

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

TAKEDA VACCINES, INC.
Petitioner,

v.

VALNEVA AUSTRIA GMBH,
Patent Owner.

Case No. IPR2023-00354
U.S. Patent No. 11,219,681

DECLARATION OF SCOTT BAILEY, Ph.D.

Mail Stop "PATENT BOARD"
Patent Trial and Appeal Board
U.S. Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	MY BACKGROUND AND QUALIFICATIONS	2
III.	OVERVIEW OF ASPECTS OF THE CLAIMS OF THE '681 PATENT AND SUMMARY OF MY ANALYSIS	6
IV.	TECHNICAL BASIS FOR MY OPINIONS	11
A.	Calculating the Total Number of Variant Nucleic Acid Sequences Encompassed by Claims 1–3, 5-15 and 21–25	11
1.	Manually Determining the Number of Variant Nucleic Acid Sequences	12
2.	The Number of Variant Nucleic Acid Sequences Recited in the Claims can be Mathematically Calculated	16
B.	Determining the Minimal Number of Variant Amino Acid Sequences Encompassed by Claims 1–3, 5-15, and 21–25	23
C.	Calculating the Minimal Number of Variant <i>Amino</i> Acid and Variant Nucleic Acid Sequences Recited in Claim 4	29
V.	LIST OF DOCUMENTS CONSIDERED	34
VI.	CONCLUSION.....	37

I, Scott Bailey, hereby declare as follows:

I. INTRODUCTION

1. I am over the age of 18 and competent to make this Declaration.

2. I have been retained as an expert witness on behalf of Takeda Vaccines, Inc. (“Petitioner”) for the above-captioned *inter partes* review (IPR). I am being compensated for my time in connection with this IPR at my standard consulting rate, which is \$750 per hour.

3. I understand that this Declaration accompanies a petition for IPR involving Claims 1–15 and 21–25 of U.S. Patent No. 11,219,681 (“the ’681 patent”) (EX1001), which resulted from U.S. Patent Application No. 16/813,862 (“the ’862 application”), filed on March 10, 2020. I also understand that the ’681 patent is currently assigned to Valneva Austria GmbH (“Valneva”). The ’681 patent states on its face page that it is a continuation of U.S. Patent Application No. 16/063,007, which I understand is a United States patent application based on International Patent Application No. PCT/EP2016/082664, filed on December 23, 2016 (EX1008). The ’681 patent states on its face page that it claims priority to a series of foreign patent applications, the earliest of which was filed on December 23, 2015. I do not offer an opinion as to whether the ’681 patent is entitled to the benefit of the filing date of any of the alleged priority applications.

The analysis and opinions provided herein are applicable irrespective of the filing date considered.

4. In formulating my opinions, I have relied upon my experience, education, and knowledge in molecular biology, genetics, and bioinformatics. EX1122, 1018 (“Bioinformatics is defined as the application of tools of computation and analysis to the capture and interpretation of biological data. It is an interdisciplinary field, which harnesses computer science, mathematics, physics, and biology[.]”).

II. MY BACKGROUND AND QUALIFICATIONS

5. I am currently an Associate Professor in the Department of Biochemistry and Molecular Biology (BMB) at Johns Hopkins Bloomberg School of Public Health. I also hold a secondary appointment in the Department of Biophysics and Biophysical Chemistry at the School of Medicine. Additionally, I have a joint appointment in the Johns Hopkins Oncology Center at the School of Medicine. Apart from my academic appointments, I am a member of The American Society for Biochemistry and Molecular Biology.

6. Since earning my Ph.D. in Biophysics in 2002 from the University of Sheffield (U.K.) and completing a postdoctoral fellowship at Yale University, I have had nearly two decades of extensive experience in molecular biology and genetic engineering. In conducting such research, I regularly rely on widely

accepted¹ Python computer program-based methods for data processing.²

Fortunately, advances in computational methods and algorithms have made it possible to automate calculations that would be cumbersome and time-consuming

¹ EX1123, 1 (“The ability to design and write computer programs is among the most indispensable skills that a modern researcher can cultivate. Python has become a popular programming language in the biosciences[.]”); EX1124, 1274, 1276 (“New multiprocessing frameworks are required to achieve rapid data analysis[.] . . . Python scripting gives the programmer access to numerous external code libraries, which have proven to be of enormous help in developing code that operates in a variety of experimental contexts.”); EX1125, 1 (“Guido van Rossum is the author of Python, an interpreted, interactive object-oriented programming language. . . . [Users] can write modules . . . [that] let you do all sorts of things: communicate with graphics libraries, data flow libraries, and all sorts of file formats.”).

² For example, my research team used the Python-based software PHENIX to complete iterative model building and refinement calculations. *See* EX1127, 535; *see also* EX1126, 213–214 (“The object-oriented Python scripting language . . . is generally most efficient for the rapid development of new algorithms.”).

to calculate by hand. Indeed, the popularity of Python as a computer programming language has led to its use in many other molecular biology applications³ (EX1126).

7. My expertise also includes a thorough knowledge and understanding of genome editing tools, and I regularly teach classes on “*Concepts of Molecular Biology*,” “*Gene Editing, Therapy and Manipulation*,” “*Proteins and Nucleic Acids*,” “*Macromolecular Structure and Function*,” “*Genes and Gene Therapy*,” “*Concept of Molecular Biology*,” “*Principles of Cell Biology*,” “*Genome Integrity*,” “*Quantitative Biology Lab*,” and “*Genome Integrity*.”

8. Since 2008, I have been a principal investigator directing research on the interactions between proteins and nucleic acids. My research has led to over 30 peer-reviewed publications. Over 20 of these publications relate directly to the molecular and biological mechanisms by which cells interact with and modify nucleic acid sequences. I also have experience using sequence comparison tools to search sequence databases for optimal local alignments to a query sequence; for example, I have extensive experience using the Basic Local Alignment Search

³ See *PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution*, at 214.

Tool (BLAST).EX1128 (“The BLAST algorithm was developed as a way to perform DNA and amino acid sequence similarity searches by an algorithm that is faster than FASTA but considered to be equally as sensitive.”); EX1129. For example, in a prior study exploring nucleotide splice-site selection by CRISPR-Cas systems, I used BLAST to identify natural target sequences necessary for our experiment protocol. EX1130, 13285; EX1176, 3412. This is the same sequence alignment tool mentioned in the ’681 patent at column 128, line 28. EX1001, 128:28.

