CLAIMS

What is claimed is:

1. A method for identifying amino acid sequences as candidate antigens in the development of a therapy against a pathogen, comprising:

obtaining a plurality of complete or partial genomic sequences of different strains of the pathogen from a data structure;

extracting, by a processor of a computing device, coding sequences from the genomic sequences;

categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

converting, by the processor, the selected coding sequences into corresponding amino acid sequences;

aligning, by the processor, the amino acid sequences;

classifying each of a plurality of portions of the aligned amino acid sequences according to a level of conservation of said portion among the different strains of the pathogen;

selecting portions of the amino acid sequences classified as conserved, comparing the selected conserved sequences to human protein sequences, and further classifying the selected conserved sequences as identical or not identical to a human protein sequence; and categorizing a selected conserved sequence not identical to a human protein sequence as a candidate antigen in the development of a therapy against the pathogen.

2. The method according to claim 1, wherein the data structure comprises contigs, and wherein obtaining the plurality of complete or partial genomic sequences of different strains of the pathogen from the data structure comprises merging, by the processor, overlapping contigs to produce at least a portion of the complete or partial genomic sequences.

3. The method according to claim 1 or claim 2, wherein the categorizing step comprises quantifying the measure of identity and the measure of coverage for each of a plurality of pairs, each of said pairs comprising an extracted coding sequence and a reference sequence.

4. The method according to any one of claims 1 to 3, wherein the categorizing step comprises computing, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and the subject sequence.

5. The method according to claim 4, wherein the computing step comprises creating a matrix of said measures of similarity and rendering a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.

6. The method according to claim 5, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.

7. The method according to any one of claims 1 to 6, wherein the measure of identity comprises number of mutations.

8. The method according to any one of claims 1 to 7, wherein the measure of coverage comprises percent coverage.

9. The method according to any one of claims 1 to 8, wherein the measure of identity comprises calculating E- value.

10. The method according to any one of claims 1 to 9, wherein categorizing the selected conserved sequence as a candidate antigen further comprises determining the presence or

absence of one or more amino acid domains in the selected conserved sequence.

11. The method according to any one of claims 1 to 10, wherein categorizing the selected conserved sequence as a candidate antigen further comprises determining whether the candidate antigen corresponds to a protein that is secreted or is exposed within a membrane and/or cell wall of the pathogen.

12. The method according to any one of claims 1 to 11, wherein categorizing the selected conserved sequence as a candidate antigen further comprises determining the presence of a transmembrane domain in a selected conserved sequence.

13. The method according to any one of claims 1 to 12, wherein the therapy comprises a vaccine and the method further comprises non-clinically evaluating the candidate antigen for immunogenicity.

14. The method according to claim 13, wherein ~~the evaluating step comprises administering~~ a polypeptide comprising the candidate antigen is to be administered to an animal.

15. The method according to any one of claims 1 to 14, wherein the therapy comprises an antibody therapy, and the method further comprises producing an antibody or fragment thereof that specifically binds to an epitope on the candidate antigen.

16. The method according to any one of claims 1 to 15, wherein the pathogen is a virus.

17. The method according to claim 16, wherein the virus is methicillin-resistant Staphylococcus aureus (MRSA), Hepatitis B Virus (HBV), influenza, or Ebola virus.

18. The method according to claim 16, wherein the virus is a coronavirus.

19. The method according to claim 18, wherein the coronavirus is Severe Acute Respiratory Syndrome-associated coronavirus (SARS-CoV), Severe Acute Respiratory Syndrome

coronavirus 2 (SARS-CoV-2), or Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV).

20. The method according to any one of claims 1 to 15, wherein the pathogen is a bacterium.

21. The method according to claim 20, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.

~~22. A method of identifying one or more putative escape mutations after administration of a therapeutic agent to one or more subjects for treatment of a pathogen infection, comprising:~~

~~obtaining a plurality of complete or partial pathogen genomic sequences isolated from one or more subjects after administration of the therapeutic agent to each subject;~~

~~extracting, by a processor of a computing device, coding sequences from the genomic sequences;~~

~~categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;~~

~~selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;~~

~~converting, by the processor, the selected coding sequences into corresponding amino acid sequences;~~

~~aligning, by the processor, the amino acid sequences;~~

~~identifying, in the aligned amino acid sequences, one or more amino acid variants more frequent in the aligned amino acid sequences than in a reference, said one or more amino acid variants being one or more putative escape mutations.~~

~~23. The method according to claim 22, wherein the reference comprises one or more complete or partial pathogen genomic sequences representative of a canonical pathogen~~

sequence, one or more clinical strains of the pathogen, one or more earlier samples of pathogen from one or more of the subjects administered the therapeutic agent, or one or more samples of pathogen from subjects not administered the therapeutic agent.

24. The method according to claim 22 or claim 23, further comprising a step of determining whether one or more of the putative escape mutations decreases binding affinity of the therapeutic agent with a reference polypeptide.

