Next-gen sequencing – a disruptive technology to transform freedom-from-disease (FFD) surveillance

PA DURR[1]*, J ROBERTS[2]
[1]CSIRO-Australian Animal Health Laboratory, Geelong, Australia; [2]CSIRO Health & Biosecurity, Canberra, Australia

*peter.durr@csiro.au

Abstract
Next-gen sequencing ("NGS") is rapidly transforming many areas of animal health, from diagnostics to surveillance. As such it is a "disruptive technology", wherein the challenge is to find new uses to tackle established, and often difficult problems. Here we show an example of its use in providing additional support for a freedom-from-disease ("FFD") survey which demonstrated that Australia is free from the bee disease caused by deformed wing virus. This is the first report we are aware of in which NGS has been used for FFD surveillance.

Keywords: *Bee diseases; deformed wing virus; next-gen sequencing; sacbrood virus*

Introduction
Currently animal health surveillance is being subjected to two major "disruptive technologies" which will overturn established dogmas and norms. The first is "Big Data" which uses databases and high-performance computing to enable the analysis and integration of complex datasets. The potential of this is well recognised in the animal health surveillance community, particularly with respect to the analysis of animal movement datasets.

The second of these disruptive technologies is the sequencing of pathogen genomes, and more specifically the use of next generation ("next gen") high-throughput sequencing. Although this technology is only about 10 years old, already it is enabling the low cost and rapid sequencing of the whole genomes of viruses and bacteria.

The advent of next-gen sequencing ("NGS") has created the situation where the capability of *generating* data is now far ahead of the *use* of this data, and thus the challenge is how to make it relevant. Here we tackle the particular problem of the use of NGS - and existing data stored in large sequence repositories such as *GenBank* - to assist in the specific surveillance problem of proving a country free of viral pathogens.

To make this practical, we will use as an example of a recent survey we undertook to prove Australia to be free (or not) from several viral bee pathogens. Furthermore, although our focus is on the use of sequence data in FFD surveillance, an important objective is to relate this to conventional approaches, and as such we will also describe the FFD survey from which the samples were obtained to undertake the NGS.

Materials and methods
The 2014-2015 national bee pathogen survey
Australia is fortunate in being one of the few countries that is free of the *Varroa destructor* mite. There is general confidence in this statement, as the disease caused by the mite, "varroosis", is readily recognisable to beekeepers on account of the high mortality it causes upon introduction to naïve hives. However, a survey of eastern Australian bees in the 1980s – using serological methods – did show that they were infected with a range of generally non-pathogenic viruses, including sacbrood virus (SBV), black queen cell virus (BQCV), chronic bee paralysis virus (CBPV) and Kashmir bee virus (KBV) (1). On the other hand, there are some potentially pathogenic viruses, particularly deformed wing virus (DWV) and slow bee paralysis virus (SBPV) which have never been identified in Australia, but whose status was uncertain. As well, there are two pathogenic spore forming microsporidians, *Nosema apis and N. ceranae*, which have been identified in Australia (2), and for which there is an industry need to better understand risk-factors which might lead to increased losses in hives. Thus in undertaking the survey there were multiple objectives. In addition, the survey had resource restrictions, and needed to be undertaken by a single scientist, albeit working in close co-operation with local bee-keepers.

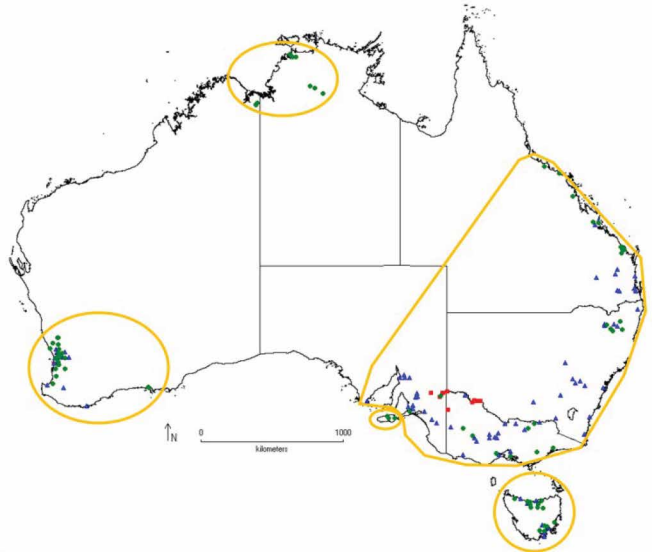
Amongst these multiple objectives, the most important was establishing the proof of freedom of DWV and SBPV – as this would support the industry's export of queens and packaged bees – and thus the survey was designed along established FFD survey protocols.

