INFECTION AND IMMUNITY, Apr. 2000, p. 2009–2015 Vol. 68, No. 4
0019-9567/00/$04.00+0

Molecular Epidemiologic Approaches to Urinary Tract Infection Gene Discovery in Uropathogenic *Escherichia coli*

LIXIN ZHANG, BETSY FOXMAN, SHANNON D. MANNING, PATRICIA TALLMAN, AND CARL F. MARRS*

Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan

Received 19 October 1999/Returned for modification 7 December 1999/Accepted 29 December 1999

Urinary tract infection (UTI) is one of the most frequently acquired bacterial infections. The vast majority of UTIs are caused by a large, genetically heterogeneous group of *Escherichia coli*. This genetic diversity has hampered identification of UTI-related genes. A three-step experimental strategy was used to identify genes potentially involved in *E. coli* UTI transmission or virulence: epidemiologic pairing of a UTI-specific strain with a fecal control, differential cloning to isolated UTI strain-specific DNA, and epidemiologic screening to identify sequences among isolated DNAs that are associated with UTI. The 37 DNA sequences initially isolated were physically located all over the tester strain genome. Only two hybridized to the total DNA of the sequenced *E. coli* K-12 strain; eight sequences were present significantly more frequently in UTI isolates than in fecal isolates. Three of the eight sequences matched to genes for multidrug efflux proteins, usher proteins, and pathogenicity island insertion sites, respectively. Using population characteristics to direct gene discovery and evaluation is a productive strategy applicable to any system.

Urinary tract infection (UTI) is one of the most frequently acquired bacterial infections; *Escherichia coli* accounts for as much as 90% of all UTIs seen among ambulatory patients (20). Certain O:K:H serotypes and virulence factors occur more frequently in urinary isolates than in fecal isolates, suggesting that uropathogenic *E. coli* strains are different from normal bowel inhabitants (18). However, they are also a diverse group: 20 O:K:H serotypes have been associated with pyelonephritis (18). In addition, when first-time UTI isolates were grouped by the presence or absence of nine putative UTI virulence genes, 36 groups were observed (14).

Uropathogenic strains of *E. coli* are believed to display a variety of virulence properties that assist in colonization of host mucosal surfaces and in circumventing host defenses to allow invasion of the normally sterile urinary tract (18). A limited number of virulence factors, including adhesins, siderophores, toxins, capsules, and a protease, have been implicated as important traits allowing uropathogenic *E. coli* to cause disease (1, 7, 8, 12, 18, 21). Nonetheless, no one factor or set of factors is known to uniquely identify uropathogenic *E. coli*. Except for the putative virulence factors CNF1 and OmpT, no new virulence genes have been identified in the last decade. Although considerable insights into some of the above-mentioned virulence determinants have been accumulated, our understanding of the bacterial pathogenesis of UTI is rudimentary due to the limited number of known virulence determinants.

In general, identifying virulence genes among uropathogenic *E. coli* strains usually relies on the individual phenotypic-trait-to-gene approach. Pathogenic phenotypes and their associated assays are the basis for the identification and isolation of the responsible virulence genes. This functional genomic approach to searching for virulence genes is obviously limited to the known determinants. Virulence genes can also be identified by positional cloning and/or sequencing. With the rapid advances in sequencing technology, efforts have been made to identify virulence gene candidates by direct sequence analysis. In several cases, pathogenicity has been correlated with the presence of genes encoding virulence factors organized on large blocks, called pathogenicity islands (PAIs) (17). Straightforward DNA sequence analysis of PAIs has revealed some candidate genes important in UTIs that were not previously known to exist in uropathogenic *E. coli* (19, 27). Large-scale sequencing is still a laborious and costly process, even on the scale of PAIs (11 to 190-kb). Moreover, virulence genes are not necessarily confined within PAIs. Searching for virulence gene candidates through sequence analysis is limited to known virulence determinants, since the gene candidates are initially identified by their homology to known virulence genes. Finally, the pathogenic relevance of gene candidates cannot be established until functional and epidemiologic evaluations have been done.

We propose a molecular epidemiologic approach to UTI virulence gene discovery that uses epidemiologic information to select candidates for subtractive cloning and to evaluate the importance of newly identified genes. Virulence phenotypes of *E. coli* are largely determined by their gene compositions. These virulence phenotypes dictate epidemiologic and clinical outcomes of infection with disease-causing *E. coli* strains, measureable by their association with particular epidemiologic and/or clinical characteristics, which are the population-level phenotypes. By comparing *E. coli* strains with specific epidemiologic features to *E. coli* strains without comparable epidemiologic features, we expect that the virulence genes responsible for the pathogenic features will be found in the genetic difference between the two types of *E. coli* isolates. This report describes our strategy and the experiments used to identify DNA sequences potentially associated with UTI virulence, transmission, and duration of colonization.

