EHEC Genomics: Past, Present, and Future

SHAH M. SADIQ,1 TRACY H. HAZEN,2 DAVID A. RASKO,2 and MARK EPPINGER1

1Department of Biology, and South Texas Center for Emerging Infectious Diseases, University of Texas at San Antonio, San Antonio, TX 78249; 2Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201

ABSTRACT This article examines the role of genomics in the understanding and identification of O157:H7 enterohemorrhagic Escherichia coli (EHEC). We highlight the development of novel molecular typing systems that are based on the genomic sequence that has been generated for this pathotype. The genomic comparisons of EHEC to other E. coli strains highlight the close relatedness of the O157 and O55 isolates and also identify other non-O157 clades of isolates that appear to have a different genomic history. Analysis within the EHEC isolates must be completed on a fine scale using whole-genome sequence-based approaches to assess both the conserved and lateral acquired gene content. The plethora of genomic data for EHEC isolates has provided the ability to examine this pathotype in detail, which has provided opportunities for novel surveillance, detection, and diagnostics.

INTRODUCTION O157:H7 is the most common enterohemorrhagic Escherichia coli (EHEC) serotype in North America, and it has been the principal causative agent of numerous food-poisoning outbreaks worldwide (1, 2). Initially E. coli O157:H7 was recognized as a human pathogen in 1982 when it was isolated from 47 persons in two states who had developed bloody diarrhea after consuming hamburgers contaminated with this organism (3). Since then, E. coli O157:H7 has emerged as a major enteric pathogen, capable of causing localized infections and large outbreaks of gastrointestinal disease (4). Data accumulated from 1982 to 1996 showed that approximately two-thirds of the 3,000 cases of E. coli infections from 139 recognized outbreaks were associated with the ingestion of contaminated food products, whereas 22% of the reported cases were from direct person-to-person transmission and 10% were from drinking water (5). Surveillance data have demonstrated a high prevalence of E. coli O157:H7 among cattle and their environment, but a relatively low incidence of human infection. This supports the potential hypothesis that a subset of E. coli O157:H7 harbored in cattle may be responsible for the majority of human disease (6). To minimize or eradicate adverse effects on public health, the E. coli O157:H7 lineage has been the focus of numerous epidemiological, microbiological, genomic, forensic, and diagnostic studies. Overall, it is estimated that E. coli O157:H7 alone causes more than 76,000 infections and 61 deaths in humans due to severe complications annually in the United States (7). Symptoms include a range of gastrointestinal morbidities, such as severe abdominal cramping accompanied with little or no associated fever and a watery diarrhea that leads to severe bloody diarrhea (8). Although many infected individuals remain asymptomatic, approximately 15 to 20% of people infected with EHEC present severe enough symptoms to require hospitalization. In such severe cases, patients display renal dysfunction known as hemolytic-uremic syndrome.

Received: 25 October 2013, Accepted: 18 November 2013, Published: 27 June 2014

Editors: Vanessa Sperandio, University of Texas Southwestern Medical Center, Dallas, TX, and Carolyn J. Hovde, University of Idaho, Moscow, ID


Correspondence: Mark Eppinger, mark.eppinger@utsa.edu; David A. Rasko, drasko@som.umaryland.edu

© 2014 American Society for Microbiology. All rights reserved.
(HUS), hemorrhagic colitis, and central nervous system failure with potentially lethal outcomes (9–11).

The nomenclature surrounding EHEC and other Shiga-toxin containing E. coli (STEC) can be confusing; hence, the molecular definitions for EHEC and STEC that we employ in this article are outlined in Fig. 1. The definitions are based on key virulence factors of the EHEC, STEC, and enteropathogenic E. coli (EPEC) pathogenic variants or pathovars (1, 2). We briefly introduce virulence factors as a way to define the pathogens and isolates we are examining. The three virulence factors are the eae gene that encodes intimin and is used as a surrogate marker for the locus of enterocyte effacement (LEE region); the bfpA gene, used as a marker for the bundle-forming pilus operon that has been described as essential in the binding to the epithelial cells; and the stx gene (in this case, this means any Shiga toxin genes). Figure 1 demonstrates how these three genes separate the EPEC, EHEC, and STEC isolates. EHEC isolates contain both the eae and stx genes, whereas STEC isolates contain only the stx gene. The EPEC can be separated into two groups: typical EPEC, containing both eae and bfp, and atypical EPEC, containing only the eae gene. One can quickly identify the potential flaws in these molecular definitions, especially since each of these virulence factors is encoded on a mobile element: stx on the phage; eae on the LEE genomic island, and bfp on the EAF plasmid. The ease with which each or many of these genes can be lost has not been defined in many strains, and thus, the lack of any of these features may prevent the proper categorization of the isolate to the appropriate pathovar. This highlights the crucial need for the development of more stable genetic biomarkers from genomic information that is addressed later in this article.

