Nov 23, 2021

Shine: To explore specific, sensitive and conserved biomarkers from massive microbial genomic data within intrapopulations.

Cong Ji[1]*, Junbin (Jack) Shao[1]*

1 Liferiver Science and Technology Institute, Shanghai ZJ Bio-Tech Co., Ltd. Shanghai, China

*Corresponding authors
Email: c_ji@liferiver.com.cn; junbin_shao@liferiver.com.cn

Abstract

Background

To improve the quality of nucleic acid detection reagents, we provided a new strategy, Shine, to explore specific, sensitive and conserved biomarkers from massive microbial genomic data within intrapopulations in order to improve detection sensitivity and accuracy. It is obvious that the more comprehensive genomic data are, the more effective the detection biomarkers.

Results

We demonstrated that our method could detect undiscovered multicopy conserved species-specific or even subspecies-specific target fragments, according to several clinical projects. In particular, this approach was effective for any pathogenic microorganism even in incompletely assembled motifs. Based on our strategy, the detection device designed with quantitative PCR primers and probes for systematic and automated detection of pathogenic microorganisms in biological samples may cover all pathogenic microorganisms without limits based on genome annotation. On the website https://bioinfo.liferiver.com.cn, users may select different configuration parameters depending on the purpose of the project to realize routine clinical detection practices.

Conclusions

It is recommended that our strategy is suitable to identify shared universal phylogenetic markers with few false positive or false negative errors and to automate the design of minimal primers and probes to detect pathogenic communities with cost-effective predictive power.

Keywords

Biomarker, Specific, Sensitive, Conservative, Design

Introduction

The testing and rapid detection of pathogenic organisms is a crucial undertaking related to health, safety and wellbeing, especially for the early detection of pathogens, which is important for diagnosing and preventing diseases[1-3]. While the landscape of diagnostics is rapidly evolving, polymerase chain reaction (PCR) remains the gold standard of nucleic acid-based diagnostic assays, in part due to its reliability, flexibility and wide deployment[4]. Obviously, the process of developing an emergency-use molecular-based laboratory-developed test (LDT) would be useful to other laboratories in future outbreaks and would help to lower barriers to establishing fast and accurate diagnostic testing in crisis conditions[4]. Nevertheless, the Deoxyribo Nucleic Acid (DNA) concentrations of pathogenic microorganisms in biological samples are mostly very low and close to the detection limit, so pathogen detection has become one of the most challenging aspects in clinical applications[5]. Traditional PCR or real-time PCR often lack detection sensitivity[6, 7]. Other methods, such as two-step nested PCR, may have better sensitivity, but they are not feasible for routine tests and present a high risk of contamination[8]. Thus, these methods are time consuming and costly and have poor accuracy, so it is necessary to explore biomarkers with high performance to improve the quality of reagents.

Since viruses lack shared universal phylogenetic biomarkers, a rise or drop in the concentrations of single biomarkers is not sufficient for accurate prediction of viral/bacterial community-acquired pneumonia, with overlap to varying extents depending on the marker cutoff values, detection methods, analysis, and desired specificity and sensitivity[9]. Although automatic identification of species-specific repetitive DNA sequences and their utilization for detecting microbial organisms by MultiMPrimer3 have been well presented[10], the website is limited by the lack of customized settings, especially for clinical applications. For instance, if unknown microorganisms cause epidemic outbreaks[11], the pathogenic microorganism database will be updated continuously, which may cause the original probe primer design to fail to cover epidemic pathogenic microorganisms, affecting the quality of nucleic acid detection reagents. To greatly improve the predictive power of detection, biomarker combinations have become the primary choice in many studies[12-14]. However, this approach may not be cost effective and could cause several experimental mistakes in actual mechanical processes in many clinical settings. Therefore, the importance of exploring minimal biomarkers with primers and probes to improve the detection sensitivity and accuracy at any time for any pathogen cannot be overestimated.

Generally, a common way to confirm suitable biomarkers as template regions for designing primers and probes for pathogenic microorganisms is to select specific plasmid[15] and 16S Ribosomal RNA (16S rRNA) sequences[16]. On the one hand, 16S RNA gene sequence analysis can be routinely used for the identification of mycobacteria and lead to the recognition of novel pathogens and non-cultured bacteria[17-20] because rRNA genes exist in all microbial genomes and there are often multiple copies, which can improve detection sensitivity. However, few studies have

reported consensus quantitative definitions of genera or species based on 16S rRNA gene sequence data. Several studies have highlighted that rRNA and other marker genes cannot be directly used to fully predict the functional potential of the bacterial community[21]. In fact, not all rRNA genes are species specific, i.e., rRNA genes cannot meet the requirements of species specificity or even subspecies specificity because the sequences of rRNA genes are too conserved to distinguish, especially between closely related species or even between strains of different subtypes of the same species. On the other hand, plasmid-mediated gene transfer plays an important role in the mobilization and dissemination of antibiotic resistance genes and in the spread of degradative pathways and pathogenicity determinants of pathogens[22]. However, we must note that not all microorganisms have specific-species plasmids and that some microorganisms even have no plasmids. That is, the mechanisms and selective pressures causing mosaic plasmids do not occur evenly over all species, and plasmids may provide different levels of potential variation to different species that are abundant and unevenly distributed across prokaryotic taxa[23]. Therefore, it has not been confirmed that plasmid DNA is species specific, especially because of the high similarity of plasmid DNA between some different species. Plasmids cannot universally test the species without plasmids by plasmid design. Hence, many clinical laboratories still have to validate the quality of assays by other primers and probes since plasmid PCR testing has obviously high risks of false positive or false negative errors. Overall, neither selecting a specific plasmid nor rRNA to design primers and probes for pathogenic microorganisms is the best choice.

On the basis of previous studies on comparative analysis of molecular sequence data, such as those using Molecular Evolutionary Genetics Analysis (MEGA5)[24] or Phylogenetic Analysis by Maximum Likelihood (PAML4)[25], which involved reconstructing the evolutionary histories of species and inferring the selective forces shaping the evolution of genes and species, it is also essential to practice comparative genomics in routine tests and rapid detection of pathogenic organisms for improved performance. Here, we demonstrated the Shine strategy based on comparative genomics to explore specific, sensitive and conserved biomarkers from massive microbial genomic data within populations. We hypothesized that the more comprehensive genomic data are, the more effective detection biomarkers. We aimed to show a design strategy to improve the quality of nucleic acid detection reagents, which has been validated by several clinical projects. In particular, it is available for any pathogenic microorganism even in incompletely assembled motifs. Our method could detect undiscovered multicopy universal species-specific or even subspecies-specific target fragments as design templates and automate the production of the best and minimal primer and probe sets that covered all publicly epidemic pathogenic microorganisms.