25. The method according to any one of claims 22 to 24, wherein the data structure comprises contigs, and wherein obtaining the plurality of complete or partial genomic sequences of different strains of the pathogen from the data structure comprises merging, by the processor, overlapping contigs to produce at least a portion of the complete or partial genomic sequences.

26. The method according to any one of claims 22 to 25, wherein the categorizing step comprises quantifying the measure of identity and the measure of coverage for each of a plurality of pairs, each of said pairs comprising an extracted coding sequence and a reference sequence.

27. The method according to any one of claims 22 to 26, wherein the categorizing step comprises computing, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and the subject sequence.

28. The method according to claim 27, wherein the computing step comprises creating a matrix of said measures of similarity and rendering a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.

29. The method according to claim 28, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.

30. The method according to any one of claims 22 to 29, wherein the measure of identity

~~comprises number of mutations.~~

~~31. The method according to any one of claims 22 to 30, wherein the measure of coverage comprises percent coverage.~~

~~32. The method according to any one of claims 22 to 31, wherein the measure of identity comprises calculating E-value.~~

~~33. The method according to any one of claims 22 to 32, comprising evaluating one or more of:~~

~~coding sequences of a nucleic acid that encodes a protein associated with the pathogen; conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen;~~

~~non-conserved sequences of a nucleic acid that encodes a protein;~~

~~conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.~~

~~34. The method of any one of claims 22 to 33, wherein each portion of an amino acid sequence comprises one or more amino acid positions.~~

~~35. The method according to any one of claims 22 to 34, wherein the pathogen is a virus.~~

~~36. The method according to claim 35, wherein the virus is Methicillin-resistant *Staphylococcus aureus* (MRSA), Hepatitis B Virus (HBV), influenza, or Ebola virus.~~

~~37. The method according to claim 35, wherein the virus is a coronavirus.~~

~~38. The method according to claim 37, wherein the coronavirus is Severe Acute Respiratory Syndrome-associated coronavirus (SARS-CoV), Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), or Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV).~~

~~39. The method according to claim 38, wherein the coronavirus is SARS-CoV-2.~~

~~40. The method according to any one of claims 22 to 39, comprising evaluating a coronavirus spike (S) protein [e.g ., MERS, SARS-CoV, or SARS-CoV2 spike (S) protein] or a receptor binding domain (RBD) thereof.~~

~~41. The method according to any one of claims 22 to 40, wherein the therapeutic agent comprises an antibody.~~

~~42. The method according to claim 41, wherein the antibody binds SARS-CoV-2.~~

~~43. The method according to claim 42, wherein the antibody binds SARS-CoV-2 spike protein.~~

~~44. The method according to any one of claims 41 to 43, wherein the antibody comprises at least one antibody, heavy chain (HC), light chain (LC), heavy chain variable region (HCVR), light chain variable region (LCVR), heavy chain complementarity determining region (HCDR), or light chain CDR (LCDR) according to Table 3.~~

~~45. The method according to any one of claims 22 to 34, wherein the pathogen is a bacterium.~~

~~46. The method according to claim 45, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.~~

2247. ~~A method of administering~~Use of a therapeutic agent for the manufacture of a medicament for treatment of a pathogen infection in~~to~~ a subject in need thereof, ~~comprising:~~wherein

~~selecting~~ a conserved portion of an amino acid sequence is to be selected by:

obtaining a plurality of complete or partial genomic sequences of different strains of the

pathogen from a data structure;

extracting, by a processor of a computing device, coding sequences from the genomic sequences;

categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

converting, by the processor, the selected coding sequences into corresponding amino acid sequences;

aligning, by the processor, the amino acid sequences;

classifying each of a plurality of portions of the aligned amino acid sequences according to a level of conservation of said portion among the different strains of the pathogen; and

selecting a conserved portion of the aligned amino acid sequences; ~~and administering the therapeutic agent~~ wherein the medicament is to be administered to a subject if a complete or partial pathogen genomic sequence isolated from the subject encodes the conserved portion of an amino acid sequence, wherein the therapeutic agent selectively binds the conserved portion of the amino acid sequence.

~~48. The method according to claim 47, wherein the data structure comprises contigs, and wherein obtaining the plurality of complete or partial genomic sequences of different strains of the pathogen from the data structure comprises merging, by the processor, overlapping contigs to produce at least a portion of the complete or partial genomic sequences.~~

~~49. The method according to claim 47 or claim 48, wherein the categorizing step comprises quantifying the measure of identity and the measure of coverage for each of a plurality of pairs, each of said pairs comprising an extracted coding sequence and a reference sequence.~~

50. The method according to any one of claims 47 to 49, wherein the categorizing step comprises computing, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and the subject sequence.

51. The method according to claim 50, wherein the computing step comprises creating a matrix of said measures of similarity and rendering a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.

