Microbial genome sequencing on the Applied Biosystems SOLiD® and Ion Torrent PGM™ NGS platforms

Craig Cummings, Ph.D.
Sr. Staff Scientist, Bioinformatics
Life Technologies
Foster City, CA
craig.cummings@lifetech.com

Outline

- Life Technologies sequencing portfolio
- Genome sequencing: resequencing or *de novo* assembly?
- *De novo* sequencing with the SOLiD® System
  - *Listeria* spp.
  - *Salmonella enterica* subsp*. enterica*
- Resequencing with the SOLiD® System
  - 2009-2010 U.S. *Salmonella* Montevideo outbreak
- Rapid whole-genome sequencing on the Ion Torrent PGM™ Sequencer
  - Recent German *E. coli* O104:H4 outbreak

Life Technologies Sequencing Portfolio

3730 DNA Analyzer

3500 Genetic Analyzer

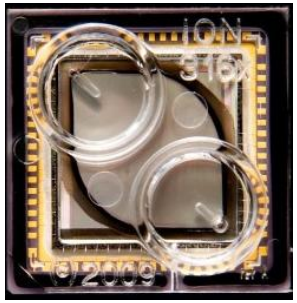
Capillary electrophoresis

SOLiD® 4 System

5500 Genetic Analyzer

Next-generation sequencing

Ion Torrent PGM™ Sequencer



Ion 316™ Chip

Semiconductor sequencing

Comparison of Life Technologies sequencing platforms

	3500xl Genetic Analyzer	Ion Torrent (316™ chip)	5500 Genetic Analyzer
Max throughput per run	< 24 kb	> 100 Mb	> 10 Gb (per lane)
Bacterial genome coverage per run	< 0.01X	> 20X	> 2,000X (or 20X on 100 genomes)
Prep time	~ 4 hrs (PCR and cycle seq)	1 day (library and ePCR)	> 1 day (library and ePCR)
Run time	Up to 2 hours	2 hours	7 days
Read length	Up to 1,000 bp	> 100 bp	60 bp
Key features	Gold standard accuracy	Rapid genome sequencing	Vast throughput, high accuracy

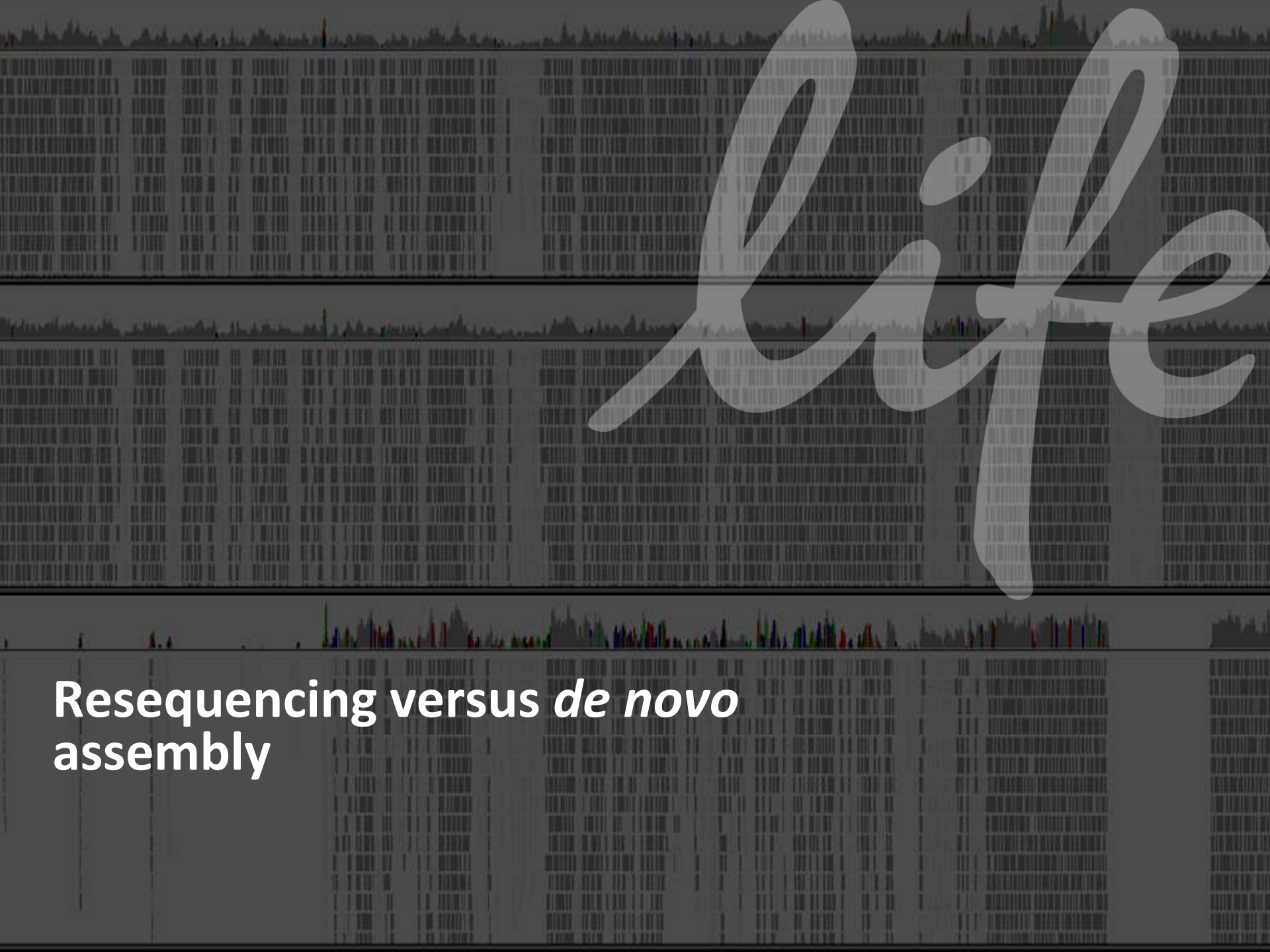
Life Technologies provides a suite of tools for rapid identification and screening of foodborne pathogens

Next-gen sequencers for rapid sequencing of novel pathogen strains

Capillary Electrophoresis for strain genotyping by multi-locus sequence typing (MLST) & sequencing of virulence genes

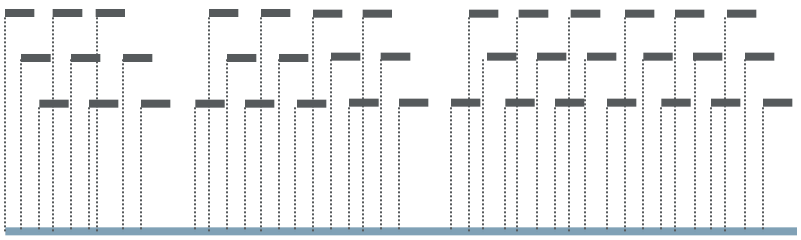
Real Time PCR Systems for rapid screening of large numbers of samples

TaqMan® Assays for sensitive detection of specific pathogens



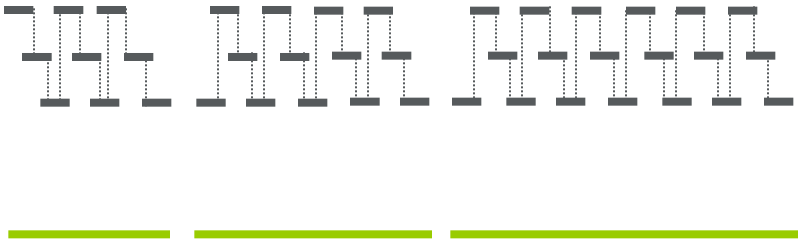
Resequencing versus *de novo* assembly

Referenced versus *de novo* assembly



Referenced assembly

Reads are mapped to a previously sequenced reference genome and differences are found



de novo assembly

Reconstruction of a genome sequence from reads alone, without any reference sequence

Referenced assembly only works for very closely related strains

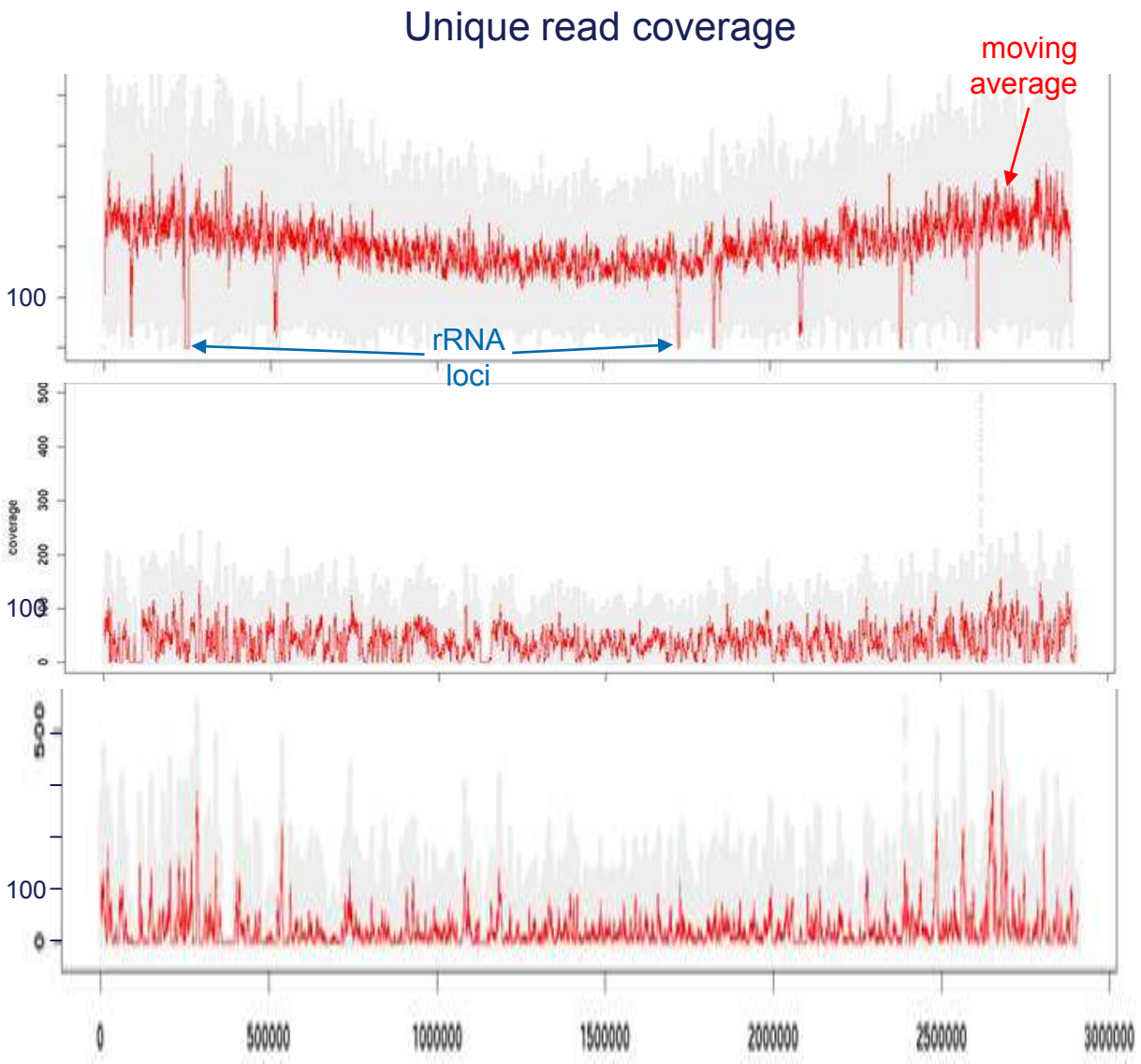
Same species (95% iden)

L. mono F2365 vs. F2365 (median = 200X)

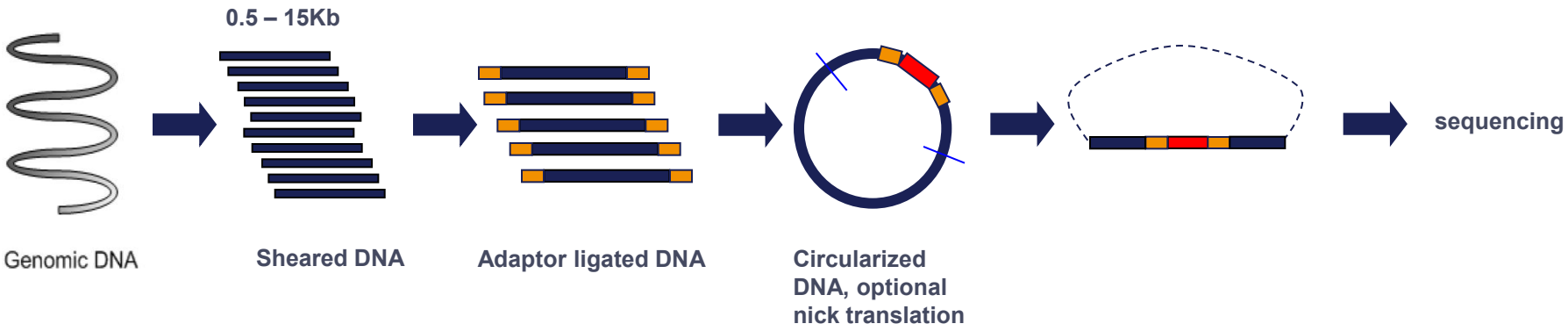
L. mono F2-208 vs. F2365 (median = 36X)

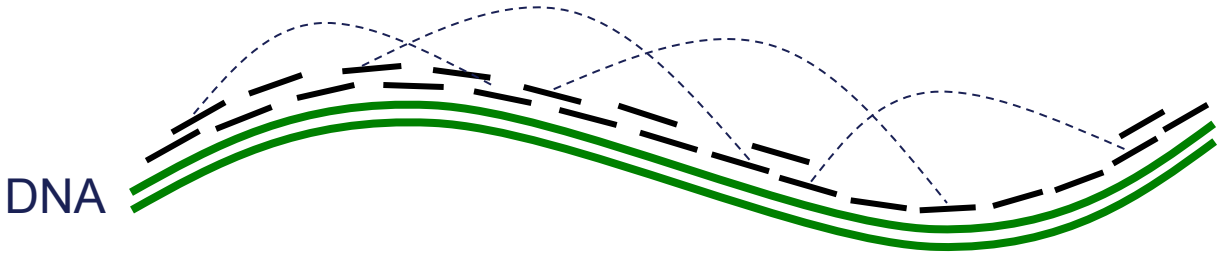
Sister taxa (88% iden)

