NEWS & ANALYSIS

GENOME WATCH

What has high-throughput sequencing ever done for us?

Julian Parkhill

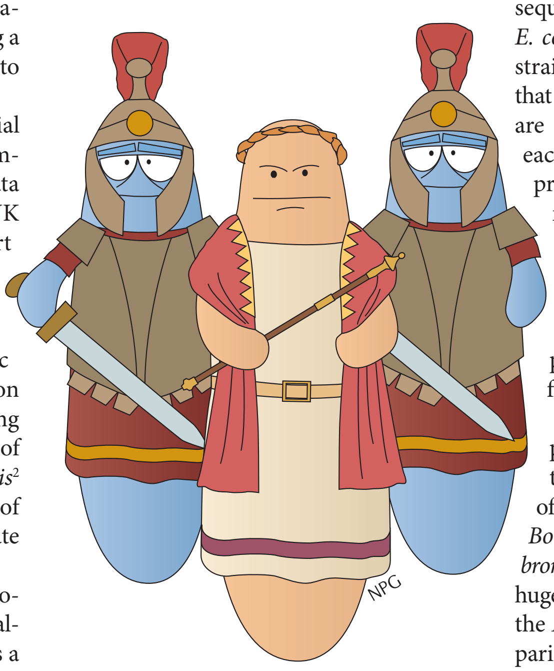
This month's Genome Watch looks back over the past 10 years and highlights how the incredible advances in sequencing technologies have transformed research into microbial genomes.

Genomics has always been a field in constant revolution, but the past decade has possibly seen the most radical changes to date. Ten years ago, microbial genome sequences were hand-crafted things, painstakingly sewn together and lovingly coloured in. Then, over the next few years, several technology companies came and parked their tanks on the lawns of the genome centres. Foremost of these were Solexa (now Illumina), Life Technologies and 454. The new sequencing technologies promised — and delivered — significantly higher throughput and lower cost, although the trade-off for this was shorter reads and a reduction in the cost-effectiveness of manual genome contiguation and correction, essentially mandating a strategic move from few finished genomes to many draft genomes.

Around the turn of the century, microbial genomes could take 1 year or more to complete; generating the original shotgun data cost tens or hundreds of thousands of UK pounds, and therefore the manual effort involved in gap closure and error correction was worthwhile. Genomes for sequencing were chosen carefully, with the first projects focussing on pathogenic bacteria and culminating in the publication of a specific paper for each genome. Among the genomes published in 2003 were those of *Porphyromonas gingivalis*[1], *Bacillus anthracis*[2] and uropathogenic *Escherichia coli*[3], each of which illustrates a key aspect of the state of microbial genomics at that time.

P. gingivalis is a human oral pathogen associated with periodontal disease, and the analysis of the genome of strain W83 (REF. 1) is a good example of the importance of reference-genome sequencing. The genome allowed the reconstruction of the full metabolic capacity of the organism, and the identification of pathways and surface structures that might contribute to host–pathogen interactions and virulence. The key function of these initial projects was to generate reference data and hypotheses for further testing, and indeed this particular paper has been cited >175 times, indicating its importance to the field.

The *B. anthracis* genome[2] also represented a reference strain (Ames) and provided a similar overview of the physiology and virulence mechanisms of the organism, but in this case the project was tightly tied up with geopolitics and the investigation of the American anthrax attacks of 2001. Indeed, publication of the *B. anthracis* str. Ames genome in 2003 was actually foreshadowed by a 2002 article comparing this genome with that of the *B. anthracis* strain isolated from the first victim of those attacks[4]. The 2002 study identified just a few discriminatory polymorphisms between the two strains and represented the first use of genomics in a forensic investigation. The genomics work in this investigation by The Institute for Genomic Research eventually extended to 13 high-coverage or complete genomes, representing a huge effort and cost using the capillary machines available at the time.



However, one of the defining findings of the *B. anthracis* genome comparisons was the lack of diversity between strains. A third single-genome paper published at the end of 2002, detailing the genome of uropathogenic *E. coli* str. CFT073, was one of the first to demonstrate the enormous diversity present within a bacterial species[3]. A comparison of this genome with those of the previously sequenced non-pathogenic laboratory strain, *E. coli* str. K12, and an enterohaemorrhagic strain, *E. coli* O157:H7 str. EDL933, showed that only 39% of the protein-coding sequences are common to all three strains. Although each genome has between 4,288 and 5,060 protein-coding genes, only 2,996 are common to all three strains, with between 585 and 1,623 being unique to each strain. This incredible diversity encompasses horizontally acquired genes and operons responsible for the differing pathogenic potentials of the strains, emphasizing the functional importance of this variation.

The signs of the end of single-genome papers were already present in 2003, with the publication of the complete genomes of three related species (*Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*) in one paper[5]. In contrast to the huge diversity evident from the comparison of the *E. coli* genomes, the *Bordetella* spp. comparison indicated that the highly pathogenic species *B. pertussis* (the agent of whooping cough) evolved from the less pathogenic broad-host-range species *B. bronchiseptica* almost entirely by genome reduction and gene loss. Specific genetic degradation underlying both a loss of host range and an increase in pathogenicity was identified in the *B. pertussis* genome, suggesting an evolutionary trajectory that is typical of the trajectories of several other host-restricted pathogens.

