1416

Comparison of multilocus variable-number tandem-repeat analysis (MLVA) and whole genome sequencing (WGS) to confirm or refute *Clostridium difficile* Infection (CDI) case clusters

W Fawley,[1] D Eyre,[2] D Griffiths,[2] D Crook,[2] T Peto,[2] AS Walker,[2] MH Wilcox[1]

[1]Leeds Teaching Hospitals Trust & University of Leeds, UK, [2]NIHR Oxford Biomedical Research Centre, UK

Abstract

Background
MLVA and WGS have increased discriminatory power for *C. difficile* epidemiology compared with most other typing methods, but notably these currently analyse different parts of the bacterial genome. A formal performance comparison of MLVA vs WGS for assessment of suspected nosocomial CDI case clusters (sCDICCs) does not yet exist.

Methods
Isolates from 61 sCDICCs, affecting 2-41 patients in 31 UK hospitals involving 11 PCR-ribotypes, underwent 7-locus MLVA and Illumina WGS. Cases from each sCDICC shared a common PCR-ribotype and close epidemiological links. MLVA and WGS data from all possible pairs of isolates within each sCDICC were compared. Results from both techniques were then used independently to establish investigation outcome: (i) single outbreak, (ii) no transmission, and (iii) a mixture of related and unrelated isolates. Differences of >10 summed tandem repeat differences (STRD) by MLVA and >2 single nucleotide variants (SNVs) by WGS were used to exclude direct transmission.

Results
Classification of direct transmission by MLVA and WGS was concordant for 1190/1488 (80%) within sCDICC pairs. For discordant pairs, 229 (15%) had ≥2 SNVs but ≤10 STRD, and 69 (5%) had ≤2 SNVs but ≥10 STRD. Discordant pairs had higher numbers of MLVA variable loci (LV) than concordant pairs, supporting the more diverse measure in each type of discordant pairs; median (IQR) LV between pairs with ≥2 SNVs but ≤10 STRD and ≤2 SNVs but ≥10 STRD were 3 (3-4) and 3 (2-4), respectively, compared with 1 (0-2) between concordant pairs (p=0.001). Conclusions regarding whether or not sCDICCs comprised real clusters matched for 58/61 (95%) investigations. Discordant conclusions were due to only a single case in each of the 3 remaining investigations.

Conclusions
MLVA and WGS findings were largely concordant, and returned very similar conclusions when applied to outbreak investigations. Tools for MLVA analysis are currently more refined for service use; however future WGS platforms may bring additional benefits, eg. virulence or antimicrobial resistance prediction. Both MLVA and WGS currently offer enhanced discrimination over other genotyping tools.

Introduction

Multilocus variable-number tandem-repeat analysis (MLVA) (1-3) and whole genome sequencing (WGS) of *C. difficile* (4,5) both offer increased discrimination over other typing schemes commonly employed for investigating suspected nosocomial CDI case clusters (sCDICCs).
It is acknowledged that MLVA and current WGS methodologies analyse completely separate parts of the *C. difficile* genome and they may therefore have differing performance as tools for *C. difficile* surveillance and epidemiology. A formal performance comparison of these two techniques does not currently exist.
MLVA measures differences in the tandem-repeat copy number at 7 or more genetic loci in order to differentiate between isolates. Such differences are reported in terms of the calculated summed tandem-repeat difference across all loci (STRD). Evidence has led to the consensus that isolates ≤2 STRD apart should be regarded as indistinguishable and isolates 3-10 STRD apart within diverse strain collections should be regarded as closely related (1,6).
WGS can be used to compare single nucleotide variants (SNVs) between isolates across the non-repetitive core genome, which accounts for ~80% of the 4.3 million base pair *C. difficile* 630 reference genome (4,5,7). ≤2 SNVs have been previously described as consistent with transmission (8).
The aim of this study was to assess the utility of MLVA versus WGS for outbreak investigation and compare their individual performances across a collection of genuine hospital sCDICCs.

Methods

Isolates from sixty-one sCDICCs involving 300 patients from 31 UK hospitals (submitted to the *C. difficile* Ribotyping Network for England and Northern Ireland (CDRN) between June 2007 and July 2011) were investigated with both MLVA (3) and WGS. Cases in individual sCDICCs shared a common ribotype and strong epidemiological links.

MLVA

- Tandem repeat numbers were determined for seven loci present on the *C. difficile* genome (designated A6, B7, C6, E7, F3, G8 and H9) (9).
- Repeat regions were amplified using a common PCR protocol with adjustments made for PCR-ribotypes 078 and 017 to guarantee efficient primer annealing for some loci (10).
- Amplified fragments were sized using multi-coloured capillary electrophoresis in combination with a size marker, and tandem repeat numbers were calculated.
- Tandem repeat loci from selected isolates were sequenced to verify accurate assignment of repeat numbers (data not shown).
- The differences between isolates were calculated as the summed tandem-repeat difference (STRD) at all 7 loci.

WGS

- DNA from all *C. difficile* isolates under analysis were prepared for sequencing using standard Illumina (San Diego, CA) and adapted protocols.
- Samples were sequenced at the Wellcome Trust Centre for Human Genetics, Oxford, UK, on the Illumina HiSeq 2000 platform.
- Paired sequence reads (~ 100base pairs) were mapped using Stampy v1.0.17 to the *C. difficile* 630 reference genome (Genbank: AM180355.1), CD630 (7).
- This approach resulted in ~80% of the 4.3 million base pair genome available for analysis.
- Single nucleotide variants (SNVs) were identified across all mapped non-repetitive sites using SAMtools (version 0.1.18) (11).

