UNIVERSITY OF CALIFORNIA, MERCED

Using Complete Genome Sequences to Predict Important Phenotypes

A thesis submitted in partial satisfaction of the requirements
for the degree of Master of Science

in

Microbiology Genomics

by

Heliodoro Cárdenas

Committee in charge:

Dr David Ardell, Chair
Dr Miriam Barlow
Dr Pilar Francino
Dr Monica Medina

2014

Signature Page

The Thesis of Heliodoro Cárdenas is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Dr. Miriam Barlow

Dr. Monica Medina

Dr. Pilar Francino

Dr. David Ardell, Chair

Date

University of California, Merced
2014

Dedication

In recognition of my family, I dedicate my work to my parents Raul and Ofelia Cárdenas, my grandfather Pablo Cárdenas, who I never had the opportunity meet, but always dreamt of a family member of reaching a higher education degree, my grandmother Maria Espinoza who was always a great inspiration with her tenderness and love, and my beautiful family; my wife Paulina and kids, Julian, Zoe, and Chloe. I am eternally grateful for your love and care. I love you.

Epigraph

"Success is not measured by what you accomplish, but by the opposition you have encountered, and the courage with which you have maintained the struggle against overwhelming odds."

Orison Swett Marden

Table of Contents

List of Abbreviations

AIEC – Adherent-invasive *Escherichia coli*
B. *anthracis* – *Bacillus anthracis*
BLAST – Basic Local Alignment Tool
bp – base pairs
DNA – Deoxyribonucleic acid
EAEC – Enteroaggregative *Escherichia coli*
E. coli – *Escherichia coli*
EHEC - Enterohemorrhagic *Escherichia coli*
EIEC – Enteroinvasive *Escherichia coli*
EPEC – Enteropathogenic *Escherichia coli*
ETEC – Enterotoxigenic *Escherichia coli*
ExPec – Extraintestinal pathogenic *Escherichia coli*
MSA - Multiple Sequence Alignment
M. turberculosis – *Mycobacteruim tuberculosis*
MST – Minimum Spanning Tree
NCBI – National Center for Biotechnology Information
NHI – National Institutes of Health
NR – Non-redundant
pINV – Invasiveness plasmid
PCR - Polymerase Chain Reaction
rpm – revolutions per minute
SNP – Single-nucleotide polymorphism

List of Figures

List of Tables

Acknowledgements

I would like to acknowledge Dr. Miriam Barlow for her unconditional support as my graduate advisor and mentor. Through several drafts and many long nights, her mentorship has been invaluable.

I would also like to acknowledge Dr. David Ardell for his support as the chair of my committee. His support had a great impact on my experience as a graduate student.

I would also like to acknowledge Dr. Pilar Francino and Dr. Monica Medina for their help and support in my research work.

Chapter 1 and 2, in full, is a reprint of the material as it appears in PLoS ONE 2013. Hall BG, Cardenas H, Barlow M, 2013.

Vita

1994	Associate in Arts, Chabot Community College, Hayward, CA
1998	Bachelor in Arts, University of California, Santa Cruz, CA
1997 - 1998	Sequencing Specialist, Incyte Pharmaceuticals Inc., Palo Alto, CA
1998 - 2007	Lawrence Berkeley National Laboratory, Berkeley, CA
2007 - Present	Director of Core Genomics Facility, University of California, Merced, CA
2012 - Present	Teacher Assistant, University of California, Merced, CA
2014	Masters in Science, Universtiy of California, Merced, CA

PUBLICATIONS

Hall BG, Cárdenas H, Barlow M. Using Complete Genome Comparisons to Identify Sequences Whose Presence Accurately Predicts Clinically Important Phenotypes. PLoS ONE 2013; 8(7): e68901. doi:10.1371/journal.pone.0068901

International Human Sequencing Consoritum (2001). The Human Genome. Nature 2001; 409(861-921).

Abstract of the Dissertation

In clinical settings it is often important to know not just the identity of a microorganism, but also the danger posed by that particular strain. For instance, *Escherichia coli* can range from being a harmless commensal to being a very dangerous enterohemorrhagic (EHEC) strain. Determining pathogenic phenotypes can be both time consuming and expensive. Here we propose a simple, rapid and inexpensive method of predicting pathogenic phenotypes on the basis of the presence or absence of short homologous DNA segments in an isolate.

Our method compares completely sequenced genomes without the necessity of genome alignments in order to identify the presence or absence of the segments to produce an automatic alignment of the binary string that describes each genome.

Analysis of the segment alignment allows identification of those segments whose presence strongly predicts a phenotype. Clinical application of the method requires nothing more that PCR amplification of each of the set of predictive segments.

Here we apply the method to identifying EHEC strains of *E. coli* and to distinguishing *E. coli* from *Shigella.* We show *in silico* that with as few as 8 predictive sequences, if even three of those predictive sequences are amplified the probability of being EHEC or *Shigella* is >0.99. The method is thus very robust to the occasional amplification failure for spurious reasons.

Experimentally, we apply the method to screening a set of 98 isolates to distinguishing *E. coli* from *Shigella*, and EHEC from non-EHEC *E. coli* strains and show that all isolates correctly identified.

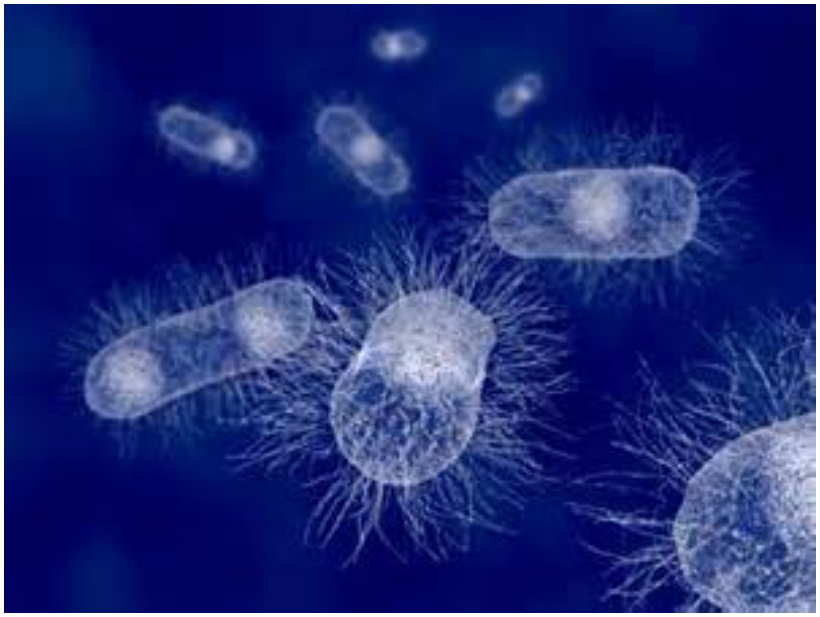
Chapter 1: Introduction to Historic Importance of Genomic Sequencing

The discovery of the double helix structure of DNA in 1953 by James Watson and Francis Crick was a significant moment in the history of biology that later propelled the scientific world into the challenge of deciphering the genetic code of life. The human genome project was initiated by the U.S. government in 1990 with a 15-year plan and a $3 billion budget to unravel the 3 billion letters of DNA that compromise the human genome. The great significance of the project undertaken by the government later become a fierce competitive race between a private company, Celera Genomics, directed by J. Craig Venter whose main focus was to patent genes for the sole purpose of creating a business revolution. Despite the scientific confrontation between the two sectors, the aggressive sequencing race ended in June 2000 when scientists declared the completion of a draft sequence of the 3.2 billion DNA units. An amicable consensus between the government and private parties was reached to make the genomic data public with the intervention of President Bill Clinton, who followed with a ceremony at the White House to announce the greatest scientific milestone aside from man landing on the moon.

President Bill Clinton (center), J. Craig Venter (left), and Francis Collins (right)

Genomic sequencing has contributed immensely to the genetic field, from gene discoveries to species identity. The latter is of great importance because many organisms share a similar genotype, but have very distinct phenotypes. *Shigella* and enterohemorrhagic *Escherichia coli* (EHEC) are two bacterial species that represent a clinical challenge to identify. Like many other similar genotype organisms, these two species share a large number of genes (core genes) and some genes that are present in one strain, but not the other (accessory genes) (Hall BG et al., 2010; Hiller NL et al., 2007; Hogg JS et al., 2007; Tettelin H et al., 2005). This genetic variation in bacterial species is the result from massive rearrangements coupled with massive gain/loss of large DNA

segments (Hall et al., 2010). The large amount of bacterial genomic data available has allowed for the phylogenetic analysis of E. coli leading to the understanding of its evolution and adaptation. When comparing evolutionary phylogenies and similarity clusters, a concept emerges that *Shigella* may have originated from at least three distinct ancestors and evolved into a single clade via horizontal gene transfer (Zhang et al., 2012).



Shigella sonnei

Predicting the phenotypes of *Shigella* and EHEC by determining unique sequence identification within these bacterial enteric pathogens can be essential. Determining pathogenic phenotypes can be both time consuming and expensive. Culture-based detection of bacterial enteric pathogens is time consuming and may require up to 96 hours to generate a definitive result. However, the use of molecular detection methods in combination with overnight faecal enrichment has the potential to reduce the time to diagnosis by at least 50% (O'Leary et al., 2009).

Enterohemorrhagic *Escherichia coli*

Identifying unique genomic sequences from each phenotype and creating a PCR assay can be a rapid diagnostic test. Using a novel method for identifying unique genomic sequences called the "Bop method", a total of 8 unique sequences were identified for both *Shigella* and EHEC from a pool of 47 E. coli and 8 *Shigella* completely sequenced genomes. The purpose of this study was validate the accuracy of identifying those unique segments whose presence strongly predicts a phenotype and creating an efficient PCR assay to correctly identify either *Shigella* or EHEC.

Chapter 2: Enterohemorrgahic *E. coli* and *Shigella* Identification

2.1 Introduction

Clinically important bacterial phenotypes can be difficult, expensive and time-consuming to determine. Phenotypic assays require expensive reagents and growth of the bacteria. PCR assays, which detect genotype, do not always correctly indicate phenotype. However, genotype and phenotype in organisms as simple as bacteria should have strong correlations, and whatever disconnect exists between genotype and phenotype likely results from an incomplete capture of genotype in which too few genetic features are considered.

Most bacterial species are characterized by a "pan-genome" in which there is a set of genes that are present in all members of the species (core genes) and a large set of genes each of which is present in some, but not all, members of the species (accessory genes) (Hall BG et al., 2010; Hiller NL et al., 2007; Hogg JS et al., 2007; Tettelin H et al., 2005). The major fraction of variation among bacterial genomes of the same species derives not from base substitutions, but from massive rearrangements coupled with massive gain/loss of large DNA segments. The fraction of the genome that is accessory genes ranges from ~8% (*B. anthracis, M. tuberculosis*) to ~35% (*E. coli*) (Hall, unpublished results based on analysis of 22 species using methods described in (Hall BG et al., 2010).

That variance in genetic content makes bacterial genomes particularly difficult to compare because genome comparison requires comparing homologous DNA sequences. To ensure comparisons among homologous bases, genes are typically aligned by one of a variety of multiple sequence alignment (MSA) methods that introduce into the alignment gaps that are intended to represent historical insertions or deletions (indels). That approach reflects the assumptions that individual genes evolve primarily by base substitutions and indels, and it works well for sequences which meet those assumptions. MSA is not robust when those assumptions are violated. MSA methods generally fail when inversions, transpositions and many indels occur.

To overcome those difficulties in bacterial genome alignment, we have developed and applied a novel approach which we have named the “Bop” method (Hall BG et al., 2013). The Bop method produces a description of each genome as a binary string that indicates the presence or absence of each bop that is to be found among the strains that are analyzed. Because bops are short homologous sequences, the set of binary strings constitutes an automatic alignment of the set of genomes with respect to the presence/absence of those bops. The bop alignment can be used directly to estimate relationships among the strains.

The most common way to estimate the relationships among organisms is by phylogenetic analysis, but phylogenetic analysis is not always appropriate for the set of organisms being compared or for the data that is used to characterize those organisms. Phylogenetic trees are used to estimate the relationships of organisms to hypothetical ancestors, and thereby to each other. The branches on a phylogenetic tree are intended to reflect the number of mutations that separate an organism from its hypothetical ancestor,

or one of those ancestors from its immediate ancestor. There are two fundamental assumptions involved in phylogenetic analysis, neither of which applies to these data sets. First, characters that are shared between a pair of individuals are assumed to be identical by descent. Deviations from that assumption, collectively called *homoplasies*, may arise from convergence, incomplete genetic isolation, etc. and are indications of loss of phylogenetic signal. For these data there is no implication that shared characters are identical by descent. A pair of strains that have the same bop may do so because they inherited that bop from a common ancestor, because of exchange with a distantly related individual, or because each has acquired the same plasmid or phage-borne bop. Similarly, individuals that lack a particular bop may do so because of inheritance or because each has independently suffered loss of that bop. The nature of the data makes the use of phylogenetic analysis inappropriate. Second, phylogenetic analysis assumes that the individuals are genetically isolated from each other; i.e. they do not exchange genetic information and inheritance is strictly vertical. As a result a node, internal (ancestral node) or external (extant individual), can have only one ancestor. That assumption certainly does not apply to most microbial species where genetic exchange among individuals is common, and in particular does not apply to *E. coli* which is known to undergo considerable genetic exchange (Wirth T et al., 2006). These considerations indicate that phylogenetic analysis is not appropriate for estimating the relationships among these genomes.

A more appropriate approach to this problem is a *spanning tree,* which is a subset of a fully connected graph in which there is a single path from any node to any other node. A *minimum spanning tree* (MST) is the shortest spanning tree of all the possible spanning trees. Depending on the order in which the nodes are considered it is possible for there to be more than one MST (Salipante SJ and Hall BG, 2011). MSTs are widely used in microbial epidemiology to represent relationships among strains. MSTs make no assumptions about identity by descent or the absence of genetic exchange. MSTs are based only on identity by state, thus the relationships that are diagramed only indicate the overall similarities among the individuals.

To obtain an MST, the bops are combined into segments, where a segment is a contiguous series of bops that have identical distributions among the set of genomes, and a new binary string is written that describes the presence/absence of each segment in the genome.

The segment alignment can then be used to determine how tightly each segment is associated with a phenotype of interest. Segments that are always present (or always absent) in strains with the phenotype of interest, and are always the opposite when the phenotype is not expressed are candidates for amplification by PCR to estimate the probability that an unknown strain exhibits that phenotype.

Given the abundance of complete genome sequence data for numerous strains expressing the same phenotype, we thought it might be possible to identify multiple genetic markers and to establish the probabilities of certain phenotypes being expressed based on the presence or absence of those genetic markers. We assumed that many clinically important phenotypes are determined less by variation within homologous sequences (SNPs) than by the presence or absence of accessory genes within the genome.

Identification of accessory sequences associated with pathogenic phenotypes requires comparison of genomic sequence obtained from strains whose phenotypes are known.

