



National Institute for Public Health
and the Environment
Ministry of Health, Welfare and Sport

The 100k genomes project And preliminary WGS results

Eelco Franz

Centre for zoonoses and
environmental microbiology

NRL The Netherlands

Eelco.franz@rivm.nl



The 100k Foodborne Pathogen Genome Sequencing Project (100kGP)

- The lack of food-related bacterial genomes is hindering advancements to improve the safety and security of the world food supply
- University of California Davis (UC Davis), funded by FDA → 100kGP
- 10 HiSeq 2000 instruments at UC Davis
- A small number of whole genomes will be “finished” to completion for use as reference genomes (*de novo* reconstruction)
- The vast majority of isolates will be sequenced and assembled as draft genomes for eventual public release (need mapping to reference)





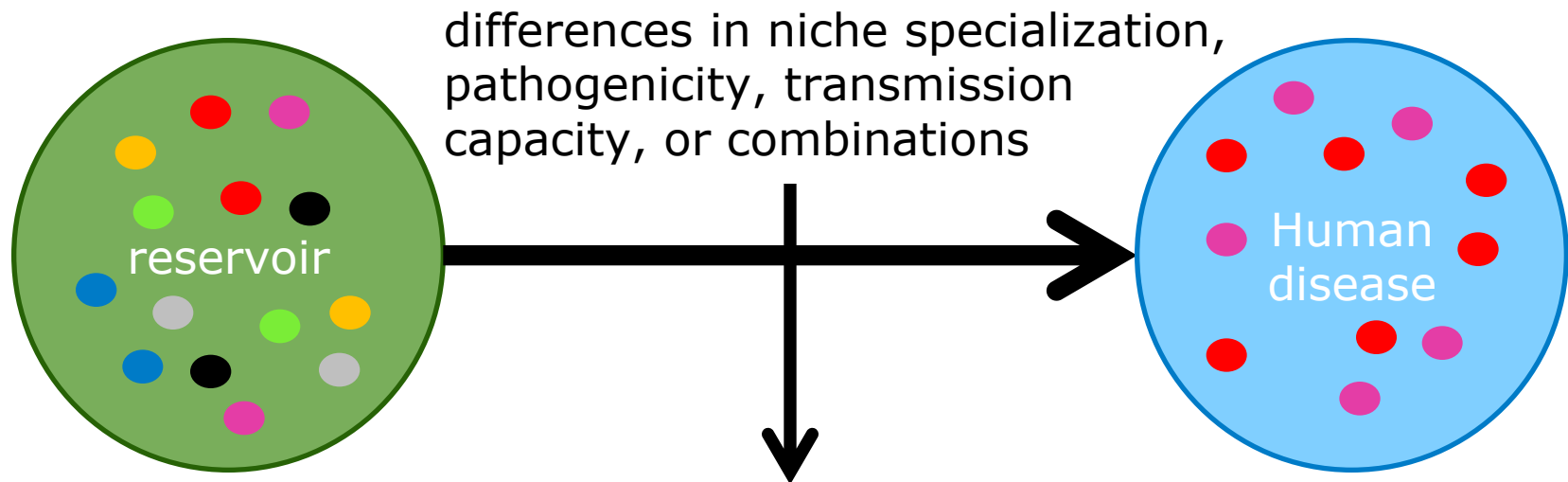
Outcomes of the 100kGP

- Genomic catalog of some of the most important outbreak organisms that impact human health (sequencing for free)
- Public access to the database after allowing for an academic publication window by submitting partners
- A culture bank will be created and maintained at UC Davis within the School of Veterinary Medicine
- Outbreak assistance: providing the basis of new diagnostic testing methods that leverages the latest in genomics technology to provide time-critical diagnostic assistance
- Assisting in systematic definition of biomarker gene sets associated with persistence, serotype diversity, geographic dispersion, antibiotic resistance, pathogenesis, and host association
- In principle open for all institutes



Our interest in STEC genomics

Understanding the factors differentiating genotypes and phenotypes that form a significant human health risks from those posing lower risk