MATERIALS AND METHODS

***E. coli* collections and characteristics.** We used *E. coli* isolates from two different collections. The first consists of 364 urinary *E. coli* isolates collected from a cohort of 304 women aged 18 to 39 years with their first UTI followed until their second UTI at the University of Michigan and University of Texas at Austin Health Services from 1992 to 1995. The second collection consists of 393

* Corresponding author. Mailing address: Department of Epidemiology, University of Michigan, 109 Observatory St., Ann Arbor, MI 48109. Phone: (734) 647-2407. Fax: (734) 764-3192. E-mail: cfmarrs@umich.edu.

fecal *E. coli* isolates collected from 395 consecutively sampled college women who used the gynecology clinic at the University of Michigan Health Service during February and March 1996. All isolates were collected and processed as described previously (13). We have type and duration of symptoms, urinalysis and urine culture results, sociodemographic information, and sexual and medical histories (UTI history and recent antibiotic use) associated with the UTI isolates and sociodemographic information and sexual and medical histories associated with the fecal isolates.

All *E. coli* isolates in these collections have been screened for the presence of nine different virulence genes by dot blot hybridization with 13 DNA probes (13, 28). Some of them have also been analyzed by restriction fragment length polymorphism and pulsed-field gel electrophoresis (PFGE) (14). All isolates were previously assigned virulence signatures based on the presence or absence of all nine virulence genes (13) and had pathotypes based on associated and linked virulence genes (28).

For DNA subtraction, we chose a strain from a group of UTI isolates with virulence signature 100111000 positive for *fim*, *aer*, *kpsMT*, and *ompT*. The tester strain, 366-11, is most representative, based on the PFGE band pattern, of a group of temporally clustering first-time UTI isolates within this virulence signature. The driver strain, F320-62, is a fecal isolate with the same virulence signature but with a band pattern different from that of the tester strain. Strain TOP10F′ (Invitrogen, San Diego, Calif.) was used as the host strain for recombinant clones. *E. coli* K-12 strain MG1655 was used as a control for the hybridization (4).

Differential cloning by subtraction PCR. DNA sequences unique to the desired UTI strain, 366-11, were identified by differential cloning through genomic subtraction between the tester (366-11) and driver (F320-62) using a commercial kit (Clontech PCR-Select bacterial genome subtraction kit). This procedure is based on the suppressive subtractive hybridization method (11, 16). After two rounds of hybridization, the tester-specific DNAs are preferentially amplified by two rounds of PCR amplifications. Subtraction between tester strain 366-11 and driver strain F320-62 was performed according to the manufacturers' protocol. However, a high-copy-number plasmid DNA unique to the tester strain was added to the driver DNA pool to suppress its overrepresentation in the final tester-specific sequences. We cloned the secondary PCR products from the subtraction into plasmid vector pCR2.1 with a TOPO TA cloning kit from Invitrogen (Carlsbad, Calif.). These fragments were designated subtracted PCR (sPCR) fragments. They were purified from the recombinant clones by PCR with two primer pairs: pair M13 and T7, which flank the linker (clone insert) region of pCR2.1 (22 cycles of 94°C for 30 s, 58°C for 30 s, and 74°C for 2 min), and pair primer1 and primer2R (supplied with the kit) at all ends of all sPCR fragments (22 cycles of 94°C for 30 s, 68°C for 30 s, and 74°C for 2 min). The PCR products were then spotted on dot blot membranes and probed with labeled genomic tester DNA and labeled driver DNA. Those sPCR fragments that represent DNA regions of the tester strain not present in the driver strain were selected from the hybridization results. They were labeled and used to probe the tester and driver and *E. coli* K-12 to further verify their specificities. Duplicated sPCR clones were removed by cross hybridization among all selected sPCR fragments.

Other molecular techniques. Routine molecular techniques were performed by standard procedures according to the methods of Sambrook et al. (25) and the manufacturers' instructions. DNA probes were prepared according to the previously established procedure (13, 28). Dot blot DNA hybridization, Southern hybridization, and PFGE were also performed according to previously described procedures (13, 14). Sequencing of the double-stranded DNA was performed at the University of Michigan Molecular Biology Core Facility using an Applied Biosystems model 373A automated sequencer.

Data analysis. Differences between groups were tested by the χ^2 test or, for smaller sample sizes, the Fisher exact test. Excel software (Microsoft, Redmond, Wash.) was used for data entry and statistical analysis. Software packages from DNAStar (Madison, Wis.) were used for DNA sequence and amino acid sequence analysis and comparison.

Nucleotide sequence accession number. The GenBank accession numbers of partial nucleotide sequences for eight cloned DNA fragments are AF215925 (P13′), AF215926 (P37), AF215927 (P69), AF215928 (P5), AF215929 (P10), AF215930 (P24′), AF215931 (P64), and AF215932 (P68).

RESULTS

A three-step experimental strategy was used to identify genes potentially involved in *E. coli* UTI transmission or virulence: epidemiologic pairing of a UTI-specific strain with a fecal control, differential cloning to isolate UTI strain-specific DNA, and epidemiologic screening to identify sequences among isolated DNA that are associated with UTI.

