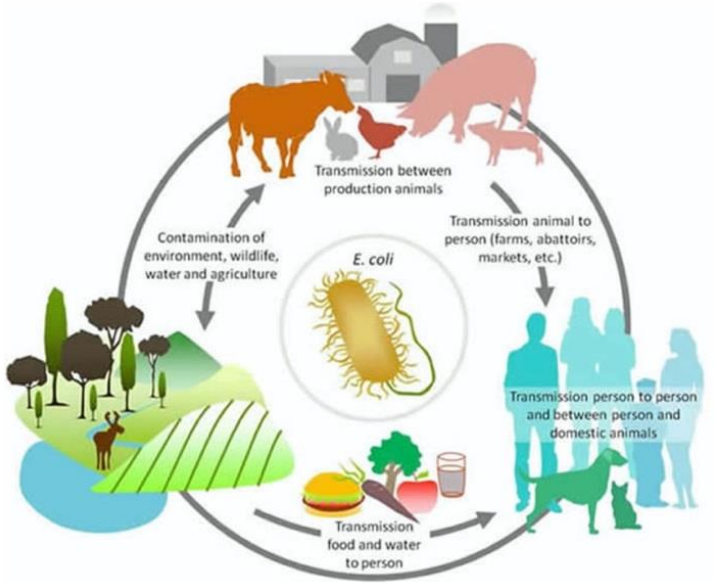
Evaluation of Whole Genome Sequencing (WGS) for Rapid Identification and Molecular Characterization of *Escherichia coli* Isolates from Food and Environment

Nicolas Lopez[1], Claudia Diaz[1], Guodong Zhang[2], and Li Maria Ma[1]
[1]Oklahoma State University, Stillwater, OK, [2]Food and Drug Administration, College Park, MD

Why do we care about *E. coli?*



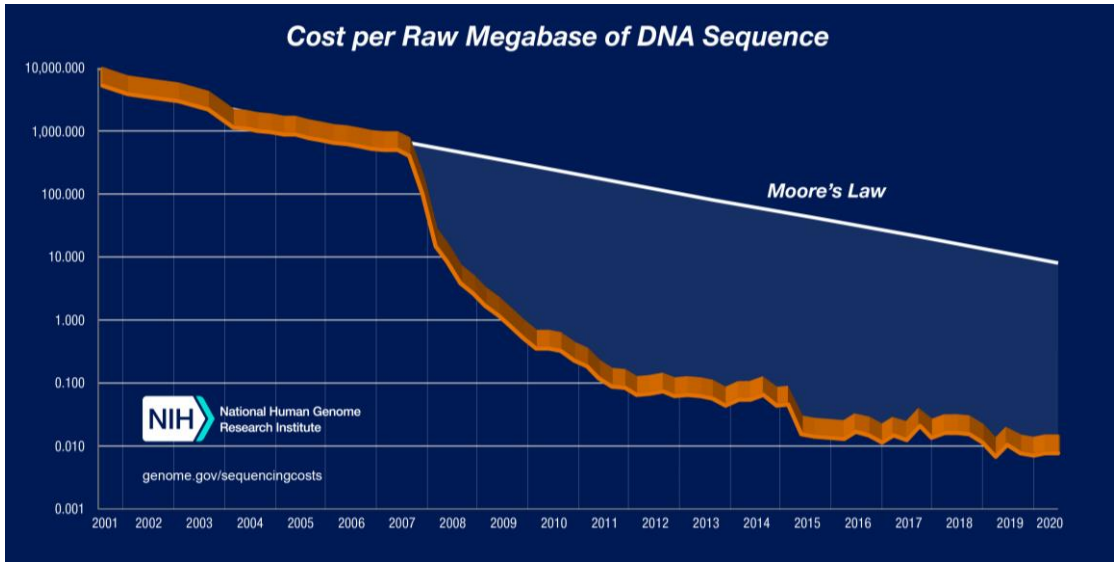
How to identify bad *E. coli*?