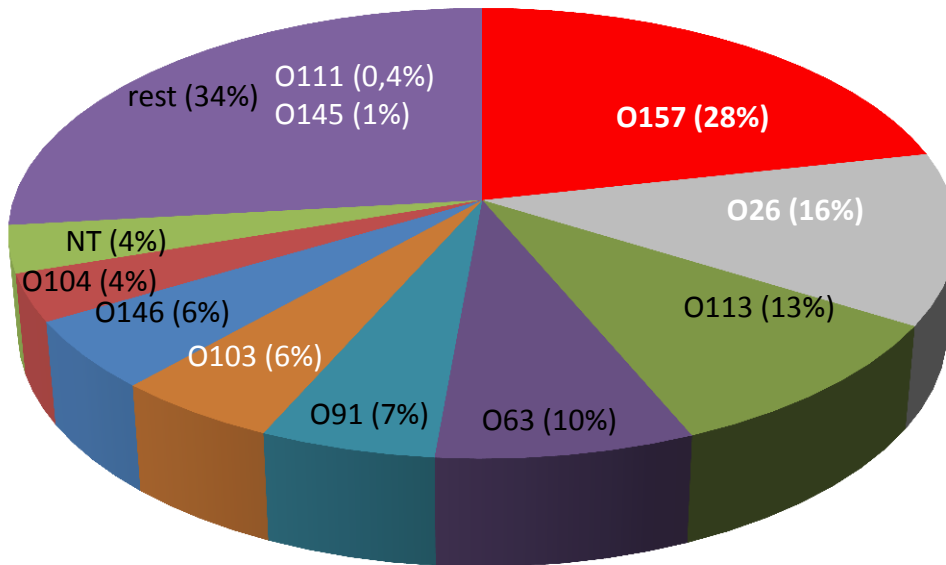


- Molecular typing
- Comparative genomics
- Comparative physiology (metabolism, stress-resistance, virulence)



NL contribution to the 100kGP

STEC serotypes in The Netherlands (2011)



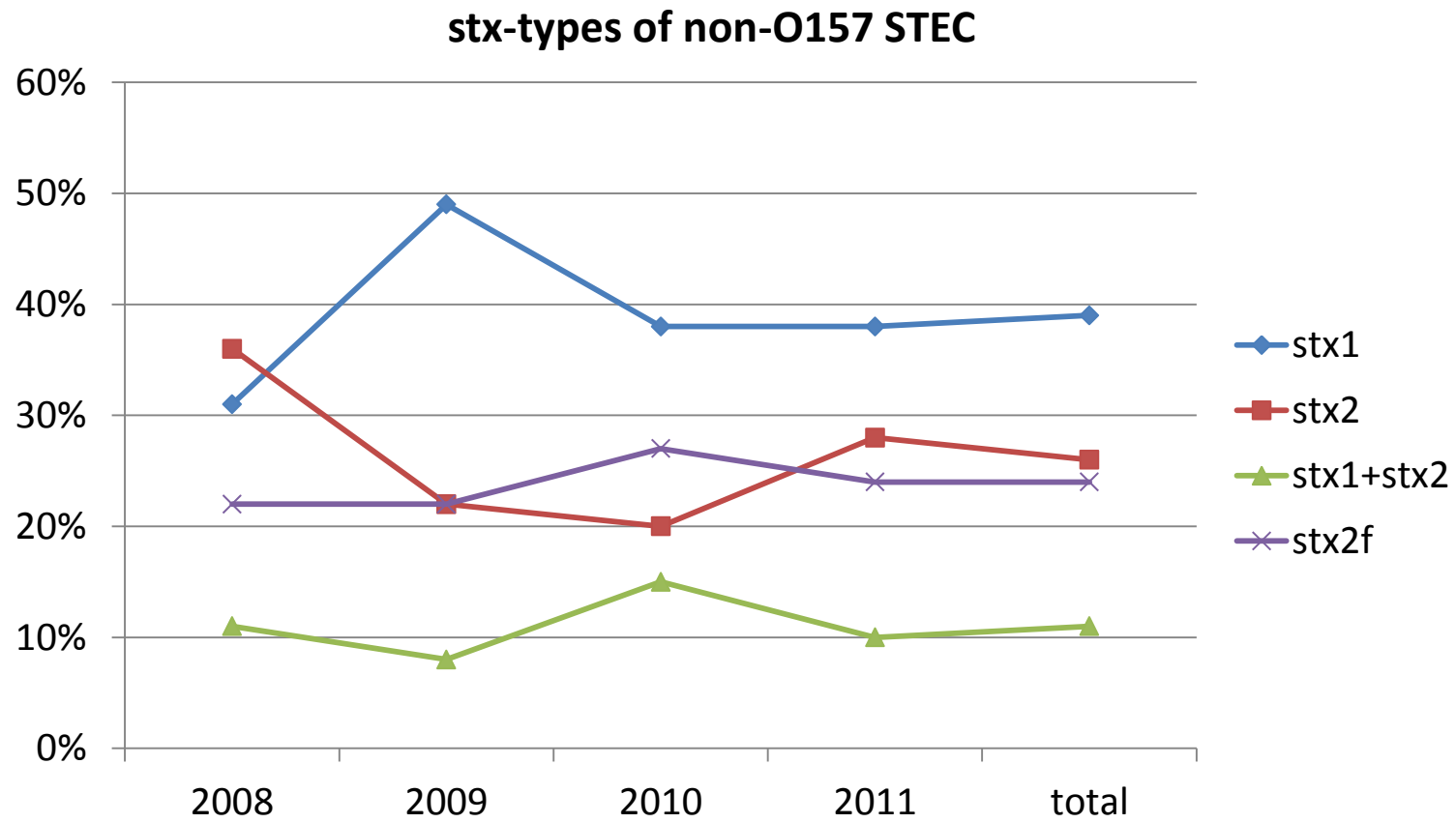
O157 incidence approx. 0.3
Total STEC incidence approx 1.4