Materials and Methods

Specificity

The pathogenic genomic data were derived from public databases, such as the National Center for Biotechnology Information (NCBI) Assembly database[26], Global Initiative on Sharing All Influenza Data (GISAID)[27, 28], ChunLab's public data and analytics portal (EzBioCloud)[29], Eukaryotic Pathogen Database Project (EuPathDB)[30], Giardia Genomics Resources (GiardiaDB)[31], Trichomonas Informatics Resources (TrichDB)[31], and Fungal & Oomycete Informatics Resources (FungiDB)[32], which either contained completely assembled pathogenic genomes or incompletely assembled motifs. The defined populations were specific species or subspecies, and the control group was all the other species or subspecies of the same classification excluding the defined populations. As shown in Figure 1b, to identify the specific regions in the microorganism target fragments, 1) the microorganism target fragments were compared with the whole genome sequences of one or more comparison strains one to one, and fragments for which the similarity exceeded the preset value were removed to obtain the plurality of residual fragments as first-round cut fragments T1-Tn, wherein n was the integer greater than or equal to 1; 2) then, the first-round cut fragments T1-Tn were compared with whole genome sequences of the remaining comparison strains, and fragments for which the similarity exceeded the preset values were removed to obtain the collection of residual cut fragments as the candidate specific regions of the microorganism target fragments; and 3) the specific regions were then verified and obtained to determine whether the candidate specific regions met the following requirements: a) searching in public databases[33] to find whether there were other species for which the similarity values to the candidate specific region was greater than the preset value; and b) comparing the candidate specific regions of the whole genome sequences of the comparison strains to find whether there were fragments with the similarity greater than the preset values. If the candidate specific regions did meet the above requirements, the candidate specific regions were considered the specific regions of the microorganism target fragments.

Sensitive

To identify the multicopy regions in the microorganism target fragments illustrated in Figure 1c, 1) for searching candidate multicopy regions, internal alignments were performed on the microorganism target fragments, and searching for the regions corresponding to the to-be-detected sequences for which the similarity met the preset values as candidate multicopy regions, the similarity was the product of the coverage rates and matching rates of the to-be-detected sequence; 2) for verifying and obtaining the multicopy regions, the median values of the copy numbers of the candidate multicopy regions were obtained, including a) determining the positions of each candidate multicopy region on the microorganism target fragments; b) obtaining the numbers of other candidate multicopy regions covering the positions of each base of the to-be-verified candidate multicopy regions; and c) calculating the median values of the copy numbers of the to-be-verified candidate multicopy regions. The other candidate multicopy regions mentioned above refer to candidate multicopy regions other than the candidate multicopy regions to be verified. The target fragments of microorganisms may be chains or multiple incomplete motifs. If the median copy

numbers of the candidate multicopy regions were greater than 1, the candidate multicopy regions were recorded as multicopy regions. The preset value of the similarity could be determined as needed. The recommended preset value of similarity had to exceed 80%. If the region where the similarity met the preset value contained different motifs, the region was divided based on the original motif connection points and divided into different subregions to determine whether the subregions were candidate multicopy regions. The coverage rate = (length of similar sequence/(end value of the to-be-detected sequence – starting value of the to-be-detected sequence +1)) %. The matching rates referred to the identity values when the to-be-detected sequences were aligned with themselves. The identity values of the two aligned sequences could be obtained by software such as needle[34], water[35] or blat[36]. The length of similar sequences referred to the number of bases in which the matched fragments occupied the to-be-detected sequences when the to-be-detected sequence was aligned with other sequences, that is, the length of the matched fragments.

Conservative

As presented in Figure 1a, to obtain species-specific consensus sequences of microorganisms, 1) for searching for candidate consensus sequences, specific sequences of target strains belonging to the same species were clustered based on the clustering algorithm[37] to obtain a plurality of candidate species-specific consensus sequences; and 2) for verifying and obtaining primary-screened species-specific consensus sequences, whether the candidate species-specific consensus sequences met the following conditions remapped by mafft was determined[38]. Herein, the strain coverage rates met the preset values, and the effective copy numbers met the preset values. If the candidate species-specific consensus sequences met all the above conditions, it was determined that the candidate species-specific consensus sequences were species-specific consensus sequences; the percentage of strain = (number of target strains with the candidate species-specific consensus sequence/total number of target strains) * 100%. The effective copy numbers, i.e., weighted average copy numbers were calculated according to formula (I), where n was the total number of copy number gradients of the candidate species-specific consensus sequences; Ci was the copy number corresponding to the i-th candidate species-specific consensus sequence; Si was the number of strains with the i-th candidate species-specific consensus sequence; and Sall was the total number of target strains. Formula (I) refers to the summation of Ci (Si/Sall), where i ranges from Cmin to Cmax, and the number of i is n. Cmin is the minimum copy number of all candidate species-specific consensus sequences. Cmax is the maximum copy number of all candidate species-specific consensus sequences.

$$\sum_{i=0}^{n} Ci * (\frac{\mathrm{Si}}{Sall}) \qquad \text{(I)}$$

Design

Based on the above various combinations of different submodules, the final candidate species-specific consensus sequences could be compared to the whole genomes of all target strains to calculate the percentage of strains and effective copy numbers of the candidate species-specific consensus sequences. Designing the templates of the primary-screened species-specific consensus sequence and achieving the best sets of primers and probes were performed as follows: 1) we obtained the candidate probes and primers by Primer3[39] or Beacon Designer™; 2) the sequences of the candidate probes and primers were aligned to the whole genome of all target strains; 3) the strain coverage rates corresponding to the sequences of each probe and primer were calculated; and 4) the candidate probes and primers for which the strain coverage rates met the preset values were screened, and the primary-screened species-specific consensus sequences corresponding to the screened candidate probes and primers were chosen as the final species-specific consensus sequences.

Results

The illustration of our strategy

We developed a de novo genome alignment-based pipeline to explore the original and specific multicopy biomarkers of the defined intrapopulations to cover all the members. If either repetitive regions or specific regions were preferred, the result was split into two selections and then processed in the other modules separately. Each selection was finally focused on searching for consensus sequences and designing the best primer and probe sets. Correspondingly, it was necessary to perform double-check validation in every module, as shown in Figure 1a. One of the important details was common block deletions used to search the specific regions, and each genome of the target strains was compared with every genome of the control strains for N calculations. Common block deletions lasted for X generations with multiple threads to search specific regions or subspecific regions for M target strains, as illustrated in Figure 1b. The other key point was searching repetitive regions with different copy numbers in every target strain and extracting potential repeats for validation by remapping and statistically summarizing the mean copy numbers and variations for each repeat, and the rest were discarded. Finally, to be conservative, the maximum values of the strain coverage rate were achieved as much as possible with the fewest consensus sequences. All the logic modules were verified multiple times.

