CHAPTER 1.1.7.

STANDARDS FOR HIGH THROUGHPUT SEQUENCING, BIOINFORMATICS AND COMPUTATIONAL GENOMICS

INTRODUCTION

High throughput sequencing, bioinformatics and computational genomics (*HTS-BCG*) *in animal health and food safety investigations should be used in accordance with standards for laboratory testing, just as any other laboratory tool or procedure. As HTS-BCG is a relatively new procedure, the purpose of this chapter is to assist laboratories by defining standards that will allow inclusion of the capability into a laboratory's scope of operations in a way in which the users of the results can have confidence.*

A. GENERAL CONSIDERATIONS

Sequence information is playing an increasingly important role in the diagnosis and management of microbial infections, including in the characterisation of infectious agents, their possible phenotypic characteristics and their epidemiology. Consequently it is incumbent on laboratories to adopt policies and practices for generating, analysing and managing genomic sequence data that are based on accurate information and rigorously interpreted.

Increasingly, for full identification and characterisation of a microorganism, there is an expectation that essential features of its genome should be described. For viruses, this may be the whole genome, while for bacteria and parasites, it may be only partial sequences. However as sequencing technology is developing so rapidly, within a short time whole genome sequences for these larger microorganisms may also be routinely generated after suitable bioinformatics procedures have been developed.

The standards described here apply to the generation of genomic sequence data during investigations of infections of single animals, animal populations and their environment. They apply to the generation, management and use of such data within the accepted practices of veterinary investigations and within a laboratory's quality assurance system.

B. THE CONDUCT OF VETERINARY INVESTIGATIONS INCORPORATING HTS-BCG

Sequence data of microorganisms, such as is generated by HTS or metagenomics approaches, is only a tool, although a powerful one, to assist in the investigation of issues regarding animal health and food safety. Appropriate experts should perform the analysis of sequence data. The interpretation of that data in relation to the disease investigation should be led by suitably qualified veterinarians, consistent with the standard requirements for diagnosis of animal disease.

The sequence and sequence analysis of infections associated with cases, outbreaks and investigations of animal disease and food safety by laboratories should be recorded and analysed together with all other information relating to the reporting and recording of such cases and outbreaks. These data should be considered a necessary part of such reports and records.

HTS-BCG can be deployed for a range of purposes in the detection of infectious agents and their characterisation, either in biological material such as diagnostic or surveillance specimens or propagated in cultures or as isolates. For primary diagnostic applications, the users of the technology should consider the

purposes of their testing in relation to the normal purposes of testing as defined in Chapter 1.1.6 *Principles and methods of validation of diagnostic assays for infectious diseases*. HTS-BCG may also be applied as a confirmatory assay for organisms detected in some other primary assay, or to provide additional characterisation of such organisms.

Further to these general purposes of testing, HTS-BCG offers specific opportunities for:

i) The detection, identification and characterisation of previously unidentified microorganisms;

ii) The improved diagnosis of known diseases;

iii) The improved diagnosis of emerging or re-emerging diseases with known or unknown aetiology;

iv) The development of single 'universal' diagnostic assays, able to identify any potential pathogen;

v) The simultaneous and quick detection of multiple agents in diseases with multifactorial aetiologies;

vi) The increased capability to study the evolutionary dynamics of pathogens at the farm, local, national and global levels;

vii) The deeper understanding of the epidemiology of infectious diseases and the phylogeography of infectious agents;

viii) The enhanced traceability of infectious diseases and modes of pathogen transmission including applications in forensic epidemiology;

ix) More extensive characterisation of 'populations' of known pathogens (e.g. relevant minority strains, escape mutants) that in turn facilitates the design of better vaccines, antivirals, etc.;

x) Better links between pathogen genotype and phenotypes enabled through full genome sequence of multiple strains (including reference strains) of a single agent.

C. STANDARDS FOR THE USE OF HTS-BCG

1. Selection of a technology platform or service

Laboratories may choose to establish a HTS-BCG capability in-house, contract commercial suppliers of services or submit specimens to designated Reference Centres.

Where the laboratory establishes its own capability there are a number of commercially available sequencing platforms for the purpose of generating sequence information from test samples. The choice of platform should be based on a consideration of the intended purpose or combination of purposes as outlined in Section B above.

