nature biotechnology

MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island

Philip M Ashton[1,6], Satheesh Nair[1,6], Tim Dallman[1], Salvatore Rubino[2,3], Wolfgang Rabsch[4], Solomon Mwaigwisya[5], John Wain[5] & Justin O'Grady[5]

Short-read, high-throughput sequencing technology cannot identify the chromosomal position of repetitive insertion sequences that typically flank horizontally acquired genes such as bacterial virulence genes and antibiotic resistance genes. The MinION nanopore sequencer can produce long sequencing reads on a device similar in size to a USB memory stick. Here we apply a MinION sequencer to resolve the structure and chromosomal insertion site of a composite antibiotic resistance island in *Salmonella* Typhi Haplotype 58. Nanopore sequencing data from a single 18-h run was used to create a scaffold for an assembly generated from short-read Illumina data. Our results demonstrate the potential of the MinION device in clinical laboratories to fully characterize the epidemic spread of bacterial pathogens.

Short read, high-throughput, next-generation sequencing (NGS) technology has transformed our understanding of microbiology and is poised to become an integral tool in epidemiology[1]. Although the utility of whole genome sequencing (WGS) for public health infection control is clear, adoption in clinical microbiology laboratories has been limited[2]. This is partly because short-read technologies cannot unambiguously assemble repetitive elements that are longer than sequencing read-length into a single contig. This assembly problem generates multiple contigs and leaves gaps in whole genome assemblies. It is particularly difficult to correctly assemble regions in which genes have been acquired by horizontal gene transfer, such as resistance and pathogenicity islands, and prophage[3], owing to their inherent repetitive nature or the flanking of these elements by repetitive insertion sequences. Analyzing these regions is essential for determining key characteristics such as antibiotic resistance profiles and for identifying highly pathogenic variants of many bacterial species[4]. Currently gap closure requires extensive, post-sequencing, laboratory-based analysis, which can take several months and makes the results irrelevant for clinical diagnostics and for guiding public health interventions.

Sequencing technology that generates long reads, capable of spanning repetitive sequences and closing gaps in short read data, is commercially available (Pacific Biosystems PacBio RS II) but has significant capital cost outlay, a very large laboratory footprint and is technically demanding. DNA sequencing using nanopore technology is an alternative method for producing long-read sequence data but has been a specialized research tool until very recently[5] and is not, as of December 2014, available commercially. The recent distribution of the MinION by Oxford Nanopore Technologies Ltd. in an early-access program (named the MinION Access Programme) has made it possible to evaluate the utility of long-read sequencing using a device that resembles a large USB memory stick.

There were an estimated 26.9 million cases of typhoid fever in 2010 (ref. 6) with a very high proportion of those cases in urban slums[7]. A recent emergence of a globally distributed multidrug-resistant (MDR) *Salmonella enterica* serovar Typhi (*S.* Typhi) haplotype, H58, has been observed contributing to a reduction in genetic diversity of extant *S.* Typhi[8–12]. At Public Health England, Salmonella Reference Service, in Colindale, UK, we have observed a similar increase in isolates of MDR *S.* Typhi phage type E9 variant from patients with a travel history to the Indian subcontinent. The routine adoption of WGS technologies to identify and type *Salmonella* isolates here allowed these to be characterized as H58 harboring multiple resistance elements including, *strA*, *strB*, *sulI*, *sulII*, *dfrA7* and bla_{TEM-1} (ref. 13) encoded on Tn10 and Tn9. The specific resistance plasmid (plasmid PST6 (incHI1)) typical of H58 isolates was, however, not present, raising the possibility that an antibiotic resistance island has integrated into the H58 chromosome.

Here, we report a hybrid assembly of combined MinION and Illumina HiSeq data to identify the structure and insertion site of a chromosomal antibiotic resistance island in *S.* Typhi H58, which, despite many "whole genome" sequencing projects[14], has not been previously characterized.

Two *S.* Typhi H58 strains (H125160566 and 08-0446) were sequenced using the Illumina HiSeq, and SNP typing was used to confirm haplotype[12]. *De novo* assembly of Illumina sequence for strain 08-0446 (ENA accession number ERR668456) resulted in 143 contigs, an N50 (a statistical measure of average length of a set of sequences) of 124 kbp and average genome coverage of 78× (374 million bases of >Q30 data). *De novo* assembly of strain H125160566 (ENA accession number ERR668457) resulted in 86 contigs, an N50 of 154 kbp and

[1]Gastrointestinal Bacteria Reference Unit, Public Health England, Colindale, London, UK. [2]Department of Biomedical Sciences, University of Sassari, Sassari, Italy. [3]Department of Infection and Immunity, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia. [4]National Reference Centre for Salmonellae and other Enterics, Robert Koch Institute, Wernigerode, Germany. [5]Norwich Medical School, University of East Anglia, Norwich, UK. [6]These authors contributed equally to this work. Correspondence should be addressed to J.O'G. (justin.ogrady@uea.ac.uk) or T.D. (tim.dallman@phe.gov.uk).

