Using Comparative Genomics to Investigate the Diversity of Pathogenic *E. coli*

Tracy H. Hazen, Ph.D.

June 16, 2016

ASM Microbe 2016: DIY Genome Workshop

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE
INSTITUTE FOR GENOME SCIENCES

Outline

- Large Scale Comparative Genomics Tools:
  - Phylogenomics
  - Using BLAST Score Ratio (BSR) for *in silico* detection of gene-based differences
- Examples of *E. coli* Comparative Genomics:
  - Genomics of the AEEC: EPEC and EHEC
  - Comparing EPEC from clinical outcomes of differing severity

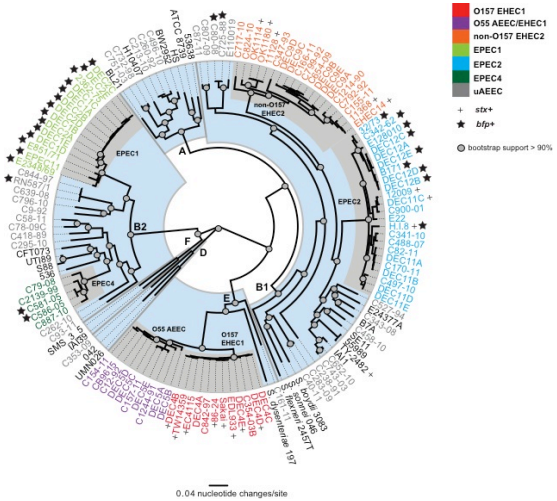
Large Scale Comparative Genomics Tools

- Phylogenomics
  - Conserved regions of genomes are used to construct phylogenies
    - Whole-genome alignments
    - SNP-based
- BLAST Score Ratio (BSR)
  - *In silico* detection of the presence of genes of interest, or the comparison of entire genomes to locate conserved or unique genes
    - BSR (Rasko *et al.* 2005. BMC Bioinf. PMID: 15634352)
    - Large Scale-BLAST Score Ratio (LS-BSR) (Sahl *et al.* 2014. PeerJ. PMID: 24749011)

Phylogenomics

- Whole-genome alignments
  - Smaller numbers of genomes (<100)
  - Mugsy (Angiuoli & Salzberg. 2011. Bioinf. PMID: 21148543)
- SNP-based analysis
  - Large number of genomes (>100)
  - In-Silico Genotyper (ISG) (Sahl *et al.* 2015. BioRxiv)
- Phylogeny construction
  - RAxML (Stamatakis. 2006. Bioinf. PMID: 16928733) or FastTree 2 (Price et al. 2010. PLoS One. PMID: 20224823)

Hazen *et al.* 2014. AAC (PMID: 24914121)



Hazen *et al.* 2013. PNAS (PMID: 23858472)

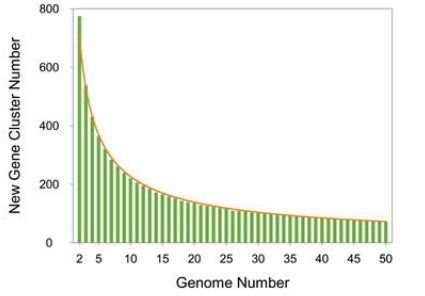
BSR and LS-BSR

- BLAST Score Ratio (BSR) or Large Scale-BLAST Score Ratio (LS-BSR)
  - *In silico* detection of known genes in genome sequences
  - Predict and classify genes to "gene clusters" and detect the presence of the gene clusters in a large set of genomes using BSR

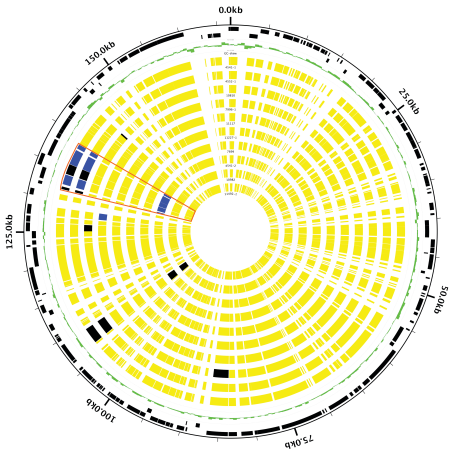
Rasko *et al.* 2005. Visualization of comparative genomic analyses using BLAST score ratio. BMC Bioinf. (PMID: 15634352)
Sahl *et al.* 2014. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ. (PMID: 24749011)

Presentation of BSR/LS-BSR Data

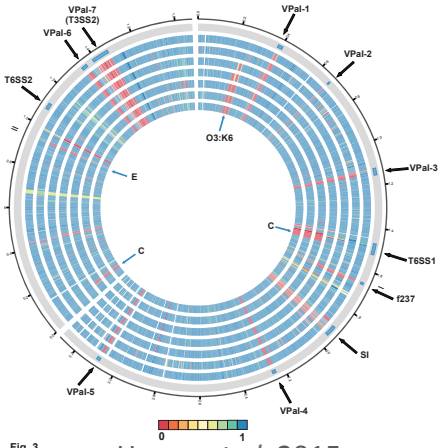
- Pangenome Analysis: visualize distribution of core vs. non-core using pangp (http://pangp.big.ac.cn/)
- Circular Displays: (using Circos or Circleator) genomic similarity relative to a reference



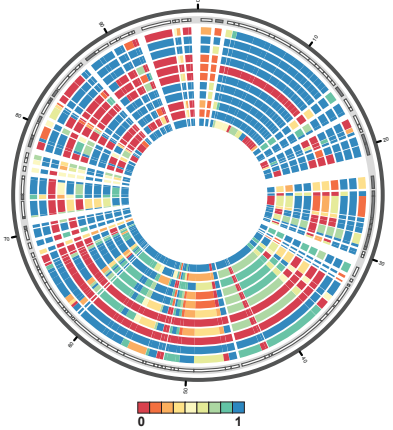
http://pangp.big.ac.cn/



Hazen *et al.* 2014. AAC
(PMID: 24914121)



Hazen *et al.* 2015.
Front. Microbiol.
(PMID: 25852665)



Hazen *et al.* 2015
Infect. Immun.
(PMID: 26238712)

Circleator: Crabtree *et al.* 2014. Bioinf. (PMID: 25075113)

