

Microbial bioinformatics 2020

Mark J. Pallen*

Microbiology and Infection Unit, Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK.

Summary

Microbial bioinformatics in 2020 will remain a vibrant, creative discipline, adding value to the ever-growing flood of new sequence data, while embracing novel technologies and fresh approaches. Databases and search strategies will struggle to cope and manual curation will not be sustainable during the scale-up to the million-microbial-genome era. Microbial taxonomy will have to adapt to a situation in which most microorganisms are discovered and characterised through the analysis of sequences. Genome sequencing will become a routine approach in clinical and research laboratories, with fresh demands for interpretable user-friendly outputs. The “internet of things” will penetrate healthcare systems, so that even a piece of hospital plumbing might have its own IP address that can be integrated with pathogen genome sequences. Microbiome mania will continue, but the tide will turn from molecular barcoding towards metagenomics. Crowdsourced analyses will collide with cloud computing, but eternal vigilance will be the price of preventing the misinterpretation and overselling of microbial sequence data. Output from hand-held sequencers will be analysed on mobile devices. Open-source training materials will address the need for the development of a skilled labour force. As we boldly go into the third decade of the twenty-first century, microbial sequence space will remain the final frontier!

Where will microbial bioinformatics be in 2020? Well, let us start by looking back. The last two decades have seen astounding progress in our ability to sequence microbial genomes (Loman and Pallen, 2015). Microbial bioinformatics has by and large kept pace with the resulting data deluge, now clearly emerging as

distinctive discipline in its own right, driven forward by an enthusiastic global community of dedicated microbial bioinformaticians (Loman and Watson, 2013). We can expect this community to continue to grow in the coming years, as microbiologists across the world grapple with established and emerging challenges, including antimicrobial resistance, microbial biodiversity, understanding microbial communities and their genes (microbiomes), synthetic biology and the adoption of genome sequencing as a routine approach in the clinical and research laboratories (Cameron *et al.*, 2014; Koser *et al.*, 2014; Brown *et al.*, 2015; Luheshi *et al.*, 2015; Shanahan, 2015).

It is worth stressing that harnessing bioinformatics to the study of microbial genes, genomes and metagenomes clearly does provide a distinctive challenge – rather than taking aim at the fixed, relatively tractable target of a human, animal or plant genome, instead, here, we have to deal with genomic information derived from thousands of microbial pathogens, millions of commensal species and as many as a billion environmental microbial species (Locey and Lennon, 2016): a distributed and dynamic system of countless billions of genes, many orders of magnitude larger than the human gene set. The resulting deluge of sequence data plainly brings the problems of big data to microbial bioinformatics (Eisenstein, 2015).

Of course, some things are going to stay the same as we approach 2020. Expert microbial bioinformaticians are still primarily going to run command-line programmes on the Linux operating system, typically using pipelines built from open-source software glued together with homebrewed scripts, although these will be written in python rather than Perl (Myhrvold, 2014) or maybe in a yet-to-be-devised scripting language. However, one should not rule out a role for commercial software packages, particularly for applications requiring accredited standard operating procedures. And, unfortunately, in 2020, there is still likely to be a dynamic tension between bioinformatics as an enabling and supporting technology for microbial genomics and bioinformatics as a scientific discipline in its own right, with consequent uncertainties reflected in the career structure and progression for microbial bioinformaticians (Pevzner, 2004; Watson, 2013).

As we approach the end of the current decade, there will be ever more microbial genomes and metagenomes and it remains uncertain whether databases and search strategies will be able to cope. Even in 2016, there is no

Received 29 June, 2016; accepted 8 July, 2016.

*For correspondence. E-mail: m.pallen@warwick.ac.uk; Tel. (+24) 7615 0392.

Microbial Biotechnology (2016) 9(5), 681–686

doi:10.1111/1751-7915.12389

Funding Information

No funding information provided.

© 2016 The Author. *Microbial Biotechnology* published by John Wiley & Sons Ltd and Society for Applied Microbiology.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

easy way to download and search the metagenomic data accumulated by humankind, while BLAST searches of NCBI's supposedly non-redundant database are beginning to strain under the weight of so many identical or near-identical sequences from commonly sequenced species. And this is only going to get worse – for example, by 2020, we are going to have hundreds of thousands, if not millions of genome sequences from key bacterial species, such as *Escherichia coli* or *Mycobacterium tuberculosis*. New approaches to data storage and analysis are going to be required – for example, development of truly non-redundant BLAST databases.

