAGACGACCAAGA CAGT
99	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTATCAAAGCCCTTAAAGAGTGGTAGGGGCAAAA GTCTGAAGCGTCCTTACTTAACTGGAGTATCTGAGATGGCCTTAATCCGCTTAGGTCTTTAATTT TATCCCTTAATGAACATTCCCTGCACTCTATGTCTTCGGGCACATCATGTAGTAGACGACCAAGA CAGT
100	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGATGTAGCAGACGGGCTAAGAGTTTCAAACCCT CTAAGGATCACTACAAACAAGAGAGAGAGACAATCCTCTCTTTTGTCTTGTCATTGTGTTTCAA ACCCTCTAAGGATCACTACAAACATCTTTAACATAGATACCCACATCATGTAGTAGACGACCAA GACAGT
101	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACCGGACGTTGTGATCACGGGTACCTTGATCTGGT ACTCAAAGGTTTGCCCCCGTGAAGTCTGGTACATGGCTAGACACGTCACTCCATTCGAGGGACA TTCGAAGTTAGAGAAGGGCAGAGCGATACATCAGATATATCACATCATGTAGTAGACGACCAA GACAGT
102	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTCTTTTCTCTACTAATTCTCCTCACGAGATCTCTAA ACATTCTTGCTGAAAGAGGATCCAAACCTAATGTAGGTTCGTCAAGCAATAAAATTGGAGGATC AGTTATTAATGCTCTTGCTAAGGCTAGTTTCCTCTGCATCACATCATGTAGTAGACGACCAAGAC AGT
103	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGATTTTGCCATCATTAAAAACAACAATTTGATCACC CATAGTCATAGCTTCTAATTGATCGTGAGTTACATAAATACTTGTGGTGTTTAACATACGGTGAA TATTTACAATTTCTCTTCGCATGTTTTCTCTTAGTTTAGCACATCATGTAGTAGACGACCAAGAC AGT

104	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATCTTTCAATTCTCGAAAGAAAAGGTTACAAGTC TCATAGATTTATTCCTCTTCACTGTTGTACGTTGGCAGCTAGAGAGAGTTTAGATTATGAGAAAA TTAAGAGAATATATGAGGATTCGTTTTCTTGGTTTAAGTCACATCATGTAGTAGACGACCAAGA CAGT
105	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTAATTGATTTTCCTGTACCATGTGGTAAAACAACG CTACCTCTTAATTGTTGATCTGCTTTTCTAGTATCAAGATTTAATCTAAAAGCTAAATCAACTGA AGCATCAAATTTTGTATAAGAAGTTTTTTTCACTAATTCCACATCATGTAGTAGACGACCAAGAC AGT
106	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGGTTTTCCCGTGAACTAATAAACACCTACTGGAG CCAAGAACGGGTCAGAATTGATGGAATAAACGTTGCGGAGAATGAAATTAATTTGTACATCAG AGACATTGATGACAACGGTGACCCTATACAGTCAACTATACCACATCATGTAGTAGACGACCAA GACAGT
107	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTAATGGAAAGTATGCTTTAGATACCTTCTGGAAC GCTATCTCACTTGGCGGGAATTCAGATATGGAGAGTAAATTAAGGGATCTGGAAGTAAAGTTAA TGTCGTTAATCTATTTAAATGAGTCACCATTAAAATCACCCACATCATGTAGTAGACGACCAAG ACAGT
108	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATAATATGTTAGAGGTAGAATTTCTTTGTGATAGA ATATTATTGATGAATGATGGAAGAGAATTAGCATTAGGAAAACCTAAGGAACTGGTAAAGGAT ACAGAATCTAAGAATCTTGAAGAGGTTTTCCTTAAACTTGTCACATCATGTAGTAGACGACCAA GACAGT
109	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTTACTTCATCTCTCAAGATAAGGGTAATAAGTTC ACTTCAAATATCTGGTCTTATCGCAAGTTGATTGAGGCTATAGTGTATAAGCTCTATGAGTATGG TATAAACGTGTTCCTCGTTGTAGAGTATAACACTTCACGCACATCATGTAGTAGACGACCAAGA CAGT
110	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTCTAGGTTTTAATTCTTCAACTGCTTCAAATACT AGCTTACTGTAGTTATCTGCCCTCATGTTAGGATATATATCTGGAATATAAGGAGGTTGATGAGT TATAAGAAGTGGATGAAATTGTTGTCACACACTCCCCTACACATCATGTAGTAGACGACCAAGA CAGT
111	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTACCTCTTCGGCCTTGTACCAACGTACCCCTGATA CAAGTTCCAAGCAGAGATGGAAAACTCGAAGATGGTATCACCCAAGATGAGATACGATATCAA TGAAGGCGAGCCTAGGTACAAGTAAAGGGATACCACGAGAGCACATCATGTAGTAGACGACCA AGACAGT
112	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGTAAGCGTTTCCTACCCTCGAGAGGGCCATCCT GGTGGTGAGGAAGTCGTCGAAGTGGGCTAAGTAAAAAGCGAAGATCTCGACCCACAATTACCT CCTCCTGTACACCAGGAATACCCCTATCAGGATAGAGATACCACATCATGTAGTAGACGACCAA GACAGT
113	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGCGTCCGGGTCGCGGCCGGGGACGACCGTCTTG ACGAAGTCGGTCGACCCCTCGTCGGTCGAGATGGTCGTCACCTCGGTGTCGAGGCCGTACGTTT CGAGCGCGTCGCGTACCAGTTCGCCGTCCGCGTCGGGACGGCACATCATGTAGTAGACGACCAA GACAGT
114	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATGTACTCGTTCCAGAAGGTGAGTTCGCTCCCCTC GATTTCGACCTCGCCCACGTCGAAGCCGCCGGTCGTTTCGAGCGCGAACGACTCGACGGGACCG ACGAGCGAAACTTCGCCGCCGAGCACGTCGGCGACGCGTTCACATCATGTAGTAGACGACCAA GACAGT
115	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGATGCGCTCGGGCTTGTAGGACTCCCCGAGGGC GTCCTTGTTGGTGAAGACGTTTTGTTTTCGCTCGAACCGGCGCATTAGCGTCGGTCCGTTGTAGC GTCCCCTTATTTAAAACCCCGATTTCATCTGATTCATGTCACATCATGTAGTAGACGACCAAGAC AGT
116	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCACGGTCCGCGACGTGAATCGGGCGTTCCAGTCG GCGTTCGGCTACGACGCCGACGACGTGGTCGGAAGCGACCTCCTCGGGCGAATCGTGCCCCCGG TGCCGGACCCGGACCCGGTGCCGGAACCGGGGGACGACGAGCACATCATGTAGTAGACGACCA AGACAGT

117	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGTCCGCGAGTTCATCCTGAACGTCGTCCCGCTGT CGCCCGGCGAGGAGCGCGGGGCGGGCTACGCCATCTACACCGACATCACGGAGCGGAAGACCC GCGAAAGCGAGCTAGAGCGACAGAACGAGCGATTGGAGGAGCACATCATGTAGTAGACGACC AAGACAGT
118	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGAGACCGGCGACGAGGTGCGCTTCGACACCGCC GAGCGGGCGCTCGAACAGATGGAGGAACTCATCGACGACCTGCTGTCGCTCGCCCGTCGCGGC CAACTGGTCGACGAGACGGAGCGCGTCGACCTCGGGGCGGTCCACATCATGTAGTAGACGACC AAGACAGT
119	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACGAACTCGTCGGTGAACATCTCGTCTTCCGGGGAG CCCGCCGCTCATGGCCTGCCCCCGCCGTAAGCTGCTGCATAAACCCGCTCCAAAATATACGGAT CATTCACCCCTTGGAATCGCTCAATCAGATCAATGTACACCACATCATGTAGTAGACGACCAAG ACAGT
120	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGCGTACATTCCCCCTAAGCGGCTCCCAATATACAG ACGCCGGTTAACGACAGCTGGCGACCCTGTGATCTCAGTACCGGTGTCGAATGACCACATCAGC TTGCCTGTCCGTGCATGGAGTTCGTATACGTACCCGTCGTCACATCATGTAGTAGACGACCAAG ACAGT
121	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGATAGATGAGCCGATCAGAGATCGCTGGTGAGTT GGTAATTGTCCCGACATAGACACGCCAACGTTCTGTTCCATCTGCTGCGTCGTAGGTCGCGAGA TACGGCCAGCCACCAACATACACAATCCCATCGACGAGGACCACATCATGTAGTAGACGACCA AGACAGT
122	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATACACCACCCCATCAGCAACAACTGAATCATGATT AAGTATCGCACCAGCATCGTAGCGCCAGCGTTCACTGCCAGTGGTGCTATCGAATGCATAGAAG ATATGCTCCTAATCGCCAATATCAGTACTTCACAAAGCCGCACATCATGTAGTAGACGACCAAG ACAGT
123	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGACGAGGAGAGGGGCGAGTACATCTGCACGCTT ACGGGAGAGGTAGTTGAGGAGACGGTTATAGATACAGGGCCCGAATGGAGGGCTTACACACCT GAGGAGAGGACCCGCAGAAGCCGCGTGGGCAGCCCGCTTACCCACATCATGTAGTAGACGACC AAGACAGT
124	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTCGATGGCTGCGGCAGCTGTCTATGCTGCCTGCC GTATACGCGGCATACCCAGGAGTATAGACGACATAGCGGAGGTCGTGAAGGGTGGCCGTAAGG AGGTTGCCCGCTGCTACCGCCTCATAGTCCGCGAGCTGAAGCACATCATGTAGTAGACGACCAA GACAGT
125	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGAGTCTTTTGTCACACCGCAGAGGCGTAGCGCT GCAGAGCAGGAGCCCAAGCCTACTGCCAACATAGAGAACATAGTGGCTACAGTATCCCTCGAC CAGACTCTAGACCTGAACCTCATAGAGAGGAGCATACTGACCACATCATGTAGTAGACGACCA AGACAGT
126	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTCGCCTGGGTTAAGAGGATGTTCGGCCTCTCCAA GGCGGGTCACGGAGGCACGCTGGACCCGAAGGTCACCGGCGTCCTCCCCGTAGCCCTGGAGGA AGCAACCAAGGTCATAGGCCTGGTGGTGCACACGAGCAAGGCACATCATGTAGTAGACGACCA AGACAGT
127	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTGGGCGAGATCTACCAGAGGCCGCCGCTCCGCA GCAGTGTTAAGAGAAGCCTCCGCGTCAAGAGGATATACGAGATAGAGCTGCTGGAGTACAACG GCAGGTACGCGCTCATGAGGGTGCTCTGCGAGGCCGGCACATCACATCATGTAGTAGACGACC AAGACAGT
128	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCTGGAAGAACGAGGGCAAGGAGGACCTGCTGCG GAGCTACATCAAGCCCGTCGAGTACGCCGTGAGCCACCTGCCCAAGATAGTTATACGCGATACC GCGGTGGACGCCATAGCCCATGGCGCGAACCTCGCGGTGCCCACATCATGTAGTAGACGACCA AGACAGT
129	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGGAGACCCCAAGGTGACCGGCGTCCTACCAGTGG GGCTCGCCAACAGCACCAAGGTCATTGGTAATGTTATACATAGTGTTAAAGAATACGTGATGGT TATACAGCTCCACGGCGATGTAGCCGAGCAGGATTTAAGAACACATCATGTAGTAGACGACCA AGACAGT

130	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAGAGGGAAAGACTGTAGCTTTCATTCCTAGGCAC GGAAAGAGACACAGAATACCTCCACATAAGATAAATTATAGAGCTAATATATGGGCATTAAAA GAACTAGGAGTGAAATGGGTCATCTCAGTTTCTGCCGTAGGACACATCATGTAGTAGACGACCA AGACAGT
131	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGAGGGAGCTCAGGAGGACTCGCACGGGGCCCTAC AGGGAGGATGAGACACTTGTAAGGCTCCAGGACGTCAGCGAGGCCCTGCTCCTGTGGAGGAGC AACGGGGATGAGAGGTATCTTAGACGCATCGTGCTACCCGTTCACATCATGTAGTAGACGACCA AGACAGT
132	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAACATCTATCGCCCACCTCCCGAAGATAATGATC TTGGATACAGCTGTCGACGCCATAGCACATGGTGCCAACCTGGCTGCCCCAGGCGTCGCCAGGT TAACCAGGAACATCGCGAAGGGTAGTACCGTAGCGATCCTCACATCATGTAGTAGACGACCAA GACAGT
133	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGCTATCCCCGTGTACAGCATGGTGGGGGTGCCGA TGCCCGGGTAGAACTTGGTGACGCTCTCCAGCTTCTCGAGGACGGTTTCCTTGGGGAGGCTCGC GGTGTCCACGAGGGTTATCGCGTCCTCGGCGCCGTCGCCGCACATCATGTAGTAGACGACCAAG ACAGT
134	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAGGACGCGAAGAGCGCGGTGGATGTGGACGCGC CGCCGCACACGTAGCCGTCGAGGTAGCGCGGAACCATCGGCGACATCAGCCCCACGACGCGAC CCGAGGCGTTGCCGAGGATCACGTCGAGCGTCACGCGCGGCACACATCATGTAGTAGACGACC AAGACAGT
135	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGACACCGTGCCGTTGCCCTCCTCTAAGTAGTCG GAAAGCCTCATCCGCGACTCCAGCTTCGCCACCGGCTCCTCGAGCAGGAGGAGGACGCGGTTG ATGCGGTAGGACGCACTGCCCGCCTCCAGCACCGCGCCGTCCACATCATGTAGTAGACGACCAA GACAGT
136	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTATGGTGTAGAACGGGTCGTTGCGGAGCCAGCCT GGCGGCACGTACCGGTCGTCCGCTATCGCCAGCGATCTCTCGAAGAGGTCGAGGTAGGCGGAC GCGTTGGCGAACGCCCCGTGTATCACGACGTCTATCCCGCCCACATCATGTAGTAGACGACCAA GACAGT
137	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATAGGTTTCAGGTATTGATAATGCATAGGAGGTT TTTAAAACCTTGAGCCGCATAGTCTTCTGGATGGGCGAGAGACATGGTTAAGTATAAGTGCGGC AGGTGCGGATACGTCTTCGACGACGAGGAGATGAAGAGGACACATCATGTAGTAGACGACCAA GACAGT
138	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTACGCCGGGTGCGTAGGAGGGCTCGAGTACATC CATGTCTATACTGATGTATGTTTTACCCAGGTCGCCTAGTGCCAGGGGTCCCTTTAACGCTTCCA GGATAGAGTACACGGTGACGTCTCTAGTCTTCTTCAAGAACACATCATGTAGTAGACGACCAAG ACAGT
139	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTACTAGCGTGTCAACGGAGCTCTTCAACGCCTTTA CTATTGGATAGGTTATAAGGTGCTCGCCTCCGAGGAATCCCAGGAGCATGCCGGGATACTCGTC TACAACGCCTTTCACCACGTCACCTATGATTCTTAAAGAGCACATCATGTAGTAGACGACCAAG ACAGT
140	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATAGGTGACATGGGGTTTCCCATTGACTCTATAAA GCCGTATCCTTTAAGCGGAGTGCAATTGGTCTACGCTTTGCTTAACAACAGGTATTTCCTACCGG GTAGAGAGGGCTCGCTCATAGCTTTAGGTAGCGTGACGGCACATCATGTAGTAGACGACCAAG ACAGT
141	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGTATCTCACCGCTTGTCACCATAGTATCCCTCAGG TACTCCAGTATTCTTGAGAGAAACGCACCTAAGCCGGATCTCAGGTTTGAATCCATAAGAACTA TGAGTGAAGCGGGATTGAAGCCCCTGCTGTTTCTAAGACCCACATCATGTAGTAGACGACCAAG ACAGT
142	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAAGGGAGATAGAGAAACGCATCAAAATACCCTTG GGGAAACTGCGTGCAGGGGTTCAATATGGAGTAGAGGTCTCAGACATAAAGGAGAAGATAGCT GCTTACGCTAGGAGGAAGGGGCTTAAATACTTCCCATCGGCACACATCATGTAGTAGACGACCA AGACAGT

143	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTGAACCTCGTGCCCGGCTCTAAGTCGTGAGGGCT TGCAACATAGGTGGGGAGGAACCCGAGCAACGGGTAAGAAGACAGGATAAGCGGTATCGCTAT GAAGAGGGCTGAGAAAAGGACATATACTCCTGAGCCCGTCCCACATCATGTAGTAGACGACCA AGACAGT
144	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAACATGCCTTCCCCGTCTATATAGACCCAGTAGA GTTTAAAAACTTAACCAGAGACGGCTTGTGAGCCGGATCTCTCCCCCGCTAGGCCCTGGATTGG GCTCGCTCCTCCTGGGACCCCGGCCTCCACATGCTCGGGACACATCATGTAGTAGACGACCAAG ACAGT
145	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTGAAGGGCTCGGCTACCCTGAAGACGGGCTTCT GCGCGACCGCCGCGTACTCCGCCGTGGAGCGGTAGAAGAGCGAGGCTGTCTCCGTGAGCCTGA CCATTCCGTACAGGGCGACTGCGACGAGCACTATGACTGCGACACATCATGTAGTAGACGACCA AGACAGT
146	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTCAAGGTGCTGATGCCGAAGGCGACTTTCGACAC CGACGATGCCGCCGACGCCCTGGCCATTGCCATCTGCCACGCGCATCACCGGCACAGTGTTGCC TATAGGATGGCGCTGGCCGGATAAGTTTGTTCTTGACCTGTCACATCATGTAGTAGACGACCAA GACAGT
147	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTCGGTTCGGCAATAAGTAATACCAACGAGGTATT ACCATGCGCGTGACCAGCAAAGGCCAAGTGACGATCCCAAAGGAGATACGGGATCATTTGGGG ATTGGGCCGGGCTCCGAGGTGGAGTTCGTGCCCACAGACGACACATCATGTAGTAGACGACCA AGACAGT
148	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGATCATATGGCCGGCACGTTGGACTTGGGAGG CATGACAACGGACGAGTATATGGAGTGGCTGAGGGGTCCACGTGAAGATCTCGACATTGATTG ACACAAATGTCCTGATCGATGTTTGGGGTCCTGCCGGACAGGCACATCATGTAGTAGACGACCA AGACAGT
149	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGGTGTATTTTACACACCTGGACAGCCAGCATATG ATGCTAGCACTCGGTGTCCCCTTATCACGGTTTCCCGCATTGTAAAGTTTTCGCGCCTGCTGCGC CCCGTAGGGCCTGGATTCATGTCTCAGAATCCATCTCCGCACATCATGTAGTAGACGACCAAGA CAGT
150	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTGGAGCCTGTTAGTTGTTACAGGTTCACCGGTTGT CGGAGTATTCAGATCATTGAGCCAGCAGTTGATGGCTGCCTGTAGTTCACTGGTTGTGATGTAA GCTGCTCCATCGGAATCAACATCGTTCCATGGGTTCCAGTCACATCATGTAGTAGACGACCAAG ACAGT
151	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACGGTCTTGCTTTCTCCTGAATCCATTTCACCTGTCC AGACCCATTCATAGCGGTTAGCTTCACTGAGGTTCTGCTTGAAGACACCGTCATCATTGTTAGAT GAGGTTATTGTCCAGCCGGCAGGAATGACTTCTTCGAACACATCATGTAGTAGACGACCAAGAC AGT
152	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTCAGCAGCTCTTCATAGAAGTTCTGGTTTGCAATA TCCCTCTGGGCAATGACAGGGTAGTCGACTTCGTTTGCAGTCAGGTGGACTGCATACAGGGACT TGCTGATGTCCGGGGTATATCCACTGTGAGGAGCATAGTACACATCATGTAGTAGACGACCAAG ACAGT
153	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACCCGTCAGTCGTGACGTCCTCCGCTCCTCCTATGC TATCTCCACACACCCACTCACGTTCTTGCTTCTTTACTACACCCTCTTTATTCAGCTCTTCGAGAA CATTATTAATGTGACCCTTAGAGATATATTCATTATACCACATCATGTAGTAGACGACCAAGAC AGT
154	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGCCTCCTCAAGCGACTGCTTAAACCCAATTACAT CTGATTTATCCTTTATTTTAGGGCCTATAGAATCTATGAATAATTCGGCGATTCTTATTATTTCTA AAACCAATTCGTCTGTTTTGAGTGGTGTGCCTTCTTCACACATCATGTAGTAGACGACCAAGAC AGT
155	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATCCCATGCATTTTCATAATAATCGGAATTCAAAT CCTCTATATTGAATTTTATCTTAACATTTGACATAATCATTTTCTCCTTACAGAAGAGATCCAGC TAAGCTTACTCATAAATGGTAGTACCATGCCAATATTGGCACATCATGTAGTAGACGACCAAGA CAGT

156	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTAGCCCGCACCTTCCTCTGGTTTAGCACCAGCGG TCCCCACAGAGTACCCATCATCCCGAAGGATATGCTGGCAACAGTGGGCACGGGTCTCGCTCGT TGCCTGACTTAACAGGATGCTTCACAGTACGAACTGACGACACATCATGTAGTAGACGACCAAG ACAGT
157	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTGATAGGCCGCAGATTCATCCTAAGGCGCCGGA GCTTTTGACCACAGAACATTCCAGTATCTATGGTATATCTGGAATTATCACCAGTTTCCCGGTGT TATGCCAGACCTTAGGGCAGATTATCCACGTGTTACTGAGCACATCATGTAGTAGACGACCAAG ACAGT
158	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTTTGGCTTGATACTAATAAAAGCACAGCTAAAAT GAAAATAAGCCGATATTTGTGATTCATGCAACTCACCCTTTTCTACATAAACAAAATACTAACC CGAAAACCGAAATTGAAATTAATGCAGAGAAACCAGGTGACACATCATGTAGTAGACGACCAA GACAGT
159	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTAACGGCACCAACAGTTATTATATTTTTAGCAGTC CCGGGTGAAGTAATTATGGAATAGTTGTTAGAATTACTGTTCTTATTACCAGCTGATTTGAAAGC AATTATACCTGCATCACGAATTGCAGCATCATAATATTCCACATCATGTAGTAGACGACCAAGA CAGT
160	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCTCAGACGACTGAAAAAGCAACGATTGGAATAA TAGGGGGTTCTGGGCTCTATGATCCTGGTATTTTGACTAACAGCAGAGAAATAAAAGTATATAC ACCCTATGGGGAACCTAGCGATTTGATAACGATAGGTAACACACATCATGTAGTAGACGACCA AGACAGT
161	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCAGAACAGGTTCCTTCTATTGGATATTCATCTTC GGCTGCAGTTGCAGGAAGAGTAAGGATATATACTACGGTCTTGCTTTCTCCTGAATCCATTTCAC CTGTCCAGACCCATTCATAGCGGTTAGCTTCACTGAGGTCACATCATGTAGTAGACGACCAAGA CAGT
162	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTCCTCCACGCATTTGTTGTGGTGCTGATGGCGTA TTCTCTGGAATTTGGGATGATTCTGGAAATCCATCCTCAGACACTTCAGATATTTTAGTCTTACT TCCAGCGTTTAATTGAACCTTACCTTTAAAAGCAGTAGTCACATCATGTAGTAGACGACCAAGA CAGT
163	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAACTTACCTTATCAGTGTCATTAAGCATATTGCT TCCAAGACCCATTGAAGCACTTACATCGTTGATACACAGGTGCCAGGAATAGTATTCCTCAGTC TCACTATAATCCTCGTTGGTGTAGCCTTCAAGAGAGTCAACACATCATGTAGTAGACGACCAAG ACAGT
164	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTTAAGCAATTCTTCGGATGAAAGATGGCGCTCTA TAGGAATTTGTTCTGGTCTAGCCATAAGGCATTATTTGTACTTAATTAGTAATAAATGTTTAGTT AATGACTATAAATCTGCAATTGGAGTCTCAAATTTTCAACACATCATGTAGTAGACGACCAAGA CAGT
165	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAACATGAAGGATGTGTGTAAGAGGAAACGTTATTA ACAGACGTAATCAGGAGGATAGTTATGCCCTAAAAACAGCAGAGTTAAGGTTTAAAAATAAGA TAAGAACTCAGTTGAGGTTTATCCATTAATCCCATTAATCCTCACATCATGTAGTAGACGACCA AGACAGT
166	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACTTTCTAAAAGCGCTTGGAGCACGTATCAGGTCAA GTCTTTCAACCTTAAATGCTGCCAGTGCCGTAAGTAGTGCAGTTATGTTGCTTATTGAAACAAAC AACTTAGCCCACTTATTACCTCTTGTCAGTGTTTTTGATCACATCATGTAGTAGACGACCAAGAC AGT
167	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATCCGCTGATATATCCTGGGGATATAGATCGCTC TGAAATGGTTACATCTATCGGTTTTAAGGACAGTTCCAACACTATTGGACCTTGCAGCTATGAC AGGAATAATCTGTTTATCGAGCACAGTTGAATTTGACCTACACATCATGTAGTAGACGACCAAG ACAGT
168	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCAATACCTAATTCTTTCCTTAGAGTGCTATTTTGAT TGAATTCCCTCAGGAAAGATTCAAAATTTAAGTAGCCGAGCTTACATCTTGAAATTTCCATCTTT ATTATGTTGCTCAGGCTTAATGCTTCTAAGTATGGGTTCACATCATGTAGTAGACGACCAAGAC AGT

169	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGATATCCTTTGAAATTCTCGTAATTGCTGAAGGCC ACTACTTCATCAGGTCTGATGCAATCTTTAATCTGAACATTGCTTTCTGAGGTCTTAGGAATAAT CCTGTAAGGGAGTCGGATATTGTTCGTTAAGATGCTCTTCACATCATGTAGTAGACGACCAAGA CAGT
170	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATAGAGGGACCTAGATTTTCAACGAGGGCAGAAAG TAGAATTTGGAGGGAAGTTTATAAAGCCGATATCATAGGGATGACTTTAGTTCCAGAAGTAAAT TTAGCTTGCGAAATGCAAATGTGCTATGCAACAATTGCGATCACATCATGTAGTAGACGACCAA GACAGT
171	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTCTTCAGCATAGTACCAGCTTATGTTGTCACCATC GTTCAGTACGTTACCACCAAGTCCACTGCCTGCAGCTACATCATTAATGTACAGGAACCAGGCA TAGTAACCGCCAGTTGATATGTAGTCTTCACCTTCGATTCCACATCATGTAGTAGACGACCAAG ACAGT
172	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACTCTCCATCATGACAGCCAGATCGGTCATAGCATC GATTGTGTACTCTTCGTCGGGATTGTTGTATGGAATGAACTTATAGTTCTCACCTGCTACCTGAT CCACTGTCATTTCTGCAAGAGTCTGCACTGTGGTAATTCCACATCATGTAGTAGACGACCAAGA CAGT
173	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATATTCCGTATTTCTTATCAAACCGATCGTGAAGAT TTGACAAAGGCTTAACTTTAGGGCTCCACTTCTCATTATTAGCCTTAGAATATAAAGCGTAACCG TAAGCCTGAGGAACGTAAAGCTTAGGAGATTCAATCCCGCACATCATGTAGTAGACGACCAAG ACAGT
174	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAAAATTAGCCGAAGGCTTCCCATTACCGAAAAAG TCGTTTATTAGCTCTTCATCCTTCTTCTCCACGTCCGCCCATTCCTCTCCTTCCCTTGGAATTTTAA GCTCGTCCCAGCTGACTCTTATGGGCAATTCAATATCCCACATCATGTAGTAGACGACCAAGAC AGT
175	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATAAACTTTTGATATAACCTTGCCTAATTTGATA TCATAGCTTATGTTTGGCGCTATCCCCCACTTGTAGAGGGTCGCGTTATATTCTCTAATAGCAAG AGAGATACAAGATTCGTTAACGTTATTTATATCACTCTCCACATCATGTAGTAGACGACCAAGA CAGT
176	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCGGAGGAATCTATCATATTAAACCTCCTCAAAAT CGCCTCCTCTTGATTGCTTAAAGGCTGTGAATTACAAAGCTTATTTAATGCGTCCCAAAGCGTTA AGTAATAATTATTTATATTAAACACTACTATTTCAGTAGCACATCATGTAGTAGACGACCAAGA CAGT
177	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCCTCCTCAATTCAATTGGACTGAAGGAGGGTAC GTTCTGGAAAACAGAGCGTAAAAGAGATATAGAACGTAGTATACACATAGCTGGAAAAAGAAC AATCATTAAGACAATAAAGAACTTTATGGAAAAGAGTAGAACACATCATGTAGTAGACGACCA AGACAGT
178	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGTGTAAAGGTTGTATAATTCAAGCCTCAGAACAT TTCGAACTCCTTACAAAATCGTTTAAACTTTCTAAGGCATAAATTTACTAGAAATTGTCATTTAT GAGAATGTAACTATATAGATGGTAAAATTATTAATCCTCCACATCATGTAGTAGACGACCAAGA CAGT
179	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCTGAAAAATAGGTTCGATCCGCCTCCTCACTTCT TCTCCTTCTTGCCCTCGGCCTCGGAGGAGGCCTCTATTCCCAGCTTCTTGGCCTCCTCCTCGGTCG TCATGAACAGGCTAGTCCTCTGCCTTCCGCCCATGCTCCACATCATGTAGTAGACGACCAAGAC AGT
180	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACCTAGCCTTACGCACAGCCCTCTCCACAACCTCC TCAAGCTTATCCCAGTCAATAGAGCTCATTACAAGTTAACCACGCCCACCTTTAATATAAACCTT TACCCCTCGTGGCAATTAACTTTAACCGCTACTCCGGTGCACATCATGTAGTAGACGACCAAGA CAGT
181	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGGCCCTTAGACCTCTGCCCATGCTTAGGCGCTTAC CCACACCTATTAGTACGGCGCCAATGCCCACGGCCATGAAGTACATTAAGGCACCCATGGTTGC ACCGTAGAGTGCCGTGAATGTTCCGTAGAATACACCGGCCCACATCATGTAGTAGACGACCAAG ACAGT

182	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGGCGAATCTGTCGAGCTCCATGACGTCCACAGA GCCGCCGAACTTGGCCGAGAATCTATCGGCCTGGGCGGTGCGCCTCCCTATCAGCAAAACCCTG GGCGCCGTCAGTAGCGCGACGGCCCTGGCGATTCCCCTGGCCACATCATGTAGTAGACGACCAA GACAGT
183	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCAGGTAGGATCTGGCCGAGAGGGAGGACGCCGC GCTGTTGTGCTCCGGGAACCCTAGAGTCACGACCGCCTTGACGCCTATACGTTCGGCGTATTCA GCGACGGCGGCGCCGGTGCCGCCCGTCAGCGTGACGGCAAGCACATCATGTAGTAGACGACCA AGACAGT
184	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCAAGAGAATACATTTTTGATGATAAGAGAAGCTT GTGGCATACTTTCTTAGGCTTTATTTCAGCATTCACTTTAGCGTATTCTATCGTTATTTTGCTATT GTTCACATTGTATCAAGTGAGAGAAAGAGAGAAGCCAACCACATCATGTAGTAGACGACCAAG ACAGT
185	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGAATCAAAGGAGTGGTGTAAAGATGGAGAGAAA AAAAGGTTGGCATCCTATTTATGTGAGTGAAGCGGTTTTAAGTAAGTTAGATAAAGAGAGAGA AGAAATTAAAGAAGAATTAGGTATTCCAAAGGAAGAGAATTTGCACATCATGTAGTAGACGAC CAAGACAGT
186	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCAGCATAAAAGACGGTTTCACGGGCCAAAGCC TAAGCGGCGTAACGGTGAAAGAAGGAGATACGGTTTTGGGCACGATTGACGACGGCGGGACGC TGGAGCTCACGAGGGGCACTCACACCTTGACTTTCGAGAAGCCACATCATGTAGTAGACGACCA AGACAGT
187	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTGATGTTATAGAAGTCCGCAAGGACGGCTCTGTCA TCTCGCCCGAGGGTGGGAAATACTATCTCGGCGACATAAGCGGCCCGACACAAATTAGCATCA AGTTCAAGGCCGGCGCGGTGGGAACCCACGGCTTCACTATCCACATCATGTAGTAGACGACCAA GACAGT
188	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTCCCTCAACCTTCGCGGGGAGAACGGCGCGGAG TACTGGACGGGCTACGCGGACGCGCTGGAAGACCTGTTGAAGAAAATCCAGAGGCGGGAGGTG AGGGCATGAGAAGGTATTGTTACATCACGTGGGGATGGATCACACATCATGTAGTAGACGACC AAGACAGT
189	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGCGCCGGGAGGTGAGGGCATGAGTGAGGAATTG ATGTTTGGTCGTGTCGTGGAGTATGTTCAGCATAGTTTCTACAAGAAACCGTTTCCTCTTGGCAG TGAGCTCAAGAATGCAGTAGAGAAGGTTATGGAAACAGGACACATCATGTAGTAGACGACCAA GACAGT
190	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGTCAGAGCCCACGTGGCAACTTTTGAGGTTCTGA CAAAAGACTATGTTCGTGAGAAATACAAAGACATCATAGAGTTCATGAGGGAGAAAGGGACAG TATCGAGAAAGGAACTGCGGAAGAAGTTCTTCTTGCTTGCTCACATCATGTAGTAGACGACCAA GACAGT
191	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTACCTCAAAATACAGAATCATATTTTACAATCGCT TGGAAATATTAATATCAACAATACGCAAGTCCAAATTAACGTCCCTGGCAAACAGGTGACAATT TATACCCACGAAATACTAGATAACGCCAAAAAGGCACTCGCACATCATGTAGTAGACGACCAA GACAGT
192	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTTGTATACTTAGATCAGGAAATGGAGCTAAAAG GCACTATCAAGAAGACAAAAGATTCCTGGAGAGAAACATTTAAAGAGTACTCCAAGACAGACA GCGAATATCTAATAAATTACAGACTGTTTTCAATACTCCCTCCACATCATGTAGTAGACGACCA AGACAGT
	Core Sequence
193	CAAATGCTTTCAGTGGTTTCCAATGCCCTGACCAGCCTGTACTCAACTAAGGGATATGACGTAA CATTCAGTGACCTGATCGCCGCCATTCAGGCAATGAAGGGCTACGATGACAGCGCAAACGCTA AACTCGTCGTGGA
194	AGTCGGAGCTTATAACCACAGATCCTGAAGTATTGAGAAAGAGAAGGGGATGGTGAGATAAAC ATGAGGTGTGNNNAAAGCTGACAATATGTACGCGTGCTTGAAGGACACTGTTGTGAAGGAAAG GTATCCTGTCGCTA