Based on expert opinion, five geographical regions ("metapopulations") were defined, between each there is restricted movement, but within each it was assumed there is relatively free mixing, as bee-keepers transport their hives to favourable foraging sites (Figure 1). To estimate the number of beekeepers to sample per region, hives per apiary and bees per hives, we used standard two-stage FFD probability formula calculations. As none of the tests could be assumed to be have perfect sensitivity (Se) or specificity (Sp), the calculations followed those of Cameron and Baldock (3), i.e. a two-stage design allowing for imperfect test(s), with the first stage being apiaries and the second being hives.

For the FFD design modelling we used the R package, "FFD" (4), and this was done separately in an iterative manner for each geographic region where we aimed to find the optimum

compromise between cost and the survey sensitivity. For most of the regions, apiaries were individually visited, but for the east coast region, advantage was taken of the congregation of apiaries during almond pollination (Figure 1).

Figure 1. The five "meta-population" regions on which the FFD survey was designed, showing the location of apiaries sampled (green circles), multi-apiary sites sampled during almond pollination (red squares) and their originating locations (blue triangles).



Molecular testing using RT-PCR and NGS

The test system for the survey was fully molecular using (1) reverse transcriptase (RT-PCR); and (2) NGS using the Illumina HiSeq 2500 platform. The RT-PCR was undertaken on pooled hive samples and the NGS was done on regional pools. Positive RT-PCR products were confirmed by Sanger sequencing. NGS data was quality trimmed and mapped to reference virus genomes using the CLC Genomics Workbench. Consensus sequences were manually inspected for genome coverage and similarity to mapped reference genomes in *Genbank* using the *BLASTn* program (https://blast.ncbi.nlm.nih.gov/).

Freedom-from-disease estimations

For the Freedom from disease (FFD) analysis, we calculated the negative predictive value (NPV) for the detection of the viruses for each region for a number of scenarios, depending on a range of possible within and between apiary values. Furthermore, we calculated the NPV for three different "prior belief of freedom" scenarios, ranging from 0.5 (no prior belief) to 0.99 (a very high prior belief). All FFD calculations were done using the "Analysis of simple 2-stage freedom survey" web form which is part of the "*EpiTools*" suite of epidemiological calculators (http://epitools.ausvet.com.au/).

Results

The field part of the survey ran from August 2013 to April 2015, during samples were taken from 1,240 hives belonging to 155 apiaries (Figure 1).

Five honey bee viruses were detected in Australia during the survey, but not DWV or SBPV. Freedom from disease (FFD) analysis showed good statistical support for this result, with >0.99 probability FFD for many of the scenarios tested for each sampling region. BQCV was the most common virus detected followed by Lake Sinai virus 1 (LSV1), SBV, Israeli acute paralysis virus (IAPV) and Lake Sinai virus 2 (LSV2). BQCV and SBV are long established honey bee viruses and were common in most regions. IAPV and the LSVs are more recently identified from overseas studies and are new records for most regions. Only 17% of samples were virus-free and 56% of samples were infected by more than one virus.

NGS confirmed the RT-PCR virus results with DWV and SBPV not detected in any sample, although a small number of short virus sequences was found that were 77% – 85% similar to DWV and 74% – 90% similar to SBPV. This level of identity is generally too small to be considered the same virus, but could represent variant strains (Table 1).

Table 1. Total sequence reads and maximum identity to reference genomes for honey bee viruses detected by NGS. Viruses above the line were detected by RT-PCR, and those beneath it were not detected.

Virus	Sequence reads	Max. identity
BQCV	169,818,518	93%
LSV1	5,534,196	94%
SBV	39,059,201	93%
IAPV	47,606,462	95%
LSV2	3,769,689	94%
DWV	551	85%
SBPV	74	90%

Discussion

Proving FFD is one of the most challenging areas of animal health surveillance on account of the general difficulty of "proving a negative", and in one of the earliest guides to statistical animal health surveillance (5), it was even stated that "To prove that a disease is not present in a population, particularly if the disease could be present at a low level, requires the testing of all susceptible individuals in the population". This is clearly impossible for most animal populations, and to overcome this initial attempts were to apply a rigorous sampling methodology based on the design prevalence concept and random sampling of individuals. Later refinements were (1) "two-stage sampling" wherein herds/flocks are the first stage and then individuals (3); and (2) adjusting for imperfect sensitivity and specificity of the FFD test (6).