It has been noted that recent outbreaks, especially those associated with green leafy vegetables, have been associated with increased virulence, as measured by the number of individuals that present in health care facilities. Three E. coli O157:H7 outbreaks in 2006 from ingested fresh produce, referred to as spinach, Taco Bell, and Taco John outbreaks (http://www.cdc.gov/ecoli/), captured the attention of both the public health and lay communities. For example, the spinach outbreak caused illness in 199 people from 26 states after they ingested fresh spinach contaminated with E. coli O157:H7 (12). Three deaths were attributed to the spinach outbreak: two elderly women and a 2-year-old child. Among the ill, 51% were hospitalized, and in 16% (12) of the cases, infection progressed to HUS and kidney failure. The high number of patients hospitalized and the high rate of kidney failure suggest that this outbreak was due to a more virulent strain of E. coli O157:H7. The isolates for the most recent outbreaks have been examined by genome sequencing as a part of the epidemiological and microbiological examination of the outbreak.

The introduction of microbial sequencing has opened up the opportunity for comparative analysis of many genomes to identify regions of the genome that may be associated with the greater virulence described above. The sequencing of the E. coli MG1655 isolate was the beginning of E. coli genome sequencing (13); this was followed by the sequencing of two EHEC O157:H7 isolates, EDL933 (14) and the Sakai isolate (15). The sequencing of these two isolates and the associated comparative analysis provided significant insights into two key points of the evolution of E. coli in general and EHEC specifically: (i) the isolates of EHEC were closely related and the analysis could distinguish differences in each group, and (ii) a significant amount of diversity within and between the EHEC isolates existed, but there was 1 Mb of DNA that was unique in the EHEC isolates that was not present in the laboratory-adapted K-12 isolate. This represented approximately 20% of the genome that was unique in the pathogen, providing ample

**FIGURE 1.** Figure depicts the molecular differences that define each of the attaching and effacing E. coli. The eae gene encodes the intimin protein on the LEE region; the bfp gene in this case is the presence of the bundle-forming pilus operon, and the stx gene encodes the Shiga toxin. These three features are classically used to define the pathotypes, including EHEC. doi:10.1128/microbiolspec.EHEC-0020-2013.f1
opportunity for functional characterization. Continued genome sequencing of *E. coli* and EHEC specifically has resulted in significant insights into the evolution of these important pathogens, but with the advent of new sequencing technologies, we no longer sequence one or two prototype isolates; rather, we sequence “collections” or “outbreaks” to define the molecular markers of these pathogens. This article highlights where we started with the molecular characterization of *E. coli* and EHEC and where we are today with high-throughput technologies.

**RESERVOIRS FOR HUMAN HEALTH**

Cattle are recognized as a main reservoir of STEC O157:H7. The significant differences in host prevalence, transmissibility, and virulence phenotypes among strains from bovine and human sources are of major interest to the public health community and livestock industry (16). Genomic analysis revealed divergence into three lineages: lineage I and lineage II/III strains are commonly associated with human disease, whereas lineage II strains are overrepresented in the asymptomatic bovine host reservoir (17). Growing evidence suggests that genotypic differences between these lineages, such as polymorphisms in Shiga toxin subtypes and synergistically acting virulence factors, are correlated with phenotypic differences in virulence, host ecology, and epidemiology (18). To assess the genomic plasticity on a genomewide scale, the whole genome of strain EC869, a bovine-associated *E. coli* O157:H7 isolate, was sequenced. Comparative phylogenomic analysis of this key isolate enabled placement of the bovine lineage II strains within the genetically homogeneous *E. coli* O157:H7 clade (18). Identification of polymorphic loci that are anchored both in the chromosomal backbone and horizontally acquired regions allowed association of bovine genotypes with altered virulence phenotypes and host prevalence. Polymorphic markers are valuable in the development of a robust typing system critical for forensic, diagnostic, and epidemiological studies of this emerging human pathogen.