52. The method according to claim 51, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.

53. The method according to any one of claims 47 to 52, wherein the measure of identity comprises number of mutations.

54. The method according to any one of claims 47 to 53, wherein the measure of coverage comprises percent coverage.

55. The method according to any one of claims 47 to 54, wherein the measure of identity comprises calculating E-value.

56. The method according to any one of claims 47 to 55, comprising evaluating one or more of:

coding sequences of a nucleic acid that encodes a protein associated with the pathogen;

conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen;

non-conserved sequences of a nucleic acid that encodes a protein;

conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.

57. The method of any one of claims 47 to 56, wherein each portion of an amino acid sequence comprises one or more amino acid positions.

58. The method according to any one of claims 47 to 57, wherein the pathogen is a virus.

59. The method according to claim 58, wherein the virus is Methicillin-resistant *Staphylococcus aureus* (MRSA), Hepatitis B Virus (HBV), influenza, or Ebola virus.

60. The method according to claim 58, wherein the virus is a coronavirus.

61. The method according to claim 60, wherein the coronavirus is Severe Acute Respiratory Syndrome-associated coronavirus (SARS-CoV), Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), or Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV).

62. The method according to claim 61, wherein the coronavirus is SARS-CoV-2.

63. The method according to any one of claims 47 to 62, comprising evaluating a coronavirus spike (S) protein [e.g ., MERS, SARS-CoV, or SARS-CoV2 spike (S) protein] or a receptor binding domain (RBD) thereof.

64. The method according to any one of claims 47 to 63, wherein the therapeutic agent comprises an antibody.

65. The method according to claim 64, wherein the antibody binds SARS-CoV-2.

66. The method according to claim 65, wherein the antibody binds SARS-CoV-2 spike protein.

67. The method according to any one of claims 64 to 66, wherein the antibody comprises at

least one antibody, heavy chain (HC), light chain (LC), heavy chain variable region (HCVR), light chain variable region (LCVR), heavy chain complementarity determining region (HCDR), or light chain CDR (LCDR) according to Table 3.

68. The method according to any one of claims 47 to 57, wherein the pathogen is a bacterium.

69. The method according to claim 68, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.

70. A method for selecting a therapeutic agent for treatment of subjects infected with a pathogen, comprising:

obtaining a plurality of complete or partial genomic sequences of different strains of the pathogen from a data structure;

extracting, by a processor of a computing device, coding sequences from the genomic sequences;

categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

converting, by the processor, the selected coding sequences into corresponding amino acid sequences;

aligning, by the processor, the amino acid sequences;

classifying each of a plurality of portions of the aligned amino acid sequences according to a level of conservation of said portion among the different strains of the pathogen, thereby identifying a conserved portion of a coding sequence representative of the pathogen; and selecting a therapeutic agent that binds the conserved coding sequence as a treatment for subjects infected with the pathogen.

71. The method according to claim 70, wherein the data structure comprises contigs, and wherein obtaining the plurality of complete or partial genomic sequences of different strains of the pathogen from the data structure comprises merging, by the processor, overlapping contigs to produce at least a portion of the complete or partial genomic sequences.

72. The method according to claim 70 or claim 71, wherein the categorizing step comprises quantifying the measure of identity and the measure of coverage for each of a plurality of pairs, each of said pairs comprising an extracted coding sequence and a reference sequence.

73. The method according to any one of claims 70 to 72, wherein the categorizing step comprises computing, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and the subject sequence.

74. The method according to claim 73, wherein the computing step comprises creating a matrix of said measures of similarity and rendering a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.

75. The method according to claim 74, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.

76. The method according to any one of claims 70 to 75, wherein the measure of identity comprises number of mutations.

77. The method according to any one of claims 70 to 76, wherein the measure of coverage comprises percent coverage.

78. The method according to any one of claims 70 to 77, wherein the measure of identity comprises calculating E-value.

79. The method according to any one of claims 70 to 78, comprising evaluating one or more of:

coding sequences of a nucleic acid that encodes a protein associated with the pathogen; conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen;

non-conserved sequences of a nucleic acid that encodes a protein;

conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.

80. The method of any one of claims 70 to 79, wherein each portion of an amino acid sequence comprises one or more amino acid positions.

81. The method according to claim 80, wherein the method further comprises non-clinically evaluating the therapeutic agent as a vaccine or component thereof.

82. The method according to claim 81, wherein the evaluating step comprises administering the therapeutic agent to an animal.

83. The method according to any one of claims 70 to 82, wherein the pathogen is a virus.

84. The method according to claim 83, wherein the virus is Methicillin-resistant *Staphylococcus aureus* (MRSA), Hepatitis B Virus (HBV), influenza, or Ebola virus.

85. The method according to claim 83, wherein the virus is a coronavirus.

86. The method according to claim 85, wherein the coronavirus is Severe Acute Respiratory Syndrome-associated coronavirus (SARS-CoV), Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), or Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV).