9. I have served as an editor to the *Proceedings of the National Academy of Sciences* (2016), *Methods of Enzymology* (2019), and *eLife Sciences* (2020). In addition, I frequently conduct peer review for publications related to molecular biology techniques and genome editing, including for *Nature*; *Molecular Cell*; *Biochemistry*; *Cell*; *Genes and Development*; *Methods*; *Journal of Structural Biology*; *Nature Biotechnology*; *Nature Chemical Biology*; *Nature Reviews Microbiology*; *Nature Structural and Molecular Biology*; *Science*; and *Trends in Biochemical Sciences*.

10. In addition to my educational training and professional research experiences highlighted above, I have kept abreast of advances in the field of molecular biology by reading scientific literature, conferring with colleagues in the

field, attending and presenting at scientific conferences, and presenting at invited lectures.

11. I have provided a copy of my *curriculum vitae* as EX1079.

III. OVERVIEW OF ASPECTS OF THE CLAIMS OF THE '681 PATENT AND SUMMARY OF MY ANALYSIS

12. The '681 patent has 26 claims, and Claims 2–15, and 21–25 depend directly or indirectly from independent Claim 1. EX1001, 445:34–446:55, 448:11–26. I do not offer any opinions regarding Claims 16–20 or 26.

13. Independent Claim 1 from the '681 patent is presented below:

1. A Zika virus vaccine comprising a Zika virus having an RNA genome corresponding to the DNA sequence provided by SEQ ID NO: 72 or a variant nucleic acid having at least 80% identity to SEQ ID NO: 72, wherein said Zika virus vaccine is capable of stimulating a neutralizing antibody titer greater than 15 in at least 70% of vaccinated subjects, wherein the neutralizing antibody titer is determined using a microneutralization assay (MN₅₀).

EX1001, 445:34–41.

14. Claim 4 from the '681 patent is presented below:

4. The Zika virus vaccine according to claim 1, wherein the Zika virus comprises an E protein having an amino acid sequence provided by any one of SEQ ID Nos: 14-69, or a variant amino acid sequence that is at least 95% identical to any one of SEQ ID Nos: 14-69 and able to pack a virulent Zika virus.

Id., 445:48–53.

15. Claim 21 from the '681 patent is presented below:

21. The Zika virus vaccine according to claim 1, wherein said variant nucleic acid has at least 90% identity to SEQ ID NO: 72.

Id., 448:11–13.

16. Claim 22 from the '681 patent is presented below:

22. The Zika virus vaccine according to claim 1, wherein said variant nucleic acid has at least 95% identity to SEQ ID NO: 72.

Id., 448:13–15.

17. Claim 23 from the '681 patent is presented below:

23. The Zika virus vaccine according to claim 1, wherein said variant nucleic acid has at least 99% identity to SEQ ID NO: 72.

Id., 448:16–18.

18. Claim 1 and dependent Claims 2–15 and 24–25 refer to SEQ ID NO: 72 or a “variant nucleic acid having at least 80% identity to SEQ ID NO: 72[.]” *Id.*, 445:36–37. Similarly, Claims 21, 22, and 23 refer to SEQ ID NO: 72 or a “variant nucleic acid [having at least 90%, at least 95%, or at least 99%] identity,” respectively, to SEQ ID NO: 72. *Id.*, 448:11–19.

19. The ’681 patent states that SEQ ID NO: 72 has 100% overlap with the published sequence of ZIKV strain H/PF/2013, and encodes the amino acid sequence set forth in SEQ ID NO: 73. *Id.*, 117–124, 129:24–30, 129:46–51.

20. I have been asked to determine the number of “variant nucleic acid[s] having [at least 80%, at least 90%, at least 95%, or at least 99%] identity to SEQ ID NO: 72,” as stated in Claims 1–3, 5–15, and 21–25 of the ’681 patent. *Id.*, 445:34–47, 445:54–446:55, 448:11–22. I have also been asked to determine the minimal number of variant nucleic acid sequences recited in Claim 4 of the ’681 patent.

21. I have also been asked to determine the minimal number of variant amino acid sequences encoded by the variant nucleic acids recited in Claims 1–15 and 21–25.

22. This Declaration describes the methodology I used to make these determinations. As discussed in further detail below, because the number of such variant nucleic acids and variant amino acid sequences is so large, I transcribed mathematical models into a Python-based computer code to carry out the calculations.

23. Using Python-based computer code (“Python Script 1”) (EX1174), as described below in paragraph 42, I have determined that the number of “variant nucleic acid[s] [having at least 80%, at least 90%, at least 95%, or at least 99%] identity to SEQ ID NO: 72,” as stated in Claims 1–3, 5-15, and 21–25 of the ’681 patent (EX1001, 445:34–47, 445:54-446:55, 448:11–22) are:

- **$3.88 \times 10^{3,366}$** variant DNA sequences are at least **80%** identical to SEQ ID NO: 72 (Claims 1–3, 5-15, 24, and 25);
- **$4.45 \times 10^{2,032}$** variant DNA sequences are at least **90%** identical to SEQ ID NO: 72 (Claim 21);
- **$7.91 \times 10^{1,182}$** variant DNA sequences are at least **95%** identical to SEQ ID NO: 72 (Claim 22); and

- **1.57×10^{310}** variant DNA sequences that are at least **99%** identical to SEQ ID NO: 72 (Claim 23).

As I show below in Section IV.B, these variant DNA sequences encode more than 6,032,020 full-length variant amino acid sequences, as determined using the Python Script 2 computer code I wrote. (EX1175).

24. Using a similar Python-based computer code I wrote (“Python Script 3”) (EX1173), as described below in Section IV.C, I have also determined that the minimal number of “variant amino acid sequence[s] that [are] at least 95% identical to any of SEQ ID Nos: 14–69,” as stated in Claim 4, is as follows:

- **2.44×10^{71}** variant amino acid sequences are at least 95% identical to SEQ ID NOs: 14–15 and 69, each of which is 498 amino acids in length;
- **9.74×10^{73}** variant amino acid sequences are at least 95% identical to SEQ ID NOs: 27–29, each of which is 500 amino acids in length; and
- **1.19×10^{74}** variant amino acid sequences are at least 95% identical to SEQ ID NOs: 16–26 and 30–68, each of which is 504 amino acids in length.

In sum, the minimal number of variant amino acid sequences recited in claim 4 is **2.17×10^{74}** (i.e., $2.44 \times 10^{71} + 9.74 \times 10^{73} + 1.19 \times 10^{74} = 2.17 \times 10^{74}$). And

because each variant amino acid sequence is encoded by at least one variant nucleic acid sequence, the minimal number of variant nucleic acid sequences recited in claim 4 also is 2.17×10^{74} .