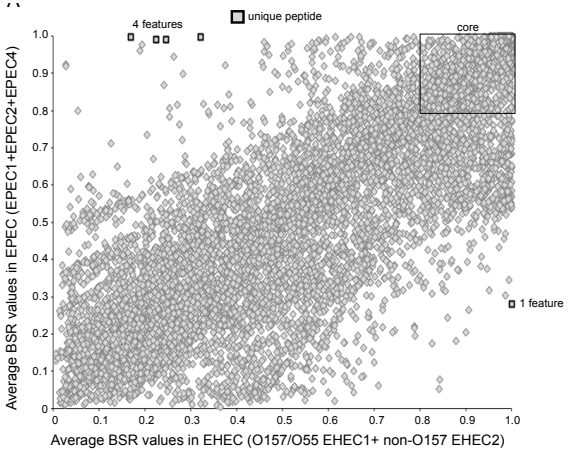
Circos: Krzywinski *et al.* 2009. Genome Res. (PMID: 19541911)

Presentation of BSR/LS-BSR Data

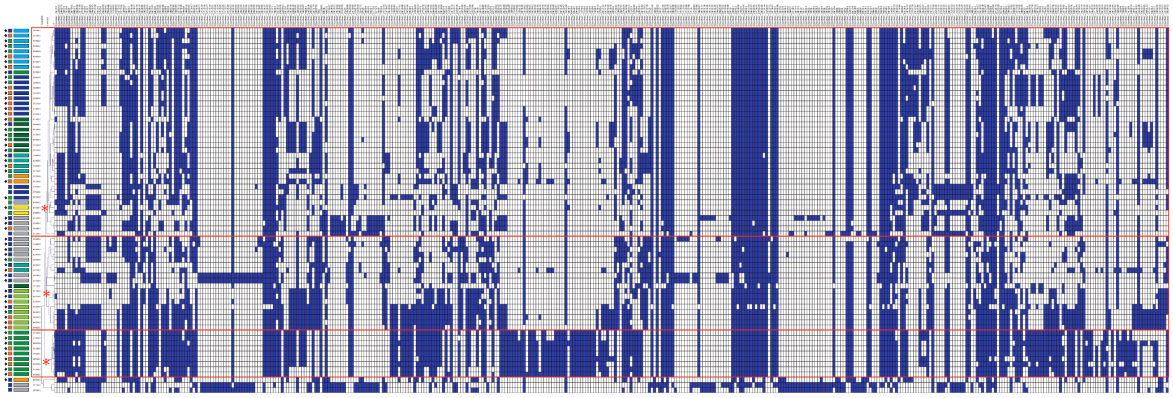
- *In Silico* Gene Distributions:
  - Summary Tables
  - Scatter Plots
  - Heatmaps (TM4/MeV or R-gplots, etc.)

Phylogenomic Groups[a]	Clinical Outcome[b]	No. Genomes	No. of Clusters Present by Percentage of Genomes: 100%[c]	≥50%[c]	<50%[c]	100% only[d]	No. of Clusters in ≥1 genome: LI only	NSI only	LI and NSI only	NSI and AI only	AI only
All isolates	All	70	1,316	5,945	8,467	-	-	-	-	-	-
	LI	24	2,867	6,134	8,278	0	455	-	-	-	-
	NSI	23	2,907	5,922	8,490	0	-	841	-	-	-
	LI+NSI	47	2,110	5,961	8,451	0	-	-	1,984	-	-
	NSI+AI	46	1,632	5,943	8,469	0	-	-	-	2,571	-
	AI	23	2,412	5,858	8,554	0	-	-	-	-	1,199
Phylogroup E	All	2	6,152	6,788	7,624	-	-	-	-	-	-
	LI	0	-	-	-	-	-	-	-	-	-
	NSI	0	-	-	-	-	-	-	-	-	-
	LI+NSI	0	-	-	-	-	-	-	-	-	-
	NSI+AI	2	6,152	6,788	7,624	-	-	-	-	-	-
	AI	2	6,152	6,788	7,624	-	-	-	-	-	-
Phylogroup B1	All	24	3,368	6,497	7,915	-	-	-	-	-	-
	LI	10	4,331	6,766	7,646	0	197	-	-	-	-
	NSI	8	4,252	6,595	7,817	0	-	952	-	-	-
	LI+NSI	18	3,801	6,512	7,900	0	-	-	1,714	-	-
	NSI+AI	14	3,577	6,560	7,852	0	-	-	-	1,562	-
	AI	6	4,288	6,651	7,761	0	-	-	-	-	384
Phylogroup A	All	5	3,750	5,891	8,521	-	-	-	-	-	-
	LI	1	6,405	6,405	8,007	213	213	-	-	-	-
	NSI	3	4,268	5,590	8,822	3	-	945	-	-	-
	LI+NSI	4	4,181	6,456	7,956	431	-	-	1,655	-	-
	NSI+AI	4	3,834	6,291	8,121	84	-	-	-	1,798	-
	AI	1	5,892	5,892	8,520	335	-	-	-	-	335
Phylogroup B2	All	39	2,773	6,188	8,224	-	-	-	-	-	-
	LI	13	4,009	6,234	8,178	0	436	-	-	-	-
	NSI	12	4,225	6,417	7,995	0	-	142	-	-	-
	LI+NSI	25	3,463	6,259	8,153	0	-	-	985	-	-
	NSI+AI	26	3,080	6,181	8,231	0	-	-	-	1,385	-
	AI	14	3,431	6,178	8,234	0	-	-	-	-	984

Hazen *et al.* 2016. Nat. Microbiol.



Hazen *et al.* 2013. PNAS (PMID: 23858472)



Hazen *et al.* 2016. Nat. Microbiol.

