March 23th, 2022

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

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

Since DNA concentrations of pathogenic microorganisms in biological samples were mostly very low and close to the detection limit with time consuming and costly with poor accuracy, pathogen detection was lack of shared universal phylogenetic repetitive biomarkers and utilization has become one of the most challenging aspects in clinical applications. Limited by failing to cover updated epidemic testing samples, some service was hard to practice in clinical applications without customized settings in order to lower the quality of nucleic acid detection reagents. Furthermore, in order to maintain the high conservations, biomarker combinations may not be cost effective and could cause several experimental mistakes in actual mechanical processes in many clinical settings. With the limit of recent developing technology and knowledge, 16s rRNA was too conservative to distinguish the closely related species, and likely mosaic plasmid was always not effective because of its unevenly distributed across prokaryotic taxa.

Results

Here we provided a new strategy, Shine, to explore specific, sensitive and conserved biomarkers from massive microbial genomic data within intra-populations in order to improve detection sensitivity and accuracy. Distinguished with simple concatenations with blast-based filtering, our method was constructed by a de novo genome alignment-based pipeline to explore the original and specific repetitive biomarkers of the defined intra-population to cover all members to detect newly discovered multi-copy conserved species-specific or even subspecies-specific target probes and primers sets, which has been successfully applied on a number of clinical projects and have the overwhelming advantages of quantitative systematic and automated detection of all pathogenic microorganisms without limits of genome annotation and incompletely assembled motifs. Based on our strategy, users may select different configuration parameters depending on the purpose of the project to realize routine clinical detection practices on the website https://bioinfo.liferiver.com.cn with easy-to-go registration.

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 high effective primers and probes to detect pathogenic communities with cost-effective predictive power.

Keywords

Biomarker, Specific, Sensitive, Conservative, Design

Background

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 deoxyribonucleic 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]. For instance, either traditional PCR or real-time PCR often lack detection sensitivity[6, 7], or 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]. In consequence, 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.

Accuracy and sensitivity have been played the important role in quality of nucleic acid detection reagents, that accuracy mainly due to high specificity and conservation. Contributed to lack of 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 shared universal phylogenetic and repetitive biomarkers with primers and probes to improve the detection sensitivity and accuracy at any time for any pathogen cannot be overestimated.

However, the common way to confirm suitable biomarkers as template regions for designing primers and probes for pathogenic microorganisms is to select 16S ribosomal RNA (16S rRNA) sequences[15] and specific plasmid[16]. On the one hand, although 16S RNA gene sequence analysis can be routinely used for the identification of

mycobacteria and lead to the explore novel pathogens and non-cultured bacteria[17-20] because of existing in all microbial genomes and there are often multiple copies related with detection sensitivity. Few studies have reported consensus quantitative definitions of genera or species based on 16S rRNA gene sequence data. And 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]. The main reason is 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 still doesn't work that 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]. The principal reason is 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]. At last, 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 the specific plasmid nor 16S rRNA to design primers and probes for all the specific pathogenic microorganisms is the appropriate drop-in strategy.

On the basis of above challenges, we propose a new strategy Shine which is also essential to practice comparative genomics in routine tests and rapid detection of pathogenic organisms for improved performance, while previous studies involved in reconstructing the evolutionary histories of species and inferring the selective forces shaping the evolution of genes and species 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]. Hypothesized that the more comprehensive genomic data are, the more effective detection biomarkers. To explore specific, sensitive and conserved biomarkers from massive microbial genomic data within populations, we aimed to show a design strategy to improve the quality of nucleic acid detection reagents, which has been validated by several clinical projects.

Results

The illustration of our strategy

Generally, we developed a de novo genome alignment-based pipeline to explore the original and specific multi-copy biomarkers of the defined intra-populations to cover all the members. Either repetitive regions or specific regions were preferred, raw data was divided into two selections, i.e., Prioritizing multi-copy regions for big genomes and Prioritizing specific regions for small genomes. And then it would be processed in the other modules separately. That was to say, if the first direction, it would be processed into the next module of searching for specific regions. If the second direction, it would be pushed into the next module of searching for multi-copy regions. Finally, it was 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, distinguished with blast-based filtering method. Common block deletions lasted for X generations with multiple threads to search specific regions or sub-specific regions for M target strains, as illustrated in Figure 1b. 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 target comparison strains into M groups, each group included a plurality of the same comparison strains; 2) simultaneously comparing the first-round divided fragment T^1 with the whole genome sequences of comparison strain 1 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 T^1; 3) simultaneously comparing the previous-round candidate sequence library of the first-round divided fragment T^1 with whole genome sequences of comparison strain 2 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 T^2; 4) operations from the second-round candidate sequence library were repeated until the N-round candidate sequence library were obtained as the candidate specific sequence library of the N-round divided fragments Tn; 5) the collection of all the candidate specific sequence libraries of the N-round divided fragments were the candidate specific regions. The further validation comprised two parts of scoring and filtering the potential specific regions: blast to all organisms and compare genomic with control strains, in order that we could catch the specific regions with the stringent quality standards.

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 repeats, and

the rest were discarded in Figure 1c. Here each repeats were filtered into the classified summation so that we could get the minimal accurate sets of repetitive sequences from the whole genome. When identifying multi-copy regions in microorganism target fragments, the motifs are connected together before searching for candidate multi-copy regions, in which the microorganism target fragments often have multiple incomplete motifs.

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. Demonstrated in Table2 and 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. Designing the templates of the primary-screened species-specific consensus sequence and achieving the best sets of primers and probes were performed.

Finally, operational tasks could 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 with easy-to-go registration. 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 multi-copy regions vs. prioritizing specific regions), calculation of target strain and alarm threshold similarities, and primer probe design parameters.

Application and practice on a series of clinical projects

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.

Compared with the previous method, our strategy was highly accurate and sensitive, and undiscovered multi-copy regions could be identified which demonstrated in Table 2. For instance, it was clear that IS6110 was identified by Shine, that the motif was 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[26, 27]. Likely, comparing with known 16s rRNA genes, i.e., IS1081 and PGRS of *Mycobacterium tuberculosis* complex (MTBC), IS1663 of *Bordetella pertussis* (BP), IS1001 of *Bordetella pertussis* (BP) and *Bordetella parapertussis* (BPP), RepMP 4/5/1 of *Mycoplasma pneumonia* (MP), cagA of *Helicobacter pylori* (HP) were all consistent with our study. Interestingly, IS1002 was present in both *Bordetella pertussis* (BP) and *Bordetella parapertussis* (BPP) strains isolated from humans detected by Shine, inconsistent with the reported copy numbers of recent study [28], that only 47.2% of strains in BP and BPP have IS1002 specific region with 0.52 the effective copy number heterozygous in taxa. In case of selecting IS1002 as the designing templates, there would be produced obviously false negative errors that IS1002 could not be covered across all BP and BPP strains. Besides we cought several cases with false negative errors, for example the conservation of IS481 of BP and BPP was only 93.2% with 1.2 effective copy numbers. It was also clear that GBSil of *Streptococcus agalactiae* would not be core templates to cover across all *Streptococcus agalactiae*, since that the conservation of GBSil was only 28.6%.

At the same time, it was clear that the performance of our strategy on reducing false positive rate for sample sets of identified newly discovered species or subspecies specific multi-copy regions shown in Table 3, such as *Mycobacterium tuberculosis, Bordetella pertussis, Bordetella parapertussis, Mycoplasma pneumonia, Streptococcus agalactiae, Helicobacter pylori, Legionella pneumophila, Candida auris.* Here we demonstrated the distribution of copy numbers of de novo biomarkers by the weighted average copy numbers and a series of sequences of identified newly discovered species or subspecies specific multi-copy regions as the potential designing templates, in order to improving to avoid false positive rate of testing for any sequencing pathogenic organisms. It should be very competitive strategy to catch the newly discovered biomarkers.

Advantages of quantitative systematic and automated detection

Since 16S rRNA genes are not limited to whether there was a whole genome sequence that was not always multi-copy, some rRNA genes in the closely related species could not be distinguished from each other. Likely, not all plasmids have specificity and universality and are unevenly distributed across prokaryotic taxa. However, our method is highly accurate and sensitive and could be capable of detecting newly discovered multi-copy universal species-specific and even subspecies-specific target fragments, covering all publicly epidemic pathogenic microorganisms. 16s rRNA genes such as IS481, RepMP 2/3, GBSil, could not be able to cover all pathogenic target microorganism genomes to avoid producing false negative rates and lowering the quality of the nucleic acid detection reagents presented in Table 2. Yet we tried to explore a series of newly discovered specific, sensitive and conserved biomarkers database to improve the performance of clinical kits and reducing the false positive

rates in Table 3. In all, our study 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. In particular, it 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.

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. Also 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. 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 re-comparing the sequences of the original probes and primers to the updated sequences. This result could reflect whether the sequence of the original probes and primers could cover the newer strains.

Whether genome annotation is present was not a limiting factor or not, the purpose of accurate detection with specific settings may cover all pathogenic microorganisms, including bacteria, viruses, fungi, amoebas, cryptosporidia, flagellates, microsporidia, piroplasma, plasmodia, toxoplasma, trichomonas and kinetoplastids. This method is also available for whole-genome sequencing data generated by new technologies such as third-generation sequencing.

Discussion

To explore specific, sensitive and conserved biomarkers from massive microbial genomic data within intra-populations 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[29]. Despite its preliminary characteristics, this study still could be more clearly promoted and explored in the future such as follows.

Assumed that the more comprehensive genomic data are, the more effective the detection biomarkers, the necessary of the ability to identify specific regions in microorganism target fragments cannot be overestimated. As we known, 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 [30], which will help to fill the evolutionary gaps of RNA viruses[31]. 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[32]. Additionally, several studies have reported multiplex real-time PCR assays for detecting four microorganisms relevant to community-acquired pneumonia (CAP) infections[33] 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[33].

Herein, we generated a more significant biomarker dataset, which was validated by several clinical experiments, as described in Table 1-3. All the results support that our strategy is robust for detecting effective biomarkers, which seems that specificity 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[34, 35]. 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[36, 37]. Furthermore, 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[38]. 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 could be performed at a single temperature in less than one hour with minimal equipment[39]. As a consequence, 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.

To ensure covering all pathogenic microorganism genomics and effectively identify the specific species or subspecies to avoid lowering the quality of nucleic acid detection reagents, of course, 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. 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 pre-set 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 pre-set 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. Our strategy has been successfully applied on a number of clinical projects and have the overwhelming advantages of quantitative systematic and automated detection of all pathogenic microorganisms without limits of genome annotation and incompletely assembled motifs described above.

Based on the incomplete motifs, 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. Comparing with 16s rRNA, de novo biomarkers by the weighted average copy numbers and a series of sequences of identified newly discovered species or subspecies specific multi-copy regions could be cought as the potential designing templates, in order to improving to avoid false positive rate of testing for any sequencing pathogenic organisms.

In addition to the above characteristics, 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[40].

Materials and Methods

Acquisition of raw data.

The pathogenic genomic data were derived from public databases, such as the National Center for Biotechnology Information (NCBI) Assembly database[41], Global Initiative on Sharing All Influenza Data (GISAID)[29, 42], ChunLab's public data and analytics portal (EzBioCloud)[43], Eukaryotic Pathogen Database Project (EuPathDB)[44], Giardia Genomics Resources (GiardiaDB)[45], Trichomonas Informatics Resources (TrichDB)[45], and Fungal & Oomycete Informatics Resources (FungiDB)[46], Pathosystems Resource Integration Center database (PATRIC) [47] 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. 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.

To search for the specific regions.

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[48] to find whether there were other species for which the similarity values to the candidate specific region was greater than the pre-set value; and b) comparing the candidate specific regions of the whole genome sequences with 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.

To search for the multiple copies regions.

The target fragments of microorganisms may be chains or multiple incomplete motifs. When identifying multi-copy regions in microorganism target fragments, the motifs were connected together before searching for candidate multi-copy regions, in which the microorganism target fragments often have multiple incomplete motifs. The motif

was caused by incomplete splicing of short read lengths under existing second-generation sequencing conditions. To identify the multi-copy regions in the microorganism target fragments illustrated in Figure 1c, 1) for searching candidate multi-copy 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 multi-copy 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 multi-copy regions, the median values of the copy numbers of the candidate multi-copy regions were obtained, including a) determining the positions of each candidate multi-copy region on the microorganism target fragments; b) obtaining the numbers of other candidate multi-copy regions covering the positions of each base of the to-be-verified candidate multi-copy regions; and c) calculating the median values of the copy numbers of the to-be-verified candidate multi-copy regions.

The other candidate multi-copy regions mentioned above refer to candidate multi-copy regions other than the candidate multi-copy regions to be verified. If the median copy numbers of the candidate multi-copy regions were greater than 1, the candidate multi-copy regions were recorded as multi-copy 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 sub-regions to determine whether the sub-regions were candidate multi-copy 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[49], water[50] or blat[51]. 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. In the preferred embodiment, the 95% confidence interval of the copy numbers of the candidate multi-copy 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 which indicates the credibility of the measured values of the measured parameters.

To search for the conservative regions.

To obtain species-specific consensus sequences of microorganisms as presented in Figure 1a, 1) for searching for candidate consensus sequences, specific sequences of target strains belonging to the same species were clustered based on the clustering algorithm[52] 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[53]. 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.

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

To design the best sets from the appropriate templates above.

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[54] 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.

Conclusions

Above all, the Shine strategy was presented to explore specific, sensitive and conserved biomarkers from massive microbial genomic data within intra-populations. 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 newly discovered multi-copy 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.

Declarations

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 material

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

Competing interests

We were funded by Liferiver Science and Technology Institute of Shanghai ZJ Bio-Tech Co, Ltd. The authors declare that they have no competing interests.

Author information

Affiliations

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

Cong Ji, Junbin (Jack) Shao

Authors' contributions

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

Corresponding author

Correspondence to Cong Ji and Junbin (Jack) Shao.

Acknowledgment

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.

Funding
Not applicable.

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. van Hattem, J.M. and B. de Wever, *16S rRNA sequence analysis: application and pitfalls.* Ned Tijdschr Geneeskd, 2019. **163**.
16. Antipov, D., et al., *Plasmid detection and assembly in genomic and metagenomic data sets.* Genome research, 2019. **29**(6): p. 961-968.
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. 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.
27. 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.
28. 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.
29. Shu, Y. and J. McCauley, *GISAID: Global initiative on sharing all influenza data - from vision to reality.* Euro Surveill, 2017. **22**(13).
30. 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.
31. Zhang, Y.Z., et al., *The diversity, evolution and origins of vertebrate RNA viruses.* Curr Opin Virol, 2018. **31**: p. 9-16.
32. 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.
33. 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.
34. El Moutaouakil, A., et al., *Review: Graphene-based biosensor for Viral Detection*. 2020.
35. Vermisoglou, E., et al., *Human virus detection with graphene-based materials.* Biosens Bioelectron, 2020. **166**: p. 112436.
36. Pena-Bahamonde, J., et al., *Recent advances in graphene-based biosensor technology with applications in life sciences.* J Nanobiotechnology, 2018. **16**(1): p. 75.

37. Das Jana, I., et al., *Development of a copper-graphene nanocomposite based transparent coating with antiviral activity against influenza virus.* BioRxiv, 2020.
38. 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.
39. 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.
40. Gorbalenya, A.E. and S.G. Siddell, *Recognizing species as a new focus of virus research.* PLoS pathogens, 2021. **17**(3): p. e1009318-e1009318.
41. Kitts, P.A., et al., *Assembly: a resource for assembled genomes at NCBI.* Nucleic Acids Res, 2016. **44**(D1): p. D73-80.
42. 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.
43. 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.
44. 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.
45. 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.
46. Basenko, E.Y., et al., *FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes.* J Fungi (Basel), 2018. **4**(1).
47. Davis, J.J., et al., *The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities.* Nucleic Acids Res, 2020. **48**(D1): p. D606-D612.
48. Stephen F.A, et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
49. 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.
50. Smith TF and W. MS, *Identification of common molecular subsequences.* Journal of Molecular Biology, 1981. **1**(147): p. 195-197.
51. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
52. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-1.
53. 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.
54. Untergasser, A., et al., *Primer3--new capabilities and interfaces.* Nucleic acids research, 2012. **40**(15): p. e115-e115.

Figures, Tables and Supplementary Information

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 was divided into two directions, 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 searching for the multicopy regions preferentially.

Table 1. Final sets of primers and probes by our strategy on different species or subspecies of coronaviridae. HKU1 represents Human coronavirus HKU1; OC43 represents Human coronavirus OC43; NL63 represents Human coronavirus NL63; 229E represents Human coronavirus 229E; MERS represents Middle East respiratory syndrome coronavirus; SARS-CoV represents severe acute respiratory syndrome coronavirus; SARS-CoV-2 represents severe acute respiratory syndrome coronavirus 2.

Table 2. The performance of our strategy on reducing false negative rate for sample sets of identified multi-copy regions compared with known 16S rRNA genes. Here the distribution of copy numbers of de novo biomarkers separately means Sample ID; Cluster ID; Mean copy numbers (Minimum copy numbers - Maximum copy numbers); The copy number corresponding to the i-th candidate species-specific consensus sequences (The number of strains with the i-th candidate species-specific consensus sequences); Markers; Conversation [summation of numbers of covered strains/ total number of target strains]; The weighted average copy numbers.

Table 3. The performance of our strategy on reducing false positive rate for sample sets of identified newly discovered species or subspecies specific multi-copy regions. Here the distribution of copy numbers of de novo biomarkers separately means Cluster ID; Distribution: The copy number corresponding to the i-th candidate species-specific consensus sequences (The number of strains with the i-th candidate species-specific consensus sequences); Markers; Percent of strains [summation of numbers of covered strains/ total number of target strains]; Weighted average copy numbers.

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

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

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

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

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

Supplementary Table6. Cluster sets of identified undiscovered 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.

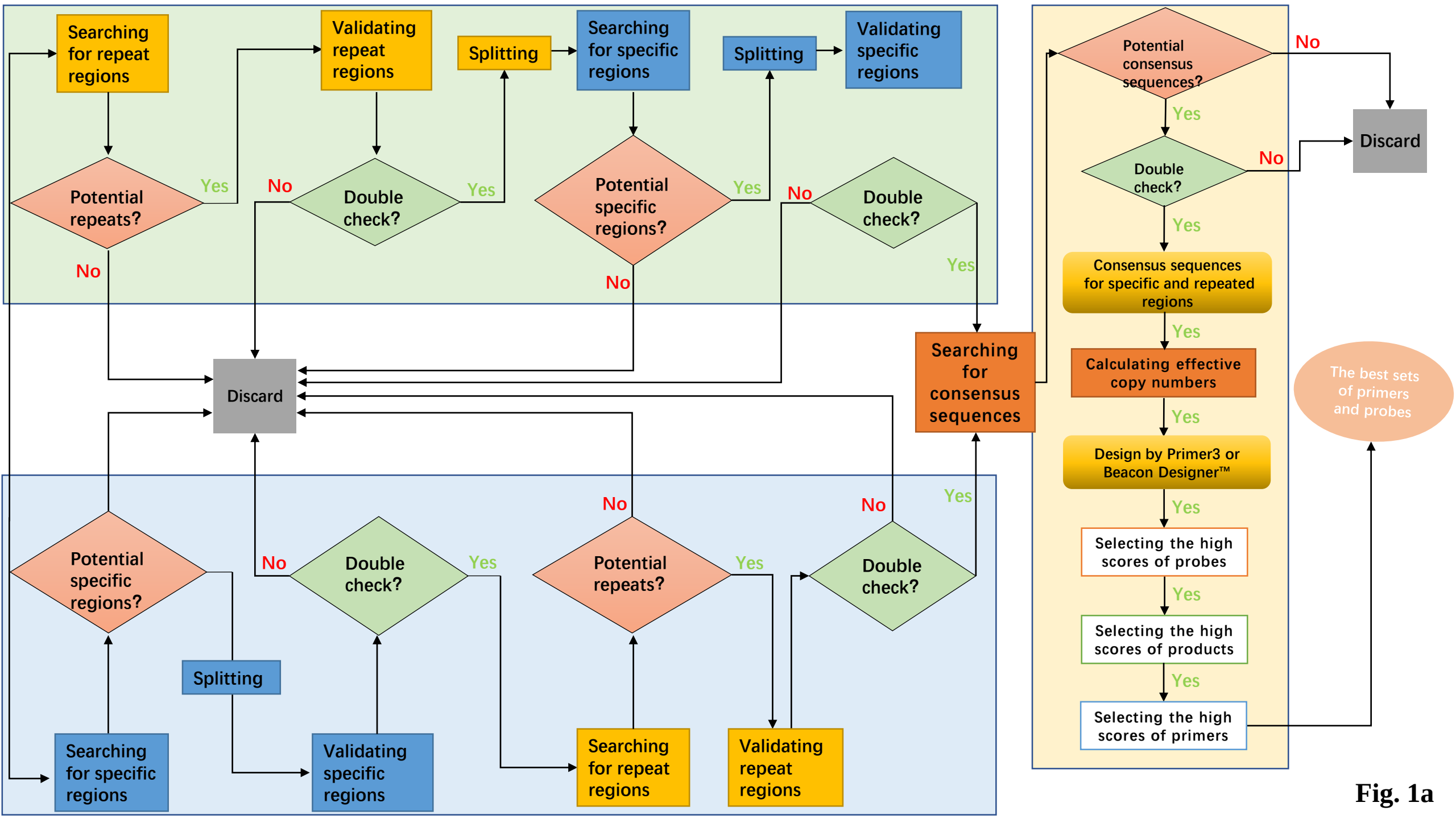

Searching for repeat regions
Validating repeat regions
Splitting
Searching for specific regions
Splitting
Validating specific regions
Potential repeats?
Yes
No
Double check?
No
Yes
Potential specific regions?
Yes
No
Double check?
No
Yes
Discard
Searching for consensus sequences
Potential consensus sequences?
No
Yes
Double check?
No
Yes
Discard
Consensus sequences for specific and repeated regions
Yes
Calculating effective copy numbers
Yes
Design by Primer3 or Beacon Designer™
Yes
Selecting the high scores of probes
Yes
Selecting the high scores of products
Yes
Selecting the high scores of primers
The best sets of primers and probes
Potential specific regions?
No
Double check?
Yes
Potential repeats?
No
Yes
Double check?
No
Yes
Splitting
Searching for specific regions
Validating specific regions
Searching for repeat regions
Validating repeat regions

Fig. 1a

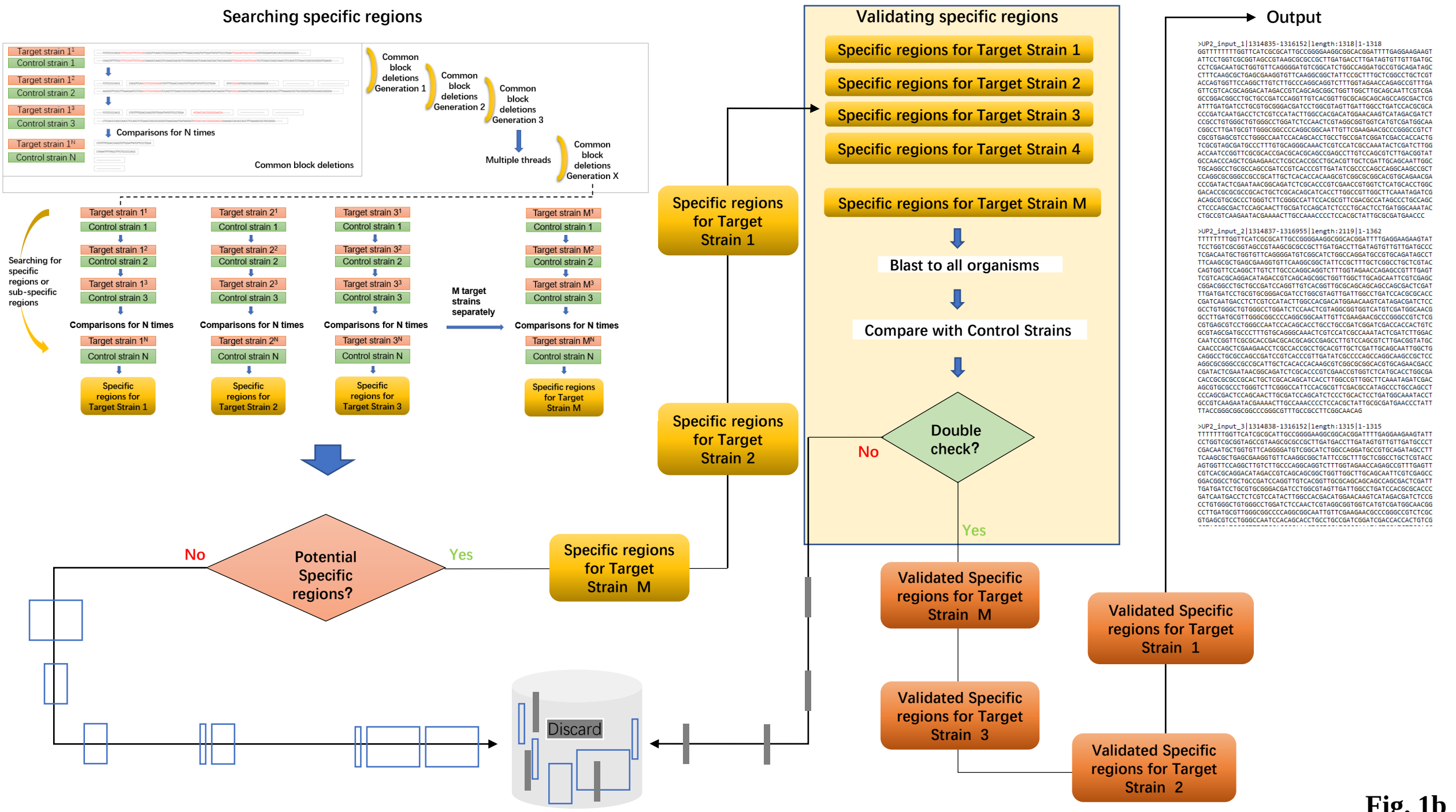

Searching specific regions
Target strain 1¹
Control strain 1
Target strain 1²
Control strain 2
Target strain 1³
Control strain 3
Target strain 1ᴺ
Control strain N
Comparisons for N times
Common block deletions
Common block deletions Generation 1
Common block deletions Generation 2
Common block deletions Generation 3
Multiple threads
Common block deletions Generation X
Searching for specific regions or sub-specific regions
Target strain 2¹
Target strain 2²
Target strain 2³
Target strain 2ᴺ
Target strain 3¹
Target strain 3²
Target strain 3³
Target strain 3ᴺ
Target strain M¹
Target strain M²
Target strain M³
Target strain Mᴺ
M target strains separately
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
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

Fig. 1b

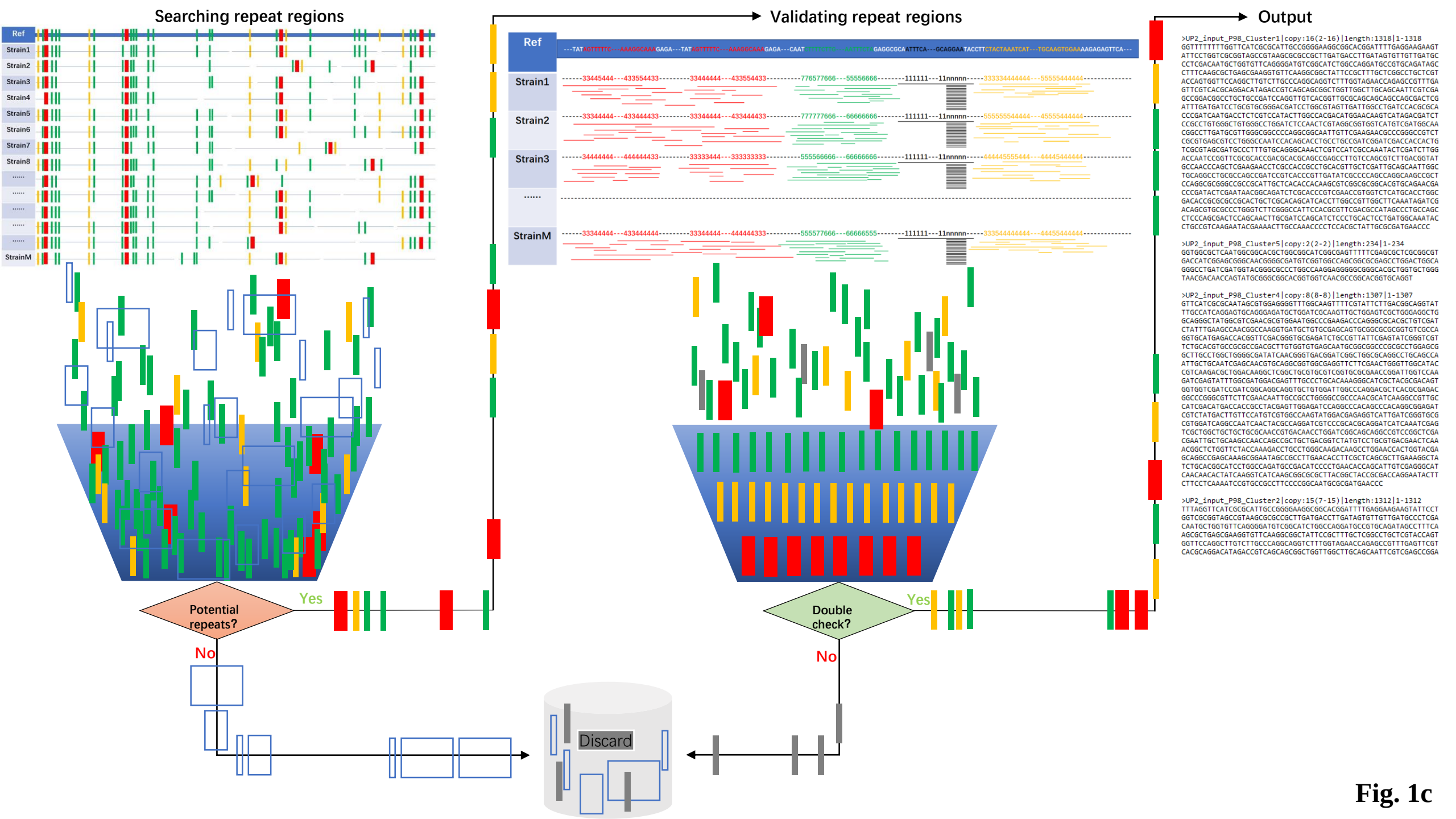

Searching repeat regions
Validating repeat regions
Output
Ref
Strain1
Strain2
Strain3
Strain4
Strain5
Strain6
Strain7
Strain8
StrainM
Potential repeats?
Yes
No
Double check?
Yes
No
Discard

Fig. 1c

Table 1. Final sets of primers and probes by our strategy on different species or subspecies of coronaviridae.

