

Detection of mixed-strain infections by FACS and ultra-low input genome sequencing

Mária Džunková^{a,b,c,d}, Andrés Moya^{a,b,c}, Xinhua Chen^e, Ciaran Kelly^e, and Giuseppe D'Auria^{a,b,c,f}

^aDepartment of Genomics and Health, Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO-Public Health), València, Spain; ^bCIBER in Epidemiology and Public Health (CIBEResp), Madrid, Spain; ^cInstitute for Integrative Systems Biology (I2SysBio), The University of Valencia and The Spanish National Research Council (CSIC)-UVEG, València, Spain; ^dAustralian Centre for Ecogenomics, The University of Queensland, St Lucia, Australia; ^eDivision of Gastroenterology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, USA; ^fSequencing and Bioinformatics Service of the Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO-Public Health), València, Spain

ABSTRACT

The epidemiological tracking of a bacterial outbreak may be jeopardized by the presence of multiple pathogenic strains in one patient. Nevertheless, this fact is not considered in most of the epidemiological studies and only one colony per patient is sequenced. On the other hand, the routine whole genome sequencing of many isolates from each patient would be costly and unnecessary, because the number of strains in a patient is never known *a priori*. In addition, the result would be biased by microbial culture conditions.

Herein we propose an approach for detecting mixed-strain infection, providing *C. difficile* infection as an example. The cells of the target pathogenic species are collected from the bacterial suspension by the fluorescence activated cell sorting (FACS) and a shallow genome sequencing is performed. A modified sequencing library preparation protocol for low-input DNA samples can be used for low prevalence gut pathogens (< 0.1% of the total microbiome). This FACS-seq approach reduces diagnostics time (no culture is needed) and may promote discoveries of novel strains. Methodological details, possible issues and future directions for the sequencing of these natural pan-genomes are herein discussed.

ARTICLE HISTORY

Received 18 April 2018
Revised 30 August 2018
Accepted 11 September 2018

KEYWORDS

FACS-seq; *C. difficile*; mixed-strain infection; low-input DNA sequencing; epidemiology

Introduction

Whole genome sequencing technologies have continuously become more efficient, and thus they have the potential to become a common diagnostic tool allowing accurate identification of pathogenic strains. The most habitual approach in the epidemiological studies is to sequence one colony per patient. However, recent studies showed that if only one colony per patient is screened, only as few as 25 % of cases can be linked to the previously isolated strains within the same hospital.^{1–3} This discrepancy is caused by the presence of multiple strains in a single patient. These mixed-strain infections can impair the exclusion of transmission and determination of the outbreak origin.

As the number of strains in a patient is never known *a priori*, the routine whole genome sequencing of many isolates from each patient would be costly and unnecessary. In addition, as different strains often have very different growth requirements, it may be

impossible to isolate all the different strains from one patient. One option for obtaining genome sequences of all strains in their natural proportion without culture would be the metagenomic sequencing of the biological sample. However, such an approach would not be feasible for pathogens which form very small portion of the total microbiome.

Clostridium difficile infection (CDI) is one of the numerous diseases for which co-infection by multiple strains is often reported.^{4–6} CDI has nosocomial origin and its severity and mortality is increasing.⁷ *Clostridium difficile*, recently renamed *Clostridioides difficile*,⁸ forms only 0.0001–0.1 % of the infected microbiome.^{9,10} Its strains have different growth rate and culture medium requirements.¹¹ As the isolation of *C. difficile* is usually done with a single culture media, the mixed-strain infections often remain undetected.

Recovery of multiple *C. difficile* genomes directly from the metagenomic sequences would

be very difficult because of its low proportion in the gut microbiome. However, the fluorescence activated cell sorting (FACS) can be used for enrichment of the target bacterial species prior to sequencing.¹² While this FACS-seq method is mostly used for genomic characterization of uncultured bacteria, it has not been used in clinical microbiology yet. However, FACS-seq can have high importance in studies focused on low-prevalent pathogenic bacterial species whose strains have very different culture requirements.

Facs-seq of *C. difficile* cells

As an example for this commentary/view article, we performed the FACS-seq of *C. difficile* with one faecal sample from a patient who was hospitalized in May 2014 at the Beth Israel Deaconess Medical Center of the Harvard Medical School, Boston, MA, USA. The patient was diagnosed to be CDI positive by routine Illumigene assay (Meridian Biosciences). The study was approved by the institutional review board of BIDMC, and a written informed consent was obtained from the participant. The proportion of *C. difficile* in the patient's

faecal sample was 0.1% (quantified by Qiagen qPCR kit # BPID00110AF targeting *C. difficile* specific 16S rDNA sequences).