Identification of specific regions

The most striking finding of this method was the contribution of specific, sensitive, and conservative biomarkers for each species or subspecies, especially those available for microbial genomes. First, the obvious advantage of our strategy was that it was capable of detecting species-specific or even subspecies-specific target fragments that contained forward primers, reverse primers and probes separately in several projects, such as Human coronavirus HKU1 (HKU1), Human coronavirus OC43 (OC43), Human coronavirus NL63 (NL63), Human coronavirus 229E (229E), Middle East respiratory syndrome (MERS) coronavirus, severe acute respiratory syndrome coronavirus (SARS-CoV) and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

Notably, if there were no hits with the above biomarker genes or probes and no annotation, the sets were defined de novo, as presented in Table 1, and were obviously distinguished from other species or subspecies.

To accelerate the comparison, in a preferred embodiment, the first-round divided fragments T1-Tn were respectively compared with whole genome sequences of the remaining comparison strains by group iterations, as shown in Figure 1b: 1) dividing the remaining comparison strains into P groups, each group included a plurality of comparison strains; 2) simultaneously comparing the first-round divided fragment Tn with the whole genome sequences of each comparison strain in the first group one to one and removing fragments for which the similarity exceeded the preset value, the plurality of residual fragments was obtained as the first-round candidate sequence library of the first-round divided fragment Tn; 3) simultaneously comparing the previous-round candidate sequence library of the first-round divided fragment Tn with whole genome sequences of each comparison strain in the next group one to one and removing fragments for which the similarity exceeded the preset value, the plurality of residual fragments was obtained as the next-round candidate sequence library of the first-round divided fragment Tn; 4) operations from the first-round candidate sequence library were repeated until the Pth-round candidate sequence library were obtained as the candidate specific sequence library of the first-round divided fragments Tn; 5) the collection of all the candidate specific sequence libraries of the first-round divided fragments were the candidate specific regions. The method further comprised comparing selected adjacent microorganism target fragments one to one; if the similarity after comparison was lower than the preset values, an alarm was issued, and screening conditions corresponding to the target strains were displayed. Abnormal data and redundant data caused by human errors could be filtered. The target fragments of microorganisms could be the whole genomes of microorganisms or their gene fragments.

Identification of multi-copy regions

Second, compared with the previous method, our strategy was highly accurate and sensitive, and undiscovered multicopy regions could be identified which demonstrated in Figure 1c. For example, it was clear that IS6110 was identified by Shine, as shown in Table 2. The motif is an insertion element found exclusively within members of the *Mycobacterium tuberculosis* complex (MTBC), which has become an important biomarker in the identification of MTBC species[40, 41]. IS1002 is present in both *Bordetella pertussis* (BP) and *Bordetella parapertussis* (BPP) strains isolated from humans and was also detected by Shine, consistent with a recent study[42]. Finally, it could be necessary to cover all pathogenic target microorganism genomes to avoid lowering the quality of the nucleic acid detection reagents presented in Table 2. Therefore, our strategy was more flexible for customized settings to obtain the most conserved biomarkers, primers and probes provided by the continuous updating of massive microbial genomic data. Since 16S rRNA genes are not limited to whether there was a whole genome sequence that was not always multicopy, some rRNA genes

in the closely related species could not be distinguished from each other. It is likely that not all plasmids have specificity and universality and are unevenly distributed across prokaryotic taxa. In short, our method was more comprehensive than limited selection of plasmids or 16S rRNA genes as template regions, as repetitive, specific and universal target fragments could be found even in incompletely assembled motifs in any case.

Identification of conservative regions

As demonstrated in Supplementary Table1-8, the distribution of each cluster was summarized by the count numbers in parentheses of all the copy number out of parentheses. And it was obvious to pick up the target cluster sequences as the design templates by the percentage of strains and weighted average numbers of copy numbers detailed in Materials and Methods. In summary, the detection device based on our strategy, Shine, designed with quantitative PCR primers and probes for systematic and automated detection of pathogenic microorganisms in biological samples may cover all pathogenic microorganisms, including bacteria, viruses, fungi, amoebas, cryptosporidia, flagellates, microsporidia, piroplasma, plasmodia, toxoplasma, trichomonas and kinetoplastids. However, whether genome annotation is present was not a limiting factor. Operational tasks can be submitted by providing the names of the target strains and the comparison strains or by uploading sequence files locally on the website https://bioinfo.liferiver.com.cn. Therefore, users may select different configuration parameters depending on the purpose of the project. The configuration parameters mainly include the name of the workflow, target species, comparison species, uploaded local fasta files, target fragment length, species specificity, repeated region similarity, target fragment strain distribution, host sequence filtering, priority scheme (prioritizing multicopy regions vs. prioritizing specific regions), calculation of target strain and alarm threshold similarities, and primer probe design parameters. As a consequence, it was suitable for identifying shared universal phylogenetic biomarkers with few false positive or false negative errors and automating the design of minimal primers and probes to detect the pathogenic community with cost-effective predictive power.

Discussion

We demonstrated a new strategy, Shine, to explore specific, sensitive and conserved biomarkers from massive microbial genomic data within intrapopulations to improve detection sensitivity and accuracy. Several clinical projects have been carried out by devices based on Shine. Unfortunately, it should be noted that this study examined only limited public genomic data, and we are still looking forward to promoting collaboration with more organizations on the basis of open sharing of data and respect for all rights and interests[28]. Despite its preliminary characteristics, i.e., specific, sensitive and conservative, this study can be clearly described and explored in the future for several reasons, as follows.

The first aspect involved the ability to identify specific regions in microorganism target fragments. The biodiversity and evolution of vertebrate RNA viruses has expanded dramatically since the beginning of the millennium, and it has been reported that more