IV. TECHNICAL BASIS FOR MY OPINIONS

A. Calculating the Total Number of Variant Nucleic Acid Sequences Encompassed by Claims 1–3, 5-15 and 21–25

25. I have been asked to calculate how many variant nucleic acid sequences are encompassed by the '681 patent claims' recitation of "variant nucleic acid[s]" having a specified percent identity to SEQ ID NO: 72, as recited in Claim 1 (and dependent Claims 2, 3, 5-15, 24, and 25) and Claims 21–23. In other words, how many variant nucleic acids have:

Claim 1: "at least 80% identity to SEQ ID NO: 72";

Claim 21: "at least 90% identity to SEQ ID NO: 72";

Claim 22: "at least 95% identity to SEQ ID NO: 72"; or

Claim 23: "at least 99% identity to SEQ ID NO: 72."

EX1001, 445:36–37, 448:11–18.

26. The '681 patent states, "the terms 'identical' or percent 'identity' in the context of two or more nucleic acids or amino acid sequences refer to two or more sequences or subsequences that are the same." EX1001, 127:33–35. I have used this definition of "identity" in performing the calculations discussed herein.

1. Manually Determining the Number of Variant Nucleic Acid Sequences

27. To illustrate how one can determine how many variant nucleic acid sequences are at least 80% identical to a given sequence (i.e., a reference sequence), I first provide various examples that illustrate the principles and methodology used to determine the number of variant sequences.

28. As an example, I have calculated the number of variants that are 80% identical to the reference sequence “ATGCA.” In this example, substituting a single nucleotide in the 5-nucleotide length sequence with a different nucleotide results in a variant that is 80% identical to the reference sequence “ATGCA.”⁴ As shown below, any one of the 5 nucleotides in the sequence ATGCA can be substituted (“N” represents the position of the substituted nucleotide):

<u>N</u> TGCA	A <u>N</u> GCA	AT <u>N</u> CA	ATG <u>N</u> A	ATGC <u>N</u>
---------------	----------------	----------------	----------------	---------------

For a nucleic acid composed of the nucleotides A, T, G, and C, the nucleotide “N” in the variant nucleic acid represents any one of the three alternative nucleotides (e.g., the nucleotide A can be substituted with the nucleotide C, G, or T). As an

⁴ When 1 of 5 nucleotides in the sequence is substituted with a different nucleotide, 4 of the 5 nucleotides (i.e., 80%) remain the same as the reference sequence, resulting in 80% identity to the reference sequence.

example, I have listed below the variants having at least 80% identity to the reference sequence “ATGCA.” Here, I have substituted the nucleotide at each 1 of the 5 positions in the sequence (i.e., 80% identical) with each of the three remaining nucleotides, as shown below for the reference sequence ATGCA:

Position of substituted nucleotide (“N”)	<u>N</u> TGCA	A <u>N</u> GCA	AT <u>N</u> CA	ATG <u>N</u> A	ATGC <u>N</u>
Variant nucleic acids that is 80% identical to the reference sequence	<u>T</u> TGCA	A <u>A</u> GCA	AT <u>A</u> CA	ATG <u>A</u> A	ATGCT <u>T</u>
	<u>G</u> TGCA	A <u>G</u> GCA	AT <u>T</u> CA	ATG <u>T</u> A	ATGCC <u>G</u>
	<u>C</u> TGCA	A <u>C</u> GCA	AT <u>C</u> CA	ATG <u>G</u> A	ATGCC <u>C</u>

Having considered every possible nucleotide substitution at every possible position (“N”) in the sequence, there are **15 variant nucleic acid sequences** (shown above) that are 80% identical to “ATGCA.”

29. The methodology described above can also be used to determine the number of variant sequences that have a different percent identity to the reference sequence. For example, to determine how many variant sequences are 60% identical to “ATGCA,” I have substituted two nucleotides at a time. The other 3 of 5 nucleotides (i.e., 60%) remain identical to the reference sequence. In this example, the potential positions of the substitutions are again represented by “N”:

<u>N</u> NGCA	<u>N</u> TG <u>N</u> A	A <u>N</u> NCA	A <u>N</u> GC <u>N</u>	AT <u>N</u> C <u>N</u>
<u>N</u> TNCA	<u>N</u> TGC <u>N</u>	A <u>N</u> GNA	AT <u>N</u> NA	ATG <u>N</u> N

Again, for a nucleic acid composed of the nucleotides A, T, G, and C, the nucleotide “N” in the variant sequence represents a substitution with any one of the remaining three nucleotides. The positions of the substituted nucleotides (shaded row) and the variant nucleic acid sequences having 60% identity to the reference sequence “ATGCA” (non-shaded rows) are shown below.

<u>N</u> NGCA	<u>N</u> TG <u>N</u> A	A <u>N</u> NCA	A <u>N</u> GC <u>N</u>	AT <u>N</u> C <u>N</u>	<u>N</u> TNCA	<u>N</u> TGC <u>N</u>	AT <u>N</u> NA	AT <u>N</u> C <u>N</u>	ATG <u>N</u> N
ATGTG	TTGCT	ATACG	CTGCC	CCGCA	ACTCA	ATGAG	ATTTA	AGGAA	AGTCA
AGGTA	ATACC	TAGCA	ATTGA	ATATA	ATGGG	CTGCT	ATGAT	ATACT	ACGAA
GTACA	ATCTA	ACGGA	ATCAA	GTGGA	AATCA	ATCGA	ATAGA	CTGGA	CTTCA
CTCCA	GTTCA	TTACA	TTGTA	AAACA	GTGTA	GAGCA	TTGCC	ACGTA	GTGCC
ATGTT	AAGTA	ATTCT	ATCCC	AGGCC	ACGCG	ACACA	TTTCA	ACGCT	ATGAC
AAGGA	ATGGC	ATCCG	ATTAA	AGGCT	GTGAA	ACCCA	ATGGT	TGGCA	ACGCC
AAGCG	GGGCA	AAGCT	AACCA	AAGCC	GCGCA	TTGAA	CTGAA	GTGCT	TTGCG
TTCCA	ATTCT	ATTCC	GTGCG	GTCCA	TTGGA	ATGTC	CTACA	CTGCG	CTGTA
AGACA	AGCCA	AAGAA	ATCCT	AGGCG	ATAAA	CAGCA	CGGCA	AGGGA	TCGCA

Thus, **90 variant sequences** are 60% identical to “ATGCA.”

