charles river

Summary

The sequencing methods used by Charles River include automated data assembly, analysis and review of data quality, genetic distance measurement, and result interpretation. These elements, with a validated and relevant reference library, ensure the most accurate microbial identification results.

MICROBIAL SOLUTIONS

Click to learn more

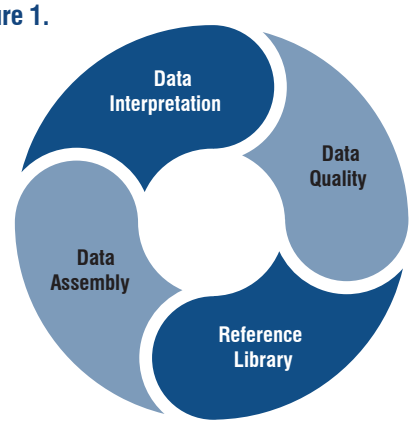
Attributes of the AccuGENX-ID® Sequence-Based Microbial Identification Method

Significance of Microbial Identification on Environmental Monitoring Programs

When bacterial or fungal isolates are recovered from a production facility, it is extremely important to accurately identify the organism to the species level in a timely manner. Reliable microbial identification systems are needed for consistent and accurate results to permit tracking of organisms, complete investigations, and avoid delays in product release. Methods that yield inaccurate, inconsistent, or no identification are neither useful for tracking isolates to their source, nor for generating trending reports. Ultimately, this can lead to a false sense of control, misdirected remediation efforts, and additional operational costs to the organization. Accurate methods for identification will yield a reliable historical database that allows for sound data comparisons, interpretations, and mitigation of risk.

There are four main areas that specifically impact the accuracy of sequence-based microbial identifications: 1) the relevancy and depth of coverage of the reference library, 2) the quality of the data generated by the system, 3) the assembly and analysis of the data and calculation of the distance measurements, and 4) interpretation of the results (Figure 1). The methods used by Charles River can be traced back to the methods used by taxonomists, and these methods analyze microbial sequence data through a conventional assembly process. This process includes auto-assembly, review of the data, and a phylogenetic analysis to generate an identification.

Figure 1.



There are four main areas that specifically impact the accuracy of sequence-based microbial identifications: 1) the quality of the data generated by the system 2) the relevancy and depth of coverage of the reference library 3) the assembly and analysis of the data and 4) interpretation of the results.

Increasing Accuracy and Reliability

Comparative sequencing of the ribosomal RNA genes (rRNA) is a genotypic technology that is the gold standard for the taxonomic classification of bacteria and fungi. It is also the most accurate and reproducible method for identifying unknown microorganisms. Although the science behind the technology is widely accepted, there are still a number of variables that can impact its implementation and the achievement of an accurate identification. The analysis of DNA sequence data is a complicated process, and the overall performance of the different ribosomal DNA (rDNA) identification systems or services being used today is not uniform. There are numerous steps that require an experienced scientist to make decisions in order to correctly analyze and interpret the data. These critical steps include the evaluation of DNA sequence quality, DNA sequence assembly, and phylogenetic interpretation. Higher quality results are obtained when key steps are performed by an experienced microbial phylogeneticist and adhere to the taxonomic process. The accuracy of an identification is dependent not only on the methods used to generate, analyze, and interpret the data, but also significantly on the library database used as a reference. For almost 20 years, the Charles River Accugenix® group has been identifying microorganisms using comparative rDNA sequencing. Our experience, in conjunction with the improvements we have made to the analysis process and our validated, relevant libraries, allows us to provide the highest quality identification results for unknown isolates.

Impact of the Reference Library

Recognizing that the reference library is a key component for microbial identification, Charles River gives utmost priority to routinely maintain and annually re-qualify its validated microbial DNA libraries with the goal of providing the most accurate identifications for all bacterial and fungal samples. Superior performance, reliability, and relevancy of microbial identification systems require libraries that exhibit breadth and depth of coverage of isolates that are important to the sterile and non-sterile manufacturing industries. Differences in the reference libraries against which you compare your data can affect the frequency of correct microbial identifications. In order to correctly identify a large percentage of the unknown isolates in manufacturing environments, the library must contain DNA sequences for the organisms most likely to be encountered. Charles River has documented the organisms found in these facilities around the world, increasing the relevancy of our libraries.