Species/ Subspecies	Gene	Forward Primer Sequence	Reference ID	Forward Primer Position	Forward Primer Conservation	Reverse Primer Sequence	Reference ID	Reverse Primer Position	Reverse Primer Conservation	Probe Sequence	Reference ID	Probe Position	Probe Conservation
HKU1	ORF1ab	CTGTGAATTAGAACTTATTTGTGA	DQ415907	5649	33.30%	CAGTCTTTGCTAATGTACCA	DQ415907	5743	100.00%	AACAGCATCAACACCAACACGACTT	DQ415907	5715	100.00%
OC43	S	TGGTGTTAATTTCAATGTAGA	KX344031	26278	70.60%	GCAGATCTACTGGARGC	KX344031	26367	98.20%	TACTACATTCGCTGCCTAGACAACCT	KX344031	26348	100%
NL63	replicase polyprotein 1ab	GCTGGTTTATGTGTTGTTTG	JQ765564	15138	81.70%	GTACCAAATACATGATCATATGC	JQ765564	15244	100.00%	TAGGCTTACGCAGACAATCACCACAA	JQ765564	15204	98.30%
229E	replicase polyprotein 1ab	TGGTGTTGTGTGYTAYA	KT253325	14653	80.00%	CACCATTCTGATAATATAAAGTRG	KT253325	14744	56.70%	ATCAGCAATGTAACCAAGTTCAGCAT	KT253325	14704	100%
MERS-CoV	S	GGCTAAAGATAAAGTCAAYGAG	MF598714	31517	98.70%	GGGCATTTACAACAAARG	MF598714	31614	98.50%	TCAAGGCACAATCCAAGCGTTCTG	MF598714	31543	98.70%
SARS-CoV	ORF1ab	CCTCAGGCTGAAGTAGA	AY394986	18941	100%	CCTCTATTTTGTAAGCTTTGTC	AY394986	19010	100%	CATGGCTGAGCATCGTAGAACTTCC	AY394986	18984	100%
COVID-19	ORF1ab	GTGTCAACCTATACTGTTACTA	EPI_ISL_402124	7959	100%	GCATCAAACATTTTAACTGC	EPI_ISL_402124	8043	100%	ACTTCCGCACTATCACCAACATCAGA	EPI_ISL_402124	8022	100%

Notes: HKU1 represents Human coronavirus HKU1; **OC43** represents Human coronavirus OC43; **NL63** represents Human coronavirus NL63; **229E** represents Human coronavirus 229E; **MERS** represents Middle East respiratory syndrome coronavirus; **SARS-CoV** represents severe acute respiratory syndrome coronavirus; **SARS-CoV-2** represents severe acute respiratory syndrome coronavirus 2.

Degenerate bases: R=A/G，Y=C/T，M=A/C，K=G/T，S=C/G，W= A/T，H= A/C/T，B= C/G/T，V=A/C/G， D=A/G /T,N=A/C/G/T.

Table 2. The performance of our strategy on reducing false negative rate for sample sets of identified multi-copy regions compared with known 16S rRNA genes.

Target Organisms	Known rRNA Genes	Known Copy Numbers	Alignment of sequences between De Novo Biomarkers and known rRNA Genes	Distribution of Copy Numbers of De Novo Biomarkers
Mycobacterium tuberculosis	IS6110	10,15	GCA_001652545.1 ----------tgaaccgccccggtgagtccggagactctctgatctgagacctca AJ242908.1 tggcgcgtgtagtagtgaaccgccccggtgagtccggagactctctgatctgagacctca ********** GCA_001652545.1 gccggcggctggtctctggctttgagcgtagtaggcagcctcgagttcgaccggcgggac AJ242908.1 gccggcggctggtctctggcgttgagcgtagtaggcagcctcgagttcgaccggcgggac ********** ********** GCA_001652545.1 gtcgccgcagtactggtagaggcggcgatggttgaaccagtcgacccagcgcgcggtggc AJ242908.1 gtcgccgcagtactggtagaggcggcgatggttgaaccagtcgacccagcgcgcggtggc ********** GCA_001652545.1 caactcgacatcctcgatggaccgccagggcttgccgggtttgatcagctcggtcttgta AJ242908.1 caactcgacatcctcgatggaccgccagggcttgccgggtttgatcagctcggtcttgta ********** GCA_001652545.1 taggccgttgatcgtctcggctagtgcattgtcataggagcttccgaccgctccgaccga AJ242908.1 taggccgttgatcgtctcggctagtgcattgtcataggagcttccgaccgctccgaccga ********** GCA_001652545.1 cggttggatgcctgcctcggcgagccgctcgctgaaccggatcgatgtgtactgagatcc AJ242908.1 cggttggatgcctgcctcggcgagccgctcgctgaaccggatcgatgtgtactgagatcc ********** GCA_001652545.1 cctatccgtatggtggataacgtctttcaggtcgagtacgccttcttgttggcgggtcca AJ242908.1 cctatccgtatggtggataacgtctttcaggtcgagtacgctttcttgttggcgggtcca **********.********** GCA_001652545.1 gatggcttgctcgatcgcgtcgaggaccatggaggtggccatcgtgga---------- AJ242908.1 gatggcttgctcgatcgcgtcgaggaccatggaggtggccatcgtggaagcgacccgcca **********	GCA_001652545.1_246_copy:8(8-10) 0(7)|1(63)|2(43)|3(23)|4(25)|5(4)|6(3)|7(6)|8(5)|9(11)|10(7)|11(11)|12(14)|13(17)|14(14)|15(23)|16(17)|17(13)|18(8)|19(4)|20(1)|21(9)|22(2)|23(3)|24(4)|25(2)| * 97.9%[332/339] **8.37463126843658**
Mycobacterium Africanum		20		
Mycobacterium bovis		1,20		

Mycobacterium tuberculosis	IS1081	5,6	GCA_000658535.1 ----------cgcgggccttaagccagctt X84741.1 ccgaaggatcacgcgaggaaccttcactcgtacaccacgtccctggccttggcctggtgt * * ******.. *..*. * GCA_000658535.1 taggcccagctggagccgacggcgctgtcggtttgtgccatgttgttgccggcagcctgc X84741.1 caggcccagctggagccgacggcgctgtcggtttgcgccatgttgttgccggcagcctgc .**********.********** GCA_000658535.1 accttctgcccgtgggcgttggcctgctcgtagatcacctggaagttacggcccagctgg X84741.1 accttctgcccgtgggcgttagcctg-tcgtagatcacctggaagttacggcccaactgg **********.***** **********.**** GCA_000658535.1 gtaatgaacccctggcaggccgccgaaccggcgccgccccaaaagtcactcgcggtcaac X84741.1 gtaatgaacccctggcaggccgccgaaccggcgccgccccaaaagtcactcgcggtcaac ********** GCA_000658535.1 acatcagaaatgatggcctgatgctcggcctccagcaacccggccagagcgcggatcatg X84741.1 acatcacgaatgatggcctgatgctcggcctccagcaacccggcctgagcgcggatcatg ****** .********** ********** GCA_000658535.1 gcgccat---------- X84741.1 gcgccatgatcgtcgacatccccgaattgatagttgatggtcatcgaacctgttctcctt *******	GCA_000658535.1_410_copy:5(2-5) 0(1)|1(2)|2(6)|3(28)|4(65)|5(236)|7(1)| * 99.7%[338/339] **4.55752212389381**
Mycobacterium Africanum		5,6		
Mycobacterium bovis		5,6		
Tuberculosis complex	PGRS	26,30	Cluster12423 ----------tgtt Z34263.1 atgcggaccacggtatctacagactcggtgaatccggactcagccggggttgccgttgtt **** Cluster12423 gcctttgccgccggtggcaccgtcgaggcccttggcgccgtcggcgccggtggcgccgtt Z34263.1 gcctttgccgccggtggcaccgtcgaggcccttggcgccgtcggcgccggtggcgccgtt ********** Cluster12423 gacgccgccgctaccgccggccccgcctttcccgccggccgcaccggtcccgccaacacc Z34263.1 gacgccgccgctaccgccggccccgcctttcccgccggccgcaccggtcccgccaacacc ********** Cluster12423 gccgtcaccgcccgtgccggcggcgccgggggtaccggcattgaaaatggccccaccgtt Z34263.1 gccgtcaccg-ccgtgccggcggcgccgggggtaccggcattgaaaatggccccaccgtt ********** ********** Cluster12423 gccg-ccgccgccgccggtcccgccagccccgccggcgccgccgacgcctccggcaccac Z34263.1 gccgcccgccgccgccggtcccgccagccccgccggcgccgccgacgcctccggcaccac **** ********** Cluster12423 cgggttggtcatcagggcctcaagcggat---------- Z34263.1 cggccccgcc-----ggccccaaactggccgaatatcccgcgcccaccgccgtccccgcc *** .. *.* ****.***.* *..	CP054014.1|3361564-3361841|2|copy:5(3-5)|length:278|1-278

Bordetella pertussis; Bordetella parapertussis	IS1002	6; 4--8; 9; 90	GCA_001509915.1 ---ctaggtcgagcccttcttgttaacagggtgatgcaacagaactggacgatgaggcag Z54268.1 ctcctaggtcgagcccttcttgttaacagggtgatgcaacagaactggacgatgaggcag ********** GCA_001509915.1 gccagccaggcggcaggggtgagcctgcgcaccggttacaagtggcttgcccgcttccga Z54268.1 gccagccaggcggcaggggtgagcctgcgcaccggttacaagtggcttgcccgcttccga ********** GCA_001509915.1 agcgaaggtctggacggcctgcttgaccgcagttctcgcccgcaccgcagcccgaaggcc Z54268.1 agcgaaggtctggacggcctgcttgaccgcagttctcgcccgcaccgcagcccgaaggcc ********** GCA_001509915.1 tgtgcgcccgagcaggttgagcacttcgcccagcagcgccgccagcgtctgccgctgtgg Z54268.1 tgtgcgcccgagcaggttgagcacttcgcccagcagcgccgccagcgtctgccgctgtgg ********** GCA_001509915.1 cgcatcgcgcgtgaagccggccgcagcctggccaccgtggcgcggtatatggaacgcatc Z54268.1 cgcatcgcgcgtgaagccggccgcagcctggccaccgtggcgcggtatatggaacgcatc ********** GCA_001509915.1 ggcctgagtcgactcgcctcgttggagccgccggcgccggtgcgccgctatgaacgcgcc Z54268.1 ggcctgagtcgactcgcctcgctggagccgccggcgccggtgcgccgctatg-acgcgcc **********.********** ******* GCA_001509915.1 agcccaggcgagctgctgcacatcgacaccaagcggctgggccgcatccggggcgtgggc Z54268.1 agcccaggcgagctgctgcacatcgacagcaagcggctgggccgcatccggggcgtgggc ********** ********** GCA_001509915.1 catcgcatcaccggcgatcgggcccagaaccgcaaccgcggcattggctgggacgctgtg Z54268.1 catcgcatcaccggcgatcgggcccagaaccgcaaccgcggcattggctgggacgctgtg ********** GCA_001509915.1 cacctggccatcgatgacttctcgcgcgtgtcctttgctcgcatcctcgacgatgaaggc Z54268.1 cacctggccatcgatgacttctcgcgcgtgtcctttgctcgcatcctcgacgatgagggc **********.*** GCA_001509915.1 ggcgatcagtgcgctgagttcctgcgccaagccaccgcctactacgccagtctgggcgtg Z54268.1 ggcgatcagtgcgctgagttcctgcaccaagccaccgcctactacgccagtctgggcgtg **********.********** GCA_001509915.1 cgcatcgaccgcgtgatgaccgacaacggcagcggctacgtctccaggacctttcgggcg Z54268.1 cgcatcgaccgcgtgatgaccgacaacggcaggggttacgtctccaggacctttcgggcg ********** **.********** GCA_001509915.1 gtgtgcgtggagctgggaatccgtcacatccgcacccgcccctacacccccaagaccaac Z54268.1 gtgtgcgtggagttgggaatccgacacatccgaacccgcccctacacctccaagaccaac **********.********** ******** **********.********** GCA_001509915.1 ggcaaggccgagcgcctcgtgcagacctgcttacgggagtgggcatacgccaggccctac Z54268.1 gtcaaggccgagcgcctcgtgaagacctgcttacgggagtgggcatacgccaggccctac * ********** ********** GCA_001509915.1 acgagctcagccgaacgacaggctgccttgcagcccttcattgaccgctacaactgg--- Z54268.1 acgagctcagacgaacgacaggctgccttgcagcccttcattgaccgctacaactggtgt ********** **********	GCA_001509915.1_74_copy:3(3-3) **0(179)|1(144)|2(15)|3(1)|** **47.2%[160/339] 0.52212389380531**

Bordetella pertussis; Bordetella holmesii	IS481	239; 100; >50; 8--10	AB670737.1 agcgccaacaccctgctgctggtgcagacgccactaggtgtgaagattcaataggttgta GCA_001985665.1 ----------ctaggtgtgaagattcaataggttgta ********** AB670737.1 tgcatggttcatccgaaccggatttgagaaactggaaatcgccaacccccccagttcactc GCA_001985665.1 tgcatggttcatccgaaccggatttgagaaactggaaatcgccaacccccccagttcactc ********** AB670737.1 aaggagcccggccggatgaacacccataagcatgcccgattgaccttcctacgtcgactc GCA_001985665.1 aaggagcccggccggatgaacacccataagcatgcccgattgaccttcctacgtcgactc ********** AB670737.1 gaaatggtccagcaattgatcgcccatcaagtttgtgtgcctgaagcggcccgcgcctat GCA_001985665.1 gaaatggtccagcaattgatcgcccatcaagtttgtgtgcctgaagcggcccgcgcctat ********** AB670737.1 ggggtcaccgcgccgactgtgcgcaaatggctgggccgcttcctggctcagggccaggcg GCA_001985665.1 ggggtcaccgcgccgactgtgcgcaaatggctgggccgcttcctggctcagggccaggcg ********** AB670737.1 ggcttggccgatgcgtcctcgcgcccgacggtctcgccccgagcgattgcgccggccaag GCA_001985665.1 ggcttggccgatgcgtcctcgcgcccgacggtctcgccccgagcgattgcgccggccaag ********** AB670737.1 gcgctggctatcgtggagctgcgccgcaagcggctgacccaagcgcgcatcgcccaggcg GCA_001985665.1 gcgctggctatcgtggagctgcgccgcaagcggctgacccaagcgcgcatcgcccaggcg ********** AB670737.1 ctgggcgtgtcagccagcaccgtcagccgcgtcctggcccgcgccggtctgtcgcacctg GCA_001985665.1 ctgggcgtgtcagccagcaccgtcagccgcgtcctggcccgcgccggtctgtcgcacctg ********** AB670737.1 gccgacctggagccggccgagccggtggtgcgctacgagcatcaggcccccggcgatctg GCA_001985665.1 gccgacctggagccggccgagccggtggtgcgctacgagcatcaggcccccggcgatctg ********** AB670737.1 ctgcacatcgacatcaagaagctgggacgtatccagcgccctggtcaccgggtcacgggc GCA_001985665.1 ctgcacatcgacatcaagaagctgggacgtatccagcgccctggccaccgggtcacgggc **********.********** AB670737.1 aaccgacgcgataccgttgagggggccggctgggacttcgtcttcgtggccatcgatgac GCA_001985665.1 aaccgacgcgataccgttgagggggccggctgggacttcgtcttcgtggccatcgatgac ********** AB670737.1 cacgcccgcgtggccttcaccgacatccaccccgacgagcgcttccccagcgccgtccag GCA_001985665.1 cacgcccgcgtggccttcaccgacatccaccccgacgagcgcttccccagcgccgtccag ********** AB670737.1 ttcctcaaggacgcagtggcctactaccagcgcctgggcgtgaccatccagcgcttgctc GCA_001985665.1 ttcctcaaggacgcagtggcctactaccagcgcctgggcgtgaccatccagcgcttgctc ********** AB670737.1 accgacaatggctcggcctttcgcagccgcgccttcgccgcgctgtgccatgagctgggc GCA_001985665.1 accgacaatggctcggcctttcgcagccgcgccttcgccgcgctgtgccatgagctgggc ********** AB670737.1 atcaagcaccgctttacccgaccttaccgcccacagaccaatggcaaggccgaacgcttc GCA_001985665.1 atcaagcaccgctttacccgaccttaccgcccacagaccaatggcaaggccgaacgcttc ********** AB670737.1 atccagtcggccttgcgtgagtgggcttacgctcacacctaccagaactcccaacaccga GCA_001985665.1 atccagtcggccttgcgtgagtgggcttacgctcacacctaccagaactcccaacaccga ********** AB670737.1 gccgatgccatgaaatcctggctacaccactacaactggcatcgaccccaccaaggcatc GCA_001985665.1 gccgatgccatgaaatcctggctacaccactacaactggcatcgaccccaccaaggcatc ********** AB670737.1 gggcgcgctgtacccatctccagactcaacctggacgaatacaacctattgacagttcac GCA_001985665.1 gggcgcgctgtacccatctccagactcaacctggacgaatacaacctattgacagttcac ********** AB670737.1 aactaggcagcgcggcgacctttacccttgccaacaaggacggcaaggtcgatatcggta GCA_001985665.1 acctagcccgcg---------- * **** * ***	GCA_001985665.1_53_copy:26(2-26) **0(23)|1(229)|2(84)|3(1)|4(2)| *** **93.2%[316/339] 1.20353982300885**

Bordetella pertussis	IS1663	17	NZ_CP025371.1:c ------gtggctgatagcagtttgttatctgcaggaggccaagtccatgttttcataggt GCA_001509915.1 ggcgcagtggctgacagcagtttgttatctgcaggaggccaagtccatgttttcataggt ********.********** NZ_CP025371.1:c attgatgtatcaaaggccaagctggattgcacgttgctgacggctgaggctgacaagtgt GCA_001509915.1 attgatgtatcaaaggccaagctggattgcacgttgctgacggctgaggctgacaagtgt ********** NZ_CP025371.1:c aagaccaaggtggtggtcaacaccgccgcgggcgtacaagccctgctggcgtggtgcgcc GCA_001509915.1 aagaccaaggtggtggtcaacaccgccgcgggcgtacaagccctgctggcgtggtgcgcc ********** NZ_CP025371.1:c aaacacggcgcacagccggcgcagttgcacgccatcctggagcctacgggtctgtatcac GCA_001509915.1 aaacacggcgcacagccggcgcagttgcacgccatcctggagcctacgggtctgtatcac ********** NZ_CP025371.1:c gagcaagcggccacggcgctgccccaggtccgggtctctttggtcaatcccgcccaggcc GCA_001509915.1 gagcaagcggccacggcgctgccccaggtccgggtctctttggtcaatcccgcccaggcc ********** NZ_CP025371.1:c cgggactttgccaaggccttggcgctgcgctccaaaaatgatgcgctcgacagctacgtg GCA_001509915.1 cgggactttgccaaggccttggcgctgcgctccaaaaatgatgcgctcgacagctacgtg ********** NZ_CP025371.1:c ctggctcgctatggacagacgctgagcccggcgctgtggcacccggcgcccttgcatgcg GCA_001509915.1 ctggctcgctatggacagacgctgagcccggcgctgtggcacccggcgcccttgcatgcg ********** NZ_CP025371.1:c cgtcaactgcgcgccttgttgacgcgacgcgaggcgctgagcaaggatctgttgcgtgag GCA_001509915.1 cgtcaactgcgcgccttgttgacgcgacgcgaggcgctgagcaaggatctgttgcgtgag ********** NZ_CP025371.1:c ctcaatcgcaaagagaagagccagttcagcccctcggcgcccttggtcgatggttccatc GCA_001509915.1 ctcaatcgcaaagagaagagccagttcagcccctcggcgcccttggtcgatggttccatc ********** NZ_CP025371.1:c gacaaggccatcgcgttcttgcgcgaacagatcaaacaaatcgagcgggcgatcgatcag GCA_001509915.1 gacaaggccatcgcgttcttgcgcgaacagatcaaacaaatcgagcgggcgatcgatcag ********** NZ_CP025371.1:c cacatcgacaaccaccccgacctcaagcaagactgcgagctgctgaactccatccccgcc GCA_001509915.1 cacatcgacaaccaccccgacctcaagcaagactgcgagctgctgaactccatccccgcc ********** NZ_CP025371.1:c atcgggcctcaggccggcaacgccatcctggccgtcatgcacaatcggcatatcgactcc GCA_001509915.1 atcgggcctcaggccggcaacgccatcctggccgtcatgcacaatcggcatatcgactcc ********** NZ_CP025371.1:c gcccagagcctggccgcctatctcggggtggtccctgtgcagcgccaatccggcagcagt GCA_001509915.1 gcccagagcctggccgcctatctcggggtggtccctgtgcagcgccaatccggcagcagt ********** NZ_CP025371.1:c ctgaacagctgcgcacgcctgtccaaagccggcccctcccaggtgcgcgccacgttatac GCA_001509915.1 ctgaacagctgcgcacgcctgtccaaagccggcccctcccaggtgcgcgccacgttatac ********** NZ_CP025371.1:c atggcggccctggttgggacccgccacaacccccacatccgcgccctttaccagcgcctg GCA_001509915.1 atggcggccctggttgggacccgccacaacccccacatccgcgccctttaccagcgcctg ********** NZ_CP025371.1:c ctcaaagcaggaaaaagcaaaaaggccgcgctgggcgcggccatgagaaaacaggtgcat GCA_001509915.1 ctcaaagcaggaaaaagcaaaaaggccgcgctgggcgcggccatgagaaaactggtgcat ********** ******* NZ_CP025371.1:c ctgtgcttcggggtcctcaaaaaccgcatcccctaccagcccaattacgccatgaatggt GCA_001509915.1 ctgtgcttcggggtcctcaaaaaccgcatcccctaccagcccaattacgccatgaacggt **********.*** NZ_CP025371.1:c tga---------- GCA_001509915.1 tgactgccaagacggtatctgtccccgcgggga ***	GCA_001509915.1_17_copy:5(2-6) 0(1)|15(1)|16(82)| * 98.8%[83/84] **15.797619047619**

Bordetella parapertussis; Bordetella bronchiseptica	IS1001	20; 22 1--7; 0;	X66858.1 ----------ggttcatcgcgcaataacgtggagggggtttggcaa-ttttcgtattct GCA_004008355_0 aggcaaaaaaaaggttcatcgcgcaatagcgtggagggggtttggcaagttttcgtattct ************************************* X66858.1 tgacggcaggtatttgacatcaggagtgcagggagatgctggatcgcaagttgatggagt GCA_004008355_0 tgacggcaggtatttgccatcaggagtgcagggagatgctggatcgcaagttgctggagt ********** ********** ****** X66858.1 cgctgggaggctggcagggctatggcgtcgaacgcgtggaatggcccgaagacccagggc GCA_004008355_0 cgctgggaggctggcagggctatggcgtcgaacgcgtggaatggcccgaagacccagggc ********** X66858.1 gcacgctgtcgatctatttgaagccaacggccaaggtgatgctgtgcgagcagtgcggcg GCA_004008355_0 gcacgctgtcgatctatttgaagccaacggccaaggtgatgctgtgcgagcagtgcggcg ********** X66858.1 cgcggtgtcgccaggtgcatgagaccacggttcgacgggtgcgagatctgccgtattcg GCA_004008355_0 cgcggtgtcgccaggtgcatgagaccacggttcgacgggtgcgagatctgccgttattcg ********** ****** X66858.1 agtatcgggtcgttctgcacgtgccgcgccgacgcttgtggtgtgagcaatgcggcggcc GCA_004008355_0 agtatcgggtcgttctgcacgtgccgcgccgacgcttgtggtgtgagcaatgcggcggcc ********** X66858.1 cgcgcctggagcggcttgcctggctggggcgatatcaacgggtgacggatcggctggcgc GCA_004008355_0 cgcgcctggagcggcttgcctggctggggcgatatcaacgggtgacggatcggctggcgc ********** X66858.1 aggcctgcagccaattgctgcaatcgagcaacgtgcaggcggtggcgaggttcttcgaac GCA_004008355_0 aggcctgcagccaattgctgcaatcgagcaacgtgcaggcggtggcgaggttcttcgaac ********** X66858.1 tgggttggcataccgtcaagacgctggacaaggctcggctgcgtgcgtcggtgcgcgaac GCA_004008355_0 tgggttggcataccgtcaagacgctggacaaggctcggctgcgtgcgtcggtgcgcgaac ********** X66858.1 cggattggtccaagatcgagtatttggcgatggacgagtttgccctgcacaaagggcatc GCA_004008355_0 cggattggtccaagatcgagtatttggcgatggacgagtttgccctgcacaaagggcatc ********** X66858.1 gctacgcgacagtggtggtcgatccgatcggcaggcaggtgctgtggattggcccaggac GCA_004008355_0 gctacgcgacagtggtggtcgatccgatcggcaggcaggtgctgtggattggcccaggac ********** X66858.1 gctcacgcgagacggcccgggcgttcttcgaacaattgccgcctggggccgcccaacgca GCA_004008355_0 gctcacgcgagacggcccgggcgttcttcgaacaattgccgcctggggccgcccaacgca ********** X66858.1 tcaaggccgttgccatcgacatgaccaccgcctacgagttggagatccaggcccacagcc GCA_004008355_0 tcaaggccgttgccatcgacatgaccaccgcctacgagttggagatccaggcccacagcc ********** X66858.1 cacaggcggagatcgtctatgacttgttccatgtcgtggccaagtatggacgagaggtca GCA_004008355_0 cacaggcggagatcgtctatgacttgttccatgtcgtggccaagtatggacgagaggtca ********** X66858.1 ttgatcgggtgcgcgtggatcaggccaatcaactacgccaggatcgtcccgcacgcagga GCA_004008355_0 ttgatcgggtgcgcgtggatcaggccaatcaactacgccaggatcgtcccgcacgcagga ********** X66858.1 ttgatcgggtgcgcgtggatcaggccaatcaactacgccaggatcgtcccgcacgcagga GCA_004008355_0 ttgatcgggtgcgcgtggatcaggccaatcaactacgccaggatcgtcccgcacgcagga ********** X66858.1 tcatcaaatcgagtcgctggctgctgctgcgcaaccgtgacaacctggatcggcagcagg GCA_004008355_0 tcatcaaatcgagtcgctggctgctgctgcgcaaccgtgacaacctggatcggcagcagg ********** X66858.1 ccgtccggctcgacgaattgctgcaagccaaccagccgctgctgacggtctatgtcctgc GCA_004008355_0 ccgtccggctcgacgaattgctgcaagccaaccagccgctgctgacggtctatgtcctgc ********** X66858.1 gtgacgaactcaaacggctctggttctaccaaagacctgcctgggcaagacaagcctgga GCA_004008355_0 gtgacgaactcaaacggctctggttctaccaaagacctgcctgggcaagacaagcctgga ********** X66858.1 accactggtacgagcaggccgagcaaagcggaatagccgcctttgaacaccttcgctcagc GCA_004008355_0 accactggtacgagcaggccgagcaaagcggaatagccgcctttgaacaccttcgctcagc ********** X66858.1 gcttgaaaggctatctgcacggcatcctggccagatgccgacatcccctgaacaccagca GCA_004008355_0 gcttgaaaggctatctgcacggcatcctggccagatgccgacatcccctgaacaccagca ********** X66858.1 ttgtcgagggcatcaacaacactatcaaggtcatcaagcggcgcgcttacggctaccgcg GCA_004008355_0 ttgtcgagggcatcaacaacactatcaaggtcatcaagcggcgcgcttacggctaccgcg ********** X66858.1 accaggaatacttcttcctcaaaatccgtgccgccttccccggcaatgcgcgatgaacc- GCA_004008355_0 accaggaatacttcttcctcaaaatccgtgccgccttccccggcaatgcgcgatgaaccc ********** X66858.1 ------ GCA_004008355_0 aaaaaa	GCA_004008355_0_copy:8(2-8) 22(21)|23(1)|24(1)| * 100.0%[23/23] **22.1304347826087**

Mycoplasma pneumoniae	RepMP 2/3	10	FJ604100.1 gctgatcacccccgggttggcttgaacgccgcaagacgttggtaacatcgttgtcagtgg GCA_002090215.1 --tgatcacccccgggttggcttgaacgccgcaagacgttggtaacatcgttgtcagtgg ********** FJ604100.1 caccagcttcagcttccagctcggcgggtggttagttacgttcacggactttatcaaacc GCA_002090215.1 caccagcttcagcttccagctcggcgggtggttagttacgttcacggactttatcaaacc ********** FJ604100.1 ccgcgctggttacctcgggctccagttaacgggtctcgatgtaagtgaagcgacccaaag GCA_002090215.1 ccgcgctggttacctcgggctccagttaacgggtctcgatgtaagtgaagcgacccaaag ********** FJ604100.1 ggagttaatttgggccaagcggccctgagcggcctttcgtggcagttgggtcaaccggct GCA_002090215.1 ggagttaatttgggccaagcggccctgagcggcctttcgtggcagttgggtcaaccggct ********** FJ604100.1 gggccgcgtggagagtgtgtgggatttcaagggggtgtgggcggatcaagctcagttggc GCA_002090215.1 gggccgcgtggagagtgtgtgggatttcaagggggtgtgggcggatcaagctcagttggc ********** FJ604100.1 cgcgcaagcagctacaagtagtaccaccaccaccgcaacaggggctaccttaccggagca GCA_002090215.1 cgcgcaagcagctacaagtagtaccaccaccaccgcaacaggggctaccttaccggagca ********** FJ604100.1 cccgaatgccctcgcgtaccaaattagctataccgacaaggattcgtacaaggcttccac GCA_002090215.1 cccgaatgccctcgcgtaccaaattagctataccgacaaggattcgtacaaggcttccac ********** FJ604100.1 tcaaggttcgggtcaaaccaattcccaaaacaattcgccctacctccattttattaaacc GCA_002090215.1 tcaaggttcgggtcaaaccaattcccaaaacaattcgccctacctccattttattaaacc ********** FJ604100.1 taagaaagtcgaaagcacgacccaactcgaccagggcttaaaaaacctgttggaccccaa GCA_002090215.1 taagaaagtcgaaagcacgacccaactcgaccagggcttaaaaaacctgttggaccccaa ********** FJ604100.1 ccaggttcgcaccaagctgcgccaaagctttggtacagaccattccacccagccccagcc GCA_002090215.1 ccaggttcgcaccaagctgcgccaaagctttggtacagaccattccacccagccccagcc ********** FJ604100.1 ccaatcgctcaaaacaacgacaccggtatttgggaggagtagtggtaacctcagtagtgt GCA_002090215.1 ccaatcgctcaaaacaacgacaccggtatttgggaggagtagtggtaacctcagtagtgt ********** FJ604100.1 gtttagtggtgggggtgctggagggggttcttcaggctcaggtcaa GCA_002090215.1 gtttagtggtgggggtgctggagg---------- **********	GCA_002090215.1_89_copy:6(3-7) **0(1)|1(9)| * 90.0%[9/10] 0.9**

Mycoplasma pneumoniae	RepMP 4	8	MF668982.1 aaaaatgatctcgccaacgctcccattaagcggagcgaggagtcgggtcagtccgtccaa GCA_002128125.1 ----atgatctcgccaacgttcccattaagcggggcgaggagtcgggtcagtccgtccaa **********.**********.********** MF668982.1 ctcaaggcggacgattttggtactgccctttccagttcgggatcaggcggcaactccaat GCA_002128125.1 ctcaaggcagccgacttcggcaccaccccatccagttcgggatcaggcggcaactccaac ********.* ***.**.**.**..***. **********. MF668982.1 cccggttcccccaccccctgaaggccgtggcttgcgactgagcaaattcacaaggacctc GCA_002128125.1 cccggttcccccaccccctgaaggccgtggcttgcgaccgagcaaattcacaaggacctc **********.********** MF668982.1 cccaaatgatccgcctcgatcctgattctgtacgatgcgccttatgcgcgcaaccgtacc GCA_002128125.1 cccaaatgatccgcctcgatcctgattctgtacgatgcgccttatgcgcgcaatcgtacc **********.****** MF668982.1 gccattgaccgcgttgatcacttggatcccaaggccatgaccgcgaactatccgcccagt GCA_002128125.1 gtcattgatcgcgttgatcacttggatcccaaggtgatgaccgcgaactatccgcccagt *.******.**********. ********** MF668982.1 tgaagaacgcccaagtgaaaccaccacggtttgtgggactgaaaggcgcgcgatgttttg GCA_002128125.1 tgaagaacgcccaagtgaaaccaccacgggttgtgggactgaaaggccagagatgttttg ********** ********** * ********* MF668982.1 ctccaaaccaccgggttcttcaacccgcgccgccaccccgagtggtttgatggcgggcag GCA_002128125.1 ctccaaaccaccgggttcttcaacccgcggcgccaccccgagtggtttgaccagggccaa ********** **********. . ** **. MF668982.1 acggtcgcggataacgaaaagaccgggtttgatgtggataactctgaaaacaccaagcag GCA_002128125.1 gcggtcgcggataatacccaaaccgggtttgatacagatgacaccgataataaaaaacaa .**********.. *.**********...***.** *.** **.* **.**. MF668982.1 ggctttcaaaaggaagctgactccgacaagtcggccccgatcgccctcccgtttgaagcg GCA_002128125.1 ggctttcaaaaggaagctgactccgacaagccggccccgatcgccctcccgtttgaagcg **********.********** MF668982.1 tacttcgccaacattggcaacctcacctggttcgggcaagcgcttttggtgtttggtggc GCA_002128125.1 tacttcgccaacattggcaacctcacctggttcgggcaagcgcttttggtgtttgg---- **********	GCA_002128125.1_490_copy:6(2-7) **1(8)|2(2)| * 100.0%[10/10] 1.2**