Non-O157 in general more mild
sequela

1. How do non-O157 strains relate (on genome basis) to O157 and to each other?
2. Test case for extracting typing information *in silico* (virulence profile, antibiotic resistance, MLST)



Special interest: *stx2f* STEC is common in NL





stx2f serotypes

	With stx2f / O-type	% of stx2f (n=87)
O2	2 / 4	2%
O2:H-	1 / 1	1%
O2:H6	1 / 2	1%
O2:H29	0 / 1	0%
O16:H5	1 / 1	1%
O35:H19	1 / 1	1%
O63:H6	47 / 47	54%
O73:H18	1 / 2	1%
O96:H7	1 / 1	1%
O101:H-	1 / 3	1%
O113	15 / 30	17%
O113:H-	2 / 3	2%
O113:H4	0 / 7	0%
O113:H6	12 / 12	14%
O113:H7	1 / 1	1%
O113:H21	0 / 7	0%
O121:H5	1 / 1	1%
O125:H6	12 / 12	14%
O132:H34	4 / 6	5%
O NT:H6	1 / 2	1%
total	87 / 110	100%

stx2f O63 isolates:

- ☐ eae-positive
 - ☐ Relatively low number of additional STEC virulence factors
 - ☐ Mainly phylogroup B2 (most STEC A/B1)
- Relation to other B2 (ExPEC strains)?
- Genomic
 - Phenotypic



Strains for 100kGP

The Netherlands (n=100) and Italy (n=100)

Top-5 STECs, supplemented with strains of specific interest (O?, O63)

N.T.	O111	O120	O157	O157	O157:H7	O26	O26	O26:H11	O63:H6
O?	O111	O121	O157	O157	O157:H7	O26	O26	O26:H11	O63:H6
O?	O111	O121	O157	O157	O157:H7	O26	O26	O26:H11	O63:H6
O?	O111	O128	O157	O157	O157:H7	O26	O26	O26:H6	O63:H6
O103	O111	O145	O157	O157	O157:H7	O26	O26:H11	O63:H6	O91
O103	O111	O145	O157	O157	O157:H7	O26	O26:H11	O63:H6	O91
O103	O111	O145	O157	O157	O157:H7	O26	O26:H11	O63:H6	O91:H-
O103:H11	O111	O145	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:H-
O103:H2	O111	O145	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:H-
O103:H2	O111	O145	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:H-
O103:H2	O111:H-	O145	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:H14
O103:H2	O111:H-	O145	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:H14
O103:H2	O111:H19	O145:H-	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:H14
O103:H2	O111:H8	O145:H-	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:H14
O103:H2	O111:H8	O145:H-	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:H21
O103:H25	O113	O145:H-	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:H26
O103:H-roug:H	O113	O145:H28	O157	O157:H7	O157:H7	O26	O26:H11	O63:H6	O91:NM
O111	O113	O157	O157	O157:H7	O26	O26	O26:H11	O63:H6	
O111	O113	O157	O157	O157:H7	O26	O26	O26:H11	O63:H6	
O111	O113	O157	O157	O157:H7	O26	O26	O26:H11	O63:H6	



NL O157 strains for 100kGP

Include strains showing strong differential environmental survival

< 200 days survival

Strain	LSPA	<i>rpoS</i> mutation
M619	I	T433G (Y145D)
M623	I/II	T402G (F134L)
M626	I/II	488, insert 1bp
M627	I/II	396, insert 12bp
M628	I/II	T383A (I128N)
M629	I	654, deletion 13bp
M630	I/II	97, deletion 4bp
M631	II	75, deletion 1bp
M632	II	G376A (G126R)
M634	I/II	A917C (Q306P)
M636	I/II	C601T (Q201Stop)