Outline

- Large Scale Comparative Genomics Tools:
  - Phylogenomics
  - Using BLAST Score Ratio (BSR) for *in silico* detection of gene-based differences
- Examples of *E. coli* Comparative Genomics:
  - Genomics of the AEECs: EPEC and EHEC
  - Comparing EPEC from clinical outcomes of differing severity

Enteropathogenic *E. coli* (EPEC)

- What is EPEC?
  - Pathovar of diarrheagenic *E. coli*
  - Causes diarrhea among young children (<12 months) in developing countries
  - A few serogroups (O127, O55, O111, O119, O128)
  - Originally characterized in a couple of MLST-based lineages (EPEC1, EPEC2), and additional lineages (EPEC3-EPEC6) have been identified

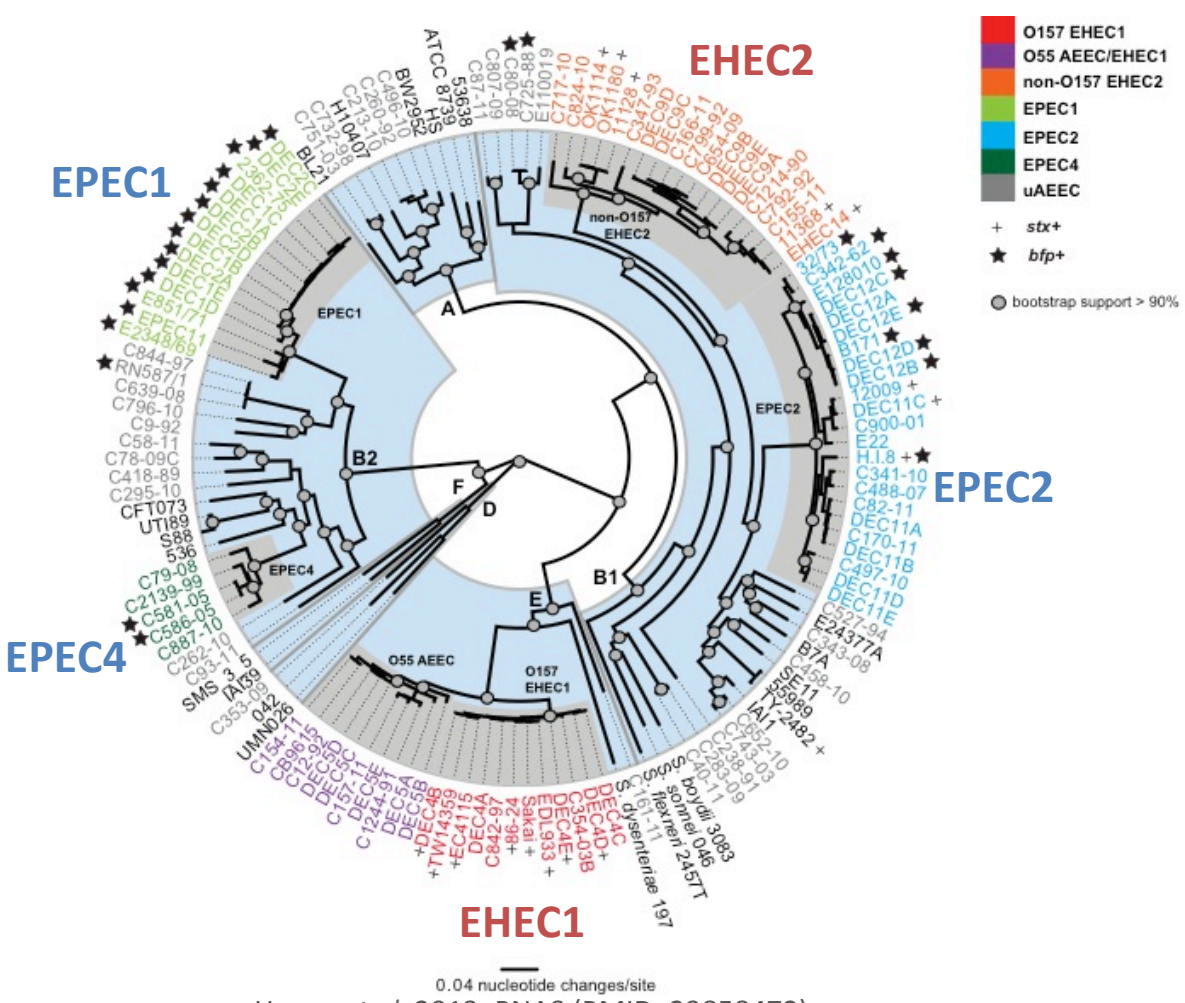
- How is EPEC identified?
  - Presence of the Locus of Enterocyte Effacement (LEE), encoding an adherence factor (intimin) and a type III secretion system (T3SS)
  - Absence of Shiga-toxin genes
  - Typical EPEC: Presence of Bundle-Forming Pilus (BFP)
  - Atypical EPEC: Absence of BFP

AEEC: EPEC and EHEC

- Attaching and Effacing *E. coli* (AEEC)
  - *E. coli* that carry the LEE region
  - Enteropathogenic *E. coli* (EPEC) can also contain the bundle-forming pilus (BFP)
  - Enterohemorrhagic *E. coli* (EHEC) contain Shiga-toxin
    - Leading cause of severe foodborne diarrheal illness in US
    - Consumption of undercooked beef, contaminated vegetables, and cattle are a reservoir
    - Bloody or non-bloody diarrhea, cases of hemolytic uremic syndrome (HUS)
    - O157:H7 (EHEC1) and non-O157 (EHEC2)

Kaper *et al.* (2004) Nat. Rev. Microbiol.

