THE LANCET Infectious Diseases

Supplementary webappendix

This webappendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Bozio CH, Vuong J, Dokubo EK, et al, for the Liberian Meningococcal Disease Outbreak Response Team. Outbreak of *Neisseria meningitidis* serogroup C outside the meningitis belt—Liberia, 2017: an epidemiological and laboratory investigation. *Lancet Infect Dis* 2018; published online Oct 15. http://dx.doi.org/10.1016/S1473-3099(18)30476-6.

Supplemental Methods

Metagenomic analysis

Metagenomic analysis was conducted to identify *Neisseria meningitidis* (Nm) genome sequences in eight non-oral specimens (3 cardiac blood, 2 plasma, 1 blood, 1 urine, and 1 vitreous humor fluid), from six cases (Figure 3). These specimens were selected based on rt-PCR results that indicated abundant Nm DNA in these specimens. Extracted DNA from these specimens was used to generate Nextera XT libraries, according to manufacturer's instructions, along with a no-DNA control. Libraries were sequenced on an Illumina MiSeq, generating 250bp paired-end reads. Reads were deduplicated using BBTools clumpify v37·41 (B. Bushnell [http://sourceforge.net/projects/bbmap/]) and trimmed of adapters and low-quality base calls using Cutadapt v1·8·3.[1] Human sequences were removed by mapping to hg19[2] with Bowtie v2·2·9.[3] The remaining read pairs were matched to bacterial species using k-SLAM v1·0.[4]

To identify the lineage of Nm present in each specimen, metagenomic reads identified as Nm were compared to whole genome data of 141 diverse isolates. These 141 isolates were selected to represent the sequence diversity of 4,810 genomes in the CDC Nm collection. This set of representative isolates was constructed by first selecting the most diverse pair of genomes based on Mash distances (v1.1, k=32, s=10,000), then iteratively adding the genome that was most distant from the isolates already in the representative set. The final set of 141 isolates had a Mash distance threshold of 0·529%; each pair of isolates within the set differed by >0·529%, while all isolates outside of the set were <0·529% different from one of the isolates in the set. The 141 isolates included 140 sequence types (STs); 92 STs belonged to 35 different clonal complexes and the remaining 48 STs were not assigned to any clonal

complex. Among this collection, there were 15 NmC isolates, with 15 STs (ST-11, 66, 212, 337, 344, 2976, 3779, 5323, 6281, 7151, 8797, 8798, 10217, 11579, and 12817).

All Nm reads were mapped to the serogroup C, ST-11 reference genome FAM18[5] using Snippy[6] and sequence similarity was assessed at the 178,772 polymorphic positions identified among the 1,472,670 positions that had base calls in all 141 isolate genomes (not including FAM18). Pairwise sequence similarity at polymorphic positions among the 141 isolate genomes ranged from 78% to 94% with a mean of 83%.

To assess the similarity of metagenomic DNA to each of the 141 isolate genomes (Figure 3), reads identified as *Neisseria* or Nm by k-SLAM were mapped to FAM18 and base calls were tallied using SAMtools v1·4·1[7] "mpileup" to identify positions where all base calls were identical to the isolate genome. Sequence similarity was defined as the number of polymorphic positions with matching base calls divided by the total number of polymorphic positions where the specimen reads identified a single base.

To evaluate bias and variability of these sequence similarity estimates, we simulated the detection of each of the 141 isolate genomes by randomly sampling a set of 250 read pairs for each genome and then calculating their sequence similarity to the other isolate genomes. Percent similarity calculated from the down-sampled reads tended to be slightly higher than the percent similarity calculated from the full set of polymorphic positions; the median difference between the two calculations was 0.4 percentage points (95% of differences between -1·76 and 2·64 percentage points). Conversely, comparison of the sets of 250 read pairs to their own isolate genome produced similarity calculations slightly lower than 100%; median similarity was 99·87% (95% of values between 99·29% and 99·98%). Together, these results

demonstrate that a small number of reads is sufficient to distinguish closely related genomes from distantly related genomes.

Data analysis was performed with SciPy v0·18 (E. Jones, E. Oliphant, P. Peterson, et al. SciPy: Open Source Scientific Tools for Python, 2001 [http://www.scipy.org/]) and BioPython v1·68.[8]

Supplemental Table 1. Allelic profiles of the clinical specimens compared with related sequence types

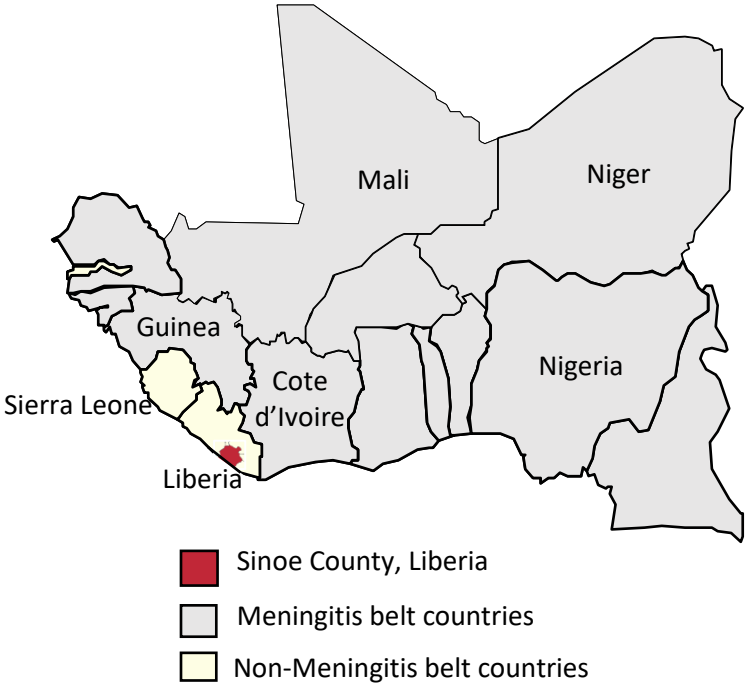
ID	Nm Serogroup	ST	CC	*abcZ*	*adk*	*aroE*	*fumC*	*gdh*	*pdhC*	*pgm*
Case 7*	C	UD	10217	12	5	4	7	187	UD‡	UD‡
NA†	Nongroupable	9367	10217	12	5	4	7	187	2	120
NA†	C	10217	10217	12	5	4	643	187	2	120

* Patient specimen tested using Sanger sequencing.

† NA = Not applicable

‡ UD = Undetermined

Supplemental Figure 1: Map of West African countries in the meningitis belt, relative to Sinoe County, Liberia



References

1. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 2011;17(1):10-2.

2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

3. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.

4. Ainsworth D, Sternberg MJE, Raczy C, Butcher SA. k-SLAM: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. Nucleic Acids Res. 2017;45(4):1649-56.

5. Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T, et al. Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. PLoS Genet. 2007;3(2):e23.

6. Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, et al. Prospective Whole-Genome Sequencing Enhances National Surveillance of Listeria monocytogenes. J Clin Microbiol. 2016;54(2):333-42.

7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

8. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422-3.