L. marthii S4-120 vs. F2365 (median = 7X)

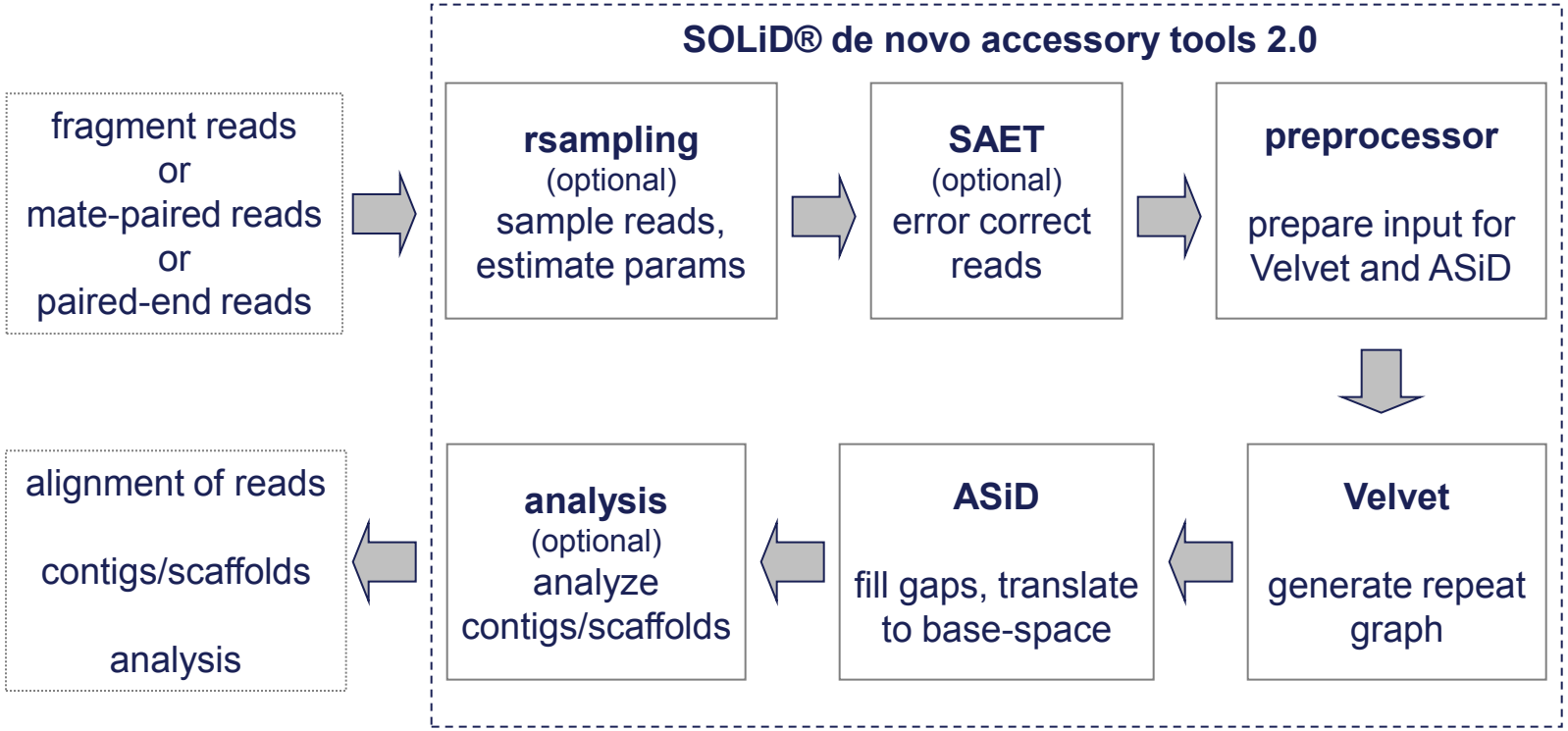


SOLiD® mate-paired library construction



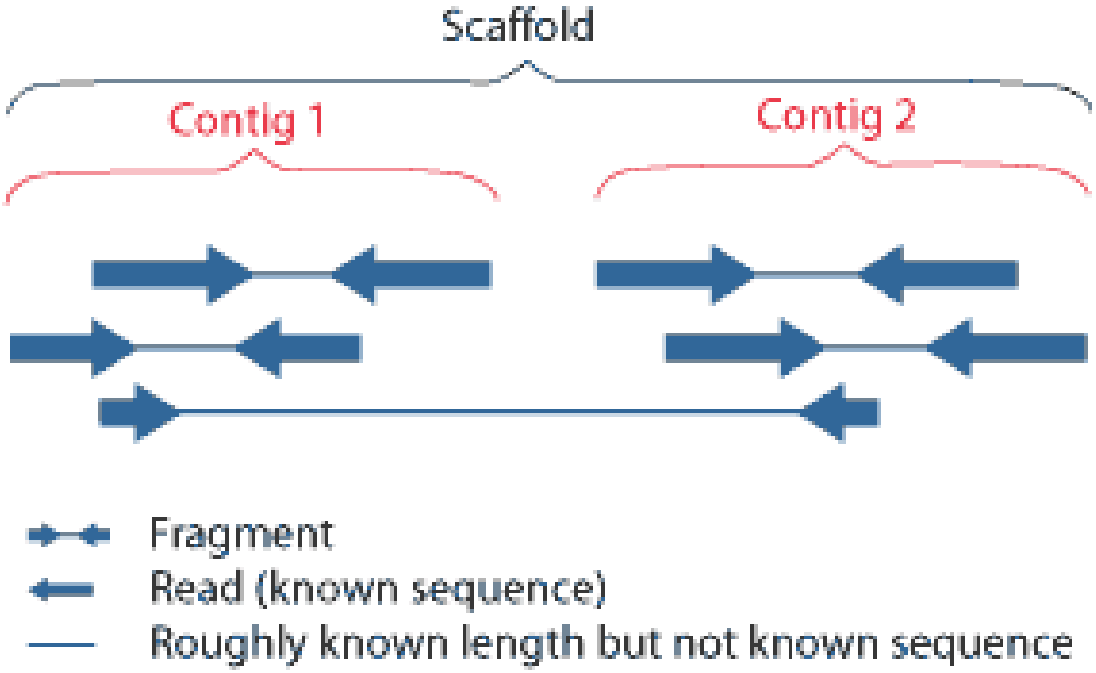


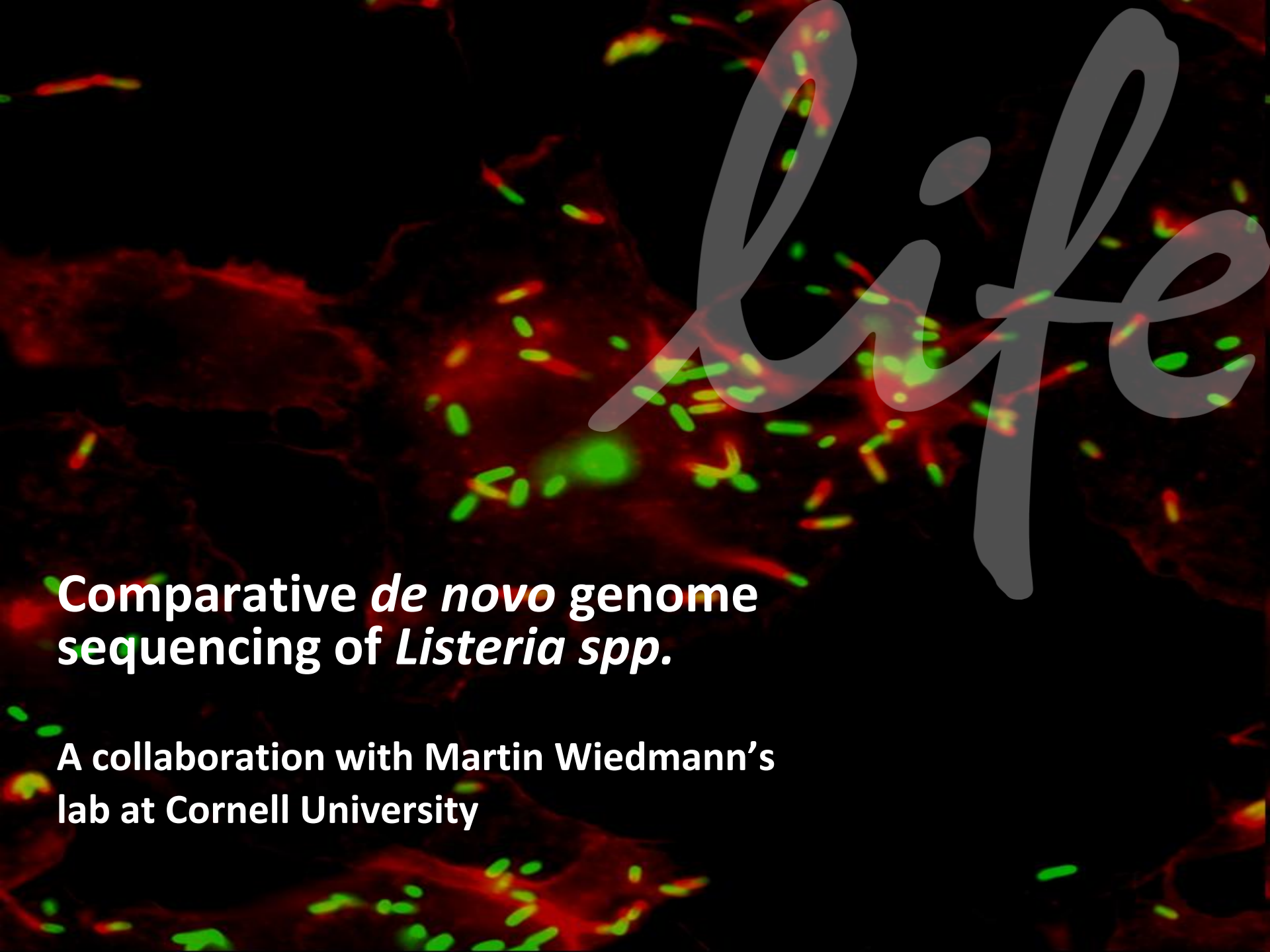
De novo assembly pipeline 2.0



- Takes as input raw data generated by SOLiD platform, reads in csfasta format and quality file if available
- Outputs assembled contigs in base-space and color space, as well as, contigs/scaffolds analysis, and their comparison with reference (if provided)
- Addition of pre-assembly error correction by SAET reduces error rate from 3-5% to sub 1%, that enables de novo assembly and increases N50 contig by factor of 3.
- Addition of post-assembly gaps filling between contigs in scaffolds by ASiD increases N50 contig length by another factor of 3 (only for paired-end and mate-paired).

Contigs and scaffolds





Comparative *de novo* genome sequencing of *Listeria spp.*

A collaboration with Martin Wiedmann's lab at Cornell University

Listeria background

- The genus Listeria is composed of eight species
  - *L. monocytogenes, L. innocua, L. marthii, L. welshimeri, L. seeligeri, L. ivanovii, L. rocourtiae, and L. grayi*
- Two species, *L. monocytogenes* and *L. ivanovii* are known pathogens and cause listeriosis in human and warm-blooded hosts
- These pathogenic species are closely related to non-pathogenic species
  - *L. monocytogenes* is closely related to *L. innocua* and *L. marthii*
  - *L. ivanovii* is closely related to *L. seeligeri*
- *L. monocytogenes* is a foodborne pathogen that causes nearly 2,500 illnesses per year in the US, primarily affecting pregnant women, newborns, and adults with weakened immune systems

Motivation

- Only a three species had previously been sequenced
  - *L. monocytogenes* (4 strains)
  - *L. innocua*
  - *L. welshimeri*
- Determine core and accessory genomes of this genus
- Identify genes that are correlated with virulence
- Define molecular signatures for the design of detection assays for sensitive and specific food safety testing

Listeria strains sequenced

Strain designation	Source, geographic origin, lineage	Genome size (Mb)	Hemolytic activity	Pathogen	Genbank acc.
L. monocytogenes					
F2365	food, listeriosis outbreak, CA, USA, 1985, lineage I	2.91	+	+	AE017262
▶ FSL R2-574	same strain as F2365, lineage I	2.87	+	+	
EGD-e	lab strain derived from isolate of rabbit, England, 1924, lineage II	2.94	+	+	AL591824
▶ FSL F2-208	blood, human listeriosis case, USA, 1999, lineage III	3.20	+	+	
HCC23	naturally avirulent serotype 4a strain from catfish, USA, lineage III	2.98	+	-	CP001175
CLIP80459	human epidemic, France, 1999, lineage I	2.91	+	+	FM242711
L. marthii					
▶ FSL S4-120	soil, forest, NY, USA, 2001	2.87	-	-	
L. innocua					
CLIP11262	food, Morocco	3.01	-	-	AL592102
▶ FSL S4-378	puddle of water, NY, USA, 2002	3.09	-	-	
▶ FSL J1-023	obtained from Qualicon, exact origin unknown	2.91	+	-	
L welshimeri					
SLCC5334	decaying vegetation, USA	2.81	-	-	AM263198
L. ivanovii					
▶ FSL F6-596	food, France	3.10	+	+	
L. seeligeri					
▶ FSL N1-067	food processing plant, NY, USA	3.09	+	-	
▶ FSL S4-171	urban environment, NY, USA, 2001	2.89	-	-	