Epidemiologic pairing and tester and driver selection. In order to mazimize the potential of subtraction to identify genes important in uropathogenesis or transmission, the initial pairing was made between UTI strains with a virulence signature that appeared to be time and space clustered and fecal isolates with the same virulence signature. Figure 1 shows a PFGE of 10 clustered UTI isolates (tester collection) and 7 fecal isolates (part of the driver collection). UTI isolates from this single virulence signature were more similar to each other than to corresponding fecal isolates. Since subtraction is done between two strains, we had to select a representative strain from each of the two groups as the tester and driver. A semiquantitative measure was used to identify the most average strain from both the tester and driver collections. We counted all the bands on the gel across strains and ranked them by the frequency of their appearance within the group. The strain that had the greatest number of these frequently shared bands was selected as the tester. This same process was performed to identify the driver among fecal strains, excluding those identified as part of the tester clonal group by PFGE. The driver strain was selected from the largest clonal group (three strains) among the remaining fecal strains. Strains 366-11 (Fig. 1a, lane 2) and F320-62 (Fig. 1b, lane 5) were chosen as tester and driver (Fig. 1). They have identical virulence signatures but very different PFGE patterns.

Differential cloning through subtractive hybridization. After choosing the tester and driver strains, the next step was to identify tester strain-specific DNA. The genomic-subtraction kit from Clontech allowed us to preferentially amplify tester-specific DNA. We purified sPCR fragments from the subtraction, identified and verified tester-specific sPCR fragments, and removed duplicated sPCR fragments. Among the first 102 independently cloned sPCR fragments, 46 (45%) contained strain 366-11 (tester)-specific DNA. Among these 46, 37 (80%) were unique DNA fragments as determined by cross hybridization. Only two fragments hybridized to the sequenced genome of *E. coli* strain K-12 (4).

We mapped the 37 unique tester-specific sPCR sequences to specific *Not*I fragments of separated genomic DNA of strain 366-11 by PFGE and Southern blot hybridization. Figure 2 shows the locations and numbers of these sPCR fragments on *Not*I fragments of the tester 366-11 genome. These 37 sPCR fragments are located on 15 of the total of 16 identifiable *Not*I fragments.

Distribution of sPCR fragments by sources, pathotypes, and signatures. The relevance of identified tester-specific sequences to UTI was first evaluated by screening our collections by epidemiologically defined characteristics. The most obvious criterion is their power to differentiate UTI and fecal isolates. *E. coli* pathotype 5, defined in an earlier study (28), might be a mixture of uropathogenic and nonuropathogenic *E. coli*. The major virulence signatures within pathotype 5 are listed in Table 1. Because the tester and driver were selected from a virulence signature within pathotype 5 (28), this group of *E. coli* isolates was the logical choice for initial epidemiologic screening of sPCR fragments. We hybridized the 37 sPCR probes to 94 pathotype 5 UTI isolates that contained almost all the pathotype 5 isolates from our first- and second-time UTI collections and to 94 pathotype 5 isolates from the fecal collection representing approximately 80% of all pathotype 5 isolates from our fecal collection.

We described the percentage of each sPCR sequence among Pathotype 5 UTI isolates and fecal isolates by virulence signature. Almost all the sPCR sequences occurred more frequently in UTI than in fecal isolates with our original tester and driver strain signature; however, the frequency of many sPCR sequences did not differentiate among pathotype 5 strains with other virulence signatures. A limited number of probes are tester strain signature specific, since their presence was largely

