AN ASSESSMENT OF POTENTIAL FALSE POSITIVE *E. COLI* PYROPRINTS IN THE CPLOP DATABASE

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Engineering with a specialization in Bioengineering

by

Skyler Alexander Gordon

February 2017

COMMITTEE MEMBERSHIP

TITLE: An Assessment of Potential False Positive *E. coli* Pyroprints in the CPLOP Database

AUTHOR: Skyler Alexander Gordon

DATE SUBMITTED: February 2017

COMMITTEE CHAIR: Trevor Cardinal, Ph.D.
Associate Professor of Biomedical Engineering

COMMITTEE MEMBER: Christopher Kitts, Ph.D.
Professor of Biology

COMMITTEE MEMBER: Dr. Michael Black, Ph.D.
Professor of Biology

COMMITTEE MEMBER: Dr. Jennifer VanderKelen, Ph.D.
Research Associate

ABSTRACT

An Assessment of Potential False Positive *E. coli* Pyroprints in the CPLOP Database

Skyler Alexander Gordon

The genetic information found in each species of organism is unique, and can be used as a tool to differentiate at the molecular level. This has caused rapid genotyping methods to become the cornerstone of a new area of research dependent on reading the genome as a form of identification. One of these specific identification methods, known as pyroprinting, relies on the small variation of DNA sequences within the same species to develop a unique, reproducible fingerprint. By simultaneously pyrosequencing multiple polymorphic loci within the ribosomal operons known as the intergenic transcribed spacers, a reproducible output is obtained, known as a pyroprint, which can be used like a fingerprint to identify that organism. This section of the genome not only differs between species but also between isolated bacteria within that species, allowing for the differentiation of species subtypes, referred to as strains. While this is a viable method for generating reproducible fingerprints from individual strains it may be possible to obtain identical fingerprints from non-identical organisms. The following report uses direct sequence comparison and *in silico* pyrosequencing of *E. coli* isolates housed in the Center for Applications in Biotechnology at California Polytechnic State University, San Luis Obispo that have matching pyroprints to show that it is possible to receive near identical pyroprints from non-identical sequences of intergenic transcribed spacers. Although the exact likelihood and cause of this false positive result remains undetermined due to limitations in the sequencing method, its existence questions the accuracy of using pyroprints of the ITS regions as a method of strain classification.

ACKNOWLEDGMENTS

This project would not have been possible without Dr. Michael Black and Dr. Christopher Kitts. I am deeply grateful for their involvement in this project, and for letting me be a part of it. I would like to thank Colin Adams for being a part of this project, it was a pleasure working with him. I would also like to thank Dr. Jennifer VanderKelen for her guidance, as well as Ellen Krippaehne, Olivia Origel**,** and Ashley Safari for their assistance in the laboratory.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

INTRODUCTION

MICROBIAL SOURCE TRACKING

Microbial source tracking (MST) was originally developed to help determine the source of microbial contaminants, commonly fecal in origin, found in large bodies of water [1]. Fecal contaminants are identified by the presence of either fecal coliforms or molecules specific to feces in a water sample [1], [2]. Once one of these identifiers is located, it can be analyzed to determine the origin. MST methods designed to identify based on species specific fecal chemicals or micro-organisms are known as library independent methods. The library dependent form of MST, commonly known as bacterial strain typing (BST), requires the analysis and cataloging of isolated fecal bacteria based on phenotypic or genotypic criteria. BST determines the specific subtype of a bacterial species, also known as a strain [3]. Because these libraries of bacterial strains are created using the feces of known host organisms as the source of bacteria, the unique collection of bacterial strains that comprise each host species microflora can be used to identify a specific host. This type of categorization allows for unknown bacterial strains to be compared against a database that will return the most likely host species, and therefore the fecal source, of that particular strain.

ITS REGIONS OF THE *E. COLI* GENOME

To create maximum variation in a genomic BST technique, the region of DNA analyzed must be common to all the bacteria being assessed but have slight differences between each strain. The regions of DNA used in this particular experiment are between the 16S, 23S and 5S sections of ribosomal genomic DNA, and are referred to as the intergenic transcribed spacer regions or ITS1 and ITS2 (Figure 1). These ITS regions are important for assisting the development of secondary structure in the ribosome, and ITS1 in particular can often contain sequence code responsible for the formation of certain tRNA [4]. Both are partially conserved regions of DNA found within the highly conserved sections of ribosomal operons. This allows for consistent amplification of the variable ITS regions by using primers directed towards the highly stable sections of the ribosomal operon.

Each version of ribosomal DNA found in a genome is referred to as an allele. The ribosomal DNA sequence is known to occur seven times in the *E. coli* genome, each allele having the possibility of containing a unique set of ITS regions. This increases the sensitivity of using ITS regions as a BST method by utilizing the many different allele ratios that could possibly exist. It is also possible that individual alleles within this ratio may also be over represented based on their distance from the origin of replication found in the genome. If some bacteria happen to be in the process of DNA replication, the alleles that are closer to the origin of replication may be found at higher frequency than what is typically present in the genome. While this would increase the variation of using the ITS regions as a form of bacterial strain typing, it could also increase the possibility of incorrectly cataloging an isolate [5]

PYROSEQUENCING

The DNA sequencing technique of pyrosequencing utilizes small flashes of light, produced in a luciferase reaction, to determine the sequence of a DNA strand. This reaction is driven by the release of inorganic pyrophosphate molecules during the addition of nucleotides by DNA polymerase. After a series of enzyme reactions, each pyrophosphate results in a flash of light being released. The intensity of light depends on the number of nucleotides incorporated to the new DNA strand. By dispensing one type of

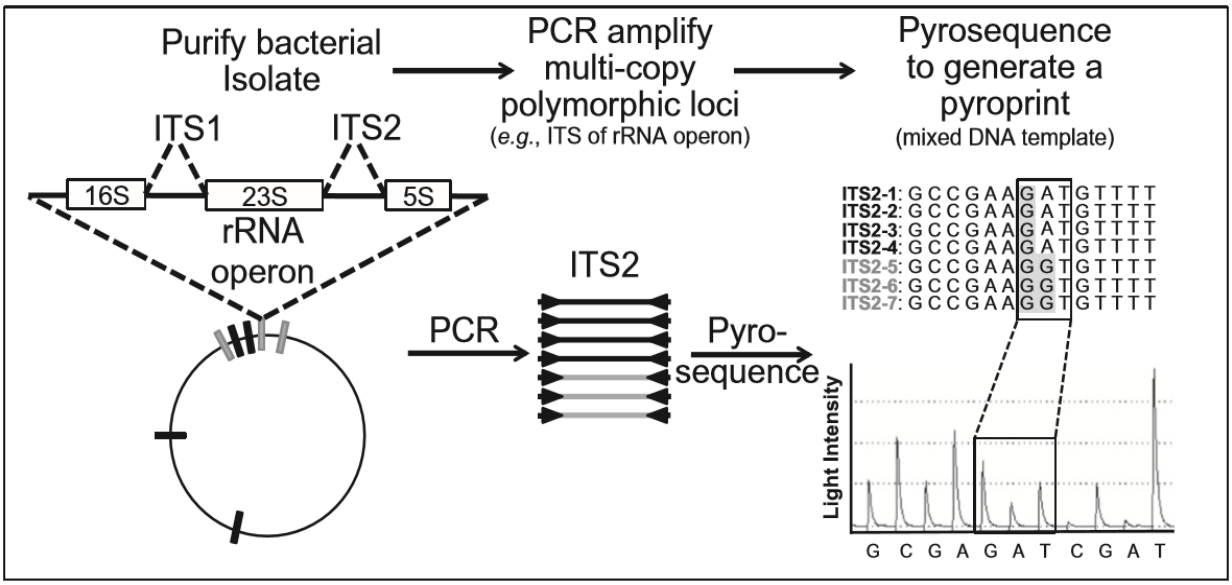


Figure 1: Pyroprinting Overview - The ratio of the seven alleles (seen here as 3:4 in grey and black) in the *E. coli* genome will present characteristic pyroprints based on the differences in sequence and ratios between strains. (Black et al., 2014)

nucleotide at a time (G, A, T, or C), the sequence of a DNA strand can be determined from the intensity

and sequence of the flashes of light. This dispensation sequence is performed by a pyrosequencer, programmed to release each type of nucleotide in a specific order. Once a single type of nucleotide has been incorporated and the flash of light, or lack thereof, has been recorded the enzyme apyrase present in the solution will remove any remaining free nucleotides. The next nucleotide in the program can then be released and the DNA sequence can be analyzed in this sequence-by-synthesis method. The resulting peaks of light can be referred to as a pyrogram. This method is usually applied to a population of DNA strands that are all identical [6]**.**

PYROPRINTING AS A BACTERIAL STRAIN TYPING METHOD

The genetic strain typing technique assessed in this experiment is a method known as pyroprinting [7]. This technique uses the simultaneous pyrosequencing of multiple polymorphic loci to generate a signature unique to a specific bacterial strain (Figure 1). This unique microbial fingerprint can then be cataloged in a database for further use. The pyroprint generated is dependent on the segment of genomic DNA used, and in this experiment the DNA amplified is the ITS regions of ribosomal DNA found in *E. coli*. Each region (ITS1 or ITS2) is analyzed under a unique dispensation sequence of nucleotides, designed *in silico,* to optimize the variability between different strains [8].

The Cal Poly Library of Pyroprints (CPLOP) database defines each strain of *E. coli* based on a combination of the pyroprints from both ITS1 and ITS2 [9]. Each pyroprint taken from a new isolate is compared to all of the existing pyroprints in the database using Pearson correlation. If these pyroprints do not match any previously cataloged entries, the isolate is defined as a new strain and assigned a specific 'cluster' number. Isolates within the same strain, and therefore identical species subtype, are determined to be the same if the pyroprint taken from both ITS regions are ≥99% similar by Pearson correlation [7]. All isolates that exhibit this unique level of matching are cataloged under the same strain type for database organization.

DETERMINING THE DATABASE QUALITY

All matching isolates are given a number and placed into what is known as a cluster in order to catalog them as a specific strain type. Ideally, each strain would be comprised of several isolates taken

from the same host species. Although the pyroprinting process is very sensitive, some *E. coli* in the CPLOP library were harvested from different host species and identified as the same strain. While the possibility exists that the same bacterial strain may occur in several different host species, this result could also be caused by incorrect cataloging of isolates, due to a false positive pyroprint correlation [8]. The source of this false positive result could be ITS alleles that are close in sequence and yet not identical (allowing for less than a 1.0% difference in the pyrograms), or sequences that are not close but several differences, masked by the dispensation of nucleotides in the pyrosequencing program, allow for the pyroprints to come out similarly. The frequency of obtaining a false positive match is currently unknown, making the accuracy of the database impossible to determine. This report evaluates the ITS regions of isolates suspected to present false positive results in an attempt to gain an experimental false positive frequency. The actual genetic sequences from isolates grouped into the same strain type were obtained and analyzed to determine if the alleles of the ITS regions were actually identical or if they were falsely cataloged.

THE ILLUMINA SEQUENCING PLATFORM

To obtain the genomic sequences of the ITS1 and ITS2 regions of the ribosomal operons, an Illumina sequencing method was employed. This is a sequence-by-synthesis process that works in a way similar to pyrosequencing. DNA sequences, typically unknown in nature, have small segments of DNA added to either end. These segments of DNA, referred to as adapters, are known sequences that are designed to allow for amplification and sequencing of the unknown portions of DNA. Adapters are placed onto sequences, first by the addition of an adenosine nucleotide to the 3' end of either DNA strand followed by association with a complimentary thymine nucleotide on the adapter. These two segments of DNA can then be ligated together. Because the adapters are added to both ends of a DNA strand, the Illumina sequencing process occurs from both ends. The final sequences are then combined using either barcode identification on the adapters or by sequence overlaps [10]

PROJECT OVERVIEW

The ITS regions of 97 different isolates, comprising 28 different strains of *E. coli*, were sequenced in an effort to determine if they have been appropriately cataloged. The hypothesis is that this genetic analysis will determine if isolates with matching pyroprints also have ITS regions that match at a genetic level. If isolates have matching pyroprints but non-identical genomic sequences, these will be considered to be falsely cataloged in the database and represent a false positive pyroprint match.

METHODS AND MATERIALS

DATABASE ISOLATE SELECTION

E. coli with matching strain types, and originating from different host organisms, were selected for sequencing from a frozen collection housed in the Center for Applications in Biotechnology at Cal Poly. Isolate selection was focused on groups that contained the maximum variation of host species, as well as on hosts known to commonly interact (i.e. Human and Canine / Human and Bovine, etc.). Isolates taken from the same host were avoided to reduce the possibility of bacterial clones being analyzed [3]. All 97 isolates were assigned a specific four-nucleotide barcode. One isolate was chosen to be amplified twice with different barcodes in order to create a positive control, for sequence matching and analysis.

