BIOINFORMATICS: HOW TO STANDARDISE AND ASSEMBLE RAW DATA INTO SEQUENCES. WHAT DOES IT MEAN FOR A LABORATORY TO USE SUCH TECHNOLOGIES?

Joseph Hughes, PhD

MRC – University of Glasgow Centre for Virus Research, Room 117, Stoker Building, Garscube Estate, Bearsden Road, Glasgow G61 1QH, United Kingdom
Tel: +44-0141 330 ext 4019; joseph.hughes@glasgow.ac.uk

Over the past decade, a number of different sequencing technologies have produced various types of high throughput sequence (HTS) data. The decreasing cost and faster turnaround time to produce these data means that pathogen whole-genome sequencing can now cross the divide between research and the practice of diagnostics. HTS data holds the potential to transform our understanding of the global spread of antimicrobial resistance and to trace the spread of diseases. However, one aspect that should not be ignored when dealing with HTS data is bioinformatics analysis. The sheer volume of data often means that data analysis cannot easily be sustained by the wet-lab researcher alone but requires capacity building in computer skills and bioinformatics expertise.

The main steps in bioinformatics analysis of HTS data involve cleaning/filtering the data, sequence assembly and genome annotation. Three examples of OIE notifiable animal diseases (equine influenza, avian influenza and infection with ranavirus) will be used to illustrate some of the veterinary applications of HTS data: whole viral genome reconstruction, characterisation of intra-host variability, and virus discovery. These examples will be used to illustrate the multitude of software that has been developed for sequence assembly. Whilst the availability of these different methods help with the analysis of HTS data, the algorithms for quality control and genome assembly sometimes present limitations that require custom scripts or in-house resolution of bioinformatic problems caused by specific needs. Some of the problems associated with HTS data analysis will be discussed.

Guidelines for the quality control and assembly of raw sequence data into whole genome sequences of pathogens will be presented as well as the accepted standards for genome projects.