195	ATGCAGTATATGGCAGGTGCGAGCAACTGCTTTACCATGGGTGCCATGGTGCAGAACGGGAGA AGCTCCCTCTACTGGAAGGTTAAGGACGCAAATTTCTGGTCCAAGACTTTCGAGAGCAAGTTGC GCGTTCTGGGGCT
196	GGCTGCGGAAGCCAGCAAGGAACCTTGTCTCTTCAGGAGAGGCAGGATATACGTCACATTCAAT GTAAGGAGGGTGGCACTGAGCGGTGCGGGACAGTTGACTGCCGCCAGGACTTATGCCATAGGC AACACCGATGCGA
197	GATGAGGTTAAGGGGATATGCGAGAGATTCGGCAAGGTCGTGGATGCCGTGAACTCTCCTGTGC TTACCGAGAGTAATGCCTCGTACAGGAATGCGGGGCTGGTGCGTGCCAGGTTCAACTGGGACTA CATCAGGCCCGA
198	TGGTCTCCACGGAAGGCTACATGGACAGGGCAATAGGCGTCCAGGATATCGGCTACCTGTTCTG GCAGGCAGGTCCCACCGCAATGAAGGATATGAGAATTTACAACGGTCCCGGTGGTCTGATCGTT CTGCCTTTCTAT
199	CTTTCTGCGCGCAACAGGCTGGCCAAGGGACTTCCGAAGAGCCTGGACATGTTTGCCAGCGTGG AAGGTCGTGACCTTGGGTACGATCCGAGGTACATAACAGAGGAAGATTACAAGACCATTATGA CCAAGGCCCGTCT
200	GGCATGATAGATGAGGATGGCTACGAGGTACCGAAGGGTGAAGACCCCAACGACCCCAGAAGT GCACACACCTTTGGTTGGGTCGACCAATCAGATGGAGGCACATCCAATGGTGGCATGCAGTCCG GTGGGAGCTCTCA
201	CTTAAGGACGGGGACGTGACAAAGCTGTATTCTCAGGACGATTACCTCAGGGTCAGCAGGCTCA AGTTCAGCGAGAATCCGATGCTTGGCATCGTCAAGAATACGGATGGCACAGGGGAGGTTATAG GTCCGTCCTTTGC
202	AGGGACCTTCTGTGGAACATAATCTCGGGTGCCCTGAATGCGGGAAGGGAACAGCTCTACGGG GATGCATTCGGCGGTCCTAAGATAGAGCAGTACGTGAAAGCACTCACGCAGGTGCTGTATGACC TGTCTGTCAACAG
203	TGAAGGAACGCGTCATCGTCCACAAGAGCACTACGAGGAACGAAGTCCTTGAAGAGTTCAAGG CGTCTAAGGAGCCGAAGGTCCTATTTGCGATAAAGATGGAAGAGGGTACGGATTTCAGGGATG ACCAGGCAAGGTGG
204	CAGATATTGGTCAAGACTCCTTACCAGGATCTGGGAGACGAGTGGGTCCGCCTCCATAGGGAGA AGATGGGACGGAGATGGTACGAGATATCCGCCCTCCAGCAGGTCATCCAGGCGAGCGGCAGGA TAATGAGGAACGA
205	CAGGGACTGGGGAGACACCTACGTCCTTGACATGAACGCCATGAAGCTCATCCGCATGTACGA AAAGGAATGCCCCCGCTGGTTTTTGAAGAGGTTGAAACTATGACGCATCACAATATCACCTTCC CCGTCCCTCCCGA
206	GCAATGTCCGCAACAGTTGACAACGTTGCACGGGTCGCGGGATGGCTCAGGGCGACTGCAGTCT CCAGCGATTTCAGGCCCGTCACCCTGAAGAAGTACGTCCTTACCCCCCGCCATATCATGGATGA GAAAGGGGATAC
207	CGAGATACAGCAGATGCTGGGGCGCGCCGGGAGGGCGAAATACGATTCCATGGGCTACGGCTA CATCTGCTCATCCGACGTTCACCTCCAGGACGTGTATAAGACGTACGTTCATGGCCGTCTGGAG AGCGTAAAATCAA
208	GGCACTGGACGAGTTCTTCTCCACCACGCTGGCACGCCACGAGGGTGCCCGTCTGGAAGAATGG ATAGACAACAGCCTTGTCTTCCTGCAGGACAACGACATGATAGTCGGGGGACGTTCCTTCACGG CTACCCCCTTCG
209	CCATCCTTGCGGACTGGATAGACGAGAAGCCCGAAAGCGACATCGTCAATAAATACAACATCT GGCCTGCCGACCTGAGGAGCAGGGTTGAGTTGGCCGAATGGCTCTCGCATTCCCTTTACGAGAT CTCGAGGGTCCTG
210	GGCGACTATTCCTACGTCAGCGTTGCGGAATATTTCTCCAGCTCAAGGATAATAGCCACTACCG CTTCCCCCGGTGGCGACAGGGAGAAGATAAACGAGATCATGCGCCACCTGAGAATAGAGAACC TTGAGGTGAGGGA
211	ATCGACGCTCCAGCTGTTCAGGGACGGTGCGGTCAGGATACTCGTAGCAACGCAGGTTGGGGA GGAAGGACTGGACGTACCGGCTGCAGATACCGTCATATTCTACGAGCCGGTGGCAAGCGAGGT CCGCTCAATCCAGA

212	CGCGGTCCTCTTTCCCAGCTTCAGTCCTTTGGCTTTCCTATGTTCTGCTGGTACTCTTCCCATTGC TCTCTCTGTTGTTTCTCCTGGCTTTTCCTGAAGTTTTCGAGGGCTTCGTCCATGCTGTCATAAGCG TTATGCGA
213	CGAACTCCTTGACGCTCCTGGAGATGACCAGTTTCTCTATCTCCACCCTCCCGCTCTTCATGTCC GATATTATTTTCCTCGCCCTTCTCAGTGCCTCGTCCACGTTCCGGTCGAGCACGAGATTGAACAT TTCCATGAGT
214	CCCTTTCCTCGAGGTCTTTGTCTATTCCCTTCACCGTTTCCCTGGCCCAAGCAGTGATGCTGGAC CCGATACTGGGATCGGTAAACCTGTAGAAGCTTGATGCAAACACTCCGTAGAACGAATTCATAA GCACCTTCACC
215	CAGTTCTCCTTCCACCCCGTCCGCCTCTTCTATGACGGTCGCACTTTCCGTCGAGCAGACGCCCT ACGGCTGGATGGATCAGTATCCCTCGTCGGTTGTTGCCCATGTCACGGGAGGCATCCCTCCCTA CGCCTATCACT
216	CAATCTGGAGGCTTTCGCCGTATCCGTCGAAGTCACGGACTCATCGGGGCATTCAGTCTCAGGT GCAATCATGATCAACTACGGTTCCATAGACCTCTCCCCGTTTGGATACATGGTAACTTTGATTTT TCCGGTGATCA
217	GAACATCCATTGCTGACCATCACGATCGATGGCTCAAAGGACACGTTCAAAACTGGCGATGTCC TGGAATGGTTGACCGAAAGTGACATCTCAAACATGCACAATGTTGCGTCCTTCACAAAATCTCT CCTGAGGATAGT
218	GCATGCTGCCCGATGGCCAATGGTTCGGGGAGGTCATTGGGAAGGACGTGCAGGGAAATCCCT ACGGCATTGATTATACAATGTGGTTGCCGTTTAACACCTACGTTAGGGATAAGCTCAGTTACAA TAGTTGGGGGAAG
219	CCAGCATTCCTCCTCTGAGGGAGTTCGGAAGGTAAAATCTCCTGATGACATCTGAGTCCCTGGC GCCCATTGTCTTGGCTGAGTAGACTTCCATTTTACTTGTCGTGCTGCCAGTTGAAAATGCAAGTA CTGCTATCATC
220	GTTCCCTGCAGGAATGATTGTTCAACTGCACTCGTCAGTACATGATAGAACAATCTTGAAGCTG ACAGATGGAGCGCATAGAAGATAAGGAATGCAAGAGGTACCAGTACTAGTACCACTGCATATA TTCCTGCGTATAG
221	CGATTGCAGGAACGTGGAAAGTGTGCGGTTAATGTATGACTCATTGCTCACATCGAATCGATAC AGAAGACCGCTGTTTGCTGCCAGGTAGGAATCTATGTTATTCAGGCCAACAACCACATCGTATC CCGCCCCCTTTG
222	AGGATTGAACCGGTTGGTGTCAACGTAGAATCCTCATTGACCCGCCACATAAAACTGAACATTC CAATTGTGTCGTCTCCTATGGATACGGTCTCTGAGGCAGATATGGCAATTGCACTAGCAAGACT CGGTGGTATTGG
223	CTTATTATACGCGACCTTTACACTGTAAGCCCGGAAACACCTGTTGACGATGCAATCCGTACTAT GAGGGAGAAGCGAATCGCTGGGCTCCCAGTGATATTGAACGGCAAACTTGTCGGAATACTTAC GAACAGGGACAT
224	GGTACGGCAAGATAGGCTCAGGGAAATTTGTACCAGAGGGAGTTGAAGGAGCAGTTCCGTACA AAGGTAAAGTTGCAGATGCAGTCTTTCAATTGATCGGGGGCCTGAAGTCGGGGATGGGGTATAC TGGCTCGCCCACA
225	GGTGGAAGCGTTGAGGAGTTTGTCACTCTATCGAGGAGAGTGGAGGCAGCGGGATTCGACAAG GTCGAGCTCAATTTGTCCTGCCCACACGTTCAGGGAGTTGGATCCGAGGTAGGACAGGATGTAG GTCTTGTAGAAGA
226	GACACTTATAGACAGGCTAGACAAGAAGACGAAGACAAGGATATTCTTCTCACTTGAGCGATT GATGAAGTGCGGCATAGGGATTTGTGACAGTTGCAGCATCAACGGCATCCGGGTATGCAAGGA CGGAACAATTTTCG
227	CTTCGCAACTGCAAAGAGGTAGCTTCTGGATGCTTCCCTGGAACTATCCCTACATTGCTGTTATC TTACTAGTGGTACTGATTTATGCAGCAATAGAGGACCTTAGGAAGAGGAAAATAACAACTATA ACCTTCCTTGCA
228	GTGACAGTTGGAACTGGTCTATCTCCCCGGTATTTTAATAAGTTTATAGGCGTAGCAAAGGCAT ATACGACAAGAGTAGGGGAGGGGATATTTCCTACTGAGATGTTTGGGGAAGAGGCAGATAGAC TTAGAACCCTAGG

229	GAAGAAGACTTAAAGGATTTAGGTAGAGAGCTTAAGGTACCAAGAAGACCGTTCAAAAAGTTA ACGCATAGAGAAGCTGTTNATATATTGAGATCTCATGGCATAAAAGCAAGTTATGAACATGAG ATACCTTGGGAAGC
230	ACGGGGAGGCTGTCTCAGGAGCTGAAAGAGAATATAGAGCGGAGAAGGTTATTGAGAGGATGA GAGCTACTGGTGAGAACCCTGCAAAATACGGTTGGTACATTGAAATGTTGAAATATGGTATTCC GCCGAGTGCAGGG
231	ATATGCAGATTTAGATGAGATTATAGGGGTTGCATCTAAGGCAGGAATAGATTGCATAACTATA GATGGGTCAGAAGGTGGAACAGGTATGAGCCCTATAGCTGCGATGAGAGAACTAGGATATCCA ACGCTAGTATGTC
232	GGACACGAAATTGCTGAAGCAGCTGGCTCAACATGGTATATCGACAATTTCTGGGATAAACTCA AAGAGGGCTGTGTAGCATATCTAAACATAGATTCACCTGGATTAAAAGATGCAACAAGATATAT CGCTTACGCGTC
233	GTAACTTCTGGAAACGCCCAATCAAAACAGATCATGACACCAAAGCTAAAATTATCTTCCCTAA TAGCTTCTATAGGTGTATCTCCAGGTTGAAATATTAGCTTCTCTTTGGCAAATAAGTGAAGTTTC CTATACTTTCC
234	CCAGATAGCCCAATAGCATCAATTTCCGTTGCAATAATAGGTACAGTACACAAAGAACACGTAA TTTTCAGCGACACTGCAAATACAGGCGACTTAATAATTTTTGCCATAGATCTCGATGGAACATTT CACCCTAAGTT
235	GTTCTAATTCCTCTCTTACAGCTTTAAAAGCAATCACAGCAGATTCCAAAATATCATCCATATCA TCCAGAGCTATAATAACACCTCTTGAAGTTTTCCCAATCTTATGCCCACTTCTTCCAACTCTTTG AACCAAACGA
236	GTAACTTGTCTGGGAGACATATATTGGACAACTAAATCAACGGTTCCTACATCAATCCCTAACT CCATAGATGATGTACAAATAAGACCTTTCAACTCACCGTCTTTAAATAACCTTTCAACTTCTATA CGAACATCTCT
237	ACTCAATGAACCATGATGCACATCAATACTTAGATTAGGATCGTATAAGTGAAGCCTAGAAGCT AGTATCTCAGCTATTTCACGAGTGTTTACAAAAGTAAGCATAGAGCGGCTCTTTTCTAATAACTC AACCAATACCC
238	CAGTTAAATCATCTTAACTCACAAATATTAAGGCTTTAATTTCTGAGGGAGTGCAAAATGAAAA CTGACGTAGTAATAGTAGGTGCAGGGCCCGCAGGCATGTTTGCTGCACATGAATTGGCAACTAA ATCTAATCTGAA
239	AAAAATAGCCAAGGATCCAAAATTCCGTGTATATACAAAAACCTTCGATGACCTTACACGTGTA TTTTGCGTTAATTATCGAGGCTTCGTCGTCCAAGAAGTCTACGGAGATATCGTTGGTGTTAACGG CCACACTCTAA
240	TCAAACAAAAATCTGAAAATGCCAATTTTGCATTTCTAGTTCGAGTTGAACTCACCGAACCGCT TGAAGACACAACCGCCTACGGATTCTCAATAGCCAAATTAGCAACTACCATAGGTGGAGGAAA ACCAATTCTTCAA
241	CGAGATACTGAATTTCCAAAACTCAAAGGATATAGAATTGTTAGAATCGCAACACATCCGCAAG TTATGAGCATGGGACTAGGAAGTGAAGGGTTGTCAAAACTTTGCCAAGAAGCCGAAAAGAGAG GACTAGATTGGGT
242	CGAAGTTTTTATCCTCCTCGGTCCAAGTCACACTGGTTACCCAGGCGTTGGAATAATGACAGAA GGCATCTGGAAAACTTCTTTAGGAGAAATATCAATAGATGAAACTCTCTCGAATACTATTTTAA ATAATTGTGACC
243	TGACACACTACGGCACCTACTATGGATACACACCAGCTGGTGTTGAACCATTAACCAAAGTTTT AGAATGGATATACCAGACGGACAAACAAGTTATTGAGAGAATTAAAAGATTAGATGGAGCAGG AGTAATAGAATAT
244	CTGAAAAGTTCATTCCAATTGTTAAATCGCCATCTTGGAAACACGGCACAAGAAAAGGGAAAG GATTTAGCATCGGTGAGATTAAAGCAGCCGAGATAGATATTAGTATGGCAGTTAAACTCGGTAT ACCCATTGATAAA
245	GGGAATAATAATTAAAATAATGTGGCACACCTTTTAGCTTCTTTTCATCTCATATTTTCAAAGAA GCCTTCCAGGTGTGCCTCATCGGTGTCCCCCGCTGCGGAGACACGGTATCATCGTATCCGCCGA AGGAAACTCAA

246	GACATTGCCTATCAATTACTTCAAGCCGGAATGCAAGTTCCCGGTTTCAGAAGGTCGCCAAAGA TAATAGAAAGAATTTTAGAAAGATATATTCCAACAGTCACCGTACTAGGCGGCATTATTGTAGG ATTAATAGCTGC
247	TGTCGTTCAGGGAGGTATAAAAATGCCAGAACCACGCTACCGGTCAAGGTCTTTAAGAAGACG ATACGTACACACACCTGGAGGAAAAACCGTCATCCATTACAGGAGAAAAAAACCTGACGTTGC AAAATGCGCATTAT
248	GTGGTCAACCTCTCAGAGGAATTCCCAGACTAAGGCCAGGAGAATTCAGAAAGTTGACAAAAA GTCAACGAAGACCAGAGAGACCTTTCGGTGGATATCTATGCCACAAATGCTTAGCAATGGAAAT CAAGAAAGCTGTT
249	ATAGGATGAATCTAACTGGGGCGACCCGGTAGATAACTGAGAGTGTAGGAGGTGAAATAATTG AGCGCAATAGAAGTAGGTAGAATATGTGTTAAAACTAGTGGAAGAGAAGCAGGAAGAAAGTG CGTTATTGTTGAAAT
250	ACACCATTTCCTAATATTTTAGTAACTAGATATGTTTGTTATAGTATTAGGGTGAAGTATTTGTA TGAAAGAAAGTTGCCATCAGACATTAAAAGAGAGATTCTAGTAAAAAGTGAAGCAGAAACTGA CCCTGCTTATGG
251	CACATGAGAGAACTTAGAAGAACACGTACAGGACCCTTTAAAGAAGATGAAACCCTAGTAACT CTTCACGATGTAGTTGATGCTTACTATTTTTGGAAGGAAGATGGAGAAGAAGAATTTCTACGAA AAGTCATACAACC
252	AATGGAAAAGGGTTTAGAACACCTACCTCACATTTGGATTAGAGATTCTGCTGTAGATGCAATA TGCCATGGGGCAAACTTAGCAGCTCCTGGTGTTGTAAAACTTCATGACGGTATATCACCTGGAG ACTTAATAGTAA
253	CGCTGATCATACATGTGCATTGTCTTTAAATACACTAGTAACGTTAATAATATCTAGCAATTTTA GATAAAAATAACTAGCAGTGCCGGGGTAGCCAAGTGGACTACAGGCCTTATACCGGTTAGGGC GCGGGCCTGGAG
254	CATGCCTTAACGAGAGGCATGGGATGGGGGAGCTGTGAGCCCCCCGAACCGGCAGATGAGGGG AAGGGTGCAAAGCATCCCTTAACGCCGGAAGCTCCCGACTTCAGTCGTGGAGCAGCTCACTGCT TTGACGAAAGGTT
255	GAACTTGCAAGGAAGGCCGGTGTTGATTATGAGACAAAGCTGTTGGTCAGGGGCAAGGAACCG GCTGAGGACATAATAGAATTTGCTGACGAGATCAGGGCAAGTCTCATTGTAATAGGGGTTAGG AAGAGGAGACCCGC
256	TCCAGAAGAGATTCAAAGCTCTCGTATTCAATGTCCCCACCAAATTTCTGGTCGCGCTCAATTTT GACTTTACCAAAAGCGGGGAAAACGTAGTGCTTTGCTAGGTCTATTATCGGATTTCCTTCTACA ACCTTTGGCGG
257	GATTTGCTCATTTTCTCCCCGTCGAGTCCTGAGATTATCGGCGTATGGATGCAGATCGGTGCCTT GTAACCGAGGGCCGGCAGATTCTCCCTTGCGAGCATGTGGATCTTTCTCTGATCTATTCCACCAA CCGCCACATC
258	TCCGGGAGTTGCAGAACCAAGCATGGAAATTGCTAGAGATCCCGAAAAGGTTTACGAGTACAC GAATAAGTGGAACACGGTTGCAATTATCACTGATGGCTCGAGGGTCTTGGGACTGGGCAACATC GGTGCGATGGCTT
259	GTGGTGTTATCAAGAGGGAATATATTGCTCAGATGGCAGAGGATCCGATAGTCTTTGCCTTATC AAACCCGGTGCCTGAGATCTATCCGCAGGAGGCAAAGGAAGCCGGAGCCAGGATCGTAGGAAC TGGTAGGAGCGAC
260	GGGATCTGTTAGTATGGCATTCAGAGCCTTTATGTCCTCATCGGTAAGCTTGTCCGATGGCAGAT CGTATTTCACGATGTCTGAAGGAGTAACTCCGAGAAACTTCGCTTCTGGTGTCGCAAGATACTC CGAGAGATGCG
261	TGAGTGCGGCTTACTCTGCACTGTGCGAGATCGATGAGGTCGTTGTTGTTGCCCCCATAACGCA GATGAGCGGAGTGGGGAGGAGCATATCCATAATGCGGCCGGTTCGTTTTTTCGAGCTCGAAATA GATGGCATGAGG
262	AGGGGAAGGGAGTACTACTGGATTCATGGGGTGGAAGTCGAAAGCGCTGAGCCTGGAACGGAC ATACACGCACTCAGAAACGGGTATGTCTCCATTACACCGATATCCTTAAATGCAACTTCGGACT GCGAAGCTTTAAG

263	ATAGTTTTATGGAGGGTGGTTGGACATGAATGAAAGGGCAAAGAAGGTCATTCTTATTGTGGAT GACGATTTGGCTCTGCTTGAAGCTCTTGAACTGATGCTTCGAGGCAAGTATGAGGTTGTGAAGG TGACAAATGGGA
264	ATGTCGATTCCGAAATAGCAGGGAGCAATTATCGGTGGGCTTCCGACCCTTAAATGGATTTCCT TCGCTCCCGCCTTTCTTATCATGTCGACTATTCTTTTGGATGTTGTTGCCCGCACAATGCTGTCGT CAACCAGCAC
265	ACTTTCTGAGGGAAAAACATTGTTGCTTATCCTAAAGAGTTTACAAGCAAGAAGCTGGAAACAA ACTCTGGATGTTATTAATTTAGAGCCTGCAGCAGCATATACAATGTTTAGAGCGGCAATAAAGA AACTATACAAAG
266	GTGGTTGAGAGGCTGCTTGAAGGCATTGCAAAGAATGAAAGGGTAGCTTACGGATTGGAGGAG GTTAGGAGGGCAAAAGAGTATGGAGCAATTGAGGTTCTGTTGGTTTCAGATGACTTCCTGCTCA CCGAGCGTGAGAA
267	TCGCTTCGAGATTCCTGATAGGAGTGGGAGTTGCCGGGGTTTACGTGCCTACGATAAAAATAAT ATCCGTCTGGTTCAGGCAGAATGAGTTTGCAACTGCTACTGGGATTCTTTTCGCGATTGGAAATC TAGGAGCGATT
268	GAGGTATCGCCTACTTAGAGAGTTCGTAAAGTCGGAGATATTGGAGGAAGTTAAATTTGAAAAC GTTGTGGACGAGTACTGGGTTGCGGAACCATTCATAAAGATCATAATTTTTGAGGATCTCGAAA ACCAGAAATTGA
269	CTAATCCGATTATCGATTCTACGCTTCCTGATGGTAGCAGGCTTCAGGCTACCCTAGGAACAGA AATTACACCTAGAGGCTCGAGCTTCACGGTGAGAAAATTTACAACCCAGCCACTGACCCCGTTA GATCTAGTGAGG
270	CAAAATTATATCGATAGAGGATACCAGAGAGATAAAGCTCCATCATGAGAACTGGCTGGCTCA GGTGACGAGAACGGGGATAGGAGAGCAGGAAATTGACATGTATGACCTTCTCAAAGCCGCCTT GAGACAGAGACCGG
271	GAATCAGTTTGTTAAATGGGATGCGAAGAAAAATTCGCATGTTGAGGTAGGGATTCCGAAAAA GCTAGAGAAAATCGCGATGTCGAGAGTGGACGATGCTTACGCGGAGCTGGAAAGAAGAAGGA GGTATTTGGAGTGGA
272	TCAGTGAAGTTAGCACGGAATTCGAAAGGATAGTGGTTCTCGTTGAAATGGGAGAGGATTTGG AAAGCGCAATGAGGTTTGTTGCAGAAACAACTCCCTCAGAGAGGCTCAGGGTTTTTCTGGAGAA CTTTATTGATGTG
273	GCTGGAGCGGGAGGCGTATCAACGCTTGCCCTCAATCCGTTACCCGAAGTTCCAGAATACTTTG AGTATTTCCAGTCCGAATAGAAGCAGAGCACCTCTCGATCGACTAGAGTCTTTCTGCTAGCTCTT GCACCCTCATC
274	GCGGAAATCTCTGCTGAAAACACCTTGACTTTTTCTTCGTATATCTCCCATTCCATCAGGCACCA CCAACTTTGGTCCTGCAAAGAGTCATCGGTGCCCCATCTGCTACGGGAACGATCTGAAAGGCTT TACCACAGAAT
275	TCCGGTTGCAGGATTGGTCTCCCCACCTCTCGAGCCTATGAGGAATACCCCATTCCTGCAGAGCT CGAGAAGCTCTTCGAATTCAAGATCCCCCTTCTGCAGGAATGTGTTGCTCATTCTGACAATCGG AAAAGCAACTC
276	AACTCCTCGATTGTTGGGTCATCGATTATTGTCACGTTCTCTCCTGCAATTCTCTCTCCAATCTTT CCAGCAAGAACGCTGTTTTCCTGCAGAACGTGATCTGCCTCGACCGCATGCCCGAAAGCTTCGT GAATAAAAAC
277	CTTTCTTCAAGAATGCTTTCTGCGGCGATAAGCCCAGTAACAGCCGCTCCAACTATTCCCCTGCT TATTCCGGCTCCATCGCCAATTGCATAGATGTACGGTATGCTTGTCCTCATCTTCTCGTCAACCT TAAGCTTCAA
278	TGCTGGATTTTTCTTTGGCCTTGCTGTGGCCGTTGACTAGACAAAAGTCGCCGTACTCCTCTCTT ATAACCCAGCCCCTCGGGCAGGTGCAGAACGTGCGCATGTAGTCGTCATGCCTCTGTGTGATTA TTCTCAGCTTT
279	TTGCCTTGGAATTTTCCGCCACTTCGATCTTATACTTTTTTACCCATTTTTCCAGCCAGTCGGCAC CGCTCCTCCCAACTGCAATTATGAGTTTGTCGTAGCCGAACTTGTCCCCATCGTTCGTCTTCACG ATCTTTTCT

280	AATCACCACCGACTGTGAAGCTCGAAGGATAGTTGGGGTTGGCATAATTCAGCTTTCCATCCGA AAGTCCTCCAGCACCACCCACACCAGAAGTAATGTTGCAGGGATCGCATTTCTTGCAATAGCTT TGCGAAAGGTCA
281	TTAACCAACCTCTTTCGCATCAAAATCCCAACTGCGGCATCCGTTATCAGCGTTACATCGATTCC ATCTTTCATAAGCTCGTAGCAGGTGAGCCTAGAGCCTTGGTTCAGCGGCCTCGTTTCGCAGGCG AAAACCTTTAC
282	TTATCGAGTTAATAGCTATCAGTGTTGCTATTACGATCGTTGCGATCCCATCAAAGATGTTATGA CCGAAGGAGATAGCAATAATGCCAAATATCGCTGCAAGCGTCGATAAGGAGTCGTTAAAACTC TCAAACATCACT
283	CTCCCTTCTAAGCTTCGTGATATCTGCATTGCCAATATCAACTAGAAATTCGATTGAGATAAGCT TGTCTCTTGCGGTTAAGCTTGTTCTCTCGATATTTATACCGAAATTTAGCAATACACCCGTGATA TCTCTCACGA
284	AAGCGGGGCTTTTGCCTTTCCAATTCCGCCGCAACCAACCGTTGGACTTATCAAACCGGAACCT TTCAACTCCGAGATTAAAGAGCCTGGCTCCTTATCGTGCTTAATAGCAATTTCTACAATGTCTTC CCCGCATACAA
285	CTAGTTCTTGGTTTTCGTCGACGTTGACCTTGTAGAACTCTACATCTGGAAACTCCTTTGAAAGC TTTTCGAGCACTGGGCTGAGATACCTGCACGGCATGCACCAGTCGGCGTAGAAGTCAACAACAA CAAGCTTATCC
286	CTCCGATCGTCTTTAAAGCTTGCAAGTCTAAATCCTCGCCCCAGGGAATTTCCTGGGATTTTCTC GCAATCTCTATCGCCGAAGTATAGGTTATCCTCGGGAATGGTATCTCGGGGACTTCGAGCTTTA GTTCGAGAATA
287	GAGGTTTTGTCCCTATTGGGTTTCTCATTGCCTGCAGCATTTCTTCTCTGCTCAGAGCTCTGCAGC CATCGCCTTTCATTCTTAAAATGCTAACCTCCCAATCATCCGGAAAATCGAGCTCTATTTCCCTA TCCTGCCAG
288	TGTTTACAGGCTGGTGGGTGGGGAAAGGAGTGTTAAGGGCAAAAGGAGTGTAAGCAAGTTCAG GGTTGCGATTGCGATTCTTCTGGCATTCATTCTGATATATCCTACATACCGCATAGCCGAGATTC AAAGCAGTGGGG
289	CAGGGTCAGGAGGATTCACGAGATAGAAGTCCTCGAGGTGAGAGGCAGGTTCGCGCTTATAAG GGTTCTCAGCGACCCCGGCACGTACATGAGGAAGCTGGCCCACGACATCGGGCTATTGCTCGGA GTAGGTGCACACA
290	GAAGTCAGTCATAGAATCAATGGTGTATTCTTCATCAGGGTTATTATACGGAATGAACTTATAG TTCTCACCTGCTACCTGATCCACTGTCATTTCTGCAAGAGTCTGCACTGTGGTAATTCCACCTTCT TCCATCCGGG
291	AGTAAGGGAATCAATGTCTTCCATTGCTGTAAGGGTTACTGTTACCTTTGTAGAAGTCAGACCG TAATTGGTCAGCAGCTCTTCATAGAAGTTCTGGTTTGCAATATCCCTCTGGGCAATGACAGGGT AGTCGACTTCGT
	Primer and Core Sequence
292	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAAATGCTTTCAGTGGTTTCCAATGCCCTGACCAGC CTGTACTCAACTAAGGGATATGACGTAACATTCAGTGACCTGATCGCCGCCATTCAGGCAATGA AGGGCTACGATGACAGCGCAAACGCTAAACTCGTCGTGGACACATCATGTAGTAGACGACCAA GACAGT
293	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTCGGAGCTTATAACCACAGATCCTGAAGTATTG AGAAAGAGAAGGGGATGGTGAGATAAACATGAGGTGTGNNNAAAGCTGACAATATGTACGCG TGCTTGAAGGACACTGTTGTGAAGGAAAGGTATCCTGTCGCTACACATCATGTAGTAGACGACC AAGACAGT
294	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATGCAGTATATGGCAGGTGCGAGCAACTGCTTTACC ATGGGTGCCATGGTGCAGAACGGGAGAAGCTCCCTCTACTGGAAGGTTAAGGACGCAAATTTCT GGTCCAAGACTTTCGAGAGCAAGTTGCGCGTTCTGGGGCTCACATCATGTAGTAGACGACCAAG ACAGT

