2007

Interrogating genomic diversity of *E. coli* O157:H7 using DNA tiling arrays

Scott A. Jackson
U.S. Food and Drug Administration

Mark K. Mammel
U.S. Food and Drug Administration

Isha R. Patel
U.S. Food and Drug Administration

Tammy Mays
U.S. Food and Drug Administration

Thomas J. Albert
NimbleGen Systems Inc.

See next page for additional authors

Authors
Scott A. Jackson, Mark K. Mammel, Isha R. Patel, Tammy Mays, Thomas J. Albert, J. Eugene LeClerc, and Thomas A. Cebula

Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International 168 (2007) 183–199
Forensic Science International
www.elsevier.com/locate/forsciint

Interrogating genomic diversity of *E. coli* O157:H7 using DNA tiling arrays

Scott A. Jackson [a], Mark K. Mammel [a], Isha R. Patel [a], Tammy Mays [a], Thomas J. Albert [b], J. Eugene LeClerc [a], Thomas A. Cebula [a,*]

[a] *Division of Molecular Biology, Office of Applied Research and Safety Assessment, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, Laurel, MD 20708, USA*
[b] *NimbleGen Systems Inc., Madison, WI 53711, USA*

Received 24 April 2006; received in revised form 20 June 2006; accepted 21 June 2006
Available online 28 August 2006

Abstract

Here, we describe a novel microarray-based approach for investigating the genomic diversity of *Escherichia coli* O157:H7 in a semi-high throughput manner using a high density, oligonucleotide-based microarray. This microarray, designed to detect polymorphisms at each of 60,000 base-pair (bp) positions within an *E. coli* genome, is composed of overlapping 29-mer oligonucleotides specific for 60 equally spaced, 1000-bp loci of the *E. coli* O157:H7 strain EDL933 chromosome. By use of a novel 12-well microarray that permitted the simultaneous investigation of 12 strains, the genomes of 44 individual isolates of *E. coli* O157:H7 were interrogated. These analyses revealed more than 150 single nucleotide polymorphisms (SNPs) and several deletions and amplifications in the test strains. Pyrosequencing was used to confirm the usefulness of the novel SNPs by determining their allelic frequency among a collection of diverse isolates of *E. coli* O157:H7. The tiling DNA microarray system would be useful for the tracking and identification of individual strains of *E. coli* O157:H7 needed for forensic investigations.

Keywords: Tiling DNA microarray; *Escherichia coli* O157:H7; Single nucleotide polymorphism

1. Introduction

Typical foodborne pathogens might be used as bioterrorist agents in an attack on the food supply. Reliable and valid methods are therefore needed for the tracking and identification of individual strains of foodborne pathogens, a requirement for the forensic investigation of such an attack [1–3]. Properties of *Escherichia coli* O157:H7, i.e., its low infectious dose, ease of transmissibility, and its known incidence in non-intentional outbreaks of human disease, make this pathogen a potential formidable biothreat agent. First recognized as a human pathogen in 1982 [4], *E. coli* O157:H7 has developed into a major enteric pathogen, capable of causing large outbreaks of gastrointestinal disease. The primary clinical manifestation of an *E. coli* O157:H7 infection is hemorrhagic colitis, which may progress into hemolytic-uremic syndrome [5]. An estimated 75,000 cases of *E. coli* O157:H7 infections occur annually, making it the principal serotype of enterohemorrhagic *E. coli* isolated from patients in the United States [6]. Infections with this serotype usually have a foodborne etiology [7]. For instance, CDC data from 1982 to 1996 indicated that about two-thirds of 3000 cases of illness in 139 outbreaks were associated with contaminated food sources, whereas 22% of cases were from person-to-person transmission (usually child care centers) and 10% were from recreational and drinking water [8].

The population genetics of extant *E. coli* O157:H7 strains circumscribed the inheritance and roles of phage-, plasmid-, and chromosomal-encoded genes in the virulence of *E. coli* O157:H7. Early multi-locus enzyme electrophoresis (MLEE) analyses suggested a clonal inheritance for O157:H7 [9] and subsequent multilocus sequence typing (MLST) of selected housekeeping genes indeed showed a marked similarity among *E. coli* O157:H7 strains that were distinguishable by

Abbreviations: SNP, single-nucleotide polymorphism; MNP, multi-nucleotide polymorphism; CNP, copy number polymorphism

* Corresponding author. Tel.: +1 301 210 6158; fax: +1 301 210 6093.
E-mail address: Thomas.Cebula@fda.hhs.gov (T.A. Cebula).

0379-0738/$ – see front matter
doi:10.1016/j.forsciint.2006.06.079

pulsed-field gel electrophoresis (PFGE) patterns [10]. A much richer genetic diversity within *E. coli* O157:H7 has been depicted by a variety of other molecular analyses, nonetheless, including octamer-based genome scanning [11], extensions of pulsed-field gel electrophoresis (PFGE) [12], whole genome PCR scanning [13], and comparative genomic microarray [14]. The inherent plasticity of the *E. coli* O157:H7 genome coupled with the transiting of prophages would suggest that other, yet unidentified, loci have likely contributed to the ultimate pathogenesis of this organism.

The complete genome sequences of two independent outbreak strains of *E. coli* O157:H7 have been determined [15,16]. An in silico comparison of these sequences reveals numerous single nucleotide polymorphisms and multi-base insertions and deletions that distinguish these two *E. coli* O157:H7 strains. The multiplicity of unique sites within these two isolates suggests that it might be possible to discriminate individual strains of O157:H7 without extensive genome sequencing. Not a trivial task, however, is discerning the locations of these novel polymorphic regions using conventional molecular biological techniques. In order to probe the genomic diversity of different strains of *E. coli* O157:H7, we investigated the use of a novel DNA tiling array that allows the interrogation of approximately 60 kb of genomic DNA from individual strains, at single base-pair resolution in a semi-high throughput fashion. Herein, we report the utility of this tiling array in discriminating between and among individual isolates of *E. coli* O157:H7.

2. Materials and methods

2.1. Bacteria

In order to assess the genetic diversity within extant populations of *E. coli* O157:H7, 44 independent isolates were assessed by a variety of microbiological and molecular criteria. To ensure an adequate sampling of diversity, isolates were selected that differed in geographical, temporal, and source origins (Table 1). Strains EDL933 and MG1655 are *E. coli* O157:H7 and K-12 strains, respectively, for which the genome sequences are available [15,17]. O157:H7 Sakai is a Japanese outbreak strain, and likewise the genome sequence is available [16]. Microbiological strain validation assays were performed on either Luria broth supplemented with 2.5 μg/ml potassium tellurite or on sorbitol MacConkey agar (SMAC, Difco) supplemented with 50 μg/ml 4-methylumbelliferyl-β-D-glucuronide (MUG). For further characterization of each strain, MAMA multiplex PCR [18] was used to determine the presence of genes associated with *E. coli* O157:H7.

2.2. DNA tiling arrays

All microarrays used here were manufactured by NimbleGen Systems Inc. (Madison, WI) using a maskless array synthesis (MAS) technology for in situ synthesis of DNA oligonucleotides directly onto glass microscope slides [19–21]. The tiling arrays were designed by dividing the genome of strain EDL933 into 60 equally spaced segments (Fig. 2). 29-mer oligonucleotides were designed to tile across a contiguous 1000-bp region of each segment with an overlap of 24 bases. Ambiguous sequence (e.g. N, R, Y, etc.) in the published genome sequence of EDL933 were not incorporated into the oligonucleotide probes present on the array, but are represented as single nucleotide deletions (that are not apparent in the resulting hybridization profiles).

2.3. DNA isolation and labeling

Strains were grown overnight in Luria broth and genomic DNA was isolated using the Qiagen DNeasy kit following the manufacturer's recommendations. Genomic DNA (5 μg) was fragmented to an average size of 100 bp by treatment with 0.1 Units of DNase I for 10 min at 37 °C in 1× One-Phor-All buffer (Amersham Biosciences, Piscataway, NJ). Heat inactivation of the DNase I enzyme was performed at 95° C for 15 min. Fragmented DNA was treated with 60 Units of terminal transferase (Promega, Madison, WI) in the presence of 40 μM biotin-N6-ddATP (Perkin-Elmer, Wellesley, MA) at 37 °C for 90 min. Inactivation of terminal transferase was effected by incubating at 95 °C for 15 min.

2.4. Microarray hybridization, post-hybridization, and staining

Tiling arrays were hybridized with 100 ng of biotinylated DNA in the presence of NimbleGen Hybridization Buffer in a total volume of 6 μl. Before application to the array, the samples were heated to 95 °C for 5 min, cooled to 45 °C, and centrifuged for 5 min at 12,000 × *g*. After application of DNA, arrays were placed in a NimbleGen 12-well hybridization apparatus and incubated at 45 °C for 14–16 h in an incubator oven. Arrays were washed with 5 ml of stringent wash buffer [100 MES, 0.1 M NaCl, 0.01% (v/v) Tween 20] at 47.5 °C, rinsed briefly in non-stringent wash buffer [6× SSPE, 0.01% (v/v) Tween-20], followed by two 5-min washes in stringent wash buffer at 47.5 °C. Afterwards, arrays were returned to non-stringent wash buffer. The arrays were stained with a solution containing Cy3–streptavidin conjugate (Amersham Biosciences, Piscataway, NJ) for 10 min, and washed with non-stringent wash buffer. The Cy3 signal was amplified by secondary labeling of the DNA with biotinylated goat anti-streptavidin (Vector Laboratories, Burlingame, CA). Subsequently a non-stringent wash was used to remove the secondary antibody, and the array was restained with a Cy3–streptavidin solution. The stain solution was removed, and the array was washed in non-stringent wash buffer. The arrays were washed for 15 s in 0.2× SSC, followed by a wash in 0.05× SSC for 15 s, and then immediately dried by centrifugation.

2.5. Data extraction and analysis

Hybridized microarrays were scanned using an Axon GenePix® 4200A scanner at 5 μm resolution using the 532 nm laser. Fluorescent intensities of each feature were extracted utilizing NimbleScan™ software (NimbleGen Systems Inc.), and all subsequent data analyses were performed using MS Excel. Ratios of probe intensities of the reference strain (EDL933) and the probe intensities of each test strain were determined, and the ratios were plotted against genome position for each probe, thus yielding a hybridization profile for each strain.

2.6. Conventional sequencing

A PCR reaction was performed by adding 5 μl DNA template, 1.3 μl each of forward and reverse primers at 10 mM (Table 3A), 5 μl 10× PCR buffer with 1.5 mM $MgCl_2$ (Perkin-Elmer), 5 μl of 2.5 mM dNTPs (Pharmacia), and 1.5 U Taq DNA polymerase (Promega) in a 50 μl reaction. Amplification was performed in a PTC-200 thermal cycler (MJ Research) under the following conditions: initial denaturation at 94 °C for 5 min, 94 °C for 1 min, 53 °C for 1 min, and 72 °C for 1.5 min (35 cycles); final incubation at 72 °C for 10 min. Products from PCR amplification were purified and concentrated using Qiaquick spin columns (Qiagen, Valencia, CA). Nucleotide cycle sequencing was performed in both directions directly on purified PCR templates via automated Sanger dideoxy-chain termination methods (Amplicon Express, Pullman, WA).