Those interested in microbial epidemiology and microbial population genetics, whether in the research or clinical context, are going to have to cope with the transition from systems based on a handful of gene sequences per organism [e.g. multilocus sequence typing (Maiden, 2006)] to whole-genome approaches (Perez-Losada *et al.*, 2013; Ashton *et al.*, 2016; Pankhurst *et al.*, 2016). Some activities, such as manual curation and annotation of sequences or metadata by the individual enthusiast or by a dedicated research community, are just not sustainable during the scale-up to the million-microbial-genome era. Instead, machine learning and artificial intelligence may have to fill the gap (Yip *et al.*, 2013). And, sadly, the problem of lack of continuity of funding for databases and other bioinformatics resources is probably not going to be solved in the next few years (Parkhill *et al.*, 2010).

After a period of lively competition (Loman *et al.*, 2012), the marketplace for high-throughput sequencing has recently settled into a state of near-monopoly, with Illumina short-read sequencing dominating the field. While this technology may be highly suited to applications such as re-sequencing of genomes, where attention is focused on detection of single nucleotide variants, it is poorly able to cope with the riotous diversity of microbial genomes and metagenomes, particularly when looking at mobile genetic elements or accessory genomes (Stoesser *et al.*, 2014). Single-molecule long-read technologies are already available at the time of writing in 2016 (e.g. from Pacific Biosystems or Oxford Nanopore), but are still waiting in the wings, despite progress in showing proof of principle applications (Loman *et al.*, 2015; Quick *et al.*, 2015, 2016) and developing bioinformatics tools dedicated to these approaches (Loman and Quinlan, 2014; Rhoads and Au, 2015; Watson *et al.*, 2015). It remains unclear how far this will change in the coming years – will existing long-read technologies take centre stage; or will new players enter the market? Whatever happens, both established and novel sequencing approaches are going to drive the development of new bioinformatics tools. Similarly, existing and new laboratory techniques focused on single-cell genomics and transcriptomics (Lasken and McLean, 2014) or

approaches to the functional genomics of microbes, such as RNA-Seq (Creecy and Conway, 2015) or Tn-Seq (Kwon *et al.*, 2016), will create a continuing demand for new software.

Microbial genomics and metagenomics are hurtling full-steam ahead into the clinical arena and into efforts to map the global landscape of microbial biodiversity (Pallen *et al.*, 2010; Didelot *et al.*, 2012; Robinson *et al.*, 2013; Kyrpides *et al.*, 2014; Brown *et al.*, 2015; Luheshi *et al.*, 2015; Spang *et al.*, 2015). In both settings, it is clear that microbial taxonomy, with its polyphasic approach that requires laboratory culture and phenotypic investigation, is already broken and simply will not cope with an era in which most microorganisms are going to be identified and characterized through the analysis of macromolecular sequences (Chun and Rainey, 2014; Ramasamy *et al.*, 2014; Thompson *et al.*, 2015; Baltrus, 2016). Let us hope a new taxonomy is born by 2020, driven by – and driving – an explosion of creativity in the bioinformatics of microbial diversity (Varghese *et al.*, 2015). Similarly, we can expect new opportunities and challenges arising from synthetic biology's desire to shift from merely reading to actively writing DNA sequences, whether in the creation of synthetic microorganisms or in novel approaches to data handling and storage (Goldman *et al.*, 2013; Boeke *et al.*, 2016; Hutchison *et al.*, 2016).

The collision between microbial bioinformatics and human health care has already led to the development of new tools and this creative clash of disciplines is going to transform the outlook for bioinformaticians. Here, we are likely to see improvements in the tools for analysing microbial genomic epidemiology – for example, in tackling the growing realization that the cellular populations of pathogens, just like those of cancers, may well be clonal but that does not mean that they are necessarily homogeneous (Jamal-Hanjani *et al.*, 2015; Paterson *et al.*, 2015). New models and new software will also need to recognize the problem of within-host pathogen diversity and fact the pathogen phylogenies do not map simply on to transmission chains (Didelot *et al.*, 2014; Gardy, 2016). But we hope that even if, as some have suggested, within-host bacterial diversity makes it harder to reconstruct transmission networks, this will no longer pose a insuperable problem in 2020 (Worby *et al.*, 2014).