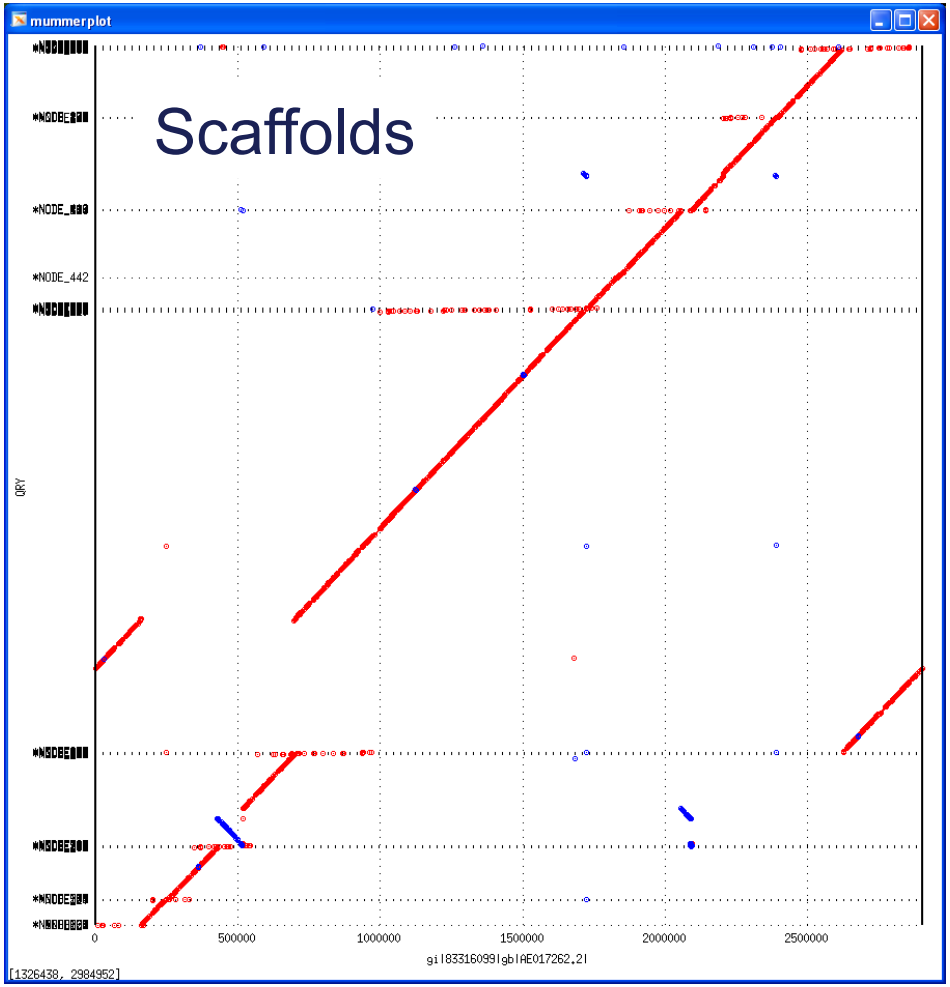
▶ **Sequenced by SOLiD® in this study**

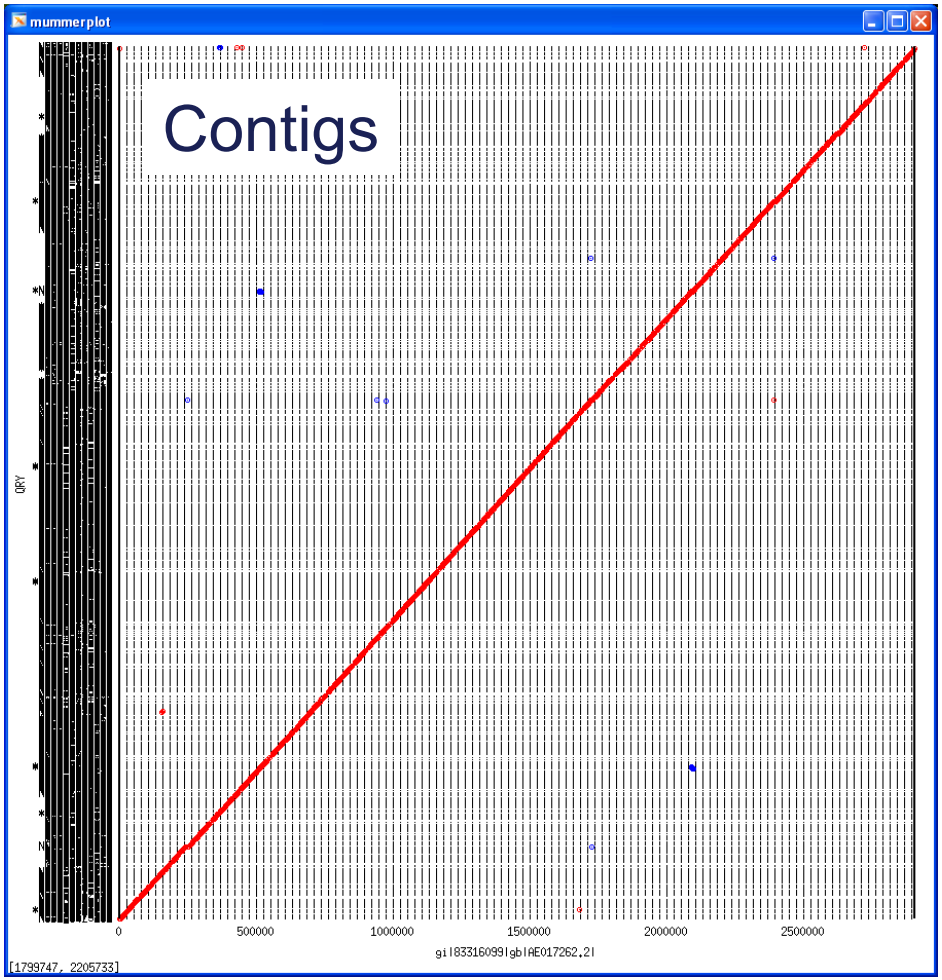
Assembly of *Listeria* contigs from 25 bp mate-paired reads

Strain	Scaffolds		Contigs		Est. #
	num	N50	num	N50	ORFs
F2-208	1,437	49,992	2,531	2,639	2,910
S4-120	404	257,992	925	7,850	2,724
S4-378	896	102,515	1,837	4,230	2,885
J1-023	324	247,625	790	9,133	2,737
N1-067	343	282,765	785	10,831	3,017
S4-171	216	226,677	868	5,655	2,820
F6-596	601	95,455	1,463	5,168	2,919

* Only contigs >= 100 bp

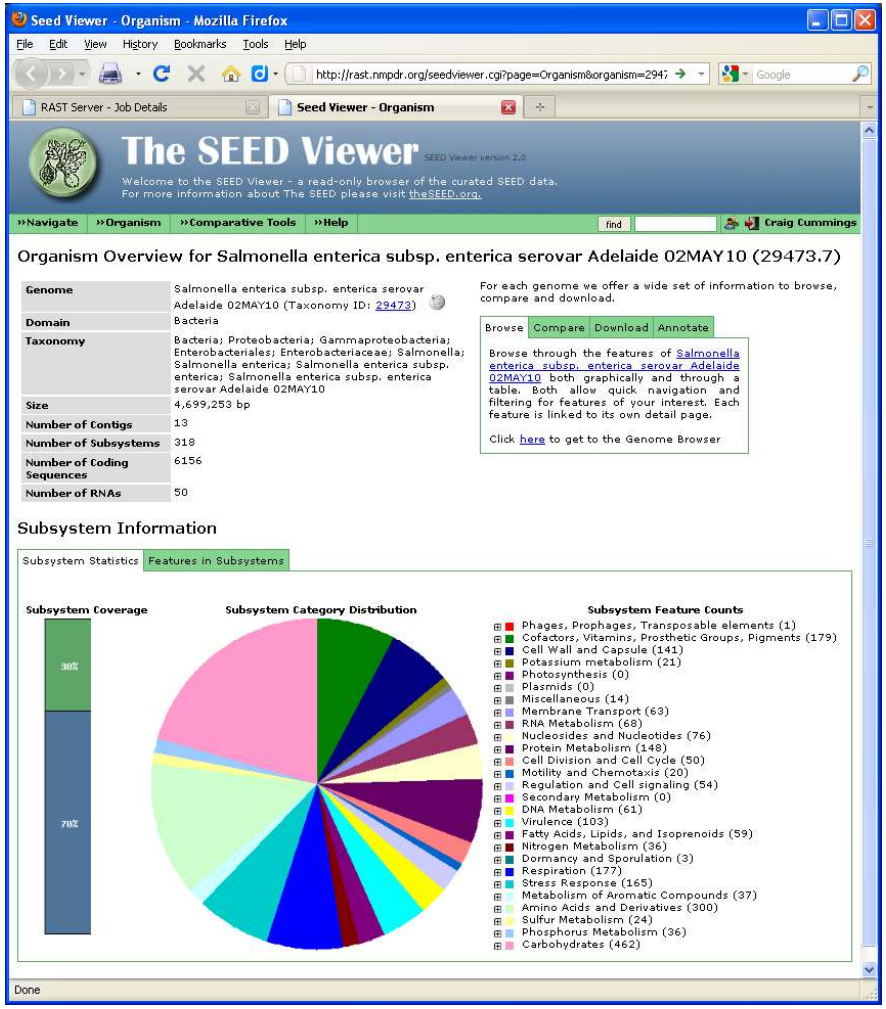
Alignment of F2365 assembly to *L. monocytogenes* reference





Weighted average identity = 99.64%
Percent of genome covered = 98.09%

Automated annotation using the RAST server

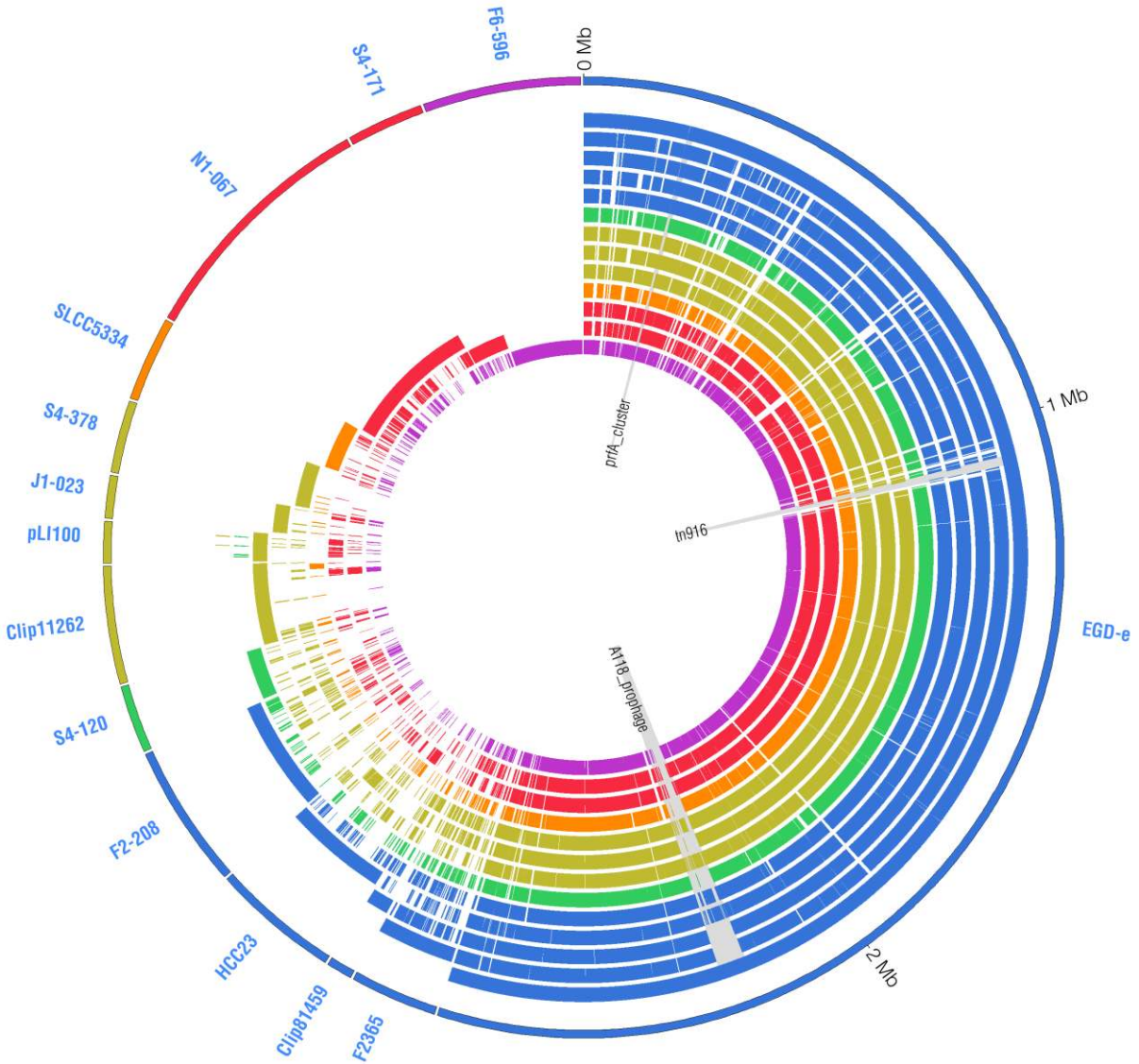


http://rast.nmpdr.org

- Call tRNA and rRNA genes
- Gene prediction:Glimmer2
- Establish phylogenetic context
- Search for homologous genes present in closest relative, copy their annotation, and adjust start of genes to match closest relative relatives
- Use matching genes to train Glimmer2 and recall protein encoding genes
- BLAST remaining genes against non-redundant protein DB (FIGfam) and clean up remaining gene calls

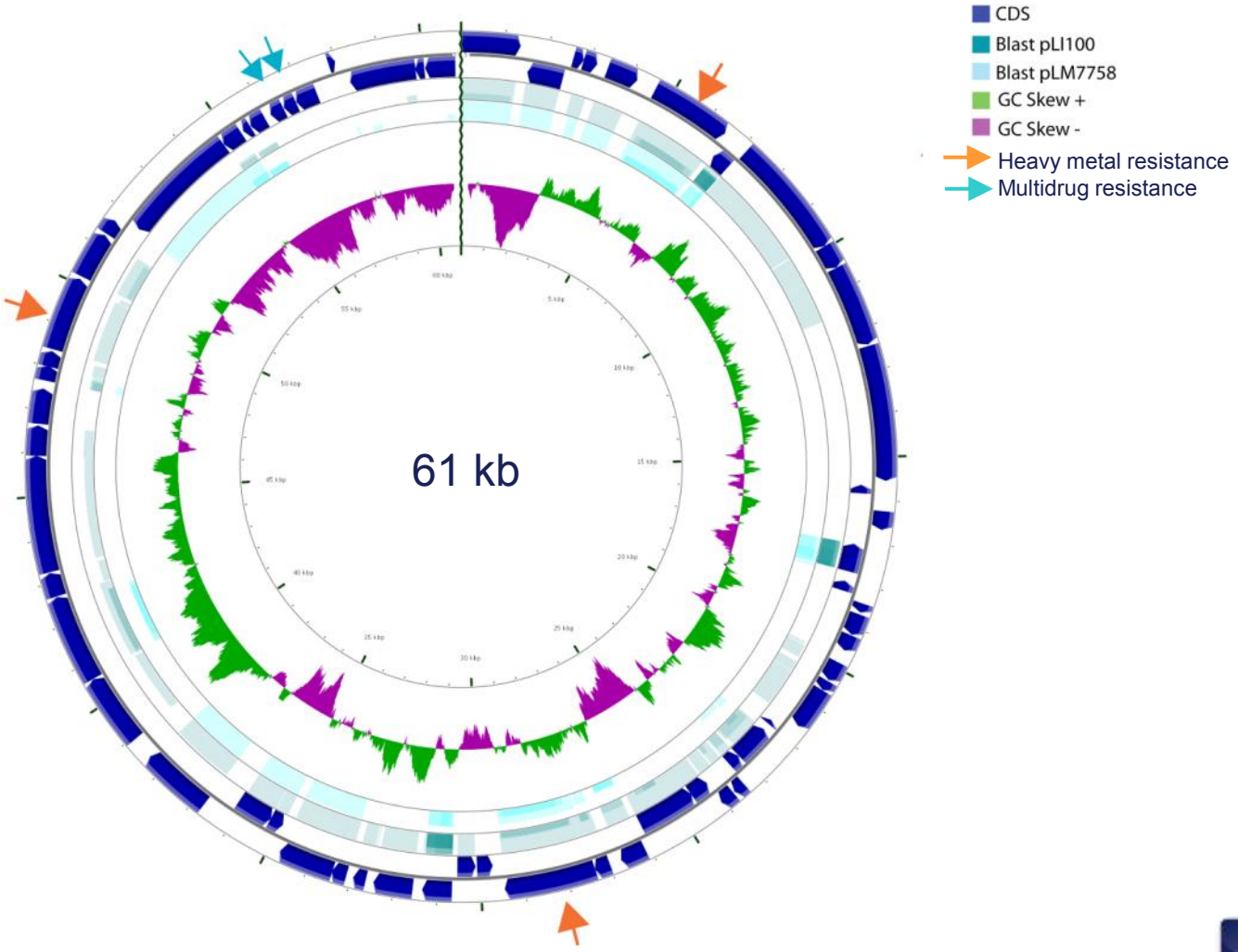
Genome content of *Listeria* species

- Comparative gene content of 13 *Listeria* strains representing six species
- Includes assemblies of seven previously uncharacterized strains

