

Manual Reference Method Versus Commercial Automated Software for Data Analysis and Result Interpretation of 16S Bacterial Sequences

Beth Burke*, Anne Buboltz, Emily Huang, Melissa Ruch and Douglas Smith

Abstract

Currently, 16S ribosomal DNA (rDNA)-based sequencing is the "gold-standard" for identifying environmental microorganisms. While it is the most accurate bacterial identification method, the overall performances of current 16S rDNA identification systems are not uniform. By comparing the performance of Accugenix's 16S rDNA identification method to another commercially available system, we pinpoint several key factors contributing to variation in 16S rDNA identification systems. First and foremost, the breadth of the microbial reference library greatly impacts the accuracy of identifications – Accugenix's larger library, which exhibits thorough coverage of isolates relevant to the pharmaceutical and biotechnology industries, outperformed all other commercially available databases. Moreover, we reveal that the software used to analyze 16S sequences also affects the accuracy of microbial identifications – Accugenix's manual reference method, which enables editing of base-caller errors that typically occur near the end of sequences or polymorphic 16S sequences, is more precise than fully automated sequence analysis methods. Additionally, the direct DNA distance measurements used by Accugenix are also more accurate than quality score methods used by others. Details regarding how library coverage and software parameters affect identification accuracy are reviewed.

Introduction

The publication of all new bacterial species required a 16S DNA sequence as of 1994. In 1998, the first commercial sequence-based identification system became available, required MAC software and was based on a semi-automated assembler and manual correction of base caller errors. In 1999, Accugenix started providing sequenced-based identifications to the pharmaceutical industry, while developing extensions to the original commercial libraries. In 2005, the second commercial version (MicroSEQ® v 2.0) was released by Applied Biosystems for use on PC Windows, that utilized a completely different assembly method and algorithm to analyze sequences. In addition, bacterial and fungal libraries supported on this system were not primarily targeted to the pharmaceutical EM market. By 2007, Accugenix had created its own full coverage libraries focusing primarily on organisms relevant to environmental monitoring programs. The software used by Accugenix allows for a manual base-caller editing approach, which yields high accuracy, full length consensus sequences for report generation. Factors affecting accuracy of identifications include library coverage for intended use, the method of sequence comparisons, and interpretation guidelines. Precision is determined by how the software handles insertions and mixed bases, detection of base calling errors and length of the read sequence.

Methods

To identify factors contributing to the variation in 16S rDNA identification systems, the performance of Accugenix's 16S rDNA system was compared to another commercially available system, MicroSEQ® v 2.1. We compared the library species entries of these systems by analyzing reports for all 262,699 unknown bacterial isolates received at Accugenix since 2006.

To perform a more detailed analysis, the 16S rDNA systems were used to analyze forward and reverse pairs of raw sequence data (.abi files) from 444 bacterial isolates that passed QC standards and were collected from aseptic and sterile manufacturing environments over a two-month period. Initially, sequences were analyzed using either the Accugenix manual reference method, which includes a manual assembly and performs searches against the validated Accugenix proprietary sequence library 04APR11, or the MicroSEQ® v 2.1 system, which includes an automated assembly and performs searches against the MicroSEQ® v 2.2 bacterial library.

Next, these sequences were assembled using either the Accugenix manual method or the MicroSEQ® v 2.1 automated assembly and the consensus sequence was searched using the MicroSEQ® v 2.1 system with the validated Accugenix proprietary sequence library 04APR11. Samples with the same genus/species first choice match or those in a group with identical sequences were considered a species match for this study. Those that did not match at the species level were considered discrepant and placed in one of several categories: genus match, no match and no result. The Analysis Protocol for MicroSEQ® defines a "Minimum Clear Length" of 400 base pairs (or 80% library length). In contrast, all sequences analyzed by Accugenix's method were full length equal to the library entry.