Mycoplasma pneumoniae	RepMP 5	7，9	HQ698831.1 aagagagcgaaagaagcagtatgccttctatcgcgattatctgctcaaccaagaaaatat GCA_002090215.1 aagagagcgaaagaagcagtatgccttctatcgcgattatctgctcaaccaagaaaatat ********** HQ698831.1 ccgaaaaatttacggagcaaatatacctttcgaaacttttcaagtaaaggatatttgtga GCA_002090215.1 ccgaaaaatttacggagcaaatatacctttcgaaacttttcaagtaaaggatatttgtga ********** HQ698831.1 aattagaagaggtcgagcgatcactaaagcgtatattaggaataatcctggagaaaatcc GCA_002090215.1 aattagaagaggtcgagcgatcactaaagcgtatattaggaataatcctggagaaaatcc ********** HQ698831.1 agtttattccgcagctaccacaaatgatggtgagttaggacgtattaaagattgcgattt GCA_002090215.1 agtttattccgcagctaccacaaatgatggtgagttaggacgtattaaagattgcgattt ********** HQ698831.1 tgatggtgaatacatcacttgaactacaaatggttatgctggagtagtttttatcgtaa GCA_002090215.1 tgatggtgaatacatcacttgaactacaaatggttatgctggagtagtttttatcgtaa ********** HQ698831.1 tggcaagtttaatgcatcacaagattgcggtgttttaaaggttaaaaataaaaaaatctg GCA_002090215.1 tggcaagtttaatgcatcacaagattgcggtgttttaaaggttaaaaataaaaaaatctg ********** HQ698831.1 caccaaattcctatccttttttattaaagatagaagcaccaaaattcgttcataatctagc GCA_002090215.1 caccaaattcctatccttttttattaaagatagaagcaccaaaattcgttcataatctagc ********** HQ698831.1 ttcaagaccaaaactaagccaaaaagtaatggctgaaatcgaactctcttttccacctct GCA_002090215.1 ttcaagaccaaaactaagccaaaaagtaatggctgaaatcgaactctcttttccacctct ********** HQ698831.1 agaaattcaagaaaagattgctgatattctctttgcctttgaaaaactttgcaatgattt GCA_002090215.1 agaaattcaagaaaagattgctgatattctctttgcctttgaaaaactttgcaatgattt ********** HQ698831.1 agtagaaggtattcctgctgaagtcgaaatgcgtaaaaagcaattggattactatcagaa GCA_002090215.1 agtagaaggtattcctgctgaagtcgaaatgcgtaaaaagcaattggattactatcagaa ********** HQ698831.1 cttcttatttaattgagtgcaagaacaaaaaactcaattagagcaaattatgtaatgctt GCA_002090215.1 cttcttatttaattgagtgcaagaacaaaaaactca---------- **********	GCA_002090215.1_316_copy:4(2-4) 4(10)| * 100.0%[10/10] **4**

Mycoplasma pneumoniae	RepMP 1	14	X13093.1 cagttcggacgactgctgcaagtagtagttttctcaacctctttcagaaaggagttattg GCA_002090215.1 --gttggcagcactgctgcaagtagtagttttctcaacctctttcagaaaggagttattg *** * * ********** X13093.1 cttaatgaaagaaaaaactccctttaaaaacgaaaaagaattccatgacatggtgaaaaa GCA_002090215.1 cttaatgaaagaaaaaactccctttaaaaacgaaaaagaattccatgacatggtgaaaaa ********** X13093.1 gaccaaaaagggcacttttcaggttgatacatcgttaatccagagaacaacagtgtgga GCA_002090215.1 gaccaaaaagggcacttttcaggttgatacatcgttaatccagagaacaacagtgtgga ********** X13093.1 attctctggcaatttcaatagacaatttaaacttaacaaannnnttataccagtgaacac GCA_002090215.1 attctctggcaatttcaatagacaatttaaacttaacaaaccggttataccagtgaacac ********** ********** X13093.1 ggaatacgtaacgcgtaaagagttcaacgagtacaaagactcaaacgaccaacgactcat GCA_002090215.1 ggaatacgtaacgcgtaaagagttcaacgagtacaaagactcaaacgaccaacgactcat ********** X13093.1 aaagattgaaaccactttggctgcccaaggtgaacaaataagaatccagggtgagcaaat GCA_002090215.1 aaagattgaaa---------- **********	GCA_002090215.1_182_copy:8(3-8) 10(2)|11(8)| * 100.0%[10/10] **10.8**
Streptococcus agalactiae	GBSi1	2	AJ292930.1 ggttcagacccaatcgctcctgcgaaacagccattgttcagctacttgagtatttaaacg GCA_002073795.1 ----------gctacttgagtatttaaacg ********** AJ292930.1 atggctacgagtggattgtggacattgacttggaaaagttcttcgatactgttccgcaag GCA_002073795.1 atggctacgagtggattgtggacattgacttggaaaagttcttcgatactgttccgcaag ********** AJ292930.1 acagattgatgtccctggttcataatatcattcaagatggcgatacggagtcactgattc GCA_002073795.1 acagattgatgtccctggttcataatatcattcaagatggcgatacggagtcactgattc ********** AJ292930.1 gtaagtacttccattcgggagttgttattaacggacagcgacataagactttagtcggga GCA_002073795.1 gtaagtacttccattcgggagttgttattaacggacagcgacataagactttagtcggga ********** AJ292930.1 cacctcaaggcgggaatctatcacccctcctatctaatattatgcttaatgagttagaca GCA_002073795.1 cacctcaaggcgggaatctatcacccctcctatctaatattatgcttaatgagttagaca ********** AJ292930.1 aaggggttggaaaagcgaggtcttcgctttgtccgttacgccgatgactgtgtcatcactg GCA_002073795.1 aaggggttggaaaagcgaggtcttcgctttgtccgttacgccgatgactgtgtcatcactg ********** AJ292930.1 tcggaagcgaagcagctgctaagcgggtcatgcattcggtcagtagctatattgagaagc GCA_002073795.1 tcggaagcgaagcagctgctaagcgggtcatgcattcggtcagtagctatattgagaagc ********** AJ292930.1 gattagggttgaaagtcaacatgactaagaccaagattgtcagaccgaacaaactcaaat GCA_002073795.1 gattagggttgaaagtcaacatgactaagac---------- **********	GCA_002073795.1_71_copy:3(3-3) 0(5)|3(1)|20(1)| 28.6%[2/7] **3.28571428571429**

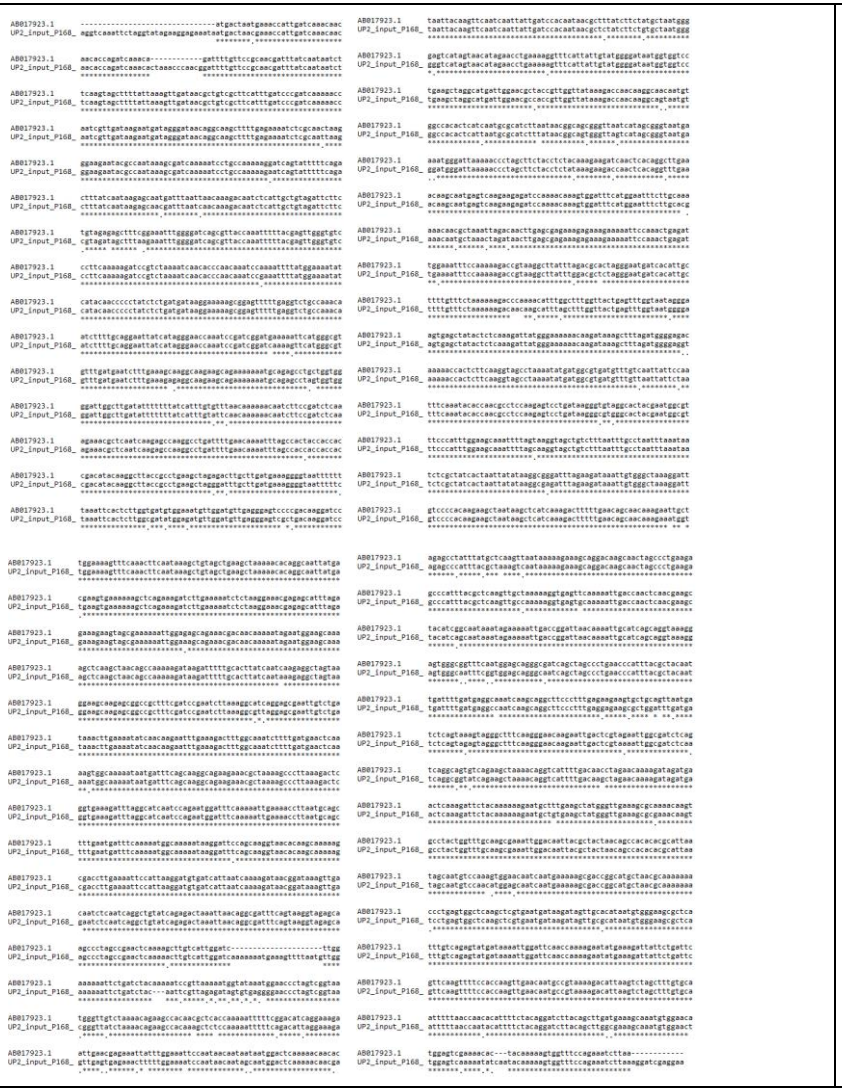
Helicobacter pylori	cagA	1.4-3.7		FR666836.1_ copy:2(2-2)

Table 3. The performance of our strategy on reducing false positive rate for sample sets of identified newly discovered species or subspecies specific multi-copy regions.

Target Organisms	Distribution of Copy Numbers of De Novo Biomarkers	Sequences of identified newly discovered species or subspecies specific multi-copy regions.
Mycobacterium tuberculosis		

ClusterID	Distribution	Marker	Percent of strains	Weighted Average Copy
Cluster198	0:3|1:51|2:40|3:35|4:21|5:4|6:3|7:5|8:5|9:8|10:7|11:11|12:10|13:21|14:18|15:20|16:24|17:13|18:6|19:3|20:2|21:7|22:3|23:2|24:2|25:2|27:1|28:1|	99.1% [336/339]	8.831858407
Cluster102	0:3|1:51|2:40|3:35|4:21|5:4|6:3|7:5|8:5|9:8|10:7|11:11|12:10|13:21|14:18|15:21|16:23|17:13|18:6|19:3|20:2|21:7|22:4|23:2|24:2|25:2|27:1|28:1|	99.1% [336/339]	8.829858702
Cluster454	0:3|1:51|2:40|3:35|4:21|5:4|6:3|7:6|8:5|9:8|10:7|11:11|12:10|13:21|14:18|15:21|16:23|17:12|18:7|19:3|20:2|21:7|22:3|23:2|24:2|25:2|27:1|28:1|	99.1% [336/339]	8.817109145
Cluster170	0:3|1:52|2:41|3:34|4:21|5:4|6:3|7:6|8:5|9:8|10:7|11:11|12:17|13:20|14:18|15:21|16:24|17:11|18:7|19:3|20:2|21:7|22:3|23:2|24:2|25:2|27:1|28:1|	99.1% [336/339]	8.808258587
Cluster439	0:3|1:52|2:43|3:33|4:20|5:4|6:3|7:7|8:6|9:8|10:9|11:11|12:17|13:20|14:18|15:20|16:25|17:11|18:6|19:4|20:2|21:7|22:3|23:2|24:2|25:2|27:1|28:1|	99.1% [336/339]	8.784217769
Cluster138	0:3|1:52|2:40|3:36|4:20|5:4|6:3|7:7|8:5|9:8|10:9|11:13|12:18|13:22|14:18|15:21|16:20|17:14|18:9|19:3|20:2|21:8|22:3|23:1|24:2|25:2|27:1|28:1|	99.1% [336/339]	8.74262537
Cluster363	0:3|1:54|2:41|3:33|4:20|5:4|6:4|7:6|8:6|9:8|10:9|11:11|12:18|13:20|14:18|15:22|16:23|17:12|18:9|19:4|20:2|21:7|22:3|23:1|24:4|25:1|27:1|28:1|	99.1% [336/339]	8.737463127
Cluster289	0:3|1:54|2:41|3:32|4:21|5:4|6:4|7:6|8:6|9:8|10:9|11:12|12:18|13:22|14:17|15:22|16:23|17:12|18:9|19:4|20:2|21:7|22:3|23:2|24:2|25:1|27:1|28:1|	99.1% [336/339]	8.734512174
Cluster76	0:3|1:52|2:43|3:33|4:20|5:4|6:3|7:7|8:6|9:8|10:10|11:11|12:17|13:21|14:19|15:23|16:20|17:12|18:8|19:3|20:2|21:7|22:4|23:1|24:3|25:1|27:1|28:1|	99.1% [336/339]	8.713864307
Cluster599	0:3|1:50|2:42|3:32|4:22|5:4|6:3|7:8|8:3|9:7|10:9|11:11|12:14|13:19|14:17|15:23|16:24|17:11|18:9|19:4|20:2|21:8|22:2|23:1|24:4|25:1|28:1|27:1|29	99.1% [336/339]	8.613588222
Cluster632	0:3|1:50|2:40|3:22|4:25|5:7|6:3|7:8|8:3|9:14|10:9|11:11|12:18|13:17|14:14|15:24|16:17|17:13|18:8|19:4|20:1|21:9|22:2|23:2|24:3|25:3|28:1|	99.1% [336/339]	8.592620084
Cluster160	0:3|1:50|2:40|3:22|4:26|5:6|6:1|7:8|8:4|9:13|10:9|11:11|12:18|13:17|14:18|15:23|16:17|17:14|18:8|19:4|20:1|21:10|22:2|23:2|24:3|25:3|28:1|	99.1% [336/339]	8.57817091
Cluster606	0:3|1:50|2:40|3:22|4:26|5:6|6:1|7:8|8:4|9:13|10:9|11:11|12:18|13:17|14:14|15:24|16:17|17:15|18:8|19:3|20:2|21:9|22:2|23:2|24:3|25:3|28:1|	99.1% [336/339]	8.572271386
Cluster537	0:3|1:50|2:40|3:22|4:27|5:5|6:1|7:8|8:3|9:14|10:9|11:11|12:18|13:18|14:13|15:23|16:17|17:13|18:8|19:4|20:1|21:8|22:3|23:2|24:3|25:3|28:1|	99.1% [336/339]	8.568221534
Cluster101	0:3|1:50|2:40|3:22|4:27|5:5|6:1|7:8|8:4|9:13|10:8|11:11|12:18|13:18|14:13|15:23|16:17|17:13|18:8|19:4|20:1|21:9|22:3|23:2|24:3|25:3|28:1|	99.1% [336/339]	8.566371691
Cluster615	0:4|1:57|2:40|3:23|4:25|5:6|6:2|7:5|8:5|9:12|10:5|11:12|12:14|13:17|14:14|15:24|16:17|17:15|18:8|19:2|20:2|21:9|22:2|23:3|24:3|25:2|28:1|	99.1% [335/339]	8.526371156
Cluster2	0:4|1:57|2:42|3:33|4:21|5:4|6:3|7:9|8:3|9:10|10:6|11:12|12:18|13:18|14:18|15:18|16:21|17:12|18:8|19:3|20:2|21:7|22:3|23:3|24:3|25:1|26:1|28:1|	99.1% [335/339]	8.525073746
Cluster535	0:3|1:50|2:47|3:31|4:22|5:4|6:2|7:9|8:2|9:9|10:9|11:11|12:11|13:20|14:16|15:23|16:20|17:14|18:9|19:4|20:3|21:8|22:4|23:1|24:3|25:1|28:2|29:1|	99.1% [336/339]	8.454277286
Cluster397	0:3|1:50|2:47|3:31|4:22|5:4|6:2|7:9|8:2|9:9|10:9|11:11|12:11|13:20|14:16|15:23|16:20|17:14|18:9|19:4|20:3|21:8|22:4|23:1|24:3|25:1|28:2|29:1|	99.1% [336/339]	8.454277286
Cluster398	0:3|1:50|2:47|3:31|4:22|5:4|6:2|7:9|8:2|9:9|10:9|11:11|12:11|13:20|14:16|15:23|16:20|17:14|18:9|19:4|20:3|21:8|22:4|23:1|24:3|25:1|28:2|29:1|	99.1% [336/339]	8.454277286
Cluster182	0:3|1:60|2:46|3:32|4:21|5:4|6:2|7:9|8:2|9:11|10:9|11:11|12:10|13:20|14:17|15:22|16:23|17:11|18:9|19:4|20:3|21:8|22:4|23:1|24:3|25:1|26:1|28:2|29:1|	99.1% [336/339]	8.42778781
Cluster605	0:2|1:65|2:45|3:29|4:24|5:5|6:1|7:9|8:4|9:15|10:4|11:6|12:13|13:19|14:16|15:23|16:19|17:14|18:9|19:4|20:3|21:6|22:3|23:2|24:3|25:1|28:3|	99.4% [337/339]	8.392330383
Cluster260	0:4|1:57|2:44|3:38|4:21|5:2|6:6|7:8|8:4|9:8|10:7|11:10|12:14|13:21|14:17|15:26|16:18|17:12|18:9|19:2|20:4|21:6|22:4|23:1|24:2|25:1|26:1|28:1|	99.1% [335/339]	8.374631268
Cluster246	0:7|1:63|2:43|3:23|4:25|5:4|6:3|7:6|8:5|9:11|10:7|11:11|12:14|13:17|14:14|15:23|16:17|17:13|18:8|19:4|20:1|21:9|22:2|23:3|24:4|25:2|	97.9% [332/339]	8.374631268
Cluster315	0:4|1:64|2:47|3:28|4:24|5:3|6:2|7:4|8:7|9:11|10:5|11:8|12:12|13:16|14:15|15:23|16:16|17:20|18:9|19:3|20:1|21:10|22:2|23:3|24:3|25:2|28:2|	99.1% [335/339]	8.356921153
Cluster67	0:3|1:66|2:46|3:27|4:24|5:4|6:1|7:6|8:4|9:13|10:5|11:6|12:14|13:18|14:15|15:23|16:18|17:16|18:9|19:3|20:2|21:8|22:3|23:3|24:3|25:2|28:2|	99.1% [336/339]	8.339233038
Cluster395	0:3|1:66|2:46|3:27|4:24|5:4|6:1|7:6|8:4|9:12|10:5|11:6|12:14|13:18|14:15|15:23|16:18|17:16|18:9|19:3|20:2|21:8|22:3|23:3|24:3|25:2|28:2|	99.1% [336/339]	8.339233038
Cluster217	0:3|1:66|2:47|3:27|4:24|5:3|6:2|7:5|8:6|9:12|10:6|11:7|12:11|13:17|14:16|15:25|16:16|17:16|18:9|19:3|20:1|21:10|22:2|23:3|24:3|25:2|28:2|	99.1% [336/339]	8.327403628
Cluster212	0:3|1:62|2:48|3:28|4:24|5:5|6:2|7:5|8:4|9:18|10:4|11:6|12:14|13:18|14:16|15:29|16:17|17:16|18:9|19:4|20:0|21:12|22:3|23:3|24:4|25:1|	98.5% [334/339]	8.324483776
Cluster514	0:3|1:66|2:47|3:27|4:23|5:4|6:2|7:5|8:6|9:12|10:5|11:6|12:13|13:18|14:15|15:23|16:18|17:16|18:9|19:3|20:1|21:9|22:2|23:3|24:3|25:2|26:2|	99.1% [336/339]	8.318584071
Cluster88	0:5|1:62|2:46|3:27|4:24|5:5|6:2|7:5|8:4|9:14|10:4|11:7|12:13|13:17|14:16|15:26|16:17|17:16|18:8|19:4|20:0|21:12|22:3|23:3|24:4|25:1|	98.5% [334/339]	8.315634218
Cluster667	0:3|1:66|2:47|3:27|4:24|5:3|6:2|7:5|8:6|9:12|10:5|11:7|12:12|13:17|14:16|15:23|16:18|17:16|18:9|19:3|20:1|21:9|22:2|23:3|24:3|25:2|28:2|	99.1% [336/339]	8.315634218
Cluster581	0:5|1:67|2:48|3:27|4:25|5:3|6:1|7:4|8:6|9:10|10:5|11:8|12:12|13:14|14:15|15:23|16:17|17:16|18:9|19:3|21:11|22:1|23:3|24:3|25:2|26:2|	98.5% [334/339]	8.292587021
Cluster558	0:4|1:64|2:48|3:33|4:18|5:3|6:6|7:5|8:4|9:13|10:7|11:8|12:10|13:20|14:18|15:29|16:19|17:11|18:6|19:1|20:5|21:6|22:3|23:1|24:2|25:1|26:1|27:1|1	98.8% [335/339]	8.14158292
Cluster84	0:3|1:7|2:2|3:25|4:3|5:20|6:283|7:19|9:1|	99.1% [336/339]	5.578221239
Cluster83	0:3|1:7|2:2|3:25|4:3|5:20|6:283|7:19|9:1|	99.1% [336/339]	5.578221239
Cluster82	0:3|1:7|2:2|3:25|4:3|5:20|6:283|7:19|9:1|	99.1% [336/339]	5.578221239
Cluster294	0:3|1:7|2:2|3:25|4:3|5:20|6:283|7:19|9:1|	99.1% [336/339]	5.578221239
Cluster460	0:18|1:23|2:36|3:31|4:13|5:7|6:15|7:8|8:1|9:1|10:1|11:10|	94.4% [320/339]	4.65718142
Cluster410	0:1|1:2|2:6|3:28|4:88|5:238|7:1|	99.7% [338/339]	4.557522124
Cluster459	0:2|1:9|2:27|3:290|4:14|	99.4% [337/339]	2.908584672
Cluster512	0:2|1:9|2:28|3:272|4:18|	99.4% [337/339]	2.872356342
Cluster652	0:2|1:9|2:40|3:272|4:17|	99.4% [337/339]	2.867256637
Cluster586	0:1|1:9|2:40|3:283|4:6|5:1|	99.7% [338/339]	2.849557522
Cluster634	0:2|1:8|2:51|3:261|4:17|	99.4% [337/339]	2.83485826
Cluster16	1:7|2:45|3:286|4:1|	100.0% [339/339]	2.828208555
Cluster187	0:2|1:9|2:53|3:262|4:13|	99.4% [337/339]	2.81120944
Cluster245	0:2|1:38|2:59|3:179|4:39|5:12|	99.1% [336/339]	2.734512174
Cluster372	0:9|1:15|2:62|3:241|4:9|5:2|6:2|	97.6% [331/339]	2.713864307
Cluster474	0:4|1:21|2:80|3:239|4:14|5:1|	98.8% [335/339]	2.710914454
Cluster604	0:3|1:29|2:101|3:196|4:9|	99.1% [336/339]	2.687315634
Cluster330	0:5|1:38|2:81|3:179|4:23|5:11|6:1|	98.5% [334/339]	2.828318584

```
>GCA_000649655.1        87      copy:2(2-2)|21586|Segment1|1291948-1292977|1|length:1030|1-1030
TACTTTGGTGGCTTGTATCTCTTGTAGTCCGCTTTGACGGGGTGGTGGTCAGGTACGGTGGCCTCGGGAGAGGCTGGAGGGCTCGACGTTTGGGCTGAGTGTCTGGGCCCGTGAAAGAGATCGTCTGCTCCAGCTTTGTCTCCTGAACTGACCCGGTTTAGGGAATTGGT
>GCA_000673095.1        600     copy:2(2-2)|25417|Segment3|3134891-3135146|1|length:256|1-256
TTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACTTGAACACCGCCGATACCTATTTGGTCGGGAGTGATAAAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACCGGACTTGATCGACGCGAACCTGTTTCCGTCCCCTCTCGGGGTTTTG
>GCA_001673395.1        239     copy:2(2-2)|74389|Segment6|2329395-2329655|1|length:261|1-261
TCACGAACAGTGCGCAATCGGCGGCCCGCAACCCGGCCCGCAGCTCACCCACGAAGTCGGCGTACCCAGGGGTGTCGACGAGGTTGACCTTGATGCCGTCGTAAGCCAGCGAGGCGACCGCAAGGCCCACCGAGCGCTGTTGCCGGATCTCCGCCTCGTCGAAGTCGCAGA
>GCA_002087305.1        623     copy:3(3-3)|99097|Segment10|747298-747684|1|length:387|1-233
ATTACCTGGTGGTCAGCGGTCGTGGGAGGTTTTCAGATGTCGTTTGTGATCGCGGTCCCGGAGACTATCGCGGCGGCGACGGATCTAGCCGATCTCGGCTCGACGATCGCTGGGGCCAACGCGGCTCGGCCGGCCGGTGCCGAT
>GCA_000651675.1        499     copy:2(2-2)|21967|Segment5|2893564-2894085|1|length:522|1-522
GGGCGATCTGATCTACGTGGTGTCCGAGGGCGAGTTCGAGATTGTGCGCGAATTGGCCGACCGGCGGTGAGGAATTGGTCAAAACCGCCGCGCCTGGGGACTACTTCGGTGAAATCGGCGTGCTGTTTCACCTGCCACGCTCGGCAACGGTACGGGCTCGCAGCGACGCGAC
>GCA_001652545.1        556     copy:6(5-9)|74549|Segment4|4223807-4224555|1|length:749|1-566
CGTCCAGCGGCCCTGGCGCGGGTTTCGACTAGCCAACCGTTACGTCTGAACCATCCCCGGCTAGCAGATGCCGCTGGGAATCGGCCGCGCGCGGGACCGGATTGAACCGCCCCGGCATGTCCGGAGACTCCAGTTCTTGGAAAGGATGGGGTCATGTCAGGTGGTTCATCGA
>GCA_000653935.1        375     copy:2(2-2)|22413|Segment1|1785560-1786044|1|length:785|1-785
CCCGCCGACAGCTCCAACACCAACATCACCATTCGGCACTGCTCCCATGGTGCTAAGTTCATCTTCAACAAGTAGCGGCCCGACGACGCTCCAACACCAACATCACCATTCGGCACTGCTCCCATGCCGCCCGGCCCACCCCCCGGGTACCGTCTCACCACCCCTCCC
>GCA_001597645.1        417     copy:2(2-2)|50827|Segment10|4286163-4286389|1|length:227|1-227
AGAATCCACCCTGCCCGGCACCGCGGCGGTCAATCGCGGCAGCCGCGACCATGCCCACCGTCACACCGGCCGCCCATCCAGGCCGGCGCCGACCGGCCAGCCGTCCTGGGCGCCCAATAGCGGCCCGATGCCGGCCAGCGGACCCACTCCCACCCCGCAGTACTACCAGGGCGG
>GCA_001673125.1        242     copy:5(5-6)|74318|Segment27|4085224-4085436|1|length:213|1-213
CTACCCACCGGCCGGTGCCCGGGCGTCAAGCCACACCCGGGCGTGAGCTCCGCAACCCCTGGGGTGCTCGGGCGGCGAATCTGGAGGCTAACAAGGACCGCGCCATTGCTCTTCAGTCCGAGGGAGATTTACAACAAATACATACACTATCTGACCGGATGTGAGCACTTCTT
>GCA_002087375.1        254     copy:8(2-8)|98529|Segment19|3306896-3307150|1|length:255|1-255
ATCCTGTTTGAGCCGCTCGATGCACATGCGAGCGGCGTCGAACGCGGCCGCGCGGTGAGGAGCCGAAGCACCGGATGACAACCGCCGCATCACCGATGCGCAATTCACCGGTCCGGTGTGCCACGGCAACTCGCACACCGTCGGCCTGTCGTTCACACTCTTCGATGATGTC
>GCA_000658535.1        418     copy:5(2-5)|23053|Segment3|3829120-3829387|1|length:268|1-268
CGCGGGCCTTAAGCCAGCTTTAGGCCCAGCTGGAGCCGACGGCGCTGTCGGTTGTGCCATTGTTGCCGGCAGCCTGCACCTTCTGCCGTGGGCGTTGGCCTGCTCGTAGATCACCTGGAAGTTACGGCCCAGCTGGGTAATGAACCCCTGGCAGGCCGCCGAACCGG
>GCA_000649675.1        265     copy:2(2-2)|21509|Segment7|241754-242046|1|length:293|1-293
NCGCGCGCCAAGCCCACCTTCATCAACGACACCAGATCGTTCCAGGCCGCGTCGAATTCCGGCTCACCGCTCCGCCTCCGGCCGCTGTGGGTCGATCCGGTGCCGAAAAAGCAACTCATTACGGTAGACGTTGCCAACACCGGCGATCACCGTTTGGTCCATCAAGAGC
>GCA_001184735.1        321     copy:2(2-2)|38075|Segment2|1924525-1924764|1|length:240|1-240
GGGCGTCGCAACCCGGCCTACCACGGTGAGGTCCACCACGGTAACGCCGTGGACAACCACCCACCGTACCGACATCAACGACCTCACACTGGCTCGCGGCCCCGACAATCGCCTTGTCGAAAAAGGCTGGAAAACCCGCAAGAACGCCAAAGGCGACACTGAATGGCTACCG
>GCA_000651655.1        358     copy:2(2-2)|21964|Segment1|1880369-1880680|1|length:312|61-312
GACGCTGTGGAGCTGGTGGCACACGGCCGTGACCGGGTTTCGGAGTTCCCGGCGGATCGCGGCTGGGATGTGGCGGGGTTGTTTGACCCCGATCCCGACGCGGTAGGCAAGTCGTATACCCGGTGCGGCGGGTTCTTGACGGATGTCGCCGGTTTTGACGCCGAGTTTTTC
>GCA_000672895.1        545     copy:3(2-3)|25391|Segment6|3850238-3850871|1|length:634|226-634
ACAGGTATTGACAATTGCTCATTCAGGCTCAAATTGCCTGTGCCGCATGATGGTTGGCGTTAAAGCGTGAGGACCTGCCGGATGAATTGACGGGAGCGTAGCTGTGACCGCGGTCCCGTCACCCGTGGCAGGGAAACAGTTGCAGTGTGTACTATCCGCCCTAGACTGCCGC
>GCA_001673395.1        240     copy:2(2-2)|74389|Segment8|2601736-2602454|1|length:729|1-555
CAGACCGGGCGGGCTCGAGCGATCACCGGTGCCCAAGGCGACCACGGCGCCTTCGCGCACCTCCGCGAACTTGCGCTGCCACCAGCTCGACCTCCGTCGGAAAGATGTTGCGTCCCGCGATGGAGATGACTTCCTTCGCGCGACCACACACCACCAGGCCGCCAGGCGCAAGT
>GCA_002087375.1        145     copy:6(6-6)|98529|Segment4|1186966-1187204|1|length:239|1-239
GGGGACGGAAACCCAACCTCACCGCCTGCTGGGTGAGACGTGCTCGCCGCGAGTCGTCAGACCCAAAACCCTGAACCGCCCCGGCATGTCCGGAGACTCCAGTTCTTGGAAAGGATGGGGTCATGTCAGGTGGTTCATCGAGGAGGTACCCGCCGGAGCTGCGTGAGCGGG
>GCA_000652135.1        639     copy:2(2-2)|22041|Segment6|2803004-2803394|1|length:391|1-299
NNNNGCTGGATGAGCGGCGTTACGTGCTGATATCGATGACTCAGGTCTTTGCGCTGGCGTTGGTGTTGTTGTCGCTGGAGTTCGTCTCCGCGCGGCGACATCTCTGATCCGTCGCTATGGACAGCTCGGCGCATTGAATGTCGTTGGGCAGGTGGTGGGTGGCTACCACCA
>GCA_001652545.1        678     copy:3(2-3)|74549|Segment3|1081665-1081890|1|length:226|1-226
GCGTCAGATGGGTATAAGAGGCAAGGCCACAGGCAGGGCACCACGCCTTCTTCCAAGCTGGGTGGCCTGTGGAAGACGGTCTCGCCGCATCTGTCGCCGATCAGCAACATGGTGTCGATGGCCAACAACCACGTGTCGATGACCAACTCGGGTGTGTCGATGACCAACACCT
>GCA_000653035.1        448     copy:3(2-4)|22191|Segment5|3942494-3943774|1|length:1281|169-506
CCAGCACGCGTTGATCAACCAACTTGACGCCCAAGCGAGGAAGAACTGGGCGGCACGCTGTGCTGCCGCCTGGCCAACCCGTTACGCATCACCAAGCCCGACGCCCGCCCGGCGCATCGCCGACGCCGCCGATCTCGGACCTCGTCGAGCACTCACCGGTGAACCGCT.
>GCA_000649675.1        639     copy:2(2-2)|21509|Segment1|2469179-2469468|1|length:290|1-290
GCTAAGGCATGTGCGGGCTCAGCCGAAGACTTCGGTCTCAGCCAGGGCCTCCGTCAGCCTGCGTGCCCCATCGGTGAACTGCCAGACGGTGTGCTCGATTACGGCGGCTGTGTCGCGGCGGCGCAGCGCGGCGATCAGCTGCCGATGACTGTTCACCGCGTCCGGCGCCCA
>GCA_002003715.1        483     copy:2(2-6)|92384|Segment1|4180433-4180671|1|length:239|14-239
AACGCCGGGTCGTACAACACGGGCAGCTTCAATGTCGGTGACACCAACACCGCCATCGCGAACTCGGGCAATCTGTCAGCACCGCGCCGCCTCATGCGGGCAACTACGAAGGCCTATTCAGCTATTCCTACAGCCTCGACGTT
>GCA_000653035.1        583     copy:2(2-2)|22191|Segment4|3655727-3656044|1|length:318|1-318
TCGGGCCCTGGTCTGTTGTCGGCAGCGGCTGGTAGTATCGAAAGTATGTTTGATCAGGTCGGGGGGCGCATGCCTTCACCGGAGGCGATCGCTCATTTTGATGAGCGGTTTGAATGCCATGCTCCGCGGACCACGAGGGTGTCGGCGGCGTTCATCGATCGGATCTGCTCGG
>GCA_000673095.1        454     copy:2(2-5)|22931|Segment1|100382-101528|1|length:1147|1-1147
TGAACCGCCCCCGGCATGTCCGGAGACTCCAGTTCTTGGAAAGGATGGGGTCATGTCAGGTGGTTCATCGAGGAGGTACCCGCCGGAGCTGCGTGAGCGGGCGGTGCGGATGGTCGCAGGATCCGCGGTCAGCACGATTCGGAGTGGGCAGCGATCAGTGAGGTGGCCGCGT
>GCA_000653535.1        52      copy:4(2-4)|22314|Segment1|1513859-1515384|1|length:1526|1087-1437
TGATGAGCCCGCCACCAATTGCGACCTGGATCATACGATCCCGTATGCCGCTGGTCGCCCACCCATGCGTCGAACCTGAAATGTTGGGGATGGCGTGATCAACAGCTACCCGACGGCACCCTGATTTTGACCTCCCCG
>GCA_001673395.1        480     copy:5(4-6)|74389|Segment32|4205069-4205529|1|length:461|1-461
GGACGCCGAACTAGCTCGATCCAGCGCCGACCCGCACCCCACTACCGGCATCTGAAGGTGAGCCAGAGGACGCCGTTCGACCAGGAAGAACCGTGGCCGCACGGGTCACCCGGGCACACCCAACCGGGCCGTGGCAAGTGCCGACTACCTGAAGAATCCCGAAAGTCCTACACC
>GCA_000893135.1        380     copy:3(2-3)|22684|Segment10|1808893-1808382|1|length:511|1-511
GGTATTCCGTTGAGCCTGAACGCGCTGGGTGGTGTCGGTCCGATTGTGGTGCCGGGATATTACTATTCCTGGTATTCCGTTGAGCCTGAACGCGCTGGGTGGTGTCGGTCCGATTGTGGTGCCGGGATATTACTATTCCTGGTATTCCGTTGAGCCTGAACGCGCTGGGTGGT
>GCA_000655515.1        283     copy:14(3-14)|22664|Segment17|2156730-2158296|1|length:1559|1-205
GCCCCACCAACGCCCCCCGAACTACTCACACCCAAACCAGTGGCGGGAACAGTAATAGGCGGAACATTGATTGGCCCCACCAACGCCCCCCGAACTACTCACACCCAAACCGATGGCGGGAACAGTAATAGGCGGAACATTGATCGGCCCCCACCAACGCCCCCCGAACTACTCA
>GCA_000653135.1        117     copy:3(2-3)|22216|Segment2|2246026-2246866|1|length:841|1-352
GGGCTGGGCGGCCGCCGGCGAGCGAAAGGCCCTTGAACATCGAGTGCAGGGTGCTGGCCATCGACGACACCCGAGTTGGTCATCGACACGTGGTTGTTGAGCATCGACACGATGTTGCTGAGCGGCGACAGATGCGGCGAGATGGCTTTCCAGAGTTCACTCAGTTGGTCGAA
>GCA_000672915.1        660     copy:4(2-4)|25392|Segment1|1637416-1638070|1|length:655|286-655
CTGTTGGCTACACCGATCGATGACCCGATGCGCGCCACATCCGAGGCCGCCGCGGCCACGCTCCGGGGTTACGATCACCAACGACATCACAGTCCACCCGCCACGCCCCTGCCCCTTCGGCAGGTCACACTCCTGCCAGATAAGGGTCGCGCCGCCACCTTGTCCGATT
>GCA_000649655.1        263     copy:5(3-5)|21586|Segment6|1671151-1671536|1|length:386|1-386
TCGTAGCGCCACGGCAACCCGCCGTCGGCGACCTGGGTACCGACCCCTGCCGGTGTATAGAAGGCCGGTATGCCCATCCCTCCGGCCCGCAACCGCTCGTCGGCGACCTGGGTACCGACCCCTGCCGGTGTATAGAAGGCCGGTATGCCCATCCCTCCGGCCCGCAACCGC
```