30. To determine how many variant sequences are *at least* 60% identical to the reference sequence, I also considered the number of variants that have only a single nucleotide substitution, in addition to those having two nucleotide

substitutions. As shown above in paragraph 28, the variant sequences that have only a single nucleotide substitution are 80% identical to the reference sequence and are therefore encompassed by the phrase “at least 60% identical.” To determine the number of variant sequences that are *at least* 60% identical to the reference sequence, I summed the number of variant sequences having a single nucleotide substitution (i.e., 15 variant sequences, as discussed in paragraph 28) to the number of variant sequences having two nucleotide substitutions (i.e., 90 variant sequences, as discussed in paragraph 29): $(15) + (90) = 105$. Thus, there are **105 variant sequences** that are *at least* 60% identical to “ATGCA.”

31. The same approach described above can be used for sequences of any length. For example, to determine how many variant sequences are 90% identical to the 10-nucleotide length sequence “ATGCAGTAGC,” I followed the same steps outlined above, in paragraphs 28–29, substituting one nucleotide at a time:

NTGCA GTAGC	ANGCA GTAGC	ATNCA GTAGC	ATGNA GTAGC	ATGCN GTAGC	ATGCA NTAGC	ATGCA GNAGC	ATGCA GTNGC	ATGCA GTANC	ATGCA GTAGN
TTGCA GTAGC	AAGCA GTAGC	ATACA GTAGC	ATGAA GTAGC	ATGCT GTAGC	ATGCA ATAGC	ATGCA GAAGC	ATGCA GTTGC	ATGCA GTAAC	ATGCA GTAGA
GTGCA GTAGC	AGGCA GTAGC	ATTCA GTAGC	ATGTA GTAGC	ATGCG GTAGC	ATGCA TTAGC	ATGCA GGAGC	ATGCA GGGGC	ATGCA GTATC	ATGCA GTAGT
CTGCA GTAGC	ACGCA GTAGC	ATCCA GTAGC	ATGGA GTAGC	ATGCC GTAGC	ATGCA CTAGC	ATGCA GCAGC	ATGCA GTCGC	ATGCA GTACC	ATGCA GTAGG

Thus, **30 variant sequences** are 90% identical to “ATGCAGTAGC.”

2. The Number of Variant Nucleic Acid Sequences Recited in the Claims can be Mathematically Calculated

32. The examples provided above use reference sequences that are sufficiently short such that the number of variant sequences can reasonably be determined by manually writing out all of the variant sequences. However, as these examples illustrate, the number of variant sequences increases when one uses a longer nucleic acid sequence, when one substitutes more than a single nucleotide, or when one specifies *at least* a given percent identity rather than an exact percent identity. In principle, I can use the process described above to manually write out the sequences of all variants that satisfy a specified percent identity to a reference sequence that is hundreds or thousands of nucleotides in length. However, doing so may take many years—if not lifetimes—depending on the length of the reference sequence and the percent identity required. Moreover, manually writing out the variant sequences is unnecessary for determining the *number* of variant sequences that have at least a specified percent identity (e.g., at least 80% or at least 99% identity) to a reference sequence, because the number of variant sequences can be mathematically calculated.

33. Typically, a person having an understanding of the fundamentals of molecular biology and having 1-2 years of experience in bioinformatics (EX1122) in 2015 (and continuing through today) would have been able to calculate the

number of variant nucleic acid sequences having at least a specified percent identity to a reference sequence by using k -combination math. In other words, such a person could use k -combination math to calculate the number of variant sequences having at least 80% identity (or at least 90%, at least 95%, or at least 99% identity) to SEQ ID NO: 72 as recited in claims 1-3, 5-15, and 21-25 of the '681 patent, and I did so as described herein.

34. k -combination math is a well-known methodology for determining the number of ways to choose k items from n items without repetition and without order. EX1131, 209-10. Here, k -combination math allows one to calculate the number of variant sequences with n -nucleotides and k number of substitutions, and thus it allowed me to calculate the number of “variant nucleic acid[s] having [at least 80%, at least 90%, at least 95%, or at least 99%] identity to SEQ ID NO: 72,” as stated in the '681 patent claims. EX1001, 445:34–47, 445:54-446:55, 448:11–22.

35. The *notation* used to denote the total number of combinations is ${}_nC_k$, which reads as the number of combinations of n elements selected k at a time. EX1131, 210. For example, if there are a total of n nucleotides in a sequence in which one wants to substitute k nucleotides: ${}_nC_k$ = the number of variants of a sequence that are n nucleotides in length with k mutations at a time.

36. The *number of combinations* for selecting k mutation sites at a time for a sequence that is n -nucleotides long is given by the formula:

$${}_nC_k = \frac{n!}{k!(n-k)!}$$

where $n!$, $k!$, and $(n-k)!$ are read and computed as factorials. *Id.* This combination formula determines the number of different sequences of nucleotides for a given number of mutations. Thus, for a sequence 5 nucleotides in length with 2 mutation sites, there are $\frac{5!}{2!(5-2)!}$, or 10 combinations of positions (mutation sites) for 2 substitutions, as shown below using the hand-written method, followed by the algorithm method:

NN_ _ _	N_ _N_	_NN_ _	_N_ _N	_ _N_N
N_N_ _	N_ _ _N	_N_N_	_ _NN_	_ _ _NN

Or

$$\frac{(5)(5-4)(5-3)(5-2)(5-1)}{[(2)(2-1)][(3)(3-2)(3-1)]} = \frac{(5)(1)(2)(3)(4)}{[2][(3)(1)(2)]} = \frac{120}{12} = 10$$

37. Next, I account for all the possible nucleotide substitutions for each combination of mutation sites. For example, for a single nucleotide mutation site, there are three possible nucleotide substitutions: for example, an “A” can be substituted with a “C,” “T,” or “G.” Thus, for k -mutations, there are 3^k possible nucleotide substitutions for each combination of mutation sites. For two nucleotide

mutations, there are 3^2 , or 9 possible nucleotide combinations (i.e., CT, CG, CC, TT, TG, TC, GT, GG, and GC).