**TYPING AND GENETIC ANALYSIS OF EHEC OUTBREAKS WITH PREGENOMIC METHODOLOGIES**

Unlike other *E. coli* serotypes, the O157:H7 lineage is distinguished by its genetically highly homogeneous population structure, comparable to clonal microbial species such as *Yersinia pestis* (19) or *Bacillus anthracis* (20). With potentially lethal and widespread outbreaks, a large-scale and in-depth survey of genetic and architectural polymorphisms is a crucial prerequisite to obtain insights into the natural pathogenome evolution and extent of bacterial disease virulence genotypes. Genetic heterogeneity among O157:H7 EHEC strains has been established by using a broad panel of targeted- and whole-genome-based typing assays to determine diversity and evolutionary relationships among EHEC isolates, such as multilocus sequence typing (21), octamer- and PCR-based genome scanning, (22, 23), phage typing (24–26), multiple-locus variable-number tandem repeat analysis (27, 28), microarrays (29), microarray-based comparative genome hybridization (17), nucleotide polymorphism assays (19, 30, 31), pulsed-field electrophoresis (PFGE) (32), subtractive hybridization (33, 34), and optical mapping (12, 35, 36).

**Pulsed-Field Gel Electrophoresis**

PFGE has been widely employed as a molecular typing method in epidemiological investigations of EHEC (32, 37). According to differences in the XbaI PFGE patterns, EHEC O157:H7 isolates are classified into different types (types I to V and ND) (38). Alterations in PFGE patterns after restriction digest are results of genomic rearrangements driven by recombination of prophages and mobile elements (39, 40). PFGE, though considered a first-line molecular screening tool, provides insight into the genome structure and changes in restriction patterns and potentially identifies spontaneous genomic rearrangements or recombination of mobile elements (39, 40). In short-term epidemiological studies, it may be that the PFGE pattern will be useful as an inclusion or exclusion criterion when *E. coli* O157:H7 isolates are examined over a defined time window in an outbreak situation. However, even now, we are seeing the greater use of whole-genome sequencing for the investigation of outbreaks of *E. coli* (41–46).

**Multilocus Sequence Typing**

Sequenced-based methods, such as multilocus sequence typing (MLST), have been powerful subtyping tools in molecular epidemiology. These methods have the advantage of being easily standardized and automated. MLST was first developed for *Neisseria meningitidis* in 1998 to overcome the poor reproducibility between laboratories applying older molecular typing schemes (47). The principle behind the MLST scheme is to identify internal nucleotide sequences of approximately 400 to 500 bp in multiple housekeeping genes. Unique sequences (alleles) are assigned a random integer number, and a unique combination of alleles at each locus,
an “allelic profile,” specifies the sequence type. In this new era of high-throughput sequencing, it may be more rational to use whole-genome sequence data for typing. Two MLST schemes exist for *E. coli*: *E. coli* scheme 1, which employs seven genes (*adk, fumC, gyrB, icd, mdh, purA, recA*) (48), and *E. coli* scheme 2, which employs eight genes (*dinB, icdA, pabB, polB, putP, trpA, trpB, uidA*) (49). Though these MLST methods are generally congruent, there are some differences for some strains.

These MLST systems first determined that EHEC isolates are genetically highly clonal but could also be separated into multiple clades, suggesting that there are multiple evolutionary paths to generate a fully virulent EHEC isolate (50).

### Multiple-Locus Variable Number Tandem Repeat Analysis

As a result of the poor discriminatory ability of MLST for *E. coli* O157:H7, it was decided to target short tandem repeats, which are areas of the bacterial genome that evolve rapidly. Targeting of these elements, which often vary in number among different strains of the same species (the definition of a variable-number tandem repeat), has successfully been used to discriminate between strains of prokaryotes (51). Multiple-locus variable-number tandem repeat analysis involves determination of the number of repeats at multiple loci, thereby providing a powerful tool for assessing the genetic relationships between bacterial strains of the same species. Multiple-locus variable-number tandem repeat analysis has several advantages over PFGE because, like MLST, the output is highly objective, making the data amenable to automated computer analysis for the rapid detection of outbreaks and easy to compare across laboratories.