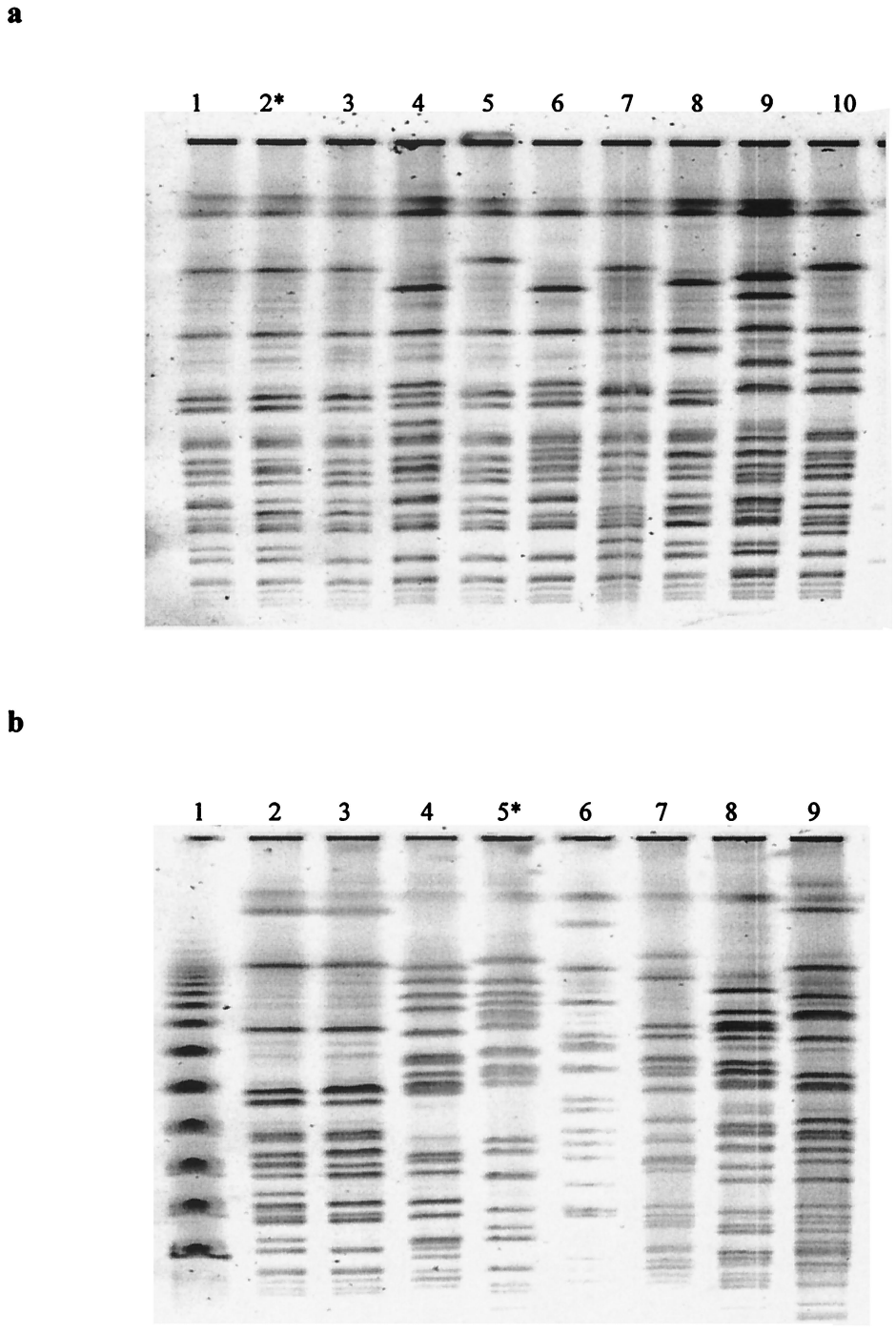

FIG. 1. (a) *Not*I PFGE patterns of UTI isolates. Lanes 1 to 7, first-time UTI *E. coli* isolates 318-11, 366-11, 373-11, 383-11, 244-11, 298-11, and 277-11 collected from patients at University of Michigan (collection 1) ordered by time (September 1993 to November 1994); lanes 8 to 10, first-time UTI *E. coli* isolates 1011-11, 1030-11, and 1031-11 collected from patients at University of Texas—Austin (collection 1) ordered by time (May 1992 to August 1994). All isolates are positive for *fim*, *aer*, *ompT*, and *kpsMT* (virulence profile 100111000). (b) *Not*I PFGE patterns of fecal isolates. Lane 1, lambda. lanes 3 to 9, fecal *E. coli* isolates 400-62, 331-61, 320-62, 250-62, 157-62, 153-62, and 58-62 collected from women presenting at the University of Michigan student health service gynecology clinic over a 2-month period (collection 4). All isolates are positive for *fim*, *aer*, *ompT*, and *kpsMT* (virulence profile 100111000). Lane 2, isolate 1161-11 representing the common PFGE pattern representative of first-time UTI isolates with the same virulence profile. *, Note the difference between the PFGE patterns in lane 5 (the selected fecal driver isolate) and lane 2, one of the UTI samples with a PFGE pattern typical of the cluster that includes lane 2 in panel a (the selected UTI tester isolate).

confined to strains with this signature. Eight of the sPCR probes differentiated between UTI and fecal strains containing signatures other than the tester strain signature. This last group is of the most interest because their differentiating powers transcend the virulence signatures. They may represent UTI genes or may be linked to markers of UTI genes rather than representing clonal differences between our tester and driver isolates.

These eight sPCR DNAs were chosen for further analysis (Table 1). Five of them (P24′, P37, P64, P68, and P69) were

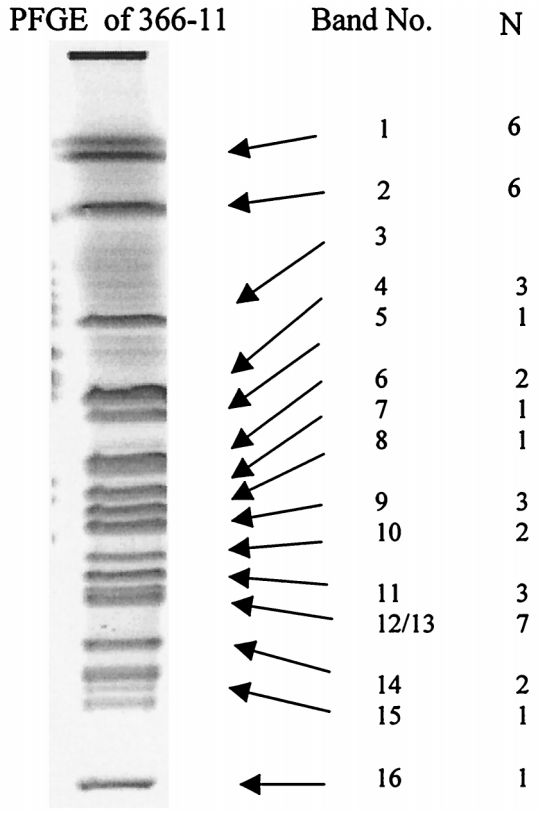

FIG. 2. Location and distribution of sPCR fragments on 366-11 genome. The DNA bands are the result of the *Not*I PFGE pattern of strain 366-11. The identifiable bands are numbered according to their sizes. The band just above band 1 is the compression zoom. N, numbers of sPCR fragments mapped to each *Not*I band.

chosen because they were found more frequently in UTI than fecal pathotype 5 isolates. P5 and P10 were chosen because they were found more frequently in UTI than fecal isolates within virulence signature 1101110000100 (positive for *fim*, *prf*, *aer*, *kpsMT*, *ompT*, and *papGad*), which occurs more frequently in UTI isolates. P13′ was chosen because its pattern of hybridization to pathotype 5 isolates was remarkably similar to the pattern seen using a probe from a new class of pilus usher genes we had previously identified (data not shown). The frequencies of these eight sPCR sequences were further examined among our entire urinary and fecal *E. coli* collections. All eight sPCR sequences are present significantly more frequently in UTI than in fecal isolates—including ones that previously could not be differentiated within the pathotype 5 strains (Table 2).