AMPLIFICATION AND BARCODING OF ITS1 AND ITS2 TEMPLATES

Isolates were removed from a -70˚C freezer and streaked directly onto petri dishes containing lysogeny broth (LB) agarose (Fisher Scientific, Waltham, Massachusetts). Samples were then sealed with parafilm and incubated for 24 hours at 37˚C. Individual bacterial colonies were used in separate PCR reactions for each ITS region, amplified with GoTaq Polymerase master mix in a 25 µL reaction (Table 1) [11]. These reactions contained primers, at a concentration of 0.2 mM, designed to modify ITS1 and ITS2 in preparation to be barcoded and combined into a single product (Table 2). This was achieved by adding identical sequences to the 5’ ends of the forward primers of ITS1 and ITS2 (barcode adapters) and complimentary regions to the 5’ end of the reverse primers (stitch adapters). Each sample was analyzed after PCR using a 1% agarose gel to confirm successful amplification of the desired product. Successful

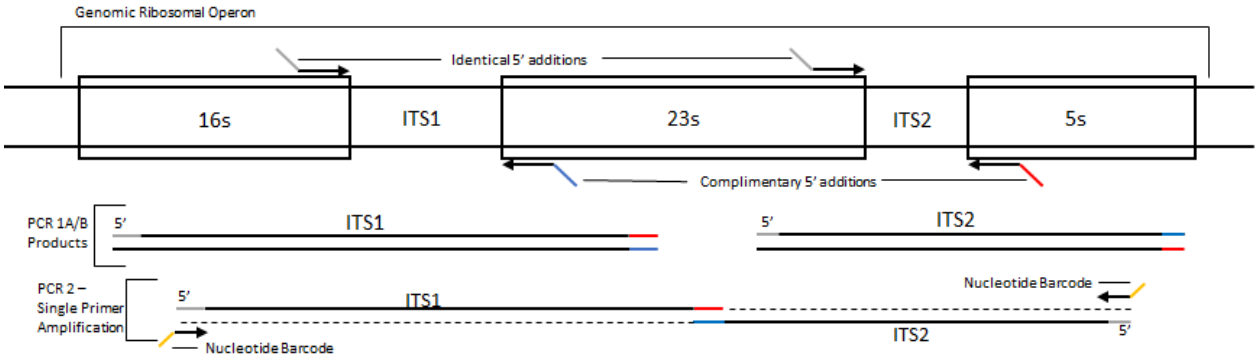


Figure 2: PCR 1 & 2 Overview

reactions were determined by the clear amplification of DNA strands approximately 350-450 basepairs in length for ITS1 and 180-250 basepairs in length for ITS2. Bands as large as 750 or 1000 bp found in ITS1 are suspected to be due to the large number of tRNA insertions that can be found in this region (Figure 3) [4]. Remaining primers were removed using UltraClean PCR Clean-Up Kit (MoBio Laboratories, Carlsbad, California) and each sample quantified using UV spectrometry (IMPLEN pearl nanospectrophotometer, Westlake Village, California). An aliquot of each sample was then diluted to 0.2 ng/µL, using molecular grade water, for a second PCR reaction performed using ~1 ng of each ITS region. This reaction (PCR2) was done to create a single product from the two ITS regions (Figure 2). The complimentary regions added in PCR 1, referred to as stitch adapters, allow ITS1 and ITS2 to bind and create a combination of the two strands as a final product. Single primer amplification was used to amplify the ITS1 and ITS2 (ITS1/2) final products and add an isolate specific barcode by targeting the identical sequences, referred to as barcode adapters, on the 5' end of each product. This process resulted in the two ITS regions being stitched, end-to-end, with isolate specific barcodes at either side of the sequence (Figure 2). The combination of ITS1 and ITS2 into a single stitched product was done to retrieve more data, by utilizing the dual directional sequencing properties of the Illumina system [10]. Phusion polymerase (New England Biolabs, Ipswich, Massachusetts) was used in this step to reduce

Table 1: Thermocycling parameters

Polymerase Chain Reaction 1A	Polymerase Chain Reaction 1B	Polymerase Chain Reaction 2
ITS1:	ITS2:	Linkage:
a) 95°C for 2 minutes	a) 95°C for 2 minutes	a) 95°C for 2 minutes
b) 95°C for 30 seconds	b) 95°C for 30 seconds	b) 95°C for 30 seconds
c) 68°C for 30 seconds	c) 56°C for 30 seconds	c) 62°C for 30 seconds
d) 68°C for 1 minute	d) 68°C for 1 minute	d) 68°C for 1 minute 30 seconds
e) 68°C for 5 minutes	e) 68°C for 5 minutes	e) 68°C for 5 minutes
f) 4°C hold	f) 4°C hold	f) 4°C hold
Steps b-d were repeated for 40 cycles	Steps b-d were repeated for 45 cycles	Steps b-d were repeated for 40 cycles

mutation and ensure proper amplification of the added barcodes [12]. These reactions were performed in 25 µL batches, with a primer concentration of 0.4 mM (Table 2). A sample of each reaction was then run on a 1% agarose gel to check for successful amplification. Proper amplification was recognized by the presence of DNA strands 600-800 bp in length. Left over primers were removed using an UltraClean PCR Clean-Up Kit (MoBio Laboratories, Carlsbad, California**)** and samples were quantified using UV spectrometry (IMPLEN pearl nanospectrophotometer, Westlake Village, California). An aliquot of each product was then diluted to 10 ng/µL using molecular grade water.

CONFIRMATION OF STITCHED PRODUCT SEQUENCE POPULATION

To ensure that final stitched products correctly represented sequence populations of both ITS1 and ITS2, several stitched products were chosen to be pyroprinted and compared to their genomic counterpart. Pyroprints taken from the ITS1/2 stitched products were compared to the pyroprints of ITS1 and ITS2 taken directly from the isolate. This process was done by taking 1 ng of the stitched ITS1/2 product and amplifying the ITS1 and ITS2 sections in two separate, 25 µL PCR reactions. These two reactions contained biotinylated reverse primers, at a concertation of 0.2 mM, and were run under the same thermocycling parameters as PCR 1A and 1B for each respective region. These reactions were then repeated, using a live bacterial colony from each isolate as source DNA. Successful amplification was determined by running samples from all four reactions on a 1% agarose gel. Proper amplification was

Table 2: PCR primers – Original ITS1 and ITS2 primers are located in bold. Primers labeled with ** are biotinylated when pyrosequencing is performed. 'NNNN' represents nucleotide place holders for the barcode unique to each isolate. A full list of these barcodes can be found in the appendix.

Primer Name	Primer Sequence (5' to 3')	Tm (°C)
PCR 1A:		
ITS 1 – F	atggtggtcggctggtggta**GGAACCTGCGGTTGGATCAC**	62.5
ITS 1 – R	**actcagaagtgaaacgccgtag**CTTCATCGCCTCTGACTGCC**	62.5
PCR 1B:		
ITS 2 – F	atggtggtcggctggtggta**ATGAACCGTGAGGCTTAACCTT**	60.3
ITS 2 – R	**ggcagtcagaggcgatgaag**CTACGGCGTTTCACTTCTGAGT**	62.1
PCR 2:		
ITS 1/2 – F/R	tatNNNNtatggtggtcggctggtggta	62.5

determined by the presence of 350-450 bp strands for ITS1 and 180-250 bp strands for ITS2. Once properly amplified, pyrosequences were performed on all four samples and pyrograms were compared using Pearson correlation.

SEQUENCING OF ITS1/2 STITCHED PRODUCTS

Final stitched products were prepared for sequencing by combining 10 µL of each 10 ng/µL ITS1/2 product into a single 1.5 mL micro centrifuge tube and then quantified using UV spectrometry (IMPLEN pearl nanospectrophotometer, Westlake Village, California). After being frozen and packed in dry ice, the mixture of stitched ITS1/2 products was sent out for sequencing on the Illumina platform (MR DNA, Shallowater, TX). This process is designed to give 150 nucleotide base pair reads from either end of the ITS1/2 sequence. Data were returned via the Illumina Basespace file-sharing network.

COMPUTER PROCESSING OF FASTQ SEQUENCES

The sequences retrieved were initially processed through a series of python scripts [13]. Original sequences were returned in FASTQ format, and converted to FASTA for analysis. All sequences were then truncated using an *in silico* pyrosequence to determine the length of an allele based on the dispensation sequence corresponding to each region [7], [8]. This modification allowed for the sequences to be analyzed as only the nucleotides that would contribute to a pyroprint. Sequences that did not have either an identifiable barcode or ITS region, were too short to determine a specific allele, or had a nonspecific nucleotide marker meant to represent the possibility of several different nucleotides (i.e. N, K, M, Y, etc.) were placed in a separate trash file for further analysis. Identical alleles for each strain were then counted and placed into a file, with the exception of sequences that added up to less than 1% of the total count for that strain. These alleles presented a low enough portion of the overall population that they were assumed to be mutations induced by PCR and placed into a separate file. Isolates were then organized into strains in order to compare the sequences originally found to have pyroprints ≥99% similar by Pearson correlation. Because the number of sequences was so large, a distributed framework was needed to process the data [13].

Isolates within a strain were compared based on the truncated ITS alleles present within the genome. This was done by inspecting the direct nucleotide sequences of the alleles found in each isolate, as well as the relative sequencing counts associated with each allele. Each allele was cross referenced to a set of alleles taken from the genomes of 51 *E. coli* isolates found in the NCBI – GenBank database to help visually assess each isolate (Personal communication, Michael Black, 2016).

SYNTHETIC PYROPRINTING

A synthetic pyroprint is a computer generated output of a pyroprint, using a population of DNA sequences and a predetermined dispensation sequence. In theory, this output should be identical to that of the experimental pyroprints if the population of sequences used in both is identical. Because of this, discrepancies between actual and synthetic pyroprints should represent a difference between sequences retrieved from the Illumina sequencing method and those used in obtaining an experimental pyroprint. To increase the accuracy of the synthetic pyroprint for each isolate, the entire population of sequences retrieved were used in this method barring those found to be under 1% of the total sequence count. These synthetic pyrograms were then compared to experimental pyrograms found in the CPLOP database using Pearson correlation.

CONTINUOUS CORRELATION

Several groups of these synthetic pyroprints were used to create a continuous correlation. This was done by taking the Pearson correlations between the synthetic and experimental pyrograms at each step of the dispensation sequence. By graphing how the synthetic pyroprints differ from the experimental pyroprints at the addition of each individual nucleotide in the pyrosequencing process, the pattern seen across the 96 dispensations may give insight to the quality of the sequencing results.

RESULTS AND ANALYSIS

GEL CONFIRMATION OF AMPLIFCAITON

All PCR amplified products were run on a 1% agarose gel, and imaged using the intercalating and UV reactive properties of ethidium bromide (Figure 3). Products were analyzed for base pair length by comparison against a 1kb GeneRuler DNA ladder (Fisher Scientific, Waltham, Massachusetts). Expected band sizes were 200-250 bp for ITS2 (Figure 3 – Far left), 450-500 bp for ITS 1 (Figure 3 – center) and 650 – 750 bp for stitched ITS1/2 (Figure 3 – Far right). Faint bands of DNA can be seen near 400 bp for ITS 2 and 800 bp for ITS1, approximately doubling the successfully amplified band, and are thought to be caused by tRNA insertions in the ITS regions.

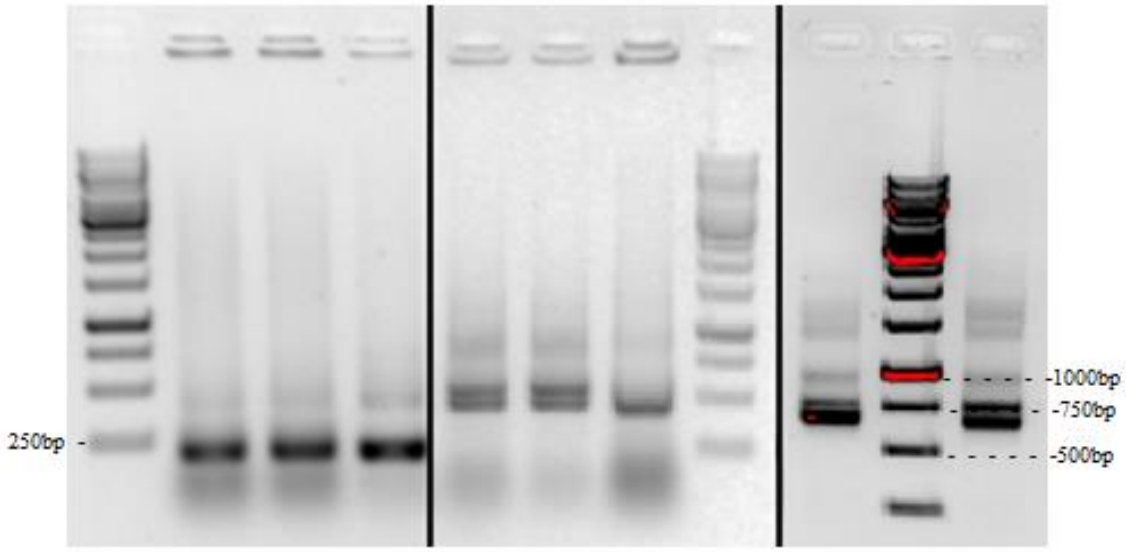


Figure 3: ITS1, ITS2, and ITS1/2 Gel Images – ITS2 is represented on the left, ITS1 in the middle, and stitched ITS1/2 on the right. Images are taken from 1.0% agarose gels run for ~45 minutes at 100 volts. DNA ladder sizes can be found in the appendix.

CONFIRMATION OF STITCHED PRODUCT SEQUENCE POPULATIONS

The pyroprints of ITS1 and ITS2 using stitched products as source DNA were compared to pyroprints of each respective region using bacterial colonies as source DNA, to confirm that the population of sequences amplified in both reactions were identical (Table 3). If differences were to occur in this experiment, it would imply that specific sequences were being selectively amplified when the ITS1 and ITS2 products were being stitched together. The data show that populations of sequences amplified using either genomic DNA or stitched ITS1/2 products as the source are near identical. This check

ensures that the population of ITS regions included in the ITS1/2 stitched product sent out for sequencing accurately represents the population of sequences used to generate experimental pyroprints for those isolates.

Table 3: Stitched Product Pyroprint Comparison – The tables below show the correlation levels between pyroprints taken from stitched source DNA and those taken from bacterial colonies as source DNA. Bacterial sources are labeled by C while stitched products are labeled by L.

