Claims

1. A method for obtaining species-specific consensus sequences of microorganisms, comprising at least:

S100, obtaining candidate species-specific consensus sequences: according to the principle of sequence similarity, clustering specific sequences of target strains belonging to the same species based on a clustering algorithm, wherein after clustering, similar specific sequences form a group, and each group corresponds to a consensus sequence, wherein a specific sequence that best represents a group in different groups is selected as a consensus sequence, and the consensus sequence in a group is the closest to all the specific sequences in the group, thereby obtaining a plurality of candidate species-specific consensus sequences; wherein the specific sequences refer to target fragments belonging to the same target strain, and a region where the target fragments are located is a specific region of the target strain;

S200, verifying the candidate species-specific consensus sequences:

judging whether the candidate species-specific consensus sequences meet the following conditions:

1) a strain coverage rate meets a first preset value;

2) an effective copy number meets a second preset value;

if the candidate species-specific consensus sequences meet all the above conditions, determining that the candidate species-specific consensus sequences are species-specific consensus sequences;

wherein,

the strain coverage rate = (number of target strains with the candidate species-specific consensus sequence / total number of target strains) * 100%;

the effective copy number is calculated according to formula (I):

$$\sum_{i=1}^{n} Ci * \left(\frac{Si}{Sall} \right) \quad \text{(I);}$$

wherein,

n is a total number of copy number gradients of the candidate species-specific consensus sequences;

Ci is the copy number corresponding to the i-th copy number gradient of the candidate

species-specific consensus sequence;

Si is the number of strains within the i-th copy number gradient of the candidate species-specific consensus sequence;

Sall is a total number of the target strains.

2. The method for obtaining species-specific consensus sequences of microorganisms according to claim 1, wherein the specific region is a specific multi-copy region.

3. The method for obtaining species-specific consensus sequences of microorganisms according to claim 1, wherein obtaining the specific region comprises:

S110, respectively comparing a microorganism target fragment with whole genome sequences of one or more comparison strains one-to-one, and removing fragments of which a similarity exceeds a third preset value, to obtain a plurality of residual fragments as first-round cut fragments T_1-T_n, wherein n is an integer greater than or equal to 1;

S120, respectively comparing the first-round cut fragments T_1-T_n with whole genome sequences of remaining comparison strains, and removing fragments of which a similarity exceeds the third preset value, to obtain a collection of residual cut fragments as a candidate specific region of the microorganism target fragment; and

S130, verifying the candidate specific region to obtain the specific region: determining whether the candidate specific region meets the following requirements:

1) searching in public databases to find whether there are other species of which a similarity to the candidate specific region is greater than the third preset value;

2) respectively comparing the candidate specific region with whole genome sequences of the comparison strains and a whole genome sequence of a host of a source strain of the microorganism target fragment, to find whether there are fragments with a similarity greater than the third preset value;

if the candidate specific region does not meet the above requirements, the candidate specific region is a specific region of the microorganism target fragment.

4. The method for obtaining species-specific consensus sequences of microorganisms according to claim 3, further comprising one or more of the followings:

a. the method is capable of distinguishing whether the source strain of the microorganism target fragment and a comparison strain belong to a same species or a same subspecies;
b. the similarity refers to a product of a coverage rate and a matching rate of the microorganism target fragment, and the coverage rate = (length of similar sequence fragment/(end value of the microorganism target fragment - starting value of the microorganism target fragment + 1))%;
c. in operation S120, the first-round cut fragments T_1-T_n are respectively compared with whole genome sequences of the remaining comparison strains by group iteration;
d. the third preset value of similarity exceeds 80%;
e. positions of bases between two to-be-compared sequences do not cross;
f. the method further comprises: S111, comparing selected adjacent microorganism target fragments one-to-one; if the similarity after comparison is lower than the third preset value, issuing an alarm and displaying screening conditions corresponding to a target strain.