### Whole-Genome Mapping

Whole-genome mapping has been used for detailed genome comparisons to differentiate closely related *E. coli* O157:H7 strains based on alterations in the chromosomal architectures (insertions, deletions, and rearrangements). The sizing and positioning of lateral acquired genomic regions in EHEC are crucial in assessing the Shiga toxin virulence status determined by *Stx* allele prevalence and respective chromosomal insertion sites of the converting prophages (12, 35, 36). Comparative map analysis reveals valid biological markers to trace evolution and also assists in genome assembly for molecular epidemiology outbreak investigations (36).

### Genome Sequence-Based High-Resolution Genotyping

#### Octamer-based genome scanning typing assay

The rapid accumulation of whole-genome sequence data for O157:H7 has allowed the development of high-resolution subtyping methods that enable inter- and intraspecies bacterial genome comparisons. Sequence-based phylogenetic assays have determined that EHEC strains comprise three highly related but distinct populations with a global prevalence that differs in genotype and host ecology. Polymorphisms were identified using high-density octamer-based genome scanning (OBGS) analysis by testing multiple OBGS primer combinations in independent reactions screening a diverse strains set. Polymorphic OBGS products that were specific to lineage I or lineage II strains excluding the polymorphisms are not found within prophage, insertion sequences, or plasmid sequences. OBGS first demonstrated that the *E. coli* O157:H7 clonal complex had diverged into two highly related lineages, designated lineages I and II, that were found to be disproportionately represented among human and bovine isolates, respectively (22). Lineage II strains are found less frequently associated with human disease due either to inefficient transmission from bovine sources or to attenuated virulence in humans and the bovine host (26).

#### Lineage-specific polymorphism typing assay

Further analyses led to a refined classification system, termed the lineage-specific polymorphism assay (LSPA), that ultimately partitions *E. coli* O157:H7 strains into three lineages, I, I/II, and II, according to a PCR-based assay and polyacrylamide gel separation by testing the repeat length at six genic and intergenic chromosomal loci (52). In reference to the EDL933 genome, the LSPA markers comprise a 9-base insertion in gene *Z5935*, a 78-base insertion in the *yheG* gene, a 9-base deletion in the *rbsB* gene, a 9-base insertion in the *rtcB* gene, and an 18-base insertion in the intergenic region spanning the *arp-iclR* genes. These techniques showed divergence into three different but interlinked lineages; lineage I and lineage I/II strains are commonly associated with human infections, whereas lineage II strains are overrepresented in the asymptomatic bovine host reservoir. Genetic evidence also suggests that distributions of genotypes between these lineages, such as polymorphisms in Shiga toxin subtypes and synergistically acting virulence factors, are correlated with phenotypic differences of...
major biological relevance in virulence, host ecology, and epidemiology. Identification of polymorphic loci that are anchored both in the chromosomal backbone and in horizontally acquired regions allowed us to associate bovine genotypes with altered virulence phenotypes and host prevalence. Numerous novel lineage II-specific genome signatures, some of which appear to be intimately associated with the altered pathogenic potential and niche adaptation within the bovine rumen and further discriminate bovine super shedders, have been cataloged (18, 26, 36).

**Whole-genome sequence typing/single nucleotide polymorphism typing**

Despite multiple genomes of this lineage having become available in the genomics era (14, 15, 18, 36, 53, 54), the biological insight into epidemiology and disease mechanism suffers from a lack of markers for accurate typing and genotype/phenotype association. High-resolution phylogenetic approaches allow the dynamics of pathogenome evolution to be followed at a high level of phylogenetic accuracy and resolution. Whereas the current molecular markers and typing assays used by public health microbiology laboratories may be adequate for routine surveillance and identification of *E. coli* O157:H7, these approaches lack the polymorphic markers and discriminatory power to study the relatedness of strains of unknown provenance, which becomes of particular importance when investigating outbreak strains that form clonal complexes with only a few genetic polymorphisms. With the increase of next-generation sequencing technologies, whole-genome sequences typing approaches, such as the discovery of single nucleotide polymorphisms (SNPs), have gained popularity. SNP typing not only provides stable genetic markers to study evolution but also offers greater phylogenetic resolution. SNP discovery approaches have yielded thousands of high-quality SNPs as critical bases for the development of a refined phylogenetic framework (36, 55, 56). SNP typing allowed the elucidation of the evolutionary origin and emergence of the pathogenic O157:H7 lineage, and the determination of the genetic relationships to the intermediate immobile O157:H(−) and ancestral EPEC O55:H7 serotypes (56). A SNP-based clade typing assay that detects SNPs in 96 loci has been applied to more than 500 clinical *E. coli* O157 strains. This resulted in a refined LSPA-6 lineage classification and further separated EHEC isolates into nine distinct clades. The frequency and distribution of Shiga toxin-converting prophages and form of clinical disease manifestation could also be elucidated (30, 57).