38. To summarize, ${}_nC_k = \frac{n!}{k!(n-k)!}$ calculates the number of position combinations and 3^k calculates the number of possible nucleotide substitutions at k amount of mutation sites for each combination. The product of these two equations gives one algorithm that determines the total number of variant sequences (${}_nC_k$) accounting for each possible nucleotide substitution (3^k).

$$[3^k] \left[\frac{n!}{k!(n-k)!} \right] = \frac{3^k * n!}{k!(n-k)!}$$

39. To illustrate the validity of the algorithm model, I return to the example in Section IV.A.1 to calculate many variant sequences are 80% identical to “ATGCA.” Recall that I manually determined there are 15 variant sequences that are 80% identical to “ATGCA.” Likewise, using the algorithm, I also calculated that there are 15 variant sequences that are 80% identical to “ATGCA”:

$$\frac{3^k * n!}{k!(n-k)!} = \frac{3^1 * 5!}{1! (5-1)!} = \frac{[3][(5)(5-4)(5-3)(5-2)(5-1)]}{[1][(4)(4-3)(4-2)(4-1)]} = \frac{(3)(120)}{(1)(24)} = \underline{\underline{15}}$$

As with the manually determined example in Section IV.A.1, the algorithm also works on a sequence of any length. For example, I determined the number of variant sequences that are 90% identical to the 10-nucleotide sequence “ATGCAGTAGC” as follows:

$$\frac{3^{k*} n!}{k!(n-k)!} = \frac{3^1 * 10!}{1! (10-1)!} =$$

$$\frac{[3][(10)(10-9)(10-8)(10-7)(10-6)(10-5)(10-4)(10-3)(10-2)(10-1)]}{[1]*[(9)(9-8)(9-7)(9-6)(9-5)(9-4)(9-3)(9-2)(9-1)]} = \frac{(3)(3,628,800)}{(1)(362,880)} =$$

$$= \underline{\underline{30}}$$

40. In the previous examples, I determined the number of variant sequences that are a certain percent identical to the reference sequence. To determine how many variant sequences are at least a certain percent identical to the reference sequence, I expanded the algorithm to sum from 1 up to the number of mutations (G) that yield the desired percent identity. Returning to the example in Section IV.A.1: To illustrate how many variant sequences are *at least* 60% identical to “ATGCA,” I summed the number of variant sequences that have two substitutions (i.e., are 60% identical) and the number of variant sequences that have one substitution (i.e., are 80% identical), because variant sequences that are 80% identical are also *at least* 60% identical. This function⁵ can be expressed as:

⁵ In mathematics, the symbol E (sigma) is generally used to denote a sum of multiple terms. This symbol is accompanied by an index that varies to encompass all the terms that must be considered in the sum. In other words: $\sum_1^G i = 1 + 2 + 3 + \dots + n$. EX1132, 1-2.

$$\sum_{k=1}^{k=G} \frac{3^k \times n!}{k! \times (n-k)!}$$

Where “G” is the number of mutated nucleotides that yield the given percent (%) identity. I calculated G from the length of the sequence (*n*) and the given percent identity with the following formula:

$$G = \frac{n*(100-\%)}{100}$$

For example, for a sequence of 5 nucleotides (*n*), to determine how many nucleotides in this sequence change (G) with 80% identity, I calculated:

$$G = \frac{5*(100-80)}{100} = 1$$

This makes sense, because mutating 1 of 5 nucleotides in a sequence leaves 4 of 5 nucleotides (i.e., 80%) identical to the original sequence. However, with sequences that are hundreds or thousands of nucleotides in length, manually calculating G is unwieldy. In those cases, as in SEQ ID NO: 72, it is more practical to use the formula for G.

41. Returning to the example in Section IV.A.1, I calculated the number of variant sequences that are at least 60% identical to “ATGCA” as follows. The formula for G shows that 2 of 5 nucleotides can be substituted while maintaining 60% identity to the reference sequence:

$$G = \frac{5*(100-60)}{100} = 2$$

I then summed the number of variants that are at least 60% identical (i.e., have 1 or 2 substitution) to the reference sequence that was 5 nucleotides in length, as described above:

$$\sum_{k=1}^{k=G=\lceil n*(100-\%) \rceil / 100} = \sum_1^2 = \frac{3^1 * 5!}{1!(5-1)!} + \frac{3^2 * 5!}{2!(5-2)!} = 15 + 90 = \underline{\underline{105}}$$

42. Finally, to automate the calculations, I wrote a Python-based computer program (“Python Script 1”) to calculate the number of variants given a specific sequence identity and length using the summation algorithm

$(\sum_{k=1}^{k=G} \frac{3^k \times n!}{k! \times (n-k)!})$ and the formula for finding $G (\frac{n*(100-\%)}{100})$.⁶ I have included a true and correct copy of the program code for “Python Script 1” as EX1174. Before using the program to perform calculations for sequences that are hundreds or thousands of nucleotides long, I validated the program by using it to solve for the same shorter-length sequences that I used in the manually determined examples discussed above. The program output the same results that I calculated manually (EX1171).

43. Using this Python Script 1 program, I have determined:

⁶ For a calculated “G” that is not a whole number, I use the floor of “G,” which is expressed mathematically as $\lfloor G \rfloor$. For example, if $G = 5.9$, then $\lfloor G \rfloor = 5$.

Claim 1⁷: There are $3.88 \times 10^{3,366}$ variant DNA sequences that are at least 80% identical to SEQ ID NO: 72.

Claim 21: There are $4.45 \times 10^{2,032}$ variant DNA sequences that are at least 90% identical to SEQ ID NO: 72.

Claim 22: There are $7.91 \times 10^{1,182}$ variant DNA sequences that are at least 95% identical to SEQ ID NO: 72.

Claim 23: There are 1.57×10^{310} variant DNA sequences that are at least 99% identical to SEQ ID NO: 72.

For comparison, there are an estimated 2×10^{23} stars in the universe, which scientists posit is a “gross underestimation.” EX1133, 4. In other words, even the number of variant nucleic acids that are *at least 99% identical* to SEQ ID NO: 72 is an enormous number.

B. Determining the Minimal Number of Variant Amino Acid Sequences Encompassed by Claims 1–3, 5-15, and 21–25

44. I have also been asked to determine the minimal number of variant amino acid sequences encoded by the variant nucleic acid sequences recited in Claims 1–3, 5-15, and 21–25.

⁷ This calculation also applies to dependent claims 2, 3, 5-15, 24, and 25.