87. The method according to claim 86, wherein the coronavirus is SARS-CoV-2.

88. The method according to any one of claims 70 to 87, comprising evaluating a coronavirus spike (S) protein [e.g ., MERS, SARS-CoV, or SARS-CoV2 spike (S) protein] or a receptor binding domain (RBD) thereof.

89. The method according to any one of claims 70 to 88, wherein the therapeutic agent comprises an antibody.

90. The method according to claim 89, wherein the antibody binds SARS-CoV-2.

91. The method according to claim 90, wherein the antibody binds SARS-CoV-2 spike protein.

92. The method according to any one of claims 89 to 91, wherein the antibody comprises at least one antibody, heavy chain (HC), light chain (LC), heavy chain variable region (HCVR), light chain variable region (LCVR), heavy chain complementarity determining region (HCDR), or light chain CDR (LCDR) according to Table 3.

93. The method according to any one of claims 70 to 82, wherein the pathogen is a bacterium.

94. The method according to claim 93, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.

95. A method for assessing conservation of portions of amino acid sequences representative of a pathogen, comprising:

obtaining a plurality of complete or partial genomic sequences of different strains of the pathogen from a data structure;

extracting, by a processor of a computing device, coding sequences from the genomic sequences;

categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

converting, by the processor, the selected coding sequences into corresponding amino acid sequences;

aligning, by the processor, the amino acid sequences; and

identifying a level of conservation of one or more portions of amino acid sequences representative of the pathogen using the aligned amino acid sequences.

96. The method according to claim 95, wherein one or more of the portions is identified as a candidate antigen in the development of a therapy against the pathogen.

97. The method according claim 95 or claim 96, wherein the data structure comprises contigs, and wherein obtaining the plurality of complete or partial genomic sequences of different strains of the pathogen from the data structure comprises merging, by the processor, overlapping contigs to produce at least a portion of the complete or partial genomic sequences.

98. The method according to any one of claims 95 to 97, wherein the categorizing step comprises quantifying the measure of identity and the measure of coverage for each of a plurality of pairs, each of said pairs comprising an extracted coding sequence and a reference sequence.

99. The method according to any one of claims 95 to 98, wherein the categorizing step comprises computing, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query

~~sequence and the subject sequence and a measure of coverage between the query sequence and the subject sequence.~~

~~100. The method according to claim 99, wherein the computing step comprises creating a matrix of said measures of similarity and rendering a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.~~

~~101. The method according to claim 100, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.~~

~~102. The method according to any one of claims 95 to 101, wherein the measure of identity comprises number of mutations.~~

~~103. The method according to any one of claims 95 to 102, wherein the measure of coverage comprises percent coverage.~~

~~104. The method according to any one of claims 95 to 103, wherein the measure of identity comprises calculating E-value.~~

~~105. The method according to any one of claims 95 to 104, comprising evaluating one or more of:~~

~~coding sequences of a nucleic acid that encodes a protein associated with the pathogen; conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen;~~

~~non-conserved sequences of a nucleic acid that encodes a protein;~~

~~conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.~~

~~106. The method of any one of claims 95 to 105, wherein each portion of an amino acid sequence comprises one or more amino acid positions.~~

~~107. The method according to any one of claims 95 to 106, wherein the pathogen is a virus.~~

~~108. The method according to claim 107, wherein the virus is Methicillin-resistant *Staphylococcus aureus* (MRSA), Hepatitis B Virus (HBV), influenza, or Ebola virus.~~

~~109. The method according to claim 107, wherein the virus is a coronavirus.~~

~~110. The method according to claim 109, wherein the coronavirus is Severe Acute Respiratory Syndrome-associated coronavirus (SARS-CoV), Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), or Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV).~~

~~111. The method according to claim 110, wherein the coronavirus is SARS-CoV-2.~~

~~112. The method of any one of claims 95 to 111, wherein the genomic sequences are SARS-CoV-2 genomic sequences and the reference sequence is a SARS-CoV-2 reference sequence.~~

~~113. The method according to any one of claims 95 to 112, comprising evaluating a coronavirus spike (S) protein [e.g ., MERS, SARS-CoV, or SARS-CoV2 spike (S) protein] or a receptor binding domain (RBD) thereof.~~

~~114. The method according to any one of claims 95 to 106, wherein the pathogen is a bacterium.~~

~~115. The method according to claim 114, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.~~

~~116. A method for identifying whether an isolated pathogen is representative of a circulating strain, comprising:~~

~~obtaining a plurality of complete or partial genomic sequences of the circulating strain~~

~~of the pathogen from a data structure;~~

~~identifying one or more conserved portions of said sequences of the circulating strain; obtaining a plurality of complete or partial genomic sequences of the isolated pathogen; and~~

~~identifying whether said isolated pathogen is representative of the circulating strain by comparing at least a portion of said sequences of the isolated pathogen against the identified one or more conserved portions of the sequences of the circulating strain.~~