Recently, Manning et al. (30) used SNP analysis to identify a group (clade 8) of hypervirulent *E. coli* O157:H7 strains found in the United States. In 2006, food-borne outbreaks involving O157 contamination of fresh produce (e.g., spinach) were associated with more severe infection, causing higher rates of HUS and more frequencies of hospitalization, demonstrating that increased virulence had been acquired (30). In silico scoring of SNPs was successfully deployed by Eppinger et al. to investigate the degree of genetic heterogeneity in strains derived from a single outbreak of human disease (36) and to establish biologically relevant markers among strains from clinical settings and the animal reservoir (36). These enriched mutational database resources will provide a robust foundation to better associate genotypic group profiles and virulence phenotypes within *E. coli* O157:H7.

**HISTORICAL VIEW OF EHEC/STEC GENOMICS**

Whereas *Helicobacter pylori* was the first organism to have multiple isolates sequenced (58, 59), EHEC was the first *E. coli* pathotype with genomes of multiple isolates available for comparison (14, 15). Considering the relatedness of these isolates, it can be argued as the true beginning of comparative genomics as we know it. Irrespective of when comparative genomics began, the scope of the analyses has changed over the years. We have transitioned from sequencing prototype isolates from a group of pathogens to sequencing large numbers of isolates in an effort to understand the population structure of the pathogen associated with human health and to explore the concept of genomic epidemiology. This section briefly covers some of the important publications on genomics associated with *E. coli* O157 isolates.

**First View of the Interpathotype Comparisons: EDL933 and Sakai Genomes**

The first genome of *E. coli* was published in 1997 by the Blattner group and signaled a change in our understanding of model organisms (13). It could be argued that the sequencing of the MG1655 genome was the first nonpathogen genome to be decoded, as it had long been known that this isolate could no longer colonize humans or cause disease in humans. The sequencing of the EDL933 genome (14) provided the opportunity to examine two genomes from the same species. These early comparisons identified that there was >1 Mb of unique genomic material, encoding 1,387 genes, associated with...
the EDL933 genome that was not in the MG1655 genome. It was assumed that the majority of this DNA, present in what was termed “O-islands,” would be associated with the ability to cause disease. Although it was true that the majority of genes that had been identified to be associated with human disease were in these O157-specific regions, including Shiga-toxin phage and the LEE region, it was clear that not all of these regions were directly associated with disease and could be considered part of the central metabolism of these organisms. It was also noted that the genome of the EDL933 isolate contained an ~400-kb chromosomal inversion when compared to MG1655 genome. It is still unclear if this genomic inversion contributes to the pathogenesis of EDL933. One key point that was highlighted in this first genomic comparison was the identification of 18 identifiable prophage regions ranging in size from ~7 kb to 62 kb that were designated BP-933[A-Q]. These prophage regions include a number of potential secreted effectors (60). At this point, a limited number of genomes were available (<20 complete and draft genomes), and comparisons were limited to mostly prototype isolates. The study by Rasko et al. highlighted that there were significant genetic differences between the EHEC isolates, but this observation could also be attributed to the fact that the annotation of different genomes could have a significant impact on the findings that were gene based (62).

Additional interesting genomes of the O157 serotype were published by Kulasekara et al. in 2009 (63); they sequenced an isolate, TW14359, that was associated with HUS. In this study they identified an additional 70 kb of genetic material that was not present in either of the reference O157 isolates. These regions encoded additional putative Type III secreted effectors and a gene for an anaerobic nitric oxide reductase, norV. It was suggested that the norV gene could be used as a marker for increased virulence leading to HUS; however, screening of large numbers of isolates suggested that the norV gene was associated with the loss of stx1, but the direct impact on pathogenesis was not confirmed. This publication is important in that it attempts to meld comparative genomics with disease presentation and gene presence or absence. This can be considered the beginning of the field we now know as genomic epidemiology.