The 16S rDNA probes were used for hybridization of *C. difficile* cells present in the faecal bacterial suspension as described by Novakova *et al.*¹³ A sample containing 90 % of faecal bacteria from the patient and 10% of cultured *C. difficile* cells (ATCC 9689) was used as an additional FACS sorting control. The FACS was performed on S3 cell sorter (Bio-Rad). The first cell selection gate was set on the red 640 nm channel (FL4) to discard organic particles present in faeces which are not fluorescent when stained by the DNA stain SYTO 62. The cells containing DNA were then visualized on the next bi-plots representing cell size (forward scatter) on the x-axis and green fluorescence on the FL1 channel (488 nm) on the y-axis (Figure 1). The final sorting gate was set by comparing the non-hybridized faecal samples with the *C. difficile* positive control.

The amount of extracted DNA from the 10,000 FACS-collected *C. difficile* cells was undetectable by Picogreen assay (Life Technologies). It means that the sample contained less DNA amount than

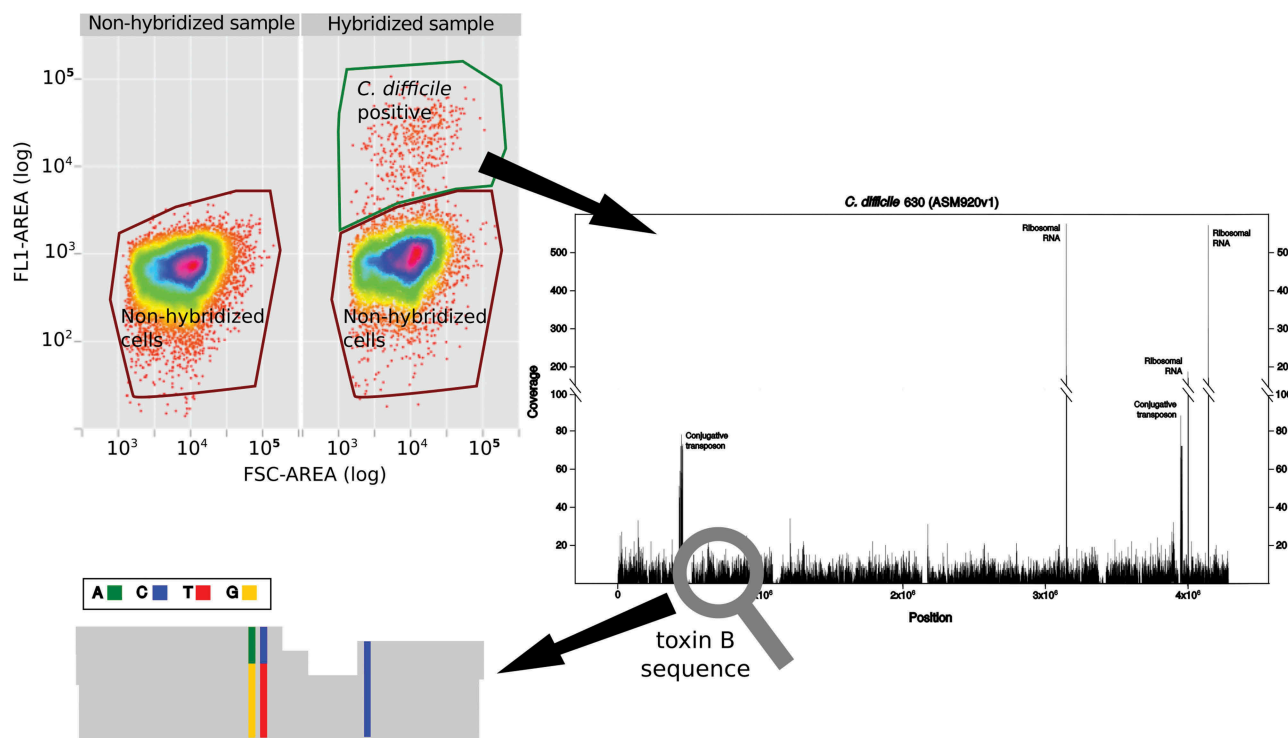


Figure 1. Work-flow chart of the detection of mixed-strain infections by FACS and ultra-low input genome sequencing.

the 50 ng required by the standard Illumina Nextera XT library preparation protocol version from the year 2014 (Ref. FC-121-1030) and it was also less than 1 ng required by the current version of the Nextera XT tagmentation protocol (Document number #15031942). Therefore, for the sequencing library preparation a modified protocol adjusted for ultra-low input DNA samples was used. The volume of the Nextera XT tagmentation mixture was reduced five times, while the volume of DNA sample was increased accordingly. The sample was sequenced with MiSeq® Reagent Kit v3.

