Illustration Courtesy by Arwa A. Abbas

PEDIATRIC CLOSTRIDIUM DIFFICILE ISOLATES FROM PHILADELPHIA ANALYZED USING MULTIPLE MINION SEQUENCING

asm microbe 2018 | SUNDAY – 1019 06/10/2018

PennCHOP MICROBIOME PROGRAM

Chunyu Zhao[1], Kyle Bittinger[1], Guanxiang Liang[1], Lisa Mattei[1], Alissa C Galgano[1], Casey E Hofstaedter[1], Sandra Montgomery[1], Sylvester Roundtree[1], Robert Baldassano[1], and Frederic D. Bushman[2]*

[1] Division of Gastroenterology, Hepatology and Nutrition, The Children's Hospital of Philadelphia [2] Department of Microbiology, University of Pennsylvania

E-mail: zhaoc1@email.chop.edu
Phone: 1-215-317-6133

Children's Hospital of Philadelphia

Penn University of Pennsylvania

ABSTRACT

Clostridium difficile infection (CDI) is a frequent sequela of antibiotic use and is common in the inflammatory bowel disease (IBD) population, even among subjects without antibiotic exposure. While much is known about C. difficile genomics in adults, it is less well investigated for pediatric IBD subjects. Here, we report whole genome sequencing of C. difficile isolates from 13 pediatric IBD patients with CDI using both Oxford Nanopore sequencing and Illumina HiSeq 2500 sequencing.

The clinical laboratory of The Children's Hospital of Philadelphia purified C. difficile strains from infected pediatric IBD subjects. Oxford Nanopore MinION sequencing was used to generate long reads for 13 isolates at the PennCHOP Microbiome Program. We performed basecalling, adapter sequence removal, chimera detection, and hybrid assembly using our in-house bacterial genome assembly pipeline, yielding near-complete genome assemblies.

BIOINFORMATICS PIPELINE - NANOFLOW

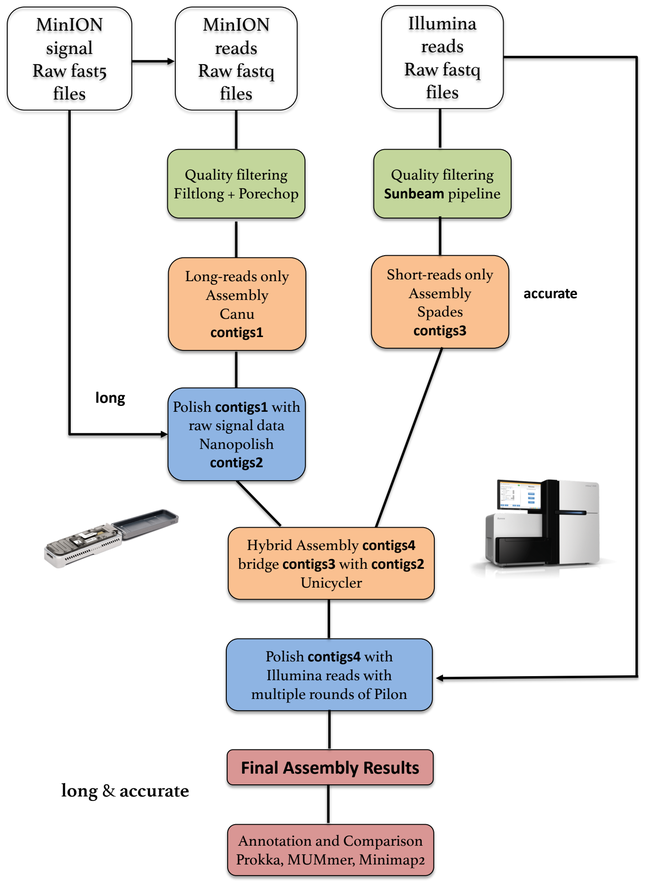


Figure 1: Nanoflow - hybrid genome assembly and evaluation pipeline. We found that assembling the genome directly from ONT long reads data resulted in less accurate gene sequences and interrupted genes. Thus, our hybrid assembly pipeline created contigs using Illumina and then stitch them together with nanopore reads. Nanoflow is implemented in Snakemake and available on https://github.com/zhaoc1/nanoflow.

RUN SUMMARY AND EVALUATION

Figure 2: Nanopore read length distribution and assembly accuracy estimation. Raw reads and assembled contigs were mapped to reference genome CD630 using Minimap2, and the sequence similarity was calculated based on the aligned region. The figures were generated by the Nanoflow pipeline.