2.7. Pyrosequencing

Twenty microliters of biotinylated PCR product was immobilized onto 3 μl streptavidin-coated Sepharose beads (Amersham Biosciences, Uppsala,

Table 1
Strains used in this study

FDA designation	Sender designation	Source	PCR-based assay					Microbiological plating assay		
			uidA[a]	*stx1*	*stx2*	*eaeA*	*ehxA*	Tellurite resistance	SMAC	MUG
93-111	93-111	Human, WA, 1995; P. Tarr	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
AB1	AB1	A. Benson	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
AB2	AB2	A. Benson	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
AB3	AB3	A. Benson	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
AB4	AB4	A. Benson	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
AB5	AB5	A. Benson	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
AB6	AB6	A. Benson	(+)	(−)	(−)	(+)	(+)	(+)	(−)	(−)
AB8	AB8	A. Benson	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC423	260	P. Feng	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC486	ATCC-43890	Human feces; ATCC	(+)	(+)	(−)	(+)	(+)	(+)	(−)	(−)
EC488	ATCC-43895	Hamburger, MI, 1982; ATCC	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC502	EC121	P. Feng	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC504	ATCC-43894	Human feces; P. Feng	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC505	ATCC-43895	Hamburger, MI, 1982; P. Feng	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC506	ATCC-43888	Human feces; P. Feng	(+)	(−)	(−)	(+)	(+)	(−)	(−)	(−)
EC507	ATCC-35150	Human feces; P. Feng	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC508	ATCC-43889	Human feces; P. Feng	(+)	(−)	(+)	(+)	(+)	(+)	(−)	(−)
EC510	4936i	Human; P. Feng	(+)	(−)	(+)	(+)	(+)	(−)	(+)	(+)
EC516	EC269	Human; P. Feng	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC535	86-01	Human, WA, 1986; P. Feng	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC536	86-17	Human, WA, 1986; P. Feng	(+)	(−)	(+)	(+)	(+)	(+)	(−)	(−)
EC552	491	Human; P. Feng	(+)	(−)	(+)	(+)	(+)	(+)	(−)	(−)
EC866	FDASea13B88	R. Buchanan	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(+)
EC868	USDA-FSIS-45750	R. Buchanan	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC869	USDA-FSIS-45753-32	R. Buchanan	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC870	CDC933	Hamburger, MI, 1982; R. Buchanan	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC871	A9218-C1	Human, AK, 1983; R. Buchanan	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC873	B6-914	Human feces; R. Buchanan	(+)	(−)	(−)	(+)	(+)	(+)	(−)	(+)
EC874	C7927	Apple cider; R. Buchanan	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC875	C984	Human feces; R. Buchanan	(+)	(+)	(−)	(+)	(+)	(+)	(−)	(−)
EC877	ENTC9490	Hamburger, 1993; R. Buchanan	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC878	F12	R. Buchanan	(+)	(−)	(+)	(+)	(+)	(+)	(−)	(−)
EC881	86-24 NaIR (psk+)	R. Buchanan	(+)	(−)	(+)	(+)	(+)	(+)	(−)	(−)
EC882	F12(PRFBE)	R. Buchanan	(+)	(−)	(+)	(+)	(+)	(+)	(−)	(−)
EC887	HB101	R. Buchanan	(+)	(−)	(−)	(−)	(+)	(+)	(−)	(−)
EC1214	CAI	Human, 2002; C. Park	(+)	(−)	(−)	(+)	(+)	(+)	(−)	(−)
EC1215	DIRKA	Human, 2000; C. Park	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC1219	CAN12	Human, Canada; C. Park	(+)	(+)	(+)	(+)	(+)	(−)	(−)	(−)
EC1220	CAN28	Human, Canada; C. Park	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC1221	CAM 10	Human, Canada; C. Park	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC1227	86-24	Human, WA, 1986; T. Whittam	(+)	(−)	(+)	(+)	(+)	(+)	(−)	(−)
EC1228	2886-75	Human, WA, 1993; T. Whittam	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(−)
EC1231	G5101	Human, CA, 1975; T. Whittam	(+)	(+)	(+)	(+)	(+)	(+)	(−)	(+)
EC1242	CVM42	Human, GA 1992; D. White	(+)	(−)	(+)	(+)	(+)	(+)	(−)	(−)

In order to verify the identity of each strain, MAMA multiplex PCR [18] was used to determine the presence of genes associated with *E. coli* O157:H7. Microbiological plating assays were performed to further validate these strains (see Section 2).

[a] uidA positive (+) results indicate that the strain carries a G residue at uidA position 92, characteristic of *E. coli* O157:H7 [18].

Sweden) in 40 μl binding buffer, pH 7.6 (10 mM Tris–HCl, 2 M NaCl, 1 mM EDTA, 0.1% Tween 20) in a 96-well plate. The plate was incubated at room temperature for 10 min with shaking (900 rpm) to keep the beads dispersed. Beads were harvested with a vacuum prep tool (Biotage AB, Uppsala Sweden) and immersed for 5 s each in 70% ethanol, 0.2 M NaOH, and washing buffer, pH 7.6 (1 mM Tris–Acetate). They were then dispensed into a 96-well plate containing 4 μl 10 mM primers (Table 3B) in 40 μl annealing buffer, pH 7.6 (20 mM Tris, 2 mM Magnesium acetate-tetrahydrate). The plate was heated on an 80 °C heat block for 2 min for annealing. Pyrosequencing was performed on an automated PSQ96MA instrument using the PSQ 96 SNP reagent kit (Biotage AB, Uppsala, Sweden) according to the manufacturer's instructions. Pyrosequencing in the SNP mode and SQA mode was carried out according to the instructions of the manufacturer. For sequencing a pool of DNA from a set of 100 *E. coli* O157:H7 strains, the concentration of each DNA sample was determined and an equimolar amount of each sample was added to the pool. Allele frequencies were determined from relative peak heights of the Pyrogram™ resulting from the pyrosequencing reactions.

3. Results

3.1. DNA tiling arrays

Several molecular techniques have been utilized to investigate the genomic diversity of *E. coli* O157:H7, although none has provided sufficient precision to assess genome-wide diversity with single base-pair resolution. We set out to sample a portion of the O157:H7 genome from many strains in order to test for the presence of single-nucleotide polymorphisms (SNPs) as well as indels that occur throughout the genome. To accomplish this goal, a novel semi-high throughput array platform was used that permitted the interrogation of relatively large regions of a bacterial genome for mapping SNPs with near single base-pair exactness. The array was designed to identify the locations of SNPs, multi-nucleotide polymorphisms (MNPs), copy number polymorphisms (CNPs), and deletions in target genomic sequences. Oligonucleotide probes were 29-mers that overlapped in sequence by 24 bases each. Selected regions of the sequenced genome of *E. coli* O157:H7 strain EDL933 were tiled at five base-pair (bp) spacing, allowing the identification of genomic anomalies that exist within these regions of a test strain relative to the EDL933 strain (Fig. 1). Like a resequencing chip, hybridization conditions for the O157:H7 chip were optimized such that single base-pair mismatches within a test genome could be detected by a decrease in hybridization signal (i.e., fluorescence) relative to the reference genome. The design of the chip is well suited for SNP discovery since the sequence of each probe is partially redundant with its neighboring probe by 24 bases; therefore, each mismatch present in the test genome would be signaled by multiple oligonucleotide probes. As a result, a true SNP might be called with confidence since it would be signaled by disruption in the hybridization of three to five contiguous oligonucleotide probes (Fig. 1). Absent these criteria determined by visual inspection of array profiles, 'false' SNPs (false positives) can be readily discriminated and eliminated. Furthermore, MNPs and CNPs can be detected by their unique hybridization signatures, which become visible by plotting the hybridization intensities of the tiled probes that span the polymorphic region.

As a first approach for assessing the genomic diversity among *E. coli* O157:H7 strains, sequences of genetic loci were chosen from evenly spaced intervals of the O157:H7 genome to develop the tiling array. Such a "random" approach assumed no prior knowledge of diversity at any locus being examined, thus precluding biases about the evolution of the *E. coli* O157:H7 genome. This stochastic approach also had the potential for discovering novel variable regions within the genome, which might serve as useful biomarkers for strain identification. Thus, a tiling array was designed by dividing the genome of EDL933 into 60 equally spaced segments (Fig. 2). From each of these segments, a contiguous 1000 bp region was selected, and oligonucleotides were designed to "tile" across each segment. This design provided 1 kb samplings at 60 evenly spaced loci around the chromosome, or about 1% of the genome. A brief description of each of the 60 loci is presented in Table 2, according to the ASAP Database annotation of the EDL933 genome [22]. Note that 1% sampling was chosen to test whether a tiling strategy employing a random sampling of a small portion of a bacterial chromosome could be used to measure diversity within individual strains of *E. coli*. While the tiling strategy might be employed to interrogate as much, or all, of the genome as necessary to distinguish a strain, our choice of 1% allowed the analysis of more strains per chip than other methods.

In order to maximize efficiency, we used a multi-well microarray format that allowed hybridization of genomic DNA from 12 different strains on a single chip. Since the tiling arrays were manufactured using a high-density, photolithographic technique, each well contained on the order of 13,000 oligonucleotide probes. Hence, approximately 60 kb of the genome from each of 12 isolates could be analyzed. As with other resequencing approaches, the design of tiling probes was based on a sequenced, reference genome. Accordingly, in each experiment, labeled DNA from the reference strain was hybridized in 1 of the 12 wells and the hybridization signals compared to those from each of 11 test strains being interrogated on the same chip. The relative signal intensities were plotted against the genome position to yield a hybridization profile for particular regions of the genome in each of the test strains.

3.2. Tiling arrays of sequenced genomes

As a proof-of-principle, genomic DNA isolated from the *E. coli* K-12 strain MG1655 was tested on the O157:H7 tiling

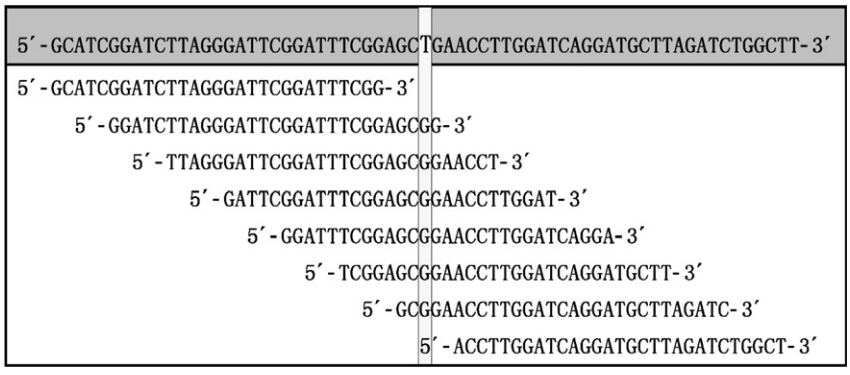

Fig. 1. Tiling array design. Tiling arrays consist of 29-mer oligonucleotide probes whose sequences overlap by 24 nt (5 nt spacing). The sequenced O157:H7 strain EDL933 was used as the reference strain.