Received 7 September; accepted 21 November; published online 8 December 2014; doi:10.1038/nbt.3103

Table 1 Read and mapping statistics

Read stats[a]

	Number of reads	Median length (bp)	Max length (bp)	Total length (Mbp)	Median Phred score	Median accuracy (%)
Total	16,401	5,412	66,748	93.4	5	68.4
Template	8,209	5,614	58,810	49.6	3	49.9
Complement	4,454	4,728	66,748	21.1	4	60.2
2D	3,738	5,943	31,630	22.7	8	84.2

Mapping stats[b]

	Number of reads aligned	Number of alignments	Total length alignments (Mbp)	Median accuracy (%)	Median gaps (%)
Total	11,278	16,337	70.8	64.2	17.8
Template	4,705	6,194	31.4	64.9	17.1
Complement	3,171	5,275	17.4	61.1	23.5
2D	3,402	4,868	22.0	72.6	14.8

[a]Read statistics for a single 18-h MinION run of *S.* Typhi H125160566, broken down by read type. [b]Statistics on mapping of reads to the H125160566 Illumina assembly.

average genome coverage of 38× (182 million base pairs of >Q30 data). When these contigs were aligned to pAKU[12] (a sequenced plasmid most closely related to pST6 found in *S.* Typhi H58) (ref. 13) using BLAST, only 21 kbp (10%) of the plasmid had matches with significant nucleotide similarity. Similar results were obtained when the H125160566 assembly was compared against pHCM1 incHI1 plasmid (found in *S.* Typhi CT18)[15], which has been hypothesized to be a source of the chromosomally encoded resistance genes in H58.

Seven Illumina contigs (>100 bp and totaling 20,974 bp) had nucleotide similarity and query coverage with pAKU >90%, including two with 100% similarity that code for transposable elements IS26 and IS1. In addition, four long contigs (15–65 kbp) had short regions of similarity (<100 bp) with pAKU toward the 5′ or 3′ ends (**Supplementary Fig. 1**). These regions included IS1 inverted repeat sequences, indicating that there were two separate IS1 insertion sites. One of these insertions disrupted the *yidA* gene, whereas the other was in an intergenic region between STY3618 and STY3619 of the CT18 reference genome. Although it is possible to determine whether an insertion site is occupied using traditional methods (PCR), it was not feasible in this case as we did not know the internal structure of the island and it was too large to span by PCR.

S. Typhi strain H125160566 was also sequenced on the MinION device for 18 h resulting in 16,401 sequencing reads (ENA accession number ERR668747—the MinION fast5 files contain both raw 'squiggle plot' data and associated base-called data) with median length 5,412 bp, a maximum length of 66,748 bp, median Phred score of 5, median accuracy of 68.4% (derived from Phred score) and a total of 93.4 Mbp of sequence data (**Table 1**). MinION sequencing produces three types of read for each DNA molecule analyzed, template (lead DNA strand), complement (complementary DNA strand) and two-direction (2D). When both template and complement data are available, an additional 2D basecall is performed, providing a consensus. There were 8,209 template, 4,454 complement and 3,738 2D reads. On average, the complement reads were shorter (median 4,728 bp) than the template and two-directional reads (5,614 bp & 5,943 bp, respectively) (**Supplementary Fig. 2**). According to the MinION basecalling software (Metrichor version 1.3.1) Phred scores (a measure of the likely accuracy of the base called data), the two-direction reads had the highest mean accuracy (83.6%), followed by the complement reads (55.6%) and the template reads (53.9%).

Two additional MinION runs were carried out, one was a repeat of H125160566 with overnight (rather than 30 min) library incubation (as recommended by Oxford Nanopore after some reagent problems caused by shipping conditions; ENA accession number ERR668746) and one with the second *S.* Typhi H58 isolate (strain 08-04776, ENA accession number ERR668983). The data from all three runs were very similar (see **Supplementary Table 1**).

All *S.* Typhi strain H125160566 MinION reads were mapped to the Illumina assembly of the same strain using the LAST sequence alignment tool[16] (**Table 1**). In our hands LAST performed better than other programs tested (BLAST and BWA mem) as it can align sequences with many mismatches and gaps, which is typical of the sequence data currently produced by the MinION. In total 68.7% (11,278/16,401) of the reads mapped at least once to the Illumina assembly, with 16,337 nonduplicate alignments giving an average coverage of 14×. Mismatches between MinION reads and the Illumina assembly were identified, giving a mean percentage accuracy for all read types of 65.6%. Complement reads had the lowest similarity (61.6%), followed by template (64.3%) then two-directional reads (71.5%). There were a large number of gaps in the alignments, with a median of 18.5% of the length of alignments consisting of indels.