	R1C	*R1L*
R1C	1	
R1L	0.995533	1

	R2C	*R2L*
R2C	1	
R2L	0.993074	1

	S1C	*S1L*
S1C	1	
S1L	0.998625	1

	S2C	*S2L*
S2C	1	
S2L	0.992188	1

GENERAL SEQUENCING RESULTS

From the 5,258,600 sequences originally received from the Illumina sequencing process, 2,031,803 sequences were determined to be valid for analysis. Over 2 million sequences were removed from the total 5.25 million due to an inability to identify a barcode or ITS region present within the sequence. Roughly 1.8 million of these removed sequences were completely unidentifiable. Although there is a possibility that these unrecognizable DNA sequences were added to the final mixture in the laboratory before being sent for sequencing, it is highly unlikely that these would be represented at such high count due to the number of PCR amplifications and dilutions performed on the samples prior to being sent out. Even if genomic DNA or contaminations were introduced during laboratory procedures, these small amounts of DNA, because they would not be amplified during PCR, would be overwhelmed by the amount of DNA created in the two separate PCR reactions. It is possible that these removed

Table 4: Invalid Sequence Example – These sequences represent common errors found in the experiment (Missing portions of barcode, mutated barcode, and mutated ITS region primer). Sequences are separated into **Barcode** / Adapter / ITS1 or ITS2 forward primer / ITS1 or ITS2

TOO_SHORT_BARCODE:
ACG/TATGGTGGTCGGCTGGTGGTA/ATGAACCGTGAGGCTTAACCTT/ACAACGCCGAAGCTGTTTTGGC

INVALID_BARCODE:
TAT***GGCC***/TATGGTGGTCGGCTGGTGGTA/GGAACCTGCGGTTGGATCAC/CTCCTTACCTTAAAGAAGC

INVALID_REGION:
TAT***AACT***/TATGGTGGTCGGCTGGTGGTA/GGAACCTGCGG**C**TGGATCAC/CTCCTTACCTTAAAGAAGCGTTCT

sequences were misreads from the sequencing process, but it remains impossible to determine the exact source of this unidentifiable DNA.

Invalid DNA sequences were inspected for the adapter ligation sequence, found just after the barcoding sequence, using BLAST nucleotide analysis (Geer et al., 2010). Using this technique, it was observed that a large number of sequences were labeled as invalid due to degradation on the 5' end. This lack of sequence data on the 5' end caused the removal of isolate barcodes, making sequence data impossible to identify. Another large number of sequences were removed due to the mutation of the barcode or the ITS region, causing these sequences to not be sortable based on these variables (Table 4). This realization calls into question the validity of some sequence data. If the barcodes are susceptible to mutation, valid barcodes could possibly be converted into other known valid barcodes and corrupt the sequence data of those isolates.

Valid sequences were inspected for integer allele ratios based on the total sequence count for each allele found within an isolate (Table 5). All isolates contained a set of sequences at high counts, presumably the true alleles in the genome, as well as a set of sequences at low counts that are most likely attributable to DNA mutation in PCR amplification. These low count alleles, often different from high count alleles by only a single point mutation, are nearly indistinguishable from high count alleles and as such cannot be ruled out as errors introduced by mutation (Table 5). To clarify which alleles were indicative of the true allele ratio, sequences were cross compared against alleles taken from 51 different *E. coli* genomes found in the GenBank database (Personal communication, Michael Black, 2016). Alleles not found in this list and representing low counts of the total sequences were considered more likely to be mutations.

Table 5: An Example of Valid Sequences for a Single Isolate – Point mutations between different alleles are represented by bold lettering, while insertions/deletions are represented by a hyphen. Counts on the right side are the raw counts of each allele from the Illumina sequencing process. The bottom two alleles may be mutations created during PCR but cannot be determined as such with certainty.

Isolate Av-049 Alleles	Count
1 - ACAACGCCGAAG**A**TGTTTTGGCGGATGAGAGAAAGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAA	3229
2 - ACAACGCCGAAG**C**TGTTTTGGCGGATGAGAGAAAGATTTTCAGCCTGATACAGATTAAATCAGAACGCA---	2267
3 - ACAACGCCGAAG**G**TGTTTTGGCGGATGAGAGA-AGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAA	1328
4 - ACAACGCCGAAG**C**TGTTTTGGCGGATGAGAGA-AGATTTTCAGCCTGATACAGATTAAATCAGAACGCA---	1322
5 - ACAACGCCGAAG**A**TGTTTTGGCGGATGAGAGA-AGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAA	331
6 - ACAACGCCGAAG**G**TGTTTTGGCGGATGAGAGAAAGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAA	325

ANALYSIS OF ISOLATE ALLELES BY CLUSTER

Valid data were visually inspected by assigning all previously sequenced alleles a specific number, corresponding to the GenBank list previously metioned, in an attempt to assess the possible allele ratio in each isolate. Most of the isolates contained alleles not found in the list retrieved from GenBank, which often differed from alleles in the list by only a single nucleotide polymorphism (SNP) or an insertion or deletion event (in-del) (Table 5). The level of similarity between these alleles, often combined with indistinguishable allele ratios (non-integer or greater than seven alleles), makes it impossible to determine if the alleles not found in the GenBank list are natural alleles or mutations introduced by PCR. Although alleles present could not, with certainty, be determined to be mutations or true genomic representations, there were obvious categories that certain strains fell into.

Of the 97 isolates sequenced, 54 were observed to have identical alleles present at very similar relative sequence counts to another isolate within the same strain (Table 6). These clusters represent isolates with matching fingerprints that also have matching genetic sequences. This is the expected result, and provides a baseline result for what should be seen.

Table 6: Expected Positive Results – Cluster 179: isolates with similar counts and the same present alleles were considered to have near identical allele ratios present in the genome. These would represent the expected result from correctly cataloged isolates.

Isolate	Alleles	Allele Count as Seen in Isolates
Ct-058	08, 07, 17	9309, 1357, 562
Dg-112	08, 07, 17	9909, 3846, 589
Dg-146	08, 07, 17	9124, 2926, 435
Hu-947	08, 07, 17, M1, M2	7557, 2470, 569, 369, 199

Another 24 isolates were observed to have similar alleles, but at such drastically different sequence counts that it would imply the genomic ITS alleles are different from others within the same cluster. This undesired result, isolates with matching fingerprints but non-identical genomic ITS alleles, presents evidence that isolates have been falsely cataloged in the CPLOP database (Table 7). Some of

these isolates, while having a much lower total sequence count than others within the cluster, still present a completely different potential allele ratio than what is to be expected.

Table 7: Expected Negative Results – Cluster 1460: isolates with non-similar counts and semi-similar alleles present were considered to have non-identical allele ratios present in the genome. This would represent the expected result from incorrectly cataloged isolates (i.e. false positive results.

Isolate	Alleles	Allele Count as Seen in Isolates
Cw-782	23, 02, 09, 24, 13, 29	5056, 3072, 2478, 2121, 283, 230
Hu-657	09, 02, 12, 24, 29	4608, 3135, 547, 520, 155

The other 19 isolates could not be so easily identified. Often, isolates within a cluster contained what appeared to be mutations, with no obvious way of determining the true alleles present within the genome (Table 8). Although these isolates often contained similar sets of alleles and at similar sequence counts, without the ability to remove high count mutant alleles they cannot accurately be determined as matching or not matching the other isolates within the same cluster.

Table 8: Undeterminable Results – Cluster 478 isolates appeared to have near similar counts, with near similar alleles present. High count mutant alleles (labeled as M*) also appeared in these groups regularly, making it impossible to determine the allele ratio present in the genome or if these were correctly or incorrectly cataloged isolates.

Isolate	Alleles	Allele Count as Seen in Isolates
Cw-1210	17, 07, M1, 28, 08, 25, M2	7204, 1611, 852, 737, 284, 263, 166
Cw-1243	17, 07, 28, M3, M1, 08, 25, M2	6341, 2069, 923, 533, 452, 278, 254, 142
Cw-1295	17, 07, 28, M3, M1, 25, 08, M2	8371, 1609, 822, 599, 489, 337, 272, 208
Cw-1460	17, 07, 28, 25, 08	7725, 1441, 648, 347, 248
Ho-025	17, 07, M4, 28, 18, M5, 08	5506, 1829, 1547, 786, 339, 261, 259
Pg-174	17, 07, 28, 08, 25	5472, 1131, 379, 174, 134

In an attempt to resolve the sequences of each isolate into a group of alleles present within the genome, the positive control was inspected in more depth (Table 9). While the alleles in ITS2 seem to present the idea that alleles with very low sequence counts can be removed from the overall inspection, data from ITS1 seems to show that not only can the same mutations be present in two different isolates, but that they can appear at drastically different counts and at times relatively high counts. This analysis of the positive control shows that not only do the sequence results seem to be corrupted by a large amount of

mutations, but that high count mutant alleles cannot be accurately distinguished from alleles found in the genome.

Table 9: Allele Results from the Positive Control – The presence of mutant alleles at drastically different counts in ITS1 of the positive control raises the question of what can be considered a true allele found in the genome and what is a mutant allele introduced by mutation. If these cannot be distinguished from one another, positively cataloged isolates cannot be accurately determined.

Isolate	Alleles by Region	Allele Count as Seen in Isolates
	ITS1	
Hu-1666 (1)	11, 07, 28, 13, 23, 17, 22, M1, M2, 08	3962, 1436, 705, 538, 393, 316, 194, 168, 133, 122
Hu-1666 (2)	11, 07, M1, M2, 28, 17, 13, 23, M3, M4, M5, 22, 08, M5, 21	2329, 1185, 959, 654, 560, 421, 356, 269, 222, 216, 201, 157, 153, 137, 111
	ITS2	
Hu-1666 (1)	29, 23, 34, 09, 02, 38, 24	3325, 3022, 1123, 556, 176, 128, 106
Hu-1666 (2)	29, 23, 34, 09, 02	3268, 2982, 1112, 383, 115

PCR SEQUENCE BIAS AND MUTATION

In the sequencing results, it is hypothesized that a large number of isolates were subjected to PCR sequence bias. Certain alleles appear at much higher counts than what is possible if considering integer allele ratios from the ITS regions, only explainable by these sequences either being over amplified during the replication process or over represented in the genome. While there is a small chance that these phenomena could be due to DNA replication events or plasmids present in the cells, this is unlikely. This selective amplification has made the ability to determine the allele ratio present in the genome impossible.

The high mutation rate in this experiment only accentuates this issue. Because a large number of mutations were seen, and the possibility that even high count alleles could be caused by mutation, determining the allele ratio in the genome of the isolate was not possible. Curiously, many of these mutations occurred in the same location in the ITS sequence. It is suspected that these are some form of preferential mutation that is dependent on the previous portions of sequence. These preferential mutations caused many sequences to differ by only a SNP or in-del, making the difference between high and low count mutant alleles nearly indistinguishable and further blurring the possible allele ratios.

To better understand the full effect of the mutation rate on sequence populations, the combined minimum mutation rate was calculated for PCR 2 and the Illumina sequencing method. A fifteen

nucleotide section of the primer used in PCR 2, located just after the barcode, was analyzed for single nucleotide polymorphisms (SNPs). This region of the sequences was amplified by Phusion polymerase during PCR 2 and then again during the Illumina sequencing process. Of the 5,258,600 total sequences, approximately 3.6%, were found to have a SNP in this specific window. This equates to a mutation likelihood of roughly 0.238% for each nucleotide. It can be assumed that the actual mutation rate is much higher, considering that this analysis does not account for more than one SNP, in-dels, the mutation rate introduced during PCR 1, or the sequences declared invalid that do not contain the 15 nucleotide region being analyzed. Based on these calculations, it can be supposed that more than 25% of all sequences returned are mutated at least once within the first one hundred nucleotides.

SYNTHETIC PYROPRINTS AS DIAGNOSTIC TOOLS

In an attempt to distinguish sequences present in the genome from those possibly introduced by mutation, a synthetic pyroprint was employed. These synthetic pyroprints were performed using all the alleles present in the sequencing results with the total count for each allele. The synthetic pyrogram of each isolate was then compared to the corresponding experimental pyrogram using Pearson correlation in the hope that isolates exhibiting potential high count mutant alleles would have lower correlation levels than those that did not (Figure 4). This was not the case, and in some instances isolates containing what were thought to be high count mutants appeared to match the database pyrograms more closely. Although these isolates appeared to match their database entry more closely, they did not match the other synthetic pyroprints within the same cluster as well, drawing into question the capability of the synthetic pyroprint to properly assess sequence accuracy.

To help determine the viability of synthetic pyroprints as a tool for determining if sequence information returned was indicative of genomic sequence populations, continuous correlations were performed and graphed based on the dispensation order for each region (Figure 5). These continuous correlations allow for the observation of how each isolate deviated from the corresponding database entry for that isolate at each individual dispensation. As the dispensation progresses, large changes in correlation can be attributed to specific locations in the sequence data. The largest changes correspond

directly to portions of alleles that contain large strands of a single repeated nucleotide (Figure 5). These sections of homopolymers are often difficult to detect using pyrosequencing and would present a fundamental difference between synthetic and experimental pyroprint outputs. Because of this, and the lack of sensitivity seen in the synthetic pyroprints to the introduction of high count alleles, it was determined that synthetic pyroprinting was not a valid form of analysis for determining the correct allele ratio present in the genome.