We used complete genome sequences in order to identify sequences whose presence is strongly associated with difficult-to-determine phenotypes. Once identified, we reasoned that amplification of such sequences would provide a rapid, reliable and inexpensive means of assessing the probabilities of those phenotypes. We have applied this method to two clinically important species of bacteria, *Escherichia coli* and *Shigella.*

Escherichia coli K12 was among the first bacteria to be completely sequenced (Blattner FR et al., 1997). Its historic role as the laboratory strain that was the center of the development of molecular biology fully justified sequencing its genome. Long regarded as a benign commensal, that perspective changed in 1982 when an O157:H7 enterohemorrhagic (EHEC) strain was shown to be responsible for the "Jack-in-the-box" outbreak that resulted in several deaths. Sequencing an O157:H7 strain (Hayashi T et al., 2001) surprised the microbial community by revealing the dramatic differences in gene content and gene arrangement between the two sequenced genomes. *E. coli* is now well understood to vary enormously with respect to pathogenicity, and a variety of pathogenic phenotypes have been described including enterotoxigenic (ETEC), adherent-invasive (AIEC), enteroaggregative (EAEC), enteropathogenic (EPEC), and extraintestinal pathogenic (ExPec). Other strains are recognized as non-pathogenic commensals, of which five (K12 and its derivatives, B and its derivatives, C, W, and "Crookes") are considered "safe" strains for general laboratory use and are classified as Risk Group 1 (Archer CT et al., 2011). As of November 2011, 46 *E. coli* strains have been completely sequenced.

Enterohaemorrhagic *E. coli* (EHEC) causes serious symptoms including lower gastrointestinal bleeding, diarrhea and colitis. The clinical importance of these strains, and the need to track outbreaks and epidemics of these pathogens means that properly identifying them is important. However current clinical assays often fail to tell them apart because they are closely related, and symptomatically similar. Because *E. coli* varies so much in pathogenicity it is often not sufficient for public health officials to simply determine whether or not *E. coli* is present, it is often important to determine the danger posed by the strains that are present. Serotyping is an important tool in evaluating risk (O157:H7 strains can be presumed to be EHEC and therefore very dangerous), but other serotypes are also EHEC and it is not necessarily the case that all members of a particular serogroup would be EHEC. Determination of the EHEC phenotype is neither rapid nor cheap. O serotyping is performed following the procedure published by Orskov et al. (Orskov I et al., 1977). H typing is performed by the method described by Machado et al. (Machado J et al., 2000).

It would be useful to identify DNA sequences that correlate very strongly with the EHEC phenotype in order to develop a PCR assay that could quickly determine the probability that a given strain is EHEC. It is well understood that many of the virulence determinants associated with the EHEC phenotype are plasmid borne, but although some plasmids are shared by EHEC strains, none is shared by all.

Shigella is another serious pathogen that infects the digestive tract and can cause abdominal cramping, lower gastrointestinal bleeding, diarrhea, and colitis and severe dehydration. *Shigella* has been treated as a separate genus because of its clinical

pathogenicity resulting in shigellosis, but it has long been considered to be part of *E. coli* (Ochman H et al., 1983; Rolland K et al., 1998). At this time eight *Shigella* strains have been completely sequenced. *Shigella* is a clinically important clade within *Escherichia coli* (Sims GE and Kim SH, 2011) that causes shigellosis (Bardhan P et al., 2010). Distinguishing *Shigella* from other *E. coli* is non-trivial, and it would be valuable to have molecular markers that would unambiguously distinguish the two.

Here we apply the Bop method to the analysis of 47 *E. coli* and 8 *Shigella* completely sequenced genomes to determine their similarity and to identify genetic sequences that correlate strongly with the phenotype of species identity for EHEC *E. coli* and *Shigella*, with the intent of creating a reliable PCR assay for rapidly identifying them.

2.2. Results

2.2.1. BopGenomes analysis of E. coli –Shigella genomes

The 47 E. coli genomes and 8 Shigella genomes were analyzed using BopGenomes. Supplementary table S1 lists the GenBank accession numbers, pathogenicity phenotypes and references for those phenotypes. Depending upon the intended downstream application of the analysis, one might include the plasmids in each strain or exclude them from the analysis. The 55 genomes were analyzed both ways. In each case the genomes were digested in silico by restriction enzyme NcoI. When plasmids were excluded, the 55 genomes included 17,469 unique restriction fragments and comprised 69,010 unique bops. When plasmids were included, there were 18,193 unique restriction fragments and 76,517 bops.

2.2.2. Clustering of E. coli-Shigella genomes based on bops

Minimum spanning trees (MST) based on the presence or absence of each bop are shown in Figures 1 and 2. For the analysis shown in Figure 1, we excluded plasmids. For Figure 2, plasmids were included. Just as a phylogenetic tree is a graph that illustrates the relationships between individuals and their hypothetical ancestors based on identity by descent, an MST is a graph that illustrates the relationships between individuals based on identity by state. On an MST each node represents an individual and nodes are connected by edges whose lengths reflect the distance between the nodes. In this case the distance between a pair of genomes is shown as the number of differences in the state of the bop (0 or 1) divided by the number of bops. A spanning tree is a subset of a fully connected graph in which there is a single path from any node to any other node. A minimum spanning tree (MST) is the shortest spanning tree of all the possible spanning trees. Depending on the order in which the nodes are considered it is possible for there to be more than one MST (Salipante and Hall 2011), but in both of these cases there was a single MST. MSTs are widely used in microbial epidemiology to represent relationships among strains.

Figures 1 and 2 have been colored to indicate the pathogenic phenotypes when those are known. In general phenotypes tend to cluster together. We define a cluster as a set of genomes such that there is a path from each member of a cluster to every other member of that cluster, and the path does not pass through any node not belonging to that cluster. When plasmids are excluded the EHEC strains fall into a single cluster, whereas

when plasmids are included they fall into two closely related clusters, one consisting of the four O157:H7 strains. In both data sets the Shigella strains fall into a single cluster, and 13 or 14 of the 16 the commensal strains fall into a single cluster.

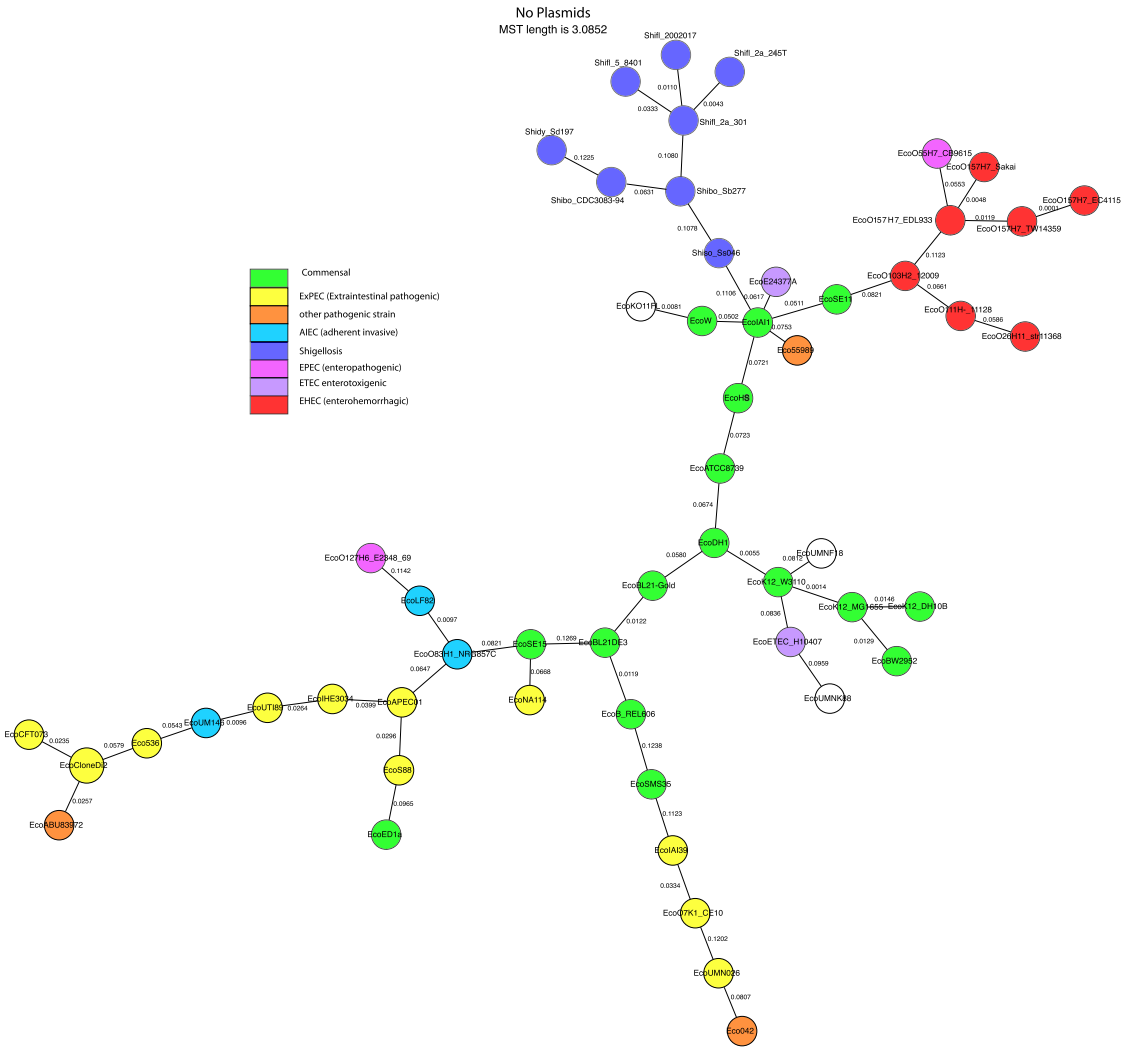


Figure 1: Minimum spanning tree based on complete genomes excluding plasmids Pathogenicity phenotypes are indicated by colors. The pathogenicity of uncolored strains is not known.

In both of these cases there was a single MST. The two MSTs do differ in some details. When plasmids are excluded (Figure 1) the three AIEC strains form a cluster; when plasmids are included (Figure 2) they do not. One of those strains (LF82) lacks plasmids, each of the others has a large (>100 kb) plasmid but the plasmids are unrelated to each other. Excluding plasmids results in the loss of information, and it seems reasonable that two strains that share a plasmid are more alike than they would be did they not share that plasmid. On the other hand, plasmids clearly move more frequently

among strains than do chromosomal genes and if one's interest is more in chromosomal similarity, then excluding plasmids makes sense.

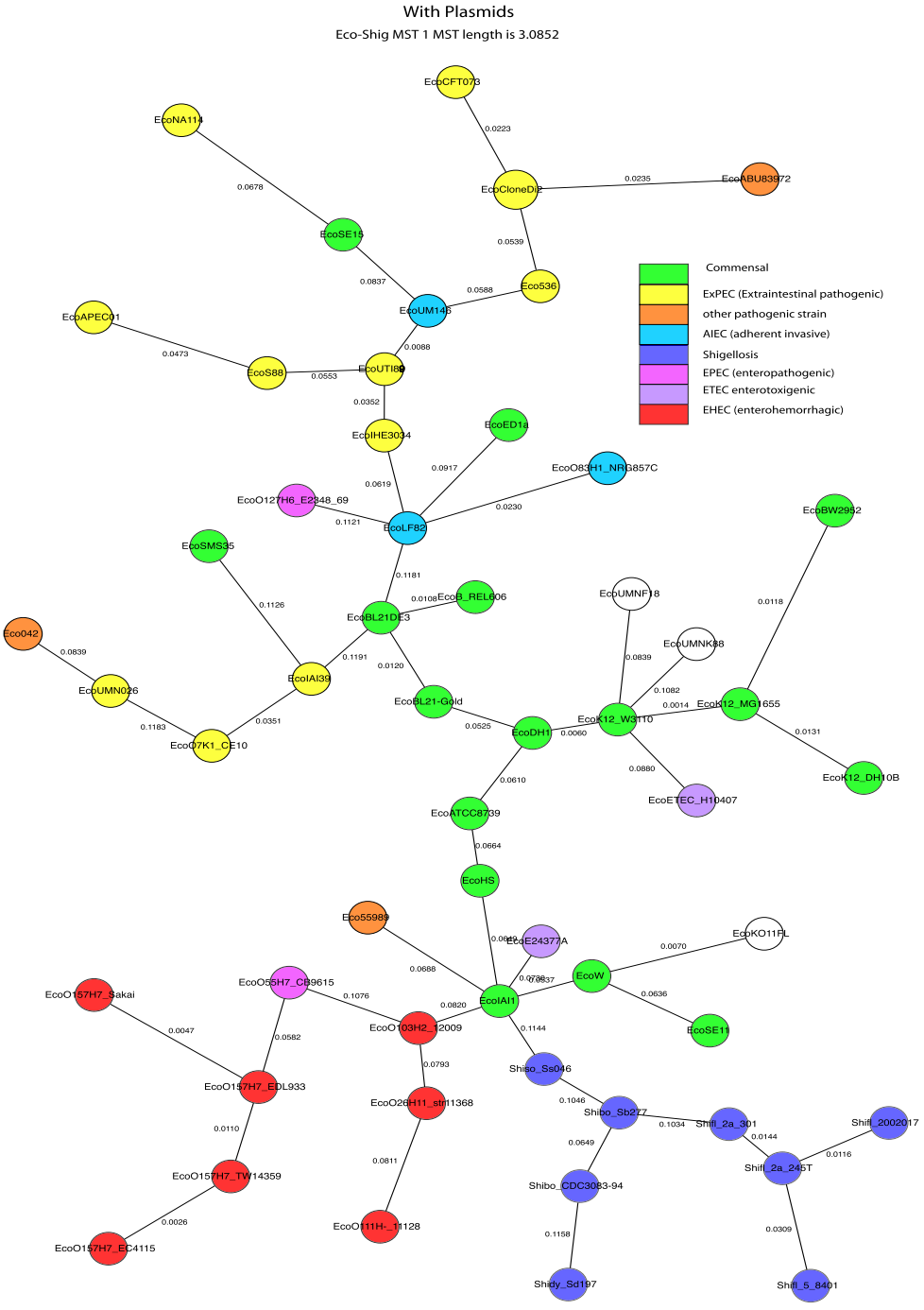


Figure 2: Minimum spanning tree based on complete genomes including plasmids. Pathogenicity phenotypes are indicated by colors. The pathogenicity of uncolored strains is not known.

The most common way to estimate the relationships among organisms is by phylogenetic analysis, but phylogenetic analysis is not always appropriate for the set of organisms being compared or for the data that is used to characterize those organisms. Phylogenetic trees are used to estimate the relationships of organisms to hypothetical ancestors, and thereby to each other. The branches on a phylogenetic tree are intended to reflect the number of mutations that separate an organism from its hypothetical ancestor, or one of those ancestors from its immediate ancestor. There are two fundamental assumptions involved in phylogenetic analysis, neither of which applies to these data sets. First, characters that are shared between a pair of individuals are assumed to be identical by descent. Deviations from that assumption, collectively called homoplasies, may arise from convergence, incomplete genetic isolation, etc. and are indications of loss of phylogenetic signal. For these data there is no implication that shared characters are identical by descent. A pair of strains that have the same bop may do so because they inherited that bop from a common ancestor, because of exchange with a distantly related individual, or because each has acquired the same plasmid or phage-borne bop. Similarly, individuals that lack a particular bop may do so because of inheritance or because each has independently suffered loss of that bop. The nature of the data makes the use of phylogenetic analysis inappropriate. Second, phylogenetic analysis assumes that the individuals are genetically isolated from each other; i.e. they do not exchange genetic information and inheritance is strictly vertical. As a result a node, internal (ancestral node) or external (extant individual), can have only one ancestor. That assumption certainly does not apply to most microbial species where genetic exchange among individuals is common, and in particular does not apply to E. coli which is known to undergo considerable genetic exchange (Wirth, et al. 2006). These considerations indicate that phylogenetic analysis is not appropriate for estimating the relationships among these genomes. MSTs, however, make no assumptions about identity by descent or the absence of genetic exchange. MSTs are based only on identity by state, thus the relationships that are diagramed only indicate the overall similarities among the individuals.