Integration of microbial genomics and bioinformatics into clinical practice will bring fresh demands that pipelines not only be credible, robust and reproducible but produce easily interpretable clinician-friendly outputs, for example, the programme Mykrobe, which analyses genomes from *Staphylococcus aureus* and *M. tuberculosis* (Bradley *et al.*, 2015). Integration of sequence data with clinical metadata will be difficult, particularly as precision medicine is going to need precise ontologies (Dugan *et al.*, 2014) – for example, in analysing a hospital

outbreak, the next generation of NHS bioinformaticians are going to have to be highly attuned to, say, the difference between a 'bed' and a 'bed space'. They will be assisted in these efforts as the 'internet of things' penetrates healthcare systems, so that a patient, an instrument or even a piece of hospital furniture or plumbing will have its own IP address and GPS-savvy chip and all providing information that can be integrated with pathogen genome sequences (Hao and Wang, 2015).

Metagenomics as a diagnostic approach is likely to move closer to routine practice (Loman *et al.*, 2013; Doughty *et al.*, 2014; Pallen, 2014; Wilson *et al.*, 2014), but reliably disentangling pathogen genomes from metagenomes – particularly if short-read technologies still dominate the field – is going to present a formidable challenge (Alneberg *et al.*, 2014).

The current mania for microbiomes looks set to continue, so new bioinformatics tools to detect 'sick microbiomes' and link them to disease states are going to be required (Forslund *et al.*, 2015). Perhaps by 2020, the tide will have turned away from molecular barcoding approaches, epitomized by what has been called the one-eyed king (Forney *et al.*, 2004), 16S ribosomal RNA gene sequences, towards the more widespread adoption of shotgun metagenomics (Jovel *et al.*, 2016). If so, new tools will be required to translate metagenomes into the standard outputs of microbial ecology (rarefaction curves, diversity indices etc.). Similarly, new tools will emerge at the interface between metagenomics, metatranscriptomics, metabolomics and systems biology (Franzosa *et al.*, 2014).

A potential concern is the growth of a wild frontier of microbial genome and microbiome analyses performed by non-experts, hand-cranking data through pipelines that they do not fully understand and then naively interpreting results without engaging the healthy scepticism of the seasoned expert (Bhatt *et al.*, 2013; Branton *et al.*, 2013; Laurence *et al.*, 2014; Salter *et al.*, 2014; Strong *et al.*, 2014; Ackelsberg *et al.*, 2015; Afshinnekoo *et al.*, 2015). Eternal vigilance is likely to be the price of containing the equivalent of microbial genomic astrology!

When it comes to provisioning of hardware and software, microbial bioinformatics is being pulled away from the archetypal self-administered server or cluster run by a single-user or single-research-group. In one direction lies the development of apps for mobile devices (Rose *et al.*, 2013; Wong *et al.*, 2013; Nguyen *et al.*, 2014), in parallel with the rise of palmtop sequencing (Quick *et al.*, 2016), so that in 2020, sequencing and analysis may well take place out in the field and/or much closer to the patient. The centralizing efforts of national or transnational projects are pulling in another direction, aimed at standardizing protocols for creating, storing and analysing microbial sequence data, particularly for healthcare, although it perhaps unlikely that such efforts will have

arrived at a stable agreed global solution by 2020 (Moran-Gilad *et al.*, 2015).

Another potential trend is the rise of crowd-sourced microbial bioinformatics analyses, performed by bioinformaticians around the globe – there have already been proof-of-principle cases (Rohde *et al.*, 2011; Gardy *et al.*, 2015) and we are likely to see more of this by 2020, particularly in response to public-health emergencies. Similarly, microbial bioinformaticians are likely to embrace cloud computing (Drake, 2014), which brings economies of scale in effort and costs and liberates end users from the hassle of maintaining systems and setting up commonly used software, while also delivering improvements in the sharing of pipelines and data, which in turn will enhance the reproducibility of bioinformatics analyses. One promising example here is the UK's Cloud Infrastructure for Microbial Bioinformatics (CLIMB) project, which provides end users in the microbiology community with access to virtual machines provided via the OpenStack open-source cloud-computing environment (Connor *et al.*, 2016).