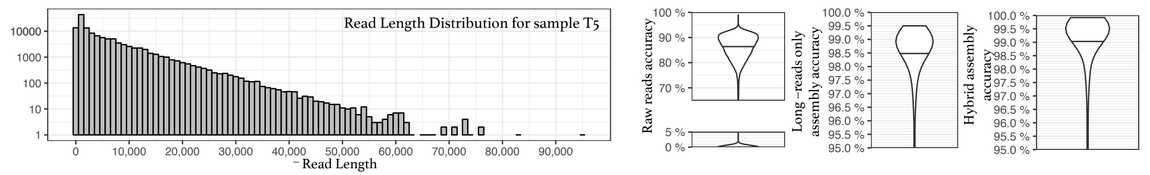


Table I: Bioinformatics report of all three nanopore runs

Sample ID	Run Number	Barcode	Study Group	Coverage	Read Counts	Average Read Len	Largest Read	ST	MLST clades**	PaLoc region***
T3	run9	barcode01	IBD C. diff	198.3	176479	5169.07	101202	2	1	19.8
T4	run9	barcode02	IBD C. diff	131.2	196654	3068.12	93243	110	1	20.4
T5	run9	barcode03	IBD C. diff	132.8	128700	4747.73	95331	43	1	19.8
T1*	run10	barcode04	IBD C. diff	194.6	187215	4781.89	122111	188	**2**	**21.8**
T6	run10	barcode05	IBD C. diff	164.8	118327	6407.34	89287	54	1	19.8
T7	run10	barcode06	IBD C. diff	99.9	78650	5842.22	124984	43	1	19.8
T9	run10	barcode07	IBD C. diff	295.5	166992	8136.31	164884	185	1	19.8
T10	run10	barcode08	IBD C. diff	109.0	90723	5527.03	115938	43	1	19.8
T8	run13	barcode07	IBD C. diff	53.4	47942	5120.57	104390	10	1	19.8
T11	run13	barcode08	IBD C. diff	103.1	79676	5951.86	100076	10	1	19.8
T12	run13	barcode09	IBD C. diff	126.6	80388	7244.45	119513	42	1	19.8
T14	run13	barcode10	Onco C. diff	135.6	79096	7887.91	125401	14	1	19.8
T16	run13	barcode12	Onco C. diff	86.1	92662	4273.20	69055	2	1	19.8

** Each isolate was assigned to one of six sequence types using the multilocus sequence typing (MLST) database of 7 housekeeping genes.

* T1: only sample with positive binary toxin genes, and belong to hypervirulent clade 2.

*** The symptoms of CDI are caused by toxins expressed from its 19.6 KB pathogenicity locus (PaLoc). Our genomic analysis revealed that this region was present in all eight isolates.

PLASMID IDENTIFICATION

Sample	Contig	Length	Depth*
T3	1	4262003	1.00x
T3	2	296440	0.89x
T3	3	78329	1.01x
T3	4	5006	2.09x
T4	1	4184498	1.00x
T5	1	4190414	1.00x
T5	2	62461	1.20x
T5	3	12488	5.16x
T5	4	6935	5.06x
T8	1	4207905	1.00x
T8	2	47546	2.70x
T8	3	12523	8.69x
T8	4	6760	7.32x
T11	1	4205065	1.00x
T11	2	132448	1.04x
T11	3	47546	2.17x
T11	4	12526	5.90x
T11	5	6795	4.34x
T12	1	4100444	1.00x
T14	1	4150459	1.00x
T14	2	77774	1.06x
T14	3	12526	6.71x
T16	1	4204038	1.00x
T1	1	4259604	1.00x
T1	2	9554	5.17x
T1	3	6865	6.55x
T1	4	6830	10.19x
T6	1	4321901	1.00x
T7	1	4231134	1.00x
T7	2	42249	2.60x
T7	3	12488	5.59x
T7	4	6830	6.41x
T9	1	4121924	1.00x
T10	1	4187195	1.00x
T10	2	46053	2.18x
T10	3	12488	5.04x

Table II: Summary of assembly results. In addition to either the complete circular or near-to-complete linear chromosome assembly, we also assembled some small circular DNA sequences. To identify whether the small assemblies are plasmids, we first annotated the assembly using Prokka, and then used Blastn to identify matching genes to all eleven existing C. difficile plasmids.

Circular DNA
Replication protein
Plasmid - like

* The depth was calculated relative to the largest contigs.