- Serotyping (O somatic antigen, H flagellar antigen).
- Identification of Virulence factors (Stx1, Stx2, eae, ehxA) and Antibiotic resistance genes.
- Phylogenetic analyses.

WGS era

Advantages

- Captures both large and small variants that might be missed with targeted approaches.
- Identifies potential causative variants for further follow-up studies of gene expression and regulation mechanisms.



Due to advances in technology it becomes cheaper and cheaper!

↓

CDC and FDA now use it

PulseNet USA
The National Molecular Subtyping Network for Foodborne Di

GenomeTrakr

Mostly used platforms for WGS:

illumina®

Oxford NANOPORE Technologies

PACBIO®

OSU

OBJECTIVE

This study was conducted to evaluate the performance of Minion sequencing for rapid identification and molecular characterization of *E. coli* isolated from food and environment.

METHODOLOGY

Isolation and WGS of *E. coli*.

- ***E. coli* isolation from pecan orchards (soil, cattle feces, and fallen in-shell pecans):** FDA Bacteriological Analytical Manual (FDA-BAM) modified protocol for detection and isolation of Shiga Toxin-producing *E. coli.*
- **Screening by multiplex PCR** (Stx1, Stx2, eae genes)
- **WGS:** 11 isolates were sequenced by:
  ONT Minion: Rapid Barcoding Sequencing kit (SQK-RBK004).
  NextSeq Illumina: Nextera XT DNA sample preparation kit with the NextSeq®500 high output kit.

Evaluation of sequencing reads yield and time-subsampling.

- **ONT Minion:** Reads a Q score ≥7 were filtered with EPI2ME (BC: Guppy). Passed reads were demultiplexed with Porechop v. 0.2.4, and finally filtered with Filtlong v. 0.2.0 using a Q score cutoff of 9. A custom Perl script was used to sub-sample the reads based on when they were obtained.
- **Illumina:** Trimmomatic v. 0.32 was used to remove barcodes and to trim the sequences with a window size of 4 and a Q score cutoff of 20.
- **Species confirmation:** Kraken2 (default mode) with a confidence threshold of 0.05 against a custom RefSeq database (Archaeal, bacterial, viral and plasmids) created on January 2020.

Rapid identification of serotype and virulence determinants using sequencing reads

Serotype prediction: Serotypefinder v. 2.0. (default mode)
Virulence determinants Databases: Stx1, Stx2 and eae sequeneces were obtained from reference genomes harboring specific variants for each gene (Genbank).
Mapping reads to determinants: KMA against the built DBs (minimun alignment score 0.75, gap open -5, gap extend -1, penalty -3, reward 1)

Evaluation of genome continuity

Genome assembly: Flye v. 2.0. for ONT Minion
Hybrid assemblies (two different approaches): Flye assembled genomes were aligned to Illumina reads with Bowtie2, then were polished with Pilon. The other approach was using Unicycler.
Genome assemblies quality assessment: QUAST, BUSCO.
Statistical analysis: N50, size of the longer contig, and the entire number of contigs were used in a distributional independent multivariate analysis.

Characterization of Virulence factors (VFs) and Antibiotic Resistance (AR) Genes.

Genome annotation: PROKKA v. 1.14.6. (default mode)
Characterization of VFs and AR genes: BLASTp search (e-value: 1e-6, percentage of identity: 95% and query coverage: 85%) against the Comprehensive Antibiotic Resistance Database (CARD) and Virulence Factor Database (VFDB).