expensive, better sampling worldwide and more powerful approaches for virus characterization are needed to help us find these divergent viruses, such as chuviruses and jingmenviruses[43], which will help to fill the evolutionary gaps of RNA viruses[44]. With the development of methods for detecting more than 100 different nucleic acid targets at one time, FilmArray made the system well suited for the molecular detection of infectious agents, and the automated identification of pathogens from their corresponding target amplicons could be accomplished by analysis of the DNA melting curve of the amplicon[45]. Additionally, several studies have reported multiplex real-time PCR assays for detecting four microorganisms relevant to community-acquired pneumonia (CAP) infections[46] in Asia; CAP is one of the most common infectious diseases and a significant cause of mortality and morbidity globally. The availability of tests with improved diagnostic capabilities potentially leads to an informed choice of antibiotic usage and appropriate management of the patient to achieve a better treatment outcome and financial savings[46]. Herein, we generated a more significant biomarker dataset, which was validated by several clinical experiments, as described in Table 1 and Table 2. All the results support that our strategy is robust for detecting effective biomarkers. It seems that specificity, sensitivity and conservation could account for this performance. Interestingly, graphene is a lightweight, chemically stable and conductive material that can be successfully utilized for the detection of various virus strains. The current state-of-the-art applications of graphene-based systems for sensing a variety of viruses, e.g., SARS coronavirus 2 (SARS-CoV-2), influenza, dengue fever, hepatitis C virus, human immunodeficiency virus (HIV), rotavirus and Zika virus, have been summarized[47, 48]. Graphene-based biosensor technology with high sensitivity and specificity could be particularly useful in the life sciences and medicine since it can significantly enhance patient care, early disease diagnosis and pathogen detection in clinical practice[49, 50]. Notably, CRISPR-Cas systems, in particular the recently discovered DNA-targeting Cas12 and RNA-targeting Cas13 systems, both possessing unique trans-cleavage activity, are being harnessed for viral diagnostics and therapies[51]. In addition, specific high-sensitivity enzymatic reporter unlocking (SHERLOCK) testing in one pot (STOP) is a streamlined assay combining simplified extraction of viral RNA with isothermal amplification and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-mediated detection, which can be performed at a single temperature in less than one hour with minimal equipment[52]. Therefore, we tentatively propose cooperating with related institutes to combine the strategy of Shine with graphene-based biosensor technology or CRISPR-Cas systems for application in pathogen sensing.

On the other hand, when identifying multicopy regions in microorganism target fragments, the motifs are connected together before searching for candidate multicopy regions, in which the microorganism target fragments often have multiple incomplete motifs. The motif is caused by incomplete splicing of short read lengths under existing second-generation sequencing conditions. There was no specific restriction on the order in which the motifs were connected together, i.e., the motifs may have been connected to the chain in random order. If the region where the similarity met the preset value

contained different motifs, the region was divided based on the original motif connection points into different subregions to determine whether the subregions were candidate multicopy regions. This method is also suitable for whole-genome sequencing data generated by new technologies such as third-generation sequencing. In the preferred embodiment, the 95% confidence interval of the copy numbers of the candidate multicopy regions was calculated. The confidence interval refers to the estimated interval of the overall parameters constructed by the sample statistics, that is, the interval estimation of the overall copy numbers of the target regions. The confidence interval reflected the degree to which the true values of the copy numbers of the target regions were close to the measurement result. The confidence interval indicates the credibility of the measured values of the measured parameters.

Finally, this approach is related to obtaining species-specific consensus sequences for microorganisms. Were these different assignments due to the fundamental nature of the approach or the result of different approaches to species demarcation by the respective specialized study groups (SGs)? For instance, HIV-1 and HIV-2 were assigned to two different species, while SARS-CoV and SARS-CoV-2 were assigned to two strains of a single species. That is, how can the position of the viral entity in the natural world be defined? In practical terms, recognizing virus species as the principal subjects of virology would also expand the scale of the spatiotemporal framework connecting studies of natural virus variation, cross-host transmission, and pathogenicity and thus contribute to the understanding and control of virus infections[53]. Here, we present a method to ensure covering all pathogenic microorganism genomics to avoid lowering the quality of nucleic acid detection reagents. Users may submit the latest sequence dataset through a user-friendly interface. The sequence update coverage rate modules may reintegrate the latest sequence dataset into the database to calculate the coverage rates by recomparing the sequences of the original probes and primers to the updated sequences. This result may reflect whether the sequence of the original probes and primers could cover the newer strains. Exceptions always occurred for highly divergent viruses, such as Sapovirus and human astrovirus, which have limited consensus biomarkers with high performance. If none of the strain coverage rates of the candidate consensus sequences reached the preset value, we had to prioritize specificity and/or sensitivity and combine the candidate consensus sequences to improve conservation, although it may not be cost-effective and could cause several experimental errors. The recommended process was in turn performed by screening the combinations with the strain coverage rate reaching the preset values and having the fewest consensus sequences, taking the screened combinations as the candidate consensus sequences, and then verifying/obtaining the primary-screened species-specific consensus sequences. Herein, the combination could be performed according to the number of consensus sequences from low to high for selection. Unless a single consensus sequence covered all the current strains, it was possible to find two consensus sequences for which the sum of the strain coverage rates of the two consensus sequences was greater than or equal to the preset value of the strain coverage rate. If it did exist, two consensus sequences were recorded in the results; if not, three consensus sequences were

combined. That is, unless there was a single consensus sequence or two consensus sequences that could meet the preset value of the strain coverage rate, it was possible to find three consensus sequences, where the sum of the strain coverage rates of the three consensus sequences was greater than or equal to the preset value of the strain coverage rate. If it did exist, the three consensus sequences were recorded in the results; if not, four consensus sequences were combined. By that analogy, infinite numbers of consensus sequences should not be combined until the consensus sequence combination that could meet the preset value of the total strain coverage rate is found and recorded in the result.

Conclusion

Above all, the Shine strategy was presented to explore specific, sensitive and conserved biomarkers from massive microbial genomic data within intrapopulations. We have proposed a design strategy to improve the quality of nucleic acid detection reagents, which has been validated by several clinical projects. Our method was highly accurate and sensitive and could be capable of detecting undiscovered multicopy universal species-specific and even subspecies-specific target fragments, covering all publicly epidemic pathogenic microorganisms. Therefore, it was suitable for identifying shared universal phylogenetic biomarkers with few false positive or false negative errors and automating the design of minimal primers and probes to detect the pathogenic community with cost-effective predictive power.

Abbreviations

PCR: Polymerase Chain Reaction
LDT: Laboratory-developed Test
DNA: Deoxyribo Nucleic Acid
RNA: Ribonucleic Acid
16S rRNA: 16S Ribosomal RNA
MEGA: Molecular Evolutionary Genetics Analysis
PAML: Phylogenetic Analysis by Maximum Likelihood
NCBI: National Center for Biotechnology Information
GISAID: Global Initiative on Sharing All Influenza Data
EzBioCloud: ChunLab's public data and analytics portal
EuPathDB: Eukaryotic Pathogen Database Project
GiardiaDB: Giardia Genomics Resources
TrichDB: Trichomonas Informatics Resources
FungiDB: Fungal & Oomycete Informatics Resources
HKU1: Human coronavirus HKU1
OC43: Human coronavirus OC43
NL63: Human coronavirus NL63
229E: Human coronavirus 229E
MERS: Middle East respiratory syndrome-related coronavirus
SARS-CoV: Severe acute respiratory syndrome coronavirus
SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

MTBC: Mycobacterium tuberculosis complex
BP: Bordetella pertussis
BPP: Bordetella parapertussis
CAP: community-acquired pneumonia
HIV: human immunodeficiency virus
CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats
CRISPR-Cas: CRISPR-associated (Cas) genes
SHERLOCK: specific high-sensitivity enzymatic reporter unlocking
STOP: SHERLOCK testing in one pot
SGs: study groups