Results

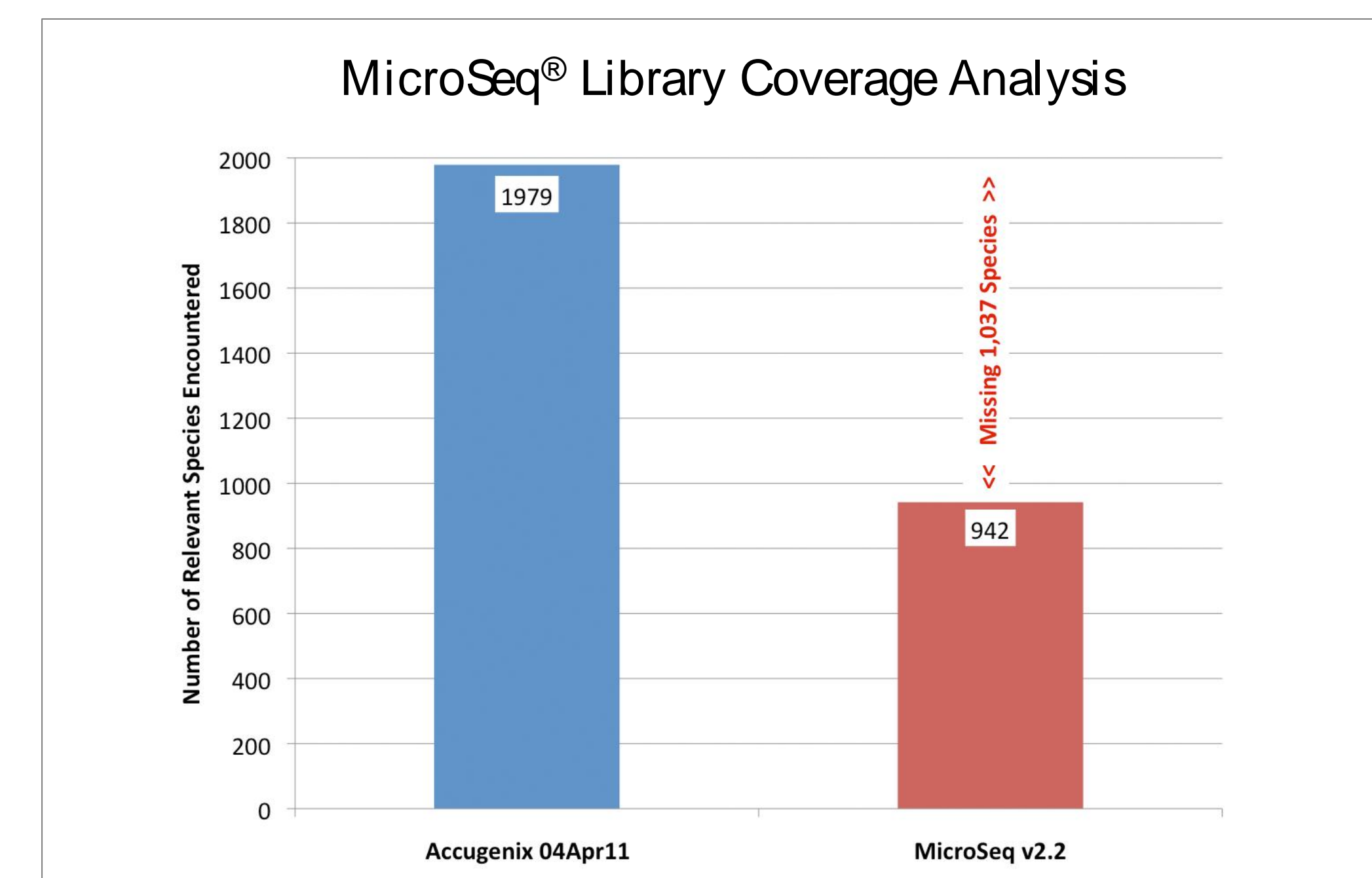


Figure 1. Relevant bacterial species encountered in Pharma EM programs by Accugenix since June 2006 (n= 262,699). Of the 1979 different species reported by Accugenix, only partial coverage is provided by the MicroSEQ® v 2.2 bacterial library.