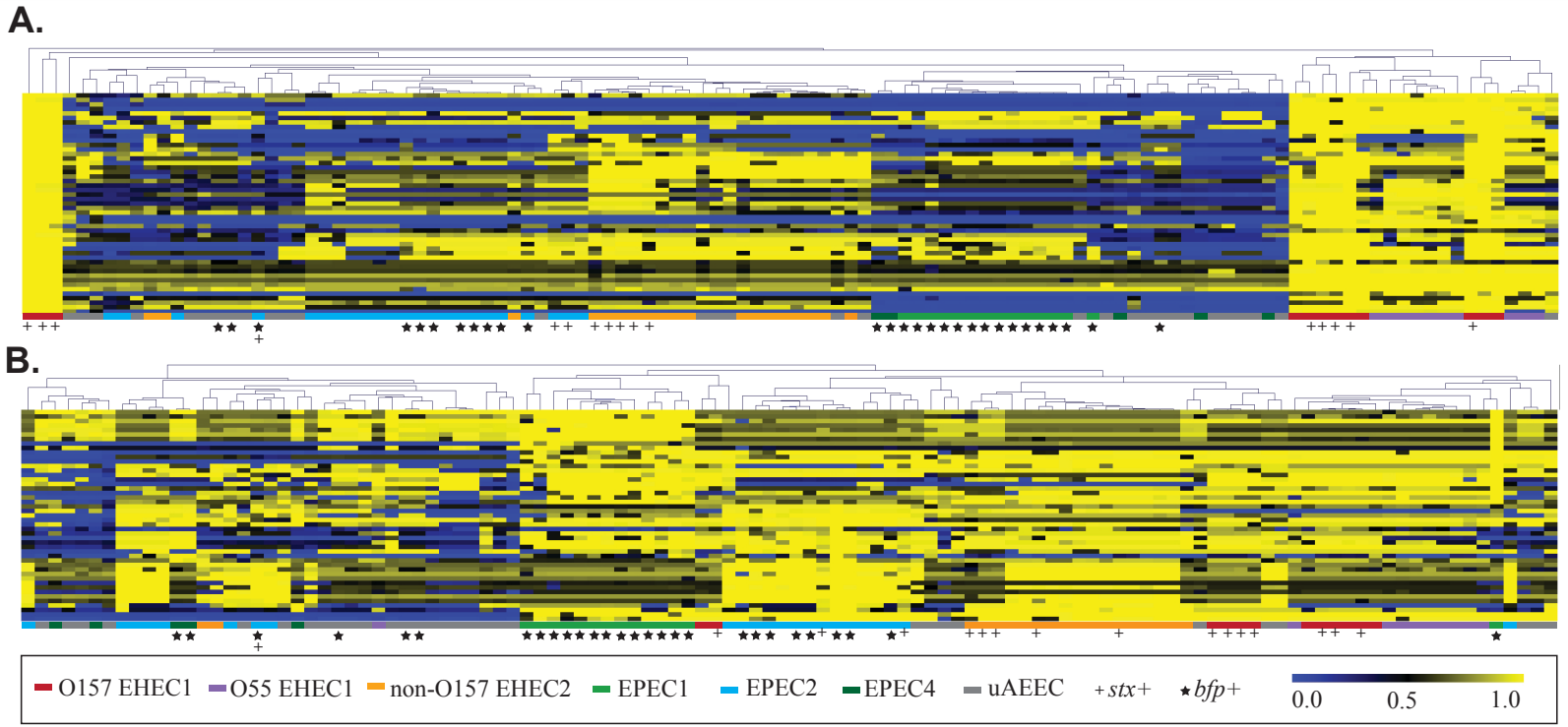
Phylogenomic Diversity of EPEC and EHEC



Hazen *et al.* 2013. PNAS (PMID: 23858472)

- 2.5 Mb whole-genome alignment (Mugsy) of 114 AEEC genomes and 24 *E. coli* and *Shigella* reference genomes
  - 102 newly described AEEC genomes
- By Virulence Gene Content:
  - 15 EHEC (LEE+/STX+/BFP-)
  - 28 Typical EPEC (LEE+/STX-/BFP+)
  - 71 Atypical EPEC (LEE+/STX-/BFP-)
- By Phylogenomic Lineage:
  - 22 EHEC1 (12 O157 EHEC & 10 O55 AEEC)
  - 19 EHEC2
  - 14 EPEC1
  - 23 EPEC2
  - 5 EPEC4
  - 31 unclassified

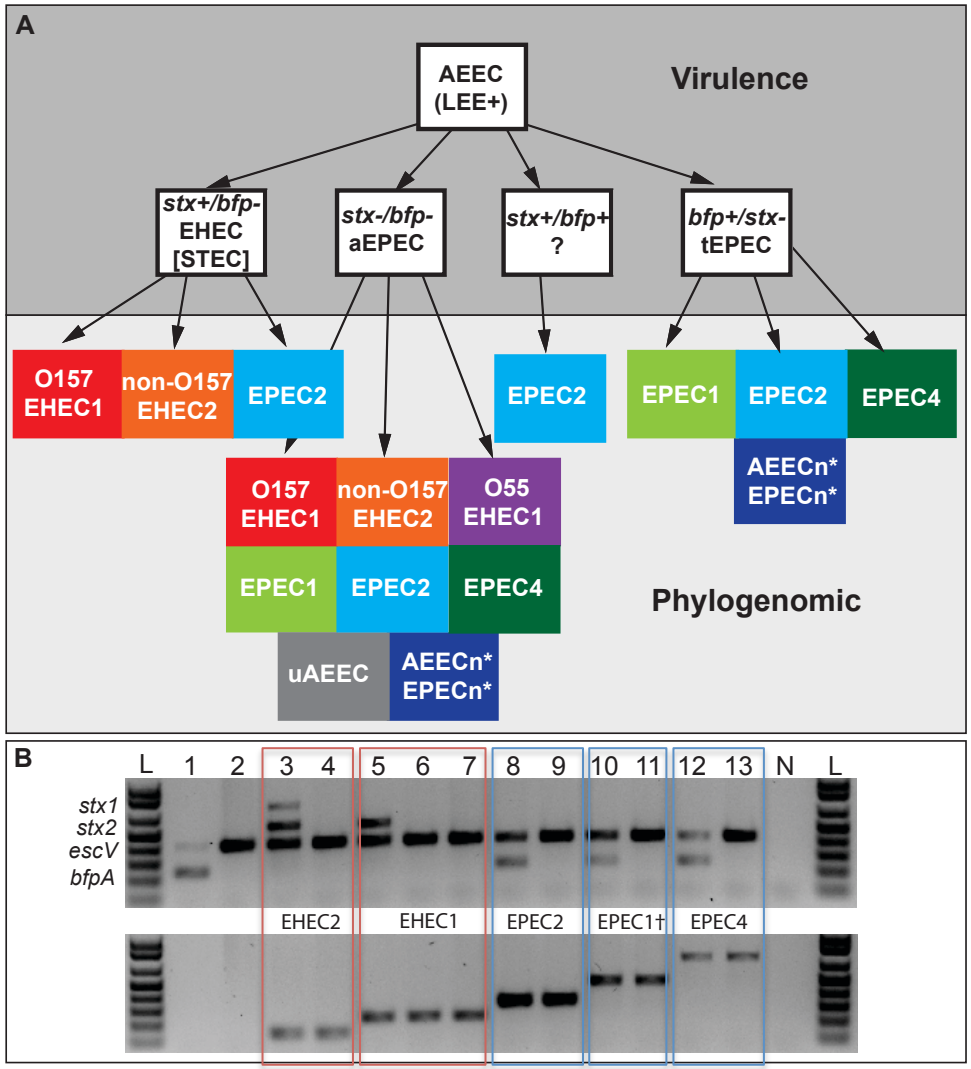
Lineage-specificity of T3SS Effectors



- *In silico* detection of O157:H7 (A), and EPEC1 and EPEC2 (B) T3SS effector proteins
- Effectors exhibited lineage-specific sequence similarity

Hazen *et al.* 2013. PNAS (PMID: 23858472)