5.The method for obtaining species-specific consensus sequences of microorganisms according to claim 4, wherein the first-round cut fragment T_n being compared with whole genome sequences of the remaining comparison strains by group iteration comprises:

S121, dividing the remaining comparison strains into P groups, each group including a plurality of comparison strains;

S122, simultaneously comparing the first-round cut fragment T_n with the whole genome sequences of the comparison strains in the first group one-to-one, and removing fragments of which the similarity exceeds the third preset value, to obtain a plurality of residual fragments as a first-round candidate sequence library of the first-round cut fragment T_n;

S123, simultaneously comparing a previous-round candidate sequence library of the first-round cut fragment T_n with the whole genome sequences of the comparison strains in a next group one-to-one, and removing fragments of which a similarity exceeds the third preset value, to obtain a plurality of residual fragments as a next-round candidate sequence library of the first-round cut fragment T_n; repeating operation S122 from the first-round candidate sequence library until a P-th-round candidate sequence library is obtained as the candidate specific sequence

library of the first-round cut fragment T_n;

wherein a collection of all the candidate specific sequence libraries of the first-round cut fragments is the candidate specific region.

6.The method for obtaining species-specific consensus sequences of microorganisms according to claim 2, wherein obtaining the multi-copy region comprises:

S140, searching for a candidate multi-copy region: performing internal alignment on a microorganism target fragment, and searching for a region corresponding to a to-be-detected sequence of which a similarity meets a third preset value as a candidate multi-copy region, the similarity being a product of a coverage rate and a matching rate of the to-be-detected sequence;

S150, verifying the candidate multi-copy region to obtain a multi-copy region: obtaining a median value of copy numbers of the candidate multi-copy region; if the median value of the copy numbers of the candidate multi-copy region is greater than 1, the candidate multi-copy region is recorded as a multi-copy region.

7. The method for obtaining species-specific consensus sequences of microorganisms according to claim 6, further comprising one or more of the followings:

a. the coverage rate = (length of similar sequence/(end value of the to-be-detected sequence - starting value of the to-be-detected sequence + 1))%;
b. when the microorganism target fragment includes multiple incomplete motifs, the motifs are connected together before searching for the candidate multi-copy region;
c. the obtaining of the median value of the copy numbers of the candidate multi-copy region includes: determining a position of each candidate multi-copy region on the microorganism target fragment, obtaining the number of other candidate multi-copy regions covering a position of each base of the to-be-verified candidate multi-copy region, and calculating the median value of the copy numbers of the to-be-verified candidate multi-copy region;
d. in operation S150, a 95% confidence interval of the copy numbers of the candidate multi-copy region is calculated; preferably, when calculating the 95% confidence interval of the copy numbers of the candidate multi-copy region, a base number of the

candidate multi-copy region serves as a sample number, and a copy number value corresponding to each base in the candidate multi-copy region serves as a sample value.

8. The method for obtaining species-specific consensus sequences of microorganisms according to claim 1, further comprising one or more of the following operations:

S300, obtaining a candidate probes and primers by designing the probes and primers for the candidate species-specific consensus sequence according to a design rule of probes and primers; aligning a sequence of the candidate probes and primers to whole genomes of all target strains, calculating a strain coverage rate corresponding to the sequence of each probes and primers, screening out the candidate probes and primers of which the strain coverage rate meets a first preset value, and taking a candidate species-specific consensus sequence corresponding to the screened candidate probes and primers as a final species-specific consensus sequence;

S400, if none of the strain coverage rates of the candidate species-specific consensus sequences in operation S200 reaches the first preset value, combining the candidate species-specific consensus sequences, screening out a combination with a strain coverage rate reaching the first preset value and having the least consensus sequence, taking the screened combination as the candidate species-specific consensus sequence, and performing S200 to verify the candidate species-specific consensus sequences.