Also in 2009, the first non-O157 EHEC genomes were published when Ogura et al. (53) published the genomes of isolates from serotypes O26, O111, and O103. These genomes were compared to the Sakai genome. This study demonstrated that the EHEC genomes were routinely larger than the MG1655 genome and contained a diverse array of phage and integrative elements, sometimes associated with the virulence gene catalog. Most importantly, it demonstrated how different serogroups and lineages of E. coli could evolve into EHEC isolates and demonstrated that there was a genomic core among similar isolates by comparing the gene content of 345 orthologous genes in all E. coli isolates that were sequenced to date. This is similar to the presentations in Fig. 2 and Fig. 3 in this article; however, we now avoid the problems caused by calling genes and curation errors by using the unannotated whole genome sequence for comparison (see below). Additionally, the study demonstrated that not all features of the genomes were the same. For example, the LEE pathogenicity island was
inserted at different locations in the genome in these isolates, but it appeared to be functional as all isolates were derived from individuals who were ill. As the cost of sequencing decreased, there started to be studies that examined large numbers of isolates. With EHEC, these large-scale sequencing projects were often linked to outbreaks of food-borne disease in the mid-2000s.

How Outbreaks Have Shaped the Sequencing of EHEC
From the early days of genomics, the majority of the focus has been on the EHEC O157:H7 isolates associated with human disease. In recent years, with the decrease in the cost of sequencing we have observed an increase in sequencing a significant number of isolates associated with disease within an outbreak setting. Some
of these studies have attempted to associate genetic changes with increases in virulence or markers that laboratories could use as identifiers of outbreak strains. The study by Eppinger et al. describes the sequencing of a large number of isolates (36) to reconstruct the “anatomy of the outbreak.” In this study the authors sequenced 25 isolates and through comparative genomics identified mutational and structural markers in the genome core and mobilome that allowed distinguishing these outbreak isolates among three almost simultaneous outbreaks that occurred in the United States as well as other O157 isolates that have been

**FIGURE 3** Phylogeny of O157 *E. coli* reference isolates demonstrating that there is variation within the O157 clade that is not observed in the larger genomic comparisons. This multi-whole genome alignment contains 4,093,272 bp of sequence. The tree is maximum likelihood with 100 bootstraps made using RAxML, as previously described (68). The color indicates distinguishable clades within this selection of 50 EHEC isolates. The additional numbers in parentheses are the clade and LSPA designations, as described in the text.

sequenced with other outbreaks and clinical sources as well as from animal reservoirs. Using a panel of >1,200 SNPs they were able to identify specific lineages of *E. coli* O157:H7. This type of study demonstrates how the genomics of a pathogen can distill the identification to a very fine scale. This is possible among the O157 isolates, since genomically they are all very similar when compared to the other *E. coli* strains (Fig. 2); however, there are significant differences when only the O157 isolates are compared (Fig. 3).

The *E. coli* outbreak in Europe in 2011 provided an opportunity to highlight the use of genomics to examine the bacterial isolates in near real time (41–44, 46). Additionally, the results provided insight into the evolution of bacterial pathogens as a species and challenged the definitions and boundaries of the STEC pathovar. The genome sequencing of the isolates associated with this outbreak and comparison to isolates that were in the database identified that the majority of the chromosome of the outbreak-associated isolate was most similar to EAEC isolate 55989 (61). There were significant changes in the genome in that it had acquired the Shiga-toxin phage, as well as an antimicrobial resistance plasmid. Neither of these features had been previously identified in EAEC isolates, but both potentially contribute to the pathogenicity of the bacteria. Thus the definition of the isolate was technically STEC, but the majority of the gene markers identified pointed toward an EAEC isolate (64). Many names were provided to this new type of isolate with this collection of virulence factors, including “entero-aggregative-haemorrhagic” *E. coli*, (46) but the most appropriate name proposed would be entero-aggregative, Shiga toxin-producing *E. coli* O104:H4. The isolates in this outbreak highlighted the issues with the current nomenclature that is largely based on the identification of virulence-associated genes that are present on mobile elements.