Bordetella pertussis

ClusterID	Distribution	Marker	Percent_of_strains	Weighted_Average_Copy
Cluster43	32(84)|	*	100.0%[84/84]	32
Cluster38	32(84)|	*	100.0%[84/84]	32
Cluster33	31(1)|32(83)|	*	100.0%[84/84]	31.98809524
Cluster159	31(1)|32(83)|	*	100.0%[84/84]	31.98809524
Cluster198	19(1)|32(83)|	*	100.0%[84/84]	31.8452381
Cluster59	18(1)|32(83)|	*	100.0%[84/84]	31.83333333
Cluster208	18(1)|32(83)|	*	100.0%[84/84]	31.83333333
Cluster90	14(1)|32(83)|	*	100.0%[84/84]	31.78571429
Cluster57	8(1)|32(83)|	*	100.0%[84/84]	31.71428571
Cluster73	4(1)|32(83)|	*	100.0%[84/84]	31.66666667
Cluster25	4(1)|32(83)|	*	100.0%[84/84]	31.66666667
Cluster160	2(1)|32(83)|	*	100.0%[84/84]	31.64285714
Cluster37	1(1)|32(83)|	*	100.0%[84/84]	31.63095238
Cluster24	1(1)|32(83)|	*	100.0%[84/84]	31.63095238
Cluster64	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster63	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster62	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster60	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster58	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster49	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster48	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster45	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster44	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster40	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster39	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster35	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster34	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster29	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster28	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster26	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster22	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster21	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster207	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster199	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster167	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster164	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster161	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster16	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster15	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster14	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster13	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster12	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster120	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster113	0(1)|32(83)|	*	98.8%[83/84]	31.61904762
Cluster53	0(1)|31(4)|32(79)|	*	98.8%[83/84]	31.57142857
Cluster27	0(1)|30(2)|31(71)|32(10)|	*	98.8%[83/84]	30.72619048
Cluster80	2(1)|17(83)|	*	100.0%[84/84]	16.82142857
Cluster77	2(1)|17(83)|	*	100.0%[84/84]	16.82142857
Cluster119	2(1)|17(83)|	*	100.0%[84/84]	16.82142857
Cluster215	15(1)|16(83)|	*	100.0%[84/84]	15.98809524
Cluster181	15(2)|16(82)|	*	100.0%[84/84]	15.97619048
Cluster1	15(2)|16(82)|	*	100.0%[84/84]	15.97619048

```
>GCA_001509915.1        89      copy:2(2-27)|66337|Segment1440|2558232-2563126|1|length:4895|1784-3019
GGCTCAGGCGCCCTGCCTGGCCGAGGCGGCGCGGCGACGCGGCCAGAAGTAGTACAGGGCCACCGCTATCTGCAGCGCCAGGAAGGCCCCCCACGCGGCCAGGTAGCCGCCGGCCGGGTAGCTGCCGGGCTCATACACGGCGTGCCGCGCGGCCATAGTTCGA
>GCA_001187485.1        11      copy:12(2-12)|43361|Segment159|3014361-3014662|1|length:302|1-302
TGGCGCCGTCCGGTCCGGGCCAGGCCAGCTGCCGATGCAGCTTTCGGTGCGTCCCTGGGTTTCCGGCTTGGGCCGGCACGGGATACATCTCGGCGGCGTCGACCAGGTTGATGCCGTGCGCCAGCGCATAGTCCAACTGTTCGTGAGCTCGCGCCTCGGTGTTCTG
>GCA_001686485.1        191     copy:3(2-3)|76341|Segment733|1845332-1851533|1|length:6202|626-2198
TACGGATCCTGGTTTCGATCAGGGTCACACACACGAGATTGAACTGAAGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGATGCTTTACACATGCAAGTCGGACGGCAGCACGGGCTTCGGCCTGGTGGCGAGTGGCGAACGGGTGAGTAATGTATCGGAA
>GCA_001187485.1        190     copy:2(2-2)|43361|Segment209|3143428-3144188|1|length:761|1-761
GTGGCCGGGCGACAACTATTTCGATACCACGCTGGTGCGCGAGCAGGTGCGGCGCGCCCTGGGCGGCTACGAAAGCCGCCTGCCCGTGCGCGGTGACGCGTTGGTGGCCAGGCTGATGGAGTCGGCCGGGAGCGGTCGGCAAGGCGCTGGGCCTGAAGGTGGGTGT
>GCA_001187485.1        56      copy:6(3-6)|43361|Segment83|2319850-2320142|1|length:293|1-293
CCATCGCCGTCATCAGCGCCCCTGCTGTTCTACCGCCACCGCGCCAACATCGCCGCCGGCTGCTCAAGGGCACCGAAAGCGCATCGGGAAGAAGAAGTAACCCGCTCAGGAGCCTGTCCCGGCGGGGACAGAAGCAACCCACTCAGGAGCCTGTCCCGGCGGGGGC
>GCA_001509915.1        114     copy:2(2-27)|66337|Segment1439|2558232-2563126|1|length:4895|1369-1773
GGCCTGGCGGCTGGGCCGCAAGAACGTGCATATCGCGTTCAGCAATTTCGCCGAGGCGTTCTATCGCATGATGAGCGAGCCGAGCACGACCAGGTGAGCGTGCCCGAATTCAATAACCTGCTGGTGCAGAACCACATCCTGGCCTCGCAGATCACGGCCGTCG
>GCA_001187485.1        148     copy:2(2-2)|43361|Segment90|2366749-2366995|1|length:247|1-247
CAGCAGGACGCGGCTCGCGCCCCAACCGGGCCGCCAGCGCTCTCTCCCAGGGCGTCGCTGGCGCCAGCGCCAACCAGCGTCCTCGCATGCGTACAGAACTCAAACATCGCCACCTCGCTCAGAACAGGCTGAAGTCCGCACGCCGACGCACGGCATCCCGCCCGCGC
>GCA_001187485.1        152     copy:11(4-11)|43361|Segment195|3023626-3023867|1|length:242|1-242
ACGGCCACGGCCAGCCGCGCCGGCGCGGGATCCAGGGCGCGCGGCCGCCACCGGCCGGCCTAAGGCCGCGGGGCATAGGTCGGAAAGAGAGGGAGCCGGCGCTGCGGGGGAGGTGGACATGGTGGCTGGCGGCGCCGGCGTGACGCGCTGGCGGGCTGG
>GCA_001985665.1        119     copy:7(7-7)|81953|Segment306|1289397-1289801|1|length:405|1-405
CTGGGGCAGCGCCGTGGCCGCTTGCTCGTGATACAGACCCGTAGGCTCCAGGATGGCGTGCAACTGCGCCGGCTGTTTGGCGCACCACGCCAGCAGGGCTTGTACGCCACCCTTGGTCTTTACGCTTTGCTCAGCCTCAG
>GCA_001187485.1        124     copy:4(2-4)|43361|Segment1|1314937-1317201|1|length:2265|392-1131
TCGCCGCCAGCAGGTTGGCCTGCACCGCGTTCTCCACGAAGCAGAAGTCGCGGCTGGTCTGGCCGTCGCCGTTGATCACGACGTCCTCGCCCTTGATCATCGCCGCCCGTCCACTTCGGAATCACCGCCGCGTACGCGCGCCGTCCGGGTCCTGGCGCTTGCCGAAC
>GCA_001988875.1        181     copy:7(3-8)|82228|Segment1323|2246290-2247080|1|length:791|1-791
TCCGGGTCTCCTTGGTCAATCCCGCCCCAGGCCCGGGACTTTGCCAAGGCCTTGGCGCTGCGCTCCAAAAATGATGCGCTCGACAGCTACGTGCTGGCCTGGCTCGCTATGGACAGACGCTGAGCCCGGCGCTGTGGCACCCGGCGCCCTTGCATGCGCGTCAACTGCGC
>GCA_001187485.1        133     copy:4(2-4)|43361|Segment128|2982307-2982782|1|length:476|1-476
TACGGCATCCATGCCGCCGGCGGGCACGCCGGCGGCGGTCGTGGATCGGCTCAACCAGGCCCTGGTGGCCGTTCCTGTCCAACGCCCAGGTCGCGCGCGGCGCTGGTGGGGCAGGGCTATGAGGTTGCGACCAGACCGCCTGATGTTGCCGGATGTTGCGCGA
>GCA_001187485.1        4       copy:7(2-7)|43361|Segment93|2422215-2422469|1|length:255|1-255
CGCGCCACCCACCGCATCCTGACCCAGAACATCGTCCTGGCGCTGGGCATCAAGGCGGTGTTCCTGGTGCTGGCGATGGCCGGCCACGCTGTGGATGGCGGTGTTCGCCGACGTCGGCGCCAGCCTGCTGGTGGTCGCCAACGGCTTGCGGCTGCTGCG
>GCA_001686485.1        75      copy:14(2-15)|76341|Segment1652|2720186-2722112|1|length:1927|1728-1927
CCACGCCGTGCTTGGCGGGCCACGGTAGGCGGACTTCGCGGCCGAGGCCACCAGGCCATGGGCCGAGGCGATGTTGATGATGCGCCCCCAGCCCTGCTTCTTCATGTGGGGCAGCGCCGCGGCCGTGCCATGGAATACCGCCGAGAGATTCAGCGCGATGATGGC
>GCA_001509915.1        108     copy:3(2-3)|66337|Segment869|1978457-1984658|1|length:6202|6002-6202
GGGGTACGCCCAGGCGGCTGGCCCCGCAACGGGCACGACAGCCGCGGCCGGGGGGCCTGTCCCTGCGGGGGAAGGATACCTGCGGCCCGGGTGCCTGTCCCCGCAGGGGGACAGGCACCTGCAGACATCCCCATCTTCTGACGAACCGCCTGCAGCGTGTGTG
>GCA_001686485.1        175     copy:2(2-2)|76341|Segment225|1244049-1246162|1|length:2114|1-429
AGCGCGACTCAGCCTGGCCTACTCGCCGGGCGTGGCCGCCCGCCTGCGAGGAGATCGTCGACGATCCGGTGAATGTGTTCCGCTACACCGGGCGGGGCAACCTGGTGGGCGTGATCACCAACGGCACGGCGGTGCTGGGGCTGGGCAACATCGGCGCGCTGGCC
>GCA_001686485.1        36      copy:3(2-25)|76352|Segment1832|2857139-2858204|1|length:1066|1-1066
TACAACCTATTGACAGTTCACACCTAGCTGTCGATTCTCAATAGGTTATTCACATCTGGGATACGCGTGATGGGTGGCTGATGTCCGAGAGCGGAGTGTGGCCGACACCAGTTGTAGCGGTCAATGAAGGGCTGCAAGGCAGCCTGTCGTTCGGCTGAGCTCGT
>GCA_001187485.1        146     copy:3(3-5)|43361|Segment60|1948920-1949207|2|length:288|1-288
CGGGCGGCGAAGGCAACCGGCCGTCCTCGCTATCGACGGGAGCGCCGTCCAGCTCGGGACGCGGAATATCGATGTGGGTCCAGCCATCGTCGATGGCCGTGCCATCGAGATCTATCGTTGTCCAGCCGTTTTCATCGTCGACGGGCCAGCCATCTAACTGTGAA
>GCA_001686485.1        141     copy:20(2-22)|76341|Segment818|193225-195314|1|length:2090|1-2090
TAGCTGTGAACTGTCAATAGGTTGTATTCGTCCAGGTTGAGTCTGGAGATGGGTACAGCGCGCCCGATGCCTTGGTGGGGTCGATGCCAGTTGTAGTGGTGTAGCCAGGATTTCATGGCATCGGCTCGGTGTTGGGAGTTCTGGTAGGTGTGAGCGTAAGCCCA
>GCA_001187485.1        203     copy:5(4-7)|43361|Segment112|2690079-2690374|2|length:296|1-296
TGCTGGGCAACCTGGTCGGCCTCATCGACTCCGGACTACCAGGGCCAGTTGATGGTCTCGACGTGGAACCGCGGCACGCAGCCGTTCACCCCTCGACCCCATGGAGCGCCTGGCGCACAACTGGTGATCGTGCCGGTACAGCAAGTGGCCTTCAACGTCGTGGAAGAC
>GCA_001985725.1        167     copy:19(11-19)|81957|Segment2919|438223-438809|1|length:587|1-587
AGCTGTGAAGATTCAATAGGTTGTATGCATGGTTCATCCGAACCGGATTTGAGAAACTGGAAATCGCCACCCCCCAGTTCACTCAAGGAGCCCCGGCCGGATGAACACCCATAAGCATGCCCGATTGACCTTCCTACGTCGACTCGAAATGGTCCAGCAATTGA
>GCA_001509915.1        174     copy:2(2-2)|66337|Segment2518|360017-360264|1|length:248|1-248
CCTGGTCGCCTTCGGCGTGCTGCTGGTGGGCATGGTGACGGGCATCGACGGCTCGGGGTTCGCCGGCCTGCGCGCTGACCGGCACCCTGTCCGGCGCCCTTGGGCCCGGTGGTGGGCATCGACCCGGCGACGCTGGCCGCGGTGGGCCAGATGGGCGCCGTATGGA
>GCA_001187485.1        0       copy:4(2-4)|43361|Segment2|1314941-1317201|1|length:2262|1-380
GGCGCCGGCGGCCTTGCCCAGTATCAGGAAGACGGCCACCCCGGAACGCACCCTTGAAAATGCGTTGGTGGTCCGGATGGATCGCGTCGAGTTTGCGGCGGATGGCTTTCAAATCAACGCAGGAATTGCGTGTACCAGGGCATGGCGGCTTCCAGCCCGCGCAGG
>GCA_001987755.1        56      copy:2(2-2)|82153|Segment36|1061843-1062036|1|length:194|1-194
GGAACACCGCCTCGGGAATGGAGTGGGTGATGAACATGACCGACTTGCGGGTCTCGCTCCAGATGCGCTGCAGCTCCACCATCATGGTCTCGCGCGTCATCGCGTCGAGCGCGGCGAACGGCTCGTCCATCAGCAGCAGGCTGGGGTCGTGAATCAGGCCGCGC
>GCA_001509915.1        21      copy:29(25-30)|66337|Segment1582|2699673-2700072|1|length:400|1-400
TAGTTGTGAACTGTCAATAGGTTGTATTCGTCCAGGTTGAGTCTGGAGATGGGTACAGCGCGCCCGATGCCTTGGTGGGGTCGATGCCAGTTGTAGTGGTGTAGCCAGGATTTCATGGCATCGGCTCGGTGTTGGGAGTTCTGGTAGGTGTGAGCGTAAGCCCA
>GCA_001187485.1        204     copy:2(2-2)|43361|Segment128|2792857-2793062|1|length:206|1-206
TCCTGTGGGGCACGGCCGGCGCGGTGCTGCTGCGCGTGCTGCTGATCGCCTTCGCGCTGGTCCTGCTCGACATTTCCGTTTCTCAAGGCCCTGGGCTGCCTGCTGCTGGTCTGGATAGGCGTCAAGCTGCTGGCCCCGCATGAGGACGCGCACGGCAACATCACC
>GCA_001988335.1        180     copy:24(6-28)|82188|Segment240|1194383-1197531|1|length:3149|1-3149
CCTAGGTGTGAACTGTCAATAGGTTGTATTCGTCCAGGTTGAGTCTGGAGATGGGTACAGCGCGCCCGATGCCTTGGTGGGGTCGATGCCAGTTGTAGTGGTGTAGCCAGGATTTCATGGCATCGGCTCGGTGTTGGGAGTTCTGGTAGGTGTGAGCGTAAGCC
>GCA_001686485.1        28      copy:28(28-28)|76341|Segment180|1177933-1178251|1|length:319|1-319
CAGCGCATGGGCGATGCGCGCTTGGGTCAGCCGCTTGCGGCGCAGCTCCACGATAGCCAGCGCCTTGGCCGGCGCAATCGCTCGGGGCGAGGACCGTCGGGCGCGAGGACCGCTCGGCCTGAGCCAGGAAGCGGCCCAGCCATTTGCGCA
>GCA_001686485.1        113     copy:29(4-29)|76341|Segment1653|3696885-3696485|1|length:321|1-321
ACAGCTAGGTGTGAAGATTCAATAGGTTGTATGCATGGTTCATCCGAACCGGATTTGAGAAACTGGAAATCGCCACCCCCCCAGTTCACTCAAGGAGCCCGGCCGGATGAACACCCATAAGCATGCCCGATTGACCTTCCTACGTCGCCTCGAAATGGTCCAGCAT
>GCA_001187485.1        126     copy:2(2-2)|43361|Segment32|1659987-1661091|1|length:1105|1-1105
TCGGGGATCAGCCTGCCGCGTCGAGCGTTGCCGTGCCCGCGGCCATGCCCGGCGCCAGGGCGACGCGGGATCCGGCGGCCCTGCGGGACGTACCCCGCAAGCCGGTGGCGCCCGCCGTGGCCTTGCCGCAGCCACTGATCGAAACCCGCATCAAGTATCTCGAC
>GCA_001686485.1        192     copy:3(2-3)|76341|Segment735|1845332-1851533|1|length:6202|3723-3956
GCACCGGGTGCTAACGTCCGGACTCAGGACTCAAGGGGAAACCAACCCAGACCGCCAGCTAAGGTCCCGAATTATCGCTAAGTGGGAAACGAAGTGGGAAGGCATAGACAGTCAGGAGGTTGGCTTGAAAGCAGCCACCCTTTAAAGAAAGCGTAATAGCTCACTGATCGA
```

```
>GCA_001985665.1        53      copy:26(2-26)|81953|Segment2184|3299433-3300742|1|length:1310|1-1310
CTAGGTGTGAAGATTCAATAGGTTGTATGCATGGTTCATCCGAACCGGATTTGAGAAACTGGAAATCGCCAACCCCCCAGTTCACTCAAGGAGCCCGGCCGGATGAACACCCATAAGCATGCCCGATTGACCTTCCTACGTCGACTCGAAATGGTCCAGCAATTGATCGC
>GCA_001509915.1        14      copy:31(26-31)|66337|Segment1004|2092533-2093244|1|length:712|1-712
AGCTGTGAAGATTCAATAGGTTGTATGCATGGTTCATCCGAACCGGATTTGAGAAACTGGAAATCGCCAACCCCCCAGTTCACTCAAGGAGCCCGGCCGGATGAACACCCATAAGCATGCCCGATTGACCTTCCTACGTCGACTCGAAATGGTCCAGCAATTGATCGCC
>GCA_001686485.1        154     copy:3(2-3)|76341|Segment734|1845332-1851533|1|length:6202|2662-3715
CATGATTTTGTTCGAAGAACTTCAAGAACTGGCTGGGCGCGCAAGCGTTTGGTCAGATGCTTTGAACTTATGAACGGCACAAGCGCGAATGAACAGCACCTATAAGACTTTAGTGTTATAGGATCAAGCGACTAAGTGCATATGGTGGATGCCTTGGCGATCACAGGCGATG
>GCA_001509915.1        27      copy:27(2-27)|66337|Segment3143|736748-738057|1|length:1310|1-1310
GGCCGCCCGCGCTGAAACGCCGGAACCGACTCGGGGTTCGATCCCGGCAGATAGCGTTGGAACAGGTACCGCATGGAACACATACACCACGACCGACAGGCTGCACAGCACGGTCGGCCACAGCAGCCACGCGATTCTTGAACCGGGTATCAAGGCAAATCCTCTTCTTGTGC
>GCA_001187485.1        71      copy:2(2-2)|43361|Segment214|877840-878271|1|length:432|193-432
GGTCACATTGGTGATGGTCGGTGCGGCCGGCGCGGTCTTGTCCACGGTATCCGCATAGGCGCGGGTGGCCTCGGGGCTGCGGTTGCCGCCCTTGTCGATGGCCTGCACGCGGAGGTCGCCGGCCAAGCCCGGCCTGCGCATCGACGTCGCGATCCGCGCCATGGCGCCATCGCGCGCCCGCGCGCGCCGCGCGCG
```

Bordetella parapertussis

ClusterID	Distribution	Marker	Percent_of_strains	Weighted_Average_Copy
Cluster6	22(21)|24(1)|25(1)|	*	100.0%[23/23]	22.2173913
Cluster5	22(21)|24(1)|25(1)|	*	100.0%[23/23]	22.2173913
Cluster4	22(21)|24(1)|25(1)|	*	100.0%[23/23]	22.2173913
Cluster22	22(21)|24(1)|25(1)|	*	100.0%[23/23]	22.2173913
Cluster13	22(21)|24(2)|	*	100.0%[23/23]	22.17391304
Cluster12	22(21)|24(2)|	*	100.0%[23/23]	22.17391304
Cluster1	22(21)|23(1)|24(1)|	*	100.0%[23/23]	22.13043478
Cluster0	22(21)|23(1)|24(1)|	*	100.0%[23/23]	22.13043478
Cluster3	0(1)|2(22)|	*	95.7%[22/23]	1.913043478
Cluster9	0(1)|1(1)|2(21)|	*	95.7%[22/23]	1.869565217
Cluster23	0(1)|1(1)|2(21)|	*	95.7%[22/23]	1.869565217
Cluster29	1(22)|3(1)|	*	100.0%[23/23]	1.086956522
Cluster8	1(22)|2(1)|	*	100.0%[23/23]	1.043478261
Cluster38	1(22)|2(1)|	*	100.0%[23/23]	1.043478261
Cluster31	1(22)|2(1)|	*	100.0%[23/23]	1.043478261
Cluster25	1(22)|2(1)|	*	100.0%[23/23]	1.043478261
Cluster2	1(22)|2(1)|	*	100.0%[23/23]	1.043478261
Cluster36	1(23)|	*	100.0%[23/23]	1
Cluster34	0(1)|1(21)|2(1)|	*	95.7%[22/23]	1
Cluster33	0(1)|1(21)|2(1)|	*	95.7%[22/23]	1
Cluster30	0(1)|1(21)|2(1)|	*	95.7%[22/23]	1
Cluster27	1(23)|	*	100.0%[23/23]	1
Cluster26	1(23)|	*	100.0%[23/23]	1
Cluster24	0(1)|1(21)|2(1)|	*	95.7%[22/23]	1
Cluster18	0(1)|1(21)|2(1)|	*	95.7%[22/23]	1
Cluster17	1(23)|	*	100.0%[23/23]	1
Cluster10	0(1)|1(21)|2(1)|	*	95.7%[22/23]	1
Cluster7	0(22)|1(1)|		4.3%[1/23]	0.043478261
Cluster40	0(22)|4(1)|		4.3%[1/23]	0.173913043
Cluster39	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster37	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster35	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster32	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster28	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster21	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster20	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster19	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster16	0(22)|4(1)|		4.3%[1/23]	0.173913043
Cluster15	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster14	0(22)|2(1)|		4.3%[1/23]	0.086956522
Cluster11	0(22)|2(1)|		4.3%[1/23]	0.086956522

```
>GCA_000636935  26      copy:2(2-2)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_10|Segment9|1537101-1537430|1|length:330|1-282
GACAACCAGTATGCGGGCGCACGCTGGTCAACGCCGGCACGGTGCAGGTTGCGCGCGATGCGAACCTGGGTGCCGCCGGCCGGGTGGCGCTTACGGCGCTTGCGCGCTGACCATCGGCGCCGGCAACGGGGCGATGTCG
>GCA_000317935  14      copy:2(2-3)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_13|Segment123|2255085-2268420|1|length:5336|3156-4341
CGGGCGTCGAGGGTCGGGCCGATCAGGCCGGTGGCTACCGCCACGCGCGGCACCATGGCGGCGTATCCGGCCAGGTCAAGGCGCGTGGCGCGCAGGTATGAGGCGGCGGCTCATTGCCGGCTCCCCGAAGGTGACGCGCTGGCCGTCTTCGGTGAGCTCGACCAGCG
>GCA_001525545  31      copy:2(2-2)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_16|Segment123|4775972-4786187|1|length:10216|9544-10096
CCGATGACGGCCCGGCCGGCGGCATGTCCAGCAAGCAGCTGGAAGACCTGCGCGTCAAGGCCCTGGCCAAGTTCGACGAGGTGAGCAAGCAGTTCGAGAAGATGCGCCAGTCCTATGAGAAGGAAGGCTATAAGTCGGACGCCTATCTCAGGCCCAGGACATCATCCAGACCGAACT
>GCA_000317935  28      copy:2(2-2)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_13|Segment41|1747064-1751055|1|length:3992|1-1220
CATCTATTCCTCCGTTCTTCCTTTTCGATTTTCTCGGCTCAGCCAGCGGTGCAGGGCCAGGGCGAGCGCCACGACGGCGCCCACGGCGATGCCCAGCAGGGCGAGGTCAGGGCATGGGCGGCCTCCTGCATCTTGGCGCCGATCTCGGCAGCGGCACGGACGATGGCGC
>GCA_001525545  10      copy:2(2-2)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_16|Segment124|566-10715|1|length:10150|1675-1936
CGCGCGCCCAGGGCATCATGGGCGAGGGCGACGCCCAGGCGGCAGCATCTACGCCCAGGCGTTCGGCCGCAATACCGAGTTCTACACCTATTACAAGAGCCTGGAAGCCTATCGGCGCCGCGTTCGGCAAAACCGGTGACGTATTGGTGGTCGATCCGACGTCGGAGTTCTTCCAGTT
>GCA_000317935  5       copy:1(0-5-19)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_13|Segment110|2132763-2133567|1|length:805|1-805
CGTCGGTGCGCGAACCGGATTGGTCCAAGATCGAGTATTTGGCGATGGACGAGTTTGCCCTGCACAAAGGGCATCGCTACGCGACAGTGGTGGTCGATCCGATCGGCAGGCAGGTGCTGTGGATTGGCCCAGGACGCTCACGCGAGACGGCCCGGGCGTTCTTCGAACAATTGCCGC
>GCA_004000355  9       copy:2(2-2)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_1|Segment144|754200-754526|1|length:327|15-312
ATGCAGGCACCGTCTCTTCCACGCCAACAACCACGACCGCAGCGGACCCGGGTGCCTGGCTCCCCGCCACACCGGCCCGCGCCACTGCGCTGCCTGCGCGGCGGGTGCCTGGCACCTTGCCGCCACCCCACGTGCCACGCTCGCAGGCGGCCATCG
>GCA_004000215  36      copy:2(2-3)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_5|Segment141|476195-476494|1|length:300|30-279
CGCCGGCACGGTGCAGGTTGCGCGCGATGCGAACCTGGGTTCGCTGCCGGCCGGCGCGGTGGCGCTCAATGGCGGCACGCTGGCGGCATCGGCGGTTTTCGAGCGCTCTCGCGGCGTGACCATCGGCGCCGGCGCGGCGGCTGGACTGGCAGGG
>GCA_000317935  13      copy:16(16-16)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_13|Segment151|2791796-2792191|1|length:396|1-396
CGTCACCCCGTTGATATCGCCCCAGCCAGGCAAGCCGCTCCCAGGCGCGGGCCGCCGCATTGCTCACACCACAAGCGTCGGCGCGGCACGTGCAGAACGACCCGATACTCGAATAACGGCAGATCTCGCACCCGTCGAACCGTGGTCTCATGCACCTGGCGACACCGCGCGCCGCACTG
>GCA_000317935  7       copy:3(2-3)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_13|Segment138|2355749-2356855|1|length:3071|-232
TGGCGGGCGACATCCGGGTGCAGGCTGCCGACAAGGCGGGCAACCGAGAGCCCCGAGACCACTCGCGCCTATGCGGATGCCGGACCATCGCCAATGTGGACCACCGACGCGAACGACGGCCGCATCACGGTGACGGGCATGGCCGAGCCTGGCG
>GCA_000475845  27      copy:2(2-3)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_3|Segment130|4271267-4271686|1|length:340|36-328
GTACTGGTTGTCGTTACCCAGCACCAGCGTGCCCGCGCCTTCCTTGATCAAGGCGCCTGCGTGCCGCCGATCAGGCCTTGCCAGTCCAGGCTCGCGCCGCTGGCCACCGACATCGCCCCGTTGCCGGCGCCGATGGTCACGCCGCGAGCGCTCGAAAAACCCGCCGATGCCGCCACGCGT
>GCA_000317935  21      copy:2(2-2)|UP2_input_P98_BPP_S1_V7_Bordetella_parapertussis_13|Segment125|2260526-2261148|1|length:623|1-423
GGCCAGGTTCGCGCCCGGCCAGGTTCGCGCGGGCCAGGTTCGCGCGGGCCAGGTTCGCGCCGGGCCAGGTCCGCGTCGGCCAGGTACGCGTCGGCCAGGTTCGCGCCGGGCCAGGTTCGCGCGGGCCAGGTTCGCGCGGGCCAGGTTCGCG
```