One final challenge for microbial bioinformatics in the run-up to 2020 is meeting the need for training and the development of a skilled labour force (Via *et al.*, 2013; Watson-Haigh *et al.*, 2013). Cloud computing may play a contribution here, in providing a standardized environment for workshops and hackathons as well as for research groups. Similarly, we can expect to see a continued rise in open-source training materials suitable for use in bioinformatics boot camps and the development of new workflow and data integration systems, such as the Genomics Virtual Laboratory (Afgan *et al.*, 2015).

In conclusion, microbial bioinformatics in 2020 will remain a vibrant, creative discipline, adding value to the ever-growing flood of new sequence data while embracing new technologies and new approaches. As we boldly go into the third decade of the 21st century, microbial sequence space will remain the final frontier!

Acknowledgements

I thank Nick Loman for reading a draft of this manuscript and providing comments.

Conflict of interest

None declared.

References

- Ackelsberg, J., Rakeman, J., Hughes, S., Petersen, J., Mead, P., Schriefer, M., *et al.* (2015) Lack of evidence for plague or anthrax on the New York City subway. *Cell Syst* **1**: 4–5.
- Afgan, E., Sloggett, C., Goonasekera, N., Makunin, I., Benson, D., Crowe, M., *et al.* (2015) Genomics Virtual

- Laboratory: a practical bioinformatics workbench for the cloud. *PLoS ONE* **10**: e0140829.
- Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., *et al.* (2015) Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst* **1**: 72–87.
- Aineberg, J., Bjamason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., *et al.* (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**: 1144–1146.
- Ashton, P.M., Nair, S., Peters, T.M., Bale, J.A., Powell, D.G., Painset, A., *et al.* (2016). Identification of Salmonella for public health surveillance using whole genome sequencing. *PeerJ* **4**: e1752.
- Baltrus, D.A. (2016) Divorcing strain classification from species names. *Trends Microbiol* **24**: 431–439.
- Bhatt, A.S., Freeman, S.S., Herrera, A.F., Pedamallu, C.S., Gevers, D., Duke, F., *et al.* (2013) Sequence-based discovery of *Bradyrhizobium enterica* in cord colitis syndrome. *N Engl J Med* **369**: 517–528.
- Boeke, J.D., Church, G., Hessel, A., Kelley, N.J., Arkin, A., Cai, Y., *et al.* (2016) The genome project-write. *Science* **353**: 126–127.
- Bradley, P., Gordon, N.C., Walker, T.M., Dunn, L., Heys, S., Huang, B., *et al.* (2015) Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun* **6**: 10063.
- Branton, W.G., Ellestad, K.K., Maingat, F., Wheatley, B.M., Rud, E., Warren, R.L., *et al.* (2013) Brain microbial populations in HIV/AIDS: alpha-proteobacteria predominate independent of host immune status. *PLoS ONE* **8**: e54673.
- Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., *et al.* (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**: 208–211.
- Cameron, D.E., Bashor, C.J., and Collins, J.J. (2014) A brief history of synthetic biology. *Nat Rev Microbiol* **12**: 381–390.
- Chun, J., and Rainey, F.A. (2014) Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int J Syst Evol Microbiol* **64**: 316–324.