The genomic studies also highlighted that the determination of the presence or absence of specific genes or gene combinations could not really predict the virulence of an isolate and that further functional characterization is required. In addition to a novel collection of virulence factors in the European outbreak isolates, it was demonstrated that, like EHEC isolates, the Shiga toxin genes were activated by the exposure to subclinical doses of antibiotics (42). As time goes on, more detailed studies of the functional virulence of these isolates will be examined and the true impact of the novel combinations of virulence factors will be identified.

**GENOMICS-GUIDED EXAMINATION OF PATHOGENESIS**

The study of the evolution of pathogenesis in EHEC has mainly focused on the traditional virulence factors, such as LEE, Shiga toxin, a limited number of adhesins, and plasmid encoded features. Molecular evidence suggests that EHEC isolates have acquired the majority of their virulence factors through horizontal gene transfer and thereby acquisition of the LEE pathogenicity-associated island and the Stx genes were two crucial steps in the evolution of EHEC O157 from a commensal ancestor (53). The LEE pathogenicity-associated island is clearly a mosaic structure, which arose from multiple recombination events with foreign DNA, as did the large EHEC-hemolysin plasmid (65, 66). LEE can be found on chromosomes of EHEC next to tRNA genes at different locations (67), suggesting that LEE has been acquired on more than one occasion. Additionally, other virulence factors found on the O157 genomes, such as fimbrial and nonfimbrial adhesins, iron uptake systems, and non-LEE effectors, are also thought to be required for the full virulence of EHEC, but their prevalence among non-O157 EHEC isolates, and in some cases among the O157 EHEC isolates, has not yet been systematically analyzed. Global genomic differences (or conservation) between O157 and non-O157 EHEC strains have recently been addressed in the study by Hazen et al. (68), which demonstrated that the O55 and O157 isolates formed closely related clades, but the non-O157/non-O55 EHEC (also known to some as EHEC2) were divergent. We demonstrate this point in Fig. 2, using a collection of reference *E. coli* isolates from each of the pathovars as well as a selection of the O157, O55, and non-O157 EHEC isolates that have been recently sequenced. This clearly demonstrates that although the virulence factor typing has grouped these isolates together, pathogenomic analysis suggests they are significantly different.

That there has been a parallel evolution and independent acquisition of the virulence factors in each of the serogroups (69) has been proposed. Hazen et al. (68) recently noted that there appears to be a pattern of secreted effectors that are common in EHEC strains when compared to other EAEC isolates. This suggests that there is a complement of effectors that are “tuned” to the EHEC genome content for optimal regulation, interaction, and secretion; however, this does not necessarily have to be linked to human virulence and further characterization of this orchestrated program is required.
Are There Genomic Differences in Recent Isolates That Have Come from Green Leaf Vegetables Rather Than Most of the Original Isolates That Were Beef Related in Some Way?

Recent Shiga-toxigenic E. coli O157:H7 outbreaks have been linked to consumption of fresh produce (12, 63). It is generally recognized that bacterial attachment to vegetal matrices would constitute the first step in contamination of fresh produce, but how and when this occurs are under some debate. Cellular appendages, such as curli fibers, and cellulose, a constituent of extracellular matrix, have been suggested to be involved in E. coli attachment and persistence in fresh produce. A comparative evaluation was conducted on the ability of STEC O157:H7 strains EDL933 and 86-24, linked to two independent food-borne disease outbreaks in humans, and their mutants deficient in curli and/or cellulose expression to colonize and to firmly attach to spinach leaf. Inoculated spinach leaves were incubated at 22°C, and at 0, 24, and 48 h after incubation loosely and strongly attached E. coli O157:H7 populations were determined. Curli-expressing E. coli O157:H7 strains developed stronger association with leaf surface, whereas curli-deficient mutants attached to spinach at significantly (P<0.01) lower numbers. Attachment of cellulose-impaired mutants to spinach leaves was not significantly different from strains that contained curli; however, the relative attachment strength of E. coli O157:H7 to spinach increased with incubation time for the curli-expressing strains. Laser scanning confocal microscopy analysis of inoculated leaves revealed that curli-expressing E. coli O157:H7 was surrounded by extracellular structures strongly immunostained with anticurli antibodies. Production of cellulose was not required to develop strong attachment to spinach leaf. These results indicate that curli fibers are essential for strong attachment of E. coli O157:H7 to spinach whereas cellulose is dispensable.