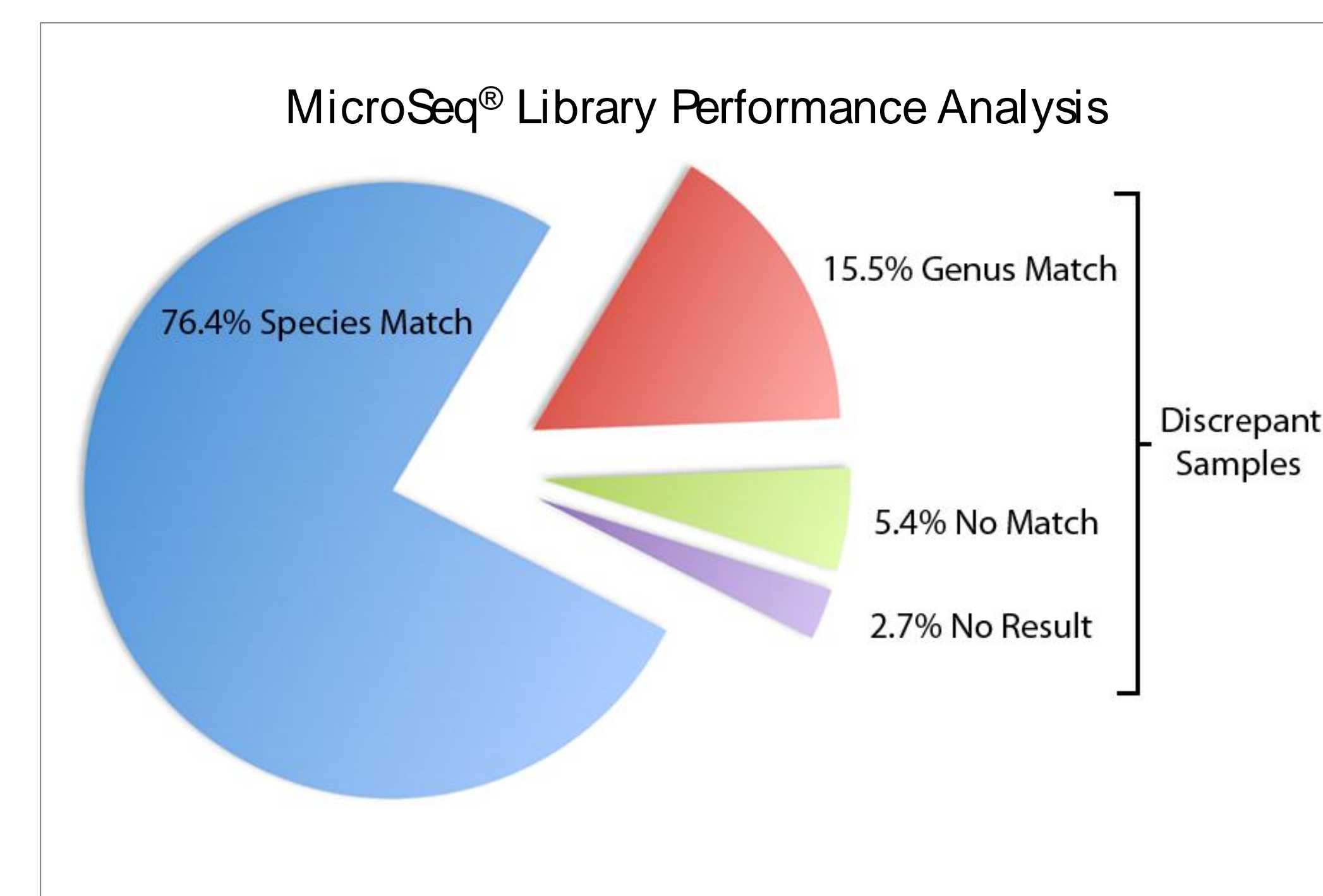


Figure 2. Discrepant results from comparing the MicroSEQ® (v 2.2 library) system with the Accugenix (04APR11 library) identification method. The same forward and reverse 16S raw sequencing data (.abi files) were analyzed by both methods. This dataset included 444 paired samples collected over a two-month period (03Aug10-05Oct10).

Discussion

Although 16S rDNA sequencing is the most accurate bacterial identification method, the overall performances of current 16S rDNA identification systems are not uniform. In this poster, we aim to present examples of this point.