Comparing the mapping-derived accuracy of MinION reads with the Phred-derived accuracy of the same reads provided insight into the reliability of Phred calling for MinION data. For H125160566, the template and complement reads had median mapping-derived accuracies of 64.3% and 61.6%, whereas the median Phred-derived accuracy for these same reads was 60.2%, demonstrating the reliability of Phred scores for these read types. For the 2D reads, however, the median mapping accuracy was 71.5%, whereas the Phred-derived accuracy was 84.2%, indicating that the Phred scoring algorithm overestimated the accuracy of these reads.

The error profile of the MinION H125160566 data set was characterized. In the 70.8 megabases of aligned nanopore sequence an indel occurred every 5.9 bases on average. Two-thirds of these indels were deletions (8.6 Mbp), whereas one-third were insertions (4.15 Mbp). The mean deletion length was 1.7 bp, whereas the mean insertion length was 1.6 bp. Both insertion and deletion lengths had a negative exponential distribution (**Supplementary Figs. 3** and **4**).

A *z*-score was calculated to identify k-mers that were deleted from mapped MinION reads at a higher-than-expected frequency. When *z*-scores for deletions of 3–6 bp were binned by GC content, the tendency of the MinION to skip k-mers containing As and Ts only or Gs and Cs only could be seen (**Fig. 1**). In particular, there were two

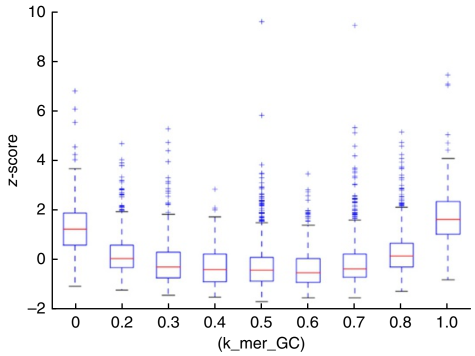

Figure 1 A box-and-whiskers plot of the *z*-score for deleted k-mers of 3–6 bp in length, grouped by the proportion of GC content (0 = 0% GC content; 1 = 100% GC content).

3-mers with a z-score > 3, GGG (z = 3.3) and CCC (z = 3.2), two 4-mers with a z-score > 3, GGGG (z = 4.4) and CCCC (z = 4), and four 5-mers with a z-score > 4, AAAAA, TTTTT, CCCCC and GGGGG. Two more diverse k-mers that had high z-scores were TAGGCA and TAGGGC, with z-scores of 9.6 and 9.5, respectively. When insertions were examined there were two 3-mers with a z-score > 3, CCC (z = 3.7) and GGG (z = 3.6), and two 4-mers with a z-score > 3, CCCC (z = 6.1) and GGGG (z = 6.0).

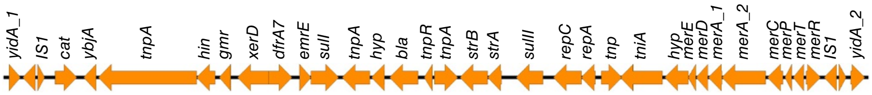

Figure 2 Genetic organization of the *S.* Typhi chromosomal resistance island. Gene names were assigned using BLAST analysis and manual annotation.

To identify substitutions, all nonconsensus bases were extracted from the MinION alignment to the Illumina assembly. This showed that A to T and T to A substitutions were approximately half as frequent as other substitutions (**Supplementary Table 2**). These data suggest that MinION currently has difficulty differentiating Gs and Cs, particularly when these bases are present in homopolymeric tracts.

We next used the long MinION reads to scaffold the H125160566 Illumina contigs with homology to pAKU (see Online Methods) to determine the structure and chromosomal insertion site of the antibiotic resistance island. Forty MinION reads that were informative as to the structure and insertion site of the island were identified and these reads were used to link the island contigs across the insertion sequences (**Supplementary Fig. 1**). There were between two and four reads spanning each break in the island assembly.

The breaks in the Illumina assembly were caused by the presence of insertion sequences in all but two cases. These additional breaks were caused by a decrease in Illumina data coverage (from an average of 38× to 1–2×) in the *merA* and *hyp* genes, resulting in 1.3-kb and 0.3-kb fragments missing in the Illumina assembly. This low coverage was associated with high GC content (>70%), a known problem with Nextera Illumina sequencing[17]. When determining the island sequence in these regions, data were derived from confirmatory PCR products rather than the erroneous Illumina-only *de novo* assembly data.

The MinION reads identified the insertion site as the *yidA* gene (**Supplementary Fig. 5**). Mapping from both ends of this disrupted gene into the island showed that it is flanked by IS1 elements. The island contains several of the resistance genes and/or elements found on IncHI1 plasmids in MDR *S.* Typhi, including *strA*, str*B*, *sulI*, *sulII* and bla_{TEM-1} (**Fig. 2**; ENA accession number PRJEB7681). The structure and insertion site of the island were confirmed by PCR followed by Sanger sequencing. This confirmation was not possible until the island structure was solved (using the hybrid assembly) because of the low-coverage and misassembled Illumina data.