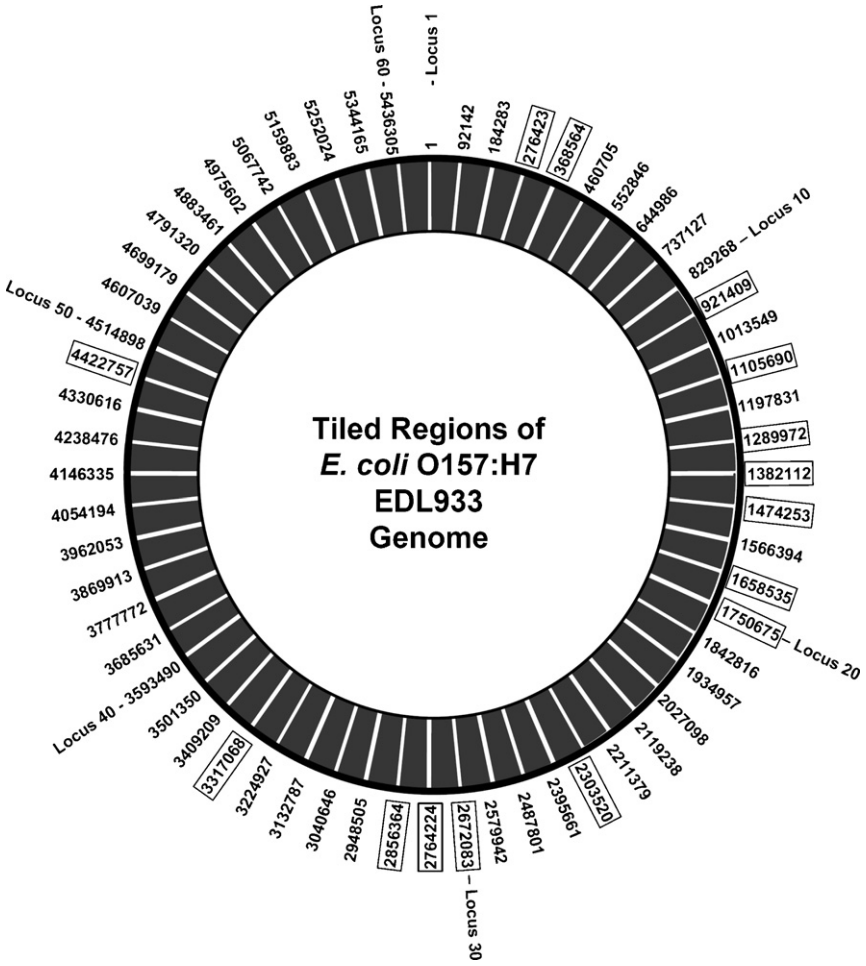

Fig. 2. Tiled regions of the *E. coli* O157:H7 EDL933 genome. A random sampling of ~1% of the EDL933 genome was accomplished by dividing the genome into 60 equally spaced regions and designing probes that are complementary to a contiguous 1000 bp region. The starting positions of each of the 1000 bp loci are shown according to the EDL933 genome coordinates. Loci shown in boxes are contained within O-islands.

array [23]. The complete sequence of the MG1655 genome [17] made possible a comparison of the tiling array data with those generated from an in silico comparison of the sequences from the two strains. This study demonstrated that the tiling array approach detected greater than 99% of the polymorphisms that exist between the two strains [23]. A similar comparison was made using the *E. coli* O157:H7 Sakai strain, for which the genome sequence is also available [16]. Fluorescently labeled genomic DNA isolated from the O157:H7 Sakai strain was hybridized to the O157:H7 tiling array and hybridization intensities, relative to those from EDL933, were plotted with respect to genome position (Fig. 3). An in silico comparison between the Sakai and EDL933 genomes indicated that at least 17 nucleotide differences were present within the 60 kb region of the genome represented on the array. When however the Sakai strain was examined with the EDL933 tiling array, using the criteria that hybridization signals from multiple contiguous oligonucleotides would be affected, these potential SNPs were not observed. Resequencing of these regions revealed that the EDL933 and Sakai genome sequences were identical at each of the 17 positions, demonstrating directly that these particular sites, rather than SNPs, were sequencing errors in either the EDL933 or the Sakai sequence.[1] These data serve to point out not only the need for independent verification of sequence data in the open databases but also the utility of tiling arrays for identifying potential sequencing errors.

The tiling array data were used to identify two loci that have differing copy numbers in the Sakai genome (Fig. 3A). That is, loci 13 and 17, which are present twice in the EDL933 genome and once in the Sakai genome, gave a significantly decreased hybridization signal relative to EDL933. The analysis also revealed that tiling arrays become limited in their ability to detect copy number polymorphisms (CNPs) as the copy number of an allele increases. For example, the DNA sequence

[1] Resequencing of particular regions of the ATCC 700927 genome revealed the following differences from the published sequence of *E. coli* O157:H7 EDL933: position no. 921,739, A → G; 921,754, C → T; 921,815, C → A; 921,816, C → T; 921,844, A → C; 921,883, T → C; 921,915, C → T; 921,987, A → C; 922,008, G → T; 922,026, T → C; 922,028, A → G; 922,038, G → A; 922,051, C → T; 922,060, C → T; 5,437,265, T → A; 5,437,292, T → A. Resequencing of regions of the ATCC BAA-460 genome revealed the following differences from the published sequence of *E. coli* O157:H7 (Sakai): 417,119, G → T.

Table 2
EDL933 loci represented on the tiling arrays used in this study

	EDL933 start position	EDL933 end position	EDL933 copy number	Sakai copy number	Annotation	Comment
1	1	1000	1	1	*thrL-thrA*	
2	92142	93141	1	1	*ilvH-fruL-fruR*	
3	184283	185282	1	1	*dgt-htrA*	
4	276423	277422	1	1	*Z0275*	O-island #7
5	368564	369563	1	1	*Z0390*	O-island #14
6	460705	461704	1	1	*yaiB-phoA*	
7	552846	553845	1	1	*aefA*	
8	644986	645985	1	1	*purE-ybbF*	
9	737127	738126	1	1	*crcA*	
10	829268	830267	1	1	*ybgQ*	
11	921409	922408	1	2	*Z0980*	O-island #36-Cryptic prophage CP-933K
12	1013549	1014548	1	1	*Z1072-Z1073*	
13	1105690	1106689	2	1	*terF-Z1178*	O-island #43
14	1197831	1198830	1	1	*mukB*	
15	1289972	1290971	3	3	*Z1380*	O-island #44
16	1382112	1383111	1	1	*Z1495*	O-island #45
17	1474253	1475252	2	1	*ureD_2-ureA_2*	O-island #48
18	1566394	1567393	1	1	*pyrC-yceB*	
19	1658535	1659534	5	4	*Z1814*	O-island #50
20	1750675	1751674	1	1	*Z1929-Z1930*	O-island #52
21	1842816	1843815	1	1	*cls-kch*	
22	1934957	1935956	1	1	*ydfG-dcp*	
23	2027098	2028097	1	1	*narZ*	
24	2119238	2120237	7	8	*Z2340-Z2342*	
25	2211379	2212378	1	1	*pspF*	
26	2303520	2304519	4	4	*Z6042-Z6043-Z6044*	O-island #71-includes cryptic prophage CP-933P
27	2395661	2396660	1	1	*slyA-Z2658-Z2659*	
28	2487801	2488800	1	1	*Z2759-Z2760-katE*	
29	2579942	2580941	1	1	*manX-manY-manZ*	
30	2672083	2673082	1	1	*Z2973-Z2974*	O-island #76-Cryptic prophage CP-933T
31	2764224	2765223	2	1	*Z3098*	O-island #79
32	2856364	2857363	1	1	*wzy-wbdN*	O-island #84
33	2948505	2949504	1	1	*molR_D-yehI*	
34	3040646	3041645	1	1	*Z3401-yeiA*	
35	3132787	3133786	1	1	*yfaA*	
36	3224927	3225926	1	1	*purF*	
37	3317068	3318067	2	2	*Z3664-gltX*	O-island #103
38	3409209	3410208	1	1	*ppk*	
39	3501350	3502349	1	1	*nadB-yfiC*	
40	3593490	3594489	1	1	*nrdE*	
41	3685631	3686630	1	1	*Z4084*	
42	3777772	3778771	1	1	*yqeH-yqeI*	
43	3869913	3870912	1	1	*ygfH*	
44	3962053	3963052	1	1	*yqhC-yqhD*	
45	4054194	4055193	1	1	*exuR-yqjA*	
46	4146335	4147334	1	1	*yrbA-yrbB-yrbC*	
47	4238476	4239475	1	1	*sun-trkA*	
48	4330616	4331615	1	1	*malP*	
49	4422757	4423756	1	1	*Z4877*	O-island #139
50	4514898	4515897	1	1	*dppB-dppA*	
51	4607039	4608038	1	1	*tdh-kbl*	
52	4699179	4700178	1	1	*Z5154-yicP*	
53	4791320	4792319	1	1	*yieM-yieN*	
54	4883461	4884460	1	1	*yigN-ubiE*	
55	4975602	4976601	1	1	*yiiP-pfkA*	
56	5067742	5068741	1	1	*htrC-thiH*	
57	5159883	5160882	1	1	*yjcC-soxS-soxR*	
58	5252024	5253023	1	1	*aspA*	
59	5344165	5345164	1	1	*pmbA-cybC*	
60	5436305	5437304	1	1	*uxuA*	

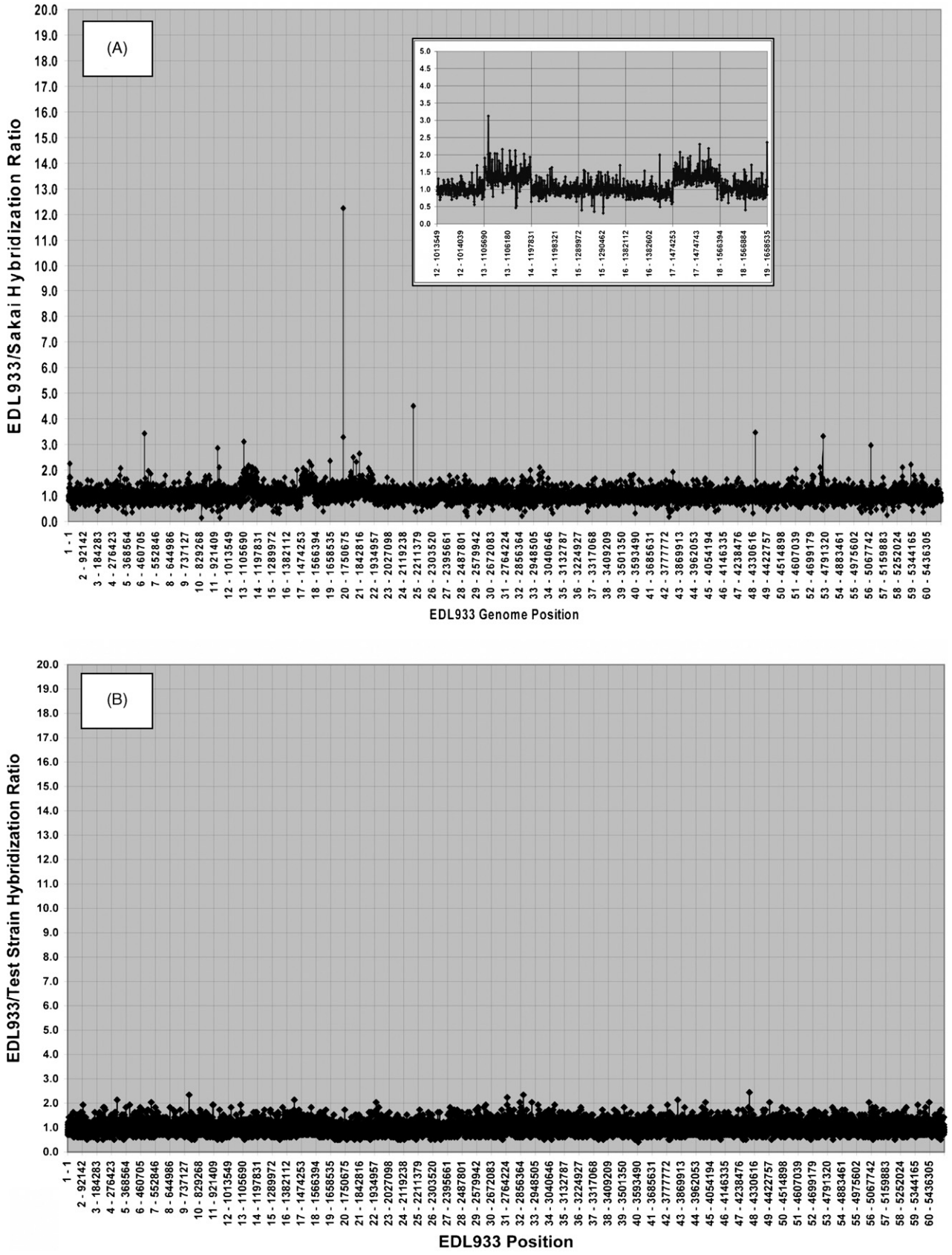

Fig. 3. (A) Comparison of the genomes of two sequenced *E. coli* O157:H7 strains (EDL933 and Sakai). The probe intensity of EDL933 relative to that of Sakai (EDL933/Sakai) is shown with respect to genome position. The inset shows an enlargement of loci 13 and 17 that contain CNPs in the Sakai strain relative to EDL933. (B) Hybridization profile for strain EC488. EC488 is an EDL933 strain acquired from the CDC in 1994.

represented by locus 24 on the tiling array is duplicated seven times in the EDL933 genome and eight times in the Sakai genome. As the magnitude of this copy number change is relatively small (1.14-fold); however, the probe intensity change was indistinguishable over the baseline fluorescence (data not shown).