Of primary concern is that the technology selected is fit for the intended purpose, that it is appropriate for the production of sequence information from the types of genome intended for study. Other considerations may take into account the time required to conduct a sequencing run, including sample preparation; ancillary equipment needed in addition to the actual sequencing device; the capital cost of the purchase and set up of all necessary equipment and the cost of annual licences or service agreements, including manufacturer's recommended maintenance schedule; the availability of supporting expertise from the supplier; the cost of reagents for a sequencing run and the likely availability of reagents in the country concerned; the staff requirements and training required to operate the equipment and to conduct the associated bioinformatic analyses and the data management requirements. Currently available systems have been reviewed (Belák *et al.,* 2013; Granberg *et al*., 2016; Marston *et al*., 2013), but new models and technologies can be expected to become available frequently.

Where a laboratory or veterinary service contracts an external provider to supply HTS-BCG services, they should ensure that the service provider meets the standards defined in this chapter.

2. Sampling and reporting

HTS-BCG is a new technological tool in the management of diseases of animals and its use should be adopted within the context of tried and accepted processes for the management of animal health and food safety including clinical or epidemiological field investigations and the sampling of animals, animal populations or other

epidemiologically relevant situations. The use of the technology should be appropriate to the purpose of the investigation, and the sampling strategy and the specimens taken should be appropriate for that investigation, based on an understanding of the pathogenesis and epidemiology of the infection under study or the likely pathogenesis and epidemiology of any novel infectious agent suspected. Such investigations should be under the supervision of appropriately qualified veterinarians.

In laboratories where HTS-BCG is used it should be managed within the context of the laboratory's quality assurance system. Hence the results of HTS-BCG must be interpreted in the context of the pathogenesis and epidemiology of the infection in the animal species under study. Results should be reported by appropriately qualified veterinary investigators with the authority to make diagnoses of animal diseases under the laboratory's quality assurance system and in the jurisdiction where the investigation is conducted.

3. Specimens and sample preparation

Specimens should be collected and submitted to the testing laboratory in accordance with the standards communicated in Chapter 1.1.2 *Collection, submission and storage of diagnostic specimens.* The normal comprehensive information regarding the individual animal, the case or reason for sampling and the relevant epidemiological information should be recorded in the laboratory's accessions processes, as for any submission to the laboratory.

As with other laboratory processes ensuring the integrity of the specimen and the samples to be tested is critical. Nucleic acids, either DNA or RNA, need to be extracted from the samples. In some cases, enrichment strategies to increase the ratio of pathogen to host nucleic acids can be used to maximise the sensitivity of the technique, however care must be taken to avoid biasing the outcome in the context of the intended purpose. Precautions to ensure the integrity and quality of nucleic acids must be followed similarly to any other molecular technique (e.g. polymerase chain reaction [PCR]) as already described in Guideline 3.2 *Biotechnology in the diagnosis of infectious diseases*[1]. Once nucleic acids are extracted from the samples, they need to be further manipulated (e.g. reverse transcription of RNA into complementary DNA) in order to be used in HTS. Different technological platforms require specific sets of reagents in order to generate the final material ("libraries") ready for sequencing. Commercial kits are available for this purpose.

HTS is an extremely sensitive technology and even few molecules of nucleic acid could be detected. Hence, precautions to avoid cross-contaminations must be followed as in the case of many other molecular techniques used to detect nucleic acids (e.g. PCR). Separation of work areas from the possibility of cross contamination with nucleic acid from other molecular investigations is an essential requirement. In addition, HTS very frequently involves "multiplexing" of several samples in a single reaction. Individual samples are "tagged" during one of the stages of sample preparation by the use of short index sequences linked to nucleic acid molecules. Index sequences must be of sufficient quality and design to be relied on as a signature for the tagged library for HTS use in order to avoid artefacts during bioinformatics analysis of sequencing data obtained.

Every application of HTS-BCG technology should include positive and negative controls appropriate to the investigation and that have been incorporated through the sample preparation processes of the sequencing run as well as the actual run on the technology platform. Appropriate controls should be used to verify each step of the procedure including nucleic acid quality, library preparation, cross-contamination (including multiplexing) sensitivity and reproducibility.

As with any other diagnostic method, confirmation of results would require resampling of the original specimen, which therefore has to be protected from cross-contamination and be stored appropriately.

4. Generation of sequence data

While HTS platforms differ widely in their details, basic principles of quality control relevant to the technology can be followed, and generic recommendations for acceptable quality metrics can be made. Suitable control measures might include the use of positive, negative and no-template controls run in replicates of the test and a quality scoring system. Sequencing quality metrics provide suitable parameters for the validation and monitoring of platform performance. Most platforms offer the possibility to spike controls in reagents and to use the control's QC metrics to monitor platform and reagent performance. Additional technology specific performance metrics can be used to monitor platform performance and to identify aberrant sequencing runs.