The second H58 *S.* Typhi strain analyzed in this study (08-04776) contained the antibiotic resistance island and an IncN plasmid. The Illumina contigs from 08-04776 contained one 7 kb contig with a dihydropteroate synthase gene (*sulI*) and there were no other *sulI* genes in the assembly. When MinION reads were mapped to this contig, they did not map contiguously from this contig to other island contigs, but rather, mapped only to the *sulI*-encoding section. Therefore, we hypothesize that 08-04776 encodes two *sulI* genes; one on the IncN plasmid and one on the chromosomally inserted island. The assembly of the Illumina data collapsed these two genes into a single copy, which represented a misassembly that confounded analysis and was only resolved using the MinION reads.

To demonstrate the utility of MinION sequence for hybrid genome assembly, the Illumina and Minion data for the H125160566 strain were assembled using SPAdes[18]. This method resolved the genome into 34 contigs, with an N50 of 319 kbp, whereas the Illumina-only assembly produced 86 contigs with an N50 of 154 kbp. The SPAdes hybrid assembly confirmed the *yidA* insertion site but failed to resolve the complete island structure with contig breaks in the *hyp* and *merA* genes, presumably due to the low coverage regions in the Illumina data. SPAdes hybrid assembly demonstrated significant improvement over Illumina-only assembly—further improvements would be possible given higher-coverage MinION data.

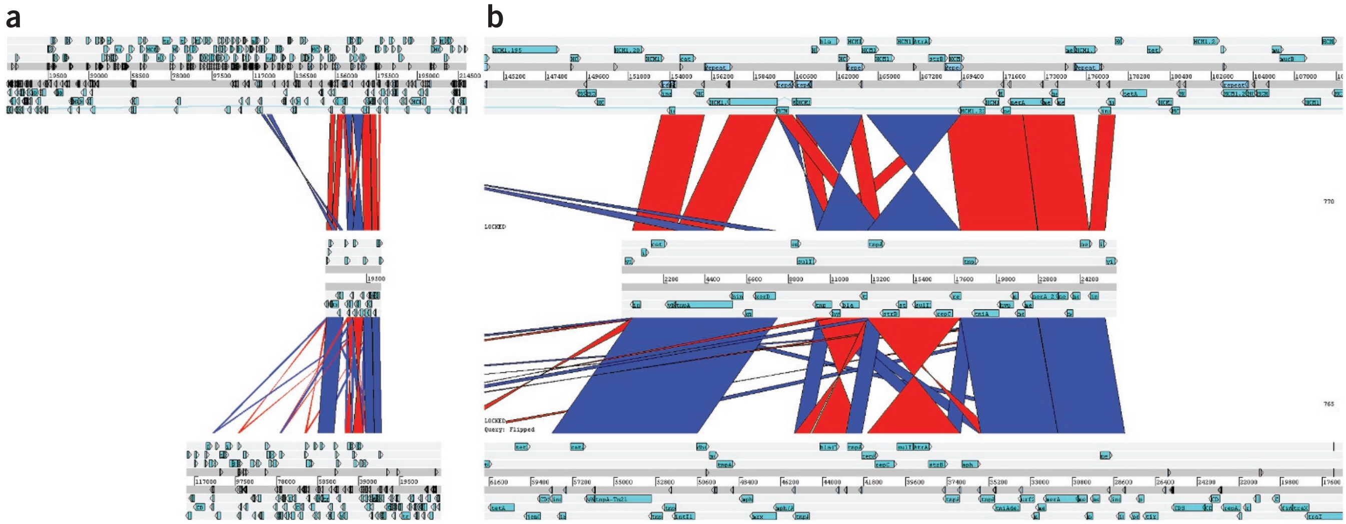

Figure 3 Comparison of the *S.* Typhi chromosomal resistance island with two closely related plasmids. The *S.* Typhi chromosomal resistance island was compared to the *S.* Typhi multi-drug resistance plasmid pHCM1 and the plasmid with the highest BLASTN similarity to the island, pRSB107, from an uncultured organism in sewage. Top, pHCM1; middle, genomic island; bottom, pRSB107. (**a**) Overview to show full-length sequence of pHCM1 and pRSB107. (**b**) Zoomed-in view to show detail of regions of similarity. Red indicates 100% similarity and blue indicates 99% similarity; the cut-off minimum was set at 800.

