Genome Medicine

MEETING REPORT

Towards the integration of genomics, epidemiological and clinical data

Victor V Solovyev[1]* and Tatiana V Tatarinova[2]

Abstract

A report on 'A Wellcome Trust Scientific Conference: Applied Bioinformatics and Public Health Microbiology 2011', Hinxton, Cambridge, 1-3 June, 2011.

The microbiome in health and disease

The 'Applied Bioinformatics and Public Health Microbiology 2011' conference was an exciting opportunity to bring together scientists from diverse disciplines, among them microbiologists, virologists, bioinformaticians and public health specialists. Reports of recent scientific and technological advances in bioinformatics and microbiology were presented, inspiring discussions centered around the applications of state-of-the art bioinformatics tools to public health. The conference opened with William Wade (King's College London Dental Institute, UK) discussing the role of the human oral microbiome in health and disease. The oral microbiome consists of more than 600 different taxa of bacteria, viruses, fungi and protozoa, with every individual person carrying a unique microbiome that can impact health and wellbeing. Indeed, most inhabitants of the human mouth are bacteria, approximately half of which remain unidentified. Different oral diseases are associated with characteristic microbiomes. Next generation sequencing (NGS) has been applied to saliva and plaque samples, leading to the discovery of new taxa that have a potential role in disease. Wade and colleagues developed The Human Oral Microbiome Database (HOMD; http://www.homd.org/), which offers user-friendly tools for viewing publicly available oral bacterial genomes.

A tsunami of sequence data

The exciting discussion on the human oral microbiome was followed by informative reports on new sequencing technologies. Jason Mayers (Ion Torrent, USA) described an Ion Torrent device that includes a semiconductor chip capable of directly translating a chemical sequence into digital information. This bench-top sequencing platform does not use light and delivers information quickly at a low cost. Geoffrey Smith (Illumina Cambridge Ltd, UK) introduced the MiSeq™ sequencing system, which generates over a gigabyte of data from 2 × 150 base pair reads in just over 24 hours. By combining MiSeq and a novel fast library generation technology (Nextera), researchers were able to cover the path from sample to answer within a day. This technology was successfully applied to identify drug-resistant and drug-sensitive bacterial strains. The parade of technological advances was joined by Andrew Kasarskis (Pacific Biosciences, USA) who presented the single molecule real-time (SMRT) sequencer. This technology provides information on the kinetics of polymerization and on modification status across a population of individual molecules. It can produce reads separated by long sequence segments and, in combination with MiSeq data, could significantly improve the assembly of complex genomes. Chinnappa Kodira (Roche Applied Science, USA) described the application of the Roche 454 sequencer for investigating *Salmonella* outbreaks, for determining mutations of HIV that were not detected by standard sequencing, and in a multi-faceted study of the hepatitis C virus. Kodira stressed the importance of transcriptome sequencing for resolving alternative splicing gene variants. Julian Parkhill (Wellcome Trust Sanger Institute, UK) discussed the use of high-throughput sequencing for drafting the whole-genome sequence of hundreds of bacterial strains simultaneously. Parkhill's team was engaged in over 100 projects on organisms ranging from humans to bacteria and collaborates widely within the UK and internationally.

Applications of NGS

A number of presentations highlighted the application of NGS techniques. Katja Lehmann (Center for Ecology and Hydrology, UK) reported that analysis pipelines can produce different results for the same NGS input data, and therefore it is dangerous to treat them as black boxes. Rebecca Jones (Animal Health and Veterinary Laboratories

*Correspondence: victor@cs.rhul.ac.uk
[1]Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK
Full list of author information is available at the end of the article

Agency, UK) evaluated the annotation accuracy of five popular genome annotation pipelines - RAST, FgenesB, MG/ER, IGS and xBASE - using manual annotations of *Salmonella typhimurium* genomes as a gold standard. She reported that 50 to 80% of bacterial proteins can be identified from NGS data. Keith Jolley (University of Oxford, UK) described the freely accessible Bacterial Isolate Genome Sequence Database (BIGSdb; http://pubmlst.org/software/database/bigsdb/).

Helena Seth-Smith (Wellcome Trust Sanger Institute, Hinxton, UK) discussed the application of NGS to analyze a major cause of sexually transmitted disease, *Chlamydia trachomatis,* which gave an insight into population structures and the evolution of the bacterium. Angela McCann (University College Cork, Ireland) presented an analysis of genetic diversity in *Salmonella enterica*, a pathogen in cattle and poultry disease, using paired-end Illumina reads to detect SNPs. The topology inferred from SNP analysis is consistent with epidemiological data and therefore NGS sequencing can be used as a predictive tool at the onset of an outbreak.