2.2.3. Predicting phenotypes and identifying sequences that do so

Because E. coli varies so much in pathogenicity it is often not sufficient for public health officials to simply determine whether or not E. coli is present, it is often important to determine the danger posed by the strains that are present. Serotyping is an important tool in evaluating risk (O157:H7 strains can be presumed to be EHEC and therefore very dangerous), but other serotypes are also EHEC and it is not necessarily the case that all members of a particular serogroup would be EHEC. Determination of the EHEC phenotype is neither rapid nor cheap. O serotyping is performed following the procedure published by Orskov et al. (Orskov, et al. 1977). H typing is performed by the method described by Machado et al. (Machado, et al. 2000).

It would be useful to identify DNA sequences that correlate very strongly with the EHEC phenotype in order to develop a PCR assay that could quickly determine the probability that a given strain is EHEC. It is well understood that many of the virulence determinants associated with the EHEC phenotype are plasmid borne, but although some plasmids are shared by EHEC strains, none is shared by all. We wondered if there are

any chromosomal DNA sequences that are common to, and perhaps exclusive to, EHEC strains. The ***BopGenomes*** analysis of the data set in which plasmids were excluded was used to identify EHEC-specific segments.

A segment is a contiguous series of bops that have identical distributions among the set of genomes. ***BopGenomes*** generates a file (.segScores) in which each strain is described by a binary string that shows the presence/absence of each segment (Methods and Figure 3). The program ***GetProbs*** was used to determine, for each segment, the probability that it is present in a set of EHEC strain and in a set of non-EHEC strains. That set of strains is known as the "training set" and consisted of four randomly chosen EHEC strains and 30 randomly chosen non-EHEC strains. A parameter, β, was calculated for each segment. β is the probability that the presence/absence of the segment is non-randomly distributed with respect to the phenotype. The output of ***GetProbs*** was used by the program PredictPhenotypes: (1) to calculate for all 55 genomes the probability that the strain is EHEC and (2) to identify the sequences that most strongly predict the EHEC phenotype.

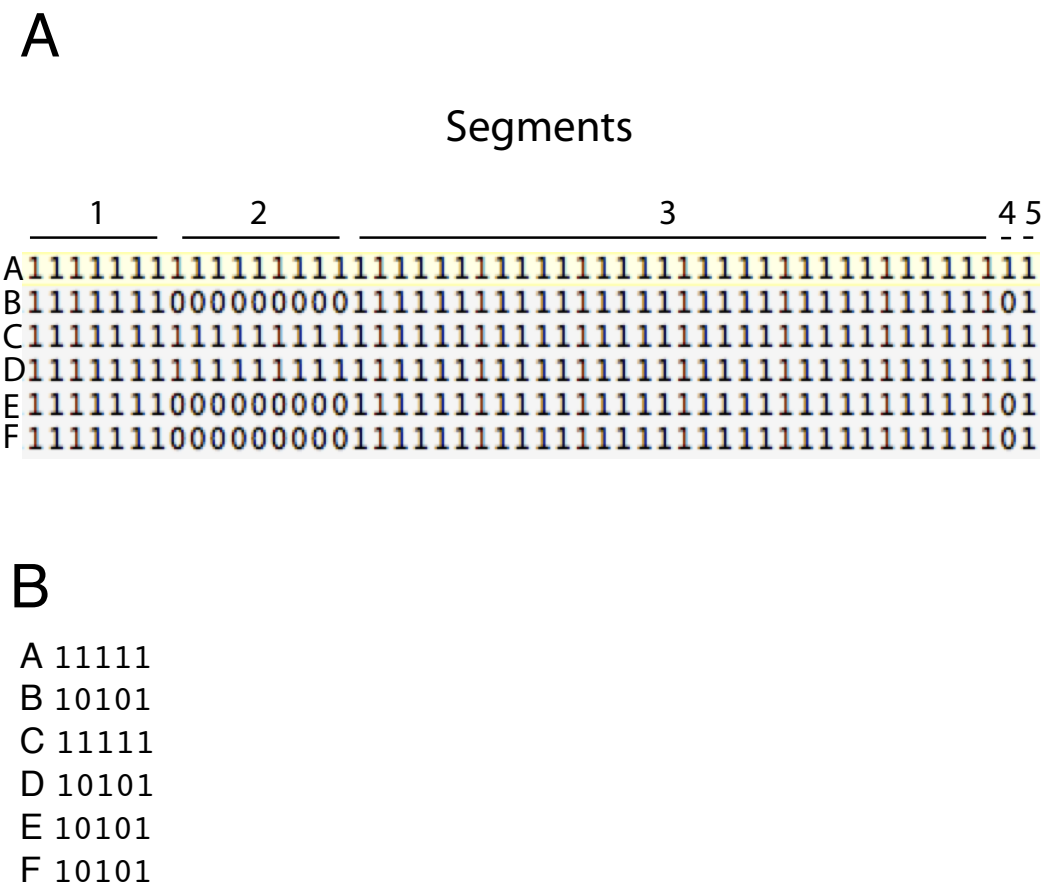


Figure 3: Determination of segments from bops. A: Part of a .scores file in which binary strings indicate the presence or absence of a series of bops for strains A-E. Numbered lines above the string show contiguous bops that are I dentically distributed among strains A-E. B: A corresponding .segScores file in which the binary strings indicate the presence of absence of the segments shown in panel A.

Of the 47 *E. coli* strains 7 are known to be EHEC and 38 are known to be non-EHEC (Table S1). To test the predictive ability of our programs 4 EHEC strains and 30 non-EHEC strains (including Shigella strains) were used as training strains to predict the EHEC phenotypes of the remaining strains (of which 3 were known to be EHEC and 16 to be non-EHEC). Given the small number of EHEC strains the predictions are expected to vary slightly depending upon the makeup of the training set. With that in mind 20 independent runs of GetProbs followed by PredictPhenotypes were carried out. In each run only sequences with a β value >0.9999999 were used to predict phenotypes and to identify predictive sequences. Twenty runs provide a reasonable fraction of the possible combinations of 4 EHEC and 30 non-EHEC strains in the training sets.

Table 1 shows the probabilities with which each EHEC was predicted to be EHEC over the 20 runs. Among the non-EHEC strains the mean probability of being EHEC was 0.0098, with the maximum probability being 0.19 for strain UMNK88. Together with the results in Table 1 this provides an excellent example of the ability of this analysis to predict phenotypes based on the presence of segments with high β values.

Table 1: EHEC strains

Strain	**Probability of being EHEC**
EcoO157H7_EC4115	1.00
EcoO157H7_EDL933	0.99
EcoO157H7_Sakai	0.99
EcoO157H7_TW14359	1.00
EcoO26H11_str11368	0.84
EcoO103H2_12009	0.85
EcoO111H-_11128	0.88

Eighteen segments had β values >0.9999999. A better test of the predictive utility of those 18 segments comes from blast searches of the non-redundant (NR) nucleotide database. Table 2 shows the results of those searches. Nearly all of the segments (except 13 and 16) show homology with more EHEC than non-EHEC hits. Eight of the segments would be useful as PCR probes. The sequences of those eight segments are given in Supplementary table S2. While most of those eight segments on their own would have insufficient predictive ability for reliable clinical assays, when combined, they are powerful predictors of phenotype. Based on the in silico specificities, any strain in which even the four least specific probes amplified would have a >99.97% probability of being an EHEC *E. coli*.

Table 2: Segments used to search the nr database

Segment #	Segment ID	Length (bp)	Number of EHEC hits[a]	Number of non-EHEC hits[a]	p[b]	Comment[e]
1	10254	734	7	0	1.0	+
2	10258	6,924	7	1[c]	0.875	+
3	10261	200	7	2[c]	0.778	
4	10263	270	7	1[c]	0.875	+
5	10314	600	7	1	0.875	+
6	10375	108	7	1[c]	0.875	
7	10380	196	7	2	0.778	
8	10391	217	7	2[c]	0.778	
9	10393	400	7	0	1.0	+
10	10396	800	7	0[d]	1.0	+
11	10398	141	7	1[d]	0.875	
12	10402	636	7	0	1.0	+
13	10545	12,387	7	12	0.368	
14	10547	1,744	7	3	0.700	
15	10549	376	7	1	0.875	+
16	10551	5,590	7	14	0.333	
17	10605	134	7	0	1.0	
18	11916	138	7	1[c]	0.875	

[a]Hits align over at least 50% of the query length
[b]p is the probability that a hit is EHEC
[c]Includes *Citrobacter rodentium* ICC168, a strain that is known to have acquired EHEC and EPEC associated sequences from *E. coli* (Petty, et al. 2010)
[d]Hits in bacteriophage genomes were not counted
[e] Plus sign indicates that length is ≥ 200 bp and p is >0.80. Segments indicated by + would be useful as PCR probes to detect EHEC strains.

2.2.4. Identification of Shigella-specific sequences

Shigella is a clinically important clade within *Escherichia coli* (Sims and Kim 2011) that causes shigellosis (Bardhan, et al. 2010). Distinguishing *Shigella* from other *E. coli* is non-trivial, and it would be valuable to have molecular markers that would unambiguously distinguish the two. We have used ***GetProbs*** to identify 8 sequences, ranging from 400 to 1536 bp, that were present only in the 8 *Shigella* strains ($\beta \geq 0.9999999999$). We used each of those sequences as queries in BLAST searches to screen the entire non-redundant database of DNA sequences. None of the sequences were present in any organism but *Shigella*, including in any of the 47 completely sequenced *Escherichia coli* strains. Our criterion for being present was that the query aligned over >50% of its length with > 80% sequence identity. All but one of the sequences was present in all eight of the completely sequenced *Shigella* strains and none were present in any other organisms including *E. coli*. Some sequences did align over

short regions with other organisms, so we trimmed the query sequences to include only the completely *Shigella*-specific regions (Table S3).

Table 3: Experimental reliability of EHEC-specific and *Shigella*-specific PCR probes

EHEC-specific PCR probes		
Probe	**Number of amplicons**	**Number of amplicons not in EHEC strain**
EHEC 1	48	0
EHEC 2	27	0
EHEC 3	0	0
EHEC 4	44	2
EHEC 5	49	0
EHEC 6	47	0
EHEC 7	0	0
EHEC 8	0	0
***Shigella*-specific PCR probes**		
Probe	**Number of amplicons**	**Number of amplicons not in *Shigella* strain**
Shi 1	15	0
Shi 2	14	0
Shi 3	13	0
Shi 4	17	2
Shi 5	16	0
Shi 6	14	0
Shi 7	14	0
Shi 8	16	0

2.2.5. Experimentally testing the reliability of EHEC-specific PCR amplification probes

We identified primers for each of the EHEC-specific sequences in Table S2 (Table S4) and screened a collection that included 56 EHEC *E. coli*, 17 non-EHEC *E. coli*, 16 *Shigella sp*., 4 *Klebsiella pneumoniae*, 1 *Klebsiella oxytoca*, 3 *Proteus mirabilis*, and 1 *Pseudomonas aeruginosa* strains. Three of the EHEC strains and two of the non-EHEC strains were among the set for which complete genome sequences are available. Genomic DNA was prepared from each strain and was used as the template for PCR reactions with each pair of primers. Four of the probes, EHEC 2, EHEC3, EHEC 7 and EHEC8, were deemed unreliable on two grounds: (a) they produced amplicons in less than half of the known EHEC strains, and (b) they failed to produce amplicons in the three EHEC strains that had been completely sequenced an in which the sequences were known to be present (Table 3). None of the four reliable EHEC probes amplified all 56 EHEC strains, and one probe, EHEC 4, amplified two Shigella strains. Thus none of the probes is, by itself, capable of reliably identifying EHEC strains.

2.2.6. Experimentally testing the reliability of Shigella-specific PCR amplification probes

We identified primers for each of the 8 trimmed *Shigella*-specific sequences in Table S3 (Table S5), and screened the same collection of bacterial strains. None of the PCR probes produced amplicons in any of the species other than *Shigella* and *E. coli*. Table 3 summarizes the results of those experiments. Only one probe amplified all 17 *Shigella* strains, and one probe, Shi 4, amplified two *E. coli* strains. Again, none of the probes is, by itself, capable of reliably identifying *Shigella*.

2.2.7. Updating the in silico results

In the time since the EHEC-specific and *Shigella*-specific sequences were identified, while the experimental PCR studies were being conducted, an additional nine *E. coli* and two *Shigella* genomes have been completed and one *E. coli* genome has been delisted by GenBank. Minimum spanning trees that have been updated to reflect those changes are shown in supplementary figures S1 and S2. The *Shigella*-specific and EHEC-specific probes (Tables S4 and S5) were used as queries in BLAST searches of the non-redundant nucleotide database. Table 4 shows that both the *Shigella*-specific and EHEC-specific probes are highly specific, but are individually insufficient to identify isolates as EHEC or as *Shigella* with $\geq$ 99.9% confidence.

Table 4: *In silico* reliability of EHEC-specific and *Shigella*-specific PCR probes

EHEC-specific PCR probes		
Probe	**Number of hits[a]**	**Number of hits not in EHEC strain**
EHEC 1	10	1
EHEC 2	10	1
EHEC 3	9	0
EHEC 4	9	0
EHEC 5	9	0
EHEC 6	9	0
EHEC 7	10	1
EHEC 8	10	1
***Shigella*-specific PCR probes**		
Probe	**Number of hits**	**Number of hits not in *Shigella* strain**
Shi 1	11	1
Shi 2	11	1
Shi 3	11	1
Shi 4	11	1
Shi 5	12	2
Shi 6	11	1
Shi 7	11	1
Shi 8	10	1

[a]Hits in a BLAST search of the GenBank non-redundant nucleotide database

2.2.8. Predicting EHEC or Shigellosis phenotypes based on amplification profiles

Practical application of this approach to predicting phenotypes means predicting a phenotype from the 'amplicon profile' of a strain. The amplicon profile is a binary string in which a 1 means that a PCR probe produced an amplicon and a 0 means that it did not. For example, *Shigella flexneri* 2a str. 2457T produces BLAST hits (in silico equivalent of an amplicon) with all *Shigella*-specific probes except probe Shi 8. Its amplicon profile is therefore 11111110. Table 5A combines the results from Tables 3 & 4 to show, for each probe, the fraction of amplicons or hits that were from non-*Shigella* or non-EHEC strains. Table 5A is based upon a total of 26 Shigella, 63 EHEC-E. coli, 65 non-EHEC *E. coli*, and 9 strains of other species.