PUBLIC HEALTH VIEW—WHAT ARE THE NEXT STEPS?

Whole bacterial genome sequencing in medical microbiology is fast becoming a reality; however, the challenge of converting the primary sequence data into useful clinical or public health action remains unmet, because experience with such data is limited. For the identification of an isolate or an outbreak strain, comparison with genomes of very closely related organisms is required. What is clear is that a large sequence database comprising a comprehensive panel of well-chosen isolates is necessary if genome sequencing is to be useful in the field. Perhaps the most comprehensive single-species genome data available thus far are for E. coli. This database became extremely useful when the outbreak in Europe occurred in 2011 (41–43). This outbreak resulted in 3,816 identified STEC infections and 54 deaths (45). It was rapidly determined that the etiological agent was E. coli, but the isolate had features of EAEC and was Shiga toxin positive but lacked the majority of features associated with EHEC (LEE, ehx, etc.), suggesting that this isolate may simply be an EAEC isolate that had acquired the Shiga toxin phage. Rapid comparative genomics highlighted that this was the case and that the isolate had also acquired an antibiotic resistance plasmid. That we had large collections of isolates already sequenced allowed the accurate and rapid determination of the genomic origin of this isolate. In the future these databases of isolate collections will become invaluable in the public health setting in transmitting information from the sequencer to the physician treatment paradigm for any patient. We are still far away from this reality, as the software for these analyses are somewhat custom, but as the benchtop sequencers become more prevalent in clinical laboratories, the methodology must follow.

CONCLUSION

The use of modern techniques like whole-genome sequence typing can benefit researchers interested in surveillance (18), forensics (70), diagnostics, and epidemiological studies of microbial pathogens and patho-genome evolution, and holds the potential to discover novel E. coli O157:H7 virulence genotypes that are intimately associated with adverse human disease outcomes. Use of these techniques is not limited to E. coli O157:H7 or EHEC as this research can provide principles and valuable benchmarking data crucial in the modeling of bacterial outbreak-associated population dynamics. Once data are available, the scientific community will be key to the analysis of the gathered sequence data. Application of postgenomic tools to the generated data can help to understand variability in gene content and activity. These technologies include DNA microarray analysis or RNAseq for gene profiling and differentiation of strains; optical mapping of chromosomes for discovery of novel gene insertions and deletions; and phenotypic microarray of pathogens’ physiological and metabolic capabilities for strain characterization. Virulence assays in host model systems allow validation of genetic predictions and association of isolate-specific gene content to outbreak-associated...
physiological and virulence capabilities. Better and more effective schemes for epidemiological studies can be developed to trace the source of the outbreaks, or ultimately link the source to the infection in outbreak investigations. The identification of subtle but defined polymorphisms would allow us to investigate the drivers of pathogenome evolution and analyze mutations that are correlated with human disease and severity. This will enable us to track and distinguish separate unrelated outbreaks and classify isolates as members of an outbreak population, and also to refine standards currently used in outbreak investigations. The potential discovery of novel polymorphisms complements current techniques used to classify strains, and is crucial in the development of rapid methods of detecting and tracking microbial outbreaks. Studies of non-O157 EHEC are currently under way by several independent research teams. These data sets will potentiate each other in gathering a more complete and refined picture of the E. coli O157:H7 pathogenome evolution and its pathogenic potential in relation to the E. coli species biology.

ACKNOWLEDGMENTS

D.A.R. and T.H.H. are supported by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under grant number U19AI090873 and funds from the State of Maryland. M.E. is supported by the Department of Biology and South Texas Center for Emerging Infectious Diseases, University of Texas at San Antonio. This review is further based upon work at the University of Texas at San Antonio supported by the Army Research Office of the Department of Defense under Contract No. W911NF-11-1-0136. We thank Sara S. K. Koenig for critically reading the manuscript and all research groups who have produced data discussed in this work.

REFERENCES


33. Sadiq et al. Escherichia coli Genomic anatomy of


36. Eppinger M, Mammel MK, Pacheco AG, Boxrud DJ, Harrison LH.