RESULTS AND DISCUSSION

Species confirmation

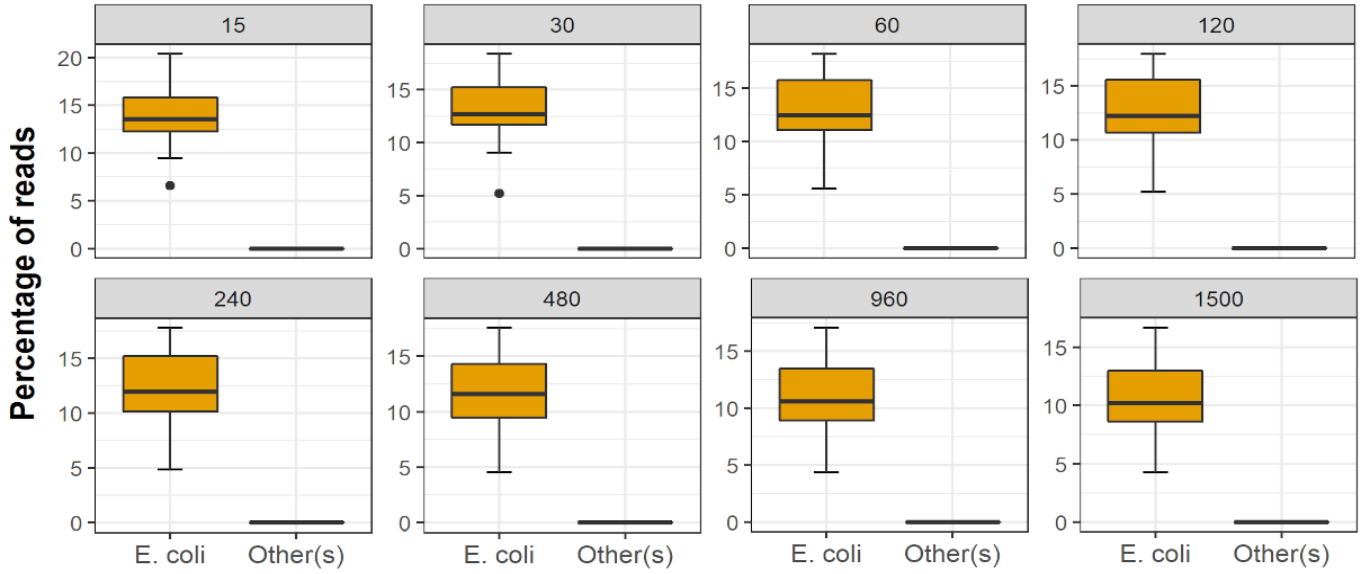


Fig 1. Species confirmation in the sub-sampled reads in the lapse of 15, 30, 60, 120, 240, 480, 960 and 1500 minutes.

Serotype prediction:

Isolate	15	30	60	120	240	480	960	1500	Illumina
G1BLF1_5	-:-	-:-	-:-	-:-	O168:-	O168:-	O168:-	O168:-	O168:H8
G1BLF2_1	-:-	-:H8	-:H8	-:H8	O168:H8	O168:H8	O168:H8	O168:H8	O168:H8
G1M0S3_4	-:-	-:-	-:-	O108:-	O108:-	O108:-	O108:-	O108:-	O108:-
G1M4F3_31	-:-	-:-	-:-	-:-	O153/O178:-	O153/O178:-	O153/O178:-	O153/O178:-	O153/O178:H19
G4M0F1_1	-:-	-:-	-:-	-:-	-:-	O8:-	O8:-	O8:-	O8:H7
G4M0F2_14	-:-	-:-	-:-	-:-	-:-	-:-	-:-	-:-	O91:H21
G5BLF1_1	-:-	-:-	O157:-	O157:-	O157:-	O157:-	O157:-	O157:-	O157:H7
G5BLF3_3	-:-	-:-	O130:-	O130:-	O130:H11	O130:H11	O130:H11	O130:H11	O130:H11
G5BLF3_8	-:-	-:-	-:-	-:-	O109:-	O109:-	O109:-	O109:-	O109:-
G5M2P3_1	-:-	-:-	O91:-	O91:-	O91:-	O91:-	O91:H21	O91:H21	-:-
G5M4F2_1	-:-	-:-	-:-	-:-	-:-	O103:-	O103:-	O103:H2	O103:H2

Table 1. Prediction of O and H antigen from subsampled reads.