45. To calculate the minimal number of amino acid sequences encoded by the variant nucleic acid sequences recited in Claims 1–3, 5–15, and 21–25, I used “Python Script 2” (EX1175). EX1175 is a true and correct copy of Python-based programming code I wrote to execute the following steps:

- (1) The Python Script 2 calculation begins when I input the open-reading frame (ORF) of the selected DNA sequence, excluding the STOP codon, as well as the number of nucleotides that are to be mutated. EX1102, 3019 (“Open reading frames (ORFs) are sequences that have potential to be translated into proteins. They are delineated by start sites, at which translation is initiated by assembly of a ribosome complex, and stop sites, at which translation is terminated and the ribosome complex disassembles.”).
- (2) The script then determines all possible variant sequences of the input open-reading frame having up to the given number of mutations. It does this by first marking the positions within the input open-reading frame that need to be mutated and then filling in these marked positions with all possible combinations of bases. For example, if the input open reading frame is “ATG” and the number of nucleotides to be mutated is 1, then this part of the script would first mark the positions (i.e., NTG, ANG and ATN) and then proceed by replacing

these mutation-site marks with all possible bases (i.e., TTG, CTG, GTG, AAG, ACG, AGG, ATA, ATT, ATC – where the position of the mutated base is underlined here for clarity).

- (3) The script then translates each read⁸ of the variant DNA sequences to their respective *amino acid* sequences using the universal codon table⁹.
- (4) The script then removes any redundant amino acid sequences to ensure that all of the variant amino acid sequences are unique.
- (5) Finally, the script reads each unique variant amino acid sequence and counts how many of them have a premature STOP codon and how many encode full-length amino acid sequences. EX1169, 3
(“Premature stop codons are those that occur within the normal coding

⁸ In DNA sequencing, a read refers to the DNA sequence from one fragment. EX1136, 1.

⁹ “The genetic information about [an] individual[] is stored in the DNA, which make up the genes. ... Genes use triplet codes (“codons”) to translate the information into proteins[.] ... The Genetic Code Tables summarize how this codon assignment is made[.]” See EX1137, 680.

sequence due to a mutation. This means that the message to create the protein of interest is incomplete; thus only truncated protein is formed.”).

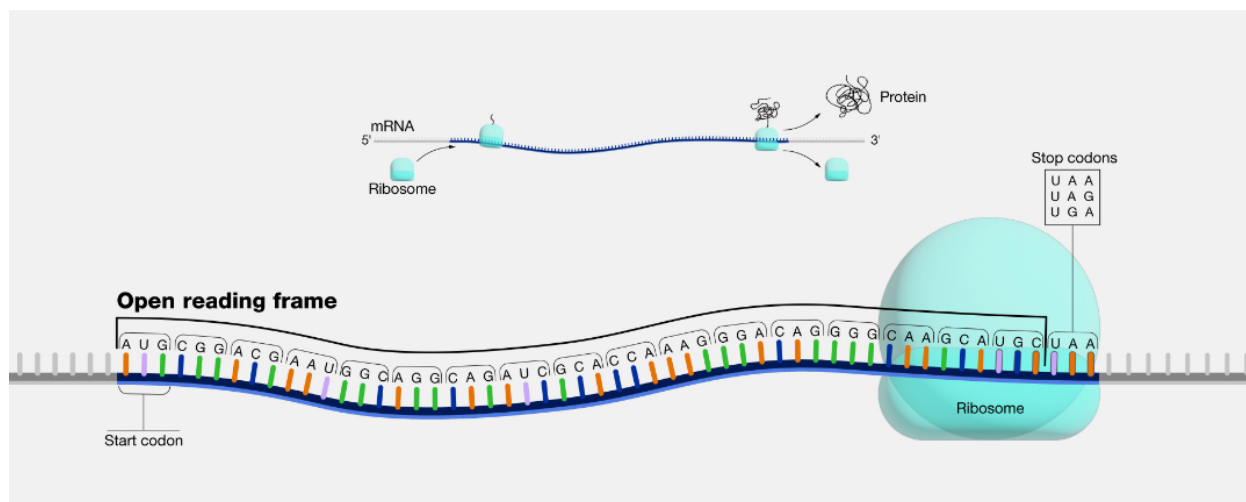


Figure 1. A ribosome “reads” the open-reading frame (ORF) of a nucleic acid sequence to create the encoded amino acid sequence. Translation begins at the START codon and ends upon reaching the STOP codon. Each tri-nucleotide group (e.g., “CGG”) represents one amino acid. EX1135, 2.

46. For example, to illustrate the enormity of the number of variant amino acid sequences encoded by the variant nucleic acid sequences recited in claims 1-3, 5-15, and 21-25, I performed a conservative calculation: I calculated the number of variant amino acid sequences arising from nucleic acid variants having up to 5 nucleotide substitutions in just the first 33 nucleotides of the open-reading frame of

SEQ ID NO: 72.¹⁰ With up to 5 nucleotide substitutions in the 10,773-long sequence of SEQ ID NO: 72, such variant nucleic acid sequences would be at least 99.95% identical to SEQ ID NO: 72. I first determined how many variant nucleic acid sequences have up to 5 nucleotide substitutions in the first 33 nucleotides of the ORF of SEQ ID NO: 72, by performing the following summation calculation:

$$\sum_{k=0}^5 \frac{3^k * n!}{k!(n-k)!} = \frac{3^0 * 5!}{0!(5-0)!} + \frac{3^1 * 5!}{1!(5-1)!} + \frac{3^2 * 5!}{2!(5-2)!} + \frac{3^3 * 5!}{3!(5-3)!} + \frac{3^4 * 5!}{4!(5-4)!} + \frac{3^5 * 5!}{5!(5-5)!} = \underline{\underline{61,139,331}}$$

These variants are a subset of the variants recited in Claims 1–3, 5-15, and 21–25.

47. To determine the number of unique *amino acid* sequences that result from the 61,139,331 variant nucleic acid sequences, I used “Python Script 2”

¹⁰ The open-reading frame of SEQ ID NO: 72 begins at nucleotide 88 with the sequence “atg aaa aac cca.” Using the universal codon table, this nucleic acid sequence encodes the amino acid sequence “Met, Lys, Asn, Pro,” just as shown at the beginning of the Zika virus amino acid (polyprotein) sequence provided as SEQ ID NO: 73 of the ’681 patent. EX1001, 415-431; 129:46-51. Exclusive of the stop codon, the open-reading frame is 10,269 nucleotides in length (3 times the length of the 3423-amino acid long sequence of the polyprotein of SEQ ID NO:73). *Id.*, 415-431. The open-reading frame ends with the sequence “gga gtg ctg taa,” where “taa” represents the stop codon.

(EX1175) to execute the steps as described above, in paragraph 45. Using this program, I determined that there are 8,842,868 variant amino acid sequences. Of those 8,842,868 variant amino acid sequences, there are 6,032,020 variant amino acid sequences *without* premature STOP codons. In other words, translating the 61,139,331 variant nucleic acid sequences arising from a ZIKV open-reading frame having up to 5 nucleotide substitutions in just the first 33 nucleotides yields 6,032,020 full-length variant amino acid sequences.