Each day, new bacterial and fungal species are being discovered, named, and published. These new species may represent organisms that have been previously encountered in manufacturing facilities, and could pose a risk to the product. In addition, the taxonomy of organisms is constantly changing. With the emergence of DNA sequencing as the gold standard for classification and identification, scientists are correcting past taxonomic errors that were caused by solely phenotypic characterization and not phylogenetic relatedness. The Accugenix® libraries are continuously curated and updated to reflect taxonomic changes and inclusion of novel organisms encountered in manufacturing environments. Serving as a contract laboratory for highly regulated manufacturing environments dictates that Charles River follows a rigorous cGMP compliant program which encompasses our original library validation procedures and continues to drive the maintenance of our libraries.[1]

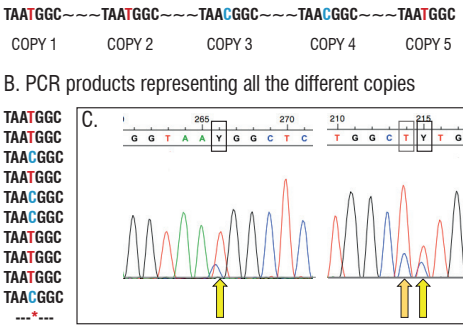
Data Analysis and Assembly

In addition to the reference library, factors affecting the accuracy of identifications also include the method of data analysis and the data interpretation guidelines. Precision is affected by the ability of the software on the sequencer to interpret the different types of data anomalies that affect base calling during the normal sequencing process, as well as by the method of reconciliation of potential errors in these base calls. These errors are caused by peak mobility shifts, base insertions or deletions, and mixed base positions in multi copy genes. Accuracy is also affected by the length of the resulting consensus sequence used for comparison to the library where the consensus sequence comprises the most commonly encountered nucleotides found at each location in the target DNA sequence alignment.

During data analysis, we first confirm that the data are of a high enough quality to generate an accurate consensus sequence. The initial data quality check is performed to classify the raw data as acceptable, not acceptable, or requires verification. Data quality is impacted by the rRNA operon itself. The rRNA operon comprises the genes for the three different-sized rRNA molecules (e.g., the 16S, 23S, and 5S rRNA genes in bacteria or the 18S, 5.8S, and 28S rRNA genes in fungi). In most species, the ribosomal operon is multi-copy, up to 15 copies in some cases. The copies are arranged in tandem on the chromosome, and often not all the copies in each species are the same sequence or same length. However, all copies of the gene are amplified since universal PCR primers are used to amplify the target rRNA regions for identification. When these PCR products are sequenced, the sequence represents a summation of all the copies. The differences in sequences between the 16S rRNA copies, for example, can result in polymorphisms or mixed-bases at different nucleotide positions. A polymorphism is seen in the raw data electropherogram as distinct nucleotides at the same position in the sequence (Figure 2). Care must be taken during data review to call the mixed base positions as they are evolutionarily important.

Figure 2.

A. Genomic organization of rRNA operons



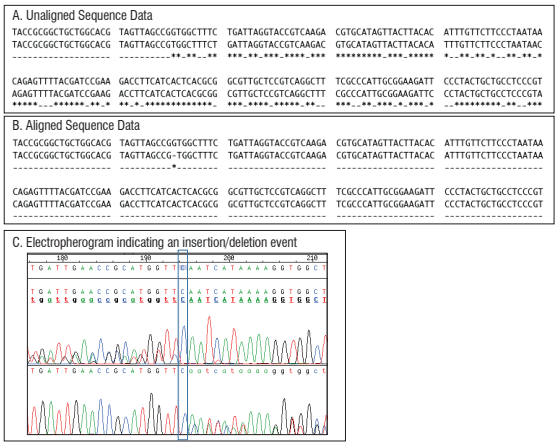
B. PCR products representing all the different copies

D.Consensus Sequences: **GGTAAYGGC TC TGGCYYTGCT**

Visualizing and determining polymorphic base calls. A) Copies of the rRNA operons are arranged tandemly in the genome, and may have different sequences. B) The gene target is amplified with PCR, and visualized via Sanger sequencing. C) Both bases can be seen at the same position in the electropherograms indicating a mixed base polymorphism. The positions indicated with the yellow arrows in the two electropherograms are correctly identified by the analysis algorithm as a "Y," a mixture of T (red peak) and C (blue peak), while the purple arrow indicates a position where the base calling software failed to recognize the polymorphism. C) Review of the data detects and reconciles these base calling errors, recording the polymorphic position of "Y" in the consensus sequences.