Sequence analysis of sPCR fragments. The sequence homology of sPCR DNAs to known genes might provide insight into their uropathogenic potential. We sequenced the eight sPCR fragments (between 360 and 650 bp) identified in Table 1 and performed sequence similarity searches. Three of these eight sequences (P10, P68, and P69) have no match to currently known sequenced genes or proteins at a 15% similarity threshold. Two of them (P5 and P64) had limited similarity to a hypothetical protein of non-*E. coli* species. P5 has 32.3% similarity to an internal region of a predicted periplasmic protein from the human pathogen *Campylobacter jejuni*, and P64 has 30.3% similarity to an internal region of a predicted protein on the Ti (tumor-inducing) plasmid of *Agrobacterium tumefaciens* that causes crown gall tumor formation in plants.

The remaining three sequences (P24′, P13′, and P37) have matches to DNA or proteins with known functions. The partial protein sequence deduced from P24′ has 60 to 75% sequence similarity to a family of multidrug efflux proteins found in many bacteria, such as *Bacillus*, *Helicobacter*, and *Mycobacterium tuberculosis*, as well as in eucaryotic organisms, including humans (6). It is most closely (76.5% similarity) related to the bicyclomycin resistance protein homolog *ydgk* of *Bacillus subtilis*. The protein sequence deduced from P13′ has similarity to a number of usher proteins that are essential components in making various pilus and adhesin structures found in uropathogens, such as type I pili, P pili, S fimbriae, and Dr adhesins. However, the DNA sequence homology of P13′ to any other known usher gene is less than 65%. The third sequence, P37, has an even more interesting match. The first half of the sequence matches 100% to an amber suppressor gene *supP*, a dispensable member of the Leu-tRNA family that exists in K-12 as *leuX*. The protein sequence deduced from the second half of this sPCR sequence has 52.6 to 96% similarity to a group of phage integrases including P4 integrase which also exists in K-12 next to *leuX*. It is possible that this sequence represents an integration of exogenous DNA at the tRNA site. This sequence may represent the remnants of an ancestral prophage or part of an inserted PAI. Since UTI isolates are significantly more likely to have the P37 sequence than fecal isolates, the latter possibility is worth investigating.

DISCUSSION

We present a strategy of bacterial gene identification and evaluation that relies on epidemiologic information for selecting isolates for study and screening of epidemiologically defined collections for evaluation of the significance of the genes identified. Our epidemiologic approach to gene discovery relies on population phenotypes (epidemiologic characteristics) to direct gene discovery. By comparing the relative frequencies of subtracted sequences among uropathogenic strains and fecal strains, we can identify sequences potentially associated with uropathogenicity. In contrast to individual-level phenotypes, population phenotypes that reflect true human-pathogen interactions can be measured by properly designed epidemiologic studies. This approach has the potential to open up a new avenue for identifying pathogenesis-related genes.

In this study, we tested our strategy using uropathogenic *E. coli*. We first matched a urinary isolate from a clustering UTI strain to a nonclustering fecal strain. We then identified UTI strain-specific DNA sequences by genomic subtraction between two strains. Among the initial 37 regions of DNA isolated, several sequences have potential importance in UTI based on epidemiologic screening and sequence analysis.

Although the selected UTI tester strain, 366-11, and the fecal driver strain, F320-62, share a virulence signature (the same nine UTI genes are present or absent), the first 37 tester strain-specific DNA fragments identified were physically located all over the tester strain genome. This indicates that genetic differences between these two strains exist over many regions of the chromosome. Only two of these 37 fragments hybridize to the total DNA of the sequenced *E. coli* strain K-12. Thus, we can reasonably assume that most of the identified UTI strain-specific DNAs are not part of *E. coli* housekeeping genes that would also be present on the K-12 genome.