References

1. Priyanka, B., R. Patil, and S. Dwarakanath, *A review on detection methods used for foodborne pathogens.* Indian Journal of Medical Research, 2016. **144**(3): p. 327-338.
2. Upadhyay, A., et al., *ZnO Nanolower-Based NanoPCR as an Efficient Diagnostic Tool for Quick Diagnosis of Canine Vector-Borne Pathogens.* Pathogens, 2020. **9**(2).
3. Fox, R.T.V., *The present and future use of technology to detect plant pathogens to guide disease control in sustainable farming systems.* Agriculture, Ecosystems & Environment, 1997. **64**(2): p. 125-132.
4. Anahtar, M.N., et al., *Development of a qualitative real-time RT-PCR assay for the detection of SARS-CoV-2: a guide and case study in setting up an emergency-use, laboratory-developed molecular microbiological assay.* Journal of Clinical Pathology, 2021: p. jclinpath-2020-207128.
5. Rajapaksha, P., et al., *A review of methods for the detection of pathogenic microorganisms.* Analyst, 2019. **144**(2): p. 396-411.
6. Thornton, B. and C. Basu, *Rapid and simple method of qPCR primer design.* Methods Mol Biol, 2015. **1275**: p. 173-9.
7. Smith, C.J. and A.M. Osborn, *Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology.* FEMS Microbiol Ecol, 2009. **67**(1): p. 6-20.
8. Lusi, E.A., et al., *One-step nested PCR for detection of 2 LTR circles in PBMCs of HIV-1 infected patients with no detectable plasma HIV RNA.* J Virol Methods, 2005. **125**(1): p. 11-3.
9. Thomas, J., et al., *Blood biomarkers differentiating viral versus bacterial pneumonia aetiology: a literature review.* Italian journal of pediatrics, 2020. **46**(1): p. 4-4.
10. Koressaar, T., K. Joers, and M. Remm, *Automatic identification of species-specific repetitive DNA sequences and their utilization for detecting microbial organisms.* Bioinformatics, 2009. **25**(11): p. 1349-55.
11. Fumian, T.M., et al., *Detection of norovirus epidemic genotypes in raw sewage using next generation sequencing.* Environ Int, 2019. **123**: p. 282-291.
12. Valim, C., et al., *Responses to Bacteria, Virus, and Malaria Distinguish the Etiology of Pediatric Clinical Pneumonia.* American journal of respiratory and critical care medicine, 2016. **193**(4): p. 448-459.
13. Elemraid, M.A., et al., *Utility of inflammatory markers in predicting the aetiology of pneumonia in children.* Diagn Microbiol Infect Dis, 2014. **79**(4): p. 458-62.

14. Naydenova, E., et al., *The power of data mining in diagnosis of childhood pneumonia.* J R Soc Interface, 2016. **13**(120).
15. Antipov, D., et al., *Plasmid detection and assembly in genomic and metagenomic data sets.* Genome research, 2019. **29**(6): p. 961-968.
16. van Hattem, J.M. and B. de Wever, *16S rRNA sequence analysis: application and pitfalls.* Ned Tijdschr Geneeskd, 2019. **163**.
17. Clarridge, J.E., 3rd, *Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases.* Clin Microbiol Rev, 2004. **17**(4): p. 840-62, table of contents.
18. Chakravorty, S., et al., *A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.* J Microbiol Methods, 2007. **69**(2): p. 330-9.
19. Matsuki, T., et al., *Development of 16S rRNA-gene-targeted group-specific primers for the detection and identification of predominant bacteria in human feces.* Appl Environ Microbiol, 2002. **68**(11): p. 5445-51.
20. Patel, J.B., *16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory.* Mol Diagn, 2001. **6**(4): p. 313-21.
21. Sevigny, J.L., et al., *Marker genes as predictors of shared genomic function.* BMC Genomics, 2019. **20**(1): p. 268.
22. Smalla, K., S. Jechalke, and E.M. Top, *Plasmid Detection, Characterization, and Ecology.* Microbiol Spectr, 2015. **3**(1): p. PLAS-0038-2014.
23. Pesesky, M.W., R. Tilley, and D.A.C. Beck, *Mosaic plasmids are abundant and unevenly distributed across prokaryotic taxa.* Plasmid, 2019. **102**: p. 10-18.
24. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Mol Biol Evol, 2011. **28**(10): p. 2731-9.
25. Yang, Z., *PAML 4: Phylogenetic Analysis by Maximum Likelihood.* Molecular Biology and Evolution, 2007. **24**(8): p. 1586-1591.
26. Kitts, P.A., et al., *Assembly: a resource for assembled genomes at NCBI.* Nucleic Acids Res, 2016. **44**(D1): p. D73-80.
27. Elbe, S. and G. Buckland-Merrett, *Data, disease and diplomacy: GISAID's innovative contribution to global health.* Glob Chall, 2017. **1**(1): p. 33-46.
28. Shu, Y. and J. McCauley, *GISAID: Global initiative on sharing all influenza data - from vision to reality.* Euro Surveill, 2017. **22**(13).
29. Yoon, S.H., et al., *Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies.* Int J Syst Evol Microbiol, 2017. **67**(5): p. 1613-1617.
30. Warrenfeltz, S., et al., *EuPathDB: The Eukaryotic Pathogen Genomics Database Resource*, in *Eukaryotic Genomic Databases: Methods and Protocols*, M. Kollmar, Editor. 2018, Springer New York: New York, NY. p. 69-113.
31. Aurrecoechea, C., et al., *GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis.* Nucleic Acids Research, 2009. **37**(suppl_1): p. D526-D530.
32. Basenko, E.Y., et al., *FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes.* J Fungi (Basel), 2018. **4**(1).