> 200 days survival

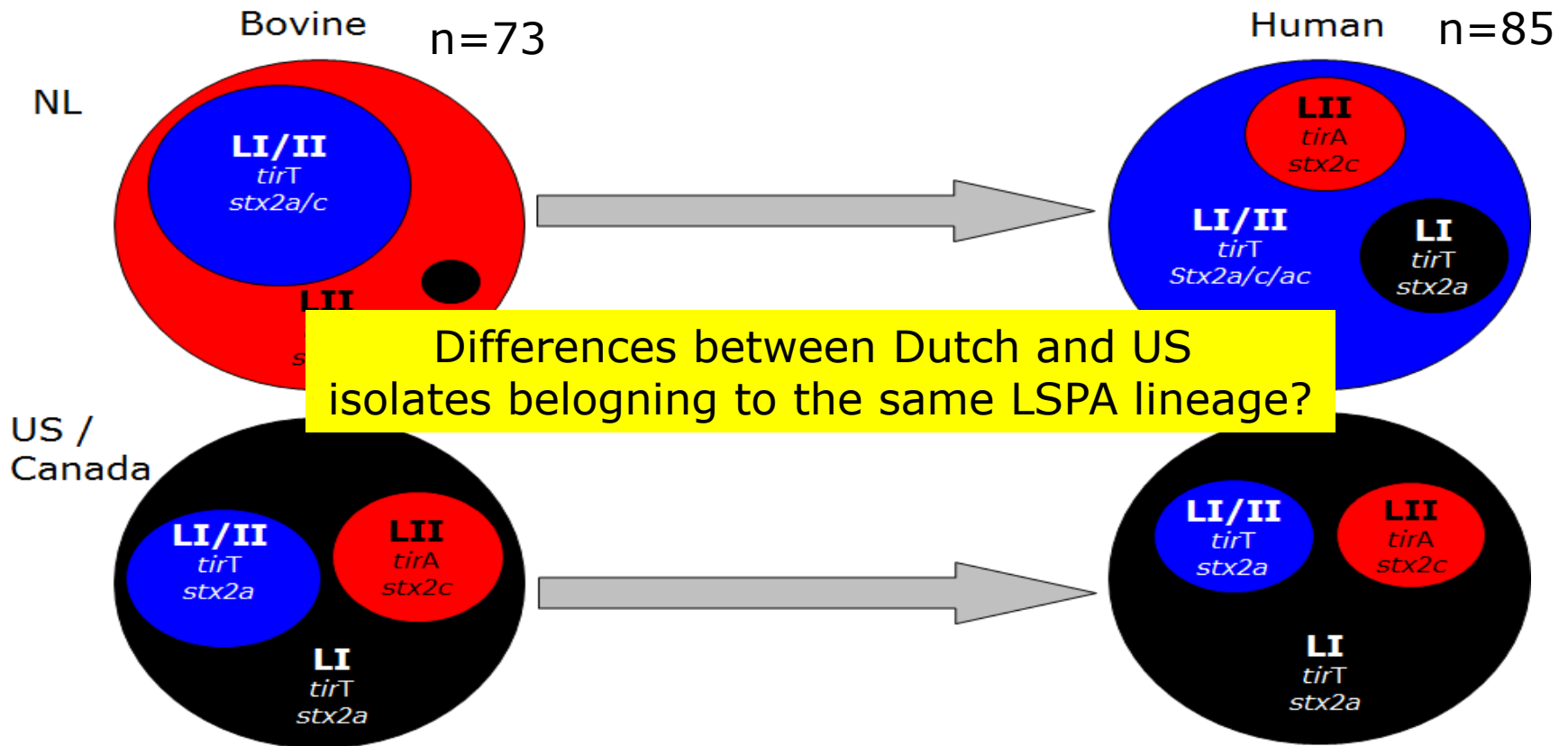
Strain	LSPA	<i>rpoS</i> mutation
M622	II	-
M625	II	-
M633	I/II	-
M635	I/II	-
M637	I/II	-
M638	I/II	-
M639	I/II	-

Other genetic differences???
→ WGS



Preliminary results WGS 0157

Previously: Lineage Specific Polymorphism Assay (LSPA) (Yang 2004)





WGS of Dutch O157

In collaboration with Central Veterinary Institute (Illumina Miseq) and United States Department of Agriculture (Roche454)(Jim Bono)



N=39

	LSPA I	LSPA I/II	LSPA II
Human	H06	H25	H2
	H07	H27	H19
	H09	H38	H24
	H13	H42	H32
	H15	H44	H51
		H48	
		H49	
		H66	
Bovine		H83	
	A42	A25	A12
		A37	A13
		A40	A16
		A45	A29
		A48	A30
		A51	A32
		A60	A34
		A62	
		A63	
		A69	
		A72	
		A76	



Mapping of O157 genomes

Mapping: the re-sequencing of a genome aims to capture information on

- Single Nucleotide Polymorphisms (SNPs),
- INsertions and DEletions (INDELs) and
- Copy Number Variants (CNVs)

between representatives of the same species, usually in cases where a reference genome already exists (not *do novo* genome construction)

Mapping possible to 4 complete genomic and pO157sequences:

Lineage I: EDL933 (NC_002655) and **Sakai (NC_002695)**

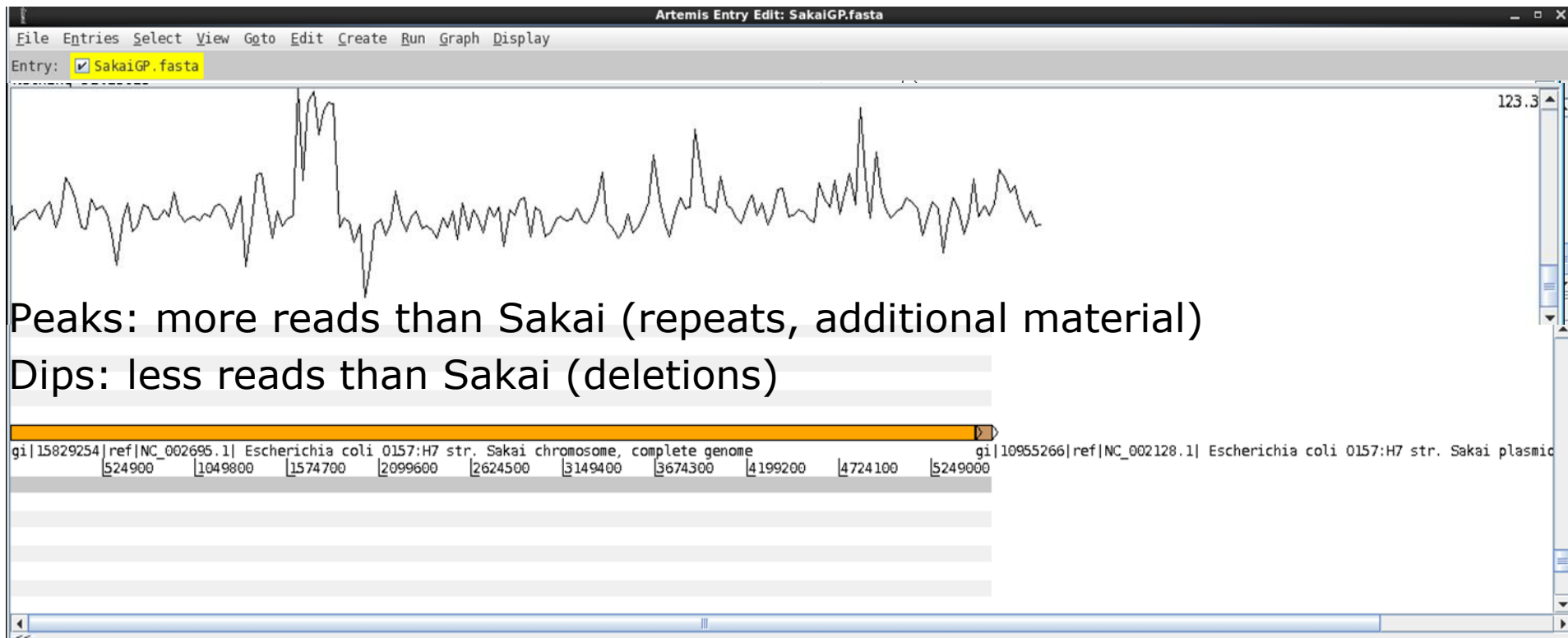
Lineage I/II: TW14359 (NC_013008) and EC4115 (NC_011353)

No complete LSPA II available (yet)



Coverage of sequences

	LI	LI/II	LII
Mapped	81.43%	80.16%	80.77%





In silico presence/absence virulence genes

Location		H06	H07	H09	H13	H15	A42	H25	H27	H42	H44	H48	H49	H66	H83	A25	A37	A40	A45	A48	A51	A60	A62	A63	A69	A75	A76	H02	H17	H19	H24	H32	H51	A12	A13	A16	A29	A30	A32	A34
Eae	Chromosome	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
espA	Chromosome	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
espF	Chromosome	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
espJ	Chromosome	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
nleB	Chromosome	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
nleC	Chromosome	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
tccP	Chromosome	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
espP	pO157	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+
etpD	pO157	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
hlyA	pO157	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
katP	pO157	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+
toxB	pO157	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+
RT*	pO157	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

* RT= reverse transcriptase



LSPA Lineage I

LSPA Lineage I/II

LSPA Lineage II

Single Nucleotide Polymorphisms (SNPs)

Sakai			Sakai		
	Genome	Plasmid		Genome	Plasmid
H06	668	43	A42	302	137
H07	891	31	A25	829	11
H09	822	35	A37	1211	24
H13	149	66	A40	1489	19
H15	147	49	A45	1418	190
H25	735	16	A48	1250	19
H27	819	14	A51	1312	21
H42	814	14	A60	969	21
H44	830	11	A62	1319	22
H48	791	12	A63	1058	50
H49	965	4	A69	568	14
H66	1781	4	A75	673	17
H83	844	13	A76	868	15
H02	2796	40	A12	1072	18
H17	1058	23	A13	1101	19
H19	935	12	A16	1101	16
H24	1605	13	A29	1226	22
H32	937	22	A30	1063	25
H51	906	16	A32	1106	22
			A34	950	24

Single Nucleotide Polymorphisms (SNPs)