295	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCTGCGGAAGCCAGCAAGGAACCTTGTCTCTTCA GGAGAGGCAGGATATACGTCACATTCAATGTAAGGAGGGTGGCACTGAGCGGTGCGGGACAGT TGACTGCCGCCAGGACTTATGCCATAGGCAACACCGATGCGACACATCATGTAGTAGACGACCA AGACAGT
296	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGATGAGGTTAAGGGGATATGCGAGAGATTCGGCAA GGTCGTGGATGCCGTGAACTCTCCTGTGCTTACCGAGAGTAATGCCTCGTACAGGAATGCGGGG CTGGTGCGTGCCAGGTTCAACTGGGACTACATCAGGCCCGACACATCATGTAGTAGACGACCAA GACAGT
297	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGGTCTCCACGGAAGGCTACATGGACAGGGCAATA GGCGTCCAGGATATCGGCTACCTGTTCTGGCAGGCAGGTCCCACCGCAATGAAGGATATGAGA ATTTACAACGGTCCCGGTGGTCTGATCGTTCTGCCTTTCTATCACATCATGTAGTAGACGACCAA GACAGT
298	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTTCTGCGCGCAACAGGCTGGCCAAGGGACTTCCG AAGAGCCTGGACATGTTTGCCAGCGTGGAAGGTCGTGACCTTGGGTACGATCCGAGGTACATAA CAGAGGAAGATTACAAGACCATTATGACCAAGGCCCGTCTCACATCATGTAGTAGACGACCAA GACAGT
299	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCATGATAGATGAGGATGGCTACGAGGTACCGAA GGGTGAAGACCCCAACGACCCCAGAAGTGCACACACCTTTGGTTGGGTCGACCAATCAGATGG AGGCACATCCAATGGTGGCATGCAGTCCGGTGGGAGCTCTCACACATCATGTAGTAGACGACCA AGACAGT
300	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTAAGGACGGGGACGTGACAAAGCTGTATTCTCA GGACGATTACCTCAGGGTCAGCAGGCTCAAGTTCAGCGAGAATCCGATGCTTGGCATCGTCAAG AATACGGATGGCACAGGGGAGGTTATAGGTCCGTCCTTTGCCACATCATGTAGTAGACGACCAA GACAGT
301	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGGACCTTCTGTGGAACATAATCTCGGGTGCCCTG AATGCGGGAAGGGAACAGCTCTACGGGGATGCATTCGGCGGTCCTAAGATAGAGCAGTACGTG AAAGCACTCACGCAGGTGCTGTATGACCTGTCTGTCAACAGCACATCATGTAGTAGACGACCAA GACAGT
302	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGAAGGAACGCGTCATCGTCCACAAGAGCACTACG AGGAACGAAGTCCTTGAAGAGTTCAAGGCGTCTAAGGAGCCGAAGGTCCTATTTGCGATAAAG ATGGAAGAGGGTACGGATTTCAGGGATGACCAGGCAAGGTGGCACATCATGTAGTAGACGACC AAGACAGT
303	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGATATTGGTCAAGACTCCTTACCAGGATCTGGGA GACGAGTGGGTCCGCCTCCATAGGGAGAAGATGGGACGGAGATGGTACGAGATATCCGCCCTC CAGCAGGTCATCCAGGCGAGCGGCAGGATAATGAGGAACGACACATCATGTAGTAGACGACCA AGACAGT
304	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGGGACTGGGGAGACACCTACGTCCTTGACATGA ACGCCATGAAGCTCATCCGCATGTACGAAAAGGAATGCCCCCGCTGGTTTTTGAAGAGGTTGAA ACTATGACGCATCACAATATCACCTTCCCCGTCCCTCCCGACACATCATGTAGTAGACGACCAA GACAGT
305	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCAATGTCCGCAACAGTTGACAACGTTGCACGGGT CGCGGGATGGCTCAGGGCGACTGCAGTCTCCAGCGATTTCAGGCCCGTCACCCTGAAGAAGTAC GTCCTTACCCCCCGCCATATCATGGATGAGAAAGGGGATACCACATCATGTAGTAGACGACCAA GACAGT
306	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAGATACAGCAGATGCTGGGGCGCGCCGGGAGGG CGAAATACGATTCCATGGGCTACGGCTACATCTGCTCATCCGACGTTCACCTCCAGGACGTGTA TAAGACGTACGTTCATGGCCGTCTGGAGAGCGTAAAATCAACACATCATGTAGTAGACGACCA AGACAGT
307	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCACTGGACGAGTTCTTCTCCACCACGCTGGCACG CCACGAGGGTGCCCGTCTGGAAGAATGGATAGACAACAGCCTTGTCTTCCTGCAGGACAACGA CATGATAGTCGGGGGACGTTCCTTCACGGCTACCCCCTTCGCACATCATGTAGTAGACGACCAA GACAGT

308	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCATCCTTGCGGACTGGATAGACGAGAAGCCCGAA AGCGACATCGTCAATAAATACAACATCTGGCCTGCCGACCTGAGGAGCAGGGTTGAGTTGGCC GAATGGCTCTCGCATTCCCTTTACGAGATCTCGAGGGTCCTGCACATCATGTAGTAGACGACCA AGACAGT
309	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCGACTATTCCTACGTCAGCGTTGCGGAATATTTC TCCAGCTCAAGGATAATAGCCACTACCGCTTCCCCCGGTGGCGACAGGGAGAAGATAAACGAG ATCATGCGCCACCTGAGAATAGAGAACCTTGAGGTGAGGGACACATCATGTAGTAGACGACCA AGACAGT
310	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATCGACGCTCCAGCTGTTCAGGGACGGTGCGGTCA GGATACTCGTAGCAACGCAGGTTGGGGAGGAAGGACTGGACGTACCGGCTGCAGATACCGTCA TATTCTACGAGCCGGTGGCAAGCGAGGTCCGCTCAATCCAGACACATCATGTAGTAGACGACCA AGACAGT
311	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCGGTCCTCTTTCCCAGCTTCAGTCCTTTGGCTTTC CTATGTTCTGCTGGTACTCTTCCCATTGCTCTCTCTGTTGTTTCTCCTGGCTTTTCCTGAAGTTTTC GAGGGCTTCGTCCATGCTGTCATAAGCGTTATGCGACACATCATGTAGTAGACGACCAAGACAG T
312	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAACTCCTTGACGCTCCTGGAGATGACCAGTTTCT CTATCTCCACCCTCCCGCTCTTCATGTCCGATATTATTTTCCTCGCCCTTCTCAGTGCCTCGTCCA CGTTCCGGTCGAGCACGAGATTGAACATTTCCATGAGTCACATCATGTAGTAGACGACCAAGAC AGT
313	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCCTTTCCTCGAGGTCTTTGTCTATTCCCTTCACCGT TTCCCTGGCCCAAGCAGTGATGCTGGACCCGATACTGGGATCGGTAAACCTGTAGAAGCTTGAT GCAAACACTCCGTAGAACGAATTCATAAGCACCTTCACCCACATCATGTAGTAGACGACCAAGA CAGT
314	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGTTCTCCTTCCACCCCGTCCGCCTCTTCTATGACG GTCGCACTTTCCGTCGAGCAGACGCCCTACGGCTGGATGGATCAGTATCCCTCGTCGGTTGTTGC CCATGTCACGGGAGGCATCCCTCCCTACGCCTATCACTCACATCATGTAGTAGACGACCAAGAC AGT
315	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAATCTGGAGGCTTTCGCCGTATCCGTCGAAGTCAC GGACTCATCGGGGCATTCAGTCTCAGGTGCAATCATGATCAACTACGGTTCCATAGACCTCTCC CCGTTTGGATACATGGTAACTTTGATTTTTCCGGTGATCACACATCATGTAGTAGACGACCAAG ACAGT
316	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAACATCCATTGCTGACCATCACGATCGATGGCTCA AAGGACACGTTCAAAACTGGCGATGTCCTGGAATGGTTGACCGAAAGTGACATCTCAAACATG CACAATGTTGCGTCCTTCACAAAATCTCTCCTGAGGATAGTCACATCATGTAGTAGACGACCAA GACAGT
317	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCATGCTGCCCGATGGCCAATGGTTCGGGGAGGTC ATTGGGAAGGACGTGCAGGGAAATCCCTACGGCATTGATTATACAATGTGGTTGCCGTTTAACA CCTACGTTAGGGATAAGCTCAGTTACAATAGTTGGGGGAAGCACATCATGTAGTAGACGACCA AGACAGT
318	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCAGCATTCCTCCTCTGAGGGAGTTCGGAAGGTAA AATCTCCTGATGACATCTGAGTCCCTGGCGCCCATTGTCTTGGCTGAGTAGACTTCCATTTTACT TGTCGTGCTGCCAGTTGAAAATGCAAGTACTGCTATCATCCACATCATGTAGTAGACGACCAAG ACAGT
319	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCCCTGCAGGAATGATTGTTCAACTGCACTCGTC AGTACATGATAGAACAATCTTGAAGCTGACAGATGGAGCGCATAGAAGATAAGGAATGCAAGA GGTACCAGTACTAGTACCACTGCATATATTCCTGCGTATAGCACATCATGTAGTAGACGACCAA GACAGT
320	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGATTGCAGGAACGTGGAAAGTGTGCGGTTAATGT ATGACTCATTGCTCACATCGAATCGATACAGAAGACCGCTGTTTGCTGCCAGGTAGGAATCTAT GTTATTCAGGCCAACAACCACATCGTATCCCGCCCCCTTTGCACATCATGTAGTAGACGACCAA GACAGT

321	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGATTGAACCGGTTGGTGTCAACGTAGAATCCTC ATTGACCCGCCACATAAAACTGAACATTCCAATTGTGTCGTCTCCTATGGATACGGTCTCTGAG GCAGATATGGCAATTGCACTAGCAAGACTCGGTGGTATTGGCACATCATGTAGTAGACGACCAA GACAGT
322	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTATTATACGCGACCTTTACACTGTAAGCCCGGAA ACACCTGTTGACGATGCAATCCGTACTATGAGGGAGAAGCGAATCGCTGGGCTCCCAGTGATAT TGAACGGCAAACTTGTCGGAATACTTACGAACAGGGACATCACATCATGTAGTAGACGACCAA GACAGT
323	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGTACGGCAAGATAGGCTCAGGGAAATTTGTACCA GAGGGAGTTGAAGGAGCAGTTCCGTACAAAGGTAAAGTTGCAGATGCAGTCTTTCAATTGATCG GGGGCCTGAAGTCGGGGATGGGGTATACTGGCTCGCCCACACACATCATGTAGTAGACGACCA AGACAGT
324	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGTGGAAGCGTTGAGGAGTTTGTCACTCTATCGAG GAGAGTGGAGGCAGCGGGATTCGACAAGGTCGAGCTCAATTTGTCCTGCCCACACGTTCAGGG AGTTGGATCCGAGGTAGGACAGGATGTAGGTCTTGTAGAAGACACATCATGTAGTAGACGACC AAGACAGT
325	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACACTTATAGACAGGCTAGACAAGAAGACGAAGA CAAGGATATTCTTCTCACTTGAGCGATTGATGAAGTGCGGCATAGGGATTTGTGACAGTTGCAG CATCAACGGCATCCGGGTATGCAAGGACGGAACAATTTTCGCACATCATGTAGTAGACGACCA AGACAGT
326	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTCGCAACTGCAAAGAGGTAGCTTCTGGATGCTTC CCTGGAACTATCCCTACATTGCTGTTATCTTACTAGTGGTACTGATTTATGCAGCAATAGAGGAC CTTAGGAAGAGGAAAATAACAACTATAACCTTCCTTGCACACATCATGTAGTAGACGACCAAG ACAGT
327	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGACAGTTGGAACTGGTCTATCTCCCCGGTATTTT AATAAGTTTATAGGCGTAGCAAAGGCATATACGACAAGAGTAGGGGAGGGGATATTTCCTACT GAGATGTTTGGGGAAGAGGCAGATAGACTTAGAACCCTAGGCACATCATGTAGTAGACGACCA AGACAGT
328	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAGAAGACTTAAAGGATTTAGGTAGAGAGCTTAA GGTACCAAGAAGACCGTTCAAAAAGTTAACGCATAGAGAAGCTGTTNATATATTGAGATCTCAT GGCATAAAAGCAAGTTATGAACATGAGATACCTTGGGAAGCCACATCATGTAGTAGACGACCA AGACAGT
329	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACGGGGAGGCTGTCTCAGGAGCTGAAAGAGAATAT AGAGCGGAGAAGGTTATTGAGAGGATGAGAGCTACTGGTGAGAACCCTGCAAAATACGGTTGG TACATTGAAATGTTGAAATATGGTATTCCGCCGAGTGCAGGGCACATCATGTAGTAGACGACCA AGACAGT
330	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATATGCAGATTTAGATGAGATTATAGGGGTTGCATC TAAGGCAGGAATAGATTGCATAACTATAGATGGGTCAGAAGGTGGAACAGGTATGAGCCCTAT AGCTGCGATGAGAGAACTAGGATATCCAACGCTAGTATGTCCACATCATGTAGTAGACGACCA AGACAGT
331	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGACACGAAATTGCTGAAGCAGCTGGCTCAACATG GTATATCGACAATTTCTGGGATAAACTCAAAGAGGGCTGTGTAGCATATCTAAACATAGATTCA CCTGGATTAAAAGATGCAACAAGATATATCGCTTACGCGTCCACATCATGTAGTAGACGACCAA GACAGT
332	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTAACTTCTGGAAACGCCCAATCAAAACAGATCAT GACACCAAAGCTAAAATTATCTTCCCTAATAGCTTCTATAGGTGTATCTCCAGGTTGAAATATTA GCTTCTCTTTGGCAAATAAGTGAAGTTTCCTATACTTTCCCACATCATGTAGTAGACGACCAAGA CAGT
333	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCAGATAGCCCAATAGCATCAATTTCCGTTGCAATA ATAGGTACAGTACACAAAGAACACGTAATTTTCAGCGACACTGCAAATACAGGCGACTTAATA ATTTTTGCCATAGATCTCGATGGAACATTTCACCCTAAGTTCACATCATGTAGTAGACGACCAA GACAGT

334	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCTAATTCCTCTCTTACAGCTTTAAAAGCAATCA CAGCAGATTCCAAAATATCATCCATATCATCCAGAGCTATAATAACACCTCTTGAAGTTTTCCCA ATCTTATGCCCACTTCTTCCAACTCTTTGAACCAAACGACACATCATGTAGTAGACGACCAAGA CAGT
335	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTAACTTGTCTGGGAGACATATATTGGACAACTAA ATCAACGGTTCCTACATCAATCCCTAACTCCATAGATGATGTACAAATAAGACCTTTCAACTCA CCGTCTTTAAATAACCTTTCAACTTCTATACGAACATCTCTCACATCATGTAGTAGACGACCAAG ACAGT
336	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACTCAATGAACCATGATGCACATCAATACTTAGATT AGGATCGTATAAGTGAAGCCTAGAAGCTAGTATCTCAGCTATTTCACGAGTGTTTACAAAAGTA AGCATAGAGCGGCTCTTTTCTAATAACTCAACCAATACCCCACATCATGTAGTAGACGACCAAG ACAGT
337	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGTTAAATCATCTTAACTCACAAATATTAAGGCTT TAATTTCTGAGGGAGTGCAAAATGAAAACTGACGTAGTAATAGTAGGTGCAGGGCCCGCAGGC ATGTTTGCTGCACATGAATTGGCAACTAAATCTAATCTGAACACATCATGTAGTAGACGACCAA GACAGT
338	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAAAAATAGCCAAGGATCCAAAATTCCGTGTATATA CAAAAACCTTCGATGACCTTACACGTGTATTTTGCGTTAATTATCGAGGCTTCGTCGTCCAAGAA GTCTACGGAGATATCGTTGGTGTTAACGGCCACACTCTAACACATCATGTAGTAGACGACCAAG ACAGT
339	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCAAACAAAAATCTGAAAATGCCAATTTTGCATTTC TAGTTCGAGTTGAACTCACCGAACCGCTTGAAGACACAACCGCCTACGGATTCTCAATAGCCAA ATTAGCAACTACCATAGGTGGAGGAAAACCAATTCTTCAACACATCATGTAGTAGACGACCAA GACAGT
340	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAGATACTGAATTTCCAAAACTCAAAGGATATAG AATTGTTAGAATCGCAACACATCCGCAAGTTATGAGCATGGGACTAGGAAGTGAAGGGTTGTC AAAACTTTGCCAAGAAGCCGAAAAGAGAGGACTAGATTGGGTCACATCATGTAGTAGACGACC AAGACAGT
341	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAAGTTTTTATCCTCCTCGGTCCAAGTCACACTGG TTACCCAGGCGTTGGAATAATGACAGAAGGCATCTGGAAAACTTCTTTAGGAGAAATATCAATA GATGAAACTCTCTCGAATACTATTTTAAATAATTGTGACCCACATCATGTAGTAGACGACCAAG ACAGT
342	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGACACACTACGGCACCTACTATGGATACACACCA GCTGGTGTTGAACCATTAACCAAAGTTTTAGAATGGATATACCAGACGGACAAACAAGTTATTG AGAGAATTAAAAGATTAGATGGAGCAGGAGTAATAGAATATCACATCATGTAGTAGACGACCA AGACAGT
343	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTGAAAAGTTCATTCCAATTGTTAAATCGCCATCTT GGAAACACGGCACAAGAAAAGGGAAAGGATTTAGCATCGGTGAGATTAAAGCAGCCGAGATA GATATTAGTATGGCAGTTAAACTCGGTATACCCATTGATAAACACATCATGTAGTAGACGACCA AGACAGT
344	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGGAATAATAATTAAAATAATGTGGCACACCTTTTA GCTTCTTTTCATCTCATATTTTCAAAGAAGCCTTCCAGGTGTGCCTCATCGGTGTCCCCCGCTGC GGAGACACGGTATCATCGTATCCGCCGAAGGAAACTCAACACATCATGTAGTAGACGACCAAG ACAGT
345	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACATTGCCTATCAATTACTTCAAGCCGGAATGCAA GTTCCCGGTTTCAGAAGGTCGCCAAAGATAATAGAAAGAATTTTAGAAAGATATATTCCAACAG TCACCGTACTAGGCGGCATTATTGTAGGATTAATAGCTGCCACATCATGTAGTAGACGACCAAG ACAGT
346	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTCGTTCAGGGAGGTATAAAAATGCCAGAACCAC GCTACCGGTCAAGGTCTTTAAGAAGACGATACGTACACACACCTGGAGGAAAAACCGTCATCC ATTACAGGAGAAAAAAACCTGACGTTGCAAAATGCGCATTATCACATCATGTAGTAGACGACC AAGACAGT

347	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGTCAACCTCTCAGAGGAATTCCCAGACTAAGG CCAGGAGAATTCAGAAAGTTGACAAAAAGTCAACGAAGACCAGAGAGACCTTTCGGTGGATAT CTATGCCACAAATGCTTAGCAATGGAAATCAAGAAAGCTGTTCACATCATGTAGTAGACGACCA AGACAGT
348	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATAGGATGAATCTAACTGGGGCGACCCGGTAGATA ACTGAGAGTGTAGGAGGTGAAATAATTGAGCGCAATAGAAGTAGGTAGAATATGTGTTAAAAC TAGTGGAAGAGAAGCAGGAAGAAAGTGCGTTATTGTTGAAATCACATCATGTAGTAGACGACC AAGACAGT
349	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACACCATTTCCTAATATTTTAGTAACTAGATATGTT TGTTATAGTATTAGGGTGAAGTATTTGTATGAAAGAAAGTTGCCATCAGACATTAAAAGAGAGA TTCTAGTAAAAAGTGAAGCAGAAACTGACCCTGCTTATGGCACATCATGTAGTAGACGACCAAG ACAGT
350	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCACATGAGAGAACTTAGAAGAACACGTACAGGACC CTTTAAAGAAGATGAAACCCTAGTAACTCTTCACGATGTAGTTGATGCTTACTATTTTTGGAAGG AAGATGGAGAAGAAGAATTTCTACGAAAAGTCATACAACCCACATCATGTAGTAGACGACCAA GACAGT
351	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAATGGAAAAGGGTTTAGAACACCTACCTCACATTT GGATTAGAGATTCTGCTGTAGATGCAATATGCCATGGGGCAAACTTAGCAGCTCCTGGTGTTGT AAAACTTCATGACGGTATATCACCTGGAGACTTAATAGTAACACATCATGTAGTAGACGACCAA GACAGT
352	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCTGATCATACATGTGCATTGTCTTTAAATACACT AGTAACGTTAATAATATCTAGCAATTTTAGATAAAAATAACTAGCAGTGCCGGGGTAGCCAAGT GGACTACAGGCCTTATACCGGTTAGGGCGCGGGCCTGGAGCACATCATGTAGTAGACGACCAA GACAGT
353	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATGCCTTAACGAGAGGCATGGGATGGGGGAGCTG TGAGCCCCCCGAACCGGCAGATGAGGGGAAGGGTGCAAAGCATCCCTTAACGCCGGAAGCTCC CGACTTCAGTCGTGGAGCAGCTCACTGCTTTGACGAAAGGTTCACATCATGTAGTAGACGACCA AGACAGT
354	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAACTTGCAAGGAAGGCCGGTGTTGATTATGAGAC AAAGCTGTTGGTCAGGGGCAAGGAACCGGCTGAGGACATAATAGAATTTGCTGACGAGATCAG GGCAAGTCTCATTGTAATAGGGGTTAGGAAGAGGAGACCCGCCACATCATGTAGTAGACGACC AAGACAGT
355	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCAGAAGAGATTCAAAGCTCTCGTATTCAATGTCC CCACCAAATTTCTGGTCGCGCTCAATTTTGACTTTACCAAAAGCGGGGAAAACGTAGTGCTTTG CTAGGTCTATTATCGGATTTCCTTCTACAACCTTTGGCGGCACATCATGTAGTAGACGACCAAGA CAGT
356	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGATTTGCTCATTTTCTCCCCGTCGAGTCCTGAGATT ATCGGCGTATGGATGCAGATCGGTGCCTTGTAACCGAGGGCCGGCAGATTCTCCCTTGCGAGCA TGTGGATCTTTCTCTGATCTATTCCACCAACCGCCACATCCACATCATGTAGTAGACGACCAAGA CAGT
357	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCGGGAGTTGCAGAACCAAGCATGGAAATTGCTA GAGATCCCGAAAAGGTTTACGAGTACACGAATAAGTGGAACACGGTTGCAATTATCACTGATG GCTCGAGGGTCTTGGGACTGGGCAACATCGGTGCGATGGCTTCACATCATGTAGTAGACGACCA AGACAGT
358	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGTGTTATCAAGAGGGAATATATTGCTCAGATG GCAGAGGATCCGATAGTCTTTGCCTTATCAAACCCGGTGCCTGAGATCTATCCGCAGGAGGCAA AGGAAGCCGGAGCCAGGATCGTAGGAACTGGTAGGAGCGACCACATCATGTAGTAGACGACCA AGACAGT
359	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGGATCTGTTAGTATGGCATTCAGAGCCTTTATGTC CTCATCGGTAAGCTTGTCCGATGGCAGATCGTATTTCACGATGTCTGAAGGAGTAACTCCGAGA AACTTCGCTTCTGGTGTCGCAAGATACTCCGAGAGATGCGCACATCATGTAGTAGACGACCAAG ACAGT

360	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGAGTGCGGCTTACTCTGCACTGTGCGAGATCGATG AGGTCGTTGTTGTTGCCCCCATAACGCAGATGAGCGGAGTGGGGAGGAGCATATCCATAATGCG GCCGGTTCGTTTTTTCGAGCTCGAAATAGATGGCATGAGGCACATCATGTAGTAGACGACCAAG ACAGT
361	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGGGAAGGGAGTACTACTGGATTCATGGGGTGGA AGTCGAAAGCGCTGAGCCTGGAACGGACATACACGCACTCAGAAACGGGTATGTCTCCATTAC ACCGATATCCTTAAATGCAACTTCGGACTGCGAAGCTTTAAGCACATCATGTAGTAGACGACCA AGACAGT
362	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATAGTTTTATGGAGGGTGGTTGGACATGAATGAAA GGGCAAAGAAGGTCATTCTTATTGTGGATGACGATTTGGCTCTGCTTGAAGCTCTTGAACTGAT GCTTCGAGGCAAGTATGAGGTTGTGAAGGTGACAAATGGGACACATCATGTAGTAGACGACCA AGACAGT
363	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATGTCGATTCCGAAATAGCAGGGAGCAATTATCGG TGGGCTTCCGACCCTTAAATGGATTTCCTTCGCTCCCGCCTTTCTTATCATGTCGACTATTCTTTT GGATGTTGTTGCCCGCACAATGCTGTCGTCAACCAGCACCACATCATGTAGTAGACGACCAAGA CAGT
364	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACTTTCTGAGGGAAAAACATTGTTGCTTATCCTAAA GAGTTTACAAGCAAGAAGCTGGAAACAAACTCTGGATGTTATTAATTTAGAGCCTGCAGCAGCA TATACAATGTTTAGAGCGGCAATAAAGAAACTATACAAAGCACATCATGTAGTAGACGACCAA GACAGT
365	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGTTGAGAGGCTGCTTGAAGGCATTGCAAAGAA TGAAAGGGTAGCTTACGGATTGGAGGAGGTTAGGAGGGCAAAAGAGTATGGAGCAATTGAGGT TCTGTTGGTTTCAGATGACTTCCTGCTCACCGAGCGTGAGAACACATCATGTAGTAGACGACCA AGACAGT
366	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGCTTCGAGATTCCTGATAGGAGTGGGAGTTGCCG GGGTTTACGTGCCTACGATAAAAATAATATCCGTCTGGTTCAGGCAGAATGAGTTTGCAACTGC TACTGGGATTCTTTTCGCGATTGGAAATCTAGGAGCGATTCACATCATGTAGTAGACGACCAAG ACAGT
367	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGGTATCGCCTACTTAGAGAGTTCGTAAAGTCGG AGATATTGGAGGAAGTTAAATTTGAAAACGTTGTGGACGAGTACTGGGTTGCGGAACCATTCAT AAAGATCATAATTTTTGAGGATCTCGAAAACCAGAAATTGACACATCATGTAGTAGACGACCAA GACAGT
368	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTAATCCGATTATCGATTCTACGCTTCCTGATGGTA GCAGGCTTCAGGCTACCCTAGGAACAGAAATTACACCTAGAGGCTCGAGCTTCACGGTGAGAA AATTTACAACCCAGCCACTGACCCCGTTAGATCTAGTGAGGCACATCATGTAGTAGACGACCAA GACAGT
369	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAAAATTATATCGATAGAGGATACCAGAGAGATAA AGCTCCATCATGAGAACTGGCTGGCTCAGGTGACGAGAACGGGGATAGGAGAGCAGGAAATTG ACATGTATGACCTTCTCAAAGCCGCCTTGAGACAGAGACCGGCACATCATGTAGTAGACGACCA AGACAGT
370	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAATCAGTTTGTTAAATGGGATGCGAAGAAAAATT CGCATGTTGAGGTAGGGATTCCGAAAAAGCTAGAGAAAATCGCGATGTCGAGAGTGGACGATG CTTACGCGGAGCTGGAAAGAAGAAGGAGGTATTTGGAGTGGACACATCATGTAGTAGACGACC AAGACAGT
371	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCAGTGAAGTTAGCACGGAATTCGAAAGGATAGTG GTTCTCGTTGAAATGGGAGAGGATTTGGAAAGCGCAATGAGGTTTGTTGCAGAAACAACTCCCT CAGAGAGGCTCAGGGTTTTTCTGGAGAACTTTATTGATGTGCACATCATGTAGTAGACGACCAA GACAGT
372	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCTGGAGCGGGAGGCGTATCAACGCTTGCCCTCAA TCCGTTACCCGAAGTTCCAGAATACTTTGAGTATTTCCAGTCCGAATAGAAGCAGAGCACCTCT CGATCGACTAGAGTCTTTCTGCTAGCTCTTGCACCCTCATCCACATCATGTAGTAGACGACCAAG ACAGT

373	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGGAAATCTCTGCTGAAAACACCTTGACTTTTTCT TCGTATATCTCCCATTCCATCAGGCACCACCAACTTTGGTCCTGCAAAGAGTCATCGGTGCCCCA TCTGCTACGGGAACGATCTGAAAGGCTTTACCACAGAATCACATCATGTAGTAGACGACCAAGA CAGT
374	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCGGTTGCAGGATTGGTCTCCCCACCTCTCGAGCC TATGAGGAATACCCCATTCCTGCAGAGCTCGAGAAGCTCTTCGAATTCAAGATCCCCCTTCTGC AGGAATGTGTTGCTCATTCTGACAATCGGAAAAGCAACTCCACATCATGTAGTAGACGACCAAG ACAGT
375	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAACTCCTCGATTGTTGGGTCATCGATTATTGTCACG TTCTCTCCTGCAATTCTCTCTCCAATCTTTCCAGCAAGAACGCTGTTTTCCTGCAGAACGTGATCT GCCTCGACCGCATGCCCGAAAGCTTCGTGAATAAAAACCACATCATGTAGTAGACGACCAAGA CAGT
376	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTTCTTCAAGAATGCTTTCTGCGGCGATAAGCCCA GTAACAGCCGCTCCAACTATTCCCCTGCTTATTCCGGCTCCATCGCCAATTGCATAGATGTACGG TATGCTTGTCCTCATCTTCTCGTCAACCTTAAGCTTCAACACATCATGTAGTAGACGACCAAGAC AGT
377	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGCTGGATTTTTCTTTGGCCTTGCTGTGGCCGTTGAC TAGACAAAAGTCGCCGTACTCCTCTCTTATAACCCAGCCCCTCGGGCAGGTGCAGAACGTGCGC ATGTAGTCGTCATGCCTCTGTGTGATTATTCTCAGCTTTCACATCATGTAGTAGACGACCAAGAC AGT
378	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTGCCTTGGAATTTTCCGCCACTTCGATCTTATACTT TTTTACCCATTTTTCCAGCCAGTCGGCACCGCTCCTCCCAACTGCAATTATGAGTTTGTCGTAGC CGAACTTGTCCCCATCGTTCGTCTTCACGATCTTTTCTCACATCATGTAGTAGACGACCAAGACA GT
379	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAATCACCACCGACTGTGAAGCTCGAAGGATAGTTG GGGTTGGCATAATTCAGCTTTCCATCCGAAAGTCCTCCAGCACCACCCACACCAGAAGTAATGT TGCAGGGATCGCATTTCTTGCAATAGCTTTGCGAAAGGTCACACATCATGTAGTAGACGACCAA GACAGT
380	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTAACCAACCTCTTTCGCATCAAAATCCCAACTGCG GCATCCGTTATCAGCGTTACATCGATTCCATCTTTCATAAGCTCGTAGCAGGTGAGCCTAGAGCC TTGGTTCAGCGGCCTCGTTTCGCAGGCGAAAACCTTTACCACATCATGTAGTAGACGACCAAGA CAGT
381	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTTATCGAGTTAATAGCTATCAGTGTTGCTATTACGA TCGTTGCGATCCCATCAAAGATGTTATGACCGAAGGAGATAGCAATAATGCCAAATATCGCTGC AAGCGTCGATAAGGAGTCGTTAAAACTCTCAAACATCACTCACATCATGTAGTAGACGACCAAG ACAGT
382	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCCCTTCTAAGCTTCGTGATATCTGCATTGCCAAT ATCAACTAGAAATTCGATTGAGATAAGCTTGTCTCTTGCGGTTAAGCTTGTTCTCTCGATATTTA TACCGAAATTTAGCAATACACCCGTGATATCTCTCACGACACATCATGTAGTAGACGACCAAGA CAGT
383	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAAGCGGGGCTTTTGCCTTTCCAATTCCGCCGCAACC AACCGTTGGACTTATCAAACCGGAACCTTTCAACTCCGAGATTAAAGAGCCTGGCTCCTTATCG TGCTTAATAGCAATTTCTACAATGTCTTCCCCGCATACAACACATCATGTAGTAGACGACCAAG ACAGT
384	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTAGTTCTTGGTTTTCGTCGACGTTGACCTTGTAGA ACTCTACATCTGGAAACTCCTTTGAAAGCTTTTCGAGCACTGGGCTGAGATACCTGCACGGCAT GCACCAGTCGGCGTAGAAGTCAACAACAACAAGCTTATCCCACATCATGTAGTAGACGACCAA GACAGT
385	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCCGATCGTCTTTAAAGCTTGCAAGTCTAAATCCT CGCCCCAGGGAATTTCCTGGGATTTTCTCGCAATCTCTATCGCCGAAGTATAGGTTATCCTCGGG AATGGTATCTCGGGGACTTCGAGCTTTAGTTCGAGAATACACATCATGTAGTAGACGACCAAGA CAGT

386	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGGTTTTGTCCCTATTGGGTTTCTCATTGCCTGCA GCATTTCTTCTCTGCTCAGAGCTCTGCAGCCATCGCCTTTCATTCTTAAAATGCTAACCTCCCAA TCATCCGGAAAATCGAGCTCTATTTCCCTATCCTGCCAGCACATCATGTAGTAGACGACCAAGA CAGT
387	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTTTACAGGCTGGTGGGTGGGGAAAGGAGTGTTA AGGGCAAAAGGAGTGTAAGCAAGTTCAGGGTTGCGATTGCGATTCTTCTGGCATTCATTCTGAT ATATCCTACATACCGCATAGCCGAGATTCAAAGCAGTGGGGCACATCATGTAGTAGACGACCA AGACAGT
388	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGGGTCAGGAGGATTCACGAGATAGAAGTCCTCG AGGTGAGAGGCAGGTTCGCGCTTATAAGGGTTCTCAGCGACCCCGGCACGTACATGAGGAAGC TGGCCCACGACATCGGGCTATTGCTCGGAGTAGGTGCACACACACATCATGTAGTAGACGACCA AGACAGT
389	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAGTCAGTCATAGAATCAATGGTGTATTCTTCATC AGGGTTATTATACGGAATGAACTTATAGTTCTCACCTGCTACCTGATCCACTGTCATTTCTGCAA GAGTCTGCACTGTGGTAATTCCACCTTCTTCCATCCGGGCACATCATGTAGTAGACGACCAAGA CAGT
390	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTAAGGGAATCAATGTCTTCCATTGCTGTAAGGGT TACTGTTACCTTTGTAGAAGTCAGACCGTAATTGGTCAGCAGCTCTTCATAGAAGTTCTGGTTTG CAATATCCCTCTGGGCAATGACAGGGTAGTCGACTTCGTCACATCATGTAGTAGACGACCAAGA CAGT
	Forward and Reverse Primers
391	TCTCCTTCTTAGCTTCGTGAGAAC
392	CTTGGTCGTCTACTACATGATGTG

Primer Binding Sequences

[0069] The 5' and 3' primer binding sequences are selected to be complementary to a SDSI 5' and 3' primer which is included in an amplification reaction and used to amplify SDSIs present in a given sample. The primer binding sites may be optimized for multiplex amplification with a set of primers used to amplify a genome for sequencing. In one example embodiments, the 5' and 3' primer binding sites have a Tm of between 55-65 ^{o}C. In one example embodiment, the 5' and 3' primer binding site are complementary to primers having SEQ ID NOS: 391 and 392.