Table 5: Probe reliabilities[a]

A Probe reliabilities			
EHEC probe	**Probability not EHEC[b]**	***Shigella* probe**	**Probability not *Shigella*[c]**
EHEC 1	0.018	Shi 1	0.039
EHEC 2	Unreliable	Shi 2	0.04
EHEC 3	Unreliable	Shi 3	0.042
EHEC 4	0.039	Shi 4	0.107
EHEC 5	0	Shi 5	0.037
EHEC 6	0	Shi 6	0.04
EHEC 7	Unreliable	Shi 7	0.04
EHEC 8	Unreliable	Shi 8	0.039

[a]Combined results from tables 3 and 4

[b] Probability that a strain with a hit or amplicon from this probe is not EHEC

[c] Probability that a strain with a hit or amplicon from this probe is not *Shigella*

While the presence or absence of an in silico hit can be considered a completely reliable indicator of the presence or absence of a query sequence in the genome, the presence or absence of an amplicon in a PCR experiment is a much less reliable indicator of the presence or absence of a sequence. The absence of an amplicon can either mean that the sequence is not present in that strain, or that the PCR reaction failed for spurious reasons; i.e. a false negative. False positives are much rare than false negatives because to be counted as a positive not only must an amplicon be present, it must be an amplicon of the correct size. In other words, we can trust the interpretation of the presence of an amplicon more than we can the interpretation of the absence of an amplicon.

If the amplification profile of an isolate that was probed with all eight *Shigella*-specific probes is 00011001 we want to know the probability that an isolate in which probes Shi 4, Shi 5 and Shi 9 produce amplicons is *Shigella*. Probe Shi 4 produced in silico hits in 11 strains Table 4) and amplicons in 17 strains (Table 3), of which a total of 3 strains were not *Shigella*. The probability that a Shi 4 hit or amplicon was not in a *Shigella* strain is therefore 0.107 (Table 5). The probability that a strain with the amplification profile 00011001 is not *Shigella* is the product of those probabilities for probes Shi 4, Shi 5 and Shi 8, or 1.54×10^{-4}. The probability that the strain is *Shigella* is therefore 0.9998.

Most *Shigella* strains produced amplicons/hits with all eight *Shigella*-specific probes, and most non-*Shigella* strains produced no hits with any *Shigella*-specific probes, i.e. the amplicon profiles were 11111111 and 00000000 respectively. Their respective probabilities of being *Shigella* are 0.99999999998 and 1.6×10^{-11} respectively. Similarly most EHEC *E. coli* had amplicon profiles of 1111 with the reliable EHEC-specific probes, while most non-EHEC *E. coli* and other species had a 0000 profile. Their respective probabilities of being EHEC are 0.99999993 and 7×10^{-8} respectively. Those strains with other profiles, and their probabilities of being EHEC or *Shigella*, are shown in Table S6.

All of the *Shigella* strains tested were identified as *Shigella* with probabilities ≥0.999. One *E. coli* strain, strain 53638, was incorrectly identified as being *Shigella* with

a probability >0.9999, giving a false positive frequency of 1 out of 75 *E. coli* screened, or 0.013 (Table S6). Two EHEC strains, RDEC-1 and RD8, were not identified as EHEC, giving a false negative rate of 0.03. No false positives, either EHEC *E. coli* or *Shigella*, were found.

For clinical applications we suggest that any strain in which the probability of being EHEC or *Shigella* is <0.999 should be rejected as having that phenotype. Because false positives do occasionally occur we suggest that any strain in which a single probe produces an amplicon should be retested.

2.3. Discussion

Shigella and *E. coli* have clinically been distinguished on the basis of a variety of biochemical and serological tests. More recently they have been distinguished on the basis of presence of the pINV plasmid and of the Shiga toxin genes, but none of these phenotypes are unique to *Shigella*, and in particular many are shared by EIEC *E. coli* strains that closely resemble *Shigella* (Lan R and Reeves PR, 2002; Pollard DR et al. 1990). Similarly, EHEC *E. coli* are classically distinguished from non-EHEC *E. coli* on the basis of serotype and presence of Shiga toxin (verocytotoxin), but again those factors are not unique to EHEC strain. The most common EHEC serotype is O157:H7, but there are other EHEC serotypes (Sims GE and Kim SH, 2011). Non-EHEC strains, including EIEC and EAEC-STEC strains produce the Shiga toxin (Ahmed SA et al. 2012). Indeed, "EHEC are sometimes difficult to identify" and "There is no single technique that can be used to isolate all EHEC serotypes" (Iowa State University, 2009). Because the infectious doses of both *Shigella* and ETEC *E. coli* are four orders of magnitude lower than that of most other pathogenic *E. coli* (Lan R and Reeves PR, 2002), rapid and reliable identification of those organisms is clinically important. The major advantage of using the Bop method to identify phenotype-specific sequences that can be used as PCR probes is not only specificity, but also that using sets of those probes allows a probability statement about the reliability of identification of EHEC and *Shigella* strains to be made. Additionally, use of such probe sets is both rapid and inexpensive.

The BOP method can be applied to any species for which a sufficient number of strains in which the phenotype of interest is known have been completely sequenced. The method is useful for estimating the relationships among the sequenced strains, but the most valuable application of the method is to identify phenotype-specific sequences that can then be used to inexpensively and quickly characterize clinical isolates by PCR. Once the bop and segment profiles of a set of sequenced genomes has been determined it requires only a few minutes to identify phenotype-specific segments among those isolates.

At present it is quite difficult to obtain phenotypic information about pathogens that have been completely sequenced. It is typically the case that the group that sequenced the strain is interested in only one phenotype (if any). For instance, the public sequence databases rarely include any information about antibiotic resistance. The availability of the BOP method makes a strong argument for developing collections of clinically important strains whose complete genome sequences have been determined and whose phenotypes are thoroughly described. The availability of such a collection would

make it possible to develop databases of their phenotypes and to use those phenotypes to identify phenotype-specific PCR probes. Appearance of a new clinically relevant phenotype could quickly be followed by characterization of that collection by a laboratory with the necessary expertise, and in turn followed by development of phenotype-specific PCR probes.

These approaches that we have introduced have the potential to apply genomic data for epidemiological and clinical analyses. We believe that the approach of using bops as input for MSTs and segments as a method for identifying sequences that predict phenotype can ultimately be implemented into clinical labs as cost and time effective methods for analyzing infective bacteria. Additionally, the programs we have developed are readily available and can be implemented by a wide range of users with common computing equipment. Our results are also of theoretical importance because they identify a new type of character, namely a bop, to perform robust analyses of genomic sequence data.

2.4. Materials and Methods

2.4.1 The BOP method

The BOP method is applied to a set of completely sequenced (closed) genomes through use of the BopGenomes program. The genomes are digested in silico with one of several restriction enzymes to produce an ordered restriction map. The selection of restriction enzymes is designed to allow adequate restriction of any genome regardless of GC content, and other sequence biases. By allowing the option to digest the genome at a variety of restriction sites, it is possible for the program to accommodate most genomes. Each fragment is given an ID that consists of its length (rounded to the nearest 100 bp) and the lengths of the fragments that flank it. Thus a 16,542 bp fragment flanked by a 4,320 bp to its left and a 1,680 fragment to its right would be identified as 4.3-16.5-1.7. That particular 16.5 kb fragment is distinguished from all other 16.5 kb fragments by the lengths of its flanking fragments. At this point that system appears sufficient to uniquely identify restriction fragments. Cases of multiple occurrences of the same fragment turn out to be duplicated regions that contain the same internal restriction fragments. Usually such regions represent multiple copies of mobile elements or phages.

BopGenomes makes a list of all unique restriction fragments; i.e. as all the genomes are considered a fragment is added to the list only if it was not already in the list. Restriction fragments cannot be used directly to assess genome content because restriction fragments are degenerate; i.e. multiple restriction fragments can include the same homologous sequence. For instance, the appearance of a new restriction site destroys an existing restriction fragment and creates in its place two fragments whose lengths sum to the length of the original fragment. In order to deal with restriction fragment degeneracy, each fragment is divided into ~200 bp sections called "bops". Most bops are exactly 200 bp, but bops at the end of a fragment may be less than 200 bp. If a bop is <100 bp it is joined to the previous bop, thus generating a bop of up to 300 bp. Although the use of restriction fragments may appear to be superfluous when analyzing completely sequenced genomes, it does serve to put the bops in most homologous regions

into the same frame. Doing that reduces the number of unique bops and dramatically reduces computation time.

After introducing the restriction sites and creating the bops, the program lists all unique bops. As each bop is considered for addition to the list it is aligned against each of the bops already in the list by the blast2seq program (NCBI). If a bop shares ≥80% sequence identity over > 50% of its length with a bop already in the list it is not added to the list. At the same time, lists of each bop in each restriction fragment are maintained. Thus, from the ordered restriction map of a genome we know which bops are present. Since we know the sequence of each bop we know the sequence information that is present in each genome.

Finally, each genome is described by a binary string in which the *i*th character indicates the presence of bop number *i* by a 1, and its absence by a 0. Note that homologous bops (those that share > 80% sequence identity) are considered equivalent. Minor variation in sequence is lost to this analysis, as is the position of sequences in the genome. The binary strings that describe the presence/absence of each bop in each of the genomes are contained in the BopGenomes output file with the extension '.scores' .

2.4.2. Clustering by Minimum Spanning Trees (MST)

Clustering by MST was carried out using the MS gold program (Salipante and Hall 2011). The pairwise distances between genomes were based on the equidistant method (Salipante and Hall 2011).

Just as a phylogenetic tree is a graph that illustrates the relationships between individuals and their hypothetical ancestors based on identity by descent, an MST is a graph that illustrates the relationships between individuals based on identity by state. On an MST each node represents an individual and nodes are connected by edges whose lengths reflect the distance between the nodes. In this case the distance between a pair of genomes is shown as the number of differences in the state of the bop (0 or 1) divided by the number of bops.

2.4.3 Predicting phenotypes and identifying predictive segments.

Bops are sufficient for estimating the genetic relationships among genomes by clustering methods, but for other downstream applications (such as predicting phenotypes) it is useful to join a series of contiguous bops that have identical distributions among the genomes into "segments". A segment is thus a series of contiguous bops that behave as a unit with respect to their presence or absence in genomes. A genome can thus also be described by a binary string that indicates the presence or absence of each of the segments. The binary strings that describe the presence/absence of each segment in each of the genomes are contained in the ***BopGenomes*** output file with the extension '.segScores' .

The program ***GetProbs*** uses a set of genomes whose phenotypes are known (the training set) to calculate, from the binary strings in a .segScores file, (1) the probability of a positive phenotype given that a particular segment is present, (2) the probability of a positive phenotype given that the segment is absent, and (3) β, the probability that the presence or absence of the segment is non-random with respect to phenotype. β is thus a

measure of the degree to which the presence or absence of a segment is associated with the phenotype. That information is saved in a file with the extension '.pp'.

The program ***PredictPhenotypes*** uses the probabilities for each segment in a .pp file to predict the probability that a genome whose phenotype is unknown has a positive phenotype. Since segments with high β values are more strongly associated with a particular phenotype, the segments that are used to predict phenotype are filtered to include only those segments with β values above a chosen threshold. The probability that a genome has a positive phenotype is the sum of the probabilities of being positive given that the segment is present for all segments that are present plus the sum of the probabilities of being positive given that the segment is absent for all segments that are absent, divided by the number of segments whose β value is above the threshold. The program lists all of the segments above the threshold β value.

2.4.4. Strains

Most of the EHEC *E. coli* and *Shigella* strains were obtained from a collection at Michigan State University that was developed by Dr. Shannon Manning. *E. coli* strains whose genomes have been sequenced were obtained from authors of the genomes sequences.

2.4.5. Genomic DNA preparation

A boiling genomic prep was used to lyse cells and extract genomic DNA. Cells were suspended in 25 µl of water and were heated to 100º for 10 minutes. The samples were then cooled and centrifuged at 1000 rpm (max speed) in a benchtop centrifuge for 1 minute and 3 µls of the supernatant each sample were used as template for 10 µl PCR reactions.

2.4.6. PCR Assays

PCR reactions were performed as follows:
2x Taq Mastermix (New England Biolabs ™) was used according to manufacturer instructions. Primers were used at a concentration of 500 pM. The reaction was run for 30 cycles with a denaturation temperature of 94ºC, annealing temperature of 65ºC, and an elongation temperature of 72ºC.

2.5 Conclusion

It is evident that predicting pathogenic phenotypes can be achieved via a simple, rapid, and inexpensive method on the basis of the presence or absence of short homologous DNA segments in an isolate. The process of identifying unique genomic segments by the Bop method enables for designing of probes specific to strains, in this case, EHEC strains of *E. coli* and *Shigella.* The fact that both strains share a similar genotype, their phenotype outcomes are very different due to genetic rearrangements and the gain and loss of DNA sequences. Our robust technique determined that with as few as 8 predictive segments and amplification of at least three of those segments, we can rest assure of a confidence level of >0.99 that it is either EHEC or *Shigella*. The

technique proposed in this paper can provide a significant clinical advancement and monetary savings in the identification of other strains that require valuable time in proper strain identification. A simple and rapid PCR diagnostic test can make a clinical difference in the survival rate of patients with infections.

Supporting Information

Table S1: Strains, accession numbers and phenotypes

Strain	ID on MSTs	Accession Number	Phenotype and reference[a]
Escherichia coli 042	Eco042	FN554766	EAEC (Crossman, et al. 2010)
Escherichia coli 536	Eco536	NC_008253.1	ExPec (Zhou, et al. 2010)
Escherichia coli 53638	Eco53638	AAKB00000000	EIEC
Escherichia coli 55989	Eco55989	NC_011748.1	EAEC (Sims and Kim 2011)
Escherichia coli ABU 83972	EcoABU83972	CP001671	Asymptomatic bacteriuria(Zdziarski, et al. 2010)
Escherichia coli APEC O1	EcoAPEC01	NC_008563.1	ExPec (Zhou, et al. 2010)
Escherichia coli ATCC 8739	EcoATCC8739	NC_010468.1	Commensal (Archer, et al. 2011)
Escherichia coli B str. REL606	EcoB_REL606	NC_012967.1	Commensal (Sims and Kim 2011)
Escherichia coli 'BL21-Gold(DE3)pLysSAG'	EcoBL21-Gold	NC_012947.1	Commensal
Escherichia coli BL21(DE3)	EcoBL21DE3	AM946981	Commensal (Sims and Kim 2011)
Escherichia coli BW2952	EcoBW2952	NC_012759.1	Commensal (Ferenci, et al. 2009)
Escherichia coli CFT073	EcoCFT073	NC_004431.1	ExPec (Zhou, et al. 2010)
Escherichia coli str. 'clone D i2'	EcoCloneDi2	CP002212	ExPec (Reeves, et al. 2011)
Escherichia coli str. 'clone D i14'	EcoCloneDi14	CP002212	ExPec (Reeves, et al. 2011)
Escherichia coli DH1	EcoDH1	CP001637	Commensal (Suzuki, et al. 2011)
Escherichia coli E24377A	EcoE24377A	NC_009801.1	ETEC