Notably, when a collection of *E. coli* O157:H7 isolates was analyzed, strain EC488 displayed a hybridization profile indistinguishable from EDL933 (Fig. 3B). Inspection of the strain history of EC488 revealed that this strain was received in 1993 from the Centers for Disease Control and represents an independent isolate of EDL933 (ATCC 43895 Batch 92-01). This result speaks to both the genome stability of the strain in storage and the use of tiling arrays in accurately reflecting similarities (and differences) among strains. As a built-in redundancy in the analysis, EC505 and EC870 were isolates of EDL933 obtained from independent sources (Table 1), and similarly showed an absence of features on the tiling array distinguishing them from the EDL933 reference strain.

3.3. SNP discovery from tiling array data

Genomic DNAs from 44 geographically and temporally varied isolates of pathogenic *E. coli* O157:H7 were analyzed by this tiling strategy. As described in Table 1, each of the strains was assayed for molecular and microbiological characters that typify O157:H7. This was essential in order to verify the identity of *E. coli* O157:H7 strains obtained from diverse sources. For molecular characterization, a multiplex PCR assay was used to examine the strains for the presence of a series of biomarkers most often found associated with this enterohemorrhagic pathogen [18]. The analysis also allowed a partitioning of strains based on the presence of Shiga-like toxin (*stx*) genes and identified strains that were defective in *stx1* (9), *stx2* (2), and *stx1·stx2* (5) among the 44 strains. The phenotypic tests applied to the strains assayed for tellurite resistance, sorbitol auxotrophy, and a MUG-negative phenotype for defective β-glucuronidase. The plating assays identified three tellurite-sensitive, one sorbitol-positive, and four MUG-positive variants in the reference set of strains.

Labeled genomic DNA was hybridized to custom tiling arrays as described in Section 2. The relative hybridization intensity of each probe was plotted with respect to its genome position to generate a hybridization profile for each strain. Putative SNPs were localized by the increased hybridization ratios (EDL933/test strain) signaled by the successive probes that encompassed the SNP site. As an example, a hybridization profile for *E. coli* O157:H7 isolate EC1231 is shown in Fig. 4. Tiling array data from this strain predicted the presence of 37 SNPs as well as several multi-nucleotide polymorphisms

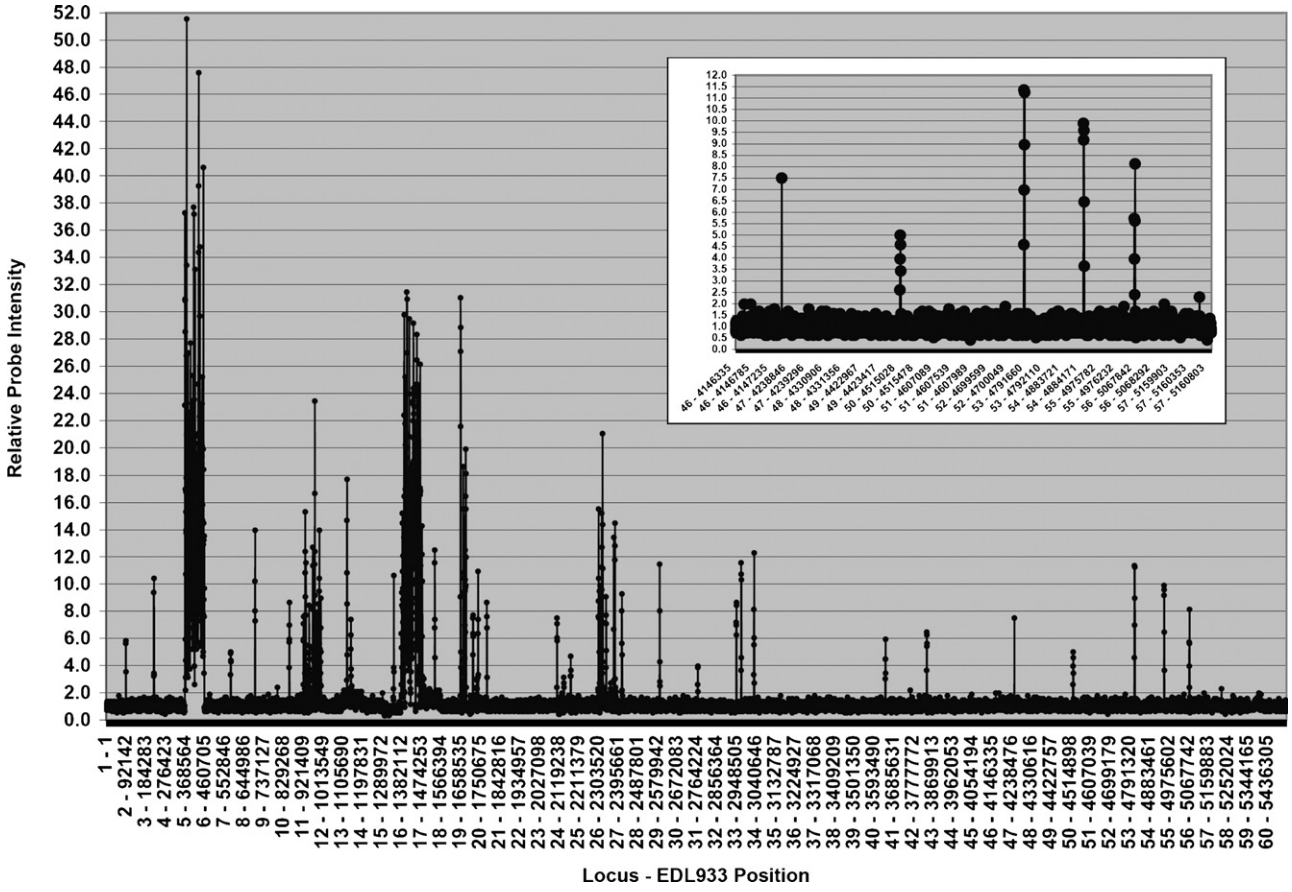

Fig. 4. Hybridization profile from a representative *E. coli* O157:H7 strain. Labeled genomic DNA from strain EC1231 was hybridized to the EDL933 tiling array and relative fluorescent intensities were determined for each probe. The inset shows the enlargement of a locus that contains four SNPs. The single probe located at position 4,238,476 was excluded from further analysis because no neighboring probe was affected.

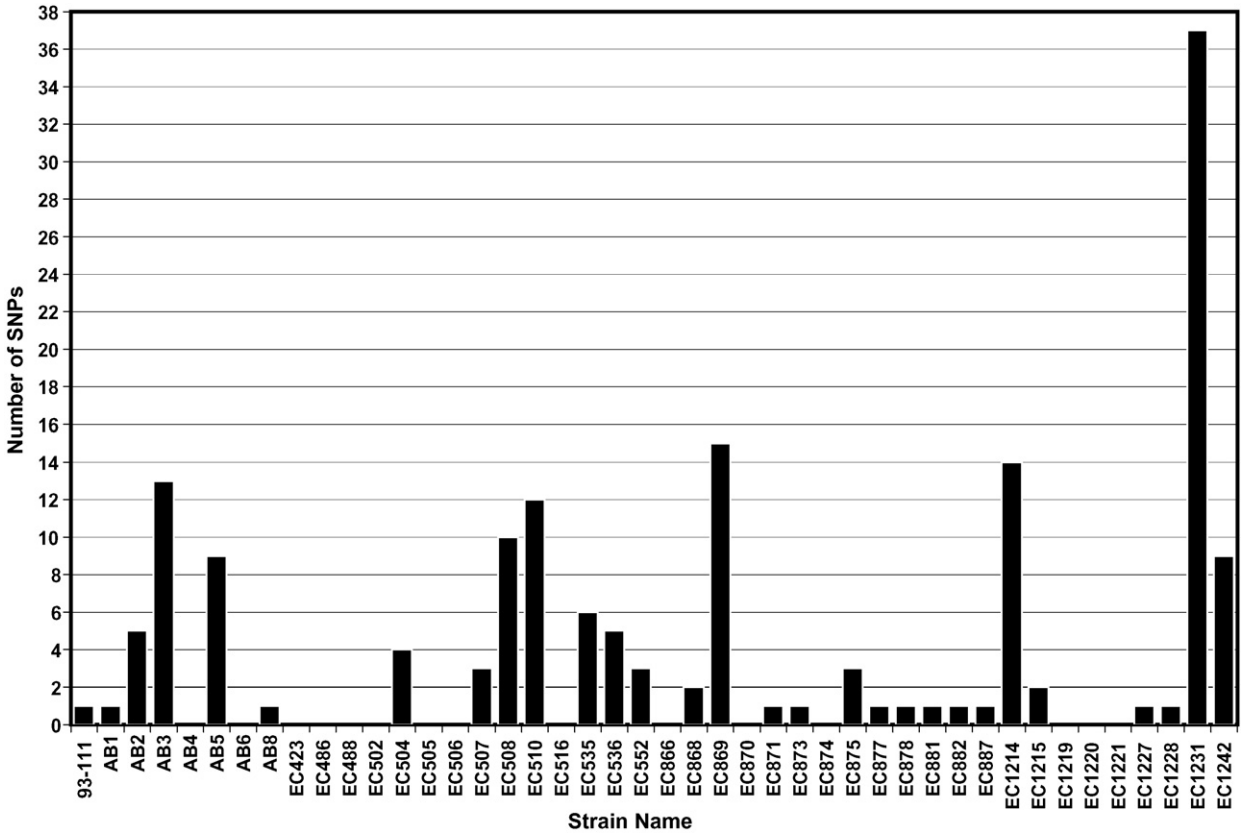

Fig. 5. The number of SNPs found per strain.