Methods of Detecting and Preventing Sample Contamination

[0070] In one example embodiment, a method of detecting and preventing contamination in one or more amplification reactions comprises adding a SDSI according to the example embodiments disclosed above to a one or more samples to be assayed. An amplification reaction is then used to amplify a target sequence in the samples. The amplification reaction will include probes and primers needed to amplify the target sequence and to amplify the SDSI. The amplicons generated from the amplification step are then used the one or more samples, sequencing the amplified samples and determining the number of reads of the SDSI from the one or more samples,

wherein detection of only a single SDSI in the sample indicates contamination free amplification of the same, and wherein detection of multiple SDSI's indicates possible contamination of the sample. Samples identified as potentially contaminated may then be discarded or marked for repeat to confirm accuracy of results.

Amplification

[0071] The present invention solves this problem by providing for the sequencing of spike-DNA sequences at concentrations that can be amplified concurrently with the nucleic acids of interest. In one example embodiment, sequencing includes extracting total RNA or DNA from a biological sample, such as a sample collected with a swab (e.g., nasal, rectal, vaginal). Methods of extracting total RNA or DNA are known in the art and commercial kits are available. The presence of a pathogen may be confirmed in a sample. Exemplary methods for confirming include PCR, RT-PCR and RT-qPCR. In certain embodiments, sequencing includes DNase treatment to remove residual DNA. In certain example embodiments, sequencing may include depletion of ribosomal RNA (rRNA). In certain example embodiments, cDNA may be prepared from total RNA using RT-PCR. In certain example embodiments, RT-PCR may be performed using random hexamer priming. In one example embodiments, a SDSI is added to each cDNA sample. The SDSI can be added to the total cDNA sample. In certain example embodiments, cDNA samples may be normalized to a constant amplification level. In certain example embodiments, real time PCR may be performed on the cDNA using one or more standard primers and a Ct value is used to normalize cDNA samples. As used herein, standard primers refer to a primer set that is used for every sample. In certain embodiments, the standard primers are directed to a region of the pathogen to be sequenced. The samples can be diluted such that all of the samples for amplification have the same Ct value in the amplification reaction. In certain embodiments, each sample is normalized to a Ct value less than 35, 34, 33, 32, 30, 29, 28, 27, 26, 25, or 24. In preferred embodiments, the samples are normalized to a Ct value of 26 to 28, preferably 27. In one example embodiment, a SDSI is added to the normalized sample used for PCR amplification of the pathogen. The cDNA may be amplified in the same reaction with pathogen specific primers and primers specific to the SDSI. Amplification may be performed in a multi-well plate (e.g., a standard PCR plate).

[0072] In certain example embodiments, the primer concentration is 100μM. In certain example embodiments, the primer concentration is between 50μM-150 μM, or between 50μM-

200μM, or between 50μM-250μM, or between 50μm-250μM or between 50μm-300μM or between 50μm-350μM or between 50μm-400μM or between 50μm-450μM or between 50μm-500μM. In certain example embodiments, the primer concentrations is between 50μm-70μM or between 70μm-90μM or between 90μm-110μM or between 110μm-130μM or between 130μm-150μM or between 150μm-170μM or between 170μm-190μM or between 190μm-210μM or between 210μm-230μM or between 230μm-250μM or between 250μm-270μM or between 270μm-290μM or between 290μm-310μM or between 310μm-330μM or between 330μm-350μM or between 350μm-370μM or between 370μm-390μM or between 390μm-410μM or between 410μm-430μM or between 430μm-450μM or between 450μm-470μM or between 470μm-490μM. In certain example embodiments, the primer concentration is between 50μm-100μM or between 100μm-150μM or between 150μm-200μM or between 200μm-250μM or between 250μm-300μM or between 300μm-350μM or between 350μm-400μM or between 400μm-450μM or between 450μm-500μM.

[0073] In certain example embodiments, a spike-in may be relatively the same length as the amplicons generated for the target organism. In one example embodiment, spike-ins are the same size and share the same priming region to ensure similar amplification performance. In certain embodiments, a spike-in for MNase-seq, ChIP-seq, and genomic DNA are around 150 nucleotides in length. In one example embodiment, a spike-in accounts for 0.1% - 3.5% reads. A spike-in to total sample ratio may be from 1,000:1 to 50:1. In one example embodiment, a spike-in includes primer binding sites on the 3' end and/or the 5' end. (Chen K., *et al.*, The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. MolCell Biol (2016) 36:662–667) The primers and primer binding sites on the SDSI may range between 15 – 40 nucleotides in length. The primer's melting temperature (T_m) may range from 40°C – 95 °C, preferably between 55-65 °C.

Sequencing

[0074] After amplification of cDNA, standard sequence library generation can be performed. In certain embodiments, sequencing comprises high-throughput (formerly "next-generation") technologies to generate sequencing reads. In DNA sequencing, a read is an inferred sequence of base pairs (or base pair probabilities) corresponding to all or part of a single DNA fragment. A typical sequencing experiment involves fragmentation of the genome into millions of molecules

or generating complementary DNA (cDNA) fragments, which are size-selected and ligated to adapters. The set of fragments is referred to as a sequencing library, which is sequenced to produce a set of reads. Methods for constructing sequencing libraries are known in the art (see, *e.g.*, Head et al., Library construction for next-generation sequencing: Overviews and challenges. Biotechniques. 2014; 56(2): 61–77). A "library" or "fragment library" may be a collection of nucleic acid molecules derived from one or more nucleic acid samples, in which fragments of nucleic acid have been modified, generally by incorporating terminal adapter sequences comprising one or more primer binding sites and identifiable sequence tags. In certain embodiments, the library members (e.g., genomic DNA, cDNA) may include sequencing adaptors that are compatible with use in, e.g., Illumina's reversible terminator method, long read nanopore sequencing, Roche's pyrosequencing method (454), Life Technologies' sequencing by ligation (the SOLiD platform) or Life Technologies' Ion Torrent platform. Examples of such methods are described in the following references: Margulies *et al* (Nature 2005 437: 376-80); Schneider and Dekker (Nat Biotechnol. 2012 Apr 10;30(4):326-8); Ronaghi *et al.* (Analytical Biochemistry 1996 242: 84-9); Shendure *et al.* (Science 2005 309: 1728-32); Imelfort *et al.* (Brief Bioinform. 2009 10:609-18); Fox *et al.* (Methods Mol. Biol. 2009; 553:79-108); Appleby *et al.* (Methods Mol. Biol. 2009; 513:19-39); and Morozova *et al.* (Genomics. 2008 92:255-64), which are incorporated by reference for the general descriptions of the methods and the particular steps of the methods, including all starting products, reagents, and final products for each of the steps.

[0075] In one example embodiment, any suitable RNA or DNA amplification technique may be used to amplify a sample and SDSI. In one example embodiment, the RNA or DNA amplification is an isothermal amplification. The isothermal amplification may be nucleic-acid sequenced-based amplification (NASBA), recombinase polymerase amplification (RPA), loop-mediated isothermal amplification (LAMP), strand displacement amplification (SDA), helicase-dependent amplification (HDA), or nicking enzyme amplification reaction (NEAR). In certain example embodiments, non-isothermal amplification methods may be used which include, but are not limited to, PCR, multiple displacement amplification (MDA), rolling circle amplification (RCA), ligase chain reaction (LCR), or ramification amplification method (RAM).

Example Applications

[0076] In one example embodiment, the present invention is used to improve any method of sequencing wherein the nucleic acids to be sequenced are amplified (i.e., amplicon-based methods). In certain example embodiments, the amplification method preferentially amplifies a contaminant nucleic acid if it is present in a sample. In preferred embodiments, samples comprising a pathogen of interest are sequenced. In more preferred embodiments, the pathogen of interest includes variants that can be clustered into families or a lineage. As used herein, the term "variant" refers to any virus having one or more mutations as compared to a known virus. A strain is a genetic variant or subtype of a virus. The terms 'strain', 'variant', and 'isolate' may be used interchangeably. In certain embodiments, a variant has developed a "specific group of mutations" that causes the variant to behave differently than that of the strain it originated from. In certain example embodiments, the families of variants are important for tracking and responding to epidemics and pandemics. For example, sequencing can be used to determine variants that are emerging as the dominant variants causing disease or are spreading more quickly. In another example, sequencing variants can be used to track community transmission and superspreading events (see e.g., Lemieux *et al.*, 2020). Variants may also include those that are resistant to a specific treatment, such as drug resistance. In certain embodiments, variants are associated with more severe disease. As used herein, the term "epidemic" refers to the rapid spread of disease to a large number of people in a given population within a short period of time or the occurrence of more cases of disease, injury, or other health condition than expected in a given area or among a specific group of persons during a particular period. For example, in meningococcal infections, an attack rate in excess of 15 cases per 100,000 people for two consecutive weeks is considered an epidemic. Epidemics of infectious disease are generally caused by several factors including a change in the ecology of the host population (*e.g.*, increased stress or increase in the density of a vector species), a genetic change in the pathogen reservoir or the introduction of an emerging pathogen to a host population (by movement of pathogen or host). Generally, an epidemic occurs when host immunity to either an established pathogen or newly emerging novel pathogen is suddenly reduced below that found in the endemic equilibrium and the transmission threshold is exceeded. An epidemic may be restricted to one location; however, if it spreads to other countries or continents and affects a substantial number of people, it may be termed a pandemic. Effective

preparations for a response to a pandemic are multi-layered. The first layer is a disease surveillance system, which includes sequencing of all variants in a population. In certain embodiments, sequencing contaminants that were amplified from a sample would provide an incorrect identification and clustering of the variants.

[0077] Any method of sequencing variants in pathogens, such as viral pathogens, is applicable to the present invention (see e.g., Lemieux *et al.*, 2020). Current sequencing methods all suffer from the risk of contamination and the user would be blind to whether the results were accurate.

[0078] In certain example embodiments, a pathogen with a DNA genome is sequenced. Sequencing may include whole genome sequencing. Whole genome sequencing (also known as WGS, full genome sequencing, complete genome sequencing, or entire genome sequencing) is the process of determining the complete DNA sequence of an organism's genome at a single time. This entails sequencing all of an organism's chromosomal DNA as well as DNA contained in the mitochondria and, for plants, in the chloroplast. "Whole genome amplification" ("WGA") refers to any amplification method that aims to produce an amplification product that is representative of the genome from which it was amplified. In certain embodiments, the SDSIs of the present invention are added at the amplification step. Non-limiting WGA methods include Primer extension PCR (PEP) and improved PEP (I-PEP), Degenerated oligonucleotide primed PCR (DOP-PCR), Ligation-mediated PCR (LMP), T7-based linear amplification of DNA (TLAD), and Multiple displacement amplification (MDA).

[0079] In certain example embodiments, the present invention includes whole exome sequencing. Exome sequencing, also known as whole exome sequencing (WES), is a genomic technique for sequencing all of the protein-coding genes in a genome (known as the exome) (see, e.g., Ng et al., 2009, Nature volume 461, pages 272–276). It consists of two steps: the first step is to select only the subset of DNA that encodes proteins. These regions are known as exons – humans have about 180,000 exons, constituting about 1% of the human genome, or approximately 30 million base pairs. The second step is to sequence the exonic DNA using any high-throughput DNA sequencing technology. In certain embodiments, whole exome sequencing is used to determine germline mutations in genes associated with disease.

[0080] In certain example embodiments, targeted sequencing is used in the present invention (see, e.g., Mantere *et al.*, PLoS Genet 12 e1005816 2016; and Carneiro *et al.* BMC Genomics,

2012 13:375). Targeted gene sequencing panels are useful tools for analyzing specific mutations in a given sample. Focused panels contain a select set of genes or gene regions that have known or suspected associations with the disease or phenotype under study. In certain embodiments, targeted sequencing is used to detect mutations associated with a disease in a subject in need thereof. Targeted sequencing can increase the cost-effectiveness of variant discovery and detection. In certain embodiments, targeted sequencing includes amplification and the SDSIs of the present invention are added at the amplification step.

[0081] In one example embodiment, the mitochondrial genome from more than one sample is sequenced. In certain embodiments, mitochondrial genome sequencing includes amplification and the SDSIs of the present invention are added at or before the amplification step. An exemplary method includes MitoRCA-seq (see e.g., Ni *et al.*, MitoRCA-seq reveals unbalanced cytosin to thymine transition in Polg mutant mice. Sci Rep. 2015 Jul 27;5:12049. doi: 10.1038/srep12049). The method employs rolling circle amplification, which enriches the full-length circular mtDNA by either custom mtDNA-specific primers or a commercial kit and minimizes the contamination of nuclear encoded mitochondrial DNA (Numts). In certain embodiments, RCA-seq is used to detect low-frequency mtDNA point mutations starting with as little as 1 ng of total DNA.

[0082] In another example embodiment, multiple displacement amplification (MDA) is used to generate a sequencing library. Multiple displacement amplification (MDA, is a non-PCR-based isothermal method based on the annealing of random hexamers to denatured DNA, followed by strand-displacement synthesis at constant temperature (Blanco et al. J. Biol. Chem. 1989, 264, 8935-8940). It has been applied to samples with small quantities of genomic DNA, leading to the synthesis of high molecular weight DNA with limited sequence representation bias (Lizardi et al. Nature Genetics 1998, 19, 225-232; Dean et al., Proc. Natl. Acad. Sci. U. S. A. 2002, 99, 5261-5266). As DNA is synthesized by strand displacement, a gradually increasing number of priming events occur, forming a network of hyper- branched DNA structures. The reaction can be catalyzed by enzymes such as the Phi29 DNA polymerase or the large fragment of the Bst DNA polymerase. The Phi29 DNA polymerase possesses a proofreading activity resulting in error rates 100 times lower than Taq polymerase (Lasken et al. Trends Biotech. 2003, 21, 531-535). In certain embodiments, the SDSIs of the present invention are added to samples and amplified during MDA or in a subsequent amplification step.

[0083] In one example embodiment, is sequencing comprises sequencing of SARS-CoV-2 variants. The scale of the SARS-CoV-2 pandemic has led to a particular focus on reducing the cost and time of amplicon-based methods, often at the cost of slightly reduced sensitivity. However, viral loads of SARS-CoV-2 can vary widely between individuals, in particular when samples are caught early in infection or follow-up sampling is needed. An open-access tiled primer set developed by the ARTIC network is the most widely used method for SARS-CoV-2 specific genome amplification followed by sequencing on either Illumina or nanopore instruments (Quick *et al.*, 2017; Tyson e*t al.*, 2020). A wide array of protocols and publications are now available that integrate these ARTIC primers with different amplification and library construction indexing strategies (Baker *et al.*, 2020; Gohl *et al.*, 2020). Approaches such as batching samples by viral load to increase sensitivity are impractical to scale to current needs, resulting in incomplete recovery of viral genomes, especially from low titer samples.

[0084] In certain embodiments, the methods described herein can be used to sequence viral samples with low viral loads. A viral load may also be interchangeably referred to as viral burden or viral titer. A viral load may be expressed in viral particles per mL, infectious particles per mL, copies per mL, or virus per mL. A low viral load may be a cycle threshold (CT) > 30 or copies per mL < 10^4. A high viral load may be a CT < 30 or par or copies per mL > 10^5. For example, viral loads lower than 10,000, 1,000, 500, 400, 300, 200, 100, 50, 40, 30, 20, 10 viral particles. In certain embodiments, a single viral particle is sequenced.

[0085] In certain embodiments, the SDSI is used to detect and prevent contamination in genomic analysis samples of pathogens. A pathogen may include viruses, bacteria, fungi, and protozoa. In certain embodiments, a virus may belong to any morphological category including helical, envelope, or icosahedral. In certain embodiments, a virus me comprise of DNA or RNA, may be single stranded or double stranded, and may be linear or circular. In certain embodiments, the genome of the virus may be one nucleic acid molecule or several nucleic acid segments. In certain embodiments a virus may belong to the family: Adenoviridae, Papovaviridae, Parvoviridae, Herpesviridae, Poxviridae, Anelloviridae, Pleolipoviridae, Reoviridae, Picornaviridae, Caliciviridae, Togaviridae, Arenaviridae, Flaviviridae, Orthomyxoviridae, Paramyxoviridae, Bunyaviridae, Rhabdoviridae, Filoviridae, Astroviridae, Bornaviridae, Arteriviridae, Hepeviridae, Retroviridae, Caulimoviridae, Hepadnaviridae, Coronaviridae. In certain embodiment, the virus is

SARS-CoV-2. (Gelderblom HR. Structure and Classification of Viruses. In: Baron S, editor. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston; 1996. Chapter 41).

[0086] In an exemplary embodiment, the pathogen sequenced is a coronavirus. As used herein, "coronavirus" refers to enveloped viruses with a positive-sense single-stranded RNA genome and a nucleocapsid of helical symmetry that constitute the subfamily *Orthocoronavirinae*, in the family *Coronaviridae* (see, e.g., Woo PC, Huang Y, Lau SK, Yuen KY. Coronavirus genomics and bioinformatics analysis. Viruses. 2010;2(8):1804-1820). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the virus causing the ongoing Coronavirus Disease 19 (COVID19) pandemic (see, e.g., Zhou, et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273). In preferred embodiments, the virus is SARS-CoV-2 or variants thereof. In preferred embodiments, the disease treated is COVID-19. SARS-CoV-2 is the third zoonotic betacoronavirus to cause a human outbreak after SARS-CoV in 2002 and Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012 (de Wit et al., 2016, SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol *14*, 523-534). While there are many thousands of variants of SARS-CoV-2, (Koyama, Takahiko Koyama; Platt, Daniela; Parida, Laxmi (June 2020). "Variant analysis of SARS-CoV-2 genomes". Bulletin of the World Health Organization. 98: 495–504) there are also much larger groupings called clades. Several different clade nomenclatures for SARS-CoV-2 have been proposed. As of December 2020, GISAID, referring to SARS-CoV-2 as hCoV-19 identified seven clades (O, S, L, V, G, GH, and GR) (Alm E, Broberg EK, Connor T, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020 [published correction appears in Euro Surveill. 2020 Aug;25(33):]. Euro Surveill. 2020;25(32):2001410). Also as of December 2020, Nextstrain identified five (19A, 19B, 20A, 20B, and 20C) (Cited in Alm et al. 2020). Guan et al. identified five global clades (G614, S84, V251, I378 and D392) (Guan Q, Sadykov M, Mfarrej S, et al. A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic. Int J Infect Dis. 2020;100:216-223). Rambaut et al. proposed the term "lineage" in a 2020 article in Nature Microbiology; as of December 2020, there have been five major lineages (A, B, B.1, B.1.1, and B.1.777) identified (Rambaut, A.; Holmes,

E.C.; O'Toole, Á.; et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology". 5: 1403–1407).

[0087] Exemplary, non-limiting variants applicable to the present invention are described below. Genetic variants of SARS-CoV-2 have been emerging and circulating around the world throughout the COVID-19 pandemic (see, e.g., The US Centers for Disease Control and Prevention; www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html). Exemplary, non-limiting variants applicable to the present disclosure include variants of SARS-CoV-2, particularly those having substitutions of therapeutic concern. **Table A** shows exemplary, non-limiting genetic substitutions in SARS-CoV-2 variants.

Table A.	
Spike Protein Substitution	**Common Pango Lineages with Spike Protein Substitutions**
L452R	A.2.5, B.1, B.1.429, B.1.427, B.1.617.1, B.1.526.1, B.1.617.2, C.36.3
E484K	B.1.1.318, B.1.1.7, B.1.351, B.1.525, B.1.526, B.1.621, B.1.623, P.1, P.1.1, P.1.2, R.1
K417N, E484K, N501Y	B.1.351, B.1.351.3
K417T, E484K, N501Y	P.1, P.1.1, P.1.2
A67V, del69-70, T95I, del142-144, Y145D, del211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F	B.1.1.529 and BA lineages

Phylogenetic Assignment of Named Global Outbreak (PANGO) Lineages is software tool developed by members of the Rambaut Lab. The associated web application was developed by the Centre for Genomic Pathogen Surveillance in South Cambridgeshire and is intended to implement the dynamic nomenclature of SARS-CoV-2 lineages, known as the PANGO nomenclature. It is available at cov-lineages.org.

[0088] In some embodiments, the SARS-CoV-2 variant is and/or includes: B.1.1.7, also known as Alpha (WHO) or UK variant, having the following spike protein substitutions: 69del, 70del, 144del, (E484K*), (S494P*), N501Y, A570D, D614G, P681H, T716I, S982A, and

D1118H (K1191N*); B.1.351, also known as Beta (WHO) or South Africa variant, having the following spike protein substitutions: D80A, D215G, 241del, 242del, 243del, K417N, E484K, N501Y, D614G, and A701V; B.1.427, also known as Epsilon (WHO) or US California variant, having the following spike protein substitutions: L452R, and D614G; B.1.429, also known as Epsilon (WHO) or US California variant, having the following spike protein substitutions: S13I, W152C, L452R, and D614G; B.1.617.2, also known as Delta (WHO) or India variant, having the following spike protein substitutions: T19R, (G142D), 156del, 157del, R158G, L452R, T478K, D614G, P681R, and D950N; P.1, also known as Gamma (WHO) or Japan/Brazil variant, having the following spike protein substitutions: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, and T1027I; and B.1.1.529 also known as Omicron (WHO), having the following spike protein substitutions: A67V, del69-70, T95I, del142-144, Y145D, del211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F, or any combination thereof.

[0089] In some embodiments, the SARS-CoV-2 variant is classified and/or otherwise identified as a Variant of Concern (VOC) by the World Health Organization and/or the U.S. Centers for Disease Control. A VOC is a variant for which there is evidence of an increase in transmissibility, more severe disease (e.g., increased hospitalizations or deaths), significant reduction in neutralization by antibodies generated during previous infection or vaccination, reduced effectiveness of treatments or vaccines, or diagnostic detection failures.

[0090] In some embodiments, the SARS-Cov-2 variant is classified and/or otherwise identified as a Variant of High Consequence (VHC) by the World Health Organization and/or the U.S. Centers for Disease Control. A variant of high consequence has clear evidence that prevention measures or medical countermeasures (MCMs) have significantly reduced effectiveness relative to previously circulating variants.

[0091] In some embodiments, the SARS-Cov-2 variant is classified and/or otherwise identified as a Variant of Interest (VOI) by the World Health Organization and/or the U.S. Centers for Disease Control. A VOI is a variant with specific genetic markers that have been associated with changes to receptor binding, reduced neutralization by antibodies generated against previous

infection or vaccination, reduced efficacy of treatments, potential diagnostic impact, or predicted increase in transmissibility or disease severity.

[0092] In some embodiments, the SARS-Cov-2 variant is classified and/or is otherwise identified as a Variant of Note (VON). As used herein, VON refers to both "variants of concern" and "variants of note" as the two phrases are used and defined by Pangolin (cov-lineages.org) and provided in their available "VOC reports" available at cov-lineages.org.

[0093] In some embodiments the SARS-Cov-2 variant is a VOC. In some embodiments, the SARS-CoV-2 variant is or includes an Alpha variant (e.g., Pango lineage B.1.1.7), a Beta variant (e.g., Pango lineage B.1.351, B.1.351.1, B.1.351.2, and/or B.1.351.3), a Delta variant (e.g., Pango lineage B.1.617.2, AY.1, AY.2, AY.3 and/or AY.3.1); a Gamma variant (e.g., Pango lineage P.1, P.1.1, P.1.2, P.1.4, P.1.6, and/or P.1.7), a Omicon variant (B.1.1.529) or any combination thereof.

[0094] In some embodiments the SARS-Cov-2 variant is a VOI. In some embodiments, the SARS-CoV-2 variant is or includes an Eta variant (e.g., Pango lineage B.1.525 (Spike protein substitutions A67V, 69del, 70del, 144del, E484K, D614G, Q677H, F888L)); an Iota variant (e.g., Pango lineage B.1.526 (Spike protein substitutions L5F, (D80G*), T95I, (Y144-*), (F157S*), D253G, (L452R*), (S477N*), E484K, D614G, A701V, (T859N*), (D950H*), (Q957R*))); a Kappa variant (e.g., Pango lineage B.1.617.1 (Spike protein substitutions (T95I), G142D, E154K, L452R, E484Q, D614G, P681R, Q1071H)); Pango lineage variant B.1.617.2 (Spike protein substitutions T19R, G142D, L452R, E484Q, D614G, P681R, D950N)), Lambda (e.g., Pango lineage C.37); or any combination thereof.

[0095] In some embodiments SARS-Cov-2 variant is a VON. In some embodiments, the SARS-Cov-2 variant is or includes Pango lineage variant P.1 (alias, B.1.1.28.1.) as described in Rambaut et al. 2020. Nat. Microbiol. 5:1403-1407) (spike protein substitutions: T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, TI027I)); an Alpha variant (e.g., Pango lineage B.1.1.7); a Beta variant (e.g., Pango lineage B.1.351, B.1.351.1, B.1.351.2, and/or B.1.351.3); Pango lineage variant B.1.617.2 (Spike protein substitutions T19R, G142D, L452R, E484Q, D614G, P681R, D950N)); an Eta variant (e.g., Pango lineage B.1.525); Pango lineage variant A.23.1 (as described in Bugembe et al. medRxiv. 2021. doi: doi.org/10.1101/2021.02.08.21251393) (spike protein substitutions: F157L, V367F, Q613H, P681R); or any combination thereof.

[0096] In certain embodiments, the pathogen sequenced is a pathogenic bacteria and may include: spirochetes; spirilla; vibrios; gram-negative aerobic rods and cocci; enterics; pyogenic cocci; and endospore-forming bacteria; actinomycetes and related bacteria; rickettsias and chlamydiae; mycoplasmas, which are groups defined by some bacteriological criteria. A pathogenic bacteria may include: *Escherichia coli, Salmonella enterica, Salmonella typhi, Shigella dysenteriae, Yersinia pestis, Pseudomonas aeruginosa, Vibrio cholerae, Bordetella pertussis, Haemophilus influenza, Helicobacter pylori, Campylobacter jejuni, Neisseria gonorrhoeae, Neisseria meningitidis, Brucella abortus, Bacteroides fragilis, Staphylococcus aureus, Streptococcus pyogenes, Streptococcus pneumoniae, Bacillus anthracis, Bacillus cereus, Clostridium tetani, Clostridium perfringens, Clostridium botulinum, Clostridium difficile, Corynebacterium diphtheriae, Listeria monocytogenes, Mycobacterium tuberculosis, Mycobacterium leprae, Chlamydia trachomatis, Chlamydia pneumoniae, Mycoplasma pneumoniae, Rickettsias, Treponema pallidum, Borrelia burgdorferi*, or a variant thereof. (Todar, K. Textbook of Bacteriology (2020) Online).

[0097] In an exemplary embodiment, the pathogen sequenced is a pathogenic fungi and may include: Aspergillus; Blastomyces; Candida; Coccidioides; Cryptococcus; Fusarium; Microsporum; Epidermophyton; Trichophyton; Histoplasma; Rhizopus; Mucor; Rhizomucor; Syncephalastrum; Cunninghamella; Apophysomyces; Lichtheimia (formerly Absidia); Eumycetoma; Pneumocystis; Trichophyton; Microsporum; Epidermophyton; Sporothrix; Paracoccidioides; Talaromyces or a variant or species thereof. (CDC)

[0098] In an exemplary embodiment, the pathogen sequenced is a pathogenic protozoa belonging to the group: Sarcodina; Mastigophora; Ciliophora; or Sporozoa defined by their mode of movement. (CDC) In certain embodiments, the pathogenic protozoa may include: Entamoeba; Trichomonas; Leishmania; Chilomonas; Giardia; Isopora; Sarcocystis; Nosema; Balantidium; Eimeria; Histomonas; Trypanosoma; Plasmodium; Babesia; or Haemoproteus or a variant or species thereof.

[0099] Further embodiments are illustrated in the following Examples which are given for illustrative purposes only and are not intended to limit the scope of the invention.

EXAMPLES

[0100] Here Applicants designed, optimized, and implemented a novel sample identification method using synthetic DNA spike-ins (SDSIs) that is broadly compatible with SARS-CoV-2 sequencing approaches and settings. Applicants implemented these SDSIs for Illumina sequencing with SARS-CoV-2 specific amplification using the ARTIC consortium's amplicon designs. To maximize epidemiological utility by increasing the number of genomes recovered from samples with low viral loads, Applicants benchmarked key amplification and library construction steps. Applicants propose a modified protocol, hereafter termed SDSI+ARTIC, that provides increased confidence in the veracity of genomes with minimal extra cost and time that can be applied to investigations of SARS-CoV-2 epidemiology and emerging viral variants (**FIG. 1**).

Example 1 – Design and in silico validation of novel amplicon spike-ins

[0101] Applicants sought to design a robust system for contamination tracing and sample tracking applicable to a wide-variety of viral sequencing strategies via known synthetic DNA sequences. Applicants envisioned that these novel synthetic DNA spike-ins (SDSIs) would consist of a uniquely identifiable sequence such that each sample in a sequencing batch could be paired with a different SDSI, enabling in-sample labeling. SDSIs should be sufficiently distinct from one another as well as common laboratory or human pathogens to ensure reliable identification. Each unique sequence is then flanked by constant priming regions so that a single additional primer set can be integrated into a multiplexed PCR to co-amplify the SDSI with the sample **(FIG 2A)**. In labeling all amplified viral genomic material in a laboratory setting, Applicants could track sample swaps and viral contamination with exquisite resolution and accuracy.

[0102] Excerpting DNA sequences from diverse, exotic archaea genomes to serve as the unique portion of the SDSI precludes false detection and cross-identification. To balance common sequencing library construction constraints, DNA synthesis costs, and providing enough sequence to be uniquely identifiable, Applicants generated SDSIs with a 140 bp stretch of variable sequence. Applicants confirmed that the various SDSIs were significantly different from each other to mitigate cross-identification; among all SDSIs, the minimum pairwise Hamming distances of the 140 bp stretch of unique sequence was 84 (mean=105; max=121). Since false detection of SDSI would occur if its sequence shared significant homology with other genetic material in a sample, Applicants based these sequences on archaea, which are divergent from organisms found in typical

laboratory or clinical settings **(Table 2)**. A permissive search performed against the entire NCBI database confirmed that 44/48 SDSI sequences had significant homology (>75% sequence identity over >75% query cover) exclusively within the domain archaea; the remaining SDSIs had homology to a handful of bacterial genuses unlikely to be found in laboratories (**Table 2**). In considering the application of these SDSIs to ARTIC SARS-CoV-2 amplicon sequencing, Applicants also specifically verified that each unique SDSI sequences were unlikely to be confused with expected COVID-19 clinical sample content, confirming that each sequence had very limited homology (nothing >50% sequence identity over >50% query cover) to both *Homo sapiens* and SARS-CoV-2. In designing these amplicon sequences Applicants also avoided extremes of GC content (range: 35-65%) in order to promote similar amplification rates across different SDSIs, as well as other potential targets of the multiplexed reaction, such as viral amplicons. Applicants specifically ensured that the SDSIs had similar GC content to ARTIC SARS-CoV-2 amplicons (**FIG. 6**).