9. The method for obtaining species-specific consensus sequences of microorganisms according to claim 8, further comprising:

S500, if none of the strain coverage rates of the candidate probes and primers in operation S300 reaches the first preset value, combining the candidate species-specific consensus sequences, screening out a combination with a strain coverage rate reaching the first preset value and having the least consensus sequence, taking the screened combination as the candidate species-specific consensus sequence, performing S200 to verify the candidate species-specific consensus sequences.

10.The method for obtaining species-specific consensus sequences of microorganisms

according to claim 8 or 9, wherein in operations S400 and S500, the combining is performed according to the number of consensus sequences from low to high for selection.

11. The method for obtaining species-specific consensus sequences of microorganisms according to claim 8 or 9, wherein when the number of target strains is updated, the original candidate probes and primers is aligned to updated whole genomes of the target strains, a coverage rate is calculated, and whether the original candidate probes and primers can cover the updated target strains is verified.

12. A device for obtaining species-specific consensus sequences of microorganisms, comprising:

a candidate species-specific consensus sequence obtaining module, configured to obtain a plurality of candidate species-specific consensus sequences by clustering specific sequences of target strains belonging to a same species based on a clustering algorithm according to the principle of sequence similarity, wherein after clustering, similar specific sequences form a group, and each group corresponds to a consensus sequence, wherein a specific sequence that best represents a group in different groups is selected as a consensus sequence, and the consensus sequence in a group is the closest to all the specific sequences in the group; wherein the specific sequences refer to target fragments belonging to the same target strain, and a region where the target fragments are located is a specific region of the target strain;

a candidate species-specific consensus sequence verifying module, configured to judge whether the candidate species-specific consensus sequences meet the following conditions:

1) a strain coverage rate meets a first preset value;

2) an effective copy number meets a second preset value;

if the candidate species-specific consensus sequences meet all the above conditions, determining that the candidate species-specific consensus sequences are species-specific consensus sequences;

wherein,

the strain coverage rate = (number of target strains with the candidate species-specific consensus sequence / total number of target strains) * 100%;

the effective copy number is calculated according to formula (I):

$$\sum_{i=1}^{n} Ci * \left(\frac{Si}{Sall} \right) \quad \text{(I);}$$

wherein,

n is a total number of copy number gradients of the candidate species-specific consensus sequences;

Ci is the copy number corresponding to the i-th copy number gradient of the candidate species-specific consensus sequence;

Si is the number of strains within the i-th copy number gradient of the candidate species-specific consensus sequence;

Sall is a total number of the target strains.

13. The device for obtaining species-specific consensus sequences of microorganisms according to claim 12, wherein the specific region is a specific multi-copy region.

14. The device for obtaining species-specific consensus sequences of microorganisms according to claim 12, further comprising the following modules for obtaining a specific region:

a first-round cut fragment obtaining module, configured to respectively compare a microorganism target fragment with whole genome sequences of one or more comparison strains one-to-one, and remove fragments of which a similarity exceeds a third preset value, to obtain a plurality of residual fragments as first-round cut fragments T_1-T_n, wherein n is an integer greater than or equal to 1;

a candidate specific region obtaining module, configured to respectively compare the first-round cut fragments T_1-T_n with whole genome sequences of remaining comparison strains, and remove fragments of which the similarity exceeds the third preset value, to obtain a collection of residual cut fragments as a candidate specific region of the microorganism target fragment; and

a candidate specific region verifying module, configured to determine whether the candidate specific region meets the following requirements:

1) public databases are searched in to find whether there are other species of which a similarity to the candidate specific region is greater than the third preset value;

2) the candidate specific region is compared with whole genome sequences of the comparison strains and a whole genome sequence of a host of a source strain of the microorganism target fragment respectively, to find whether there are fragments with a similarity greater than the third preset value;

if the candidate specific region does not meet the above requirements, the candidate specific region is a specific region of the microorganism target fragment.