Genome sequencing has provided a wealth of information on the genetics and biology of the versatile bacterium *Escherichia coli.* Ulrich Dobrindt (University of Münster, Germany) presented his NGS analysis of *E. coli* population diversity and microevolution. This approach has the potential to improve the typing of pathogenic *E. coli* variants. Rolf Kaas (Technical University of Denmark, Denmark) compared nucleotide identity amongst 171 *E. coli* strains. Kaas showed that fast- and slow-evolving genes should be treated differently in the development of a dynamic typing method that can be used in different outbreak scenarios.

The hypervirulent bacterium *Clostridium difficile* is the most serious cause of antibiotic-associated diarrhea, frequently resulting from eradication of the normal gut flora by antibiotics. Miao He (Wellcome Trust Sanger Institute, UK) described the application of whole-genome sequencing (WGS) for analysis of genetic variation among a global collection of 370 different strains of *C. difficile*. A combination of genomics data with time/space and clinical data allows the identification of both long-range transmission events between and within countries and routes of transmission within hospitals. It also makes it possible to differentiate re-lapsing disease and re-infection. This topic was continued by Tim Peto (John Radcliffe Hospital, UK). Peto demonstrated that WGS can provide more information than multi-locus sequence typing (MLST). Brendan Wren (London School of Hygiene & Tropical Medicine, UK) presented a whole-genome comparison of *Campylobacter jejuni*, *Yersinia enterocolitica* and *C. difficile* that can be useful for understanding how bacterial pathogens transmit and evolve. Understanding the mechanism of pathogenicity will ultimately lead to the development of radical interventions to stop the spread of epidemics. Oliver Pybus (Oxford University, UK) stressed that NGS should not be restricted to bacteria by describing its application for RNA viruses, including common human RNA viral pathogens such as HIV, influenza and hepatitis C. He highlighted the importance of the timely development of data analysis methods to take advantage of the exponential growth in available sequence data.

It was especially refreshing to see many excellent presentations from talented early-career female scientists. Rebecca Gladstone (Health Protection Agency, UK) gave a talk on the phenotypic and genotypic diversity of *Streptococcus pneumoniae* isolates seen during the introduction of conjugate vaccines. She emphasized that WGS can uncover relationships between strains that are not detected by other methods (such as MLST). Another remarkable young scientist, Jacqueline Chan (Centre for Systems Biology, UK), outlined the use of high-throughput sequencing to track pathogenic bacterial infections in hospitals and the use of phage therapy as an alternative to antibiotics. Jennifer Gardy (British Columbia Center for Disease Control, Canada) addressed the important issue of analyzing the dynamics of tuberculosis outbreaks. She demonstrated how a combination of epidemiologic data, WGS and social-network analysis can be used to determine the origins and transmission dynamics of the outbreak. Supang Martin (Health Protection Agency, UK) investigated the development and evolution of drug resistance in the *pol* gene of HIV-1 using phylogenetic analysis. She demonstrated the existence of recombination between viruses from different subpopulations, and discussed the complex dynamics in the development of drug resistance in patients undergoing treatments with protease, reverse transcriptase and integrase inhibitors. Pimlapas Leekitcharoenphon (Technical University of Denmark, Denmark) presented an interesting study on genomic variation in core genes of *S. enterica* genomes. The core genes are conserved in most of the subspecies and variation within them is essential for bacterial adaptation.

Future perspectives

The members of the scientific organizing committee initiated an open session in which attendees discussed which bioinformatics tools would be the best in a clinical environment. Participants indicated a strong demand for the organization of adequate data-sharing (of reference genomes and strains, for example), a desire to link relevant data types, and the need to devise standard data generation and analysis protocols. Many errors have been found in GenBank records, and it was advised that researchers should provide indicators of the quality of assembly. As remarked by Nick Loman (University of

Birmingham, UK), researchers must not blindly trust anything in bioinformatics. Another issue is that information about mobile elements of the genome is removed from or ignored in the datasets compiled from many phylogenetic studies of genome sequence variations. But mobile elements can carry drug-resistance determinants and therefore should be examined. Many scientists were concerned about the speed with which new technologies are propagated into clinics (it took approximately 10 years for PCR to reach hospitals). To facilitate the propagation of NGS into patient care, we need to offer educational bioinformatics courses (including online modules and on-site training) for clinical microbiologists. To further this field, a formalized network and organized sharing of best practice should be established.

Abbreviations

MLST, multi-locus sequence typing; NGS, next generation sequencing; SNP, single nucleotide polymorphism; WGS, whole genome sequencing.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK. [2]Division of Mathematics and Statistics, University of Glamorgan, Pontypridd CF37 1DL, UK.

Published: 27 July 2011

doi:10.1186/gm264
Cite this article as: Solovyev VV, Tatarinova TV: Towards the integration of genomics, epidemiological and clinical data. *Genome Medicine* 2011, **3**:48.