33. Stephen F.A, et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
34. B.N., S. and C. D.W., *A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins.* Journal of Molecular Biology, 1970. **48**(3): p. 443-453.
35. Smith TF and W. MS, *Identification of common molecular subsequences.* Journal of Molecular Biology, 1981. **1**(147): p. 195-197.
36. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
37. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-1.
38. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Mol Biol Evol, 2013. **30**(4): p. 772-80.
39. Untergasser, A., et al., *Primer3--new capabilities and interfaces.* Nucleic acids research, 2012. **40**(15): p. e115-e115.
40. Coros, A., E. DeConno, and K.M. Derbyshire, *IS6110, a Mycobacterium tuberculosis complex-specific insertion sequence, is also present in the genome of Mycobacterium smegmatis, suggestive of lateral gene transfer among mycobacterial species.* J Bacteriol, 2008. **190**(9): p. 3408-10.
41. Millan-Lou, M.I., et al., *Global study of IS6110 in a successful Mycobacterium tuberculosis strain: clues for deciphering its behavior and for its rapid detection.* J Clin Microbiol, 2013. **51**(11): p. 3631-7.
42. Van, D., et al., *The Differentiation of Bordetella parapertussis and Bordetella bronchiseptica from Humans and Animals as Determined by DNA Polymorphism Mediated by Two Different Insertion Sequence Elements Suggests Their Phylogenetic Relationship.* Int J Syst Bacteriol, 1996. **46**(3): p. 640-647.
43. Li, C.-X., et al., *Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses.* eLife, 2015. **4**: p. e05378.
44. Zhang, Y.Z., et al., *The diversity, evolution and origins of vertebrate RNA viruses.* Curr Opin Virol, 2018. **31**: p. 9-16.
45. Poritz, M.A., et al., *FilmArray, an automated nested multiplex PCR system for multi-pathogen detection: development and application to respiratory tract infection.* PLoS One, 2011. **6**(10): p. e26047.
46. Koo, S.H., et al., *Development of a rapid multiplex PCR assay for the detection of common pathogens associated with community-acquired pneumonia.* Transactions of The Royal Society of Tropical Medicine and Hygiene, 2021.
47. El Moutaouakil, A., et al., *Review: Graphene-based biosensor for Viral Detection*. 2020.
48. Vermisoglou, E., et al., *Human virus detection with graphene-based materials.* Biosens Bioelectron, 2020. **166**: p. 112436.
49. Pena-Bahamonde, J., et al., *Recent advances in graphene-based biosensor technology with applications in life sciences.* J Nanobiotechnology, 2018. **16**(1): p. 75.
50. Das Jana, I., et al., *Development of a copper-graphene nanocomposite based transparent coating with antiviral activity against influenza virus.* BioRxiv, 2020.
51. Freije, C.A. and P.C. Sabeti, *Detect and destroy: CRISPR-based technologies for the response against viruses.* Cell Host Microbe, 2021. **29**(5): p. 689-703.

52. Joung, J., et al., *Detection of SARS-CoV-2 with SHERLOCK One-Pot Testing.* New England Journal of Medicine, 2020. **383**(15): p. 1492-1494.
53. Gorbalenya, A.E. and S.G. Siddell, *Recognizing species as a new focus of virus research.* PLoS pathogens, 2021. **17**(3): p. e1009318-e1009318.

Ethical Approval and Consent to participate
All protocols were approved by the Liferiver Science and Technology Institute, Shanghai ZJ Bio-Tech Co., Ltd and Use Committee (Shanghai, China).

Consent for publication
Not applicable.

Availability of data and materials
All shared universal phylogenetic markers from massive microbiological genomics have been available at the website https://bioinfo.liferiver.com.cn/#/home.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by no funding.

Authors' contributions
The authors' responsibilities were as follows-JBS and CJ designed and conducted the research; CJ and JBS analyzed the data and perform the analysis; CJ wrote the paper; JBS revised the manuscript; and CJ had primary responsibility for the final content. The authors read and approved the final manuscript.

Acknowledgments
Appreciation goes to Zhang Hanyan, Xiong Lei, and Pan Daxia for their experimental validation in carrying out this study. The authors deeply thank Liu Yan, Zhang Jie, and Li Qiang for their valuable suggestions and comments on this work. Many facets of the user-interface design benefited from Niu Xingsheng, Lu Wang and Pan Yajie. We wish to express our thanks for the valuable modifications to the paper made by Shen Yilin, Zhu Lingjiao, Guo Jingjing, and Zhou Miaomiao, who helped us greatly revise this paper. All the other data supporting the findings of this study and the computational code used in this study are available from the corresponding authors upon reasonable request. Cong Ji and Junbin (Jack) Shao are named inventors on the pending PCT Patent Applications PCT/CN2020/090180, PCT/CN2020/090175, and PCT/CN2020/090177 filed by the Liferiver Science and Technology Institute of Shanghai ZJ Bio-Tech Co., Ltd., which separately describe the method and device for identifying multicopy, species-specific consensus sequences in microorganism target fragments and use thereof. The other authors declare no competing interests.

Author information
Liferiver Science and Technology Institute, Shanghai ZJ Bio-Tech Co., Ltd. Shanghai, China
Cong Ji, Junbin (Jack) Shao

Figure and Table Legends

Figure 1 Schematic map of Shine. This new strategy was used to explore specific, sensitive and conserved biomarkers to cover all members of defined intrapopulations. **1a.** The total pipeline contains two selections, i.e., to search specific regions preferentially or to search sensitive regions preferentially. **1b.** Illustration of the submodule searching for the specific regions preferentially. **1c.** Illustration of the submodule for searching for the multicopy regions preferentially.

Table 1. Sample sets for detecting species-specific or even subspecies-specific target fragments. This output includes forward primers, reverse primers and probes separately for several projects on different species or subspecies of coronavirus which meet three criterias: specificity, sensitivity and conservation.

Table 2. Sample sets of identified undiscovered multicopy regions compared with known 16S rRNA genes. This output included all the de novo multicopy fragments identified by our method and several known 16S rRNA genes for the target pathogenic microorganisms with corresponding copy numbers and conservation.

Supplementary Table1. Cluster sets of identified multicopy regions from Mycobacterium tuberculosis, Mycobacterium Africanum, Mycobacterium bovis, Tuberculosis complex.

Supplementary Table2. Cluster sets of identified multicopy regions from B. pertussis, B. parapertussis, B. holmesii.

Supplementary Table3. Cluster sets of identified multicopy regions from B. parapertussis, B. bronchiseptica.

Supplementary Table4. Cluster sets of identified multicopy regions from M. pneumoniae strain M129.

Supplementary Table5. Cluster sets of identified multicopy regions from Streptococcus agalactiae.

Supplementary Table6. Cluster sets of identified multicopy regions from H. pylori UA802, H. pylori strain PMSS1, H. pylori strain 7.13.

Supplementary Table7. Cluster sets of identified undiscovered multicopy regions from Legionella pneumophila.

Supplementary Table8. Cluster sets of identified undiscovered multicopy regions from Candida auris.

