A Comparison of the Attributes of the AccuGENX-ID Microbial Sequencing Methods versus MicroSEQ v2.0

Executive Summary
A detailed analysis of the differences between AccuGENX-ID™ and the MicroSEQ® v2.0* methodologies in obtaining an accurate identification are discussed. We focus on 4 areas that impact the accuracy of microbial identifications: 1) Library relevance, depth of coverage and validation; 2) Analyzing the quality of the data generated; 3) Data assembly analysis and distance measurements and 4) Results Interpretation. The semi-automated methods used by Accugenix can be traced back to the reference methods used by taxonomists, and these methods analyze microbial sequence data through a conventional assembly process. This process includes auto-assembly, manual review, correction of base caller errors and demonstrates a 98% accuracy and six-fold better repeatability than the fully automated method used by MicroSEQ v2.0.

Increasing Accuracy and Reliability
Comparative DNA sequencing is a genotypic technology that has been the gold standard for classification and identification of bacteria and fungi since 1994, as it is 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 which can impact the implementation of this technology to achieve an accurate identification. The analysis of DNA sequence data is a complicated process, and the overall performance of the current ribosomal DNA (rDNA) identification systems is not uniform. There are numerous steps which require an experienced scientist to make decisions in order to correctly analyze and interpret the data, including 1) DNA sequence quality, 2) DNA sequence assembly and 3) phylogenetic interpretation. Attempts to automate this process, such as with MicroSEQ v2.0 in 2005, have not been entirely successful. Higher quality results are obtained when key steps are performed by an experienced microbial phylogeneticist and adhere to the reference 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 dependent on the library database used as a reference.

For more than 10 years, Accugenix has identified more microorganisms using comparative rDNA sequencing than any other service laboratory in the industry. 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 of unknown isolates. We have previously presented a comparative study utilizing the different data analysis methods and libraries. In that study, we effectively demonstrated that our reference method has a six-fold increase in precision and that our proprietary library also provides a two-fold increase in species coverage, leading to a two-fold decrease in unidentified samples as compared with MicroSEQ v2.0'.

Key Factors
Data Interpretation, Data Quality, the Reference Library and Data Assembly are all factors that have interdependencies and impact the quality and accuracy of a microbial identification.
Genotypic Identification

Recognizing that the reference library is a key component for microbial identification, Accugenix gives utmost priority to routinely maintain and annually re-validate its microbial DNA libraries with the goal of providing accurate identifications for all customer bacterial and fungal samples that have been described as a published species. Superior performance, reliability and relevancy of microbial identification systems require libraries that exhibit breadth and depth of coverage that are important to the sterile and non-sterile manufacturing industries. 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. Accugenix has documented the organisms found in these facilities around the world and as a result, our validated, cGMP compliant libraries contain more than twice the number of relevant species than any available library supporting commercial phenotypic, proteotypic or genotypic identification systems.

Each day, new bacterial and fungal species are being discovered, named and published. On average, about one out of five of these new species has 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 technology as the new gold standard for identification and classification, scientists are correcting past errors that were solely based on phenotypic characteristics and not phylogenetic relatedness. The libraries at Accugenix 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 Accugenix follows a rigorous cGMP compliant program which encompasses our original library validation procedures and continues to drive the maintenance of our libraries.

Understandably, differences in the reference libraries against which you compare your data can affect the frequency of correct microbial identifications. Accugenix publishes a list comparing the species coverage of all major commercial systems. Using this species coverage list and the known frequency of occurrence, one can estimate the performance of these systems, and we find this estimate closely reflects actual experimental results. Additionally, our comparative studies show that the methods and software used to analyze and assemble rDNA sequences can impact the accuracy and reliability of the identification system.

Data Analysis

Factors affecting accuracy of identifications include the method of data analysis and the interpretation guidelines. Precision is affected by the ability of the software to detect all the different types of base calling errors that occur during the normal sequencing process, and how the software and/or manual intervention reconciles these base calling errors which are caused by peak mobility shifts, insertions, deletions and mixed base positions. Precision is also affected by the length of the resulting consensus sequence used for comparison to the library. Unlike the MicroSEQ v2.0 system, Accugenix uses the exact same reference method for analysis of customer samples as it does for making library entries, thereby ensuring compliance with our cGMP methodologies. Taxonomists also employ the reference method, which involves the use of a reference sequence during data analysis, assembly and interpretation, when making new species classifications. First, we confirm that the data are of a high enough quality to generate an accurate consensus sequence. The initial data quality check is performed automatically to classify the data as acceptable, not acceptable or requires a manual verification of the raw data. Based on comparative studies we have performed, we have learned that some data will automatically be rejected by the MicroSEQ v2.0 software due to constraints in their algorithms.

In most species, ribosomal targets are multi-copy genes and often not all the copies in each species are the same. The differences in sequences between the 16S copies, for example, can result in polymorphisms or mixed-bases at different nucleotide positions. A polymorphism occurs when two or more clearly distinct nucleotides exist at the same position in a species and requires both software and manual review of the raw data and editing to correct for the mixed base positions (Figure 1). Additionally, for some of the isolates analyzed, the lengths of the copies are not the same, requiring the use of software to correct for polymorphic changes caused by insertions or deletions (Figure 2). For example, data which result from an organism that contains multiple copies of the 16S rRNA gene which are not the same length would be rejected by MicroSEQ's software. The reason for this is that the data resulting from this insertion or deletion event would appear very similar to DNA sequences derived from a mixed culture. However, our experienced personnel can easily tell the difference between mixed sequences and insertions or deletions and use a proprietary software program to correct for these events. This ability allows for successful analysis and assembly of a higher number of samples and increased performance.