Quality metrics for the evaluation of the analytical performance of HTS-based tests, include:

1 Available at: http://www.oie.int/fileadmin/Home/eng/Health_standards/tahm/GUIDE_3.2_BIOTECH_DIAG_INF_DIS.pdf

i) Depth of coverage. This indicates the number of sequence reads providing information about a given nucleotide. When ongoing quality monitoring shows that the coverage depth at a given nucleotide is below the validated minimum coverage, confirmation should be provided using alternate methods (e.g. Sanger sequencing) or additional sequencing.

ii) Uniformity of coverage. This parameter describes how the depth of coverage is distributed over the test's target region(s). Deviations of uniformity of coverage from the validated range potentially indicate errors in the testing process.

iii) GC bias. The GC content (relative abundance of G and C nucleotides) of a target region affects the efficiency of sequencing reactions and will affect the uniformity of coverage. Where possible, the amount of GC bias in the test's target region(s) should be determined during validation and monitored to evaluate test performance.

iv) Base call quality scores. These are platform-derived reflections of the signal-to noise ratio and reflect the probability that the base call was correct. An acceptable raw base call quality threshold should be established during validation, and incorporated in bioinformatics filters to eliminate poor quality data during analysis.

v) Decline in signal intensity or read length. Depending on the exact application, HTS platform and chemistry, sequence reads have a typical distribution of read length and signal intensity. The expected signal intensity across reads (or read length distribution) should be established during validation and monitored for each run. Deviations in the distribution of read lengths may indicate problematic datasets.

vi) Mapping quality. This is a measure of uncertainty that a read is mapped properly to a genomic position within the target region. Acceptable values (e.g. proportion of reads mapping to the target) should be established during validation of bioinformatics workflows and the proportion of reads not mapping to the target can be monitored during each run.

vii) Internal controls. Most platforms offer the possibility to spike an internal control at very low frequency during the sequencing run. The quality metrics of those reads can to be compared to previously reported quality metrics.

5. Bioinformatics

An absolute requirement for any laboratory intending to establish a HTS-BCG capability is the employment of specialised bioinformatics skills. Even if platforms with supporting software for specific analyses in defined clinical situations were to become available the use of such packages would not remove the responsibility of the laboratory to be able to competently analyse its own data.

The bioinformatic analysis assembling the pathogen genomic sequence from the raw data and the subsequent secondary analysis are the critical elements in HTS-BCG. Hence the approaches used must be transparent and a declaration of the software packages, software versions, and reference databases or sequences used should be a component of every report of sequence analysis. Software programs used for these analyses must be readily available (commercially or open access) in order to be evaluated by the international community.

As with any laboratory procedure, attention must be given to quality assurance. The test method should include criteria for acceptance or rejection of each run based on the satisfactory analyses of the controls. Sequencing data must be documented to have satisfied minimum quality scores and coverage for each nucleotide of the assembled final consensus sequence obtained.

The appropriateness of chosen bioinformatics software for particular analyses can be evaluated through testing its performance against standard data sets containing data relating to agents expected to be present in the specimens to be tested.

6. Data management

The data generated from HTS-BCG operations are essential to reach the diagnosis or other scientific purpose of the investigation, such as agent characterisation, and are an integral component of the process. As such it is an essential requirement of laboratories to have policies, processes and supporting systems to curate, manage and store the data generated.

Different HTS technology platforms produce raw data in different formats and stage of pre-analysis, so it is necessary for laboratories to have policies and processes specific to the technology platform in use. Data

management systems will include aspects of which data to keep, and the length of time for which they will be kept, and the back-up strategies to protect against accidental loss or deliberate erasure. Metadata describing the generation and analysis of the sequence data is essential, so that the process itself can be analysed or repeated.

Where a sequence analysis leads to an output of animal health significance, especially one of trade or international significance, it is an absolute requirement that the data on which the analysis was performed be kept available for audit or confirmatory analysis for a period of time commensurate with the significance of the animal health finding. This is particularly important where the finding may be disputed. Failure to be able to produce the required data for independent analysis could be taken to invalidate the finding.

Sequence data should be stored in a manner in which there is a clear link to the metadata associated with the specimen that was the subject of the analysis. As is standard practice in laboratory investigations, such metadata includes information regarding the animal sampled, its ownership and location, and accompanying clinical and epidemiological information regarding the animal population.