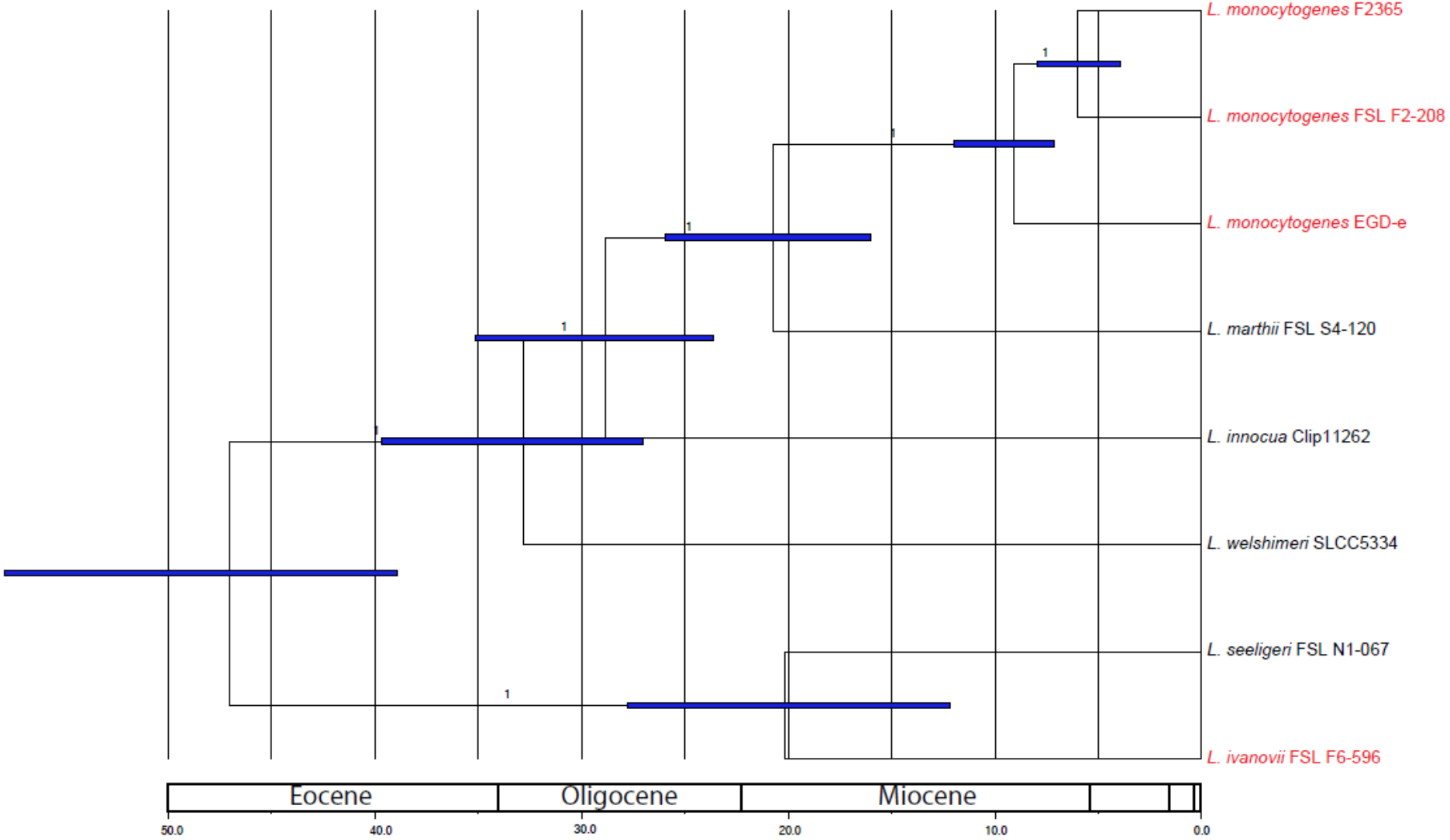


L. mono EGD
L. mono F2365
L. mono Clip81459
L. mono HCC23
L. mono F2-208
L. marthii S4-120
L. innocua Clip
L. innocua J1-023
L. innocua S4-378
L. welshimeri SLCC
L. seeligeri N1-067
L. seeligeri S4-171
L. ivanovii F6-596

Discovery of a novel plasmid in *L. seeligeri*



Evolution of pathogenicity in *Listeria*



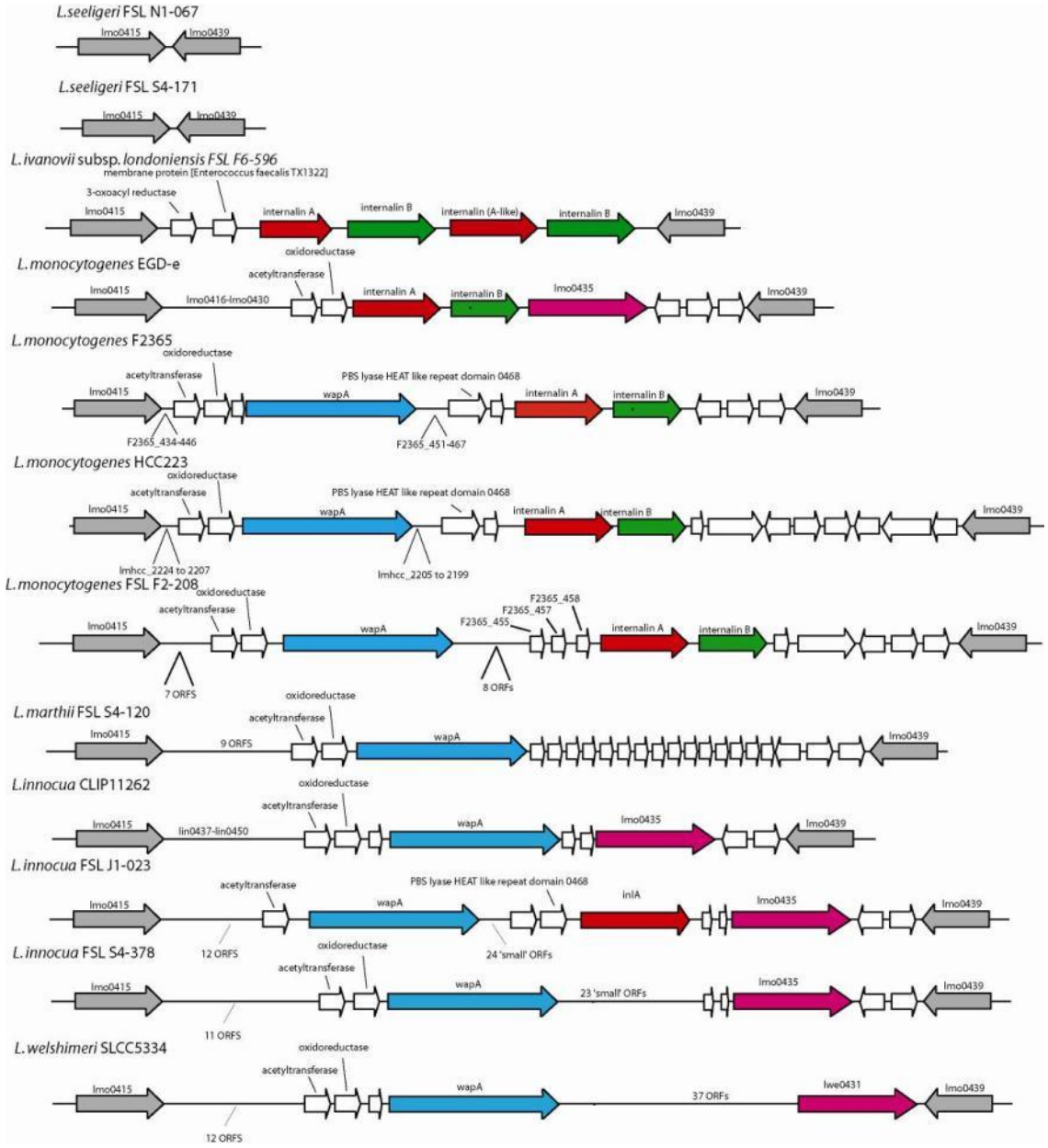
Distribution of putative virulence genes in pathogenic and non-pathogenic *Listeria*

	Pathogenic					Non-pathogenic							
Gene	**Lm**	**Lm**	**Lm**	**Lm**	**Liv**	**Lin**	**Lm**	**Ls**	**Ls**	**Lma**	**Lin**	**Lin**	**Lw**
lmo0206	+	+	+	+	-	+	+	-	-	-	-	-	-
lmo0257	+	+	+	+	+	+	+	+	+	+	-	-	+
inlH/inlC2	+	+	+	+	-	-	-	-	-	-	-	-	-
vip	+	+	+	-	-	-	-	-	-	-	-	-	-
inlA	+	+	+	+	+	+	+	-	-	-	-	-	-
inlB	+	+	+	+	+	-	+	-	-	-	-	-	-
uhpT	+	+	+	+	+	+	+	+	+	-	-	-	-
lmo0915	+	+	+	+	+	+	+	-	+	+	+	+	+
aut	+	+	+	-	+	+	+	+	+	+	+	+	+
lmo1081	+	-	-	-	-	-	-	+	-	-	-	-	-
lmo1082	+	-	-	-	-	-	-	+	-	-	-	-	-
lmo1290	+	+	+	+	+	-	+	-	+	+	-	-	-
inlC	+	+	+	+	+	-	-	-	-	-	-	-	-
bsh	+	+	+	+	-	-	+	-	-	-	-	-	-
lmo2157	+	+	+	-	+	-	-	-	-	+	-	-	-
lmo2439	+	+	+	+	+	+	+	+	-	+	+	+	+
ami	+	+	+	+	-	-	+	+	+	+	+	+	+
lmo2713	+	+	+	+	-	+	+	+	+	+	+	+	+
inlJ	+	+	+	+	-	-	-	-	-	-	-	-	-

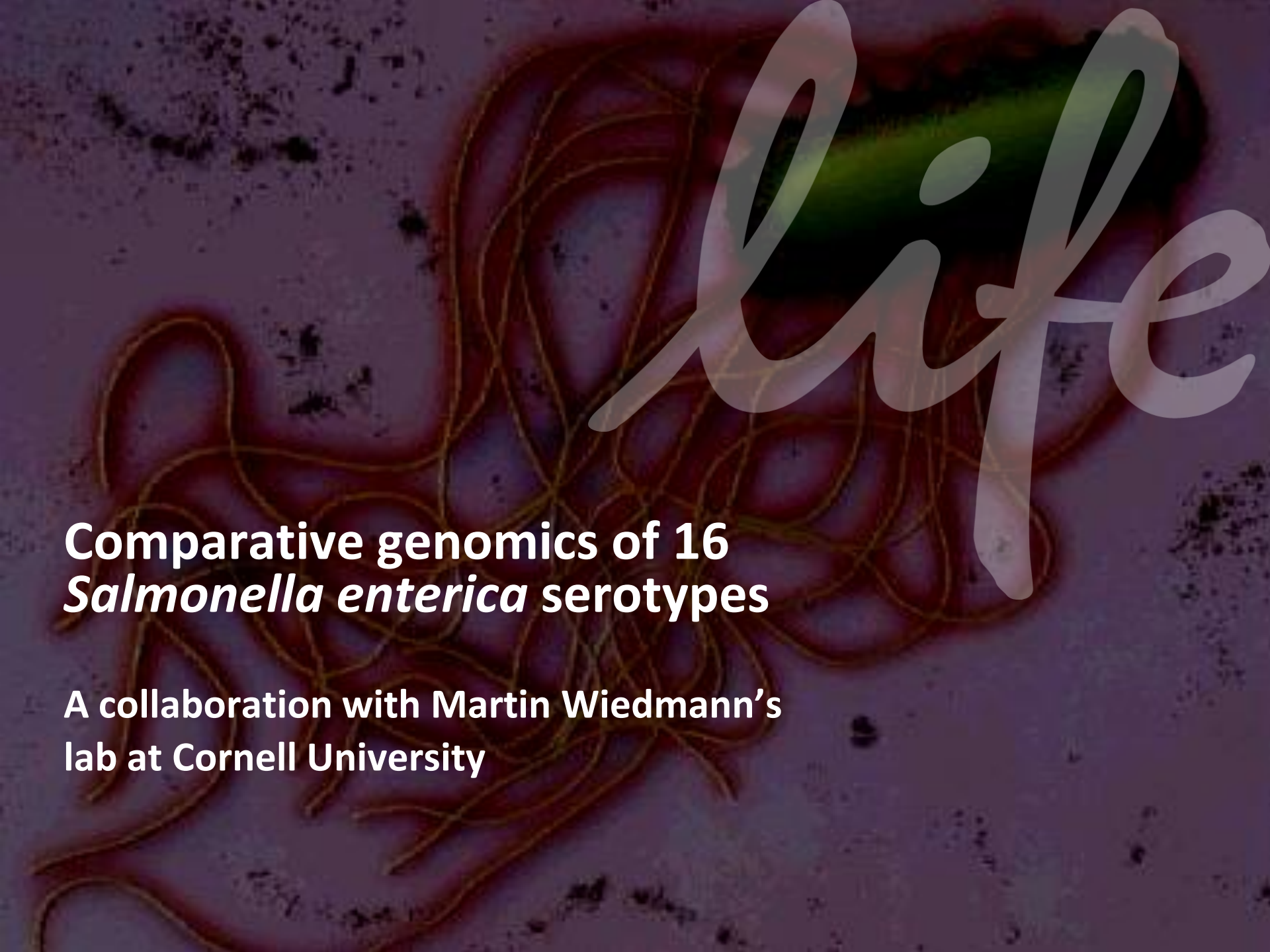
Present in all pathogenic strains

Specific for pathogenic strains

Hypervariable pathogenicity islands from assembled *Listeria* genomes: *inlAB* locus



- Internalin A
- Internalin B
- Wall-associated protein
- Peptidoglycan-bound protein