**Figure 1.** Electropherograms indicating mixed base polymorphisms. The three electropherograms clearly show mixed base positions that result from variations in the different copies of the 16S sequence. The positions indicated with the yellow arrows are correctly identified by the analysis algorithms as a Y, a mixture of C/T, while the purple arrows indicate positions containing base calling errors. Manual review detects and reconciles these errors, recording the polymorphic position of R, a mixture of A/G in the middle panel, or Y in the right panel.
Data Assembly

After evaluating the quality of the sequence data and reconciling polymorphic positions, our scientists will ensure that our analysis uses the entire sequence generated from the PCR product. No nucleotides between the PCR primer sites will be omitted from our analysis. Commercial analysis programs utilize an algorithm that will automatically remove nucleotides from the ends of the sequence until a certain quality score is reached. Depending on the overall quality of the DNA sequence, the number of nucleotides remaining can range from 50 to 450 of the approximately 500 nucleotides in the original sequence. Our experience has proven that the most accurate distance measurements, and therefore the most accurate identifications, are generated using as much of the DNA sequence as possible (Figure 3). The more of the sequence that is analyzed, the higher confidence we have in the answer. We do not produce an identification report using anything other than the entire sequence.

Distance Measurements

Distance measurements are a comparison of one sequence to another and are expressed as a percentage of nucleotide differences between the two sequences. 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 (pairwise comparison), and the percentage difference is calculated. The calculations based on genome organization and those used by the reference taxonomic method for distance measurements are the same but differ from those used by the current commercial MicroSEQ v2.0 system. As an example of the distance measurements determined by the genome organization (Figure 4), assume we have two copies of a gene target, organized in tandem, 20 bases long. In the library entry, both copies are the same. In the unknown sample genome, the sequences differ by one base in one copy (red asterisk). The distance measurement considers both copies linearly, as organized in the genome, and would be calculated to be 2.5% or 1 difference in 40. Adhering to the reference taxonomic method, where PCR amplifies both copies and creates a mixture of the two copies, we would see a mixed base call at this location that we would score as ½, and the total length of the sequenced PCR product would be 20 bases. The distance measurement by the reference method would be 2.5% (½ difference in 20), exactly the same percent difference as seen based on the genome organization. The commercial method utilized by MicroSEQ v2.0 calculates the distance as the base differences like the reference method (½), but it reduces the value by the ratio of the mismatched base quality score to the sum of all quality scores in the consensus sequence. In this example, the distance would be calculated to be 1%. By using quality scores in the distance calculation, the commercial method systematically understates distances when the sequence contains polymorphic bases. This leads to first choice identification errors and errors in family, genus or species level interpretation.
Data Interpretation

Once the DNA sequence data have been correctly analyzed, compared to a relevant, validated library of known organisms and the distance measurements calculated, the identification report requires interpretation (Figure 5). This is a critical part of the analysis process since no interpretation rules can be applied universally. In addition to the percentage of nucleotide differences, there are other attributes that need to be considered before making an identification. The most important consideration 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 actually relate to one another and are equally important in making the final interpretation. Our microbial phylogeneticists are very experienced and recognize where potential problems may lie and how they 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. The interpretation must take into consideration the genetic variability and branching order of a group of organisms in determining an identification. However, based on our experience of identifying over 1,000,000 unknown microorganisms, Accugenix has accumulated the experience and knowledge to convert this information into routine identifications.

Summary

The semi-automated methods used by Accugenix can be traced back to the reference methods used by taxonomists, and these methods analyze microbial sequence data through a conventional assembly process. This process includes auto-assembly, manual review and correction of base caller errors and demonstrates six-fold better repeatability than the fully automated method used by MicroSEQ v2.0. Table 1 presents a comparison of the key attributes of the different genotypic identification methods. The fully automated commercial method handles all sequence files in the same manner regardless of the data quality, weighs polymorphic positions differently than single copy sequences and provides the top matches, relying on the end user to determine the true identity of the organism. Accugenix’s proprietary methods compare sample sequences against full coverage proprietary libraries, resulting in conclusive data interpretation and an identification with assigned confidence levels based on phylogenetic analysis. Utilizing a method that includes manually assisted automated assembly and continuously curated and validated, proprietary libraries will allow accurate trending and tracking of isolates, as well as a high confidence in the interpretation of the result.

Table 1. Comparison of the attributes of the Accugenix AccuGENX-ID microbial identification sequencing services versus MicroSEQ v2.0.

<table>
<thead>
<tr>
<th>Service or System</th>
<th>Automated and manual verification of data quality</th>
<th>Requires effects of polymorphic positions (insertions and deletions)</th>
<th>Utilized full sequence potential of no truncation</th>
<th>Calculates distance measurements to generate a distance matrix (as of 31 Jan 2012)</th>
<th>Reduces distance measurements for an improved data set</th>
<th>Considers both the distance maternal and a log likelihood ratio for an identification tree (as of 31 Jan 2012)</th>
<th>Calculating maintenance to validated library</th>
<th>Number of species entries (as of 31 Jan 2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Method</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>AccuGENX-ID</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>4568</td>
</tr>
<tr>
<td>MicroSEQ v2.0</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>1228</td>
</tr>
</tbody>
</table>

* MicroSEQ® is manufactured by Applied Biosystems in Foster City, CA.

References: