Journal of Microbiological Methods 105 (2014) 121–129

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Pyroprinting: A rapid and flexible genotypic fingerprinting method for typing bacterial strains

Michael W. Black [a,*], Jennifer VanderKelen [a], Aldrin Montana [b,1], Alexander Dekhtyar [b], Emily Neal [a], Anya Goodman [c], Christopher L. Kitts [a]

[a] Biological Sciences Department, California Polytechnic State University, San Luis Obispo, CA, USA
[b] Computer Science Department, California Polytechnic State University, San Luis Obispo, CA, USA
[c] Department of Chemistry and Biochemistry, California Polytechnic State University, San Luis Obispo, CA, USA

ARTICLE INFO

Article history:
Received 10 May 2014
Received in revised form 17 July 2014
Accepted 19 July 2014
Available online 1 August 2014

Keywords:
Pyrosequencing
DNA fingerprinting
rRNA
Strain typing
Microbial source tracking (MST)
Escherichia coli

ABSTRACT

Bacterial strain typing is commonly employed in studies involving epidemiology, population ecology, and microbial source tracking to identify sources of fecal contamination. Methods for differentiating strains generally use either a collection of phenotypic traits or rely on some interrogation of the bacterial genotype. This report introduces pyroprinting, a novel genotypic strain typing method that is rapid, inexpensive, and discriminating compared to the most sensitive methods already in use. Pyroprinting relies on the simultaneous pyrosequencing of polymorphic multicopy loci, such as the intergenic transcribed spacer regions of rRNA operons in bacterial genomes. Data generated by sequencing combinations of variable templates are reproducible and intrinsically digitized. The theory and development of pyroprinting in *Escherichia coli*, including the selection of similarity thresholds to define matches between isolates, are presented. The pyroprint-based strain differentiation limits and phylogenetic relevance compared to other typing methods are also explored. Pyroprinting is unique in its simplicity and, paradoxically, in its intrinsic complexity. This new approach serves as an excellent alternative to more cumbersome or less phylogenetically relevant strain typing methods.

1. Introduction

As the vast expanse of bacterial diversity becomes more apparent, methods for differentiating between closely related bacterial strains are increasingly important. Bacterial strain typing is commonly employed in studies involving epidemiology (Tenover et al., 1997; Van Belkum et al., 2007), population ecology (Anderson et al., 2006; Gordon, 2003, 2010; Gordon et al., 2008), and microbial source tracking to identify sources of fecal contamination (Carlos et al., 2010; McLellan et al., 2003; Scott et al., 2002). Methods for defining strains generally use either a collection of phenotypic traits or rely on some interrogation of the bacterial genotype (Tenover et al., 1997). The collection of traits used, whether phenotypic or genotypic, serves as a fingerprint to match isolates obtained from multiple samples, thus providing evidence of a relationship between the isolates (Tenover et al., 1997; Van Belkum et al., 2007).

Genotypic fingerprinting methods proliferated over the last 20 years as researchers searched for better ways to discriminate between closely related strains in a variety of uses. Genotypic fingerprints fall into three categories: DNA banding patterns, DNA hybridization patterns, and DNA sequencing (Li et al., 2009). In general, optimally-performing typing methods use traits that remain stable for reasonable periods, are applicable to all isolates in question, discriminate well between isolates already known to be different strains, and most importantly, result in reproducible outputs: the fingerprints. Convenient typing methods are also flexible, rapid, accessible, easy to use, low cost, and amenable to computerized analysis (Li et al., 2009; Van Belkum et al., 2007). At present, no single method is universally ideal. Each method has its own strengths and weaknesses among the criteria listed above. Thus, the trade-offs among performance and convenience criteria are application-specific.

This report details a novel genotypic typing method called pyroprinting that meets most of the criteria above. Similar to multilocus sequence typing (MLST) (Maiden et al., 1998), pyroprinting uses DNA pyrosequencing to interrogate multiple polymorphic loci to increase the power of discrimination between closely related organisms. Pyrosequencing is a commonly used method of DNA sequence analysis that employs a sequencing-by-synthesis method for the detection of inorganic pyrophosphate produced when nucleotides are incorporated during DNA synthesis. Through a series of enzymatic reactions, these sequencing byproducts are used to drive a luciferase-catalyzed reaction generating a light signal proportional to the number of nucleotides incorporated (Ronaghi et al., 1998). The key difference between these

* Corresponding author at: Department of Biological Sciences, California Polytechnic State University, CA, 93420, USA. Tel.: +1 805 756 2894.
E-mail address: mblack@calpoly.edu (M.W. Black).
[1] Workday, Pleasanton, CA, USA.

http://dx.doi.org/10.1016/j.mimet.2014.07.019
0167-7012/

typing methods is that MLST evaluates each polymorphic locus independently while pyroprinting simultaneously pyrosequences the multiple loci in a single reaction. Since the existence of multiple polymorphic templates confounds the user's ability to read a nucleotide sequence, the pyroprint data instead represents the population of templates in the form of a pattern of light output peak heights. Thus, for each combination of templates, a unique genetic fingerprint may be obtained: a pyroprint.

The internal transcribed spacer (ITS) regions in the rRNA operons (rDNA) are excellent targets for pyroprinting. ITS regions have been used for distinguishing bacteria because they are highly variable, consisting of polymorphic noncoding sequences and insertions of one or more tRNA genes (Brosius et al., 1981; Jensen et al., 1993). In addition, multiple ITS alleles are often present in a single organism as bacterial genomes harbor multiple copies of rDNA. Due to the ubiquitous nature of these sequences, pyroprinting is applicable to any bacterial species that can be isolated and purified by conventional culture methods. In *Escherichia coli*, the model organism used in this study, there are seven rRNA operons present in the genome named *rrnA*, *B*, *C*, *D*, *E*, *H*, and *G*. The short, variable ITS regions are flanked by highly conserved rRNA genes that provide convenient primer sequences for both amplification and sequence analysis. By using the rDNA operons, both the type of allele and the relative frequencies of the alleles present in the genome contribute to a strain-specific pyroprint. The data generated from pyroprinting are reproducible, readily digitized and instantly amenable to computerized analysis. These qualities make pyroprinting ideally suited to tracking studies where large numbers of isolates must be tested and compared to a large database of fingerprints (Van Belkum et al., 2007).

2. Materials and methods

2.1. Isolation and purification of E. coli

Fecal samples were obtained from human subjects either directly from stool or via anal swabs. Samples were initially plated onto MacConkey Agar (DIFCO, Detroit, Michigan, USA) to obtain well-isolated colony forming units. To verify the isolation of *E. coli*, red colony forming units from the MacConkey plates were subject to additional metabolic tests. To evaluate lactose fermentation, isolates were plated onto Eosin-methylene blue agar (EMB, DIFCO) to identify those that produced a green metallic sheen. To evaluate citrate utilization, isolates were plated onto Simmons citrate agar (Oxoid, Basingstoke, Hampshire, United Kingdom) to identify those that do not grow on the medium. To evaluate tryptophanase activity, isolates were grown in tryptone broth (Neogen, Lansing, Michigan, USA) to detect indole production using Kovac's reagent (5% p-dimethylamino benzaldehyde in 75% butyl alcohol and 25% HCl). Following this series of metabolic tests, the isolates were considered to be *E. coli* and further evaluated by pyroprinting.

2.2. Construction of mock genomes

Genomic DNA was isolated from 50 *E. coli* isolates harvested from environmental samples and 5 different animal hosts using the DNeasy Tissue Kit without RNase treatment (Qiagen, Valencia, California, USA). A 4 kb region of the rRNA operons containing the ITS 1, 23S gene, and ITS 2 was amplified from each isolate using illustra puReTaq Ready-To-Go PCR Beads (GE Healthcare, Pittsburgh, Pennsylvania, USA) or a 1× final concentration of 2× Quickload Taq Master Mix (New England Biolabs, Ipswich, Massachusetts, USA) and 0.2 μM of primers 1525-F and 23-5-R (Table 1) in a total reaction volume of 25 μl. The thermocycling protocol was 95 °C for 2 minutes followed by 35 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds and 68 °C for 4 minutes. The PCR products were cloned into the pCR2.1 vector using the TOPO TA Cloning Kit, and plasmids were transformed into One-Shot TOP10 competent cells (Invitrogen/Life Technologies, Grand Island, New York, USA). Plasmids were sequenced (Sequetech, Mountain View, California, USA) in the forward and reverse directions using M13F and M13R primers. Sequences were trimmed and aligned using the EditSeq and Megalign programs (Lasergene 5, DNASTAR, Madison, Wisconsin, USA) to identify unique ITS alleles. Seven plasmids, differing from each other in the ITS 2 region by at least one base, were chosen for the mock genome analysis. All plasmids were normalized to a concentration of 0.1 μg/ml and mixed to create mock genomes that differed in the molar ratios of the alleles.

Table 1
Primers and dispensation sequences used in this study.

Primer name[a]	Sequence	T_m (°C)
16-23-F[b]	GGAACCTGCGGTTGGATCAC	60
16-23-R-biotin	[Bio~5′]CTTCATCGCCTCTGACTGCC	59
23-5-F	ATGAACCGTGAGGCTTAACCTT	57
23-5-R-biotin[c]	[Bio~5′]CTACGGCGTTTCACTTCTGAGT	57
23-5-Seq	GAGGCTTAACCTT	37
1525-F	GGTTGGATCACCTCCTTACC	55
ITS template	**Dispensation order**	**Length**
ITS1	CCTCTACTAGAGCG 20(TCGA)TT	96
ITS2[d]	AACACGCGA 21(GATC)C	94

[a] ITS1-16-23; ITS2 = 23-5.
[b] Forward primer also used as ITS1 sequencing primer.
[c] Nonbiotinylated version (23-5-R) used for cloning alleles.
[d] Original dispensation was AACACGCGA 23(GATC)GAA.

2.3. Pyroprinting the rDNA internal transcribed spacers

Primers for amplifying the ITS regions and pyrosequencing were based on conserved regions in rDNA cassettes of fully sequenced *E. coli* genomes retrieved using Integrated Microbial Genomes (IMG v.3.3; (Markowitz et al., 2012)). ITS sequences were aligned using CLUSTALW (Larkin et al., 2007) and regions of conservation near the variable ITS were noted. Pyromark Assay Design software (V2.0, Qiagen) was used to select primer sequences optimized for pyrosequencing. The regions between the 16S and 23S genes (ITS1) and the 23S and 5S genes (ITS2) were amplified in separate PCR reactions either directly from *E. coli* colonies, from genomic DNA purified from *E. coli* cultures, or from mock genomes prepared using purified plasmids carrying sequenced ITS alleles. For colony PCR, a small amount of cell paste consisting of approximately 2×10^7 cells from an isolated colony was directly added to the complete PCR mix. For genomic DNA, the DNA was isolated from *E. coli* using the ZR Fungal/Bacterial DNA MiniPrep (Zymo Research, Irvine, California, USA) with final concentrations representing between 1×10^4 and 1×10^6 cells. The mock genomes were prepared by mixing together plasmids with known ITS sequences to represent allelic ratios of an *E. coli* genome with 7 copies of the rRNA operons (see below). Each reaction contained 50 pg of plasmids, which is equivalent to the number of rRNA operons represented in 8.7×10^5 purified *E. coli* genomes. PCR was performed using 0.2 μM of each respective forward and reverse primer (Table 1) and 1× Quickload Taq Master Mix (New England Biolabs) in a total reaction volume of 25 μl. The thermocycling protocol for the ITS1 locus was 95 °C for 2 minutes followed by 40 cycles of 95 °C for 30 seconds, and 68 °C for 90 seconds (served as both annealing and extension) with a final extension at 68 °C for 5 minutes. The protocol for the ITS2 locus was 95 °C for 2 minutes followed by 45 cycles of 95 °C for 30 seconds, 56 °C for 30 seconds, and 68 °C for 60 seconds with a final extension at 68 °C for 5 minutes. The appropriate length and quality of the PCR products were confirmed by agarose gel electrophoresis. The 15-20 μl of the remaining PCR product was pyrosequenced using the Pyromark Q24 pyrosequencer as described in manufacturer's protocols (Qiagen) using ITS-defined sequencing primers and dispensation sequences (Table 1). The

sequencing data are reported in an XML file, including the peak of the light output within the timeframe of each dispensation. A pyroprint is defined as the vector of peak heights keyed to the dispensation sequence used in the pyrosequencing reaction.

2.4. Analysis of pyroprint data

Each *E. coli* isolate has two ITS regions that are independently pyroprinted. Given a pair of *E. coli* isolates I_1 and I_2, each represented as a pair of pyroprints (one per ITS region), $I_1 = (p1, p2)$, $I_2 = (q1, q2)$, we determine whether I_1 and I_2 belong to the same strain by comparing the paired pyroprints. Pyroprint $p1$ is compared to $q1$, while $p2$ is compared to $q2$. The comparison of each pair results in a binary similar/not similar decision. Isolates I_1 and I_2 are then considered to belong to the same strain if their pyroprints from both ITS regions were deemed similar, i.e., if $p1$ is similar to $q1$ and $p2$ is similar to $q2$. Pairwise similarities between pyroprints were evaluated using the Pearson product moment correlation. A cumulative analysis of correlations between two pyroprints after a given dispensation was also performed for some of the trials. Starting with a string of peak heights corresponding to the first three dispensations, pairwise correlations were calculated against a reference pyroprint for each successive dispensation.

2.5. Phylogenetic typing

Initial phylogroup assignments to A, B1, B2 or D were based on the presence or absence of the ChuA, YjaA and TspE4C2 gene products via triplex PCR using 1 μM of each primer (ChuA.1, ChuA.2, YjaA.1, YjaA.2, TspE4C2.1, and TspE4C2.2 (Clermont et al., 2000)) and 1× Quickload Taq Master Mix (New England Biolabs) in a total reaction volume of 25 μl. The thermocycling protocol was 94 °C for 5 minutes followed by 30 cycles of 94 °C for 30 seconds, 55 °C for 30 seconds, and 72 °C for 30 seconds. Isolates that were placed in phylogroups A and D were retested with a quadruplex method to determine the presence of two new phylogroups C or E, respectively. Colony PCR was performed using 1 μM each of the ChuA.1b, ChuA.2, YjaA.1b, Yja.2b, TspE4C2.1b and TspE4C2.2b primers and 2 μM each of the AceK.f and ArpA1.r primers (Clermont et al., 2013) with 1× Quickload Taq Master Mix in a total reaction volume of 20 μl. The thermocycling protocol was 94 °C for 4 minutes followed by 30 cycles of 94 °C for 5 seconds and 59 °C for 20 seconds with a final extension at 68 °C for 5 minutes. Isolates assigned to group A or C were further evaluated by a third PCR assay using 0.8 μM of each primer specific for phylogenetic group C (trpAgpC.1 and trpAgpC.2 (Clermont et al., 2013)). Isolates assigned to groups D or E were further evaluated by PCR using 0.8 μM of each primer specific for group E (ArpAgpE.f and ArpAgpE.r (Clermont et al., 2013)). The reactions were performed using 1× Quickload Taq Master Mix in a final reaction volume of 20 μl. The thermocycling protocol for these reactions was 94 °C for 4 minutes followed by 30 cycles of 94 °C for 5 seconds and 57 °C (group E) or 59 °C (group C) for 20 seconds with a final extension at 68 °C for 5 minutes.

Isolates were subjected to multilocus sequence typing (MLST) by amplifying seven gene fragments (*adk*, *fumC*, *icd*, *purA*, *gyrB*, *mdh*, and *recA*) with 1× Quickload Taq Master Mix, and 0.2 μM of each forward and reverse primers as presented on the University College Cork MLST website (Wirth et al., 2006). PCR fragments were confirmed by agarose gel electrophoresis and purified with the DNA clean and concentrator kit (Zymo Research). PCR products were sequenced (Sequetech) and allele numbers were assigned using the University College Cork *E. coli* MLST Database. The combination of alleles from each isolate was used to define the most related sequence type (ST). In addition, gene fragment sequences from each isolate were concatenated and aligned using Clustal W. TreeView was used to produce a phenogram to visualize sequence similarity (Lasergene 6 DNA Analysis software, DNASTAR).

3. Results

3.1. Interrogating the genotype by pyroprinting

Genotypic typing is based on the production of patterns, commonly referred to as fingerprints, generated by the analysis of polymorphic DNA sequences. In pyroprinting, fingerprints are generated by the simultaneous pyrosequencing of genomic regions present in multiple copies and containing slight variations. The loci chosen for this study are the two internal transcribed spacer regions of bacterial rRNA operons: ITS1 between the 16S and 23S genes and ITS2 between the 23S and 5S genes. To evaluate the strain typing capacity of pyroprinting, *E. coli* was chosen as a model system. *E. coli* is often used as a fecal indicator in microbial source tracking studies and thus numerous methods have been used to type strains within this species. When an ITS region is amplified by PCR in an *E. coli* isolate, the seven rRNA operons encoded in its genome serve as templates to produce a mixed population of ITS sequences. The PCR products are then pyrosequenced in a single reaction to produce a pattern of light output peak heights graphically represented by a pyrogram. Thus, a pyroprint is defined as this vector of peak heights keyed to the dispensation sequence used in the process.

An example of collecting pyroprint data based on the seven ITS2 alleles will be explained below (Fig. 1). The first dispensation (dGTP) provides a peak height reference for a single nucleotide incorporation event in all seven alleles. The second dispensation (dCTP) provides a peak height indicating a double incorporation. The difference in sequence between the two types of alleles present in this example is first observed in the fifth dispensation (dGTP), where the peak height is intermediate between a single and double nucleotide incorporation event because four alleles incorporate only one dGTP while the other three incorporate two. This results in the advancement of the polymerase on ITS2-5 to ITS2-7 ahead of those on ITS2-1 to ITS2-4. In the following dATP dispensation, nucleotides are only incorporated in ITS2-1 to ITS2-4, resulting in a peak height roughly half of a single incorporation event. This dispensation allows the polymerases on ITS2-1 to ITS2-4 to catch up with the other templates and continue synthesis in a coordinated manner.

When pyrosequencing unknown sequences, a cyclic dispensation of all four nucleotides is generally used. To compare pyroprints between two isolates, the pyrosequencing data must be derived from the same dispensation order. Thus, the most effective cyclic permutation for differentiating strains by pyroprinting was sought using an *in silico* pyrosequencing algorithm (Brandt et al., 2012) applied to the known sequenced *E. coli* genomes available at the start of this study. The algorithm uses the seven *E. coli* ITS templates (see example in Fig. 1) and models the relative light output from each nucleotide incorporated, producing an *in silico* pyroprint. All *in silico* pyroprints generated with a given dispensation sequence were compared using the Pearson product moment correlation. Correlations were averaged for each of the 24 possible cyclic dispensations. The dispensation sequence maximizing the differences between dissimilar ITS sequences will, in general, produce the lowest correlations between pyroprints. Therefore, for each ITS region, a dispensation sequence with a low average correlation was selected: TCGA for ITS1, and GATC for ITS2.

To efficiently read through conserved regions following the sequencing primers, consensus-directed dispensations were included at the start of the dispensation sequence for ITS1 (13 dispensations) and ITS2 (8 dispensations). The quality of nucleotide incorporation was also assessed by repeating the type of nucleotide added in the first two dispensations (two dispensations of dCTP in ITS1 and dATP in ITS2). This quality control measure was also included at the end of the dispensation sequence (Table 1), but the final replicated nucleotide was excluded from the correlation analysis.

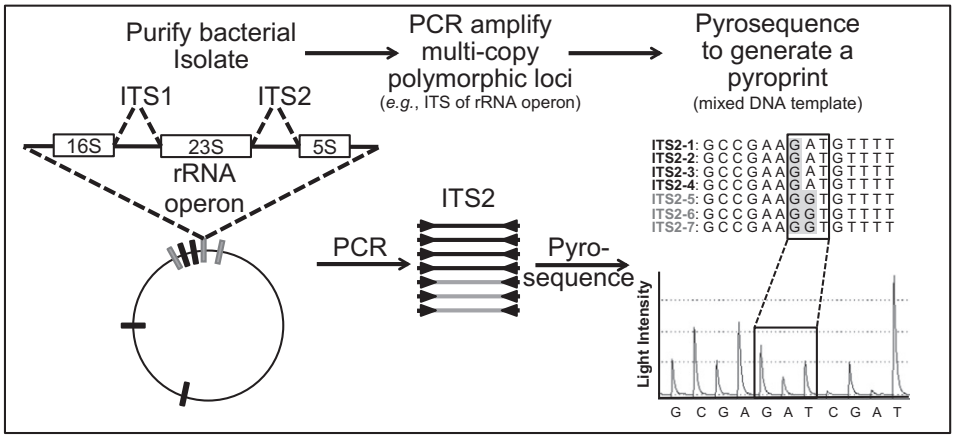

Fig. 1. Pyroprinting requires a purified bacterial isolate containing multiple copies of a polymorphic sequencing target. The rRNA operons in E. coli are present in seven copies. PCR amplification of the internal transcribed spacer (ITS) regions generates a population of mixed products that are pyrosequenced to generate a pyroprint: a pattern of light peak heights based on the allele sequences and their relative frequencies in the genome.

3.2. Defining pyroprint quality assurance parameters

The Pyromark Q24 pyrosequencing protocol used in this study includes apyrase to degrade unincorporated nucleotides and thus removes the requirement of a wash step following each nucleotide dispensation (Ronaghi, 2001). By removing a wash step, the read lengths are limited to a maximum of approximately 100 dispensations due to the accumulation of byproducts (Mashayekhi and Ronaghi, 2007). To evaluate the impact of dispensation length on reproducibility, replicate pyroprints from a single *E. coli* isolate were evaluated by tracking cumulative changes in correlation over the length of 104 dispensations (Fig. 2). Pyroprints of the ITS2 locus from six selected replicates were correlated with a reference pyroprint from the same isolate. These data were used to define 93 dispensations as the maximum length for the ITS 2 locus, as a downward trend in correlation is observed following this dispensation (vertical dotted line, Fig. 2). A maximum reproducible length of 95 dispensations was determined for the ITS1 region using a similar approach (data not shown).

Pyroprint reproducibility may also be affected by other factors, such as machine error. For example, unusually low correlations in the consensus region at the start of the dispensation order (see Rep-1 pattern, Fig. 2) may be caused by incomplete dispensation of nucleotides. Although it is possible to have strains carrying polymorphisms in the conserved region, none of the pyroprints with low initial correlations have been reproducible. Similar dispensation errors may be responsible for the steep drops in cumulative correlations observed in Rep-2 and Rep-4. These declines occur immediately following dispensations with either double peak or shoulder peak outputs (additional peaks occurring between dispensations, noted by asterisks). There were no issues identified in Rep-3 or Rep-5 to explain the differences between these replicates and Rep-6, which had the highest overall correlation to the reference pyroprint. Additional issues affecting reproducibility were identified by software-generated quality control warnings in Qiagen's PyroMark Q24 Sequencing system including: low peak heights, wide peak heights and dispensation errors (data not shown).

3.3. Determining thresholds for matching pyroprints

Pearson's correlation is an appropriate measure of similarity between pyroprints since the data is continuous and must be normalized

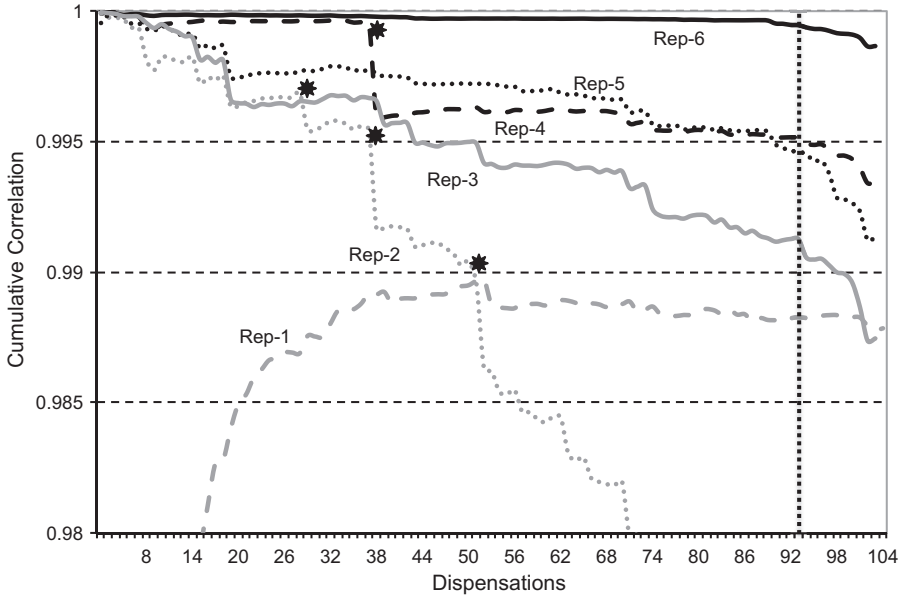

Fig. 2. Cumulative correlation plot to evaluate the quality of pyroprint data generated from the ITS2 locus. Each trace represents the cumulative correlations between one of six replicates (Rep-1 to Rep-6) and a seventh replicate used as reference. Asterisks (*) in Rep-2 and Rep-4 traces denote positions where double peaks were detected in the pyrograms. The vertical dotted line denotes a 93-dispensation cut-off for minimizing signal degradation.

before comparison. Correlation thresholds for determining a match between two pyroprints should minimize both false positive and false negative probabilities. Thus, determining a threshold requires a quantitative understanding of pyroprint reproducibility and a theoretical capacity for discriminating between closely related strains. Pyroprint reproducibility was evaluated using three randomly selected *E. coli* isolates (A, B, and C in Table 2). Theoretically, all replicate pyroprints from a single isolate should return pairwise Pearson's correlations of 1.0. However, the observed correlations between replicates varied between 0.96 and 1.0 due to natural variations in process, including biases introduced during PCR and the pyrosequencing reaction. Replicate correlations were analyzed at three thresholds: ≥0.97, ≥0.99, and ≥0.995 (Table 2). Over 95% of replicate correlations for isolates B and C were ≥0.99 at each ITS locus, suggesting this could be a good threshold to minimize false separation of isolates from the same strain. Only 90% of replicate correlations from isolate A's ITS2 were ≥0.99, even when additional replicates were included. The difference between these isolates suggests that reproducibility is influenced by the sequences of the ITS alleles and/or the number of different alleles within a genome. To estimate the incidence of this kind of false negative from the ITS2 locus, 85 *E. coli* isolates were pyroprinted in triplicate at ITS2. All three replicate correlations were ≥0.99 for 76 isolates (89.4%), two out of three replicates correlated at ≥0.99 for another 8 isolates, and only one isolate produced replicates with no correlations over 0.99. Thus, using the 0.99 correlation threshold, 19 out of 255 pairwise correlations (4.3%) resulted in falsely separating replicates (i.e., producing false negatives) and only one isolate out of 85 (1.2%) consistently produced pyroprints that failed to meet this threshold.

To investigate the capacity of pyroprinting to differentiate *E. coli* isolates with ITS allele sequences and ratios known to be similar, a series of mock genomes was created *in vitro* using plasmids containing cloned and sequenced ITS alleles. Three ITS2 alleles were chosen and combined in appropriate molar ratios to simulate the ITS complement of two *E. coli* genomes (Fig. 3). The ITS2 locus combination from each mock genome was pyroprinted in triplicate. The pyrogram shows the light output peak heights for the first 25 dispensations from each mock genome. At dispensation 10, the difference in allelic content between these trials is first detected, where the number of dGTPs incorporated is higher in mock genome 1 than in genome 2. The highlighted cytidine in allele 7 of genome 2 also causes a delay in the progression of DNA polymerase relative to the other alleles, resulting in a "ripple effect" that affects light output peak heights downstream of the single nucleotide polymorphism. For example, incorporation of the four dTTPs that follow this polymorphism is delayed relative to the other alleles: occurring in dispensation 20 instead of dispensation 16.

The differences in light output between the two mock genomes manifested in pairwise correlations. Using a single replicate of mock genome 1 as the reference pyroprint, the cumulative correlation analysis showed a large drop in correlation in the three replicates of genome 2 at dispensation 10 (Fig. 4A). A continued drop in subsequent correlations demonstrates the ripple effect, with a steep drop in dispensation 20 where the four dTTPs in allele 7 of mock strain 2 are incorporated. After completing the full 93 dispensations, the average pairwise correlation between triplicates was 0.9995 for mock genome 1 and 0.9979 for mock genome 2, while the average correlation between the two mock genomes was 0.9664 (Fig. 4B).

A set of 47 sequenced *E. coli* genomes was retrieved from GenBank (Benson et al., 2013) and used to evaluate ITS sequence diversity and the likelihood of matching isolates that should be different (false positives) for a given threshold. (Note that the origins of these genomes are biased toward pathogenic isolates from humans and do not represent a true level of ITS sequence diversity.) ITS sequences from these genomes were trimmed based on the sequencing start and dispensation sequence for each ITS locus (Table 1), resulting in 27 unique ITS1 alleles and 36 unique ITS2 alleles with an average length of 70 base pairs. These genomes were categorized into 35 distinct groups by a visual comparison of allele sequences and ratios. Several groups shared identical ITS1 or ITS2 loci, showing the importance of using pyroprints of both regions to increase the capacity for strain differentiation.

Using the *in silico* pyrosequencing algorithm described above, ITS sequences from representative genomes were converted into pyroprints. The 0.99 correlation threshold was enforced on each ITS region separately to determine matches, resulting in 34 distinct strains from the 35 originally identified based on sequence. In the single false positive match, the genomes differed from one another only in the frequencies of shared ITS2 alleles. The frequency difference occurred between two alleles (2 vs. 3 copies) that varied in a single nucleotide, resulting in a difference of one adenosine incorporated at dispensation 35. This polymorphism did not produce a ripple effect as described above and resulted in a very high correlation (0.9998) between the ITS2 pyroprints.

3.4. Use of pyroprint matches to cluster isolates into strains

Clustering with empirical pyroprints was performed on a set of 24 *E. coli* isolates collected from 9 different human hosts. Each isolate, identified by a letter for the host (A–H, J) and a number (1–6), was pyroprinted at both ITS loci. After enforcing the quality assurance measures described above, pairwise correlations of pyroprints from each locus and isolate were placed in a correlation matrix to facilitate strain identification (Fig. 5). This matrix emulates an algorithm developed to analyze pyroprint data from two loci (Montana, 2013; Montana et al., 2011). Correlations below the 0.99 threshold are set to zero to avoid erroneously clustering an isolate into a strain when one of the two loci does not meet the minimum threshold. This effect is illustrated in the matrix by white blocks while the gray shading represents matching correlations (≥0.99). The algorithm also enforces a second threshold to consistently seed clusters with the most similar pyroprints. Correlations over 0.995 are transformed to 1.0, represented in the matrix by dark gray shaded blocks. The algorithm then clusters isolates into strains only if the pairwise correlations for both ITS loci are over 0.99. First, clusters are formed between isolates where both ITS loci now show a correlation of 1.0 (dark gray). Clusters are populated with additional isolates if both regions have an average correlation ≥0.99 across all of the isolates already present in the cluster. The matrix emulates this process by reordering rows and columns until clusters of gray and dark gray blocks are formed along the diagonal. In situations where two isolates cluster with a third, but not with each other, the algorithm clusters the isolates with the highest average correlation. An example of this situation is cluster 6 (Fig. 5), which consists of isolates C4 and C6 (≥0.995 at both loci) instead of C5 and C6 (≥0.995 at ITS1, but not at ITS2).

The clustering parameters described above were used to sort these 24 human isolates into 12 groups based on their ITS1 and ITS2 pyroprints. Six of the isolates were unique in their combined pyroprints,

Table 2
Relative frequency of correlations between replicate pyroprints as a measure of reproducibility.

		ITS Pearson correlation[a]		
Isolate[b]	*N*	≥0.97	≥0.99	≥0.995
A	31	100.0%	94.1%	72.0%
B	12	100.0%	98.5%	84.8%
C	11	100.0%	100.0%	77.8%
		ITS Pearson correlation[a]		
Isolate	*N*	≥0.97	≥0.99	≥0.995
A	34	99.8%	90.3%	71.2%
B	12	100.0%	100.0%	95.5%
C	10	100.0%	96.4%	80.0%

[a] Percentage represents fraction of pairwise correlations between replicate pyroprints (*N*) at or above threshold.
[b] A, B, and C represent randomly chosen *E. coli* isolates.

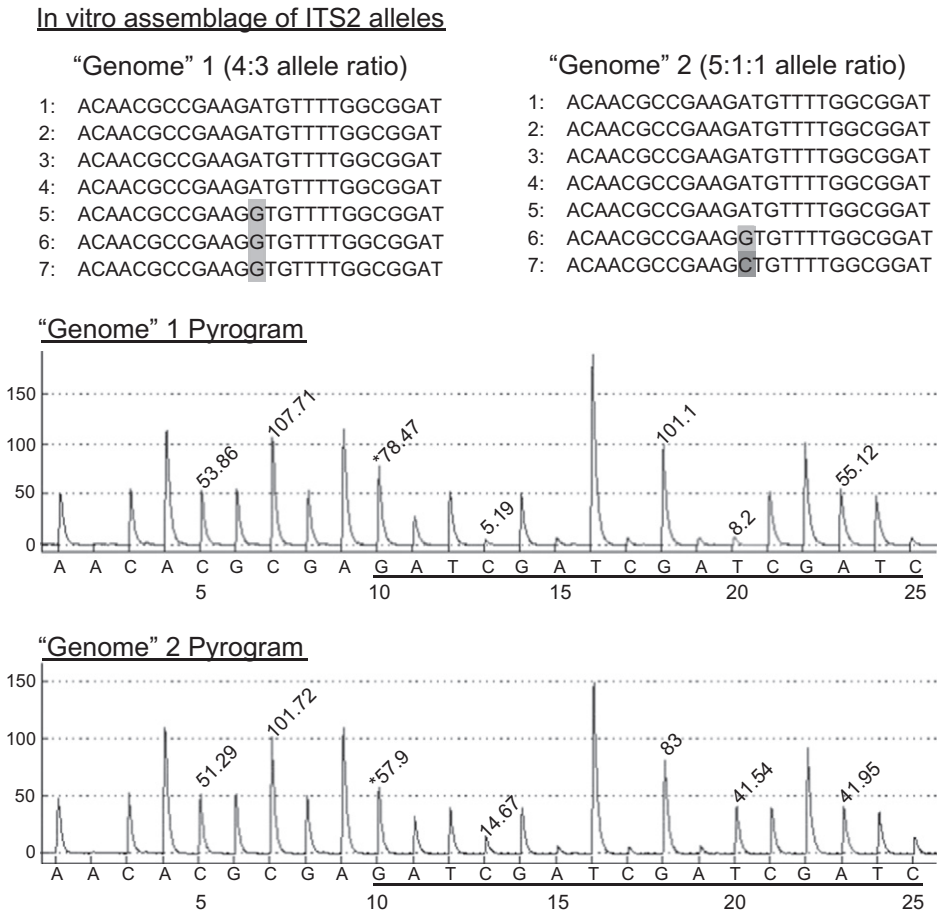

Fig. 3. Pyroprint data generated from mock genomes assembled from ITS2 plasmids. Allele differences are noted by gray shading. Traces of light peak output are shown in the pyrograms for the first 25 dispensations. The asterisks (*) over peaks at dispensation 10 (dGTP) identify the first incorporation that discriminates the mock genome pyroprints based on their allelic content. The effect of this difference is propagated throughout the remaining dispensations (underlined), as noted in the peak light output values positioned over selected peaks.

while the remaining isolates were clustered into groups of 2–6 isolates. Discrimination between isolates was reduced when the isolates were matched using only one pyroprint: the ITS1 region produced only 9 groups and ITS2 produced 10 groups.