To test the hypothesis that the antibiotic resistance island originated from an S. Typhi MDR IncHI1 plasmid, we compared it to pHCM1 and other publicly deposited sequences using BLASTN—surprisingly one of the nearest matches (88% coverage, 99% identity) was pRSB107, a plasmid that was sequenced in sewage[19]. A comparison of the structure of the island with pHCM1 and pRSB107 (**Fig. 3a**,**b**) demonstrated that, although the island shares several features with both plasmids, the lack of absolute similarity suggests it did not come directly from either, but rather, all three have a common source that is currently undefined. The evolutionary process leading to the emergence of H58 has involved bacterial host-plasmid adaptation causing a reduction in biological cost of the resistance phenotype[13] and now, as we have demonstrated, chromosomal integration of a set of genes. These genes have no obvious source, and the similarity of the island to a plasmid identified in sewage has opened up our research into the island's origins. Acquisition of this antibiotic resistance island has allowed a clonal sweep within the population of one of the classic human pathogens.

If the enormous potential of applying WGS in healthcare is to be realized, then public health and clinical laboratories must have easy access to technology capable of providing fully assembled bacterial genomes. This will enable bacterial identification, subtyping and resistance gene detection at the point of clinical need. MinION technology demonstrates potential for such applications as it is highly accessible and is already well supported by the bioinformatics community, with two software packages, poretools and poRe, released for data analysis[20,21]. In our hands MinION flow cells with >400 active nanopores (out of a total of 512) produce yields up to 150 Mbp in 48 h (maximum yield in this study was 90 Mbp) with an accuracy of ~72% for 2D reads. This is comparable to R7 MinION data published by Quick *et al.*[22] although the yield they achieved was ~400 Mb and the mean accuracy for 2D reads was higher (~80%)[22]. These differences in performance are most likely related to variations in the quality of the flow cells. These data are a considerable improvement on the 10% accuracy reported in an early MinION paper by Mikheyev *et al.*[23]. We did not make a comparison with other long read technologies such as PacBio, as this was not the purpose of our study, but the median and maximum MinION read lengths (~6 kbp and 60 kbp, respectively) were comparable to those reported for recent PacBio chemistry (~7–12 kbp average and 20–30 kbp maximum length filtered subreads) although the yield (~90 Mbp for MinION versus 350 Mbp for PacBio) and accuracy (~72% versus 85%) were lower[24]. It must be noted that, in our hands, the length of the input DNA, not the chemistry, determined read length on the MinION, and we did not attempt to maximize read length.

Further improvements are needed in MinION data quality and yield, which are promised in flow cell and library preparation kit upgrades. The recent R7.3 flow cell and SQK-MAP003 gDNA library preparation kit have increased the proportion of full 2D reads (best-quality 2D reads) from 8% to 86% (ref. 22), which results in improved accuracy (~85% for full 2D reads). Recent bioinformatic developments have allowed the *de novo* assembly of genomes from high coverage (~100×) PacBio data[25] and this may also be possible using improved MinION technology in the future.

In conclusion, the MinION data presented here have a higher error rate than other widely used long-read and short-read sequencing platforms. However, the reads produced are of comparable length to PacBio and, despite the high error rate, we have used the MinION data to solve the structure of a complex antibiotic resistance island using 'off the shelf' bioinformatic tools. Nanopore technology has the potential to create a paradigm shift in genomics; low cost, long-read sequencing can be carried out in nonspecialist laboratories, both research and service. The question is: can nanopore sequencing replace short-read sequencing technology? The answer: not yet, but it is knocking at the door.

Accession codes. ENA: ERR668456, ERR668457, ERR668747, ERR668746, ERR668983 and PRJEB7681.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

M. Day (Salmonella Reference Unit, Public Health England) for antibiotic susceptibility tests. J.O'G. and J.W. were funded by the University of East Anglia. We would like to thank Oxford Nanopore Technologies Ltd. for including us in the MinION Access Programme. We would also like to thank L. Nederbragt for a thorough review and contributions toward the presentation of **Figure 1**.