48. As discussed, these 6,032,020 variants are a subset of the variants recited in the claims, because these calculations assumed that up to 5 nucleotide substitutions are introduced into just the first 33 nucleotides of the open reading frame of SEQ ID NO: 72. These variant nucleic acid sequences are at least 99.95% identical to SEQ ID NO: 72, whereas the claims allow for many more substitutions (i.e., a lower percent identity) throughout the entirety of SEQ ID NO: 72. Accordingly, the total number of full-length variant amino acid sequences encoded by the variant nucleic acids recited in claims 1-3, 5-15, and 21-25 is substantially greater than **6,032,020**.

49. Extensive computing power was required run the Python Script 2 to determine the full-length variant amino acid sequences encoded by nucleic acid variants having up to 5 substitutions in the first 33 nucleotides of SEQ ID NO: 72. Such variants are at least 99.95% identical to SEQ ID NO: 72. As discussed in

Section IV.A.2, Claim 23 recites a higher percent identity than any other claim, and it states that the variant nucleic acid sequences have at least 99% identity to SEQ ID NO: 72. As such, there are 1.57×10^{310} variant nucleic acid sequences that are at least 99% identical to SEQ ID NO: 72 as recited in Claim 23. Running a computer program such as Python Script 2 to determine the exact number of full-length variant amino acid sequences encoded by 1.57×10^{310} variant nucleic acid sequences would require super-computing powers that are not ordinarily available to a bioinformatician. Nor is determining the exact number of full-length variant amino acid sequences necessary to illustrate that the claims include, at a minimum, nucleic acid sequences encoding more than 6 million variant amino acid sequences, as discussed above.

C. Calculating the Minimal Number of Variant *Amino Acid* and Variant Nucleic Acid Sequences Recited in Claim 4

50. Claim 4 depends from Claim 1 and further recites that “the Zika virus comprises an E protein having an amino acid sequence provided by any one of SEQ ID NOs: 14-69, *or a variant amino acid sequence that is at least 95% identical to any one of SEQ ID NOs: 14-69* and able to pack a virulent Zika virus.” EX1001; 445:48-53 (emphasis added). I have been asked to determine the minimal number of variant amino acid sequences that are at least 95% identical to any one of SEQ ID NOs: 14-69. The resulting number also provides a minimal number of

variant nucleic acid sequences having “at least 80% identity to SEQ ID NO:72,” as recited in Claim 1 and encoding a “variant amino acid sequence that is at least 95% identical to any one of SEQ ID NOs: 14-69,” as recited in Claim 4.

51. My approach to this task applies both the logical reasoning and algorithmic math underlying the approach to calculating the number of nucleic acid variants having at least a specified percent identity to a reference sequence, as discussed in Section IV.A, above. To determine the number of variant amino acid sequences that are at least 95% identical to the SEQ ID Nos. recited in Claim 4, I accounted for all the *sequence combinations* (i.e., the order of each amino acid in a sequence) and all the possible *positional combinations* (i.e., the identity of each amino acid within every sequence combination) for SEQ ID Nos. 14-69. To do so, I employed *k*-combination math executed using the “Python Script 3” computer code that I wrote (EX1173). EX1173 is a true and correct copy of the program code for “Python Script 3.”

52. For an amino acid sequence comprised of the 20 naturally occurring amino acids, there are 19 possible amino acid substitutions at each mutation site. EX1134, 3 (“Although there are hundreds of amino acids found in nature, only about 20 amino acids are needed to make all the proteins found in the human body and most other forms of life.”). Thus, for *k* amino-acid mutations, there are 19^k possible substitutions. Recall from Section IV.A.2 that to calculate the number of

ways to choose k -items from n -items without repetition and without order, I used the following equation:

$${}_nC_k = \frac{n!}{k!(n-k)!}$$

EX1131, 210.

53. Thus, k -combination math calculates the positional combinations and 19^k calculates the sequence combinations. The product of these two equations determines the total number of variant amino acid sequences that contain k -mutations:

$$[19^k] \left[\frac{n!}{k!(n-k)!} \right] = \frac{19^k * n!}{k!(n-k)!}$$

54. As above in Section IV.A.2, I expanded this algorithm to determine the total number of variant sequences that are *at least* a set percent identical (X%) by summing from 1 up to the given number of mutations (G) necessary to achieve the set percent identity:

$$\sum_{k=1}^{k=G} \frac{19^k \times n!}{k! \times (n-k)!}$$

Here, “G” is the number of mutated amino acids that yield the given percent (%) identity. I calculated G from the length of the sequence (n) and the given percent identity with the following formula:

$$G = \frac{n * (100 - \%)}{100}$$

55. To illustrate such math, I provide an example for an amino acid sequence that is five amino acids long. In this illustration, I calculate how many variant amino acid sequence are at least 60% identical to the 5-amino-acid long reference sequence. Using the formula shown just above, I determined that up to 2 substitutions (G=2) yields at least 60% identity for an amino acid sequence that is 5 amino acids long (n=5). Using the Python Script 3 program then applied k-combination math to calculate the number of variant amino acid sequences as follows:

$$\sum_1^2 = \frac{19^1 * 5!}{1!(5-1)!} + \frac{19^2 * 5!}{2!(5-2)!} = 95 + 3,610 = \underline{\underline{3,705}}$$

Thus, for a five amino acid long sequence with up to two amino acid substitutions, there are 3,705 variant amino acid sequences.

56. Using these same *k*-combination math principles and Python Script 3 (EX1173), I have determined that the number of variant amino acid sequences that are at least 95% identical to any one of SEQ ID NOs: 14–69 is:

SEQ ID NO	Amino Acid (aa) Length¹¹	Variant Amino Acid Sequences
14–15; 69	498 aa	2.44×10^{71}
27–29	500 aa	9.74×10^{73}
16–26; 30–68	504 aa	1.19×10^{74}
Total for SEQ ID NOs. 14-69		2.17×10^{74}

¹¹ EX1001, 257-405.

As an example, I have included EX1172, an example run of this program for calculating the number of variants that have at least 95% identity to a 500-amino acid sequence, such as SEQ ID NOs. 27-29. As illustrated, there are 9.74×10^{73} variants having a least 95% identity to a 500-amino acid sequence. In the same manner, I used the Python Script 3 to determine the number of variants having at least 95% identity to amino acid sequences that are 498 or 504 amino acids in length, as summarized in the table above. Accordingly, the phrase “variant amino acid sequence that is at least 95% identical to any of one of SEQ ID NOs: 14-69” encompasses 2.17×10^{74} amino-acid sequence variants.