[0103] Similarly, the design of common primers for SDSI amplicons enabled compatibility with a broad spectrum of amplicon-based sequencing reactions, including in clinical settings. To preclude off-target priming in the PCR reaction that could outcompete amplification of a primary target, Applicants limited SDSI primer homology to common organisms, particularly on the 3' end of the primer. Applicants specifically confirmed that primers were unlikely to amplify human or SARS-CoV-2 to promote SDSI primer integration into the ARTIC SARS-CoV-2 amplicon sequencing PCR reaction. Primers were compatible with ARTIC v3 primer sets, with a similar length (24 bps each) and GC content (45.8% each) (**FIG. 6**).

Example 2 – Application of spike-ins to ARTIC SARS-CoV-2 sequencing

[0104] Applicants demonstrated that the addition of SDSIs into the ARTIC multiplexed PCR provided a sample-specific internal control and did not detrimentally affect the amplification of SARS-CoV-2 RNA. SDSI primers did not produce any nonspecific amplification, including in the presence of NP swab RNA, supporting the expectation that primers shared limited homology with genomic material from clinical samples **(FIG. 6)**. All SDSIs amplified in an ARTIC SARS-CoV-2 PCR reaction with SDSI primers included, in each case yielding a single clean product of the expected size **(FIG. 6)**. Applicants next sought to ensure that inclusion of the SDSI oligo and SDSI primers did not limit amplification of SARS-CoV-2 RNA. To prevent SDSIs overtaking the

amplification and sequencing of SARS-CoV-2 amplicons, Applicants optimized the amount of SDSI added to each reaction through limited titration (**FIG. 7**). Applicants found that 1µl of a 1fM SDSI resulted in the reliable detection of the SDSI across a range of CT values (CT 20, 25, 30, 35) while the majority of reads (>96%) still mapped to SARS-CoV-2 **(Table 3; FIG. 2B)**.

[0105] Applicants performed SDSI+ARTIC sequencing on a batch of 48 SARS-CoV-2+ clinical samples to demonstrate its feasibility and utility in tracking samples and identifying contamination. After adding a different SDSI to each sample, Applicants found that 47/48 SDSIs were identified exclusively in the anticipated sample, validating the use of SDSIs as an internal control for sample tracking. One SDSI (SDSI_48) was detected in the sample that it was added to as well as a neighboring sample in the batch **(FIG. 2C)**. Applicants suspect that this represents unintentional within-batch contamination that was likely a consequence of spillover between neighboring wells. This case reveals the insidious nature of commonplace contamination and underscores the importance of this novel method for identifying it.

[0106] As shorter amplicons have been purported to yield superior recovery for low viral load samples (Antonov et al., 2005; No et al., 2019)), Applicants explored extending SDSIs to the Paragon Genomics' CleanPlex SARS-CoV-2 panel, but identified fatal shortcomings. Paragon amplicons are on average half the size of ARTIC (149bp vs 343bp), and compatible with the SDSI length 140bp. (Antonov et al., 2005; No et al., 2019) (*SARS-CoV-2 COVID-19 Coronavirus Research and Surveillance*, n.d.)(Antonov et al., 2005; No et al., 2019). However, the Paragon panel had dropout regions even in low CT samples which resulted in missed SNP calls compared to ARTIC across 5 samples (CTs = 20-37), consistent with other reports (**FIG 8A, 8B**) (Klempt et al., 2020). Although this panel did recover more of the genome in very high CT samples (>35), Applicants did not proceed with SDSI integration as the uneven and unreliable genome coverage across most clinical CTs limited Paragon's epidemiological utility (**FIG. 8C**).

Example 3 – Improving genome recovery and coverage for Illumina-based ARTIC SARS-CoV-2 sequencing

[0107] Applicants benchmarked various alterations to Illumina-based SDSI+ARTIC sequencing in order to maximize the number of complete, high-quality genomes recovered from clinically diverse samples. Higher CT samples prove especially challenging to sequence but their recovery is still of critical importance to epidemiological and clinical applications of viral

genomics. Applicants found that substituting a more processive reverse transcriptase provided the single biggest benefit. Comparing cDNA produced with Superscripts III, IV, or IV-VILO across a range of clinical CTs (low CT: <20, mid-low CT: 20-25, mid-high CT: 25-30, and high CT: >30), SSIV-VILO and SSIV produced the highest number of amplicons with at least 10X coverage across 13 samples (SSIII: 72.64%, SSIV: 81.93%, SSIV-VILO: 86.97%) **(FIG. 3A)**. These processive reverse transcriptases also displayed lower variability as measured by the percent of amplicons with <20% mean coverage (SSIII: 36.89% SSIV: 31.24% SSIV-VILO: 22.45%) **(FIG. 9A)**. Applicants also tested five DNA polymerase and conditions in the SDSI+ARTIC PCR reaction **(Methods)** and found that Q5 Hot Start High-Fidelity 2X Master Mix and KAPA reactions yielded the highest amplification (average 85.3nM and 56nM respectively) **(FIG. 9B)**.

[0108] Applicants also attempted protocol modifications to increase sequence depth uniformity in SDSI+ARTIC, which is crucial for recovering complete genomes in the fewest number of reads. When Applicants increased (2x) primer concentrations (20.8nM final) for low efficiency amplicons, Applicants observed increased coverage in these amplicons that enabled whole genome recovery for multiple samples, especially those with higher CTs **(FIG. 3B; FIG. 10; Table 4).** Other groups have also noted that alternative primers or changes in annealing temperature can reduce the formation of certain primer interactions, and Applicants suspect exploration of these avenues would further optimize SDSI+ARTIC (Itokawa et al., 2020). Applicants also attempted to recover high CT samples by increasing the number of PCR cycles and observed greater coverage uniformity with increasing cycles (**FIG. 3C**). However, at 45 cycles Applicants observed 3 SNPs that were not present in lower-amplified samples. To avoid erroneous SNP calls, Applicants decided to implement and optimize the SDSI's for a 40 cycle PCR. Additional modifications such as DNA-rehybridization steps (Mathieu-Daudé et al., 1996) or slower temperature ramp speeds had no significant effects **(FIG. 9C, 9D)**.

[0109] Applicants reduced the potential for highly amplified library contamination within the laboratory or clinical setting by scaling down (.5X) the Illumina DNA Flex library construction kit, which also reduced per sample cost without impacting performance (**Table 5; Table 6**). In benchmarking library construction methods, Applicants confirmed Nextera DNA Flex generated greater coverage depth than DNA XT (**FIG. 10**). In combination, the final suggested modifications to Illumina ARTIC sequencing include using more processive reverse transcriptases, 40 cycles of

PCR, and 2X primer concentration to recover higher CT samples, as well as a .5X scale down of Illumina DNA Flex to produce less concentrated, and thus less likely to contaminate libraries at a halved cost. Integrating these modifications into the SDSI approach may enable greater genomic surveillance in a limited number of samples.

Example 4 – SDSI-ARTIC sequencing benchmarks well against metagenomic sequencing.

[0110] Highlighting the reliability and robustness of this approach, Applicants observed high sequence correlation and superior genome recovery with SDSI+ARTIC compared to an unbiased metagenomics approach, the gold standard in generating error-free viral genomes. Applicants sequenced a small batch of six samples (CTs = 16-31) using ARTIC without SDSIs, and generated full length genomes with 100% concordance to those generated with metagenomic sequencing, indicating the accuracy of ARTIC-based sequencing methods (Lemieux et al., 2021). Applicants then resequenced 89 unique patient samples with SDSI+ARTIC that were previously sequenced using the same standard metagenomics approach (Lemieux et al., 2021) to serve as a direct comparison. The 89 samples in the validation batch consisted of diverse viral lineages and a broad range of CTs (range = 11.9-37.4; mean = 27.4) (**FIG. 12A**). SDSI+ARTIC outperformed metagenomic sequencing in terms of genome recovery, with increased median assembly lengths (29,577 bp and 4,389 bp respectively) (**FIG. 4A**), and a higher number of complete (>98%) genomes assembled (50 and 31 respectively). Applicants recovered even more partial (>80%) genomes with SDSI+ARTIC when compared to metagenomic sequencing (75 vs 36 respectively). Notably, 5 complete genomes recovered for SDSI+ARTIC had a CT above 30 (**FIG. 4B; FIG. 12B).** Applicants also assessed coverage uniformity in both methods, as increasing uniformity reduces the sequencing depth required to generate reliable genomes, thus improving throughput and efficiency. (So et al., 2018). As measured by a gini coefficient for each sample that generated an assembly, uniformity decreased in both methods above a CT of 25 but was markedly worse for metagenomics (**FIG. 12C**).

[0111] SDSI+ARTIC displayed high concordance in sequence variant identification to metagenomics, producing only two divergent SNP calls out of 331 total SNPs across 38 genomes (**FIG. 4C**). Notably, this discordance was present with both relaxed (n=3) and conservative (n=20) minimum coverage thresholds. The discordant SNPs, observed in two samples, were present in different regions. However, both were located in ARTIC primer regions and matched the primer

sequence even though primer trimming was performed and confirmed by manual inspection. Additionally, the coverage depth in the regions of the SNPs was greater than 1000X for both platforms in both samples. Applicants believe these errors likely arose during the ARTIC PCR, suggesting a discordance rate of 0.6% between amplicon-based and metagenomic sequencing. Notably these few mismatches did not result in lineage misassignment for either sample. To evaluate concordance, Applicants compared consensus sequences without down sampling using only samples that produced a full genome to make the most equivalent comparison (Methods).

Example 5 – Rapid deployment of SDSI-ARTIC sequencing confirms a suspected nosocomial transmission cluster

[0112] SDSI+ARTIC is a powerful method for public health interventions, especially as superspreading events - and clusters of cases linked to close contact settings more broadly - have become a defining feature of the SARS-CoV-2 pandemic ((Adam et al., 2020; Dearlove et al., 2020; Lemieux et al., 2021; Wong & Collins, 2020)). Viral genomes can reveal whether these clusters are linked through transmission, based on shared viral sequences, providing useful information for public health interventions. Such outbreak investigations of single cases leading to many are distinguishable due to low viral sequence variation but requires higher levels of confidence to ensure such a pattern has not occurred due to laboratory contamination. To demonstrate the utility of the novel SDSIs and modified protocol, Applicants applied the method to investigate a putative cluster of 14 SARS-CoV-2 cases from Massachusetts General Hospital (MGH), for which the infection control unit had suspicion of a nosocomial outbreak. Applicants sequenced 24 samples; 14 samples believed to be part of the cluster based on traditional contact-tracing, 8 unlinked samples and 2 negative controls.

[0113] The SDSI+ARTIC method enabled fast and confident identification of a nosocomial cluster, with samples processed within 24 hours and final genomes assembled within 52 hours of bio-sample receipt. Applicants assembled 14 complete genomes (>98% complete) of which 9 were from cluster-associated samples. Those samples that did not yield a full genome were those with lower viral loads (CT > 30). Phylogenetic analysis showed that samples from the cluster were genetically highly similar and clustered together (**FIG. 5A, 5B**) to the exclusion of other samples from Boston around the same time, strongly suggesting that this cluster did reflect transmission within the hospital. One sample, MA-MGH-02834, differed from other cluster-associated samples

by 18-19 consensus-level variants suggesting that this infection was likely acquired separately and not as part of the same nosocomial transmission. Analysis of the SDSIs confirmed that genome sequence similarity was not the result of cross-contamination from highly amplified final libraries (**FIG. 5C**).

Example 6 – Discussion on novel amplicon spike-ins

[0114] As the SARS-CoV-2 pandemic intensifies and new genomic variants continue to emerge, it is imperative to build robust experimental confidence into genomic surveillance data interpretation. Here, Applicants report a novel design and implementation of Synthetic DNA Spike-ins (SDSI) as an essential component for tracking and tracing contamination, a potential confounder in amplicon-based sequencing methods of SARS-CoV-2. The in-silico design generated robust synthetic targets at low costs while mitigating inter-spike-in sequence homology as well as homology with human, SARS-CoV-2, and common laboratory reagents. While broadly applicable to most amplicon-based approaches, as a proof-of-principle Applicants coupled the SDSIs to an improved ARTIC amplicon sequencing protocol yielding faster throughput with an overall reduced cost compared to existing Illumina DNA Flex-based protocols.

[0115] SDSIs can readily be adopted by laboratories and platforms of all sizes with only minor changes to existing methodologies, little additional cost per sample ($0.006), and no interruption to standard workflow methodologies. Additional synthetic targets could be designed using the same principles to expand into 384 well formats and beyond. Primer sites could also be modulated for integration with new advancements in amplicon sequencing, like tailed primer approaches (Gohl et al., 2020). More broadly, standardizing controls across the viral surveillance community will increase accuracy and integrity of SARS-CoV-2 genomic data worldwide. These SDSIs not only enable profiling of in-batch contamination, but also laboratory-wide detection as their presence in other data (amplicon, metagenomic, qPCR, or otherwise) would indicate a tagged amplification and thus contamination. Moreover the approach is applicable to both Illumina and Nanopore sequencing platforms as well as any other existing or future tiled amplicon panel, such as those previously used for Zika, Ebola, and other recent outbreaks (Quick et al., 2016) (Metsky et al., 2017). SDSIs could serve as a broad tool for tracing potential contamination across a plethora of fields that employ amplicon based genomic sequencing, such as food safety, species identification or environmental sampling.

[0116] In optimizing the SDSI+ARTIC protocol Applicants tested and incorporated a number of cost and time saving adjustments. Modifications that can be used include implementing liquid handlers in high volume settings such as public health laboratories. Additional methodological improvements could allow for direct PCR amplification of SARS-CoV-2 using primers with indexing adapter compatible ends (Baker et al., 2020; Gohl et al., 2020) or the inclusion of unique molecular identifiers to understand intra-host variation. The SDSIs were designed to be compatible with such potential future approaches. Applicants note that there is still considerable non-uniformity in per-amplicon coverage for samples with low viral loads highlighting the need for methods that can confidently capture this information. A recent update to the ARTIC protocol for nanopore suggests that a change in the annealing temperature from 65°C to 63°C can reduce dropout of amplicon 64 (Tyson et al., 2020), a particularly poorly performing amplicon. The results show that 2x primer concentration for a subset of underperforming amplicons improved performance, and matching primer concentrations with amplicon efficiency would likely yield more uniform coverage (**Table 4**). Alternative approaches for the recovery of genomes from samples with low viral load include the use of targeted enrichment approaches (Houldcroft et al., 2017; Metsky et al., 2019) are more costly and time-consuming.

[0117] Amplicon based sequencing methods fill a critical need for rapid turn around and full genome recovery for epidemiological surveillance where SNP identification is crucial. While benchmarking the modified protocol against the gold standard approach of metagenomics Applicants observed discordant SNPs were rare (2/331). This emphasizes the need for caution and replication of libraries for highly important samples. Other commercial amplicon-based designs such as those by Paragon Genomics are significantly faster workflows and use smaller size amplicons, but the ARTIC primer set results in better overall coverage for the majority of samples (up to CT=35) and genome accuracy. Applicants believe subsequent generations of amplicon-based sequencing will address this pressing need pushing cost down while increasing genomic surveillance accuracy, which is sorely needed in the public health setting. The rapid deployment of SDSI+ARTIC confirming a nosocomial infection cluster further emphasizes the utility of the SDSIs to confidently identify samples of high genetic similarity.

The potential emergence of SARS-CoV-2 immune and vaccine escape variants underscores the ongoing necessity of accurate, reliable, and accessible genome sequencing. The modifications and

suggestions build upon a remarkable global genomic surveillance response that has developed new tools for the rapid sequencing of viral genomes at an unprecedented rate. In light of the latest surges in SARS-CoV-2 cases globally and the emergence of more transmissible lineages and variants of concern that are rising in frequency in multiple continents, continual innovation in these protocols to improve their efficiency, cost-effectiveness and reliability are essential to meet the growing need for genomic surveillance of SARS-CoV-2. Moreover, stringent sample tracking and contamination detection strategies must become a standard practice, maximizing the utility of genomic data and its increasing importance for shaping public health interventions.

Example 7– Design and characterization of synthetic DNA spike-ins for AmpSeq

[0118] Applicants designed a simple and flexible system for sample tracking and contamination tracing using a core uniquely identifiable DNA sequence flanked by constant priming regions that satisfy several design requirements. This design allows in-sample tracking through the addition of a different SDSI to each sample during sample processing. Following sequencing, the data can be analyzed for both the presence of the expected SDSI and any other SDSI, illuminating both sample misassignment and contamination with high resolution and accuracy (**FIG 13**). Applicants focused the initial design on highly stable DNA oligos that would be added to sample cDNA and could capture contamination at or after the critical viral amplification step, including contamination generated during amplification and in handling amplified material. By using a longer unique core sequence, as compared to a short barcode system, these SDSIs are compatible with both tagmentation- and ligation-based sequencing approaches. The constant priming regions mean that only a single primer pair needs to be added into the existing multiplexed PCR step to co-amplify all SDSIs with the primary reaction target(s) **(FIG 14A)**. In particular, Applicants sought to design a system that could be integrated into diverse amplicon-based viral sequencing approaches. 96 distinct DNA sequences from the genomes of diverse, uncommon archaea serve as the core portion of each SDSI, precluding false detection and cross-identification **(Table 1, Methods)**. By using extremophilic archaea, the designs maximized evolutionary distance from common human pathogens. To avoid false positive results the core SDSI sequences should be sufficiently distinct from one another, as well as sequences commonly found in laboratories and clinical samples. A permissive BLASTn search performed against the entire NCBI database confirmed that the unique SDSI core sequences had limited homology

outside the domain archaea, specifically to genera unlikely to be found in laboratories (**FIG 18A**). While this limited homology outside of the domain archaea maximized the potential for broad applications, Applicants also confirmed that none of the core sequences shared significant homology with *Homo sapiens* or known viral genomes (**Methods**). Applicants considered significant homology as >90% sequence identity over 50 bps, as library construction can result in the generation of small fragments. Similarly, Applicants confirmed that all SDSIs were significantly different from each other to prevent misidentification; among all pairwise combinations of SDSIs, the greatest homology occurred between SDSI 14 and 18, which had 15 mismatches over 66 bps (**FIG 18B**). Sequencing of the SDSIs confirmed that each of the 96 constructs resulted in a robust and specific signal of mapped reads (**FIG 14B**).

[0119] Applicants selected a pair of primers and corresponding priming regions on each SDSI that are highly specific and show reliable amplification across SDSIs and under standard PCR conditions. Using Primer-BLAST, Applicants predicted that these sequences had limited homology to common organisms and thus were unlikely to amplify nonspecific templates that could outcompete amplification of a primary target. Experimentally Applicants confirmed that the SDSI primers did not produce any nonspecific amplification, including in the presence of cDNA from a nasopharyngeal (NP) swab sample **(FIG 19A)**. The primer pair also had a common length (24 bps), GC content (45.8%), and melting temperature (62℃ and 63℃, respectively in the SDSI+AmpSeq protocol), ensuring their compatibility with many multiplexed PCR reactions, including the most widely used SARS-CoV-2 amplicon sequencing strategy (artic.network/) (**FIG 19B**). Since each SDSI was identically sized and shared a priming region, a similar amplification rate was expected across all SDSIs. Applicants avoided extremes of GC content in SDSI amplicons (range: 33-65%) in order to promote similar amplification rates across different SDSIs and to viral amplicons (e.g., the GC content of the SARS-CoV-2 genome is roughly 37±5%) [19] (**FIG 19C**). Applicants confirmed experimentally that all SDSIs amplified in an ARTIC SARS-CoV-2 PCR reaction with SDSI primers included, in each case yielding a single clean product of the expected size (**FIG 19D**). Furthermore, Applicants observed that GC content did not significantly bias the number of SDSI reads detected in clinical samples (**FIG 19E**).

Example 8 – Validation of an SDSI+AmpSeq SARS-CoV-2 sequencing approach

[0120] Applicants determined that the addition of SDSIs into the ARTIC multiplexed PCR did not detrimentally affect or otherwise alter the amplification of SARS-CoV-2 cDNA from clinical samples. First, to prevent SDSIs from overtaking the amplification and sequencing of SARS-CoV-2 amplicons, Applicants optimized the amount of SDSI added to each reaction through limited titration. Using a randomly selected SDSI (SDSI 49), Applicants found that the highest concentration tested, 600 copies/μL, resulted in reliable SDSI detection with >96% of reads still mapping to SARS-CoV-2 and no apparent alteration in coverage across the genome **(FIG 20A,B)**. Applicants then validated the specificity of the 96 selected SDSIs in a batch of clinical samples to confirm that there was no unpredicted cross-mapping, misidentification, or significant differences in amplification rate (**FIG 15A**). To assess more precisely how the addition of SDSIs would affect SARS-CoV-2 genome sequencing in clinical samples, Applicants processed 14 samples, spanning a range of CT values (CT range = 25-33), with both the standard ARTIC and SDSI+AmpSeq methods. For each amplicon, across all samples, there was no significant difference in coverage between the ARTIC and SDSI+AmpSeq conditions (**FIG 15B**). Even in samples with low viral loads (CT>30), Applicants found that there were no significant differences in amplicon coverage (**FIG 21A**). Additionally, within the 14 samples processed with and without an SDSI, Applicants see a 100% genome concordance rate illustrating the addition of the SDSIs does not impact recovery of accurate genomes.

[0121] As extensive PCR can result in the propagation of numerous types of errors, such as DNA polymerase base substitution errors, PCR recombination events, template switching, and thermocycling induced DNA damage, Applicants further compared SARS-CoV-2 genome concordance between the SDSI+AmpSeq method and unbiased, metagenomic sequencing[9,10,20]. Applicants performed SDSI+AmpSeq on a batch of 89 unique patient samples previously sequenced with unbiased metagenomics[21]. The samples consisted of diverse viral lineages and a broad range of viral loads (CT range = 11.9-37.4; mean = 27.4) with the more sensitive amplicon sequencing method generating more complete genomes at higher CTs (**FIG 22A-D**). Applicants assessed the coverage uniformity between the methods, as increasing uniformity reduces the sequencing depth required to generate reliable genomes[22]. Applicants found that unbiased sequencing had more uniform coverage up to a CT of 25 (N=31, Gini Coefficient =

0.240 ± 0.046 (unbiased) vs 0.428 ± 0.026 (SDSI+AmpSeq)), while SDSI+AmpSeq generated more uniform coverage for samples above a CT of 25 (N=39, Gini Coeff = 0.766 ± 0.265 (unbiased) vs 0.554 ± 0.124 (SDSI+AmpSeq)) (**FIG 22E)**. For the 37 samples that assembled a full genome in both methods, only two out of 332 total single nucleotide variants (SNVs) identified compared to the reference (Wuhan-Hu-1) were divergently identified by SDSI+AmpSeq **(FIG 15C**). Each SNVs was observed in only one sample, and both fell within an ARTIC primer region, despite primer trimming during analysis; this suggests that PCR error from the ARTIC protocol may have contributed to the discrepancy[23]. Manual inspection of one SNV, (a C9565T mutation in unbiased sequencing) indicated the presence of intra-host variation in both methods with a variant allele frequency of 39.4% (SDSI+AmpSeq) and 59.2% (unbiased sequencing). Overall, the discordance rate between SNV calling for SDSI+AmpSeq and unbiased sequencing was 0.6%, a percentage that is reasonable with SNV rates and sequencing based errors. Consistent with previous reports from other groups, ARTIC amplicon sequencing maintains a high level of concordance at the consensus genome level[10], even with the addition of SDSIs.

[0122] Applicants explored a number of other technical modifications to the ARTIC amplicon sequencing protocol in order to improve genome recovery, limit contamination points, and enhance reproducibility of the SDSI approach. Foremost, increasing cDNA length by use of more processive reverse transcriptases improves amplicon coverage (**FIG 23A,B**). Amplification of ARTIC amplicons and SDSIs by Q5 Hot Start High-Fidelity 2x Master Mix results in higher amplification (**FIG 23C, Table 7**). Applicants found that increasing (2X) primer concentrations (20.8nM final) for poor performing amplicons increased coverage in these amplicons, even enabling whole genome recovery for multiple samples supporting that primer rebalancing can enable greater coverage[24,25] **(FIG 23D, FIG 24, Table 4).** Applicants then explored the effects of different numbers of PCR cycles, DNA-hybridization steps, and temperature ramp speeds. Both DNA-hybridization steps and temperature ramping provided no significant changes in amplification **(FIG 23E,F)**. Although it may lead to a potential increase in erroneous SNV calls[23], additional cycles of PCR can be beneficial for low viral load samples by increasing genome coverage uniformity **(FIG 23G).** Using a standardized cDNA input, Applicants found that the DNA Flex library preparation kit resulted in an increased depth of coverage for the SARS-CoV-2 genome across all CT values tested, compared to Nextera XT (**FIG 23H**). To further mitigate the

risk of contamination from such highly amplified libraries, Applicants took advantage of the self-normalizing feature of the DNA flex kit and found that limiting the tagmentation beads by scaling down (.5X) all components of the DNA Flex library construction reagents restricted library over-amplification. Notably, this limitation did not impact final library size distributions or SDSI amplification, while having the desired effect of generating final sequencing libraries at half their original concentrations (**Table 8**). This approach also had the added benefit of nearly halving the library construction cost per sample (**Methods**). Applicants have summarized the results of the optimizations within the full SDSI+AmpSeq protocol (benchling.com/s/prt-R95g0tCxKOeCAqn8lAk3); additionally, Applicants have found that the SDSIs can be easily integrated with numerous protocol alterations.

Example 9 – Implementation of SDSIs to SARS-CoV-2 clinical samples at scale

[0123] The SDSI+AmpSeq method is compatible with a range of viral CTs, SARS-CoV-2 lineages, origin of the patient sample, and laboratory in which the pipeline is implemented demonstrating that this is a robust and flexible approach that can be readily implemented for surveillance. A half plate of SDSIs were used at two large-scale sequencing facilities, the Broad Institute and Jackson Laboratories (JAX), for SDSI+AmpSeq SARS-CoV-2 surveillance across a total of 6,741 clinical samples and controls (JAX: N=3,838; Broad: N=2,903). Individual batches typically consisted of 92 clinical samples with 4 designated water controls. Clinical samples were largely from Maine, Massachusetts, and Rhode Island from December 2020 to July 2021 and covered a wide range of viral CT values (CT 8.4-39.9) and pango lineages (77 total lineages) (**FIG 16A**). The SDSI+AmpSeq method worked robustly despite minor implementation differences in protocols between the two laboratories including alterations in the cDNA synthesis enzymes (SSIV vs Lunascript), CT normalization implementation, and library construction approaches (0.5X Illumina DNA Flex vs Illumina COVID-Seq) (**Methods**).

[0124] The SDSI+AmpSeq is a tractable and easily-implemented method for genome quality control when applied to high-throughput processing of clinical samples. Across thousands of clinical samples, the SDSIs performed consistently and reliably (**FIG 16B,C**). The mean percentage of SDSI reads that mapped to the expected SDSI was above 95% for all SDSIs in both laboratories (**FIG 16B**). This demonstrated that across a large set of highly variable clinical samples, there were no systemic issues of misidentification for specific SDSIs. Additionally,

across all samples from both institutions, the percentage of all SDSI reads in SARS-CoV-2 positive samples averaged 3.71% (90% of samples fell between 0.002-9.989%) (**FIG 16C**). Each SDSI consumed roughly the same read percentage, with no SDSI consistently absent or regularly taking up more than 10% of the sample reads, supporting the prediction that the unique constructs amplified at similar rates. Importantly, this low, but consistent percentage of reads mapping to SDSIs allows for their implementation without needing to greatly increase sequencing depth. Across batches, SDSIs also take up roughly similar shares of the reads, indicating that the SDSI+AmpSeq method is consistent over time. Notably, the SDSIs performed well with and without prior normalization of cDNA based on CT, however normalizing did increase the percentage of SDSI reads **(FIG 21B, FIG 16B left, Methods)**. Normalization of viral CT may provide an additional level of quality control that is especially important for labs with limited sequencing capacities.

Example 10 – SDSI+AmpSeq provides highly confident genome sequencing and analysis

[0125] SDSIs enable detection of sample swaps and contamination events that occur in large scale batch processing which may otherwise go undetected. In a controlled experiment, Applicants demonstrated that the SDSI+AmpSeq approach provides a feasible method to accurately detect contamination. Applicants mixed two SDSIs at various ratios prior to the ARTIC PCR and found that those SDSI ratios were reflected in the sequencing output (**FIG 17A**). With evidence of SDSI's robust detection, uniqueness, and ability to detect intentional contamination, Applicants proceeded to use them to identify sample swaps and contamination in large batch processing. Across thousands of SARS-CoV-2 samples processed, SDSIs detected in samples to which they were not intentionally added allowed for identification of multiple key modes of error (**FIG 17B**). As plotted, a plate without contaminating events or sample swaps should display a simple diagonal pattern with 1-1 matching of expected and observed SDSIs. In some cases, off-diagonal events occur in clear patterns, enabling speculation on the nature of the contamination, clearly demonstrating the utility of SDSIs as an internal control and in-sample label. Applicants observed cases where a plate was likely inverted when SDSI+AmpSeq pool 1 was mixed with pool 2 (**FIG 15B**). The SDSI+AmpSeq approach allows researchers to detect entire flawed batches that may not have been flagged with standard controls (as in the case with the plate inversion where water controls in plate corners would not have been affected). In another example, SDSIs were detected

unexpectedly throughout a batch, indicating that SDSI (and possibly SARS-CoV-2 and other genetic material) contaminated a common reagent.

[0126] SDSI+AmpSeq also enables fine-resolution insight into sample processing errors with high specificity. In one example, SDSI counts indicated columns were unintentionally mixed together (**FIG 17B**). Here, in-sample labeling in all wells allowed researchers to confidently move forward with analyses on unaffected samples. In other cases, samples are associated with both the expected SDSI and SDSIs that were expected in neighboring samples. This indicates a potential spillover event or pipetting errors. Again, genomes generated from samples with suspicious SDSI profiles can be investigated further, and potentially removed from analyses and/or reprocessed. Applicants recommend manual curation of genomes assembled from any samples with <95% of SDSI reads mapping to the expected SDSI. This level of impurity is likely attributable to sample processing contamination, given minimal baseline crosstalk from sources like indexing primer or oligo synthesis observed (Methods, **FIG 25**). Moreover, these patterns of contamination events identified via use of SDSI+AmpSeq illuminated key sources of error in processing pipelines and provided an opportunity to improve processing fidelity in subsequent batches.

[0127] To demonstrate the application of the SDSIs for confident interpretation of sequencing data Applicants used SDSI+AmpSeq to investigate a putative SARS-CoV-2 cluster from Massachusetts General Hospital (MGH) for which the Infection Control Unit suspected nosocomial transmission, a context in which both sample swaps and contamination could easily undermine findings. Applicants sequenced 22 samples with SDSI+AmpSeq (14 samples suspected to be part of the cluster based on epidemiological contact-tracing and 8 unlinked samples as controls), within 24 hours and final genomes were assembled within 52 hours of biosample receipt. Of the 11 samples that Applicants assembled genomes from that were suspected to be part of the cluster, 10 were genetically highly similar (0-1 consensus nucleotide difference) (**FIG 17C)** and distinct from other samples from Massachusetts around the same time **(FIG 26)**, strongly suggesting that this cluster did arise from nosocomial transmission. Analysis of the SDSIs confirmed that genome sequence similarity among cluster-associated samples was not the result of cross-contamination **(FIG 17C)**. Indeed, 23/24 libraries (22 patient samples and 2 water controls) contained >95% SDSI-mapped reads corresponding to the expected SDSI. One sample that was not part of the cluster (MA_MGH_02845) showed 18% of reads from a second SDSI,

which was added to a different sample in the batch (MA_MGH_02839). Applicants re-sequenced both samples implicated in the contamination event. Applicants confirmed that the two genome sequences for MA_MGH_02845 were 100% concordant, and no genome was assembled for MA_MGH_02839 in either attempt, likely due to its very low viral load (CT = 37). This example illustrates how SDSIs can be used to isolate and validate only those samples implicated in contamination events and altogether increase confidence in cluster investigations.

[0128] To further increase the confidence in AmpSeq methods for viral genomics, Applicants sought to capture contamination and sample swaps that might occur before the cDNA stage. Applicants explored the feasibility of modifying the SDSI approach to enable synthetic RNA spike-ins (SRSI) from the same constructs, which could be added to clinical sample RNA to provide end-to-end quality control. For a subset of SDSIs, Applicants included a T7 promoter site to enable in-vitro production of these constructs as RNAs. For two clinical samples representing low (20) and mid (26) CTs, Applicants detected reads from the RNA spike-ins added directly to extracted viral RNA as a proof of principle (**FIG 27**). Notably, this approach did not require any additional protocol modifications, and Applicants therefore expect it to be a highly versatile and user-friendly method when deployed at scale for complete end-to-end sample tracking.

