GENOME WATCH

Sherlock Genomes — viral investigator

Sarah E. Smith and Rachael S. Wash

This month's Genome Watch highlights how deep sequencing technologies have vastly reduced the time and prior knowledge needed to generate viral genomes.

Deep sequencing technologies and state-of-the-art bioinformatics techniques have revolutionized the way that RNA viruses, a notoriously variable group of pathogens, can be identified and characterized. Traditionally, specific information about the viral genome was required to design primers for the reverse transcription and amplification of viral RNA prior to sequencing. In addition, a reference genome was often used to assemble short read sequences into a complete genome. However, recent methodological advances have negated some of these prerequisites.

Gall *et al.*¹ developed a method to generate full genomes of HIV-1. They designed a 'pan-HIV-1' reverse transcription PCR primer set, based on sequences available in the Los Alamos HIV Sequence Database, that could give rise to four overlapping amplicons from any HIV-1 virus. The addition of multiplex identifier (MID) adaptors (tags that enable the source of each read to be identified) meant that many samples could be sequenced simultaneously on deep sequencing platforms. By using *de novo* assembly of short reads, the authors showed that this method could generate full genomes for viruses within all four major HIV-1 genetic groups. This shows the potential of deep sequencing for high-throughput studies involving HIV-1 genomes with a broad range of sequence diversity.

The high mutation rate of HIV-1 means that an infected individual can be host to many viral genome variants. Some of these can confer drug resistance, which in HIV-1 can involve mutations

in different genes along the whole genome. The protocol described above offers a read depth that would allow the detection of low-frequency genotypes, facilitating the analysis of rare mutations associated with drug resistance, unlike capillary sequencing, which offers low read coverage.

Even when almost nothing is known about the agent of a viral infection, deep sequencing technologies can identify a pathogen and produce the full genome, fast. In 2012, a Saudi Arabian patient was admitted to hospital suffering from acute severe pneumonia of unknown cause. Preliminary diagnostics indicated that he was infected by a coronavirus (CoV), a single-stranded RNA virus that can infect many species, with bats being an important zoonotic reservoir. Six CoV species have been detected in humans, although only one was known to cause severe disease: severe acute respiratory syndrome (SARS)-CoV.

When SARS-CoV began infecting people in 2002, it took a large team to generate a full genome by capillary sequencing. First, primers were designed for conserved regions of known CoV species, followed by further primer design as the sequence of SARS-CoV was generated. By contrast, in 2012 Van Boheemen *et al.*² used random priming to amplify viral RNA isolated from the Saudi Arabian patient, and then deep-sequenced the genome. Using a massively parallel approach, whereby all the randomly amplified template DNA was sequenced simultaneously on one plate, the team pieced together 90% of the genome in a single run, with an average read depth of 1,006-fold. Primers could then easily be designed to fill in small gaps by capillary sequencing; the whole process took just a few days.

Knowing the sequence of the whole genome

enabled the authors to further characterize this new virus, named human CoV (HCoV)-EMC2012. Phylogenetic analysis was carried out on sequence alignments of genome portions, including those encoding the replicase and other structural proteins, and this revealed that the virus was a novel CoV. The authors showed that HCoV-EMC2012 is most closely related to two members of the C lineage of betacoronaviruses: bat CoV HKU4 and bat CoV HKU5. However, the conserved synteny between HCoV-EMC2012 and these two bat CoV isolates is well below the 90% threshold for members of the same species. More recently, Annan *et al.*³ screened faecal samples from >4,000 bats across Africa and Europe, with the hope of finding viruses related to HCoV-EMC2012. Sequencing facilitated the detection of betacoronaviruses in 25% and 15% of tested *Nycteris* and *Pipistrellus* bats, respectively. Phylogenetic analysis of these viruses indicated that HCoV-EMC2012 originated from bats.

Fast, whole-genome sequencing of viruses is important for improved diagnostics, accurate geographical mapping of viral spread, and drug treatments. As sequencing technologies continue to improve, so will the value of whole-genome sequencing to both clinical and basic virology research.

Sarah E. Smith and Rachael S. Wash are at the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

e-mail: microbes@sanger.ac.uk

doi:10.1038/nrmicro2979

1. Gall, A. *et al.* Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J. Clin. Microbiol.* **50**, 3838–3844 (2012).
2. van Boheemen, S. *et al.* Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *mBio* **3**, 00473–12 (2012).
3. Annan, A. *et al.* Human betacoronavirus 2c EMC/2012-related viruses in bats, Ghana and Europe. *Emerg. Infect. Dis.* Mar 2013 (doi:10.3201/eid1903.121503).

Competing interests statement

The authors declare no competing financial interests.