57. Claim 4 depends from Claim 1, which allows for additional variations in the claimed nucleic acid sequence (*viz.*, “or a variant nucleic acid having at least 80% identity to SEQ ID NO: 72”). Such language from claim 1 significantly increases the number of variants beyond the 2.17×10^{74} defined by claim 4. For example, using the Python Script 2, I determined that even a single nucleotide substitution (*i.e.*, 99.99% identity) in the open-reading frame of SEQ ID NO: 72 yields 30,807 variant nucleic acid sequences that encode 20,921 amino acid sequences, 19,816 of which lack premature stop codons. I have included EX1193 that shows the output of Python Script 2 for this example.

58. Thus, given claim 1’s recitation of at least 80% identity to SEQ ID NO: 72 combined with claim 4’s recitation of at least 95% identity to SEQ ID NOs: 14-69, claim 4 recites an enormous number of variants that is at least 2.17×10^{74} amino acid sequences and at least 2.17×10^{74} nucleic acid sequences. Determining the *precise* number of total variant nucleic acid or amino acid sequences recited in claim 4, would require extraordinary supercomputing power that is generally not available to a bioinformatics expert. And doing so is unnecessary to establish that claim 4 recites an extraordinarily large number of variant amino acid and nucleic acid sequences (at least 2.17×10^{74}).

V. LIST OF DOCUMENTS CONSIDERED

<i>Exhibit #</i>	<i>Description</i>
1001	U.S. Patent No. 11,219,681 to Barbero Calzado et al. (“’681 patent”).
1008	Barbero Calzado, J., et al., International Patent Appl. Publ. No. WO 2017/109225 A1, “Zika Virus Vaccine” (filed Dec. 23, 2016; published June 29, 2017)
1018	European Patent Appl. No. 16182845.4 (filed Aug. 4, 2016)
1079	<i>Curriculum Vitae</i> of Scott Bailey, Ph.D.
1102	Singh, U. and Wurtele, E.S., “orfipy: A Fast and Flexible Tool for Extracting ORFs,” <i>Bioinformatics</i> , 37(18): 3019–3020 (2021)
1122	A. Bayat, <i>Science, Medicine, and the Future: Bioinformatics</i> , 324 BMJ 1018, 1018 (2002).
1123	Berk Ekmekci et al., <i>An Introduction to Programming for Bioscientists: A Python-Based Primer</i> ,” 12(6) PLOS COMPUT. BIOL. 1, 1 (2016).

Inter Partes Review of U.S. Patent 11,219,681
Declaration of Scott Bailey, Ph.D.

Exhibit #	Description
1124	N.K. Sauter et al., <i>New Python-Based Methods for Data Processing</i> , 69 ACTA CRYST. 1274, 1274, 1276 (2013).
1125	Bill Venners, <i>The Making of Python: A Conversation with Guido van Rossum, Part I</i> , ARTIMA (Jan. 13, 2003), https://www.artima.com/articles/the-making-of-python .
1126	P.D. Adams et al., <i>PHENIX: A Comprehensive Python-based System for Macromolecular Structure Solution</i> , 66 ACTA CRYST. 213, 213–214 (2010).
1127	R.P. Hayes et al., <i>Structural Basis for Promiscuous PAM Recognition in Type I-E Cascade from E.Coli</i> , 25 NATURE 530, 535 (2016).
1128	<i>BLAST: Basic Local Alignment Search Tool</i> , NIH NAT’L LIB. MED., https://blast.ncbi.nlm.nih.gov/Blast.cgi .
1129	D.W. Mount, <i>Using the Basic Local Alignment Search Tool (BLAST)</i> , CSH PROTOC. (July 1, 2007).
1130	H. Chen, J. Choi & S. Bailey, <i>Cut Site Selection by the Two Nuclease Domains of the Cas9 RNA-guided Endonuclease</i> , 289 J. BIO. CHEM. 13284, 13285 (2014).
1131	PREM S. MANN, <i>INTRODUCTORY STATISTICS</i> 209–10 (7 th ed. 2010).
1132	<i>Summation Notation</i> , COLUMBIA U., http://www.columbia.edu/itc/sipa/math/summation.html .
1133	A. Harvey & E. Howell, <i>How Many Stars Are in the Universe?</i> , SPACE (Feb. 11, 2022), https://www.space.com/26078-how-many-stars-are-there.html .
1134	M.J. Lopez & S.S. Mohiuddin, <i>Biochemistry, Essential Amino Acids</i> , in STAT PEARLS, at 3 (2022).
1135	Oleg Shchelochkov, <i>Open Reading Frame</i> , NIH NAT’L HUMAN GENOME RES. INST., https://www.genome.gov/genetics-glossary/Open-Reading-Frame (last updated Oct. 26, 2022).
1136	<i>Read</i> , GENOMICS EDUC., https://www.genomicseducation.hee.nhs.uk/glossary/read/ (last updated on May 31, 2019).

Inter Partes Review of U.S. Patent 11,219,681
Declaration of Scott Bailey, Ph.D.

<i>Exhibit #</i>	<i>Description</i>
1137	See M. Kurnaz, T. Bilgin & I. Kurnaz, <i>A Statistical Analysis of the Robustness of Alternate Genetic Coding Tables</i> , 9 INT. J. MOL. SCI. 679, 680 (2008).
1169	“Stop Codon,” excerpt from G. Davies et al., <i>Molecular Therapies for Cystic Fibrosis</i> , in <i>Kendig’s Disorders of the Respiratory Tract in Children</i> (9th ed., 2019), available at, https://www.sciencedirect.com/topics/neuroscience/stop-codon#:~:text=Premature%20stop%20codons%20are%20those,as%20truncated)%20protein%20is%20formed
1171	Bailey Declaration Example 1
1172	Bailey Declaration Example 2
1173	Bailey Declaration Python Script 1
1174	Bailey Declaration Python Script 2
1175	Bailey Declaration Python Script 3
1176	A. Ramachandran et al., <i>Processing and Integration of Functionally Oriented Prespacers in the Escherichia coli CRISPR System Depends on Bacterial Host Exonucleases</i> , 295 J. BIO. CHEM. 3403, 3412 (2020).
1193	Bailey Declaration Example 3

VI. CONCLUSION

59. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Respectfully submitted,

A handwritten signature in black ink, appearing to read "S Bailey", is written over a horizontal line.

Scott Bailey, Ph.D.

Date: December 14, 2022