Analysis

- WGS and MLVA results from all pairs of patients within each outbreak were compared. Transmission between pairs of cases was excluded by WGS and MLVA where pairs of cases were separated by >2SNVs or >10STRDs respectively (Table 1).
- Each of the 61 potential outbreaks were classified as i) all cases belonging to a single outbreak, ii) all cases being unrelated, or iii) a mix of related and unrelated cases, in order to investigate whether the conclusion reached in each potential outbreak investigation based on MLVA would be changed using WGS (Table 2).

Results

		MLVA	
		STRD ≤10	STRD >10
WGS	SNVs ≤2	945 (64%) LV: 1 (0-2)	69 (5%) LV: 3 (2-4)
	SNVs >2	229 (15%) LV: 3 (3-4)	245 (16%) LV: 4 (4-5)

Table 1. Classification of 1488 pairs of same PCR-ribotype cases across 61 UK sCDICCs using WGS and MLVA.

Overall 1190/1488 (80%) pairs of patients had concordant results on MLVA and WGS, supporting possible transmission between 945(64%) pairs and excluding transmission between 245(16%). However, for 229(15%) pairs, MLVA supported possible transmission but WGS did not (≤10 STRDs but >2 SNVs). Similarly, for 69(5%) of pairs WGS supported possible transmission but MLVA did not (≤2 SNVs but >10 STRDs). Discordant pairs had higher numbers of MLVA LVs than concordant pairs (median (IQR) LV between pairs with ≥2 SNVs but ≤10 STRD and ≤2 SNVs but ≥10 STRD were 3 (3-4) and 3 (2-4), respectively, compared with 1 (0-2) between concordant pairs (p=0.001).

		MLVA		
		Single outbreak	No transmission	Mixture
WGS	Single outbreak	33	0	1[a]
	No transmission	0	10	1[b]
	Mixture	1[c]	0	15

Table 2. Comparison of outbreak classification across 61 UK sCDICCs using WGS and MLVA.

Concordant findings are shaded in grey. Each of the 3 discordant classifications could be explained by a discordant finding in only one case in the investigation:

a) PCR-ribotype 106 (12 cases), single case 13 STRDs apart from closest case but 2 SNVs (all other cases within 7 STRD and 2 SNVs)
b) PCR-ribotype 015 (7 cases), single pair of cases with 4 STRDs but 12 SNVs (all other cases >10 STRD and ≥4 SNVs apart)
c) ribotype 027 (3 cases), single case 16 SNVs apart from closest case but 10 STRDs (all other cases within 4 STRD and 0 SNVs).

Discussion

- Despite analysing completely separate regions of the bacterial genome, conclusions reached for outbreak investigations using MLVA and WGS were largely concordant.
- This indicates that in the majority of cases the rate of genetic change in both the repetitive and non-repetitive regions of the *C. difficile* genome are generally comparable. An understanding of this is critical as current WGS technologies are unable to assay those areas of the genome currently exploited for surveillance purposes (e.g. in PCR-ribotyping, MLVA). This is a consequence of the longer lengths of repetitive regions typically found at these genetic loci compared with those that can currently be resolved by WGS (~100bp).
- As WGS read lengths continue to increase, it is likely to become possible to perform additional *in silico* typing 'equivalent' to existing *C. difficile* typing techniques, such as PCR-ribotyping, MLVA (on current and novel tandem-repeat loci), and multi-locus sequence typing (MLST), and to exploit other repetitive regions on the genome, on the same convenient platform. This would be of great benefit for our understanding of global CDI epidemiology, which is currently impeded by different regional surveillance programs employing varied typing techniques.
- Both MLVA and WGS currently require specific expertise in the consistent production and accurate interpretation of data. However, end-to-end commercial solutions for MLVA are available, and WGS pipelines are likely to become increasingly automated. The reagent costs of MLVA and WGS in this study were comparable (US $40 - $65 per sample). Both assays required similar amounts of time associated with laboratory processing (~16 hours of hands on time needed to process 96 samples, or ~10 minutes per sample).

Conclusions

- **Both MLVA and WGS offer similar enhanced discrimination over other established genotyping methods in CDI outbreak investigation, at similar per sample cost and laboratory time (according to current prices and technology).**
- **WGS currently offers additional benefits including *in silico* determination of virulence factors and antimicrobial resistance. Future WGS technologies aim to extend into regions of repetitive sequence, offering better coverage of the bacterial genome.**

References

1. Marsh J.W., O'Leary M.M., Shutt K.A. *et al.* 2006. *J Clin Microbiol.* **44**, 2558-2566.
2. Killgore G., Thompson A., Johnson S. *et al.* 2008. *J Clin Microbiol.* **46**, 431–437.
3. Fawley W.N., Wilcox M.H. 2011. *J Clin Microbiol.* **49**, 4333–4337.
4. Eyre D.W., Golubchik T., Gordon N.C. *et al.* 2012. *BMJ Open.* **2**, e001124.
5. Didelot X., Eyre D.W., Cule M.L. *et al.* 2012. *Genome Biol.* **13**, R118.
6. Marsh J.W., O'Leary M.M., Shutt K.A. *et al.* 2010. *J Clin Microbiol.* **48**, 412–418.
7. Sebaihia M., Wren B.W., Mullany P. *et al.* 2006. *Nat Genet.* **38**, 779–786.
8. Eyre D.W., Cule M.L., Walker A.S. *et al.* 2012. *Lancet.* **380**, S12.
9. van den Berg R.J., Schaap I., Templeton K.E. *et al.* 2007. *J Clin Microbiol.* **45**, 1024–1028.
10. Bakker D., Corver J., Harmanus C., *et al.* 2010. *J Clin Microbiol.* **48**, 3744–3749.
11. Li H., Handsaker B., Wysoker A., *et al.* 2009. *Bioinformatics.* **25**, 2078–2079.