Comparison of DNA Distance Calculations from MicroSEQ® and Accugenix 16SrDNA Systems

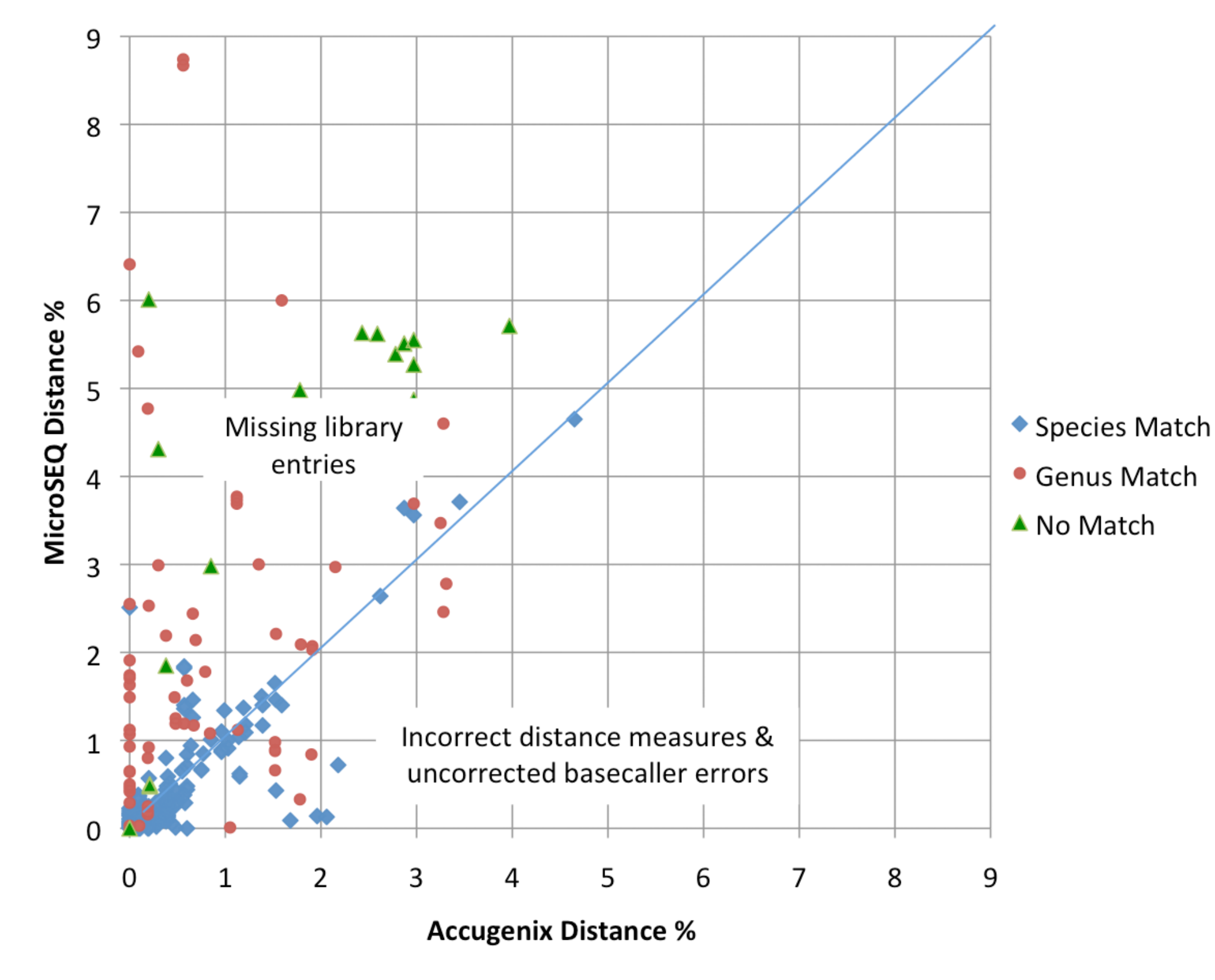


Figure 3. For each sample, the DNA distance was calculated by both the MicroSEQ® v 2.1 system using the v 2.2 library (vertical axis) and the Accugenix 16S rDNA system using the 04APR11 Bacterial Library. This dataset included 444 paired samples collected over a two-month period (03Aug10-05Oct10).