Mycoplasma pneumoniae

ClusterID	Distribution	Marker	Percent_of_strains	Weighted_Average_Copy
Cluster22	11(10)|	*	100.0%[10/10]	11
Cluster164	11(10)|	*	100.0%[10/10]	11
Cluster141	11(10)|	*	100.0%[10/10]	11
Cluster9	10(2)|11(8)|	*	100.0%[10/10]	10.8
Cluster472	10(2)|11(8)|	*	100.0%[10/10]	10.8
Cluster468	10(2)|11(8)|	*	100.0%[10/10]	10.8
Cluster188	10(2)|11(8)|	*	100.0%[10/10]	10.8
Cluster187	10(2)|11(8)|	*	100.0%[10/10]	10.8
Cluster182	10(2)|11(8)|	*	100.0%[10/10]	10.8
Cluster0	10(2)|11(8)|	*	100.0%[10/10]	10.8
Cluster66	10(10)|	*	100.0%[10/10]	10
Cluster16	10(10)|	*	100.0%[10/10]	10
Cluster92	9(2)|10(8)|	*	100.0%[10/10]	9.8
Cluster512	9(2)|10(8)|	*	100.0%[10/10]	9.8
Cluster43	8(1)|10(9)|	*	100.0%[10/10]	9.8
Cluster446	8(2)|10(8)|	*	100.0%[10/10]	9.6
Cluster421	9(10)|	*	100.0%[10/10]	9
Cluster34	9(10)|	*	100.0%[10/10]	9
Cluster6	8(10)|	*	100.0%[10/10]	8
Cluster51	8(10)|	*	100.0%[10/10]	8
Cluster457	8(10)|	*	100.0%[10/10]	8
Cluster413	8(10)|	*	100.0%[10/10]	8
Cluster35	8(10)|	*	100.0%[10/10]	8
Cluster299	8(10)|	*	100.0%[10/10]	8
Cluster198	8(10)|	*	100.0%[10/10]	8
Cluster19	8(10)|	*	100.0%[10/10]	8
Cluster18	8(10)|	*	100.0%[10/10]	8
Cluster13	8(10)|	*	100.0%[10/10]	8
Cluster241	7(1)|8(9)|	*	100.0%[10/10]	7.9
Cluster156	7(1)|8(9)|	*	100.0%[10/10]	7.9
Cluster420	7(2)|8(8)|	*	100.0%[10/10]	7.8
Cluster189	7(2)|8(8)|	*	100.0%[10/10]	7.8
Cluster149	7(2)|8(8)|	*	100.0%[10/10]	7.8
Cluster134	7(2)|8(8)|	*	100.0%[10/10]	7.8
Cluster21	6(1)|7(1)|8(8)|	*	100.0%[10/10]	7.7
Cluster27	7(8)|8(2)|	*	100.0%[10/10]	7.2
Cluster82	7(10)|	*	100.0%[10/10]	7
Cluster67	7(10)|	*	100.0%[10/10]	7
Cluster54	7(10)|	*	100.0%[10/10]	7
Cluster532	7(10)|	*	100.0%[10/10]	7
Cluster509	7(10)|	*	100.0%[10/10]	7
Cluster388	7(10)|	*	100.0%[10/10]	7
Cluster257	7(10)|	*	100.0%[10/10]	7
Cluster247	7(10)|	*	100.0%[10/10]	7
Cluster239	7(10)|	*	100.0%[10/10]	7
Cluster232	7(10)|	*	100.0%[10/10]	7
Cluster230	7(10)|	*	100.0%[10/10]	7
Cluster204	7(10)|	*	100.0%[10/10]	7
Cluster173	7(10)|	*	100.0%[10/10]	7
Cluster125	7(10)|	*	100.0%[10/10]	7
Cluster102	7(10)|	*	100.0%[10/10]	7
Cluster248	6(1)|7(9)|	*	100.0%[10/10]	6.9

```
>GCA_002090295.1        314     copy:2(2-2)|95688|Segment371|241910-242027|1|length:118|1-118
TTAGTGATGAAATTTAAGTATGGTGCCATTGTTTTCAGTGGTCTTTTAGGAGTCTCTGCCATTTTAGCTGCTTGTGGTACACGCGGTAAATTTGACCAAATTGATGATGGCAAAATTA
>GCA_002090215.1        254     copy:2(2-2)|95634|Segment753|57396-58562|1|length:1167|1-1167
AAACATTAAATTTTCTTTGTTGGTTGTAGTTGTTAGGTTGTTTAGTCAAAGATCCACCTTTTCGTAATAGTTAATTAAGTAATTTAGTTAGTTTCTTTTTATCCAATTTTAAACAGACTAAAAAATCCGCTTTCGGTATAAAGCGTTTCTACTAACTTTATTTGCGTG
>GCA_002090215.1        5       copy:5(4-7)|95634|Segment203|181875-182200|1|length:326|1-326
CCCCACCCCCTGAAGGCCGTGGCTTGCGACTGAAGCAAATTCACAAGGACCTCCCCAAATGATCCGCCTCGATCCTGATTCTGTACGATGCGCCTTATGCGCGCAACCGTACCGCCATTGACCGCGTTGATCACTTGGATCCCAAGGCCATGACCGCGAACTATCCGCC
>GCA_002128125.1        485     copy:4(3-6)|95827|Segment207|180148-180389|1|length:242|1-242
TTCCACCTCCACCACCCAACGTGGGGGTTCGTCAAATGAAAACAAAATCAAGGCGTTGCAGGTGGCGGTGAAAAAGAAATCCGGGAGTCAGGGCAACTCCGGTGACCAAGGCACCGAACGAGGTGGAACTTGAATCTAATGATTTAGCCAACGCCCCGATTAAACGGGG
>GCA_002090215.1        85      copy:2(2-2)|95634|Segment882|767898-767365|1|length:268|1-268
TTTGATCCATAAGGTGTACTTTCAACCTGCTAAAAGAGCGTTAACTGTTTTGTTAGGATCAGAGTTATAGCCATCTTTCAGAATGTTGATAACGTTCTTTTTGTTCAATTCACCTTTATTTGCTTCAAAGTATTTTTGCGCGCTACTTAAACTCGATATAACCTGGC
>GCA_002090215.1        289     copy:5(2-6)|95634|Segment302|195605-195814|1|length:210|1-210
TCTGATCTAAAAAGTCTTCTTAAAGAAACAATTAACAAGATATTGATTTATCTAAAACAATCAAAACTTCAGAAGGCAAACAACAGATAATAACTTCTTTAAAAAAGTCTTACGAAGTTAATCCGAAAGATACAACAAAATTGTTGTTAGACGCATGAAAACAGAGC
>GCA_002090215.1        16      copy:8(3-8)|95634|Segment207|183164-183300|1|length:137|1-137
CAAACTGATCCCACTTCTCCCCCACGCTTTCCCGTTTCTCCACCCGGGTTCAACCTTGTGGGGTCGGTGCTCGACCAGGTGTTGGATTATGTGCCCTGGATTGGGAATGGGTACAGGTATGGCAATAACCACCGGGGC
>GCA_002090215.1        531     copy:5(5-6)|95634|Segment585|443515-443764|1|length:250|1-250
ACCTCGCCGGGCGTTGTACGACCTCGACTTTTCGAAGTTAAACCCGCAACGCCCACGCGCGACCAAACCGGGCAGATCACCTTTAACCCCTTTGGCGGCTTTGGTTTGAGTGGGGCTGCACCCCAACAGTGAAACGAGGTCAAAAACAAGGACCCCGTCGAGGTGGC
>GCA_002090215.1        204     copy:8(7-8)|95634|Segment254|186418-186543|1|length:126|1-126
ACAACCCCGGACCCAAAGCCGTCAACCGGCTTTAAGCTCGATAAGGGCCGCGCGTACCGGAAGCTGATGAAAGTTGACCGGTGTATGAACCCCTGGATTCGACCAAGGAGGGGGAAGGGGGAAGGATG
>GCA_002090215.1        372     copy:8(5-8)|95634|Segment436|337072-337208|1|length:137|1-137
ATTTTGTTGTCACATATAAAGTCTTAGTTCAATTTATTTAGTGGAGGAGCAAAATGCAAATCTTTAATTTTTGAAATTGATCAGTAAAGAGTCTGATCCCTGTTGACAACCGAGTTGCAAGTGTTAGTTATTTTTTA
>GCA_002090215.1        20      copy:7(6-7)|95634|Segment101|132324-132467|1|length:144|1-144
CCCAAAAAAACCTCACCAACAACCCCGGACCCAAAGCCGTCACCGGCTTAAGCTCGATAAGGGCCGCGCGTACCGGAAGCTAAATGAAGCTTGACCGGTGTATGAACCCCTGGATTCGACCAAGGAGGGGGAAGGGGGAAGGATG
>GCA_002090215.1        216     copy:2(2-4)|95634|Segment793|608604-608722|1|length:119|1-119
AGCAAATTAAAGAACTTCAAATCGAGCAAAAAGCTCAGGGCGAAACCTTAAAGCTAATCTTGCAAACGCTGCAAAAGATGAGCGATCGCTTGGACAAGATGGAAGTAAAAATGGACAAA
>GCA_002090215.1        18      copy:8(4-8)|95634|Segment224|183897-184033|1|length:137|1-137
CCAAAAATGGTGCCTACACCGACTTACATTCGGACCAAACCAAACTGAACCTCCCCGCTTACGGTGAGGTGAATGGGTTGTTGAATCCGGCGTTGGTGGAAACCTATTTTGGGAACACGCGAGCGGGTGGTTCGGGGT
>GCA_002096035.1        122     copy:2(2-2)|95739|Segment882|705222-705334|1|length:113|1-113
AATTTTTAAGACTTTTGGTTTTTAACAAAGGTAAATTGTTGGTCGTTGCTTTGCTTAATTTTGCTCAAACAAATGGGTTTTATGTTGTTTAGCAAATTGAGCGTAGTACTTTTT
>GCA_002090215.1        306     copy:5(5-5)|95634|Segment310|196084-196243|1|length:160|1-160
TGGACAAGAAACTACAAGAAATTTAGAGTGAATTGAGTTGGTGCTTCAATTAGCTGATCATTAAAAGGCAAAGATGATGCTAGTGAAAAAAGTGTGAAACAATTTTTAGATAGCTATGCTAATAACACATCAGGGTACTCATCAGATATTAATCTCTTT
>GCA_002090215.1        172     copy:8(4-8)|95634|Segment250|186250-186602|1|length:353|1-353
GCTCCCTCTCAAAACCCTCCAAGACCTCCTCGTCGAACAACCCGTGACCCCTTACACCCCGAATGCGGGGTTGCTCAGGATACGGTTCATTTTGGTTCGGGTCAAGGAATCGAGTTGGAATTCCCAACGTTCCCAAAAAGGCCTTAAAA
>GCA_002128125.1        398     copy:5(2-6)|95827|Segment829|608137-609005|1|length:869|1-869
AATAAAAAGTCATCTTATCAAGGCGAAAGACGGACATCTAATGAATTGAAATCAATTACTTAATTAAATATTCCAAGTAAATAAAAGAGATTAATATCTCTTGAGTAGCCGCTGGTATTATTAGCATAGCTACTCAAAAATTGTTTCACACTTTTTTGACTAACATCAT
>GCA_003890215.1        350     copy:2(2-3)|95634|Segment174|171621-171727|1|length:107|1-107
TGAGTGGGTGGCTTGTGGGGCAGTTACCAAGCACGAGTGACGGGAAACACCTCCTCCACCAACAACCTCGCGCCTAATACTAACGGGGAATGAGGTGGTGGGGGTG
>GCA_002090315.1        482     copy:3(3-3)|95696|Segment120|164621-164778|1|length:158|1-158
AAGATAAGATTAATAATTGTCATGTTTGGGAAGAAGAGTTACCTGATGGTAGCTACGACATGGGATTTAATGGCAATTTCAACCATATGGAAAAACGAAAAAGTGGTTATGTTACCCAAAAGCAGTTTAGCGAGTTCAAAGATGCCAACAATCAGCGT
>GCA_002090215.1        259     copy:2(2-2)|95634|Segment116|15256-15395|1|length:140|1-140
ATCCTGATACGCCAGATGTTTCTGAAGGCCCTAAAGAACAAGGTGGTCAGAAAGATCCAGGTGGTCCGGATGATCCCAGTGAAGGAAAGCAGAAAGTGTCCAAACCAGATAAATACGTAACCCACCGTGAGCTTGATGAA
>GCA_002095995.1        235     copy:2(2-2)|95715|Segment48|125974-126227|1|length:254|1-254
ACAGTACACCAAATGCTTTGTTTACCTCAGTTTTTGCTGCAGGTGATAGTAACTTTGATAACTTCTTTTATAAGGTTAAAGATGGTCGCGCTGACTTTAGTAACTTTAACGAAAACGGTACCTCTTACAAAAATCTGGAAAAGGTTTTTAATGACTATAAAAAACTCA
>GCA_002090215.1        209     copy:5(2-5)|95634|Segment589|444026-444307|1|length:282|1-282
AACTTGAAACCAACAACCTCGCCAACCGCCCCGATTAAACGGGGCTCCAATAACAACGCAAGTCCAACTCAAGGCGGAGCGATTTTGGTACTGCCCCTTCCAGTTCGGGATCAGGCACCCAAGATGGCACCCCCACCCCCTGAACGCCGTGGTTAACGACTGAGCAAA
>GCA_002090215.1        346     copy:6(5-6)|95634|Segment106|132918-133042|1|length:125|1-125
CCTCCACGTCCCAACTCGCTTACTTAAATGGCCAGATCGTTGTGATGGGTTCAAACGCGGTACCGAGTTTGTGGTATTGGGTCGTGGATGAGCGAACCACGTCCGGGGAGGGCGACTTGATGAGCG
>GCA_002090215.1        352     copy:4(4-4)|95634|Segment206|182228-182343|1|length:116|1-116
GCTTTCAAAAGGAAGCTGACTCCGACAAGTCGGCCCCGATCGCCCTCCCGTTTGAAGCGTACTTCGCCAACATTGGCAACCTCACCTGGTTCGGGCAAGCGCTTTTGGTGTTTGGT
>GCA_002090215.1        198     copy:8(6-8)|95634|Segment439|337101-337224|1|length:124|1-124
CAATTTATTTAGTGGAGGAGCAAAATGCAAATCTTTAATTTTTGAAATTGATCAGTAAAAGAGTCTGATCCCTGTTGACAACCGAGTTGCAAGTGTTAGTTATTTTTTAAAAAATCAAAAGAAAG
>GCA_002090215.1        51      copy:7(6-8)|95634|Segment210|183172-183455|1|length:284|1-284
TCCCACTTCTCCCCCACGCTTTCCCGTTTCTCCACCGGGTTCAACCTTGTGGGGTCGGTGCTCGACCAGGTGTTGGATTATGTGCCCTGGATTGGGAATGGGTACAGGTATGGCAATAACCACCGGGGCGTGGATGATATAACCGCGCCTCAAACCAGCGCGGGGTCG
>GCA_002128125.1        392     copy:3(2-3)|95827|Segment366|248302-248490|1|length:189|1-189
ATGAAATTTAAGTATGGTGCCATTGTTTTCAGTGGTCTTTTAGGAGTCTCTGCCATTTTAGCTGCTTGTGGTACACGCGGTAAATTTGACCAAATTGATGATGGCAAAATTAAGCTAGCATCTTCGTTAACTTCCAAAGGTGCCGCTAATGCTTTACAAGCAATTGTG
>GCA_002090215.1        170     copy:7(2-7)|95634|Segment104|132897-133573|1|length:677|1-677
TCATCACCCAGCTTTATCATACCTCCACGTCCCCAACTCGCTTACTTAAATGGCCAGATCGTTGTGATGGGTTCAAACGCGGTACCGAGTTTGTGGTATTGGGTCGTGGATGAGCGAACCACGTCCGGGGAGGGCGACTTGATGAGCGAAGACACATTTGAACTTCGGCA
>GCA_002090215.1        49      copy:4(2-5)|95634|Segment534|439818-440123|1|length:306|1-306
AATAACAAAAGTGCCAGCAATTAATCTATAAAGGTTAGTCGCTGGTATTTTCATTTTATTTGAAATAAATAAAAAAGCCATCCTATCAAGGAGAAAGACAGACAGCTAATGAATTGAAATCAATTACTTAATTAAATATTCCAAGTAAATGAAAGAGATTAATATCTGAT
>GCA_002090215.1        416     copy:7(6-7)|95634|Segment409|249307-249458|1|length:152|1-152
CCCGCACCCTCGAAGGCCGTGGCTTGCGACTGAAGCAAATTCACAAGGACCTCCCCAAATGATCCGCCTCGATCCTGATTCTGTACGATGCGCCTTATGCGCGCAATCGTACCGCCATTGATCGCGTTGATCACTTGGATCCCAAGGTGATGA
>GCA_002090215.1        65      copy:6(5-8)|95634|Segment361|245824-245961|1|length:138|1-138
TCGGCGTTGGTCTCAAAGCCAACGTCCAAGCCACGCTCGGTGGCAGTCAGACGACGACTACAGGCGGCTCCAAATGATCCACCCTCGACAAAGCCAACCTCCAGCCCTGAACGGGGGCGGGGTGAAGGAATGATAAGA
```

```
>GCA_002090215.1        131     copy:6(5-6)|95634|Segment664|498697-498892|1|length:196|1-196
CCAACTCGCTTACTTAAATGGCCAGATCGTTGTGATGGGTTCAAACGCGGTACCGAGTTTGTGGTATTGGGTCGTGGATGAGCGAACCACGTCCGGGAGGGCGACTTGATGAGCCCACACCGAGTTGAACTGGGGCACCGACAAGCAGAAGCAGTTTGTGGAGAATCAGCTTGGTTTTA
>GCA_002090215.1        280     copy:7(2-8)|95634|Segment209|183169-184269|1|length:1101|1-1101
TGATCCCACTTCTCCCCCACGCTTTCCCGTTTCTCCACCGGGTTCAACCTTGTGGGGTCGGTGCTCGACCAGGTGTTGGATTATGTGCCCTGGATTGGGAATGGGTACAGGTATGGCAATAACCACCGGGGCGTGGATGATATAACCGCGCCTCAAACCAGCGCGGGGTCGTCCACGCGG
>GCA_002090215.1        301     copy:4(3-4)|95634|Segment853|614897-615110|1|length:214|1-214
TTAGTGTTTATGAAATTTAAGTATGGTGCCATCTTTTTTAGTGGCTTTCTAGGTCTTTCTGCCATTTTGGCTGCCTGTGGTACAAAGGGTAAATTTGATCAAGTTGATGATGGCAAGATTAAATTGGCTTCCTCTTTAACTTCTAAAAGTGCATCAAAAGCCTTACAAGCAATTGTTAA
>GCA_002090215.1        526     copy:2(2-3)|95634|Segment128|166677-166798|1|length:122|1-122
AACTTGAATCTAATGATTTAGCCAACGCCCCGATTAAACGGGGCTCCAATCCAGCAAGTCCAACGCAAGCCGAGCTTGCGCACCACCCCATCCAGTTCGGAATCTGGTCAATCAGGCAC
>GCA_002090295.1        151     copy:6(6-7)|95688|Segment757|569407-569520|1|length:114|1-114
TCGTAAGACTTTTTTAAAGAAGCTATTATGTTCTTTTTGCCTTCTTTAGTTTTAATTGCTTTAGACAAATCAATGTCTTTACCAGTGGTTTCTTTAATAAGATTTCTAAGATCT
>GCA_002128125.1        395     copy:6(6-7)|95827|Segment344|196144-196381|1|length:238|1-238
AAGTACTGGAATCGCTACCCCGTGCTGCTGGTGGTGTTTCAGTCCAAGCAGCGACCCGGCAATCGTCTTCCACAAATGAAGATCTACCGAATGTAATCACCCAGTTATACCATACTTCAACCGCCCAACTCGCTTACTTAAATGGCCAGATCGTTGTGATGGGTTCGAACGCGGCGAACGTCCGTGCGGAACGTCCGA
>GCA_002090235.1        388     copy:4(4-5)|95658|Segment771|588501-588658|1|length:158|1-158
GCTGATTGTTGGAATTTTTGAAATCATTAAACTGCTTTTGGGTAACGTAACCACTTTTTCGTTTTTCCATATGGTTTAAATTGCCATAAAAGCCCATGTCATAGCTGCCATCAGGTAGATCTTCGATTCAGGCTTTTGATGTTTTAATCTTATCTTTT
>GCA_002128125.1        393     copy:5(4-6)|95827|Segment440|320322-320669|1|length:348|1-348
CTTACGAAGTTAATCCGAAAGATACAACAAAATTGTTGATAGCGCCTGAAAACAGAGTTTTGAAGAGGGTAAATTAGGTATTCGGGATTTTGATTTTGACCATGTAGCTTATCCTCAAACAAATGATCCTTTCACGATGGAGCGAAAGGTTGATCATTTTCAAATGACATACCAAAGT
>GCA_002090215.1        391     copy:5(4-5)|95634|Segment311|196084-196273|1|length:190|1-190
TGGACAAGAAACTACAAGAAATTTAGAGTGAATTGAGTTTGGTGCTTCAATTAGCTGATCATTAAAAGGCAAAGATGATGCTAGTGAAAAAAGTGTGAAACAATTTTTAGATAGCTATGCTAATAACACATCAGGGTACTCATCAGATATTAATCTCTTTAGTTATTTGGATATTTAA
>GCA_002090215.1        120     copy:6(6-6)|95634|Segment197|181199-181476|1|length:278|1-278
CCACCTTTGCGTTACGCCGGTATGACCTCGCCGGCGCGCCTTATACGACCTCGATTTTTCGAAGTTAACCCGCAAACGCCCACGCGCGACCAAACCGGGCAGATCACCTTTAACCCCTTTGGCGGCTTTGGTTTGAGTGGGGCTGCACCCCAACAGTGAAACGAGGTCAAAAACAAG
>GCA_002090215.1        92      copy:9(2-9)|95634|Segment179|178613-178877|1|length:265|1-265
TTATTTTGAATCTTTGTAAGTCGTTGGTCGTTTGATCTTTGTACTCGTTAAACTCTTTGCGCGTTACGTATTCCGTGTTCACTGGTATAACCGGTTTGTTAAGTTTAAATTGTCTATTGAAATTGCCAGAGAATTCCACACTTTTGTTTTCTGGATTAATGATGTATCAGCCCGAAAA
>GCA_002090215.1        54      copy:7(6-7)|95634|Segment202|181875-182021|1|length:147|1-147
CCCCACCCCCTGAAGGCCGTGGCTTGCGACTGAAGCAAATTCACAAGGACCTCCCCAAATGATCCGCCTCGATCCTGATTCTGTACGATGCGCCTTATGCGCGCAACCGTACCGCCATTGACCGCGTTGATCACTTGGATCCCAAGGC
>GCA_002090295.1        481     copy:5(5-5)|95688|Segment656|495523-495626|1|length:104|1-104
CTTGGATACCAGTGCAAGCGCTTAAAATTGTTCCACATAACAAAAATAAAGAACCAAAAGCAAACCTTGACAAAAATTTCTTTTTGCGCATCAACAGTAAAAAT
>GCA_002128125.1        489     copy:4(3-5)|95827|Segment11|114681-115064|1|length:384|1-384
ACACCGCCCCAGGCCTTGCACCAGATGGGGGTGATAGTGCCATCTTTGGAAACGTGACCTGGTAAACCAAGTACTGGAATCGCTACCCGTGCTGCTGGTGGTTTTCAGTCCAAGCAGCGACTCGGCTCGACTTCTACAAATGAAGATCTGCCGAATGTAATAACAAAGTTATACCACACC
>GCA_002128125.1        547     copy:3(2-4)|95827|Segment866|702593-704077|1|length:1485|1-1485
CAAAGTCCCTAGCAGCAAAAGCGGTTTTGGGAATTACTGTGGTTTGGACATTAACAAAGGCTTGAGCATTATTGGGCAACTTAACTTGACTAGGTTCGACAAAATCACGAACGTGAGTGTATTTGCCTATTCGAAAAGCGGTCGTTCAATCGGCACTGGTATCAGAGTTTTTATAAGGA
>GCA_002090215.1        117     copy:5(2-7)|95634|Segment645|497041-499261|1|length:2221|1-2221
CCTGAACCCCGTGGTTAACGACTGAGCGAAATTCACAACGACCCGCCAAATTCGCCGCCTCGATCCTGATCCTGTACGATGCGCCTTATGCGCGCAACCGTACCGCCATTGACCGTGATCACTTGGATCCCAAGGTGATGACCGCGAACTATCCGCCCAGTTGAAGAACGCCCAAG
>GCA_002128125.1        512     copy:8(7-8)|95827|Segment191|177185-177328|1|length:226|1-226
TTTGTTCACCTTGGGCAGCCAAAGTGGTTTCAATCTTTATGAGTCGTTGGTCGTTTGAGTCTTTGTACTCGTTAAACTCTTTGCGCGTTACGTATTCCGTGTTCACTGGTATAACCGGTTTGTTAAGTTTAAATTGTCTATTGAAATTGCCAGAGAATTCCACACTTTTGTTTTCTGGA
>GCA_002090235.1        373     copy:3(2-3)|95658|Segment463|338878-339012|1|length:135|1-135
CTTTGTTAATAAGGCGGTTCTTACTCTTCTAACTTTCAAAAGTTTCACCAGTTAGCTTACAGTATCTCCTTACCTCTGGTTTCTATTATTTCATTTGCTGGTAATAGTGCTAAGCGGCTTAAAGTTTGGTGATAA
>GCA_002090215.1        75      copy:3(2-5)|95634|Segment794|608745-609031|1|length:287|1-287
TGGAAGTAAAATGGACAAAATGGAAAAAAGAATAGACAAACTTGAGTCCAAATAACCAATGCTCAGAAATAAAAATACCAACAGTGAAACCTAAAGATGCGGCGCTGTTGGTATTTTTTTCAATTCACATAAATAAAAAAGCCATCTGCCCAAGGTGAAAAAAACAGACGGCTTTTTG
>GCA_002090215.1        584     copy:5(3-5)|95634|Segment898|779610-779814|1|length:205|1-205
TTTAATTGCTTTAGACAAATCAAAATCTTTGCCAGTTGTTTCTTTAAGAAGATTCCTAAGATTGGCTTGGATCCATGTGCAAGCACTTAATATTGTTCCACATAACAAAAACAAAGAACCGAAGGAAATCCTTGACAAAAGTTTCTTTTTACGCATTAACATAAAAAATAATGCATGAT
>GCA_002090215.1        195     copy:7(3-8)|95634|Segment228|184103-184665|1|length:563|1-563
TGATCACCCCCGGGTTGGCTTGAACGCGCCCCAGGACGTCGGTAACCTCGTTGTCAGTGGCACCACGGTGAGCTTCCAGCTCGGCGGGTGGCTGGTCAACCTTCACGGACTTTGTCAAACCCCGCGCGGGTTACCTCGGTCTCCAGTTAACGGGCTTGGATGCAAGTGATGCGACGCAGCGCC
>GCA_002128125.1        258     copy:3(2-3)|95827|Segment358|190875-190213|1|length:139|1-139
AAATTTCAAGGGTTAAACATCTACAAACTTTGTTAATATTTCTCTAAATTTCGGGTTTCTAGTGTTTATGAAATTTAAGTACGGCGCTATTGTTTTTTGGGCGTTTCGGCCATTTTAGCTGCCTGTGGG
>GCA_002090235.1        94      copy:4(2-5)|95658|Segment223|182187-182389|1|length:203|1-203
CCGGGTTTGATGTGGATAACTCTGAAAACACCAAGCAGGGCTTTCAAAAGGAAGCTGACTCCGACAAGTCGGCCCGGATCGCCCTCCCGTTTGAAGCGTACTTCGCCAACATTGGCAACCTCACCTGGTTCGGGCAAGCGCTTTTGGTGTTTGGTGGCAATGGCCATGTTACCAAGTCG
>GCA_002090215.1        152     copy:3(2-9)|95634|Segment152|170338-170519|1|length:182|1-182
CGAAACTACACCAATTTCGCGAGTGCTAAGGGCAGTACGAACCAGCAGGGTTCAACAACGGGTGGATCAGCGGGCAACCCCGACTCGTTAAAGCAGGATAAGGCTGATAAAAGTGGTGATTCGGTAACGGTTGCAGAAGCCACATCGGGCGACAACCTGACGAATTACACCAACCTCC
>GCA_002090215.1        534     copy:2(2-2)|95634|Segment880|739001-739376|1|length:376|1-376
AAATCTAGCTCAATTTCTGCTAAAATTCCTTGCATTAATTTAGGAATAGGGCAAGCATAATTAACAAATTTAGGTGTTTCTAATCTAAGTGCATAAGCTAAAAACTTAGGATTAATTTCATCACTTTTTACCTTAAGAATTCCACAAACATTGGTTGCGTTAAATTGACCGTTACGATA
```