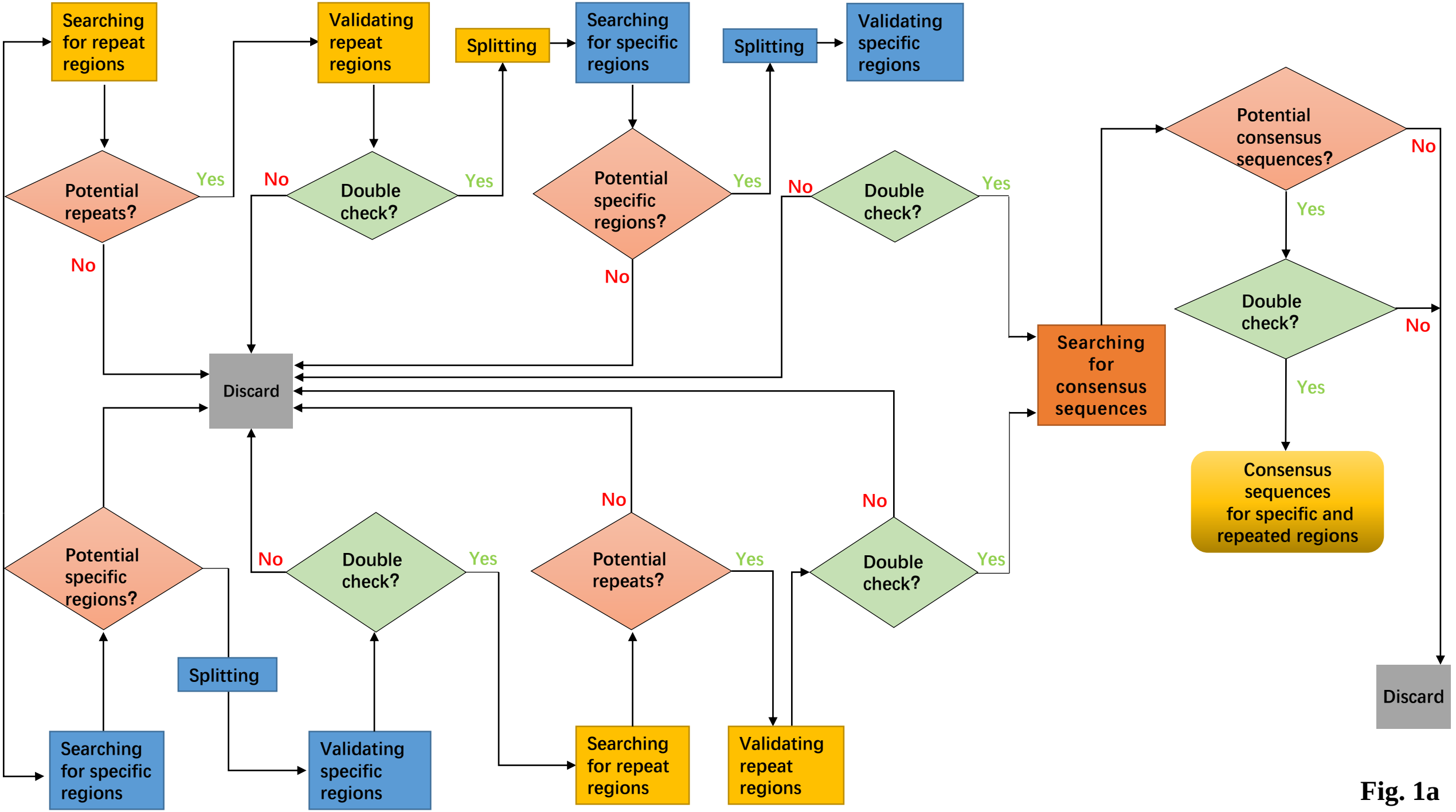
Searching for repeat regions
Validating repeat regions
Splitting
Searching for specific regions
Splitting
Validating specific regions
Potential repeats?
Yes
No
Double check?
No
Yes
Potential specific regions?
Yes
No
Double check?
No
Yes
Potential consensus sequences?
No
Yes
Double check?
No
Yes
Searching for consensus sequences
Discard
Consensus sequences for specific and repeated regions
Potential specific regions?
No
Double check?
No
Yes
Potential repeats?
No
Yes
Double check?
No
Yes
Splitting
Searching for specific regions
Validating specific regions
Searching for repeat regions
Validating repeat regions
Discard

Fig. 1a

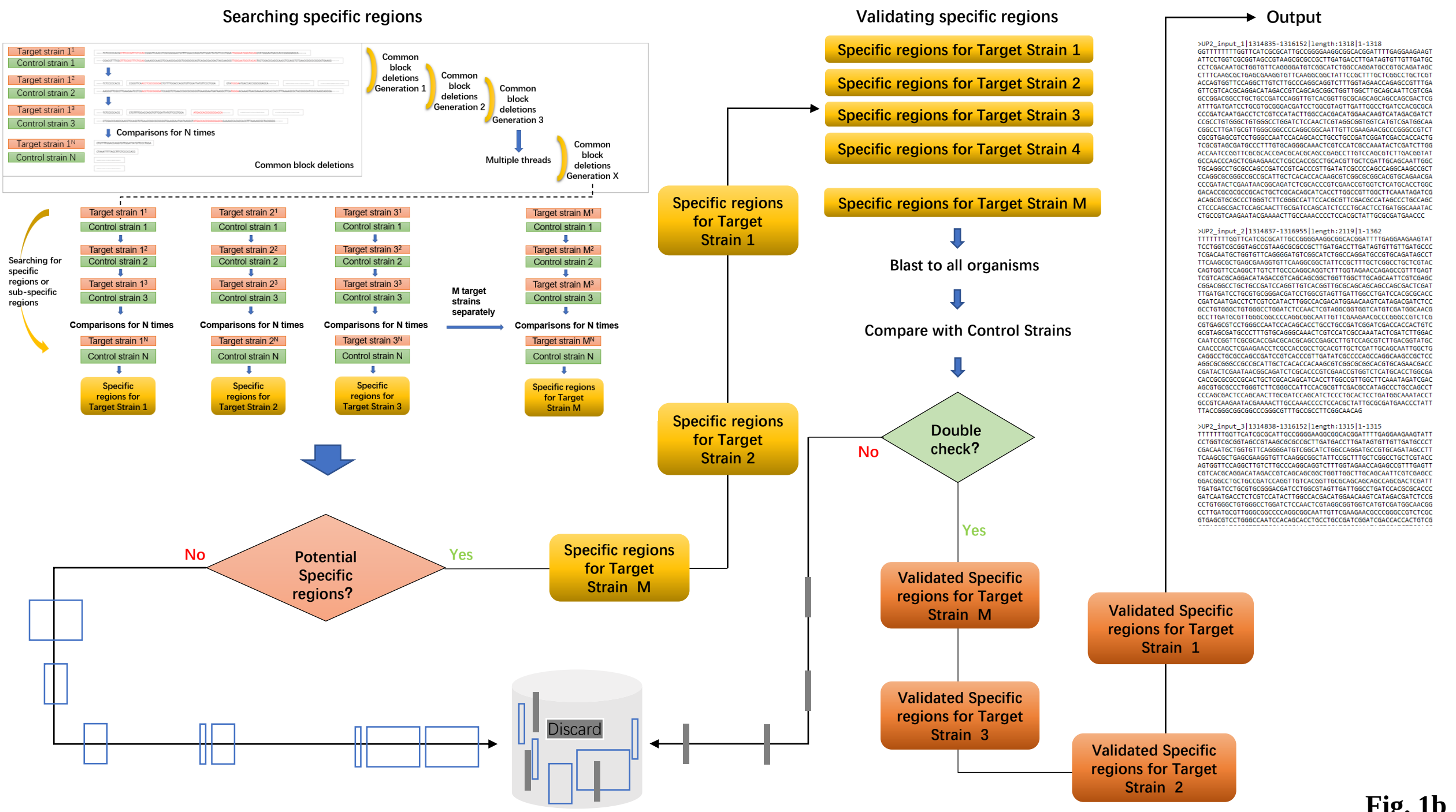
Searching specific regions
Target strain 1¹
Control strain 1
Target strain 1²
Control strain 2
Target strain 1³
Control strain 3
Target strain 1ᴺ
Control strain N
Comparisons for N times
Common block deletions
Common block deletions Generation 1
Common block deletions Generation 2
Common block deletions Generation 3
Multiple threads
Common block deletions Generation X
Searching for specific regions or sub-specific regions
Target strain 2¹
Control strain 1
Target strain 2²
Control strain 2
Target strain 2³
Control strain 3
Target strain 2ᴺ
Control strain N
Target strain 3¹
Control strain 1
Target strain 3²
Control strain 2
Target strain 3³
Control strain 3
Target strain 3ᴺ
Control strain N
M target strains separately
Target strain M¹
Control strain 1
Target strain M²
Control strain 2
Target strain M³
Control strain 3
Target strain Mᴺ
Control strain N
Comparisons for N times
Specific regions for Target Strain 1
Specific regions for Target Strain 2
Specific regions for Target Strain 3
Specific regions for Target Strain M
Potential Specific regions?
No
Yes
Specific regions for Target Strain M
Specific regions for Target Strain 2
Specific regions for Target Strain 1
Discard
Validating specific regions
Specific regions for Target Strain 1
Specific regions for Target Strain 2
Specific regions for Target Strain 3
Specific regions for Target Strain 4
Specific regions for Target Strain M
Blast to all organisms
Compare with Control Strains
Double check?
No
Yes
Validated Specific regions for Target Strain M
Validated Specific regions for Target Strain 3
Validated Specific regions for Target Strain 2
Validated Specific regions for Target Strain 1
Output
>UP2_input_1|1314835-1316152|length:1318|1-1318
>UP2_input_2|1314837-1316955|length:2119|1-1362
>UP2_input_3|1314838-1316152|length:1315|1-1315