Comparison of DNA Distance Calculations from MicroSEQ® and Accugenix 16SrDNA Systems With Library Correction

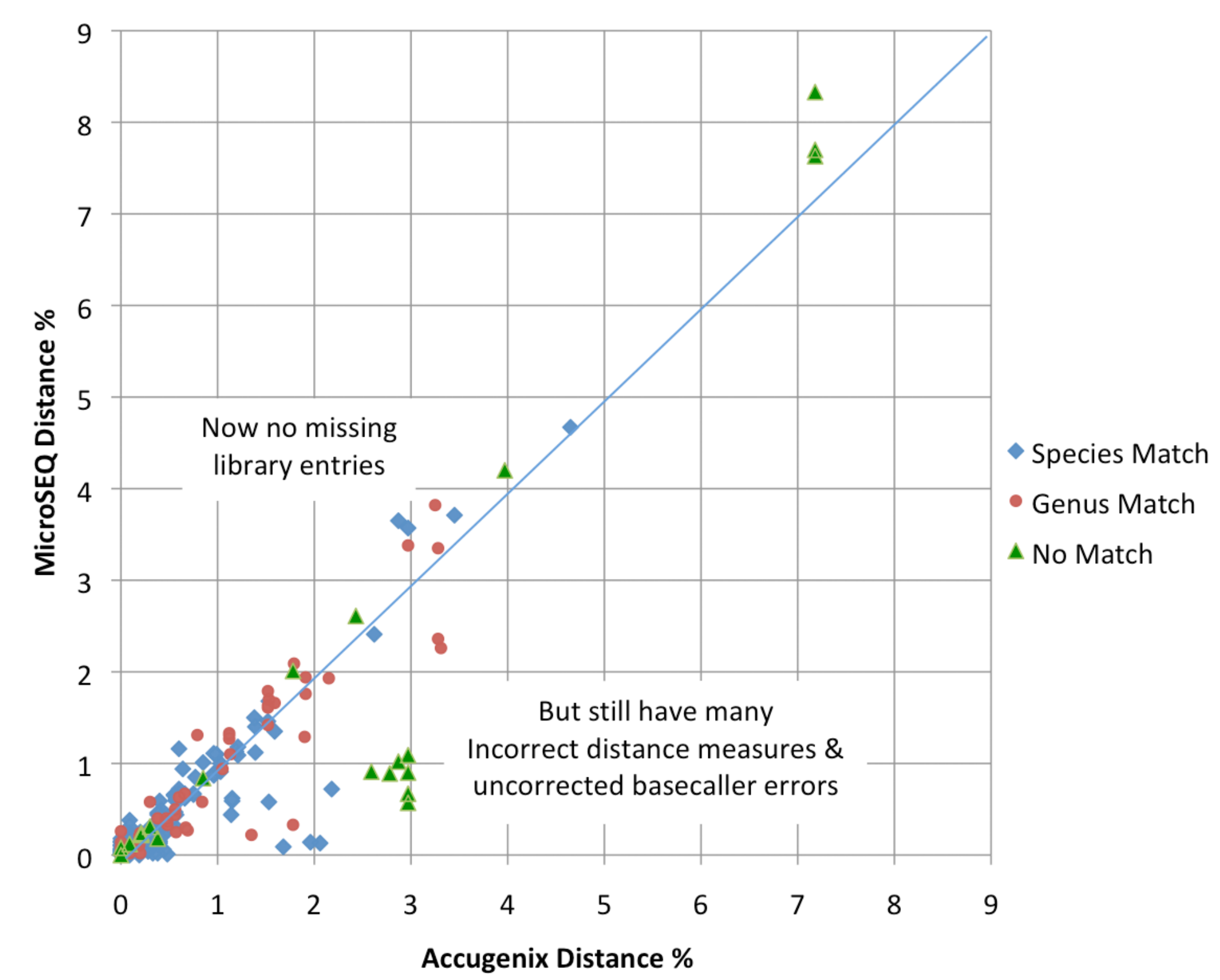


Figure 4. For each sample, the DNA distance was calculated by both the MicroSEQ® v 2.1 system and the Accugenix system. In contrast to Figure 3, all samples were then searched against Accugenix's 04APR11 Bacterial Library; thus, eliminating this as a source of differences. This dataset included 444 paired samples collected over a two-month period (03Aug10-05Oct10).