Streptococcus agalactiae

ClusterID	Distribution	Marker	Percent_of_strains	Weighted_Average_Copy
Cluster99	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster89	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster87	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster86	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster81	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster80	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster74	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster7	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster68	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster19	0(1)|1(2)|3(2)|4(1)|7(1)|	*	85.7%[6/7]	2.714285714
Cluster67	0(1)|1(2)|3(3)|4(1)|	*	85.7%[6/7]	2.142857143
Cluster53	0(1)|1(2)|3(3)|4(1)|	*	85.7%[6/7]	2.142857143
Cluster91	2(7)|	*	100.0%[7/7]	2
Cluster77	2(7)|	*	100.0%[7/7]	2
Cluster75	2(7)|	*	100.0%[7/7]	2
Cluster63	2(7)|	*	100.0%[7/7]	2
Cluster56	2(7)|	*	100.0%[7/7]	2
Cluster51	2(7)|	*	100.0%[7/7]	2
Cluster30	2(7)|	*	100.0%[7/7]	2
Cluster27	2(7)|	*	100.0%[7/7]	2
Cluster25	2(7)|	*	100.0%[7/7]	2
Cluster45	0(1)|1(1)|2(4)|3(1)|	*	85.7%[6/7]	1.714285714
Cluster32	0(1)|1(1)|2(4)|3(1)|	*	85.7%[6/7]	1.714285714
Cluster64	0(1)|1(2)|2(4)|	*	85.7%[6/7]	1.428571429
Cluster31	0(1)|1(2)|2(4)|	*	85.7%[6/7]	1.428571429
Cluster15	1(4)|2(3)|	*	100.0%[7/7]	1.428571429
Cluster6	0(1)|1(3)|2(3)|	*	85.7%[6/7]	1.285714286
Cluster2	0(1)|1(3)|2(3)|	*	85.7%[6/7]	1.285714286
Cluster26	0(1)|1(5)|2(1)|	*	85.7%[6/7]	1
Cluster98	0(5)|6(1)|8(1)|		28.6%[2/7]	2
Cluster97	0(6)|2(1)|		14.3%[1/7]	0.285714286
Cluster96	0(6)|26(1)|		14.3%[1/7]	3.714285714
Cluster95	0(6)|26(1)|		14.3%[1/7]	3.714285714
Cluster94	0(6)|26(1)|		14.3%[1/7]	3.714285714
Cluster93	0(6)|2(1)|		14.3%[1/7]	0.285714286
Cluster92	0(4)|2(1)|8(1)|10(1)|		42.9%[3/7]	2.857142857
Cluster9	0(5)|8(1)|10(1)|		28.6%[2/7]	2.571428571
Cluster90	0(6)|2(1)|		14.3%[1/7]	0.285714286
Cluster88	0(6)|2(1)|		14.3%[1/7]	0.285714286
Cluster85	0(6)|2(1)|		14.3%[1/7]	0.285714286
Cluster84	0(6)|2(1)|		14.3%[1/7]	0.285714286
Cluster83	0(6)|2(1)|		14.3%[1/7]	0.285714286
Cluster82	0(6)|2(1)|		14.3%[1/7]	0.285714286
Cluster8	0(2)|1(1)|2(1)|5(1)|10(1)|11(1)|		71.4%[5/7]	4.142857143
Cluster79	0(5)|4(1)|16(1)|		28.6%[2/7]	2.857142857
Cluster78	0(6)|9(1)|		14.3%[1/7]	1.285714286
Cluster76	0(5)|1(1)|2(1)|		28.6%[2/7]	0.428571429
Cluster73	0(6)|26(1)|		14.3%[1/7]	3.714285714
Cluster72	0(5)|3(1)|16(1)|		28.6%[2/7]	2.714285714
Cluster71	0(5)|3(1)|20(1)|		28.6%[2/7]	3.285714286
Cluster70	0(6)|16(1)|		14.3%[1/7]	2.285714286
Cluster69	0(6)|3(1)|		14.3%[1/7]	0.428571429

```
>GCA_001275545.2        48      copy:5(2-6)|62292|Segment23|473548-475020|1|length:1473|1-1473
AGTACCGGATAAAGATAAATATGATCCAACAGGAGGGGAAACAACAGTACCCCCAAGGGACACCAGTTTCAGATAAAGAAATCACAGACTTAGTTAAGATTCCAGATGGCTCAAAAGGGGTTCCGACAGTTGTTGGTGAACGTCCAGATACTAACGTTCCTGG
>GCA_002073795.1        60      copy:4(3-5)|86319|Segment7|1366977-1367842|1|length:866|1-866
CATCTTCGGCTGCTTTCACGGCTGCTTCAGCTGCGGCTTCTGCGGCATCTTGGGAATCTGGTTTGCCGTTGCTATCCGCATCGTTGACGGTTACTTCTGATGTCGTTACTTGATCAAGACGTGCTTTCAACGCTTCTTTGACTGGGCCTTCTGGTAAGCTATC
>GCA_001275545.2        16      copy:5(4-6)|62292|Segment25|473553-474774|1|length:1222|1-1222
CGGATAAAGATAAATATGATCCAACAGGAGGGGAAACAACAGTACCCCCAAGGGACACCAGTTTCAGATAAAGAAATCACAGACTTAGTTAAGATTCCAGATGGCTCAAAAGGGGTTCCGACAGTTGTTGGTGAACGTCCAGATACTAACGTTCCTGGAGATCA
>GCA_001932715.1        91      copy:2(2-2)|76649|Segment1|1193719-1194930|1|length:1212|1-386
TATACCAAGTCCTGCAAGAAGCTCTCTCCAATCATCGCAGTATCAACACCAAAGTTATCCGCAACTGTTGCTGAAATATCTGCGAAATGCCCTACTGGGATAAGTCCGTTTCCTGTGAAAGATGGGCTATAGGCTAAGAGAAGGAATGTATTCTCTTGTATGG
>GCA_001932715.1        6       copy:2(2-2)|76649|Segment6|613235-613802|1|length:568|1-568
TATTTTTCTAACTTTTATAAATGTTAAATTCTCACGACAGATGTGATATATTTTGCTATTTTTTCAGGACTTTCAGTAAAATTATTATCTATCCAATTTAAAATGATACTTTCAATTGTTGAGATATAAAATATCAAGAGCGAAAGAATATGGAATATTGATAT
>GCA_000736485.1        29      copy:19(2-19)|17821|Segment1|1019400-1020952|1|length:1553|429-887
CTCACGATAGCCATCTCGAGTAGTCGTTGAATAGGATAAGGGGTGATTTTCTGGACAGAGGTAAACATCATAATACTCATCATAAATAAAGTCTTTTAGCCGAAGCATACCCCTTTTTACTACGAGGTCTAGTATAAGAAAAAACTGGTATGATACCTATAGT
>GCA_001275545.2        14      copy:2(2-2)|62292|Segment17|1217486-1218585|1|length:1100|477-1100
GTTTGTTTAGATCCCGATTCCCCGTCGGACTTTTTTTGCTATCCTCCTGTGATATAATGACATTATCATTACTGAAAGGAGGGATAAGTTATGAATTGGAATCAAATCATCGTTACTTTTTTAACTGCCTGTGTGCCTGCTTTGATTGCTTATTTAACCAGTCA
>GCA_001275545.2        42      copy:2(2-2)|62292|Segment5|115087-122282|1|length:7196|6357-7196
CTTCTTAATTTATAATATAATTATACAAGAAAATCTTGTATTTGTAAAGTATTTTTAAAATATTATTTGTTTTTTAAAAAGAATTTGTTTTTAAAATGAATTTTATTGTAATATAAAAATAAAAAGGAGATTAGAAATGGAATTGTACGAATTTGTCGGAGA
>GCA_002073695.1        51      copy:2(2-2)|86334|Segment1|1304259-1305470|1|length:1212|514-815
CTCATGAGGATATCATTCCTCTAGAAGAACTCTACCGTATTTGTGAATATGCTCGTAGTATCACAATGGAGCGTCCAGCTCTCCTTGGTCGTATTATCGCTCGTCCATATGTAGGTGAACCAGGCAACTTTACGCGTACAGCCAACCGTCACGACTATGCAG
>GCA_002073695.1        21      copy:13(10-23)|86334|Segment12|1370852-1374258|4|length:3399|473-777
AACTAGATGCGGTTACGTATTGACTAGAGGGAAGGTTCACCAATGTAAGATATCATGCAGACAACGAAAAAAGAAAGGAATACGCACATGTCAGAATTGCTAGATAAGATATTATCTCGGAACAATATGCTCGAAGCTTACAAGCAAGTGAAATCAAATAAAGC
>GCA_002073695.1        83      copy:2(2-2)|86334|Segment259|451176-459802|1|length:8627|8146-8562
AGAAGTTAAAACACCAACTGAGACGTTTAAATGGAATCAACTTGGAGAAGCTATATCAGCATTTTACTCAGACCTTACGCTAGTCTAACCATCAAAGAACTTACGGCTCAACTCAATGATAATGTAGATGCTATAGGTAACATGATTTGATTCAATGAGCGACGA
>GCA_001275545.2        76      copy:2(2-2)|62292|Segment1|110675-111018|1|length:344|1-344
AAACTTACAAACGCAAAAACGGTTTTGAAACAGAGTATAACTTTTTTGCTCTTGACGATGACGGAGTTGGTTGGCGCTTTTATAAGCAGTTTAAATTTCAAGCGGATAGGAACTATCACAATCCTTTCTCAATGAACAGAAGTAGAGCAGTTGAGATTTTCCC
>GCA_001275545.2        31      copy:2(2-2)|62292|Segment19|1910656-1910959|1|length:304|1-304
CCTGTCATCTCATTTTTTTGAAAGTCGATAAACTCAATTTTTTTTATTTGATTACCATTTGTTAACTCAACATGTTGTTTTATGTAATTGACAATTCTATCTTGTTCCTGTTTCCTTGATGATGCTTCTGTTGTTGCAAATAAGTTTTTTCATGGCTAGTCCTC
>GCA_001275545.2        84      copy:2(2-2)|62292|Segment9|1184927-1192122|1|length:7196|2967-7196
TTTAGTGATTGCACGCATGTCCCCTCCCTATTTTTTTGGCTCGTGGTAAGTCAATGCTTGCTCACTGTCTGAAAGACCTTCGGTTGTTGGGTCTGTCACAACTCCAAGTAATACCAAAAGCGTTACTGCTGTGTTTTGCAATATCCGCAATATTTGACGGTAAT
>GCA_002025025.1        34      copy:16(15-16)|82762|Segment43|807501-809986|1|length:2486|1-2486
AGCAACGGCAAACCAGATTCCCAAGATGCCGCAGAAGCCGCAGCTGAAGCAGCCGTGAAAGCAGCCGAAGATGCCGCCCAAGCAGGCAAAGATAAGAAAGCTGAAGTTGAAGCAGATGGTGTTGTTAATCCAGATGAGAAATCAGCCGTAGATGGCTTGAACC
>GCA_001275545.2        56      copy:2(2-2)|62292|Segment21|1962925-1964136|1|length:1212|396-860
ACCATGACTATTTGATTTGTTATGTCCCATATCATGATTAATGCCTGAACCATTAAAAATATCATTGATTTTACCAACAGCATAAGTATCGATGCCAGCTTGATCTAGTTTATTTAATACGTATCTTCAAATGGTGATACTGCATAGTCTGTGACGATTGGC
>GCA_000736485.1        26      copy:2(2-2)|17821|Segment89|1467469-1467840|1|length:372|1-372
CTACGCATGCTCCGCTGATTCTGTAGGATTAGAATGAGATGCATCTGTTGATAACACTGCTGAAGTTGACGTATCTGTCGCAAGACTTGTTACATCTGAAGTAGGCTGTGTAACAGTAACTTCCGGTTTTACACCAGCTGTTTGTGGTACCTCAGTACTCA
>GCA_001275545.2        12      copy:2(2-2)|62292|Segment10|1192531-1193219|1|length:689|1-689
TACGCCTCTACTGTCACGTTCAAAACGTGGTTAAAGTCGCTTGCCTTAACGATTAGCGATGGCCCAACATCAATTAGCGAGTCACCATTTTTATAAAAATAAACAGCATCGTAATCTTTCCCATTTTTTTGCAGGTTAGGCGTTAATACAGACTCACCAACCC
>GCA_002025025.1        49      copy:16(15-16)|82762|Segment44|807501-810574|1|length:3074|1-3074
AGCAACGGCAAACCAGATTCCCAAGATGCCGCAGAAGCCGCAGCTGAAGCAGCCGTGAAAGCAGCCGAAGATGCCGCCCAAGCAGGCAAAGATAAGAAAGCTGAAGTTGAAGCAGATGGTGTTGTTAATCCAGATGAGAAATCAGCCGTAGATGGCTTGAACC
>GCA_001275545.2        41      copy:2(2-2)|62292|Segment4|115087-122282|1|length:7196|4261-6349
CATACCGCTTGCAGTGGTTGCTTTTTTTAGTTATCCGATTTATTTACTTTTTGGAAAGGAGGAGTAAATGGCAACTTATCAAGAATATAAAAGCAGGTCAAATGGTAACGCTTATGATATTGATGGTTCGTTCGGTGCACAATGCTGGGACGGTTATGCAGAT
>GCA_001275545.2        55      copy:2(2-2)|62292|Segment15|1211131-1211614|1|length:484|1-484
TATCCCCCATTTCCAGTCAGCTCAGCAATCCGTTTGTCTGTCTAGCTCTATCACTAGCACGTTTAAGCTGCTTTTGTGTCCGTAATTCTGTAATTTGTAGCTTGTAGTGGTCTTGCAGTGCGACGTTTAAAATAGATATAGCAA
>GCA_002073695.1        35      copy:13(5-23)|86334|Segment11|1370851-1374251|1|length:3401|1766-2357
GCCCTCGCAAAGCGAGGACTGGTCAGTTGCTTAGATTACTATCTTGAACGACATGCGTTAAAAGTTAGTTGAACCGCCGTATGCCGAACGGCACGCATGCTGGGGTGTGAGAGGGGCTAGAGATTAGCCTCTACTCGATTTCTTACCGAATTCTCCCATAAACCA
>GCA_001275545.2        19      copy:2(2-2)|62292|Segment20|1912811-1913145|1|length:335|6-335
CAAATTAAATTAACTCCTGAAGTACTCCGTGCGTCAGCACAAAAATACACAATAGGATCACAATCAATTACAGATGTGTTAACTGCTTTAACACAAGAGCAAGCAGTTATTGATGAGAATTGGGGTGGTACAGCATTTGACAGTTTTGAAGCACAGTTCAATC
>GCA_000736485.1        3       copy:21(8-21)|17821|Segment127|1605963-1607506|6|length:1544|182-694
CTCTTGCGTCAAATTGATGAAGCGATAGATTTTCATTTATTTATGACTTAGTGAAAGACAGTTATTGTGCAGATAACGGTCGTCTAGTTTGGATCCGGTCATGTTAGTAAAAATCCCCCATGATTCAATGTCTCTTTGGCATTCGTTCCATGCGTCAAACCA
>GCA_002073695.1        23      copy:8(6-10)|86334|Segment218|2095646-2097813|1|length:2168|1-2168
GTTCCATCTGGATAAGTTACATCAACCGTTACTTTATGATCTCCAGGAACGTTAGTATCTGGACGATCACCAACAACTGTCGGAACCCCTTTTGAGCCATCTGGAATCTTAACTAAGTCTGTGATTTCTTTATCTGAAACTGGTGTCCCTTGGGGTACTGTTT
>GCA_001275545.2        47      copy:2(2-2)|62292|Segment31|94465-94838|1|length:374|1-374
ATGCTTGTTGTGGAAGTCGCTTGTTTTGGTTTGACAAAAACGAAGAACACACTACTTTTATGGATATCAGACAAGAAAGTTTTGATATTCATGGGAAGCATGTAAATGTTGTTCCTGACGTAGTTGGAGATTTCAGGGATATTCCTTTTGAAGACAATACTT
>GCA_002025025.1        7       copy:4(2-4)|82762|Segment13|1576753-1578310|1|length:1558|280-902
TCAACCAAAGACTGCCTAGACTGTCCGCTACTAAGAAAAGCTCCCTATGTTTGTAATGGGTGTCCAAAGAGAAGGATCAACTGTGGTTACAAGAAGAGCCTTCTATCTTGCTAAGCAAGCTCAAAGAAACTATGAAAAACTTTTAGTTGAATCTAGAGAAGGA
>GCA_002025025.1        99      copy:4(2-4)|82762|Segment11|1537597-1539142|1|length:1546|268-729
TCAACCAAAGACTGCCTAGACTGTCCGCTACTAAGAAAAGCTCCCTATGTTTGTAATGGGTGTCCAAAGAGAAGGATCAACTGTGGTTACAAGAAGAGCCTTCTATCTTGCTAAGCAAGCTCAAAGAAACTATGAAAAACTTTTAGTTGAATCTAGAGAAGGA
>GCA_002073695.1        25      copy:2(2-2)|86334|Segment260|4989094-5001151|1|length:1212|223-815
ATCCAAGTGGCTACCGCGACGAAGTTACCAAGAAGTATCACTTGGTAAAGATACAATGACAGGGCATTGGGATACTCATGGCTTAAATATTACAGAGCCATTTGGATACTTTCTGGAATGGTTTTCCAGAAGATATCATCACAAAGATTGAGGATTTTTCAGGACC
>GCA_001275545.2        57      copy:2(2-2)|62292|Segment30|93605-94088|1|length:484|1-400
AAGCAATCGGAAACATATATCAAAACAGCGAATTACTAGAAAGCGTGGAAGGATGACAAAAGAAGAAGTAATTGCATTTCTGACAGAACAGCGTGATTTGCGGCTTATCGGATATGAGTGGGGAAAAGACAATCTGTCCGATTTTGAGAGATGGCAGTTAGCC
>GCA_002073795.1        19      copy:6(3-6)|86319|Segment15|1042791-1044132|1|length:1342|904-1290
AACTCTTTCCTGTTATCCTAACGGATAATGGTGGGGAATTCGCTAGAGTGGATGATATTGAGATAGATTTTGTGGACAGTCTCAACTCTTCTTTTGTGACCCTAATCGGTCTGACCAGAAAGCAAGAATCGAGAAGAATCATACCCCTCGTGAGAGACATAC
```

```
>GCA_000736485.1        63      copy:2(2-2)|17821|Segment91|149031-150242|1|length:1212|54-731
GCGGCACCAGATGCCAATGATTTTGTCAATGCAGGAGTTCCTGATGGTGCTTCAGACACGCTTGGCCACATTTCAAAAACAGTTGGACTTGCGGTGCCAAACATGGCTAAAATCGGACTTGGTAATATTCCTCGTCCCCAAGCGCTTAAGACCGTCCCTGCAG
>GCA_002073795.1        20      copy:2(2-2)|86319|Segment65|671478-672129|1|length:652|1-652
AGTGATACAATAGTTATACAAAAATGATTTGAGGTGATATTATGTCAACAGTAGCAGTAAGAGTTGATGATCAATTAAAAAGATGATGCTACAGAACTTTTTCAATCCTTAGGGTTGGATATGAGCACAGCAGTAAAAATGTTTTTGATTCAATCTGTTAAAAC
>GCA_002073795.1        87      copy:7(2-7)|86319|Segment53|320838-322275|1|length:1438|547-1071
CTCTTGTTATCATAGAGTGTCCCTCTTAATGACCTGTATGTGTTTAGCAGTTTCGATAGCTGTCTTAGAATCCATCAGTTTGGCAAAGATGAAGTTACAGAAGGCGACATTGAAGGTGAGAAGGACCTTTCCACCAATCCGTCCAATAACCGTGTCCATTTCA
>GCA_001275545.2        43      copy:2(2-2)|62292|Segment7|1184927-1192122|1|length:7196|1-838
TATTATAACAAAAAAGAACCTTAATATGGTATCATTTGCTATCATTGCCCAATCACTTGATAAATCAACATTTTACAACCTTACATTATCTGGTCAATTACACTGATTTTAAAAACAACGTAGTAAATAACGTAGTAATTTTTGCGCTGGGGAGCATCTGATT
>GCA_001275545.2        97      copy:2(2-2)|62292|Segment3|115087-122282|1|length:7196|1-4228
CAGATGGGACAGTTTGATTTTAGTGTGTCCGATAGCGCAAACAAGCGCTACATTGGGCAATATACTGACTACGATGCAATAGACTCAAACGACCCTAAAAAATATCGCTGGACTGATATGATTGGGACGGTAGTCGTTGGAACTAATAACCTAATCAATGGTAA
>GCA_001275545.2        13      copy:2(2-2)|62292|Segment13|1208673-1209008|1|length:336|1-336
CTTTCACGAGCGCTTTTTTATTGTTCTCCTTTCTTCTGCTGACTTATTTTTGTTGTTAAATTGTCGAGTATTAAATTTTTAGTTTTGCGTCAGCACTTTATTTGACAGATTACGCTTGTATAATCATCTGTGAGCGATAACAGACTTTAGATTTTTATC
>GCA_001275545.2        65      copy:2(2-2)|62292|Segment21|1111522-1113072|1|length:1551|1-1551
TGTTATTGCGAAAACTTAAATCTAGAGGTATTGCAATGCATATAAAGCCACTAAAGCGATGTCGTTTGAAGAGTATCTTGTAGCATTTGACATTTTAAACTGGGGAGCTTTGACAATTGGGACAAACGAAGTAAATGACAAAAAACTTACT
>GCA_001275545.2        44      copy:2(2-2)|62292|Segment16|1211756-1212448|1|length:693|160-693
GACCACGCTCTAAATTTTGGTATCATAACTCACCTCTCAAAAAATACTCTGCATCACTTTTAGCGATTAAGCTATCACGATAAACAATAGCTTCGTCTTTTGGTCTTAAACTCTTTGTCTTTATAAACAGTAGGCAACACTCGTCCGCCAATGTGCTCATAAA
>GCA_002025005.1        8       copy:11(7-15)|82755|Segment26|454751-457812|4|length:2262|1-2262
GTGACACCAAAACCAGTACCGGATAAAGATAAATATGATCCAACAGGAGGGGAAACAACAGTACCCCCAAGGGACACCAGTTTCAGATAAAGAAATCACAGACTTAGTTAAGATTCCAGATGGCTCAAAAGGGGTTCCGACAGTTGTTGGTGATCGTCCAGATA
>GCA_002025025.1        101     copy:16(12-16)|82762|Segment46|807501-812000|3|length:4500|1-4500
AGCAACGGCAAACCAGATTCCCAAGATGCCGCAGCTGAAGCAGCCGTGAAAGCAGCCGAAGATGCCGCCCAAGCAGGCAAAGATAAGAAAGCTGAAGTTGAAGCAGATGGTGTTGTTAATCCAGATGAGAAATCAGCCGTAGATGGCTTGAACC
>GCA_000736485.1        94      copy:19(19-19)|17821|Segment11|1019407-1020946|8|length:1540|422-840
CTCACGATAGCCATCTCGAGTAGTCGTTGAATAGGATAAGGGGTGATTTTCTGGACAGAGGTAAACATCATAATACTCATCATAAATAAAGTCTTTTAGCCGAAGCATACCCCTTTTTACTACGAGGTCTAGTATAAGAAAAAACTGGTATGATACCTATAGT
>GCA_002025025.1        69      copy:14(9-16)|82762|Segment41|805481-811692|1|length:6212|1-6212
CGCAGCTGAAGCAGCCGTGAAAGCAGCCGAAGATGCCGCCCAAGCAGGCAAAGATAAGAAAGCTGAAGTTGAAGCAGATGGTGTTGTTAATCCAGATGAGAAATCAGCCGTAGATGGCTTGAACGATGTCACACGTTGGAAAAGGTACAGCAACTCCGTTA
>GCA_000736485.1        0       copy:3(2-3)|17821|Segment13|1049121-1049539|1|length:419|1-419
ATGTTCTAGAGCGTCGTCAACGTGATGCGGATAACAAGAGCCAAGGCAATGTTCTAGAACGTCGTCAACGCGATGTAGAAAACAGAAGCCAAGGCAATGTTCTAGAGCGTCGTCAACGCGATGCGGATAACAAGAGCCAAGGCAATGTTTTAGAGCGCCGCCA
>GCA_002025005.1        86      copy:3(2-3)|82755|Segment10|1764234-1765774|1|length:1541|369-893
CTTCTATCTTGCTAAGCAAGCTCAAAGAAACTATGAAAAACTTTTAGTTGAATCTAGAGAAAGGAATCCCTTTGAACAAAAGAGACCTTCTGGAAAATAGATAGAGTGCTTTCTAATGGGGTCAAGAAGGGCCAACGTATCTATCATATCCTCAAAACCAACGA
>GCA_000736485.1        64      copy:2(2-2)|17821|Segment272|412175-412500|1|length:326|1-326
TCCAAGGTTCTCTTTCAGGTATTCTCAACCAACATTCAACACTTTAACAAACAAACAACGCAGCAGCGACACCGCAAGTATCAAGCCCCAAAGCAGAACGTCAGTTAACAGATTCAGAACTTTATGAAAGTGCCCAAAAACAAGTATTACCAAAATATATCC
>GCA_001932715.1        32      copy:2(2-2)|76649|Segment3|1686454-1686782|1|length:329|1-329
TGACGTTCTGCTTTGGGGCTTGATACTTGCGGTGTCACTGCTGCGTTTTGTTGTTTGTAAAGTGTTGGAATGCTTGGAAGAGAACCTTGGATATATTTTGGTAATACTTGTTTTTGGGCACGTTCATAGATTTCTGAATCTGTTAACTGAC
>GCA_000736485.1        73      copy:17(4-17)|17821|Segment108|1585294-1586839|1|length:1546|99-695
GCTATGTTCCACAAAGAAAATCCTGACTATAATCGAAATCAAGTTGGTTTCTATAGTTTAGATGAGTTGGTTCCCAAAGACCATCTCTTGCGTCAAATTGATGAAGCGATAGATTTTTCATTTATTTATGACTTAGTGAAAGACAGTTATTGTGCAGATAACC
>GCA_002073795.1        102     copy:3(3-3)|86319|Segment1|1124922-1126783|1|length:1862|405-875
TGTGTCTGTAAAACAAACGGTCTGATAGAACTAGATGCGGTTACGTATTGACTAGAGGAAGGTTCACCAATGTAAGATATCATGCAGACAACGAAAAAAGAAAGGAATACGCACATGTCAGAATTGCTAGATAAGATATCATCTCGGAACAATATGCTCGAAG
>GCA_000736485.1        27      copy:2(2-2)|17821|Segment365|924581-925792|1|length:1212|826-1212
TGATAAAAACAATGGGGTTATCAGAGTTTGAAAAAGGCTTCTCCCTTTACTAATTTGGTTGACTTTGATGCTCTCTATGGCTATCGTGATTGTCTTCATGAGTTTGATGAGCGTCTTCCAGAGATCATTTCAGCAATGCGTG
>GCA_000736485.1        2       copy:2(2-2)|17821|Segment90|1467958-1468513|1|length:556|1-556
GGTGATAGTATATTGGTAAGTTCAGATAACACTAATTTAAAATTTTTAAAAGCTTTCTCAGAATTGTTAAAGATGAGGAGCTTTGAGCAAATTAAAGTTAGCGATTTAGCAAAAAAAGCTCGACTTAGTAGAAGATCTTTCTATAATCATTACAATTCAAAAC
>GCA_002073695.1        38      copy:2(2-2)|86334|Segment264|643940-652561|1|length:8622|7434-8622
TTTAGTGAAGCAAAAGTCATCTTTACTTCACTTATCTATTTCTTAATGAAGCGAAACCTGTTTTGGCTTCACTTGTTTTTTATTATGTGAAGCAGAGTTGGATTTCGCTTCACTTAGCGACTGGTAAGTGACTTATCATTTATACTTTATTTTATACTGAA
>GCA_002025005.1        33      copy:12(8-13)|82755|Segment15|452057-454550|1|length:2494|1-2494
CCAGTACCGGATAAAGATAAATATGATCCAACAGGAGGGGAAACAACAGTACCCCAAGGGACACCAGTTTCAGATAAAGAAATCACAGACTTAGTTAAGATTCCAGATGGCTCAAACGGCGTACCGAAAGTAGTAGGAAACCGTCCGAATACAGACGTACCA
>GCA_002025005.1        58      copy:12(8-15)|82755|Segment17|452057-455800|2|length:2944|1-2944
CCAGTACCGGATAAAGATAAATATGATCCAACAGGAGGGGAAACAACAGTACCCCAAGGGACACCAGTTTCAGATAAAGAAATCACAGACTTAGTTAAGATTCCAGATGGCTCAAACGGCGTACCGAAAGTAGTAGGAAACCGTCCGAATACAGACGTACCA
```