The sequences have been deposited in European Nucleotide Archive database with study accession number PRJEB20472. Quality filtered reads were mapped to the *C. difficile* reference strain 630 (GenBank assembly accession GCA_000009205.1) by *Bowtie2* using the default parameters for “very-fast” mapping of only the most similar reads, requiring that the entire read align from one end to the other.¹⁴ The resulting mapping was inspected by the Integrative Genomics Viewer software.¹⁵ The coverage plots of the reference genome revealed that there was a high sequence variability in the areas of ribosomal genes and conjugative transposones, due to contamination by few other bacterial species in the sorted sample.

The further analysis of the single nucleotide polymorphism sites (SNPs) focused only on the toxin B gene which is specific to virulent *C. difficile* strains only.⁷ If other bacterial species are investigated by the same approach, a subset of genes specific to that particular species or its core genome, should be taken into account. In our case, the toxin B region (7,098 bp) contained three SNP variants. The obtained shotgun reads contained in the position 789,657 bp either the reference guanine or a novel adenine. In the position 789,660 bp it had either the reference thymine or a novel cytosine, and in the position 789,681 bp a cytosine (Figure 1). These nucleotide changes were synonymous substitutions. In order to find out whether these variants have been previously sequenced by other studies, the toxin B sequences containing the novel non-reference variants were aligned to the “nr” database of NCBI. The sequences matched with 100 % identity the sequence of the toxin B sequence types B07 and B08 isolated in China in

September. 2014¹⁶ However, as no whole genome sequence is available for this Chinese isolate, we cannot conclude that the patient was infected by this particular strain.

Future directions

The current high-throughput sequencing and cell sorting technologies provide new ways to study the natural genetic diversity of strains of selected pathogens that are present in a single patient at a very low abundance. Current improvements of the sequencing library preparation protocols allow the sequencing of samples containing only a few hundreds bacterial cells. The FACS-seq based approach presented here recovers multiple strains at their natural proportion and so facilitates epidemiological tracking of an outbreak. In addition, it may lead to discovery of novel pathogenic strains with unusual culture conditions requirements. Moreover, as there is no need to culture the cells of interest, this method can reduce diagnostic time, especially in cases of pathogens which produce colonies after several days or weeks of culture, such as *Legionella pneumophila* or *Mycobacterium tuberculosis*.¹⁷

In this pilot study, the proportion of *C. difficile* in the patient’s faecal microbiome was as low as 0.1%. We performed a shallow genome sequencing of 10,000 *C. difficile* cells, which was, however, sufficient for detection of possible genetic variants among the multiple strains in one patient’s sample. According to Li *et al.*,¹⁸ a reference genomes coverage of as few as 4x may be sufficient to detect differences among strains in metagenomes, however, the required minimal coverage may vary in distinct projects depending on the microbial community complexity and species similarity. After the presence of multiple SNPs in a sample are found in low genome coverage, the samples can be then sequenced more deeply to obtain a genome coverage of at least a hundred times. Such a high coverage would allow recovery of nearly complete genomes of multiple strains. According to Rinke *et al.*,¹⁹ a successful Illumina sequencing run can be achieved with as few as 100 femtograms of DNA (the DNA amount contained in 100–1000 bacterial cells) using a modified library preparation protocol. Confirmation of multiple strains in

single patients by the FACS-seq approach may help to resolve links in an epidemiological network of disease outbreak detection obtained by common bioinformatic tools.²⁰

However, it is important to note that despite a FACS-separated sample being enriched for the target pathogen species, it may also be contaminated by other bacterial species. Such contamination does not necessarily mean that the hybridization probes were non-specific. If the taxonomic composition of the contaminating species in the FACS-separated sample is the same as the taxonomic composition of the original non-FACS-separated sample, the source of the contamination may be the flow cytometer itself. The contamination of the FACS equipment is quite common and therefore, a rigorous cleaning procedure of the FACS equipment should be performed before the cell sorting.²¹ The reads belonging to the contaminating species may map weakly to the target pathogen genome forming peaks with extremely high coverage (as occurred in the present study in the 16S ribosomal gene regions and conjugative transposomes). The SNPs observed in these regions should not be considered for further analysis.