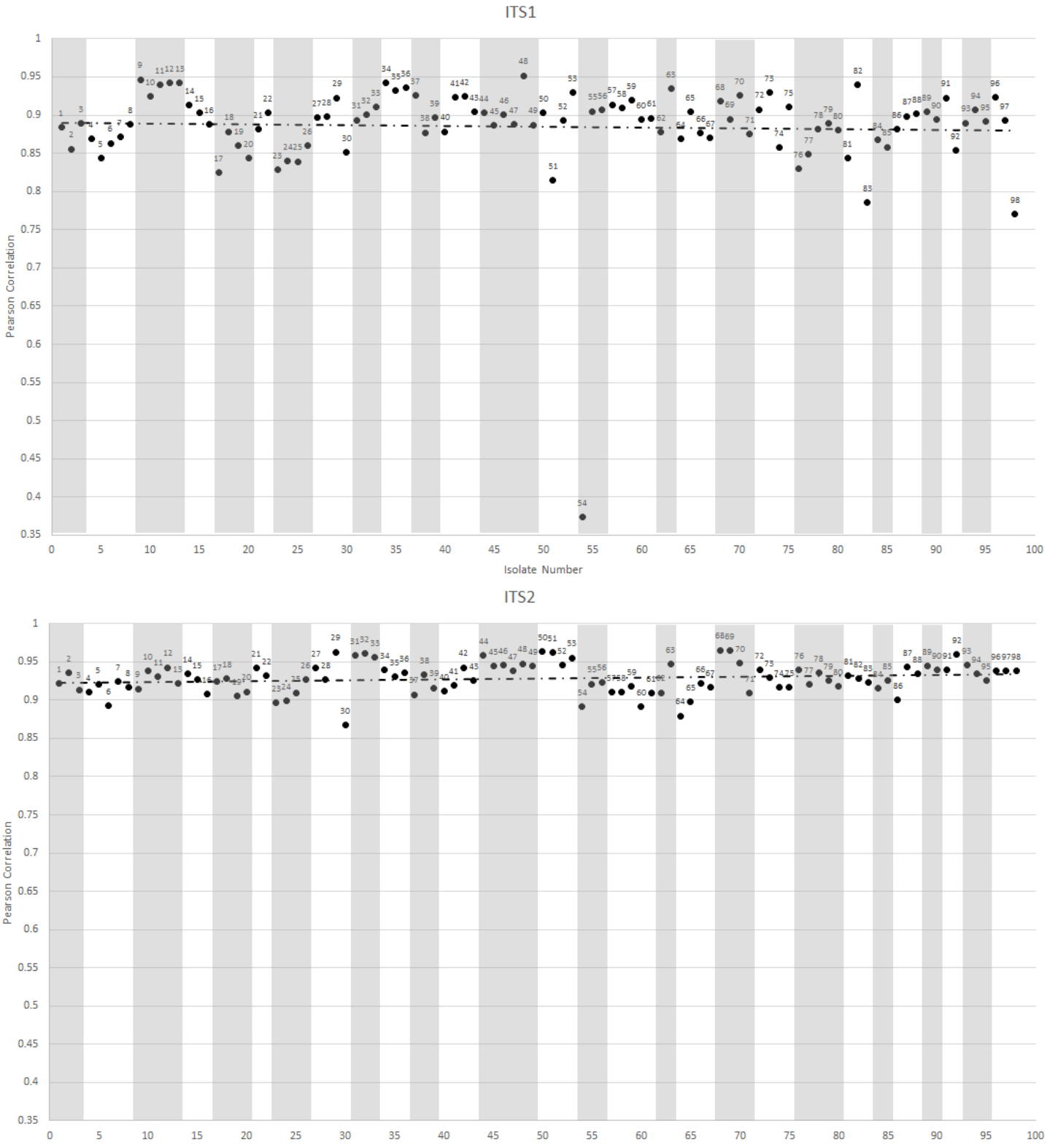
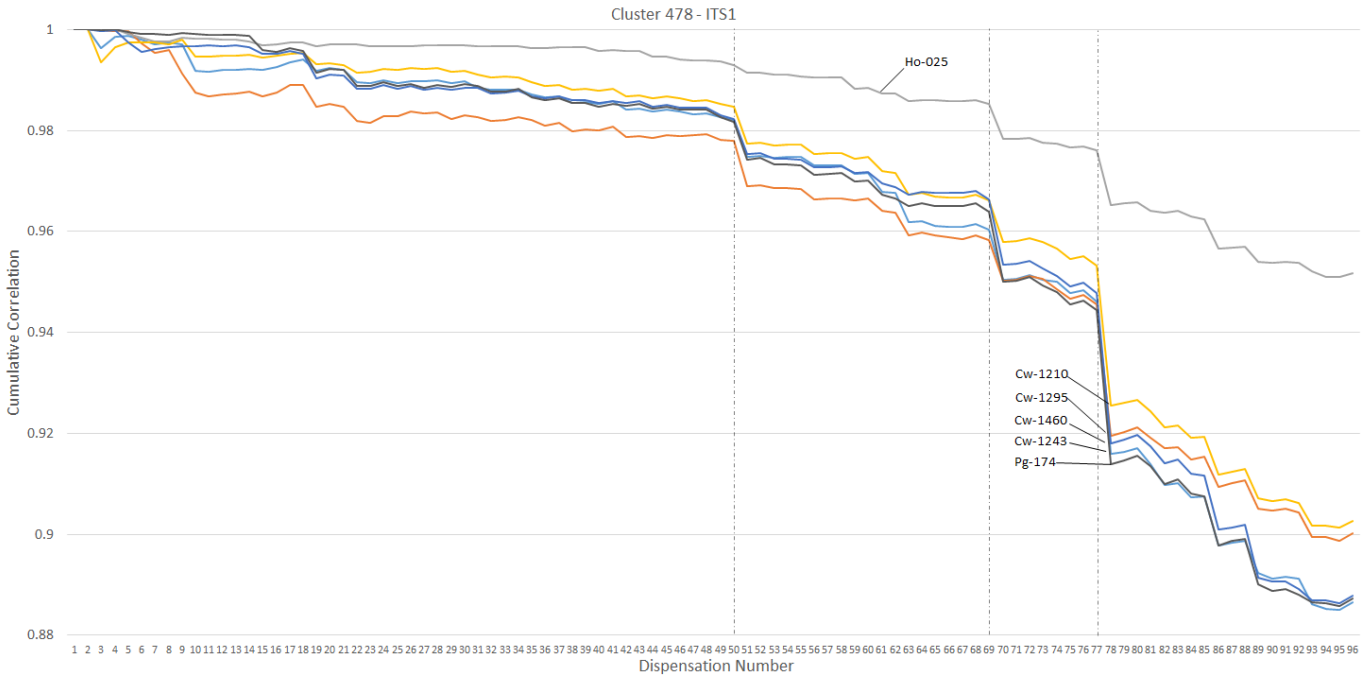


Figure 4: Experimental-Synthetic Pyroprint Correlations – All isolates found within a single strain are separated by the shaded regions on the graph. Isolate names can be found in the appendix, referenced by number.



CTCCTTACCTTAAAGAAGCGTTCTTTGAAGTGCTCACACAGATTGTCTGATAGAAAGTG-**AAAA**-GCAAGGCGT
CTCCTTACCTTAAAGAAGCGTACTTTGCAGTGCTCACACAGATTGTCTGATG-**AAAA**-TGAGCAGTAAAA
CTCCTTACCTTAAAGAAGCGTTCTTTGAAGTGCTCACACAGATTGTCTGATGAAAATGAGC-**A**-GTAAAACCTCT

Figure 5: Continuous Correlation Example – While the correlation pattern is similar between different isolates, the alleles are not. The alleles listed represent the largest population of sequences in every isolate. Bold lettering corresponds to the nucleotides added in dispensation 77, where the largest change in correlation is seen. All large correlation changes coincide with homopolymer regions.

DISCUSSION

The practice of MST was specifically designed to make locating the source of microbial contaminants both easier and cheaper [1]. If the process is very accurate, fewer tests would need to be performed and the testing would become quicker and more cost effective. While library dependent MST methods make determining the source of a microbial contamination easier, because they are so dependent on the development of databases there is a certain investment cost that goes along with establishing each library-based method. This is known before constructing each database, but the *in vitro* false positive frequency associated with each method can only be determined once the database has already been established. Although the accuracy of the database must be assessed in order to determine if it is valid for use, this may not be possible using current methods of sequencing.

To determine the accuracy of the database, a true frequency of finding a false positive must be assessed. While actual false positives were found in the database, due to concerns with mutation rates and PCR bias true allele ratios present in the genome could not be found for all isolates. Without knowing the quality of the sequencing data, there is no way to accurately determine if the isolates found to be unique are truly different from others found within the same strain. For the sake of this experiment, the true alleles in the genome may not be what needs to be analyzed, but instead the true population of sequences used in the pyroprinting process.

Because the pyroprints from stitched ITS1/2 amplicons were the same as those taken directly from colonies of that same isolate, the non-integer allele ratios created during PCR amplification is somehow preserved across multiple rounds of PCR. This result determines that experimental pyroprints may not be using a population of sequences that directly represents the allele ratio present in the genome, and that the non-integer allele ratios present in the sequence data is to be expected.

Considering this, it is also possible that the mutation rate seen in these isolates is also to be expected. Mutations seem to occur at specific, preferential locations in the genetic code and the possibility exists that these are consistently produced during PCR amplification before pyrosequencing. If this is true, the unique alleles thought to be introduced during PCR could be accentuated by the high

number of cycles performed during PCR. The mutation rate of the Illumina process may be accentuating these mutants, but the high correlation rates of the synthetic pyroprints seems to imply that these either do not have a large effect on the final product, or are also included in the experimental pyroprinting process. Because the synthetic pyroprint does not appear to be sensitive to the introduction of novel alleles to a particular isolate within a cluster, it remains to be seen if this is an accurate way of comparing sequence results to the experimental pyroprints found within the database.

CONCLUSION

To determine false positive pyroprints, the sequence data could only be analyzed in two ways. Either by comparing the genomic allele ratios of isolates within a strain, or by comparing isolate sequence populations to the experimental pyroprints in the CPLOP database via synthetic pyroprinting. Due to non-integer allele ratios and a large number of mutations introduced by PCR, the allele ratios of each isolate could not be accurately determined. Isolates thought to be false positives could not definitively be determined as unique when synthetic and experimental pyroprints were compared, due to the lack of sensitivity of *in silico* synthetic pyroprints to the introduction of novel, high count alleles. Although several clear false positives were recognized in the data, for these reasons this report was not able to accurately assess the potential for receiving a false positive strain comparison in the CPLOP database.

REFERENCES

[1] D. Stoeckel and V. Harwood, “Performance, Design, and Analysis in Microbial Source Tracking Studies.,” *Applied and Environmental Microbiology*, Apr-2007. [Online]. Available: http://aem.asm.org/content/73/8/2405.full. [Accessed: 12-Jul-2016].

[2] M. Hajj-Mohaman *et al.*, “Wastewater micropollutants as tracers of sewage contamination: analysis of combined sewer overflow and stream sediments.,” *NCBI*. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25189851.

[3] A. Van Belkum *et al.*, “Guidelines for the validation and application of typing methods for use in bacterial epidemiology,” *European Society of Clinical Microbiology and Infectious Diseases*. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17716294.

[4] Carlos R. Osorio, Matthew D. Collins, Jesus L. Romalde, and Alicia E. Toranzo, “Variation in 16S-23S rRNA Intergenic Spacer Regions in Photobacterium damselae: a Mosaic-Like Structure,” *ASM*, Sep. 2004.

[5] Griffiths, AJF, Miller, JH, and Suzuki, DT, “An Introduction to Genetic Analysis, 7th Edition: Mechanisms of DNA Replication,” 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK21862/.

[6] Gong, M, Foo, SH, Lin, L, Liu, ET, Gharizadeh, B, and Goel, S, “Pyrosequencing enhancement for better detection limit and sequencing homopolymers.,” 15-Sep-2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20833128.

[7] M. Black, J. Vanderkelen, A. Montana, and C. Kitts, “Pyroprinting: A Rapid and Flexible Genotypic Fingerprinting Method for Typing Bacterial Strains.,” *Research Gate*, Oct-2014. [Online]. Available: https://www.researchgate.net/publication/264463301_Pyroprinting_A_Rapid_and_Flexible_Genotypic_Fingerprinting_Method_for_Typing_Bacterial_Strains.

[8] D. Brandt, A. Montana, B. Somers, M. Black, A. Goodman, and C. Kitts, “Pyroprinting Sensitivity Analysis on the GPU.,” *IEEE*, Oct-2012. [Online]. Available: http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1242&context=csse_fac.

[9] K. Webb, J. Soliman, A. Montana, and C. Ricketts, “CPLOP Database,” *Cal Poly Library of Pyroprints*. [Online]. Available: http://cplop.cosam.calpoly.edu/. [Accessed: 10-Jun-2016].

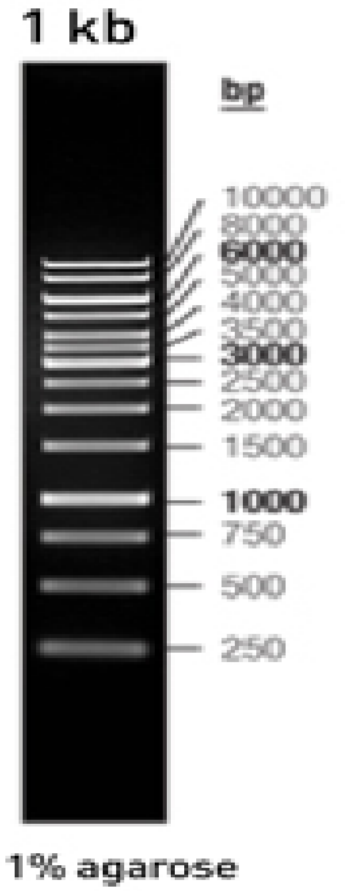
[10] Illumina, “Illumina Sequencing Technology,” *The Dictionary of Genomics, Transcriptomics and Proteomics*, Oct-2010. [Online]. Available: http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf.

[11] Promega, “GoTaq Hot Start Polymerase,” *Promega Corporation*. [Online]. Available: https://www.promega.com/resources/protocols/product-information-sheets/g/gotaq-hot-start-polymerase-protocol/. [Accessed: 25-Jun-2016].