Comparative genomics of 16 *Salmonella enterica* serotypes

A collaboration with Martin Wiedmann's lab at Cornell University

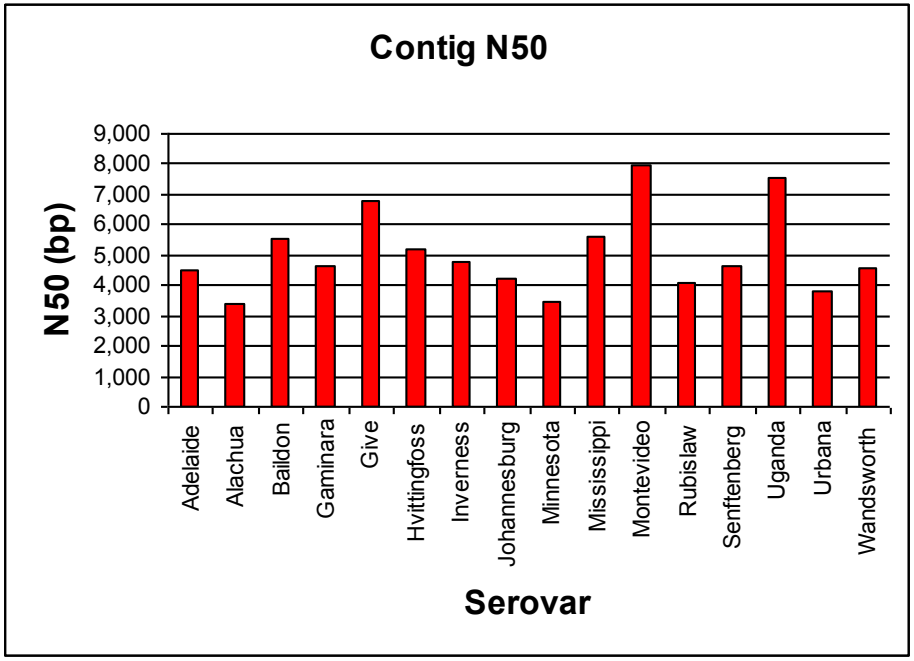
Why sequence more *S. enterica* genomes

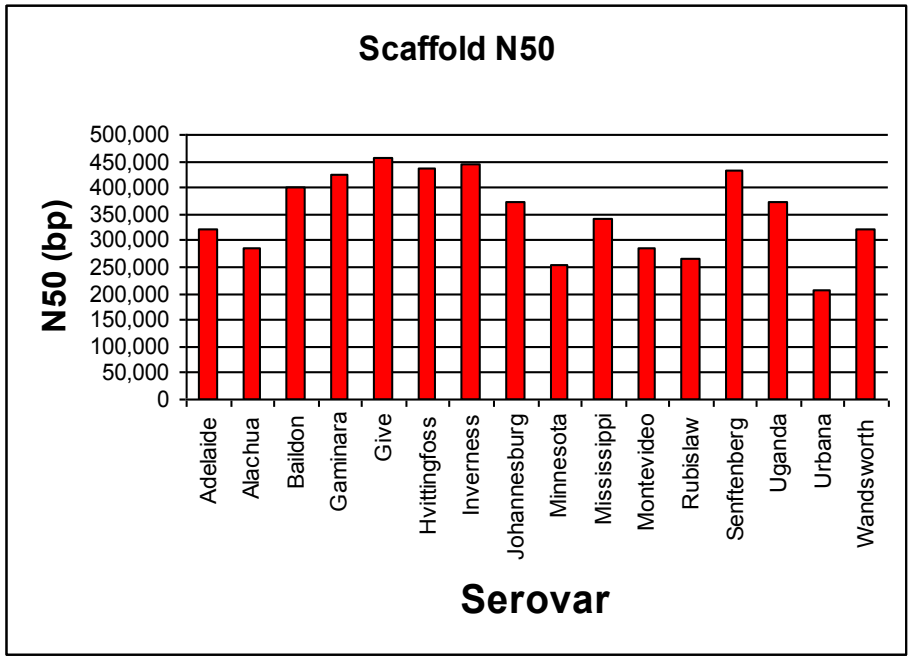
- Only 21 of the 2,500 *S. enterica* serotypes (those causing the majority of human disease) had been previously sequenced
  - These represent only four (B, C, D, and E) of the more than 40 known O-serogroups
- The repertoire of pathogenecity islands and plasmids, which have been implicated in the acquisition of virulence and drug-resistance, has not been extensively explored across this species
- Goals
  - Explore genomic diversity across a broad sample of strains representing a wide variety of serotypes
  - Obtain sequences of serotype determining regions to guide development of molecular serotyping assays
  - Develop sensitive and accurate tests for detection of *Salmonella* in food safety applications

16 S. enterica serotypes selected for one SOLiD® system sequencing run

Serotypes Sequenced (16)	Serogroup	
Adelaide	O (O:35)	
Alachua	O (O:35)	
Baildon	D2 (O:9,46)	
Gaminara	I (O:16)	◄ Closely related
Give	E1 (O:3,10)	◄ Closely related
Hvittingfoss	I (O:16)	◄
Inverness	P (O:38)	
Johannesburg	R (O:40)	
Minnesota	L (O:21)	
Mississippi	G (O:13)	
Montevideo	O:54	◄ Atypical plasmid-borne O-antigen locus
Rubislaw	F (O:11)	
Senftenberg	E4 (O1,3,19)	◄
Uganda	E1 (O:3,10)	◄
Urbana	N (O:30)	
Wandsworth	Q (O:39)	

Salmonella assembly statistics: 2 x 25bp



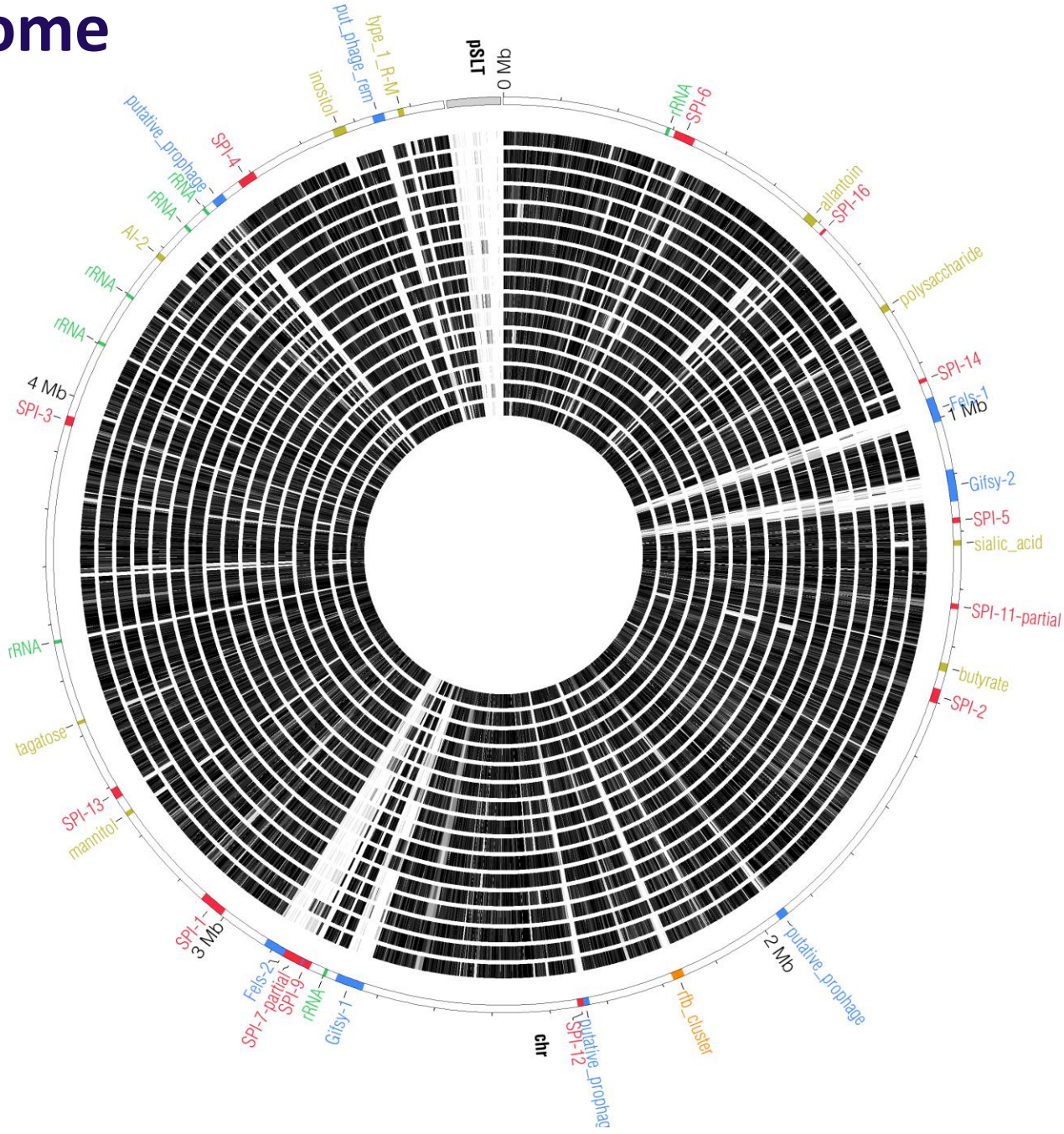


Min	Max	Median
3,408	7,978	4,640

Min	Max	Median
206,877	455,389	356,858

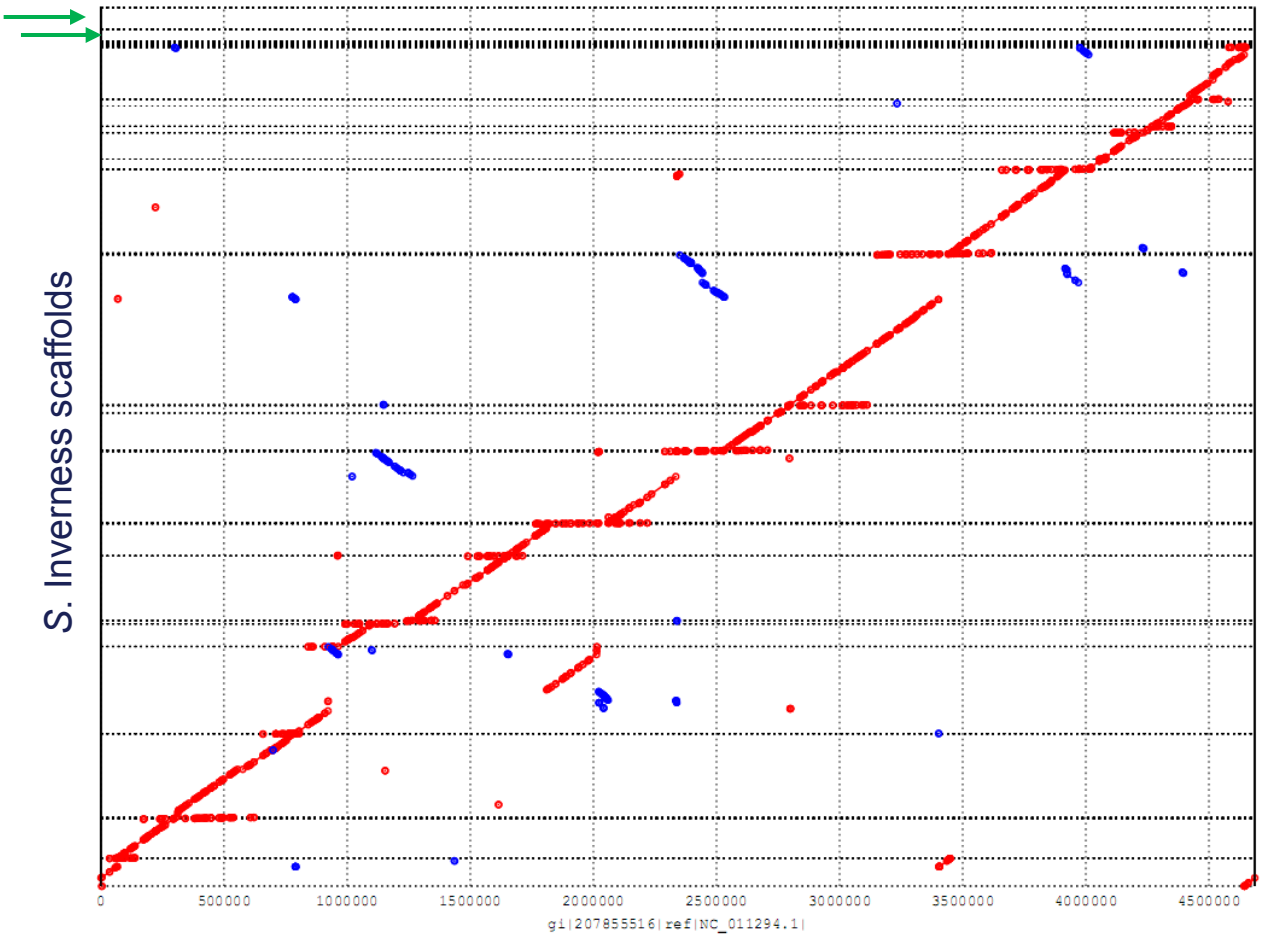
Diversity in genome content

- Gene presence and absence in 16 *S. enterica* strains relative to *S.* Typhimurium LT2



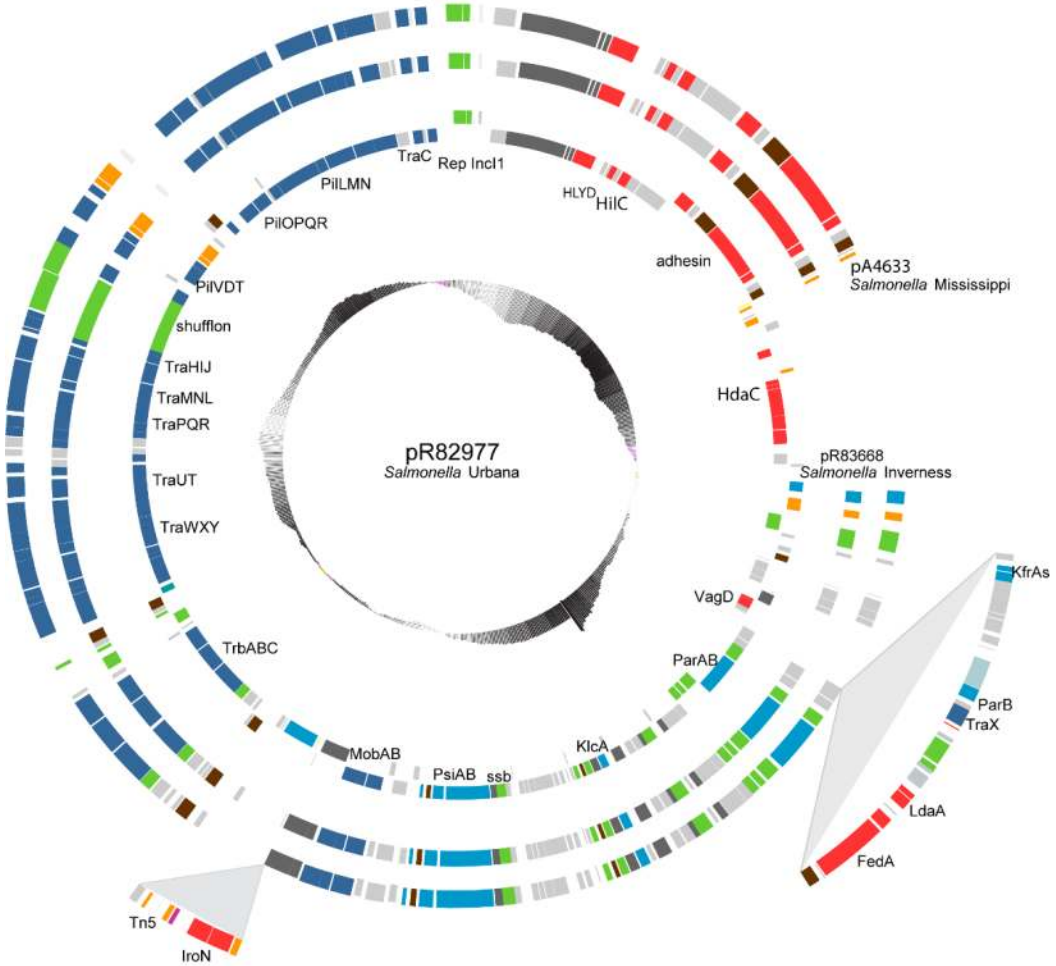
Plasmids discovered in *Salmonella* strains

Plasmids were recognized as large scaffolds in the genome assemblies that did not match a closely related reference genome.