# NL strains	CHROM	PLASMID	# NL strains	CHROM	PLASMID
40	37	0	20	20	0
39	9	0	19	13	1
38	13	0	18	15	0
37	19	0	17	11	0
36	5	3	16	19	1
35	6	2	15	28	0
34	70	0	14	18	1
33	22	0	13	28	0
32	22	0	12	45	0
31	19	3	11	43	0
30	14	0	10	159	1
29	5	0	9	86	1
28	16	0	8	220	2
27	16	1	7	201	2
26	10	0	6	180	2
25	20	1	5	255	10
24	44	0	4	431	10
23	28	1	3	969	32
22	14	1	2	2020	108
21	51	0	1	7222	280



Common SNPs in Dutch strains

Position	Locus tag	Start	End	Length	start_end	Gene	Product
40	205552 ECs0201	224944	225975	1031	224944..225975	metN	DL-methionine transporter ATP-binding protein
40	375317 ECs0357	373223	374911	1688	373223..374911		choline dehydrogenase
40	415750 ECs0392	415575	416234	659	415575..416234		carbonic anhydrase
40	431214 ECs0405	430715	431524	809		hpD	2-keto-4-pentenoate hydratase
40	469824 intergenic						
40	474603 ECs0447	474591	477734	3143			exonuclease Sbc
40	517475 ECs0486	516716	517663	947			cytochrome o ubiquinol oxidase subunit II
40	616828 intergenic						
40	648904 ECs0589	648484	649005	521			HypoProt
40	730801 ECs0654	729393	730925	1532			citrate lyase subunit alpha
40	745039 ECs0670	744109	745320	1211			D-alanyl-D-alanine carboxypeptidase
40	1987879 ECs2007	1987410	1989833	2423			BigB-like protein
40	2497693 ECs2521	2496634	2497995	1361			para-aminobenzoate synthase component I
40	3637135 ECs3645	3636098	3637399	1301		imA	23S rRNA 5-methyluridine methyltransferase
40	3744736 ECs3743	3744593	3745783	1190			Aspartate/ornithine carbamoyltransferase
40	3800637 ECs3788	3800463	3801458	995			arginine/ornithine transport system ATPase
40	3957788 ECs3954	3957087	3958607	1520			aerotaxis sensor receptor
40	4170227 ECs4154	4170219	4171508	1289			16S rRNA methyltransferase GidB
40	4213984 ECs4216	4213917	4216460	2543		irB	nitrite reductase (NAD(P)H) subunit
40	4451098 ECs4425	4450607	4452298	1691			phosphoethanolamine transferase
40	4538018 ECs4494	4537153	4538178	1025		dh	L-threonine 3-dehydrogenase
40	4617900 ECs4583	4617332	4617985	653	4617332..4617985		type III secretion system protein
40	4619636 ECs4587	4619553	4619771	218	4619553..4619771		HypoProt
40	4705415 ECs4671	4705284	4707113	1829	4705284..4707113		glucosamine--fructose-6-phosphate aminotransferase
40	4714723 ECs4680	4714680	4715495	815	4714680..4715495		ATP synthase F0F1 subunit A
40	5106168 ECs5024	5107225	5109648	2423	5107225..5109648	plsB	glycerol-3-phosphate acyltransferase



Some SNPs are lineage-biased

		REF	H06	H07	H09	H13	H15	A42	H25	H25_2	H27	H42	H44	H48	H49	H66	H83	A25	A37	A40	A45	A48	A51	A60	A62	A63	A69	A75	A76	H02	H17	H19	H24	H32	H51	A12	A13	A16	A29	A30	A34		
34 18711	A	#	#	#	#	#	#	#	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C		
34 53715	C	#	#	#	#	#	#	#	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T		
34 180160	T	#	#	#	#	#	#	#	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C		
34 289210	A	#	#	#	#	#	#	#	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G		
34 331010	C	#	#	#	#	#	#	#	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A		
34 958598	T	#	#	#	#	#	#	#	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G		
34 1976735	T	#	#	#	#	#	#	#	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C		
34 2343203	G	#	#	#	#	#	#	#	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A		
10 102481	C	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	T	T	T	T	T	#	T	T	T	#	T	T	
10 206849	C	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	T	T	T	T	T	#	T	T	T	#	T	T
10 660142	C	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	T	T	T	T	T	#	T	T	T	#	T	T
10 1494111	C	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	T	T	T	T	T	#	T	T	T	#	T	T
10 3405292	G	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	A	A	A	A	A	#	A	A	A	#	A	A
10 4514692	T	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	C	C	C	C	C	#	C	C	C	#	C	C
8 10685	C	#	#	#	#	#	#	#	#	#	#	#	#	#	T	T	#	#	T	#	T	T	T	#	T	T	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
8 503881	C	#	#	#	#	#	#	#	#	#	#	#	#	#	T	T	#	#	T	#	T	T	T	#	T	T	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
8 1067081	C	#	#	#	#	#	#	#	#	#	#	#	#	#	T	T	#	#	T	#	T	T	T	#	T	T	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
8 1372768	C	#	#	#	#	#	#	#	#	#	#	#	#	#	T	T	#	#	T	#	T	T	T	#	T	T	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
8 1372768	C	#	#	#	#	#	#	#	#	#	#	#	#	#	T	T	#	#	T	#	T	T	T	#	T	T	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
8 1443786	G	#	#	#	#	#	#	#	#	#	#	#	#	#	A	A	#	#	A	#	A	A	A	#	A	A	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
8 1638536	T	#	#	#	#	#	#	#	#	#	#	#	#	#	G	G	#	#	G	#	G	G	G	#	G	G	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#



Existing SNP-lists

Majority of our SNP also identified by others (but not all!)