[12] P. McInerney, P. Adams, and M. Hadi, “Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase,” *Molecular Biology International*, May-2014. [Online]. Available: https://www.researchgate.net/publication/265418094_Error_Rate_Comparison_during_Polymerase_Chain_Reaction_by_DNA_Polymerase. [Accessed: 25-Jun-2016].

[13] C. Adams, “Using Hadoop to Identify False Positives in Bacterial Strain Typing from DNA Fingerprints,” Jul. 2016.

APPENDIX



Fisher Scientific 1kb ladder used in gel images

Raw correlation between synthetic and database pyroprints
Database entries are determined by a 'DB' after the strain name

Cluster 9

ITS1

	Av009	*Cw053*	*Pg085*	*Av009DB*	*Cw053DB*	*Pg085DB*
Av009	1					
Cw053	0.99996	1				
Pg085	0.999971	0.999942	1			
Av009DB	0.884095	0.883222	0.882468	1		
Cw053DB	0.856593	0.855656	0.854888	0.997691	1	
Pg085DB	0.890444	0.889576	0.888863	0.998965	0.995041	1

ITS2

	Av009	*Cw053*	*Pg085*	*Av009DB-1*	*Av009DB-2*	*Cw053DB*	*Pg085DB*
Av009	1						
Cw053	0.989132	1					
Pg085	0.981605	0.99825	1				
Av009DB-1	0.921384	0.925368	0.923266	1			
Av009DB-2	0.922087	0.927525	0.925549	0.999225	1		
Cw053DB	0.918323	0.935567	0.936521	0.99361	0.9951	1	
Pg085DB	0.897505	0.913305	0.913528	0.993118	0.994775	0.996762	1

Cluster 19

ITS1

	Av020	*Ct036*	*Cw377*	*Cw963*	*Cw990*	*Av020DB*	*Ct036DB*	*Cw377DB*	*Cw963DB*	*Cw990DB*
Av020	1									
Ct036	0.998172	1								
Cw377	0.99984	0.998933	1							
Cw963	0.999731	0.997984	0.999771	1						
Cw990	0.99851	0.994396	0.998042	0.998947	1					
Av020DB	0.868837	0.860292	0.868545	0.873846	0.877845	1				
Ct036DB	0.852802	0.84419	0.85265	0.85833	0.862459	0.998639	1			
Cw377DB	0.86316	0.855692	0.86321	0.868429	0.8717	0.999026	0.998485	1		
Cw963DB	0.865897	0.857073	0.865564	0.871151	0.875211	0.995872	0.995751	0.996922	1	
Cw990DB	0.878866	0.87037	0.878673	0.884059	0.888103	0.999055	0.997514	0.998787	0.996253	1

ITS2

	Av020	*Ct036*	*Cw377*	*Cw963*	*Cw990*	*Av020DB*	*Ct036DB*	*Cw377DB*	*Cw963DB*	*Cw990DB*
Av020	1									
Ct036	0.999462	1								
Cw377	0.999559	0.999765	1							
Cw963	0.999878	0.999157	0.999499	1						
Cw990	0.999785	0.99917	0.999594	0.999974	1					
Av020DB	0.910401	0.91353	0.912226	0.911317	0.911896	1				
Ct036DB	0.916117	0.920044	0.918814	0.916682	0.917356	0.997959	1			
Cw377DB	0.889586	0.893982	0.892234	0.890276	0.890966	0.997136	0.995241	1		
Cw963DB	0.923469	0.926294	0.925057	0.924191	0.924681	0.998108	0.997581	0.993922	1	
Cw990DB	0.91574	0.918511	0.91732	0.916588	0.917105	0.997798	0.996517	0.994851	0.999443	1

Cluster 48

ITS1

	Av049	*Cw626*	*Cw780*	*Cw893*	*Pg065*	*Av049DB*	*Cw626DB*	*Cw780DB*	*Cw893DB*	*Pg065DB*
Av049	1									
Cw626	0.999788	1								
Cw780	0.999493	0.999677	1							
Cw893	0.998961	0.998464	0.997567	1						
Pg065	0.998345	0.997933	0.996611	0.999668	1					
Av049DB	0.946657	0.946477	0.943225	0.946589	0.949013	1				
Cw626DB	0.924971	0.925006	0.920998	0.923861	0.927535	0.993749	1			
Cw780DB	0.94393	0.943674	0.939904	0.943911	0.946881	0.997358	0.997007	1		
Cw893DB	0.942338	0.94238	0.938422	0.942655	0.945979	0.997158	0.996743	0.998906	1	
Pg065DB	0.939712	0.93943	0.93616	0.939619	0.942115	0.998987	0.994022	0.996931	0.996451	1

ITS2

	Av049	*Cw626*	*Cw780*	*Cw893*	*Pg065*	*Av049DB*	*Cw626DB*	*Cw780DB*	*Cw893DB*	*Pg065DB*
Av049	1									
Cw626	0.998679	1								
Cw780	0.999459	0.999825	1							
Cw893	0.999762	0.998907	0.999547	1						
Pg065	0.997989	0.993662	0.995547	0.997019	1					
Av049DB	0.914771	0.914239	0.914764	0.915008	0.912364	1				
Cw626DB	0.938531	0.93781	0.938405	0.938547	0.93646	0.996522	1			
Cw780DB	0.931365	0.930042	0.930855	0.93132	0.930007	0.997659	0.99786	1		
Cw893DB	0.942502	0.941908	0.94242	0.942547	0.939683	0.991787	0.997114	0.995478	1	
Pg065DB	0.925715	0.926347	0.926415	0.925667	0.922154	0.998041	0.997214	0.997076	0.992501	1

Cluster 107

ITS1

	Ck022	*Cw1043*	*Pg063*	*Ck022DB*	*Cw1043DB*	*Pg063DB*
Ck022	1					
Cw1043	0.994111	1				
Pg063	0.989126	0.982823	1			
Ck022DB	0.913428	0.928703	0.903366	1		
Cw1043DB	0.880538	0.903376	0.869087	0.995358	1	
Pg063DB	0.887298	0.903831	0.888083	0.994451	0.992919	1

ITS2

	Ck022	*Cw1043*	*Pg063*	*Ck022DB*	*Cw1043DB*	*Pg063DB*
Ck022	1					
Cw1043	0.99569	1				
Pg063	0.997235	0.996104	1			
Ck022DB	0.934343	0.929205	0.929262	1		
Cw1043DB	0.924133	0.926367	0.920987	0.995906	1	
Pg063DB	0.908405	0.905097	0.907797	0.994987	0.993875	1

Cluster 120

ITS1

	Ck035	*Cw099*	*Hu615*	*Ck036*	*Ck035DB*	*Cw099DB*	*Hu615DB*	*Ck036DB*
Ck035	1							
Cw099	0.993717	1						
Hu615	0.971236	0.989273	1					
Ck036	0.995238	0.984986	0.953347	1				
Ck035DB	0.824965	0.850004	0.845566	0.833874	1			
Cw099DB	0.854835	0.878453	0.873696	0.862352	0.995587	1		
Hu615DB	0.842024	0.863671	0.860222	0.847164	0.99497	0.994962	1	
Ck036DB	0.835198	0.859429	0.854565	0.843995	0.998294	0.998115	0.99637	1

ITS2

	Ck035	*Cw099*	*Hu615*	*Ck036*	*Ck035DB*	*Cw099DB*	*Hu615DB-1*	*Hu615DB-2*
Ck035	1							
Cw099	0.996641	1						
Hu615	0.993886	0.999141	1					
Ck036	0.957374	0.963297	0.962857	1				
Ck035DB	0.923977	0.921504	0.916168	0.912512	1			
Cw099DB	0.925162	0.927607	0.923021	0.91314	0.993387	1		
Hu615DB-1	0.892131	0.894057	0.889961	0.890288	0.988814	0.988336	1	
Hu615DB-2	0.906983	0.909633	0.905511	0.899054	0.990103	0.997108	0.9894	1
Ck036DB	0.923087	0.920769	0.915345	0.911143	0.999659	0.994515	0.989216	0.991378

Cluster 153

ITS1

	Ct028	*Hu074*	*Ct028DB-1*	*Ct028DB-2*	*Hu074DB*
Ct028	1				
Hu074	0.995816	1			
Ct028DB-1	0.881372	0.891288	1		
Ct028DB-2	0.865724	0.872131	0.988931	1	
Hu074DB	0.892103	0.902701	0.994036	0.985202	1

ITS2

	Ct028	*Hu074*	*Ct028DB-1*	*Ct028DB-2*	*Hu074DB*
Ct028	1				
Hu074	0.997938	1			
Ct028DB-1	0.941727	0.933723	1		
Ct028DB-2	0.910382	0.900713	0.994096	1	
Hu074DB	0.939387	0.931804	0.996386	0.993106	1

Cluster 179

ITS1

	Ct058	*Dg112*	*Dg146*	*Hu947*	*Ct058DB*	*Dg112DB*	*Dg146DB*	*Hu947DB*
Ct058	1							
Dg112	0.992408	1						
Dg146	0.996217	0.999259	1					
Hu947	0.995504	0.997566	0.998468	1				
Ct058DB	0.828533	0.842662	0.838748	0.838861	1			
Dg112DB	0.82581	0.839754	0.835893	0.835946	0.998823	1		
Dg146DB	0.828781	0.84333	0.839168	0.839433	0.999217	0.997905	1	
Hu947DB	0.850545	0.864018	0.860466	0.86045	0.997106	0.995892	0.997611	1

ITS2

	Ct058	*Dg112*	*Dg146*	*Hu947*	*Ct058DB*	*Dg112DB*	*Dg146DB*	*Hu947DB*
Ct058	1							
Dg112	0.999943	1						
Dg146	0.999856	0.99998	1					
Hu947	0.999756	0.999793	0.999758	1				
Ct058DB	0.896185	0.894812	0.893966	0.896742	1			
Dg112DB	0.900633	0.899414	0.898661	0.901262	0.998734	1		
Dg146DB	0.911561	0.91027	0.909476	0.911963	0.997062	0.996749	1	
Hu947DB	0.925836	0.924748	0.924076	0.926544	0.996018	0.995995	0.997068	1

Cluster 247

ITS1

	Cw1002	Cw999	Sp079	Hu143	Cw1002DB	Cw999DB	Sp079DB-1	Sp079DB-2	Sp079DB-3	Hu143DB-1	Hu143DB-2
Cw1002	1										
Cw999	0.999693	1									
Sp079	0.982059	0.982623	1								
Hu143	0.920618	0.924772	0.940742	1							
Cw1002DB	0.896998	0.901791	0.877821	0.842402	1						
Cw999DB	0.892796	0.897919	0.872024	0.839158	0.999604	1					
Sp079DB-1	0.899324	0.904786	0.890025	0.856249	0.99603	0.995407	1				
Sp079DB-2	0.917315	0.92286	0.910207	0.878295	0.992373	0.991461	0.998108	1			
Sp079DB-3	0.92727	0.932223	0.922563	0.886795	0.9895	0.988183	0.995664	0.998502	1		
Hu143DB-1	0.902476	0.907074	0.888106	0.851213	0.996717	0.995873	0.996724	0.994192	0.993094	1	
Hu143DB-2	0.894069	0.899139	0.875519	0.84195	0.998963	0.998842	0.996572	0.99311	0.99071	0.998146	1
Hu143DB-3	0.904587	0.909273	0.888544	0.851337	0.996834	0.996089	0.997389	0.995162	0.99363	0.999379	0.998138

ITS2

	Cw1002	*Cw999*	*Sp079*	*Hu143*	*Cw1002DB*	*Cw999DB*	*Sp079DB-1*	*Sp079DB-2*	*Hu143DB-1*	*Hu143DB-2*	*Hu143DB-3*	*Hu143DB-4*	*Hu143DB-5*
Cw1002	1												
Cw999	0.999812	1											
Sp079	0.994054	0.991989	1										
Hu143	0.919255	0.911508	0.953003	1									
Cw1002DB	0.941613	0.942306	0.931466	0.850467	1								
Cw999DB	0.926492	0.927077	0.916884	0.839749	0.998204	1							
Sp079DB-1	0.928407	0.927392	0.929468	0.875367	0.991435	0.99215	1						
Sp079DB-2	0.965168	0.964825	0.96296	0.894105	0.984166	0.980299	0.985552	1					
Hu143DB-1	0.941166	0.941709	0.931655	0.853115	0.999242	0.997901	0.992725	0.985428	1				
Hu143DB-2	0.952019	0.952342	0.943791	0.867058	0.99809	0.994475	0.991594	0.987867	0.998761	1			
Hu143DB-3	0.949807	0.950347	0.940513	0.860716	0.998431	0.995094	0.991177	0.987555	0.99904	0.999597	1		
Hu143DB-4	0.947381	0.947972	0.93778	0.857283	0.998408	0.995584	0.990994	0.987515	0.999101	0.99919	0.999719	1	
Hu143DB-5	0.948074	0.948553	0.938923	0.86008	0.997514	0.994225	0.990072	0.987344	0.998381	0.998878	0.999464	0.999631	1
Hu143DB-6	0.947176	0.947597	0.938551	0.860856	0.998655	0.996002	0.992169	0.98784	0.999177	0.999233	0.999583	0.999601	0.99909

Cluster 248

ITS1

	Cw1001	*Hu053*	*Sp063*	*Cw1001DB-1*	*Cw1001DB-2*	*Hu053DB-1*	*Hu053DB-2*	*Sp063DB-1*	*Sp063DB-2*
Cw1001	1								
Hu053	0.962659	1							
Sp063	0.950292	0.997435	1						
Cw1001DB-1	0.882386	0.849034	0.845711	1					
Cw1001DB-2	0.893628	0.864777	0.861768	0.99824	1				
Hu053DB-1	0.882593	0.863905	0.859522	0.99307	0.993524	1			
Hu053DB-2	0.908434	0.900282	0.896329	0.987468	0.991457	0.995619	1		
Sp063DB-1	0.877084	0.855284	0.852847	0.995153	0.996397	0.996557	0.992104	1	
Sp063DB-2	0.916765	0.911487	0.910763	0.981593	0.98849	0.986214	0.994457	0.988959	1