AUTHOR CONTRIBUTIONS

P.M.A., S.N., T.D., J.W. and J.O'G. conceived the study, performed the analysis and wrote the first draft of the manuscript. J.O'G. and S.M. performed the MinION sequencing. P.M.A. and T.D. performed the bioinformatics analysis. S.N. performed the PCR analysis and coordinated the Illumina sequencing. P.M.A., T.D., S.R., W.R., J.W. and J.O'G. analyzed the resistance island structure and insertion site and devised the figures. All authors contributed to editing and data analysis of the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Underwood, A.P. *et al.* Public health value of next-generation DNA sequencing of enterohemorrhagic *Escherichia coli* isolates from an outbreak. *J. Clin. Microbiol.* **51**, 232–237 (2013).
2. Wain, J. & Mavrogiorgou, E. Next-generation sequencing in clinical microbiology. *Expert Rev. Mol. Diagn.* **13**, 225–227 (2013).
3. Thomson, N. *et al.* The role of prophage-like elements in the diversity of Salmonella enterica serovars. *J. Mol. Biol.* **339**, 279–300 (2004).
4. Livermore, D.M. & Wain, J. Revolutionising bacteriology to improve treatment outcomes and antibiotic stewardship. *Infect Chemother.* **45**, 1–10 (2013).
5. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
6. Buckle, G.C., Walker, C.L. & Black, R.E. Typhoid fever and paratyphoid fever: Systematic review to estimate global morbidity and mortality for 2010. *J. Glob. Health* **2**, 010401 (2012).
7. Wain, J., Hendriksen, R., Mikoleit, M., Keddy, K. & Ochiai, R. Typhoid fever. *Lancet* doi:10.1016/S0140-6736(13)62708-7 (21 October 2014).
8. Roumagnac, P. *et al.* Evolutionary history of *Salmonella* typhi. *Science* **314**, 1301–1304 (2006).
9. Kariuki, S. *et al.* Typhoid in Kenya is associated with a dominant multidrug-resistant Salmonella enterica serovar Typhi haplotype that is also widespread in Southeast Asia. *J. Clin. Microbiol.* **48**, 2171–2176 (2010).
10. Holt, K.E. *et al.* Temporal fluctuation of multidrug resistant salmonella typhi haplotypes in the mekong river delta region of Vietnam. *PLoS Negl. Trop. Dis.* **5**, e929 (2011).
11. Holt, K.E. *et al.* High-resolution genotyping of the endemic Salmonella Typhi population during a Vi (typhoid) vaccination trial in Kolkata. *PLoS Negl. Trop. Dis.* **6**, e1490 (2012).
12. Holt, K.E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* **40**, 987–993 (2008).
13. Holt, K.E. *et al.* Emergence of a globally dominant IncHI1 plasmid type associated with multiple drug resistant typhoid. *PLoS Negl. Trop. Dis.* **5**, e1245 (2011).
14. Le, T.A. *et al.* Clonal expansion and microevolution of quinolone-resistant Salmonella enterica serotype typhi in Vietnam from 1996 to 2004. *J. Clin. Microbiol.* **45**, 3485–3492 (2007).
15. Phan, M.D. *et al.* Variation in Salmonella enterica serovar typhi IncHI1 plasmids during the global spread of resistant typhoid fever. *Antimicrob. Agents Chemother.* **53**, 716–727 (2009).
16. Frith, M.C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11**, 80 (2010).

17. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).
18. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
19. Szczepanowski, R. *et al.* The 120 592 bp IncF plasmid pRSB107 isolated from a sewage-treatment plant encodes nine different antibiotic-resistance determinants, two iron-acquisition systems and other putative virulence-associated functions. *Microbiology* **151**, 1095–1111 (2005).
20. Watson, M. *et al.* poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* doi:10.1093/bioinformatics/btu590 (29 August 2014).
21. Loman, N.J. & Quinlan, A.R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30**, 3399–3401 (2014).
22. Quick, J., Quinlan, A.R. & Loman, N.J. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *GigaScience* **3**, 22 (2014).
23. Mikheyev, A.S. & Tin, M.M.Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Res.* **14**, 1097–1102 (2014).
24. Kim, K.E. *et al.* Long-read, whole genome shotgun sequence data for five model organisms. Preprint at http://biorxiv.org/content/early/2014/10/23/008037 (2014).
25. Chin, C.S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).

ONLINE METHODS

Bacterial isolation, biochemical and serological identification. *S.* Typhi strain H125160566 was isolated in 2012 from a patient returning from Bangladesh and sent to the Salmonella Reference Service, Public Health England. *S.* Typhi strain 08-04776 was isolated in 2008 at the Robert Koch Institute, Wernigerode, from a patient returning from Iraq. The sequenced *S.* Typhi strain CT18 (ref. 26) and *S.* Paratyphi A strain AKU 12601 (ref. 27) were used as controls in the PCR experiments. The isolates were biochemically typed and serotyped in accordance to the White-Kauffmann-le Minor Scheme[28].

Phage typing. *S.* Typhi strains were typed according to the phage-typing scheme described by Callow[29].

Antimicrobial resistance. *S.* Typhi strains were tested for resistance to antimicrobials using standard disc diffusion methods for Enterobacteriaceae in accordance with the European Committee on Antimicrobial Susceptibility Testing (EUCAST, Version 3.1) guidelines. The control strain used was *Escherichia coli* ATCC 25922.

DNA extraction and quantification. Chromosomal DNA used for PCR and sequencing was isolated using the Wizard genomic DNA purification kit (Promega) according to the manufacturer's instructions.

DNA was quantified using the Quant-iT dsDNA High Sensitivity Assay (HS) Kit (Life Technologies), which enables high-throughput quantitation of samples on the GloMax Multi+ Detection System (Promega).

Illumina DNA sequencing. Extracted DNA was prepared using the NexteraXT library preparation method according to the manufacturer's instructions and sequenced with a standard 2 × 101 base protocol on a HiSeq Instrument (Illumina, San Diego).