Escherichia coli ED1a	EcoED1a	NC_011745.1	Commensal (Zhou, et al. 2010)
Escherichia coli ETEC H10407	EcoETEC_H10407	FN649414	ETEC (Zhou, et al. 2010)
Escherichia coli HS	EcoHS	NC_009800.1	Commensal (Zhou, et al. 2010)
Escherichia coli IAI1	EcoIAI1	NC_011741.1	Commensal (Zhou, et al. 2010)
Escherichia coli IAI39	EcoIAI39	NC_011750.1	ExPec (Zhou, et al. 2010)
Escherichia coli IHE3034	EcoIHE3034	CP001969	ExPec (Moriel, et al. 2010)
Escherichia coli str. K-12 substr. DH10B	EcoK12_DH10B	NC_010473.1	Commensal (Zhou, et al. 2010)
Escherichia coli str. K-12 substr. MG1655	EcoK12_MG1655	NC_000913.2	Commensal (Sims and Kim 2011)
Escherichia coli str. K-12 substr. W3110	EcoK12_W3110	AP009048	Commensal (Sims and Kim 2011)
Escherichia coli KO11FL	EcoKO11FL	CP002516	Commensal (Turner, et al. 2012)
Escherichia coli LF82	EcoLF82	CU651637	AIEC (Wine, et al. 2009)
Escherichia coli NA114	EcoNA114	CP002797	ExPec (Avasthi, et al. 2011)
Escherichia coli O103:H2 str. 12009	EcoO103H2_12009	NC_013353.1	EHEC (Sims and Kim 2011)
Escherichia coli O104:H4 2009EL-2050	EcoO104H4_2009EL-2050	CP003297.1	Other pathogen
Escherichia coli O104:H4 2009EL-2071	EcoO104H4_2009EL-2071	CP003301.1	Other pathogen
Escherichia coliO104:H4 2011C-3493	EcoO104H4_2011C-3493	CP003289.1	Other pathogen
Escherichia coli O111:H- str. 11128	EcoO111H-_11128	NC_013364.1	EHEC (Sims and Kim 2011)
Escherichia coli O127:H6 str.	EcoO127H6_E2348_	NC_011601.1	EPEC (Zhou, et al. 2010)

E2348/69			
Escherichia coli O157:H7 str. EC4115	EcoO157H7_EC4115	NC_011353.1	EHEC (Sims and Kim 2011)
Escherichia coli O157:H7 str. EDL933	EcoO157H7_EDL933	NC_002655.2	EHEC (Sims and Kim 2011)
Escherichia coli O157:H7 str. Sakai	EcoO157H7_Sakai	NC_002695.1	EHEC (Sims and Kim 2011)
Escherichia coli O157:H7 str. TW14359	EcoO157H7_TW14359	NC_013008.1	EHEC (Sims and Kim 2011)
Escherichia coli O157H7 str. TW14588	EcoO157H7_TW14588	CM000662.1	EHEC (Kulasekara, et al. 2009)
Escherichia coli O26:H11 str. 11368	EcoO26H11_str11368	NC_013361.1	EHEC (Sims and Kim 2011)
Escherichia coli O55:H7 str. CB9615	EcoO55H7_CB9615	NC_013941.1	EPEC (Zhou, et al. 2010)
Escherichia coli O55:H7 str. RM12579	EcoO55H7_RM12579	CP003109.1	EPEC (Kyle, et al. 2012)
Escherichia coli O7:K1 str. CE10	EcoO7K1_CE10	CP003034	ExPec
Escherichia coli O83:H1 str. NRG 857C	EcoO83H1_NRG857C	CP001855	AIEC (Allen, et al. 2008)
Escherichia coli P12b	EcoP21B	CP002291.1	????? (Liu, et al. 2012)
Escherichia coli S88	EcoS88	NC_011742.1	ExPec (Zhou, et al. 2010)
Escherichia coli SE11	EcoSE11	NC_011415.1	Commensal (Zhou, et al. 2010)
Escherichia coli SE15	EcoSE15	AP009378	Commensal (Toh, et al. 2010)
Escherichia coli SMS-3-5	EcoSMS35	NC_010498.1	Commensal (Zhou, et al. 2010)
Escherichia coli UM146	EcoUM146	CP002167	AIEC (Krause, et al. 2011)
Escherichia coli UMN026	EcoUMN026	NC_011751.1	ExPec (Zhou, et al. 2010)

Escherichia coli UMNK88	EcoUMNK88	CP002729	????????
Escherichia coli UTI89	EcoUTI89	NC_007946.1	ExPec (Zhou, et al. 2010)
Escherichia coli W	EcoW	CP002185	Commensal
Escherichia coli Xuzhou21	EcoXUZhou21	CP001925.1	EHEC (Xiong, et al. 2012)
Shigella boydii CDC 3083-94	Shibo_CDC3083-94	NC_010658.1	(Pupo, et al. 2000)
Shigella boydii Sb227	Shibo_Sb277	NC_007613.1	(Yang, et al. 2005)
Shigella dysenteriae Sd197	Shidy_Sd197	NC_007606.1	(Yang, et al. 2005)
Shigella flexneri 2002017	Shifl_2002017	NC_004741.1	(Ye, et al. 2010)
Shigella flexneri 2a str. 2457T	Shifl_2a_245T	NC_004741.1	(Wei, et al. 2003)
Shigella flexneri 2a str. 301	Shifl_2a_301	NC_004337.2	(Jin, et al. 2002)
Shigella flexneri 5 str. 8401	Shifl_5_8401	NC_008258.1	(Nie, et al. 2006)
Shigella flexneri 5a str. M90T	Shifi_M90T	CM001474.1	(Onodera, et al. 2012)
Shigella sonnei Ss046	Shiso_Ss046	NC_008258.1	(Yang, et al. 2005)
Shigella sonnei 53G	Shiso_53G	HE616528.1	

[a]If no reference is given the genome is a direct submission and the phenotype is taken from the GenBank file annotation

Table S2: Sequences of segments that are useful as probes to detect EHEC *E. coli* strains

>10254
TGGCGCGGGGAGAGTAGTCGATGAATAAACAATACGAACTGGTTGTAAAATG
AATATTTCTAACTGAAAAAACGTTCCATGAGGTGAGAAAAGGTCACAGGCAA
TCAATAACAGGACGTGATGAAAGACCCTTGCATTTGTGCGCTTTCTCTTTAGA
TAGCAGCAGATACTGAAAATCTGAGTTGTCGGGGAGTCAGGGATACAGCTGT
GCAAGAGTTGGTCATTGTGATTCCATTGAAATCCTGTATGCCATGAAGGGCA
GGATTTTATGGCTACCTGAGCTTTGGTGATAGTAAGTTGAAAATTCGCATTTT
TTGCTGACATGCGTAACGAGAATCCCATAAGCAGGGAGGACTTAATTCTTCA
TTAACCCATGCGTTGATATTATGTTTCAGCCGTTGAAGCATCAGCGGTGTTAA
TGTTGTGGTAATAATATCCAGCGTTTTATGTGAGATCTTACCGTAAGGGTCTG
CAAGAATGCTGCTTGTTGCTTCGTTATTATCTGCCATCAGAAGAAGTAACTCT
GATTTAACGTTTTCTGTCATTAGTTGTAAAAATCTTCTGCGCAAACTTTCTTTA
CTGTTCATTTATATGGCTTCATTTGTTGTAATCTGCTGCGTCTCAAGGGATATG

TTTATGAGAGCGACCATGAGTGTTGGATTATATACCTAACATATCAAGGGATT
AGAAATCGATAAATCCCCATGAACGAAAAAATAAAATACGGCCTGTCG

>10258
TCCGGGGAATATTTGTTAGATAAAAAAGAGGAGATAATTCAATAGGGAGTTA
AATTAATGCCGATAAATCTGACATCTTATTTGTGGGTACAGGGACAGAAAGT
TGTCCCGGCAGTTGTGTTTTCTAATTTTACTTTAGTGGACTAAGTAAAAAGGA
GTGAGATAAATGCTGCCCACTACAAATATCTCTGTAAATTCTGGAGTAATATC
TTTTGAAAGTCCTGTAGATTCACCATCTAACGAGGATGTTGAAGTTGCCCTCG
AAAAGTGGTGCGCTGAGGGAGAATTTAGCGAAAATCGTCATGAGGTTGCATC
AAAAATACTTGATGTTATAAGTACTAATGGAGAGACTTTATCAATCAGTGAG
CCAATAACAACATTACCAGACTTGCTTCCAGGTTCTCTGAAAGAACTGGTTTT
GAATGGATGTACAGAGCTTAAATCAATAAACTGCTTACCCCCCAACTTATCTT
CATTAAGTATGGTTGGATGCTCATCATTAGAGGTTATAAATTGCAGCATACCT
GAAAATGTCATTAATTTATCTTTATGCCATTGTAGTTCTTTGAAACATATAGA
AGGTTCCTTTCCTGAGGCACTCAGAAATTCCGTATATTTAAATGGCTGTAATT
CATTAAATGAATCGCAATGTCAATTCCTTGCATATGATGTCAGTCAAGGCCGT
GCCTGCCTGAGCAAAGCTGAGCTTACTGCTGACTTAATTTGGTTGTCGGCTAA
CCGAACGGGTGAAGAGTCTGCTGAAGAATTGAATTACTCTGGATGTGACTTG
TCAGGTCTAAGTCTTGTAGGGCTGAATTTATCATCAGTAAATTTTTCTGGAGC
AGTGCTTGATGATACAGATCTCAGGATGAGTGATTTGTCTCAGGCTGTATTGG
AAAACTGTTCTTTTAAAAACTCGATTTTGAATGAATGTAATTTTTGTTATGCT
AATTTATCTAATTGTATTATTAGGGCTTTGTTTGAAAACTCTAATTTCAGCAAT
TCCAATCTTAAAAATGCATCATTTAAAGGATCTTCATATATACAATATCCTCC
AATTTTGAACGAGGCTGATTTAACAGGAGCTATTATAATTCCTGGAATGGTTT
TAAGTGGTGCTATCTTAGGTGATGTAAAGGAGCTCTTTAGTGAAAAAAGTAA
TACCATTAATCTAGGAGGGTGTTACATAGATCTATCTGACATACAGGAAAAT
ATATTATCTGTGTTGGATAACTATACAAAATCAAATAAATCAATTTTATTGAC
TATGAATACATCTGATGATAAGTATAACCATGATAAAGTAAGGGCCGCTGAA
GAACTTATCAAAAAAATATCTCTTGACGAATTAGCGGCGTTCCGGCCCTATGT
TAAGATGTCTTTGGCTGATTCATTTAGTATTCATCCTTATTTGAACAACGCAA
ATATACAGCAATGGCTCGAGCCTATATGTGATGACTTTTTTGATACTATAATG
TCTTGGTTTAATAATTCAATAATGATGTATATGGAGAATGGTAGTTTATTGCA
GGCAGGGATGTATTTTGAGCGACATCCAGGTGCGATGGTATCTTATAATAGTT
CCTTTATACAAATTGTAATGAATGGTTCACGGCGTGATGGAATGCAGGAACG
ATTTAGGGAACTCTATGAAGTATATTTAAAAAATGAAAAAGTTTATCCTGTCA
CACAGCAGAGTGATTTTGGATTGTGCGATGGCTCTGGGAAGCCTGACTGGGA
TGATGATTCCGATTTGGCTTATAACTGGGTTTTGTTATCATCACAGGATGATG
GTATGGCAATGATGTGTTCTTTGAGTCATATGGTTGATATGTTATCTCCTAAT
ACATCAACTAACTGGATGTCCTTTTTTTTATATAAGGATGGAGAAGTTCAAAA
TACATTTGGGTATTCATTGAGCAATCTTTTTTCTGAATCATTTCCAATTTTCAG
TATTCCTTATCATAAAGCTTTTTCCCAGAATTTCGTTTCTGGTATTCTGGATAT
ACTCATTTCTGATAATGAACTCAAAGAGAGATTTATTGAGGCACTTAATTCCA
ATAAATCAGATTATAAAATGATTGCTGATGATCAGCAAAGGAAACTTGCCTG
TGTCTGGAATCCCTTTCTTGATGGTTGGGAACTGAACGCTCAGCATGTAGATA

TGATTATGGGGAGCCATGTATTGAAAGATATGCCACTAAGAAAACAGGCTGA
AATATTATTTTGTTTAGGGGGGGTTTTCTGTAAATACTCATCGAGTGATATGT
TTGGTACAGAGTATGATTCTCCTGAGATTCTACGGAGATATGCAAATGGATTG
ATTGAACAAGCTTATAAAACAGATCCTCAGGTATTTGGCTCAGTTTATTATTA
CAATGATATTTTAGACAGGCTACAAGGAAGAAATAATGTTTTTACTTGTACCG
CTGTGCTGACTGATATGCTAACGGAGCATGCAAAAGAATCTTTTCCTGAAAT
ATTTTCATTGTATTATCCTGTTGCGTGGCGTTGATTTAGAGACCATGGATGAA
TATTATTGTAACACTGTCTTTTTAGGTTTGCACATGTTCAGTGGGAACATTTAT
TGCTCCGTTGTTATTATGTCTTGGTTTAGTGGCATGAGCCGAATGTTCTTAAA
ATTTACAGTGTCGAAGATGAAAGAGTACGATAGAACTCGTTGTGATTCGATT
TGTCTGATATGGTGGTAATATATGAATAGGATACTGCATATATCGTAGTGCTT
GAGGATGTTGATTAGGGCATATGATTTTTATATTTTTTTTGAGCAACGTTTTA
AGGGAAAATTTACATATGACAACTTTGTCTGAAGAGTATCTTTGTTGTCAGTA
TTTTTTTTGAATGTGATAATTGATTTTGTTTTGCTGTGAATGGTCACTATAGAT
GAAATATGATTTTAATTACAACGAAAATTATGTTCGTTATTTCTGGGCAAATC
GTGGAGAATCATATTTTATGAATTGATATTCAGATTAATATGTTTTTGTTACT
ATAGTAATATGGCGTAATTAATGATTGTTTTGAAAAGACTCTCAACTGGCTTT
TATTCATTGAATAGTGCGTTATAAGAGGAAGTGGAATTTTAATGAAAATAAC
AAACTATATACTGCCAACAAGTCGTACTCATGGTTCATTCTCAACTATAAAAT
CATGGGACACAATGAATTATATTAAACATTTAATCAGACATACAAATGACCC
TGTATTTGAAGAACAATTTTATAAAATAACACAATCTCATATTGACTTTGACA
AAAGAGCTAAAGATGAAAAAAATGACACCATTAACATTTATGATAACTTTTT
CTATTCATCTAATGATGATCTTGATTCTAAAATTAGAAGTATGTTAAATAATT
TATATGAGAAAAGCTTAACTTTCCGAAGAATCATTAATTATTATGTGAAGGA
AATAAACTTAAGTGATTATGGCTTTCTAAAATGTAAGATTTTACCAGCATATG
CTTATAACTATGAGATGGAAAATGATGCCCCCCCAAAAATACTAATTCCAAT
TGACCATGATTTAAATTTTATTGATGCAAAATATAATGGAGAAACTTATCGGG
GAAATGAAGAGTTTGCTATTAATCTTTTTCTGCAGCATATATTACATAATGAC
ATACAAGAACAAACATCGATAGACTTATACACGAGCATAATAAATAAAGAGT
TGGATAGCAATAGAAAATCATACAATAATGAAATTTTTAACAATTTCTCTTTT
GATAAGTCTGTAAAGTTGAATTCATATAACTATATTGCAGATGATATAGAGC
AAGTAATCGATAAAGGAAGCAAAGTTCAATTGGAGGTATATAATTTATTATC
CGAAGAAAAGATATTTGAACATAAAATTATGAATAATTGGACAAGGAGCATA
AAAAATATATTGACGACATATTTGTTTATGTCATCAGGAGCGGTGACAGCCA
GAAATGTTCAAACCTTTTCTCCAACAATAAATAATGAGTCAAGGATTCGATTG
CCGAGAGCATTGCCAGTAGGCCATCCATATCCTGAGGAACATAAGGCTTCTG
GTTTCTCCCCTTTTATGATGGGGGGGCTGAGTGGTGATATTCTTCCGGAAATT
TTAACGGGGAATGGACCATCTATATTTTTTAACGGAAAACATAATAACCAAC
ATGATGGAGCTTTTGGAAAAATAATAGATTTTACCCAAAATGGAAATAAAAT
AAGTGCAAAAGATAAAGAAATAATAAAAAGATATATTTTTGATAAGATCAAT
GTTTTGATTAAAGAGTATTTCATTAGAACTGGTAAAAATTCTCATACCCCATT
TGAAGTTTTTATAAAGGAGCGATTATTTAATCAATATGATATTTTTAAAACAT
TGGCTAGAGATATATTGGCACACCCATTAGTAATATATGATGCGGGTTACAA
AAATTATCATGAGTCATTAAATGCTGCTATTGCAATAAACTCTAGACCATTAC
AAGAAATACATTATGGTGATGTTTTATATCATTATCATAAAAATGACATCTCT