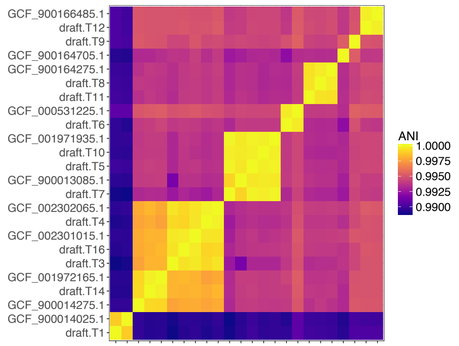
ANTIBIOTIC RESISTANCE GENES

Table III: Antibiotic resistance genes annotation for T6. We used protein BLAST to identify homologous genes in the CARD databases (ident > 75%)

Gene Name	Ident	Card Gene Coverage	Start Pos	End Pos	Gene Annotation
draft.T6_00451	100.0	1	402599	404038	bifunctional aminoglycoside N-acetyltransferase **AAC(6')-Ie**/aminoglycoside O-phosphotransferase **APH(2'')-Ia**
draft.T6_00458	97.8	0.98	408257	408994	23S rRNA (adenine(2058)-N(6))-methyltransferase **Erm(B)**

HEATMAP OF AVERAGE NUCLEOTIDE IDENTITY

Figure 3: Pairwise genome comparisons based on ANI value. Reference Genome Candidates: the nearest neighbors of each draft genome on the phylogenetic tree.



PHYLOGENETIC AND PANGENOME ANALYSIS

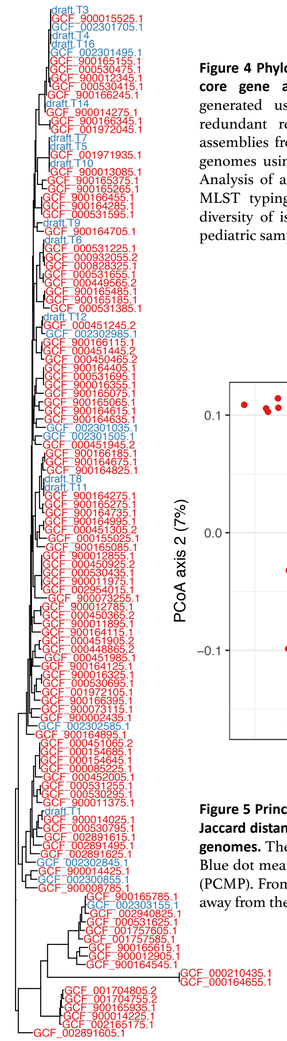


Figure 4 Phylogenetic tree of 114 C. difficile strains using 2139 core gene alignments. A maximum likelihood tree was generated using FastTree. For selection of the 114 non-redundant representative strains, we took all C. difficile assemblies from the RefSeq database, filtered out low-qualify genomes using CheckM, and clustered with 99% ANI cutoff. Analysis of an expanded set of gene sequences recapitulated MLST typing results, and demonstrated the broad genetic diversity of isolates from pediatric IBD patients. Blue means pediatric samples.

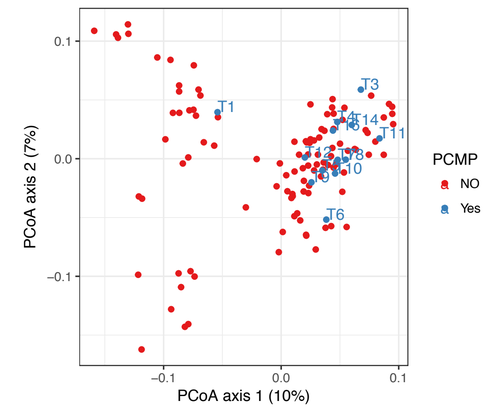


Figure 5 Principle Coordinate Analysis (PCoA) plot based on the Jaccard distance of the presence/absence of core and accessory genomes. The Pangenome analysis was carried out using Roary. Blue dot means samples from PennCHOP Microbiome Progrm (PCMP). From the PCoA plot, we could see that T1 is clustered away from the other 12 samples

SUMMARY

- ✓ **We described a hybrid genome assembly pipeline using Nanopore and Illumina sequencing data, and Nanoflow was implemented in Snakemake.**
- ✓ **For 13 C. difficile isolates, we assembled the draft genomes near-to complete, for both chromosome and plasmids; and annotated the PaLoc region and binary toxin genes.**
- ✓ **C. difficile in pediatric IBD in Philadelphia was dominated (12/13) by clade 1, with 1/13 isolates from the hypervirulent clade 2.**
- ✓ **Analysis of the core genomes recapitulated MLST typing results, and demonstrated the broad genetic diversity of isolates from pediatric IBD patients.**