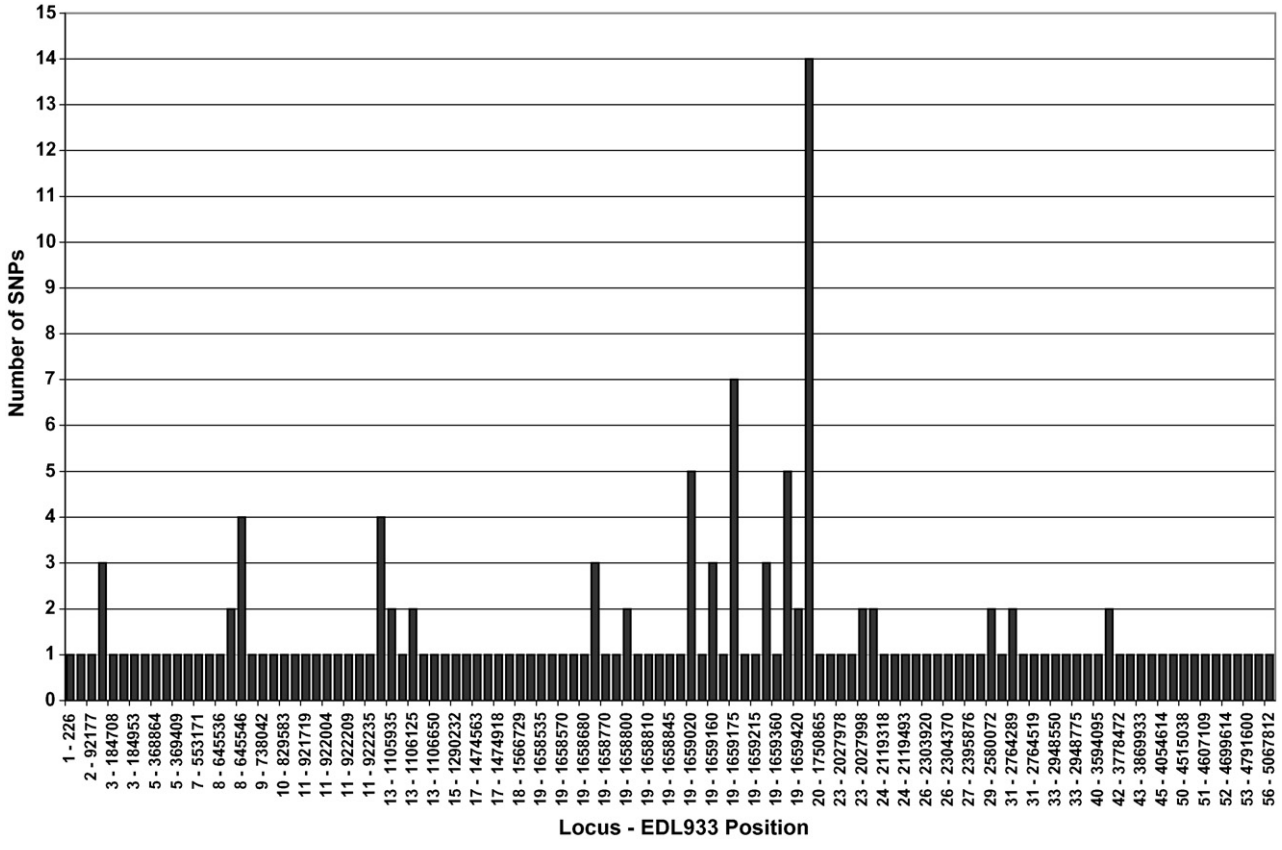

Fig. 6. The frequency and location of SNPs. The bar graph shows the population frequency of each SNP among the 44 strains interrogated.

(MNPs) (described below). The Fig. 4 inset shows a particular portion of the hybridization profile in which four SNPs can be clearly identified. In addition, at another site (located at position 4,238,476, see Fig. 4), a single isolated probe signaled an eight-fold difference from the baseline intensity, though flanking probes for the site did not show differences in hybridization. The latter result emphasizes the relative power of the overlapping probe strategy to distinguish false positives from true SNPs.

Fig. 5 shows the number of SNPs (relative to the EDL933 genome) identified in 44 *E. coli* O157:H7 strains by DNA tiling arrays. The interrogation of 60 kb of DNA in each strain, over the 2.64 Mb of genomes interrogated, resulted in the detection of 164 SNPs. The distribution of SNPs was non-uniform among strains. For example, strain EC1231 contained 37 SNPs, while 15 of the strains (34%) showed no SNP within the 60 kb of DNA analyzed. That strain EC1231, which had the most SNPs of the strains examined, is an atypical MUG^+ variant of O157:H7 (Table 1) is likely significant, but a conclusion is confounded by the observation that the three other MUG^+ strains in the collection, EC510, EC873, and EC866, showed 12, 1, and no SNPs, respectively.

Of the 164 SNPs detected, 113 SNP positions appeared at unique positions in the genome. The distribution of these SNPs with respect to EDL933 genome position showed that here, too, their distribution was non-uniform (Fig. 6). For example, locus 19 (starting bp 1,658,535) contained 61 (37%) of the 164 total SNPs detected. This locus spans gene Z1814 and is contained within O-island #50, which represents the cryptic prophage 933N [12]. Furthermore, this locus is present in multiple copies throughout the EDL933 genome and each of the five alleles differs due to the presence of multiple SNPs. The abundance of SNPs observed at this locus may result from the multiple alleles present in strains of O157:H7.

To test for the presence of SNPs detected by tiling array profiles of O157:H7 strains, sequence analysis was carried out in the indicated strains. Either conventional sequence analysis (Table 3A) or SQA pyrosequencing (Table 3B) was directed to the locus of the predicted SNP(s) as indicated by tiling results. In most cases, base changes were identified within 29 bases from the start position of the 29-mer oligonucleotide probe and were 5–25 bases from the start position of the probe as assayed by conventional sequence analysis (Table 3A) and six to 24 bases from the start position of the probe assayed by SQA pyrosequencing (Table 3B). Though the tiling strategy signals genomic anomalies quite accurately, false positives do occur as evidenced by putative SNPs detected in strains AB2 and EC536 at probe positions 1,106,650 and 1,567,059, respectively, for these could not be verified by sequence analysis (data not shown).

Table 3
Identification of SNPs in *E. coli* O157:H7 strains

Strain	Probe position	Position identified	Nucleotide change	Pool %[a]	Pyrosequencing primer[b] 5′ → 3′
(A) Conventional sequencing					
EC508	184693	184718	T → C	20	TATTAAGCCTGTCATTATCA
EC868	553036	553059	G → A	2	ACCAGCTGGCTGTTATAAGA
EC1231	553171	553185	G → T	8	GGCTGGGACGTAAGG
EC508	645546	645555	T → C	24	TGACTACCGCATTTTG
EC508	1105925	1105942	G → C	18	CAGTAAGCGTACAGCCT
EC508	1106125	1106138	A → C	5	CGCTCGTCATCTCAA
EC1231	1474718	1474329	G → T	<2	CGTCTGGAGATATGGG
EC1231	1474943	1474956	T → C	<2	TAATCTGCGTTGCCA
EC1231	2764299	2764312	A → C	14	TATGAAGATCCGGGG
EC1231	2764467	2764467	A → G	ND[c]	
EC1231	2764778	2764778	A → G	ND[c]	
EC1231	2948775	2948796	A → G	4	AGCTGACCAAAGGCG
EC1231	2949425	2949431	C → A	2	CATTTATCATAATTCCAGGT
EC1231	3778472	3778477	A → T	10	TGAATCCTCTTCTCGG
EC1231	3778472	3778486	A → C	11	TGAATCCTCTTCTCGG
(B) SQA pyrosequencing					
AB5	276668	276684	G → A	37	CCATTCCAGAGTTGCTT
EC508	368864	368886	C → T	6	TCTTATTCCCGGCAG
AB2	738042	738048	C → T	2	GGATGGGAAAGTACCTG
EC869	922209	922228	C → T	15	GGTGAATGGTGTCACG
EC869	2027998	2028014	SND[d]	21	CACCGGCTAATGTCAG
EC1231	2580072	2580095	A → T	<2	TAACGTATTGTGTTGATTAT
EC869	4054614	4054624	G → A	<2	CGATTCACCAGGAATAA
EC508	4699614	4699632	C → T	<2	CGCAAAGAAGGCGTT
EC1231	4791585	4791606	T → C	3	GCGTTCTTCTTCTAATAGC
EC1231	4884216	4884240	C → T	8	CCTTGCTGATATCAATGAA

[a] A mixture of DNA was prepared by combining (pooling) equal quantities of DNA from 100 individual *E. coli* isolates and was used for pyrosequencing assays in order to estimate the allele frequency of each SNP.
[b] Pyrosequencing primer for determining the SNP's allele frequency.
[c] Not determined.
[d] Single nucleotide deletion.

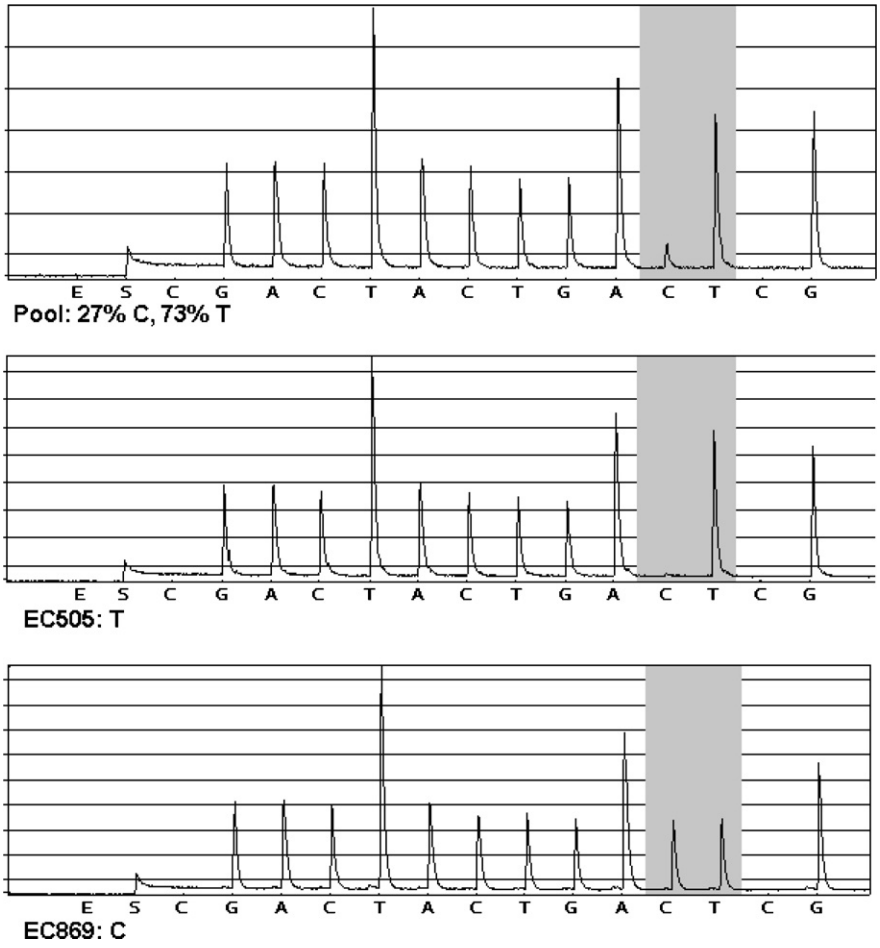

Fig. 7. Pyrogram™ showing SNP allele frequency among a population of strains. A SNP (C/T) located at position 184,718 and the following base (T) are shown in the shaded box. The bottom panel shows strain EC869 with C at the SNP position followed by T. The middle panel shows strain EC505 with T at the SNP position followed by T, read as a double peak. The top panel determined the frequency of the SNP alleles in a pool of DNAs from 18 strains.