TTGGGAGTAGATACTCTTTACGGGAGGGAAAGTTTTGATATTGTACTGGATGC
AATGAACGTATATAGAAAAAGCAAAAAAATGAGAGTTATTTCCAATAATGAG
ATGAAAAAAAGCATTAAAATATCTGAATTAGTTATCCATAATATTATAAAGA
AAGGATTGACTAATTGTTTGCTTAAAAAGGATGTTCTTAATGCCAGATATGAT
CTTATTAGAGATATTCTTCGATATTCTTTAAATATACGACAGGGAATTAAACA
TGATGATGTTAATAGAATAGCGGAAAATATAATAAAAAAGTATGGTATAACT
GAGGGTATGAATCCTAAACCTAGGAATGCCAGAATATCTAAAGAATTGCTTT
TATTAGCTGTTGATAGACAGATTGAGTGGGCGAAAAAACATTTTATAACAAA
AGATGTATTGGAAAATGTTGTGTCAAAATGTGATTTATCATCTATCTTTAATG
TTAATAAAGTGCTTCAGAATACTATTCTTGAGTTTGTCCATGAAATTAATAAT
ATATCATCTGCTCGCTGGATGTCAAAATCAGAAAAGAATAATAAACAAAAAG
AGGCAATAGAAAAGTTCAAAAAAGAAGTATCCCATATGAATGGCGGGCAGC
AGTTTATTTGGGGGTTTGATAAGGTTATTCAAGAAGGCTTAAGTGGATTGATT
GAGTTAAGTATCGATATTAATGATAGTACAAATCATCGTGATAAGTCTTCTCT
TTCTCCTGATGGGAGAGCTGTGTTACATTTTTTAGGTACAATTTGGAATATGG
CGATGGGAGCTGTACCTGGTTATAATGCATTGTCTGGTGTAAGTAGCATTTTA
CATAGTGCTATAGTAAAAGAATCTAGCAATATCTGTGATTATATTCAGGGGG
CTGTACGTATTGGAATGGACTTTGTTCCAGGCACTCGCTCTGACTTACATAGC
CGTTCGCTGCAGATAAAATATGAGGCTTTGAAGCATATAGAAAAAAACATTA
ATGATAATATTATTTATCATCCGAGTAATAATGCTAATTTCTATTCTGTAATTG
AGTCAATTGATGGTAATGACTTTATATATAACGAAAAACAATCTAAAATATT
AGAAATGAAACAGGATCGTGGGGGGAATAGATATAGTGCAGTAGATCTTAA
CTCTTCTAAGTATGGGTATTATGAGAAAGTTGGCGGTGGTTTTTATAGATATA
TAGAATCCTTTAACCCCATATCTTCAGAGACACCAAATAAAATAGTCTACAA
GGGGGAATCAGTAGATTTAACTAAGGAGCCAAATTCGGAATTATATTCAGGT
AGGTATTCTATAAATAACAAACAGGTTAATGTTTATTTCTTTCGTGACGCTGA
TGGTACATTTTATAAATCAGAAGGTCTTCATGGTGGGGGAGTTATTAGATACA
TAGATAAACCGTATTCTCAGTTAAGAGAAGGAGATATTGGGTATGATGAGGA
TTTGTTGGATATATACGATGATTCTCCGGTGCTTGAAGACACGTTGCCTGCTT
TATCTTCTGAAATAGTACCAACTCCAGAACATAGTATTAAACAAATTTATTCG
AAAATTAAGGAGGGGCACATAGAACTGTCCGATTCAGACATCATATTGTGTC
GCGGCACAACCGGTATTCAAGCTGAAAATATCGTTGAATATAAAACTGCTGG
AGGGTTTCCTGATTCAAATCCAAATGTAAAAGCACCAGATGAATATATGGCA
CAACAGCAGGTACGTATTGGAAGAATATTGCCTGAATACACATCGGATCTTA
GCGTTGCTGATCGGTTTAGTCGTGAGCATTATCTAATAGTTGTTAAAGTAAAG
GCAAAATATATCACACGAGGAAGTGTTACAGAGAGTGGTTGGGTTATAGATA
AGACCGCACCTGTTGAACCACTTGCGATAATTGATAGAACTTTTGGTATGAA
GGAAAATATCTCAATGGTAAATGCATCGAAATAGTTTTTTTACAATCTATGTC
CTGCCTCCTCTGGTAAAAACGATGCTACATCTTTGAGATGTTGCACGGCAGTA
CGGTGTTGACCGATATATAGTGAAGTACACATCGGTCAACGAATTACCATTG
TCAGCAATATCATCCTGATAAACTCAGTACTCGTGAGCCGCTTGATGACGGG
CGGAATAGCCCAAAAGGTAAACGATCCGCATGGCAGTCCGTCGGAAAGCTG
CTGCTTCATACGACCGCTTAAACCGTCAGTTAGTGTCAGTATCGCTGAAGATC
AGCTTCTTTTGCTGATTTACTCTGTTTTTACCTGCCCTGATGAAATACTCTTTA

ATCATAACATTGATTATATCGAAAATATATTTTTTGCTATCATTAAGAATATT
TATATGTGGAACCGGAGTATTTTGGACGG

>10263
CAGCAAAGCCAAGGTGTCAATATTAGAAAAGAGAGTTATATCTCAAAATCAA
CAGACTGTTGATTTTAACTTTTGCGAATACTTTCTCCAGTCCGTCATGCATGG
ATTAGGATTGCTCATTTTTATACCATTTGTTGTTTTTTATAGCGATTTGCTAAA
AGGGAGAAGAAAAAACGTTCTCCAAAATTAGATTGCAACTGTTTGATTTTAT
GGATAGAGGCATGCTGTTTACTAGACAGTAAAAAGCATGCCAAATTTGCTAT
TAACCTT

>10314
GAGTTAGGTGAGTATCTTTATTCAGGAAATGTCATAAAACTCAGTCAGTTATC
AATTCGTTACCTACCCAATATCAACTCAATCTCATTAATAGAGACAAAACAG
AGTTTGTTGCTACATCGATTATATTCAGATGAAGTACTTCAGAGAAATGGAAC
ACTTATCCCGACACCACTACATGAAGAAAAATCAATTCCCGCTGCCAATATA
AAAACAATGCTCAACAACATACCAACTTACAAAATGTTACCGCCATTCACAG
AAACACAAGGTAATTGTTCTTCTGGCGCAGCCACGTTTTTACGCAAATCAGGC
GCCGAAGAAAAAGATATTCTTGCATGTAGCCCCCGAAATTATGGGCTGCATC
ATAACATAAAAACATGGGACCCCTTGGTTAGAAATTAAGGATCCAAATATTA
TATATTTCATATAAAGCAAGGAAAATATTGCCTGAATAATTGTTCAGGCAAT
ATATCCTTACATGGCACTATTATAATAAACTATTAATATAAAAACACACCAAC
AGAAAAAATTAAGCATCACTTGCAACAAAGGCTTCTTTTTTTGAATCAAAGT
GACATTCGTCTTTTCTCATGATCA

>10393
ACTTCTCTATTCAGGGAAGCCCGTCTGATACAAAGCTTCTATTTCTCGTTAAA
TATTCTCCATTCTGGCGCCAAGCCCTATCTATAGCTGATTGAAAACAACTCTA
AAAACGCAATAATATCAAATAGTTTATTCCTACAGCCGATCATTTTGAGTAAA
ATATTGACGAGAATTTCACAAAAGGTAGATTTCATGCAAACGGATAAATTAG
ACTATCACATAAAGTTCTTATAAGGTTATCTATATGATAAATAGCATTAATTC
TTTTTTTTCCAGCATTCCTCGCAGTATATCAAGTGTTACGCGTAACAGTAGCTT
CACTGCCTCACAGCACAAAAGCACACCTAACACGGTAAAAACCAGCTCACCT
CTTTCTCCCAGCAACAGTCCTGCCAGTGCA

>10396
AACAGGTCGAAGTAATACCCGCACATTACCTCCAGACGCGCTGACGGTAGCA
GGCATGTGTCCGTGGCAGATGTGCACGAACAGGAAGATATACAGAAACGGTC
CAGGTCAGACGATCAGCGTTCAGACTCCGCTCCACACGGACACCGCGACGCA
GATACGCCTCTTGAAGCATATCTGCCTCATCGATCGTACAGAACAGATAGTG
AAACCAGCCATACTGAGGCGCACGAAAACGCCTCCCCTGCTTAATTTCCGGG
TCGGCTTCAGAATTGTGGGATTTTATGTGTTGTGTCATCGGATTCTCCGGTGA
CAGCAGGTGTCAGTTGTTCAGGCTGACTGCGCGAATTGTAAGGCAATACGCC
GGAATGTACAAACAGAAAACCCGTCAGTAAGACGGGCTTAACAAGCAGGGG
CGGTTACTTTAATAATTTCAGTGCCTTTACATCAACTTCAACACTGCTCAGGT

CTTTATCAATTTCACCCTCAATTCTTACTTTGTCTTTCGGAGAAACATTCTGCC
CGGCCCATACGCTGTCATCAATATCCGTGACAATTGTCCCGCTATTGTCACGA
AACTCATAACGTTCATCACCCACTTTTTTAACGATGCTCCCTTCAAGGATAAC
CCATGCATCATCCTTCAGTTCTTTTGCCTGCGCTACTGTTGAACGCTCTGCTTC
TGGCCCCTGGAAACCACCCTGCTGTGCAAAAGCGCCAAAAGACACACCGGA
AATAAGTGCTGCAATCAATACCTTTTTCATTCATAGTCCTCTTTCAGAGATGA
ACATTCAAACAGC

>10402
TACTCACTGTTAACCTCCTGCAACGCTACACGATACGCCTTCTTTATCCACGC
CTTACTGCCATATAATTTAGTCTTCATAATAAACACACCTGCACGACTCGCCG
ATATCCCCGGACAGGTTAACAGCACAGCATCCACCACACGGTTATGCTTCCG
AAACTCCATTACAGTACTGCTGATAACCACCTGCCCCACCGGGCCGTAATCCT
GATACAGGATTTTCACGCAGACACCCTCCTGTCGAAATAAACGTAGTTATTCA
CTATGCGCAGCGGCATGCCTAATTTTCTGGCAATTTCCCTTCTTTGCATGCCTC
TCTGATGCAGTTGCCGCGCCAGCTCAATATCACGCTGAGAATATTTTGCCGAC
GGGTGAAAATCACCACGTAACATCATGCTGATGCCCAGCTCCCGTGCCTTCG
TTCTCACTGCCGCTTCAGTTCGTCCGATAAGTGCGCCAACGCTTTTTACCTTCC
TCGTTCCCGCACACTGCCGGAGTATCATGATTTCCGCCCGGCACCACGTCTTC
CACCCACTCACCGCTGCTGTTCTCTGGTGGCGGTAATATCCCGGAGAATATCC
CGGCACTTGTTCAGCTCCCGCAGCGCGGCGCAGACTCGCTCCCACTTCTGAAC

>10549
TGAGTACCTGTGCGTCAGGGTAACGCAATCAGTTCAGGTTTAGCGGGGTGTC
AAGTCGGTTTGGCTAACCGACCGCAGAATTAACTGTACCTTTAAAGGTCGCC
CCCGATATATCTGCCTGTCAGACATATCGCCGATTCGGTATCTCACTACATAA
AACTCCTTAAAAAACAATGCGATTCTGTTGGCACGAAGCTTTCAGGAAGTCA
GTCCGGAGAAGCGCCAGTACGATACAGAATGAGGTTGAAAAAAACGTTCAC
GTCTGGTCACTTCTTCAAATGTTAAAAAACATCTCGTAAAGAGCCAGTTTATG
CGGGTTCGTTGAGTAGACGTGCGAGCGTCACTTTGTCAGCTTTAACGCTGACG
AAAGTGACGC

Table S3: Sequences of segments that are useful as probes to detect *Shigella* strains

>15843T
AAGCAAATGCCTCTTAGGAATTAATCCTTTATAATCAATTAATTACAAGTGGT
GTTTCGAAATCGTTCGAAATTAATTCTAAATGTCTTTCCATTCCCTCCCTCGTC
CGTCACGGTATTTATCCGTCATGCTGGATGAACGATGTCCAAGGAGCGCCTG
AGTGAAATCACTGCCTTTTTCTGCTGAGTGCAGGCGACCGGATAAACTTCGG
ATTTCATGAAATGGCGGTGGGTCACCTTCCCATTTCAATCCGGATGTGTCTCG
GGCAATCCTGAACTGTGCTGCGATTGTCCTGGCTGTCTTCCCTCCGAGTAACT
TTTCGTTTTTACCGTTTATTTTTTTCAAGTTATCCAGTACATCGGTAAGTGTTA
AATTCAATCTGGAAATCGTGGTGGGGAGA

>15897T
AACAGAAGTCCCGGGATAATCTGGGGCTGAAAAGTGCGGCCACGATGGAAG
CACAGAGCGACATTTACGACCGGACAAAAGGCCGTCTGGCGATACCCGGCGC
ATTCGGCTTTGGGTGTGCTTTTCTGCCTGAAGATGTTATCCGTTTTGACACTAA
GAGTGATTTCCAGGCCTGGGTAAGGAATGCGCTGCCAGGTGAATATTCCGTT
GCTGGCCCCTACGACATCATCATACCCGACACACGGTTTGAAGGGGTGCTCA
GCATCCGGTGGACTGATGCACGCCCTGAGACAACAGAACCGCGGTACAGAG
CCAAATCCCTTACTTTTTACGGCATTAACGGCCCCATTTATCACACCCGCTAC
TGCTACTGGCCCATATCCAGACTGACT