ITS2

	Cw1001	*Hu053*	*Sp063*	*Cw1001DB-1*	*Cw1001DB-2*	*Hu053DB-1*	*Hu053DB-2*	*Hu053DB-3*	*Hu053DB-4*	*Sp063DB-1*
Cw1001	1									
Hu053	0.866153	1								
Sp063	0.885016	0.997354	1							
Cw1001DB-1	0.788992	0.93361	0.929697	1						
Cw1001DB-2	0.958801	0.88459	0.900985	0.887736	1					
Hu053DB-1	0.761965	0.923261	0.915241	0.994185	0.869576	1				
Hu053DB-2	0.746911	0.907795	0.899783	0.992779	0.86388	0.998635	1			
Hu053DB-3	0.774234	0.923684	0.917533	0.995533	0.882496	0.998259	0.998133	1		
Hu053DB-4	0.831523	0.961422	0.958935	0.984028	0.916575	0.981558	0.97746	0.984852	1	
Sp063DB-1	0.798713	0.938093	0.935102	0.997703	0.895929	0.994475	0.992739	0.996863	0.988308	1
Sp063DB-2	0.82669	0.957592	0.956342	0.986684	0.914491	0.982986	0.979521	0.98618	0.998935	0.990602

Cluster 369

ITS1

	Cw1111	*Cw1132*	*Hu095*	*Cw1111DB*	*Cw1132DB*	*Hu095DB*
Cw1111	1					
Cw1132	0.999241	1				
Hu095	0.960455	0.954294	1			
Cw1111DB	0.942151	0.938065	0.919752	1		
Cw1132DB	0.936207	0.932098	0.916292	0.999364	1	
Hu095DB	0.932973	0.928576	0.935494	0.992889	0.993986	1

ITS2

	Cw1111	*Cw1132*	*Hu095*	*Cw1111DB*	*Cw1132DB*	*Hu095DB*
Cw1111	1					
Cw1132	0.999683	1				
Hu095	0.99884	0.997948	1			
Cw1111DB	0.939302	0.938222	0.937737	1		
Cw1132DB	0.931434	0.930742	0.929654	0.998128	1	
Hu095DB	0.9344	0.933044	0.935597	0.996941	0.995358	1

Cluster 413

ITS1

	Cw1151	*Pg052*	*Rt071*	*Cw1151DB*	*Pg052DB*	*Rt071DB-1*	*Rt071DB-2*
Cw1151	1						
Pg052	0.993719	1					
Rt071	0.999597	0.994976	1				
Cw1151DB	0.925321	0.900669	0.919379	1			
Pg052DB	0.898865	0.877143	0.892407	0.993093	1		
Rt071DB-1	0.903194	0.876828	0.89658	0.997356	0.995789	1	
Rt071DB-2	0.901659	0.874674	0.895856	0.996205	0.992213	0.997681	1

ITS2

	Cw1151	*Pg052*	*Rt071*	*Cw1151DB*	*Pg052DB-1*	*Pg052DB-2*	*Pg052DB-3*	*Rt071DB*
Cw1151	1							
Pg052	0.997617	1						
Rt071	0.999155	0.999034	1					
Cw1151DB	0.907198	0.907503	0.907636	1				
Pg052DB-1	0.903659	0.905594	0.90408	0.995909	1			
Pg052DB-2	0.906107	0.910349	0.90806	0.997001	0.997606	1		
Pg052DB-3	0.929994	0.93346	0.931321	0.994692	0.994163	0.996848	1	
Rt071DB	0.914227	0.915334	0.916161	0.995471	0.993585	0.994721	0.993412	1

Cluster 421

ITS1

	Cw1159	*Cw140*	*Pg010*	*Pg033*	*Cw1159DB*	*Cw140DB*	*Pg010DB-1*	*Pg010DB-2*	*Pg010DB-3*	*Pg010DB-4*	*Pg010DB-5*	*Pg010DB-6*
Cw1159	1											
Cw140	0.997587	1										
Pg010	0.990426	0.992035	1									
Pg033	0.992297	0.992734	0.999592	1								
Cw1159DB	0.878408	0.889055	0.868046	0.870972	1							
Cw140DB	0.913312	0.923177	0.902701	0.905157	0.993864	1						
Pg010DB-1	0.905118	0.916048	0.904647	0.905957	0.993434	0.994672	1					
Pg010DB-2	0.889658	0.901044	0.888104	0.889551	0.995554	0.994079	0.998911	1				
Pg010DB-3	0.896351	0.907117	0.89432	0.895905	0.995333	0.995539	0.999024	0.999526	1			
Pg010DB-4	0.878358	0.890078	0.876009	0.877511	0.996257	0.991787	0.997415	0.999353	0.998423	1		
Pg010DB-5	0.918447	0.928643	0.918984	0.9201	0.989236	0.994776	0.997965	0.996072	0.997164	0.993236	1	
Pg010DB-6	0.910564	0.920939	0.909947	0.911248	0.992137	0.995613	0.999247	0.998283	0.999006	0.996157	0.999018	1
Pg010DB-7	0.900085	0.910838	0.899104	0.900431	0.99406	0.994199	0.999336	0.999346	0.999378	0.998174	0.997668	0.999204
Pg010DB-8	0.896085	0.907198	0.895216	0.896505	0.994411	0.994163	0.999284	0.999548	0.999348	0.998524	0.997276	0.999065
Pg010DB-9	0.892595	0.903823	0.891567	0.892892	0.994603	0.992796	0.998839	0.999179	0.998755	0.998707	0.996047	0.997932
Pg010DB-10	0.88305	0.894094	0.881094	0.882555	0.995242	0.992369	0.997395	0.999127	0.998828	0.9992	0.994478	0.997013
Pg010DB-11	0.874563	0.886064	0.872224	0.873734	0.995743	0.990621	0.996414	0.998715	0.997953	0.999374	0.992507	0.995433
Pg010DB-12	0.870584	0.882553	0.868876	0.870249	0.995133	0.989286	0.995807	0.998293	0.997005	0.999284	0.991253	0.9944
Pg010DB-13	0.916684	0.926559	0.917019	0.918461	0.974539	0.982473	0.985038	0.981892	0.982515	0.978164	0.987175	0.985418
Pg010DB-14	0.92495	0.934137	0.924039	0.925527	0.987667	0.995111	0.996464	0.994067	0.995853	0.99049	0.998652	0.998216
Pg010DB-15	0.90434	0.915185	0.90442	0.905785	0.978626	0.982301	0.98703	0.984777	0.98456	0.982361	0.986754	0.986094
Pg010DB-16	0.8971	0.90782	0.89533	0.896854	0.995218	0.994798	0.99889	0.998881	0.998988	0.997932	0.997276	0.998621
Pg010DB-17	0.783519	0.798215	0.77699	0.778919	0.95246	0.937283	0.945779	0.952582	0.947919	0.95732	0.935628	0.942384
Pg033DB	0.907491	0.917214	0.90481	0.906647	0.992646	0.99655	0.997323	0.996966	0.998451	0.994692	0.997934	0.998542

	Pg010DB-7	*Pg010DB-8*	*Pg010DB-9*	*Pg010DB-10*	*Pg010DB-11*	*Pg010DB-12*	*Pg010DB-13*	*Pg010DB-14*	*Pg010DB-15*	*Pg010DB-16*	*Pg010DB-17*
Cw1159											
Cw140											
Pg010											
Pg033											
Cw1159DB											
Cw140DB											
Pg010DB-1											
Pg010DB-2											
Pg010DB-3											
Pg010DB-4											
Pg010DB-5											
Pg010DB-6											
Pg010DB-7	1										
Pg010DB-8	0.999682	1									
Pg010DB-9	0.999289	0.99925	1								
Pg010DB-10	0.998576	0.998767	0.999081	1							
Pg010DB-11	0.997762	0.998057	0.998779	0.99955	1						
Pg010DB-12	0.997149	0.99748	0.9984	0.999157	0.999513	1					
Pg010DB-13	0.984003	0.983585	0.983116	0.978473	0.9776	0.975363	1				
Pg010DB-14	0.995842	0.995602	0.994193	0.992661	0.990297	0.988357	0.986275	1			
Pg010DB-15	0.986164	0.985894	0.986156	0.981699	0.981842	0.979973	0.998703	0.984742	1		
Pg010DB-16	0.999097	0.999241	0.999112	0.998596	0.997907	0.997187	0.983928	0.996176	0.985949	1	
Pg010DB-17	0.94944	0.949791	0.953959	0.955852	0.958209	0.962277	0.92001	0.930877	0.927521	0.948676	1
Pg033DB	0.997469	0.997391	0.996429	0.99625	0.994807	0.992909	0.98433	0.998174	0.9843	0.997885	0.939414

ITS2

	Cw1159	*Cw140*	*Pg010*	*Pg033*	*Cw1159DB*	*Cw140DB*	*Pg010DB-1*	*Pg010DB-2*	*Pg010DB-3*	*Pg010DB-4*	*Pg010DB-5*	*Pg010DB-6*
Cw1159	1											
Cw140	0.996478	1										
Pg010	0.998095	0.998558	1									
Pg033	0.997442	0.998687	0.999946	1								
Cw1159DB	0.911503	0.906546	0.910932	0.910372	1							
Cw140DB	0.918949	0.919154	0.920117	0.919971	0.9959	1						
Pg010DB-1	0.896569	0.892116	0.896803	0.896331	0.997382	0.994114	1					
Pg010DB-2	0.908546	0.903733	0.908353	0.90781	0.997311	0.994806	0.998808	1				
Pg010DB-3	0.911094	0.906364	0.911	0.910458	0.997403	0.995136	0.998349	0.999782	1			
Pg010DB-4	0.929129	0.92492	0.928988	0.928453	0.99632	0.9966	0.995135	0.997816	0.998269	1		
Pg010DB-5	0.925759	0.921492	0.925859	0.925359	0.996146	0.995924	0.99513	0.997876	0.998487	0.999397	1	
Pg010DB-6	0.932851	0.928342	0.932541	0.931986	0.994458	0.994504	0.992707	0.996675	0.997419	0.999221	0.999262	1
Pg010DB-7	0.925923	0.921671	0.92597	0.925478	0.997384	0.996493	0.995477	0.997759	0.998363	0.999271	0.999272	0.99878
Pg010DB-8	0.943191	0.93846	0.942575	0.941963	0.992065	0.992552	0.988145	0.992745	0.994023	0.997716	0.997578	0.99882
Pg010DB-9	0.8714	0.866614	0.871305	0.870786	0.991767	0.987224	0.997462	0.995026	0.993899	0.987652	0.988043	0.984777
Pg010DB-10	0.893453	0.88894	0.893828	0.893366	0.997087	0.99338	0.999439	0.99768	0.9973	0.993886	0.993796	0.991217
Pg010DB-11	0.896763	0.892085	0.896865	0.896365	0.9973	0.993961	0.999544	0.998132	0.997712	0.994778	0.994515	0.992166
Pg010DB-12	0.908679	0.904035	0.908812	0.908312	0.998102	0.995429	0.999108	0.9989	0.998928	0.997209	0.997036	0.99539
Pg010DB-13	0.916369	0.912447	0.916751	0.916306	0.997999	0.996696	0.997955	0.998768	0.998973	0.998656	0.998318	0.997294
Pg010DB-14	0.933134	0.929659	0.933607	0.933173	0.990084	0.990414	0.988133	0.992275	0.993259	0.995178	0.99451	0.994914
Pg010DB-15	0.934474	0.930802	0.93453	0.934054	0.991675	0.991952	0.990619	0.995471	0.996015	0.997707	0.997184	0.998579
Pg010DB-16	0.935391	0.931698	0.935644	0.935185	0.987262	0.986851	0.984434	0.989064	0.990237	0.992287	0.99159	0.992351
Pg010DB-17	0.938105	0.934183	0.938012	0.9375	0.992128	0.992673	0.990459	0.995221	0.995887	0.998253	0.997617	0.998939
Pg033DB	0.925364	0.921037	0.92541	0.924899	0.996796	0.99613	0.995072	0.997698	0.998424	0.999358	0.999373	0.999236

	Pg010DB-7	*Pg010DB-8*	*Pg010DB-9*	*Pg010DB-10*	*Pg010DB-11*	*Pg010DB-12*	*Pg010DB-13*	*Pg010DB-14*	*Pg010DB-15*	*Pg010DB-16*	*Pg010DB-17*
Cw1159											
Cw140											
Pg010											
Pg033											
Cw1159DB											
Cw140DB											
Pg010DB-1											
Pg010DB-2											
Pg010DB-3											
Pg010DB-4											
Pg010DB-5											
Pg010DB-6											
Pg010DB-7	1										
Pg010DB-8	0.997681	1									
Pg010DB-9	0.987772	0.977752	1								
Pg010DB-10	0.994681	0.987054	0.997268	1							
Pg010DB-11	0.995323	0.988164	0.997121	0.999876	1						
Pg010DB-12	0.997806	0.992446	0.994511	0.999015	0.999245	1					
Pg010DB-13	0.998922	0.99508	0.992225	0.997772	0.998215	0.999442	1				
Pg010DB-14	0.994061	0.993406	0.978548	0.986107	0.986859	0.990637	0.992398	1			
Pg010DB-15	0.996625	0.996979	0.983153	0.989	0.990061	0.993246	0.995477	0.995567	1		
Pg010DB-16	0.991676	0.99158	0.973696	0.982508	0.98322	0.987612	0.989468	0.999314	0.993422	1	
Pg010DB-17	0.997194	0.997872	0.981971	0.988847	0.989965	0.993458	0.99581	0.996039	0.999769	0.993993	1
Pg033DB	0.999526	0.997942	0.987777	0.994232	0.994889	0.997429	0.998771	0.994272	0.997288	0.99149	0.997699

