GENOME WATCH

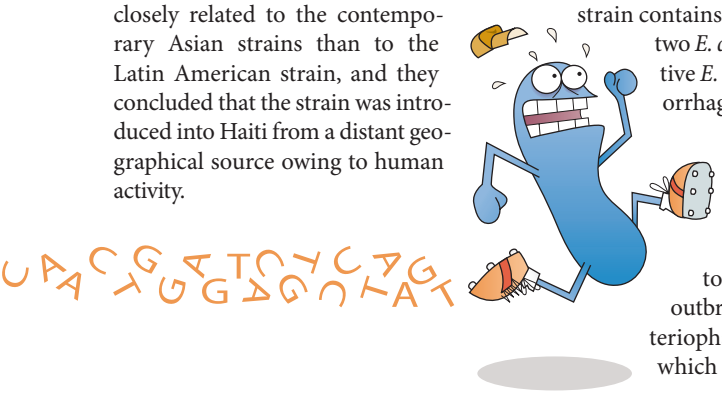
Real-time sequencing

Thomas D. Otto

This month's Genome Watch describes the impact of next-generation sequencing on the 'real-time' analysis of pathogen genomes during outbreaks.

It has been shown previously that second-generation sequencing generates enough data to allow the assembly of large numbers of bacterial genomes. But, depending on the technology used, the time from sampling to assembly can take up to several weeks. To be relevant for clinicians, the turnaround needs to be no more than a few days, but would ideally be a few hours. Recently, several groups have demonstrated that these sorts of timescales are becoming realistic for sequencing and analysing samples from outbreaks. This is because sequencing technologies are continually improving, in terms of price, sequencing speed or library preparation time.

The first example of this faster turnaround comes from an analysis of the recent cholera outbreak in Haiti, in which >93,000 people were infected and >2,100 died[1]. In this analysis, the authors sequenced two *Vibrio cholerae* samples from Haiti, one from Peru and two from Asia. The results were compared with 23 previously sequenced *V. cholerae* strains by building a phylogenetic tree based on 1,588 conserved orthologous genes. The authors showed that the Haitian samples are more closely related to the contemporary Asian strains than to the Latin American strain, and they concluded that the strain was introduced into Haiti from a distant geographical source owing to human activity.



The primary significance of this work is the time that was needed to generate the data; less than 24 hours were required, using the PacBio *RS* sequencing system from Pacific Biosciences, to generate coverage levels of 28–60 times for the five genomes. Moreover, the authors showed that running the machine for just 3 hours would have produced enough coverage of the genomes to identify the key variants that were used in the comparative analysis. However, the authors did not specify the time taken for the analysis. This might become a bottleneck in the future, as a strain can only be quickly genotyped if reference genomes exist, and ensuring that the correct conclusions are drawn may remain time consuming.

The second example comes from this year's outbreak of bloody diarrhoea and haemolytic uraemic syndrome (HUS) caused by *Escherichia coli*, which resulted in >50 deaths and >4,000 cases of infection in Germany. Five different teams sequenced samples from patients (using 454-pyrosequencing, Illumina sequencing by synthesis, Ion Torrent PGM and PacBio *RS* sequencing[2]), leading to two publications to date.

One group sequenced and assembled the samples from the outbreak in less than 62 hours using Ion Torrent PGM sequencers[3]. This group suggested that the outbreak strain contains virulence determinants from two *E. coli* pathotypes: enteroaggregative *E. coli* (EAEC) and enterohaemorrhagic *E. coli* (EHEC). The second pathovar is usually associated with HUS. Phylogenetic analysis revealed that the backbone of the outbreak strain is more closely related to EAEC strains, but that the outbreak strain had acquired a bacteriophage encoding the Shiga toxin, which is more commonly found in EHEC strains. Furthermore, the sequenced samples carry multidrug resistance genes that are commonly found on plasmids. The authors proposed that the outbreak strain is derived from an EAEC progenitor. These findings are in general agreement with those of another group[4] that sequenced two samples using 454-pyrosequencing.

During this outbreak, several groups were able to generate genomic data sets, and more publications are likely to arise. A large proportion of these data was rapidly made publicly available, so the community was able to compare the samples with existing *E. coli* strains, propose reasons for the virulence and drug resistance attributes of the strain and speculate about its origin — all in less than 1 week.

In conclusion, these sequencing projects show that, in the near future, it is feasible that clinicians will be able to access the genomic content of an outbreak strain in close to real time. The main restriction might be the cost of sequencing, which will need to continue falling as it has in recent years.

Thomas D. Otto is at the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.
e-mail: microbes@sanger.ac.uk

doi:10.1038/nrmicro2638

1. Chin, C. S. *et al.* (2011). The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33–42 (2011).
2. Github. *E. coli* O104:H4 genome analysis crowdsourcing. *Github* [online] https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki (accessed 24 Jul 2011).
3. Mellmann, A. *et al.* Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* **6**, e22751 (2011).
4. Brzuszkiewicz, E. *et al.* Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). *Arch. Microbiol.* 29 Jun 2011 (doi:10.1007/s00203-011-0725-6).

Competing interests statement
The author declares no competing financial interests.

© 2011 Macmillan Publishers Limited. All rights reserved