Illumina data analysis. *S.* Typhi strains were confirmed as being Haplotype 58 from the Illumina sequencing data by mapping to the *S.* Typhi CT18 reference genome (NC003198) and determining the presence of haplotype 58 specific SNPs[12].

Illumina reads were assembled using SPAdes v3.1.1 (ref. 18) with the '—careful' flag and k-mers of 21, 33, 55 and 77. The resulting contigs were compared against the IncHI1 plasmid pAKU (NCBI accession AM412236) using BLAST within the BioPython framework[30] and all contigs that contained a region that matched pAKU with an E value of less than 1×10^{-20} were extracted from the assembly. The alternative IncHI1 plasmid pHCM1 (NCBI accession AL513383) was also used as the reference and the same regions of the same contigs matched the plasmid.

MinION library preparation. The Oxford Nanopore MinION Genomic DNA Sequencing Kit was used to prepare the *S.* Typhi DNA libraries according to the manufacturer's instructions. Briefly, kit reagents were either thawed on ice (tether, DNA CS, HP adaptor, adaptor mix, ligation buffer, HP motor and fuel mix) or at room temperature (wash buffer, elution buffer and EP buffer), and once thawed, all reagents were kept on ice until required. One microgram of DNA was diluted to 85 µl in molecular grade water and sheared using a Covaris g-TUBE according to the manufacturer's instructions by centrifuging through the column in both directions at 3,600*g* for 60 s (sample was pulse centrifuged at 4,000*g* if all liquid had not passed through the column). End-repair was performed on the sheared DNA using NEBNext End Repair Module (New England BioLabs, cat. no. E6050) by adding 10 µl reaction buffer and 5 µl enzyme mix and incubating for 30 min at 20 °C. The end-repaired DNA was then purified using 1× volume (100 µl) Agencourt AMPure XP beads (Beckman Coulter Inc., cat. no. A63880) according to the manufacturer's instructions. Samples were incubated on a magnetic rack (Invitrogen MagnaRack, cat. no. CS15000) for 3 min and washed twice in 200 µl 70% ethanol while still on the magnet. DNA was eluted in 25.2 µl of molecular grade water. dA-tailing was then performed using NEBNext dA-tailing module (New England BioLabs, cat. no. E6053) in a total of 30 µl by adding 3 µl buffer and 1.8 µl Klenow Fragment to the clean DNA. The sample was incubated in a thermal cycler for 30 min at 37 °C. A ligation reaction was then performed by adding the following reagents in order: 50 µl dA-tailed DNA; 10 µl adaptor mix; 10 µl HP adaptor; 50 µl Blunt/TA DNA ligase (New England BioLabs M0367). The reaction was gently pipetted up and down to mix and incubated at room temperature for 10 min. The adapted-ligated DNA was purified using 0.4 × volume (40 µl) Agencourt AMPure XP beads according to the manufacturer's instructions but using the Oxford Nanopore-supplied wash buffer (150 µl × one wash only making sure to remove all wash buffer) and elution buffer (25 µl). The tether was annealed by adding 10 µl tether to the 25 µl ligated DNA and incubating for 10 min at room temperature. Finally, 15 µl HP motor was added to the reaction (50 µl total volume) and incubated for 30 min (or overnight) at room temperature. The library was then ready for sequencing on the MinION (see below).

MinION sequencing. During the last half-hour library preparation incubation step the MinION was connected to the computer (a PC that met the requirements for running the MinION and associated software: Windows 7; USB3; SSD; i7 processor) via USB3; the MinION flow cell (R7 flow cell chemistry) was removed from its packaging and inserted into the MinION; the MinKNOW software was opened; a flow cell quality control program was run to assess pore activity; and the flow cell was equilibrated by pipetting two aliquots of EP buffer into the flow cell, incubating for 10 min after each addition. Immediately after the library preparation reaction was complete, 6 µl library, 140 µl EP buffer and 4 µl fuel mix were combined in a fresh tube, mixed by gentle pipetting and loaded onto the MinION according to the manufacturer's instructions. The 48-h sequencing protocol was chosen and the sequencing reaction was started.

MinION data analysis. Once the sequencing run had begun, the Metrichor program was started and the MinION automatically uploads resulting raw data to the Metrichor platform for base calling (workflow r7 2D Basecalling rev 1.3.1), although all data required for base calling are available to the user. The base calling via Metrichor returns a series of 'fast5' files, which contain, among other things, 'events' (i.e., perturbations of current through the pore) and sequence data in fastq format with associated quality scores. Percentage accuracy of the reads was calculated from the phred score[31]. The fast5 data is stored in Hierarchical Data Format v5 (HDF5), which can be viewed via a graphic user interface application such as HDFView (http://www.hdfgroup.org/products/java/release/download.html) or using a package such as h5py for Python (http://www.h5py.org/). Fast5 files from Metrichor were parsed and the sequence extracted in fasta and fastq format using the Poretools v 0.3.0 library[21], commands 'poretools fasta' and 'poretools fastq'.