Cluster 478

ITS1

	Cw1210	*Cw1243*	*Cw1295*	*Cw1460*	*Ho025*	*Pg174*	*Cw1210DB*	*Cw1243DB*	*Cw1295DB*	*Cw1460DB*	*Ho025DB*	*Pg174DB*
Cw1210	1											
Cw1243	0.996916	1										
Cw1295	0.999271	0.994908	1									
Cw1460	0.994173	0.987597	0.992939	1								
Ho025	0.971436	0.974361	0.967816	0.973884	1							
Pg174	0.994147	0.988702	0.992319	0.999684	0.97345	1						
Cw1210DB	0.902621	0.910152	0.89817	0.908338	0.951271	0.907429	1					
Cw1243DB	0.875888	0.886633	0.870302	0.88032	0.935047	0.879533	0.996971	1				
Cw1295DB	0.905601	0.914842	0.900149	0.909952	0.954248	0.909549	0.997797	0.996201	1			
Cw1460DB	0.881463	0.889082	0.876569	0.887779	0.936274	0.886485	0.997743	0.99786	0.995807	1		
Ho025DB	0.89007	0.895627	0.886397	0.896172	0.951746	0.894154	0.994877	0.993352	0.992308	0.994567	1	
Pg174DB	0.882806	0.891605	0.877994	0.888239	0.93868	0.887225	0.998425	0.998519	0.995592	0.998497	0.994979	1

ITS2

	Cw1210	*Cw1243*	*Cw1295*	*Cw1460*	*Ho025*	*Pg174*	*Cw1210DB*	*Cw1243DB*	*Cw1295DB*	*Cw1460DB*	*Ho025DB*	*Pg174DB*
Cw1210	1											
Cw1243	0.999296	1										
Cw1295	0.999917	0.999139	1									
Cw1460	0.999993	0.999173	0.999892	1								
Ho025	0.999915	0.999545	0.999784	0.999888	1							
Pg174	0.999906	0.999513	0.999748	0.999885	0.999997	1						
Cw1210DB	0.957971	0.957545	0.956874	0.958085	0.957807	0.95795	1					
Cw1243DB	0.943518	0.945042	0.9424	0.943458	0.94372	0.943812	0.995459	1				
Cw1295DB	0.947091	0.947115	0.946001	0.947164	0.946997	0.947125	0.998195	0.9961	1			
Cw1460DB	0.938406	0.938413	0.937245	0.938489	0.938262	0.938399	0.995642	0.996586	0.997594	1		
Ho025DB	0.947657	0.947947	0.946421	0.947727	0.947623	0.947768	0.99797	0.997327	0.998911	0.998414	1	
Pg174DB	0.944628	0.944714	0.943402	0.944715	0.944595	0.944747	0.996559	0.995017	0.997851	0.997364	0.998096	1

Cluster 602

ITS1

	Cw1324	*Ho023*	*Ho032*	*Pg104*	*Cw1324DB*	*Ho023DB*	*Ho032DB*	*Pg104DB*
Cw1324	1							
Ho023	0.947067	1						
Ho032	0.994298	0.973716	1					
Pg104	0.992182	0.907293	0.978683	1				
Cw1324DB	0.903202	0.821872	0.896364	0.923683	1			
Ho023DB	0.89624	0.814433	0.889807	0.918205	0.9979	1		
Ho032DB	0.898242	0.818694	0.893067	0.920495	0.99678	0.99912	1	
Pg104DB	0.90836	0.831446	0.903973	0.92983	0.996655	0.994717	0.994892	1

ITS2

	Cw1324	*Ho023*	*Ho032*	*Pg104*	*Cw1324DB*	*Ho023DB*	*Ho032DB*	*Pg104DB*
Cw1324	1							
Ho023	0.997142	1						
Ho032	0.997431	0.999744	1					
Pg104	0.997275	0.999786	0.999748	1				
Cw1324DB	0.9633	0.962705	0.963897	0.961858	1			
Ho023DB	0.95773	0.962331	0.963018	0.961076	0.996553	1		
Ho032DB	0.938944	0.943993	0.945298	0.94287	0.993616	0.996795	1	
Pg104DB	0.952064	0.956001	0.956637	0.955176	0.995228	0.996068	0.996621	1

Cluster 758

ITS1

	Cw147	*Hu050*	*Sp022*	*Cw147DB*	*Hu050DB*	*Sp022DB*
Cw147	1					
Hu050	0.498527	1				
Sp022	0.474651	0.996375	1			
Cw147DB	0.373356	0.879009	0.863835	1		
Hu050DB	0.385594	0.904874	0.894619	0.994173	1	
Sp022DB	0.383114	0.915196	0.907149	0.991384	0.997523	1

ITS2

	Cw147	*Hu050*	*Sp022*	*Cw147DB*	*Hu050DB*	*Sp022DB*
Cw147	1					
Hu050	0.956415	1				
Sp022	0.98541	0.989996	1			
Cw147DB	0.891708	0.915929	0.912904	1		
Hu050DB	0.876264	0.921264	0.9094	0.994555	1	
Sp022DB	0.900633	0.924703	0.923602	0.990672	0.993254	1

Cluster 871

ITS1

	Cw170	*Cw180*	*Pg067*	*Rt002*	*Sw029*	*Cw170DB*	*Cw180DB*	*Pg067DB-1*	*Pg067DB-2*	*Rt002DB*	*Sw029DB-1*	*Sw029DB-2*
Cw170	1											
Cw180	0.999939	1										
Pg067	0.993261	0.993473	1									
Rt002	0.999323	0.999658	0.993537	1								
Sw029	0.980834	0.982019	0.989889	0.984148	1							
Cw170DB	0.913804	0.912299	0.927351	0.908776	0.908633	1						
Cw180DB	0.911216	0.909711	0.925275	0.90617	0.906989	0.999818	1					
Pg067DB-1	0.892818	0.891723	0.913037	0.889376	0.898871	0.996091	0.99653	1				
Pg067DB-2	0.899694	0.898459	0.919308	0.895814	0.902513	0.996948	0.997163	0.999045	1			
Rt002DB	0.900179	0.898413	0.913516	0.894331	0.891845	0.998513	0.998667	0.995467	0.996124	1		
Sw029DB-1	0.886962	0.885733	0.907462	0.883028	0.896191	0.995068	0.995627	0.99859	0.997585	0.994917	1	
Sw029DB-2	0.875603	0.874046	0.893264	0.870498	0.878778	0.993883	0.994728	0.996609	0.995472	0.995144	0.997224	1
Sw029DB-3	0.878917	0.877537	0.896923	0.874372	0.884384	0.994327	0.995312	0.997157	0.995607	0.995023	0.997619	0.998524

ITS2

	Cw170	*Cw180*	*Pg067*	*Rt002*	*Sw029*	*Cw170DB*	*Cw180DB*	*Pg067DB*	*Rt002DB*	*Sw029DB-1*	*Sw029DB-2*
Cw170	1										
Cw180	0.999995	1									
Pg067	0.997706	0.997543	1								
Rt002	0.996931	0.996767	0.999405	1							
Sw029	0.999526	0.99953	0.998105	0.99732	1						
Cw170DB	0.909945	0.910076	0.906844	0.907787	0.909426	1					
Cw180DB	0.910189	0.910352	0.906556	0.907443	0.909782	0.99945	1				
Pg067DB	0.917512	0.917561	0.917765	0.918608	0.918006	0.997647	0.997207	1			
Rt002DB	0.889205	0.889279	0.889661	0.891332	0.889877	0.994524	0.994042	0.995985	1		
Sw029DB-1	0.892793	0.892901	0.891321	0.892278	0.89337	0.995351	0.995322	0.995602	0.998764	1	
Sw029DB-2	0.908898	0.908984	0.907442	0.908253	0.909254	0.99851	0.998145	0.998573	0.997083	0.998065	1

Cluster 1002

ITS1

	Cw301	*Hu184*	*Cw301DB-1*	*Cw301DB-2*	*Hu184DB-1*	*Hu184DB-2*
Cw301	1					
Hu184	0.942939	1				
Cw301DB-1	0.840777	0.898692	1			
Cw301DB-2	0.877736	0.934289	0.987443	1		
Hu184DB-1	0.837814	0.90415	0.994538	0.978933	1	
Hu184DB-2	0.871395	0.934989	0.989026	0.985456	0.994389	1

ITS2

	Cw301	*Hu184*	*Cw301DB-1*	*Cw301DB-2*	*Hu184DB-1*	*Hu184DB-2*	*Hu184DB-3*	*Hu184DB-4*
Cw301	1							
Hu184	0.947058	1						
Cw301DB-1	0.909701	0.946025	1					
Cw301DB-2	0.886659	0.95817	0.983436	1				
Hu184DB-1	0.903418	0.946802	0.995034	0.980032	1			
Hu184DB-2	0.902551	0.944195	0.99498	0.979847	0.999468	1		
Hu184DB-3	0.904628	0.946447	0.994922	0.979445	0.999714	0.999558	1	
Hu184DB-4	0.886337	0.945396	0.985027	0.992554	0.987981	0.987722	0.987551	1

Cluster 1062

ITS1

	Cw361	*Hu045*	*Sw022*	*Sw034*	*Cw361DB*	*Hu045DB*	*Sw022DB*	*Sw034DB*
Cw361	1							
Hu045	0.995317	1						
Sw022	0.996017	0.999727	1					
Sw034	0.996037	0.996822	0.998042	1				
Cw361DB	0.869297	0.881097	0.876261	0.86971	1			
Hu045DB	0.893696	0.903871	0.899837	0.894489	0.995624	1		
Sw022DB	0.86955	0.881725	0.876773	0.869824	0.999569	0.995229	1	
Sw034DB	0.868807	0.880527	0.875678	0.869815	0.998981	0.994119	0.999323	1

ITS2

	Cw361	*Hu045*	*Sw022*	*Sw034*	*Cw361DB*	*Hu045DB*	*Sw022DB*	*Sw034DB*
Cw361	1							
Hu045	0.986495	1						
Sw022	0.985557	0.993166	1					
Sw034	0.995309	0.997024	0.995707	1				
Cw361DB	0.879376	0.894158	0.899699	0.893793	1			
Hu045DB	0.871983	0.898259	0.899199	0.892317	0.994756	1		
Sw022DB	0.888493	0.914456	0.921312	0.909895	0.990562	0.993767	1	
Sw034DB	0.902769	0.916592	0.922305	0.916471	0.995819	0.993167	0.995196	1

Cluster 1071

ITS1

	Cw370	*Hu733*	*Sp043*	*Sw058*	*Cw370DB-1*	*Cw370DB-2*	*Hu733DB-1*	*Hu733DB-2*	*Sp043DB-1*	*Sp043DB-2*	*Sw058DB*
Cw370	1										
Hu733	0.983687	1									
Sp043	0.981913	0.997725	1								
Sw058	0.988372	0.980869	0.978936	1							
Cw370DB-1	0.8894	0.868667	0.868035	0.870964	1						
Cw370DB-2	0.918247	0.900818	0.900626	0.90047	0.99586	1					
Hu733DB-1	0.89795	0.888972	0.890176	0.883097	0.993327	0.994109	1				
Hu733DB-2	0.905276	0.894917	0.895776	0.88957	0.995419	0.997232	0.995473	1			
Sp043DB-1	0.891606	0.878933	0.879792	0.875445	0.9973	0.996168	0.995926	0.998651	1		
Sp043DB-2	0.928428	0.924135	0.92535	0.914082	0.987547	0.995306	0.994604	0.995702	0.992737	1	
Sw058DB	0.879621	0.868457	0.869137	0.874795	0.990996	0.987856	0.992708	0.990437	0.992504	0.98517	1

ITS2

	Cw370	*Hu733*	*Sp043*	*Sw058*	*Cw370DB-1*	*Cw370DB-2*	*Hu733DB-1*	*Hu733DB-2*	*Sp043DB-1*	*Sp043DB-2*	*Sw058DB*
Cw370	1										
Hu733	0.999397	1									
Sp043	0.998663	0.999655	1								
Sw058	0.9987	0.99906	0.998629	1							
Cw370DB-1	0.887985	0.888241	0.886854	0.889586	1						
Cw370DB-2	0.965491	0.965932	0.964885	0.966832	0.970253	1					
Hu733DB-1	0.883571	0.885401	0.884635	0.886274	0.996601	0.965435	1				
Hu733DB-2	0.963183	0.964648	0.964019	0.965372	0.970518	0.999217	0.967591	1			
Sp043DB-1	0.891664	0.893689	0.893456	0.894447	0.997293	0.970706	0.996669	0.972596	1		
Sp043DB-2	0.946754	0.948567	0.948737	0.949382	0.978299	0.9963	0.975685	0.997704	0.980567	1	
Sw058DB	0.905595	0.906197	0.90488	0.909522	0.996272	0.977222	0.992667	0.977296	0.995834	0.982134	1