Phylogenomics-Based Typing Assay



- By virulence gene content many isolates were atypical EPEC that were in EHEC or EPEC phylogenomic lineages
- Lineage-specific markers
- PCR multiplex for virulence genes and lineage-specific markers

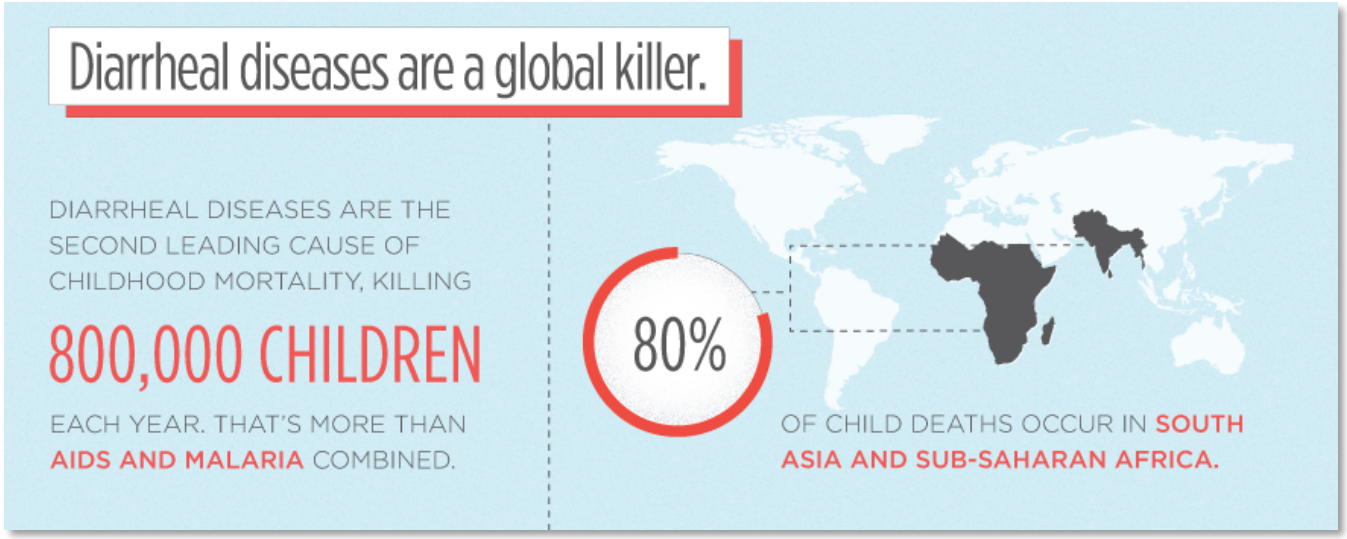
Phylogenomic-based typing along with the virulence gene detection provides insight into the other virulence factors that an isolate may carry that are lineage-specific, which is especially informative for atypical EPEC.

Hazen *et al.* 2013. PNAS (PMID: 23858472)

Outline

- Large Scale Comparative Genomics Tools:
  - Phylogenomics
  - Using BLAST Score Ratio (BSR) for *in silico* detection of gene-based differences
- Examples of *E. coli* Comparative Genomics:
  - Genomics of the AEECs: EPEC and EHEC
  - Comparing EPEC from clinical outcomes of differing severity

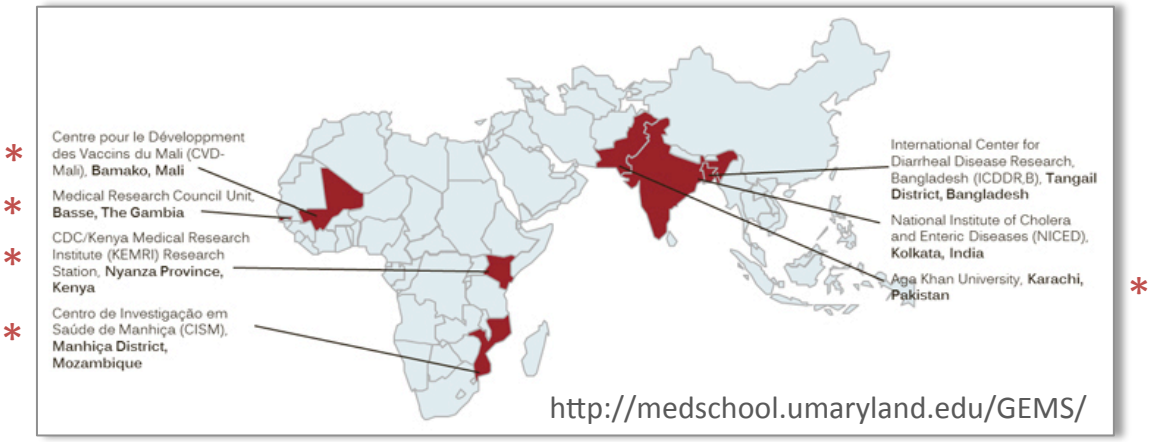
Global Enteric Multicenter Study Sites



- Three year matched case-control study of children 0-59 months
- Seven sites in Africa and Asia
- 9,439 children with moderate-to-severe diarrhea (MSD) and 13,129 without diarrhea
- Most cases of MSD: rotavirus, *Cryptosporidium,* enterotoxigenic *E. coli* (ETEC), and *Shigella*
- Associated with increased risk of death: typical EPEC (infants 0-11 months) and *Cryptosporidium* (toddlers 12-23 months)