>15898T
GAAAATAAATATAACCACAGAAGATATTATTTACAGAATCGTGGCGAGCTCT
GTCCGCAACAGATGGGGAGACCCTGACATTGGCGGGCTGATTATTGCTGCGT
ACCAGGGAGAAGCTGACGGTGATAAAGTCATCAGACTTGTCAGGGGGCAGTC
ATACAGAGGCTCACGACTGGGACCGGTGGGGATTTCAGTGCCCAGTACTCCC
ACCGGAACGTATATAGCATCCCCACAATTTTTCATTACGGGATGTTCAGAGCA
TTCATTACCGGGGTCATATTGCGCCCTGTCCGGGGTGCCGGATGCTCATGTCT
CTGGCGCAATGCCCGGGCTTTTTATTCGCACATCGTGAGGAATGCACCGTGG
AAATTAAAAAAACCATTAATCCCCGTTATACCGA

>15862T
TGTGTAGTTGGTTTTCCCCAACGTTTAAAAGCCTCAGATTTTGAGGAAGGGGC
GGTAGCTCTTGTAATCTGTTATCAAAAGCACTTAGTTCCTGTAGTTCACATGG
TAATTCTGGTAAGGCAGTAAGCTGGTTGTTGCCAACAGAAAGTTTTTGTAGA
GCCGGGGGAAGATCAGGTAGCGTCTCCAGGCTATTGAGAAGGGCTGAGAGT
GACTGCAAGGAAGATGGAAGAGAAGGTAAACATGTTAAAACTCTGTTAAGT
GAAACATCAAGAGCGACCAGGTGAGGCGGAAGAGCCGGGAGCCTGCTCAGT
CTGTTATCGCTGGCTTTAAGTACAGTTAAGGACGGAGGGAGTTCTGGCAGAG
AGCGTAGCTCATTGCTAGAGATGTTAAGTTCTTGTATGTGCGGGGGCAGGTAT
GGGAGAGAGCGTAATCTGAGTAAACTTAAATTGAGGGCTGGCTCTTGAAAAG
CCAGACATATTTTCAGTAATCGAACTGCCTGTGTTCGGTCTTCTGTTGCAGCA
CCTTCCTTGGCCCAGTTATCCCAGATGCGGTGATAATCCGTGATATTTTGCTC
TTCTACGGATAAACGGGATATACATC

>15901T
ACGCTATCAGAACTGAGGCCAAGTATGCTTTCCGGAATTGCCGTGATCCGGTT
CTGATTAAGCCAGAATATTCTTATAATATGATCATCTCTTTCTGGAAAATCTG
GTATTACTTCCAAAGCATTTCTGGCTGCATCAAGTAATTCCAATGACATTGGT
AACCTCGGAAGTACAGACAGGTGATTGTCACTTACATTAATATATTCTAGTGA
CGCAGGTAATTCAGGAAGTGCGAACAAATGGTTATCACTCACATTTATATATT
CCAGCGACACTGGTAATTCAGGAAGTGCAGATAATTGATTACTGCTTGCATTC
AGCTCTTTCAATGCCCTTGGTAGCTCGGGGAGCATTGATAGTTGGTTATTGCT
TACATTAATTTCATCAAGATTGTCAGGCAATCGTGGTAGAGATCTGAGACCTA
AACAAGATAAGTCCAACGATGTTTCACTGTTCTCCAGACATAATTGGAGCCG
GGTAAAAGCAGTTTCCCTGTTTTCTCCGGAAATGCTGTTTTTAGTCCATTCAA

CCCATTCGGAGAGATAATTATTGTGAACATTGTCGATTGATGTAGTTCTGTAA
AAAGAGACGTTTCCAGTGGATAGGGGGGGATTATTTACAGGAAGCATAATAA
CCTCGCAGAAGGATATCCTGATAAAATGTGATACTGGGAAATAATAAGCTAA
AATAAAATATTTCAGGATTGAAATTTATTGTTTTCATTTTAAGGAAGTGATGT
TGATTTAACTGTAAAAAAACGATACTATTTTCTTAGATGGTAAATGTCTCGCC
TGTGCTATATTTTGTTTTTCGGTTAGTTAAATATCGTACGTTCATATTTGAACG
TTCTGCCGGAATGCATTATCAATAGAGGTAAAGTCGCAACCCCAAATCGTAA
AG

>15961T
CGGTTAACATATTGTCGATTGCCCATATGGAGTGAATGTTGTGAGGTAGAGG
GGGGAGTTCACTAAGAAGATTCCCTATTGCGCTAATCTCTTGTAAAGAAAAT
GGTAAAGGTGGTAAAACTGCCAGCCCATTACATGATACATCTAATGTTTCCA
GTAGTTCTGGTAAAACAGGAAGAGAACATAATTGATTATTTGAGACATGAAG
CTCCTTCAAAGATATGGGGAGTGTGGGTAGTGTGATTAGTTGATTGTGGGAC
GCATTCAATAATTTAAGTCCTTGAGGCAAAGCAGGCAGTTCAATAAGTCTGTT
ATAGCTGACATTAAGCTGTGTAAGGGACGCAGGCAATGGGGAGATTAAGCTT
AAATTATTTTTACTTATATTAATTGATTTAATTCCCGGGGGGATTTCAGGTAA
TGTTGTCAGGCCTAATTCAGACAAGTCTAGGTTCGTCTCTTGGTTTTGTAGAC
ATGATACTAGTCGCTGAAAAGCGATGTCTCGTTGTTCTTCTTGTATGCGGTTA
TTTTTCCATTCAGTCCAATGGGTTAGATAACTTTCATATGCGCGGCTAGTGTC
GATTGAATAAGTGGAGAACGAATTTGAAATTAATCTGTGATTGTTATTTGTCG
GGAGCATAAATATCAGGGCGTATTCTTTATGTTGAGAGGTGTTGGATTCTTTT
TTATCTGCTTTGCAATGTTATTGGTTCCCTTCAGCAGTAGACAGGAACTTCTG
GAGGAGGCGGATATATAGTTTGTGTAAGCATATTAACTCCATATGTTATATAT
TGAAGAACTTCTGCTTTACACTATCCTAAACTGGATCGGTTTAATGTAAAAAA
ACGATACTTTTTGGGGGAGGGGGCAATTTTTCCTTAAGGTATGAAAAAAGGA
GCTGAAGCTATATAATAGCTATTAATGCCACTGATATCAGTGAATCATGTATA
TAAAAAA

>15983T
ACAGAGACTCCCTGATCCGTGAGGACAGAGGATTGTCTTCGAGGATGATAGT
GCAGGTCGGATCAAGGCTAAGTATATTTTCCGGAATGTGTGTGATGCGATTCT
CGCGGCACCGGAAAAATATCTCGGTTTCCTCTGAGTGATGATTTCTTACAGGT
ACGGCTGGTAGGCTTTCCAGAAGATTAGTACTTACATCGAGCGCTTCCAGTG
ATTCAGGTAACTCAGGAAGAAATGTCAGCTGGTTATTTCTTACTGAGAGCACT
TCCAGCGATGTAGGTAATTCAGGAAGCATGGTTAGCTGATTGTTATCTGCATT
AATATATTCCAGCAATGCAGGCAATTCAGGAAGCATGGTTAGTTGGTTGTTAT
CTACATCAAGATGTTTCAGAGATGCGGGTAATTCAGGAAGTGTTGACAGGTG
ATTGTCACAGGCGTCAAGGTATTCCAGCGATGCTGGCAATTCTGGTAATGAT
ATTAGGGCATTCTGAGTAATTTCCAGAACAGTGAT

>16002T
GTATCTTACTCAAGACTAATGTCTTCTGGTATAAATCCTCTGCATTGTGAATA
TGTAAAACATCAAAAAAAGATGAACCAAATATACGAAAAAAGCAAATATAC

TTACCAGGATAGAAAGCATGATATAACACTCATTCCCTTATGGCGAATATTCT
GAAACACCCTCCTCATTAGGAGAGATACATTAAATACAATTTATTCACAATCT
AATTCCATATCCATATGCGAGAAATTAATATGCTCAGAAATATTTCATCCTGT
TTATTTCCACATATCAGCACAATTACATCCCCCAACCATTATTTGTCCGAATG
GGATGATTGGGAGAAACAGGGGTTACCGGAAGAACAGCGTACTGAGGCGGT
AAGAAGACTTCGTGCATGTCTTACCTCTAAGGGGCATAAACTGGACCTGCGA
GCCTTGGCGCTTTCCTCGTTACCTGTACTCCCTGCTTGCATTAAAAAGCTTGAT
GTGAGCTGTAATAAATTAACCATCCTTACTGATCTACCTGAAAATATTAAAGA
ACTTATTGCAAGAGATAATTTCTTAACACATATATCTGCATTACCACATTATC
TAATAACTTTGGATGTGTCCGAAAATCAATTAGAGAATCTGCCGTTATTACCA
GACACCATCAAATCACTAAGCGCAGAGTATAATAGGTTATCCACACTGCCTT
CATTACCCTTGAATTTAAAAAAACTTGAGGTTAGGAACAACGAACTGCAAAC
TCTTCCATCTCTGCCTTCTAATCTTAAGATACTTAAGGTTGCGCACAACCATCT
TACTGAACTGCCCCCTTTACCTAGGAGACTGCAACTTCTTTTTGCATATAGCA
ATAGATTAAGCAACTTACCAAACATCCAAGAAAATATTATCATGAGAAGATT
TTTTTATTTTGAAAACAACCAAATAACTACAATCCCGACAAATCTTTTTCGTT
TAGATCCTCATATAACTATTGAGATTGCAAATAACCCCTTATCAGATCAAACT
CTGCTATTCTTAATACAGCAAACTTCGGTTCCAAATTTTAACGGGCCTCAGTT
TCGTATTTCCCTGTCAGACCAAAACAGACTGTTTTTACGCCAGATGTTGCCGC
AAAATTTACATTCGCGCCATATCAGAGTCATCACTGAAGGGGGGCAGAACTT
TCAGATCCCCCCTCTTCCCGAAACTGTGGCAGCCTGGTTTCCTGAAGCAGATC
GTCGGGAGGTTTCTACACAATGGACTTCTTTTTCCACCGAGGAGAATTCCCGG
GCATT

Table S4. EHEC specific amplification probes

Probe	Primer Sequence	Sequence ID	Primer Location	Amplicon Length
EHEC 1				
Forward Primer	5' TGGCGCGGGGAGAGTAGT 3'	10254	1.18	509bp
Reverse Primer	5' ATGGCAGATAATAACGAAGCAACA 3'		509…486	
EHEC 2				
Forward Primer	5' GCAAAGCCAAGGTGTCAAT 3'	10263	3…21	220bp
Reverse Primer	5' TGCCTCTATCCATAAAATCAAAC 3'		222…200	
EHEC 3				
Forward Primer	5' CCCGCTGCCAATATAAAAACAA 3'	10314	196…217	405bp
Reverse Primer	5' TGATCATGAGAAAAGACG 3'		600…583	
EHEC 4				
Forward Primer	5' GGAAGCCCGTCTGATACAAA 3'	10393	15…34	367bp
Reverse Primer	5' CTGGGAGAAAGAGGTGAGC 3'		381…363	
EHEC 5				
Forward Primer	5' CCATACTGAGGCGCACGAAAAC 3'	10396	216…237	479bp
Reverse Primer	5' TCCAGGGGCCAGAAGCAGAG 3'		694…675	
EHEC 6				
Forward Primer	5' ACGCCTTCTTTATCCACGCCTTAC 3'	10402	35…58	538bp
Reverse Primer	5' GGGATATTACCGCCACCAGAGAAC 3'		572…549	
EHEC 7				
Forward Primer	5' GGCTAACCGACCGCAGAAT 3'	10549	63…81	267bp
Reverse Primer	5' CTACTCAACGAACCCGCATAAACT 3'		329…306	
EHEC 8				
Forward Primer	5' GGCGTTGATTTAGAGACCA 3'	10258	2508…2526	449bp
Reverse Primer	5' TTGCCCAGAAATAACGAAC 3'		2956…2938	

Table S5. *Shigella* specific amplification probes

Probe	Primer Sequence	Sequence ID	Primer Location	Amplicon Length
Shigella 1				
Forward Primer	5' ATGGGATGATTGGGAGAAACA 3'	16002T	314…334	447bp
Reverse Primer	5' AGATTAGAAGGCAGAGATGGAAGA 3'		760…737	
Shigella 2				
Forward Primer	5' CCGGGGGAAGATCAGGTAGC 3'	15862T	160…179	435bp
Reverse Primer	5' TATCCCGTTTATCCGTAGAAGAG 3'		594…572	
Shigella 3				
Forward Primer	5' ATATGGAGTGAATGTTGTGAGGTA 3'	15961T	25…48	549bp
Reverse Primer	5' CTAGCCGCGCATATGAAAGTTA 3'		573…552	
Shigella 4				
Forward Primer	5' CTCGGAAGTACAGACAGGTGATTG 3'	15901T	163…186	483bp
Reverse Primer	5' CTTCTGCGAGGTTATTATGCTTCC 3'		645…622	
Shigella 5				
Forward Primer	5' CCGCAACAGATGGGGAGAC 3'	15898T	55…73	338bp
Reverse Primer	5' ACGGGGATTAATGGTTTTT 3'			
Shigella 6				
Forward Primer	5' ATTCGGCTTTGGGTGTGCTTTTCT 3'	15897T	104…127	207bp
Reverse Primer	5' CTGTACCGCGGTTCTGTTGTCTCA 3'		310…287	
Shigella 7				
Forward Primer	5' GGAGCGCCTGAGTGAAAT 3'	15843T	150…167	158bp
Reverse Primer	5' AGGGAAGACAGCCAGGACAATC 3'		307…286	
Shigella 8				
Forward Primer	5' GGATTGTCTTCGAGGATGATAGTG 3'	15983T	30…53	419bp
Reverse Primer	5' CTGGAATACCTTGACGCCTGTGAC 3'		448…425	

Table S6: Amplification profiles

EHEC-specific probes		
Strain[a]	**Amplification Profile**[b]	**Probability EHEC**
Escherichia coli O55:H7 RM12579	1000	0.98
***Escherichia coli* E32511**	1110	>0.9999
***Escherichia coli* G5101**	1110	>0.9999
***Escherichia coli* 5905**	1110	>0.9999
***Escherichia coli* DEC8C**	1011	>0.9999
***Escherichia coli* DEC10B**	1011	>0.9999
***Escherichia coli* DEC10C**	1011	>0.9999
***Escherichia coli* DEC9F**	1100	>0.9998
***Escherichia coli* VP30**	1011	>0.9999
***Escherichia coli* RDEC-1**	1000	0.98
***Escherichia coli* MT#10**	0111	>0.9999
***Escherichia coli* M103-19**	1011	>0.9999
***Escherichia coli* MI01-88**	1011	>0.9999
***Escherichia coli* MI05-14**	1011	>0.9999
***Escherichia coli* DA-21**	1011	>0.9999
***Escherichia coli* RD8**	0000	0
***Escherichia coli* DA-5**	0011	0.9999
***Escherichia coli* IH 16**	0111	>0.9999
***Shigella*-specific probes**		
Strain[a]	**Amplification Profile**	**Probability *Shigella***
Shigella sp. 2770-51	11011111	>0.9999
Shigella sp. K-147	10001001	>0.9999
Shigella sp. 3554-77	00011001	0.9998
Shigella flexneri 2457T	11111110	>0.9999
Escherichia coli 53638	11111101	>0.9999
Escherichia coli H10407	00010000	0.893
***Escherichia coli* EC4115**	00010000	0.893
Sodalis glossinidius str 'morsitans DNA'	00000010	0.96
Citrobacter rodentium ICC168	00001000	0.963

[a]Boldface strains are EHEC

[b]EHEC amplification profiles based on probes EHEC1, EHEC4, EHEC5, EHEC6

[c]Although probes EHEC 5 and EHEC 6 did not amplify, either experimentally or *in silico*, in any non-EHEC strains these probabilities are based on the assumption that the probability of doing so is 0.01.