Helicobacter pylori

ClusterID	Distribution	Marker	Percent_of_strains	Weighted_Average_Copy
Cluster8	1(3)|2(8)|3(6)|4(3)|5(1)|	*	100.0%[21/21]	2.571428571
Cluster107	1(6)|2(9)|3(6)|	*	100.0%[21/21]	2
Cluster66	1(5)|2(15)|3(1)|	*	100.0%[21/21]	1.80952381
Cluster24	1(6)|2(13)|3(2)|	*	100.0%[21/21]	1.80952381
Cluster42	0(3)|1(2)|2(13)|3(3)|	*	85.7%[18/21]	1.761904762
Cluster299	0(2)|1(4)|2(12)|3(3)|	*	90.5%[19/21]	1.761904762
Cluster133	0(4)|1(3)|2(11)|3(3)|	*	81.0%[17/21]	1.619047619
Cluster55	1(13)|2(5)|3(3)|	*	100.0%[21/21]	1.523809524
Cluster40	0(4)|1(5)|2(9)|3(3)|	*	81.0%[17/21]	1.523809524
Cluster262	1(10)|2(11)|	*	100.0%[21/21]	1.523809524
Cluster236	0(4)|1(7)|2(6)|3(4)|	*	81.0%[17/21]	1.476190476
Cluster109	0(4)|1(5)|2(11)|3(1)|	*	81.0%[17/21]	1.428571429
Cluster333	0(1)|1(13)|2(7)|	*	95.2%[20/21]	1.285714286
Cluster3	0(3)|1(11)|2(5)|3(2)|	*	85.7%[18/21]	1.285714286
Cluster211	0(4)|1(10)|2(4)|3(3)|	*	81.0%[17/21]	1.285714286
Cluster158	0(3)|1(12)|2(3)|3(3)|	*	85.7%[18/21]	1.285714286
Cluster354	0(1)|1(14)|2(6)|	*	95.2%[20/21]	1.238095238
Cluster25	0(4)|1(9)|2(7)|3(1)|	*	81.0%[17/21]	1.238095238
Cluster214	1(17)|2(4)|	*	100.0%[21/21]	1.19047619
Cluster2	0(4)|1(10)|2(6)|3(1)|	*	81.0%[17/21]	1.19047619
Cluster20	0(4)|1(10)|2(6)|3(1)|	*	81.0%[17/21]	1.19047619
Cluster150	0(3)|1(11)|2(7)|	*	85.7%[18/21]	1.19047619
Cluster41	0(1)|1(16)|2(4)|	*	95.2%[20/21]	1.142857143
Cluster355	0(4)|1(11)|2(5)|3(1)|	*	81.0%[17/21]	1.142857143
Cluster32	0(3)|1(13)|2(4)|3(1)|	*	85.7%[18/21]	1.142857143
Cluster250	1(18)|2(3)|	*	100.0%[21/21]	1.142857143
Cluster227	1(18)|2(3)|	*	100.0%[21/21]	1.142857143
Cluster201	1(19)|2(1)|3(1)|	*	100.0%[21/21]	1.142857143
Cluster141	0(3)|1(12)|2(6)|	*	85.7%[18/21]	1.142857143
Cluster62	1(19)|2(2)|	*	100.0%[21/21]	1.095238095
Cluster53	0(4)|1(12)|2(4)|3(1)|	*	81.0%[17/21]	1.095238095
Cluster373	0(3)|1(13)|2(5)|	*	85.7%[18/21]	1.095238095
Cluster350	0(4)|1(12)|2(4)|3(1)|	*	81.0%[17/21]	1.095238095
Cluster287	0(3)|1(13)|2(5)|	*	85.7%[18/21]	1.095238095
Cluster216	1(19)|2(2)|	*	100.0%[21/21]	1.095238095
Cluster205	0(4)|1(12)|2(4)|3(1)|	*	81.0%[17/21]	1.095238095
Cluster200	0(4)|1(12)|2(4)|3(1)|	*	81.0%[17/21]	1.095238095
Cluster19	0(4)|1(12)|2(4)|3(1)|	*	81.0%[17/21]	1.095238095
Cluster155	1(19)|2(2)|	*	100.0%[21/21]	1.095238095
Cluster135	0(1)|1(17)|2(3)|	*	95.2%[20/21]	1.095238095
Cluster134	0(3)|1(13)|2(5)|	*	85.7%[18/21]	1.095238095
Cluster117	0(1)|1(18)|2(1)|3(1)|	*	95.2%[20/21]	1.095238095
Cluster95	1(20)|2(1)|	*	100.0%[21/21]	1.047619048
Cluster85	1(20)|2(1)|	*	100.0%[21/21]	1.047619048
Cluster84	1(20)|2(1)|	*	100.0%[21/21]	1.047619048
Cluster83	1(20)|2(1)|	*	100.0%[21/21]	1.047619048
Cluster81	1(20)|2(1)|	*	100.0%[21/21]	1.047619048
Cluster77	1(20)|2(1)|	*	100.0%[21/21]	1.047619048
Cluster72	1(20)|2(1)|	*	100.0%[21/21]	1.047619048
Cluster68	0(4)|1(12)|2(5)|	*	81.0%[17/21]	1.047619048
Cluster6	1(20)|2(1)|	*	100.0%[21/21]	1.047619048
Cluster5	1(20)|2(1)|	*	100.0%[21/21]	1.047619048

```
>GCA_001189695.1          200      copy:2(2-4)|43475|Segment32|438528-439046|1|length:519|1-519
GTCCGCTCCAAAACCATAAGTCCACACATCAGAAGCCGAGTTGAAGAAGCTAGATTTAATAAACGCATGGTTGTAATCAAAAAGCCGTAATACCTAGCACCCCACCTCTCTTTTTTGGCCAAAGAATTGTTTGTAGCCCACTTGAATGCCGATCCCATT
>GCA_001248105.1          162      copy:2(2-3)|46457|Segment2|1031252-1031938|1|length:687|1-687
TTATCAAAATTCTATTCTCATTAAGTATTGCTTATAAAATTCAAAAGATGGTTTCATTATTACAAAAATCAACGCTTCAAAGATAAATAGGTTAAAATACCCCCAAAATCCTTTTTTTTTTTTTTTGAAATACAAATAGTCATGGTAAAATCCAATCAGT
>GCA_001189695.1          55       copy:3(2-3)|43475|Segment52|897234-897555|1|length:322|1-322
TTTAAAAAGGAAAGGCTTTAAAAGAAATCCCTTAAAAAGGGAGTTTCACAAAAGCTCAATACGCAAACACATAGTTGAGATACACGCTATAAAGCCTGCGGTATTGGAGTTGAGTGCCTAGCAAAGAATAGTAATTTGTGTTGATCGTGGGGATCTTCA
>GCA_001248035.1          197      copy:2(2-2)|46452|Segment24|326701-327063|1|length:363|1-363
CTGAGCCATCATCTCTCCTTTGAATTTAACTCTGTAAGCCCCACTGATTTAAAAATGGGGCTTATGGAATCAAATGGTTGAAAAATGTTGCGCGGGTTAATGAAAGGGGGAATAATAGAGAGAATAATAAAAAGTAAAGGATTTTTAAGCGGATTTTTG
>GCA_001248035.1          293      copy:2(2-2)|46452|Segment14|1554998-1555341|1|length:344|1-344
ATCATTCTTTCTACCTTTATTTTTATCTTTTTGTAACCGCTGAAGAAAATGGGGCGTATGTGAGCGTGGGGTTTGAATATTCTATCAGTCATGCGATCCAACACAATAACCCTTTTTTGAATCAAGAACGCATCCAAACCATTTCTAAACGCTCAAAATAA
>GCA_001189695.1          316      copy:2(2-2)|43475|Segment17|1474835-1475135|1|length:301|1-301
GGCTAGCACTTGGTTTGCAAGTAACAATTTCACCCCTTTCAATCAAGTCAAGAGCCACACGATTTTTCAGTTACAAGGAAAATTTGGCGTTCGTTATAATAGCGATGAATACGATATTGATCGCTATGGCGATGAAATATATCTTGGGGGTTCTAGTGT
>GCA_001051445.1          342      copy:2(2-3)|38517|Segment91|730208-730653|1|length:446|1-446
TTGTTTTTACAGGCTAAGATTTCTTATGGCATAGGCAAGCGGCTAATGCTGGGGGGGTGAGCGTGAGCGGCTTAGAAATGGCACAAAATGCGAGCATGCACCCTTGGAGTTTTGAAGTGGTGGATGCGAAATTACACCACATCATGAAAGAGATTTATA
>GCA_001446295.2          305      copy:3(3-3)|81079|Segment14|1523759-1524180|2|length:422|1-422
CTTAAAGACAAGATACGATTTTAAATTTAATTTTCATTAAAAACAACATAAAAACCAAAAACAGAACCTTTTCAAGTTTGGTTTTTTTTTTTTTTTGGATAAAATAGCTTTAAAAAAGGGTGTTTAATTTTTTATGACTTCAGGTTCAAGCCATTCTCTTAA
>GCA_001189755.1          288      copy:2(2-2)|43477|Segment71|9877-10267|1|length:391|1-391
TACTATCCCTATTTATTGGGGGAGTCCTAGCGTGGMGAAAGATTTTAACCCTAAAAGTTTTGTGAATGTCCATGATTTCAAAACTTTGATGAAGCGATTGATTATATCAGATACTTGCATGCGCATCAAAACGCTTATTTAGACATGCTCTATGAAAA
>GCA_001189755.1          187      copy:2(2-2)|43477|Segment13|1326399-1326783|1|length:385|1-385
AAATTTCTTCTTGCTCACATCTTCTAAAATATCTTGGGAGATCCCATCAATTTTAGTGCTGACTTGTAAAGAATGATTAGCGATAGTCAAATTATCCTGGACATCTTGTTGCAAGGCGTTAATGGAGTTGTGGATGTTTTCTACGCCCTTATTTTGCAT
>GCA_001051445.1          1        copy:2(2-2)|38517|Segment20|1186325-1186743|1|length:419|1-419
TTAGCCCTAACCAGCCAAAAAAAACCGTTACCGCCAAAACCGCCCTGAAACGCTTAAAAACAGCTACAATACCCCCGGATAGCCAGCGTCTTGATGATTTATTTGAGAAAGGAATCAAACACCAAGAAGTTTTAAAACTCAATGAGATCAAATCTTCTA
>GCA_001051445.1          179      copy:2(2-2)|38517|Segment37|1296520-1297124|1|length:605|1-605
TGGCAGTGGGTATGAATTTGTGAATGAAATTTCTGGGGGCGTGATCCATAAAGAATATATCCCTGCGGTGGATAAGGGTATCCAAGAAGCGATGCAAAATGGCGTTTTGGCAGGCTATCCGGTGGTGGATTTTAAAGTTACCCTTTATGATGGGAGCTA
>GCA_001051445.1          81       copy:2(2-2)|38517|Segment77|61-397|1|length:337|1-337
GATTTTTCTCCTTTGAAAAAATTAGGCACTACTAAAACTTTTATACTAACATATTTTCATTGAAATAACCCTAAATTCATTTATATCTAAAAGTTATCCGACCCTTGTCTAAGCTATAGGGCGTAAGCTCTAGCTTGACCCTATCGCCTAAAGCAAT
>GCA_001248035.1          351      copy:2(2-2)|46452|Segment31|587686-587985|1|length:300|1-300
GCTCCACGCTGAAGACGACGGCTTTTACATGAGCGCGGGCTATCAGATCGGTGAGGCCGCTCAAATGGTGAAAAACACCAAAGGCATTCAAGATCTTTCAGAAAACTATGAAAAGCTGAACAACCTTTTGAACAATTACAGCACCCTAAACACCCTGAT
>GCA_001189775.1          152      copy:3(2-3)|43479|Segment3|1886073-1887182|1|length:310|1-310
AAAGTGTTTGAAAGATCTCCCTAAAGACTTACAAAGCGATATTTTAGCTAAAGAGAGCCTGAAAGCTTATAAAGACTGCGCATCTCAAGCCAAAACTGAAGCTGAAAAAAAAGATGCGAGAAATTACTCACCCCTGAAGCGAAAAAACTTTTAGAAGA
>GCA_001446395.2          213      copy:2(2-2)|81035|Segment12|1310071-1319545|1|length:475|1-475
TAAGCCTAACGCCTCTTCTTAACTCTTGCTCTAAAAGTTTTTTGCTTCAGGCGTGAGTAACTCTCACAAGCTCTCCTTTCTCTTCATTTCTAGCTTGAGAGAGGCAGTCTTTATAAGCTTTCAGGCTCTCTTTAGCTTCTTCTAAAAGCTTTTTCT
>GCA_001189775.1          193      copy:2(2-2)|43479|Segment42|249951-250313|1|length:363|1-363
ATGCAATAGTTATTTTTTCTCCATCATTGTTTAAATCCTTGTAATAACCATACAAATCCCTCCCTGAATTTATCACTGGATACATTCCCTTATTTTTTACTTCGCTCACTTCGCTTACTTTGAGTGTTTTTTTATTTGTGCTTTCACACATCTCCCCCA
>GCA_001248105.1          115      copy:2(2-2)|46457|Segment31|1512646-1513574|1|length:929|1-929
AACCCCTTTACAAAGTAAATGGGGGGGGGAGTTATTCTACCTCGCCCTCCGCTTGATTTATCTCCTTCCTGTAACAATACCCTCTCCACAATCACATTACCCTCCTCATCATAATCCTCATCTCTGAAACCATAAGACCAAAGATTATCGGCTAAAA
>GCA_001051445.1          256      copy:2(2-2)|38517|Segment54|201754-202335|1|length:582|1-582
AACTTTAGCAGAAGAGGCAAAAGAATTGATCGCTGAAGCGGGATTTGACAGATTCTATGGCGCTAGACCCTTAAAGCGCGCGCTCTATGAAATGGTAGAAGACAAGCTCGCTGAACTCATTTTAGAGGATAAAATTAAAGAAAATGACAGCGTGGCGTT
>GCA_001189765.1          346      copy:2(2-3)|43478|Segment18|1569812-1569512|1|length:331|1-331
AATCGCTACCCACCCCAATAAGTCCAAAACATCAGAAGCTGAGTTAAAGAAATTGGATTTGATATAGGCGTGGTTGTAATCAAAGAAACCATAATACCTTAACCCCCCAAGTCCTCTTTTTGCCAAAAAATTGCTTATAGCCCGCTTGCACCCCAAAGCCAT
>GCA_001189695.1          92       copy:4(3-4)|43475|Segment37|565078-565555|1|length:478|1-478
CTTTTTGCAAGTCTTKRGGAATTTTTCAAGCACTCGTTCGTTCTTCATCGTTTAGCGTTTTCGACTCAACTTTTTTTCGCTTCAGGGGTGAGTAATTTCTCGCATTCTTGTTTCTCTTTTTTCATTCCTAGCT
>GCA_001051445.1          143      copy:2(2-2)|38517|Segment55|243816-244406|1|length:591|1-591
TAGTGAATAGCGGCACAGAAGCGACCATGAGCGCGATACGACTCGCTAGAGCTTATAGCCAAAAAGAGCGATTTGATCAAGTTTGAAGGGTGCTATCATGGGCATAGCGACTCCTTATTGGTGAAAGCGGGTAGCGGGTGTGCGACTTTTGGATCCCCTT
>GCA_001189765.1          325      copy:2(2-2)|43478|Segment32|543731-544082|1|length:352|1-352
CATCATGCAATTAAGAGACATTATCCATGAATACGGGCATACTTTAGGCTATACGCATAACGGGAACATGACTTATCAGAGAGTGCGCATGTGTGAAGAAAGCAACGGGCCAGAAGAGCGCTGTAAGGGCGGGAAGATAGAGCAAGTGGATGGGAAAGA
>GCA_001248105.1          133      copy:3(2-3)|46457|Segment56|664582-665165|1|length:584|1-584
CCCCTTTAAAAAAGAAAAGTTTCAAAGAAACCCCTAAAAAGGGAGTCCCACAAAGAGCCAAACATCAGTAAGCGAAACACGCTATAGAGCCTTCTGTATTTGAGTTCAGCCCCCATAAAGGAATAATAGTTGGTGTTGATGGTA
>GCA_001051445.1          3        copy:3(3-3)|38517|Segment28|1196405-1196717|1|length:313|1-313
ACTTCTAAAGCAGCAAATTCTTTAAAAGGCTCTGTAATTTGGTGGAGTTTTGCTTCATCTTCTTCATCCAAACCTACCACTAAAGCGTATAGAGTGTAACGAATTTTTTTATTGGCTTTAGCGATGCTTTCTAGCAAGGAATAAATGCACGCGCCCGCA
>GCA_001189695.1          30       copy:6(4-7)|43475|Segment43|570952-571384|1|length:353|1-353
TTAAGCTTTCAATTGAGTKGCYCACAAAGATAAAAAAGAGCCCAAAAACCTGAGTTAAGCTCTCAATTAGTGGCTCACNAAAAGATAAAAAAGAGCCCAAAAAACCTGAGTTAAGCTNCTCAATTAGTGGCTCACAAAGATAAAAAAAGAGCGCCAAAAAA
>GCA_001189755.1          97       copy:2(2-2)|43477|Segment64|963634-964163|1|length:530|1-530
TAGGATCAGGCGTTTGAAAGAATCTAAAATCTTTGATTCGCAGTCTAGTAATGATCAAGAATTGCAAAAAAGCGCTAATAGCAACCTGGATTTAGACCCTATCGGCAACGCCATGCCCCAAACTTTAGCTAAAACAGAGGCTAAAGAAACTCAAACAGA
>GCA_001189755.1          190      copy:2(2-2)|43477|Segment11|12-756|1|length:745|1-745
CTTGTGACGTCTCTTGCTCTTTGATTTGAGCGTTCTGCAATTCTTGTTGTAAAATCTGATTGCACATGCTCTGTAAATCGTTTTGAGCGTGATTTCTTTCTGTTCTAAGGTTTGTATCACTTGCTTTAATTCGGCTCTCTCTGCTCTCTCTCAAAGCTT
>GCA_001051445.1          312      copy:2(2-2)|38517|Segment94|755246-756010|1|length:765|1-765
TTTTTGGCATCACGCATGATAACGCTTGTTTGCACCCCGGCTTGAGGCATGAATAATCGGTGGGCATTTTAATGCTCGCTAAAAGCGAATGTTTTTTTAAAATTTCAACATTGGATTTTAAAGCTTGCCCGCTCCCTGCGCTATCTTGGATAATGATC
>GCA_001189695.1          261      copy:2(2-2)|43475|Segment22|275570-276167|1|length:598|1-598
TTATTTCATCCAAAACAACGCTATAAAATCATTCTGCACAGCAACCCCAACGAACCATCAGATCTAGTCTTTGGCAATCCTTTGGAACAAGCCAGAAAAATCTTATCGTATCAAAACACCAAAAGAGTGTTTTACACCGGCGAAAATGAAGCGCCTAA
>GCA_001446295.2          69       copy:2(2-2)|81079|Segment2|1110016-1116393|1|length:378|1-378
TAATATTGACTTGACATTAAACACATAGGCTAAAAAGCCAGCGTTAAAAAAAAGAATGATTTTAGAGAAATGAGCGGATGGGGTGGGTAAAAAAGTTTTTATTGTTTTGATGCATGGGATTTCCTAAATTTAAAAATAACTAAAGATTAACATGTTTTTA
```

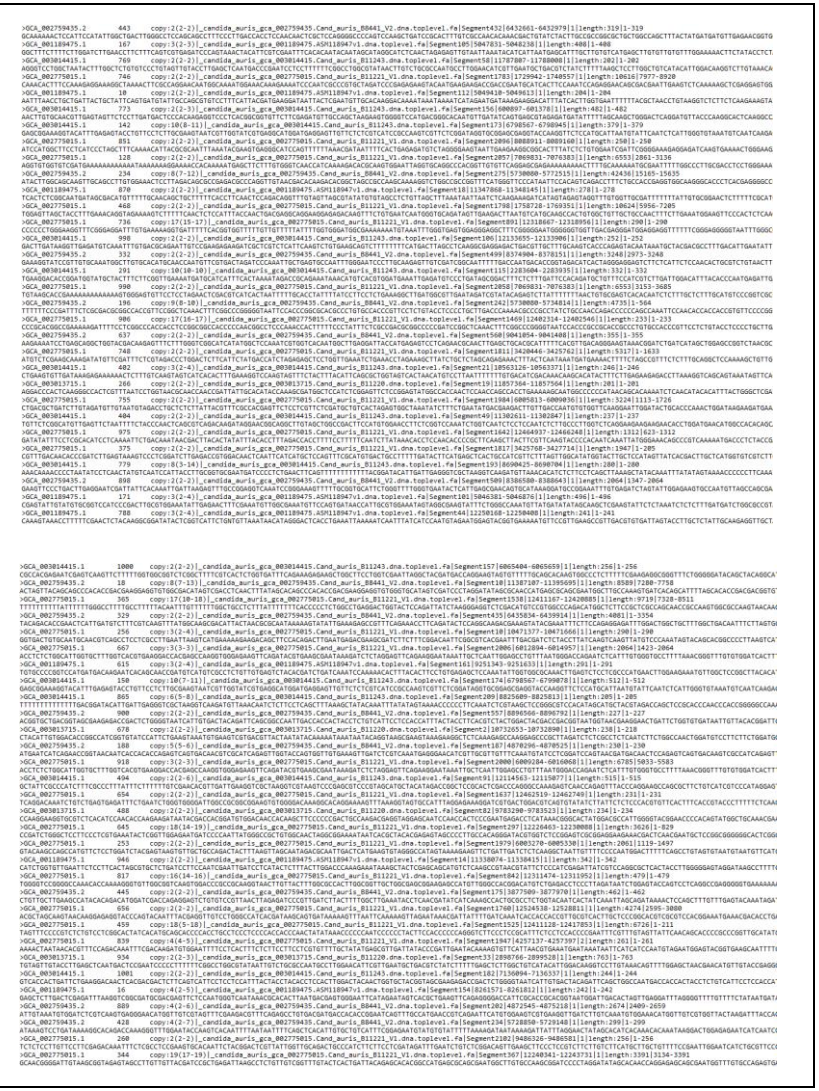
```
>GCA_001189755.1          191      copy:3(3-4)|43477|Segment70|965358-965761|1|length:404|1-404
AGCTTCCTTTTTTTCTTGTTCTAGTTCTATTTGTTCACTAACATCACCAACGCTACAAGCGGCTAATAATAAACTTGTCGCTATGGTTAATCCTGAATATTTCCAACCAATTGATTTGATTCTCTTTTTCAGCTTGTTTGGATTTATCTTGGACTTTATCGTC
>GCA_001248035.1          61       copy:3(3-4)|46452|Segment53|905449-906075|1|length:627|1-627
CAATTTTTCGCATTCTTTTTTTTCAGCTTAGTTTTGGCTTGAGATACGCAGTCTTTATAAGCTTTCAGGCTCTCCTTTAGCTTCTCAAAAGTTTTTTCGCTTCAGGCGTGAGTAATTTCTCGCATTCTTTTTTCTCTTTTTCATTCCTAGCTCTTGAAAC
>GCA_001051445.1          257      copy:2(2-2)|38517|Segment56|285757-286541|1|length:785|1-785
AAGCGCTCTTTCAAACGAAACGCCATAGCACGCACGGCACGGCGGGTAAATACTGGATAGCGAGCGCGTCTTGATGATGGTGTTGCTGCTTATCATAGGGCGTTCAAAACCTTTTCTATATTTTTTCGCTTGTTTCGCTTGTTTGAGTTTTTTGAATTTCTTTGGGC
>GCA_001189755.1          264      copy:2(2-2)|43477|Segment29|1449521-1449821|1|length:300|1-300
GTGATTTTTCGCAACTTTATTCAACGGCACTTGATTAAACCACTTTAAGAACTTTTCGCTATTGAAAGGCGTAAGGAACTTCGTTGCCTTTTTCACTCCTTAAGGTTACACGATTAAACCCAATACTAATGACTTTAGCCTCCACCACTTTGTCGTTTATCGT
>GCA_001248105.1          120      copy:2(2-2)|46457|Segment46|653331-653637|1|length:307|1-307
GAAACTAGAAAGCGAAAGGTTAGAAAGCGAACAGGAATTTTTAAAGGCTAAAGAGCAAGAAATGCAACGCCAAAGCGCGCGTTAAAAAAGAAATTAGAACACGAGCGAGGCAATGCGGGCAACATTGAAAGCCAGACTAAAATAGAAGTAGGAGAGGATATACC
>GCA_001446295.2          70       copy:2(2-2)|81079|Segment12|1337224-1337586|1|length:363|1-363
TTTTAAGCTTAGTGCTTTCAATCATTATTGCGTTATTTGTGTATAAAATGCCTGCTCAAATGGTGAGCGCGAGTTTTTTCTATGGCTTTCTTTATGGCTTATGGCCGATCGCATGGATTGTGATCGCTGCGATTTTTCTCTACAACCTTTCAGGTAAATCCG
>GCA_001189775.1          268      copy:2(2-2)|43479|Segment26|1373693-1374195|1|length:503|1-503
TGTGTATAAAXGGAGTGGTGCTGTTATAGTTGTTGGTAAAGCTTTTTAAATAGTTTAAAGAGCTATTGACCAGCTCAATGTTTGGCGTTTTGAATGCTTTTGAATGCTTTTGTTTAACGAAGCCACTTGATTTAAATACTGGTTTAATTGCTCGTATTTGTCGTTCAA
>GCA_001248105.1          368      copy:2(2-2)|46457|Segment17|1267127-1267570|1|length:444|1-444
TTTACATTGATCCGCCTTATAACACGCAAAACAATGAATTTATTTATGCGGATAATTTCAAACGCTCCAGTTGGCTCTCTATGATGGAAAACCGCCTAGATCTTGCGCATTCCTTATTGAATGATAAAGGCGCGATATTTGTGAGCATTGATGACAACGAAC
>GCA_001278525.1          375      copy:2(2-2)|62574|Segment40|599655-599956|1|length:302|1-302
CAATCCAACTAAAAAAGTTTTTGCTTCCTCTTTGGGGTGTTTGGGGGGTTAAGGGGCTTATACAATAGCTACTATGTTTTCAATCAAGTCAAAGGAAGCGATTAGATATAGTTACCGGGTTTAATTACCGCTACAAGCATTCTAAATATTCTATAGG
>GCA_001248035.1          366      copy:2(2-4)|46452|Segment38|767904-769219|1|length:1316|1-1316
GCCCCTTTTTAAAAAGAGGTTTTTTAATAAGCGAACACGTAATTGAGATACACGCTATAAAGCCTTCGGTATTCTAGTTTGGTGTCAGGAAAGGAATAGTAGTTCGTGTTAATGGTAGGGATTTTCACGCCCAATTCCATGCCATGTTGAGCCGCATGATG
>GCA_001051445.1          75       copy:4(2-4)|38517|Segment65|1136540-1137423|1|length:884|1-884
AAGCAATATAGAGACTAATAATCAAATAAAAGTAGAACAAGAACAACAAAAGACAGAACAGGAAAACAAAAGACAGAACAAGAAAACAAAAGACACTTAAAACACAAAAAGATTTCATTAAAGACCTAGAACAAAATTGCAAAGAAAATCATGGTCAATT
>GCA_001248105.1          201      copy:2(2-2)|46457|Segment4|1157913-1158243|1|length:331|1-331
ACCCTACCGCCGCGTTGATCCGCTAAAAGCACGGCTTGATAGGCGGGGGAGTTGGCTTTACCGCCGATCAAATTCTTCGCCTCGGACTCAGAGGATAATTTGATGAGGGTGTTTAGGGTGCTGTATC
>GCA_001248105.1          47       copy:3(3-3)|46457|Segment19|1355257-1355786|1|length:530|1-530
AGGCTAACACCAAACAAAAAGCACCAAAAAAGCGCTAAAAATCAAAAAATGACAAAATTTTTCCCAAATGACAAAAAAAAAAAAAGCGCTTTTATGTATAATGCTCTAAATACATTCCAAATGTATTCCAATGAATGGCCAATGAATGTATAAAATCCCTA
>GCA_001278525.1          57       copy:2(2-2)|62574|Segment10|1455155-1455620|1|length:466|1-466
AACGCTTCAGATCCCCTTTATTGGGCGTTAATGTGAAAATAGGCTATCAAATTACTTTAATGATTTCATAGGGTTGGCTTATTATGGCATTGTGAAATACAACTATTCTAAAGCCCTTAATCAAAAATTCCAGCAATTGAGCTATGGTGGGGGGATAGATT
>GCA_002093735.1          249      copy:2(2-2)|97042|Segment12|499735-500416|1|length:682|1-682
GAATTAGACAAGAAATTAGACGCTGATAGCATCTACTCATTAAAAAACATCGTCGCTAAAAACCTTAATTTTGATAAGGCTACTCATCCTAATGTAGGCGATAGTAATTTAGCGCTTTTAATGTTCAACATGCCAAGGACCAAAACGCAAAGCGAAGTAGAATTA
>GCA_001248105.1          203      copy:2(2-2)|46457|Segment42|52409-52805|1|length:397|1-396
TTTTTTGGCGTCTTTTTATCTTTGTGAGCCACTAATTGAGAGCTTAACTCAGGTTTTTTGGCGTCTTTTTTATCTTTGTGAGCCACTAATTGAGAGCTTAACTCAGGTTTTTTGGCGTCTTTTTTATCTTTGTGAGCCACTAATTGAGAGCTTAACTCAGG
>GCA_001189775.1          234      copy:2(2-2)|43479|Segment27|1382118-1382505|1|length:388|1-388
GCCTGTAGGGTTATTTTTATTTGGAGTAAGTGGTGATGAAATCCACTAACAAATCTATCCCCCCACCATAGCTTAATTGCTGGACTTTTTGGTTAGTGGCTTTAGCGTAATTGTATTTGATGATGCCATAATAAGCCAACCCTATGAAATCATTGAAGTAGTT
>GCA_001248105.1          353      copy:2(2-2)|46457|Segment38|319676-320067|1|length:392|1-392
CTTACACCCCATTAAGACAGAGCATGAATTGGTGCGTTTTGGGGCAAAACATGATGGGGGCTATTTGATACCTAAGGATTTAAGGGGATTAGAACGCTCTTTTCTCCAGGTGTGGGAGGTGAAAGCGCGTTTGAAGAAGATTTTTATCGTCAATGCAAACT
>GCA_001051445.1          184      copy:2(2-2)|38517|Segment103|956708-957289|1|length:582|1-582
RGGGTTAAAAHAATANWTTAAATTAATTTAATTTAATTTATTAAAAAAACAAAAAAAGCAAAAAAAGAAAGCCTTTCAAGNWTTNCGGGTTTTTTTTTTTTTTT
```