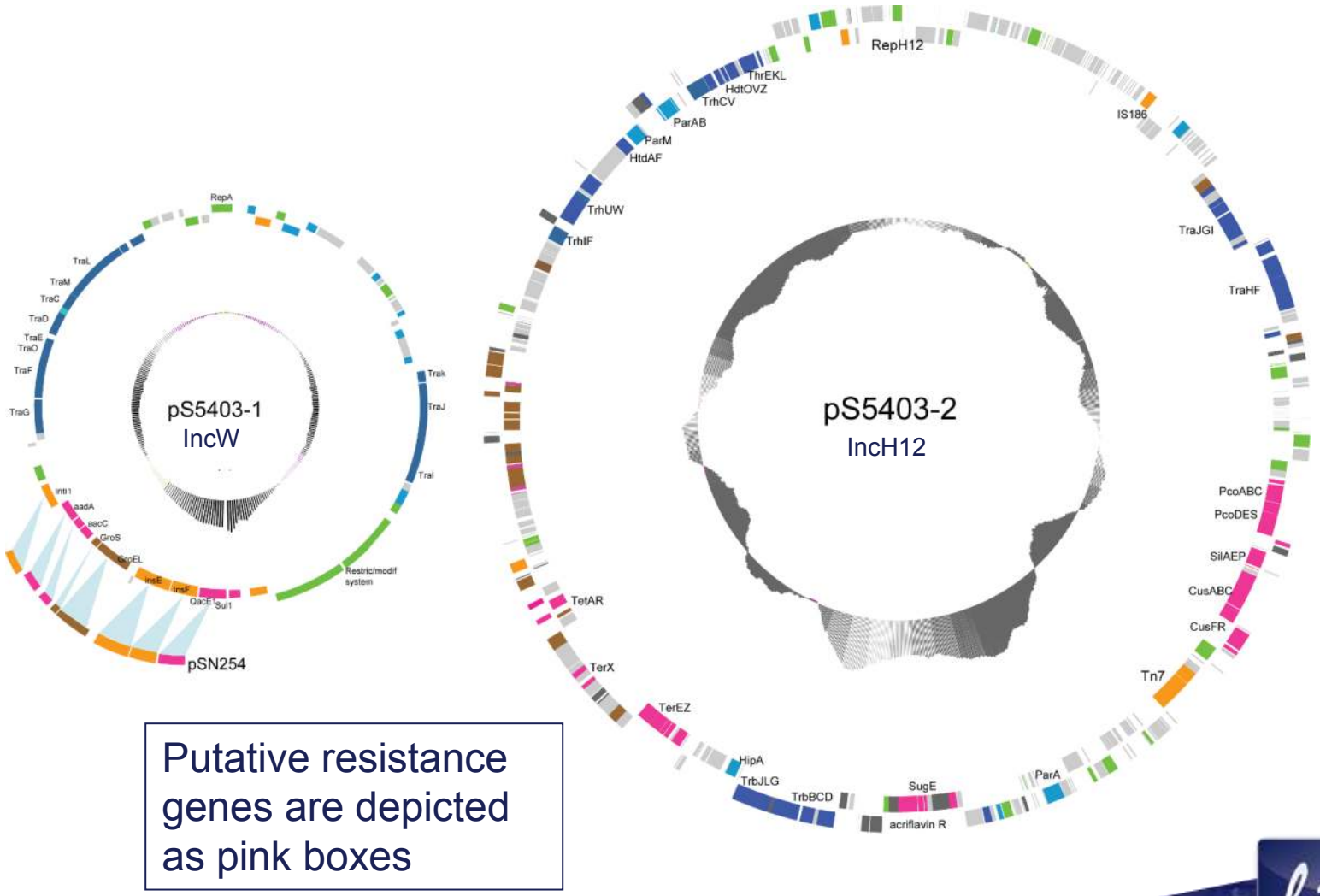


Novel IncI1 type plasmids from three *Salmonella enterica* serotypes encode virulence functions

- Putative virulence genes are depicted as red boxes
- May be helper plasmids for mobilization of integrative transposons in *S.* Inverness and *S.* Urbana



Two plasmids in *S.* Montevideo encode drug resistance genes

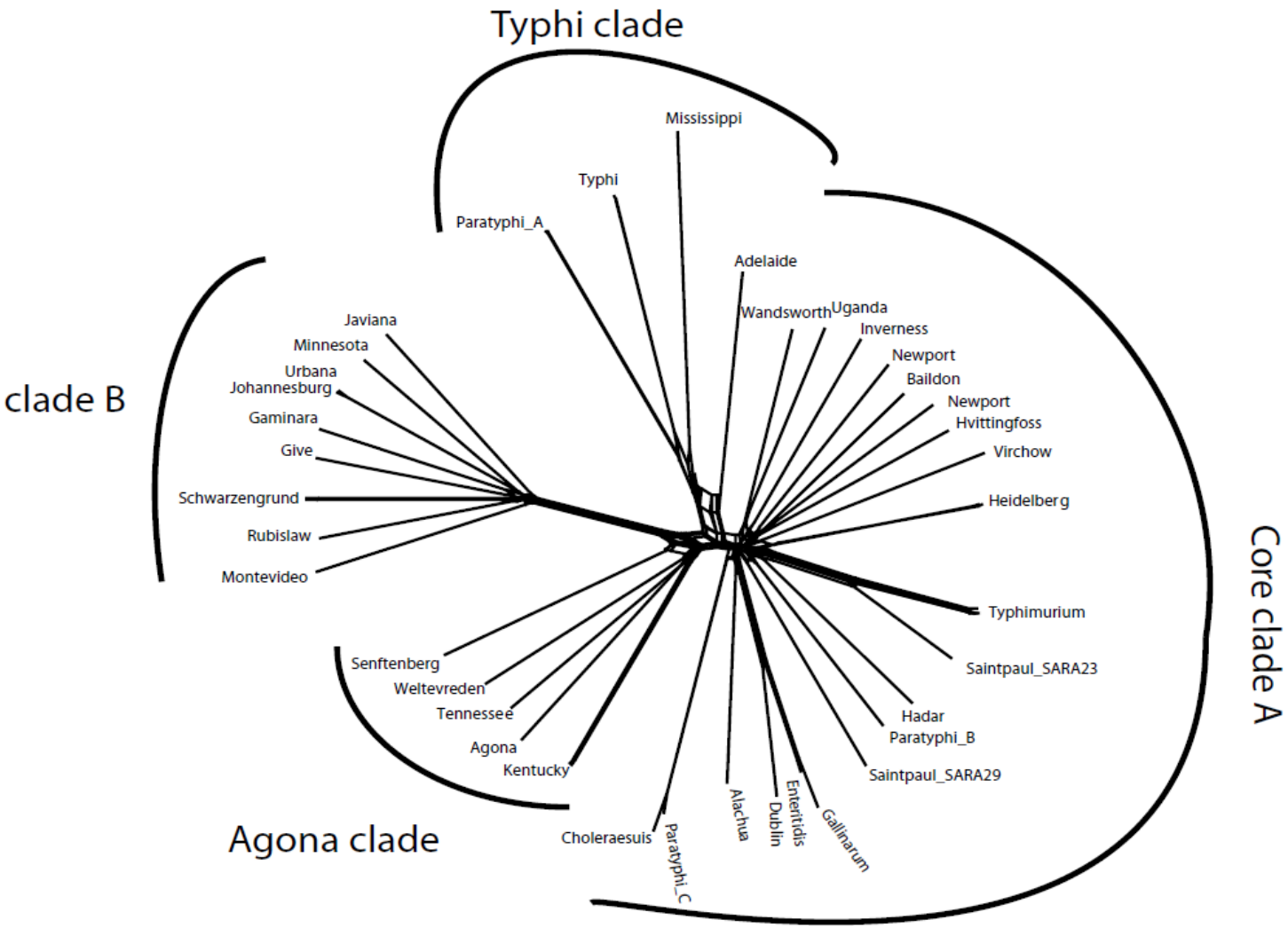


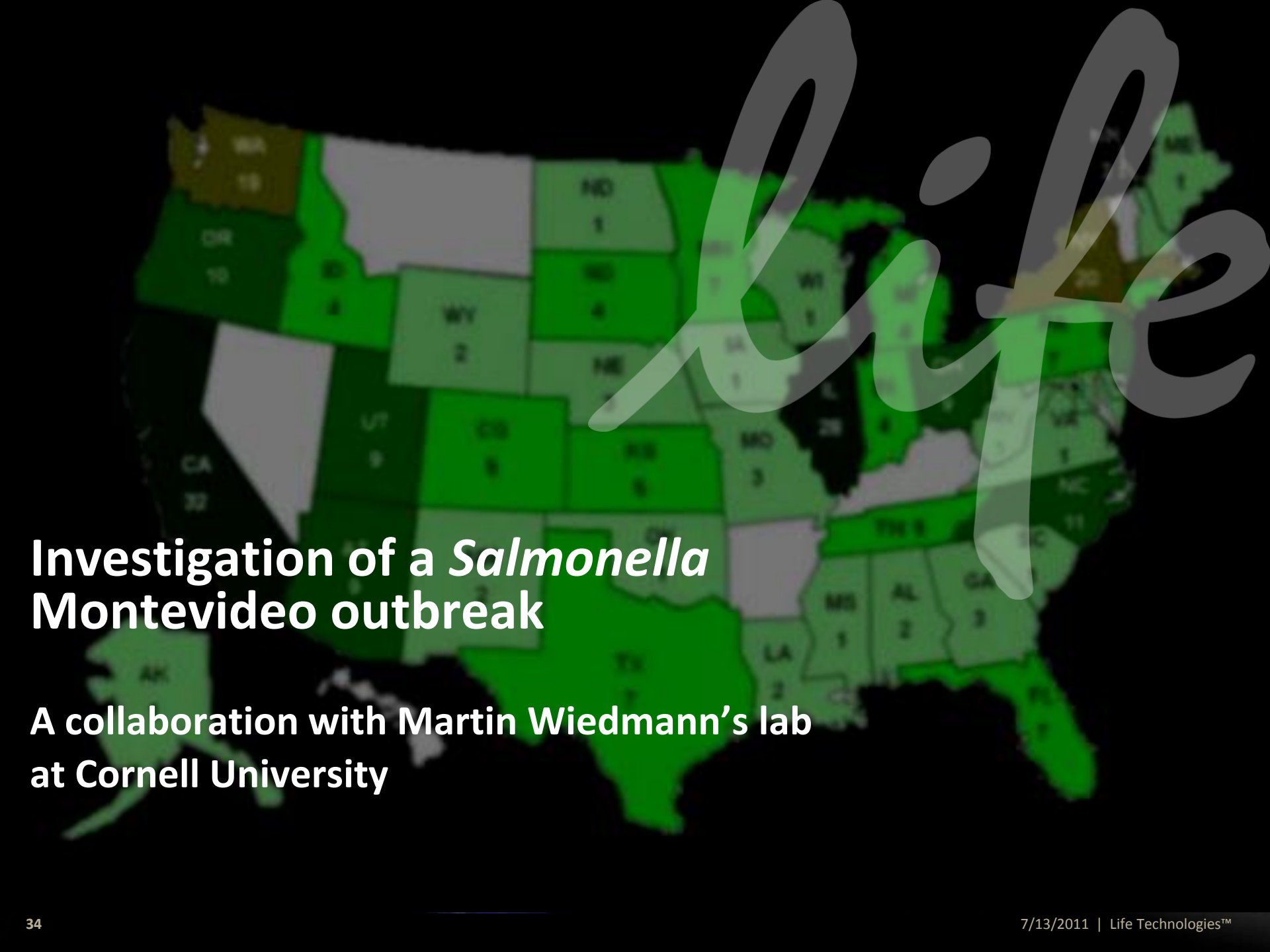
Putative resistance genes are depicted as pink boxes

Genomic organization of the O-antigen loci of 16 *Salmonella*

Matt Raineri

Population structure of *S. enterica* subsp. *enterica*



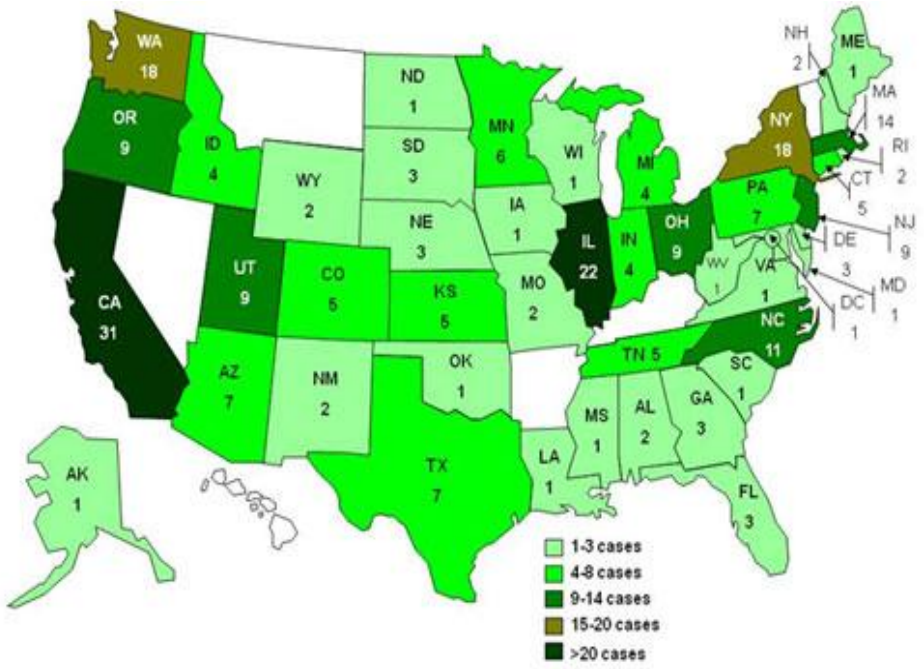


Investigation of a *Salmonella* Montevideo outbreak

A collaboration with Martin Wiedmann's lab at Cornell University

2009-2010 *S.* Montevideo outbreak

- 272 persons sickened in 44 states and DC
- Linked to black and red pepper in Italian-style meats
- PFGE typing resolve outbreak strains into two closely related types, one of which is identical to the most common sporadic isolate