Bono et al. in MBE 2012

Phylogeny of Shiga Toxin-Producing *Escherichia coli* O157 Isolated from Cattle and Clinically Ill Humans

Eppinger et al. in PNAS 2011

Genomic anatomy of *Escherichia coli* O157:H7 outbreaks

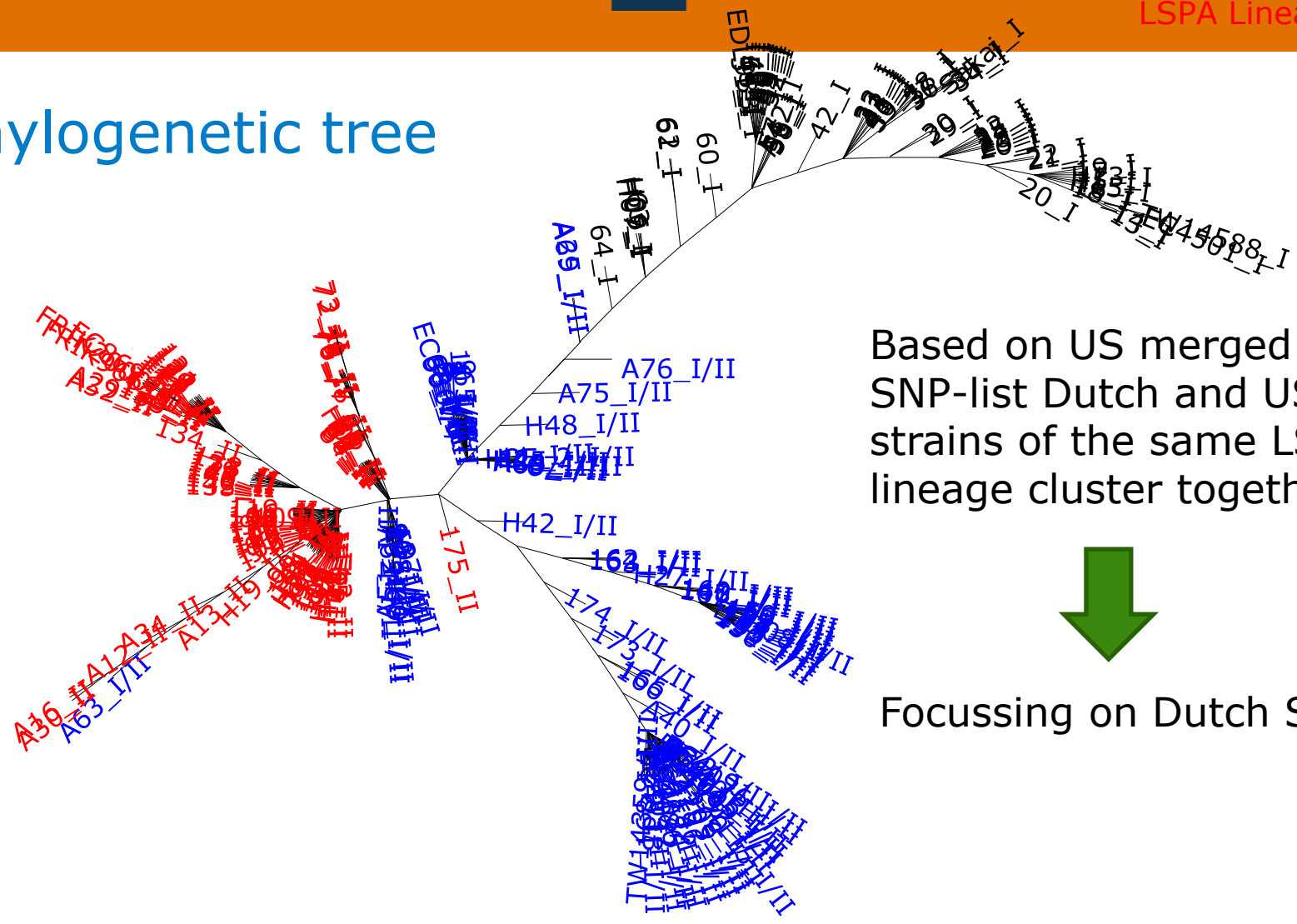
Merged to a list of 368 SNPs

Phylogenetic tree of Dutch (n=39) and US (n=183)
O157 strains based on merged SNP-list (Dutch SNPs yet
to be included)