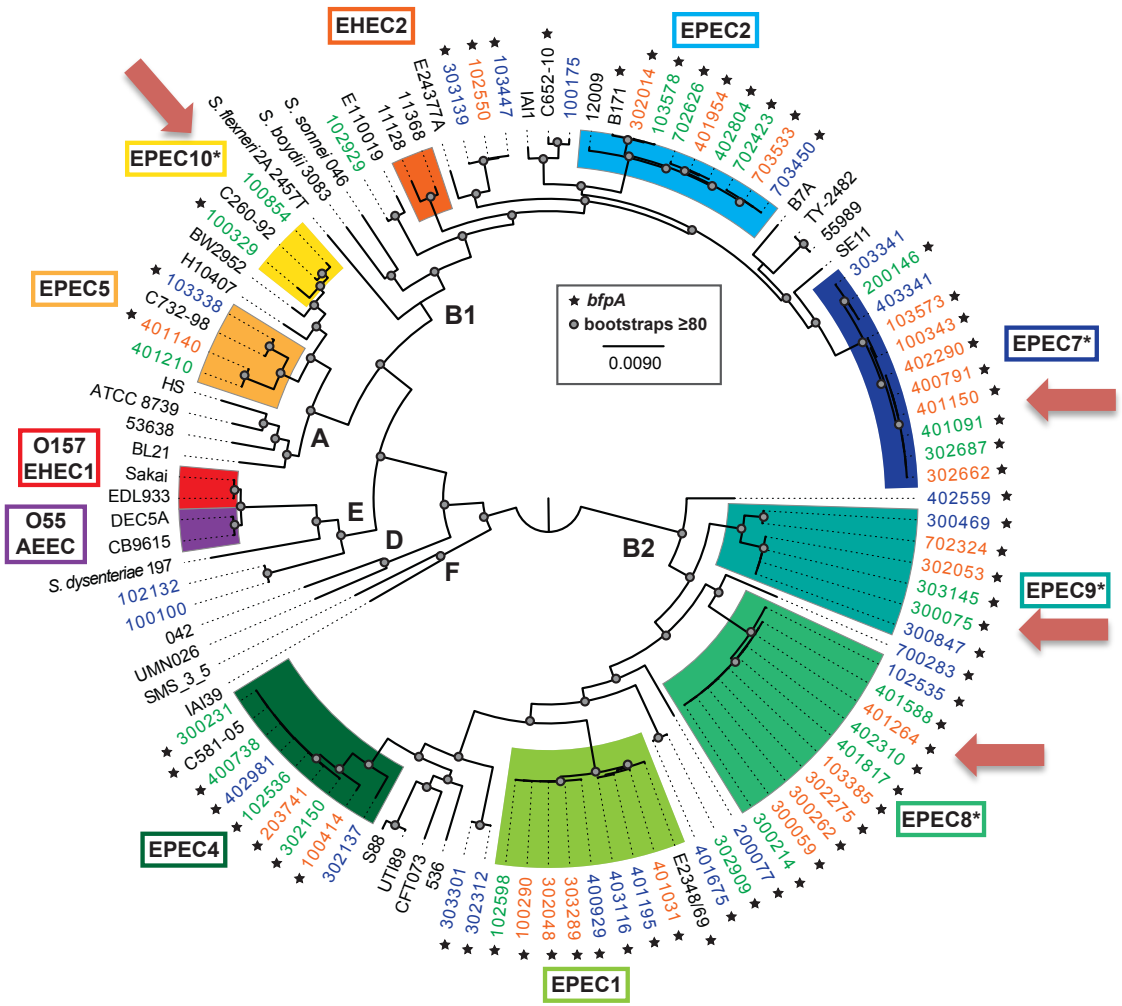
Kotloff *et al.* (2012) Clin. Infect. Dis.
Kotloff *et al.* (2013) Lancet

ERIN CRC Project on EPEC



- Enterics Research Investigational Network (ERIN) Cooperative Research Centers (CRC)
- Identify factors associated with virulence of typical EPEC
- Kenya, The Gambia, Mozambique, Pakistan, and Mali
- 70 total EPEC isolates:
  - Lethal infections (LI): 24 total (all typical EPEC)
  - Non-lethal symptomatic infections (NSI): 23 total (20 typical EPEC, 3 atypical EPEC)
  - Asymptomatic infections (AI): 23 total (17 typical EPEC, 6 atypical EPEC)

Phylogenomic Diversity of EPEC



- 820 Kb whole-genome alignment
- 38% (27/70) of the ERIN EPEC genomes in lineages EPEC1-EPEC5
  - 10 LI, 10 NSI, 7 AI
- 41% (29/70) comprised four new lineages (EPEC7, EPEC8, EPEC9, EPEC10)
  - 13 LI, 11 NSI, 5 AI
- 20% (14/70) are unclassified AEEC
  - 1 LI, 2 NSI, 11 AI

Overall, there was no correlation with clinical outcome and phylogenomic lineage, or by geography.

Isolate ID colors: orange (lethal; LI), green (non-lethal; NSI), blue (asymptomatic; AI)

Hazen *et al.* 2016. Nat. Microbiol.