~~117. The method according to claim 116, wherein identifying one or more conserved portions of said sequences of the circulating strain comprises:~~

~~extracting, by a processor of a computing device, coding sequences from the genomic sequences;~~

~~categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;~~

~~selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;~~

~~converting, by the processor, the selected coding sequences into corresponding amino acid sequences;~~

~~aligning, by the processor, the amino acid sequences; and~~

~~classifying each of a plurality of portions of the aligned amino acid sequences according to a level of conservation of said portion among the aligned amino acid sequences.~~

~~118. The method according to claim 116 or claim 117, wherein the data structure comprises contigs, and wherein obtaining the plurality of complete or partial genomic sequences of different strains of the pathogen from the data structure comprises merging, by the processor, overlapping contigs to produce at least a portion of the complete or partial genomic sequences.~~

~~119. The method according to any one of claims 116 to 118, wherein the categorizing step comprises quantifying the measure of identity and the measure of coverage for each of a plurality~~

of pairs, each of said pairs comprising an extracted coding sequence and a reference sequence.

120. The method according to any one of claims 116 to 119, wherein the categorizing step comprises computing, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and the subject sequence.

121. The method according to claim 120, wherein the computing step comprises creating a matrix of said measures of similarity and rendering a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.

122. The method according to claim 121, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.

123. The method according to any one of claims 116 to 122, wherein the measure of identity comprises number of mutations.

124. The method according to any one of claims 116 to 123, wherein the measure of coverage comprises percent coverage.

125. The method according to any one of claims 116 to 124, wherein the measure of identity comprises calculating E-value.

126. The method according to any one of claims 116 to 125, comprising evaluating one or more of:

coding sequences of a nucleic acid that encodes a protein associated with the pathogen;
conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen;

non-conserved sequences of a nucleic acid that encodes a protein;

~~conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.~~

~~127. The method of any one of claims 116 to 126, wherein each portion of an amino acid sequence comprises one or more amino acid positions.~~

~~128. The method according to any one of claims 116 to 127, wherein the pathogen is a virus.~~

~~129. The method according to claim 128, wherein the virus is Methicillin-resistant *Staphylococcus aureus* (MRSA), Hepatitis B Virus (HBV), influenza, or Ebola virus.~~

~~130. The method according to claim 128, wherein the virus is a coronavirus.~~

~~131. The method according to claim 130, wherein the coronavirus is Severe Acute Respiratory Syndrome-associated coronavirus (SARS-CoV), Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), or Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV).~~

~~132. The method according to claim 131, wherein the coronavirus is SARS-CoV-2.~~

~~133. The method according to any one of claims 116 to 132, comprising evaluating a coronavirus spike (S) protein [e.g ., MERS, SARS-CoV, or SARS-CoV2 spike (S) protein] or a receptor binding domain (RBD) thereof.~~

~~134. The method according to any one of claims 116 to 127, wherein the pathogen is a bacterium.~~

~~135. The method according to claim 134, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.~~

~~136. A method for identifying a mass-to-charge ratio of a peptide representative of a~~

pathogen, comprising:

obtaining a plurality of complete or partial genomic sequences of different strains of the pathogen from a data structure;

extracting, by a processor of a computing device, coding sequences from the genomic sequences;

categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

converting, by the processor, the selected coding sequences into corresponding amino acid sequences; and

determining the mass-to-charge ratio of one or more of the amino acid sequences or portions thereof.

137. The method according to claim 136, wherein the data structure comprises contigs, and wherein obtaining the plurality of complete or partial genomic sequences of different strains of the pathogen from the data structure comprises merging, by the processor, overlapping contigs to produce at least a portion of the complete or partial genomic sequences.

138. The method according to claim 136 or claim 137, wherein the categorizing step comprises quantifying the measure of identity and the measure of coverage for each of a plurality of pairs, each of said pairs comprising an extracted coding sequence and a reference sequence.

139. The method according to any one of claims 136 to 138, wherein the categorizing step comprises computing, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and

the subject sequence.

140. The method according to claim 139, wherein the computing step comprises creating a matrix of said measures of similarity and rendering a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.

141. The method according to claim 140, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.

142. The method according to any one of claims 136 to 141, wherein the measure of identity comprises number of mutations.

143. The method according to any one of claims 136 to 142, wherein the measure of coverage comprises percent coverage.

144. The method according to any one of claims 136 to 143, wherein the measure of identity comprises calculating E-value.

145. The method according to any one of claims 136 to 144, comprising evaluating one or more of:

coding sequences of a nucleic acid that encodes a protein associated with the pathogen; conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen;

non-conserved sequences of a nucleic acid that encodes a protein;

conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.

146. The method of any one of claims 136 to 145, wherein each portion of an amino acid sequence comprises one or more amino acid positions.

147. The method according to any one of claims 136 to 146, wherein the pathogen is a virus.

148. The method according to claim 147, wherein the virus is Methicillin-resistant *Staphylococcus aureus* (MRSA), Hepatitis B Virus (HBV), influenza, or Ebola virus.