Among eight sequenced sPCR fragments, only three (P24′, P13′, and P37) have sequence matches to known *E. coli* genes. Based on their distribution among our *E. coli* collections and their sequence homology, all three have potential importance in UTI. P24′ has sequence similarity to a family of multidrug

TABLE 1. Examples of sPCR probes hybridized to pathotype 5 isolates

Pathotype 5 virulence signature[a] and group	Hybridization [no. (%)] to probe:							
	P5	P10	P13′	P24′	P37	P64	P68	P69
1000110000000								
UTI (n = 23)	14 (61)	13 (57)	18 (78)	21 (91)	1 (4)	1 (4)	2 (9)	7 (30)
Fecal (n = 27)	18 (67)	17 (63)	23 (85)	20 (74)	2 (7)	1 (4)	2 (7)	6 (22)
1001110000000[b]								
UTI (n = 15)	11 (73)	11 (73)	13 (87)	13 (87)	8 (53)	13 (87)	12 (80)	4 (27)
Fecal (n = 30)	19 (63)	16 (53)	20 (67)	21 (70)	13 (43)	24 (80)	20 (67)	4 (13)
1100110000100								
UTI (n = 13)	13 (100)	12 (92)	12 (92)	13 (100)	0 (0)	1 (8)	3 (23)	1 (8)
Fecal (n = 19)	18 (95)	18 (95)	19 (100)	15 (79)	1 (5)	0 (0)	1 (5)	3 (16)
1101110000100								
UTI (n = 25)	21 (84)	9 (36)	8 (32)	12 (48)	20 (80)	19 (76)	18 (72)	16 (64)
Fecal (n = 8)	5 (63)	1 (13)	2 (25)	4 (50)	5 (63)	5 (63)	6 (75)	3 (38)
All other signatures								
UTI (n = 18)	11 (61)	10 (56)	14 (78)	15 (83)	4 (22)	9 (50)	7 (39)	6 (33)
Fecal (n = 10)	6 (60)	4 (40)	6 (60)	5 (50)	2 (20)	5 (50)	3 (30)	3 (30)
Total pathotype 5								
UTI (n = 94)	70 (74)	55 (59)	65 (69)	74 (79)	33 (35)	43 (46)	42 (45)	34 (36)
Fecal (n = 94)	66 (70)	56 (60)	70 (74)	65 (69)	23 (24)	35 (37)	32 (34)	19 (20)
Probe location[c]	11	11	1	10	2	4	1 and 4	2

[a] Hierarchically defined groups based on the associations among all genes. Those with the strongest associations combined with known biologic evidence for physical linkage were classified as a pathotype (28). 1 and 0, presence and absence, respectively, of the following genes (in order): *fim*, *prf*, *sfa*, *aer*, *kpsMT*, *ompT*, *hly*, *cnfI*, *drb*, *capIII*, *papGad*, *prsGj96*, and *papGj96*.
[b] Virulence signature of tester and driver strains.
[c] PFGE band of tester (366-11) DNA cleaved with *Not*I and hybridized.

efflux systems found in many bacteria. Acquisition of a multidrug efflux system by a bacterium may decrease its susceptibility to a broad spectrum of antibiotics. If P24′ is indeed part of a functional efflux system, it is possible that the higher frequency of P24′ sequence found in UTI isolates than in fecal isolates could be the result of selection pressure due to the increased use of antibiotics among women with UTI. A recent study has shown an increased prevalence of antimicrobial resistance among *E. coli* organisms isolated from patients with community-acquired cystitis (15). Thus, antibiotic resistance genes could have potential clinical implications in the treatment of UTI.

The second sequence, P13′, has sequence homology to several known pilus usher genes. However, its DNA sequence similarity to any other known usher genes is less than 65%. The prevalence of P13′ sequence among our *E. coli* collections is over 50%, and its distribution does not match any of the other four classes of adhesins we examined. Therefore, P13′ may indicate the presence of a new class of adhesin structure which has not yet been identified. Further study is needed to illustrate its structure and function and to evaluate its importance in UTI.

The third sequence, P37, matches both the tRNA gene *lexU* and the phage P4 integrase gene *intB*. P37 hybridized to the *E. coli* strain K-12, and these genes are located next to each other at 96.9 min in the K-12 genome (3, 4). The *intB* gene and one other phage-related gene are the only remnants of an ancestral prophage (3, 4, 23). Nothing about this region in K-12 would seem to explain why it is found more frequently in the UTI isolates than in fecal isolates. However, an interesting possibility is that this region may serve as an integration site for a PAI present in at least some of the UTI isolates.

At least three known examples of P4-like integrase genes next to tRNA genes serving as such integration sites were found, including two in other species. The *vap* (virulence-associated protein) genes of *Dichelobacter nodosus* (the agent of sheep footrot) insert into and duplicate the 3′ end of the tRNA gene, *serV*, located next to a P4-like integrase gene (10). In *Mesorhizobium loti*, a 500-kb symbiosis island integrates into and duplicates the 3′ end of a Phe-tRNA gene (26). However, the most relevant is PAI II, found in uropathogenic *E. coli* strain 536. PAI II is located at *leuX* at map position 97 of the *E. coli* chromosome (5, 17). It is characterized by the presence of an α-hemolysin gene cluster (*hly*) as well as a *prf* (P-related fimbria) determinant (17). Interestingly, many P37-positive isolates in our *E. coli* collections do not carry any *hly* and *prf* sequences. If there are PAIs integrated at this site among these strains, they are probably different from PAI II of strain 536 in gene composition.