LSPA Lineage I
LSPA Lineage I/II
LSPA Lineage II

Phylogenetic tree



Based on US merged
SNP-list Dutch and US
strains of the same LSPA
lineage cluster together

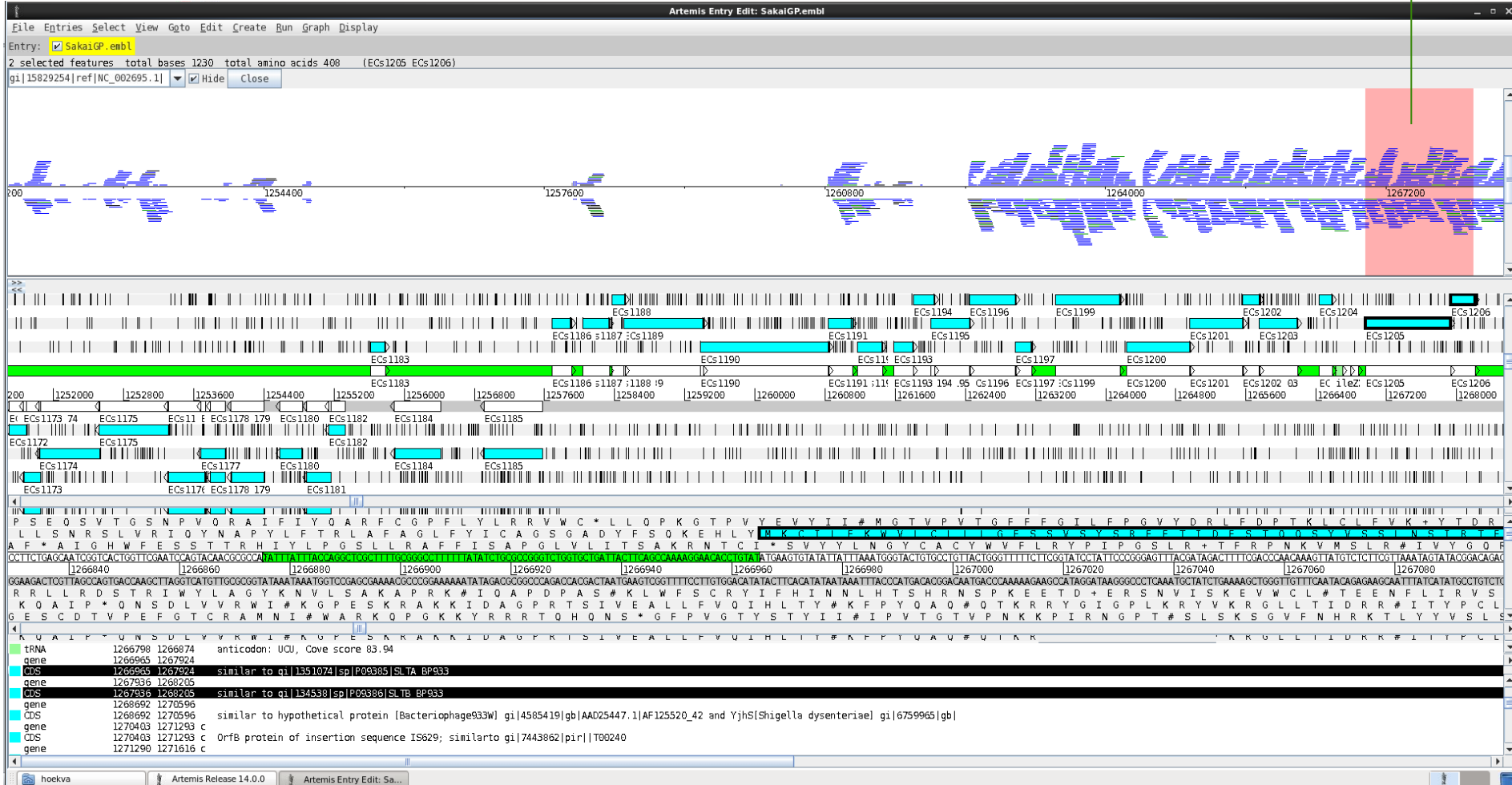
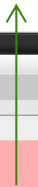


Focussing on Dutch SNPs



Insertion and deletions (*stx*-phages)

stx2a





Conclusions

- Whole-genome sequencing gives enormous opportunities in diagnostics, surveillance, evolutionary en epidemiological studies
- Likely to replace conventional typing methods
- The 100kGP is an unique opportunity to build an extensive database of foodborne pathogen genome sequences
- Comparative genomics is helpfull, but the phenotype is causing disease!
 - Small genetic differences may lead to large phenotypic differences! (survival/growth, stress resistance, virulence)

Thank you!

Angela van Hoek (RIVM)
Henk Aarts (RIVM)

Alex Bossers (WUR – CVI)
Frank Harders (WUR-CVI)

Jim Bono (USDA)