Example 11– Discussion on SDSI+AmpSeq

[0129] Amplicon-based sequencing methods crucially empower rapid, full genome recovery for emerging SARS-CoV-2 variant surveillance; however, robust tools are needed to ensure accuracy in genomic data. SDSI+AmpSeq is a reliable technique for detecting key modes of contamination, addressing this critical gap in standard controls and practices. SDSIs do not compromise genome quality, have been successfully deployed in thousands of clinical samples, and are in use across multiple laboratories with differing protocols. These SDSIs revealed numerous instances of sample swaps and contamination, many of which would go unnoticed with standard batch-level controls. SDSIs further provide critical confidence in the interpretation of clusters of identical genomes, a renewed challenge in the surveillance of more transmissible variants. The common primer design of the SDSI approach enables them to be readily applied to multiple short amplicon designs and sequencing strategies, adding only minor changes to existing protocols and minimal additional cost.

[0130] SDSIs overcome multiple modes of error in the production of amplicon-based genomic sequencing data and are a critical component of quality control measures. The approach is most effective when adopted fully within a laboratory setting and thus Applicants propose routine use of the SDSI+AmpSeq method to flag laboratory-wide contamination. Applicants have implemented SDSI's across diverse approaches and provide an extensively tested protocol with ARTIC v3 and Illumina-based tagmentation. It can also be applied to other sequencing pipelines, though this potentially requires further optimization. The pathogen-exclusion design criteria allows the 96 validated SDSIs to be immediately incorporated into other tiled amplicon panels, such as existing ones for Zika, Ebola, and other viruses of epidemic potential[26,27].

[0131] The SDSI-labeling paradigm is broadly applicable to many amplicon-based needs: amenable to a variety of technical enhancements, flexible to remaining error modes, and expandable to additional targets. One could apply the same design parameters to expand the set of SDSIs, such as to 384 well formats. Additionally, uniquely permuted sets of any size could be created for specific sample batches. To design larger panels of SDSIs, Applicants could use artificial core sequences, rather than excerpting from archaea. Primer sites could also be easily adapted for integration with new advancements in amplicon sequencing, like tailed primer approaches or new primer schemes[28–32]. In its current implementation, the SDSIs detect contamination or workflow errors that occur during and after amplification, but not issues arising at the RNA or cDNA generation stage, and act qualitatively, rather than quantitatively. Further refinement of the RNA spike-in approach could address other modes of contamination, enabling end-to-end sample tracking at scale. Future work improving quantification and SDSI analysis pipelines may enable them to serve as within sample controls, since samples or batches with outlier SDSI read counts may reveal missing or defective PCR components, incomplete mixing, thermocycling issues, or other types of experimental error.

The integration of SDSIs can mitigate a critical vulnerability of amplicon-based sequencing while preserving the many advantages, increasing the robustness of its use across laboratory and clinical settings. Adoption of controls across the viral surveillance community would increase accuracy and integrity of genomic data worldwide. Looking forward, SDSIs could serve as a crucial component in improving data integrity in amplicon based genomic sequencing beyond infectious disease surveillance, such as food safety, species identification and environmental sampling.

Example 12 – Methods

SDSI design and in silico validation

[0132] Applicants designed synthetic DNA fragments that each contained a 140 bp unique sequence and constant priming regions. Core SDSI sequence homology to sequences from various organisms was predicted by a permissive BLAST search (blastn; 5000 max targets; E=10; word size=11; no mask for low complexity). Applicants considered homologies identified with this BLASTn search described above that were additionally >50 bps (>35% query cover) and >90% sequence identity to be significant homologies. For all 96 selected SDSIs, there were no such significant homologies when results were filtered to all *Homo sapiens* (taxid:9606) or viral (taxid:10239) sequences in the NCBI database. For significant homologies to bacterial or eukaryotic sequences in the NCBI database (excluding archaea: taxid:2157), Applicants report both the SDSI and the genus it mapped to in each case (**FIG 18A**). Using the same BLASTn parameters, Applicants also mapped SDSIs against a custom database including SDSI core sequences and found no significant homologies between SDSIs. As there were no significant homologies between SDSIs and human, virus, or other SDSI sequences, Applicants noted the maximum alignment scores for any non-significant homology identified and reported these (**FIG 18B)**.

[0133] Applicants confirmed that SDSI primers and amplicons were predicted to amplify specifically and consistently with ARTIC v3 amplicons. Applicants used Primer-BLAST to predict 50-5000 bp amplicons produced on templates in the entire nr database; no amplicons were identified. Applicants calculated the length and GC content of SDSI primers and full SDSI amplicon sequences and ARTIC v3 primers and amplicons using Geneious Prime (2019.2.1) and compared their distributions (**FIG 19B-C**). ARTIC and SDSI primer melting temperatures were matched and calculated using the New England Biolabs online calculator (tmcalculator.neb.com).

SDSI experimental validation

[0134] Applicants sought to validate in silico predictions for the performance of the SDSI primers and amplicons. Applicants ordered primers (IDT) (oligo sequences in **Supplementary Data File 1**) and performed qPCR using the Q5 Hotstart 2x Mastermix, with 500nM SDSI primers and 0.17X SYBR Gold (ThermoFisher #S11494), and without ARTIC primer pools. Applicants performed this assay in triplicate in 10μL reactions on a QuantStudio 6 with the following cycling

conditions: 95°C for 30 seconds, followed by 35 cycles of 95°C for 15 seconds and 65°C for 5 minutes. Applicants tested 4 conditions: (1) 0.5µL of an SDSI gene block (IDT) (1pM), (2) 0.5µL of an SDSI gene block + 0.5µL of cDNA from an NP swab, (3) 0.5µL of cDNA from an NP swab, and (4) no template to detect any nonspecific amplification of the primers (**FIG 19A**). Applicants performed PCR on each SDSI oligo, using the standard SDSI+AmpSeq PCR conditions (benchling.com/s/prt-R95g0tCxKOeCAqn8lAk3), then ran the PCR products on a 2.2% agarose gel to confirm that these primers amplified the SDSIs and that the product was clean and of the expected size (**FIG 19D**).

[0135] Applicants ordered unique oligos as TruGrade ultramers (IDT), then resuspended and stored them at 10µM in water (oligo sequences in **Table 1**). Further characterization for identification of 96 SDSIs was achieved by direct PCR amplification with primers containing the constant SDSI handle and an Illumina P5/P7 adapter followed by sequencing with a MiSeq Nano 2x150bp kit (Illumina #MS-102-2002). SDSI reads were quantified by mapping each SDSI against other SDSIs with the align_and_count_multiple_report wdl implemented in Terra, as described below, and purity and sequence fidelity of SDSIs was achieved by calculating the percentage of reads mapping to each SDSI out of total SDSI reads (**FIG 14B**). Given these same data, Applicants explored the SDSI mapping stringency threshold. Applicants determined whether each SDSI was uniquely identified over a range of SDSI stringency thresholds (0.01%-50% of SDSI reads mapping, with a step size of 0.01%) (**FIG 25**). Applicants tested 142 total unique SDSIs; all SDSIs amplified successfully with high sequence fidelity and purity (>95% of reads mapped to the expected SDSI in the experiment described above). The final set of 96 SDSIs were chosen after first pass validation in a combination of clinical sample amplification tests, GC cutoffs, and sequence homology cutoffs. SDSIs excluded because of poor amplification or impurity in clinical sample processing were not retested to determine whether error was technical or biological.

Sample collection and study design

[0136] Research was conducted at the Broad Institute with an exempt determination from the Broad Office of Research Subjects Protections and with approval from the MIT Institutional Review Board under protocol #1612793224. Samples were obtained from Massachusetts General Hospital (MGH), Massachusetts Department of Public Health, the Rhode Island Department of Public Health and the Broad Institute Clinical Research Sequencing Platform. Samples from

Massachusetts General Hospital (MGH) fall under Partners Institutional Review Board under protocol #2019P003305. Samples were secondary-use or residual clinical and diagnostic specimens (referred to collectively throughout as clinical samples), obtained by researchers under a waiver of consent. All samples were nasopharyngeal or anterior nares swabs in a stabilizing medium (e.g., MTM or VTM). These unique biological materials are not available to other researchers as they are human patient samples from clinical excess material and thus are of limited volume. Samples sequenced at Jackson Laboratories (JAX) were approved under protocol 2020-NHSR-019-BH.

Viral CT Determination

[0137] Viral cycle threshold (CT) for all samples sequenced at the Broad Institute were obtained using the CDC RT-qPCR assay with the N1 probe as previously described[21]. Viral CTs for samples sequenced at JAX were obtained from various providers and thus the RT-qPCR assays used are variable.

CT Normalization

[0138] CT normalization was performed by first setting a desired mock viral CT and calculating the difference between this desired mock viral CT and the measured viral CT of a given sample, rounding to the nearest whole number. Applicants next calculated the number of doublings required for the mock viral CT (assuming 100% PCR efficiency) and multiplied this by the volume of cDNA input to be used for the normalization. The final volume of water used to dilute the cDNA was the doubling factor minus the volume of cDNA input. An example calculation is illustrated below:

Example of CT Normalization:

N= Difference between actual and mock
X= Volume (μL) of cDNA to use for normalization
DF=Doubling factor is $X(2^N)$
Volume water for dilution (μL)=DF-X

Actual viral CT=23
Desired mock viral CT=27
N=27-23=4
X=1μL
DF= $1(2^4)$
Volume water for dilution (μL) = 16-1 = 15μL

Add 1µL of cDNA to 15µL nuclease free water.

[0139] This CT normalization was done for certain method development samples which are described throughout the manuscript as being "mock diluted" or "normalized to CT X". The nosocomial cluster was normalized to CT 27. The majority of batch data generated at the Broad Institute underwent CT normalization to CT 25. Batch data from JAX did not undergo CT normalization. CT normalization of the cDNA prior to the ARTIC PCR should reduce the potential for generating excessively large libraries from very high viral load samples, keep the percentage of SDSI reads in a detectable range (**FIG 21B**), and further reduce the need for additional normalization steps later in the pipeline.

cDNA generation and ARTIC amplification optimization

Reverse transcriptase

[0140] Applicants tested reverse transcriptase enzymes using extracted RNA from four SARS-CoV-2 positive clinical samples (CTs = 13.9, 23.9, 29.6, 33.6) **(FIG 23A,B)**. Applicants added 2µL of purified DNase treated RNA as input into SuperScript III (Thermo #18080093), SuperScript IV (Thermo #18091050), or SuperScript IV VILO (Thermo #11756500). Superscript IV (SSIV) reactions incubated at room temperature for 10 minutes, followed by 50°C for 60 minutes and an inactivation step at 80°C for 10min. Superscript IV VILO shared the same protocol, but with a temperature of 85°C for the inactivation step. Applicants input 2.5µL of cDNA for ARTIC pool #1 PCR under standard conditions for 40 cycles. Applicants then tested the resulting pool #1 using the scaled down Illumina DNA Flex library construction (as described in Methods below) and sequenced on Illumina Miseq (V2 reagent kit) with 2 x 150 bp paired end sequencing.

ARTIC PCR enzyme

[0141] Applicants tested PCR enzyme efficiency using extracted RNA from SARS-CoV-2 positive clinical samples followed by cDNA generation using SuperScript IV and diluted the resulting cDNA to a mock CT value of 35 for standardization across all PCR enzyme tests. Applicants set up the standard ARTIC PCR pool #1 and pool #2 using an input of 2.5µL, altering only the PCR enzyme and corresponding buffer. Applicants tested NEB Q5 Hot Start High-fidelity 2x Master Mix (Q5 2X MM) (NEB #M0494L), NEB Q5 Hot Start High-fidelity 2x Master Mix plus .01% SDS, NEB Q5 Ultra II Master Mix (NEB #M0544L), KAPA HiFi HotStart (Roche

#KK2601), and KOD Hot Start DNA polymerase (Sigma-Aldrich #71842) **(FIG 23C)**. Applicants quantified the resulting ARTIC PCR amplicons using a High Sensitivity DNA Qubit kit, then input 25ng from each pool (50ng total) into scaled down Illumina DNA Flex library construction. The resulting libraries (except Q5 plus .01% SDS, which had no visible product using the Tapestation D1000 High Sensitivity Kit) were quantified and pooled on Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

Rehybridization PCR

[0142] Applicants optimized PCR cycling conditions on mock CT 35 cDNA (generated as described above) using standard ARTIC PCR primer conditions. Applicants performed a catch-up/rehybridization PCR under the following conditions: 98°C for 30s, 95°C for 15s then 65°C for 5 min (10 cycles), 95°C for 15s then 80°C for 30s then 65°C for 5 min (2 cycles), 95°C for 15s then 65°C for 5 min (8 cycles), 4°C hold **(FIG 23E)**. Applicants quantified the resulting ARTIC PCR amplicons using a High Sensitivity DNA Qubit kit, then input 25ng from each pool (50ng total) into scaled down Illumina DNA Flex library construction. Applicants then quantified these libraries and pooled on Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

Cycle Test

[0143] Applicants further optimized ARTIC PCR by modifying PCR cycle numbers. Extracted RNA from six SARS-CoV-2 positive clinical samples ranging from CT 27-37 were converted to cDNA with Superscript IV and amplified under standard ARTIC PCR reaction components (with Q5 2X MM) modifying the final number of cycles of PCR from 35, 40 and 45 **(FIG 23G)**. Applicants quantified cDNA and used at a standard 50ng of input for scaled down Illumina DNA Flex Library Construction, then quantified the resulting libraries and pooled on Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

Ramp Test

[0144] Applicants used mock CT 35 cDNA to test the effect of decreased ramp speed on genome recovery and coverage. ARTIC PCR conditions for this experiment were 98°C for 30 seconds, followed by 40 cycles of 95°C for 15 seconds and 65°C for 5 minutes with a cooling and heating ramping speed of 3°C/s. Applicants tested a slow ramp PCR protocol with the ramp speed reduced to 1.5°C/s **(FIG 23F)**. Libraries were constructed with Illumina DNA Flex and were sequenced on Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

Primer Concentration Optimization

[0145] Under standard ARTIC protocol conditions, Applicants ordered lyophilized ARTIC v3 primers from IDT and resuspended in water at 100μM each. Pool #1 primers consisted of all odd numbered amplicons whereas pool #2 primers consisted of all even numbered amplicons. To generate the 100μM pool #1 primer stock, Applicants combined 5μL of each 100μM pool #1 primer, and repeated this protocol for the even numbered primers to give a 100μM pool #2 primer stock. Applicants selected a total of 20 amplicons as regions of low coverage from previous sequencing data (**Table 4**). Low coverage amplicons were present in both pools, with 11 coming from pool #1 and 9 coming from pool #2. For the primer 2X pools, Applicants spiked in primers for the corresponding amplicons at 2X the concentration (20.8nM final) of the other primers in the pool. For these low coverage primers, Applicants used 10μL of the 100μM stock rather than 5μL. Applicants diluted both the original and 2X primer pools 1:10 in nuclease free water to generate a 10μM working stock. Applicants then selected 8 samples with varying CT values to determine if selectively increasing primer concentrations reduced amplicon dropout **(FIG 23D)**. Applicants used the SDSI+AmpSeq protocol (without the SDSI or SDSI primers) and processed each sample with both the original primer pool, as well as the 2X primer pool, then sequenced these 16 samples on an Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing. Only 6 of the 8 samples generated complete genomes (>98%) in both conditions and were used for further analysis.

CT normalization experiment

[0146] The CT normalization experiment was performed by taking four individual clinical samples (CT = 18-25) with four randomly selected SDSIs and either not normalizing the cDNA or normalizing to CT 25, 26, or 27 prior to the ARTIC PCR **(FIG 21B)**. Samples were processed with the standard SDSI+AmpSeq protocol described below and were sequenced on a NextSeq 500 Mid Output Kit v2.5 (300 Cycles).

Illumina DNA Flex

[0147] Applicants performed a head-to-head comparison of standard Illumina Nextera DNA Flex and Nextera XT (Illumina #FC-131-1096) library construction kits **(FIG 23H)**. The Nextera XT protocol was performed as previously described[21,33]. Both library construction methods were compared on post ARTIC v1 PCR amplicons from clinical samples. In short, applicants amplified samples with a range of SARS-CoV-2 viral CT values (CTs = 22.9, 26.2, 30.3) with ARTIC v1

primers, producing 400 bp size fragments. Applicants then quantified amplicons from each ARTIC primer pool and pooled in equal molar concentrations. Standard Nextera DNA Flex input was 100ng (50ng from each pool) and 1ng (.5ng from each pool) for Nextera XT. Applicants quantified and pooled the resulting libraries before sequencing on an Illumina Miseq (V2 reagent kit) with 2 x 150 paired end sequencing.

[0148] Applicants optimized Illumina DNA Flex library construction (Illumina #20018705) construction with the goal of reducing normalization steps, cost and increasing throughput. Applicants scaled down (.5X) Illumina DNA Flex throughout the standard Illumina sequencing protocol, also scaling down sample input for a total of 50ng (25ng from each primer pool). Due to the CT normalization step, applicants removed the pre-DNA Flex DNA concentration and pooling step. Applicants used 1-2µL of post ARTIC PCR amplicon as input into the scaled down DNA Flex library construction and performed post library construction quantification and pooling with more uniform library size and concentration, further reducing time and cost of pooling libraries for sequencing. This protocol was used for all method development experiments, the cluster investigation, and a portion of the batch data generated from both the Broad Institute and JAX.

SDSI+AmpSeq SDSI titration in ARTIC SARS-CoV-2 sequencing

[0149] To determine an optimal concentration for SDSIs in ARTIC SARS-CoV-2 sequencing, applicants diluted SDSI 49 to 0.6, 6, 60, and 600 copies/µL (1, 0.1, 0.01, and 0.001fM); 1µL of SDSI 49 was added to 5µL of cDNA, to be split to 2x3µL for each ARTIC pool (**FIG 20, Table 1**). SDSI primers were added to each ARTIC pool with a final concentration of 40nM. The cDNA from one clinical sample (MA_MGH_00195; CT =16) was mock diluted to CT 20,25,30, and 35 for this experiment using the protocol described within the CT normalization section. Based on the results of this experiment, SDSIs were used at 6e2 copies/µL (1fM) for all method development data. Batch processing modifications to this approach from the Broad Institute and Jackson Laboratories are detailed below.

SDSI+AmpSeq Protocol

[0150] Full protocol details can be found here: benchling.com/s/prt-R95g0tCxKOeCAqn8lAk3 **(FIG 13)**. In short, cDNA synthesis is performed on 2.5µL of DNAse-treated viral RNA with SSIV following the manufacturer's protocol with an extension of the 50°C incubation from 10 minutes to 60 minutes. An additional cDNA normalization step can be

performed (see above) or one can move directly into the ARTIC PCR by taking 5μL of cDNA and mixing this with 1uL of a 1fM SDSI (equal to 600 copies/μL). After mixing, split into 2 x 3μL aliquots and add ARTIC primer pool 1 or pool 2, as well as 1μM of the spike-in forward and reverse primers (40nM final concentration in the ARTIC pool). The ARTIC PCR conditions were 98°C for 30 seconds, followed by 40 cycles of 95°C for 15 seconds and 65°C for 5 minutes. Pool 1 and pool 2 PCR reactions were combined and taken through library construction with scaled down Illumina DNA Flex.

Broad Institute Sample Processing

[0151] The batch data from the Broad Institute was generated using SDSI+AmpSeq with minor modifications **(FIG 16)**. In short, SSIV was used for cDNA synthesis. Q5 2X MM was used for the ARTIC PCR which was run for 35 cycles. The SDSIs were spiked in at 6e3 copies/μL and the SDSI specific primers were added to each ARTIC pool at a final concentration of 40nM. Library construction was performed either with the scaled down Illumina DNA Flex (previously described) or COVID-seq (Illumina #20043675). Samples were sequenced on a NovaSeq 6000 SP Reagent Kit v1 (300 cycles) or v1.5 kits (300 cycles), or NextSeq 500 v2 kit (300 cycles).

[0152] The GC percent for each SDSIs and percent SDSI reads over total reads correlation for SDSI (2-48) was performed with the samples sequenced at the Broad Institute (N=2,903) **(FIG 19E)**. A linear regression was used to evaluate significance (p-value = 0.8160).

Jackson Laboratory Sample Processing

[0153] Data generated at Jackson Laboratory (JAX) used two different protocols publicly available here: github.com/tewhey-lab/SARS-CoV-2-Consensus **(FIG 16)**. All samples included 6e2 copies/μL of SDSIs and the SDSI specific primers were added to each ARTIC pool at a final concentration of 4nM. Samples processed from December 2020 to April 2021 used Lunascript (NEB #E3010) for cDNA synthesis and Q5 2X MM for the ARTIC PCR which was run for 35 cycles. These samples used scaled down Illumina DNA Flex for library construction. Samples sequenced after April 2021 used the standard COVID-seq protocol. All samples were sequenced on a NextSeq500 using paired 75 bp reads by the Genome Technology group on Jackson Laboratory's Bar Harbor campus. The validation of all SDSIs in clinical samples (**FIG 15A**) was performed with this protocol and is presented as the percent of SDSI reads over the total of all reads for each sample.

[0154] Of note, the SDSIs (used at the lowest recommended concentration of 6e2 copies/uL) were reliably detected in the samples sequenced at JAX. This reliable detection however is also dependent on the sequencing depth used by the institution.

SDSI impact on genome recovery

[0155] For +/- SDSI experiments testing impact on recovery of viral genomes, fourteen clinical samples spanning a range of CTs (CT = 17.6-30) were selected **(FIG 15B, FIG 21A)**. Samples were CT normalized and split after cDNA synthesis into 2 x 5µL aliquots. Samples below CT 20 were normalized to CT 25 and samples between CT 20-25 were normalized to CT 26. Fourteen randomly selected SDSIs were used with each sample receiving either an SDSI (600 copies/µL) and the SDSI specific primers (40nM final concentration in the ARTIC pool) or just the ARTIC pool 1 and pool 2 mastermix with additional nuclease free water and no SDSI primers. Samples were processed according to the SDSI+AmpSeq protocol using scaled down Illumina DNA Flex for library construction, sequenced on a NextSeq 500 Mid Output Kit v2.5 (300 Cycles) and analyzed as described below.

[0156] Statistical analysis for the plus/minus SDSI experiment involved analysis of the mean coverage for all 98 amplicons for the full sample set with a two-tailed Mann Whitney t-test and multiple comparison two-stage step-up Benjamini, Krieger, and Yekutieli test with FDR set to 5%. All 98 amplicons were found to be not significantly different (p-value > 0.05) between the plus and minus SDSI group. Samples were also separated into three CT bins (CT <27 (n=4), 27-29 (n=6), CT>30 (n=4)) and this test repeated for each CT bin. This analysis also revealed that there was no significant difference (p-value > 0.05) in the mean coverage across any amplicon for any CT bin.

Intentional SDSI contamination experiment

[0157] The intentional contamination experiment used SDSI 87 and SDSI 94 (SDSI 87: SDSI 94). The SDSIs were mixed at five different proportions (100:0, 75:25, 50:50, 25:75, and 0:100) **(FIG 17A)**. Each condition was performed in duplicate. All validation experiment samples were processed according to the SDSI+AmpSeq protocol using scaled down Illumina DNA Flex for library construction. Samples were processed with the standard SDSI+AmpSeq protocol and sequenced on a NextSeq 500 Mid Output Kit v2.5 (300 Cycles).

Production and application of synthetic RNA spike-ins (SRSI)

[0158] Applicants ordered SDSI oligos with minor modifications to enable in-vitro transcription of RNAs (including a T7 promoter upstream of the SDSI amplicon, as well as 17 bps of constant sequence within the primer region) (Twist Bioscience) (sequences in attached **Sup Data File 1**). For two SDSIs (SDSI 1 and SDSI 4) applicants in-vitro transcribed RNA using a T7 transcription kit (NEB E2050), quantified by RNA screen tape (Agilent 5067-5579 and 5067-5580), then diluted in water to 10fM (6,000 copies/μL), 1fM (600 copies/μL), 100aM (60 copies/μL), and 10aM (6 copies/μL). Applicants added 1μL of SRSI at each concentration directly to 5 μL of RNA from two patient samples with high and intermediate viral loads, respectively, and prepared sequencing libraries using the SDSI+AmpSeq protocol (without the SDSI addition step at the cDNA stage). For the sample with a high viral load, applicants performed a dilution at the cDNA stage (diluting 32-fold for a mock Ct of 25 rather than 20). Reads mapping to unique SDSI sequences and SARS-CoV-2 were quantified using the align_and_count_multiple_report and assemble_refbased wdls respectively, and % SDSI/total reads was reported (**FIG 27**).

Computational analysis workflow

[0159] Applicants analyzed sequencing data on the Terra platform (app.terra.bio) using viral-ngs 2.1.28 with workflows that are publicly available on the Dockstore Tool Repository Service (dockstore.org/organizations/BroadInstitute/collections/pgs). Samples were demultiplexed using the demux_plus workflow with a spike in database file for the SDSIs. Applicants performed any separate analyses to quantify read counts, including those for SDSIs, with the align_and_count_multiple_report workflow with the relevant database. For most analyses involving direct comparisons between samples, applicants performed downsampling to the lowest number of reads passing filter with the downsample workflow. Applicants performed assembly using the assemble_refbased workflow to the following reference fasta: *www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta*. Applicants used iVar version 1.2.1 for primer trimming on all samples followed by assembly with minimap2 set to a minimum coverage of either 3, 10, or 20, skipping deduplication procedures. The computational pipeline for all samples sequenced at JAX is publicly available at the following: github.com/tewhey-lab/SARS-CoV-2-Consensus.

[0160] Samples from the batch data were subset in the following way for analysis. All samples with a present SDSI were used for the percent of SDSI reads out of the sum of all SDSI reads analysis (JAX: N=3,838, Broad: N=2,903). Samples with known experimental contamination errors or where the dominant (>50%) SDSI was not the correct SDSI were removed. For the percent of SDSI reads over the total of all sequenced reads analysis (JAX: N=3,093, Broad: N=2,670), non-template controls (waters) and clinical samples with no detectable viral load (CT>40 or not detected via qPCR as described above) were removed from analysis.

Metagenomic sequencing and comparison

[0161] Metagenomic sequencing data and genome assemblies used for the comparison of amplicon-based sequencing were prepared, sequenced, analyzed as described previously,[21] and the data are publicly available at NCBI's GenBank and SRA databases under BioProject PRJNA622837. Applicants prepared amplicon sequencing libraries from the sample RNA extract following the SDSI+AmpSeq protocol (**FIG 13**). In order to increase sample throughput and bypass an additional more laborious quantification step post the ARTIC PCR, applicants normalized cDNA samples that had a high viral load (CT<27) to a CT of 27. To prepare for the ARTIC PCR, applicants transferred 5µL of the normalized cDNA to a new plate and added 1µL of a SDSI (600 copies/µL). After mixing, applicants transferred 3µL to a new plate, added ARTIC PCR pool #1 mastermix and pool #2 mastermix to the respective plates, and on a thermal cycler incubated at 98°C for 30s, followed by 40 cycles of 95°C for 15s and 65°C for 5min. Applicants then combined in equal molar amounts of amplified samples for a total of 50ng and processed through .5X Illumina Flex library construction pipeline. Applicants sequenced the concordance data set on a NovaSeq 6000 SP Reagent Kit v1 (300 cycles) and analyzed as detailed in the methods below. For SNV analysis, the coverage depth over each divergent SNV was greater than 1000X for both platforms, and both SNV calls persisted at relaxed (n=3) and conservative (n=20) minimum coverage thresholds. Primer trimming using iVar version 1.2.1 was manually confirmed.

Suspected nosocomial cluster investigation

[0162] Applicants received NP swab samples in UTM and extracted RNA from 200µL of biosample as previously described[21]. Applicants prepared amplicon sequencing libraries as described above and analyzed them as detailed in the methods below. A pairwise distance was calculated between all partial genomes (>80% complete), excluding gaps, to determine whether

samples were likely to be the result of nosocomial transmission **(FIG 17C)**. Applicants calculated the proportion of reads that mapped to a given SDSI out of all reads that mapped to any SDSI. Data has been made available in both the Short Read Archive and NCBI GenBank under Bioproject PRJNA622837. GenBank accessions for SARS-CoV-2 genomes from this set of samples are MW454553 - MW454562.

[0163] For phylogenetic tree reconstruction applicants placed the suspected nosocomial cluster in a broader genomic context by performing a subsampling of the genome sequences available in GISAID (as of January 26 2021) **(FIG 26)**. Applicants used the sarscov2_nextstrain workflow to perform a Massachusetts-weighted subsampling of samples from 1 January 2020 - 1 November 2020. Applicants' subsampled dataset included 3146 sequences; 1449 samples from Massachusetts, 1425 samples from elsewhere in the United States and 283 from other countries. Applicants constructed a maximum likelihood tree using iqtree with a GTR substitution model and edited and interpreted the tree in Figtree v1.4.4.

Data presentation

[0164] Data analysis and graphing was performed using R Statistical Software (version 1.3.959; R Foundation for Statistical Computing, Vienna, Austria), GraphPad PRISM (version 9.0.2; GraphPad Software, La Jolla California USA, www.graphpad.com) and Python (version 3.7). Applicants created original figures using BioRender (BioRender.com).

Code availability

[0165] Viral genomes were processed using the Terra platform (app.terra.bio) using viral-ngs 2.1.1 with workflows that are publicly available on the Dockstore Tool Repository Service (dockstore.org/organizations/BroadInstitute/collections/pgs). Downstream analyses were performed using Geneious or standard R packages. Custom scripts used to generate figures are available upon request.

Methods

Data availability

[0166] Sequences and genome assembly data are publicly available on NCBI's Genbank and SRA databases under BioProject PRJNA622837. GenBank accessions for SARS-CoV-2 genomes newly reported in this study are MW454553 - MW454562.

References

1. Washington, N. L. et al. Genomic epidemiology identifies emergence and rapid transmission of SARS
CoV-2 B.1.1.7 in the United States. medRxiv (2021) doi:10.1101/2021.02.06.21251159.

2. Walensky, R. P., Walke, H. T. & Fauci, A. S. SARS-CoV-2 Variants of Concern in the United States—Challenges and Opportunities. JAMA vol. 325 1037 (2021).

3. Wang, P. et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature 593, 130–135 (2021).

4. Focosi, D., Tuccori, M., Baj, A. & Maggi, F. SARS-CoV-2 Variants: A Synopsis of In Vitro Efficacy Data of Convalescent Plasma, Currently Marketed Vaccines, and Monoclonal Antibodies. Viruses 13, (2021).

5. Wang, P. et al. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. Cell Host Microbe 29, 747–751.e4 (2021).

6. Naveca, F. et al. SARS-CoV-2 reinfection by the new Variant of Concern (VOC) P. 1 in Amazonas, Brazil. virological. org (2021).

7. Organization, W. H. & Others. Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health, 8 January 2021. (2021).

8. COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk. An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet Microbe 1, e99–e100 (2020).

9. Chiara, M. et al. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Brief. Bioinform. (2020) doi:10.1093/bib/bbaa297.

10. Charre, C. et al. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. Virus Evol 6, veaa075 (2020).

11. Rausch, J. W., Capoferri, A. A., Katusiime, M. G., Patro, S. C. & Kearney, M. F. Low genetic diversity may be an Achilles heel of SARS-CoV-2. Proceedings of the National Academy of Sciences of the United States of America vol. 117 24614–24616 (2020).

12. Endo, A., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott, S., Kucharski, A. J. & Funk, S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Res 5, 67 (2020).

13. Lagerborg, K. A., Watrous, J. D., Cheng, S. & Jain, M. High-Throughput Measure of Bioactive Lipids Using Non-targeted Mass Spectrometry. Methods Mol. Biol. 1862, 17–35 (2019).

14. Boja, E. S. & Rodriguez, H. Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins. Proteomics 12, 1093–1110 (2012).

15. Chen, K. et al. The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. Molecular and Cellular Biology vol. 36 662–667 (2016).

16. Illumina COVIDSeq Test. emea.illumina.com/products/by-type/ivd-products/covidseq.html.

17. Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 21, 1543–1551 (2011).

18. Quail, M. A. et al. SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. BMC Genomics 15, 110 (2014).

19. Dilucca, M., Forcelloni, S., Pavlopoulou, A., Georgakilas, A. G. & Giansanti, A. Codon usage and evolutionary rates of the 2019-nCoV genes. Cold Spring Harbor Laboratory 2020.03.25.006569 (2020) doi:10.1101/2020.03.25.006569.

20. Potapov, V. & Ong, J. L. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLOS ONE vol. 12 e0169774 (2017).

21. Lemieux, J. E. et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science 371, (2021).

22. So, A. P. et al. A robust targeted sequencing approach for low input and variable quality DNA from clinical samples. NPJ Genom Med 3, 2 (2018).

23. Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 20, 8 (2019).

24. Pipelines R&D, D. N. A. et al. COVID-19 ARTIC v3 Illumina library construction and sequencing protocol v5. protocols.io (2020) doi:10.17504/protocols.io.bibtkann.