Additionally, for some of the isolates analyzed, the lengths of the rRNA genes are not the same. There has been either an insertion or deletion of bases during evolution. This requires the same care during data review to capture these events (Figure 3). For example, data which result from an organism that contains multiple copies of the 16S rRNA gene, or ITS rRNA region, which are not the same length due to an insertion or deletion event appears very similar to DNA sequences derived from a mixed culture. However, our experienced data review scientists can easily tell the difference between sequences from mixed cultures and sequence data that contain insertions or deletions, and use a proprietary software to reconcile these events. This ability allows for successful analysis and assembly of a higher number of samples and increased performance.

Figure 3.



Example of base calling errors that result from an insertion or deletion event in one copy of the 16S or ITS region. A) Unaligned data results in many mismatches (indicated by asterisks). B) Aligned data compensates for the insertion/deletion and indicates only one mismatch. C) Image of the raw sequence data showing effect of an insertion/deletion (at blue box) on peak patterns making the data on either side of the event appear "mixed."

Evaluating the quality of the sequence data, and reconciling polymorphic positions, ensures that our analysis uses the entire sequence generated from the PCR product. No nucleotides between the PCR primer sites are omitted from our analysis and a full-length consensus sequence is generated. Our experience has shown that the most accurate distance measurements, and therefore the most accurate identifications, are generated using the full-length DNA sequence.

Distance Measurements

Bacterial and fungal taxonomy reflects a phylogenetic classification scheme based on the sequence from the 16S rRNA gene or ITS region as visualized by dendrograms, or trees, whose structure is based on different models of evolution. We use the same process for our identifications. Once a consensus sequence is generated, it is compared to our reference library using the basic local alignment search tool (BLAST). The search determines the closest library reference matches to the unknown sequence. These reference library entries enter in the phylogenetic analysis pipeline with the calculation of percent differences and genetic distances.

First, the sequences are aligned to minimize the absolute number of differences between the two sequences. Next, the sequences are compared at every nucleotide position, a pairwise comparison, and the percentage difference is calculated. The percent differences reflect the absolute number of bases that are different between the unknown and the reference across the complete sequence. A simple example: six base differences across a 500 bp gene region is 1.2% different (6/500*100=1.2%). The most closely related reference sequences and the unknown sample are then analyzed according to a distance matrix model. Distance measurements are a comparison of one sequence to another and are also expressed as a percentage. Distance matrix models are based on what is expected to give rise to differences between present day sequences by comparing them, counting differences, and applying corrections for superimposed mutations. The real number of changes can be underestimated by just looking at the difference seen in the present sequences. Thus, the differences are transformed to distances using models such as Jukes Cantor. The evolutionary distance measurements are calculated and the Neighbor Joining tree, the model used by Charles River and other phylogeneticists, is constructed using these data. The tree is the core to the identification report and requires interpretation.

Data Interpretation

Interpretation of the phylogenetic tree is a critical part of the analysis process since no interpretation rules can be applied universally. There is no phylogenetic standard for family, genus, or species demarcation. There is also no universal agreement on a threshold value that constitutes a definitive species or species delineation. Because bacterial genera do not evolve at the same speed, and because many organisms were misclassified as the same or different species prior to the evaluation of the rRNA gene sequence, it is necessary to use different threshold values depending on the genus in question. Thus, the threshold value varies by the taxonomic group as some species are too closely related, or there is little information of normal strain to strain variation within a species or inter-operon variation within a strain. The most important consideration in making an identification is the phylogenetic tree. A phylogenetic tree is a visual representation of the genetic variability between the most closely related organisms in the library to the unknown. Both the distribution and the branching order indicate how organisms relate to one another and are important in making the final interpretation.