Using the sequences from 262,699 unknown bacterial isolates received at Accugenix since 2006, we compared the library databases used by two 16S rDNA identification systems – Accugenix's 16S rDNA identification system and MicroSEQ® v 2.1 (Figure 1). Approximately 52% of all the bacterial species that are most relevant to the pharmaceutical and biotechnology industries are missing from the MicroSEQ® v 2.2 bacterial library. This deficiency in the library will limit the successful identification of many bacteria frequently observed in aseptic and sterile manufacturing environments.

Using the raw data sequence files from 444 unknown bacterial isolates received at Accugenix over a two-month period, a more detailed comparative analysis of these 16S rDNA identification methods was conducted. We show that 105 (23.6%) of the identifications made by these two methods were discrepant (Figure 2). Twelve (2.7%) of the 444 sequences were rejected by the MicroSEQ® v 2.1 system, due to short sequences caused by polymorphic insertions or deletions. In contrast, all of these insertions were corrected by Accugenix's identification system. Of the remaining discrepancies, 20.9% were overwhelmingly due to missing entries in the MicroSEQ® v 2.2 Bacterial Library, leading to only a genus level identification or no match. This clearly demonstrates that Accugenix's Bacterial Library results in more than a two-fold reduction in the number of unidentified strains.

Theoretically, if both 16S rDNA identification methods were equivalent, all isolates should identify equivalently with the same distance measurement calculated for every sequence. Figure 3 shows a comparison of the percent similarity to the closest identification match for 444 of the unknown sequences. If all data points fell on the 45° trend line, it would indicate that the identification systems are performing similarly. However, a number of data points do not fall on the trend line, suggesting that the methods are not performing equally. Data points lying above the 45° trend line represent samples whose species library entry is missing from the MicroSEQ® v 2.2 Bacterial Library. In contrast, data points lying below the 45° trend line represent samples where the MicroSEQ® v 2.1 system substituted actual DNA distances with base quality scores. Overall, this data strongly suggests that 16S rDNA identification methods do not perform equally. Moreover, our results indicate that missing library entries, uncorrected base caller errors and variation in the calculation of distance measurements contribute to variation in 16S rDNA identification methods.

Understandably, differences in 16S bacterial libraries can affect microbial identifications. However, our data also suggest that the software used to analyze 16S sequences may also impact the accuracy and reliability of 16S rDNA systems. To confirm this possibility, we analyzed these 444 sequences in the same manner as Figure 3, except that all samples were searched against the Accugenix Bacterial Library (04APR11) (Figure 4). By performing the analysis this way, the effect of missing library entries is effectively abrogated, meaning that any continued variation is due to differences between Accugenix's manual reference method and the MicroSEQ® v 2.2 automated assembly method. In Figure 4, data points lie below the 45° trend line, suggesting that MicroSEQ® v 2.2 automated assembly method leads to decreased 16S rDNA identification performance. Thus, 16S rDNA identification systems that include manually assisted automated assembly, which allows for manual editing of base-caller errors of polymorphic (mixed) base calls, outperforms commercial systems with automated assembly methods.

Conclusion

Accugenix provides a superior method of data analysis, incorporating relevant, up-to-date libraries over the MicroSEQ® ID Analysis Software Version 2.1 (with Bacterial Library v 2.2). The inherent problems with the MicroSEQ® library and the method used to assemble the sequencing runs and determine the distance calculations are revealed when its performance is compared to Accugenix's 16S rDNA identification system. At best, MicroSEQ® can only identify 76.1% of organisms to the species level, due to invalid and missing library entries. MicroSEQ® also lacks the ability to provide sequences of a quality high enough to accurately trend and track relevant organisms, due to uncorrected base-caller errors, truncated consensus sequences and distance measurement calculations.

Utilizing a method that includes manually assisted automated assembly and continuously updated, proprietary libraries will allow accurate trending and tracking of strains, as well as a high confidence in result interpretation.

