Rapid whole genome sequencing investigation of a familial outbreak of E. coli O121:H19 with a sheep farm as the suspected source

Robert Söderlund[1,2], Cecilia Jernberg[3], Christine Källman[2], Ingela Hedenström[3], Erik Eriksson[2], Erik Bongcam-Rudloff[1], Anna Aspán[2]✉
[1]SLU Global Bioinformatics Centre, Swedish University for Agricultural Sciences, Uppsala, Sweden
[2]National Veterinary Institute, Uppsala, Sweden
[3]Swedish Institute for Communicable Disease Control, Solna, Sweden

Motivation and Objectives

Finding the source of an outbreak of zoonotic bacterial disease requires comparison of patient isolates to isolates found in food, animal or environmental samples. The current gold standard for this is pulsed field gel electrophoresis (PFGE), but multi-locus VNTR analysis (MLVA) has also been gaining popularity in recent years. Methods with too low discriminatory power can lead to false positives which are highly disruptive for food producers and costly for society, while excessively discriminatory markers can be affected by genetic changes which have occurred within an outbreak, causing false negatives. The recent emergence of benchtop high-throughput sequencing instruments has made whole genome sequencing (WGS) of bacteria a viable alternative to the currently available methods in terms of speed and cost, while potentially providing far more reliable genetic comparison data.

In 2012, verotoxin 2 (vtx2)-positive *E. coli* O121:H19 was isolated from a Swedish EHEC patient and two asymptomatic family members. A single isolate of O121:H19 vtx2$^+$ was found in samples taken from sheep at a farm where the patient had patted the animals. PFGE and MLVA analyses were inconclusive with limited variation both within the family and between the family and animal isolates. To resolve this and evaluate WGS as a tool for routine molecular epidemiology, the genomes of the three familial outbreak isolates, the isolate from the farm and two unrelated patient isolates were sequenced and compared.

Methods

MLVA was performed according to a generic *E. coli* protocol targeting a total of 10 loci. PFGE was performed according to PulseNet standard laboratory protocols (pulsenetinternational.org). For sequencing, crude lysates of bacteria harvested from agar plates were treated with Qiagen RNAseA and DNA extracted using a Qiagen EZ1 Biorobot. The DNA was quantified using the Qubit BR kit, and the integrity was checked by slab gel electrophoresis. Libraries for sequencing were prepared using the Nextera XT DNA Sample Preparation Kit with indexing, using 1 ng of extracted DNA as starting material. Libraries were verified with an Agilent Bioanalyzer HS DNA kit and run on a Illumina MiSeq instrument using the MiSeq V2 500 cycle run kit (2*250bp). The generated sequences were assembled using MIRA with contig filtering based on size and coverage. SNP discovery with the publicly available MT#2 draft genome sequence (GenBank AGTJ01000000) as reference was performed using MUMmer with further quality filtering, and SNP states were extracted from all contig sets using BLAST+ with the input and output treated by custom R scripts. Phylogenetic trees were drawn using the neighbor-net algorithm in SplitsTree. Regions of interest, e.g. known virulence factors in O121 and other types of EHEC as well as targets for multi locus sequence typing (MLST) and clustered regularly interspaced short palindromic repeats (CRISPR) typing were extracted from the contig sets. The time required to complete the full process was approximately one working week, with limited hands-on time.

Results and Discussion

The sequenced draft genomes had estimated sizes around 5.5 Mbp with an average coverage of 33x - 44x, and N50 values of 55 kbp - 96 kbp. Comparison of the contig sets identified 369 high quality SNPs in regions conserved in all included isolates. Analysis of the SNP data strongly supported a recent common origin for two of the three outbreak isolates, one of which was from the patient. These differed by a single SNP located in the coding region of an uncharacterized protein. However, the third outbreak isolate as well as the sheep isolate were distinct from each other and

Table 1. Characteristics of the sequenced O121:H19 isolates based on whole genome sequencing. *Symptomatic patient, reference for SNP similarity comparison

Source	wzx/wzy/ fliC	CRISPR	MLST	SNP similarity [%] *	Vero-toxins	Secondary virulence factors						
						eae	tir	hlyA	toxB	espP	efa1 OI-122	terB/ iha OI-48
Outbreak isolate 1*	O121:H19	CB8124	ST655	(100)	vtx2a	ε	+	+	+	+	+	+ / -
Outbreak isolate 2	O121:H19	CB8124	ST655	99.7	vtx2a	ε	+	+	+	+	+	+ / -
Outbreak isolate 3	O121:H19	CB8124	ST655	80.8	vtx2a	ε	+	+	+	+	+	+ / -
Sheep farm isolate	O121:H19	CB8124	ST655	62.4	vtx2a	ε	+	+	+	+	+	+ / -
Unrelated pat. isolate A	O121:H19	CB8124	ST655	94.3	vtx2a	ε	+	+	+	+	+	+ / -
Unrelated pat. isolate B	O121:H19	CB8124	ST655	67.2	vtx1a vtx2a	ε	+	+	+	+	+	+ / -
MT#2 (ref)	O121:H19	CB8124	ST655	45.0	vtx2a	ε	+	+	+	+	+	+ / -

from the isolate from the patient (Table 1). In fact, the outbreak patient isolate was far more similar to one of the unrelated reference isolates. Thus, there was no evidence that the sheep farm was the source of the infection.

Regions of interest were extracted from the generated sequences to produce backward compatible typing data traditionally produced by PCR or Sanger sequencing (Table 1). This analysis supported the PFGE and MLVA typing in indicating that all four isolates in the outbreak investigation belonged to the same clone of O121:H19. The prevalence of this clone in ruminants and asymptomatic humans in Sweden merits further investigation. In retrospect, sufficient typing data to exclude the farm as the source of the outbreak was produced by either PFGE or MLVA alone. However, the low total variation combined with the surprising outcome meant that WGS data was necessary for a definite answer. To see if multiple strains of O121:H19 were present at the farm, sampling of the animals was repeated at a later date, but on this occasion no O121 could be found. The source of the infection remains unknown.

The emergence of quick and affordable lab methodology combined with standardized data analysis workflows will see WGS taking an increasingly important role in the routine work of veterinary and public health authorities in the next few years.

Acknowledgements

This study was financially supported by the Swedish Civil Contingencies Agency. The Bo Segerman Group, SVA is thanked for sharing bioinformatics software and hardware. The authors also thank the staff at the SVA and SMI EHEC laboratories for excellent technical asssistance.