Cluster 1149

ITS1

	Cw471	*Cw784*	*Dg114*	*Rt076*	*Cw471DB*	*Cw784DB-1*	*Cw784DB-2*	*Dg114DB*	*Rt076DB-1*	*Rt076DB-2*
Cw471	1									
Cw784	0.999329	1								
Dg114	0.989401	0.988769	1							
Rt076	0.991083	0.992761	0.981962	1						
Cw471DB	0.907387	0.907856	0.890562	0.918342	1					
Cw784DB-1	0.899207	0.899569	0.882804	0.911532	0.998881	1				
Cw784DB-2	0.929632	0.929872	0.914692	0.940779	0.996419	0.995513	1			
Dg114DB	0.875227	0.876336	0.85805	0.888868	0.995722	0.996859	0.987925	1		
Rt076DB-1	0.88385	0.884813	0.868408	0.90249	0.995597	0.997135	0.991727	0.996776	1	
Rt076DB-2	0.889628	0.890662	0.874372	0.910828	0.993818	0.994561	0.992904	0.991698	0.997385	1

ITS2

	Cw471	*Cw784*	*Dg114*	*Rt076*	*Cw471DB*	*Cw784DB*	*Dg114DB*	*Rt076DB*
Cw471	1							
Cw784	0.999608	1						
Dg114	0.998988	0.999195	1					
Rt076	0.997783	0.997338	0.996069	1				
Cw471DB	0.939092	0.939402	0.937917	0.937992	1			
Cw784DB	0.928786	0.929466	0.928116	0.927343	0.995336	1		
Dg114DB	0.916423	0.916541	0.916532	0.915348	0.994531	0.995634	1	
Rt076DB	0.913715	0.91293	0.911346	0.916267	0.991493	0.995537	0.994926	1

Cluster 1200

ITS1

	Cw522	*Sp053*	*Sp056*	*Sp093*	*Sw044*	*Cw522DB*	*Sp053DB*	*Sp056DB*	*Sp093DB*	*Sw044DB*
Cw522	1									
Sp053	0.992172	1								
Sp056	0.991845	0.999686	1							
Sp093	0.99155	0.999756	0.999827	1						
Sw044	0.991575	0.999801	0.999842	0.999925	1					
Cw522DB	0.829429	0.847827	0.845167	0.850465	0.847092	1				
Sp053DB	0.826136	0.848252	0.84577	0.851119	0.847866	0.997482	1			
Sp056DB	0.865028	0.884028	0.88201	0.886786	0.883805	0.995086	0.996257	1		
Sp093DB	0.866671	0.886371	0.884451	0.889255	0.886295	0.993944	0.995392	0.999529	1	
Sw044DB	0.863425	0.880137	0.879042	0.883277	0.880537	0.990895	0.990651	0.994673	0.994854	1

ITS2

	Cw522	*Sp053*	*Sp056*	*Sp093*	*Sw044*	*Cw522DB*	*Sp053DB*	*Sp056DB*	*Sp093DB*	*Sw044DB*
Cw522	1									
Sp053	0.997996	1								
Sp056	0.998442	0.999908	1							
Sp093	0.998738	0.999799	0.999975	1						
Sw044	0.998382	0.999969	0.999979	0.999924	1					
Cw522DB	0.939849	0.935597	0.93619	0.936674	0.936142	1				
Sp053DB	0.923888	0.92062	0.921431	0.921907	0.921204	0.994707	1			
Sp056DB	0.938279	0.934529	0.935515	0.936075	0.935212	0.996671	0.998131	1		
Sp093DB	0.92752	0.923931	0.924824	0.925345	0.924568	0.994055	0.996999	0.996364	1	
Sw044DB	0.921275	0.917413	0.918302	0.918833	0.918055	0.995195	0.999248	0.997481	0.996624	1

Cluster 1237

ITS1

	Cw559	*Hu079*	*Rt108*	*Cw559DB*	*Hu079DB*	*Rt108DB*
Cw559	1					
Hu079	0.914445	1				
Rt108	0.996665	0.890131	1			
Cw559DB	0.843534	0.916272	0.813832	1		
Hu079DB	0.852563	0.939904	0.825511	0.991368	1	
Rt108DB	0.819088	0.914064	0.785931	0.993731	0.99003	1

ITS2

	Cw599	*Hu079*	*Rt108*	*Cw559DB*	*Hu079DB*	*Rt108DB*
Cw599	1					
Hu079	0.953743	1				
Rt108	0.958672	0.998435	1			
Cw559DB	0.931508	0.947544	0.952501	1		
Hu079DB	0.913055	0.928257	0.930377	0.993831	1	
Rt108DB	0.905826	0.916659	0.922711	0.990769	0.994799	1

Cluster 1449

ITS1

	Cw772	*Hu1030*	*Hu1614*	*Cw772DB*	*Hu1030DB-1*	*Hu1030DB-2*	*Hu1614DB*
Cw772	1						
Hu1030	0.985165	1					
Hu1614	0.986212	0.992883	1				
Cw772DB	0.867856	0.854064	0.848189	1			
Hu1030DB-1	0.869386	0.858303	0.854687	0.995955	1		
Hu1030DB-2	0.869386	0.858303	0.854687	0.995955	1	1	
Hu1614DB	0.892828	0.883703	0.881792	0.990382	0.995143	0.995143	1

ITS2

	Cw772	*Hu1030*	*Hu1614*	*Cw772DB*	*Hu1030DB*	*Hu1614DB*
Cw772	1					
Hu1030	0.999402	1				
Hu1614	0.999319	0.99916	1			
Cw772DB	0.915882	0.913574	0.91393	1		
Hu1030DB	0.92675	0.925151	0.924904	0.99551	1	
Hu1614DB	0.900739	0.898485	0.900209	0.993677	0.9938	1

Cluster 1460

ITS1

	Cw782	*Hu657*	*Cw782DB*	*Hu657DB*
Cw782	1			
Hu657	0.981312	1		
Cw782DB	0.898686	0.913615	1	
Hu657DB	0.890688	0.901625	0.995335	1

ITS2

	Cw782	*Hu657*	*Cw782DB*	*Hu657DB-1*	*Hu657DB-2*
Cw782	1				
Hu657	0.998575	1			
Cw782DB	0.942944	0.939973	1		
Hu657DB-1	0.933816	0.934631	0.995363	1	
Hu657DB-2	0.924593	0.925002	0.994362	0.99622	1

Cluster 1507

ITS1

	Cw841	*Hu142*	*Cw841DB*	*Hu142DB-1*	*Hu142DB-2*	*Hu142DB-3*	*Hu142DB-4*
Cw841	1						
Hu142	0.972961	1					
Cw841DB	0.904584	0.895571	1				
Hu142DB-1	0.875254	0.884451	0.990742	1			
Hu142DB-2	0.870738	0.879887	0.990905	0.997205	1		
Hu142DB-3	0.874643	0.879391	0.993458	0.995553	0.998724	1	
Hu142DB-4	0.887807	0.894423	0.993845	0.995665	0.99762	0.997845	1

ITS2

	Cw841	*Hu142*	*Cw841DB*	*Hu142DB-1*	*Hu142DB-2*	*Hu142DB-3*	*Hu142DB-4*	*Hu142DB-5*	*Hu142DB-6*
Cw841	1								
Hu142	0.988674	1							
Cw841DB	0.944163	0.925255	1						
Hu142DB-1	0.942576	0.935129	0.991122	1					
Hu142DB-2	0.944616	0.933144	0.992468	0.999168	1				
Hu142DB-3	0.943136	0.932308	0.992147	0.999543	0.99971	1			
Hu142DB-4	0.945771	0.936252	0.99063	0.998747	0.999255	0.998854	1		
Hu142DB-5	0.947546	0.937728	0.991568	0.998922	0.999413	0.999079	0.999595	1	
Hu142DB-6	0.949682	0.939504	0.991466	0.998424	0.999119	0.998644	0.999805	0.999533	1

Cluster 1533

ITS1

	Cw867	*Hu148*	*Cw867DB*	*Hu148DB-1*	*Hu148DB-2*	*Hu148DB-3*
Cw867	1					
Hu148	0.963728	1				
Cw867DB	0.922598	0.855018	1			
Hu148DB-1	0.911025	0.853005	0.996635	1		
Hu148DB-2	0.911571	0.853483	0.997197	0.999263	1	
Hu148DB-3	0.910043	0.85156	0.997335	0.999596	0.999556	1

ITS2

	Cw867	*Hu148*	*Cw867DB*	*Hu148DB-1*	*Hu148DB-2*	*Hu148DB-3*	*Hu148DB-4*	*Hu148DB-5*	*Hu148DB-6*
Cw867	1								
Hu148	0.99793	1							
Cw867DB	0.939667	0.938702	1						
Hu148DB-1	0.945879	0.949639	0.995894	1					
Hu148DB-2	0.949324	0.953192	0.995075	0.999589	1				
Hu148DB-3	0.955827	0.959392	0.994112	0.999096	0.999477	1			
Hu148DB-4	0.942342	0.946232	0.995829	0.999302	0.999137	0.997988	1		
Hu148DB-5	0.947516	0.951191	0.99554	0.999426	0.999581	0.998973	0.999533	1	
Hu148DB-6	0.952679	0.956185	0.993596	0.998614	0.99934	0.999241	0.9982	0.999126	1

Cluster 1565

ITS1

	Cw899	*Hu070*	*Pg045*	*Cw899DB*	*Hu070DB*	*Pg045DB*
Cw899	1					
Hu070	0.997096	1				
Pg045	0.998987	0.997325	1			
Cw899DB	0.889793	0.89376	0.895871	1		
Hu070DB	0.903875	0.907025	0.909399	0.998593	1	
Pg045DB	0.88608	0.889736	0.891993	0.998492	0.997741	1

ITS2

	Cw899	*Hu070*	*Pg045*	*Cw899DB*	*Hu070DB*	*Pg045DB*
Cw899	1					
Hu070	0.999991	1				
Pg045	0.999996	0.999979	1			
Cw899DB	0.946168	0.945835	0.946365	1		
Hu070DB	0.935173	0.934858	0.935353	0.998541	1	
Pg045DB	0.925693	0.925381	0.925864	0.996182	0.998759	1

Cluster 1594

ITS1

	Cw928	*Hu1666(1)*	*Hu1666(2)*	*Cw928DB*	*Hu1666DB*
Cw928	1				
Hu1666(1)	0.97709	1			
Hu1666(2)	0.849449	0.887306	1		
Cw928DB	0.923737	0.902035	0.778333	1	
Hu1666DB	0.913136	0.892517	0.7704	0.994293	1

ITS2

	Cw928	*Hu1666(1)*	*Hu1666(2)*	*Cw928DB*	*Hu1666DB-1*	*Hu1666DB-2*	*Hu1666DB-3*
Cw928	1						
Hu1666(1)	0.999984	1					
Hu1666(2)	0.999516	0.999642	1				
Cw928DB	0.938377	0.938401	0.938083	1			
Hu1666DB-1	0.930018	0.929984	0.929352	0.998107	1		
Hu1666DB-2	0.931812	0.931809	0.931423	0.998069	0.999452	1	
Hu1666DB-3	0.938504	0.938515	0.938123	0.998904	0.99915	0.999308	1

Size of each cluster

Each isolate within a cluster is derived from a different known host organism

Number of isolates in a cluster	Number of clusters
2	6
3	10
4	6
5	5
6	1
Total number of isolates	97

Barcodes used for isolate specific identification

gatc	agtc	ctga	tcag	cagg	cgaa	catt	gtcc
gact	actg	ctag	tacg	ggta	gaat	gtta	gtaa
gtac	agct	cgat	tgca	gcga	acat	tcta	cgca
gcat	atgc	cgta	tgac	gcgt	acag	tctg	cgct
gcta	atcg	cagt	tcga	ggca	aact	ttca	ccga
gtca	acgt	catg	tagc	ggct	aacg	ttcg	ccgt
cgga	caat	cttg	gcca	gctg	acta	tcgt	cgac
cggt	caag	ctta	gcct	gcag	acga	tcat	cgtc
cgtg	cata	ctat	gcac	ggac	aatc	ttac	ccag
cgag	caga	ctgt	gctc	ggtc	aagc	ttgc	cctg
ctgg	ctaa	cgtt	gacc	gtcg	atca	tact	cagc
gtgc	atac	agta	cacg	gacg	agca	tgct	ctgc
gagc	agac	tatc	ctcg	aatg			

All raw data can be retrieved by contacting Christopher Kitts at California Polytechnic State University ckitts@calpoly.edu

1	Av009	47	Cw1460
2	Cw053	48	Ho025
3	Pg085	49	Pg174
4	Av020	50	Cw1324
5	Ct036	51	Ho023
6	Cw377	52	Ho032
7	Cw963	53	Pg104
8	Cw990	54	Cw147
9	Av049	55	Hu050
10	Cw626	56	Sp022
11	Cw780	57	Cw170
12	Cw893	58	Cw180
13	Pg065	59	Pg067
14	Ck022	60	Rt002
15	Cw1043	61	Sw029
16	Pg063	62	Cw301
17	Ck035	63	Hu184
18	Cw099	64	Cw361
19	Hu615	65	Hu045
20	Ck036	66	Sw022
21	Ct028	67	Sw034
22	Hu074	68	Cw370
23	Ct058	69	Hu733
24	Dg112	70	Sp043
25	Dg146	71	Sw058
26	Hu947	72	Cw471
27	Cw1002	73	Cw784
28	Cw999	74	Dg114
29	Sp079	75	Rt076
30	Hu143	76	Cw522
31	Cw1001	77	Sp053
32	Hu053	78	Sp056
33	Sp063	79	Sp093
34	Cw1111	80	Sw044
35	Cw1132	81	Cw559
36	Hu095	82	Hu079
37	Cw1151	83	Rt108
38	Pg052	84	Cw772
39	Rt071	85	Hu1030
40	Cw1159	86	Hu1614
41	Cw140	87	Cw782
42	Pg010	88	Hu657
43	Pg033	89	Cw841
44	Cw1210	90	Hu142
45	Cw1243	91	Cw867
46	Cw1295	92	Hu148

93	Cw899		
94	Hu070		
95	Pg045		
96	Cw928		
97	Hu1666(1)		
98	Hu1666(2)		