Most of the initial 37 tester strain-specific DNA sequences identified by the new approach are absent on the sequenced K-12 strain. K-12 is a nonpathogenic laboratory strain, and its genome size of 4.6 Mb is among the smallest in the *E. coli* population (*E. coli* genome size can vary between 4.5 and 5.5

TABLE 2. Hybridization of eight sPCR probes with genomic DNA from urinary and fecal *E. coli* isolates

sPCR probe	Hybridization [no. (%)][a]		P[b]
	UTI (n = 350)	Fecal (n = 349)	
P13′	222 (64)	149 (43)	<0.0001
P37	206 (59)	123 (35)	<0.0001
P69	113 (32)	89 (26)	0.047
P5	257 (74)	218 (63)	0.002
P10	207 (59)	129 (37)	<0.0001
P24′	268 (77)	192 (55)	<0.0001
P64	203 (58)	114 (33)	<0.0001
P68	218 (62)	134 (38)	<0.0001

[a] Urinary isolates are from women with first-time UTI; fecal isolates are from women without UTI infections (28).
[b] Comparison between UTI and fecal strains using chi-square test.

Mb [2, 4]). Thus, it is not surprising that many *E. coli* genes, especially virulence-related genes, are not found on K-12. There is certainly a need to sequence other pathogenic *E. coli* strains. Given the genetic diversity among *E. coli* strains, even a few more complete sequences will not fully reveal the richness of *E. coli* genetic information. A better approach will be to target strain-specific DNA for selective sequencing. PAIs are good candidates for identifying virulence genes through direct sequencing. However, PAIs still require large-scale sequencing efforts and the ability to identify and track PAIs in the first place.

The positional-cloning approach to identifying disease genes has been the workhorse for human disease gene identification over the past few decades. Many techniques used for humans are not applicable to bacteria due to differences in genetics. Comparing chromosomal structures between strains to identify possible virulence loci, such as PAIs, has been proposed as a positional approach to identifying genes related to *E. coli* pathogenesis (24). This approach may be complicated by the plasticity of the bacterial genome. In addition to horizontal gene transfers, chromosomal rearrangements are also very common. The uncertainty of the chromosomal structure of *E. coli* may limit the success of the positional-cloning approach. The epidemiologic-subtraction approach used in this study enabled the identification of sequences potentially important for UTI without regard to the genetic location of the DNA within the bacteria. It could also identify PAIs based on the distribution of larger numbers of sPCR fragments to the same genomic location. Our approach provides a direction for positional cloning and sequence analysis. Furthermore, by the nature of the epidemiologic approach, genes identified through this process can be immediately evaluated for their roles in UTI pathogenesis and transmission at the population level. However, to define their functions and verify their importance in UTI, further in vitro and in vivo studies are needed. It is important to note that our method will not identify important UTI genes that are present in both the UTI driver and fecal tester strains.

Our approach is possible because we have *E. coli* collections from well-defined populations with detailed epidemiologic and clinical information associated with each isolate. Epidemiologically defined *E. coli* collections provide meaningful subtraction comparison groups. In the history of infectious disease research, disease outbreaks have always provided the best sources for identifying pathogens and their associated virulence genes. The identification of human immunodeficiency virus and enterohemorrhagic *E. coli* O157:H7 provides two good examples. The classification of several classes of *E. coli* causing diarrheal disease as separate entities was not possible until clones and virulence factors associated with epidemics were identified. Ultimately, at least six different pathogenic processes were identified at the molecular level (22). The first epidemiologic pairing chosen in our subtraction and described in this paper was between *E. coli* isolates involved in time-space UTI clustering and the matching fecal control. However, many other potential pairings can be made within our other collections. For example, pairings can be made between *E. coli* strains from women with recurring UTI and women with nonrecurring UTI to isolate *E. coli* risk factors for recurrent infections. Pairings can also be made between *E. coli* strains shared by heterosexual partners and fecal strains that are not shared to identify factors important in transmission.

This approach is not limited to the discovery of UTI-related genes in *E. coli*. The general principle of using population characteristics to direct gene discovery and evaluation is an expansion of the traditional phenotype-to-gene approach and is applicable to any system.

ACKNOWLEDGMENTS

This work was supported by National Institute of Digestive and Kidney Diseases Grants DK35368 (B.F.) and DK031388 (B.F.).