3.5. Phylogenetic relevance of pyroprint data

Pyroprinting can reproducibly discriminate between *E. coli* isolates at a subspecies level; however, the phylogenetic relevance of pyroprint clustering must be evaluated. Currently used gold standards for determining phylogenetic relationships include phylogroup analysis (Clermont et al., 2000, 2013) and MLST (Wirth et al., 2006). If pyroprinting clusters accurately represent strains of *E. coli*, then members of a pyroprint cluster should be in the same phylogroup, and perhaps even the same MLST group.

The 24 human-derived *E. coli* isolates clustered by pyroprinting in Fig. 5 were sorted into four phylogenetic groups based on the banding patterns resulting from a triplex PCR (Clermont et al., 2000). Isolates that fell into phylogroups A or D were retested with a quadruplex method to determine if they actually belonged to newly designated phylogroups C or E (Clermont et al., 2013). All isolates in a given pyroprint cluster were in the same phylogroup (Fig. 6). As expected, pyroprinting was more discriminatory as several pyroprint clusters could be found within a given phylogroup.

The 24 *E. coli* isolates were also analyzed by sequencing portions of the *adk*, *fumC*, *icd*, *purA*, *gyrB*, *mdh* and *recA* genes for MLST analysis. The sequences were evaluated by concatenating the individual gene fragments and aligning them to produce a phenogram based on DNA sequence identity. All pyroprint clusters showed MLST sequence identity at a level of 99.8% or higher (Fig. 6). The sequence type (ST) for each isolate was determined by allele typing each amplified gene fragment using the MLST database for *E. coli*. In no case were the isolate MLST alleles identical to those already in the database, suggesting that novel *E. coli* strains were isolated. In pyroprint clusters 6 and 9, the isolates came from the same human host (hosts C and B, respectively) and had the same ST assignment (ST2166 and ST58, respectively). Four pyroprint clusters (1, 2, 5, and 11) contained isolates obtained from different hosts, indicating that this method could be used to track transmission events. Isolates from pyroprint cluster 11 were all grouped into a unique sequence type (ST10) while those in clusters 1 and 2 were grouped together by MLST (ST131). The observation that a total of three pyroprint clusters (1–3) were grouped into ST131 suggests that pyroprinting possesses a higher level of discrimination than MLST for the alleles used in this analysis. However, this is not true for all of the sequence types, as pyroprint cluster 5 was split between ST2619 and ST95 based solely on the *gyrB* allele designation. Six of the 24 isolates did not cluster with others via pyroprinting; isolate C3 grouped into the ST131 group and C5 grouped into ST2166 by MLST analysis, while the remaining four isolates were in unique ST groups. Overall, 92% of the isolates showed agreement between pyroprint clusters and MLST sequence types.

3.6. Performance characteristics of pyroprinting

Pyroprinting may be performed using PCR products directly amplified from isolated colonies or from genomic DNA purified from *E. coli*

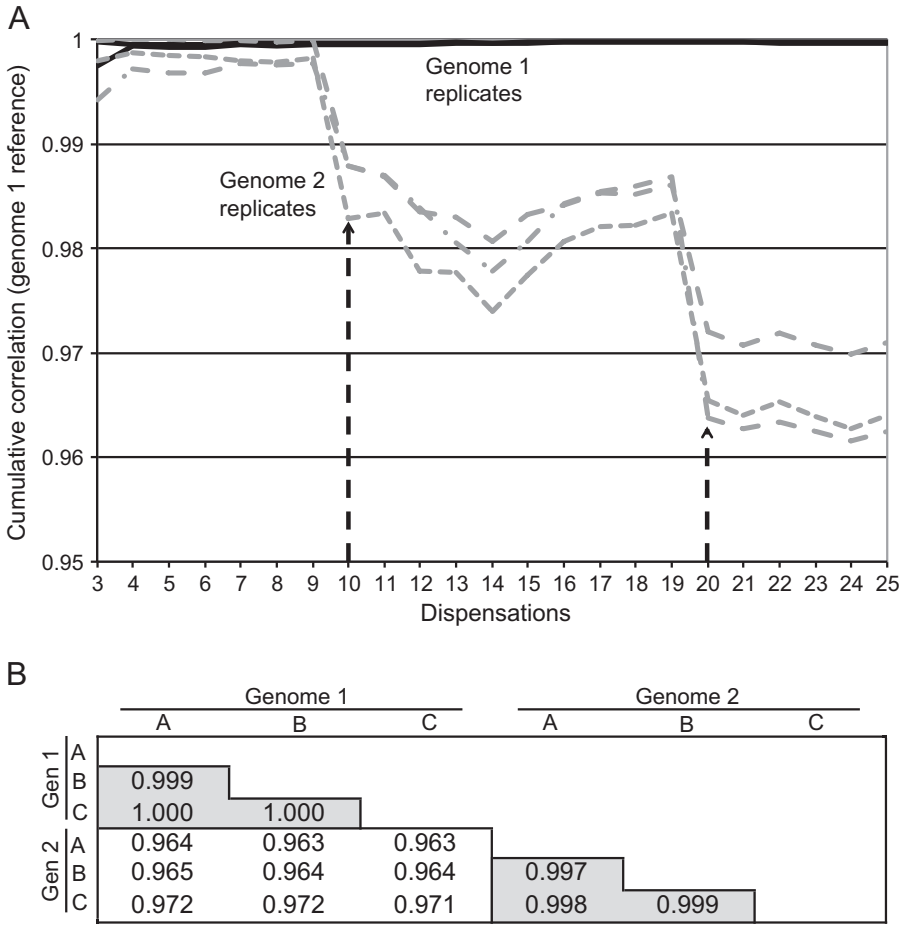

Fig. 4. Pearson correlation analysis of mock genomes generated using ITS2 plasmids. (A) Cumulative correlation analysis between a replicate of mock genome 1 (reference strain) and the remaining replicates of genome 1 (black lines), or triplicate pyroprints of genome 2 (gray lines). The hatched arrows identify dispensations resulting in large differences in light output peak heights. (B) Correlation matrix between replicate pyroprints (A, B, and C) of the mock genomes (Gen-1 and Gen-2) based on pyroprints resulting from 93 dispensations.

cultures. A comparison between these two approaches was performed on a single isolate using approximately 2×10^7 cells for colony PCR and three dilutions of purified genomic DNA that were the equivalent of 1×10^4 to 1×10^6 cells per reaction. The pyroprints generated from PCR products derived from these two methods resulted in pairwise correlations >0.99 between all of the trials (data not shown).