The potential contamination issues may be solved by metagenome binning. The shotgun reads are assembled and then mapped back to the assembled contigs. The characteristics (e.g. coverage, k-mer frequencies and GC content) of the contigs are compared and then clusters of similar contigs (so called bins) are formed.²² Bins belonging to the target pathogenic strains should be easily distinguished from the contaminating species. The genetic diversity of these strains may be assessed, including presence of plasmids or mobile genetic elements.

It is also important to mention that the fluorescently hybridized cells can be distributed by the flow cytometry equipment one by one into 384 well plates, so their whole genomes can be analyzed separately. In this approach, *Phi* polymerase adds random oligomers to the whole DNA molecule which results in amplification of the whole genome hundreds of times.²³ Despite improving chemistry of the whole genome amplification, some regions of the bacterial genome can be accidentally omitted and the amplified sequences may contain polymerase proofreading errors. However, single-cell resolution can provide important information on large genomic differences between

strains, such as mobile genetic elements.²⁴ Therefore, single-cell genomic approach could be used for confirmation of unusual genomic rearrangements detected by sequencing of bulks of cells.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank John Tigges and Vasilis Toxavidis from Flow Cytometry Core Facility of Beth Israel Deaconess Medical Center (BIDMC) of Harvard Medical School (MA, USA) for helping us with setting-up of the flow cytometry sorting. We also thank Kelsey Shields and Joshua Hansen from the Department of Gastroenterology of BIDMC for the help with sample collection and Dr. Nuria Jiménez from sequencing laboratory of FISABIO – Public Health, Valencia, Spain for the sequencing of the samples. We also want to thank Lauren Barker for English language corrections.

Funding

This work was supported by the Boehringer Ingelheim Fonds [Travel Grant 2014]; Generalitat Valenciana [PrometeoII/2014/065]; Instituto de Salud Carlos III [CP09/00049]; Instituto de Salud Carlos III [PIE14/00045]; Ministerio de Economía y Competitividad [SAF 2012-31187]; Ministerio de Economía y Competitividad [AC15/00022]; Ministerio de Economía y Competitividad [SAF2015-65878-R]; Ministerio de Economía y Competitividad [SAF2013-49788-EXP]; Ministerio de Educación, Cultura y Deporte [FPU2010].

ORCID

Mária Džunková  <http://orcid.org/0000-0002-1765-0697>

References

1. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L, et al. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. *BMJ Open*. 2012;2(3): e001124+. doi:10.1136/bmjopen-2012-001124.
2. Didelot X, Eyre D, Cule M, Ip C, Ansari M, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty EM, et al. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol*. 2012;13(12):R118+. doi:10.1186/gb-2012-13-12-r118.