Our microbial phylogeneticists are very experienced and recognize where potential problems may lie, and how the problems could make the interpretation confusing. One problematic area is that the phylogenetic classification of many organisms is currently incorrect. Organisms are misclassified and misnamed, creating a very complex situation. A further problematic area results when the phylogenetic analysis of the ribosomal gene regions is not sufficient for species resolution. This is true of the *Burkholderia cepacia* complex. There are many organisms that are too closely related and have a high degree of

conservation in the ribosomal RNA regions such that 16S or ITS2 sequencing is not sufficient to make the classification any more accurate than at the group or complex level (our "species*" confidence level, which indicates that "*The unknown matches two or more closely related species"). Sequencing information from other genetic markers such as protein-coding gene targets like gyrB, recA, or TUB2 can be used to increase discrimination. However, in many cases, species resolution cannot be achieved even with protein-coding gene sequencing due to intricate or ill-defined taxonomy. We then have to accept an identification at the group or complex level (species* confidence) as this is the most accurate answer.

Our microbial phylogeneticists use a combination of the genetic variability, branching order of the neighbor joining tree, and knowledge of the interspecies variation when interpreting the reports and assigning the taxonomic confidence level. Based on our experience of identifying over 1,000,000 unknown microorganisms, Charles River has the experience and knowledge to convert this complex information into routine identifications.

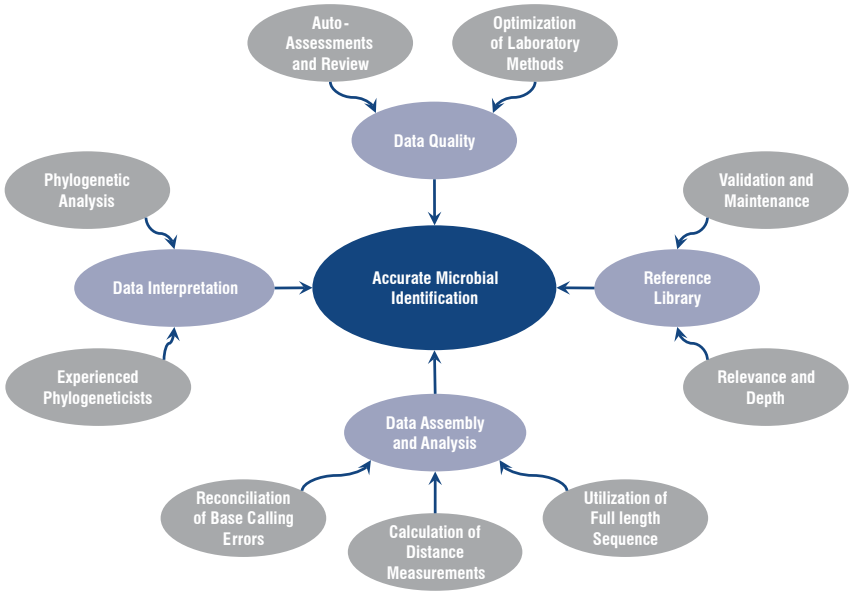
Summary

The sequence-based phylogenetic identification methods used by Charles River for our AccuGENX-ID service can be traced back to the methods used by taxonomists, and these methods analyze microbial sequence data through a conventional process. This process includes auto-assembly, review of the data, and a phylogenetic analysis to generate an identification. The proprietary methods employed by Charles River compare sample sequences against full-coverage proprietary libraries, result in conclusive data interpretation, and an identification with assigned confidence levels based on phylogenetic analyses. Utilizing these scientifically proven methods, and validated, continuously curated, proprietary libraries provide the most accurate microbial identifications (Figure 4). These data, in turn, will allow for accurate tracking and trending of isolates, as well as high confidence in the interpretation of the results.

Reference

1. Creating and Maintaining Validated Microbial Identification Libraries. 2018. Charles River Technical Note.

Figure 4.



Factors that impact the accuracy of a microbial identification include data quality, data assembly and analysis, data interpretation, and the reference library.

charles river

askcharlesriver@crl.com • www.criver.com