Pyrosequencing was used to determine the allele frequencies of the verified SNPs in a pool of DNA from test sets of strains in order to assess the usefulness of novel SNPs for characterizing the diversity in a population of *E. coli* O157:H7 strains (Fig. 7). The distribution of SNPs and allele frequencies were quantified using a pool of 100 *E. coli* O157:H7 strains (Table 3), as results of assays of individual strains from a test set and those of pooled samples are comparable [23].

3.4. Deletions, substitutions, and copy number polymorphisms

In addition to detecting SNPs, hybridization profiles identified the sites of genomic deletions and CNPs that occurred in test strains relative to EDL933. While deletions were apparent due to the dramatic decrease in hybridization intensity along contiguous probe sites, they could often be distinguished from the more modest effect on the relative hybridization intensity caused by CNPs. As several EDL933 loci represented on the array are duplicated elsewhere in the EDL933 genome, CNPs occur in the form of a deletion or amplification of loci in the test strain, respectively increasing or decreasing the relative hybridization intensity (EDL933/test strain) in the tiled region.

Visual inspection of the hybridization profiles revealed the presence of CNPs or multi-nucleotide polymorphisms (MNPs, i.e., deletions, rearrangements, or substitutions of sequence) in the *E. coli* O157:H7 strains analyzed. As CNPs and MNPs could not be distinguished from the tiling array profiles in all cases, we thus classified them together in this analysis. Of 60 loci represented on the tiling array, 10 (17%) contained CNPs or MNPs in one or more of the strains tested (Fig. 8). These 10 variable loci were located within O-islands, i.e., genomic regions of *E. coli* O157:H7 strain EDL933 that are not present in the *E. coli* K-12 MG1655 genome [24]. Six of these loci were highly variable as measured by their absence in 12 (27%) of the strains tested.

Locus 13 (Fig. 9A, beginning at EDL933 position 1,105,960) and locus 17 (Fig. 9A, beginning at EDL933 position 1,474,253) are contained within O-islands #43 and #48, respectively, and each were present as single copy CNPs in 28 (64%) of the strains tested. O-islands #43 and #48 are duplications of the same prophage in EDL933, while other O157:H7 strains, including the Sakai isolate, have a single copy [24]. The genes encoded on the prophage include an integrase, phage proteins, tellurite resistance, and urease gene products. Locus 13 encompasses part of the terF gene that is part of the

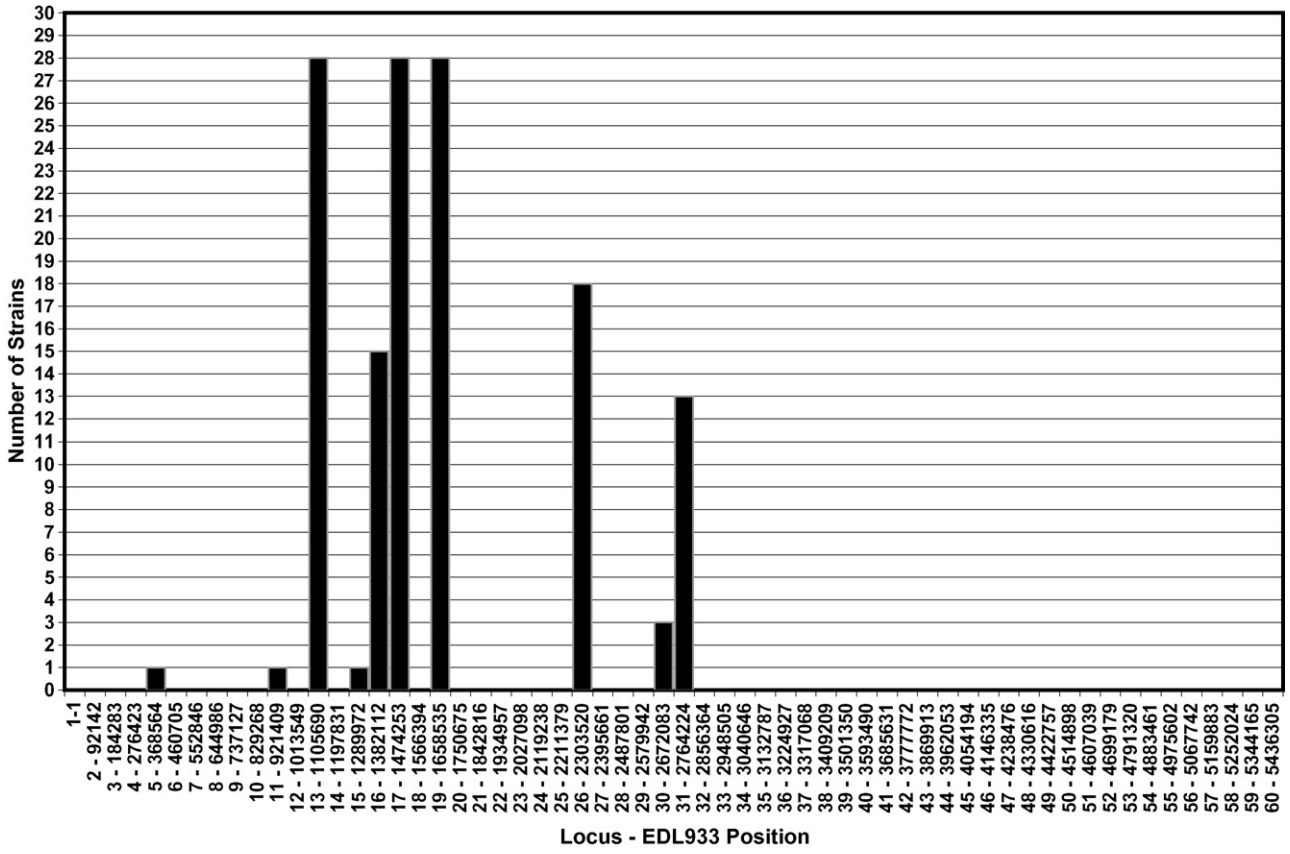

Fig. 8. The frequency and location of CNPs and MNPs. The bar graph shows the population frequency of each polymorphism among the 60 loci examined.

tellurite resistance operon, though *terF* itself is not essential in conferring tellurite resistance [25]. Oligomers on the tiling array complementary to O-island #43 were specific for genes *terF* and *Z1178*, whereas oligomers complementary to O-island #45 were specific for the genes *ureD_2* and *ureA_2*. The *terF* genes are separated from the urease genes by approximately 30 kb on the island(s), and notably, the hybridization profiles always showed the two loci present in the same copy number. These data demonstrate a linkage between the *terF* and urease genes, suggesting that the gain and loss of these gene markers involve transiting of an intact prophage.

Locus 16 (Fig. 9A, beginning at EDL933 position 1,382,112) is contained within O-island #45 and encompasses gene *Z1495* for an unknown protein encoded by bacteriophage BP-933W. This prophage contains genes for the Shiga-like toxin II production (*Z1464* and *Z1465*) in strain EDL933, which contribute to the pathogenesis of O157:H7. This locus is present once in EDL933 and appears to be either partially deleted or hypervariable in 15 (34%) of the strains tested. It is noted that nine of these strains tested positive for the *stxII* genes via multiplex PCR (Table 1). Changes in locus 16 relative to the presence of the *stxII* genes are consistent with the results of Ohnishi et al. [13] and Shaikh and Tarr [26] that, in the large majority of O157:H7 strains they tested, the *stxII* bacteriophage varied in structure and occupied different integration sites than the Sakai and EDL933 reference strains.

Locus 26 (Fig. 9B, beginning at EDL933 position 2,303,520) is contained within O-island #71 and covered genes *Z6042*, *Z6043*, and *Z6044* that encode unknown proteins of cryptic prophage CP-933P. The entire set of genes is duplicated four times in both the EDL933 and Sakai genomes. The first 500 bp of this 1 kb examined was either hypervariable or present as a CNP in 18 (44%) of the strains tested, suggesting that gene *Z6042* encoded by prophage CP-933P is a variable prophage gene.

Locus 31 (Fig. 9C, beginning at EDL933 position 2,764,304) is contained within O-island #79 and covered gene Z3098, which encodes a putative head-tail preconnector protein of prophage CP-933 U. This locus is present twice in EDL933 and once in the Sakai genome. MNPs and CNPS are found in 13 (33%) of the strains examined. Five of the strains examined contained this locus in a higher copy number than that of EDL933.

Other CNPs and/or MNPs were found within strains at locus 5 (position 368,564), locus 11 (position 921,409), locus 15 (position 1,289,972), and locus 30 (Fig. 9B, position 2,672,083). In summary, 136 sites were identified as either MNPs or CNPs among the strains tested. Of the 44 strains, 34 strains (85%) were found to have a CNP or MNP present in at least one of the 60 loci examined (Fig. 10).

3.5. Detection of hypervariable regions

Tiling array data identified several regions of the O157:H7 genome that appeared to be hypervariable. We defined regions as being hypervariable when more than 10 consecutive probes

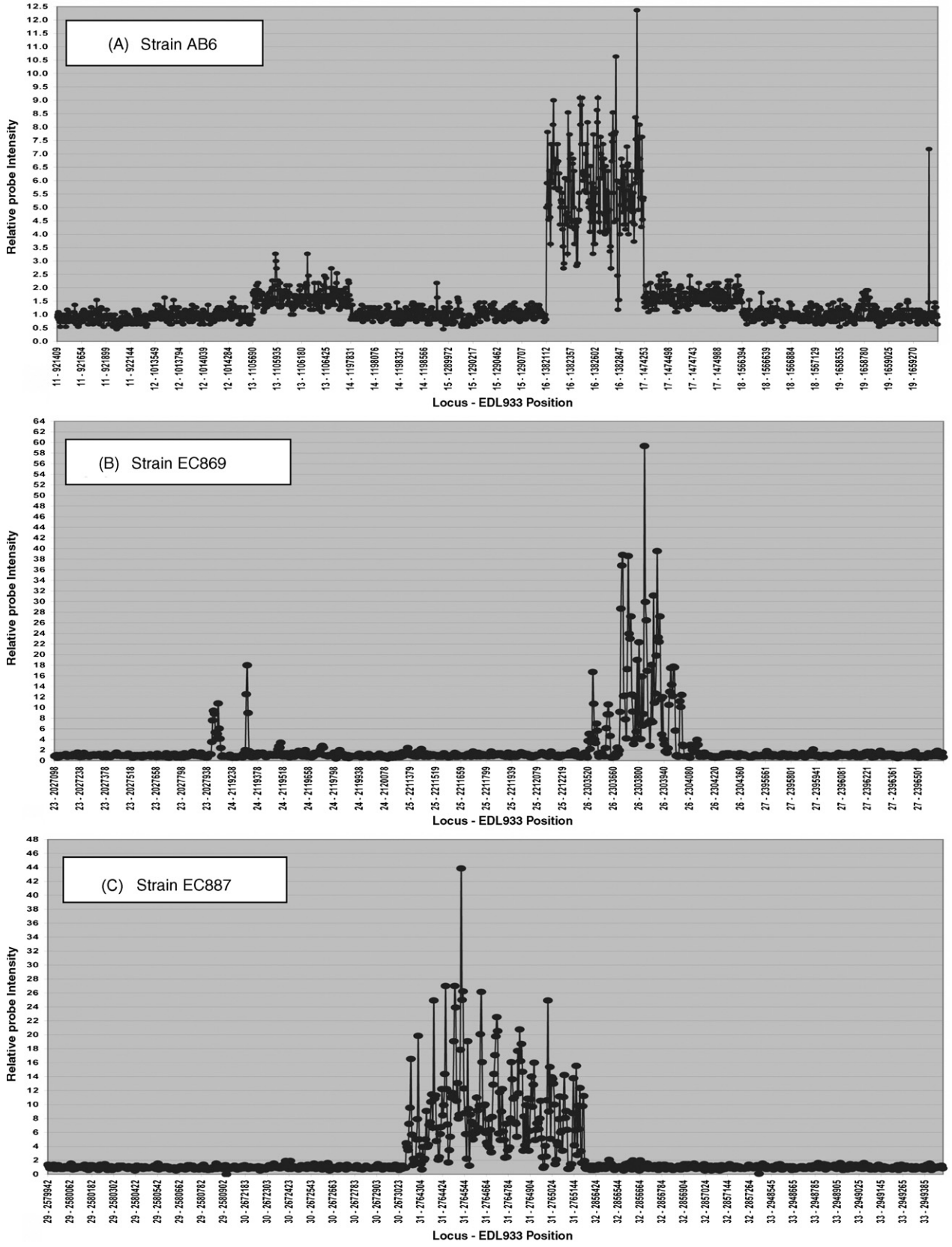

Fig. 9. The effect of CNPs and MNPs on hybridization profiles. CNPs and MNPs are indicated according to their position in the EDL933 genome and by locus number (Table 2). (A) Loci 13 and 17 are present as CNPs while locus 16 is absent in strain AB6; (B) locus 26 is hypervariable in strain EC869; (C) locus 31 is absent in strain EC887.