A final landmark paper from early 2004 presented the first metagenomic analysis of a microbial community, that from an acid-mine drainage environment[6]. Given the expense of genome sequencing using capillary approaches, attempting to sequence multiple genomes simultaneously from the environment was brave, albeit considerably facilitated by the low-complexity nature of the community. The data obtained allowed the assembly of near-complete genomes of a bacterium (*Leptospirillum* group II) and an archaeon (*Ferroplasma* type II), along with partial genomes from three other organisms. Using these data, the interconnected metabolic pathways and genetics of the key players in the community were reconstructed.

Ten years later, microbial genomics has branched out in many different directions, driven by the ease and low cost of generating vast volumes of sequence data. This has driven the focus away from finished individual genomes and towards large numbers of draft sequences to identify fine variation, as well as towards using the power of high-throughput sequencing to sample individual cells and whole populations. At the same time, the introduction of bench-top, rapid-turnaround machines has allowed real-time sequencing to be used in clinical investigations.

Whereas the first genome sequence of *P. gingivalis* was that of a cultured reference strain, the latest was generated from a single cell isolated from a hospital sink biofilm. In this study[7], an automated platform was used in which single cells are flow-sorted into 384-well plates, with the genomes then being amplified and subsequently identified via 16S rRNA gene sequencing. Genomes of particular interest were sequenced on the 454 platform, and a near-complete genome of *P. gingivalis* was obtained. Further analysis of this genome identified novel SNPs in the fimbrial subunit (one of the primary virulence factors of this pathogen), along with much larger-scale changes in the capsule biosynthesis and CRISPR (clustered regularly interspaced short palindromic repeats) loci, both of which are potentially involved in biofilm formation.

Comparative genomics has moved from comparing single isolates representative of whole species or pathovars to deep comparisons within highly related, expanding clones. In a recent example, approximately 300 genomes of the hospital-acquired pathogen *Clostridium difficile* ribotype 027 were sequenced[8]. This fine-scale analysis showed that this ribotype is actually composed of two lineages that arose in the early 1990s in North America and subsequently spread globally, probably driven by the early acquisition of resistance to fluoroquinolone antibiotics through point mutation. The depth of the study was such that the routes and timescale of both between- and within-country transmissions could be identified.

Similarly, the ability to use metagenomics has exploded from the analysis of single, simple environments[6] to the sequencing of such large numbers of samples that they can be used as data points in association studies. At the end of 2012, faecal metagenomes from 345 individuals in a case-control study were sequenced to identify variations in the gut microbiome associated with type 2 diabetes (T2D)[9]. The analysis identified >50,000 genetic markers associated with T2D, and these could be grouped into 47 linkage groups probably representing 47 different microbial species, the abundance of which correlated with gut dysbiosis in the T2D group.

The genomic component of the forensic investigations into the anthrax attacks took many months, if not years. However, the high-throughput sequencing technologies now available lend themselves well to rapid-turnaround times using relatively inexpensive bench-top machines. This has led to the growing use of genomics for clinical investigations, such as identifying the transmission routes of pathogens during infectious-disease outbreaks. In one example, a retrospective analysis of a methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak in a special-care baby unit turned into a real-time analysis when the outbreak recurred during the investigation[10]. Within 48 hours, whole-genome sequencing established that the recurrence was due to re-emergence of the original outbreak strain, and implied that asymptomatic carriage by staff might be behind the recurrence. Further screening led to the identification of a colonized staff member, and genome sequencing demonstrated their linkage to the outbreak, again within 48 hours. This study and others like it are leading the way to the routine real-time use of microbial whole-genome sequencing in public health, an application that was unthinkable with the best technology available only 10 years ago.

So, apart from reducing costs, increasing speed, providing access to the entire range of genetic variation from within whole populations to between single cells, and facilitating intervention at the clinical level, what has high-throughput sequencing ever done for us? Clearly, this question answers itself, but where will we go in the next 10 years? It is likely that microbial whole-genome sequencing will be in routine use in outbreak investigations, at both the hospital and community levels. It also seems probable that the use of sequencing will go beyond this and will be used next for the prediction of antibiotic sensitivity and later for rapid syndromic diagnosis. At the research level, genome sequencing will continue its progress from a heroic stand-alone project to an underpinning technology that will be used to support and enable hypothesis-driven science in numerous different ways: population genetics, association studies, environmental exploration, saturation mutagenesis, transcriptomics, DNA–protein interactions, experimental evolution and many others. The future is bright; the future is genomics.

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

doi:10.1038/nrmicro3112
Published online 27 August 2013

1. Nelson, K. E. *et al.* Complete genome sequence of the oral pathogenic bacterium *Porphyromonas gingivalis* strain W83. *J. Bacteriol.* **185**, 5591–5601 (2003).
2. Read, T. D. *et al.* The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **423**, 81–86 (2003).
3. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 17020–17024 (2002).
4. Read, T. D. *et al.* Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* **296**, 2028–2033 (2002).
5. Parkhill, J. *et al.* Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nature Genet.* **35**, 32–40 (2003).
6. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
7. McLean, J. S. *et al.* Genome of the pathogen *Porphyromonas gingivalis* recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. *Genome Res.* **23**, 867–877 (2013).
8. He, M. *et al.* Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nature Genet.* **45**, 109–113 (2013).
9. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
10. Harris, S. R. *et al.* Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.* **13**, 130–136 (2013).

Competing interests statement
The author declares competing financial interests: see Web version for details.