15. The device for obtaining species-specific consensus sequences of microorganisms according to claim 14, further comprising one or more of the following:

a. the device is capable of distinguishing whether the source strain of the microorganism target fragment and a comparison strain belong to the same species or the same subspecies;
b. the similarity refers to a product of a coverage rate and a matching rate of the microorganism target fragment, and the coverage rate = (length of similar sequence fragment/(end value of the microorganism target fragment - starting value of the microorganism target fragment + 1))%;
c. in the candidate specific region obtaining module, the first-round cut fragments T_1-T_n are respectively compared with whole genome sequences of the remaining comparison strains by group iteration;
d. the third preset value of similarity exceeds 80%;
e. positions of bases between two to-be-compared sequences do not cross;
f. the first-round cut fragment obtaining module further includes a raw data similarity comparison submodule, to compare selected adjacent microorganism target fragments one-to-one; if the similarity after comparison is lower than the third preset value, an alarm is issued and the screening conditions corresponding to a target strain are displayed.

16. The device for obtaining species-specific consensus sequences of microorganisms according to claim 15, wherein when a first-round cut fragment T_n is compared with whole genome sequences of the remaining comparison strains by group iteration, the candidate specific region obtaining module includes the following submodules:

a comparison strain grouping submodule, configured to divide the remaining comparison strains into P groups, each group including a plurality of comparison strains;

a first-round candidate sequence library obtaining submodule, configured to simultaneously compare the first-round cut fragment T_n with the whole genome sequences of the comparison strains in the first group one-to-one, and remove fragments of which the similarity exceeds the third preset value, to obtain a plurality of residual fragments as a first-round candidate sequence library of the first-round cut fragment T_n;

a candidate specific region obtaining submodule, configured to simultaneously compare a previous-round candidate sequence library of the first-round cut fragment T_n with whole genome sequences of the comparison strains in a next group one-to-one, and remove fragments of which the similarity exceeds the third preset value, to obtain a plurality of residual fragments as a next-round candidate sequence library of the first-round cut fragment T_n; the candidate specific region obtaining submodule is repeated from the first-round candidate sequence library until a P-th-round candidate sequence library is obtained as a candidate specific sequence library of the first-round cut fragment T_n;

wherein a collection of all the candidate specific sequence libraries of the first-round cut fragments is the candidate specific region.

17. The device for obtaining species-specific consensus sequences of microorganisms according to claim 13, further comprising the following modules for obtaining a multi-copy region:

a candidate multi-copy region searching module, configured to perform internal alignment on a microorganism target fragment, and search for a region corresponding to a to-be-detected sequence of which a similarity meets a third preset value as a candidate multi-copy region, the similarity being a product of a coverage rate and a matching rate of the to-be-detected sequence;

a candidate multi-copy region verifying module, configured to obtain a median value of copy numbers of the candidate multi-copy region; if the median value of the copy numbers of the candidate multi-copy region is greater than 1, the candidate multi-copy region is recorded as a multi-copy region.

18. The device for obtaining species-specific consensus sequences of microorganisms according to claim 17, further comprising one or more of the following:

a. the coverage rate = (length of similar sequence/(end value of the to-be-detected sequence - starting value of the to-be-detected sequence + 1))%;

b. when the microorganism target fragment includes multiple incomplete motifs, the motifs are connected together before searching for the candidate multi-copy region;

c. the candidate multi-copy region verifying module further includes a candidate multi-copy region copy number median value obtaining submodule, to determine a position of each candidate multi-copy region on the microorganism target fragment, obtain the number of other candidate multi-copy regions covering a position of each base of the to-be-verified candidate multi-copy region, and calculate the median value of the copy numbers of the to-be-verified candidate multi-copy region;

d. the candidate multi-copy region verifying module is further configured to calculate a 95% confidence interval of the copy numbers of the candidate multi-copy region; preferably, when calculating the 95% confidence interval of the copy numbers of the candidate multi-copy region, a base number of the candidate multi-copy region serves as a sample number, and a copy number value corresponding to each base in the candidate multi-copy region serves as a sample value.