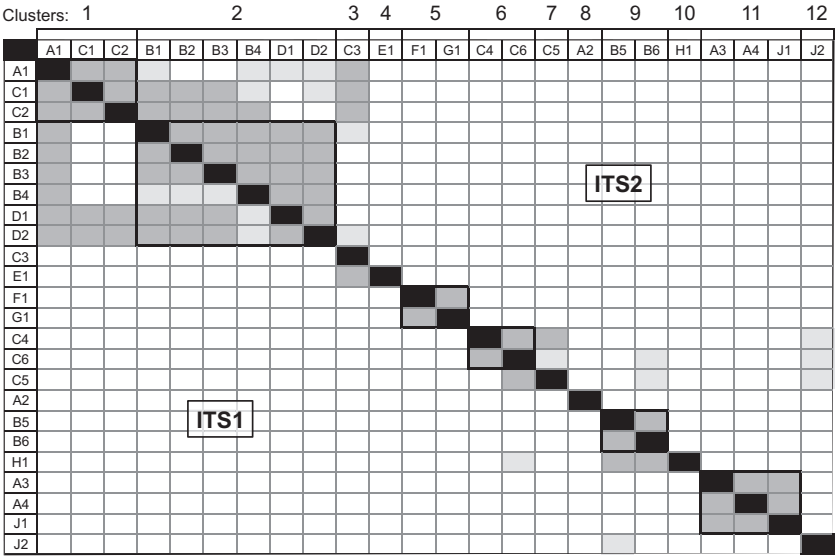

Fig. 5. Pearson correlation matrices for 24 human isolates (A1-J2) based on pyroprints of ITS1 and ITS2. Dark gray shading represents pairwise correlation ≥0.995. Light gray shading represents pairwise correlations between 0.99 and 0.995. Absence of shading represents pairwise correlations <0.99. Isolates were clustered into 12 strains around the central diagonal as shown by black shading. Clusters containing more than one isolate are identified in boxes (clusters 1, 2, 5, 6, 9, and 11).

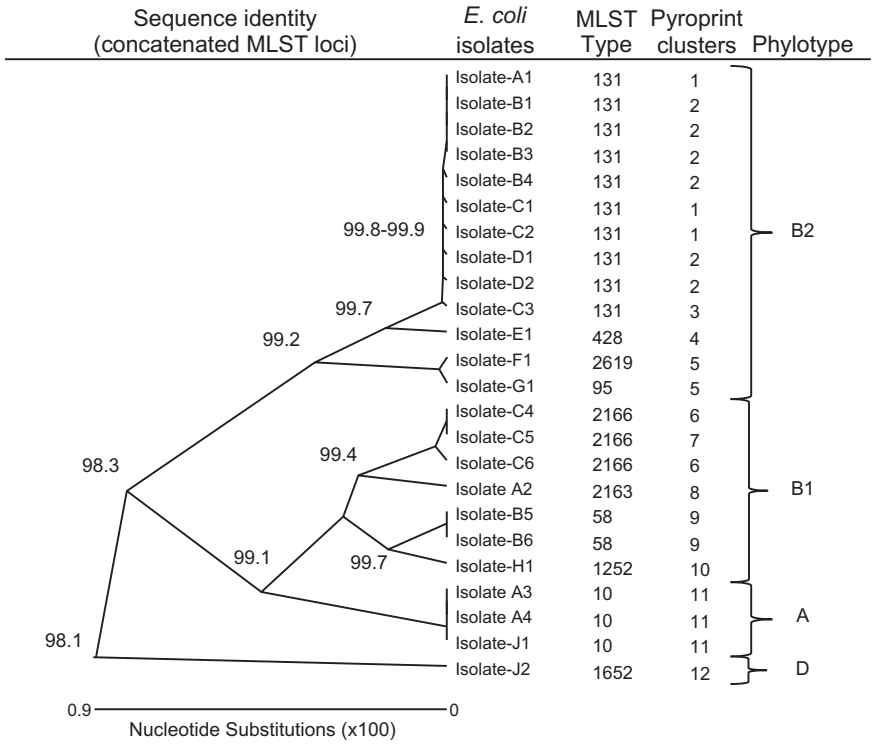

Fig. 6. Phylogenetic relevance of pyroprint clusters. Comparison of the pyroprint clusters of the 24 human-derived *E. coli* strains (Fig. 5) to a phylogram constructed from MLST sequences, the relevant sequence type generated by MLST, and the phylotype of each isolate (A, B1, B2, and D). Numbers at the branch points on the phylogram show percent identity of the aligned sequences.

The pyrosequencing runtime is under 2 hours for 24 samples. Because pyroprinting requires no DNA extraction or purification before PCR, pyroprint data from both ITS loci can be obtained for a set of 36 isolates by one technician in a single day using a single Pyromark Q24 machine. Qiagen's Pyromark Q96 (a 96-well version of the machine) could increase the rate to 144 isolates per day. The reagent costs for pyroprinting, from PCR to pyrosequencing, currently adds up to less than $4 per pyroprint. Pyroprint data is automatically digitized and partially quality controlled by the Pyromark Q24 software.

4. Discussion

Many genotyping methods already exist for differentiating bacteria at subspecies level (reviewed in Wolska and Szwe, 2012). While some, such as RAPD, BOX or ERIC, are rapid and inexpensive, reproducibility and ease of digitizing data are suboptimal and the results may not be phylogenetically relevant. Others, such as ribotyping, AFLP, PFGE and MLST, can be highly discriminating and reproducible, but are not cost or time effective for use on large numbers of isolates. In contrast, pyroprinting meets most of the van Belkum criteria for good typing methods (Van Belkum et al., 2007). Primers used for the PCR step rely on conserved ribosomal gene sequences that ensure all isolates can be typed and the method can be easily extended to include broad categories of bacteria with multiple rRNA gene copies. In addition, the sequence-based nature of pyroprinting uniquely allows for estimations of false positive rates, which will improve as more genome sequences become available.

Pyroprinting is reproducible. Unlike many existing genotyping methods, pyroprinting enjoys the benefit of internal quality control parameters. The pyroprinting protocol described above provides limits for peak height data and duplicate dispensations at the start and finish to ensure accurate results. In addition, the PyroMark Q24 software provides flags for common errors in pyrosequencing. Analysis of replicate pyroprints indicated a high level of confidence (90–95%) in the correlation threshold of 0.99 for matching pyroprints. Most isolates tested (89%) produced triplicate matching pyroprints at this threshold and only 1% of isolates could not produce any matching triplicate pyroprints. While a lower matching threshold could reduce the rate of false negatives, 0.99 was chosen as it ensured an estimated false positive rate of less than 3%. For most applications a false positive result is more problematic than a false negative, so the 0.99 threshold for matching makes sense.

Pyroprinting is highly discriminating. Although each region sequenced for pyroprinting is small (on average about 70 nucleotides) slight changes in allele sequences or ratios can produce a disproportionately large difference in the final pyroprints. Recall from Fig. 1 that polymerases moving along different templates can fall out of phase with one another following a single nucleotide polymorphism between two alleles. This phase effect, combined with the nucleotide dispensation order, can result in a "rippling" of peak height differences between isolates, beyond the expected difference at a single peak height in the pyroprint (Figs. 3 and 4). Including pyroprints from different target regions can selectively enhance the discriminating power of the clustering algorithm. By using both ITS1 and ITS2 pyroprints, discrimination between the 24 human *E. coli* isolates was increased by 25% compared to ITS1 alone or 17% compared to ITS2 (Fig. 5). Pyroprinting with two loci resulted in approximately the same level of discrimination as sequence typing using the MLST database.