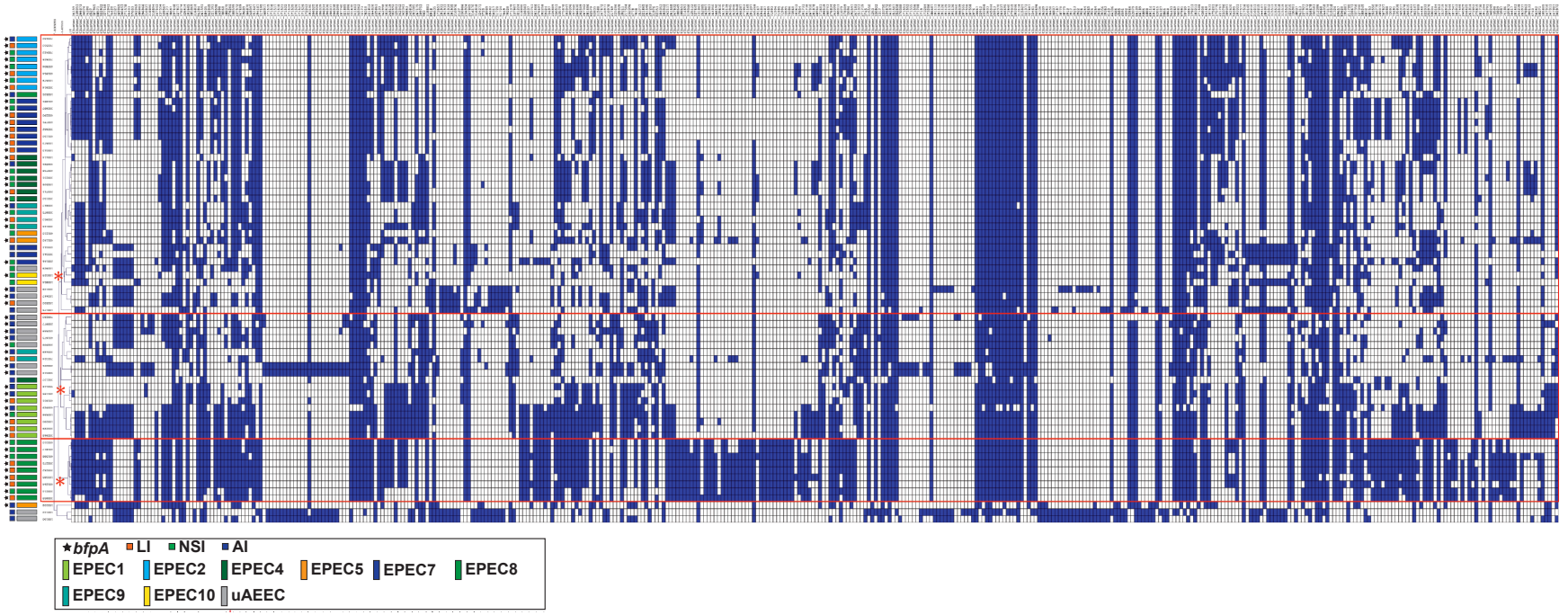
EPEC1 and EPEC2 Virulence Factors

Symptomatic Group	Phylogenomic Lineage	Specimen Id	Child Id	BFP	*perA*	T3SS Effectors: EPEC1 (E2348/69, n=21)	T3SS Effectors: EPEC2 (B171, n=26)	T3SS Effectors: O157:H7 EHEC (Sakai, n=49)	T2SS (E2348/69)	T6SSA (B171)	T6SSM (B171)
LI	EPEC1	100290	101003810	+	+	19 (90)	14 (54)	16 (33)	+	-	-
LI	EPEC1	302048	301043618	+	+	19 (90)	13 (50)	15 (31)	+	-	-
LI	EPEC1	303289	301071084	+	-	18 (86)	13 (50)	14 (29)	+	-	-
LI	EPEC1	401031	411001161	+	-	19 (90)	8 (31)	12 (24)	+	-	-
NSI	EPEC1	102598	102010781	+	+	19 (90)	14 (54)	16 (33)	+	-	-
AI	EPEC1	401195	408900255	+	-	17 (81)	11 (52)	13 (27)	+	-	-
AI	EPEC1	403116	405902255	+	+	20 (95)	13 (50)	15 (31)	+	-	-
AI	EPEC1	400929	408900628	+	+	20 (95)	13 (50)	15 (31)	+	-	-
								No. Genomes	8	0	0
								Prevalence (%)	100	0	0
LI	EPEC2	302014	301038524	+	+	9 (43)	23 (88)	17 (35)	-	+	+
LI	EPEC2	401954	405006609	+	+	8 (38)	22 (85)	15 (31)	-	+	+
LI	EPEC2	703533	702026545	+	-	1 (5)	9 (35)	4 (8)	-	+	+
NSI	EPEC2	103578	101042706	+	-	4 (19)	16 (62)	7 (14)	-	+	+
NSI	EPEC2	402804	411006247	+	+	9 (43)	24 (92)	18 (37)	-	+	+
NSI	EPEC2	702423	702012431	+	+	9 (43)	23 (88)	16 (33)	-	+	+
NSI	EPEC2	702626	707022601	+	+	6 (29)	22 (85)	15 (31)	-	+	+
AI	EPEC2	703450	702901552	+	+	8 (38)	23 (88)	15 (31)	-	+	+
								No. Genomes	0	8	8
								Prevalence (%)	0	100	100

- The EPEC1 (B2) and EPEC2 (B1) genomes have lineage-specific T3SS effector profiles
- EPEC1 genomes have T2SS but not T6SS, whereas EPEC2 genomes have two T6SSs but not T2SS

Hazen *et al.* 2016. Nat. Microbiol.

Genes with Greater Prevalence in Symptomatic vs. Asymptomatic Genomes



- 428 gene clusters were significantly ($p<0.05$) more associated with symptomatic (lethal and non-lethal) compared to asymptomatic isolates
- Hierarchical clustering analysis by genome separated the isolates into three main groups (I, II, and III):
  - Group I: All EPEC8, and from symptomatic outcomes (LI and NSI)
  - Group II: All phylogroup B2, 7 (39%) symptomatic vs. 11 (61%) asymptomatic
  - Group III: All three phylogroups, 31 (78%) symptomatic vs. 9 (22%) asymptomatic

Hazen *et al.* 2016. Nat. Microbiol.

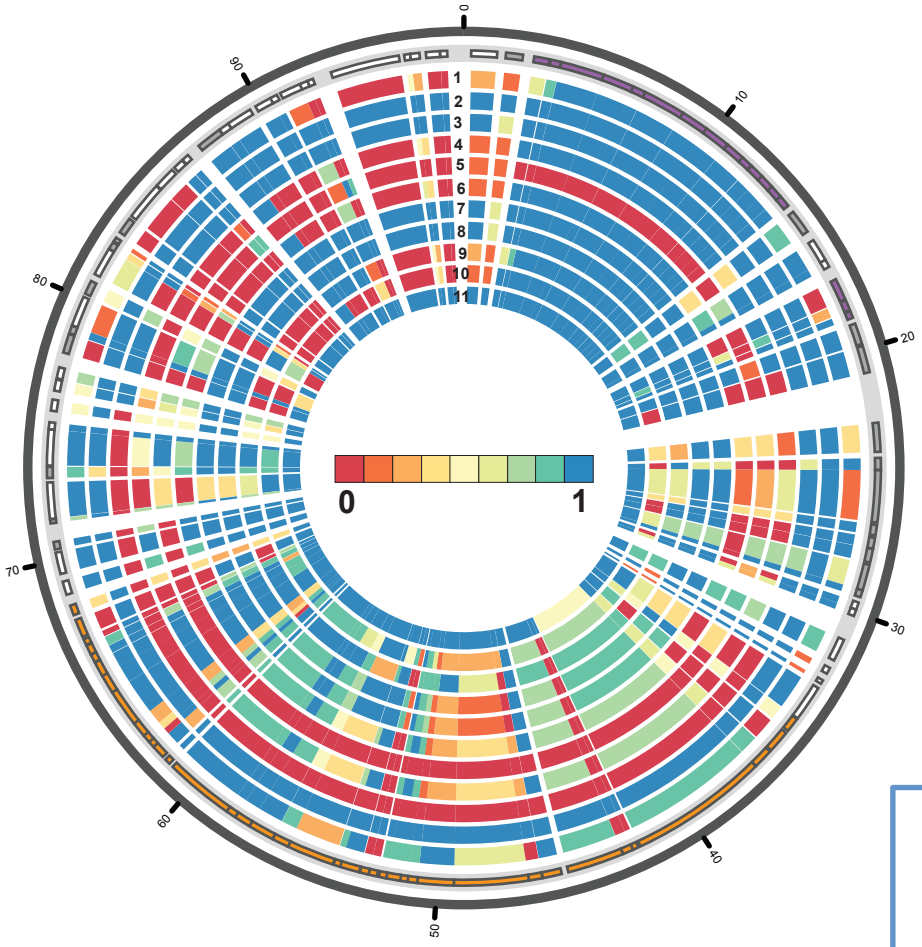
Gene Clusters By Clinical Outcome

- Gene clusters detected in association with a particular clinical outcome:
  - In ≥1 LI genome and no other genomes:
    - O-antigen biosynthesis genes, autotransporter, putative hemolysin, hypotheticals
  - In ≥1 AI genome and no other genomes:
    - Some T3SS effectors, hypotheticals

*Functional characterization required

Hazen *et al.* 2016. Nat. Microbiol.