Nevertheless, FFD surveys using statistically designed surveys are unsettling in their lack of certainty. This is illustrated with a well-designed and executed survey undertaken in NZ to prove FFD from IAPV (7). This used both risk-based and a formal randomised survey in the North and South Island respectively. IAPV was not detected in either survey, with the overall

probability of absence determined to be 92% for New Zealand as a whole. Anyone familiar with interpreting FFD surveys would consider this to be a robust result, especially given the rigor of the sampling as described in the paper. However, there is potentially a communication challenge for biosecurity managers in that, despite a reasonably well resourced survey, there remains an 8% chance that the virus is actually present in the country and was not detected.

Given this problem, there has been a move to take a "total evidence" approach to interpreting FFD surveys. This in fact is how the OIE currently assesses evidence of certifying a country free from certain "officially recognised" diseases, including foot-and-mouth disease. This requires previously infected countries present not only the results of a formal survey, but also additional data on the surveillance system, the strength of veterinary services etc. The total "dossier" is then assessed by a panel of experts with the experience to question any shortcomings in the assertion of freedom.

As in our example, the use of NGS has considerable potential to deliver corroborative evidence, as via the read-count data, it provides additional support to supplement the survey test. In this case, it can clearly be seen that those viruses that were unambiguously positive to the RT-PCR test also gave very high read counts (Table 1). By contrast, for those viruses where the RT-PCR was negative (DWV and SBPV), some counts were detected, but these were low, and were not closely related to reference genomes.

To date we are unaware of another FFD survey which has used NGS data in this way, and thus the methodology is still exploratory. However, already its potential can be seen, particularly as a possible technique to establish relatively low-cost, ongoing monitoring following a FFD survey. For example, NGS might be combined with syndromic surveillance, whereby beekeepers are invited to submit bees for NGS testing when mortality events occur. In such instance, NGS might be used to rapidly detect the possible causative virus (or other pathogens) but indirectly also provide ongoing evidence of FFD of other diseases.

Nevertheless, considerable thought will be needed to translate such concepts into robust NGS-based surveillance systems, and as a warning we provide the following example. During 2006-07 in northeastern USA, a syndrome of very high colony losses occurred, which was termed "colony collapse disorder" ("CCD"). A university-led team undertook a metagenomics approach, whereby they sampled bees from unaffected and affected colonies and undertook NGS (8). This found a higher proportion of reads to IAPV in affected colonies, and this led them to conclude that this might be the agent responsible. Furthermore, the sequence data associated the isolates to those recorded from Australia, which indicated that CCD might have been caused by the importation of bees. However, subsequent studies established that this apparent association was spurious, and that CCD is a complex multi-causal syndrome, of which infection with IAPV is only a minor component (9).

From this example, some may conclude that metagenomics approaches to surveillance are fraught with dangers; however, a more considered learning is that NGS surveillance needs to be seen as a complementary tool, and not one to replace well established epidemiological and surveillance practices like careful case definitions, sample size calculations, random sampling etc. Our experience with using NGS for a FFD survey supports this later approach.

Finally we note, as a matter of semantic correctness, that the divide between the two disruptive technologies (i.e. NGS and Big Data) is not a strict one, as the former, in producing large volumes of data can be considered an example of the latter. However, Big Data is increasingly accepted to be defined not solely about the volume of data, but also by other attributes, with adding "value" a key objective (10). We propose that this adding value attribute is where NGS has most potential.

References

1. **Hornitzky MAZ.** Prevalence of Virus-Infections of Honeybees in Eastern Australia. J*ournal of Apicultural Research* 26(3), 181-185, 1987
2. **Giersch T *et al.*** Nosema ceranae infects honey bees (Apis mellifera) and contaminates honey in Australia. *Apidologie* 40(2), 117-123, 2009
3. **Cameron AR, Baldock FC.** Two-stage sampling in surveys to substantiate freedom from disease. *Preventive Veterinary Medicine* 34, 19-30, 1998
4. **Kopacka I, Hofrichter J, Fuchs K.** Exact alpha-error determination for two-stage sampling strategies to substantiate freedom from disease. *Prev Vet Med* 109 (3-4), 205-12, 2013
5. **Cannon RM, Roe RT.** *Livestock disease surveys: a field manual for veterinarians*. Canberra: AGPS, p35, 1982
6. **Cannon RM.** Sense and sensitivity - designing surveys based on an imperfect test. *Preventive-Veterinary-Medicine* 49(3-4), 141-163, 2001
7. **McFadden AM *et al*.** Israeli acute paralysis virus not detected in Apis mellifera in New Zealand in a national survey. *Journal of Apicultural Research* 53(5), 520-527, 2014
8. **Cox-Foster DL *et al.*** A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* 318(5848), 283-287, 2007
9. **van Engelsdorp D *et al.*** A survey of honey bee colony losses in the U.S., fall 2007 to spring 2008. *PLoS One* 3(12), e4071, 2008
10. **Gandomi A, Haider M.** Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 35(2), 137-144, 2015