Pyroprints are phylogenetically relevant. Pyroprinting was used to cluster 24 *E. coli* isolates into 12 groups that were consistent with strain typing by phylogroup analysis. When compared to MLST analysis, 92% of the isolates were similarly clustered by pyroprinting (Fig. 6). In addition, all pyroprint clusters corresponded to groups with >99.8% sequence identity in the MLST gene sequences. Thus, although pyroprinting does not produce sequence data, Pearson's correlation of pyroprints appears to conform to relationships derived from comparison of DNA sequences at loci distant to the ITS region. This preliminary result indicates that the concordance of sequence based phylogeny and pyroprint clustering could prove interesting for analysis of large collections of *E. coli*, resulting in rapid estimates of phylogenetic relatedness among strains.

In conclusion, pyroprinting is a rapid, inexpensive, reproducible, automatically digitized, highly discriminating and phylogenetically relevant method for typing bacterial isolates. The authors have developed pyroprinting as a tool for creating a microbial source tracking (MST) library of *E. coli* to aid in discovering the origins of environmental contamination by fecal matter (Scott et al., 2002; Stoeckel and Harwood, 2007). The characteristics of low cost, high reproducibility and ease of digitizing are essential for this application. Pyroprinting could reduce the cost of a library-based MST investigation by over 60%. Studies are currently underway to build a database containing pyroprints of *E. coli* obtained from a wide range of hosts and environments. Pyroprinting is also ideal for collecting data on interspecies transmission of *E. coli* strains. Validation of transmission models requires the evaluation of reasonably large host populations and multiple isolates per individual host; thus a rapid and inexpensive method is essential.

Acknowledgments

This work was supported by the W. M. Keck Foundation, the California State University Program for Education and Research in Biotechnology, and the National Science Foundation (#DUE-1140828). We would like to thank Alice Hamrick for support in preliminary reproducibility studies and Ryan Mitchell for contributing to the phylogroup assays.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mimet.2014.07.019.

References

Anderson, M.A., Whitlock, J.E., Harwood, V.J., 2006. Diversity and distribution of *Escherichia coli* genotypes and antibiotic resistance phenotypes in feces of humans, cattle, and horses. Appl. Environ. Microbiol. 72, 6914–6922. http://dx.doi.org/10.1128/AEM.01029-06.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res. 41, D36–D42. http://dx.doi.org/10.1093/nar/gks1195.

Brandt, D., Montana, A., Somers, B., Dekhtyar, A., Lupo, C., Black, M., Goodman, A., Kitts, C., 2012. Pyroprinting sensitivity analysis on the GPU. 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). Presented at the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 951–953 http://dx.doi.org/10.1109/BIBMW.2012.6470279.

Brosius, J., Dull, T.J., Sleeter, D.D., Noller, H.F., 1981. Gene organization and primary structure of a ribosomal RNA operon from *Escherichia coli*. J. Mol. Biol. 148, 107–127. http://dx.doi.org/10.1016/0022-2836(81)90508-8.

Carlos, C., Pires, M.M., Stoppe, N.C., Hachich, E.M., Sato, M.I., Gomes, T.A., Amaral, L.A., Ottoboni, L.M., 2010. *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. BMC Microbiol. 10, 161. http://dx.doi.org/10.1186/1471-2180-10-161.

Clermont, O., Bonacorsi, S., Bingen, E., 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl. Environ. Microbiol. 66, 4555–4558.

Clermont, O., Christenson, J.K., Denamur, E., Gordon, D.M., 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ. Microbiol. Rep. 5, 58–65. http://dx.doi.org/10.1111/1758-2229.12019.

Gordon, D.M., 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. Microbiology 149, 3575–3586. http://dx.doi.org/10.1099/mic.0.26486-0.

Gordon, D.M., 2010. Strain typing and the ecological structure of *Escherichia coli*. J. AOAC Int. 93, 974–984.

Gordon, D.M., Clermont, O., Tolley, H., Denamur, E., 2008. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. Environ. Microbiol. 10, 2484–2496. http://dx.doi.org/10.1111/j.1462-2920.2008.01669.x.

Jensen, M.A., Webster, J.A., Straus, N., 1993. Rapid identification of bacteria on the basis of polymerase chain reaction-amplified ribosomal DNA spacer polymorphisms. Appl. Environ. Microbiol. 59, 945–952.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948. http://dx.doi.org/10.1093/bioinformatics/btm404.

Li, W., Raoult, D., Fournier, P.-E., 2009. Bacterial strain typing in the genomic era. FEMS Microbiol. Rev. 33, 892–916. http://dx.doi.org/10.1111/j.1574-6976.2009.00182.x.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. 95, 3140–3145.

Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N.N., Kyrpides, N.C., 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 40, D115–D122. http://dx.doi.org/10.1093/nar/gkr1044.

Mashayekhi, F., Ronaghi, M., 2007. Analysis of read length limiting factors in pyrosequencing chemistry. Anal. Biochem. 363, 275–287. http://dx.doi.org/10.1016/j.ab.2007.02.002.

McLellan, S.L., Daniels, A.D., Salmore, A.K., 2003. Genetic characterization of *Escherichia coli* populations from host sources of fecal pollution by using DNA fingerprinting. Appl. Environ. Microbiol. 69, 2587–2594. http://dx.doi.org/10.1128/AEM.69.5.2587-2594.2003.

Montana, A., 2013. Algorithms for Library-based Microbial Source Tracking. Masters Theses Proj. Rep.

Montana, A., Dekhtyar, A., Neal, E., Black, M., Kitts, C., 2011. Chronology-sensitive hierarchical clustering of pyrosequenced DNA samples of *E. coli*: a case study. 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Presented at the 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 155–159 http://dx.doi.org/10.1109/BIBM.2011.99.

Ronaghi, M., 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. 11, 3–11. http://dx.doi.org/10.1101/gr.150601.

Ronaghi, M., Uhlén, M., Nyrén, P., 1998. A sequencing method based on real-time pyrophosphate. Science 281 (363), 365.

Scott, T.M., Rose, J.B., Jenkins, T.M., Farrah, S.R., Lukasik, J., 2002. Microbial source tracking: current methodology and future directions. Appl. Environ. Microbiol. 68, 5796–5803. http://dx.doi.org/10.1128/AEM.68.12.5796-5803.2002.

Stoeckel, D.M., Harwood, V.J., 2007. Performance, design, and analysis in microbial source tracking studies. Appl. Environ. Microbiol. 73, 2405–2415. http://dx.doi.org/10.1128/AEM.02473-06.

Tenover, F.C., Arbeit, R.D., Goering, R.V., 1997. How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists Molecular Typing Working Group of the Society for Healthcare Epidemiology of America. Infect. Control Hosp. Epidemiol. Off. J. Soc. Hosp. Epidemiol. Am. 18, 426–439.

Van Belkum, A., Tassios, P.T., Dijkshoorn, L., Haeggman, S., Cookson, B., Fry, N.K., Fussing, V., Green, J., Feil, E., Gerner-Smidt, P., 2007. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin. Microbiol. Infect. 13, 1–46.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C.J., Ochman, H., Achtman, M., 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. 60, 1136–1151. http://dx.doi.org/10.1111/j.1365-2958.2006.05172.x.

Wolska, K., Szwe, P., 2012. Genotyping techniques for determining the diversity of microorganisms. In: Caliskan, M. (Ed.), Genetic Diversity in Microorganisms. InTech.