149. The method according to claim 147, wherein the virus is a coronavirus.

150. The method according to claim 149, wherein the coronavirus is Severe Acute Respiratory Syndrome-associated coronavirus (SARS-CoV), Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), or Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV).

151. The method according to claim 150, wherein the coronavirus is SARS-CoV-2.

152. The method according to any one of claims 136 to 151, comprising evaluating a coronavirus spike (S) protein [e.g ., MERS, SARS-CoV, or SARS-CoV2 spike (S) protein] or a receptor binding domain (RBD) thereof.

153. The method according to any one of claims 136 to 146, wherein the pathogen is a bacterium.

154. The method according to claim 153, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.

155. A method for identifying an amino acid sequence as a candidate antibiotic resistance marker, comprising:

obtaining a plurality of complete or partial plasmid sequences of a pathogenic bacterium from a data structure;

extracting, by a processor of a computing device, coding sequences from the plasmid sequences;

categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent

identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

converting, by the processor, the selected coding sequences into corresponding amino acid sequences;

aligning, by the processor, the amino acid sequences;

classifying each of a plurality of portions of the aligned amino acid sequences according to a level of conservation of said portion among the plurality of plasmid sequences;

selecting portions of the amino acid sequences classified as conserved; and categorizing a selected conserved sequence as a candidate antibiotic resistance marker.

156. The method according to claim 155, further comprising identifying the candidate antibiotic resistance marker as a candidate according to one or more additional criteria comprising a presence of a transmembrane domain in a selected sequence.

157. The method according to claim 155 or claim 156, wherein the data structure comprises contigs, and wherein obtaining the plurality of complete or partial plasmid sequences from the data structure comprises merging, by the processor, overlapping contigs to produce at least some of the complete or partial plasmid sequences.

158. The method according to any one of claims 155 to 157, wherein the categorizing step comprises quantifying the measure of identity and the measure of coverage for each of a plurality of pairs, each of said pairs comprising an extracted coding sequence and a reference sequence.

159. The method according to any one of claims 155 to 158, wherein the categorizing step comprises computing, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and

the subject sequence.

160. The method according to claim 159, wherein the computing step comprises creating a matrix of said measures of similarity and rendering a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.

161. The method according to claim 160, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.

162. The method according to any one of claims 155 to 161, wherein the measure of identity comprises number of mutations.

163. The method according to any one of claims 155 to 162, wherein the measure of coverage comprises percent coverage.

164. The method according to any one of claims 155 to 163, wherein the measure of identity comprises calculating E-value.

165. The method according to any one of claims 155 to 164, comprising evaluating one or more of:

coding sequences of a nucleic acid that encodes a protein associated with the pathogen; conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen;

non-conserved sequences of a nucleic acid that encodes a protein;

conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.

166. The method of any one of claims 155 to 165, wherein each portion of an amino acid sequence comprises one or more amino acid positions.

167. The method according to any one of claims 155 to 166, wherein the bacterium is a

Staphylococcus species or a Pseudomonas species.

168. A method for identifying one or more conserved portions of coding sequences representative of a plasmid, comprising:

obtaining a plurality of complete or partial plasmid sequences of a pathogenic bacterium from a data structure;

extracting, by a processor of a computing device, coding sequences from the plasmid sequences;

categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

converting, by the processor, the selected coding sequences into corresponding amino acid sequences;

aligning, by the processor, the amino acid sequences; and

classifying each of a plurality of portions of the amino acid sequences according to a level of conservation of said portion among the plurality of plasmid sequences, thereby identifying one or more conserved portions of coding sequences representative of the plasmid.

169. The method according to claim 168, wherein the data structure comprises contigs, and wherein obtaining the plurality of complete or partial plasmid sequences from the data structure comprises merging, by the processor, overlapping contigs to produce at least some of the complete or partial plasmid sequences.

170. The method according to claim 168 or claim 169, wherein the categorizing step comprises quantifying the measure of identity and the measure of coverage for each of a plurality of pairs, each of said pairs comprising an extracted coding sequence and a reference sequence.

~~171. The method according to any one of claims 168 to 170, wherein the categorizing step comprises computing, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and the subject sequence.~~

~~172. The method according to claim 171, wherein the computing step comprises creating a matrix of said measures of similarity and rendering a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.~~

~~173. The method according to claim 172, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.~~

~~174. The method according to any one of claims 168 to 173, wherein the measure of identity comprises number of mutations.~~

~~175. The method according to any one of claims 168 to 174, wherein the measure of coverage comprises percent coverage.~~

~~176. The method according to any one of claims 168 to 175, wherein the measure of identity comprises calculating E-value.~~

~~177. The method according to any one of claims 168 to 176, comprising evaluating one or more of:~~

~~coding sequences of a nucleic acid that encodes a protein associated with the pathogen; conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen;~~