These reads were mapped to the *de novo* assembly of the H125160566 Illumina data using the LAST aligner with the paraments 'lastal –s 2-T 0 –Q 0 –a 1' (ref. 16). Miscalled bases (assuming no SNPs and perfect alignment) were determined by parsing the alignment output. If two alignments overlapped by >75%, the alignment with the lower E value was discarded.

To investigate whether some k-mers have a higher occurrence of being missed by the MinION detection, all deletions of length k (where $k = 1$ to 6) in the MinION reads were extracted from a SAMtools pileup ('samtools mpileup –BQ0 –f <reference.fasta> <sorted.bam>') file derived from mapping of the MinION reads to the Illumina assembly of the same strain. The distribution of deletion lengths was calculated, as was the frequency of each k-mer (normalized as the number of deletions of a k-mer proportional to the total occurrences of that k-mer) in the H125160566 lllumina assembly. The z-score for each k-mer deletion was calculated according to $z = x - \mu/\sigma$, where x is the proportion for a particular k-mer, μ is the mean and σ is the s.d. for the proportion of deletions across all k-mers of that length. A similar process was used to analyze the inserted sequences, with the exception that their frequency was not normalized against the reference genome. Substitutions were identified from the SAMtools pileup alignment of MinION reads against the H125160566 Illumina assembly.

A hybrid Illumina MinION assembly was generated using SPAdes as above (v3.1.1 (ref. 18) with the '—careful' flag, k-mers of 21, 33, 55 and 77) with the concatenated MinION reads for that isolate input under the–PacBio flag as recommended by the authors.

Determination of island structure and chromosome insertion site. MinION reads in fasta format were mapped to the Illumina contigs of H125160566 that had matched the pHCM1 and pAKU sequence using LAST. The LAST output was transformed to BLAST format using the maf-convert.py script bundled with the LAST aligner. The BLAST results were then parsed to obtain all reads that had mapped to at least three of the pHCM1- H125160566 contigs. This process is described in more detail in the associated Github repository (https://github.com/flashton2003/MinION_analysis), and the codes are present in **Supplementary Software**. Briefly, the output of the LAST mapping is parsed and read-level and hit-level parameters are produced. The read-level parameters produced are: the number of matches to different contigs; and the read length. The hit-level parameters produced are: the Illumina contig matched to; relative orientation of MinION read and Illumina contig; LAST score; hit length; number of positive matches in the hit; number of indels in the hit; the query (MinION read) start position; the query stop; the subject (Illumina contig) start; and the subject stop. Hits are grouped by MinION read, and reads that map to more than one contig are selected. The hits are ordered by the query start, thus making it easy to see the order and orientation of contigs reflected in the MinION reads.

Confirmatory PCR. Primers used to confirm the structure and point of insertion of the island were designed according to the H125160566 sequence (**Supplementary Table 3**).

PCR was performed in a volume of 40 µl containing 20 µl of FideliTaq PCR master mix 2X (Affymetrix, USB), 17 µl water, 1 µl of 10 µM of each primer and 1 µl of 100 ng/µl DNA.

Amplification was performed in a MWG Gradient Thermal Cycler programmed as follows: initial denaturation at 95 °C for 30 s and then 29 cycles of denaturation (30 s, 94 °C), annealing (1 min, 57 °C) and elongation (7 min, 68 °C), with a final elongation step (7 min, 68 °C). The amplified product was electrophoresed at 100 V on a 1.6% agarose gel (type II medium electrophoresis grade) in a 0.5 × TBE buffer (45 mM Tris, 45 mM boric acid, 10 mM EDTA (pH 8)). Following electrophoresis the gel was stained in ethidium bromide (1 µg/ml) and photographed under UV light.

Sequencing of PCR amplicons. PCR amplicons were purified by the ExoSAP-IT cleanup system according to the manufacturer's instructions (Affymetrix/USB). Sanger sequencing of these amplicons was performed using the same PCR primers at the Genomic Service Unit, Public Health England. Analysis was performed using BLAST programs and database services produced by the National Center for Biotechnology Information, Bethesda, MD (http://www.ncbi.nlm.nih.gov/).

26. Parkhill, J. *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**, 848–852 (2001).
27. Holt, K.E. *et al.* Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. *BMC Genomics* **10**, 36 (2009).
28. Grimont, A. & Weill, F. *Antigenic Formulae of the Salmonella Serovars* 9th edn. (World Health Organization, Geneva, 2007).
29. Callow, B. A new phage-typing scheme for *Salmonella typhi-murium*. *J. Hyg. (Lond.)* **57**, 346–359 (1959).
30. Cock, P.J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
31. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 175–185 (1998).

doi:10.1038/nbt.3103