25. Lam, C. et al. Sars-CoV-2 Genome Sequencing Methods Differ In Their Ability To Detect Variants From Low Viral Load Samples. J. Clin. Microbiol. JCM0104621 (2021).

26. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228–232 (2016).

27. Metsky, H. C. et al. Zika virus evolution and spread in the Americas. Nature 546, 411–415

(2017).

28. Gohl, D. M. et al. A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. BMC Genomics 21, 863 (2020).

29. Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R. & Kuroda, M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. PLoS One 15, e0239403 (2020).

30. Tyson, J. R. et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv (2020) doi:10.1101/2020.09.04.283077.

31. VarSkip: VarSkip multiplex PCR designs for SARS-CoV-2 sequencing. (Github).

32. primer_schemes/nCoV-2019 at master · artic-network/artic-ncov2019. github.com/artic-network/artic-ncov2019.

33. Wong, F., & Collins, J. J. (2020). Evidence that coronavirus superspreading is fat-tailed. In *Proceedings of the National Academy of Sciences* (Vol. 117, Issue 47, pp. 29416–29418). doi.org/10.1073/pnas.2018490117.

34. Adam, D. C., Wu, P., Wong, J. Y., Lau, E. H. Y., Tsang, T. K., Cauchemez, S., Leung, G. M., & Cowling, B. J. (2020). Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. In *Nature Medicine* (Vol. 26, Issue 11, pp. 1714–1719). doi.org/10.1038/s41591-020-1092-0.

35. Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18.

36. Antonov, J., Goldstein, D. R., Oberli, A., Baltzer, A., Pirotta, M., Fleischmann, A., Altermatt, H. J., & Jaggi, R. (2005). Reliable gene expression measurements from degraded RNA by quantitative real-time PCR depend on short amplicons and a proper normalization. *Laboratory Investigation; a Journal of Technical Methods and Pathology*, *85*(8), 1040–1050.

37. *Artic Network*. (n.d.). Retrieved February 2, 2021, from artic.network/

38. Baker, D. J., Kay, G. L., Aydin, A., Le-Viet, T., Rudder, S., Tedim, A. P., Kolyva, A., Diaz, M., de Oliveira Martins, L., Alikhan, N.-F., Meadows, L., Bell, A., Gutierrez, A. V., Trotter, A. J., Thomson, N. M., Gilroy, R., Griffith, L., Adriaenssens, E. M., Stanley, R., … O'Grady, J. (2020). CoronaHiT: large scale multiplexing of SARS-CoV-2 genomes using Nanopore

sequencing. In *Cold Spring Harbor Laboratory* (p. 2020.06.24.162156). doi.org/10.1101/2020.06.24.162156.

39. Dearlove, B., Lewitus, E., Bai, H., Li, Y., Reeves, D. B., Gordon Joyce, M., Scott, P. T., Amare, M. F., Vasan, S., Michael, N. L., Modjarrad, K., & Rolland, M. (2020). A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. In *Proceedings of the National Academy of Sciences* (Vol. 117, Issue 38, pp. 23652–23662) doi.org/10.1073/pnas.2008281117.

40. Houldcroft, C. J., Beale, M. A., & Breuer, J. (2017). Clinical and biological insights from viral genome sequencing. *Nature Reviews. Microbiology*, *15*(3), 183–192.

41. Klempt, P., Brož, P., Kašný, M., Novotný, A., Kvapilová, K., & Kvapil, P. (2020). Performance of Targeted Library Preparation Solutions for SARS-CoV-2 Whole Genome Analysis. *Diagnostics (Basel, Switzerland)*, *10*(10). doi.org/10.3390/diagnostics10100769.

42. Mathieu-Daudé, F., Welsh, J., Vogt, T., & McClelland, M. (1996). DNA Rehybridization During PCR: The "C o t Effect" and Its Consequences. *Nucleic Acids Research*, *24*(11), 2080–2086.

43. Metsky, H. C., Siddle, K. J., Gladden-Young, A., Qu, J., Yang, D. K., Brehio, P., Goldfarb, A., Piantadosi, A., Wohl, S., Carter, A., Lin, A. E., Barnes, K. G., Tully, D. C., Corleis, B., Hennigan, S., Barbosa-Lima, G., Vieira, Y. R., Paul, L. M., Tan, A. L., … Matranga, C. B. (2019). Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nature Biotechnology*, *37*(2), 160–168.

44. No, J. S., Kim, W.-K., Cho, S., Lee, S.-H., Kim, J.-A., Lee, D., Song, D. H., Gu, S. H., Jeong, S. T., Wiley, M. R., Palacios, G., & Song, J.-W. (2019). Comparison of targeted next-generation sequencing for whole-genome sequencing of Hantaan orthohantavirus in Apodemus agrarius lung tissues. *Scientific Reports*, *9*(1), 16631.

45. Popa, A., Genger, J.-W., Nicholson, M. D., Penz, T., Schmid, D., Aberle, S. W., Agerer, B., Lercher, A., Endler, L., Colaço, H., Smyth, M., Schuster, M., Grau, M. L., Martínez-Jiménez, F., Pich, O., Borena, W., Pawelka, E., Keszei, Z., Senekowitsch, M., … Bergthaler, A. (2020). Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Science Translational Medicine*, *12*(573). doi.org/10.1126/scitranslmed.abe2555.

46. Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T. F., Beutler, N. A., Burton, D. R., Lewis-Ximenez, L. L., de Jesus, J. G., Giovanetti, M., Hill, S. C., Black, A., Bedford, T., Carroll, M. W., Nunes, M., … Loman, N. J. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols*, *12*(6), 1261–1276.

47. *SARS-CoV-2 COVID-19 Coronavirus research and surveillance*. (n.d.). Retrieved February 12, 2021, from www.paragongenomics.com/product/cleanplex-sars-cov-2-panel/

Sethuraman, N., Jeremiah, S. S., & Ryo, A. (2020). Interpreting Diagnostic Tests for SARS-CoV-2. *JAMA: The Journal of the American Medical Association*, *323*(22), 2249–2251.

48. Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C. V., Boyd, O., Loman, N. J., McCrone, J. T., Gonçalves, S., … The COVID-19 Genomics UK (COG-UK) consortium. (2021). Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. In *bioRxiv*. medRxiv. doi.org/10.1101/2020.12.30.20249034.

Tables

Table 2. Non-limiting set of designed spike-ins.

Oligo ID	**Sequence to order (5' to 3')**	**Nonarchaeal genuses with significant homology***
SDSI forward primer	TCTCCTTCTTAGCTTCGTGAGAAC (SEQ ID NO: 391)	n/a
SDSI reverse primer	CTTGGTCGTCTACTACATGATGTG (SEQ ID NO: 392)	n/a
SDSI 1	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGACCGGACGTTGTGATCACGGG TACCTTGATCTGGTACTCAAAGGTTTGCCCCCGTGAAGTCTGGTACATGGCT AGACACGTCACTCCATTCGAGGGACATTCGAAGTTAGAGAAGGGCAGAGC GATACATCAGATATATCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 393)	none
SDSI 2	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTTAATGGAAAGTATGCTTTAGA TACCTTCTGGAACGCTATCTCACTTGGCGGGAATTCAGATATGGAGAGTAA ATTAAGGGATCTGGAAGTAAAGTTAATGTCGTTAATCTATTTAAATGAGTC	none

	ACCATTAAAATCACCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 394)	
SDSI 3	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATAATATGTTAGAGGTAGAATTTCTTTGTGATAGAATATTATTGATGAATGATGGAAGAGAATTAGCATTAGGAAAACCTAAGGAACTGGTAAAGGATACAGAATCTAAGAATCTTGAAGAGGTTTTCCTTAAACTTGTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 395)	none
SDSI 4	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGTCTAGGTTTTAATTCTTCAACTGCTTCAAATACTAGCTTACTGTAGTTATCTGCCCTCATGTTAGGATATATATCTGGAATATAAGGAGGTTGATGAGTTATAAGAAGTGGATGAAATTGTTGTCACACACTCCCCTACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 396)	none
SDSI 5	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGTAAGCGTTTCCTACCCTCGAGAGGGCCATCCTGGTGGTGAGGAAGTCGTCGAAGTGGGCTAAGTAAAAAGCGAAGATCTCGACCCACAATTACCTCCTCCTGTACACCAGGAATACCCCTATCAGGATAGAGATACCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 397)	none
SDSI 6	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCACGGTCCGCGACGTGAATCGGGCGTTCCAGTCGGCGTTCGGCTACGACGCCGACGACGTGGTCGGAAGCGACCTCCTCGGGCGAATCGTGCCCCCGGTGCCGGACCCGGACCCGGTGCCGGAACCGGGGGACGACGAGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 398)	none
SDSI 7	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGCGTCCGCGAGTTCATCCTGAACGTCGTCCCGCTGTCGCCCGGCGAGGAGCGCGGGGCGGGCTACGCCATCTACACCGACATCACGGAGCGGAAGACCCGCGAAAGCGAGCTAGAGCGACAGAACGAGCGATTGGAGGAGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 399)	none
SDSI 8	ACAGTTCTCCTTCTTAGCTTCGTGAGAACACGAACTCGTCGGTGAACATCTCGTCTTCCGGGGAGCCCGCCGCTCATGGCCTGCCCCCGCCGTAAGCTGCTGCATAAACCCGCTCCAAAATATACGGATCATTCACCCCTTGGAATCGCTCAATCAGATCAATGTACACCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 400)	none
SDSI 9	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGCGTACATTCCCCCTAAGCGGCTCCCAATATACAGACGCCGGTTAACGACAGCTGGCGACCCTGTGATCTCAGTACCGGTGTCGAATGACCACATCAGCTTGCCTGTCCGTGCATGGAGTTCGTATACGTACCCGTCGTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 401)	none
SDSI 10	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATACACCACCCCATCAGCAACAACTGAATCATGATTAAGTATCGCACCAGCATCGTAGCGCCAGCGTTCACTGCCAGTGGTGCTATCGAATGCATAGAAGATATGCTCCTAATCGCCAATATCAGTACTTCACAAAGCCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 402)	none
SDSI 11	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTGGAGTCTTTTGTCACACCGCAGAGGCGTAGCGCTGCAGAGCAGGAGCCCAAGCCTACTGCCAACATAGAGAACATAGTGGCTACAGTATCCCTCGACCAGACTCTAGACCTGAACCTCATAGAGAGGAGCATACTGACCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 403)	none
SDSI 12	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTCGCCTGGGTTAAGAGGATGTTCGGCCTCTCCAAGGCGGGTCACGGAGGCACGCTGGACCCGAAGGTCACCGGCGTCCTCCCCGTAGCCCTGGAGGAAGCAACCAAGGTCATAGGCCTGGTGGTGCACACGAGCAAGGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 404)	none
SDSI 13	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTGGGCGAGATCTACCAGAGGCCGCCGCTCCGCAGCAGTGTTAAGAGAAGCCTCCGCGTCAAGAGGATATA	none

	CGAGATAGAGCTGCTGGAGTACAACGGCAGGTACGCGCTCATGAGGGTGC TCTGCGAGGCCGGCACATCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 405)	
SDSI 14	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGCTGGAAGAACGAGGGCAAGG AGGACCTGCTGCGGAGCTACATCAAGCCCGTCGAGTACGCCGTGAGCCAC CTGCCCAAGATAGTTATACGCGATACCGCGGTGGACGCCATAGCCCATGGC GCGAACCTCGCGGTGCCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 406)	none
SDSI 15	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGGAGACCCCAAGGTGACCGGC GTCCTACCAGTGGGGCTCGCCAACAGCACCAAGGTCATTGGTAATGTTATA CATAGTGTTAAAGAATACGTGATGGTTATACAGCTCCACGGCGATGTAGCC GAGCAGGATTTAAGAACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 407)	none
SDSI 16	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAGAGGGAAAGACTGTAGCTTT CATTCCTAGGCACGGAAAGAGACACAGAATACCTCCACATAAGATAAATT ATAGAGCTAATATATGGGCATTAAAAGAACTAGGAGTGAAATGGGTCATC TCAGTTTCTGCCGTAGGACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 408)	none
SDSI 17	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGAGGGAGCTCAGGAGGACTCG CACGGGGCCCTACAGGGAGGATGAGACACTTGTAAGGCTCCAGGACGTCA GCGAGGCCCTGCTCCTGTGGAGGAGCAACGGGGATGAGAGGTATCTTAGA CGCATCGTGCTACCCGTTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 409)	none
SDSI 18	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAACATCTATCGCCCACCTCCC GAAGATAATGATCTTGGATACAGCTGTCGACGCCATAGCACATGGTGCCAA CCTGGCTGCCCCAGGCGTCGCCAGGTTAACCAGGAACATCGCGAAGGGTA GTACCGTAGCGATCCTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 410)	none
SDSI 19	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGCTATCCCCGTGTACAGCATG GTGGGGGTGCCGATGCCCGGGTAGAACTTGGTGACGCTCTCCAGCTTCTCG AGGACGGTTTCCTTGGGGAGGCTCGCGGTGTCCACGAGGGTTATCGCGTCC TCGGCGCCGTCGCCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 411)	none
SDSI 20	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAGGACGCGAAGAGCGCGGTG GATGTGGACGCGCCGCCGCACACGTAGCCGTCGAGGTAGCGCGGAACCAT CGGCGACATCAGCCCCACGACGCGACCCGAGGCGTTGCCGAGGATCACGT CGAGCGTCACGCGCGGCACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 412)	none
SDSI 21	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTATGGTGTAGAACGGGTCGTT GCGGAGCCAGCCTGGCGGCACGTACCGGTCGTCCGCTATCGCCAGCGATCT CTCGAAGAGGTCGAGGTAGGCGGACGCGTTGGCGAACGCCCCGTGTATCA CGACGTCTATCCCGCCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 413)	none
SDSI 22	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCCTACGCCGGGTGCGTAGGAGG GCTCGAGTACATCCATGTCTATACTGATGTATGTTTTACCCAGGTCGCCTAG TGCCAGGGGTCCCTTTAACGCTTCCAGGATAGAGTACACGGTGACGTCTCT AGTCTTCTTCAAGAACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 414)	none
SDSI 23	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTACTAGCGTGTCAACGGAGCTC TTCAACGCCTTTACTATTGGATAGGTTATAAGGTGCTCGCCTCCGAGGAAT CCCAGGAGCATGCCGGGATACTCGTCTACAACGCCTTTCACCACGTCACCT ATGATTCTTAAAGAGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 415)	none
SDSI 24	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCATAGGTGACATGGGGTTTCCCA	none

	TTGACTCTATAAAGCCGTATCCTTTAAGCGGAGTGCAATTGGTCTACGCTTT GCTTAACAACAGGTATTTCCTACCGGGTAGAGAGGGCTCGCTCATAGCTTT AGGTAGCGTGACGGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 416)	
SDSI 25	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGTATCTCACCGCTTGTCACCAT AGTATCCCTCAGGTACTCCAGTATTCTTGAGAGAAACGCACCTAAGCCGGA TCTCAGGTTTGAATCCATAAGAACTATGAGTGAAGCGGGATTGAAGCCCCT GCTGTTTCTAAGACCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 417)	none
SDSI 26	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAAGGGAGATAGAGAAACGCAT CAAAATACCCTTGGGGAAACTGCGTGCAGGGGTTCAATATGGAGTAGAGG TCTCAGACATAAAGGAGAAGATAGCTGCTTACGCTAGGAGGAAGGGGCTT AAATACTTCCCATCGGCACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 418)	none
SDSI 27	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTGTGAACCTCGTGCCCGGCTCTA AGTCGTGAGGGCTTGCAACATAGGTGGGGAGGAACCCGAGCAACGGGTAA GAAGACAGGATAAGCGGTATCGCTATGAAGAGGGCTGAGAAAAGGACATA TACTCCTGAGCCCGTCCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 419)	none
SDSI 28	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGAACATGCCTTCCCCGTCTATA TAGACCCAGTAGAGTTTAAAAACTTAACCAGAGACGGCTTGTGAGCCGGAT CTCTCCCCCGCTAGGCCCTGGATTGGGCTCGCTCCTCCTGGGACCCCGGCCT CCACATGCTCGGGACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 420)	none
SDSI 29	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTCGGTTCGGCAATAAGTAATA CCAACGAGGTATTACCATGCGCGTGACCAGCAAAGGCCAAGTGACGATCC CAAAGGAGATACGGGATCATTTGGGGATTGGGCCGGGCTCCGAGGTGGAG TTCGTGCCCACAGACGACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 421)	mesorhizobium; neorhizobium
SDSI 30	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTCGATCATATGGCCGGCACGTT GGACTTGGGAGGCATGACAACGGACGAGTATATGGAGTGGCTGAGGGGTC CACGTGAAGATCTCGACATTGATTGACACAAATGTCCTGATCGATGTTTGG GGTCCTGCCGGACAGGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 422)	mesorhizobium; neorhizobium; rhizobium; neorhizobium; aminobacter; sinorhizobium; shinella;
SDSI 31	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCAGGTGTATTTTACACACCTGGA CAGCCAGCATATGATGCTAGCACTCGGTGTCCCCTTATCACGGTTTCCCGC ATTGTAAAGTTTTCGCGCCTGCTGCGCCCCGTAGGGCCTGGATTCATGTCTC AGAATCCATCTCCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 423)	'uncultured bacteria'
SDSI 32	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCGTAGCCCGCACCTTCCTCTGGT TTAGCACCAGCGGTCCCCACAGAGTACCCATCATCCCGAAGGATATGCTGG CAACAGTGGGCACGGGTCTCGCTCGTTGCCTGACTTAACAGGATGCTTCAC AGTACGAACTGACGACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 424)	'uncultured bacteria'
SDSI 33	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAAACTTACCTTATCAGTGTCAT TAAGCATATTGCTTCCAAGACCCATTGAAGCACTTACATCGTTGATACACA GGTGCCAGGAATAGTATTCCTCAGTCTCACTATAATCCTCGTTGGTGTAGCC TTCAAGAGAGTCAACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 425)	none
SDSI 34	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTTAAGCAATTCTTCGGATGAA AGATGGCGCTCTATAGGAATTTGTTCTGGTCTAGCCATAAGGCATTATTTGT ACTTAATTAGTAATAAATGTTTAGTTAATGACTATAAATCTGCAATTGGAG	none

	TCTCAAATTTTCAACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 426)	
SDSI 35	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAACATGAAGGATGTGTGTAAGA GGAAACGTTATTAACAGACGTAATCAGGAGGATAGTTATGCCCTAAAAAC AGCAGAGTTAAGGTTTAAAAATAAGATAAGAACTCAGTTGAGGTTTATCCA TTAATCCCATTAATCCTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 427)	none
SDSI 36	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTATCCGCTGATATATCCTGGGG ATATAGATCGCTCTGAAATGGTTACATCTATCGGTTTTAAGGACAGTTCCA ACACTATTGGACCTTGCAGCTATGACAGGAATAATCTGTTTATCGAGCACA GTTGAATTTGACCTACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 428)	none
SDSI 37	ACAGTTCTCCTTCTTAGCTTCGTGAGAACATATTCCGTATTTCTTATCAAAC CGATCGTGAAGATTTGACAAAGGCTTAACTTTAGGGCTCCACTTCTCATTAT TAGCCTTAGAATATAAAGCGTAACCGTAAGCCTGAGGAACGTAAAGCTTA GGAGATTCAATCCCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 429)	none
SDSI 38	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTAAAATTAGCCGAAGGCTTCCC ATTACCGAAAAAGTCGTTTATTAGCTCTTCATCCTTCTTCTCCACGTCCGCC CATTCCTCTCCTTCCCTTGGAATTTTAAGCTCGTCCCAGCTGACTCTTATGG GCAATTCAATATCCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 430)	none
SDSI 39	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCCGGAGGAATCTATCATATTAA ACCTCCTCAAAATCGCCTCCTCTTGATTGCTTAAAGGCTGTGAATTACAAA GCTTATTTAATGCGTCCCAAAGCGTTAAGTAATAATTATTTATATTAAACAC TACTATTTCAGTAGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 431)	none
SDSI 40	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCCTCCTCAATTCAATTGGAC TGAAGGAGGGTACGTTCTGGAAAACAGAGCGTAAAAGAGATATAGAACGT AGTATACACATAGCTGGAAAAAGAACAATCATTAAGACAATAAAGAACTT TATGGAAAAGAGTAGAACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 432)	none
SDSI 41	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCGTGTAAAGGTTGTATAATTCA AGCCTCAGAACATTTCGAACTCCTTACAAAATCGTTTAAACTTTCTAAGGC ATAAATTTACTAGAAATTGTCATTTATGAGAATGTAACTATATAGATGGTA AAATTATTAATCCTCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 433)	none
SDSI 42	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGGCTGAAAAATAGGTTCGATCC GCCTCCTCACTTCTTCTCCTTCTTGCCCTCGGCCTCGGAGGAGGCCTCTATT CCCAGCTTCTTGGCCTCCTCCTCGGTCGTCATGAACAGGCTAGTCCTCTGCC TTCCGCCCATGCTCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 434)	none
SDSI 43	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTTCAGCATAAAAGACGGTTTC ACGGGCCAAAGCCTAAGCGGCGTAACGGTGAAAGAAGGAGATACGGTTTT GGGCACGATTGACGACGGCGGGACGCTGGAGCTCACGAGGGGCACTCACA CCTTGACTTTCGAGAAGCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 435)	none
SDSI 44	ACAGTTCTCCTTCTTAGCTTCGTGAGAACCTGATGTTATAGAAGTCCGCAA GGACGGCTCTGTCATCTCGCCCGAGGGTGGGAAATACTATCTCGGCGACAT AAGCGGCCCGACACAAATTAGCATCAAGTTCAAGGCCGGCGCGGTGGGAA CCCACGGCTTCACTATCCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 436)	none
SDSI 45	ACAGTTCTCCTTCTTAGCTTCGTGAGAACTCTCCCTCAACCTTCGCGGGGAG AACGGCGCGGAGTACTGGACGGGCTACGCGGACGCGCTGGAAGACCTGTT	none

	GAAGAAAATCCAGAGGCGGGAGGTGAGGGCATGAGAAGGTATTGTTACAT CACGTGGGGATGGATCACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 437)	
SDSI 46	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGAGCGCCGGGAGGTGAGGGCAT GAGTGAGGAATTGATGTTTGGTCGTGTCGTGGAGTATGTTCAGCATAGTTT CTACAAGAAACCGTTTCCTCTTGGCAGTGAGCTCAAGAATGCAGTAGAGAA GGTTATGGAAACAGGACACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 438)	none
SDSI 47	ACAGTTCTCCTTCTTAGCTTCGTGAGAACAGGTCAGAGCCCACGTGGCAAC TTTTGAGGTTCTGACAAAAGACTATGTTCGTGAGAAATACAAAGACATCAT AGAGTTCATGAGGGAGAAAGGGACAGTATCGAGAAAGGAACTGCGGAAG AAGTTCTTCTTGCTTGCTCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 439)	none
SDSI 48	ACAGTTCTCCTTCTTAGCTTCGTGAGAACGTACCTCAAAATACAGAATCAT ATTTTACAATCGCTTGGAAATATTAATATCAACAATACGCAAGTCCAAATT AACGTCCCTGGCAAACAGGTGACAATTTATACCCACGAAATACTAGATAAC GCCAAAAAGGCACTCGCACATCATGTAGTAGACGACCAAGACAGT (SEQ ID NO: 440)	none

Table 3. SDSI and Viral Read Percentages

Mock CT	Viral reads %	Spike-in reads %
20	99.56	0.18
25	99.19	0.38
30	98.47	1.11
35	99.65	3.17

Table 4. Exemplary ARTIC v3 Primers and Primers Spiked in at 2X

Name	Pool	Sequence	Length	%GC	Spiked in at 2X
nCoV-2019_1_LEFT	nCoV-2019_1	ACCAACCAACTTTCGATCTCTTGT (SEQ ID NO: 441)	24	41.7	
nCoV-2019_1_RIGHT	nCoV-2019_1	CATCTTTAAGATGTTGACGTGCCTC (SEQ ID NO: 442)	25	44.0	
nCoV-2019_2_LEFT	nCoV-2019_2	CTGTTTTACAGGTTCGCGACGT (SEQ ID NO: 443)	22	50.0	
nCoV-2019_2_RIGHT	nCoV-2019_2	TAAGGATCAGTGCCAAGCTCGT (SEQ ID NO: 444)	22	50.0	
nCoV-2019_3_LEFT	nCoV-2019_1	CGGTAATAAAGGAGCTGGTGGC (SEQ ID NO: 445)	22	54.6	
nCoV-2019_3_RIGHT	nCoV-2019_1	AAGGTGTCTGCAATTCATAGCTCT (SEQ ID NO: 446)	24	41.7	
nCoV-2019_4_LEFT	nCoV-2019_2	GGTGTATACTGCTGCCGTGAAC (SEQ ID NO: 447)	22	54.6	

nCoV-2019_4_RIGHT	nCoV-2019_2	CACAAGTAGTGGCACCTTCTTTAGT (SEQ ID NO: 448)	25	44.0	
nCoV-2019_5_LEFT	nCoV-2019_1	TGGTGAAACTTCATGGCAGACG (SEQ ID NO: 449)	22	50.0	
nCoV-2019_5_RIGHT	nCoV-2019_1	ATTGATGTTGACTTTCTCTTTTTGGAGT (SEQ ID NO: 450)	28	32.1	
nCoV-2019_6_LEFT	nCoV-2019_2	GGTGTTGTTGGAGAAGGTTCCG (SEQ ID NO: 451)	22	54.6	
nCoV-2019_6_RIGHT	nCoV-2019_2	TAGCGGCCTTCTGTAAAACACG (SEQ ID NO: 452)	22	50.0	
nCoV-2019_7_LEFT_alt0	nCoV-2019_1	CATTTGCATCAGAGGCTGCTCG (SEQ ID NO: 453)	22	54.6	X
nCoV-2019_7_RIGHT_alt5	nCoV-2019_1	AGGTGACAATTTGTCCACCGAC (SEQ ID NO: 454)	22	50.0	X
nCoV-2019_8_LEFT	nCoV-2019_2	AGAGTTTCTTAGAGACGGTTGGGA (SEQ ID NO: 455)	24	45.8	
nCoV-2019_8_RIGHT	nCoV-2019_2	GCTTCAACAGCTTCACTAGTAGGT (SEQ ID NO: 456)	24	45.8	
nCoV-2019_9_LEFT_alt4	nCoV-2019_1	TTCCCACAGAAGTGTTAACAGAGG (SEQ ID NO: 457)	24	45.8	X
nCoV-2019_9_RIGHT_alt2	nCoV-2019_1	GACAGCATCTGCCACAACACAG (SEQ ID NO: 458)	22	54.6	X
nCoV-2019_10_LEFT	nCoV-2019_2	TGAGAAGTGCTCTGCCTATACAGT (SEQ ID NO: 459)	24	45.8	
nCoV-2019_10_RIGHT	nCoV-2019_2	TCATCTAACCAATCTTCTTCTTGCTCT (SEQ ID NO: 460	27	37.0	
nCoV-2019_11_LEFT	nCoV-2019_1	GGAATTTGGTGCCACTTCTGCT (SEQ ID NO: 461)	22	50.0	
nCoV-2019_11_RIGHT	nCoV-2019_1	TCATCAGATTCAACTTGCATGGCA (SEQ ID NO: 462)	24	41.7	
nCoV-2019_12_LEFT	nCoV-2019_2	AAACATGGAGGAGGTGTTGCAG (SEQ ID NO: 463)	22	50.0	X
nCoV-2019_12_RIGHT	nCoV-2019_2	TTCACTCTTCATTTCCAAAAAGCTTGA (SEQ ID NO: 464)	27	33.3	X
nCoV-2019_13_LEFT	nCoV-2019_1	TCGCACAAATGTCTACTTAGCTGT (SEQ ID NO: 465)	24	41.7	
nCoV-2019_13_RIGHT	nCoV-2019_1	ACCACAGCAGTTAAAACACCCT (SEQ ID NO: 466)	22	45.5	
nCoV-2019_14_LEFT_alt4	nCoV-2019_2	TGGCAATCTTCATCCAGATTCTGC (SEQ ID NO: 467)	24	45.8	X
nCoV-2019_14_RIGHT_alt2	nCoV-2019_2	TGCGTGTTTCTTCTGCATGTGC (SEQ ID NO: 468)	22	50.0	X
nCoV-2019_15_LEFT_alt1	nCoV-2019_1	AGTGCTTAAAAAGTGTAAAAGTGCCT (SEQ ID NO: 469)	26	34.6	X

nCoV-2019_15_RIGHT_alt3	nCoV-2019_1	ACTGTAGCTGGCACTTTGAGAGA (SEQ ID NO: 470)	23	47.8	X
nCoV-2019_16_LEFT	nCoV-2019_2	AATTTGGAAGAAGCTGCTCGGT (SEQ ID NO: 471)	22	45.5	
nCoV-2019_16_RIGHT	nCoV-2019_2	CACAACTTGCGTGTGGAGGTTA (SEQ ID NO: 472)	22	50.0	
nCoV-2019_17_LEFT	nCoV-2019_1	CTTCTTTCTTTGAGAGAAGTGAGGACT (SEQ ID NO: 473)	27	40.7	X
nCoV-2019_17_RIGHT	nCoV-2019_1	TTTGTTGGAGTGTTAACAATGCAGT (SEQ ID NO: 474)	25	36.0	X
nCoV-2019_18_LEFT_alt2	nCoV-2019_2	ACTTCTATTAAATGGGCAGATAACAACTGT (SEQ ID NO: 475)	30	33.3	X
nCoV-2019_18_RIGHT_alt1	nCoV-2019_2	GCTTGTTTACCACACGTACAAGG (SEQ ID NO: 476)	23	47.8	X
nCoV-2019_19_LEFT	nCoV-2019_1	GCTGTTATGTACATGGGCACACT (SEQ ID NO: 477)	23	47.8	
nCoV-2019_19_RIGHT	nCoV-2019_1	TGTCCAACTTAGGGTCAATTTCTGT (SEQ ID NO: 478)	25	40.0	
nCoV-2019_20_LEFT	nCoV-2019_2	ACAAAGAAAACAGTTACACAACAACCA (SEQ ID NO: 479)	27	33.3	
nCoV-2019_20_RIGHT	nCoV-2019_2	ACGTGGCTTTATTAGTTGCATTGTT (SEQ ID NO: 480)	25	36.0	
nCoV-2019_21_LEFT_alt2	nCoV-2019_1	GGCTATTGATTATAAACACTACACACCCT (SEQ ID NO: 481	29	37.9	X
nCoV-2019_21_RIGHT_alt0	nCoV-2019_1	GATCTGTGTGGCCAACCTCTTC (SEQ ID NO: 482)	22	54.6	X
nCoV-2019_22_LEFT	nCoV-2019_2	ACTACCGAAGTTGTAGGAGACATTATACT (SEQ ID NO: 483)	29	37.9	
nCoV-2019_22_RIGHT	nCoV-2019_2	ACAGTATTCTTTGCTATAGTAGTCGGC (SEQ ID NO: 484)	27	40.7	
nCoV-2019_23_LEFT	nCoV-2019_1	ACAACTACTAACATAGTTACACGGTGT (SEQ ID NO: 485)	27	37.0	
nCoV-2019_23_RIGHT	nCoV-2019_1	ACCAGTACAGTAGGTTGCAATAGTG (SEQ ID NO: 486)	25	44.0	
nCoV-2019_24_LEFT	nCoV-2019_2	AGGCATGCCTTCTTACTGTACTG (SEQ ID NO: 487)	23	47.8	X
nCoV-2019_24_RIGHT	nCoV-2019_2	ACATTCTAACCATAGCTGAAATCGGG (SEQ ID NO: 488)	26	42.3	X
nCoV-2019_25_LEFT	nCoV-2019_1	GCAATTGTTTTTCAGCTATTTTGCAGT (SEQ ID NO: 489)	27	33.3	
nCoV-2019_25_RIGHT	nCoV-2019_1	ACTGTAGTGACAAGTCTCTCGCA (SEQ ID NO: 490)	23	47.8	
nCoV-2019_26_LEFT	nCoV-2019_2	TTGTGATACATTCTGTGCTGGTAGT (SEQ ID NO: 491)	25	40.0	