Legionella pneumophila

ClusterID	Distribution	Marker	Percent_of_strains	Weighted_Average_Copy
Cluster66	7(1)|12(1)|23(1)|28(1)|30(4)|31(1)|	*	100.0%[9/9]	24.55555556
Cluster67	5(1)|12(1)|21(1)|27(1)|28(2)|29(2)|30(1)|	*	100.0%[9/9]	23.22222222
Cluster491	4(1)|11(1)|20(1)|26(1)|27(2)|29(2)|30(1)|	*	100.0%[9/9]	22.55555556
Cluster65	6(1)|9(1)|19(1)|26(2)|28(4)|	*	100.0%[9/9]	22
Cluster351	5(1)|10(1)|18(1)|23(2)|24(1)|27(2)|29(1)|	*	100.0%[9/9]	20.66666667
Cluster180	0(1)|2(1)|13(1)|15(1)|19(1)|20(1)|21(2)|22(1)|	*	88.9%[8/9]	14.77777778
Cluster280	4(2)|6(1)|8(1)|9(2)|11(1)|12(2)|	*	100.0%[9/9]	8.333333333
Cluster111	3(1)|4(1)|6(1)|7(3)|9(3)|	*	100.0%[9/9]	6.777777778
Cluster253	3(2)|4(2)|5(5)|	*	100.0%[9/9]	4.333333333
Cluster208	1(1)|2(1)|3(1)|4(2)|5(4)|	*	100.0%[9/9]	3.777777778
Cluster185	1(1)|2(3)|3(1)|4(1)|5(2)|6(1)|	*	100.0%[9/9]	3.333333333
Cluster250	1(2)|3(4)|4(2)|6(1)|	*	100.0%[9/9]	3.111111111
Cluster364	0(1)|2(2)|3(3)|4(1)|5(2)|	*	88.9%[8/9]	3
Cluster100	1(3)|3(5)|4(1)|	*	100.0%[9/9]	2.444444444
Cluster186	1(2)|2(2)|3(5)|	*	100.0%[9/9]	2.333333333
Cluster678	1(4)|3(4)|4(1)|	*	100.0%[9/9]	2.222222222
Cluster31	1(3)|2(1)|3(5)|	*	100.0%[9/9]	2.222222222
Cluster137	1(1)|2(5)|3(3)|	*	100.0%[9/9]	2.222222222
Cluster120	2(7)|3(2)|	*	100.0%[9/9]	2.222222222
Cluster60	1(2)|2(4)|3(3)|	*	100.0%[9/9]	2.111111111
Cluster54	1(2)|2(4)|3(3)|	*	100.0%[9/9]	2.111111111
Cluster631	1(5)|3(3)|4(1)|	*	100.0%[9/9]	2
Cluster533	2(9)|	*	100.0%[9/9]	2
Cluster71	0(1)|1(5)|4(3)|	*	88.9%[8/9]	1.888888889
Cluster205	0(1)|1(2)|2(3)|3(3)|	*	88.9%[8/9]	1.888888889
Cluster197	1(5)|3(4)|	*	100.0%[9/9]	1.888888889
Cluster73	1(3)|2(5)|3(1)|	*	100.0%[9/9]	1.777777778
Cluster581	1(5)|2(1)|3(3)|	*	100.0%[9/9]	1.777777778
Cluster565	1(2)|2(7)|	*	100.0%[9/9]	1.777777778
Cluster455	1(5)|2(1)|3(3)|	*	100.0%[9/9]	1.777777778
Cluster409	1(2)|2(7)|	*	100.0%[9/9]	1.777777778
Cluster90	1(6)|3(3)|	*	100.0%[9/9]	1.666666667
Cluster88	1(6)|3(3)|	*	100.0%[9/9]	1.666666667
Cluster649	1(6)|3(3)|	*	100.0%[9/9]	1.666666667
Cluster429	1(3)|2(6)|	*	100.0%[9/9]	1.666666667
Cluster344	1(3)|2(6)|	*	100.0%[9/9]	1.666666667
Cluster248	1(3)|2(6)|	*	100.0%[9/9]	1.666666667
Cluster148	1(3)|2(6)|	*	100.0%[9/9]	1.666666667
Cluster743	1(5)|2(3)|3(1)|	*	100.0%[9/9]	1.555555556
Cluster738	1(4)|2(5)|	*	100.0%[9/9]	1.555555556
Cluster70	0(1)|1(5)|3(3)|	*	88.9%[8/9]	1.555555556
Cluster68	0(1)|1(5)|3(3)|	*	88.9%[8/9]	1.555555556
Cluster624	0(1)|1(2)|2(6)|	*	88.9%[8/9]	1.555555556
Cluster591	1(4)|2(5)|	*	100.0%[9/9]	1.555555556
Cluster56	1(4)|2(5)|	*	100.0%[9/9]	1.555555556
Cluster532	1(4)|2(5)|	*	100.0%[9/9]	1.555555556
Cluster331	1(4)|2(5)|	*	100.0%[9/9]	1.555555556
Cluster176	0(1)|1(2)|2(6)|	*	88.9%[8/9]	1.555555556
Cluster693	1(5)|2(4)|	*	100.0%[9/9]	1.444444444
Cluster437	1(5)|2(4)|	*	100.0%[9/9]	1.444444444
Cluster397	1(5)|2(4)|	*	100.0%[9/9]	1.444444444
Cluster165	1(5)|2(4)|	*	100.0%[9/9]	1.444444444

```
>GCA_001549915.1        418      copy:2(2-2)|68365|Segment31|2599805-2601004|1|length:1200|1-890
AGGGGGGGGTGCTTAAGTACAGACAATGAGGCTTCCCTATGATTTTTGGAATTTTAAACCTAACGTTTTGGGGATATGTTTTAGCAACTTTTGTTCTAACTCAAATTACCATTGCTTCGGTCACTCTTTATTTGCATCGTTATCAAACTCATCGTGCTTTAACCATGCACCCAATA
>GCA_001969405.1        587      copy:3(3-3)|81478|Segment23|2193387-2193745|1|length:359|1-359
GTAGTAGTTCAAACTTGATCTGACAGTTGCCGGTTTTTCAGAAGTGTCTGTCAGGTCAAATTCAGTTTATAAATTTTCCATCCAGATTTGTTTGCTATCCATCAAAGTTTGATCGGAGTGCTGCCACAACACATTTTGCCTTGATGGGTTCGCTCATTATTATAGTAATGAAGG
>GCA_000500125.1        174      copy:2(2-2)|57904|Segment136|3648559-3648892|1|length:334|1-334
CCTGGCTAATGACTGCAATTTGACCTGGTTCTGTCTTAAAATGTGTAACATGCTATAAGATGCATAAGCTACTTCAATTCGGTCATTCCAGGCTGGAAAATTCTGAACCAGATTGGCATGAAAATCATCTAAAGCCGCGAATTCTTCTGGTGTCATTTCTTTATTTAATGCCACA
>GCA_001989475.1        695      copy:2(2-3)|81805|Segment11|1626742-1627040|1|length:299|1-299
AGTTTAATTATTGTTTTAACAAAGGAGATACCTCACAACGCAGCATTAATACAATAGTAGCGCGTCTTATGAACAAAATAGGTCTTAAGGAGGTGTTTTTCAAGACAGATAAAGATTATGTTGACTGGATAGAAGACGTAATTTCCGCAAGAAAAAAAGCGGCTTTATCTCCCAGC
>GCA_001549915.1        247      copy:2(2-3)|68365|Segment7|14957-16138|1|length:1182|1-1182
ATCATGGTCTTGCCCATACAGTGCGAACGATGGCTTGTGCTGAGGTAATGATAGAAGAAGCACGTAAAGCTCAACTGAGAGGAGAAACATTAGACAAAGCAAAAAATGGACAGACTTTAGCCGATGTAACGCCGGAAGAACTGAAAAAAACTCTCATCGCACAAGCCTTTTCGTC
>GCA_000468235.1        596      copy:2(2-2)|55886|Segment100|2670000-2670875|1|length:876|352-602
CGCTCGTATCTAGCAGAGCTGGCCTTAGCGTTGCGTACCATGATTACTGAAGCCAATGCACTGGAAGAAGAGCAATTTACTCTGGCCGATGAAACACTGTTAATCGAAGTCTTAAGTGATGCCATCTTAACGTCATTTAATCATGATATCCCGCAAATGCTAACTGAGCATATTGT
>GCA_000465915.1        138      copy:2(2-2)|18663|Segment307|768998-769688|1|length:691|1-691
GGTTACAAAAATTTTCTAAAAAACACTGTTAGTCCTGTTTTGGAAAATAGTCTTAGTGGACTCATGGAGAGTGACCCAGGACACTTTAAGTAAAGCCTTAGCAGCATTTCCACCCGACGGACAATGGTCAGCCTTCAATTTTGAAGAAGCAAGGCAAGCCAAAACACAGATGGATG
>GCA_000465915.1        132      copy:7(7-8)|18663|Segment252|3654001-3654214|1|length:214|1-214
TTTTCTCGAGCTATGGCACTTATCATGTCTTCAATGAGCTATAGTAAACAAGAAACTTACATCCATGCTCTTGAGTAAACCCGGGTACATCCCCATATTTATGTTCATAGACACGTTTGATCAAACTCGCTTGTGCACACCCGTGTAGATCGTTCCATTTGCGTTTATTCGCCATGA
>GCA_000465915.1        133      copy:3(3-3)|18663|Segment257|3656170-3656433|1|length:264|1-264
TAAGCTTTTGGTGGGTGCTCCGGCTTCTGGTTTGGTGGATACGGCAACCAATCAGGCCATCGTGTTGTCTGCCAATGCGGGCGGCACGGAAGTGTTGGGCAAGAACACCAATGGGGATGTGGTGTTTAAGGTTTTATTAACTGCCAGCGATGGGGATGTGGAAGTCTTTCAATACC
>GCA_000468235.1        325      copy:2(2-3)|55886|Segment131|3611248-3611569|1|length:322|1-322
CATAAAATCCCTTTATTCCAAGGCTAAATCACGCCTCTTTATTAGAGCTTCTATCCCATTATAGTGGGATAGAAAAGTCTATTGATGTTGCATCCCCCATTGGCCGCTTTCAACTGGCTCTCTGAATCAATGAAGGAAATTAGCAGACGTAATAATTTGATCGCCCTCCTGAACCCC
>GCA_000468235.1        151      copy:2(2-2)|55886|Segment90|2637314-2637566|1|length:253|1-253
ACATAAGAAAAGCCAAAGTCAGAAAACCCCCGCACAGCAACAACCTTTGGTGCTGATAAAAGCGCACGTGACAATAAGTGACTTGATCTTCCATCAGATCGGGCGAGCGCCCCATCCAGGTCGTATACACAATGACTTGAGTGTCTGAAATGTCAGGCAATGCGTCCATACC
>GCA_000500125.1        241      copy:2(2-2)|57904|Segment79|2605931-2606189|1|length:259|1-259
TGTTTAATATAACGCCAAACAACCCAAAACGCCATAAGTAAAATGACACTTGCAGTAGCCAGTGCCCAGTTGCGGTGAATCACTTTTACGGCGTGCGCCATAGCACCATGTTTTACCGTATAAAAGGCATAAAACCCAGCGGACACGGTGGTCATCGTGCTCAATGCGGCCAGCCA
>GCA_001549925.1        678      copy:6(3-6)|68366|Segment101|805617-805798|1|length:182|1-182
AAGTCCTGAGATAAAGATTTATATAAGAACCAGTCGATATCCAGGGTACGTTCGGGCTGAAGGAATGCGTAAGCATCCTCTCGAAGCCTTGGCATGAAACATCAGATCTGATACAAGGCTTCGAGACGGCATAAATGCTCCTTAGCCCGATCGGTTCGAGATAACAACTGATTTTA
>GCA_001549915.1        670      copy:3(2-3)|68365|Segment35|2619530-2619801|1|length:272|1-272
GGACTTTCATAAAAGCATGGAAGTCCGGAACAAACGTTGAACTTTTACGAATGCAATCTGCTGATAATTTGGAAAAAATGATAGTCATTCTATTCTTTTGTTGCAATTCGATTACTTCAACTCAAAGAGTATTTTGAACAAGACTACCAACTTTTAAAAGCGGAACATGTTTGCATCC
>GCA_001549915.1        253      copy:7(4-7)|68365|Segment86|3466101-3466240|1|length:140|1-140
AATTAACGAGAGTTATGGATAAAGATTATATAAAGAAGCAGTTGATATCCAGGAGCCGTTCAGGCTGAGGGATGCATAAGCATCGTCTCGAAGCCTTGGCACGAAACATTAAATTCCATACAAGGCTTCGAGACAGCATTA
>GCA_001639045.1        188      copy:2(2-2)|73559|Segment27|2921332-2921897|1|length:566|1-566
CAGTCAGGGGATACAAAATACCCGCGGCCACCGGTACACCCCAATACATTATAAATAATGCAAAAAAGAGATTTTGACGGATGTTATTCATGGTGCTTTCTGACAAATGACGTGCCTTGGCAATGCCACGCAAGCTGCAGCAAGTAATACCGGCACTCTCAATAGCTACA
>GCA_000468235.1        469      copy:2(2-2)|55886|Segment103|2854089-2864291|1|length:203|1-203
TTTGCAAAAGTGGCTGTTGGATTCTTGTAATTATACTCCCATTCAGAACCCACTTCAGTGTAAGGCTGAGATAAGCGTGTTTTATACACATATTCCGCCACTTCCAACACATCAGAATCTAGAATAATAGGCACAATGGTCTCCTTCCATAAACAAAATGAGGATTTATTATATAA
>GCA_000468235.1        593      copy:3(2-4)|55886|Segment66|2316581-2317110|1|length:530|1-530
TTAGAAGCAAGCTCTTACATATTTCCCAACACCATTTGCAGGAGAATGCGAGCAAGCACCCTGCAAAATGCCTGGGAGCTATCAAAAAGCCCGCTCAAAACTACAAACCGGATGCCATTAAAAATGGCTTTCAATCCAGTCCTGATAAAGCGAGAAGACAACCCAACATGGATGAAA
>GCA_000465915.1        216      copy:2(2-2)|18663|Segment63|1844174-1844803|1|length:630|1-630
CCCGTCCCATTAACTGATTCAGTGTGGCATATTGGCTAATACGAAACTATCCAAAAGATGTTTTTCTTGAACAATGACCAAATACTCAATGCGTCCGCGCATTGTCATCCAATTTTGAAATTTCTGCGGCATGATATCGATTATGGATAACGTCAACCAATTCCCGGTTAATCCGT
>GCA_001549925.1        424      copy:2(2-3)|68366|Segment26|1866752-1867962|1|length:1211|1-1211
TTTGTTTCTCTTCCCATTTAGCTCCAATTTTGTTTGGAAGTGCTTCTTTTTCAGTAGAGTGAAACTCGGTGTCTTTTCTCATGTTGCATTAAGGTATCGATGTTTTCTCATGTCGATGTCATCAAGCAGTGAGCCTACAACAGCCATCTCGCATTTCGCTAACAAATCGCGGAAGAA
>GCA_000465915.1        295      copy:3(2-3)|18663|Segment128|198261-198404|1|length:224|1-224
TGGAGCGGTCCCCCATGATTTTGAGTTAATTGAATAGAACGTGTCATTCAGCATGCTAAAGGGACCCGTCGATCAGCAGCTAATGGGACCCGCCTAAAATGTCAAAACGGTACCTATCAAACTTGTCCACTTTTAGTTAATTATACATGTTATTTTCGCTTTTTTTGTTTCTCGCTCTC
>GCA_001549925.1        181      copy:2(2-2)|68366|Segment17|1482795-1483127|1|length:333|1-333
TAAGCTTGGTGGGTGCTCCGGCCTCGGGTTTGGTGGATACGGCCACCAACCAGGCGATCGTTTGTCTGCCAATGCGGGCGGTACGGAAGTGTTGGGCAAGAATGGCAATGGGGATGTGGTGTTTAAGGTGGCATTAAACGCCAGCAATGGGGATGTGGAAGTTTTTCAATACC
>GCA_001753145.1        744      copy:13(12-13)|76290|Segment82|688501-689680|1|length:1180|1-1180
GGGAGCCAGTTCCAATTACGCGGCTCAGGAAGGCGATATTGAATACGGGCCTGGTCTTTTTGAACGCCACCGCCAGTGACGACGGATGGGGGATGGGGGACACGGCGACCGCGCCAGCGCGTCTTTGGATTTGGGCGGCAATGTGAAGTTTGATGACGACGGTCCGAGCGTGACGATGGCGG
>GCA_001549925.1        280      copy:8(5-6)|68366|Segment104|805656-805757|1|length:102|1-102
CCAGGGTAGCTTCGGGCTGAGGAGATGCGTAAGCATCCTCTCGAAGCCTTGGCATGAAACATCAGATCTGATACAAGGCTTCGAGACGGCATAAATGCCTCC
>GCA_001549915.1        733      copy:2(2-2)|68365|Segment17|1590414-1600410|1|length:1997|1-1997
TATGCTTTGGTGAGCCGCACAGCAAGCCCTGAAACCTTGCGCACGATTTCCGACTGGACTATTCGTCCACGCCTGCTGGCTTTGGGCGGAGTCGCTTCAGTAGTTTCCCTCGGCGGAGAAGTAAAGCAATATCAAGTACAGCTTGATCCCAAGCGCCTGCTGGCTTATGGCATTAC
>GCA_000500125.1        333      copy:2(2-2)|57904|Segment59|2312723-2313009|1|length:377|1-377
ACAGACAACATCATGGTCTTTCCCATACTTTACGTACAATGGCTATGCGGAATTGATTGTTGAAGAAGCGCGTAAAGCTAAATTACGGGGTGAAACATTAGGGAAATTCAAAGACGGACGTACTATTGCTGATGTTACCCCTGAAGAACTGAAAAAAATCATGATTGCCCAGGCC
>GCA_001549915.1        64       copy:5(2-5)|68365|Segment76|3453386-3453512|1|length:127|1-127
ATATTTTATGCTTCTTCCGGGGAAATTTCTTATTCCAGATGACTTAAAATAGGTCAGGACTTACGAAAAAACCCAAGATCGAGGCAAAAATGTTTTTAATGAGGAGTTTAGATAAACTAAATGA
>GCA_001989475.1        285      copy:2(2-2)|81805|Segment32|2386069-2386883|1|length:815|1-373
CAATCAGCATATTTTGAATGGCTTTAACCACTATCCCGCCTCTGTTTCACCCATACGCAAATACAAATCAAACGTCCTCTGGCCTTCAAGAATACTGGATACCGTCACACCATTAAGAAGGATTTGAATATCCTCGGAAATTTGACCAATATTAACACCATAGCGAGCCGCTTTCTCT
>GCA_000465915.1        457      copy:6(5-7)|18663|Segment278|654058-655380|1|length:1323|1-1323
TAATCTCTGTACGGTACGATTTTGGGACCGTCGTTGCGAGGAGCATAGCGACGAACAATCCAGTGACCTTAAAATTGCTAATCTTCACAACAGTTTTCGTAAAGATCTTCCCAATAAGGGTTAGCTTTTCAATTAAAGCCAATTTCTTTTTTCTTGACCCGCCTTAAGTTC
>GCA_001989475.1        284      copy:2(2-2)|81805|Segment2|1151386-1152869|1|length:1484|1-1484
AGGAGCCCCAATAAATGTTTACTCGCCTCATTGCCTCATTGCTGGTCTTTACATAATCGCCCTTGATTTTGGCGTTAACGAATTATTCTTTGTTTACTGGGTGGTTCAAAGGAGAAAGTATTGATTACCAGGCCCTTTGATTAACATGATCCAGTGACCAGAGT
>GCA_000468235.1        141      copy:2(2-2)|55886|Segment10|1569064-1569300|1|length:237|1-237
AAAATTTTTAGTGGTGAGAATATTTTATTTTAGTGATGAAGTAGTTTTTTTAAGAGGATTGCGTCTTCTACACTGTTTTTATCTTTTATAATAGTCTTTTTTGATTTCTACTTTTTCAAATCCCATTGTTTCATAAAGATGTAATGCAGCGCTGTTACTTGGCCTTACCTCTAATAT
```

```
>GCA_000465915.1        1        copy:6(5-6)|18663|Segment13|1051931-1052222|1|length:292|1-292
TTGAAGTAGCTCACTGATTCATTGATAATCATATTTTTTCCGCTAAGAAACGACATGGAGTCGTGAATGAGTATCACTGAGCTGTCAGATATTTTAAATGGTTATTTTTCATGGAACAAGTCGCGAATCGAGTGTTTTGCGACGATGTTAATCTCTT
>GCA_000500125.1        169      copy:2(2-2)|57904|Segment81|2608300-2609267|1|length:968|1-968
ACTATGCTTCAAAATAGCACGATGGTGCAAACGAACGACCTTAGCCGATACAAAGATTCCTAAATGCCCAACGGAAGGATGAAGCAGATACACCACTCGCTGTTTGGCCTTCTTCAATGCTTGTGTCGTAGGATAAATTGTTCTTATCCAATGTAAAG
>GCA_000468235.1        597      copy:2(2-2)|55886|Segment101|2718693-2719011|1|length:319|1-319
GGCAAAAACCGAGCCAATGATTTAAAAAATGAGCTTAATAAAGCACAAGATTTAAAGACTATGCTTGATGTTGTCCAAAATCATTTCGCAAATAACTCCCGATTACATAATCATTCCCTGGATAGCTACCTGCTAAGAGCCATGCATAAAGATTTTAA
>GCA_000465915.1        137      copy:3(3-4)|18663|Segment292|759823-760192|1|length:370|1-370
ATATCCGTTTGGGAGAAAGTAAAATGCCTAAATATGTAGAAGGGGTAGAATTAACTCAAGAAGGCATGCATGCTATTTTTGCTCGGATTGGGACATCCCGAAATTAAAAGCGGTACTATTTATAATGGAGTACCTACTATTGATCAAGAAGCTCTCGAC
>GCA_000468235.1        663      copy:2(2-2)|55886|Segment31|1993901-1994240|1|length:340|1-340
ATGTTTGCGTCTTTGATGTATTTAGTCGATTGTTCCTCGGACTGGTCATAATGGTATTCTTTTTTGTGTAATTGAATAAAAACGGCAGGGTCAAGATCAAAATCTTTGGATTTCTTAAGTGCTTGCTCCATGGCCTGATAGATTGCCCCATCG
>GCA_001549915.1        732      copy:3(2-3)|68365|Segment13|1577680-1578028|1|length:349|1-349
TTAATCCGGGGTCATTGGGGAATAACAGAAAGAGTAGAGGCTATGGGTAATTGGGCTAATAAATGTATTTTTATAGCCTTCGCCTGATGTTTTACCCAGTGGGTCAAAGAGAAAGTATTGATTACCAGGCCCTTTGATTAACATGATCCAGTGACCAGAGT
>GCA_001549915.1        668      copy:3(2-3)|68365|Segment14|1577780-1578127|1|length:348|1-348
GTCAAAGAGAAAGTATTGATTACCAGGCCCCTTTGATTAACATGATCCAGTGACCAGAGTCTCTTTTAGGGCCCACTCCTGTTAAAATGGGCATGAAGCCTTGCCTGTTCAGGGCTTCTGTGTCAATAGTCGGTACTCCATTATAAATGCTTCCGCTGG
>GCA_000468235.1        541      copy:2(2-2)|55886|Segment75|2323852-2324247|1|length:396|1-396
CCCTAGTTCCAGTCAAAGTGACCCTTGATGGGAAAACAAAAACAGGAGAGAACCGATATATCTTTACCTTTGTTGCCGTAAAAAGCCCTGACTTTATACCACGACATGAAAGCGGCTATGCGGTTGAACCGTCGTTTTAAGAATGCAAGCTACCAAGTTA
>GCA_001549925.1        423      copy:2(2-2)|68366|Segment88|785436-785560|1|length:125|1-125
CATTAGAAGCATTTAACTTTTTATAAAGTGAGTAAAAAACTACTGATACAAGCTTTTTCTTATATCAGTAATTATGCCTTAGATTGAGGAAATTATAATTGTGTAAAATTATGACGGTATCGTTC
>GCA_001753145.1        195      copy:2(2-2)|76290|Segment61|3364645-3365449|1|length:805|1-805
ATTGTGCATAGCGTTGAAGTAGTTAAACCACATGTAACCAACGTCTAAAGTCAAATCGCCTTGAGCCATTGCGTAAGTGTAGTCAGCGCCAAGCTTAGCTTCCAACTCAGGAACGATGGCATTTTTAGAACCAGTAACGAAACCTATTCCATCATAGA
>GCA_000468235.1        321      copy:2(2-2)|55886|Segment88|2634380-2635275|1|length:896|1-896
TTTTTAGAACCAATGGTCAAAAGATATTTGAAACCAGATAACCGTCTCAGTTGCTCCTAAAAAGTTAATTTCATCTTGCCTCGTTAAGTCTGCTGTTTTTTTCAAATCCACTTGCCAACGCCACATCACATAAACCACAGGAATCACAATTAAAGTC
>GCA_000500125.1        337      copy:2(2-2)|57904|Segment84|2610267-2610499|1|length:233|1-233
GATTATCAAGACGCTGGCCAATTGACTGACATAGCTTAGATTTTGCTGCAGTGTTTTTTCTGTTCCTCTGATAAATCCTTGCTTAAAGCAGGTAATGCGTTGACTAAATCACGAATGGCAAATGCATGCTGGTGTATTGTCGTTTAAATCATCACTA
>GCA_000468235.1        601      copy:2(2-2)|55886|Segment136|3622967-3623214|1|length:248|1-248
TAACCCAATCTTGCCACGCATAAGCCTGCTTTTTGACATCGCTCTATAAATTCACTGTTTGATACGGGAGCAATTATCGACTTTGCAGATGGACCATATAGAATGAGGCTATTTCTTTGCAACAAATAATAATTAATAAGCCAGTCAGTCCCCGCCTGT
>GCA_000465915.1        643      copy:3(2-3)|18663|Segment155|3201308-3201569|1|length:262|1-262
GTGAAGTACATTCTGGCTGTTTTATCATACCCTGTAGGAATCCTTCTAAGTGGTGATATAACCTTTTTTCGACCAGTTATTATAAACGATTTATAAACATCTTTTCCAGTTGCCATGGTACGGTGGCAAGTCACACCATGTCGCACCAGCACAAAGCATCG
>GCA_001969405.1        567      copy:3(3-3)|81478|Segment25|2193879-2194585|1|length:707|335-707
GCGCTGTTGGCCATGAGCAGGGTATTCTATAGCATATTCCTTTACTGCTAGTTCAATAGACTCATCAACTCGATTCTTGTGATTAGGTTTTCTGCGAGTTTGATCGAATAATGCATCAACTCCACCAGACTCTACTGCTGATTTGTAACGATAAAAAG
>GCA_001969405.1        510      copy:2(2-2)|81478|Segment93|3041666-3041950|1|length:285|1-285
AAAGGAGACCGAGGGGAGGCAGGACCTCAGGGATTACCAGGCCCAAAAGGAGACAGAGGGGGAGGCAGGACCGCAGGGATTACCCGGGCCGAAAGGAGACCAAGGGGAGGCAGGGCCGCAGGGATTACCAGGACCTAAAGGAGAACCGAGGGGAAGCAGG
>GCA_001969405.1        278      copy:11(11-12)|81478|Segment153|704289-704743|1|length:455|1-455
TAAGCGTGGCTCAATACGAATCTATTAAACACCCAACAGGTGGGGCTTCTTATGACGAAGCCGTGAATTTGTCCGGCAAGATCAACGCGGTGGTGACAGTGACCGACGGCGATGGGGATGTGGCCACACAAGCGATAGGCATAGGCATAGGCGATGCGATTGTT
>GCA_000465915.1        445      copy:2(2-2)|18663|Segment279|1842379-1843298|1|length:920|1-920
AACTACAACTTGAGGAGGGGCAAATTCAGGAAAAAACATCTACCGGCATTTGCTTTAGAGTATAACCACCCAACAAACAAAGAATAATCGTTAATGCCAAAATCAAGAGACGATTATGCAAAGACCAGGCAATAAGGCGAGGTAAACATTTAATGATGCT
>GCA_001969405.1        202      copy:3(2-3)|81478|Segment22|2193381-2194591|1|length:1211|1-1211
TAATTTGTAGTTCAAACTTGATCTGACAGTTGCCGGTTTTTCAGAAGTGTCTGTCAGGTCAAATTCAGTTTATAAATTTTTCCATCCAGATTTGTTTGCTATCCATCAAAGTTTTGTATCGGAGTGCGTCCACAACACATTTTGCCTTGATGGGTT
>GCA_001639045.1        743      copy:3(3-3)|73559|Segment60|815993-816256|1|length:264|1-264
TAAAAATGGCTTTCAATCCAGTTCTGATAAAGCGAGAAGACCACCCAATATGGATGAAATTGCGGCTGCCCGCATCATTCAGCAAATCATGTCTAATCCTGATTGCATTATTGACGATCATGTGCGTATCAATGGCCAAGAACTTAATGAAAAATTCTT
>GCA_000465915.1        578      copy:2(2-2)|18663|Segment236|3598532-3598782|1|length:251|1-251
TACTAATATTTTGATGCATGATCCTACTGCTATCTACCATTTTGATGAGCATGATATTTATTTGCCAATGGTGGTGCTTGAAGAATTGGATAGTCATAAAACAGGGGACATCCGAATTGGCGAGAAACGTGAGGCAAAGCAATGCGACGTTGGTAGAGC
>GCA_001753145.1        199      copy:8(8-8)|76290|Segment107|697950-701012|1|length:3063|1-3063
TACGCCTTGAGTGTGTCCGCCCAAGGGGTGGATTCCGGCCTTGATAACAACGGCAACAACATCTACCTGTACAACATCGCAGGTTCTGTGGTGGGTTCAACATCAGCCACGCAGGCGGGTATTACCACAGGCAACACCATCTTTTCATTGGGCGTGAA
>GCA_001549925.1        740      copy:2(2-2)|68366|Segment41|2617171-2617385|1|length:215|1-215
AATGTGATGAGTTAGGGTTGTTTTGATTTCCCCTTCGTTTATTTCAATGCCTTCACCTGCCTTGTAGGTTCCGGCGGGGCCTTGATCTCCTTTAGGGCCTGCTGGCCCCATGGGTCCGGGTGCTCCTTGGGGTATGGAGCTGATTTGTGTTTTGTAAGG
>GCA_001549925.1        680      copy:2(2-2)|68366|Segment109|809637-810455|1|length:819|1-819
TTCACTTTTTTCTTTTTTCATTGTAAATTTTAGCAGTTAGATCTAAACCTTCGGCGATTTTATTTAACTCCACATCCAATAATGGCTTGGAAGAATATTTTGACCAGGTAATTGCACTGAGGTTCTCGTGTTGCAGATTATTACGAGAAAGCAATCTTA
```

Candida auris

ClusterID	Distribution	Marker	Percent_of_strains	Weighted_Average_Copy
Cluster538	6(1)|7(2)|10(1)|13(1)|	*	100.0%[5/5]	8.6
Cluster151	6(2)|7(1)|8(1)|12(1)|	*	100.0%[5/5]	7.8
Cluster206	2(1)|7(2)|9(1)|12(1)|	*	100.0%[5/5]	7.4
Cluster843	1(1)|2(1)|3(1)|10(1)|20(1)|	*	100.0%[5/5]	7.2
Cluster236	0(1)|2(1)|4(1)|10(1)|20(1)|	*	80.0%[4/5]	7.2
Cluster28	4(1)|5(1)|6(1)|9(1)|11(1)|	*	100.0%[5/5]	7
Cluster16	5(1)|6(2)|7(1)|11(1)|	*	100.0%[5/5]	7
Cluster896	1(2)|4(1)|8(1)|20(1)|	*	100.0%[5/5]	6.8
Cluster706	1(2)|4(1)|8(1)|20(1)|	*	100.0%[5/5]	6.8
Cluster39	1(2)|4(1)|8(1)|20(1)|	*	100.0%[5/5]	6.8
Cluster959	1(2)|3(1)|8(1)|20(1)|	*	100.0%[5/5]	6.6
Cluster544	1(2)|2(1)|9(1)|19(1)|	*	100.0%[5/5]	6.4
Cluster960	1(2)|2(1)|8(1)|19(1)|	*	100.0%[5/5]	6.2
Cluster895	1(2)|2(1)|8(1)|19(1)|	*	100.0%[5/5]	6.2
Cluster86	1(2)|2(1)|8(1)|19(1)|	*	100.0%[5/5]	6.2
Cluster802	1(2)|2(1)|8(1)|19(1)|	*	100.0%[5/5]	6.2
Cluster731	1(2)|2(1)|8(1)|19(1)|	*	100.0%[5/5]	6.2
Cluster36	1(2)|2(1)|8(1)|19(1)|	*	100.0%[5/5]	6.2
Cluster111	1(2)|2(1)|8(1)|19(1)|	*	100.0%[5/5]	6.2
Cluster649	1(2)|2(1)|7(1)|19(1)|	*	100.0%[5/5]	6
Cluster569	1(2)|2(1)|7(1)|19(1)|	*	100.0%[5/5]	6
Cluster349	1(3)|8(1)|19(1)|	*	100.0%[5/5]	6
Cluster962	0(1)|1(2)|8(1)|19(1)|	*	80.0%[4/5]	5.8
Cluster905	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster893	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster735	1(2)|2(1)|6(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster72	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster71	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster65	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster554	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster550	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster545	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster352	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster234	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster226	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster223	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster219	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster215	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster214	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster213	1(3)|7(1)|19(1)|	*	100.0%[5/5]	5.8
Cluster91	0(1)|1(2)|7(1)|19(1)|	*	80.0%[4/5]	5.6
Cluster902	1(3)|7(1)|18(1)|	*	100.0%[5/5]	5.6
Cluster82	1(3)|7(1)|18(1)|	*	100.0%[5/5]	5.6
Cluster803	0(1)|1(2)|8(1)|18(1)|	*	80.0%[4/5]	5.6
Cluster738	1(3)|6(1)|19(1)|	*	100.0%[5/5]	5.6
Cluster454	1(2)|2(1)|6(1)|18(1)|	*	100.0%[5/5]	5.6
Cluster377	0(1)|1(2)|7(1)|19(1)|	*	80.0%[4/5]	5.6
Cluster360	0(1)|1(2)|7(1)|19(1)|	*	80.0%[4/5]	5.6
Cluster356	1(3)|6(1)|19(1)|	*	100.0%[5/5]	5.6
Cluster354	1(2)|2(1)|6(1)|18(1)|	*	100.0%[5/5]	5.6
Cluster350	1(3)|7(1)|18(1)|	*	100.0%[5/5]	5.6
Cluster211	0(1)|1(2)|7(1)|19(1)|	*	80.0%[4/5]	5.6