- Connor, T.R., Loman, N.J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., *et al.* (2016) CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. doi: <http://dx.doi.org/10.1101/064451>.
- Creecy, J.P., and Conway, T. (2015) Quantitative bacterial transcriptomics with RNA-seq. *Curr Opin Microbiol* **23**: 133–140.
- Didelot, X., Bowden, R., Wilson, D.J., Peto, T.E., and Crook, D.W. (2012) Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet* **13**: 601–612.
- Didelot, X., Gardy, J., and Colijn, C. (2014) Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol* **31**: 1869–1879.
- Doughty, E.L., Sergeant, M.J., Adetifa, I., Antonio, M., and Pallen, M.J. (2014) Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ* **2**: e585.
- Drake, N. (2014) Cloud computing beckons scientists. *Nature* **509**: 543–544.
- Dugan, V.G., Emrich, S.J., Giraldo-Calderon, G.I., Harb, O.S., Newman, R.M., Pickett, B.E., *et al.* (2014) Standardized metadata for human pathogen/vector genomic sequences. *PLoS ONE* **9**: e99979.
- Eisenstein, M. (2015) Big data: the power of petabytes. *Nature* **527**: S2–S4.
- Forney, L.J., Zhou, X., and Brown, C.J. (2004) Molecular microbial ecology: land of the one-eyed king. *Curr Opin Microbiol* **7**: 210–220.
- Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., *et al.* (2015) Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**: 262–266.
- Franzosa, E.A., Morgan, X.C., Segata, N., Waldron, L., Reyes, J., Earl, A.M., *et al.* (2014) Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci U S A* **111**: E2329–E2338.
- Gardy, J.L. (2016) Translating phylogeny into action for HIV surveillance. *Lancet HIV* **3**: e196–e197.
- Gardy, J., Loman, N.J., and Rambaut, A. (2015) Real-time digital pathogen surveillance — the time is now. *Genome Biol* **16**: 17.
- Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B., and Birney, E. (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**: 77–80.
- Hao, A., and Wang, L. (2015) Medical device integration model based on the internet of things. *Open Biomed Eng J* **9**: 256–261.
- Hutchison, C.A.R., Chuang, R.Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science* **351**: aad6253.
- Jamal-Hanjani, M., Quezada, S.A., Larkin, J., and Swanton, C. (2015) Translational implications of tumor heterogeneity. *Clin Cancer Res* **21**: 1258–1266.
- Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., *et al.* (2016) Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front Microbiol* **7**: 459.
- Koser, C.U., Ellington, M.J., and Peacock, S.J. (2014) Whole-genome sequencing to control antimicrobial resistance. *Trends Genet* **30**: 401–407.
- Kwon, Y.M., Ricke, S.C., and Mandal, R.K. (2016) Transposon sequencing: methods and expanding applications. *Appl Microbiol Biotechnol* **100**: 31–43.
- Kypides, N.C., Hugenholtz, P., Eisen, J.A., Woyke, T., Goker, M., Parker, C.T., *et al.* (2014) Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol* **12**: e1001920.
- Lasken, R.S., and McLean, J.S. (2014) Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet* **15**: 577–584.
- Laurence, M., Hatzis, C., and Brash, D.E. (2014) Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS ONE* **9**: e97876.
- Locey, K.J., and Lennon, J.T. (2016) Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A* **113**: 5970–5975.