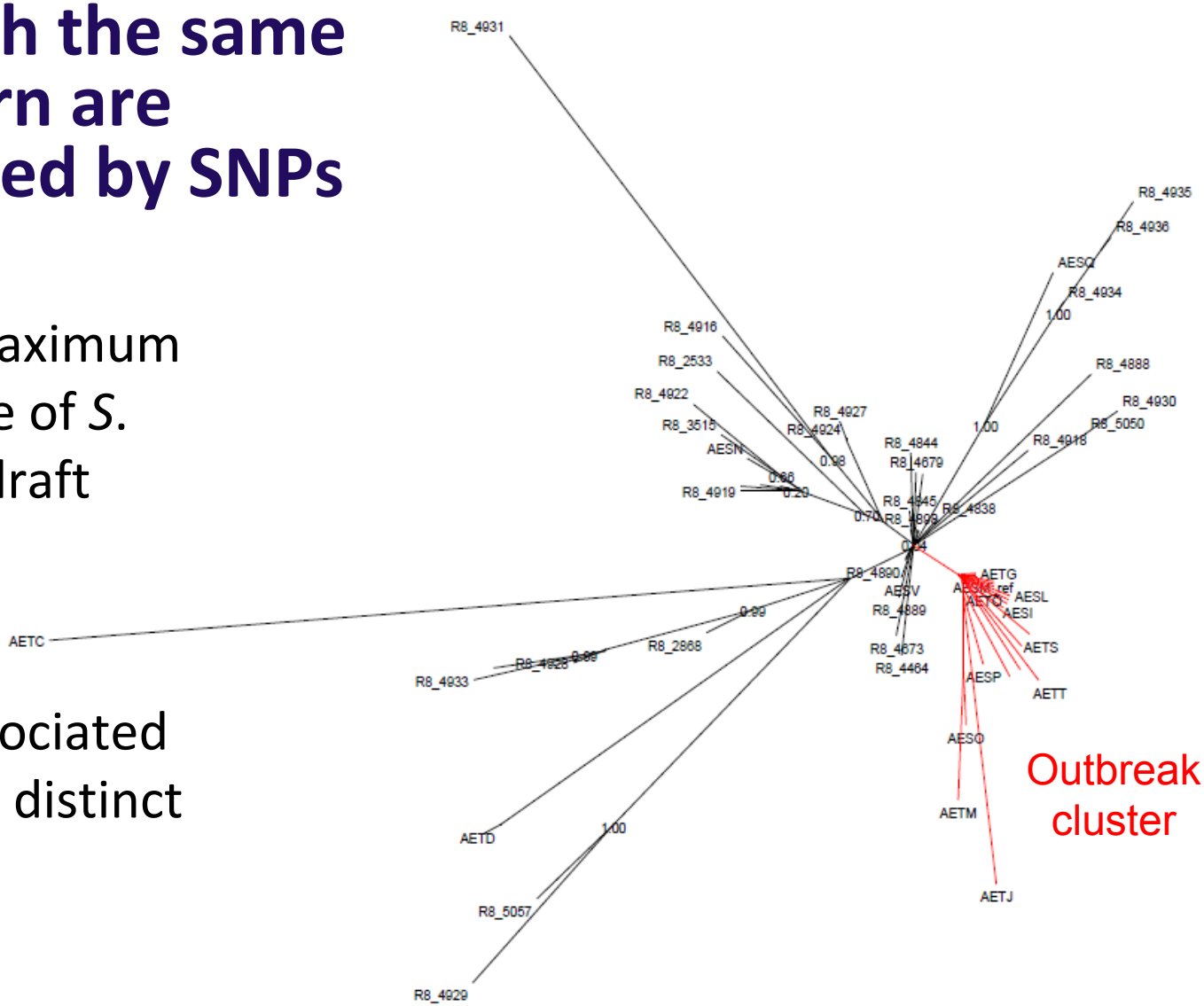
http://www.cdc.gov/salmonella/montevideo/

Whole-genome typing allows resolution of closely related strains

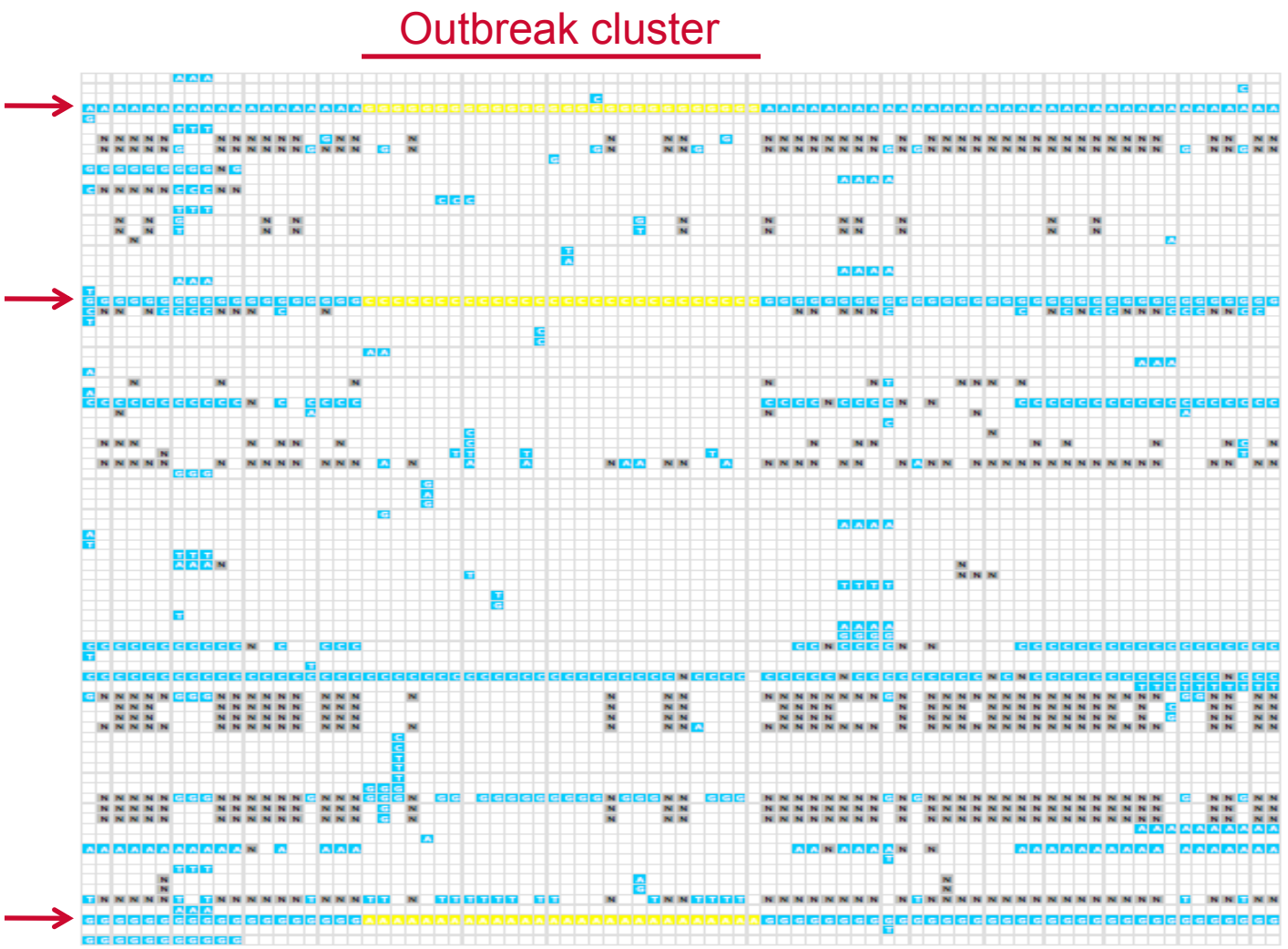
- Goal: use whole-genome resequencing to achieve increased resolution for strain detection and source tracking
- Methods
  - 47 *S*. Montevideo strains—22 presumptive case isolates and 25 isolates preceding the outbreak—were subjected to SOLiD® System fragment sequencing. All isolates had the same *Xba*I PFGE pattern.
  - Reads were mapped to a published draft *S*. Montevideo genome
  - High-confidence SNPs were identified in each strain and used for phylogenetic analyses

Isolates with the same PFGE pattern are differentiated by SNPs

- SNP-based maximum likelihood tree of *S.* Montevideo draft genomes
- Outbreak-associated strains form a distinct cluster



Identifying outbreak-specific SNPs



Sequencing of the German *E. coli* O104:H4 outbreak

A collaboration with Dag Harmsen and Alexander Mellman at the University Hospital Muenster

Timeline of Outbreak Events

Date	Event
Friday May 20th	LB226692 patient stool specimen sampled
Sunday May 22nd	Germany reports for the first time to the European Center for Disease Prevention and Control (ECDC) a significant increase in the number of patients with hemolytic uremic syndrome (HUS) and bloody diarrhoea caused by Shiga toxin-producing E. coli (STEC)
Monday May 23rd	LB226692 stool sample arrived at the Institute of Hygiene of the University of Münster, Germany
Tuesday May 24th	EHEC/STEC diagnosis of LB226692 finished and pure culture available
Wednesday May 25th	ECDC releases first epidemiologic update warning regarding STEC outbreak in Germany
Thursday May 26th	Subtyping O104:H4 of LB226692 finished and genomic DNA of LB226692 extracted (MLST & Shigatoxin Sequencing on CE)
Friday May 27th	Genomic DNA of LB226692 shipped by the Institute of Hygiene to Life Technologies

European *E. coli* Outbreak Strain Identified using Ion PGM™ Sequencer in 3 days

Rapid sequencing, de novo assembly & identification of novel microbial strains.

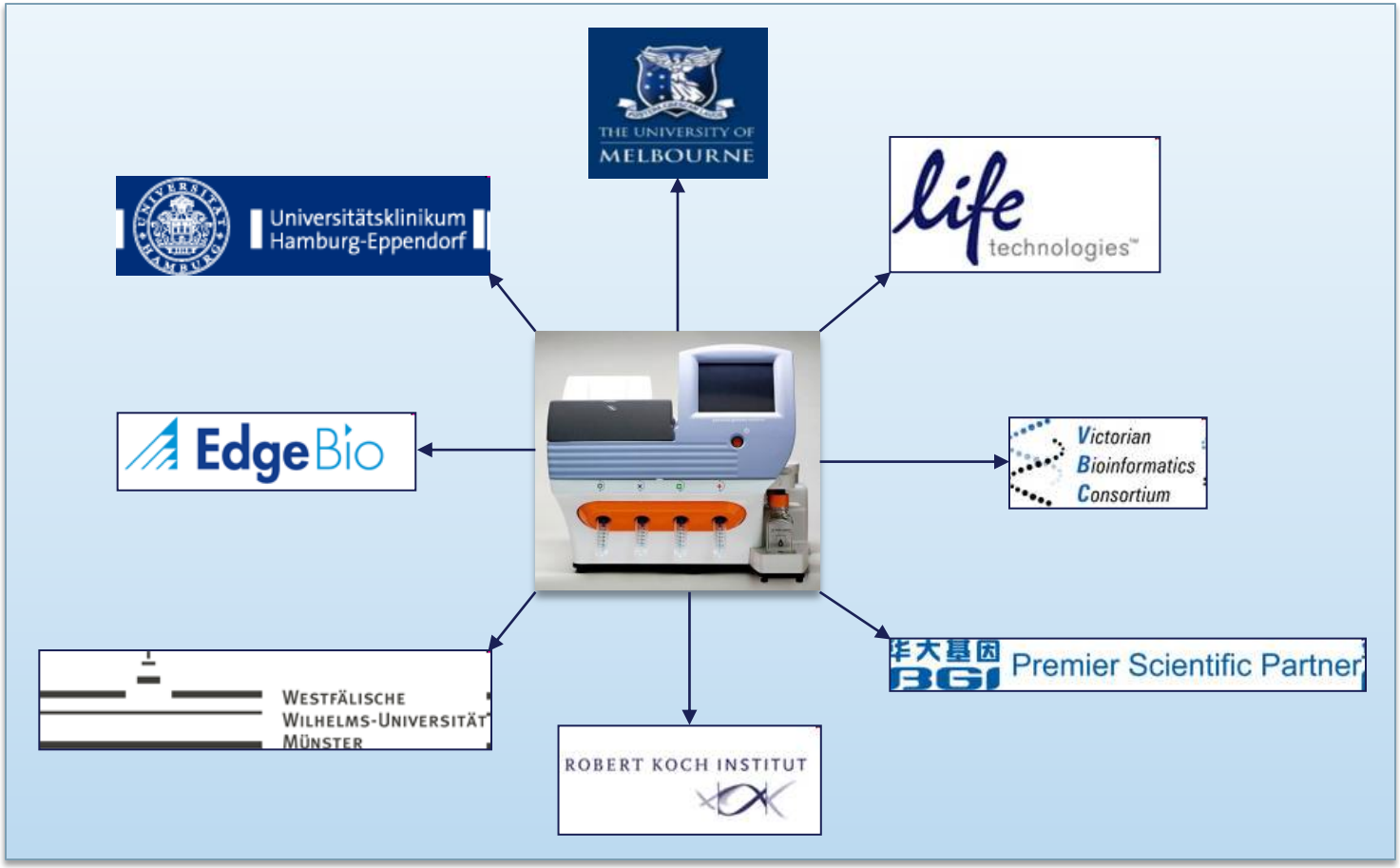
Monday May 30*	Library preparation	O104:H4 and HUSC41 samples (reference) strain libraries prepared
Tuesday May 31	Sequencing runs	0104:H4 amplified and sequenced 2 x 2 runs (Ion 314)
Wednesday June 01	Sequencing runs	0104:H4 sequenced 3 x 2 runs (Ion 314)
Thursday June 02	Analysis	Draft Genome identified, Assembled, Submitted and Released from NCBI
Friday June 03	Assay Design	TaqMan® E. coli 0104 Detection Assay Designed

*May 22 ECDC reports significant increase in patients with hemolytic uremic syndrome

"The biggest advantage [of the PGM] from my point of view as a public health official is that it's speedy, and speed is what is needed at the moment," Prof. Dr. Med Dag Harmen, University Hospital Muenster