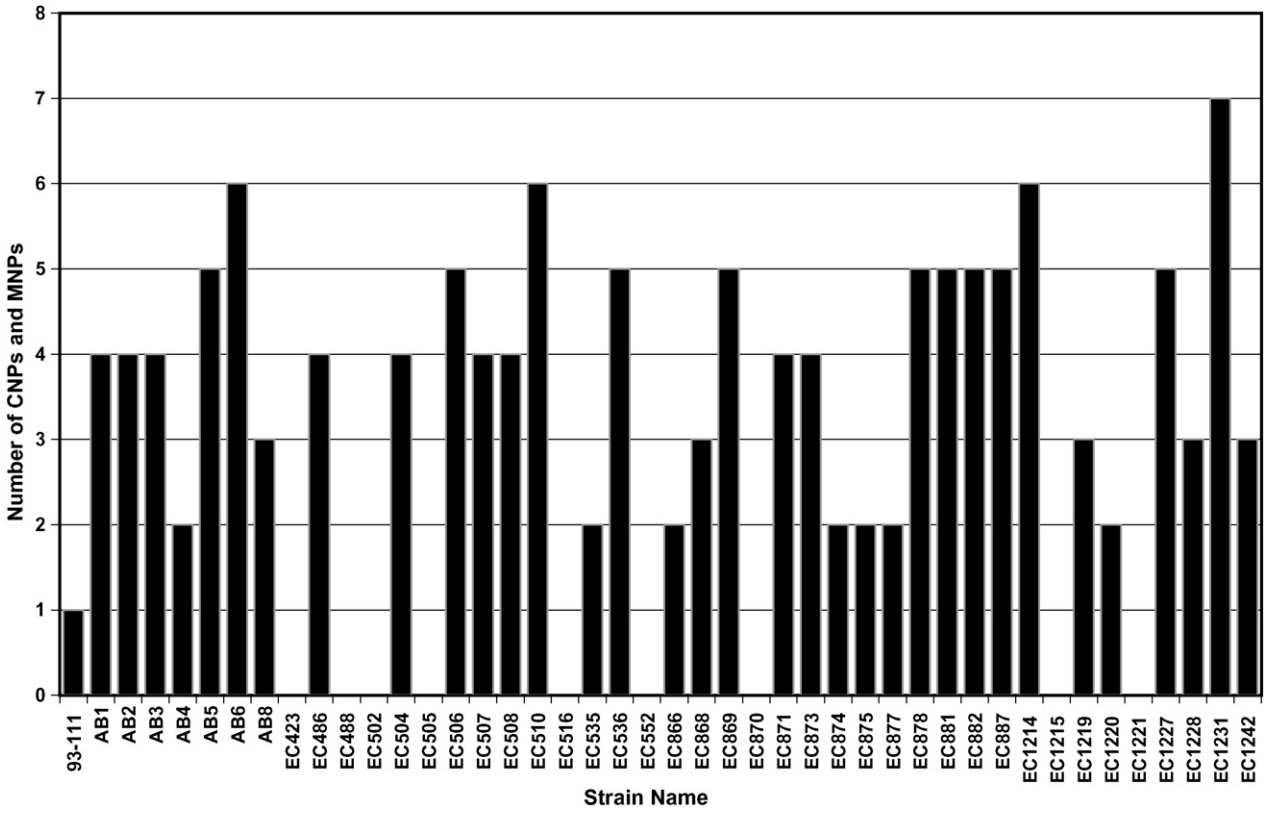

Fig. 10. Bar graph showing the number of CNPs and MNPs found per strain.

gave a hybridization ratio at least three-fold above the baseline. Examples of hybridization profiles showing hypervariable regions are given in Fig. 9A (locus 16) and B (locus 26). SNPs in close proximity, affecting the hybridization of multiple probes, prevented an accurate assessment of the number and position of SNPs in that region. Conventional sequencing was used in order to determine the extent of diversity at these loci and to understand the effect of such hypervariable regions on the hybridization signatures produced from tiled probes spanning these regions.

Locus 16 (Fig. 9A, beginning at EDL933 position 1,382,112) was hypervariable in 12 of the strains tested. Conventional sequencing of this 1-kb locus detected 32 SNPs in strain EC507 and 33 SNPs in strain EC427. Similarly, a portion of locus 26 (Fig. 9B, beginning at EDL933 position 2,303,520) was hypervariable in 12 of the strains tested and sequencing of the locus in strain EC1231 revealed 43 SNPs. It should be noted that locus 26 is present four times in the EDL933 genome, making impossible the determination of SNPs that occur at each site, as SNPs might be present in different alleles of the locus.

Locus 19 (Fig. 9B, beginning at EDL933 position 1,658,535) was also found to be a hypervariable region. This locus is contained within gene Z1814 located in O-island #50, which encodes a cryptic prophage CP-933N. While interpretation of the tiling array data from the 44 strains examined revealed approximately 61 SNPs within 26 unique positions of this locus, conventional sequencing of 619 bp in 18 strains revealed the presence of 115 SNPs in 30 unique SNP positions. Again, this discrepancy is most likely the result of multiple alleles of the locus present in O157:H7 genomes (e.g., EDL933 has five copies of locus 19). It is possible that PCR amplification of this locus resulted in the preferential amplification of a single allele and so does not represent the true polymorphic nature of this allele throughout the genome.

4. Discussion

The genomic diversity of 44 isolates of *E. coli* O157:H7 was assessed using a novel microarray-based platform. The tiling array platform consisted of approximately 13,000 29-mer oligonucleotides with overlapping sequences of 25 bases. Such overlap provided a redundancy that allowed the detection of the sites of single nucleotide change within the genomes of individual strains, as well as exposing the sites of larger genetic changes that alter the genomic architecture of individual strains. The tiling array method is therefore useful as a discovery tool for identifying the sites of polymorphisms that differentiate strains.

The tiling array approach is an extension of chip-based resequencing, introduced by Affymetrix in 1996 [27] and the use of direct hybridization of a microbial genome to a probe array to discover polymorphic markers [28]. While this tiling design cannot provide the genomic resolution of a resequencing strategy, its advantage over standard resequencing applications is that probes are utilized more efficiently to interrogate each genome position for single-nucleotide changes relative to a reference strain. This is achieved by using a spacing of five bases for each successive oligonucleotide probe, instead of one-base spacing as

used in resequencing strategies, providing a five-fold greater sampling of the genome using the same number of probes. Unlike the data provided by resequencing arrays, this probe design is limited to providing the approximate positions of SNPs, within ~25 bp, and does not reveal the identity of the nucleotide change. Subsequent analysis is needed to complement the tiling array data. We have found that the pyrosequencing reaction, using primers directed to the sites of change by tiled oligonucleotides, provides an efficient approach to obtain these data. Pyrosequencing in the SQA mode is first targeted to the site(s) of putative change in a test strain. Then, use of the reaction in the SNP mode is directed to a particular change in pooled genomes in order to establish its allele frequency in a population of strains. It should, however, be pointed out that, as both array strategies rely heavily on the known variation, new variants might be missed.

Employment of a microarray platform incorporating 12 separated wells has the added advantage of increasing throughput. Throughput is critical for examination of genomic diversity across a set of isolates representing the diversity of a pathogen group. Having the ability to interrogate 12 individual strains on a single chip made this investigation rapid and relatively economical. As compared with other methods that examine the diversity of bacterial populations, the costs of these analyses – less that 0.3 cents per base examined – are quite modest.

From the average frequency of SNPs found in the strains examined, and assuming an average size of 5.5 Mb for the O157:H7 genome, one might expect to find over 100 SNPs within the genome of each strain interrogated. While we report that tiling arrays of the 44 strains analyzed here revealed 164 SNPs in the 1% of the genome examined, this number is probably an under-representation as we were unable to resolve, by tiling array data alone, individual SNPs that occurred in hypervariable regions. The ability to identify such hypervariable regions by the tiling array procedure is a valuable feature, however, as these regions can be targeted for further investigation using conventional sequencing in order to harvest the true number and identity of individual SNPs. Because of the mosaic structure of *E. coli* genomes, the concept of average SNP distribution across these *E. coli* strains is not likely correct. Our tiling array results provide further evidence of genetic mosaicism among *E. coli* O157:H7 strains as witnessed by the non-uniform distribution of SNPs among the interrogated strains. Locus 13 contained 11 SNPs within the *terF-Z1178* region, which was the second greatest number of SNPs (10% of the total) detected in this analysis. The copy number of locus 13, part of O-island #43, was found to be highly variable based on the hybridization profiles across the locus. These observations might suggest that this locus is a mobile element that acquired single base changes as a result of frequent horizontal transfer between genomes of related species. An attractive hypothesis is that such sporadic SNP distribution, along with the observed mosaic structure of *E. coli* genomes, may largely be a result of SNP ''hitchhiking'' that occurs during horizontal transfer of larger genomic segments.

Kudva et al. [12] previously demonstrated that individual isolates of *E. coli* O157:H7 differ mainly by insertions and deletions, i.e., multi-nucleotide polymorphisms rather than single-nucleotide polymorphisms. Thus, the O157:H7 tiling chip described here is well suited for assessing the diversity of this *E. coli* pathogen because of its ability to readily detect CNPs and MNPs as well as SNPs. The tiling array data showed that 34 of the 44 strains examined (85%) had a CNP or MNP present, assessed by examination of just 1% of the genome. As many as seven individual loci were found to be deleted in an individual strain. Moreover, we found that all of the *E. coli* strains analyzed in this assay differed in MNPs only within O-islands. This finding supports the idea that strains of the *E. coli* O157:H7 have a common conserved backbone sequence and differ largely by insertions and deletions [12] and the conclusion that, when assessed at the individual strain level, the *E. coli* O157:H7 chromosome is more diverse than previously recognized [26].