[c]*Citrobacter rodentium* ICC168is known to have acquired EHEC and EPEC associated sequences from *E. coli* (Petty, et al. 2010)

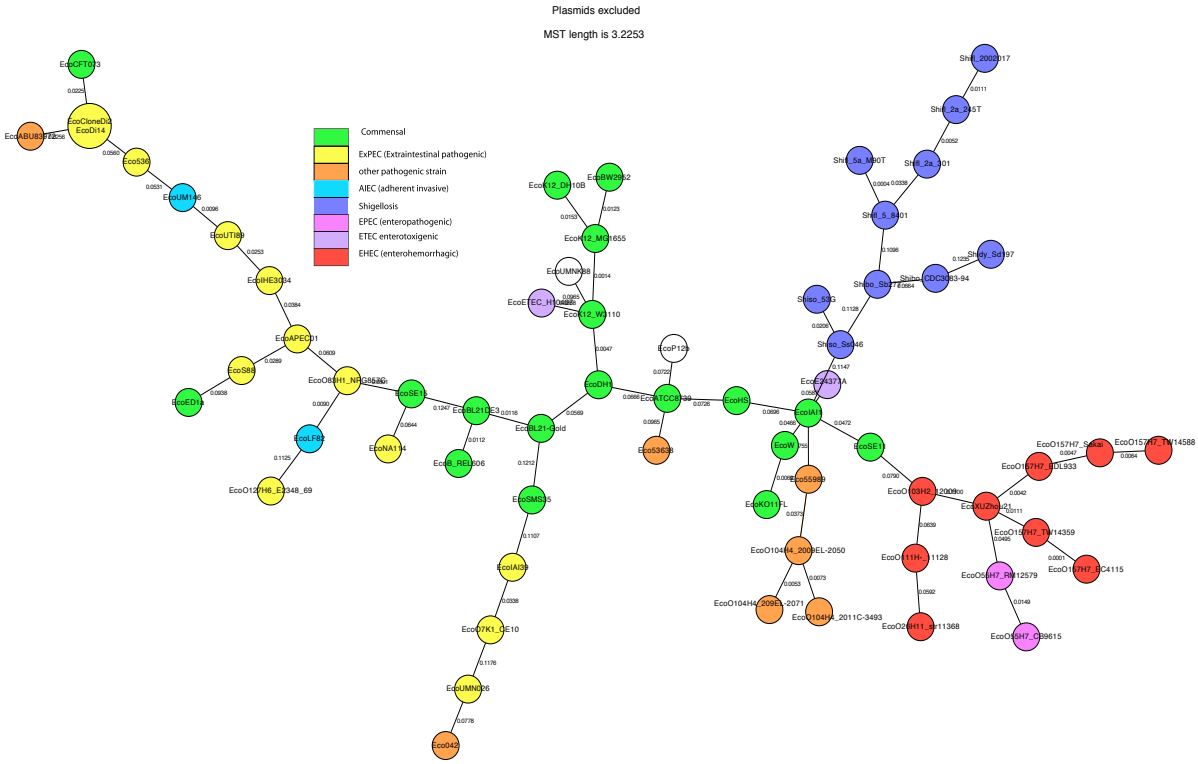


Figure S1: Updated minimum spanning tree from sequenced genomes excluding plasmids.

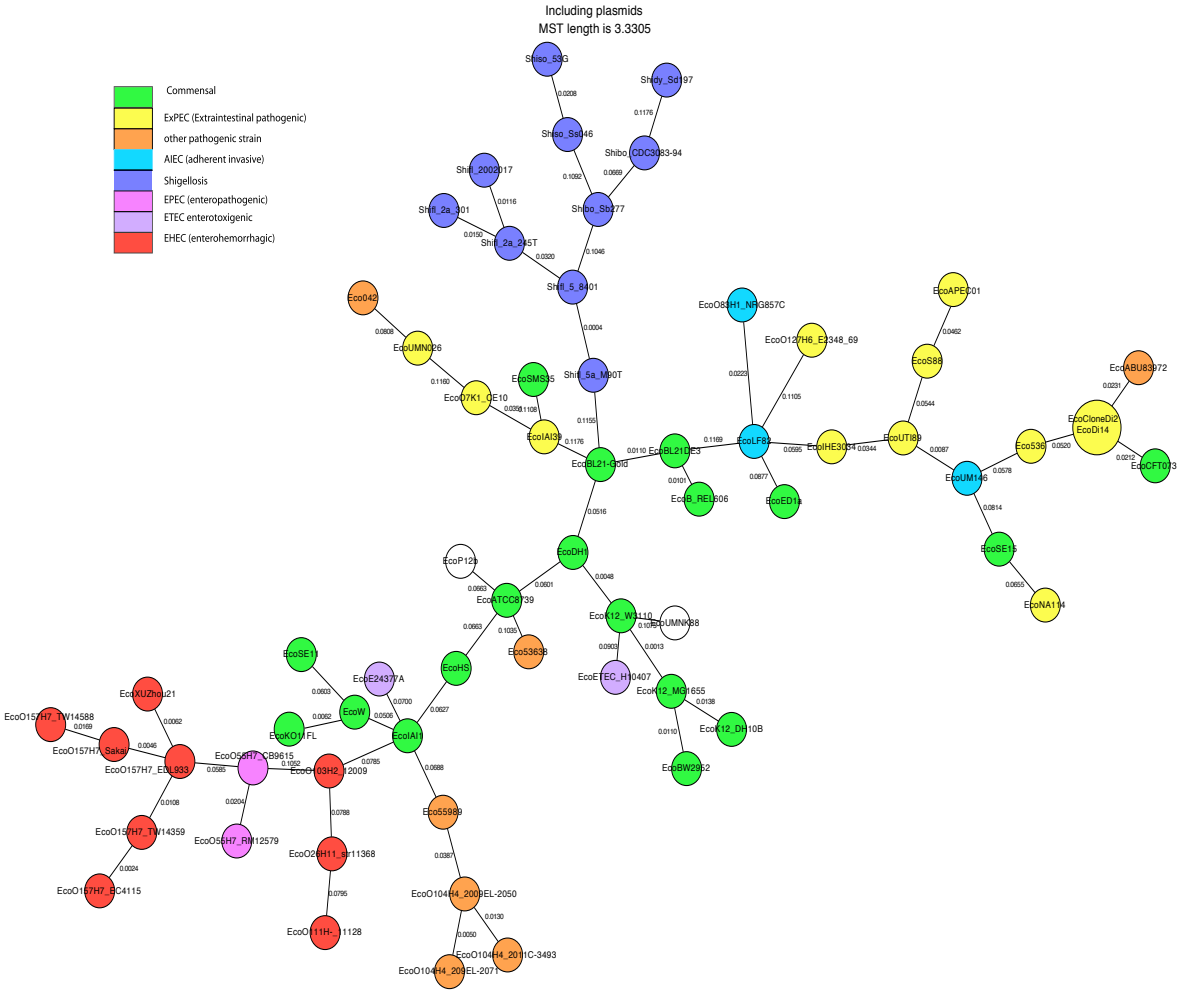


Figure S2: Updated minimum spanning tree from sequenced genomes including plasmids.

Availability of programs

BopGenomes, GetProbs, and PredictPhenotypes are part of the BopGenomes Suite that is available for Mac, Windows and Unix platforms free of charge at http://bellinghamresearchinstitute.com/software/index.html.

Research Acknowledgements

This work was supported by grant number 1R15GM090164-01A1 from the Institute of General Medical Sciences of the National Institutes of Health to M.B.

References

Allen CA, Niesel DW, Torres AG 2008. The effects of low-shear stress on Adherent-invasive Escherichia coli. Environ Microbiol 10: 1512-1525

Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, Lee SY, Nielsen LK 2011. The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. BMC Genomics 12: 9

Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N 2011. Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. J Bacteriol 193: 4272-4273

Bardhan P, Faruque AS, Naheed A, Sack DA 2010. Decrease in shigellosis-related deaths without Shigella spp.-specific interventions, Asia. Emerg Infect Dis 16: 1718-1723

Blattner FR, Plunket III G, Bloch CA, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1462

Crossman LC, Chaudhuri RR, Beatson SA, et al. 2010. A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic Escherichia coli strain H10407. J Bacteriol 192: 5822-5831

Davies, K. (2001). *Cracking The GENOME*. New York, NY: The Free Press

DeSalle, R., Yudell, M. (2005). *Welcome to the GENOME*. Canada: Wiley-Liss

Ferenci T, Zhou Z, Betteridge T, Ren Y, Liu Y, Feng L, Reeves PR, Wang L 2009. Genomic sequencing reveals regulatory mutations and recombinational events in the widely used MC4100 lineage of Escherichia coli K-12. J Bacteriol 191: 4025-4029

Hall BG, Ehrlich GD, Hu FZ 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology 156: 1060-1068

Hall BG, Kirkup BC, Riley MC, Barlow M 2012. Clustering *Acinetobacter* Strains by Optical Mapping. Mol. Biol. Evol. (submitted)

Hayashi T, Makino K, Ohnishi M, et al. 2001. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8: 11-22

Hiller NL, Janto B, Hogg JS, et al. 2007. Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. J Bacteriol 189: 8186-8195

Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD 2007. Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol 8: R103

Jin Q, Yuan Z, Xu J, et al. 2002. Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. Nucleic Acids Res 30: 4432-4441

Krause DO, Little AC, Dowd SE, Bernstein CN 2011. Complete genome sequence of adherent invasive Escherichia coli UM146 isolated from Ileal Crohn's disease biopsy tissue. J Bacteriol 193: 583

Kulasekara BR, Jacobs M, Zhou Y, et al. 2009. Analysis of the genome of the Escherichia coli O157:H7 2006 spinach-associated outbreak isolate indicates candidate genes that may enhance virulence. Infect Immun 77: 3713-3721

Kyle JL, Cummings CA, Parker CT, et al. 2012. Escherichia coli serotype O55:H7 diversity supports parallel acquisition of bacteriophage at Shiga toxin phage insertion sites during evolution of the O157:H7 lineage. J Bacteriol 194: 1885-1896

Liu B, Hu B, Zhou Z, Guo D, Guo X, Ding P, Feng L, Wang L 2012. A novel non-homologous recombination-mediated mechanism for Escherichia coli unilateral flagellar phase variation. Nucleic Acids Res 40: 4530-4538

Machado J, Grimont F, Grimont PA 2000. Identification of Escherichia coli flagellar types by restriction of the amplified fliC gene. Res Microbiol 151: 535-546

Moriel DG, Bertoldi I, Spagnuolo A, et al. 2010. Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic Escherichia coli. Proc Natl Acad Sci U S A 107: 9072-9077

Nie H, Yang F, Zhang X, et al. 2006. Complete genome sequence of Shigella flexneri 5b and comparison with Shigella flexneri 2a. BMC Genomics 7: 173

Ochman H, Whittam TS, Caugant DA, Selander RK 1983. Enzyme polymorphism and genetic population structure of *Escherichia coli* and *Shigella*. J. Gen. Microbiol. 129: 2715

O'Leary J, Corcoran D, Lucey B 2009. Comparison of the EnteriBio Multiplex PCR System with Routine Culture for Detection of Bacterial Enteri Pathogens. J Clin Micorbiol 47: 3449-3453

Onodera NT, Ryu J, Durbic T, Nislow C, Archibald JM, Rohde JR 2012. Genome sequence of Shigella flexneri serotype 5a strain M90T Sm. J Bacteriol 194: 3022

Orskov I, Orskov F, Jann B, Jann K 1977. Serology, chemistry, and genetics of O and K antigens of Escherichia coli. Bacteriol Rev 41: 667-710

Petty NK, Bulgin R, Crepin VF, et al. 2010. The Citrobacter rodentium genome sequence reveals convergent evolution with human pathogenic Escherichia coli. J Bacteriol 192: 525-538

Pupo GM, Lan R, Reeves PR 2000. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A 97: 10567-10572

Reeves PR, Liu B, Zhou Z, et al. 2011. Rates of mutation and host transmission for an Escherichia coli clone over 3 years. PLoS One 6: e26907

Rolland K, Lambert-Zechovsky N, Picard B, Denamur E 1998. Shigella and enteroinvasive Escherichia coli strains are derived from distinct ancestral strains of E. coli. Microbiology 144: 2667-2672.

Salipante SJ, Hall BG 2011. Inadequacies of minimum spanning trees in molecular epidemiology. J Clin Microbiol 49: 3568-3575

Sims GE, Kim SH 2011. Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). Proceedings of the National Academy of Sciences of the United States of America 108: 8329-8334

Suzuki S, Ono N, Furusawa C, Ying BW, Yomo T 2011. Comparison of sequence reads obtained from three next-generation sequencing platforms. PLoS One 6: e19534

Tettelin H, Masignani V, Cieslewicz MJ, et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102: 13950-13955

Toh H, Oshima K, Toyoda A, et al. 2010. Complete genome sequence of the wild-type commensal Escherichia coli strain SE15, belonging to phylogenetic group B2. J Bacteriol 192: 1165-1166

Turner PC, Yomano LP, Jarboe LR, York SW, Baggett CL, Moritz BE, Zentz EB, Shanmugam KT, Ingram LO 2012. Optical mapping and sequencing of the Escherichia coli KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the Zymomonas mobilis pdc and adhB genes. J Ind Microbiol Biotechnol 39: 629-639

Wei J, Goldberg MB, Burland V, et al. 2003. Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. Infect Immun 71: 2775-2786

Wine E, Ossa JC, Gray-Owen SD, Sherman PM 2009. Adherent-invasive Escherichia coli, strain LF82 disrupts apical junctional complexes in polarized epithelia. BMC Microbiol 9: 180

Wirth T, Falush D, Lan R, et al. 2006. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol 60: 1136-1151

Xiong Y, Wang P, Lan R, et al. 2012. A novel Escherichia coli O157:H7 clone causing a major hemolytic uremic syndrome outbreak in China. PLoS One 7: e36144

Yang F, Yang J, Zhang X, et al. 2005. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res 33: 6445-6458

Ye C, Lan R, Xia S, et al. 2010. Emergence of a new multidrug-resistant serotype X variant in an epidemic clone of Shigella flexneri. J Clin Microbiol 48: 419-426

Zhang Y, Lin K 2012. A phylogenomic analysis of Escherichia coli/Shigella group: implications of genomic features associated with pathogenicity and ecological adaptation. BMC Evolutionary Biology 12: 174

Zdziarski J, Brzuszkiewicz E, Wullt B, et al. 2010. Host imprints on bacterial genomes--rapid, divergent evolution in individual patients. PLoS Pathog 6: e1001078

Zhou Z, Li X, Liu B, et al. 2010. Derivation of Escherichia coli O157:H7 from its O55:H7 precursor. PLoS One 5: e8700

www.bioquill.comth.jpeg

www.ehec.nl-th.jpeg

laterminalrosario.wordpress.com