3. Gan M, Liu Q, Yang C, Gao Q, Luo T. Deep whole-genome sequencing to detect mixed infection of *Mycobacterium tuberculosis*. *PLoS One*. 2016;11(7):e0159029. doi:10.1016/B978-0-12-407863-5.00001-0.
4. van Den Berg RJ, Ameen HA, Furusawa T, Claas EC, van der Vorm ER, Kuijper EJ. Coexistence of multiple PCR-ribotype strains of *Clostridium difficile* in faecal samples limits epidemiological studies. *J Med Microbiol*. 2005;54(2):173–179. doi:10.1099/jmm.0.45825-0.
5. Eyre DW, Cule ML, Griffiths D, Crook DW, Peto TE, Walker A, Wilson DJ. Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in *Clostridium difficile* transmission. *PLoS Comput Biol*. 2013;9(5):e1003059+. doi:10.1371/journal.pcbi.1003059.
6. Tanner HE, Hardy KJ, Hawkey PM. Coexistence of multiple multilocus variable-number tandem-repeat analysis subtypes of *Clostridium difficile* PCR ribotype 027 strains within fecal specimens. *J Clin Microbiol*. 2010;48(3):985–987. doi:10.1128/JCM.02012-09.
7. Rupnik M, Wilcox MH, Gerding DN. *Clostridium difficile* infection: new developments in epidemiology and pathogenesis. *Nat Rev Microbiol*. 2009;7(7):526–536. doi:10.1038/nrmicro2164.
8. Lawson PA, Citron DM, Tyrrell KL, Finegold SM. Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prévot 1938. *Anaerobe*. 2016;40:95–99. doi:10.1016/j.anaerobe.2016.06.008.
9. Matsuda K, Tsuji H, Asahara A, Takahashi T, Kubota H, Nagata S, Yamashiro Y, Nomoto K. Sensitive quantification of *Clostridium difficile* cells by reverse transcription-quantitative PCR targeting rRNA molecules. *Appl Environ Microbiol*. 2012;78(15):5111–5118. doi:10.1128/AEM.07990-11.
10. Džunková M, Moya A, Vázquez-Castellanos JF, Artacho A, Chen X, Kelly C, D'Auria G. Active and secretory IgA-coated bacterial fractions elucidate dysbiosis in *Clostridium difficile* infection. *mSphere*. 2016;1(3):e00101–16. doi:10.1128/mSphere.00101-16.
11. Tschudin-Sutter S, Braissant O, Erb S, Strandén A, Bonkat G, Frei R, Widmer AF. Growth patterns of *Clostridium difficile* – correlations with strains, binary toxin and disease severity: a prospective cohort study. *PLoS One*. 2016;11(9):e0161711. doi:10.1371/journal.pone.0161711.
12. Yilmaz S, Haroon MF, Rabkin BA, Tyson GW, Hugenholtz P. Fixation-free fluorescence in situ hybridization for targeted enrichment of microbial populations. *ISME J*. 2010;4(10):1352–1356. doi:10.1038/ismej.2010.73.
13. Novakova J, Džunková M, Musilova S, Vlkova E, Kokoska L, Moya A, D'Auria G. Selective growth-inhibitory effect of 8-hydroxyquinoline towards *Clostridium difficile* and *bifidobacterium longum* subsp. *longum* in co-culture analyzed by flow cytometry. *J Med Microbiol*. 2014;63(12):1663–1669. doi:10.1099/jmm.0.080796-0.
14. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 2012;9(4):357–359. doi:10.1038/nmeth.1923.
15. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24–26. doi:10.1038/nbt.1754.
16. Du P, Cao B, Wang J, Li W, Jia H, Zhang W, Lu J, Li Z, Yu H, Chen C, et al. Sequence variation in *tcdA* and *tcdB* of *Clostridium difficile*: ST37 with truncated *tcdA* is a potential epidemic strain in China. *J Clin Microbiol*. 2014;52(9):3264–3270. doi:10.1128/JCM.03487-13.
17. Raffetseder J, Pienaar E, Blomgran R, Eklund D, Patcha Brodin V, Andersson H, Welin A, Lerm M. Replication rates of *Mycobacterium tuberculosis* in human macrophages do not correlate with mycobacterial antibiotic susceptibility. *PLoS One*. 2014;9(11):e112426. doi:10.1371/journal.pone.0112426.
18. Li SS, Zhu A, Benes A, Costea PI, Hercog R, Hildebrand F, Huerta-Cepas J, Nieuwdorp M, Salojärvi J, Voigt AY, et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science*. 2016;352(6285):586–589. doi:10.1126/science.aad8852.
19. Rinke C, Low S, Woodcroft BJ, Raina JB, Skarshewski A, Le XH. Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. *PeerJ*. 2016;4:e2486. doi:10.7717/peerj.2486.
20. Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, Ross JV, Vernon MC. Networks and the epidemiology of infectious disease. *Interdiscip Perspect Infect Dis*. 2011;2011:284909. doi:10.1155/2011/284909.
21. Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N, Dmitrieff E, Malmstrom R, Stepanauskas R, Woyke T. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc*. 2014;9(5):1038–1048. doi:10.1038/nprot.2014.067.
22. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol*. 2013;31(6):533–538. doi:10.1038/nbt.2579.
23. Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, Becraft ED, Brown JM, Pachiadaki MG, Povilaitis T, et al. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat Comm*. 2017;8(1):84. doi:10.1038/s41467-017-02128-5.
24. Ionescu D, Bizic-Ionescu M, De Maio N, Cypionka H, Grossart HP. Community-like genome in single cells of the sulfur bacterium *Achromatium oxaliferum*. *Nat Comm*. 2017;8(1):455. doi:10.1038/s41467-017-00342-9.