REFERENCES

1. **Ararwal, M., R. Wilkinson, and R. Goldstein.** 1989. Molecular epidemiology of adhesin and hemolysin virulence factors among uropathogenic *Escherichia coli*. Infect. Immun. **57:**303–313.
2. **Bergthorsson, U., and H. Ochman.** 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. Mol. Biol. Evol. **15:**6–16.
3. **Berlyn, M. K.** 1998. Linkage map of *Escherichia coli* K-12, edition 10: the traditional map. Microbiol. Mol. Biol. Rev. **62:**814–984.
4. **Blattner, F. R., 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, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277:**1453–1474.
5. **Blum, G., M. Ott, A. Lischewski, A. Ritter, H. Imrich, H. Tschape, and J. Hacker.** 1994. Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. Infect. Immun. **62:**606–614.
6. **Bradley, G., P. F. Juranka, and V. Ling.** 1988. Mechanism of multidrug resistance. Biochim. Biophys. Acta **948:**87–128.
7. **Caprioli, A., V. Falbo, F. M. Ruggeri, L. Baldassarri, R. Bisicchia, G. Ippolito, E. Romoli, and G. Donelli.** 1987. Cytotoxic necrotizing factor production by hemolytic strains of *Escherichia coli* causing extraintestinal infections. J. Clin. Microbiol. **25:**146–149.
8. **Carbonetti, N. H., S. Boonchai, S. H. Parry, V. Vaisanen-Rhen, T. K. Kornonen, and P. H. Williams.** 1986. Aerobactin-mediated iron uptake by *Escherichia coli* isolates from human extraintestinal infections. Infect. Immun. **51:**966–968.
9. **Caugant, D. A., B. R. Levin, G. Lidin-Janson, T. S. Whittam, C. Svanborg Eden, and R. K. Selander.** 1983. Genetic diversity and relationships among strains of *Escherichia coli* in the intestine and those causing urinary tract infections. Prog. Allergy **33:**203–227.
10. **Cheetham, B. F., D. B. Tattersall, G. A. Bloomfield, J. I. Rood, and M. E. Katz.** 1995. Identification of a gene encoding a bacteriophage-related integrase in a *vap* region of the *Dichelobacter nodosus* genome. Gene **162:**53–58.
11. **Diatchenko, L., Y. F. C. Lau, A. P. Campbell, A. Chenchik, F. Moqadam, B. Huang, S. Lukyanov, K. Lukyanov, N. Gurskaya, E. D. Sverdlov, and P. D. Siebert.** 1996. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc. Natl. Acad. Sci. USA **93:**6025–6030.
12. **Donnenberg, M. S., and R. A. Welch.** 1996. Virulence determinants of uropathogenic *Escherichia coli*, p. 135–174. *In* L. T. Mobley and J. W. Warren (ed.), Urinary tract infection: molecular pathogenesis and clinical management. American Society for Microbiology, Washington, D.C.
13. **Foxman, B., L. Zhang, K. Palin, P. Tallman, and C. F. Marrs.** 1995. Bacterial virulence characteristics of *Escherichia coli* isolates from first-time urinary tract infection. J. Infect. Dis. **171:**1514–1521.
14. **Foxman, B., L. Zhang, P. Tallman, K. Palin, C. Rode, C. Bloch, B. Gillespie, and C. F. Marrs.** 1995. Virulence characteristics of *Escherichia coli* causing first urinary tract infection predict risk of second infection. J. Infect. Dis. **172:**1536–1541.
15. **Gupta, K., D. Scholes, and W. E. Stamm.** 1999. Increasing prevalence of antimicrobial resistance among uropathogens causing acute uncomplicated cystitis in women. JAMA **281:**736–738.
16. **Gurskaya, N. G., L. Diatchenko, A. Chenchik, P. D. Siebert, G. L. Khaspekov, L. L. Vagner, O. D. Ermolaeva, S. A. Lukyanov, and E. D. Sverdlov.** 1996. Equalizing cDNA subtraction based on selective suppression of polymerase chain reaction: cloning of Jurkat cell transcripts induced by phytohemagglutinin and phorbol 12-myristate 13-acetate. Anal. Biochem. **240:**90–97.
17. **Hacker, J., G. Blum-Oehler, I. Muhldorfer, and H. Tschape.** 1997. Pathogenicity islands of virulent bacteria: structure, function, and impact on microbial evolution. Mol. Microbiol. **23:**1089–1097.
18. **Johnson, J. R.** 1991. Virulence factors in *Escherichia coli* urinary tract infection. Clin. Microbiol. Rev. **4:**80–128.
19. **Kao, J. S., D. M. Stucker, J. W. Warren, and H. L. T. Mobley.** 1997. Pathogenicity island sequences of pyelonephritogenic *Escherichia coli* CFT073 are associated with virulent uropathogenic strains. Infect. Immun. **65:**2812–2820.
20. **Kunin, C. M.** 1997. Urinary tract infections: detection, prevention, and management, 5th ed. Williams & Wilkins, Baltimore, Md.
21. **Lundrigan, M. D., and R. M. Webb.** 1992. Prevalence of *ompT* among *Escherichia coli* isolates of human origin. FEMS Microbiol. Lett. **97:**51–56.
22. **Nataro, J. P., and J. B. Kaper.** 1998. Diarrheagenic *Escherichia coli*. Clin. Microbiol. Rev. **11:**142–201.
23. **Pierson, L. S., III, and M. L. Kahn.** 1987. Integration of satellite bacteriophage P4 in *Escherichia coli*: DNA sequences of the phage and host regions involved in site-specific recombination. J. Mol. Biol. **196:**487–496.

24. **Rode, C. K., L. J. Melkerson-Watson, A. T. Johnson, and C. A. Bloch.** 1999. Type-specific contributions to chromosome size differences in *Escherichia coli*. Infect. Immun. **19:**230–236.
25. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
26. **Sullivan, J. T., and C. W. Ronson.** 1998. Evolution of rhizobia by acquisition of a symbiosis island that integrates into a phe-tRNA gene. Proc. Natl. Acad. Sci. USA **95:**5145–5149.
27. **Swenson, D. L., N. O. Bukanov, D. E. Berg, and R. A. Welch.** 1996. Two pathogenicity islands in uropathogenic *Escherichia coli* J96: cosmid cloning and sample sequencing. Infect. Immun. **64:**3736–3743.
28. **Zhang, L.** 1999. Molecular epidemiology of uropathogenic *Escherichia coli*. Ph.D. dissertation. University of Michigan, Ann Arbor.

Editor: P. E. Orndorff