"[The PGM] takes the shortest time to generate genomic data." Junjie Qin, BGI

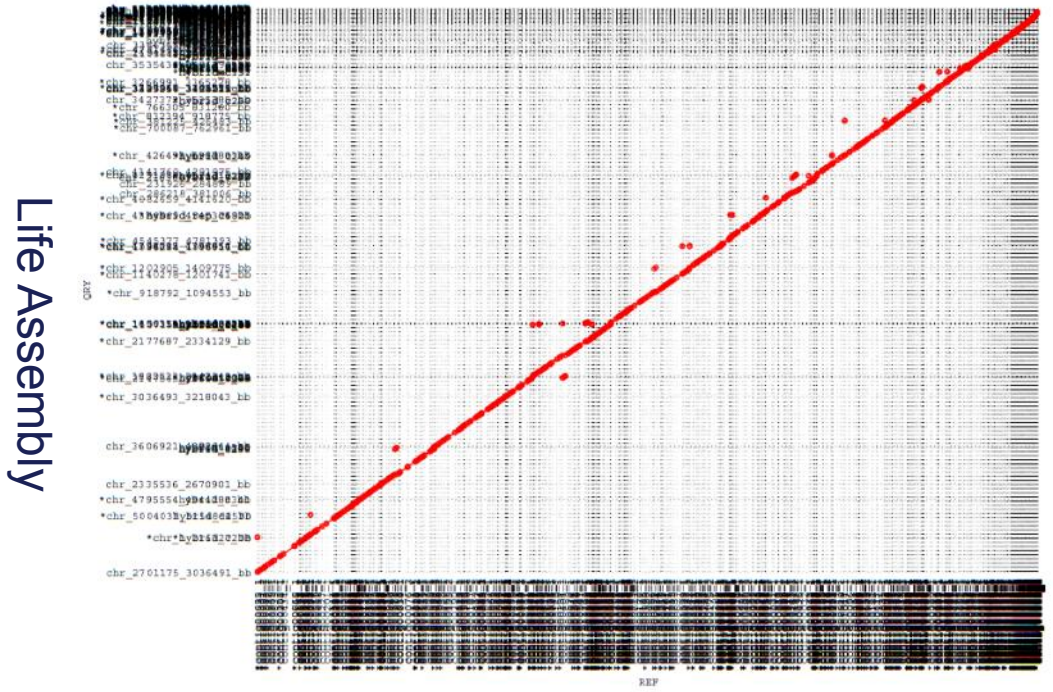
Ion data release spawns analysis frenzy



https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki

Comparison of EHEC genome assemblies by LifeTech and BGI

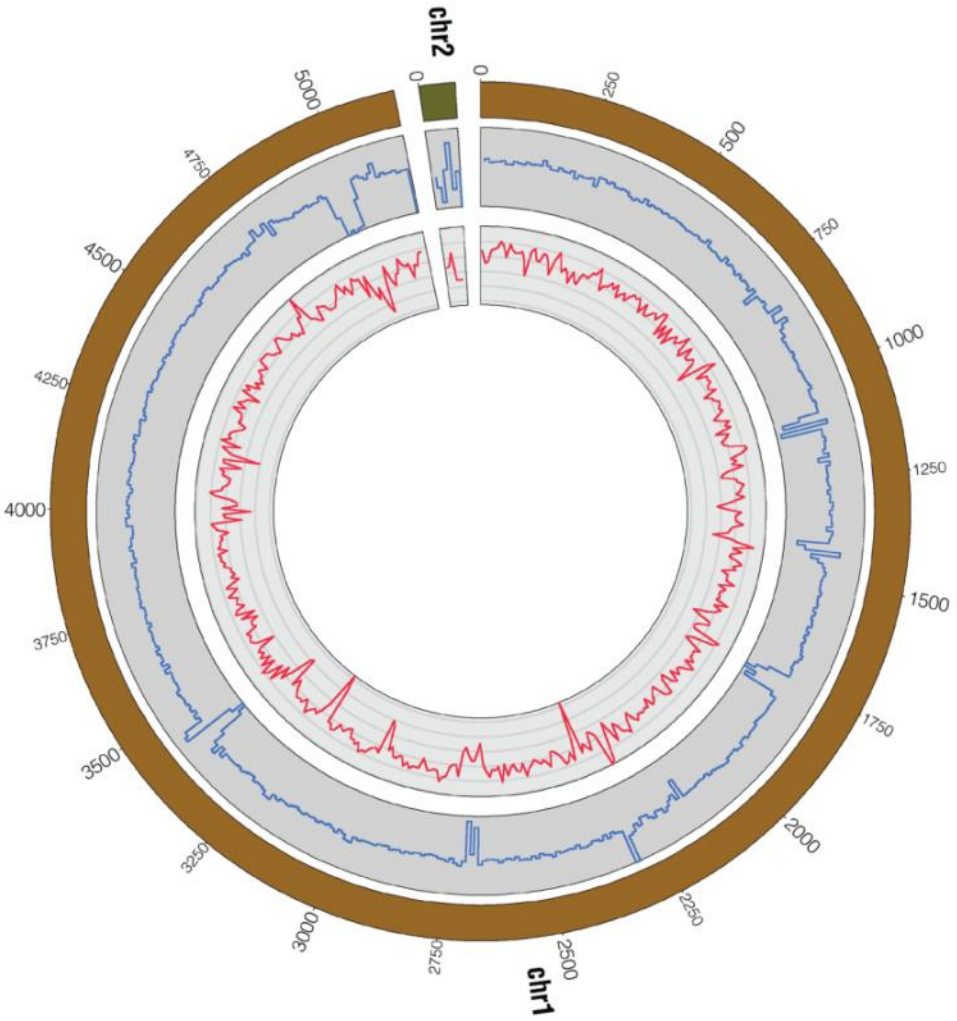
	Life Assembly	BGI Assembly
Strain sequenced	LB226692	TY-2482
perc A	24.7	24.7
perc C	25.4	25.4
perc G	25.3	25.1
perc T	24.7	24.7
Sum contig length	5,450,264	5,201,850
Num contigs	364	1,217
Mean contig length	14,973	4,274
Median contig length	762	938
N50 value	181,540	13,670
N90 value	14,537	2,689
Max	475,662	72,019



Coverage Plot of Genome Assembly

- The combination of the assembled sequence from both Hamburg/BGI & Muenster/Life Technologies strains and is aligned against the 55989 reference sequence

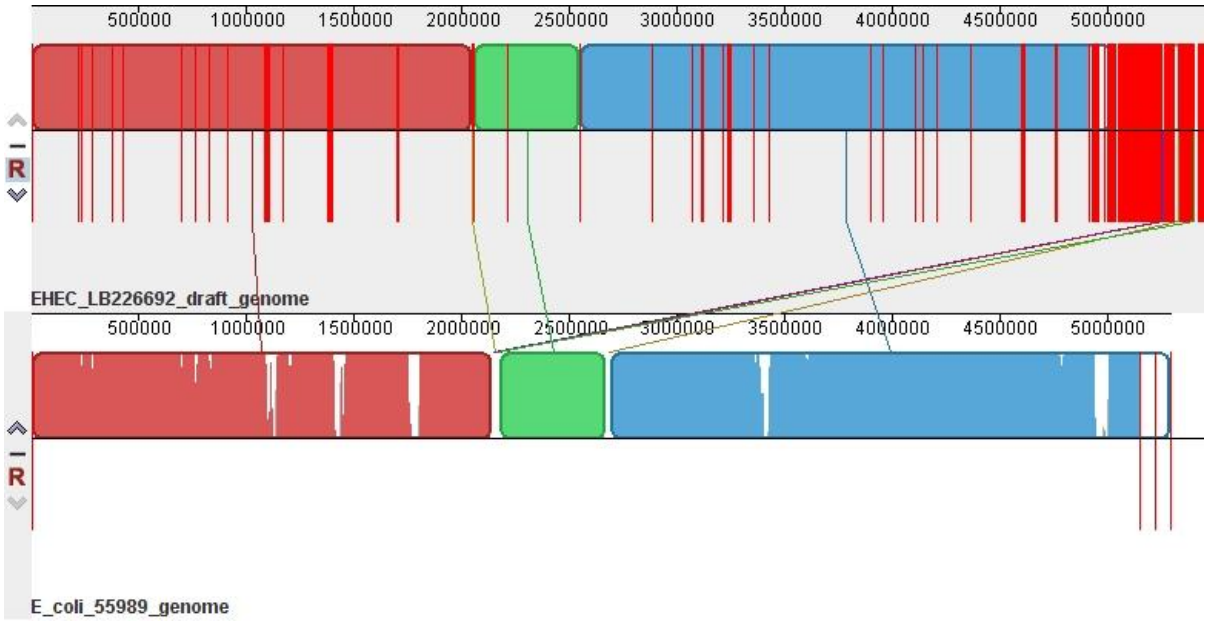
- Dips in coverage (blue line) represent differences between the strains



Genome alignment between EHEC LB226692 draft assembly and EAEC 55989 complete genome

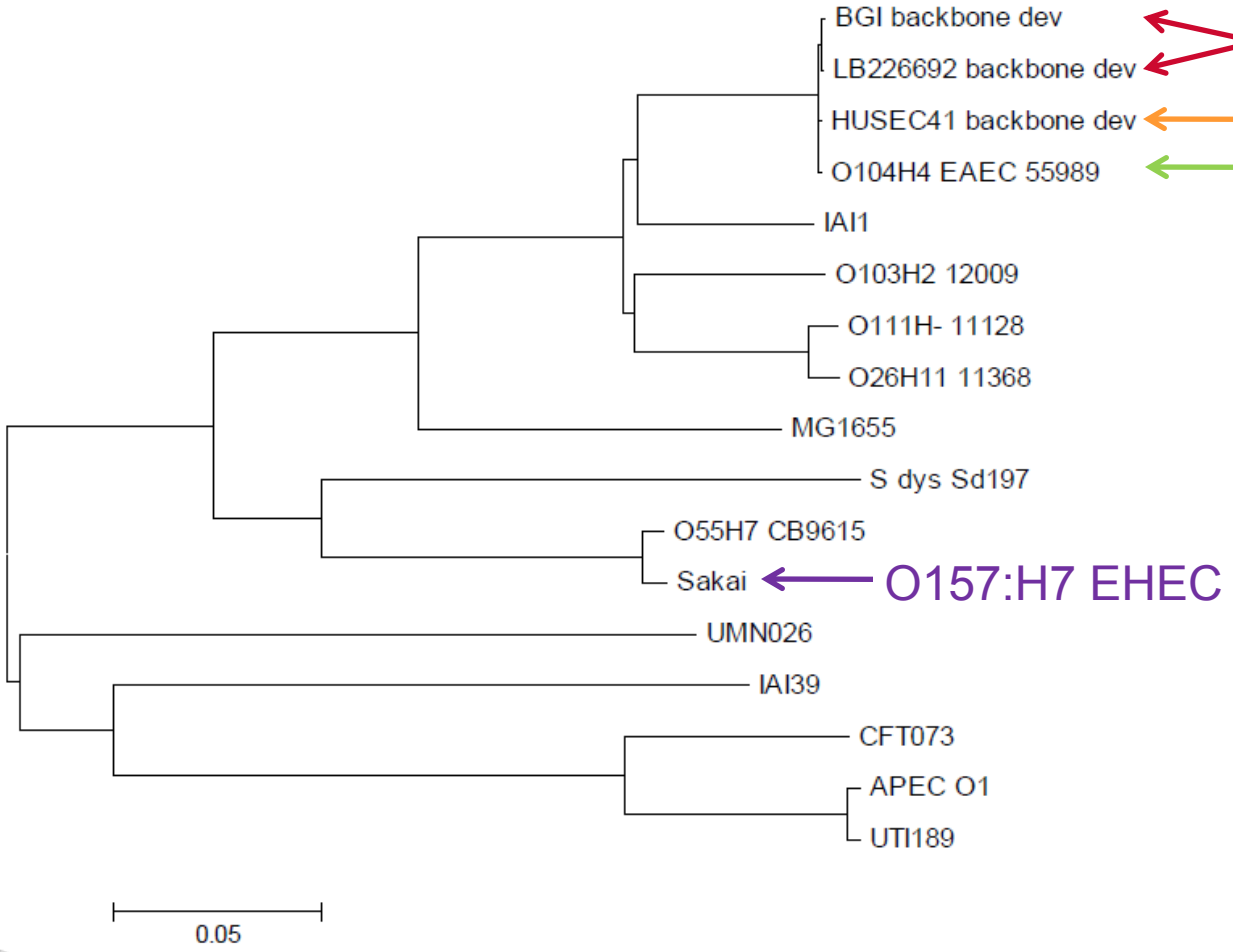
E. coli LB226692 draft assembly

E. coli 55989 complete genome



	perc A	perc C	perc G	perc T	Sum contig length	Num contigs	Mean contig length	Median contig length	N50 value	N90 value	Max contig length
Life	24.7	25.3	25.3	24.7	5,450,264	364	14,973	762	181,540	14,537	475,662

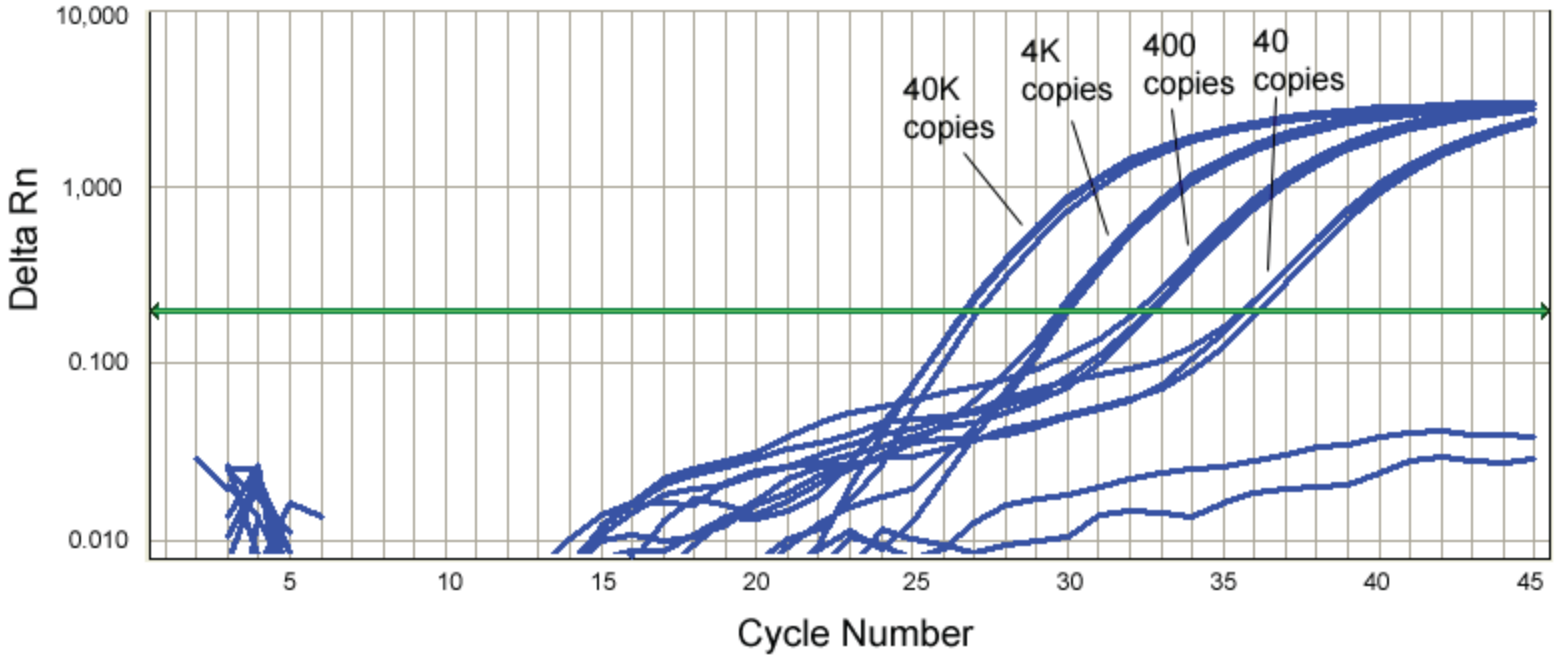
Relationship of the current outbreak strain to other *E. coli*



SNP-based phylogeny of the current O104:H4 outbreak strain with the historical STEC from 2001 and additional *E. coli* strains

Designing outbreak-specific assays

- Identification of signature sequences specific for the outbreak strain versus > 70 other E. coli strains led to rapid development of a specific assay



Acknowledgments

- **Life Technologies**
  - Lovorka Degoricija
  - Elena Bolchacova
  - Manohar Furtado
  - Olga Petrauskene
  - Rixun Fang
  - Pius Brzoska
  - Yongmei Ji
  - Steve McLaughlin
  - Simone Guenther
  - Alain Rico
  - Greg Govoni
  - Paolo Vatta

- **Cornell University**
  - Martin Wiedmann
  - Henk den Bakker
  - Andrea Moreno Switt
  - Vania Ferreira
  - Renato Orsi
  - Karin Hoelzer
  - Matt Raineiri
  - Lorraine Rodriguez-Rivera
  - Stephanie Brown

- **University Hospital Münster**
  - Dag Harmsen
  - Alexander Mellman
  - Helge Karch

Disclaimer

For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.

Craig Cummings, Ph.D.
Sr. Staff Scientist, Bioinformatics
Life Technologies
Foster City, CA
craig.cummings@lifetech.com