Rapid identification of serotype and virulence determinants using sequencing reads

Isolate	Single Colony PCR			Sequencing reads								
	Stx1	Stx2	eae	15	30	60	120	240	480	960	1500	Illumina
G1BLF1_5	-	+	-	-	-	-	-	-	-	-	-	Stx2 variant D
G1BLF2_1	-	+	-	-	-	Stx2 variant D	Stx2 variant D	Stx2 variant D	Stx2 variant D	Stx2 variant D	Stx2 variant D	Stx2 variant D
G1M0S3_4	-	-	+	-	-	-	-	-	-	-	eae β1	eae β1
G1M4F3_31	+	-	-	-	-	-	-	-	-	-	-	Stx1 S. dysenteriae
G4M0F1_1	-	-	-	-	-	-	-	-	-	-	-	-
G4M0F2_14	-	-	-	-	-	-	-	-	-	-	-	-
G5BLF1_1	-	+	+	-	Stx2 variant A	Stx2 variant A	Stx2 variant A	Stx2 variant A	Stx2 variant A	Stx2 variant A	Stx2 variant A	Stx2 variant A
G5BLF3_3	+	-	-	-	-	Stx1 S.dysenteriae	Stx1 S.dysenteriae	Stx1 S.dysenteriae	Stx1 S.dysenteriae	Stx1 S.dysenteriae	Stx1 S.dysenteriae	Stx1 S. dysenteriae
G5BLF3_8	+	-	-	-	-	-	-	-	Stx1 S.dysenteriae	Stx1 S.dysenteriae	Stx1 S.dysenteriae	Stx1 S. dysenteriae
G5M2P3_1	-	-	-	-	-	-	-	-	-	-	-	-
G5M4F2_1	+	+	+	-	-	-	-	-	Stx1 variant A, Stx2 variant A	Stx1 variant A, Stx2 variant A	Stx1 variant A, Stx2 variant A, eae ε1	Stx1 variant A, Stx2 variant A, eae ε1

Table 2. Multiplex PCR and mapping reads results for the genes Stx1, Stx2 and eae.

Evaluation of genome continuity

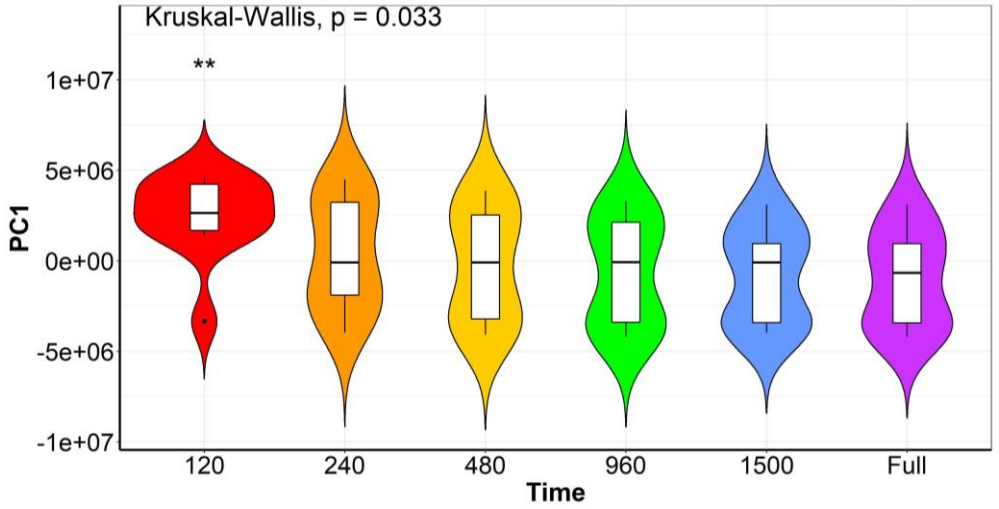


Fig 2. Distributional Independent Multivariate analysis of genome continuity from assemblies obtained using ONT Minion over time.

Characterization of Virulence factors (VFs) and Antibiotic Resistance (AR) Genes.