19. The device for obtaining species-specific consensus sequences of microorganisms according to claim 12, further comprising one or more of the following modules:

a final species-specific consensus sequence screening module, configured to obtain a candidate probes and primers by designing the probes and primers for the candidate species-specific consensus sequence according to a design rule of probes and primers, align a sequence of the candidate probes and primers to whole genomes of all target strains, calculate a strain coverage rate corresponding to the sequence of each probes and primers, screen out the candidate probes and primers of which the strain coverage rate meets a first preset value, and take a candidate species-specific consensus sequence corresponding to the screened candidate probes and primers as a final species-specific consensus sequence;

a first consensus sequence combination screening module, configured to combine the candidate species-specific consensus sequences, screen out a combination with a strain

coverage rate reaching the first preset value and having the least consensus sequence, take the screened combination as the candidate species-specific consensus sequence, and verify the candidate species-specific consensus sequences by the candidate species-specific consensus sequence verifying module if none of the strain coverage rates of the candidate species-specific consensus sequences in the candidate species-specific consensus sequence verifying module reaches the first preset value.

20. The device for obtaining species-specific consensus sequences of microorganisms according to claim 19, further comprising:

a second consensus sequence combination screening module, configured to combine the candidate species-specific consensus sequences, screen out a combination with a strain coverage rate reaching the first preset value and having the least consensus sequence, take the screened combination as the candidate species-specific consensus sequence, and verify the candidate species-specific consensus sequences by the candidate species-specific consensus sequence verifying module if none of the strain coverage rates of the candidate probes and primers in the final species-specific consensus sequence screening module reaches the first preset value.

21. The device for obtaining species-specific consensus sequences of microorganisms according to claim 19 or 20, wherein in the first consensus sequence combination screening module and the second consensus sequence combination screening module, the combining is performed according to the number of consensus sequences from low to high for selection.

22. The device for obtaining species-specific consensus sequences of microorganisms according to claim 19 or 20, further comprising:

a sequence update coverage rate module, configured to align an original candidate probes and primers to updated whole genomes of the target strains when the number of target strains is updated, calculate the coverage rate, and verify whether the original candidate probes and primers can cover the updated target strains.

23. A computer readable storage medium, which stores a computer program, wherein when

executed by a processor, the program implements the method for obtaining species-specific consensus sequences of microorganisms according to any one of claims 1-11.

24. A computer processing device, comprising a processor and the computer readable storage medium according to claim 23, wherein the processor executes a computer program on the computer readable storage medium to implement operations of the method for obtaining species-specific consensus sequences of microorganisms according to any one of claims 1-11.

25. An electronic terminal, comprising a processor, a memory and a communicator; the memory stores a computer program, the communicator communicates with an external device, and the processor executes a computer program stored in the memory, so that the terminal executes the method for obtaining species-specific consensus sequences of microorganisms according to any one of claims 1-11.

26. A use of the method for obtaining species-specific consensus sequences of microorganisms according to any one of claims 1-11, the device for obtaining species-specific consensus sequences of microorganisms according to any one of claims 12-22, the computer readable storage medium according to claim 23, the computer processing device according to claim 24 or the electronic terminal according to claim 25 for screening template sequences in PCR amplification.

27. A method for identifying microbial species, comprising: identifying whether a target strain contains a species-specific consensus sequence by means of amplification, wherein the species-specific consensus sequence is obtained by the method for obtaining species-specific consensus sequences of microorganisms according to any one of claims 1-11.

28.The method for identifying microbial species according to claim 27, further comprising one or more of the following:

a. the method is capable of distinguishing whether a source strain of the microorganism target fragment and a comparison strain belong to the same species or the same subspecies;
b. the microorganism includes one or more of bacterium, virus, fungus, amoeba,

cryptosporidium, flagellate, microsporidium, piroplasma, plasmodium, toxoplasma, trichomonas and kinetoplastid.