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 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. Use of a therapeutic agent for the manufacture of a medicament for treatment of a pathogen infection in a subject in need thereof, wherein

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; 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.

23. 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.

24. 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.

25. 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.