Diversity of the EPEC Virulence Plasmid in EPEC From Different Phylogenomic Lineages

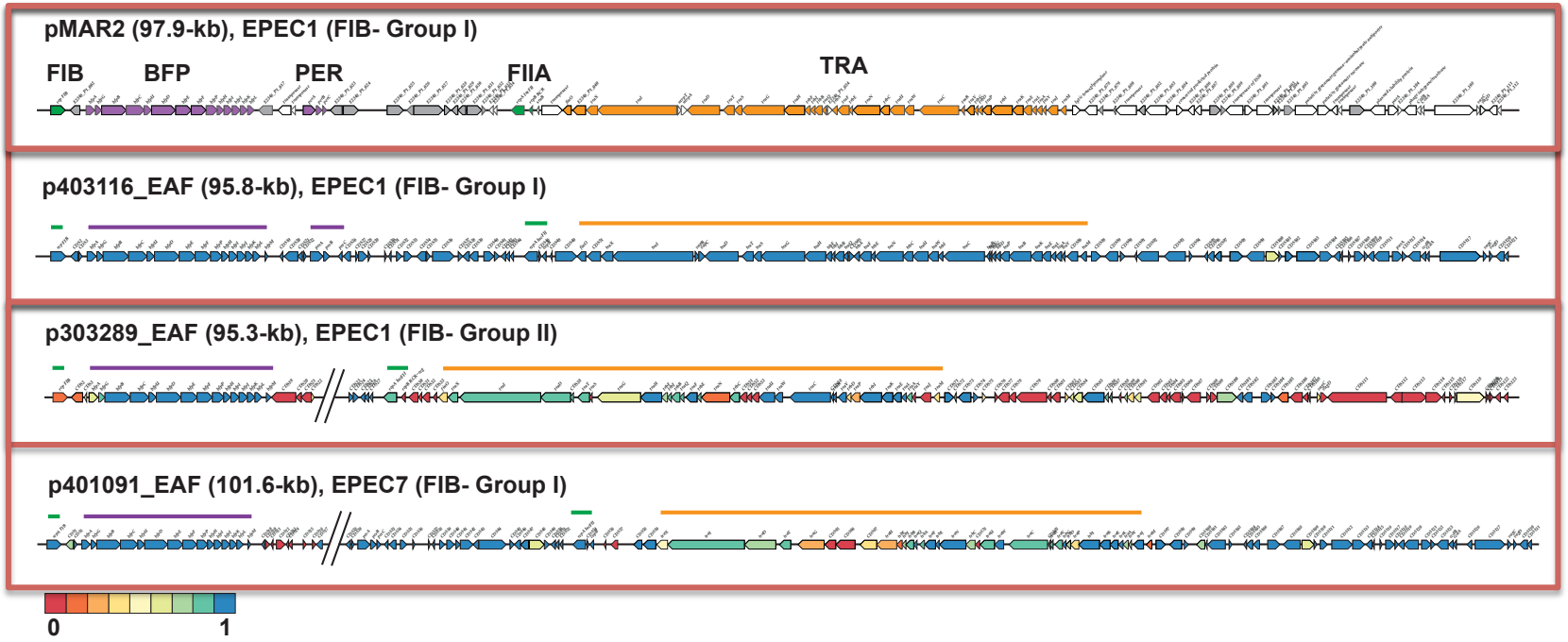


- *In silico* detection of pMAR2 genes in the genome sequences of representative EPEC isolates from diverse EPEC phylogenomic lineages
- Blue= conserved, Red= absent

Overall, the BFP genes exhibited significant similarity, but other regions of the plasmid had variable similarity for the different EPEC analyzed.

Hazen *et al.* 2015 Infect. Immun. (PMID: 26238712)

Diversity of the EPEC Virulence Plasmid in EPEC1 Genomes



- The virulence plasmid of 403116 (EPEC1) was nearly identical to pMAR2
- In contrast, the virulence plasmid of 303289 (EPEC1) was conserved in the BFP region, but was divergent over the rest of the plasmid

The virulence plasmids did not exhibit the same lineage-specificity observed for other virulence factors, suggesting the plasmids have been lost and acquired multiple times.

Hazen *et al.* 2015 Infect. Immun. (PMID: 26238712)

EPEC Genomics Summary

- Typical EPEC (LEE+/BFP+/STX-) occupied numerous phylogenomic lineages of *E. coli* and three of the *E. coli* phylogroups
- Atypical EPEC (LEE+/BFP-/STX-) were identified in lineages that contained EHEC, EPEC, or related to other pathogenic *E. coli*
- Many of the EPEC virulence factors exhibited lineage-specific sequence similarity suggesting they have been maintained in these lineages
- There were no genes detected in all of the symptomatic EPEC genomes that were not in any asymptomatic genomes; however, there were genes that were significantly more associated with one symptomatic group over another
- The EPEC virulence plasmid had greater genetic diversity than was previously described, including within lineage diversity, suggesting it has not been stably maintained and has likely been lost and acquired multiple times

Conclusions

- Whole-genome sequencing and comparative genomics are powerful tools to investigate the genomic diversity of disease-associated bacteria
- These methods can be used to identify targets for improved diagnostics, further functional characterization of virulence mechanisms, or possible vaccine development

Acknowledgements

- IGS
  - David Rasko
    - Jason Sahl (TGEN/ NAU)
    - Julia Redman
  - IFX & GRC
- University of Maryland
  - James Kaper
  - Eileen Barry
  - Michael Donnenberg
- University of Virginia
  - James Nataro

ERIN CRC: NIH U19AI090873

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE
INSTITUTE FOR GENOME SCIENCES

Posters: Friday-010, Saturday-005