nCoV-2019_26_RIGHT	nCoV-2019_2	TCCGCACTATCACCAACATCAG (SEQ ID NO: 492)	22	50.0	
nCoV-2019_27_LEFT	nCoV-2019_1	ACTACAGTCAGCTTATGTGTCAACC (SEQ ID NO: 493)	25	44.0	
nCoV-2019_27_RIGHT	nCoV-2019_1	AATACAAGCACCAAGGTCACGG (SEQ ID NO: 494)	22	50.0	
nCoV-2019_28_LEFT	nCoV-2019_2	ACATAGAAGTTACTGGCGATAGTTGT (SEQ ID NO: 495)	26	38.5	
nCoV-2019_28_RIGHT	nCoV-2019_2	TGTTTAGACATGACATGAACAGGTGT (SEQ ID NO: 496)	26	38.5	
nCoV-2019_29_LEFT	nCoV-2019_1	ACTTGTGTTCCTTTTTGTTGCTGC (SEQ ID NO: 497)	24	41.7	
nCoV-2019_29_RIGHT	nCoV-2019_1	AGTGTACTCTATAAGTTTTGATGGTGTGT (SEQ ID NO: 498)	29	34.5	
nCoV-2019_30_LEFT	nCoV-2019_2	GCACAACTAATGGTGACTTTTTGCA (SEQ ID NO: 499)	25	40.0	
nCoV-2019_30_RIGHT	nCoV-2019_2	ACCACTAGTAGATACACAAACACCAG (SEQ ID NO: 500)	26	42.3	
nCoV-2019_31_LEFT	nCoV-2019_1	TTCTGAGTACTGTAGGCACGGC (SEQ ID NO: 501)	22	54.6	
nCoV-2019_31_RIGHT	nCoV-2019_1	ACAGAATAAACACCAGGTAAGAATGAG T (SEQ ID NO: 502)	28	35.7	
nCoV-2019_32_LEFT	nCoV-2019_2	TGGTGAATACAGTCATGTAGTTGCC (SEQ ID NO: 503)	25	44.0	
nCoV-2019_32_RIGHT	nCoV-2019_2	AGCACATCACTACGCAACTTTAGA (SEQ ID NO: 504)	24	41.7	
nCoV-2019_33_LEFT	nCoV-2019_1	ACTTTTGAAGAAGCTGCGCTGT (SEQ ID NO: 505)	22	45.5	X
nCoV-2019_33_RIGHT	nCoV-2019_1	TGGACAGTAAACTACGTCATCAAGC (SEQ ID NO: 506)	25	44.0	X
nCoV-2019_34_LEFT	nCoV-2019_2	TCCCATCTGGTAAAGTTGAGGGT (SEQ ID NO: 507)	23	47.8	
nCoV-2019_34_RIGHT	nCoV-2019_2	AGTGAAATTGGGCCTCATAGCA (SEQ ID NO: 508)	22	45.5	
nCoV-2019_35_LEFT	nCoV-2019_1	TGTTCGCATTCAACCAGGACAG (SEQ ID NO: 509)	22	50.0	
nCoV-2019_35_RIGHT	nCoV-2019_1	ACTTCATAGCCACAAGGTTAAAGTCA (SEQ ID NO: 510)	26	38.5	
nCoV-2019_36_LEFT	nCoV-2019_2	TTAGCTTGGTTGTACGCTGCTG (SEQ ID NO: 511)	22	50.0	
nCoV-2019_36_RIGHT	nCoV-2019_2	GAACAAAGACCATTGAGTACTCTGGA (SEQ ID NO: 512)	26	42.3	
nCoV-2019_37_LEFT	nCoV-2019_1	ACACACCACTGGTTGTTACTCAC (SEQ ID NO: 513)	23	47.8	

nCoV-2019_37_RIGHT	nCoV-2019_1	GTCCACACTCTCCTAGCACCAT (SEQ ID NO: 514)	22	54.6	
nCoV-2019_38_LEFT	nCoV-2019_2	ACTGTGTTATGTATGCATCAGCTGT (SEQ ID NO: 515)	25	40.0	
nCoV-2019_38_RIGHT	nCoV-2019_2	CACCAAGAGTCAGTCTAAAGTAGCG (SEQ ID NO: 516)	25	48.0	
nCoV-2019_39_LEFT	nCoV-2019_1	AGTATTGCCCTATTTTCTTCATAACTGGT (SEQ ID NO: 517)	29	34.5	
nCoV-2019_39_RIGHT	nCoV-2019_1	TGTAACTGGACACATTGAGCCC (SEQ ID NO: 518)	22	50.0	
nCoV-2019_40_LEFT	nCoV-2019_2	TGCACATCAGTAGTCTTACTCTCAGT (SEQ ID NO: 519)	26	42.3	
nCoV-2019_40_RIGHT	nCoV-2019_2	CATGGCTGCATCACGGTCAAAT (SEQ ID NO: 520)	22	50.0	
nCoV-2019_41_LEFT	nCoV-2019_1	GTTCCCTTCCATCATATGCAGCT (SEQ ID NO: 521)	23	47.8	
nCoV-2019_41_RIGHT	nCoV-2019_1	TGGTATGACAACCATTAGTTTGGCT (SEQ ID NO: 522)	25	40.0	
nCoV-2019_42_LEFT	nCoV-2019_2	TGCAAGAGATGGTTGTGTTCCC (SEQ ID NO: 523)	22	50.0	
nCoV-2019_42_RIGHT	nCoV-2019_2	CCTACCTCCCTTTGTTGTGTTGT (SEQ ID NO: 524)	23	47.8	
nCoV-2019_43_LEFT	nCoV-2019_1	TACGACAGATGTCTTGTGCTGC (SEQ ID NO: 525)	22	50.0	
nCoV-2019_43_RIGHT	nCoV-2019_1	AGCAGCATCTACAGCAAAAGCA (SEQ ID NO: 526)	22	45.5	
nCoV-2019_44_LEFT_alt3	nCoV-2019_2	CCACAGTACGTCTACAAGCTGG (SEQ ID NO: 527)	22	54.6	
nCoV-2019_44_RIGHT_alt0	nCoV-2019_2	CGCAGACGGTACAGACTGTGTT (SEQ ID NO: 528)	22	54.6	
nCoV-2019_45_LEFT_alt2	nCoV-2019_1	AGTATGTACAAATACCTACAACTTGTGCT (SEQ ID NO: 529)	29	34.5	X
nCoV-2019_45_RIGHT_alt7	nCoV-2019_1	TTCATGTTGGTAGTTAGAGAAAGTGTGTC (SEQ ID NO: 530)	29	37.9	X
nCoV-2019_46_LEFT_alt1	nCoV-2019_2	CGCTTCCAAGAAAAGGACGAAGA (SEQ ID NO: 531)	23	47.8	
nCoV-2019_46_RIGHT_alt2	nCoV-2019_2	CACGTTCACCTAAGTTGGCGTAT (SEQ ID NO: 532)	23	47.8	
nCoV-2019_47_LEFT	nCoV-2019_1	AGGACTGGTATGATTTTGTAGAAAACCC (SEQ ID NO: 533)	28	39.3	
nCoV-2019_47_RIGHT	nCoV-2019_1	AATAACGGTCAAAGAGTTTTAACCTCTC (SEQ ID NO: 534)	28	35.7	
nCoV-2019_48_LEFT	nCoV-2019_2	TGTTGACACTGACTTAACAAAGCCT (SEQ ID NO: 535)	25	40.0	

nCoV-2019_48_RIGHT	nCoV-2019_2	TAGATTACCAGAAGCAGCGTGC (SEQ ID NO: 536)	22	50.0	
nCoV-2019_49_LEFT	nCoV-2019_1	AGGAATTACTTGTGTATGCTGCTGA (SEQ ID NO: 537)	25	40.0	
nCoV-2019_49_RIGHT	nCoV-2019_1	TGACGATGACTTGGTTAGCATTAATACA (SEQ ID NO: 538)	28	35.7	
nCoV-2019_50_LEFT	nCoV-2019_2	GTTGATAAGTACTTTGATTGTTACGATG GT (SEQ ID NO: 539)	30	33.3	
nCoV-2019_50_RIGHT	nCoV-2019_2	TAACATGTTGTGCCAACCACCA (SEQ ID NO: 540)	22	45.5	
nCoV-2019_51_LEFT	nCoV-2019_1	TCAATAGCCGCCACTAGAGGAG (SEQ ID NO: 541)	22	54.6	
nCoV-2019_51_RIGHT	nCoV-2019_1	AGTGCATTAACATTGGCCGTGA (SEQ ID NO: 542)	22	45.5	
nCoV-2019_52_LEFT	nCoV-2019_2	CATCAGGAGATGCCACAACTGC (SEQ ID NO: 543)	22	54.6	
nCoV-2019_52_RIGHT	nCoV-2019_2	GTTGAGAGCAAAATTCATGAGGTCC (SEQ ID NO: 544)	25	44.0	
nCoV-2019_53_LEFT	nCoV-2019_1	AGCAAAATGTTGGACTGAGACTGA (SEQ ID NO: 545)	24	41.7	
nCoV-2019_53_RIGHT	nCoV-2019_1	AGCCTCATAAAACTCAGGTTCCC (SEQ ID NO: 546)	23	47.8	
nCoV-2019_54_LEFT	nCoV-2019_2	TGAGTTAACAGGACACATGTTAGACA (SEQ ID NO: 547)	26	38.5	
nCoV-2019_54_RIGHT	nCoV-2019_2	AACCAAAAACTTGTCCATTAGCACA (SEQ ID NO: 548)	25	36.0	
nCoV-2019_55_LEFT	nCoV-2019_1	ACTCAACTTTACTTAGGAGGTATGAGCT (SEQ ID NO: 549)	28	39.3	
nCoV-2019_55_RIGHT	nCoV-2019_1	GGTGTACTCTCCTATTTGTACTTTACTGT (SEQ ID NO: 550)	29	37.9	
nCoV-2019_56_LEFT	nCoV-2019_2	ACCTAGACCACCACTTAACCGA (SEQ ID NO: 551)	22	50.0	
nCoV-2019_56_RIGHT	nCoV-2019_2	ACACTATGCGAGCAGAAGGGTA (SEQ ID NO: 552)	22	50.0	
nCoV-2019_57_LEFT	nCoV-2019_1	ATTCTACACTCCAGGGACCACC (SEQ ID NO: 553)	22	54.6	
nCoV-2019_57_RIGHT	nCoV-2019_1	GTAATTGAGCAGGGTCGCCAAT (SEQ ID NO: 554)	22	50.0	
nCoV-2019_58_LEFT	nCoV-2019_2	TGATTTGAGTGTTGTCAATGCCAGA (SEQ ID NO: 555)	25	40.0	
nCoV-2019_58_RIGHT	nCoV-2019_2	CTTTTCTCCAAGCAGGGTTACGT (SEQ ID NO: 556)	23	47.8	
nCoV-2019_59_LEFT	nCoV-2019_1	TCACGCATGATGTTTCATCTGCA (SEQ ID NO: 557)	23	43.5	

nCoV-2019_59_RIGHT	nCoV-2019_1	AAGAGTCCTGTTACATTTTCAGCTTG (SEQ ID NO: 558)	26	38.5	
nCoV-2019_60_LEFT	nCoV-2019_2	TGATAGAGACCTTTATGACAAGTTGCA (SEQ ID NO: 559)	27	37.0	
nCoV-2019_60_RIGHT	nCoV-2019_2	GGTACCAACAGCTTCTCTAGTAGC (SEQ ID NO: 560)	24	50.0	
nCoV-2019_61_LEFT	nCoV-2019_1	TGTTTATCACCCGCGAAGAAGC (SEQ ID NO: 561)	22	50.0	
nCoV-2019_61_RIGHT	nCoV-2019_1	ATCACATAGACAACAGGTGCGC (SEQ ID NO: 562)	22	50.0	
nCoV-2019_62_LEFT	nCoV-2019_2	GGCACATGGCTTTGAGTTGACA (SEQ ID NO: 563)	22	50.0	
nCoV-2019_62_RIGHT	nCoV-2019_2	GTTGAACCTTTCTACAAGCCGC (SEQ ID NO: 564)	22	50.0	
nCoV-2019_63_LEFT	nCoV-2019_1	TGTTAAGCGTGTTGACTGGACT (SEQ ID NO: 565)	22	45.5	
nCoV-2019_63_RIGHT	nCoV-2019_1	ACAAACTGCCACCATCACAACC (SEQ ID NO: 566)	22	50.0	
nCoV-2019_64_LEFT	nCoV-2019_2	TCGATAGATATCCTGCTAATTCCATTGT (SEQ ID NO: 567)	28	35.7	X
nCoV-2019_64_RIGHT	nCoV-2019_2	AGTCTTGTAAAAGTGTTCCAGAGGT (SEQ ID NO: 568)	25	40.0	X
nCoV-2019_65_LEFT	nCoV-2019_1	GCTGGCTTTAGCTTGTGGGTTT (SEQ ID NO: 569)	22	50.0	
nCoV-2019_65_RIGHT	nCoV-2019_1	TGTCAGTCATAGAACAAACACCAATAGT (SEQ ID NO: 570)	28	35.7	
nCoV-2019_66_LEFT	nCoV-2019_2	GGGTGTGGACATTGCTGCTAAT (SEQ ID NO: 571)	22	50.0	X
nCoV-2019_66_RIGHT	nCoV-2019_2	TCAATTTCCATTTGACTCCTGGGT (SEQ ID NO: 572)	24	41.7	X
nCoV-2019_67_LEFT	nCoV-2019_1	GTTGTCCAACAATTACCTGAAACTTACT (SEQ ID NO: 573)	28	35.7	X
nCoV-2019_67_RIGHT	nCoV-2019_1	CAACCTTAGAAACTACAGATAAATCTTGGG (SEQ ID NO: 574)	30	36.7	X
nCoV-2019_68_LEFT	nCoV-2019_2	ACAGGTTCATCTAAGTGTGTGTGT (SEQ ID NO: 575)	24	41.7	
nCoV-2019_68_RIGHT	nCoV-2019_2	CTCCTTTATCAGAACCAGCACCA (SEQ ID NO: 576)	23	47.8	
nCoV-2019_69_LEFT	nCoV-2019_1	TGTCGCAAAATATACTCAACTGTGTCA (SEQ ID NO: 577)	27	37.0	
nCoV-2019_69_RIGHT	nCoV-2019_1	TCTTTATAGCCACGGAACCTCCA (SEQ ID NO: 578)	23	47.8	
nCoV-2019_70_LEFT	nCoV-2019_2	ACAAAAGAAAATGACTCTAAAGAGGGTTT (SEQ ID NO: 579)	29	31.0	X

nCoV-2019_70_RIGHT	nCoV-2019_2	TGACCTTCTTTTAAAGACATAACAGCAG (SEQ ID NO: 580)	28	35.7	X
nCoV-2019_71_LEFT	nCoV-2019_1	ACAAATCCAATTCAGTTGTCTTCCTATTC (SEQ ID NO: 581)	29	34.5	X
nCoV-2019_71_RIGHT	nCoV-2019_1	TGGAAAAGAAAGGTAAGAACAAGTCCT (SEQ ID NO: 582)	27	37.0	X
nCoV-2019_72_LEFT	nCoV-2019_2	ACACGTGGTGTTTATTACCCTGAC (SEQ ID NO: 583)	24	45.8	
nCoV-2019_72_RIGHT	nCoV-2019_2	ACTCTGAACTCACTTTCCATCCAAC (SEQ ID NO: 584)	25	44.0	
nCoV-2019_73_LEFT	nCoV-2019_1	CAATTTTGTAATGATCCATTTTTGGGTGT (SEQ ID NO: 585)	29	31.0	
nCoV-2019_73_RIGHT	nCoV-2019_1	CACCAGCTGTCCAACCTGAAGA (SEQ ID NO: 586)	22	54.6	
nCoV-2019_74_LEFT	nCoV-2019_2	ACATCACTAGGTTTCAAACTTTACTTGC (SEQ ID NO: 587)	28	35.7	
nCoV-2019_74_RIGHT	nCoV-2019_2	GCAACACAGTTGCTGATTCTCTTC (SEQ ID NO: 588)	24	45.8	
nCoV-2019_75_LEFT	nCoV-2019_1	AGAGTCCAACCAACAGAATCTATTGT (SEQ ID NO: 589)	26	38.5	
nCoV-2019_75_RIGHT	nCoV-2019_1	ACCACCAACCTTAGAATCAAGATTGT (SEQ ID NO: 590)	26	38.5	
nCoV-2019_76_LEFT_alt3	nCoV-2019_2	GGGCAAACTGGAAAGATTGCTGA (SEQ ID NO: 591)	23	47.8	X
nCoV-2019_76_RIGHT_alt0	nCoV-2019_2	ACCTGTGCCTGTTAAACCATTGA (SEQ ID NO: 592)	23	43.5	X
nCoV-2019_77_LEFT	nCoV-2019_1	CCAGCAACTGTTTGTGGACCTA (SEQ ID NO: 593)	22	50.0	
nCoV-2019_77_RIGHT	nCoV-2019_1	CAGCCCCTATTAAACAGCCTGC (SEQ ID NO: 594)	22	54.6	
nCoV-2019_78_LEFT	nCoV-2019_2	CAACTTACTCCTACTTGGCGTGT (SEQ ID NO: 595)	23	47.8	
nCoV-2019_78_RIGHT	nCoV-2019_2	TGTGTACAAAAACTGCCATATTGCA (SEQ ID NO: 596)	25	36.0	
nCoV-2019_79_LEFT	nCoV-2019_1	GTGGTGATTCAACTGAATGCAGC (SEQ ID NO: 597)	23	47.8	X
nCoV-2019_79_RIGHT	nCoV-2019_1	CATTTCATCTGTGAGCAAAGGTGG (SEQ ID NO: 598)	24	45.8	X
nCoV-2019_80_LEFT	nCoV-2019_2	TTGCCTTGGTGATATTGCTGCT (SEQ ID NO: 599)	22	45.5	X
nCoV-2019_80_RIGHT	nCoV-2019_2	TGGAGCTAAGTTGTTTAACAAGCG (SEQ ID NO: 600)	24	41.7	X
nCoV-2019_81_LEFT	nCoV-2019_1	GCACTTGGAAAACTTCAAGATGTGG (SEQ ID NO: 601)	25	44.0	

nCoV-2019_81_RIGHT	nCoV-2019_1	GTGAAGTTCTTTTCTTGTGCAGGG (SEQ ID NO: 602)	24	45.8	
nCoV-2019_82_LEFT	nCoV-2019_2	GGGCTATCATCTTATGTCCTTCCCT (SEQ ID NO: 603)	25	48.0	
nCoV-2019_82_RIGHT	nCoV-2019_2	TGCCAGAGATGTCACCTAAATCAA (SEQ ID NO: 604)	24	41.7	
nCoV-2019_83_LEFT	nCoV-2019_1	TCCTTTGCAACCTGAATTAGACTCA (SEQ ID NO: 605)	25	40.0	
nCoV-2019_83_RIGHT	nCoV-2019_1	TTTGACTCCTTTGAGCACTGGC (SEQ ID NO: 606)	22	50.0	
nCoV-2019_84_LEFT	nCoV-2019_2	TGCTGTAGTTGTCTCAAGGGCT (SEQ ID NO: 607)	22	50.0	
nCoV-2019_84_RIGHT	nCoV-2019_2	AGGTGTGAGTAAACTGTTACAAACAAC (SEQ ID NO: 608)	27	37.0	
nCoV-2019_85_LEFT	nCoV-2019_1	ACTAGCACTCTCCAAGGGTGTT (SEQ ID NO: 609)	22	50.0	
nCoV-2019_85_RIGHT	nCoV-2019_1	ACACAGTCTTTTACTCCAGATTCCC (SEQ ID NO: 610)	25	44.0	
nCoV-2019_86_LEFT	nCoV-2019_2	TCAGGTGATGGCACAACAAGTC (SEQ ID NO: 611)	22	50.0	
nCoV-2019_86_RIGHT	nCoV-2019_2	ACGAAAGCAAGAAAAAGAAGTACGC (SEQ ID NO: 612)	25	40.0	
nCoV-2019_87_LEFT	nCoV-2019_1	CGACTACTAGCGTGCCTTTGTA (SEQ ID NO: 613)	22	50.0	
nCoV-2019_87_RIGHT	nCoV-2019_1	ACTAGGTTCCATTGTTCAAGGAGC (SEQ ID NO: 614)	24	45.8	
nCoV-2019_88_LEFT	nCoV-2019_2	CCATGGCAGATTCCAACGGTAC (SEQ ID NO: 615)	22	54.6	
nCoV-2019_88_RIGHT	nCoV-2019_2	TGGTCAGAATAGTGCCATGGAGT (SEQ ID NO: 616)	23	47.8	
nCoV-2019_89_LEFT_alt2	nCoV-2019_1	CGCGTTCCATGTGGTCATTCAA (SEQ ID NO: 617)	22	50.0	
nCoV-2019_89_RIGHT_alt4	nCoV-2019_1	ACGAGATGAAACATCTGTTGTCACT (SEQ ID NO: 618)	25	40.0	
nCoV-2019_90_LEFT	nCoV-2019_2	ACACAGACCATTCCAGTAGCAGT (SEQ ID NO: 619)	23	47.8	
nCoV-2019_90_RIGHT	nCoV-2019_2	TGAAATGGTGAATTGCCCTCGT (SEQ ID NO: 620)	22	45.5	
nCoV-2019_91_LEFT	nCoV-2019_1	TCACTACCAAGAGTGTGTTAGAGGT (SEQ ID NO: 621)	25	44.0	X
nCoV-2019_91_RIGHT	nCoV-2019_1	TTCAAGTGAGAACCAAAAGATAATAAGCA (SEQ ID NO: 622)	29	31.0	X
nCoV-2019_92_LEFT	nCoV-2019_2	TTTGTGCTTTTTAGCCTTTCTGCT (SEQ ID NO: 623)	24	37.5	

nCoV-2019_92_RIGHT	nCoV-2019_2	AGGTTCCTGGCAATTAATTGTAAAAGG (SEQ ID NO: 624)	27	37.0	
nCoV-2019_93_LEFT	nCoV-2019_1	TGAGGCTGGTTCTAAATCACCCA (SEQ ID NO: 625)	23	47.8	
nCoV-2019_93_RIGHT	nCoV-2019_1	AGGTCTTCCTTGCCATGTTGAG (SEQ ID NO: 626)	22	50.0	
nCoV-2019_94_LEFT	nCoV-2019_2	GGCCCCAAGGTTTACCCAATAA (SEQ ID NO: 627)	22	50.0	
nCoV-2019_94_RIGHT	nCoV-2019_2	TTTGGCAATGTTGTTCCTTGAGG (SEQ ID NO: 628)	23	43.5	
nCoV-2019_95_LEFT	nCoV-2019_1	TGAGGGAGCCTTGAATACACCA (SEQ ID NO: 629)	22	50.0	
nCoV-2019_95_RIGHT	nCoV-2019_1	CAGTACGTTTTTGCCGAGGCTT (SEQ ID NO: 630)	22	50.0	
nCoV-2019_96_LEFT	nCoV-2019_2	GCCAACAACAACAAGGCCAAAC (SEQ ID NO: 631)	22	50.0	
nCoV-2019_96_RIGHT	nCoV-2019_2	TAGGCTCTGTTGGTGGGAATGT (SEQ ID NO: 632)	22	50.0	
nCoV-2019_97_LEFT	nCoV-2019_1	TGGATGACAAAGATCCAAATTTCAAAGA (SEQ ID NO: 633)	28	32.1	
nCoV-2019_97_RIGHT	nCoV-2019_1	ACACACTGATTAAAGATTGCTATGTGAG (SEQ ID NO: 634)	28	35.7	
nCoV-2019_98_LEFT	nCoV-2019_2	AACAATTGCAACAATCCATGAGCA (SEQ ID NO: 635)	24	37.5	
nCoV-2019_98_RIGHT	nCoV-2019_2	TTCTCCTAAGAAGCTATTAAAATCACATGG (SEQ ID NO: 636)	30	33.3	

Table 5. Time and Cost Comparison of FLEX vs XT

Library Prep Kit	Cost Per Sample ($)	Time (hrs)
Illumina DNA Flex	45.96	10
Illumina Nextera XT	64.43	13.5

Table 6. Cost of SDSI+AmpSeq

Processing Step	Reagent	Vendor	Item Number	Cost (dollars)	Number of Reactions	Cost per Reaction
Biosample Extraction	MagMAX™ mirVana™Total RNA Isolation Kit	Thermo Fisher Scientific	A27828	495	96	5.16
	SSIV RT master mix	Thermo Fisher Scientific	18090050	383	50	7.66

cDNA Synthesis	Random hexamers (50ng/ul)	Thermo Fisher Scientific	N808127	91	100	0.91
	dNTPs (10nM)	Thermo Fisher Scientific	18427-013	99	100	0.99
	5x RT buffer	Thermo Fisher Scientific	18090050	x	x	x
	DTT (100mM)	Thermo Fisher Scientific	18090050	x	x	x
	Superase rnase inhibitor	Thermo Fisher Scientific	10777-019	188	125	1.50
ARTIC PCR	Q5 Hot Start High-Fidelity 2X Master Mix	New England BioLabs	M0494L	845	500	1.69
	Artic Primers Pool#1 and Pool#2	IDT		30	500	0.06
Spike-ins	Spike in Primers (Forward/Reverse)	IDT		500	1000000	0.00
	Spike-in targets n=96	IDT		5821	1000000	0.01
Post Artic Pooling Quantification	Qubit™ dsDNA HS Assay Kit	Thermo Fisher Scientific	Q32854	308	500	0.62
Library Construction	Nextera DNA flex Library Prep (n=96)	Illumina	20018705	4153	190	21.86
	Nextera index UD Set A (n=96)	Illumina	20027213	672	384	1.75
Library Quantification	High Sensitivity D1000 ScreenTape	Agilent	5067-5584	362	112	3.23
	High Sensitivity D1000 Sample Buffer	Agilent	5067-5603	59.14	112	0.53
TOTAL:						45.96

Table 7. Final Library Quantification for Enzyme Optimization Experiment

PCR Enzyme	KOD
Q5 2X MM	**Quant (nM)** 85.3
Q5 2X MM + 0.01% S	0.258
	34.5
Q5 Ultra	56.0
	8.1
KAPA	

Table 8. Library Size DNA Flex

Sample	CT Dilution	Standard DNA Flex Library Size (bp)	Standard DNA Flex .5X DNA Flex Library Concentration (nM)	.5X DNA Flex Library Concentration (nM)
MA_MGH_00109	15.39	340	332	92	54.3
MA_MGH_00110	26.39	293	271	13.4	6.84
MA_MGH_00113	31.93	211	207	3.05	1.84

***

[0167] Various modifications and variations of the described methods, pharmaceutical compositions, and kits of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific embodiments, it will be understood that it is capable of further modifications and that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are obvious to those skilled in the art are intended to be within the scope of the invention. This application is intended to cover any variations, uses, or adaptations of the invention following, in general, the principles of the invention and including such departures from the present disclosure come within

known customary practice within the art to which the invention pertains and may be applied to the essential features herein before set forth.

UNITED STATES
PATENT AND TRADEMARK OFFICE

P.O. Box 1450
Alexandria, VA 22313 - 1450
www.uspto.gov

ELECTRONIC ACKNOWLEDGEMENT RECEIPT

APPLICATION #	RECEIPT DATE / TIME	ATTORNEY DOCKET #
17/684,328	**04/04/2025 01:37:55 PM Z ET**	**BROD-5360US**

Title of Invention

AMPLICON-BASED SEQUENCING USING DNA SPIKE-INS

Application Information

APPLICATION TYPE	Utility - Nonprovisional Application under 35 USC 111(a)	PATENT #	-
CONFIRMATION #	5469	FILED BY	Susan Frattura
PATENT CENTER #	69876731	FILING DATE	03/01/2022
CUSTOMER #	198696	FIRST NAMED INVENTOR	Pardis SABETI
CORRESPONDENCE ADDRESS	-	AUTHORIZED BY	CHRISTOPHER SOUTHGATE

Documents

TOTAL DOCUMENTS: 9

DOCUMENT		PAGES	DESCRIPTION	SIZE (KB)
BI_10804_Response_NFOA_BROD_5360US.pdf		15	-	548 KB
BI_10804_Response_NFOA_BROD_5360US-A....pdf	(1-1)	1	Amendment/Request for Reconsideration-After Non-Final Rejection	117 KB
BI_10804_Response_NFOA_BROD_5360US-SPEC.pdf	(2-2)	1	Specification	121 KB
BI_10804_Response_NFOA_BROD_5360US-CLM.pdf	(3-8)	6	Claims	133 KB
BI_10804_Response_NFOA_BROD_5360US-REM.pdf	(9-15)	7	Applicant Arguments/Remarks Made in an Amendment	500 KB

BI_10804_substitute_Specification_Clean copy_BROD_5360US.pdf	110	Specification	1073 KB
BI_10804_Replacement_Drawings_BROD_5360US_Part1.pdf	45	Drawings-only black and white line drawings	25067 KB
BI_10804_substitute_Specification_marked_up_copy_BROD_5360US.pdf	110	Applicant Arguments/Remarks Made in an Amendment	1080 KB
EOT_BROD_5360US.pdf	2	Extension of Time	214 KB
BI_10804_Replacement_Drawings_BROD_5360US_Part2.pdf	14	Drawings-only black and white line drawings	6785 KB

Digest

DOCUMENT	MESSAGE DIGEST(SHA-512)
BI_10804_Response_NFOA_BROD_5360US.pdf	2724A0EFBAF7C11BAE8724F120698845431D088C7259E8328EEB1BBF173D9962778A6206DA5402394943000E6AB2F020C59426D78A2A7E61CB3B90BAAE200A8F
BI_10804_Response_NFOA_BROD_5360US-A....pdf	7F579F3EB5B9C978853FD463EF9C1EE00740231E157987A899191E9D3A44A3E49CC978D2C638CE7690367C002E8F0BC024B96EE2268398B4808CD93FEC68517C
BI_10804_Response_NFOA_BROD_5360US-SPEC.pdf	D8FCB6D12C55B9E0536EA496A297EA4F0A97559E6089484914181F6FBDA5EE72F9AC95C7F7A4FAA3FB6BED6C6F0D0E178BCA47FAEC6DA793741E040241F98504
BI_10804_Response_NFOA_BROD_5360US-CLM.pdf	2DB4CE82905D7B404B02AB2C391039DDD1AD3622DDED47798D649806CBCDA6DDCF933AB3F4F06121E4FED76BF945FFDCFED4000D1FA22330EAA0B27C728F5794
BI_10804_Response_NFOA_BROD_5360US-REM.pdf	92C35567B519A4E583FA328217E1B432000847A7C7EA3293A3632A381E3017E02023CA5AC1B750E277F7392D0BF68D81055EE9D4D565CDECF7C7501D471050F4
BI_10804_substitute_Specification_Clean copy_BROD_5360US.pdf	F80B6DC87C0A2CF5F9D052451B72FFA99E0FB9A28C7910E78AE21C7D3BF637CA31E81A67182FAF1412A4C52673CF91715CC0B1574EE35933D7178ACF04C545E7
BI_10804_Replacement_Drawings_BROD_5360US_Part1.pdf	380873A656806204A969FBAB5BEC5CB35C274601F9A59BFF5421623FD06A555F5043CE32B72032C1D8596C92291E553649FC0FD4B0AAE585971D5DE4F627E8D8
BI_10804_substitute_Specificati	AD84D44E9D96E282004758507B0A86D77E8E719360E7FF8C2

on_marked_up_copy_BROD_5360US.pdf	FAE6F26D08E0189CCFB2F590DB0A8F005707979277AD6DC27D351CDA55064552E87A2FF502BDB79
EOT_BROD_5360US.pdf	132558D373E246A6F5CD6E05260ED85D84A13CF841DB865B054C148FE8022BE1E7E8C87E7B609C08E833031563A8F6F068A1C0576AFD908B2BFC4EDA3CAF13CC
BI_10804_Replacement_Drawings_BROD_5360US_Part2.pdf	B5FC4226BF7808A2E70E1CD0496A01FF06CD7016276167A807B59DF7E46CDC332A84ED87F08124377D3B6DC34C45987FDE0744FA2CEA86FD299397B86796678E

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES
PATENT AND TRADEMARK OFFICE

P.O. Box 1450
Alexandria, VA 22313 - 1450
www.uspto.gov

ELECTRONIC PAYMENT RECEIPT

APPLICATION #	RECEIPT DATE / TIME	ATTORNEY DOCKET #
17/684,328	**04/04/2025 01:37:55 PM Z ET**	**BROD-5360US**

Title of Invention

AMPLICON-BASED SEQUENCING USING DNA SPIKE-INS

Application Information

APPLICATION TYPE	Utility - Nonprovisional Application under 35 USC 111(a)	PATENT #	-
CONFIRMATION #	5469	FILED BY	Susan Frattura
PATENT CENTER #	69876731	AUTHORIZED BY	CHRISTOPHER SOUTHGATE
CUSTOMER #	198696	FILING DATE	03/01/2022
CORRESPONDENCE ADDRESS	-	FIRST NAMED INVENTOR	Pardis SABETI

Payment Information

PAYMENT METHOD	PAYMENT TRANSACTION ID	PAYMENT AUTHORIZED BY
CARD / 2043	**E202544D38168302**	**Susan Frattura**

FEE CODE	DESCRIPTION	ITEM PRICE($)	QUANTITY	ITEM TOTAL($)
2251	EXTENSION FOR RESPONSE WITHIN FIRST MONTH, EXCEPT PROVISIONAL APPLICATIONS	94.00	1	94.00
			TOTAL AMOUNT:	**$94.00**

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C.

371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.