~~non-conserved sequences of a nucleic acid that encodes a protein;~~

~~conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.~~

~~178. The method of any one of claims 168 to 177, wherein each portion of an amino acid sequence comprises one or more amino acid positions.~~

~~179. The method according to any one of claims 168 to 178, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.~~

23~~180~~. A system for automatically identifying one or more conserved portions of coding sequences representative of a pathogen, the system comprising:

a processor; and

a memory having instructions thereon, the instructions, when executed by the processor, causing the processor to:

obtain a plurality of complete or partial genomic sequences of different strains of the pathogen from a data structure;

extract, by the processor, coding sequences from the genomic sequences;

categorize, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

select coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

convert, by the processor, the selected coding sequences into corresponding amino acid sequences;

align, by the processor, the amino acid sequences; and

classify each of a plurality of portions of the aligned amino acid sequences according to a level of conservation of said portion among the different strains of the pathogen, thereby identifying one or more conserved portions of coding sequences representative of the pathogen.

~~181. The system according to claim 180, wherein the instructions, when executed by the processor, cause the processor to compute, for each of a set of query coding sequences against a~~

set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and the subject sequence .

182. The system according to claim 181, wherein the instructions, when executed by the processor, cause the processor to create a matrix of said measures of similarity and render a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.

183. The system according to claim 182, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.

184. The system according to any one of claims 180 to 183, wherein the data structure comprises contigs, and wherein the instructions, when executed by the processor, cause the processor to obtain the plurality of complete or partial genomic sequences of different strains of the pathogen by merging, by the processor, overlapping contigs to produce at least some of the complete or partial genomic sequences.

185. The system according to any one of claims 180 to 184, wherein the instructions, when executed by the processor, cause the processor to evaluate one or more of:

coding sequences of a nucleic acid that encodes a protein associated with the pathogen; conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen;

non-conserved sequences of a nucleic acid that encodes a protein;

conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.

186. The system according to any one of claims 180 to 185, wherein the instructions, when executed by the processor, cause the processor to evaluate a coronavirus spike (S) protein [e.g,

MERS, SARS-CoV, or SARS-CoV2 spike (S) protein] or a receptor-binding domain (RBD) thereof.

187. The system according to any one of claims 180 to 186, wherein the pathogen is a virus.

188. The system according to claim 187, wherein the virus is Methicillin-resistant *Staphylococcus aureus* (MRSA), Hepatitis B Virus (HBV), influenza, or Ebola virus.

189. The system according to claim 187, wherein the virus is a coronavirus.

190. The system according to claim 189, wherein the coronavirus is Severe Acute Respiratory Syndrome-associated coronavirus (SARS-CoV), Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), or Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV).

191. The system according to claim 190, wherein the coronavirus is SARS-CoV-2.

192. The system according to any one of claims 180 to 186, wherein the pathogen is a bacterium.

193. The system according to claim 192, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.

194. A system for automatically identifying one or more conserved portions of coding sequences representative of a plasmid, the system comprising:

a processor; and

a memory having instructions thereon, the instructions, when executed by the processor, causing the processor to:

obtain a plurality of complete or partial plasmid sequences of a pathogenic bacterium from a data structure;

extract, by the processor, coding sequences from the plasmid sequences;

categorize, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

select coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

convert, by the processor, the selected coding sequences into corresponding amino acid sequences;

align, by the processor, the amino acid sequences; and

classify each of a plurality of portions of the amino acid sequences according to a level of conservation of said portion among the plurality of plasmid sequences, thereby identifying one or more conserved portions of coding sequences representative of the plasmid.

195. The system according to claim 194, wherein the instructions, when executed by the processor, cause the processor to compute, for each of a set of query coding sequences against a set of subject sequences, measures of similarity between the query coding sequence and each subject sequence, each of said measures of similarity a function of a measure of identity between the query sequence and the subject sequence and a measure of coverage between the query sequence and the subject sequence.

196. The system according to claim 195, wherein the instructions, when executed by the processor, cause the processor to create a matrix of said measures of similarity and render a graphical representation of said matrix, thereby displaying levels of conservation between the query sequences and subject sequences.

197. The system according to claim 196, wherein the graphical representation comprises one or more of a heatmap, a graph, and a phylogeny.

198. The system according to any one of claims 194 to 197, wherein the data structure

comprises contigs, and wherein the instructions, when executed by the processor, cause the processor to obtain the plurality of complete or partial plasmid sequences of a pathogenic bacterium by merging, by the processor, overlapping contigs to produce at least some of the complete or partial plasmid sequences.

199. The system according to any one of claims 194 to 198, wherein the instructions, when executed by the processor, cause the processor to evaluate one or more of:

coding sequences of a nucleic acid that encodes a protein associated with the pathogen; conserved sequences of a nucleic acid sequence that encodes a protein associated with the pathogen

non-conserved sequences of a nucleic acid that encodes a protein;

conserved domains within a particular protein associated with the pathogen; and non-conserved domains within a particular protein associated with the pathogen.