7. Validation of test systems for designated purposes

The concepts of test validation as stated in chapter 1.1.6 are broadly applicable to HTS-BCG. All procedures including sample processing (nucleic acid extraction, library preparation, tagging, target enrichment), sequencing, bioinformatics and reporting should be documented in SOPs before validation can start. Stage 1 validation data must be developed to confirm the analytic sensitivity (Se) and specificity (Sp) of the technique, and its repeatability. For sequencing based tests, analytic sensitivity can be defined as the likelihood that the assay will detect the targeted sequence variations, if present, at a given probability (e.g. 95% confidence), while analytical specificity can be defined as the probability that an assay will not detect a sequence variation when none are present at a given probability. Furthermore, each type of specimens has its own characteristics that have to be considered, e.g. nasal swab, sera or faeces. Well described samples with known concentrations of target analyte or non-target analytes and matrix components can be used to assess the analytical performance. This should include, as a minimum, serial dilutions of each type of specimen containing defined organisms to document the limits of detection of designated whole genomes or genetic sequences representative of the type for which the HTS-BCG capability will be used in the laboratory. For viral disease investigations, test specimens could be prepared to contain representative viruses of the full range of viral families from which agents may be present in test specimens of the type to be investigated in routine operations. Documentation of the laboratory's HTS-BCG system to detect these viruses will be established. The same principles apply to genetic markers, bacteria or other organisms for which the HTS-BCG capability will be used in planned routine operations. In all these runs designed to establish sensitivity and specificity, the sample preparation steps should be part of each assessment as these steps are likely to be critical to all aspects of overall test performance.

Several factors complicate the validation of NGS tests as primary diagnostic assays including:

i) The weight of the analytical and diagnostic validation required (chapter 1.1.6);

ii) The operational cost of the technology;

iii) The challenges of validation of a data analysis workflow;

iv) The high need for investment in hardware and expertise;

v) The time taken to obtain a result (currently days compared to hours for specific molecular diagnostics such as real-time PCR).

Confirmatory adjunct or secondary diagnostic assays need to be validated only for their analytical performance, e.g. analytical sensitivity and specificity, and repeatability and initial reproducibility (stage 1) and not to the full diagnostic extent (diagnostic sensitivity and specificity, stage 2).

It is recognised that it may not always be practical to produce large data sets on test performance such as would normally allow calculation of test diagnostic sensitivity and specificity, but other aspects of validation such as demonstration of test reproducibility among laboratories conducting similar investigations should be undertaken.

8. Quality assurance

Testing using HTS-BCG for the purposes of investigations of animal health and food safety should be conducted in accordance with the requirements of the laboratory's quality assurance system, the features of which will meet the standards listed in Chapter 1.1.5 *Quality management in veterinary testing laboratories.* Where the laboratory is accredited, the testing should be part of the laboratory's scope of accreditation.

Standard data sets against which the usefulness of bioinformatics software packages can be verified have been developed. Laboratories using HTS-BCG should ensure that their software packages for bioinformatics meet expected performance criteria against data standards.

Where proficiency testing strategies have been developed, laboratories using HTS-BCG should participate.

9. Interpretation of results

HTS-BCG can be used for a variety of purposes ranging from pathogen discovery to diagnosis or in-depth characterisation of known infectious agents. Consequently, the interpretation of the results obtained will be in the context of the specific clinical and epidemiological situation, reassured by satisfactory performance against all specified controls and quality assurance parameters. As with any other laboratory tests, these considerations are one among a number of parameters to be taken into account.

REFERENCES

BELÁK S., KARLSSON O.E., LEIJON M. & GRANBERG F. (2013). High-throughput sequencing in veterinary infection biology and diagnostics. *Rev. sci. tech. Off. int. Epiz.*, **32** (3), 893–915.

GRANBERG F., BÁLINT, A. & BELÁK, S. (2016). Novel technologies applied for the nucleotide sequencing and comparative sequence analysis of the genomes of infectious agents in veterinary medicine. OIE *Sci. Tech. Rev.,* **35** (1), (in press).

MARSTON D.A., MCELHINNEY L.M., ELLIS R.J., HORTON D.L., WISE E.L., LEECH S.L., DAVID D., DE LAMBALLERIE X. & FOOKS A.R. (2013). Next generation sequencing of viral RNA genomes. *BMC Genomics*, **14**, 444.

VAN BORM S., WANG J., GRANBERG F. & COLLING A. (2016). Towards validation and quality assurance of next-generation sequencing workflows in veterinary infection biology. OIE *Sci. Tech. Rev.,* **35** (1), April 2016 (in press).

*
* *

NB: There is an OIE Collaborating Centre for Viral Genomics and Bioinformatics
(see Table in Part 4 of this *Terrestrial Manual* or consult the OIE Web site for the most up-to-date list:
http://www.oie.int/en/our-scientific-expertise/collaborating-centres/list-of-centres/).
Please contact the OIE Collaborating Centre for any further information HTS-BCG.