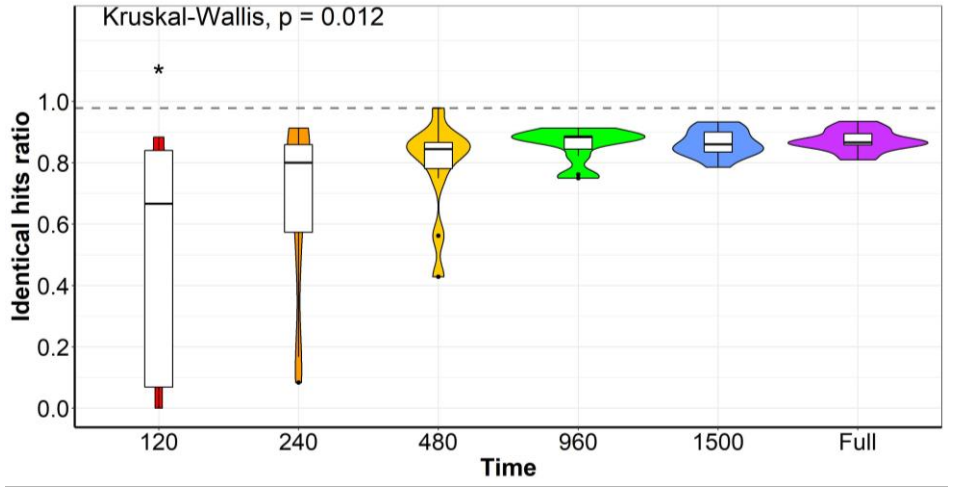


Fig 3. CARD Identical hits ratio between subsampled genomes and hybrid genomes

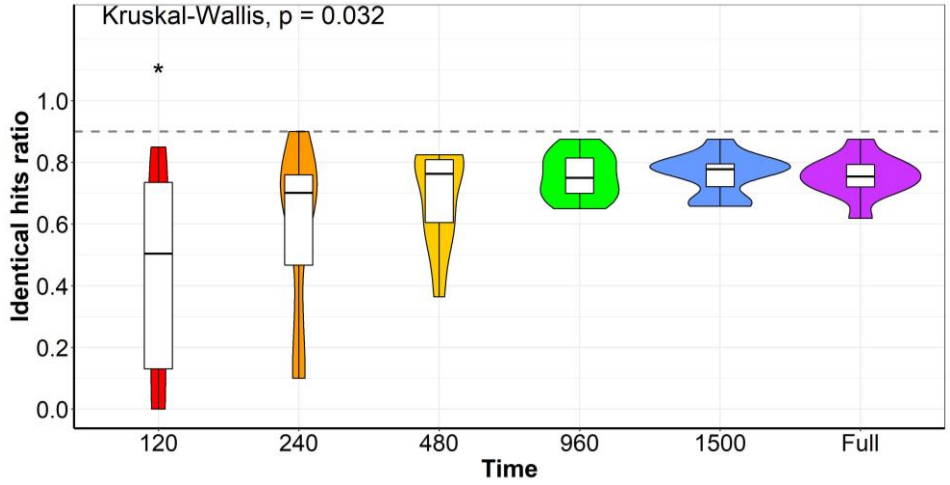


Fig 4. VFDB Identical hits ratio between subsampled genomes and hybrid genomes

Full reads dataset could only obtain an average of 87.25% and 75.51% of the hits acquired from the hybrid assemblies using the CARD and VFDB, respectively

CONCLUSIONS

Comparing to Illumina, WGS by ONT Minion offers the potential of “real-time” identification and characterization of possible foodborne bacterial pathogen isolates.

- ❑ Species level identification was achieved at 15 min of sequencing run
- ❑ O antigen prediction was achieved after 8 hrs of sequencing run
- ❑ Variants of the virulence genes were detected 24 hours of sequencing run
- ❑ ONT Minion’s slow precision limits its ability to characterize fully factors that may be relevant to the *E. coli* assessment.