- Loman, N.J., and Pallen, M.J. (2015) Twenty years of bacterial genome sequencing. *Nat Rev Microbiol* **13**: 787–794.
- Loman, N.J., and Quinlan, A.R. (2014) Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30**: 3399–3401.
- Loman, N., and Watson, M. (2013) So you want to be a computational biologist? *Nat Biotechnol* **31**: 996–998.
- Loman, N.J., Constantinidou, C., Chan, J.Z., Halachev, M., Sergeant, M., Penn, C.W., *et al.* (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* **10**: 599–606.
- Loman, N.J., Constantinidou, C., Christner, M., Rohde, H., Chan, J.Z., Quick, J., *et al.* (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* **309**: 1502–1510.
- Loman, N.J., Quick, J., and Simpson, J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* **12**: 733–735.
- Luheshi, L.M., Raza, S., and Peacock, S.J. (2015) Moving pathogen genomics out of the lab and into the clinic: what will it take? *Genome Med* **7**: 132.
- Maiden, M.C. (2006) Multilocus sequence typing of bacteria. *Annu Rev Microbiol* **60**: 561–588.
- Moran-Gilad, J., Sintchenko, V., Pedersen, S.K., Wolfgang, W.J., Pettengill, J., Strain, E., and Hendriksen, R.S. (2015) Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. *BMC Infect Dis* **15**: 174.
- Myhrvold, C. (2014). *The Fall of Perl, The Web's Most Promising Language*. <http://www.fastcompany.com/3026446/the-fall-of-perl-the-webs-most-promising-language>
- Nguyen, P.V., Verma, C.S., and Gan, S.K. (2014) DNAApp: a mobile application for sequencing data analysis. *Bioinformatics* **30**: 3270–3271.
- Pallen, M.J. (2014) Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology* **141**: 1856–1862.
- Pallen, M.J., Loman, N.J., and Penn, C.W. (2010) High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr Opin Microbiol* **13**: 625–631.
- Pankhurst, L.J., Del Ojo Elias, C., Votintseva, A.A., Walker, T.M., Cole, K., Davies, J., *et al.* (2016). Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med* **4**: 49–58.
- Parkhill, J., Birney, E., and Kersey, P. (2010) Genomic information infrastructure after the deluge. *Genome Biol* **11**: 402.
- Paterson, G.K., Harrison, E.M., Murray, G.G., Welch, J.J., Warland, J.H., Holden, M.T., *et al.* (2015) Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. *Nat Commun* **6**: 6560.
- Perez-Losada, M., Cabezas, P., Castro-Nallar, E., and Crandall, K.A. (2013) Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infect Genet Evol* **16**: 38–53.
- Pevzner, P.A. (2004) Educating biologists in the 21st century: bioinformatics scientists versus bioinformatics technicians. *Bioinformatics* **20**: 2159–2161.
- Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., *et al.* (2015) Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol* **16**: 114.
- Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., *et al.* (2016) Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**: 228–232.
- Ramasamy, D., Mishra, A.K., Lagier, J.C., Padhmanabhan, R., Rossi, M., Sentausa, E., *et al.* (2014) A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* **64**: 384–391.
- Rhoads, A., and Au, K.F. (2015) PacBio sequencing and its applications. *Genomics Proteomics Bioinform* **13**: 278–289.
- Robinson, E.R., Walker, T.M., and Pallen, M.J. (2013) Genomics and outbreak investigation: from sequence to consequence. *Genome Med* **5**: 36.
- Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N.J., Hentschke, M., *et al.* (2011) Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N Engl J Med* **365**: 718–724.
- Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res* **41**: D475–D482.
- Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., *et al.* (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12**: 87.
- Shanahan, F. (2015) Separating the microbiome from the hyperbolome. *Genome Med* **7**: 17.
- Spang, A., Saw, J.H., Jorgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., *et al.* (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**: 173–179.
- Stoesser, N., Giess, A., Batty, E.M., Sheppard, A.E., Walker, A.S., Wilson, D.J., *et al.* (2014) Genome sequencing of an extended series of NDM-producing *Klebsiella pneumoniae* isolates from neonatal infections in a Nepali hospital characterizes the extent of community-versus hospital-associated transmission in an endemic setting. *Antimicrob Agents Chemother* **58**: 7347–7357.
- Strong, M.J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., *et al.* (2014) Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog* **10**: e1004437.
- Thompson, C.C., Amaral, G.R., Campeao, M., Edwards, R.A., Polz, M.F., Dutilh, B.E., *et al.* (2015) Microbial taxonomy in the post-genomic era: rebuilding from scratch? *Arch Microbiol* **197**: 359–370.
- Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., Kyrpides, N.C., and Pati, A. (2015) Microbial species delineation using whole genome sequences. *Nucleic Acids Res* **43**: 6761–6771.
- Via, A., Blicher, T., Bongcam-Rudloff, E., Brazas, M.D., Brooksbank, C., Budd, A., *et al.* (2013) Best practices in bioinformatics training for life scientists. *Brief Bioinform* **14**: 528–537.
- Watson, M. (2013) *A Guide for the Lonely Bioinformatician*. <http://www.opiniomics.org/a-guide-for-the-lonely-bioinformatician/>

- Watson, M., Thomson, M., Risse, J., Talbot, R., Santoyo-Lopez, J., Gharbi, K., and Blaxter, M. (2015) poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* **31**: 114–115.
- Watson-Haigh, N.S., Shang, C.A., Haimel, M., Kostadima, M., Loos, R., Deshpande, N., *et al.* (2013) Next-generation sequencing: a challenge to meet the increasing demand for training workshops in Australia. *Brief Bioinform* **14**: 563–574.
- Wilson, M.R., Naccache, S.N., Samayoa, E., Biagtan, M., Bashir, H., Yu, G., *et al.* (2014) Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med* **370**: 2408–2417.
- Wong, E.D., Karra, K., Hitz, B.C., Hong, E.L., and Cherry, J.M. (2013) The YeastGenome app: the *Saccharomyces Genome Database* at your fingertips. *Database (Oxford)* **2013**: bat004.
- Worby, C.J., Lipsitch, M., and Hanage, W.P. (2014) Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput Biol* **10**: e1003549.
- Yip, K.Y., Cheng, C., and Gerstein, M. (2013) Machine learning and genome annotation: a match meant to be? *Genome Biol* **14**: 205.