The finding that strain differences affecting MNPs were restricted to O-islands is significant. It establishes that O-islands, defined as genomic regions of O157:H7 strain EDL933 that are not present in the K-12 MG1655 genome [15], are in fact specific to strain EDL933 and not the whole O157:H7 pathogroup. Even the hallmark characteristics of O157:H7, such as genes encoding tellurite resistance and Shiga-like toxins, show variability in copy number and chromosomal position. The O-islands are hotspots for DNA insertions and deletions, where such recombination is often mediated by the transiting of phages. Thus, within the recently evolved *E. coli* O157:H7, horizontal gene transfer plays the important diversifying role in changing chromosomal architecture, as well as in the shuffling of SNPs that are found associated with MNPs. It is also possible that stochastic mutations occur within these sites because of higher mutation rates in these regions of the chromosome.

It is critical to make the distinction that the polymorphisms we discovered on tiling arrays may be either synapomorphies, which serve as genetic signatures for the evolutionary relatedness of strains when found in common among strains; or they might be autapomorphies, the isolated characters most useful for distinguishing strains at the individual level [29]. Each of the polymorphisms must be analyzed in a representative set of individual strains, whose phylogeny has been well established by independent means, in order to determine the usefulness of the data either for determining evolutionary relatedness or as significant identifiers for individual strains. As a case in point, we tested a second generation tiling array designed the same as the array described here, except containing probes for informative regions of the *E. coli* genome. These included evolutionarily stable MLST genes as well as genes predicted to have high mutation rates based on comparative sequence analysis of the EDL933 and Sakai genomes. Analysis of a phylogenetically characterized set of pathogenic *E. coli* strains showed that the MLST sites on the tiling array recapitulated the MLST tree, whereas changes in fast-evolving genes, as expected, collapsed the tree. Data from fast-evolving genes are useful, however, when they are used to ''decorate'' an evolutionarily informative tree, because they segregate closely related strains and provide the resolving

power that may give individual strain discrimination. This approach, used to partition clades of *E. coli* O157:H7 strains [23], has also been discussed by Keim et al. [30].

Data of both types, synapomorphies and autapomorphies, will prove useful in providing evidence to support strain attribution for forensics purposes. The evolutionary relatedness of strains would help to establish the source of suspect strains, e.g., their genetic near-neighbors, and provide a means for assessing strain diversity across a pathogenic grouping of strains, knowledge that is requisite for making conclusions about the rarity of a particular strain. Individual strain identifiers, on the other hand, would be most useful for the direct comparison of suspect strains, data obtained from tiling arrays without extensive genome sequencing. Notably, these data would serve the microbial forensics community's need for triaging methods capable of analyzing large numbers of strains, before the decision is made to obtain whole genome sequences. After establishing the nature of informative sites, the tiling array approach is an efficient means of obtaining the needed data.

Acknowledgements

We acknowledge a Department of Homeland Security IAG #224-04-2806 for supporting work on the discrimination of *E. coli* O157:H7 strains reported here. We thank the colleagues who have supplied *E. coli* strains used in this work: Dr. Andrew Benson, University of Nebraska, Lincoln, NE; Dr. Robert Buchanan, U.S. FDA, College Park, MD; Dr. Peter Feng, U.S. FDA, College Park, MD; Dr. Choong Park, Inova Fairfax Hospital, Falls Church, VA; Dr. Phillip Tarr, Washington University, St. Louis, MO; Dr. David White, U.S. FDA, Beltsville, MD; Dr. Thomas Whittam, Michigan State University, East Lansing, MI [31,32].

References

[1] S.A. Morse, B. Budowle, Microbial forensics: application to bioterrorism preparedness and response, Infect. Dis. Clin. North Am. 20 (2006) 455–473.
[2] B. Budowle, M.D. Johnson, C.M. Fraser, T.J. Leighton, R.S. Murch, R. Chakraborty, Genetic analysis and attribution of microbial forensics evidence, Crit. Rev. Microbiol. 31 (2005) 233–254.
[3] S.E. Schutzer, B. Budowle, R.M. Atlas, Biocrimes, microbial forensics, and the physician, PLoS Med. 2 (2005) e337.
[4] L.W. Riley, R.S. Remis, S.D. Helgerson, H.B. McGee, J.G. Wells, B.R. Davis, R.J. Hebert, E.S. Olcott, L.M. Johnson, N.T. Hargrett, et al., Hemorrhagic colitis associated with a rare *Escherichia coli* serotype, N. Engl. J. Med. 308 (1983) 681–685.
[5] C. Su, L.J. Brandt, *Escherichia coli* O157:H7 infection in humans, Ann. Intern. Med. 123 (1995) 698–714.
[6] P.M. Griffin, R.V. Tauxe, The epidemiology of infections caused by *Escherichia coli* O157:H7, other enterohemorrhagic *E. coli*, and the associated hemolytic uremic syndrome, Epidemiol. Rev. 13 (1991) 60–98.
[7] P.S. Mead, P.M. Griffin, *Escherichia coli* O157:H7, Lancet 352 (1998) 1207–1212.
[8] P.M. Griffin, Epidemiology of Shiga toxin-producing *Escherichia coli* infections in humans in the United States, in: J.B. Kaper, A.D. O'Brien (Eds.), *Escherichia coli* O157:H7 and other Shiga toxin-producing *E. coli* strains, ASM Press, Washington, DC, 1998, p. 18.
[9] T.S. Whittam, I.K. Wachsmuth, R.A. Wilson, Genetic evidence of clonal descent of *Escherichia coli* O157:H7 associated with hemorrhagic colitis and hemolytic uremic syndrome, J. Infect. Dis. 157 (1988) 1124–1133.
[10] A.C. Noller, M.C. McEllistrem, O.C. Stine, J.G. Morris Jr., D.L. Boxrud, B. Dixon, L.H. Harrison, Multilocus sequence typing reveals a lack of diversity among *Escherichia coli* O157:H7 isolates that are distinct by pulsed-field gel electrophoresis, J. Clin. Microbiol. 41 (2003) 675–679.
[11] J. Kim, J. Nietfeldt, A.K. Benson, Octomer based genome scanning distinguishes a unique subpopulation of *Escherichia coli* O157:H7 strains in cattle, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 13288–13293.
[12] I.T. Kudva, P.S. Evans, N.T. Perna, T.J. Barrett, F.M. Ausubel, F.R. Blattner, S.B. Calderwood, Strains of *Escherichia coli* O157:H7 differ primarily by insertions or deletions, not single-nucleotide polymorphisms, J. Bacteriol. 184 (2002) 1873–1879.
[13] M. Ohnishi, J. Terajima, K. Kurokawa, K. Nakayama, T. Murata, K. Tamura, Y. Ogura, H. Watanabe, T. Hayashi, Genomic diversity of enterohemorrhagic *Escherichia coli* O157 revealed by whole genome PCR scanning, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 17043–17048.
[14] L.M. Wick, W. Qi, D.W. Lacher, T.S. Whittam, Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7, J. Bacteriol. 187 (2005) 1783–1791.
[15] N.T. Perna, G. Plunkett III, V. Burland, B. Mau, J.D. Glasner, D.J. Rose, G.F. Mayhew, P.S. Evans, J. Gregor, H.A. Kirkpatrick, et al., Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7, Nature 409 (2001) 529–533.
[16] T. Hayashi, K. Makimo, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C.G. Han, E. Ohtsubo, K. Nakayama, T. Murata, et al., Complete genome sequence of entero-hemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain, K-12, DNA Res. 8 (2001) 11–22.
[17] F.R. Blattner, G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, et al., The complete genome sequence of *Escherichia coli* K-12, Science 277 (1997) 1453–1462.
[18] T.A. Cebula, W.L. Payne, P. Feng, Simultaneous identification of strains of *Escherichia coli* serotype O157:H7 and their Shiga-like toxin type by mismatch amplification mutation assay-multiplex PCR, J. Clin. Microbiol. 33 (1995) 248–250.
[19] E.F. Nuwaysir, W. Huang, T.J. Albert, J. Singh, K. Nuwaysir, A. Pitas, T. Richmond, T. Gorski, J.P. Berg, J. Ballin, et al., Gene expression analysis using oligonucleotide arrays produced by maskless photolithography, Genome Res. 12 (2002) 1749–1755.
[20] T.J. Albert, J. Norton, M. Ott, T. Richmond, K. Nuwaysir, E.F. Nuwaysir, K.P. Stengele, R.D. Green, Light-directed $5' \rightarrow 3'$ synthesis of complex oligonucleotide microarrays, Nucleic Acids Res. 31 (2003) e35.
[21] C.W. Wong, T.J. Albert, V.B. Vega, J.E. Norton, D.J. Cutler, T.A. Richmond, L.W. Stanton, E.T. Liu, L.D. Miller, Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays, Genome Res. 14 (2004) 398–405.
[22] J.D. Glasner, P. Liss, G. Plunkett III, A. Darling, T. Prasad, M. Rusch, A. Byrnes, M. Gilson, B. Biehl, F.R. Blattner, N.T. Perna, ASAP, a systematic annotation package for community analysis of genomes, Nucleic Acids Res. 31 (2003) 147–151.
[23] T.A. Cebula, E.W. Brown, S.A. Jackson, M.K. Mammel, A. Mukherjee, J.E. LeClerc, Molecular applications for identifying microbial pathogens in the post-9/11 era, Expert Rev. Mol. Diagn. 5 (2005) 431–445.
[24] D.E. Taylor, M. Rooker, M. Keelan, L.K. Ng, I. Martin, N.T. Perna, V. Burland, F.R. Blattner, Genomic variability of O islands encoding tellurite resistance in enterohemorrhagic *Escherichia coli* O157:H7 isolates, J. Bacteriol. 184 (2002) 4690–4698.
[25] R. Kormutakova, L. Klucar, J. Turna, DNA sequence analysis of the tellurite-resistance determinant from clinical strain of *Escherichia coli* and identification of essential genes, Biometals 13 (2000) 135–139.

[26] N. Shaikh, P.I. Tarr, *Escherichia coli* O157:H7 Shiga toxin-encoding bacteriophages: integrations, excisions, truncations, and evolutionary implications, J. Bacteriol. 185 (2003) 3596–3605.
[27] M. Chee, R. Yang, E. Hubbell, A. Berno, X.C. Huang, D. Stern, J. Winkler, D.J. Lockhart, M.S. Morris, S.P. Fodor, Accessing genetic information with high-density DNA arrays, Science 274 (1996) 610–614.
[28] E.A. Winzeler, D.R. Richards, A.R. Conway, A.L. Goldstein, S. Kalman, M.J. McCullough, J.H. McCusker, D.A. Stevens, L. Wodicka, D.J. Lockhart, R.W. Davis, Direct allelic variation scanning of the yeast genome, Science 281 (1998) 1194–1197.
[29] P.L. Forey, C.J. Humphries, I. Kitching, R.W. Scotland, D.J. Siebert, D. Williams, Cladistics: A Practical Course in Systematics, Clarendon Press, Oxford, 1992.
[30] P. Keim, M.N. Van Ert, T. Pearson, A.J. Vogler, L.Y. Huynh, D.M. Wagner, Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales, Infect. Genet. Evol. 4 (2004) 205–213.
[31] W. Zhang, W. Qi, T.J. Albert, A.S. Motiwala, D. Alland, E.K. Hyytia-Trees, E.M. Ribot, P.I. Fields, T.S. Whittam, B. Swaminathan, Probing genomic diversity and evolution of Escherichia coli O157 by single nucleotide polymorphisms, Genome Res. 16 (2006) 757–767.
[32] D. Gresham, D.M. Ruderfer, S.C. Pratt, J. Schacherer, M.J. Dunham, D. Botstein, L. Kruglyak, Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray, Science 311 (2006) 1932–1936.