Fig. 1b

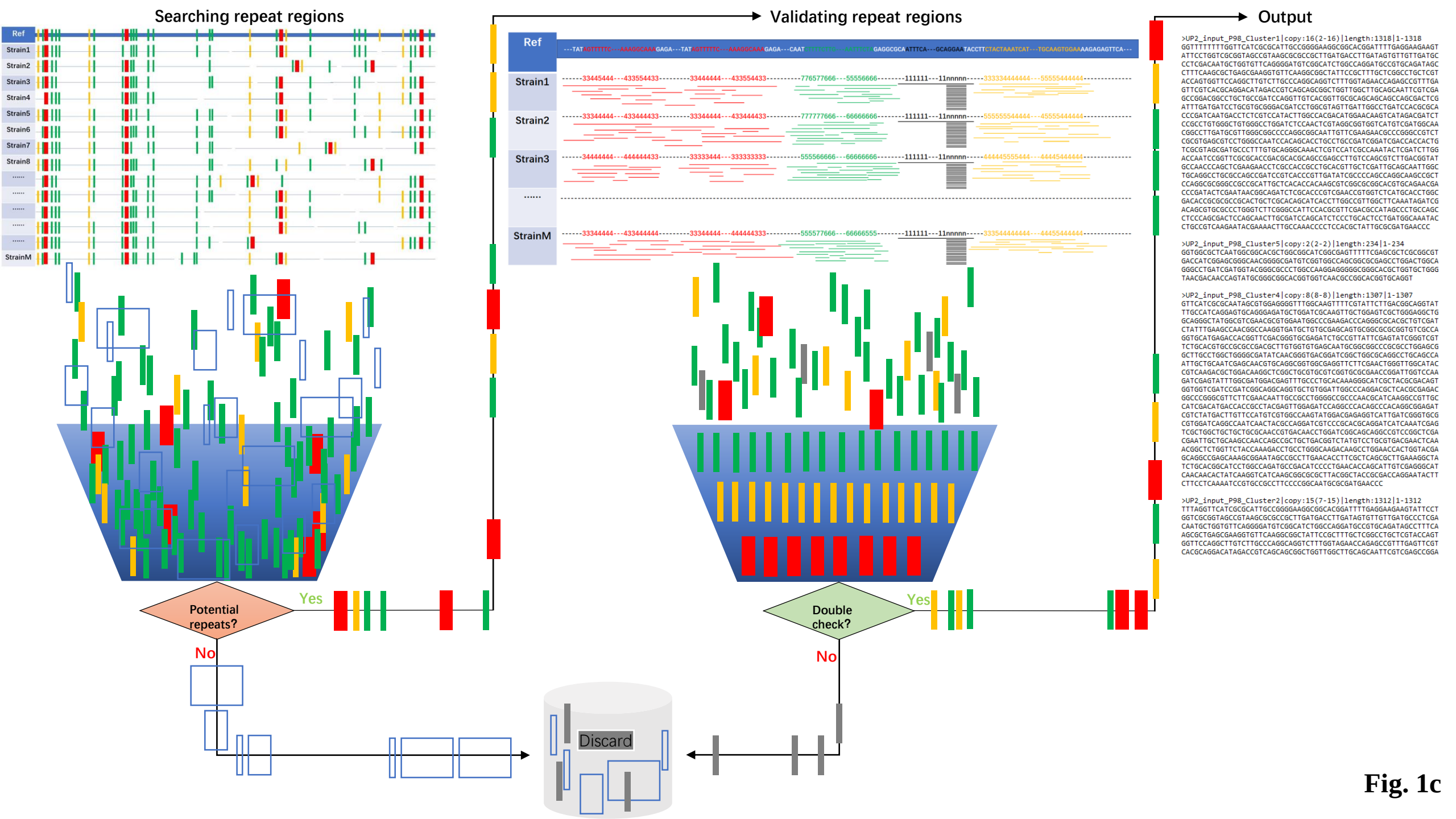
Searching repeat regions
Validating repeat regions
Output
Ref
Strain1
Strain2
Strain3
Strain4
Strain5
Strain6
Strain7
Strain8
StrainM
Potential repeats?
Yes
No
Double check?
Yes
No
Discard

Fig. 1c