200. The system according to any one of claims 194 to 199, wherein the instructions, when executed by the processor, cause the processor to evaluate a coronavirus spike (S) protein [e.g, MERS, SARS-CoV, or SARS-CoV2 spike (S) protein] or a receptor-binding domain (RBD) thereof.

201. The system according to any one of claims 194 to 200, wherein the pathogen is a virus.

202. The system according to claim 201, wherein the virus is Methicillin-resistant *Staphylococcus aureus* (MRSA), Hepatitis B Virus (HBV), influenza, or Ebola virus.

203. The system according to claim 201, wherein the virus is a coronavirus.

204. The system according to claim 203, wherein the coronavirus is Severe Acute Respiratory Syndrome-associated coronavirus (SARS-CoV), Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), or Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV).

~~205. The system according to claim 204, wherein the coronavirus is SARS-CoV-2.~~

~~206. The system according to any one of claims 194 to 200, wherein the pathogen is a bacterium.~~

~~207. The system according to claim 206, wherein the bacterium is a Staphylococcus species or a Pseudomonas species.~~

24~~208~~. A therapeutic agent for use in identifying one or more putative escape mutations after administration of the therapeutic agent to one or more subjects for treatment of a pathogen infection, the use comprising:

obtaining a plurality of complete or partial pathogen genomic sequences isolated from one or more subjects after administration of the therapeutic agent to each subject;

extracting, by a processor of a computing device, coding sequences from the genomic sequences;

categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

converting, by the processor, the selected coding sequences into corresponding amino acid sequences;

aligning, by the processor, the amino acid sequences;

identifying, in the aligned amino acid sequences, one or more amino acid variants more frequent in the aligned amino acid sequences than in a reference, said one or more amino acid variants being one or more putative escape mutations.

~~209. A therapeutic agent for use in treatment of a pathogen infection, the use comprising:~~

~~selecting a conserved portion of an amino acid sequence by:~~

~~obtaining a plurality of complete or partial genomic sequences of different strains of the pathogen from a data structure;~~

~~extracting, by a processor of a computing device, coding sequences from the genomic sequences;~~

~~categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;~~

~~selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;~~

~~converting, by the processor, the selected coding sequences into corresponding amino acid sequences;~~

~~aligning, by the processor, the amino acid sequences;~~

~~classifying each of a plurality of portions of the aligned amino acid sequences according to a level of conservation of said portion among the different strains of the pathogen; and~~

~~selecting a conserved portion of the aligned amino acid sequences; and administering the therapeutic agent to a subject if a complete or partial pathogen genomic sequence isolated from the subject encodes the conserved portion of an amino acid sequence, wherein the therapeutic agent selectively binds the conserved portion of the amino acid sequence.~~

25~~210~~. Use of a therapeutic agent for the manufacture of a medicament for identifying one or more putative escape mutations after administration of the medicament to one or more subjects for treatment of a pathogen infection, the use comprising:

obtaining a plurality of complete or partial pathogen genomic sequences isolated from one or more subjects after administration of the medicament to each subject;

extracting, by a processor of a computing device, coding sequences from the genomic sequences;

categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and

percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;

selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;

converting, by the processor, the selected coding sequences into corresponding amino acid sequences;

aligning, by the processor, the amino acid sequences;

identifying, in the aligned amino acid sequences, one or more amino acid variants more frequent in the aligned amino acid sequences than in a reference, said one or more amino acid variants being one or more putative escape mutations.

~~211. Use of a therapeutic agent for the manufacture of a medicament for treatment of a pathogen infection, the use comprising:~~

~~selecting a conserved portion of an amino acid sequence by:~~

~~obtaining a plurality of complete or partial genomic sequences of different strains of the pathogen from a data structure;~~

~~extracting, by a processor of a computing device, coding sequences from the genomic sequences;~~

~~categorizing, by the processor, the coding sequences according to a measure of identity and a measure of coverage, wherein the measure of identity comprises one or more of percent identity, percent identity over a predetermined coverage length, number of mutations, and percent mutation, and wherein the measure of coverage comprises one or more of percent coverage and coverage length;~~

~~selecting coding sequences from among the categorized coding sequences according to the measure of identity and the measure of coverage;~~

~~converting, by the processor, the selected coding sequences into corresponding amino acid sequences;~~

~~aligning, by the processor, the amino acid sequences;~~

~~classifying each of a plurality of portions of the aligned amino acid sequences according to a level of conservation of said portion among the different~~

~~strains of the pathogen; and~~

~~selecting a conserved portion of the aligned amino acid sequences; and~~

~~administering the medicament to a subject if a complete or partial pathogen genomic sequence isolated from the subject encodes the conserved portion of an amino acid sequence, wherein the therapeutic agent selectively binds the conserved portion of the amino acid sequence.~~